# Intelligence artificielle et sante prédictive : l'exemple de l'Immunoscore® dans le cancer du côlon

Clément Chollat-Namy

# THESE

PRESENTEE ET PUBLIQUEMENT SOUTENUE DEVANT LA FACULTE DE PHARMACIE DE MARSEILLE

LE JEUDI 10 JUIN 2021

PAR

M. Clément CHOLLAT-NAMY

Né le 20 décembre 1992 à Montpellier, France

EN VUE D'OBTENIR

LE DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

**TITRE :**

# INTELLIGENCE ARTIFICIELLE ET SANTE PREDICTIVE : L'EXEMPLE DE L'IMMUNOSCORE® DANS LE CANCER DU COLON

**JURY :**

Président : Monsieur le Professeur D. Berge-Lefranc

Membres : Monsieur le Docteur P. Prinderre
Monsieur le Docteur F. Devred
Monsieur le Docteur P. Chatel

Faculté de Pharmacie
Aix Marseille Université

27 Boulevard Jean Moulin – 13385 MARSEILLE Cedex 05

Tel. : 04 91 83 55 00 – Fax : 04 91 80 26 12

**ADMINISTRATION :**

| | |
|---|---|
| *Doyen :* | Mme Françoise DIGNAT-GEORGE |
| *Vice-Doyens :* | M. Jean-Paul BORG, M. François DEVRED, M. Pascal RATHELOT |
| *Chargés de Mission :* | Mme Pascale BARBIER, M. David BERGE-LEFRANC, Mme Manon CARRE, Mme Caroline DUCROS, Mme Frédérique GRIMALDI, M. Guillaume HACHE |
| *Conseiller du Doyen :* | M. Patrice VANELLE |
| *Doyens honoraires :* | M. Patrice VANELLE, M. Pierre TIMON-DAVID, |
| *Professeurs émérites :* | M. José SAMPOL, M. Athanassios ILIADIS, M. Henri PORTUGAL, M. Philippe CHARPIOT |
| *Professeurs honoraires :* | M. Guy BALANSARD, M. Yves BARRA, Mme Claudette BRIAND, M. Jacques CATALIN, Mme Andrée CREMIEUX, M. Aimé CREVAT, M. Gérard DUMENIL, M. Alain DURAND, Mme Danielle GARÇON, M. Maurice JALFRE, M. Joseph JOACHIM, M. Maurice LANZA, M. José MALDONADO, M. Patrick REGLI, M. Jean-Claude SARI |
| *Chef des Services Administratifs :* | Mme Florence GAUREL |
| *Chef de Cabinet :* | Mme Aurélie BELENGUER |
| *Responsable de la Scolarité :* | Mme Nathalie BESNARD |

## DEPARTEMENT BIO-INGENIERIE PHARMACEUTIQUE

Responsable : Professeur Philippe PICCERELLE

### PROFESSEURS

| | |
|---|---|
| BIOPHYSIQUE | M. Vincent PEYROT<br>M. Hervé KOVACIC |
| GENIE GENETIQUE ET BIOINGENIERIE | M. Christophe DUBOIS |
| PHARMACIE GALENIQUE, PHARMACOTECHNIE INDUSTRIELLE, BIOPHARMACIE ET COSMETOLOGIE | M. Philippe PICCERELLE |

## MAITRES DE CONFERENCES

| | |
|---|---|
| BIOPHYSIQUE | M. Robert GILLI<br>Mme Odile RIMET-GASPARINI<br>Mme Pascale BARBIER<br>M. François DEVRED<br>Mme Manon CARRE<br>M. Gilles BREUZARD<br>Mme Alessandra PAGANO |
| GENIE GENETIQUE ET BIOTECHNOLOGIE | M. Eric SEREE-PACHA<br>Mme Véronique REY-BOURGAREL |
| PHARMACIE GALENIQUE, PHARMACOTECHNIE INDUSTRIELLE, BIOPHARMACIE ET COSMETOLOGIE | M. Pascal PRINDERRE<br>M. Emmanuel CAUTURE<br>Mme Véronique ANDRIEU<br>Mme Marie-Pierre SAVELLI |
| BIO-INGENIERIE PHARMACEUTIQUE ET BIOTHERAPIES<br>PHARMACO ECONOMIE, E-SANTE | M. Jérémy MAGALON<br>Mme Carole SIANI |

## ENSEIGNANTS CONTRACTUELS

| | |
|---|---|
| ANGLAIS | Mme Angélique GOODWIN |

## DEPARTEMENT BIOLOGIE PHARMACEUTIQUE

Responsable : Professeur Françoise DIGNAT-GEORGE

## PROFESSEURS

| | |
|---|---|
| BIOLOGIE CELLULAIRE | M. Jean-Paul BORG |
| HEMATOLOGIE ET IMMUNOLOGIE | Mme Françoise DIGNAT-GEORGE<br>Mme Laurence CAMOIN-JAU<br>Mme Florence SABATIER-MALATERRE<br>Mme Nathalie BARDIN |
| MICROBIOLOGIE | M. Jean-Marc ROLAIN<br>M. Philippe COLSON |
| PARASITOLOGIE ET MYCOLOGIE MEDICALE, HYGIENE ET ZOOLOGIE | Mme Nadine AZAS-KREDER |

| MAITRES DE CONFERENCES | |
|---|---|
| BIOCHIMIE FONDAMENTALE, MOLECULAIRE ET CLINIQUE | M. Thierry AUGIER<br>M. Edouard LAMY<br>Mme Alexandrine BERTAUD<br>Mme Claire CERINI<br>Mme Edwige TELLIER<br>M. Stéphane POITEVIN |
| HEMATOLOGIE ET IMMUNOLOGIE | Mme Aurélie LEROYER<br>M. Romaric LACROIX<br>Mme Sylvie COINTE |
| MICROBIOLOGIE | Mme Michèle LAGET<br>Mme Anne DAVIN-REGLI<br>Mme Véronique ROUX<br>M. Fadi BITTAR<br>Mme Isabelle PAGNIER<br>Mme Sophie EDOUARD<br>M. Seydina Mouhamadou DIENE |
| PARASITOLOGIE ET MYCOLOGIE MEDICALE, HYGIENE ET ZOOLOGIE | Mme Carole DI GIORGIO<br>M. Aurélien DUMETRE<br>Mme Magali CASANOVA<br>Mme Anita COHEN |
| BIOLOGIE CELLULAIRE | Mme Anne-Catherine LOUHMEAU |

| ATER | |
|---|---|
| BIOCHIMIE FONDAMENTALE, MOLECULAIRE ET CLINIQUE | Mme Anne-Claire DUCHEZ |
| BIOLOGIE CELLULAIRE ET MOLECULAIRE | Mme Alexandra WALTON |

| A.H.U. | |
|---|---|
| HEMATOLOGIE ET IMMUNOLOGIE | Mme Mélanie VELIER |

## DEPARTEMENT CHIMIE PHARMACEUTIQUE

Responsable : Professeur Patrice VANELLE

| PROFESSEURS | |
|---|---|
| CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION | Mme Catherine BADENS |
| CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES | M. David BERGE-LEFRANC |
| CHIMIE MINERALE ET STRUCTURALE – CHIMIE THERAPEUTIQUE | M. Pascal RATHELOT<br>M. Maxime CROZET |
| CHIMIE ORGANIQUE PHARMACEUTIQUE | M. Patrice VANELLE<br>M. Thierry TERME |
| PHARMACOGNOSIE, ETHNOPHARMACOGNOSIE | Mme Evelyne OLLIVIER |

## MAITRES DE CONFERENCES

| | |
|---|---|
| BOTANIQUE ET CRYPTOGAMIE, BIOLOGIE CELLULAIRE | Mme Anne FAVEL<br>Mme Joëlle MOULIN-TRAFFORT |
| CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION | Mme Catherine DEFOORT<br>M. Alain NICOLAY<br>Mme Estelle WOLFF<br>Mme Elise LOMBARD<br>Mme Camille DESGROUAS<br>M. Charles DESMARCHELIER |
| CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES | M. Pierre REBOUILLON |
| CHIMIE THERAPEUTIQUE | Mme Sandrine ALIBERT<br>Mme Caroline DUCROS<br>M. Marc MONTANA<br>Mme Manon ROCHE<br>Mme Fanny MATHIAS |
| CHIMIE ORGANIQUE PHARMACEUTIQUE<br>HYDROLOGIE | M. Armand GELLIS<br>M. Christophe CURTI<br>Mme Julie BROGGI<br>M. Nicolas PRIMAS<br>M. Cédric SPITZ<br>M. Sébastien REDON |
| PHARMACOGNOSIE, ETHNOPHARMACOLOGIE | M. Riad ELIAS<br>Mme Valérie MAHIOU-LEDDET<br>Mme Sok Siya BUN<br>Mme Béatrice BAGHDIKIAN |

## MAITRES DE CONFERENCE ASSOCIES A TEMPS PARTIEL (M.A.S.T.)

| | |
|---|---|
| CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION | Mme Anne-Marie PENET-LOREC |
| CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES | M. Cyril PUJOL |
| DROIT ET ECONOMIE DE LA PHARMACIE | M. Marc LAMBERT |
| GESTION PHARMACEUTIQUE, PHARMACOECONOMIE ET ETHIQUE PHARMACEUTIQUE OFFICINALE, DROIT ET COMMUNICATION PHARMACEUTIQUES A L'OFFICINE ET GESTION DE LA PHARMAFAC | Mme Félicia FERRERA |

## A.H.U.

| | |
|---|---|
| CHIMIE ANALYTIQUE, QUALITOLOGIE ET NUTRITION | M. Mathieu CERINO |

## ATER

| | |
|---|---|
| CHIMIE PHYSIQUE – PREVENTION DES RISQUES ET NUISANCES TECHNOLOGIQUES | M. Duje BURIC |

**DEPARTEMENT MEDICAMENT ET SECURITE SANITAIRE**
Responsable : Professeur Benjamin GUILLET

## PROFESSEURS

| | |
|---|---|
| PHARMACIE CLINIQUE | M. Stéphane HONORÉ |
| PHARMACODYNAMIE | M. Benjamin GUILLET |
| TOXICOLOGIE ET PHARMACOCINETIQUE | M. Bruno LACARELLE<br>Mme Frédérique GRIMALDI<br>M. Joseph CICCOLINI |

## MAITRES DE CONFERENCES

| | |
|---|---|
| PHARMACODYNAMIE | M. Guillaume HACHE<br>Mme Ahlem BOUHLEL<br>M. Philippe GARRIGUE |
| PHYSIOLOGIE | Mme Sylviane LORTET<br>Mme Emmanuelle MANOS-SAMPOL |
| TOXICOLOGIE ET PHARMACOCINETIQUE | Mme Raphaëlle FANCIULLINO<br>Mme Florence GATTACECCA |
| TOXICOLOGIE GENERALE ET PHARMACIE CLINIQUE | M. Pierre-Henri VILLARD<br>Mme Caroline SOLAS-CHESNEAU<br>Mme Marie-Anne ESTEVE |

## A.H.U.

| | |
|---|---|
| PHYSIOLOGIE / PHARMACOLOGIE | Mme Anaïs MOYON |
| PHARMACIE CLINIQUE | M. Florian CORREARD |

## ATER.

| | |
|---|---|
| TOXICOLOGIE ET PHARMACOCINETIQUE | Mme Anne RODALLEC |

## CHARGES D'ENSEIGNEMENT A LA FACULTE

Mme Valérie AMIRAT-COMBRALIER, Pharmacien-Praticien hospitalier

M. Pierre BERTAULT-PERES, Pharmacien-Praticien hospitalier

Mme Marie-Hélène BERTOCCHIO, Pharmacien-Praticien hospitalier

Mme Martine BUES-CHARBIT, Pharmacien-Praticien hospitalier

M. Nicolas COSTE, Pharmacien-Praticien hospitalier

Mme Sophie GENSOLLEN, Pharmacien-Praticien hospitalier

M. Sylvain GONNET, Pharmacien titulaire

Mme Florence LEANDRO Pharmacien, Pharmacien adjoint

M. Stéphane PICHON, Pharmacien titulaire

M. Patrick REGGIO, Pharmacien conseil, DRSM de l'Assurance Maladie

Mme Clémence TABELE, Pharmacien-Praticien attaché

Mme TONNEAU-PFUG, Pharmacien adjoint

M. Badr Eddine TEHHANI, Pharmacien – Praticien hospitalier

M. Joël VELLOZZI, Expert-Comptable

Mise à jour le 23 janvier 2020

# REMERCIEMENTS

Je remercie le Docteur Prinderre pour l'attention qu'il a porté à ma thèse, ses conseils sur le sujet et sa disponibilité.

Je remercie le Professeur Berge Lefranc et le Docteur Devred pour l'intérêt qu'ils ont manifesté pour ma thèse.

Je remercie le Docteur Philippe Chatel d'avoir eu la gentillesse de bien vouloir être membre de mon jury.

Merci à l'équipe d'HalioDx : le Docteur Alexia Papadopoulos ainsi qu'Assil Benchaaben et Felipe Guimarães pour le temps qu'ils m'ont consacré et leurs explications claires.

Je remercie également le Docteur Vincent Provitolo pour son assistance.

Merci à Maël Steunou pour le temps qu'il m'a accordé.

Je remercie aussi mon père le Docteur Alexandre Chollat-Namy pour son expertise sur le sujet.

Merci à ma mère, mes grands-parents, mes oncles et tantes, ma famille, mes beaux-parents, mes beaux-frères et ma belle-sœur, mes amis pour leur accompagnement tout au long de la rédaction.

Victorine, merci pour ton affection et ton soutien sans faille, tu vois j'y suis arrivé !

***« L'Université n'entend donner aucune approbation, ni improbation aux opinions émises dans les thèses. Ces opinions doivent être considérées comme propres à leurs auteurs. »***

# TABLE DES MATIERES

# INTRODUCTION

Le 27 mars 2019, le prix Turing, l'équivalent pour l'informatique du Prix Nobel, a été remis aux Docteurs Yann LeCun, Geoffrey Hinton et Yoshua Bengio, respectivement Docteur en sciences, Docteur en intelligence artificielle et Docteur en informatique, pour leurs travaux pionniers dans les années 1990 et 2000 sur les réseaux neuronaux virtuels, qui sont à l'origine de l'explosion actuelle du développement de l'intelligence artificielle.

Ces travaux ont permis des applications très variées, de la voiture autonome aux logiciels d'autocomplétion d'*emails*, en passant par la reconnaissance faciale et vocale ou encore les diagnostics médicaux.

On peut d'autant plus reconnaitre leur travail par la détermination dont ils ont fait preuve de la moitié des années 1990 au début des années 2000, quand l'intérêt de la communauté scientifique pour l'intelligence artificielle était très faible.

C'est en 2012 qu'ils ont démontré que leurs réseaux neuronaux pouvaient être très prometteurs, quand une équipe de chercheurs menée par Hinton s'est penchée sur ImageNet, un projet de classification logicielle d'images. En apportant leur expertise au projet, le taux de reconnaissance a grimpé de 40%, et le projet est passé d'une simple banque de données à la promesse d'une vraie reconnaissance visuelle digne d'un homme.

Dans le domaine de la santé, ces technologies peuvent répondre à un des enjeux majeurs actuels qui est la capacité à prévoir et anticiper l'apparition et le développement de pathologies. Les chercheurs les destinent en effet à des prédictions toujours plus précises autant au niveau mondial, comme la prévision d'une nouvelle pandémie, qu'au niveau individuel, comme la recherche d'une prédisposition à un cancer, et elles font naitre de nouvelles perspectives de diagnostics et de traitements.

Ces nouvelles perspectives ont de grandes répercussions sur les acteurs de la santé qui se développent et créent des nouvelles entreprises et des métiers spécialisés. La pharmacie est concernée, que ce soit à l'hôpital, parmi les industriels ou en ville.

Pour développer ces procédés il a fallu plusieurs étapes pour exploiter l'information : le développement des réseaux interconnectés puis Internet, le stockage et la collecte d'un grand nombre de données aussi appelé *Big Data*, l'utilisation de cette information via des logiciels d'intelligence artificielle.

Le domaine de la santé a ainsi vu émerger la notion de santé prédictive ou médecine prédictive. L'exemple du logiciel Immunoscore est représentatif de cette nouvelle notion.

# I. INTERNET : L'ACCES A L'INFORMATION

## I.1 Définition

D'après le Larousse (*online*), Internet désigne « le réseau informatique international, qui résulte de l'interconnexion des ordinateurs du monde entier utilisant un protocole commun d'échanges de données (baptisé TCP/IP ou Transport Control Protocol/Internet Protocol et spécifié par l'Internet Society, ou ISOC) afin de dialoguer entre eux via les lignes de télécommunication (lignes téléphoniques, liaisons numériques, câble). » (1)

## I.2 Historique

ARPANET est le réseau précurseur d'Internet. C'est un projet qui a nécessité une quinzaine d'années pour émerger. Contrairement aux idées reçues, ce n'est pas un réseau militaire que l'on a ensuite donné au public mais bien l'inverse.

### I.2.1 Prémices : des années 50 à 1969

La mise en place d'ARPANET a demandé la combinaison de plusieurs évènements.

#### *I.2.1.1 La naissance de l'ARPA*

Octobre 1957. En pleine guerre froide, l'URSS envoie pour la première fois un satellite dans l'espace : Spoutnik. L'avancée technologique communiste est dure à digérer pour les USA. Le 7 février 1958, le président Eisenhower crée l'ARPA (pour Advanced Research Project Agency) à l'intérieur du département d'Etat à la Défense, pour financer et développer les STIM : Science, Technologie, Ingénierie et Mathématiques, dans le but de protéger les Etats-Unis « de toute surprise technologique ».

L'ARPA se focalise d'abord sur les missiles et la conquête spatiale, bien avant le projet d'un réseau de télécommunications entre chercheurs et militaires.

Il y eu des étapes primordiales annonciatrices d'ARPANET :

#### *I.2.1.2 Leonard Kleinrock et la transmission par paquets*

En juillet 1961, Leonard Kleinrock, un étudiant du MIT (Massachusetts Institute of Technology) publie ses travaux sur l'utilisation de la commutation par paquets pour transférer des données, et non par circuits, défendue alors par les opérateurs télécoms. (2)

En effet, les opérateurs téléphoniques américains comme AT&T utilisaient un système de transmission de l'information par circuit, ou commutation de circuits, qui était simple à mettre en place et consistait à relier physiquement par un circuit l'émetteur et le receveur, d'abord de façon manuelle (grâce aux standardistes téléphoniques), puis de façon automatique.

Elle avait néanmoins de nets désavantages pour la transmission de données. La ligne devait être établie totalement entre l'émetteur et le receveur : elle était alors occupée jusqu'à la fin de la communication même s'il y avait beaucoup de « blancs » dans l'échange, la ligne était alors sous utilisée. De plus ce système était aussi totalement vulnérable si la ligne était rompue : aucun autre chemin n'était possible.

*Figure 1 : une standardiste de l'US Air Force utilisant un standard téléphonique en 1967*

La transmission par paquets, ou commutation de paquets, est plus souple. Elle permet d'envoyer l'information sous une forme décomposée en plusieurs petits blocs de données afin de les transférer plus rapidement et plus efficacement sur les réseaux.

Chaque paquet contient un en-tête et une charge utile. L'en-tête contient 2 adresses IP réseau : l'adresse IP d'origine, à partir de laquelle le fichier de données est envoyé, et l'adresse IP de destination, à laquelle le paquet de données doit être réceptionné. Les données de l'en-tête sont utilisées par le matériel des nœuds du réseau pour aiguiller le paquet vers sa destination où la charge utile est extraite et utilisée par le logiciel d'application.

L'en-tête contient également un nombre qui identifie le nombre de paquets que contient le fichier de données entier ainsi qu'un numéro de séquence afin de remettre les paquets dans l'ordre à la réception.[1]

[1] Le numéro de séquence est nécessaire dans le mode de commutation de paquets appelé « sans connexion », ou les blocs sont routés indépendamment sur le réseau (Comme la plupart du temps sur Internet). Il existe également un mode de commutation de paquets avec connexion, en mode « circuit virtuel » qui simule un circuit entre l'envoyeur et le destinataire, il n'y a donc pas besoin de numéro de séquence. Il est considéré comme plus fiable car il n'y a pas de risque de perte de paquets d'information.

Malgré une latence et des retards pouvant être imprévisibles, la commutation par paquets pose donc les fondations d'un réseau efficace :

- l'utilisation de sa bande passante est réduite (utilisée simplement quand le paquet y transite),
- il est moins couteux à déployer que le déploiement de lignes dans un système par circuits,
- il est plus fiable. (Un nœud peut remplacer un autre nœud défaillant, et en cas de perte d'un paquet l'ordinateur destinataire peut demander son renvoi.)

#### *I.2.1.3 Douglas Engelbart et l'hypertexte*

Sous l'égide du Standford Research Institute, Engelbart dans les années 60 déposa de nombreux brevets pour « augmenter » l'interaction homme-machine : il développe la souris, l'interface graphique sur ordinateur, et l'hypertexte : des documents reliés entre eux par des hyperliens. Ses brevets lui valent d'être subventionné par l'ARPA.

#### *I.2.1.4 Le time-sharing*

Le *time-sharing,* ou temps partagé, est un terme informatique désignant la capacité d'un ordinateur à pouvoir partager ses ressources par plusieurs utilisateurs (qu'il ne faut pas confondre avec le multitâche qui correspond à la capacité d'un ordinateur à exécuter plusieurs processus en même temps). Il est développé principalement au MIT par John McCarthy, un des principaux chercheurs en intelligence artificielle (avec Marvin Mee Minsky, voir ci-après) et Fernando Corbató, informaticien ayant reçu le prix Turing en 1990.

Ces nouveaux travaux s'opposent au fonctionnement des ordinateurs de l'époque, utilisant le traitement par lots ou *batch processing* qui entrainent des temps de calculs très longs.

Au-delà des aspects techniques, cette nouvelle façon de penser l'informatique, appelée « l'informatique interactive », est débattu dans les conférences de l'époque et permet de poser les notions d'ordinateur en réseau à l'opposé d'un simple ordinateur

de calculs, pouvant être utilisé pour la communication et le partage, et qui sera au cœur d'ARPANET. (3)

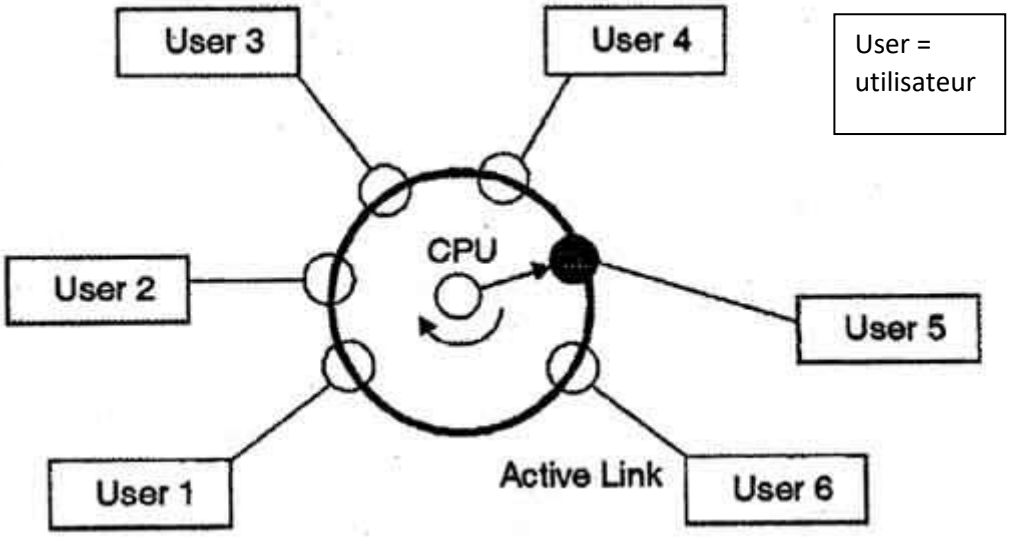



*Figure 2: Principe du time-sharing : le processeur de l'ordinateur alloue un temps imparti à chaque utilisateur, qui peuvent l'utiliser à tour de rôle. (4)*

### *I.2.1.5 Trois concepts similaires de réseaux décentralisés*

En octobre 1962, le docteur Joseph Carl Robnett Licklider, considéré aujourd'hui comme une des plus grandes figures dans l'informatique, est choisi à l'ARPA pour conduire les recherches sur une meilleure utilisation militaire de l'informatique. Il est à l'origine de l'idée d'un « réseau galactique » : un réseau d'ordinateurs connectés entre eux permettant à toute personne d'accéder rapidement à toute information ou programme où qu'il se trouve. (5)

Parallèlement la même année, avec la crise de Cuba, les autorités américaines ont besoin d'un réseau capable de résister à une attaque nucléaire. Un groupe de chercheurs du RAND (Research And Development, association non lucrative visant à développer les sciences et l'éducation aux USA) leur propose alors un système décentralisé, avec une information redondante, permettant au réseau de continuer à fonctionner même si une ou plusieurs machines est touchée(6).

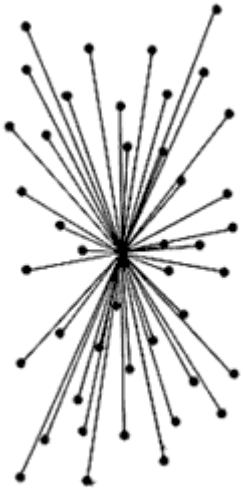


*Figure 3: Réseau centralisé*

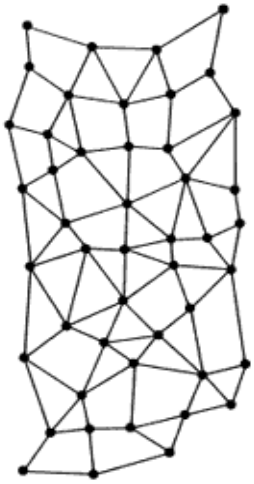


*Figure 4: Réseau distribué*

Cette idée de décentralisation du réseau a ainsi été élaborée par trois fois de façon indépendante, au MIT par Leonard Kleinrock ; par Paul Baran, un ingénieur du RAND ; et par Donald Davies un informaticien du National Physical Laboratory en Angleterre.

Ils imaginent un système où chaque machine est un nœud d'un réseau en toile d'araignée sur lequel l'information transite regroupée par paquets de données dynamiques. Quand le paquet atteint le nœud suivant, la machine le stocke puis détermine le chemin le plus court avant de le renvoyer. En cas de problème sur le trajet les paquets sont simplement détournés via un autre chemin (5) (7).

En 1967, Lawrence Roberts, un ingénieur informatique promu à la tête de l'ARPA, a été convaincu par les travaux de Kleinrock et publie ses « plans pour le réseau ARPANET » au cours d'une conférence. Lors de cette conférence les concepts similaires de réseau de Davies et Baran seront aussi publiés.

Nous pourrons noter que c'est à cause des similitudes entre les projets de transmission sécurisée même en cas de destruction partielle du réseau en cas d'attaque nucléaire de Paul Baran pour le RAND et le projet de Roberts de relier les machines des chercheurs via l'ARPA qu'est née la fausse légende selon laquelle ARPANET avait été lancé par les militaires pour créer un réseau insensible aux destructions d'une guerre nucléaire.

Ce sont néanmoins bien les projets plus universitaires de Roberts et de Davies, à qui l'on doit le terme de « commutation par paquets » qui interpellent les personnes chargées de la mise en œuvre des réseaux et permettent au projet d'ARPANET de se concrétiser 6 ans plus tard.

### I.2.2 L'ancêtre d'Internet : ARPANET

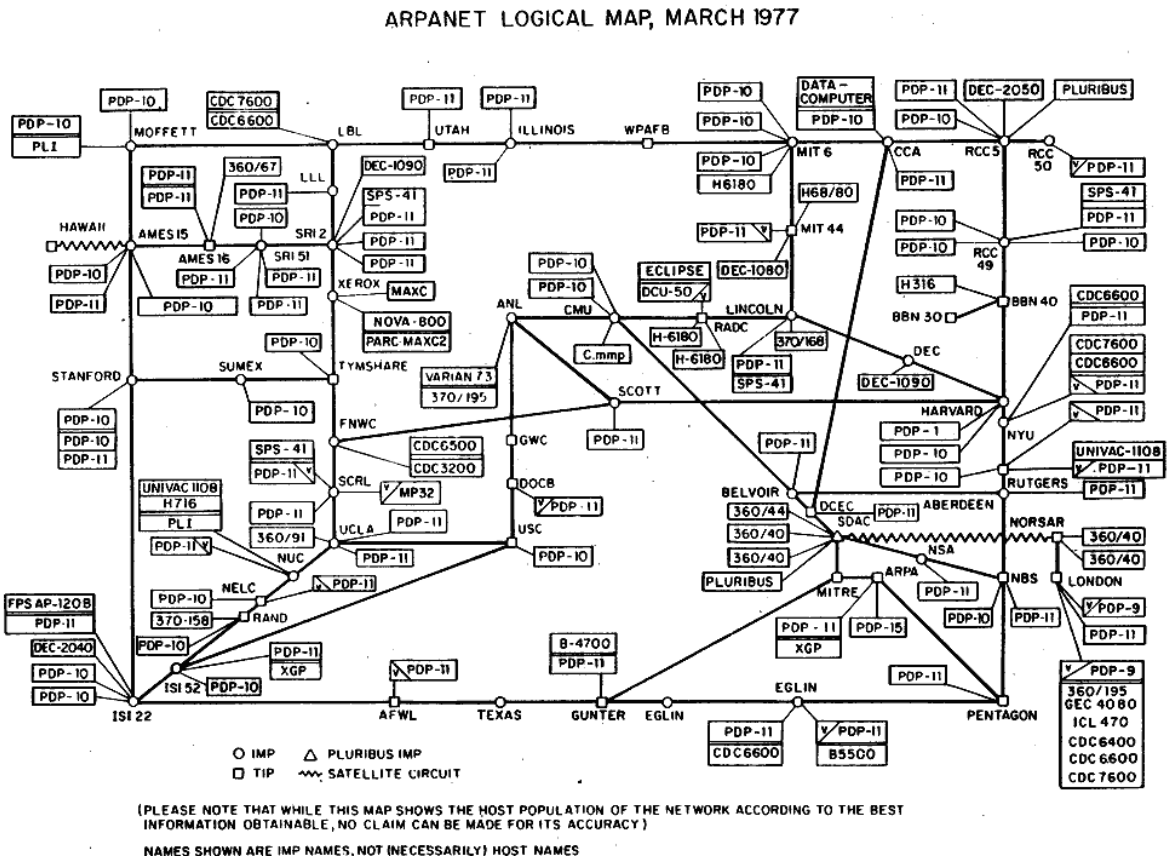



*Figure 5: Représentation du réseau ARPANET en mars 1977 (Computer History Museum)*

En 1968, la société BBN (Bolt Beranek and Newman Inc., une société de consultant en informatique du Massachussetts) remporte le contrat de l'ARPA pour réaliser les premiers équipements du réseau ARPANET : les IMP, pour Interface Messages Processors, qui servent à faire le relais entre le réseau et l'utilisateur et qui sont la première version des routeurs que l'on connait aujourd'hui.

En 1969, le réseau se dessine. Il comporte à la base 4 nœuds, situés dans des grands centres universitaires américains :

- UCLA (Université de Californie à Los Angeles)
- SRI (Institut de recherche de Stanford)
- UCSB (Université de Californie à Santa Barbara)
- L'Université de l'Utah

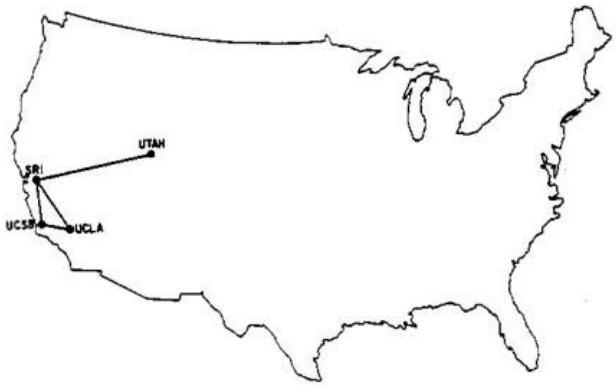



*Figure 6: Les 4 premiers nœuds de L'ARPANET (7)*

Ces nœuds étaient reliés par des câbles permettant un transport d'informations à 50 kbits par seconde, et communiquaient entre eux grâce à un protocole qui établissait la connexion entre les deux IMP : le NCP, pour « Network Control Protocol ». L'ARPANET est là.

On décide de la date conventionnelle de la « naissance d'Internet » le 29 octobre 1969, le jour où un étudiant de l'UCLA, envoie le mot « login » à l'ordinateur hôte du Stanford Research Institute (8).

A noter l'initiative de Louis Pouzin, un français ayant créé en 1971 CYCLADES, un projet de réseau global de télécommunication dont les idées ont bénéficié à l'équipe de développement d'ARPANET, mais qui fut abandonné en 1978 (9).

Entre temps nait le courrier électronique en 1971 inventé par Ray Tomlinson, toujours chez BBN, qui choisit l'arobase @ comme séparateur pour les adresses électroniques, et qui envoie d'après ses souvenirs un sommaire « QWERTYUIOP » (10).

En 1973, devant le nombre grandissant de nouveaux nœuds, il faut trouver une solution pour uniformiser l'échange d'informations entre des machines qui ne sont pas forcément toutes identiques. Pour parer à cela, la suite de protocoles TCP/IP (pour « Transmission Control Protocol/Internet Protocol ») est créée par Vint Cerf de Stanford, et Bob Kahn de la DARPA (ex ARPA).

Ainsi, qu'importe la machine, la donnée transitera toujours à travers plusieurs couches de traitements, et sera lisible pour la machine cible. Ces protocoles déterminent comment une information doit être empaquetée, adressée, transmise, routée (acheminée), et reçue. Elle apporte une solution primordiale pour le futur réseau d'Internet et est toujours utilisée aujourd'hui (11).

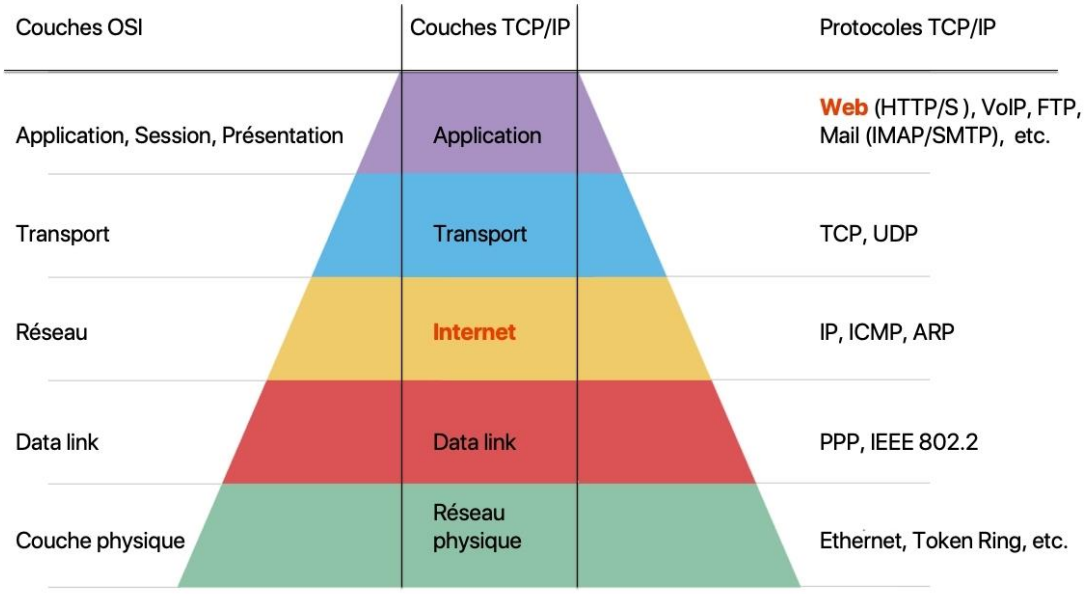



*Figure 7 : Architecture du protocole TCP/IP : couches OSI (Open Systems Interconnection : norme de communication entre ordinateurs) et protocoles TCP/IP. (d'après Larousse.fr)*

Le 1er Janvier 1983, le NCP est rendu obsolète et le TCP/IP est définitivement adopté pour ARPANET.

La même année est mis au point le DNS, « Domain Name System », permettant de gérer le nombre toujours plus grand d'appareils connectés au réseau. En effet, au début de l'ARPANET, pour ajouter une machine au réseau, il fallait appeler le Stanford Research Institute (aux horaires de bureau et hors périodes de Noël) qui enregistrait le nouveau nom et la nouvelle adresse correspondante sur un registre « hosts.txt », registre qui était ensuite distribué et mis à jour sur toutes les machines connectées. Pour résoudre ce problème, Paul Mockapetris, un ingénieur américain, remplace cette base de données unique par une base de données décentralisée sur de nombreux serveurs répartis en zone, de telle sorte que chaque machine du réseau puisse être identifiée (à travers son adresse IP) et puisse identifier à son tour la machine cible en interrogeant ces serveurs (12).

Entre le milieu des années 80 et le début des années 90, l'Europe, l'Afrique et l'Asie se rattachent au réseau.

### I.2.3 Les années 90 : web et usages modernes

En 1990, le réseau militaire se sépare du réseau universitaire : Arpanet disparait, et laisse la place au réseau que l'on connaît aujourd'hui : Internet. Néanmoins c'est véritablement avec la naissance du World Wide Web (ou simplement Web) qu'Internet s'étend au grand public.

Le Web est le fruit d'une combinaison ingénieuse de la technologie hypertexte et du TCP/IP. Ainsi grâce au protocole HTTP (*HyperText Transfert Protocol)* sur lequel est fondé le web, les serveurs sont reliés aux navigateurs par des hyperliens, qui permettent de naviguer à travers différentes pages (sites web) mises sur le réseau, à la manière d'un « livre dont vous êtes le héros », ce qui est une évolution par rapport au modèle d'ARPANET.

On doit le Web à Tim Berners Lee, chercheur au CERN à Genève, qui voyant le nombre (inhabituel pour l'époque) d'ordinateurs utilisant des infrastructures différentes et ne pouvant communiquer entre eux a l'idée de créer un « réseau des réseaux » et

développe sur un NeXT Cube le premier navigateur internet, son ordinateur devenant le premier serveur web, qui ne devait surtout pas être éteint comme l'indiquait son post-it « NE PAS ETEINDRE ».

Un premier navigateur grand public Mosaic est publié en 1993 et constitue une étape majeure dans la popularisation du web.

Aujourd'hui, une guerre commerciale est toujours menée entre les différents éditeurs pour imposer chacun son navigateur, apportant avec elle la problématique d'un monopole dans l'accès à l'information.

En 1993, le CERN ouvre le Web au domaine public avec un format Open Source.

En 1996, la voix sur IP est créée. La même année arrive IPv6, un nouveau protocole devant remplacer IPv4. Cette migration est toujours en cours à l'heure actuelle, malgré l'épuisement des adresses IPv4. Ce protocole permet un nombre illimité d'adresses IP pour répondre à la demande toujours plus forte (téléphones et objets connectés en tête).

En 1998, naît l'ICANN (Internet Corporation for Assigned Names and Numbers), organisme gérant les noms de domaines de premier niveau : l'organisation du web est ainsi finalisée.

Loin des débuts utopiques du web, en 2020 les usages du web sont concentrés autour de grandes multinationales (*GAFAM*), ce que Tim Berners Lee déplore, ainsi que Robert Cailleau : « "Le Web, c'est Facebook et du commercial, rien d'autre. Je ne veux plus y aller. » (13)

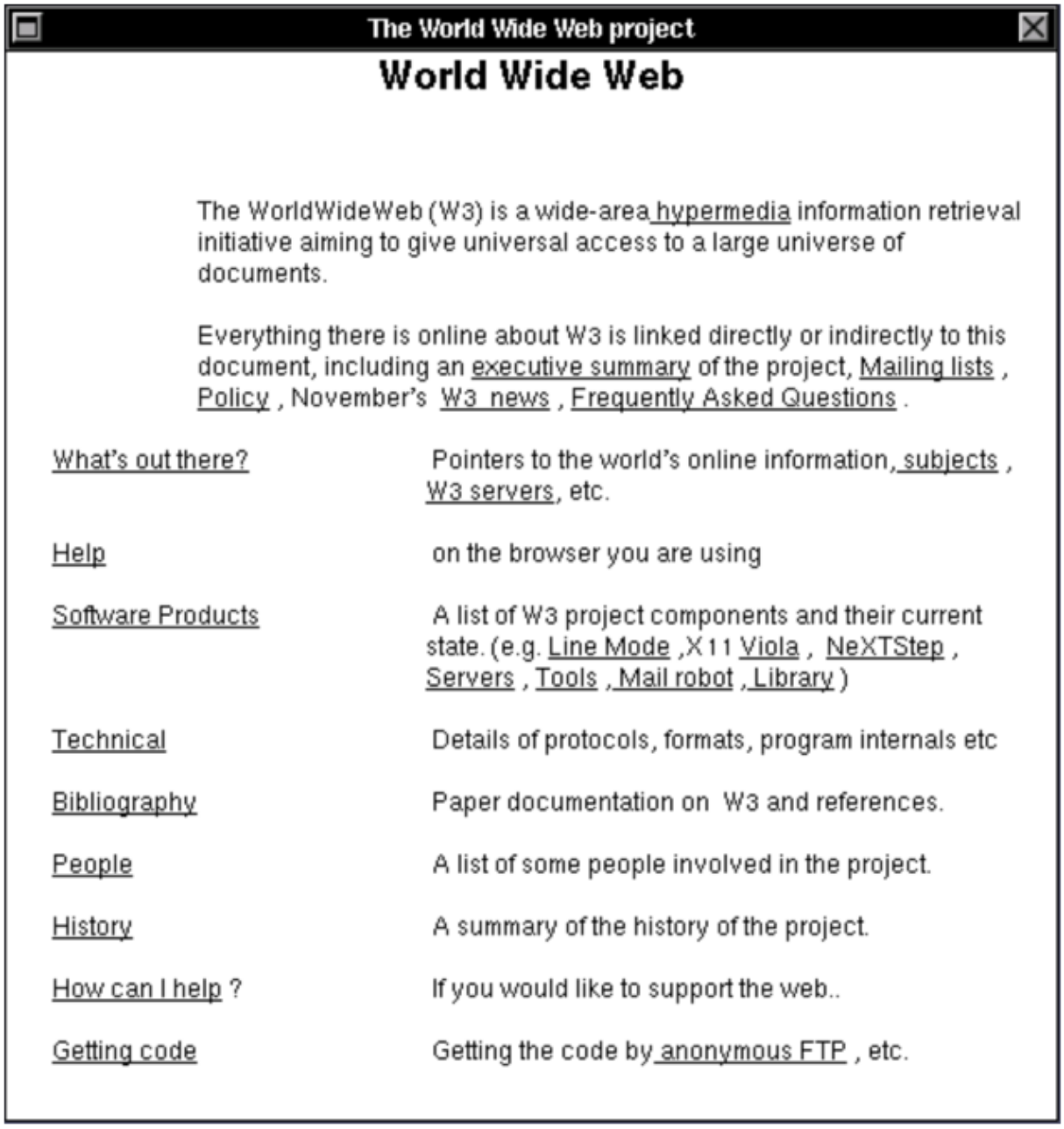



*Figure 8: La première page web, vue depuis le premier navigateur web © CERN*

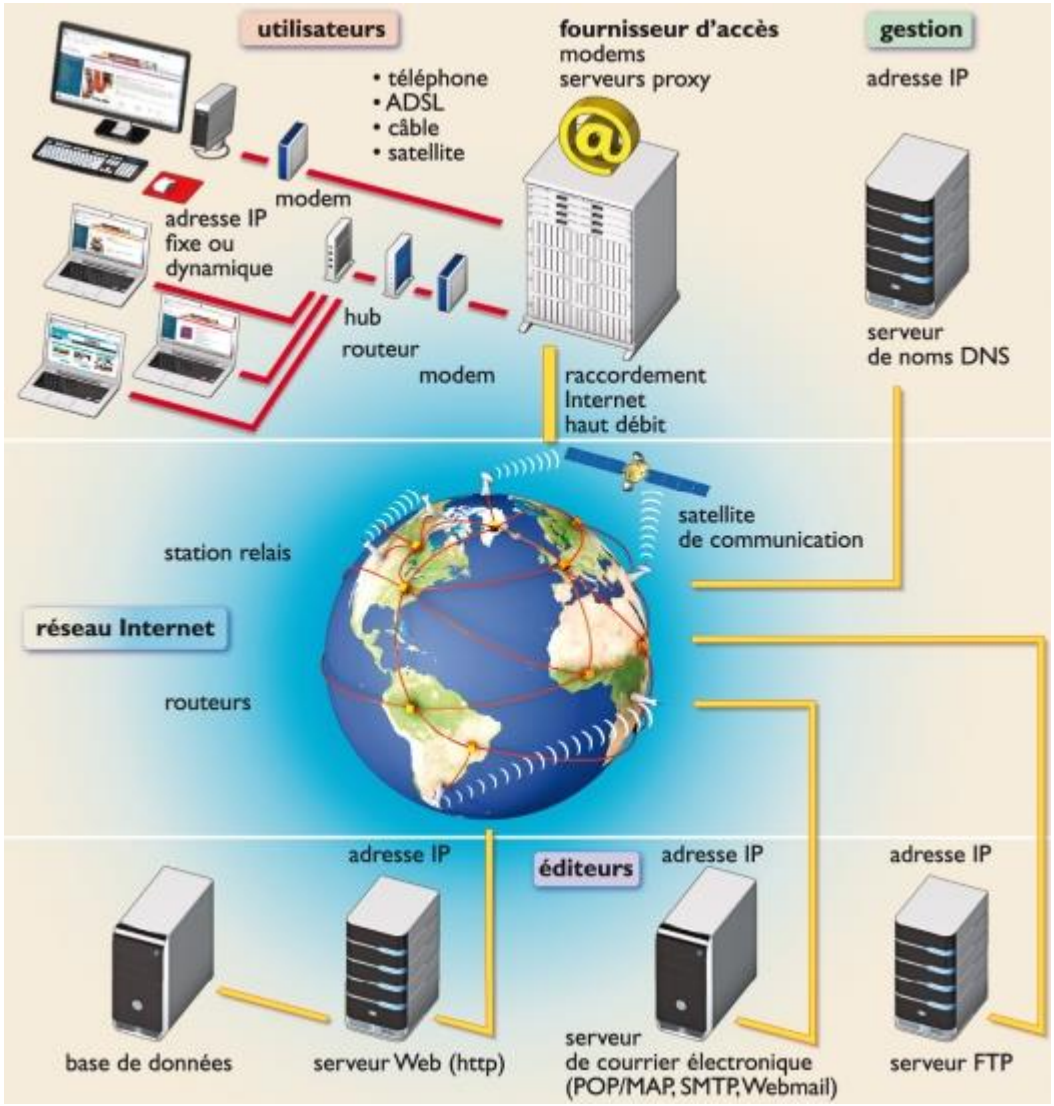



*Figure 9: La structure du réseau d'Internet (14)*

## I.3 A propos du WEB et de ses évolutions : l'Internet des Objets

A travers l'historique de la création d'Internet en point I.2, nous comprenons donc la différence entre Internet et le World Wide Web, ou « toile d'araignée à l'échelle mondiale », qui est seulement une des applications d'Internet, comme peuvent l'être le mail, ou la VoIP (voix sur IP).

En effet, le Web est un système hypertexte public se basant sur le réseau d'Internet. Avec un navigateur, le web permet de consulter différents documents (images, sons, vidéos, textes) disponibles sur des pages web, reliés entre eux par des hyperliens (3).

Il a su évoluer vers un modèle communautaire et collaboratif : c'est le Web 2.0, apparu en 2003. L'internaute, familiarisé avec la technologie d'internet, et devant l'augmentation notable des débits de données, devient acteur de celui-ci : il génère et partage du contenu sur des sites divers, comme des hébergeurs de vidéo en ligne, ou sur les réseaux sociaux.

La santé a naturellement suivi cette tendance avec ce qu'on appelle la « médecine 2.0 » ou « e-santé » (cf. chapitre IV).

Aujourd'hui si le web représente une grande partie d'Internet, d'autres usages sur ce réseau se sont fortement développés comme la communication entre les machines à l'instar des objets connectés : c'est l'Internet des Objets[2].

L'Internet des Objets est issu de l'évolution des usages du web et des équipements électroniques. Il a été défini à la suite de l'émergence au quotidien des objets connectés tels que les capteurs, smartphones, les ordinateurs portables, les DMC, les puces à RFID… qui sont capables d'interagir entre eux via l'échange de données.

L'Union internationale des télécommunications définit l'Internet des objets comme une « infrastructure mondiale pour la société de l'information, qui permet de disposer de services évolués en interconnectant des objets (physiques ou virtuels) grâce aux technologies de l'information et de la communication interopérables existantes ou en évolution »(15).

[2]L'internet des objets ou *Internet of Things (IoT)* désigne l'interconnexion entre l'Internet et des objets, des lieux et des environnements physiques.

En santé, cela a donné naissance au mIoT, ou *Medical Internet of Things*, qui est défini comme un réseau d'objets physiques et d'autres constituants, contenant de l'électronique, des logiciels, des capteurs, et une connectivité à un réseau, qui permettent à ces objets de collecter et d'échanger des données dans le secteur de la santé et du bien-être (16). Ces objets peuvent être des objets connectés simples comme une montre connectée qui suit le nombre de pas effectuées dans une journée, à des dispositifs médicaux certifiés (marquage CE) comme une pompe à insuline connectée qui facilite le suivi de traitement du diabète, ou encore des applications mobiles qui peuvent donner des conseils sur une pathologie particulière ou mettre en contact le patient avec un professionnel de santé. On estime qu'en 2020, 40% du marché de l'IoT a été relatif à la santé, pour un marché de $117 milliards de dollars (17).


*Figure 10: Publicité pour un oxymètre connecté (d'après rueducommerce.fr)*

Ainsi avec le développement de l'informatique, la technologie et la philosophie derrière la création d'Internet a permis l'émergence d'énormes flux de données, générées et échangées à travers le monde.

L'Internet des objets est en partie responsable de cet accroissement exponentiel du volume de données.

Cette ressource d'information gigantesque a fini par trouver son propre moyen de stockage et d'analyse : c'est le « big data ».

# II. LE « BIG DATA » : LA GESTION DE L'INFORMATION

## II.1 BIG DATA

### II.1.1 Définition

Le terme « big data » désigne les « grosses données », ou « données massives ». Ce sont les données produites par voie informatique : les photos, vidéos, sons, textes, logs…, qui sont toujours plus nombreuses, avec la démocratisation d'Internet d'une part, mais aussi les changements d'usages dans le grand public via les réseaux sociaux et l'apparition de nouvelles façons de se connecter comme avec les objets connectés. Autrement dit l'omniprésence du web et d'internet, l'ensemble des objets connectés, les grands projets scientifiques comme le LHC[3] du CERN sont autant de facteurs expliquant la production en constante augmentation des données.

Une donnée en informatique est composée de bits. Un bit est l'unité de mesure de base de l'information en informatique et peut prendre la valeur 0 ou 1. En général ils sont regroupés par 8 sous forme d'octets dans les programmes informatiques.[4]

En 2020, une étude estime que 40 Zettaoctets de données (soit 40 000 milliards de Go) seront générées.(18) Pour l'exemple, 150 millions d'email sont envoyés par minutes, Facebook produit 4000 To de données par jour, et 7000 To par secondes de données sont produites par le radiotélescope Square Kilometer Array.(19)

De plus la donnée peut être structurée ou non structurée[5] ; élémentaire ou sous forme de métadonnée[6], il faut donc un système adapté pour stocker, gérer, et traiter ces données.

La frontière entre Big Data et données traditionnelles se fait dans les ressources nécessaires pour le traitement de ces données. D'après le physicien Pirmin Lemberger, la distinction se fait à partir du moment où « ces données ne peuvent plus

[3] Le LHC (Large Hadron Collider) ou Grand collisionneur de hadrons est un accélérateur de particules mis en fonction en 2008 et situé dans la région frontalière entre la France et la Suisse. En 2012, il confirme l'existence du boson de Higgs.

[4] Toutefois dans certains langages informatiques comme le C, les bits peuvent être regroupés dans des bytes qui ne contiennent pas forcément 8 bits.

[5] Sans hiérarchie

[6] Une métadonnée contient des informations relatives à la donnée en elle-même, ce n'est pas une donnée élémentaire.

être traitées en un temps raisonnable ou utile par des systèmes constitués d'un seul nœud. » (20)

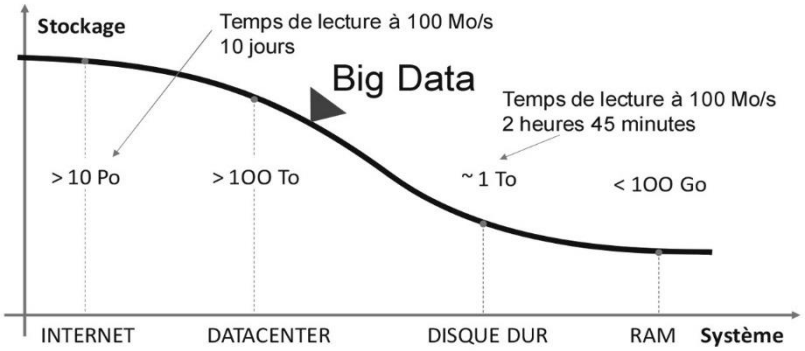



*Figure 11: Frontière du Big Data en fonction du stockage et du système gestionnaire (15)*

### II.1.2 Historique

L'avènement du big data est à mettre en relation avec l'explosion du volume des données produites, qui a déjà commencé dans les années 1980 avec le début de l'informatique personnelle mais qui a véritablement pris de la vitesse dans les années 2010 avec Internet. (cf. Figure 12)

De plus les données ne sont plus générées seulement par les internautes, mais aussi par les machines elles-mêmes, comme les véhicules ou les objets du quotidien, et également par les entreprises et organisations via l'*open data* : horaires de passages des transports en communs, statistiques démographiques…

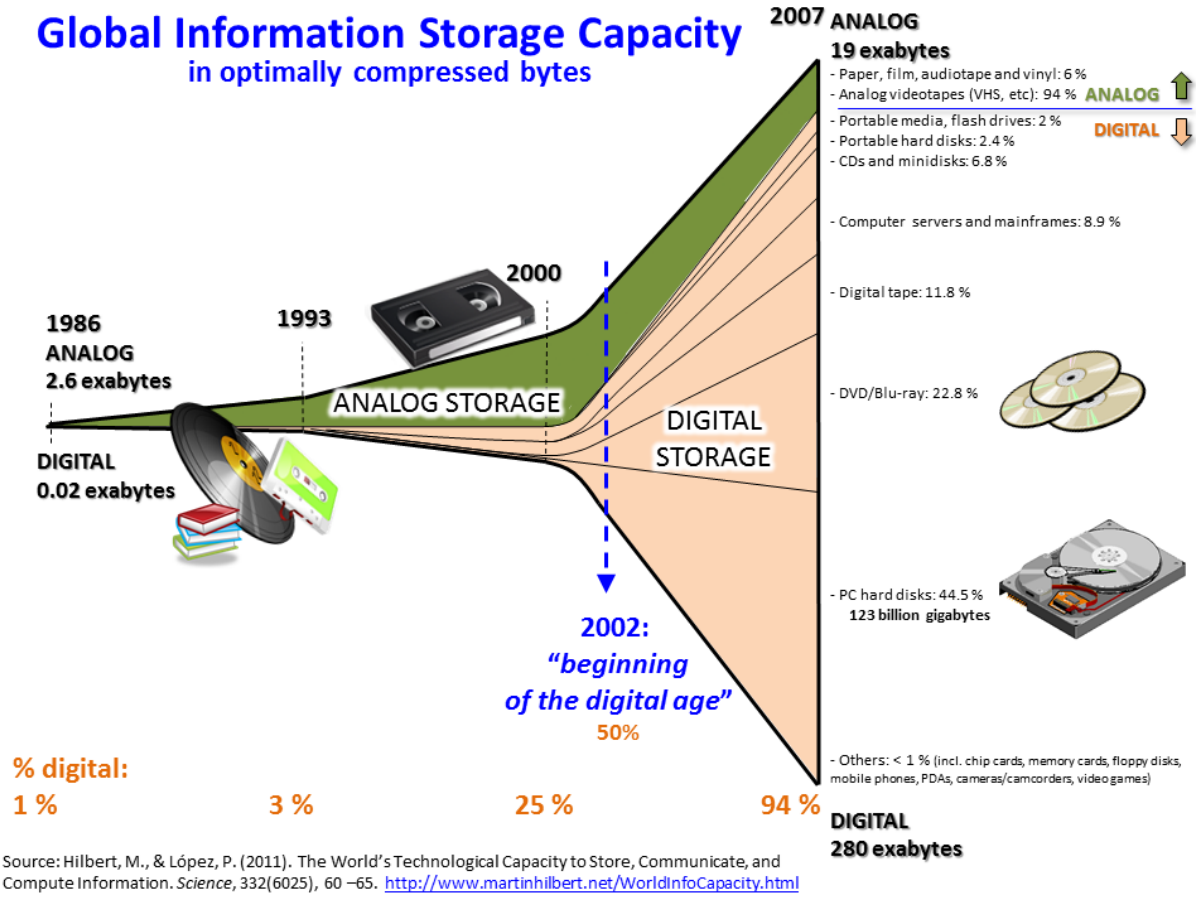



*Figure 12: Croissance de la capacité mondiale de stockage des données*

Son développement ses dernières années s'explique d'un point de vue économique avec la baisse de prix des composants, comme ceux permettant le stockage, ou ceux des nœuds dans un centre de serveurs qui peuvent maintenant être très nombreux, mais aussi ceux des processeurs (CPU) qui ont suivi la loi de Moore[7] ou de la bande passante qui est devenue moins chère et qui a permis l'émergence des solutions décentralisées. (cf. figure 13)

Il s'explique aussi par l'évolution logicielle qui a permis le traitement de cette grande masse de données. C'est le progrès de la parallélisation des traitements des données, initiées par Google avec le modèle MapReduce, puis en open source avec l'implémentation de la bibliothèque Hadoop d'Apache. Ces modèles apportent des

[7] Prédiction de l'ingénieur Moore qui a énoncé dans les années 60 et 70 que le nombre de transistors des microprocesseurs sur une puce de silicium doublerait tous les deux ans, ce qui s'est avéré. Ainsi les machines électroniques se sont miniaturisées et ont couté de moins en moins chers tout en devenant de plus en plus rapides et puissantes.

solutions aux problèmes de distribution des calculs, des données et de tolérance aux pannes.

Les systèmes de bases de données ont aussi évolué et se sont adaptés à la quantité de données à traiter et des systèmes NoSQL[8] ont vu le jour.

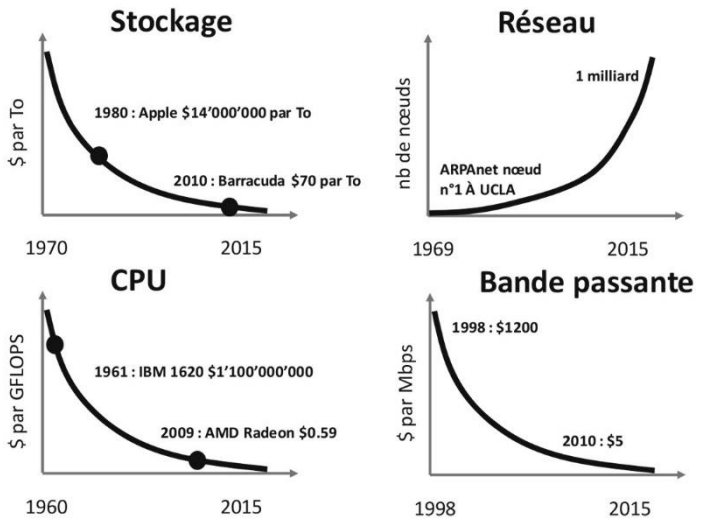



*Figure 13: Evolution des principaux composants favorisant l'expansion du Big Data (15)*

### II.1.3 Défis technologiques du big data

L'exploitation des grosses données va à l'encontre de l'informatique traditionnelle. L'informatique traditionnelle est basée sur un modèle alors que le big data vise à trouver un modèle dans les données.

Les acteurs du domaine définissent les problématiques du big data autour de 3 V : Volume, Vélocité et Variété.

Ainsi ces 3V caractérisent des phénomènes qui requièrent des réformes des outils et des méthodes utilisés dans une application pour des « petites données » :

- Le Volume des données générées nécessite une nouvelle façon de les stocker.

[8] Les systèmes NoSQL sont des systèmes de gestion de base de données adaptés à la gestion de volume importants de données, contrairement aux systèmes de gestion traditionnels.

- La Vélocité à laquelle ces données sont générées et transmises requiert de mettre en place des solutions de traitement en temps réel qui ne bloquent pas le reste de l'application.
- La Variété de formats sous laquelle se présentent les données implique de mettre en place des outils appropriés pour les acquérir, les analyser et les stocker ; ces données pouvant être structurées (documents JSON), semi-structurées (fichiers de log) ou non structurées (textes, images).

### II.1.4 Traitement de grosses quantités de données

Le traitement des *big data* implique deux modèles de calculs adaptés aux grands volumes de données :

- Le modèle dédié à la recherche d'information,
- Le modèle dédié à l'analyse de données.

Ainsi avant d'exploiter une donnée il faut d'abord la « prétraiter » : la classer, la ranger dans une catégorie grâce à des algorithmes, la rapporter à un même ordre de grandeurs par rapport aux autres (par exemple rapporter toutes les valeurs entre 0 et 1) pour faciliter son traitement.

Encore sous exploitée il y a peu, le *big data* repose sur 2 technologies centrales : le calcul à haute performance, ou *High Performance Computing* (*HPC*), et l'analyse de données à haute performance, ou *High Performance Data Analytics* (*HPDA*).

Le HPC concerne la science des ordinateurs supercalculateurs, qui ont une énorme puissance de calculs, de l'ordre de 1 pétaflops ($10^{15}$ flops), et qui permettent de calculer des simulations demandant beaucoup de ressources informatiques.

Le HPDA concerne la science de la gestion et de l'analyse de très gros volumes de données, pour traiter des gros flux de données, les comprendre et prendre les décisions adéquates.

D'après Jean-Laurent Philippe, spécialiste HPC chez Intel, on peut s'attendre à une convergence entre les techniques de HPC et HPDA, du fait des infrastructures autour des centres de calculs conçues pour la HPC qui offrent d'importantes capacités de stockage, ce qui est intéressant pour les techniques de HPDA, qui bénéficient

également de la puissance de calculs des superordinateurs pour le calcul à haute performance (21).

Un rapprochement s'effectue également entre les technologies HPC et celles de l'intelligence artificielle, étant donné la similitude de leurs applications avec le HPDA en données ou en calcul, sur des superordinateurs, sur des petits groupes de machines (*cluster*), ou bien dans des serveurs dans le nuage.

De plus l'infrastructure utilisée pour le HPC est aussi utilisée pour les technologies d'intelligence artificielle, dans un but d'économie des coûts.

Cela permet à l'IA de se développer dans des projets qui demandent d'énormes ressources en calculs et en base de données, comme pour les voitures autonomes qui doivent analyser en temps réel leur environnement et prendre une décision dans l'instant ou dans des projets médicaux pour analyser des images de lames très volumineuses (cf. V.2).

Pour que les données soient traitées correctement, les calculs doivent être réalisés par des algorithmes évolutifs. Ils doivent s'adapter aux processeurs, aux types des machines, à leur nombre, aux logiciels utilisés pour les traiter, afin d'être plus rapides et plus efficaces. Certains langages de programmation sont aussi favorisés, tels que Julia ou Python, car ils sont faciles d'accès et plus productifs. (cf. III.3)

La convergence entre HPC et IA se fera avec le développement des puissances de calculs des processeurs, chiffrées en flops[9], mais aussi avec l'évolution matérielle des processeurs, se rapprochant plus d'une architecture des réseaux neuronaux, à l'instar chez Intel des processeurs adaptés pour des calculs complexes et d'immenses bases de données, avec leur gamme Nervana Neural Network

[9] Le flops est une unité de mesure informatique qui représente le nombre d'opérations en virgule flottante par seconde que peut effectuer un processeur (ou *Floating-Point Operations Per Second,* FLOPS). Plus le nombre de flops d'un microprocesseur est élevé et plus celui-ci a une puissance de calcul élevée. Une autre unité utilisée pour mesurer la performance des processeurs est l'IPS (Instruction Par Seconde). Les nombres en virgule flottante sont l'équivalent en informatique des nombres écrits en notation scientifique : ils sont écrits avec une mantisse et un exposant.

*Ex :* $1{,}905 = 1905 \times 10^{-3}$ *où 1905 est la mantisse et 10-3 l'exposant.*

Processors (qui sera remplacée par la gamme de puces de la société Habana Labs spécialisée en IA et rachetée par Intel). (22)

Elle se fera également par l'optimisation logicielle des programmes d'IA devant travailler avec des modèles de programmation unifiés, tout en s'adaptant à des architectures *hardware* variées.

### II.1.5 Gérer des grosses quantités de données, l'exemple de MapReduce

C'est le rôle des *data architects* que de fournir aux *data scientists* les architectures qui permettront de faire tourner leurs algorithmes.

Pour traiter les ensembles de données volumineux, ces spécialistes utilisent généralement Hadoop, MapReduce et Spark.

Nous nous intéresserons à MapReduce qui est la solution logicielle qui a popularisé le traitement des *big data* (23).

MapReduce, inventé par des ingénieurs de Google en 2004 (24), est un patron d'architecture de développement informatique, à savoir une solution générale à un problème d'architecture informatique récurrent. Il sert de modèle de référence pour la conception de l'architecture d'un système ou d'un logiciel informatique, MapReduce a donc eu une grande influence sur le traitement des données. Cet outil accélère le processus de traitement des données grâce au parallélisme, c'est-à-dire à une distribution des tâches. MapReduce permet de manipuler de grandes quantités de données en les distribuant dans un *cluster*[10] de machines pour être traitées.

MapReduce, comme son nom l'indique, a deux fonctions : compiler et cartographier les données (*map*) puis réduire les données dans des classes (*reduce*). Autrement dit, la phase *map* répartit les données sur plusieurs serveurs, chacun réalisant la même tâche, puis la phase *reduce* prend les paquets de données intermédiaires et réalise la tâche suivante : additionner les paquets identiques pour aboutir au résultat final. Entre les deux phases, une « boite noire » (ou *shuffle*) regroupe les paquets

[10] Un cluster est un groupe de serveurs fonctionnant comme un seul et unique système.

suivant les caractéristiques de classification souhaitée avant de les acheminer aux serveurs suivants qui pratiqueront la phase *reduce*.

En résumé ce logiciel compile et organise les ensembles de données pour ensuite les affiner en des plus petits ensembles organisés, ce qui permet de répondre à la problématique des 3V vue précédemment.

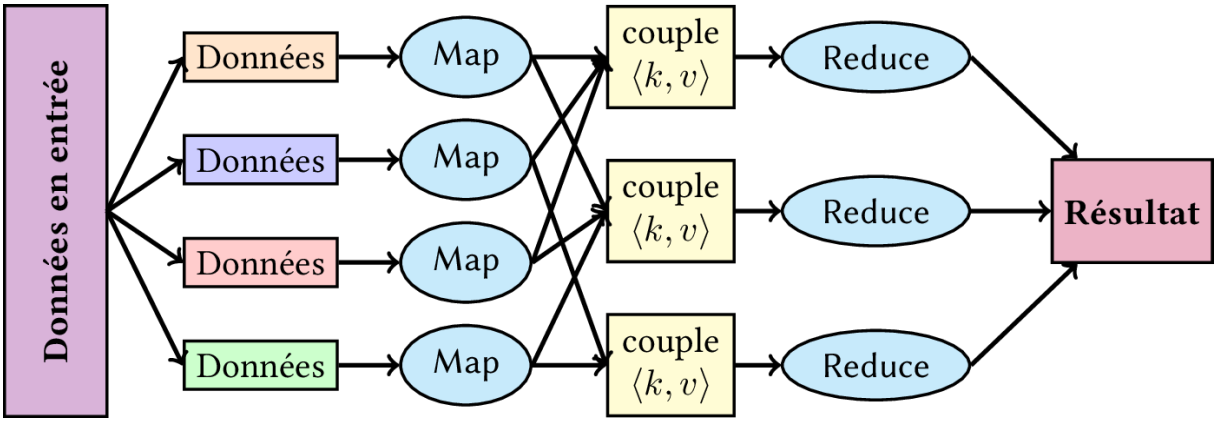



*Figure 14: Schéma de fonctionnement du MapReduce (d'après Clém IAGL)*

## II.2 Le *cloud computing* : l'accès aux données facilité

L'application la plus connue du big data est le cloud.

### II.2.1 Plateformes de cloud computing

Amazon AWS, Microsoft Azure ou Google BigQuerry sont des solutions de *cloud computing*. Ils ne sont pas compatibles entre eux, ce qui peut poser des problèmes d'interopérabilité quand des solutions en entreprise utilisent plusieurs plateformes. (cf. II.3.1).

### II.2.2 Principe du nuage

Un cloud (« nuage ») est un ensemble de matériels, de raccordements réseau et de logiciels fournissant des services qu'individus et collectivités peuvent exploiter depuis n'importe où dans le monde (25).

Le cloud consiste à louer à des tiers des ressources matérielles pour une durée déterminées. Ces ressources fournissent de la capacité de calcul (des serveurs) et de stockage (de l'espace disque). Les fournisseurs de cloud sont des hébergeurs professionnels qui se chargent de la mise en place, de l'entretien et du renouvellement du matériel.

Cette solution est la base de l'internet des objets et du développement des solutions pour les entreprises commerciales et organisations à but de recherche, qui se reposent sur les services proposés par les plateformes, que ce soit du simple stockage ou alors une utilisation de leur puissance de calcul.

Les principaux services fournis par les hébergeurs sont :

- Le IaaS : Infrastructure as a Service : c'est le service de plus bas niveau informatique qui consiste à donner accès à des machines virtuelles aux utilisateurs, qui pourront installer des systèmes d'exploitation et des logiciels.
- Le PaaS: Platform as a Service. Ici le système d'exploitation est déjà installé et l'utilisateur peut installer des applications.
- Le SaaS : Software as a Service. Dans ce type de service, les applications seulement sont mises à la disposition du consommateur.

Un avantage du cloud est son élasticité. En effet les systèmes sont capables d'agrandir ou de diminuer les capacités allouées pour un temps variable. Cela permet au *cloud* de supporter des pics de charges temporaires.

Du coté des particuliers, les avantages de cette solution sont une meilleure durée de vie par rapport au support physique (CD, clé USB, disque dur externe…), mais aussi une disponibilité constante à condition d'être connecté à Internet.

Ainsi les données sont enregistrées via Internet directement dans un espace privatif et sécurisé par un chiffrement et des identifiants de connexion, dans des data centers spécialisés dans le stockage de données. Ces structures, baptisées « fermes de serveurs » ou « data center », contiennent des armoires contenant des disques durs sur des centaines de mètres. Classiquement, l'ensemble est enfermé dans des entrepôts ou bunkers et plusieurs copies des données sont réalisées pour les sécuriser.

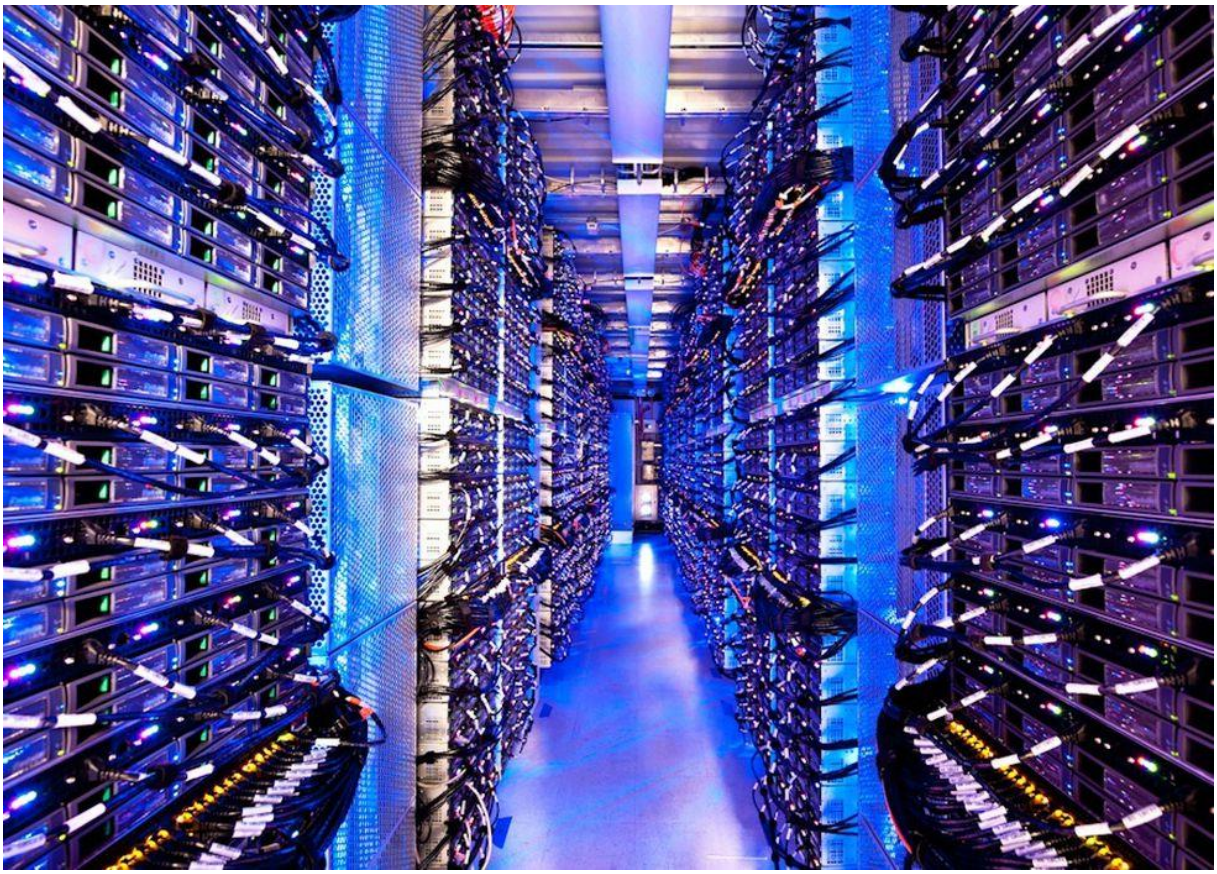


*Figure 15: Intérieur d'un data center de Microsoft dans la région de Chicago © Microsoft*

## II.3 Problématiques générales

### II.3.1 L'interopérabilité

L'interopérabilité est définie comme « la capacité d'au moins deux systèmes ou produits d'échanger des données et d'utiliser de l'information » d'après Serrano et al. (26). De la santé à l'agriculture, chaque domaine possède ses propres standards, par un fonctionnement dit « en silo » (« *data silo* »). En effet l'industrie du big data fourmille de quantités de technologies, de différents types de données et de structures, qui ne fonctionnent pas forcément correctement entre elles.

Dans une interview pour le site ISO.org, le site de l'organisation mondiale de normalisation, Klaus-Peter Eckert, un scientifique spécialisé dans les systèmes de communication ouverts, indique qu'il faut normaliser les technologies du big data (27) :

« Dans le domaine du big data et de l'analyse de données, nous avons plusieurs outils qui ont été mis au point par différentes communautés au cours des dernières années. Parallèlement, différents types d'infrastructures TI, notamment dans l'informatique en nuage, ont vu le jour indépendamment du big data. Toutes ces composantes techniques sont à notre portée mais elles ne s'articulent pas les unes avec les autres fautes d'interopérabilité. Il nous manque une architecture qui serait acceptée largement et permettrait de rassembler tous ces éléments. Et c'est justement là que les normes ont un rôle à jouer. »

Ces normes qui sont le sujet du groupe d'étude ISO/CEI JTC1, doivent participer à l'élaboration d'une plateforme du big data uniforme et favorable pour les entreprises, qui pourront agréger et analyser correctement leurs données.

### II.3.2 La propriété des données

Les données sont sous la juridiction de l'Etat où se situe le serveur.

Cela est d'autant plus problématique quand cela concerne les données de santé (cf. II.4.1), qui sont centralisées principalement autour d'acteurs propriétaires américains : en tête Amazon et Microsoft.

En France, il existe des solutions *open source* comme le service *cloud* Cryptpad de l'association InterHop qui milite pour l'utilisation des logiciels libres et une utilisation auto-gérée des données de santé à l'échelle locale (28).

Il est à noter l'initiative en 2021 du gouvernement français d'un label « cloud de confiance » qui certifie que « chaque produit numérique manipulant des données sensibles, qu'elles relèvent notamment des données personnelles des citoyens français, des données économiques relatives aux entreprises françaises, ou d'applications métiers relatives aux agents publics de l'État, devra impérativement être hébergé sur le cloud interne de l'État ou sur un cloud industriel qualifié [cloud de confiance] par l'ANSSI[11] et protégé contre toute réglementation extracommunautaire » (29).

[11] Agence nationale de la sécurité des systèmes d'information

### II.3.3 La difficulté de garantir l'accès constant aux données

La centralisation de données de plusieurs millions de personnes et d'entreprises par un seul acteur devient problématique quand celui-ci subit une défaillance.

Les exemples sont courants : que ce soit les pannes de services de Google affectant des millions d'usagers dans le monde rendant inaccessible mails, vidéos à la demande voire leur maison connectée (30), ou un incident dans la maintenance des serveurs d'Amazon qui coupe l'accès à de nombreux sites reposant sur leur service (31).

A chaque fois la perte de productivité pour les entreprises qui utilisent ces solutions est énorme.

Malgré les solutions déployées par les acteurs du *cloud*, à l'instar de la duplication des données sur différents serveurs dans des lieux géographiques différents, les problèmes sont encore courants.

### II.3.4 La sécurité des données

En stockant ses données sur des serveurs appartenant à des sociétés commerciales, on s'expose au risque de voir ses données exploitées à des fins commerciales, publicitaires… mais également au risque que la société exploitante se fasse pirater ses données et que nos informations se retrouvent en vente au marché noir, ce qui est problématique pour des informations sensibles comme les informations bancaires, les orientations politiques, ou les dossiers médicaux.

La sécurité des données repose sur la sécurité des systèmes d'information. Pour Jean François Pillou, au ministère de l'Education nationale à la délégation aux Usages de l'Internet : « Le système d'information représente un patrimoine essentiel de l'organisation, qu'il convient de protéger. La sécurité informatique consiste à garantir que les ressources matérielles ou logicielles d'une organisation sont uniquement utilisées dans le cadre prévu » (32).

La sécurité des systèmes d'information doit respecter trois fondamentaux : (33)

- Une confidentialité des données : seuls les utilisateurs autorisés peuvent accéder à leurs données. Les mesures pour garder les informations confidentielles peuvent être le chiffrement des données[12], l'accès protégé par mot de passe, l'authentification à double facteur[13], les systèmes d'accès biométriques.
- Une intégrité des données : les données doivent être exactes, complètes, et non altérées que ce soit de façon fortuite, illicite ou malveillante. Les mesures pour garantir l'intégrité peuvent être le hashage des données.[14]
- Une disponibilité des données : l'information doit être disponible rapidement quand cela est nécessaire. Pour y parvenir, on peut utiliser la redondance des données, des sauvegardes stockées dans d'autres sites, ou dans le cas du big data des clusters haute disponibilité, ou *High-availability clusters*[15].

Concernant la cyberdéfense et la protection des Etats, le contre-amiral Arnaud Coustillière, officier général à la cyberdéfense estime dans une interview pour Futura-Sciences : (34)

« le big data est une évolution qui fait irruption dans le monde de la Défense avec une multiplication de terminaux mal sécurisés, mal maîtrisés. Plus que le risque lié au stockage de données, ce qui m'inquiète, en tant que défenseur des systèmes d'information, c'est l'hétérogénéité, la taille de ce système et son nombre d'utilisateurs, avec les possibilités de failles et d'erreurs que cela implique. »

Ainsi il explique mettre en place des dispositifs pour protéger les données à plusieurs niveaux : les données personnelles des citoyens, les données numériques des

[12]Le chiffrement ou cryptage des données est un procédé de cryptographie qui réalise une conversion des données d'un format lisible à un format codé qui peut uniquement être lu ou traité avec une clé de déchiffrement.
[13] L'A2F est une technique et un processus de vérification de l'identité d'une personne par l'utilisation de deux modes d'identification.
[14] Une fonction de hachage, de l'anglais hash function (hash : pagaille, désordre, recouper et mélanger) par analogie avec la cuisine, est une fonction qui calcule une empreinte numérique servant à identifier rapidement la donnée initiale, à la manière d'une signature qui identifie une personne.
[15] Ce sont des groupes de serveurs qui peuvent supporter des utilisations qui nécessitent un minimum de temps de latence.

entreprises et organisation, et les données gouvernementales. En France, pour se protéger en amont, les nouveaux systèmes et bâtiments (bateaux, avions) doivent être homologués et répondre à certaines normes (EBIOS, ISO). Certaines systèmes complexes sont audités. Pour se protéger de menace active ou d'intrusion dans un système de l'Etat, des entités comme le centre opérationnel de l'Agence nationale de la sécurité des systèmes d'information (Anssi) ou le Centre d'analyse de lutte informatique défensive (Calid) dans le périmètre du ministère de la Défense, ou encore la Dirisi (Direction Interarmées des Réseaux d'Infrastructure et des Systèmes d'Information de la défense), qui est l'opérateur unique des systèmes d'information et de communication (SIC) de la défense, ont un rôle important et cherchent des anomalies de comportement sur les systèmes d'information. Au niveau du stockage des données, beaucoup de serveurs sont gérés dans un cloud interne par la Dirisi. Les données les plus sensibles sont traitées en interne, ce qui évite les piratages extérieurs.

Concernant les données de santé, elles nécessitent un traitement spécial détaillé au point II.4.1.

### II.3.5 Impact environnemental du stockage et des calculs dans le nuage

L'accord de Paris sur le climat, ratifié en 2015 par la presque totalité des pays du globe, vise la neutralité carbone d'ici 2050 pour ne pas dépasser les 2°C d'augmentation de la température moyenne. Cela implique une division par 6 des émissions de CO2 pour les 30 ans à venir (35).

Ainsi, alors que la contrainte climatique nous amène à viser une diminution des émissions mondiales dans les prochaines années, celles du secteur du numérique pourraient doubler d'ici 2025 pour atteindre 8 % du total.

En effet d'après l'étude du Shift Project[16], la consommation électrique globale du secteur numérique représente 14% de la consommation mondiale et ne cesse

[16] Groupe de réflexion d'experts autour de la question du changement climatique et de la raréfaction des ressources énergétiques fossiles

d'augmenter, tout comme la part d'émissions de gaz à effet de serre qui constituerait près de 4% des émissions mondiales.

L'exploitation du *Big Data*, et du numérique en général, est grand consommateur d'énergie et génère de nombreuses émissions de gaz à effet de serre.

De plus la fabrication d'un appareil est également polluante, notamment par l'exploitation de mines de métaux rares, ce qui en plus d'un coût environnemental, a aussi un coût humain ; d'autant que certaines industries comme celle du mobile produisent des nouveaux modèles tous les ans, au détriment du suivi des anciens appareils et de la qualité de leur composants, fragiles et non recyclables.

A cela s'ajoute l'augmentation de taille des contenus sur internet, qui ont suivi avec l'augmentation de la vitesse des réseaux par un effet rebond[17]. Ainsi une page web en 1995 pesait 14Ko quand elle pèse 1600Ko en 2015, soit une multiplication du poids par 115.

Certaines actions sont prises pour limiter la consommation électrique. Microsoft a immergé des serveurs de données dans l'océan au large de l'Ecosse, et a eu des résultats concluants en septembre 2020 : la consommation d'énergie des serveurs a été réduite, la casse des composants a été moins fréquente et l'installation n'a pas eu recours à de l'eau potable (37).

Google également s'engage à utiliser le maximum d'énergies renouvelables (éoliennes, solaires, hydrauliques) pour leur centre de données, comme celui qui doit être livré en 2021 au Danemark (38).

Du coté législatif, les lois du numérique portent surtout sur la protection des données.

Pourtant il est nécessaire de légiférer également sur les actions des acteurs du numérique car il est irréaliste de seulement compter sur une auto-discipline.

Le Sénat en 2019 dans sa mission « Régulation des réseaux sociaux - Expérimentation Facebook » propose une coordination à l'échelle européenne, car pour l'instant les lois sont régies par le pays où se situent les installations du siège social de l'entreprise informatique. Cela permettrait d'être plus efficace dans

[17] L'effet rebond peut être défini comme l'augmentation de consommation liée à la réduction des limites à l'utilisation d'une technologie. (36)

l'établissement de nouvelles lois, qu'elles concernent la modération de contenus ou la limitation de bande passante pour réduire l'empreinte carbone de la vidéo sur internet par exemple.

Le Sénat préconise également l'échange entre les différents acteurs : les industriels, les gouvernements, les régulateurs et la société civile ; dans une démarche de « débat public » pour favoriser des solutions cohérentes qui ne seront pas qualifiées de « liberticides ».

## II.4 Problématiques concernant les données de santé

### II.4.1 Confidentialité et souveraineté des données de santé

La possession des données est un enjeu économique. Il existe une difficulté de maintenir une souveraineté des données quand une majorité est hébergée sur des serveurs américains, détenu principalement par Amazon avec son service AWS et Microsoft avec Azure.

Ces données de santé étant sensibles, elles sont encadrées par la loi Informatique et Libertés de 1978. De plus, depuis 2016 leur garantie est renforcée par le nouveau règlement européen sur la protection des données personnelles. Ce dernier qualifie pour la première fois les données de santé comme des « données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne » (39).
Ainsi chaque hébergeur doit avoir obtenu en France un agrément préalable pour l'hébergement des données de santé à caractère personnel, conformément à l'article L. 1111-8 du Code de la Santé Publique (40).

Pourtant, l'ingérence des Etats-Unis se fait sentir en France : les lois américaines Patriot Act et le Cloud Act permettent grâce à l'extra-territorialité du droit américain, d'accéder par voie légale à des données personnelles de citoyens français (41).

Le cas du Health Data Hub permet d'illustrer cette problématique : ce projet de l'état français de centralisation des données de santé des Français pour mieux gérer l'urgence sanitaire et les connaissances sur le SARS-Cov-2, est hébergé par Microsoft Azure, la partie « *cloud* » de Microsoft.

Le 13 octobre 2020, le Conseil d'Etat estime le risque d'une surveillance par les autorités américaines, du fait de l'existence des programmes de surveillances américains (l'article 702 du « *Foreign Intelligence Surveillance Act* » (FISA) et l'*«Executive Order (EO) 12333* »), malgré les protections européennes et demande à ce que de nombreux avenants soient ajoutés au contrat d'exploitation entre la plateforme des données de santé et Microsoft pour garantir la protection des données (42).

### II.4.2 Aspects réglementaires

Des organismes existent pour encadrer les nouveaux usages des données.

En France existe la CNIL, pour Commission Nationale Informatique et Liberté, qui a été créée par la loi Informatique et Libertés du 6 janvier 1978. (43) Elle est chargée de veiller à la protection des données personnelles, ainsi que de veiller à ce que « l'informatique soit au service du citoyen et qu'elle ne porte pas atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. » Elle travaille aussi au niveau européen pour harmoniser la régulation des données personnelles et de leur traitement, comme le prévoit le RGPD.

Le RGPD, pour Règlement Général sur la Protection des Données, a été adopté au niveau européen par le parlement européen le 14 avril 2016. Pour se conformer à ce règlement, les entreprises devront recueillir un consentement éclairé des utilisateurs pour enregistrer, stocker et/ou vendre leurs données personnelles. Ils peuvent à tout moment accéder à leurs données et les modifier, exercer un droit de refus, tout comme un droit à l'oubli (44).

Ainsi le RGPD reprend un grand nombre des principes définissant le « *Privacy by design* », qui est une mesure préventive apparue dans les années 90 aux Etats-Unis, dont l'idée est d'imposer que chaque nouvelle technologie destinée à traiter les données personnelles doit être conçue de manière à offrir un haut niveau de protection des données. Pour les entreprises, mettre en place ce principe de Privacy by Design et respecter le RGPD permet de répondre aux problématiques principales du Big Data comme la fuite massive de données personnelles notamment à cause de la collecte automatisée de ces dernières (45).

## II.5 Impact dans le domaine de la santé

### II.5.1 Réduction des coûts

Le Big Data contribue grandement à la mutation du secteur de la santé, et devrait permettre de réduire les coûts pour la société en permettant une prise en charge personnalisée à chaque patient, ce qui permettra d'optimiser son parcours de soin et éviter la dépense de ressources inutiles.

Les laboratoires devraient également réduire leurs dépenses pour mettre au point des traitements dits « 2.0 » grâce à l'utilisation conjointe de dispositifs médicaux connectés et de services web pendant les phases de développement et d'essais cliniques (16). C'est l'avenir de la pharmacie (décrit en IV.2.4).

### II.5.2 Une donnée de meilleure qualité

En triant les données avec des algorithmes, l'immense quantité de données brutes disponibles peut devenir utilisable pour les laboratoires, dans le cadre de développement de nouvelles thérapies.

C'est pour cette raison que l'on assiste depuis quelques années à de nombreux rapprochements entre laboratoires pharmaceutiques « classiques » ou établissements de santé, et sociétés du numérique, que ce soit les GAFAM ou des startups, à l'image d'Apple et Novartis qui s'associe sur un essai clinique sur la

sclérose en plaques via le kit de développement ResearchKit®, ou encore le partenariat entre IBM Watson Health avec le MIT et le Memorial Sloan Kettering Cancer Center (46).

### II.5.3 Le big data moteur de la recherche

Trois ingénieurs de Microsoft ont publié The Fourth Paradigm, dans lequel ils démontrent que les données stimulent la recherche de manière croissante (47).

Pour Antoine Flahault, docteur en médecine et en biomathématique qui travaille sur l'axe « big data et santé » au Centre Virchow-Villermé, le secteur de la santé a subi de profondes mutations depuis l'émergence du big data.

« Auparavant, il était indispensable de consulter un ensemble de travaux avant d'émettre une hypothèse, de produire des données et enfin de les analyser. Aujourd'hui, la logique s'est inversée : nous cherchons à donner un sens à un gigantesque ensemble de données, mais également à tester cette signification. Le paradigme évolue et une prise de conscience s'impose de la part des chercheurs : celle de comprendre les nouvelles problématiques issues du big data et ses différents aspects. L'analyse de ces données est complexe, elle requiert des compétences spécifiques. Même les ensembles limités de données impliquent une analyse statistique pertinente sans laquelle aucun résultat cohérent n'est possible. Nous ne sommes qu'au début du processus. » (48)

Un exemple dans la recherche médicale illustre l'apport du big data en santé : pour mieux traquer les diffusions d'agents pathogènes sur le territoire, l'institut Pasteur a développé un programme PIBnet qui consiste à recueillir les génomes de souches de bactéries, virus, ou champignons et systématiquement les séquencer et les conserver sur une plateforme mutualisée avec 14 centres nationaux de références. Ainsi cette base de données permet aux scientifiques de comparer les pathogènes sur le sol français avec ceux d'autres pays, permettant de mieux anticiper la diffusion d'agents pathogènes (49).

Ainsi le big data a de nombreux champs d'action pour la recherche : la prise en charge des maladies, la prédiction des épidémies, l'amélioration la pharmacovigilance…

Pour être pertinentes à l'usage, les données du big data doivent être exploitées par des algorithmes. C'est le domaine de l'intelligence artificielle que nous abordons dans le chapitre suivant.

# III. L'INTELLIGENCE ARTIFICIELLE : L'EXPLOITATION DE L'INFORMATION

## III.1 L'Intelligence artificielle (IA)

### III.1.1 Définition

Le but de l'Intelligence Artificielle (IA) est de concevoir des systèmes capables de reproduire le comportement de l'humain dans ses activités de raisonnement. L'IA se fixe comme but la modélisation de l'intelligence prise comme phénomène (de même que la physique, la chimie ou la biologie qui ont pour but de modéliser d'autres phénomènes).

« Dans 3 à 8 ans nous aurons une machine avec l'intelligence générale d'un être humain ordinaire » … pouvait-on lire dans Life, en 1970. Aujourd'hui en 2021 nous en sommes encore loin, bien que des progrès considérables aient été faits depuis l'émergence de cette nouvelle science.

### III.1.2 Historique

#### *III.1.2.1 Les débuts de l'intelligence artificielle : 1943 – 1956*

Dans les années 50, Arthur Samuel, un Américain pionnier dans le domaine du jeu vidéo, invente un logiciel de jeu de dame capable de tenir tête à un joueur de bon niveau et est le premier à inclure un autoapprentissage (50). Les programmes de jeux vidéo ont souvent le même rôle dans la recherche en intelligence artificielle que les drosophiles ont dans la recherche en génétique : les drosophiles en génétique sont peu chères et se reproduisent vite, tandis que les jeux sont pratiques pour l'intelligence artificielle car il rend facile la comparaison entre les performances d'un ordinateur et celles d'un humain. C'est d'ailleurs souvent le mètre étalon pour mesurer les avancées dans le domaine, avec par exemple l'entreprise britannique DeepMind (rachetée par Google) et son programme AlphaGo qui a battu les meilleurs joueurs de go[18] en 2015 (52).

[18] Fin 2017, les équipe de Google développent AlphaZero, une nouvelle version d'AlphaGo plus généraliste, qui a été adapté pour jouer aux échecs et au shogi (une sorte d'échecs japonais) et est devenu quasi-imbattable en

À la suite de cette avancée, l'enthousiasme gagne les milieux scientifiques.

En plein dans la révolution de la cybernétique, la « sciences des systèmes complexes » qui a donné naissance aux sciences cognitives d'aujourd'hui, le mathématicien et cryptologue Alan Turing imagine en 1950 un test visant non pas à déterminer si une machine peut penser, mais si une machine peut *imiter* un humain (53). Il imagine ainsi un test où une personne engage une discussion avec 2 interlocuteurs, l'un étant un humain et l'autre une machine. Si la personne ne sait pas dire qui est la machine, alors le logiciel a réussi le test.

En 1956 la conférence de Dartmouth regroupe chercheurs et pionniers dans plusieurs disciplines alors balbutiantes : la cybernétique, le traitement complexe de l'information, les réseaux neuronaux formels, les modèles de prises de décisions...

La notion d'« Intelligence Artificielle » voit le jour – suggérée par John McCarthy, l'un des instigateurs du projet- et devient un domaine de recherche à part entière (54).

L'autre fondateur, Marvin Minsky, avait en 1951 construit le SNARC, la première machine à réseau neuronal, avec un réseau de 40 neurones artificiels simulant le cerveau d'un rat.

#### *III.1.2.2 L'âge d'or : 1956 – 1974*

Les programmes développés pendant ces années font des progrès remarquables pour la communauté scientifique. En effet ils peuvent résoudre des problèmes algébriques, démontrer des théorèmes géométriques ou apprendre à parler anglais.

Les financements affluent, dont celui de l'agence DARPA.

un temps encore jamais atteint. Fin 2017, AlphaZero a balayé Stockfish, champion du monde des machines d'échecs, dans un match de 100 parties (28 gains, 72 nulles, 0 défaites). Demis Hassabis (CEO de Deepmind) estime qu'AlphaZero pourra aussi être utilisé pour aider à la découverte de nouveaux médicaments ou des matériaux aux propriétés particulières. (51)

### *III.1.2.3 Le premier hiver de l'IA : 1974 - 1980*

Mais en 1970, l'intelligence artificielle n'a plus le vent en poupe : les chercheurs sont confrontés à des problèmes qu'ils ne comprennent pas.

Devant tant de promesses, les attentes trop grandes se sont transformés en des déceptions du même ordre, et les investissements des institutions qui ont investi dans la recherche en IA (DARPA , NRC, Conseil américain de la recherche, gouvernement britannique) se tarissent.

Le connexionnisme, approche scientifique qui stipule que les phénomènes mentaux peuvent être décrits à l'aide de réseaux d'unités simples interconnectées, est mis à l'écart après la critique de Minsky.

Cette désillusion s'explique par les capacités matérielles et logicielles de l'époque qui étaient trop faibles. En effet, les logiciels créés ont des performances très limitées, même pour gérer des problèmes simples, voire simplistes.

D'autre part, la puissance de calcul des ordinateurs de l'époque n'est tout simplement pas au niveau.

A cela s'ajoute des critiques, venant des institutions finançant les projets mais aussi des chercheurs universitaires eux-mêmes. Hans Moravec accuse ses collègues et leurs prédictions irréalistes et beaucoup trop optimistes. Ainsi en 1965, H. Simon déclarait « Des machines seront capables, d'ici vingt ans, de faire tout travail que l'homme peut faire », ou en 67, Minsky : « dans une génération […] le problème de la création d'une ''intelligence artificielle'' [sera] en grande partie résolue. »

### *III.1.2.4 Le boom : 1980 – 1987*

Les années 80 voient arriver des programmes d'IA appelés « système experts ». Un système expert est un programme qui fonctionne dans un cadre de connaissance bien précis, lui permettant de ne pas avoir besoin d'une « culture générale » pour donner des résultats.

C'est la première fois que l'IA se révèle utile.

Parallèlement en 1981, le ministère japonais de l'Economie, du Commerce et de l'Industrie investit massivement pour développer les ordinateurs de 5ème génération.

Ces ordinateurs sont construits dans le but de pouvoir créer des programmes qui puissent tenir des conversations, traduire, interpréter des images et raisonner comme des êtres humains.

Cette initiative fait des émules à l'étranger et bientôt le Royaume Uni et les USA démarrent des projets similaires.

Le dernier maillon de la nouvelle dynamique autour de l'AI est le regain d'intérêt pour le connexionnisme. En effet, à travers les travaux de John Hopfield et David Rumelahrt sur les réseaux neuronaux, on découvre que certaines structures en réseau sont efficaces pour apprendre et traiter une information.

Ces réseaux neuronaux connaitront le succès dans les années 90 et seront utilisés dans la reconnaissance optique de caractères et la reconnaissance vocale (55).

#### *III.1.2.5 Le second hiver de l'IA : 1987 – 1993*

Comme pour la première crise dans les années 70, l'excitation autour des systèmes experts est trop grande et la déception des investisseurs et des agences gouvernementales se fait ressentir par rapport aux résultats obtenus.

De plus, la micro-informatique s'est finalement améliorée et la puissance de calcul des ordinateurs de bureau a dépassé celle des meilleures machines spécialisées en IA, qui de surcroit ont un coût de maintenance élevé et n'ont pas de capacité d'évolution.

Entre temps à la fin des années 80, les approches dans la recherche en IA changent de paradigme, recentrée sur la robotique.

#### *III.1.2.6 Les succès de l'IA : 1993 – 2011*

La technologie de l'intelligence artificielle réalise enfin ses plus vieux objectifs, 50 ans après sa naissance. Nous pouvons à la fois l'expliquer par la puissance des processeurs qui a atteint le niveau requis mais aussi par la réorganisation de la recherche sur des problèmes bien précis, avec une très haute rigueur scientifique.

## III.2 L'IA en 2020 : le *deep learning* principalement

Depuis le milieu des années 2000, l'IA resuscite de l'intérêt. Les cartes graphiques sont au niveau (voir III.4) pour réaliser les calculs complexes nécessaires. De plus il est possible de réaliser un pré-apprentissage des réseaux de neurones profonds par des méthodes non supervisées, qui permet de réduire les difficultés d'apprentissage qu'ont rencontré les anciens algorithmes sur réseaux profonds. C'est le *deep learning*, sous-domaine du *machine learning*, représentant principal de l'IA en 2020.

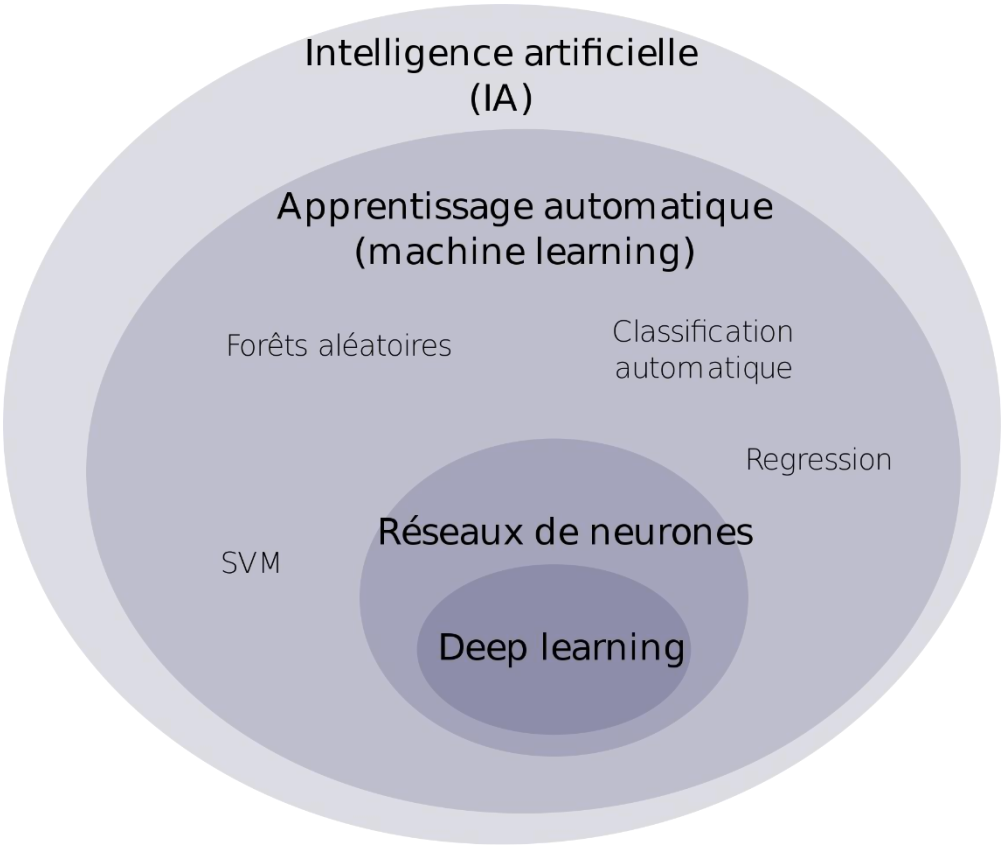



*Figure 16: Les différents champs d'application que regroupe le domaine de l'Intelligence Artificielle*

### III.2.1 Apprentissage automatique, ou *machine learning*

Le *machine learning* au sens large du terme vise à rendre capable un ordinateur d'apprendre à résoudre un problème sans que chaque règle de décision pour sa résolution ait été programmée explicitement. La machine va donc utiliser des comportements modélisés grâce à un apprentissage basé sur des données.

L'apprentissage automatique repose essentiellement sur des notions de calculs mathématiques, et notamment la théorie des probabilités et l'algèbre linéaire.

La théorie des probabilités est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude, et le fondement mathématique des statistiques (56).

L'algèbre linéaire est une branche des mathématiques qui s'intéresse à l'étude des espace vectoriels (ou linéaires), de leurs éléments les vecteurs, des transformations linéaires et des systèmes d'équations linéaires (les matrices). Elle est centrale dans presque tous les domaines mathématiques, en géométrie ou en analyse fonctionnelle, et est utilisée dans la plupart des sciences (naturelles, sociales) et des domaines de l'ingénierie (57).

Quelques applications du machine learning :

- Classifier les données
- Attribuer des valeurs aux données
- Procéder à du clustering[19]
- Permettre un filtrage collaboratif

Autrement dit l'apprentissage automatique a de nombreux intérêts dans l'aide à la décision : trier des données, segmenter une base de données, automatiser l'attribution d'une valeur, proposer des recommandations de manière dynamique, etc.

L'intérêt réside également dans la taille de la base de données qui n'a pas besoin d'être démesurée.

[19] Le clustering est une méthode d'analyse statistique utilisée pour organiser des données brutes en groupes homogènes.

Cette méthode permettra d'obtenir des résultats « classiques » tels que des données numériques, des probabilités ou une classification.

Pour pouvoir modéliser un problème à résoudre par un algorithme de machine learning, il faut d'abord traiter les données, les transformer pour les rendre exploitables et pertinentes afin de découper par étapes le problème pour l'algorithme : c'est le *feature engineering*, qui nécessite l'intervention manuelle d'un expert métier le plus souvent et qui peut nécessiter du temps.

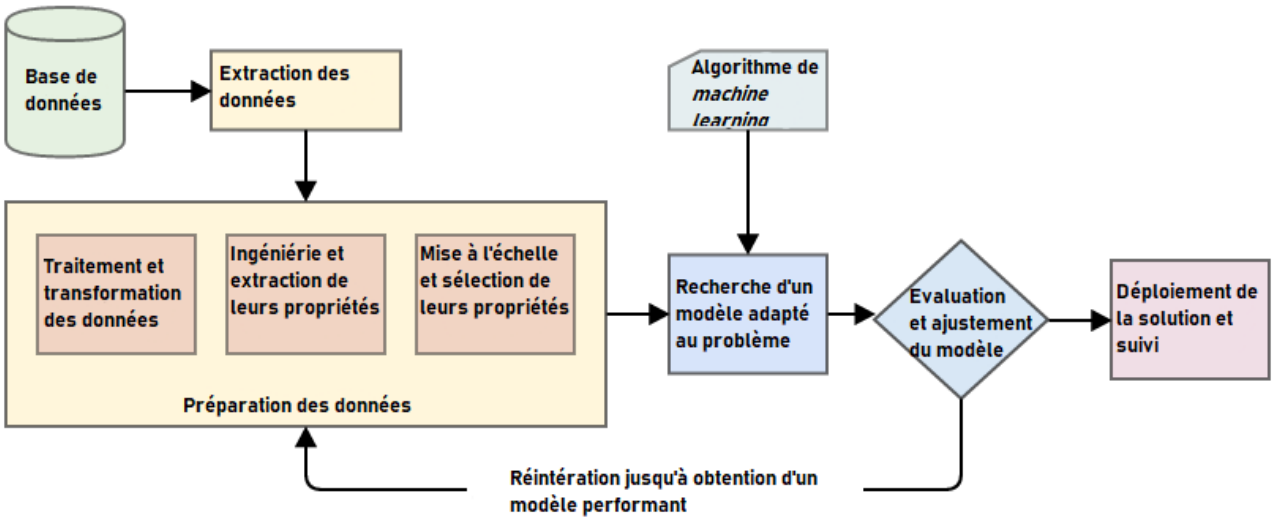



*Figure 17: Processus pour l'obtention d'un modèle en machine learning (traduit d'après Practical Machine Learning with Python, Apress/Springer)*

Ainsi pour obtenir un algorithme performant de *machine learning*, il faut suivre 5 étapes :

- Collecter les données : à partir d'un fichier, d'un capteur, d'un objet connecté etc…
- Prétraiter les données : une étape importante. Cela consiste à « nettoyer » les données brutes pour fabriquer une base de données propre. Les prétraitements sont la conversion des données, l'ignorance ou le remplissage de données manquantes, l'exclusion de données trop éloignées…
- Chercher le modèle qui convient le mieux au type de données : savoir si c'est un problème de classification pour classer des données ponctuelles, un

problème de régression pour des données continues, ou un modèle de *clustering* (association), pour associer des données en groupes.

- Entrainer et tester le modèle avec une base de données d'apprentissage. IL est à noter que la base pour l'apprentissage ne doit pas être utiliser pour évaluer ensuite l'algorithme.
- Evaluer l'algorithme : voir si les résultats répondent au problème.

Dans la phase de recherche du modèle, le modèle dépendra de si l'apprentissage est supervisé ou non supervisé.

Dans un apprentissage supervisé, nous présentons à l'IA des données qui sont déjà marquées avec le résultat voulu. Les modèles utilisés seront ceux de classification et de régression.

Dans un apprentissage non supervisé, nous présentons à l'IA des données non catégorisées. Le modèle utilisé sera celui du *clustering*.

### III.2.2 Apprentissage profond (deep learning)

Le *deep learning* n'est pas à confondre avec le *machine learning*, bien que cela soit deux concepts d'intelligence artificielle.

Le terme de « *deep learning* » est apparu dans les années 2000 (contrairement à la technologie des réseaux de neurones profonds qu'elle désigne qui est plus ancienne), quand les données exploitables sont devenues exponentielles avec la modernisation de notre société qui a connu sa 3ème révolution industrielle avec le numérique.

Cette technologie fonctionne par bio-mimétisme et reproduit le schéma des réseaux neuronaux. Partant du constat que la principale unité de calcul du cerveau était le neurone, les chercheurs en IA ont essayé de *simuler* le fonctionnement d'un cerveau humain en recréant des « réseaux neuronaux ».

Le *deep learning* s'appuie donc sur des neurones artificiels (ou formels). Un neurone formel est en fait une fonction mathématique qui reçoit une valeur d'entrée (ou *input*), la pondère avec des poids (ou coefficients) pour produire une valeur de sortie (ou *output*). Cette valeur de sortie a une fonction d'activation, à la manière d'un neurone biologique qui active le prochain neurone par courant électrique à travers son axone.

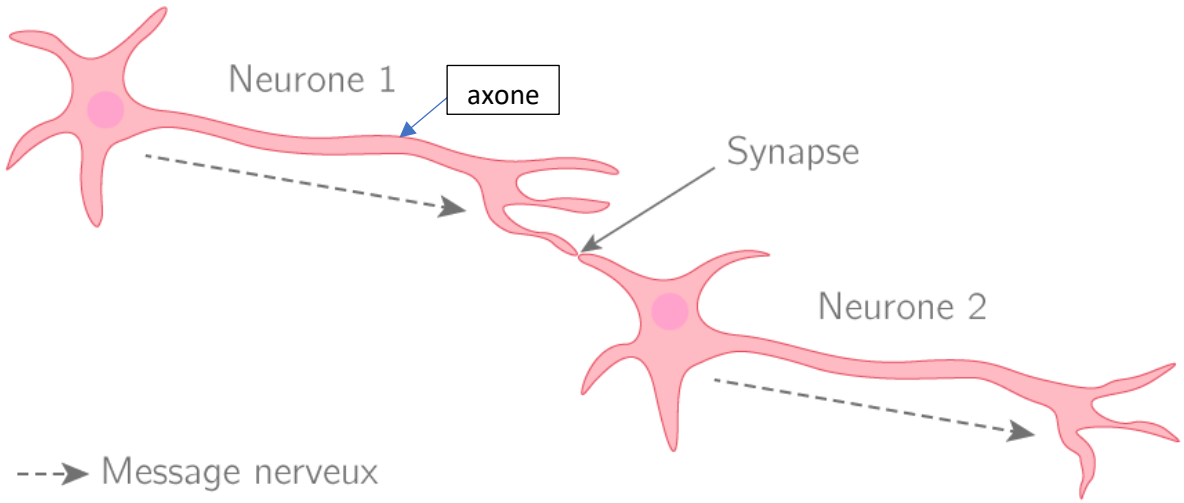



*Figure 18: Transmission de l'information neuronale (D'après Kartable.fr)*

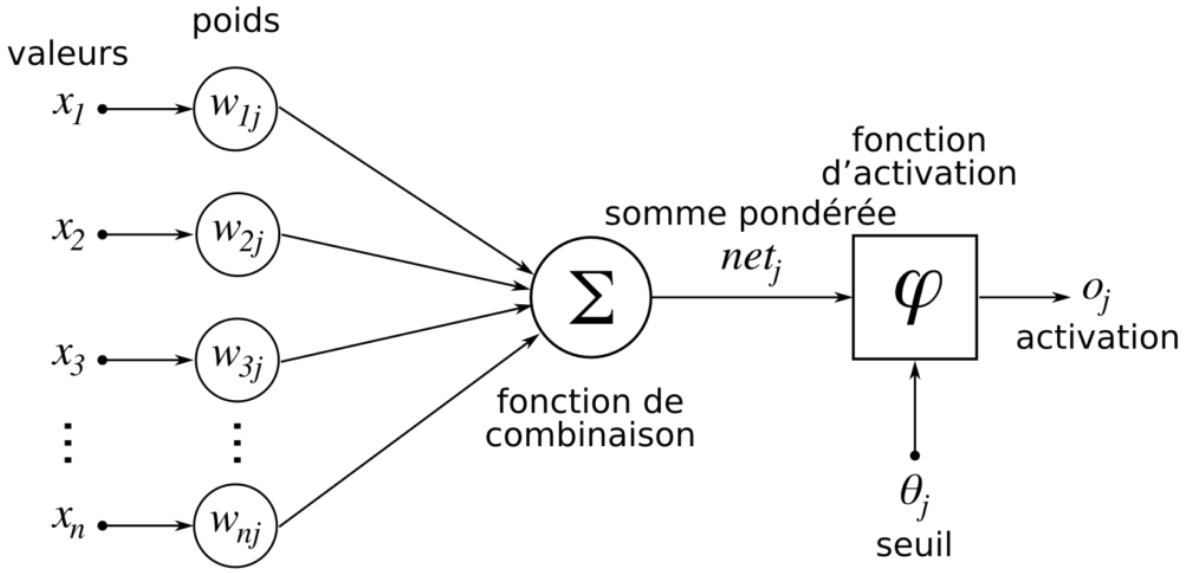



*Figure 19: Schéma d'un neurone artificiel (D'après Christoph Burgmer)*

Le premier modèle mathématique et informatique du neurone biologique est proposé en 1943.

En 1957 est inventé le perceptron, qui est un algorithme utilisant l'idée du neurone formel pour résoudre un problème de classification binaire, c'est-à-dire pour classer des éléments dans 2 classes distinctes. Les coefficients utilisés par le perceptron sont déterminés dans le cadre d'un apprentissage supervisé. L'apprentissage repose sur des principes mathématiques généraux. Le résultat de l'apprentissage est une représentation, une décision ou une transformation. Les chercheurs utilisent plusieurs neurones pour arriver à de meilleurs résultats : c'est un perceptron avec une couche de neurone, et non plus un neurone unique.

Pour résoudre des problèmes plus complexes, on ajoute encore de multiples couches à ce réseau : c'est un perceptron multi-couches.

Pour déterminer les nombreux coefficients des neurones dans un réseau neuronal multi-couches, la méthode de rétropropagation du gradient a été mise au point dans les années 80 (55) et a rendu les réseaux neuronaux utiles dans la reconnaissance de caractères, notamment sur les chèques.

Néanmoins pour des problèmes encore plus complexes, il fallait recourir à des algorithmes ayant encore plus de couches de neurones (des couches profondes), ce qui était difficile à mettre en place à cause d'un manque de puissance de calcul.

D'après Yann Le Cun, la rétropropagation du gradient constitue « la fondation du *deep learning* »(55). C'est une méthode statistique qui permet de corriger les erreurs des neurones en adaptant le poids de la décision que prend chaque neurone, en faisant en sorte que la fonction de coût, c'est-à-dire la différence entre la valeur produite et la valeur attendue, soit la plus petite possible.

C'est une rétropropagation car elle corrige en commençant par la dernière couche des neurones et en progressant vers la première couche.

Ainsi comme vu précédemment, le domaine de l'IA a connu un désintérêt jusqu'au milieu des années 2000.

En 2012, un concours organisé pour classer les 10 millions d'images dans la base de données ImageNet voit la victoire d'une équipe utilisant pour la première fois un réseau de neurones profond, et le terme de *deep learning* émerge (58). Leur algorithme AlexNet utilise 8 couches, ce qui est énorme pour l'époque, et son apprentissage est alimenté par des GPU performants et est permis par un type de réseaux multicouches particuliers : les réseaux convolutifs (cf. III.5) (55).

L'intérêt est relancé et le domaine du *deep learning,* ou apprentissage profond, se développe grandement depuis.

Le *deep learning* est donc un système d'apprentissage autonome, qui permet d'automatiser certaines tâches. Suivant les problèmes à résoudre, on utilisera des approches différentes. L'apprentissage profond se retrouve ainsi dans 3 grands domaines :

- Les tâches cognitives : qu'on utilise pour les traducteurs automatiques, l'analyse de texte, les reconnaissances d'images ou de son.
    - Pour traiter du texte il faut que l'algorithme prenne en compte la temporalité dans le texte, du fait que la compréhension de la phrase dépend de l'ordre des mots dans celle-ci. Le contexte peut être aussi à prendre en compte : il faut donc avoir compris et garder en mémoire les phrases précédentes. Les réseaux qui peuvent traiter ce type de données temporelles sont des « réseaux récurrents ». Ils existent depuis 1997, mais ont bénéficié du gain en puissance des composants informatiques et des nouvelles méthodes d'apprentissage et sont plus largement utilisés depuis 2012.
    - La reconnaissance d'image est détaillée au chapitre III.5.
- Les modèles génératifs : pour reproduire des images, vidéos suivant un style particulier, pour découvrir de nouvelles idées (design), pour augmenter artificiellement la résolution d'une image. Ce sont des GAN (*Generative Adversarial Networks)* ou réseaux antagonistes génératifs. Ces modèles utilisent deux réseaux antagonistes : l'un a pour but de générer des données aussi proches que possible de la réalité, le deuxième doit faire la distinction entre vraies données et données générées par le premier réseau. Ces deux réseaux s'entrainent en parallèle pour générer des données le plus réaliste possible, comme pour coloriser des images ou vidéos qui étaient en noir et blanc, reconstruire des photos partiellement détruites, ou encore fabriquer des visages de toutes pièces, ce qui pourrait être utile pour les illustrateurs ou les studios de jeux vidéo (59).

*Figure 20: Représentation d'un visage d'une personne qui n'existe pas, imaginée par un GAN (d'après thispersondoesnotexist.com)*

- L'interaction client ou *next best action* : dans le marketing, pour connaitre les meilleurs choix stratégiques envers un client

Les résultats obtenus ne seront pas forcément des valeurs, mais pourront être des éléments plus complexes comme des mots ou des images. (cf. III.5)

Ces technologies nécessitent une très grande base de données (se chiffrant en millions de points), pour qu'elle puisse apprendre et se perfectionner, et l'avènement du big data y a bien contribué comme évoqué précédemment, tout comme certains langages de programmation qui ont simplifié leur mise en place.

## III.3 Python : un exemple de langage de programmation utilisé pour l'apprentissage automatique

Une façon de développer des programmes utilisant l'IA est de programmer avec Python.

Python est un langage de programmation, qui s'est développé dans la communauté informatique et plus précisément dans le domaine de la science des données, la *data science*. Ce langage est multi-plateforme, et interprété, c'est-à-dire que le code source écrit par les programmeurs, qui est similaire à une langue naturelle avec un alphabet, du vocabulaire, de la syntaxe, est interprété dans un second temps par un interpréteur qui le traduit à la volée en langage machine, binaire et compréhensible par le processeur.

Son intérêt est que sa syntaxe est simple à utiliser et qu'il permet de manipuler des outils puissants.

On utilise des bibliothèques logicielles, telles que Scikit-learn ou Tensorflow, très utilisées pour l'apprentissage automatique et développées par de nombreux programmeurs notamment français (60).

Une bibliothèque logicielle est un ensemble de fonctions (appelées routines), qui sont déjà écrites dans un langage de programmation donné (ici Python), et que les programmeurs peuvent utiliser sans avoir à les réécrire.

Cela nous permet ainsi de pouvoir aborder le sujet de l'intelligence artificielle sans avoir à trop développer sur les éléments mathématiques qui l'animent, tout comme les acteurs du domaine n'ont pas à être d'éminents mathématiciens pour utiliser les algorithmes de *machine learning* et de *deep learning*.

Cette facilité d'emploi a aussi contribué à l'essor de l'intelligence artificielle.

## III.4 L'IA et les avancées matérielles

Toutes ces avancées logicielles, tant sur le plan de traitement des données (big data et IA) que sur le plan de la vitesse d'acheminement des données (internet), sont dépendantes des avancées matérielles. Elles se sont faites sur toutes les parties du PC : l'écran, la tour, l'alimentation, la carte mère, le processeur, la carte graphique, le disque dur, le lecteur/graveur, la mémoire vive (RAM), le son et bien sur les matériaux en eux-mêmes.

En plus de la disponibilité grandissante de larges bases de données, qui peuvent être traitées rapidement en les stockant dans la RAM, le développement de GPU performants (*Graphics Processing Unit*), ou processeurs graphiques, a accéléré la recherche en IA. Initialement conçus pour l'affichage graphique, l'évolution des GPU vers des circuits dédiés au calcul hautement parallèle capable de traiter plusieurs milliers de calculs en parallèle a été bénéfique pour les applications en IA.

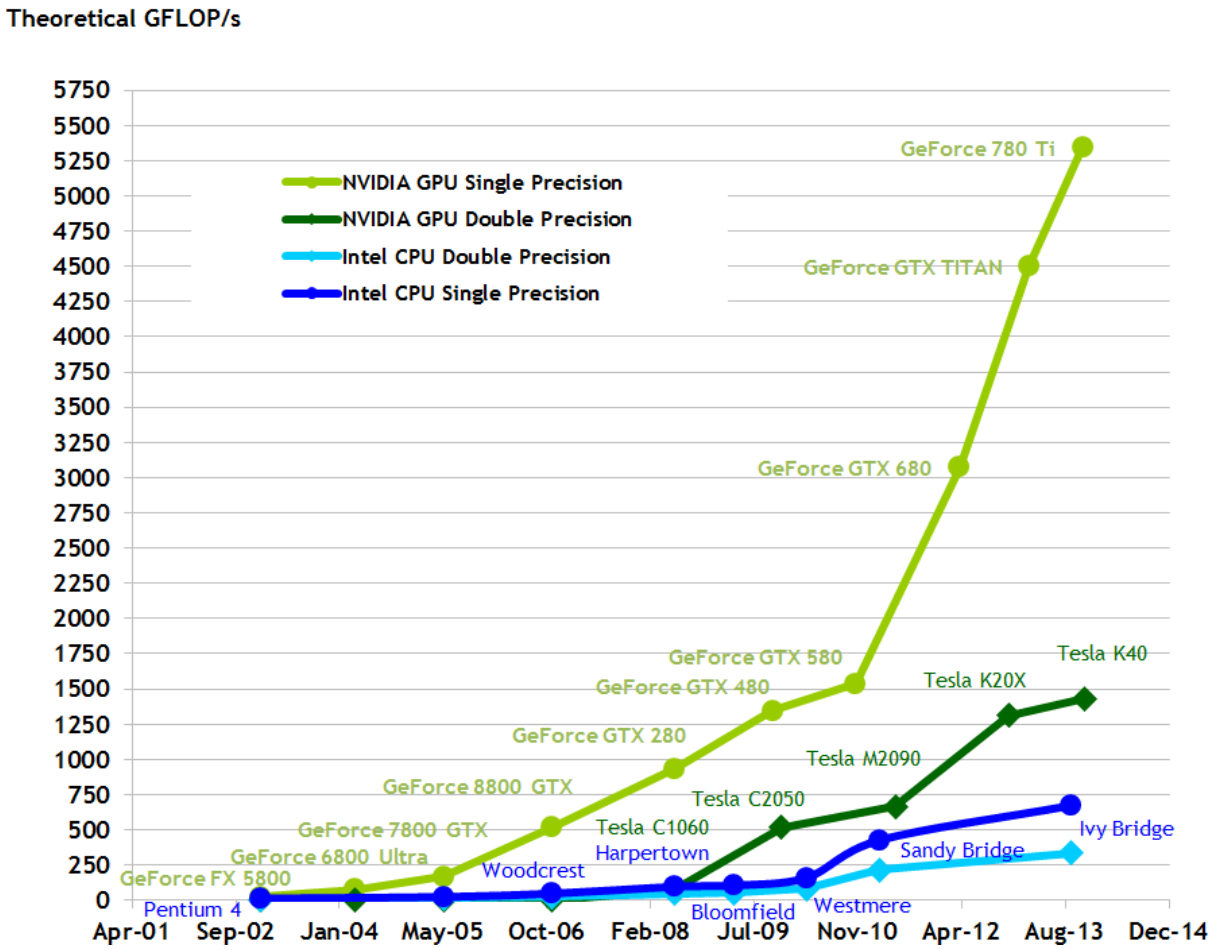



*Figure 21: Evolution des performances en GFlops des CPU Intel et GPU Nvidia (D'après Nvidia)*

Les GPU ont ainsi pu être utilisés comme des CPU pour la fabrication de superordinateurs, destinés à l'HPC, ou High Performance Computing, soit le Calcul Haute Performance qui sont destinés à des domaines très gourmands en calculs comme les simulations physiques, les prévisions météorologiques, les modélisations moléculaires, la cryptoanalyse, et les algorithmes de *deep learning*. (cf. III.2.2)

L'informatique quantique est une des possibles évolutions pour l'amélioration de la vitesse de traitement des algorithmes d'intelligence artificielle. Les ingénieurs de Google ont développé un ordinateur qui utilise des *qubits* (pour bits quantiques, différents des bits classiques), avec un processeur quantique (nommé Sycamore) capable de réaliser en 200 secondes un processus informatique qui aurait été accompli en 10 000 ans sur un ordinateur classique (61).

En effet à l'inverse des bits qui prennent une valeur fixe de 0 ou 1, les qubits possèdent deux états de bases notés $|0\rangle$ et $|1\rangle$, qui sont en superposition quantique linéaire[20], ce qui apporte une information qualitativement différente et quantitativement plus grande que celle d'un bit. L'état d'un qubit est donc une superposition quantique linéaire de ses deux états de base et une fois que plusieurs qubits sont associés dans un processus, ils sont plus rapides pour réaliser des calculs en parallèle en accomplissant moins de cycles que dans l'informatique classique, ce qui est un atout important pour l'intelligence artificielle, dont les algorithmes reposent sur des traitements parallèles.

Néanmoins la technologie en est encore à ses débuts, et des architectures de machines plus conventionnelles ont déjà réussi à dépasser les performances quantiques de la machine de Google (62)(63), qui n'exploite pas encore tout son potentiel.

[20] Et qui s'écrit comme la combinaison $\alpha \times |0\rangle + \beta \times |1\rangle$ , où α et β sont des coefficients complexes pouvant prendre toutes les valeurs possibles à condition de respecter la relation de normalisation (qui assure que le qubit est entièrement présent) : $|\alpha|^2 + |\beta|^2 = 1$.

## III.5 IA, vision par ordinateur et santé

L'analyse d'image est un champ d'application du *deep learning* qui consiste à reconnaitre les éléments sur une image, ce qui est difficile pour un ordinateur qui ne représente une image que sous forme de valeurs associées à chaque pixel la formant.

Son application aux images de pathologie constitue une « avancée importante vers la fiabilité et la reproductibilité des analyses biologiques et médicales » d'après Arnaud Abreu, Médecin infectiologue (64).

Son développement est en lien avec l'essor important de la vision par ordinateur, qui a connu une rupture de paradigme en 2012 à la suite du concours d'ImageNet et de l'arrivée des réseaux convolutifs évoquée précédemment (55).

Comme l'indique le professeur Yoshua Bengio, professeur d'informatique à l'Université de Montréal et pionner dans les développements des méthodes d'apprentissage profond, cet essor est dû à deux facteurs essentiels : l'augmentation d'un facteur 10 de la puissance de calcul des ordinateurs grâce aux processeurs dédiés au traitement d'image, ce qui a permis d'entraîner des réseaux plus grands en un temps raisonnable, et l'accès à d'énormes bases de données étiquetées, avec lesquelles les algorithmes d'apprentissage ont pu s'exercer à reconnaitre un « chat » par exemple.

### III.5.1 Principes de fonctionnement des réseaux convolutifs

Les réseaux de neurones qui analysent les images sont appelés réseaux convolutifs puisqu'ils utilisent des couches de neurones particulières appelées couches de convolutions.

Les réseaux convolutifs ont pour modèle le modèle biologique du cortex visuel chez l'homme détaillé par les Dr Hubel et Dr Wiesel, qui leur ont valu le prix Nobel de physiologie ou médecine. Ils ont inspiré les premiers chercheurs en intelligence artificielle.

Une image, d'un point de vue informatique, n’est qu'une liste de pixels, c'est à dire de nombres entiers (65). Une image d’un chat ou d’un chien est donc une donnée abstraite pour une machine.

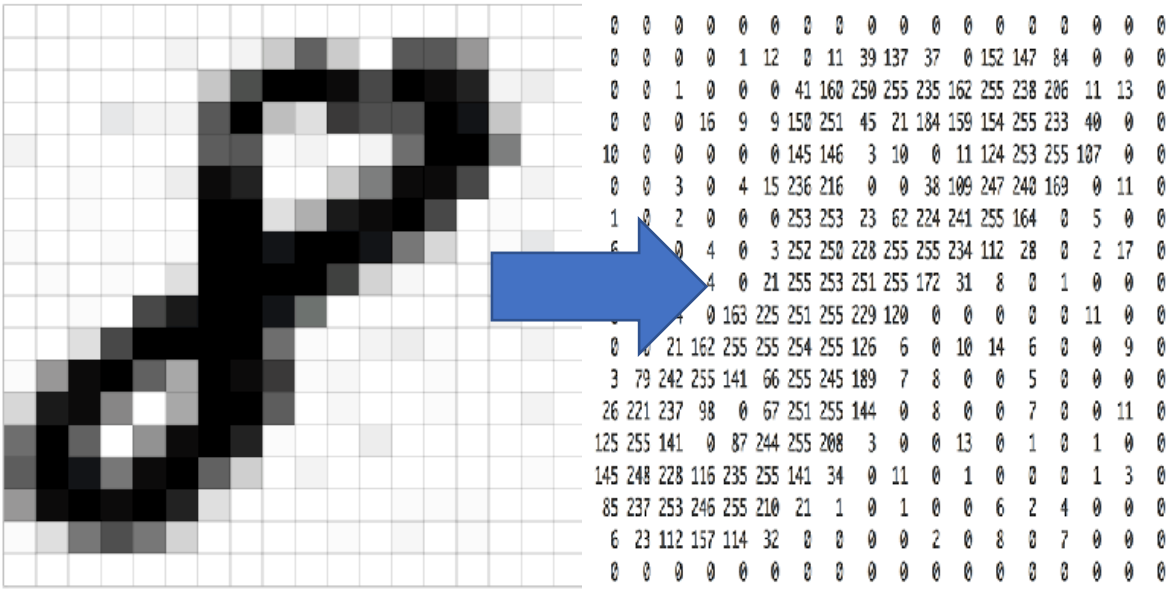



*Figure 22 : Pour un ordinateur cette image d'un "8" en nuances de gris est une grille de nombres représentant la noirceur de chaque pixel*

Le but principal du réseau convolutif est d’extraire des caractéristiques de l’image présentée. Il préserve la relation dans l’espace des pixels de l’image en analysant l’image par petites grilles de taille identique : c’est le principe de la convolution, technique qui filtre l’image en plusieurs carrés plus faciles à analyser.

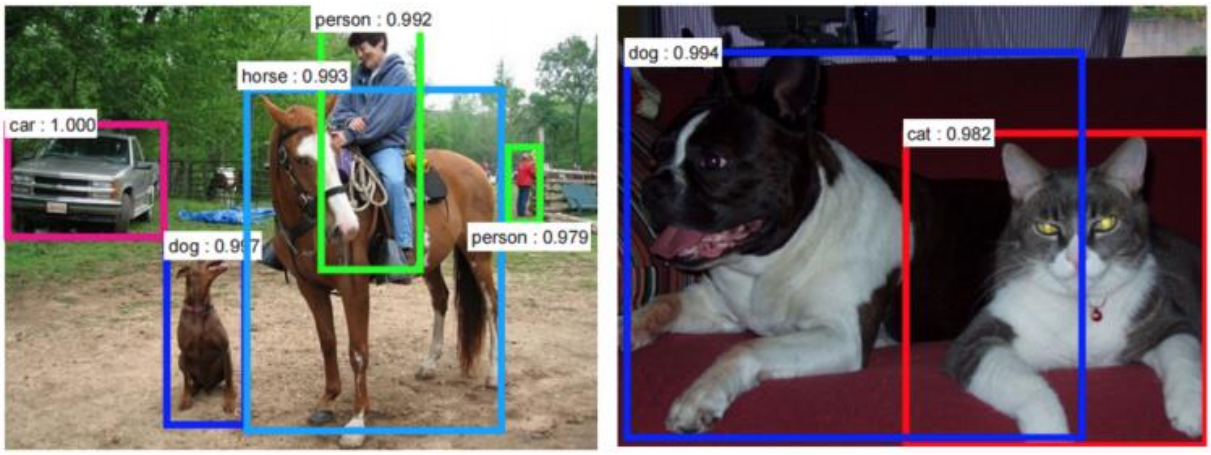



*Figure 23: Un réseau convolutif capable de reconnaitre des objets, humains, animaux (66)*

Prenons un algorithme de reconnaissance d'images, qui est utilisé dans le domaine de la santé et notamment de l'Immunoscore® décrit au chapitre V, qui devra par exemple détecter sur une photo si l'animal est un chat ou un chien.

Schématiquement, nous pouvons le représenter comme plusieurs couches de neurones qui s'activeront les unes après les autres : la première couche de neurones identifiera les pixels de la photo, la seconde identifiera les formes, la troisième les couleurs… Cela peut comporter des millions de couches.

Finalement, la dernière couche comportera 3 neurones : l'un s'activera s'il s'agit d'un chien, l'autre s'il s'agit d'un chat, et le troisième s'il ne s'agit d'aucun des deux.

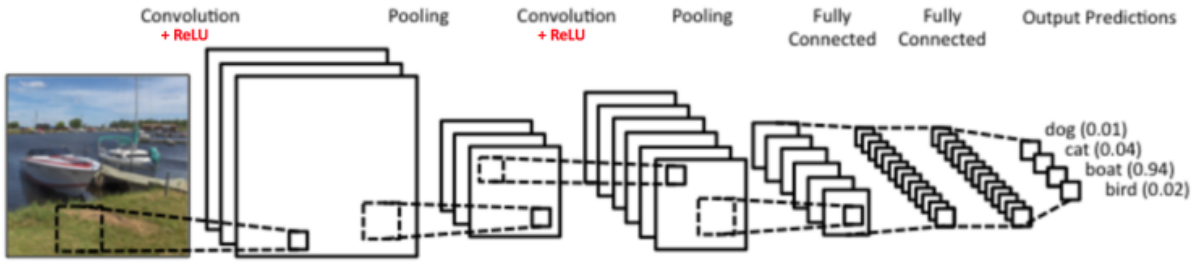



*Figure 24: Déroulement d'une analyse d'une image par un réseau neuronal convolutif*

Plus précisément, les réseaux de neurones convolutifs comportent 4 types de couches, qui peuvent être présentes plusieurs fois à travers l'algorithme :

- La couche de convolution : son but est de « scanner » l'image pour extraire les caractéristiques de l'image (ou *features* en anglais). Elle fait « glisser » une fenêtre sur l'image et calcule le produit de convolution entre la *feature* et chaque portion de l'image balayée. Plus ce produit est grand et plus la *feature* est estimée présente sur la portion de l'image. On obtient une carte d'activation, ou *feature map*, qui nous indique où sont les *features* dans l'image : plus la valeur est élevée et plus l'endroit correspondant dans l'image ressemble à la *feature*.

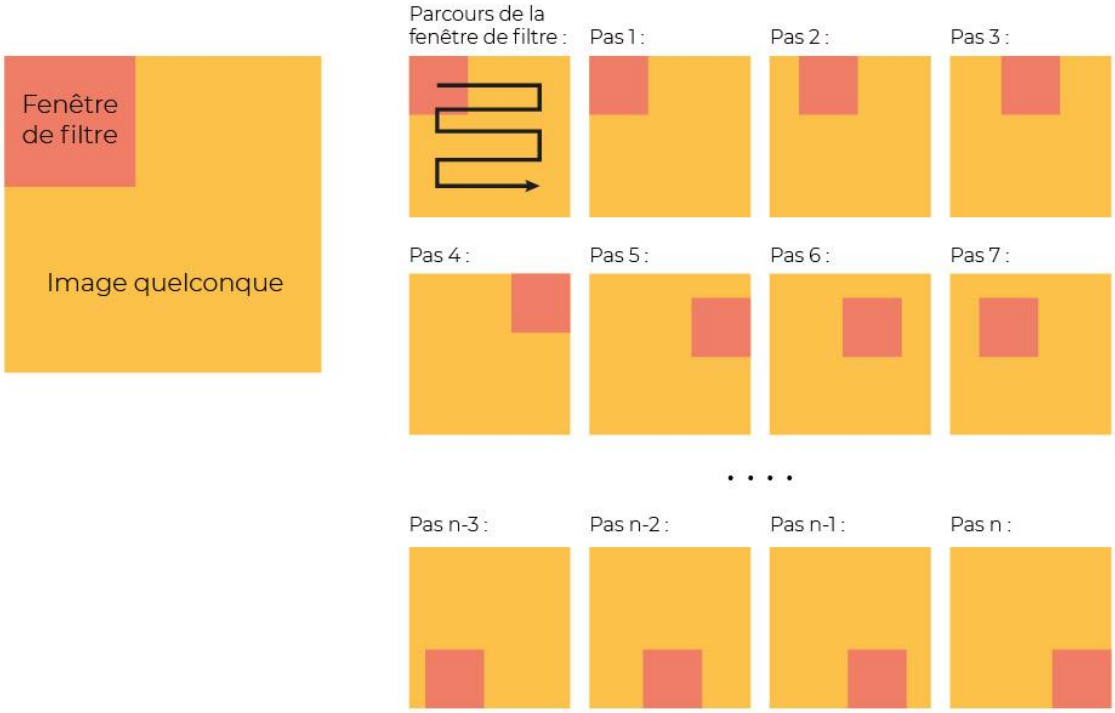



*Figure 25: Le principe de convolution (d'après OpenClassRooms.com)*

L'humain n'intervient pas pour définir quelle caractéristique discriminante choisir, c'est ce qui fait la force des réseaux convolutifs. Les poids seront ainsi déterminés par apprentissage : la rétropropagation de gradient ajuste les poids pour que les neurones des différentes couches détectent ce qui est important.

- La couche de *pooling* : elle reçoit en entrée plusieurs *feature maps*, et applique pour chacune une opération de *pooling*¸ c'est-à-dire une opération qui consiste à réduire la taille des images tout en préservant leurs caractéristiques importantes. Cela permet de réduire le nombre de paramètres et de calculs dans le réseau, ce qui améliore son efficacité et évite le phénomène de sur-apprentissage (décrit plus bas). Ainsi la couche de *pooling* rend l'algorithme moins sensible à la position des *features* : par exemple l'orientation des oreilles d'un chien ne devrait pas provoquer un changement radical dans la classification de l'image.

- La couche de correction ReLU *(Rectified Linear Units)*

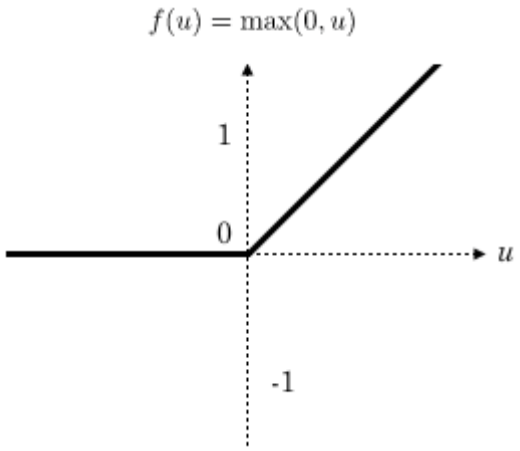



*Figure 26: Allure de la fonction ReLU*

La fonction ReLU est une fonction réelle non-linéaire définie par

$$ReLU\ (x) = max\ (0,x)$$

Comme nous pouvons le constater sur la figure 26, la fonction remplace toutes les valeurs négatives reçues en entrée par des zéros. Elle a donc le rôle de fonction d'activation, à la manière du signal de potentiel d'action d'un neurone biologique qui doit être dépassé pour que le neurone s'active. Il existe aussi les fonctions sigmoïde et tangente qui sont utilisées dans d'autres systèmes, mais la fonction ReLU est celle la plus utilisée dans les algorithmes modernes. Ainsi la *feature map* qui contient des valeurs positives et négatives (car les poids peuvent être négatifs) passe par la couche ReLu qui laisse inchangée les valeurs positives mais change à 0 les valeurs négatives. Cette transformation permet au système de détecter des motifs sur l'image.

- La couche *fully-connected*: elle est toujours la dernière couche d'un réseau de neurones, qu'il soit convolutif ou non, et permet de classifier l'image à l'entrée du réseau. Cette couche est nommée « complètement connectée » car chaque valeur d'entrée est connectée à une valeur en sortie. Chaque valeur « vote » pour une classe, et son vote a plus ou moins de poids.
  Le réseau de neurones convolutif apprend les valeurs des poids de la même manière qu'il apprend les caractéristiques discriminantes de la couche de convolution : lors de phase d'entraînement, par rétropropagation du gradient. Cette couche permettra de déterminer ce qui se trouve sur l'image, par exemple

dans la figure 23 l'algorithme estime à 99,4% qu'il y a un chien à gauche et à 98,2% un chat à droite.

### III.5.2 Exemples d'application

Comme évoqué en introduction, on doit cette découverte à Yann Le Cun, qui a publié un article fondateur en 1998 sur les réseaux convolutifs et les réseaux neuronaux.(67) Développé dans le milieu des années 80, la technologie a d'abord été implantée dans les outils de reconnaissance de chèque, avant d'être laissé de côté pendant 20 ans.

En 2020 le secteur de la santé a grandement bénéficié des avancées dans la *computer vision*.

Il a par exemple été démontré par Wu *et al.* l'efficacité d'un réseau convolutif pour détecter les tumeurs bénignes et malignes dans les mammographies. Celui-ci a été entrainé sur 1 million d'images de radiographie, et s'est montré plus fiable que les praticiens radiologues. Néanmoins, la meilleure performance était obtenue quand le système était combiné à l'expertise des radiologues (68). En effet le système permet d'éliminer les cas simples, de réduire les coûts et délais de diagnostic tout en permettant au radiologue de se concentrer sur les cas complexes, à la manière du fonctionnement de l'Immunoscore détaillé au chapitre V.

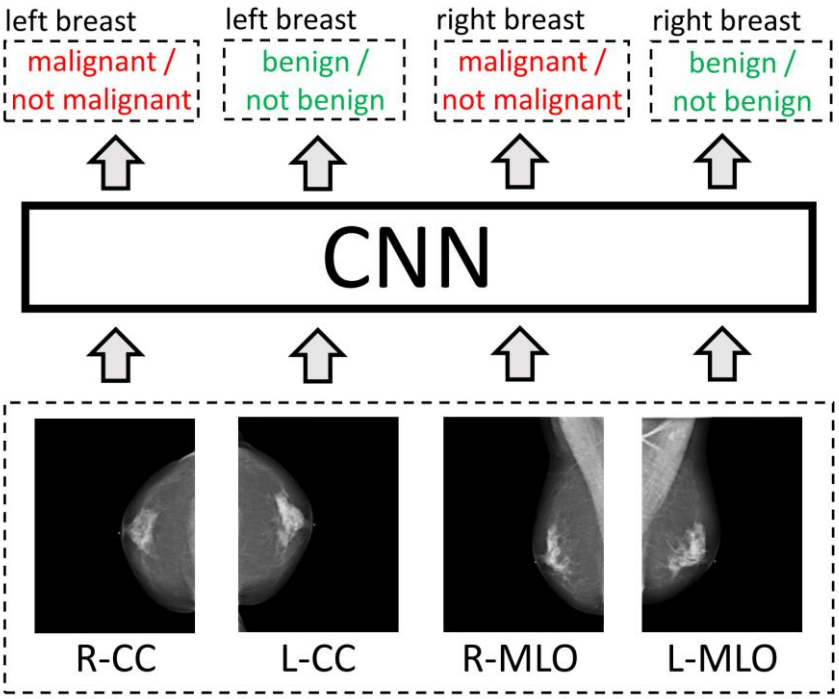



*Figure 27: représentation schématique de l'apprentissage du réseau à partir des radiographies mammaires (d'après Wu et al.)*

Le rapprochement s'opère également dans l'autre sens, depuis les acteurs de la Silicon Valley vers la santé, à travers l'internet mobile. Lors de sa conférence annuelle en 2021, Google a présenté son outil logiciel pour reconnaitre les affections cutanées que l'on prend en photo avec son smartphone.(69) Karen DeSalvo, responsable santé chez Google Health a expliqué que Google recevait 10 milliards de requêtes annuelles sur les problèmes de peaux. Leur équipe de développeurs a entrainé le modèle sur des millions d'images de peaux pathologiques, des milliers de peaux saines, et 65 000 issues du milieu clinique. Le modèle prend en compte les facteurs de l'âge, le type de peau, le sexe. Ils ont ainsi affirmé que l'état correct était identifié dans les 3 premières suggestions dans 84% des cas. Ce modèle reprend les bases de l'étude publiée dans Nature Medicine qui a indiqué que leur outil pouvait identifier 26 affections de la peau aussi bien que les dermatologistes, et même mieux que les médecins généralistes (70).

La société a publié une autre étude montrant l'aide que cela pouvait apporter aux praticiens non spécialistes (71).

La technologie de la computer vision est donc plein d'avenir, notamment dans le secteur de la santé, mais connait également des limites.

## III.6 Limites de la technologie

### III.6.1 Biais de l'IA par son apprentissage

L'étape d'apprentissage d'un modèle d'IA doit être correctement réalisée pour ne pas obtenir des résultats erronés.

Il y a en effet des risques de surapprentissage si le modèle a tendance à trop capturer d'informations, ce qui limitera sa faculté à généraliser des caractéristiques de données et donc de prédiction sur de nouveaux échantillons.

Il peut y avoir également des biais dans l'apprentissage d'un logiciel si notre base de données n'est pas optimale.

Il y a ainsi 2 types de biais par rapport à la base de données : les biais algorithmiques et les biais sociétaux (72).

Les biais algorithmiques sont causés par des bases de données biaisées. Nous pourrons citer l'exemple d'Amazon et Google qui ont reconnu que leur algorithme d'optimisation des photos étaient mal optimisées pour les personnes à peaux foncées (73).

Les biais sociétaux sont causés par les *data scientists* eux même, qui ont des préjugés lorsqu'ils codent leur application, à l'image du projet abandonné de logiciel examinateur de CV développé par Amazon qui discriminait les CV de femmes, sur la base que le logiciel avait été entrainé avec majoritairement des CV d'hommes (74).

Une des solutions pourrait venir d'un enrichissement des bases de données, comme souhaité par le collectif Women in AI (75), et également évoqué au gouvernement français. En effet sur le cas de la mixité et la diversité dans le numérique, le rapport en France de la mission Villani sur l'intelligence artificielle écrit :

« En dépit d'une féminisation lente, mais progressive des filières scientifiques et techniques, le numérique fait figure d'exception : la parité entre les hommes et les femmes est loin d'y être acquise. […] Ce manque de diversité peut conduire les

algorithmes à reproduire des biais cognitifs - souvent inconscients - dans la conception des programmes, l'analyse des données et l'interprétation des résultats. L'un des grands défis de l'IA consiste donc à parvenir à une meilleure représentativité de nos sociétés. » (76)

### III.6.2 Limites de l'automatisation et exploitation humaine

Pour l'apprentissage d'un algorithme, il est nécessaire d'engranger de grandes quantités de données. Si une entreprise a besoin de données de meilleure qualité que celles produites automatiquement par des machines, une des solutions est de faire appel à l'homme directement, ce qui peut conduire à des dérives.

Un des exemples récents est le cas de Figure Eight, qui est une société d'apprentissage automatique et d'intelligence artificielle basée à San Francisco. Elle utilise 213 millions d'humains dans le monde, payés 1 à 10 centimes d'euros la tâche. Ces tâches servent à entrainer des algorithmes d'intelligence artificielle comme l'aider à reconnaitre des piétons dans la rue ou des produits dans un magasin en ligne (77). Interrogés, ceux-ci se sentent exploités mais ne savent parfois pas faire autrement pour gagner leur vie.

Cela peut nous conduire à nous questionner sur les limites de la production de données, pour une technologie qui devrait normalement nous faciliter la vie.

### III.6.3 Menace de l'Homme et des emplois ?

De nombreuses œuvres de fiction influencé par le courant *cyberpunk* imaginent un futur dirigé par les machines, à la manière du film Terminator de James Cameron.

Néanmoins un modèle, aussi complexe soit-il, n'est pas intelligent, et reste une pure fonction mathématique, non un être doué de raison et de conscience.

Le président français Emmanuel Macron rappelait en 2018 lors de la présentation du rapport sur l'intelligence artificielle du mathématicien Cédric Villani, lauréat de la

médaille Fields : « L'IA n'est dotée d'aucun concept. Elle n'a pas de culture. Elle ne comprend rien. »

Du coté des salariés, certains emplois pourraient disparaitre s'ils étaient remplacés par un logiciel, tout comme d'autres ont ou vont être créé dans une dynamique de destruction créatrice (par exemple les analystes de données ou les architectes de la donnée) dans le cadre de la 3ème révolution industrielle du numérique (78).

Dans le domaine de la santé, praticiens et techniciens ne sont pas plus menacés. Erik Brynjolfsson, un économiste du MIT, affirme que la vitesse de pénétration d'une « technologie d'usage général » comme l'intelligence artificielle (et comme avant elle l'informatique, l'électricité, la machine à vapeur…) est limitée par le temps qu'il faut au travailleur pour apprendre à l'exploiter, ce qui peut prendre de 15 à 20 ans.

Ainsi le meilleur moyen de tirer parti de l'IA est d'investir dans l'éducation et de proposer de nouvelles formations, à l'instar de celle proposée à l'université Aix-Marseille décrite au chapitre VI.

### III.6.4 L'IA et les « blackboxes » impénétrables

Le fonctionnement de certaines solutions d'intelligence artificielle ne sont pas totalement comprises, tant dans le milieu scientifique que dans le grand public et agissent de manière opaque, à la manière d'une boite noire, ou « blackbox » qui produirait un résultat sans qu'on en connaisse les mécanismes.

Ces « blackbox » comme l'évoque le cabinet d'audit et de conseil Deloitte dans son étude sur les risques de l'ère numérique, risquent en cas de défaut de conception de conduire à des données et des décisions incorrectes, basés sur des modèles biaisés (79).

Dans le domaine médical, cela peut donner lieu par exemple à des algorithmes recalculant à la baisse des indemnités de couverture sociale sans raison, ou à un manque de confiance dans les outils mis à disposition pour les praticiens, qui ne comprennent pas les tenants et aboutissants dans la prise de décision du logiciel.

Pour éviter ces « blackbox » dans le milieu médical, des guides de bonne utilisation de l'intelligence artificielle en santé pour les développeurs émergent et stipulent que « ce qui peut être expliqué, doit être expliqué », à l'instar du guide produit par la Clinical Decision Support Coalition aux USA, une organisation visant à définir des règles pour les logiciels d'aide à la décision médicale (80).

## III.7 Perspectives d'évolution

D'après Yann Le Cun, l'IA est encore loin d'atteindre les performances d'un cerveau humain (55).

Cela s'explique déjà par la limite de puissance de calculs pour reproduire son fonctionnement. Un cerveau humain contient 86 milliards de neurones et consomme 25 watts. Pour reproduire cette architecture, il faudrait un GPU capable de $1,5 \times 10^{18}$ opérations par seconde, ce qui est encore infaisable : un GPU en 2020 est capable de $10^{13}$ opérations par seconde en consommant 250 watts.

De plus, l'apprentissage doit être encore amélioré.

En effet, les techniques d'apprentissage existantes ont leur faiblesse : l'apprentissage supervisé ne couvre qu'une infime partie des possibilités et l'algorithme peut donc être trompé facilement si nous cherchons à le faire, et l'apprentissage par renforcement, qui consiste à dire à la machine si le résultat produit est juste ou non (par exemple si un robot a bien saisi l'objet dans sa main) n'est pas adaptée aux conditions réelles. En effet si un programme comme AlphaGo a pu *simuler* des milliers de parties en parallèle pour apprendre les coups susceptibles de le mener à la victoire, on ne peut pas par exemple utiliser la même tactique pour entrainer un logiciel de conduite autonome : cela conduirait à accidenter des centaines de milliers de voiture avant que l'algorithme comprenne qu'il faut éviter les piétons, les ravins, les sens interdits… Et les simulateurs ne suffisent pas à produire une réalité parfaite pour les algorithmes.

Yann Le Cun estime que l'IA doit apprendre comme l'homme le fait, par un **apprentissage « autosupervisé** » qui serait l'idéal à atteindre.

Autrement dit elle doit être capable d'emmagasiner un grand nombre de données pour pouvoir obtenir un sens commun, à la manière de l'apprentissage d'un humain dès sa naissance. En effet d'après Emmanuel Dupoux, professeur de sciences cognitives à l'école supérieure, un bébé engrange une grande quantité de données sur le fonctionnement du monde, et cela se fait principalement par observation : par un exemple il apprend la gravité quand il lance plusieurs fois un objet au sol. Ce sens de la gravité s'acquiert vers 9 mois. Avant cette période, un objet qui semble flotter dans l'air ne le surprendra pas.

Pour acquérir ce sens commun, la machine doit disposer d'un modèle du monde réel, pour qu'elle acquiert une capacité d'anticipation et de véritable raisonnement, afin de non seulement savoir ce que sont les choses, mais également le contexte dans le monde de ces choses, ce qui est en 2020 un défi technologique. La marche à suivre des chercheurs en IA est encore incertaine.

Les applications en santé sont nombreuses pour le grand public, à l'instar des sites internet spécialisés en santé (exemple : Doctissimo), ou encore les dispositifs médicaux connectés.

Pour les professionnels, on retrouve cette technologie dans la *digital pathology*.

Nous nous intéresserons ici aux applications concernant les professionnels de santé, et notamment les médecins et l'aide au diagnostic.

# IV. LA SANTE PREDICTIVE DANS LA MEDECINE 4P : UNE UTILISATION DE L'INFORMATION MEDICALE

La santé prédictive, ou médecine prédictive, est une évolution dans la prise en charge du patient, qui a su tirer parti de la récente explosion de la quantité des données médicales disponibles. La médecine prédictive fait partie de la nouvelle vision de la médecine que l'on peut appeler « Médecine 4P ».

Avec le regain d'intérêt de l'industrie pour l'IA, nous pouvons citer beaucoup d'initiatives récentes visant à utiliser nos données de santé pour prédire nos futures pathologies.

D'après la SFMPP, la Société française de Médecine Prédictive et personnalisée, la médecine prédictive « consiste à utiliser des marqueurs, le plus souvent biologiques, pour prévenir, dépister ou traiter les maladies » (81).

## IV.1 Médecine 4P

La santé prédictive, ou médecine prédictive, est une caractéristique de la médecine moderne dite « médecine 4P ».

Ce terme a été décrit par le Dr Leroy Hood en 2011, qui évoque l'évolution d'une médecine réactive à une médecine proactive (82).

Les 4P regroupent :

- La médecine préventive : elle se concentre sur le mieux-être et non sur la maladie. Prenons l'exemple des maladies chroniques qui touchent le monde moderne pouvant être prévenues et repoussés par un mode de vie adapté.
- La médecine personnalisée : elle tient compte du profil génétique et épigénétique de chaque patient mais également de son environnement, pour proposer une prise en charge sur mesure.
- La médecine participative : qui prend en compte la rupture dans la nature de la relation entre praticien et patient, qui devient acteur de sa santé, notamment

par l'accès numérique aux informations scientifiques et au suivi de ses objets connectés. Il est parfois question d'*empowerment*[21].

- La médecine prédictive : elle s'appuie sur des modèles qui permettent d'anticiper les risques à partir d'un profil et de données personnelles. Son but est de mettre en évidence des facteurs prédictifs précoces, en se reposant notamment sur l'analyse prédictive.

Pour réaliser cette médecine 4P, le Dr Leroy Hood estime qu'il y a 2 défis majeurs : la barrière technologique et la barrière sociétale, cette dernière qui concerne patients, médecins et acteurs du médical, étant la plus difficile à lever.

Pour aller plus loin, certains parlent d'une médecine 5P voire 6P, à l'instar du LEEM, une organisation des entreprises du médicament, qui évoque une médecine de la pertinence ou de la preuve, qui doit prouver que le diagnostic aide à soigner le patient, et une médecine des parcours pluriels, qui prévoit un parcours de soin avec des acteurs multidisciplinaires.

## IV.2 La santé prédictive, un concept de la e-santé

La médecine prédictive s'inscrit dans un mouvement plus global de numérisation de la santé.

### IV.2.1 Définitions

Le terme de « e-santé » est apparu en 1999 lors du 7ème congrès international de télémédecine, qui d'après l'OMS désigne *« les services du numérique au service du bien-être de la personne »* et plus spécifiquement *« l'utilisation des outils de production, de transmission, de gestion et de partage d'informations numérisées au bénéfice des pratiques tant médicales que médico-sociales »*.

La médecine préventive est devenue médecine prédictive.

[21] Terme anglais pour « autonomisation ».

### IV.2.2 La e-santé utilisée par les professionnels de la santé

Les systèmes d'information de santé (SIS) ou hospitaliers (SIH) sont des représentants majeurs de la e-santé. Ils permettent les échanges d'informations au niveau informatique entre l'hôpital (ou la clinique) et la médecine de ville, ou entre les services d'un même hôpital. (83) Nous le retrouvons aussi en pharmacie via le Dossier Médical Partagé (DMP), la technologie de la carte vitale, ou encore la déclaration en ligne des résultats des tests antigéniques covid (84).

A côté des professionnels, les usagers et patients, qui sont devenus actifs dans leur prise en charge, génèrent et gèrent aussi leurs propres données et informations médicales, comme un « power user », à travers de nombreux dispositifs connectés : c'est la « m-santé ».

### IV.2.3 La m-santé ou santé mobile utilisée par les patients

La m-santé concerne la santé mobile. L'OMS la définit comme les « pratiques médicales et de santé publique reposant sur des dispositifs mobiles tels que téléphones portables, systèmes de surveillance de patients, assistants numériques personnels et autres appareils sans fil » (85).
Elle comprend d'une part les applications numériques pour smartphones, et d'autre part les objets connectés et les dispositifs médicaux connectés en lien avec la santé.

Concernant les applications, elles sont très utilisées sur smartphone. Pendant la pandémie de covid-19, la société spécialisée dans l'analyse des données mobiles App Annie a observé que « les consommateurs ont passé un nombre record de 113 millions d'heures à utiliser des applis de santé et de fitness pendant la semaine du 22 mars 2020 dans le monde. » (86) Attention toutefois, si certaines applications sont utiles, d'autres n'ont aucun intérêt si ce n'est un cumul brut et « bête » de données, qui ne fera qu'au pire inquiéter l'utilisateur.

Concernant les dispositifs médicaux, ils sont à différencier des simples objets connectés. En effet selon la directive européenne 93/42 CEE :

« Est considéré comme dispositif médical tout instrument, appareil, équipement, logiciel, matière ou autre article, utilisé seul ou en association, y compris le logiciel destiné par le fabricant à être utilisé spécifiquement à des fins diagnostique et/ou thérapeutique, et nécessaire au bon fonctionnement de celui-ci. Le dispositif médical est destiné par le fabricant à être utilisé chez l'homme à des fins de :

- diagnostic, prévention, contrôle, traitement ou d'atténuation d'une maladie ;
- diagnostic, contrôle, traitement, d'atténuation ou de compensation d'une blessure ou d'un handicap ;
- d'étude ou de remplacement ou modification de l'anatomie ou d'un processus physiologique ;
- maîtrise de la conception, et dont l'action principale voulue dans ou sur le corps humain n'est pas obtenue par des moyens pharmacologiques ou immunologiques ni par métabolisme, mais dont la fonction peut être assistée par de tels moyens. »

Il existe de nombreux dispositifs médicaux connectés : stéthoscope, tensiomètre, électrocardiogramme, pilulier, oxymètre, thermomètre… A côté de ceux-là existent des objets connectés *orientés santé*, telles les balances connectées ou les montres connectées, qui bien qu'utiles pour le suivi des patients, ne sont pas reconnus comme des dispositifs médicaux.

En résumé, ces appareils et applications mettent le patient comme producteur majeur de données. Elle est en lien avec le mouvement du *quantified self* prenant de l'ampleur ces dernières années, qui désigne la pratique née en Californie de la « mesure de soi » pour mieux se connaître en mesurant des données relatives à son corps et à ses activités.

### IV.2.4 L'avenir de la pharmacie, « beyond the pill »

Un nouveau moyen pour les industriels de la santé de rester compétitifs et pertinents dans la prise en charge d'une population vieillissante, instruite, et dans un marché où leurs découvertes et brevets, qui tombent dans le domaine public, ne suffisent plus, est d'élargir leur innovation au-delà du médicament lui-même. C'est le modèle du « *beyond the pill* », au-delà du seul traitement médicamenteux.

En effet les enjeux de la santé ont évolué :

- Les maladies chroniques sont la première cause de décès dans le monde d'après l'OMS (87) et nécessite un suivi renforcé.
- La défiance du grand public accentue la pression sur les laboratoires. On pourra citer l'exemple de la mise en examen de Sanofi pour homicides involontaires de nourrissons exposés à son médicament Dépakine (88)
- Le marché de la santé est exigeant et demande un niveau de preuve d'efficacité très poussé pour valider leur AMM.
- Le rendement des médicaments change : Les médicaments « blockbusters » tombent dans le domaine public, les anciens médicaments sont soumis à des limites de remboursement, la R&D coute cher pour des thérapies innovantes.
- L'écosystème de la santé a évolué avec les technologies numériques : du coté du patient, Internet a donné l'accès à des connaissances qui a réduit la fracture de l'information entre médecin et patient, donnant une place d'acteur au patient dans son parcours de soins ; les applications mobiles, objets et dispositifs médicaux connectés sont des nouveaux moyens de générer et de suivre des données de santé.

  Du coté des laboratoires, leurs collaborations avec des acteurs du secteur des NBIC[22] (Google, Microsoft, Amazon, IBM, Samsung, Philips…) engendrent des solutions mêlant logiciel, dispositif médical et médicament.

[22] Nanotechnologies, Biotechnologies, Informatique et sciences Cognitives

Les laboratoires ont été amenés à remettre le patient au centre du parcours de soin, dans une approche dite « *patient centric* », et reléguer le médicament à un élément du patient parmi d'autres, tels que son état de santé, son mode de vie, son environnement social, professionnel et territorial.

Les laboratoires utilisent alors les réseaux sociaux, proposent des formations et des supports d'information à leur patient, du coaching physique ou sur le sommeil, les suivis de traitement à distance par le praticien grâce aux objets connectés, financent des campagnes de prévention…

Plus précisément, l'Internet des Objets (évoqué en I.3) peut révolutionner les process des grands laboratoires, qui basculeront dans un modèle « Pharma IoT »(16) qui fera appel à la numérisation des produits médicaux et concepts associés, en utilisant des dispositifs médicaux connectés et services numériques comme le web et les applications sur smartphones pour le développement des traitements et le suivi des patients. Un bon exemple d'une solution « Pharma IoT » est le capteur connecté utilisé chez les patients atteints de la maladie de Parkinson, qui gérait le traitement du patient en fonction des signaux reçus, et qui a prouvé apporter une meilleure qualité de vie au porteur (89).

## IV.3 Les avancées médicales dans l'immunologie permises par la santé prédictive

En France, Jacques Ruffié, médecin et professeur au collège de France envisageait déjà en 1993 la médecine préventive par une évolution dans la prise en charge immunologique, hématologique et génétique des patients.

Dans les années 2000 les avancées du domaine de l'immunologie tant au niveau des connaissances que de l'accessibilité technique aux informations ont révolutionné les possibilités de bénéfice médical apporté au patient.

Ainsi le profil immunitaire est une donnée clé pour comprendre l'origine des pathologies, comme nous le verrons avec le rôle des lymphocytes dans le cancer du côlon au chapitre V.

Les exemples ne manquent pas. En 2019, Microsoft a développé avec la société Adaptive Biotechnologies basée à Seattle un test sanguin permettant de séquencer l'immunité d'un patient via la reconnaissance de ses lymphocytes T. Grâce aux techniques de machine learning, le logiciel en développement génère une carte des lymphocytes T du patient par rapport aux milliards de profils de lymphocytes déjà enregistrés et peut prédire les maladies potentielles de celui-ci. (90) En mars 2020 pour répondre à la crise sanitaire du COVID-19, ils continuent leur partenariat pour fournir une base de données sur la réponse immunitaire des patients contaminé par le SARS-CoV-2 en *open data* dans le but d'aider les chercheurs et les gouvernements face à ce nouveau virus (91).

D'autres initiatives sont à noter, à l'image de cet outil d'intelligence artificielle mis au point dans un but prédictif par des équipes de l'APHP et de l'Université de Paris. Son nom est iBox et son but est de prédire à long terme la survie d'un greffon rénal.

La maladie rénale chronique touche en effet une personne sur dix dans le monde, et 55% des patients touchés seront traités par dialyse, ce qui représente un coût de plusieurs milliards de dollars annuels en France et aux Etats-Unis. De plus, ceux qui recevront un greffon ont un risque de rejet aussi élevé qu'en 1999.

Pour développer cet outil les équipes ont identifié huit des paramètres associés au risque de perte de greffon dans les dix ans suivant la greffe, grâce aux données cliniques, histologiques, immunologiques et fonctionnelles de 7.500 patients en Europe et aux USA et les ont intégrés dans les paramètres de l'algorithme.

Il est maintenant intégré dans un logiciel d'aide à la décision médicale et facilite les interventions thérapeutiques, la prise de décision et le déroulement des essais cliniques, avec un gain de plus de six années en moyenne dans le développement de médicaments immunosuppresseurs.

Son développement a même permis d'envisager le même type d'outil pour les transplantations cardiaques et les maladies cardiovasculaires (92).

Ainsi la Société française de médecine prédictive et personnalisée (SFMPP), qui évalue le bénéfice médical et les bonnes pratiques de tests génétiques prédictifs pour améliorer le dépistage, la prévention et les traitements, est enthousiaste :

« Ce qui est fascinant, c'est l'affinement de la prise en charge dans les trois de domaines d'action, dépistage, prévention et traitement qui accompagne ces progrès. La démonstration de niveaux de preuve pour l'utilisation des marqueurs génomiques ou génétiques s'accélère.

Cela va des possibilités de prise en charge de risque dans le cadre de maladies héréditaires, à la prédiction de la réponse thérapeutique à des traitements conventionnels ou ciblés.

Cette révolution est en marche. Il ne faut pas la craindre, mais l'accompagner de toute la réflexion déontologique et éthique qu'elle requiert, ainsi que des évaluations médico-économiques nécessaires. »

Néanmoins l'environnement a prouvé être aussi un facteur déterminant et joue un rôle majeur dans l'apparition et le développement des maladies.

## IV.4 Les limites actuelles du diagnostic prédictif

### IV.4.1 L'influence de l'environnement

La médecine prédictive se base sur les caractéristiques du patient, mais il est nécessaire de prendre aussi en compte l'environnement.

Il existe des maladies qui s'explique par la mutation d'un seul gène, comme la chorée de Huntington, où une altération d'un gène autosomal suffit à provoquer la maladie.

Néanmoins pour les maladies multifactorielles, qui représente une majorité des pathologies, l'interaction des facteurs génétiques et de l'environnement est encore mal connue. Chaque anomalie d'un gène n'est pas suffisante pour entraîner un processus pathologique dans un contexte donné environnemental.

La connaissance précise du phénotype à partir du seul génotype est impossible, en raison du grand nombre de facteurs génétiques et environnementaux qui sont impliqués.

La médecine de demain devra donc prendre en compte un grand nombre de facteurs avant de pouvoir être réellement « prédire l'avenir » et les maladies qui surviendront

tout au long de la vie dès la naissance, en intégrant des notions de médecine environnementale.

### IV.4.2 L'interprétation des statistiques : dire tout et son contraire

Précurseurs du big data, les statistiques sont les premiers outils utilisés pour exploiter les données.

Mark Twain écrivait « les faits sont têtus, il est plus facile de s'arranger avec les statistiques ». Et ajoutait : « il y a trois sortes de mensonges : les mensonges, les sacrés mensonges, et les statistiques. »

Les statistiques, les pourcentages, résultats de sondages, peuvent être utilisés à mauvais escient ou peuvent être paradoxaux pour le bon sens et peuvent donner des interprétations contradictoires en apparence.

Ainsi nous devons examiner avec une grande attention les données chiffrées pour éviter les erreurs, en particulier dans un domaine sensible comme la santé.

Pour un risque exprimé, des paramètres doivent être bien identifiés :

- S'il s'agit d'un risque relatif ou absolu [23]
- L'échelle de temps utilisée
- L'importance du risque
- La population concernée

Dans le domaine médical les médecins doivent aussi correctement interpréter les résultats d'un test et ne pas alerter inutilement leur patient, ce qui pourrait leur nuire à terme. Cela nous amène au point suivant.

[23] Risque relatif : mesure la probabilité de survenue d'un événement dans un groupe par rapport à l'autre. Risque absolu : mesure la probabilité de survenue d'un évènement dans une population donnée et dans des conditions spécifiques.

### IV.4.3 Quel intérêt de connaitre un diagnostic alors qu'il n'existe aucune cure ?

Comme abordé précédemment, le LEEM évoque un 5ème P dans la médecine de demain, la médecine de la pertinence, à savoir la preuve de l'utilité du diagnostic. « La médecine prédictive pose un problème moral et déontologique qui n'est pas résolu. Il est compliqué de savoir si l'on doit dire à des adultes plus ou moins bien portants, des adultes-parents ou des enfants, ce qui les attend à plus ou moins brève échéance. » pouvons-nous lire dans leur rapport sur la santé en 2030 (93).

En l'absence de thérapeutique, comme dans le cas de la chorée de Huntington, le diagnostic présymptomatique ne se conçoit que chez l'adulte qui souhaite connaître réellement son statut, dans le cadre d'une prise en charge par une équipe multidisciplinaire pour assurer un suivi et accompagnement optimal (94).

Les progrès de la technique devront donc respecter l'éthique du patient et ne pas le faire souffrir inutilement, notamment les enfants que l'on pourrait dépister, comme l'indique l'avis rendu sur la question de la médecine prédictive par le Comité Consultatif National d'Ethique : « Une médecine préventive qui permettrait de prendre en charge, de manière précoce et adaptée, des enfants manifestant une souffrance psychique ne doit pas être confondue avec une médecine prédictive qui emprisonnerait, paradoxalement, ces enfants dans un destin qui, pour la plupart d'entre eux, n'aurait pas été le leur si on ne les avait pas dépistés. » (95)

Nous avons donc vu que par la mise en commun des connaissances scientifiques avec le réseau des réseaux, et l'établissement de techniques avancées d'intelligence artificielle grâce à la recherche et à l'innovation dans le *hardware*, des applications en santé sont apparues. Parmi elles, une solution pour le cancer du côlon a pu être développée : c'est l'exemple de l'Immunoscore®.

# V. UN EXEMPLE D'INTELLIGENCE ARTIFICIELLE EN SANTE : L'IMMUNOSCORE ®

Le Dr Jérôme Galon, prix de l'inventeur européen 2019, a mis au point un test permettant une meilleure estimation de l'évolution de la maladie chez des patients atteints du cancer du côlon.

## V.1 Etat des lieux sur le cancer du colon

Le cancer du côlon fait partie de la famille des cancers colorectaux.

### V.1.1 Epidémiologie : Fréquence du cancer du colon

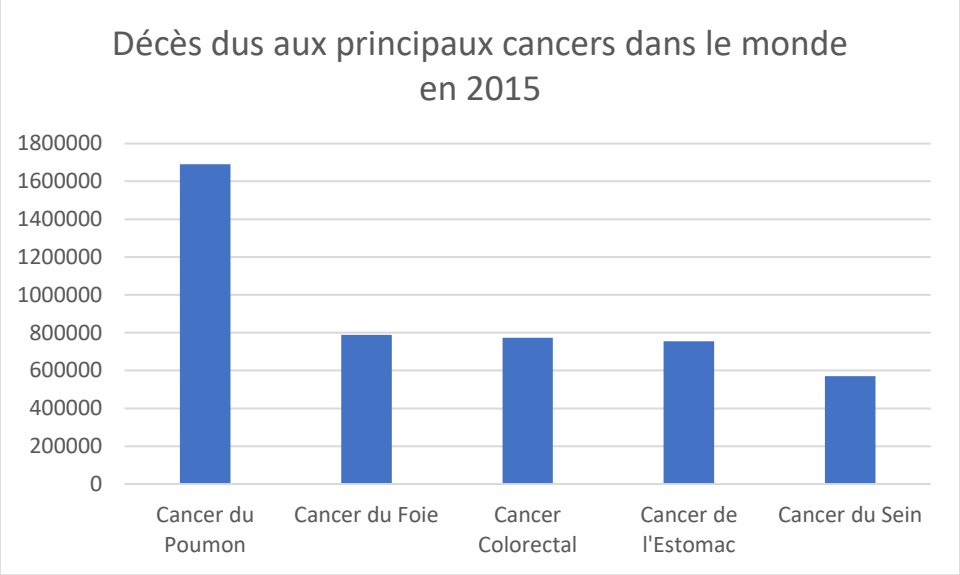



*Figure 28: Décès dus aux principaux cancers dans le monde en 2015*

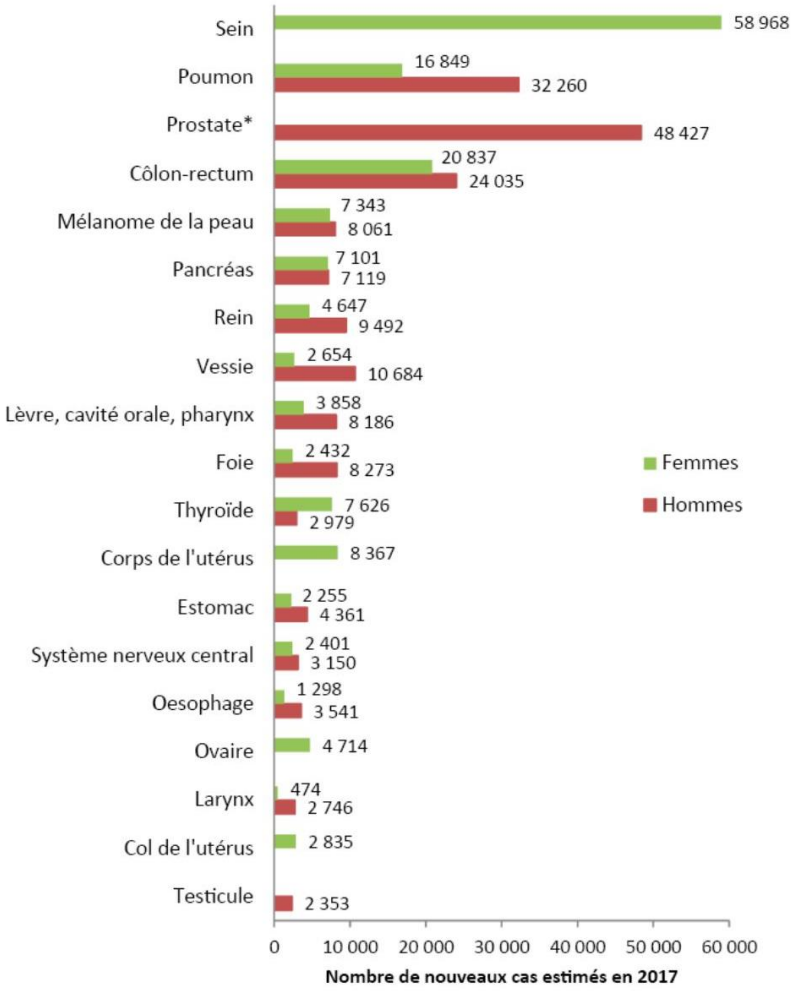
Dans le monde, le cancer colorectal est le 3ème le plus mortel avec 774000 décès en 2015 (96).

En 2018, il a été estimé 1 million de nouveaux cas avec 881 000 décès.

En France, le cancer colorectal représente 20 % de tous les cancers, et se situe en troisième place après le cancer de la prostate et celui du sein, avec plus de 43 000 nouveaux cas recensés par an en 2018.

Il représente la deuxième cause de mortalité par cancer avec 17 000 décès (97).

Tous stades confondus, la survie à 5 ans du cancer du côlon est d'environ 60 %.




* Les données de projection 2017 ne sont pas fournies pour ce cancer. Il s'agit de l'estimation pour 2013.
Source : Partenariat Francim/HCL/Santé publique France/INCa [Jéhannin-Ligier K, 2017]. Traitement : INCa 2017

*Figure 29 : Classement des tumeurs solides par incidence estimée en 2017 en France métropolitaine selon le sexe (22)*

Les facteurs de risques sont :

Exogènes :

- Le tabagisme
- Le manque d'exposition régulière au soleil
- La consommation d'alcool
- Une alimentation riche en charcuteries, grillades et aliments fumés ; ou pauvre en fibres alimentaires (fruits, légumes, céréales complètes). (A l'inverse un régime riche en fibres aurait un rôle protecteur (98).)

Endogènes :

- Le diabète de type 2 et l'obésité (apport calorique élevé et absence d'exercice physique régulier)
- Certaines maladies héréditaires et maladies inflammatoires chroniques de l'intestin comme la maladie de Crohn ou la rectocolique hémorragique
- Mutations génétiques entrainant une instabilité microsatellitaire[24] (cancer de type MSI)
- Les antécédents personnels de cancer colorectal
- La prédisposition familiale. Le risque est multiplié par 2 si un parent du 1er degré a eu un cancer colorectal.
- L'âge, supérieur à 50 ans dans la majorité des cas
- Polypose autosomique familiale

[24] Les microsatellites correspondent à des séquences d'ADN réparties sur l'ensemble du génome (séquences codantes ou non codantes) dont la structure est répétitive. Du fait de cette structure répétitive, ils sont particulièrement sujets aux erreurs de réplication en cas de défaillance du système MMR (*Mismatch Repair*). La défaillance du système MMR est responsable d'un phénotype MSI (*Microsatellite Instability*), par opposition aux cancers de type MSS (*Microsatellite Stability*).

### V.1.2 Définition

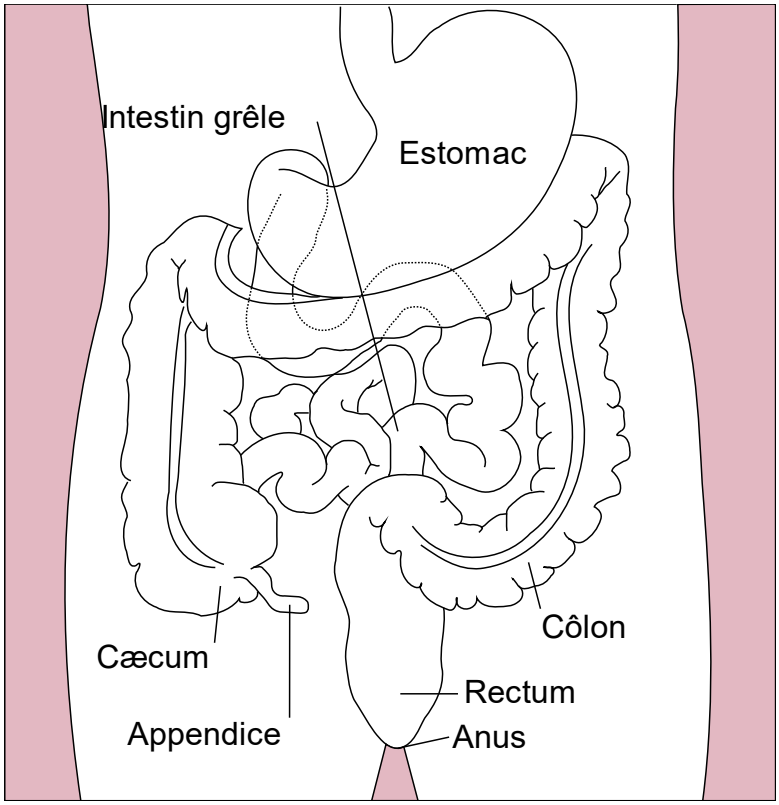



*Figure 30: Schéma du tube digestif (99)*

Le côlon forme la dernière partie du système digestif et fait suite à l'intestin grêle.

Il est constitué de plusieurs parties :

- le cæcum;
- le côlon ascendant (droit) ;
- le côlon transverse ;
- le côlon descendant (gauche);
- le côlon sigmoïde ;
- le rectum
- l'anus.

Il mesure de 1 à 1,5 mètre de long pour 4 à 8 centimètres de diamètre et est composé de plusieurs couches, de l'intérieur vers l'extérieur :

- Une muqueuse, la couche interne sans valve ni valvule, mais présentant des haustrations dites haustrations coliques,
- Une sous-muqueuse, couche intermédiaire très vascularisée et innervée
- Une musculeuse : couche externe composée des muscles lisses circulaires (couche circulaire internes) et longitudinaux (couche longitudinale externe)
- Une zone sous-séreuse entre le péritoine et la musculeuse
- Une séreuse : le péritoine qui est l'enveloppe tapissant la paroi externe du gros intestin.

Il est à noter que l'Immunoscore® ne s'applique qu'aux cancers du côlon ; le rectum non péritonisé et l'anus sont exclus de cette étude. En effet, les cancers du bas rectum (rectum non péritonisé) bénéficient de traitement particulier, pouvant inclure une radiothérapie associée à une chirurgie et/ou une chimiothérapie.

### V.1.3 Physiopathologie

Le cancer colorectal est dans 95% des cas un adénocarcinome, c'est-à-dire un type de cancer qui se développe à partir des cellules d'une glande (sein, thyroïde, prostate, etc.), de son revêtement (ovaire) ou comme ici d'un épithélium (estomac, côlon…) qui tapisse l'intérieur du colon ou du rectum. Il se développe aux dépens de la muqueuse du côlon ou du rectum et va mimer de façon anarchique l'épithélium glandulaire initial.

La dénomination du cancer dépend de la position des cellules atteintes :

- A plus de 15cm de l'entrée du rectum : c'est un cancer du côlon.
- A moins de 15 cm : c'est un cancer du rectum.

60 à 80% des cancers colorectaux se développent à partir d'un polype adénomateux.

Un polype est une excroissance de la muqueuse du côlon. Ils sont soit bénins (et n'évoluent jamais) soit adénomateux, et peuvent engendrer une tumeur cancéreuse, au cours d'une évolution lente entre 10 et 15 ans.

Dans un premier temps, une cellule dite initiatrice au sein d'une population normale (comme une muqueuse) est altérée et va subir une mutation de son ADN qui augmente sa propension à la mitose : c'est l'initiation.

Puis vient l'étape de promotion : la cellule prolifère et se multiplie, via des promoteurs. Il en existe une multitude, dont les hormones sexuelles et les facteurs de croissance. Les cellules filles ont toujours un aspect normal mais sont en trop grand nombre et forment une hyperplasie, un type de dysplasie.

À l'étape de la progression les cellules deviennent cancéreuses : elles se reproduisent de façon anarchique et peuvent perdre leur caractère différencié lié au tissu auquel elles appartenaient, changent en aspect et en forme.

La tumeur envahit progressivement le tissu environnement et s'installe durablement, pour une durée qui peut être longue.

Quand les frontières du tissu sain sont rompues (ici dans l'épithélium dès que la lame basale est rompue et dépassée), vient l'étape de la colonisation de l'organisme : la tumeur devient invasive et certaines cellules cancéreuses s'infiltrent dans les vaisseaux sanguins ou lymphatiques environnements pour venir s'implanter dans d'autres endroits du corps humains et viennent former des nouveaux foyers métastatiques.

Au niveau de la zone où nait le cancer, il existe une **réaction immunitaire** aux cellules cancéreuses : cette zone qui constitue une interface est appelée « front ».

Au niveau cellulaire, des cellules immunitaires sont dirigées contre les cellules tumorales via le mécanisme d'immunité innée de l'organisme et les détruisent.

La lyse des cellules tumorales entraine la libération de cytokines (des protéines activatrices du système immunitaire) dans l'environnement, qui incitent d'autres cellules tueuses à lyser des cellules tumorales et également engendre la libération d'antigènes que les cellules dendritiques ingèrent.

Ces cellules dendritiques, qui sont un maillon clé dans la réaction immune, s'activent, migrent vers les organes lymphoïdes secondaires et présentent l'antigène aux lymphocytes B et T qui vont intégrer l'information et à leur tour reconnaitre les cellules

cancéreuses présentant le même antigène (via deux éléments : un complexe protéique à leur surface, le Complexe Majeur d'Histocompatibilité, ou CMH ; et un ligand costimulant), se regrouper et déclencher des mécanismes de destruction ciblée.

Il y a donc des lymphocytes autour de la tumeur, et la recherche s'y intéresse aujourd'hui : ils deviennent à la fois marqueur de diagnostic (cf. ci-après) et cible d'amélioration pour des traitements anticancéreux (100).

### V.1.4 Diagnostic

Chez un patient asymptomatique, le diagnostic peut être réalisé dans le cadre de campagne ou d'une demande de dépistage.

En effet depuis 2008 en France, la population cible de 18 millions d'hommes et de femmes de 50 à 74 ans (hors critères d'exclusion) est invitée à procéder gratuitement à un test immunologique de recherche d'hémoglobine humaine dans les selles à domicile, qui sera ensuite transféré à un laboratoire pour analyse. En cas de test négatif, le patient est invité à reproduire le test 2 ans plus tard. En cas de test positif, il devra réaliser une coloscopie. La procédure a permis de détecter plus de 8000 cas de cancer colorectaux chez des individus asymptomatiques pour la période 2009-2010 (101).

Chez un patient présentant des symptômes évocateurs, le diagnostic repose sur la **coloscopie**, qui permet d'explorer la tumeur et réaliser des biopsies.

Les symptômes évocateurs sont les rectorragies, les troubles du transit d'apparition récente, les douleurs abdominales, mais aussi une anémie ferriprive chez l'homme dans tous les cas, et chez la femme sans atteinte gynécologique ou après 50 ans.

### V.1.5 Gradation du cancer

Pour établir un pronostic est utilisée la classification TNM, une classification internationale qui permet de classifier le stade d'un cancer. Par convention lorsque le monogramme TN est précédé du « p » de pathologiste, la stadification a été donnée par l'anatomo-pathologiste après examen histologique de la pièce opératoire.

La lettre T vaut pour « Tumeur » et correspond à la taille de la tumeur. Dans le cas du côlon il correspond au niveau d'invasion du cancer dans l'épaisseur de la paroi colique.

La lettre N correspond à « Node », ganglion en anglais, et indique si des ganglions lymphatiques ont été envahis ou non. L'examen d'au moins 12 ganglions régionaux est recommandé par l'UICC et l'AJCC pour établir le statut N de cette classification. (Si moins de douze ganglions lymphatiques sont examinés par le pathologiste (car non extirpés par le chirurgien lors de l'intervention) alors le nombre optimal n'est pas atteint et une réserve sur le nombre de ganglions examinés est portée.)

La lettre M est l'initiale de « Métastase » et signale leur présence ou absence.

On peut ensuite grader les scores en stade :

*Tableau 1: TNM et stades correspondants (102)*

| | |
|---|---|
| Stade I | p(1) T1-T2 N0 M0 = sous-séreuse intacte sans envahissement ganglionnaire |
| Stade II A | p(1) T3 N0 M0 = sous-séreuse atteinte sans envahissement ganglionnaire |
| Stade II B | p(1) T4 N0 M0 = séreuse franchie et/ou perforée, et/ou envahissement d'organes voisins, sans envahissement ganglionnaire |
| Stade III A | p(1) T1, T2, N1 M0 = sous-séreuse intacte avec envahissement ganglionnaire |
| Stade III B | p(1) T3,T4, N1 M0 = sous-séreuse atteinte et/ou séreuse franchie et/ou perforée, et/ou envahissement d'organes voisins, avec envahissement ganglionnaire |
| Stade III C | tous T, N2 M0 = envahissement ganglionnaire |
| Stade IV | tous T, tous N, M1= métastases à distance |

(1) p : examen anatomopathologique sur pièce opératoire.

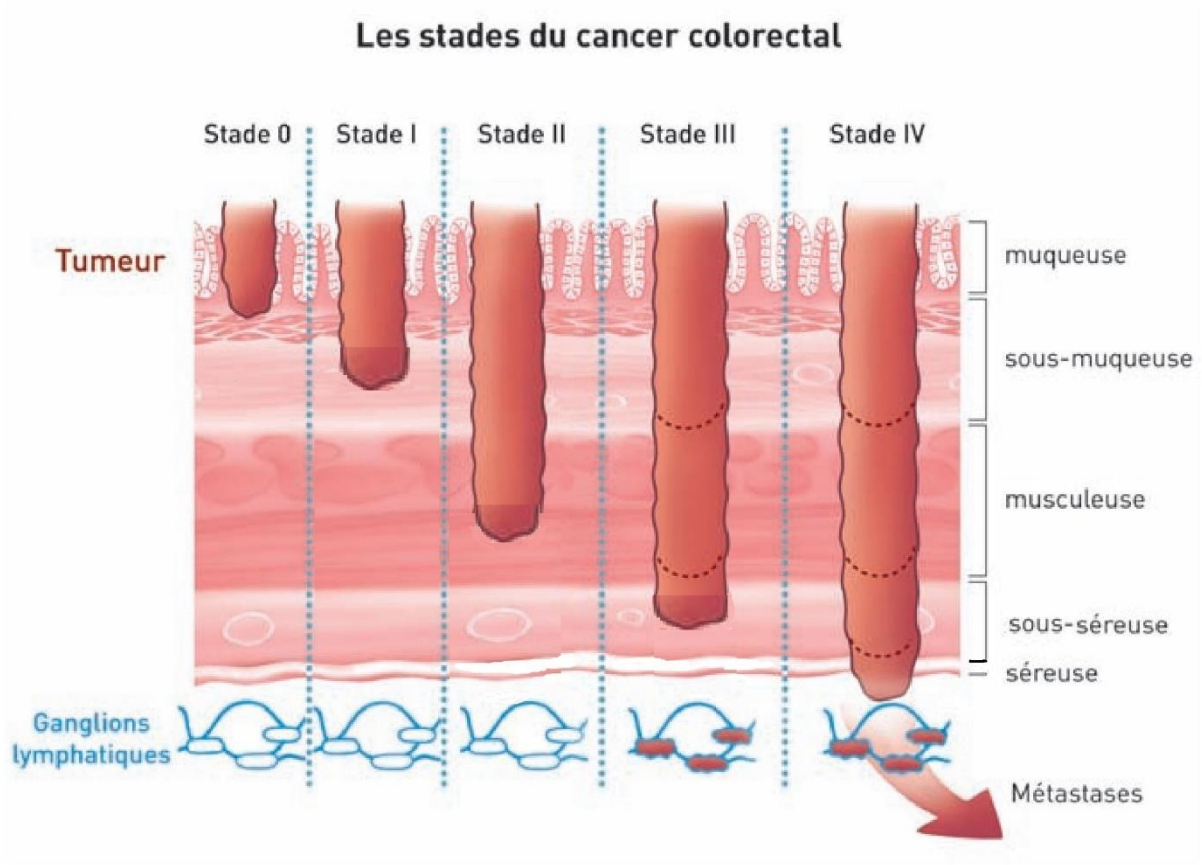



*Figure 31: Les stades du cancer colorectal (édité C. Chollat-Namy) (103)*

D'autres outils sont disponibles pour évaluer le cancer, et par conséquent son traitement adéquat. L'Immunoscore® en fait partie et fait l'objet de cette thèse. Il sera utile pour des patients qui sont gradés T2 et T3 dans la classification TNM, pour lesquels les praticiens discutent de l'intérêt d'une chimiothérapie, contrairement aux patients gradés T4 qui seront traités par chimiothérapie dans tous les cas, et les patients gradés T1 qui seront seulement traités par chirurgie. (et possiblement une radiothérapie dans le rectum non péritonisé).

### V.1.6 Facteurs prédictifs du risque métastatique

Les facteurs prédictifs sont utiles pour une population particulière qui n'a pas un cancer trop agressif et aident à déterminer la nature du traitement le plus adapté, pour éventuellement ne pas recourir à la chimiothérapie.

Les études autour du projet Immunoscore ont permis d'établir que le profil génétique d'une tumeur était moins important pour la prédiction du potentiel métastatique que l'environnement de la tumeur : la faible densité du réseau lymphatique ou sanguin ; la rareté de certaines cellules immunitaires dans et autour de la tumeur sont associées à un risque important de métastases (104) (cf. V.2).

De plus, il semblerait que le microbiote intestinal ait sa part de responsabilité. Une étude a montré que deux sous-populations de lymphocytes T CD4+ jouaient sur l'évolution du cancer colorectal. L'une est en faveur de la réaction immunitaire antitumorale, tandis que l'autre a un rôle de régulateur anti-inflammatoire et pro-tumorale : ce sont les lymphocytes Treg.

La présence d'une bactérie intestinale, Fusobacterium nucleatum, semble entrainer par une réaction inflammatoire locale l'expression de la protéine FOXP3 qui engendre la transformation des lymphocytes en lymphocytes Treg. Cette bactérie serait donc impliquée dans l'orientation pro ou antitumorale du système immunitaire dans les cancers colorectaux (105).

### V.1.7 Traitements

Il y a plusieurs traitements possibles, suivant le stade de développement de la tumeur.

#### *V.1.7.1 Chirurgie*

Elle comprend l'exérèse de la tumeur, du mésocolon attenant, et un curage ganglionnaire.

La résection colique pour cancer doit assurer une marge intestinale suffisante (5cm de chaque côté de la tumeur) et préserver une vascularisation satisfaisante.

#### *V.1.7.2 Chimiothérapie*

1) adjuvante (après chirurgie)
2) néoadjuvante (avant chirurgie)
3) en association : thérapies ciblées.

*Tableau 2: Principaux protocoles de chimiothérapie utilisés (D'après le Vidal) (106)*

| PROTOCOLES(1) | LIEU(2) | DURÉE DES CURES | INTERVALLE ENTRE LES CURES |
|---|---|---|---|
| LV5FU2<br>(acide folinique + 5FU) | HDJ ou D | 2 jours,<br>en perfusion continue | 14 jours |
| FOLFIRI<br>(LV5FU2-irinotécan)<br>(acide folinique + 5FU<br>+ irinotécan) | HDJ | 2 jours,<br>en perfusion continue | 14 jours |
| FOLFOX (LV5FU2 -<br>oxaliplatine)<br>(acide folinique + 5FU<br>+ oxaliplatine) | HDJ | 2 jours,<br>en perfusion continue | 14 jours |
| FOLFIRINOX<br>(acide folinique + 5FU<br>+ irinotécan + oxaliplatine) | HDJ | 2 jours,<br>en perfusion continue | 14 jours |
| XELOX<br>(oxaliplatine<br>+ capécitabine) | HDJ et D | Perfusion IV de 2 heures<br>+ 1 comprimé 2 fois par jour<br>pendant 14 jours | 21 jours |
| XELIRI<br>(irinotécan + capécitabine) | HDJ et D | Perfusion IV de 2 heures<br>+ 1 comprimé 2 fois par jour<br>pendant 14 jours | 21 jours |

(1) Le bévacizumab peut être ou non associé aux différents protocoles ci-dessus. Le cétuximab peut être ou non associé au LV5FU2 ou au FOLFIRI.
(2) HDJ = hôpital de jour, D = domicile

#### *V.1.7.3 Radiothérapie*

On exclura la radiothérapie qui concerne uniquement le rectum.

### V.1.8 Le prix de la chimiothérapie et l'intérêt de tests prédictifs pour s'assurer de son intérêt

Les traitements de chimiothérapie sont coûteux pour l'assurance maladie.

En effet le coût moyen d'un cancer colorectal est de 28 000€ (de 17000€ en stade I à 36000€ en stade IV), soit 1,12 milliards d'euros par an pour 40 000 nouveaux cancers (107).

Par exemple dans le cadre d'un protocole FOLFOX ou FOLFIRI, le coût est d'environ 75€ tous les 15 jours, qui peut être associé au bévacizumab (AVASTIN ®) qui coûte 2000€ par mois (108).

Le STIVARGA ® lui, coûte 2389,80 euros la boite de 84 comprimés, ce qui correspond au prix d'une cure de 21 jours à raison de 4 comprimés par jour (109).

De plus, ils entrainent de très nombreux effets indésirables ou mortels.

C'est pour cette raison que des programmes de prévention, de diagnostic et dépistage précoces ont été élaborés pour limiter l'emploi des chimiothérapies, à l'exemple de la coloscopie de prévention (évoquée plus haut).

Une fois le cancer diagnostiqué, des examens ont été développé pour aider les praticiens à choisir le traitement adapté à leurs patients. Le dépistage du déficit en DPD[25] pour le médicament Xeloda ®, qui peut être réalisée par génotypage ou phénotypage, permet par exemple de prévenir les toxicités sévères ou létales qui se produisent chez un patient en déficit partiel ou complet de cet enzyme lors d''un traitement anticancéreux à base de fluoropyrimidines (110).

Pour guider le praticien dans le choix de son traitement, il existe également le test de l'Immunoscore®, que nous étudions ici.

[25] La DPD ou l'enzyme dihydropyrimidine déshydrogénase dégrade le 5-Fluoro-Uracile (5-FU), une molécule utilisée dans certains traitements anti-cancéreux. En cas de déficit enzymatique, le 5-FU reste présent dans l'organisme et entraine des effets indésirables graves.

## V.2 L'Immunoscore, un outil de *digital pathology*

Le test Immunoscore ® a été développé par une équipe de chercheurs de l'Inserm, de l'Université Paris Descartes et de médecins de l'AP-HP. Il est destiné à des patients souffrant d'un cancer du côlon et permet une meilleure prédiction de l'évolution de la maladie, comme le démontre une étude publiée dans *The Lancet* sur plus de 2500 patients. Ainsi ce test est efficace pour prédire les patients à haut risque de récidive tumorale, et aide les praticiens à identifier la population de patients candidate à un renforcement thérapeutique après la chirurgie.

L'Immunoscore® se base sur les recherches du Dr Jérôme Galon et de son équipe qui étudiant le contexte immunitaire dans les cancers humain, ont découvert l'importance de la nature, de l'orientation fonctionnelle, de la densité et de la localisation de cellules immunitaires dans la tumeur. La conséquence de ces découvertes : la caractérisation et la quantification de la réaction immunitaire adaptative sont un meilleur prédicateur de survie que la classification traditionnelle basée sur la taille et la propagation d'une tumeur, comme la classification TNM.

En effet leurs travaux ont montré, à partir de l'analyse de la pièce opératoire de résection colique initiale, que plus le nombre de cellules immunitaires était élevé au niveau de l'interface, encore appelée front, entre cellules cancéreuses et l'hôte, plus longue était l'espérance de vie dans la survie des cancers du côlon (111).

Le test Immunoscore® permet donc par une approche de type *digital pathology* de dénombrer ces cellules immunitaires de façon précise et reproductible au niveau du front tumoral, puis de produire un score en appliquant une couche logicielle.

D'après la *Digital Pathology Association*, la *digital pathology*, ou pathologie numérique, est un processus qui consiste à numériser des lames microscopiques afin de les visualiser et d'une façon similaire à celle du microscope de les interpréter via un ordinateur (112).

Cette solution permet, contrairement à celle du microscope traditionnel :

- une accessibilité en tout lieu sur PC et tablettes via Internet sans la nécessité d'un microscope
- de consulter plusieurs lames en même temps,
- d'améliorer l'archivage et sa consultation,
- une intégration avec les systèmes d'informations,
- une exploitation des données décuplée en pouvant appliquer des algorithmes d'intelligence artificielle pour rechercher et déduire des informations (113).

Ainsi, l'étude a démontré que l'Immunoscore® est plus performant et fiable que n'importe quel autre moyen pronostic, y compris le TNM, dans l'estimation du risque de récidive dans le cancer du côlon : les patients ayant un Immunoscore élevé ont une espérance de vie de 5 à 15 ans ; ceux qui ont un score bas ont une espérance de vie de moins de 2 ans en l'absence de traitement adapté.

Les médecins peuvent donc mieux appréhender le traitement du patient et prendre des décisions plus adaptées, dans des situations particulières si le patient ne présente :

- Pas de métastase ganglionnaire
- Pas d'atteinte du péritoine
- Ni d'atteinte d'un autre organe de proximité.

En fonction des résultats de l'Immunoscore ®, l'intérêt de la chimiothérapie se discute.

Le Dr Jérôme Galon à l'origine du projet, fonde la société HalioDx et s'associe en 2018 à Keen Eye, spécialisée dans les technologies d'analyse d'images, pour développer cette méthode. Le but de cette association est de combiner une technologie basée sur l'IA stockée dans le cloud et un protocole de diagnostic déjà établi, afin d'apporter aux praticiens une solution pour poser un diagnostic plus précis et standardisés grâce à des algorithmes d'apprentissage statistique, mais également la possibilité aux partenaires et sociétés pharmaceutiques de développer des tests diagnostiques compagnons et des médicaments.

## V.3 L'approche nouvelle d'HalioDx

### V.3.1 Entretien avec Dr Alexia Papadopoulos

*Propos recueillis lors d'un entretien avec Dr. Alexia Papadopoulos, docteur en immunologie, et ingénieur dans le département Produits Pharmaceutiques chez HalioDx.*

La société HalioDx a pour origine Ipsogen, une société de biotechnologies créée en septembre 1999 et spécialisée dans la conception de biopuces, sur le campus du parc scientifique de Luminy à Marseille (114). Elle développe une plateforme technologique qui permet de développer des méthodes innovantes de diagnostic du cancer, notamment des kits de diagnostics sur les leucémies.

En 2011, Ipsogen est rachetée par Qiagen, une société néerlandaise leader mondial dans les kits de biologie moléculaire, tels que les kits d'extraction pour ADN/ARN ou les kits de diagnostics.

En 2015 l'équipe originelle d'Ipsogen quitte finalement la société et s'associe avec le Dr Jérôme Galon qui étudiait le contexte immunitaire des cancers et son impact sur la réponse du patient aux traitements pour former HalioDx, dans le but de « développer de nouvelles approches de diagnostic pour aider médecins et patients à vaincre le cancer » (115).

HalioDx comporte 3 activités principales : notre plateforme de test sur échantillons, un département Recherche et Développement chargé d'inventer des nouveaux tests et logiciels, et un département « Partnership Project » chargé du développement de kits à commercialiser avec nos partenaires, comme les kits de biologie moléculaire que l'on a développé pour Qiagen, qui sont vendus pour des praticiens hospitaliers (116). Il y a aussi une unité de production des kits, des services qualité, marketing, business…

L'objectif initial est de créer des kits de diagnostic pour l'Immunoscore ® dans le cancer du côlon.

Ces kits de diagnostic respectent les normes CE IVD, relatifs aux dispositifs médicaux de diagnostic in vitro.

- **Quel est l'objectif de l'Immunoscore ® ?**

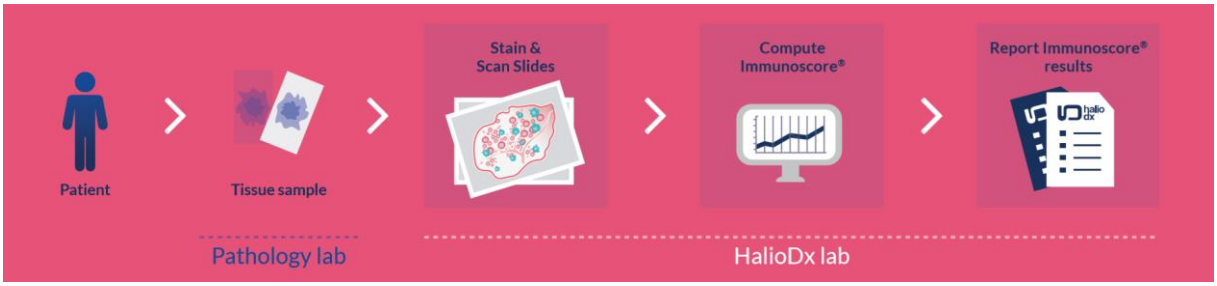



*Figure 32 : Processus des étapes de l'Immunoscore®, selon HalioDx*

L'Immunoscore est une aide au diagnostic. Il est utilisé chez les patients atteints du cancer du côlon, et cherche à identifier parmi ceux-ci une population de patients qui pourraient bien répondre à un traitement complémentaire par chimiothérapie. Le but est d'orienter le praticien vers les traitements qui sont les plus adaptés au cancer de son patient.

C'est un test de diagnostic in vitro, qui prédit le risque de rechutes chez les patients atteints de cancers du côlon localisé, en mesurant la qualité de la réponse immune de l'hôte sur le site de la tumeur.

Ce test permet de quantifier la population des cellules lymphoïdes, en particulier dans la zone d'infiltration tumorale, que l'on dénomme marge tumorale ou marge d'invasion (ou encore « front » ou « interface »). Ces cellules lymphoïdes sont des marqueurs d'adaptation et de résistance de l'hôte au cancer qui l'agresse.

Le kit de diagnostic est associé à une solution logicielle validés par les instances européennes, avec des paramètres prédéterminés pour que le test soit reproductible et fiable.

- **Comment fonctionne le logiciel ?**

Le principe est simple. Le technicien réalise deux séries d'immunodétection (réalisées à partir du matériel tumoral provenant d'une pièce opératoire fixée dans du formol, incluse en paraffine et coupée consécutivement à 5 µm d'épaisseur).

La première immunohistochimie utilise des marqueurs révélant les cellules exprimant le CD3 : des lymphocytes CD3[26], et l'autre utilise des marqueurs révélant les cellules exprimant le CD8 : des lymphocytes CD8.

Le logiciel, qui est un outil de *digital pathology*, analyse la lame du patient et propose sur l'image ce qu'il estime être la zone tumorale et la zone saine de la lame. Un opérateur retouche si besoin et valide cette proposition. Sur les lames difficiles à interpréter, le médecin anatomopathologiste contrôle les résultats du logiciel.

Le logiciel détermine ces zones par rapport à la morphologie des cellules qu'il reconnait. Actuellement nous lui apprenons à se fier aussi au marquage des cellules.

Une fois les zones de tissu sain et de tumeur déterminées, le programme détermine automatiquement la marge d'invasion, ainsi que la densité de lymphocytes CD3+ et CD8+ dans la zone d'invasion et dans la zone tumorale.

Ainsi à partir de la densité de lymphocytes CD3+ dans la zone tumorale et dans la marge d'invasion, et de la densité de lymphocytes CD8+ dans la zone tumorale et dans la marge, un algorithme du logiciel détermine l'Immunoscore ® du patient.

La sélection des zones est donc importante en amont pour avoir une répartition juste des cellules. De plus, les nouvelles études du Dr Galon et son équipe suggèrent que certaines répartitions de cellules pourraient avoir plus de poids dans le calcul de l'Immunoscore.

[26] CD = Cluster de différenciation, qui désigne les glycoprotéines membranaires servant à la classification des cellules du système immunitaire. Le CD3 est caractéristique de la population des lymphocytes T. Le CD8 est présent sur les lymphocytes T cytotoxiques.

- **Exemple de lames de colon analysées :**

**1er cas**

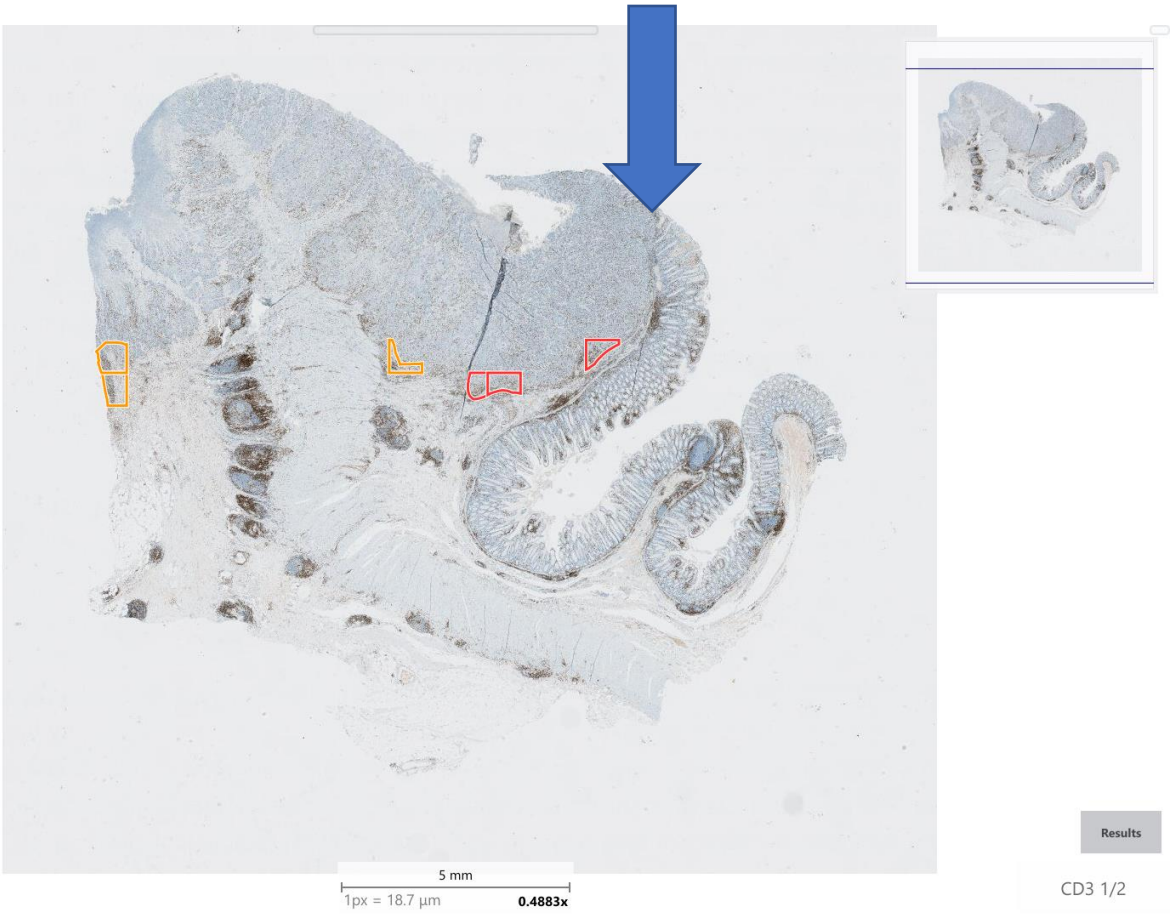



*Figure 33 : Image d'une coupe de colon*

Il s'agit d'une digitalisation d'une lame d'immunohistochimie où l'épitope CD3 a été recherché sur les cellules de la lame à l'aide d'anticorps anti-CD3.

A gauche de l'image se situe le cancer et à droite la muqueuse colique qui est normale. Une flèche délimite la zone de transition, qui sur cette diapositive est abrupte. Chez d'autres patients, la zone de transition consiste en un patchwork entre îlots épithéliaux sains et îlots tumoraux.

Nous présentons ensuite la même préparation immunohistochimique après traitement par le logiciel.

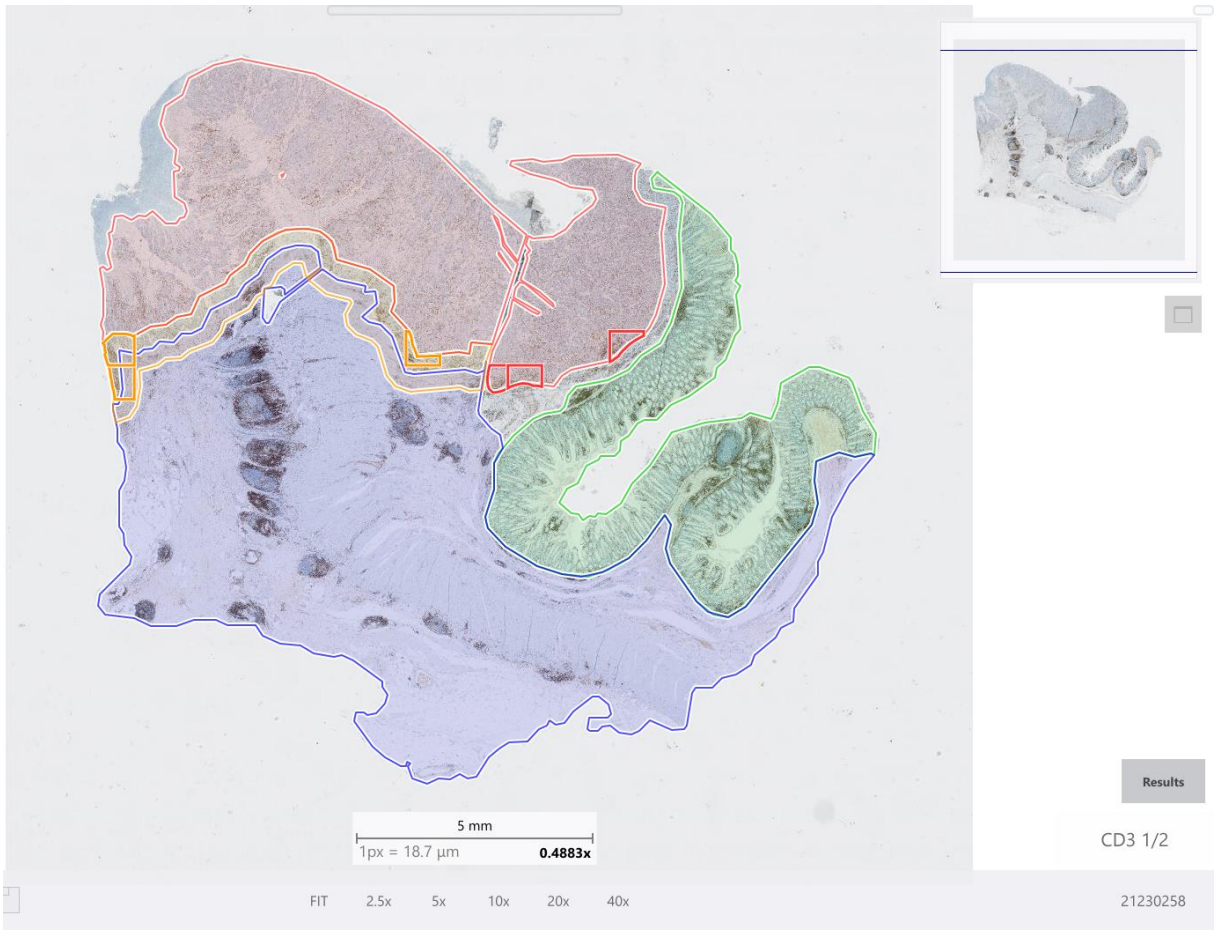



*Figure 34 : Image d'une coupe de colon analysée par logiciel*

L'image sur la figure 34 a été traitée par l'analyseur d'image avec le logiciel Immunoscore ® et contrôlée par un histotechnicien du service *testing*.

Le logiciel délimite en rouge l'aire du cancer, en vert l'aire de la muqueuse colique non tumorale (structures épithéliales ne correspondant pas à un foyer d'adénocarcinome invasif), en jaune la ligne de front qui correspond à la zone pertinente d'analyse.

En bleu se trouve la zone de tissu sain.

Les rectangles rouges ciblent les zones les plus denses en cellules.

Sur cette image la muqueuse correspond à une zone de muqueuse normale, qui n'est ni cancéreuse ni même adénomateuse.

Le logiciel a néanmoins besoin qu'on reconnaisse correctement les zones épithéliales non tumorales (les zones en vert) car elles sont exclues du calcul (ce sont des zones de silence), de même que les zones nécrosées ou purulentes, que le technicien du *testing* doit reconnaitre pour les exclure.

Après Immunoscore ®, la lame doit dans tous les cas être validée par un technicien qui valide la sélection des zones effectuée par le logiciel.

Dans un cas compliqué (présence de zones nécrosées, hémorragiques, ou d'artefacts) un pathologiste intervient et donne son expertise.

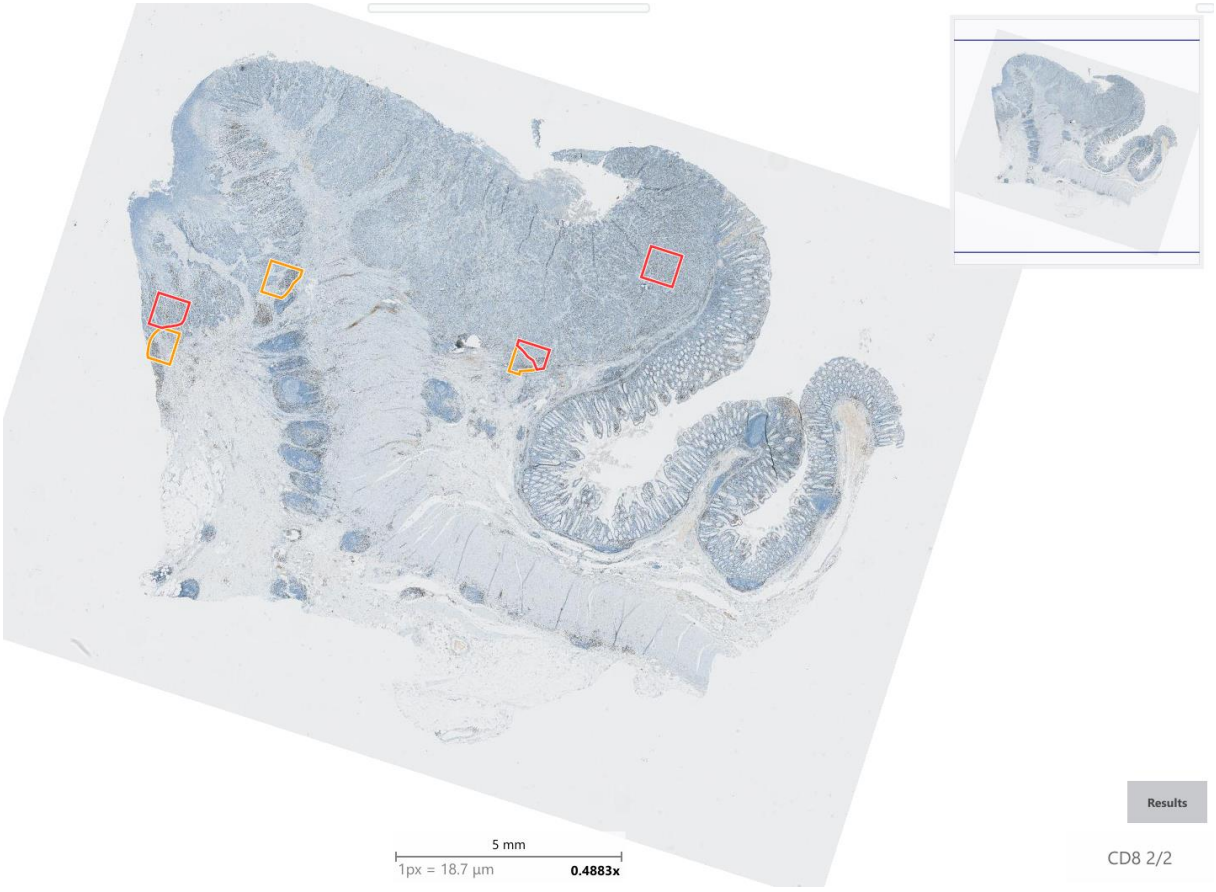



*Figure 35 : Image d'une coupe de colon*

La coupe a été traitée avec des anticorps anti-CD8.

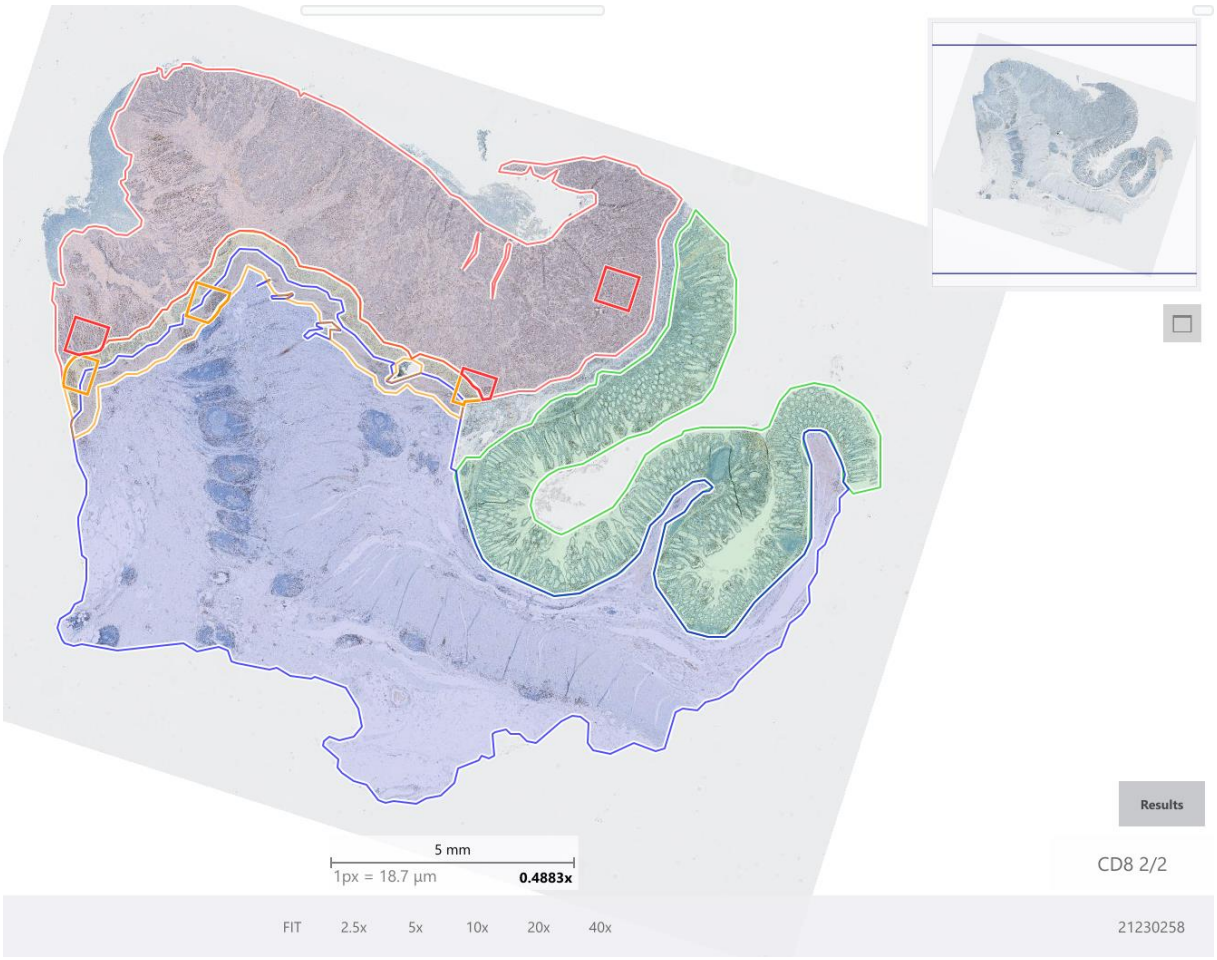



*Figure 36 : Image d'une coupe de colon analysée par logiciel*

Les images des figures 35 et 36 sont deux images d'une même zone mais traitées différemment.

Elles représentent la coupe distante d'environ 4 micromètre par rapport aux images des figures 33 et 34 du même bloc tumoral.

On constate qu'entre la figure 33 et 35 le signal est moins fort : cela signifie que les CD8 sont beaucoup moins nombreux que les CD3.

Le logiciel Immunoscore intègre les mesures réalisées sur les coupes traitées avec les anticorps anti-CD3 et les anticorps anti-CD8, calcule une moyenne et rend un résultat semi-quantitatif d'un Immunoscore compris entre 1+ et 4+.

## 2ème cas

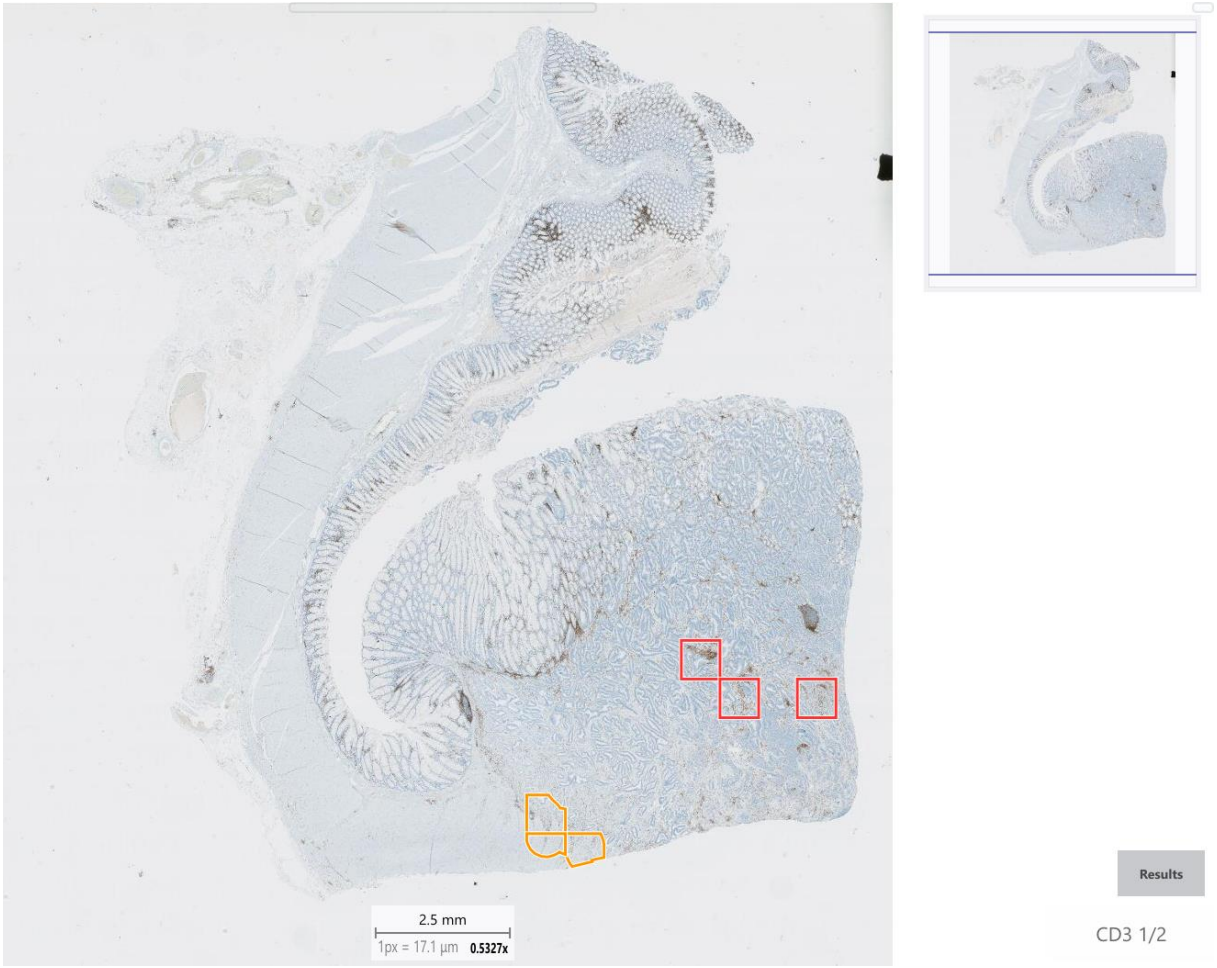



*Figure 37 : Image d'une coupe de colon*

Cette coupe est analysée avec des anticorps anti-CD3. Le résultat est natif ici.

Elle représente la présence d'un adénome, qui est une tumeur encore bénigne mais qui représente un état précancéreux.

Comme l'adénome est une tumeur bénigne elle est entourée en vert par le logiciel.

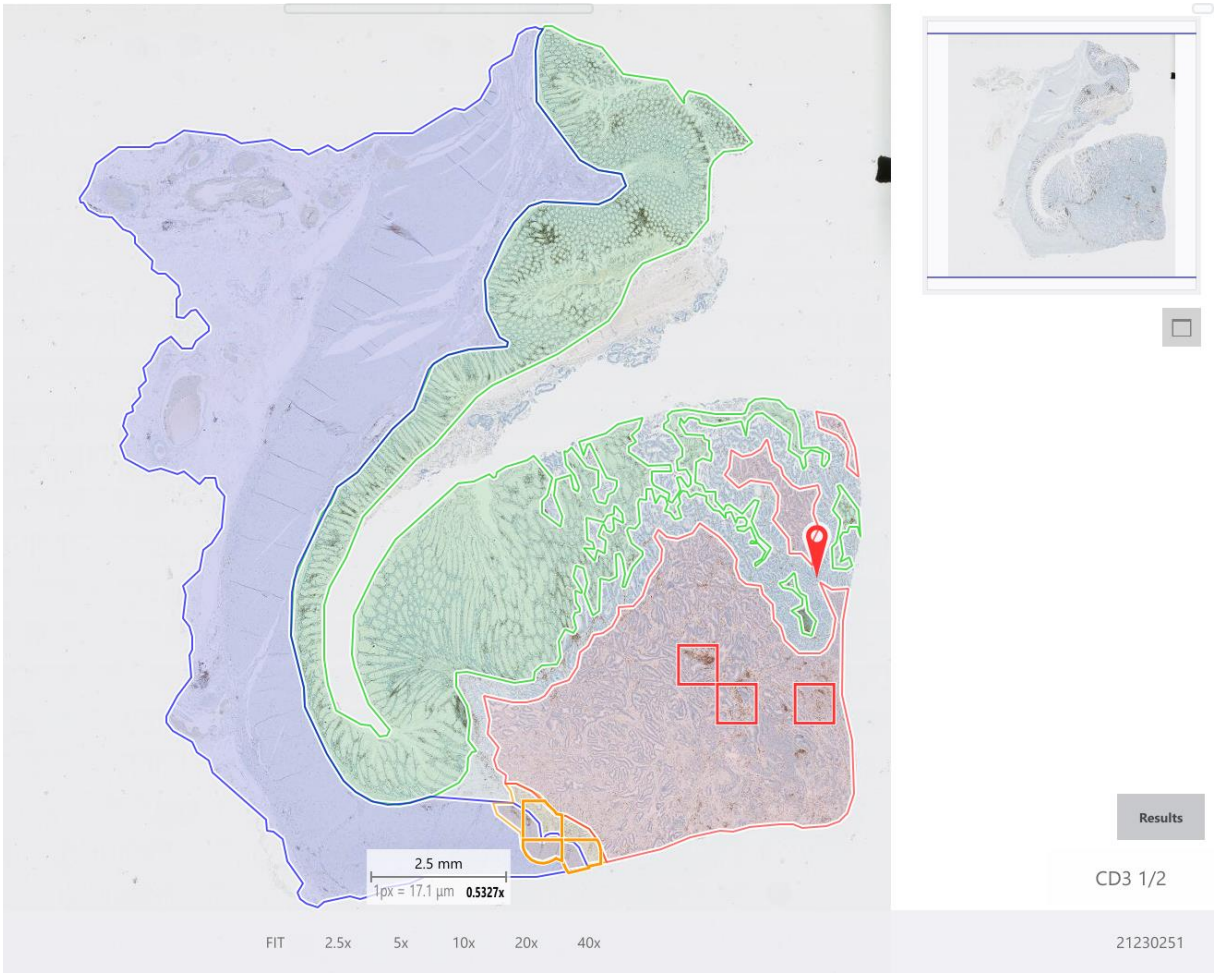



*Figure 38 : Image d'une coupe de colon, analysée par le logiciel*

Cette version avec les anticorps anti-CD3 est traitée par logiciel et corrigée par le technicien.

Sur la partie basse à droite en rouge est représenté le cancer.

Au-dessus vers la droite en vert l'adénome qui est encore bénin car cette tumeur n'a pas rompu la lame basale à partir de laquelle s'est développé le cancer.

Vers la gauche en vert, on observe uniquement de la muqueuse colique normale.

Le logiciel délimite les zones tumorales marquées en rouge, des zones épithéliales non adéno-carcinomateuses invasives qui sont colorées en vert. En jaune la ligne de front qui correspond à la zone pertinente d'analyse.

Quel que soit l'échantillon, la pertinence de l'analyse du logiciel est validée par un contrôle visuel des techniciens de l'équipe « diagnostic », le service *testing*.

En effet, le logiciel est incapable d'identifier correctement plusieurs éléments :

- les zones artefactées par la technique immunohistochimique ou la qualité de la coupe,
- les zones de nécrose tissulaire
- les foyers de suppuration qui doivent être exclues de l'analyse car il s'agit de « zone de silence »
- la différenciation parmi les zones épithéliales lésionnelles entre les zones dysplasiques (foyers adénomateux) et les zones de cancer in situ des zones où le cancer est simplement micro-invasif.

La différenciation étant souvent subtile, un médecin anatomo-pathologiste est alors sollicité par l'équipe *testing* pour délimiter au mieux les différentes zones car si l'analyse de l'échantillon est délicate le médecin individualise les zones tumorales infiltrantes et les différencie des zones saines, ou simplement adénomateuses.

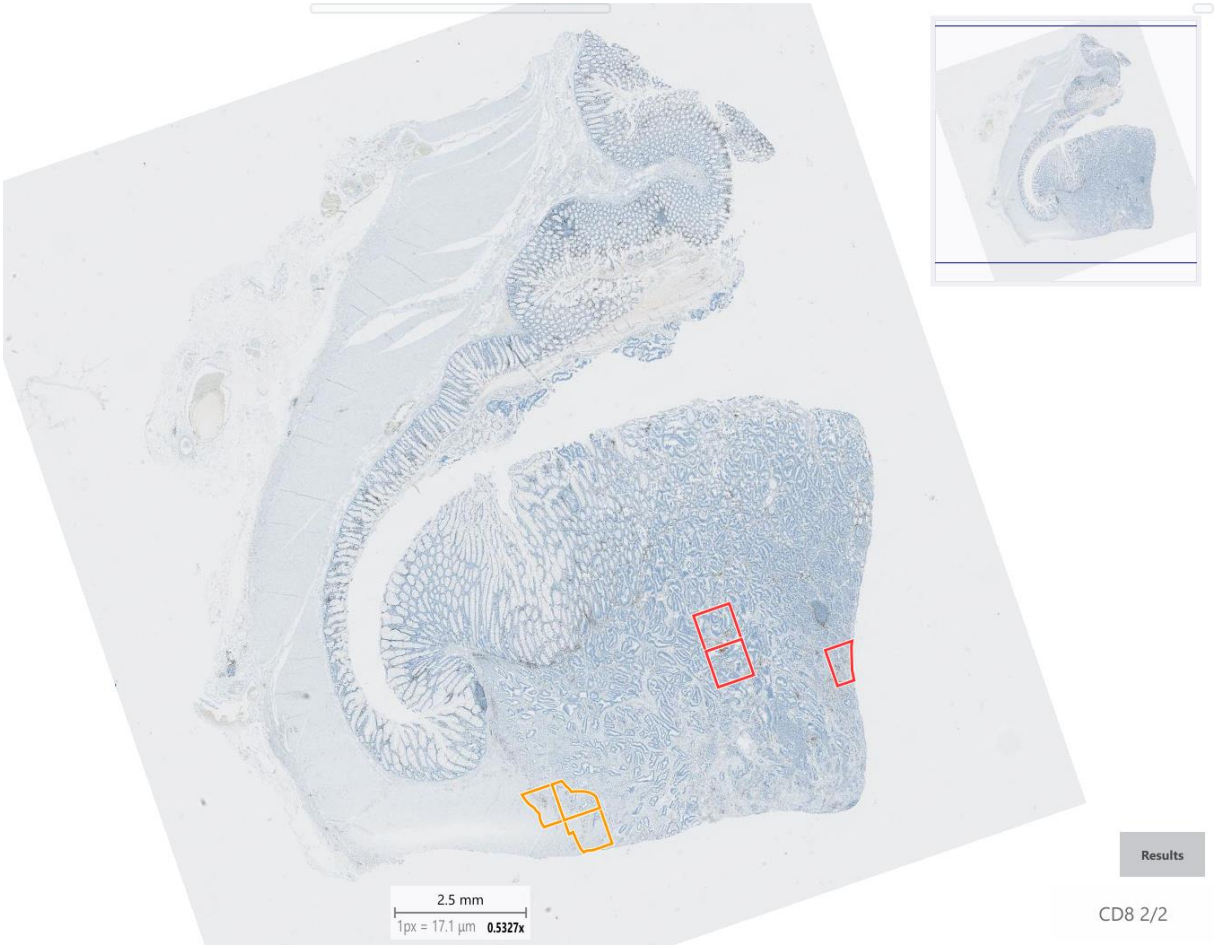



*Figure 39 : Image d'une coupe de colon*

Cette image est traitée avec des anticorps anti-CD8, et le résultat est natif.

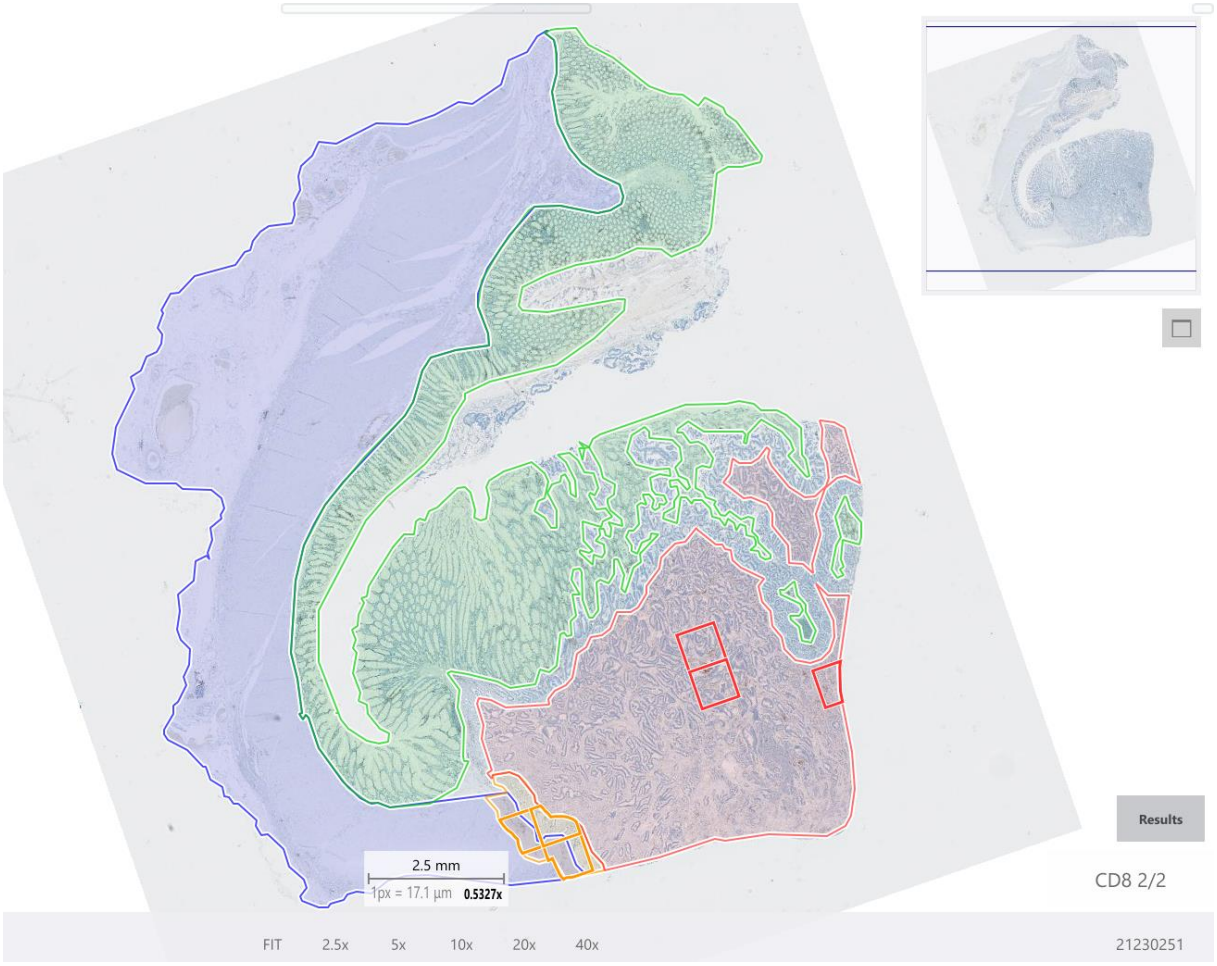



*Figure 40 : Image d'une coupe de colon, analysée par le logiciel*

Cette version avec les anticorps anti-CD8 est corrigée et revue par le technicien.

- **Dans quel cadre commercial est-il utilisé ?**

Nous vendons le kit de diagnostic aux praticiens. Le traitement de l'image se fait cependant chez HalioDx dans notre département spécialisé.

Il est prévu de proposer la solution logicielle sous forme d'une licence d'utilisation.

- **En quoi l'Immunoscore® est-il plus prédictif ?**

En clinique, certains patients avec le même TNM n'ont pourtant pas la même réponse au traitement.

L'Immunoscore est une tentative d'explication et explore l'état du système immunitaire du patient, une donnée qui n'était pas étudiée pour traiter un cancer.

Il permet de le classer en système actif (*hot* en anglais) ou inactif (*cold*).

L'intérêt est d'adapter le traitement au terrain immunitaire du patient : il n'y a pas d'intérêt à stimuler un système immunitaire non actif ou bien de recourir à un traitement lourd par chimiothérapie chez un patient ayant un système immunitaire déjà très actif.

- **Comment sont gérées les données des patients ?**

Notre service de *testing* reçoit des échantillons de patients, principalement inclus dans des essais cliniques, ou bien provenant d'échantillons commerciaux.

Les images des échantillons sont sauvegardées dans nos serveurs, malgré le poids des données qui est très lourd. Certaines sauvegardes sont faites sur des bandes qui sont scellées et envoyées chez une société spécialisée dans le stockage.

Tout est anonymisé, et nous n'avons aucune information nous permettant d'identifier un patient. Nous sommes en accord avec la loi RGPD[27] et nous sommes contrôlés par la CNIL[28].

[27] Règlement général sur la protection des données
[28] Commission nationale de l'informatique et des libertés

Le clinicien certifie que les échantillons ont été recueillis après consentement du patient.

Pour notre part nous nous assurons que nos clients ont bien recueilli les consentements.

Si un patient retire son consentement, nous supprimons toutes les données le concernant.

- **Travaillez-vous sur d'autres « scores » dans d'autres cas ?**

Oui, d'autres cas sont à l'étude au sein de notre département *testing* : les mélanomes, sarcomes, cancers ORL, cancers du sein…

Pour chaque projet de développement d'un nouveau score, la méthode est de réceptionner les échantillons dans notre département, puis de développer des séries de marquage différents, et enfin de développer des solutions logicielles qui analysent :

- la densité de marquage (la quantité de cellules exprimant le marqueur)
- l'intensité du marquage (la force d'expression du marqueur par les cellules)
- la zone des marqueurs.

Si ces études démontrent un intérêt, la prochaine étape est de développer le kit de diagnostic, soit dans notre département R&D si le demandeur est HalioDx, soit dans le département « Partnership Project » si le client est externe.

Nous sommes au début des phases cliniques pour nos clients externes et l'on a une tendance à confirmer la thèse validée sur le colon (i.e. qu'une réponse immunitaire forte est associée à un bon pronostic de rémission).

- **Utilisez-vous d'autres marqueurs que les ligands CD3 et CD8 ?**

Oui, dans un test similaire à l'Immunoscore utilisé dans le cancer du poumon : L'Immunoscore IC ® (pour Immune Checkpoint).

Il utilise CD8 et un autre marqueur : PDL1[29], un ligand qui est exprimé à la surface des cellules tumorales et immunitaire. Le logiciel utilise également d'autres signes, comme la distance entre une cellule CD8+ et une cellule PDL1+ dont l'intérêt est encore étudié.

L'outil n'est pas encore au même stade de développement que celui du cancer du côlon.

- **Quelles sont les pistes d'amélioration ?**

Dans le cas de l'Immunoscore et du colon, nous étudions la meilleure façon d'exploiter les zones. De plus notre équipe apprend au programme à distinguer une cellule tumorale d'une cellule non tumorale. En effet actuellement le logiciel sait seulement si une zone est tumorale ou non, ce qui peut manquer de précisions car à l'intérieur de cette zone se trouvent une multitude de cellules différentes, tumorales, saines, immunitaires… Pour certains besoins de nos clients nous enseignons aussi au logiciel les zones « saines » ou stroma résiduel, qui se situent dans une zone « contaminée », ce qui nous permettra aussi de limiter nos recours à l'expertise du médecin anatomopathologiste.

Présentement dans le cas du test concernant le poumon, le programme détecte la présence ou l'absence du ligand PDL1 à la surface des cellules. Ce ligand peut être exprimé par des cellules tumorales ou immunitaires. A l'avenir nous souhaitons que le programme distingue quel type cellulaire sécrète le PDL1. Nous espérons que ce score du poumon encore à l'état de recherche soit utilisé dans un kit de diagnostic approuvé et commercialisable.

Nous améliorons aussi constamment la vitesse de traitement, via l'amélioration des algorithmes et de l'*hardware* (voir ci-dessous l'entretien avec l'équipe de

[29] Program Death Ligant 1, marqueur de cellules qui ont pour rôle de freiner le système immunitaire. Les cellules tumorales le secrètent et freinent les cellules de l'immunité.

développement). Ainsi sur les dernières versions du logiciel, la durée de traitement d'une image est réduite à 4 min, contre 3 heures auparavant.

### V.3.2 Entretien avec l'équipe de développeurs

*Propos recueillis par entretien avec une partie de l'équipe de développeurs d'HalioDx, Felipe Guimarães ingénieur en sciences informatiques, et Assil Benchaaben ingénieur en imagerie médicale et traitement d'images en général.*

- **Quelle est votre formation et votre poste chez HalioDx ?**

Felipe Guimarães : - Je suis ingénieur en sciences informatiques, et je travaillais précédemment en recherche et développement au Brésil. Ici en France je travaille dans le développement informatique (*back-end* et *front-end* dans les interfaces web), stockage en ligne (*cloud*) et calcul informatique. Je suis en équipe avec Lolita, développeuse *front-end*.

Assil Benchaaben : - J'ai une licence en génie biomédical, que j'ai complétée avec un master en imagerie médical (qui concerne les technologies en radiologie : IRM, scanner ; et le traitement et design de *workflow*[30] d'images médicales) et un master informatique en traitement d'image en général. Je travaille depuis 6 ans en *digital pathology* (précédemment à Montpellier).

[30] Gestion électronique des processus métiers : modélisation et gestion informatique de l'ensemble des tâches à accomplir et des différents acteurs impliqués dans la réalisation d'un processus métier. Un processus métier représente les interactions sous forme d'échange d'informations entre divers acteurs comme des hommes, des applications ou des processus tiers.

- **Sur quels projets travaillez-vous concernant l'Immunoscore ?**

Nous avons plusieurs projets pour l'Immunoscore.

Premièrement nous avons eu le projet de ne plus utiliser l'application en local sur nos machines et de décentraliser l'application d'Immunoscore sur le web dans le *cloud* afin de déporter les calculs de traitement d'images sur des serveurs.

Pour ce projet de plateforme dans le *cloud* et pour adapter les logiciels aux besoins de la *digital pathology*, nous sommes en collaboration avec KeenEye, une entreprise parisienne qui développe une plateforme logicielle pour le stockage d'imagerie médicale. (117)

Pour cette opération il est important de noter que toute une infrastructure est nécessaire pour le calcul, le stockage, et la visualisation des données.

Deuxièmement nous avons adapté le *workflow* de l'Immunoscore local avec quelques modules supplémentaires pour apporter plus de précisions à l'Immunoscore actuel.

Nous utilisons toujours deux lames immunomarquées par des anticorps anti-CD3 et anti-CD8. Nous réalisons 3 détections différentes :

- Détections des cellules
- Détection des artefacts, ce qui est nouveau,
- Détection des ROI (acronyme de *« region of interest »* ou régions d'intérêts : tissus, tumeurs et épithéliums)

Il y a donc 3 applications qui fonctionnent en parallèle, pour 3 résultats différents.

Les détections des artefacts et ROI sont basées sur des techniques de *deep learning*, ce qui est une nouveauté. En effet nous utilisions antérieurement des algorithmes d'intelligence artificielle standards (de type *random forest*). Ces techniques de *deep learning* se sont révélés plus efficaces en matière de détection et de prédiction que les algorithmes standards.

L'outil supprime maintenant automatiquement les zones contenant des artefacts sans besoin d'une intervention humaine et ne garde que les régions d'intérêts. Ensuite

comme avec le test classique, la densité des cellules est calculée dans ces régions d'intérêts ainsi que la marge d'invasion, et le programme procède au même calcul de l'Immunoscore, et ce de la même manière pour la lame CD3+ et la lame CD8+.

- **Quelles technologies utilise le logiciel ?**

Les technologies employées par le logiciel sont le *deep learning* et la vision par ordinateur (ou en anglais *computer vision*). Elles ont des avantages par rapport à des technologies plus anciennes.

Avec le *deep learning*, l'avantage est de pouvoir s'affranchir des limites que nous devions auparavant établir pour le logiciel de reconnaissance. L'algorithme est conçu pour apprendre : nous donnons un ensemble de jeux de données et un ensemble de résultats et le programme apprendra lui-même avec la technique du *deep learning* comment arriver du postulat de départ à celui de l'arrivée. Il commence d'une manière aléatoire et parvient à un résultat via un nombre de règles (*features*).

Il existe une boucle d'apprentissage pour améliorer le résultat au fur et à mesure en exploitant un réseau de neurones. Un neurone regroupe à la fois *l'input* (le résultat d'entrée en informatique), la fonction de transformation mathématique, le poids que l'on donne à cette fonction et *l'output* (le résultat de sortie). *L'output* peut être *l'input* d'un autre neurone. Le but est de minimiser l'erreur entre *l'input* et *l'output*.

On peut employer le terme de « *blackbox* » car nous ne connaissons pas exactement la combinaison de *features* qui permettent d'obtenir le bon résultat (même si des technologies récentes permettent de l'étudier).

Avec la technologie de vision par ordinateur, qui permet aux ordinateurs de « voir » comme un être humain, nous sommes capables de designer des règles de décision pour obtenir un résultat. C'est un humain avec son expertise mathématique qui a le contrôle total d'un point de vue mathématique, ce qui est un avantage par rapport au *deep learning* car l'humain peut valider la formule mathématique utilisée pour prendre la décision.

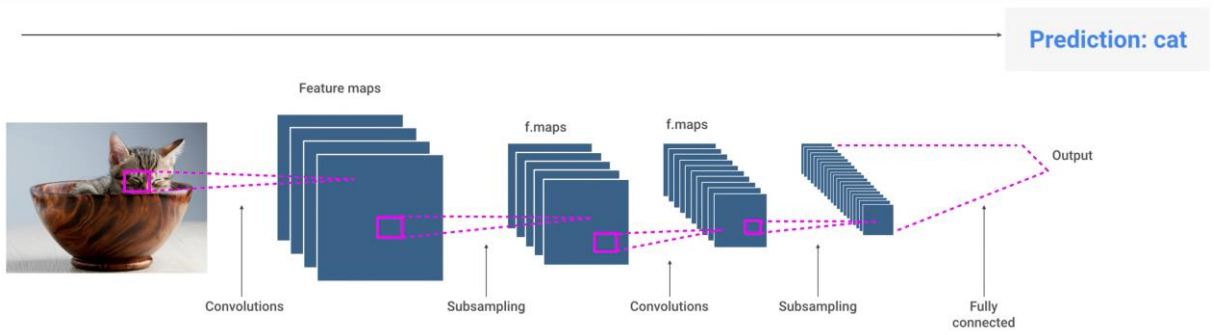



*Figure 41: L'algorithme utilise plusieurs couches de filtres pour obtenir un résultat correct © Google Cloud Tech*

- **Avez-vous dû créer des modèles statistiques ?**

Nous n'avons pas eu besoin de designer les calculs statistiques, car ils ont déjà été créés dans la bibliothèque (PyTorch) que l'on utilise. Nous concevons seulement le processus de l'algorithme en utilisant les fonctions de la bibliothèque.

- **Comment l’apprentissage du logiciel est-il possible ?**

Nous prendrons l’exemple de la détection des ROI et des artéfacts.

Pour que le logiciel ait cette capacité à les détecter nous avons procédé à un apprentissage sur 900 lames avec annotations manuelles par les pathologistes des régions d’intérêt.

Les 900 lames se composent de :

- 700 lames pour l’apprentissage du logiciel.
- 100 lames pour la validation de la performance du modèle.
  Ici le test est effectué sur des lames qui sont déjà annotées par un pathologiste et nous observons la correspondance des surfaces des zones entre celles annotées par l’humain et celles annotées par le logiciel.
  Nous procédons à une correction si nécessaire sur l’algorithme pour obtenir les mêmes résultats qu’avec l’homme et valider l’apprentissage ; avec une marge d’erreur, ou taux d’exactitude, tolérable. (Dans notre cas une exactitude de 90%). Cette marge d’erreur est calculée par validation statistique.
- 100 lames de test : le logiciel ne les a jamais vues. C’est à partir de ces 100 lames tests (également annotées par des pathologistes) que nous pouvons juger de la performance de l’algorithme. Ces lames ne font pas partie de l’apprentissage mais permettent de s’assurer que le modèle est reproductible sur une base de données qu’il n’a jamais rencontrée.

- **Quels sont les moyens d’améliorer la puissance de l’outil ?**

Nous sommes toujours dans une optique d’amélioration du modèle.

Actuellement nous avons un taux d’exactitude de 90%, ce qui veut dire que neuf fois sur dix le logiciel fera la bonne proposition et qu’une fois sur dix l’intervention de l’utilisateur sera nécessaire. Nous souhaitons augmenter ce taux d’exactitude, dans le but que l’utilisateur modifie le moins possible les zones détectées pour gagner en temps et en productivité.

Pour y parvenir, nous allons entreprendre un nouvel apprentissage sur 900 autres lames pour être plus précis et pour intégrer le nouveau marquage que nous utilisons.

900 lames sont suffisantes car la taille de chaque photographie de lame est très importante, et cette base de données d'apprentissage peut encore être augmentée grâce à des manipulations comme des translations ou des rotations de la lame par exemple. Ce sont des méthodes d'augmentation de *dataset* (bases de données).

- **Quelles sont les limites de l'automatisation ?**

Même si nous sommes arrivés à des niveaux de performances élevés, le praticien devra dans tous les cas valider visuellement la proposition du logiciel. Le tout automatique est encore loin.

De plus, nous devons garder à l'esprit que les règles de décisions prises par un modèle ne sont pas les mêmes que celles prises par un pathologiste.

En effet un modèle peut très facilement être induit en erreur. Si un logiciel de reconnaissance d'image reconnait un canapé sur une photo d'un canapé, il ne reconnaitra pas forcément ce canapé sur cette même photo si l'on ajoute un éléphant en fond.

[Citons également l'exemple de chercheurs développant un nouvel algorithme multi-neuronal répliquant le fonctionnement du cerveau humain face à divers stimuli visuels qui ont découvert que leur système peut être trompé en écrivant simplement quelque chose sur l'image (118).]

Certes, il existe des technologies qui analysent ces boites noires, mais elles sont encore en développement. Nous ne pouvons donc faire entièrement confiance à un algorithme.

En conclusion, quand la base de données n'est pas contrôlée et que l'algorithme manque de contexte, il faut être vigilant et veiller à ce que l'algorithme ne se trompe pas de voie.

Dans notre cas, nous sommes en environnement très contrôlé, avec validation des bases de données par les pathologistes.

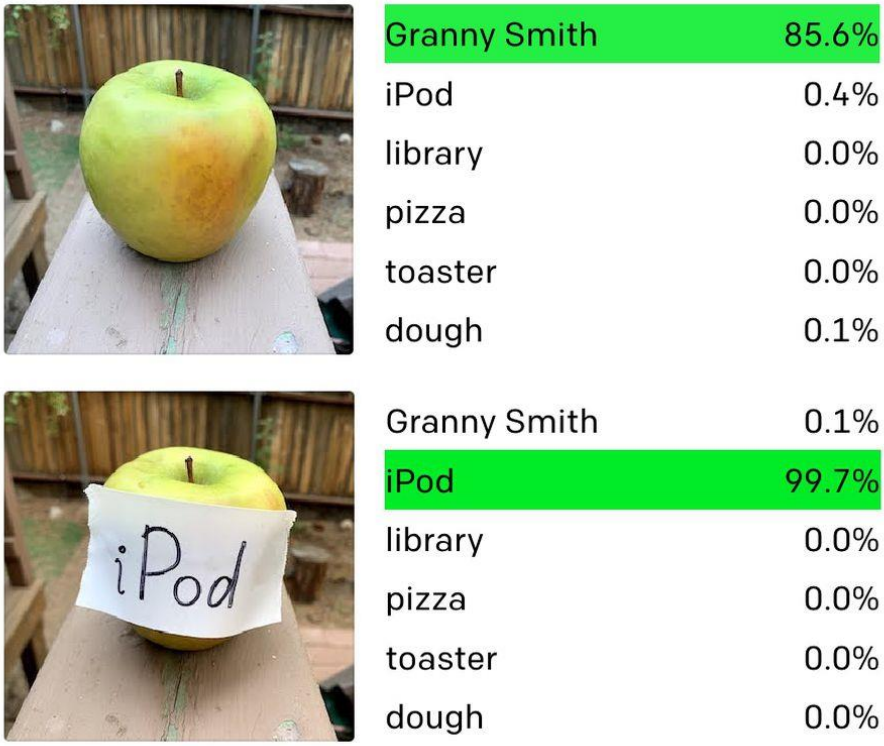



*Figure 42: Un algorithme de reconnaissance d'images est trompé par le simple ajout d'un texte écrit sur une feuille*

- **Quelles sont les infrastructures nécessaires ?**

Nous utilisions auparavant nos propres ordinateurs et serveurs sur site à Luminy, avec une version de logiciel développée par Definiens une société spécialisée dans l'identification des biomarqueurs tumoraux[31], rachetée par AstraZeneca (119). Les temps de traitement, notamment pour le comptage des cellules et la détection des régions d'intérêt étaient très longs, d'autant plus si nous souhaitions analyser plusieurs lames simultanément.

Nous avons délocalisé l'analyse des images sur des serveurs distants et utilisons une nouvelle architecture ainsi qu'un nouvel algorithme : les temps de traitements ont été réduits à 4 minutes. Notre infrastructure est maintenant modulable par rapport à nos besoins : nous pouvons utiliser plus ou moins de ressources informatiques en fonction de notre charge de travail.

[31] La technologie propriétaire de Definiens, baptisée 'phénomique des tissus', a été mise au point par le Professeur Gerd Binnig, Prix Nobel de physique en 1986.

Les serveurs que nous utilisons se situent à Paris, dans le *data center* d'Amazon que nous exploitons via leur service AWS (Amazon Web Services[32]).

Nous utilisons ceux de Paris pour rester en France, afin de respecter la souveraineté des données et les contraintes réglementaires, notamment la loi RGPD et la norme ISO 13485 (relatifs aux exigences des systèmes de management de la qualité concernant les dispositifs médicaux). Le respect de ces réglementations conditionne la validité de nos résultats.

- **Y a-t-il des formalités à réaliser auprès des autorités compétentes ?**

En France une déclaration à la CNIL est nécessaire, et le logiciel doit respecter dans le cadre de la réglementation médicale les normes ISO CE IVD. Le logiciel doit également respecter des normes de développement pour pouvoir être commercialisé.

Sur le marché américain où la firme souhaite s'implanter, il faut respecter les normes HIPAA (*Health Insurance Portability and Accountability Act*) et le déclarer à la FDA (*Food & Drug Administration)*.

On peut ajouter qu'il n'existe pas encore de normes spécifiques concernant l'IA, notamment sur les règles à établir pour remplacer un pathologiste.

- **Quelle est la place du praticien et des techniciens dans la supervision de l'Immunoscore ?**

Nous ne nous passerons pas des professionnels médicaux, mais à terme nous souhaitons que les étapes manuelles d'annotation et de comptage deviennent automatisées, pour nous dispenser d'un technicien pour le comptage des cellules, et que le travail du pathologiste soit encore plus spécialisé.

Cette évolution rejoint celle de la *digital pathology* en général.

[32] Plateforme *cloud* lancée par Amazon en 2006 qui regroupe une centaine de services répartis en diverses catégories telles que le stockage *cloud*, la puissance de calcul, l'analyse de données, l'intelligence artificielle, le développement logiciel…

# VI. DES NOUVELLES PERSPECTIVES POUR LE METIER DE PHARMACIEN

## VI.1 Le pharmacien “augmenté”

Devant les tournants technologiques que rencontrent les industries notamment technologiques et médicales, mais plus largement toute la société, l’Université Aix-Marseille évoque un développement des études en pharmacie vers l’ingénierie informatique.

Un exemple concret que pourrait apporter l’intelligence artificielle à l’officine dans le cadre de la PDA (Préparation des Doses à Administrer) :

Après avoir préparé les doses à administrer, il est nécessaire de contrôler les sachets contenant les médicaments. Aujourd’hui, nous pouvons le faire manuellement avec un contrôle visuel, mais cela nécessite beaucoup de temps et de ressources humaines, sans compter les erreurs éventuelles.

Il existe aussi des solutions de contrôle automatique via un robot qui déroule les sachets et les scanne les uns après les autres. Or ce robot rejette encore beaucoup de sachets simplement parce que les comprimés ne se présentent pas de la façon qui a été pré enregistrée par photographies, ce qui est chronophage car nécessitant encore l’intervention humaine.

Une solution serait alors d'intégrer une capacité d'apprentissage et d'expérience via la technologie des réseaux profonds pour que le système puisse apprendre à terme toutes les positions et situations (comprimés côte à côte ou l'un sur l'autre, sachet moins plat) qui validerait une prise.

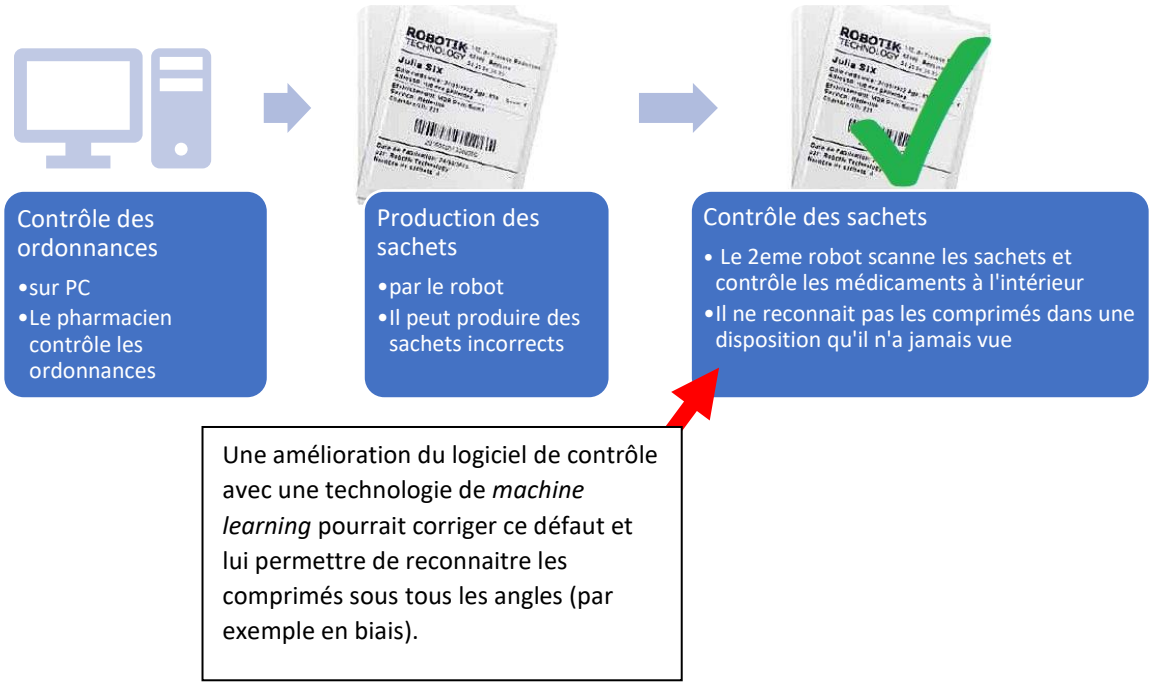



*Figure 43: Processus de la fabrication des sachets d'un pilulier*

## VI.2 Devenir « *data scientist* » ?

C'est le « métier le plus sexy du XXI[e] siècle » d'après le Harvard Business Review, l'une des plus prestigieuses revues consacrées au management (120).

Il consiste à exploiter les données accumulées (les « données massives », ou *big data*) des sociétés qui souvent aujourd'hui ne savent qu'en faire, pour extraire des nouvelles connaissances.

Le plan Big Data du ministère de l'industrie de 2014 a prévu une explosion des besoins pour ce métier : 130 000 en 2020, et 3 millions d'emplois en Europe autour de cette science de la donnée (121).

Il existe des formations mêlant mathématiques appliquées, statistique et informatique au niveau master, associées à l'environnement professionnel des laboratoires et des industriels à la pointe, comme par exemple la formation MVA (Mathématiques, Vision, Apprentissage) à l'ENS Paris-Saclay (122).

Le *data scientist* regroupe plusieurs métiers :

- Le concepteur d'algorithmes, nécessitant des bagages en maths, informatique scientifique et modélisation ;
- Le développeur informatique, qui utilise les algorithmes et des méthodes d'apprentissage, et les intègre dans des programmes informatiques ;
- Le spécialiste métier, qui manipule les outils parfaitement et les adapte au contexte d'utilisation, comme les transports, le marketing, ou la santé justement.

L'enjeu serait alors d'associer l'expertise et l'aisance avec les problématiques de santé du pharmacien à des connaissances plus techniques, par exemple via un double cursus.

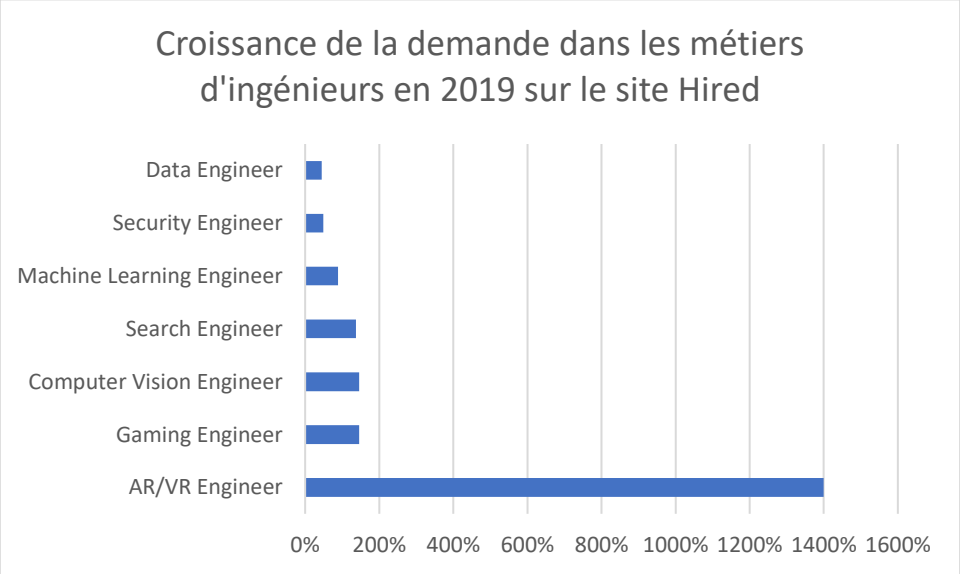



*Tableau 3: Croissance de la demande dans les métiers d'ingénieurs en 2019, D'après Hired, site de recherche d'emplois*

L'industrie est en demande : le site de recherches en métiers spécialisés Hired a vu en 2019 l'explosion de la demande pour les métiers d'ingénieurs en nouvelles technologies, avec dans le top 7 les domaines de la vision par ordinateur, la recherche, le machine learning, la sécurité informatique, et la donnée.

## VI.3 Mise en place dans l'Université Aix-Marseille

Des initiatives en ce sens naissent à l'Université Aix-Marseille, sous l'impulsion de professeurs et d'étudiants souhaitant développer le lien entre les sciences informatiques et médicales.

*Propos recueillis lors d'un entretien avec Maël Steunou, étudiant en pharmacie en filière Industrie à l'Université Aix-Marseille.*

- **Quelle est votre formation ?**

Je suis passionné d'informatique, formé en autodidacte, en 5eme pharmacie filière industrie.

Cela fait deux ans que je code sous Python avec la librairie scikit-learn : une librairie pour laquelle a été développé beaucoup de logiciels qui utilisent de l'IA. J'utilise surtout le *machine learning* et peu le *deep learning* car d'autres bibliothèques sont plus spécialisées.

J'utilise des programmes classifieurs[33], qui sont très utilisés en *machine learning*.

- **Y a-t-il des exemples de programmes que vous avez développés avec ces technologies ?**

J'ai pu en développer en stage au département oncologie de la faculté du Dr Joseph Ciccolini, qui étudiait les patients sous capécitabine, un anticancéreux utilisé dans les cancers colorectaux, de l'estomac et du sein. Une partie de la population montre une déficience enzymatique et sont incapable de métaboliser correctement le médicament, causant une grande toxicité (123).

Le but des chercheurs était d'isoler la population de patients en déficit enzymatique. Pour cela l'équipe a dosé l'enzyme par injection d'uracile qui est dégradée par la même enzyme et a enregistré le taux résiduel.

Mon rôle était d'établir la meilleure valeur limite qui définissait la déficience enzymatique chez les patients, à partir de données multicentriques, et également si

[33] De l'anglais *classifier* : programme informatique qui classe, qui trie

d'autres paramètres entraient en jeu, comme l'âge, le poids, ou d'autres rapports quantitatifs de molécules.

Pour cela j'ai standardisé les données des patients (dans une démarche dite de *feature engineering* dans le but d'optimiser les données à traiter*)*, puis je les ai entrées dans un modèle d'algorithme (déjà créé dans une bibliothèque), un *random forest classifier*[34], c'est-à-dire un classificateur qui entraine tous les arbres de décision en même temps, de manière aléatoire.

Le logiciel ainsi programmé permet de détecter 95% des patients à risque de développer un effet indésirable hématologique de stade 3 ou 4.

- **Quelles sont les études ou les initiatives qui se sont développées récemment ? Avec quels organismes ?**

Nous avons créé la spécialité Python orienté *data science*. J'initie au code des étudiants en 4ème année de pharmacie, et nous travaillons sur des projets qui répondent aux problématiques soulevées par des professeurs en pharmacie, en se basant sur des méthodes statistiques et Python.

J'ai formé une association en septembre 2019 de soutien informatique. Les étudiants sont en groupe sur des projets (comme coder un site web, analyser des données, ou programmer une application mobile).

Nous avons également un partenariat avec l'école d'ingénieurs Centrale Marseille avec un master professionnalisant que les ingénieurs peuvent faire en cours de cursus. Il y a un projet pour intégrer des pharmaciens sur des projets de programmes d'IA. Il faut des connaissances poussées en mathématiques pour les tests d'entrées, de fait certaines matières sont à approfondir dans le cursus de Pharmacie.

[34] Forêts d'arbres décisionnels : Technique d'apprentissage automatique par arbre de décision. Technique d'apprentissage supervisé qui permet de généraliser d'un échantillon vers la population générale.

- **Quelles sont les valeurs ajoutées au pharmacien par rapport à un *data scientist* classique ?**

Dans le processus de création de modèles, l'avis important est celui de l'expert dans le domaine concerné, pour pouvoir choisir les données importantes et d'amener le projet vers là où c'est intéressant.

L'intérêt d'un pharmacien formé aux sciences de la donnée est de pouvoir discuter et comprendre des ingénieurs dans les technologies de l'information (*IT*), tout en leur apportant notre savoir-faire dans la santé et la pharmacie. Dans l'industrie pharmaceutique le pharmacien peut ainsi saisir les problématiques des ingénieurs, superviser les projets, et être l'intermédiaire avec le client.

- **Quelles perspectives sont à envisager pour le pharmacien ?**

Les industries des *Big Pharma* sont très intéressées, de par le volume de données à exploiter, comme AstraZeneca qui a ouvert un pôle *data science* et IA avec des pharmaciens responsables de la plateforme (124). [Il y a également des jeunes entreprises comme Iktos, une entreprise française se basant sur l'IA pour designer de nouvelles molécules (125).]

En officine, les opportunités se développent plus lentement, mais je l'ai vu par exemple chez IBM, qui a été approché par l'équipe de développement du logiciel de gestion d'officine SmartRx pour intégrer dans leur logiciel une solution d'aide au pharmacien dans l'éducation thérapeutique du patient en utilisant l'IA.

L'intelligence artificielle pourrait donc arriver par le logiciel de gestion d'officine. Elle serait utile pour affiner les commandes ou encore dans l'analyse d'ordonnance.

# CONCLUSION

Si les sciences de l'intelligence artificielle, après de multiples passages à vide, semblent s'être cette fois-ci ancrées dans nos vies depuis les années 2010 avec la révolution des réseaux neuronaux et du *deep learning*, une véritable intelligence simulée, digne d'un animal ou d'un être humain, est encore loin.

Ce domaine a néanmoins contribué de façon notable dans de nombreux secteurs, et notamment dans celui de la santé.

Les professionnels de la santé médicales ont en effet élaboré une nouvelle forme de pratique de la médecine avec la médecine 4P, et ont pu laisser la place pour le développement de techniques issues des nouvelles technologies informatiques.

Parmi ces techniques, l'Immunoscore inventé par le Dr Jérôme Galon est aujourd'hui un nouvel outil dans la prise en charge des cancers du côlon, et permet d'aider les praticiens à choisir au mieux le traitement le plus adapté aux particularités de chaque patient.

Dans le même temps, l'intégration des algorithmes d'intelligence artificielle dans l'industrie pharmaceutique se précise : la mise au point de médicament par les industriels, mais aussi l'évolution des formations comprenant ces nouvelles technologies, permettent de mieux appréhender la médecine de demain et d'offrir des nouveaux champs de recherche, toujours dans le but d'améliorer la condition de vie humaine, à l'instar de la volonté des premiers chercheurs à vouloir mettre en commun leurs recherches qui sera à l'origine des réseaux d'Internet et du *big data*, qui bien que sujets à des dérives sécuritaires ou totalitaires, décuplent l'information disponible et propulsent la recherche pour l'Humanité.

# MEDIAGRAPHIE


1. Encyclopédie Larousse en ligne - Internet abréviation de INTERnational NETwork réseau international [Internet]. [cité 20 mars 2019]. Disponible sur: https://www.larousse.fr/encyclopedie/divers/Internet/125060

2. Leonard Kleinrock, « Information Flow in Large Communication Nets », RLE Quarterly Progress Report, Massachusetts Institute of Technology, Juillet 1961 [Internet]. [cité 17 févr 2020]. Disponible sur: https://www.lk.cs.ucla.edu/data/files/Kleinrock/Information%20Flow%20in%20Large%20Communication%20Nets.pdf

3. Serres A. Quelques repères sur l'émergence d'ARPANET. Termin Technol L'information Cult Société. 2001;(86):23-37.

4. Time sharing operating systems | PadaKuu.com [Internet]. [cité 11 mai 2021]. Disponible sur: https://padakuu.com/article/28-time-sharing-operating-systems

5. Brief History of the Internet [Internet]. Internet Society. [cité 12 févr 2020]. Disponible sur: https://www.internetsociety.org/internet/history-internet/brief-history-internet/

6. Monica 1776 Main Street Santa, California 90401-3208. Paul Baran and the Origins of the Internet [Internet]. [cité 11 févr 2020]. Disponible sur: https://www.rand.org/about/history/baran.html

7. Baran P. Reliable Digital Communications Systems Using Unreliable Network Repeater Nodes: [Internet]. 1960 [cité 11 févr 2020]. Disponible sur: https://www.rand.org/pubs/papers/P1995.html

8. Leonard Kleinrock's Home Page [Internet]. [cité 5 oct 2019]. Disponible sur: https://www.lk.cs.ucla.edu/internet_first_words.html

9. Lebrument C, Soyez F. Louis Pouzin: L'un des pères de l'internet. Economica; 2018. 170 p.

10. Official Biography: Raymond Tomlinson | Internet Hall of Fame [Internet]. [cité 21 sept 2020]. Disponible sur: https://www.internethalloffame.org//official-biography-raymond-tomlinson

11. Cerf VG, Kahn RE. A Protocol for Packet Network Intercommunication. 1974;(5):13.

12. Why Does the Net Still Work on Christmas? Paul Mockapetris | Internet Hall of Fame [Internet]. [cité 9 oct 2019]. Disponible sur: https://internethalloffame.org/blog/2012/07/23/why-does-net-still-work-christmas-paul-mockapetris

13. Robert Cailliau, l'oublié du Web [Internet]. Le Soir Plus. 2018 [cité 25 sept 2020]. Disponible sur: https://www.lesoir.be/170563/article/2018-07-30/robert-cailliau-loublie-du-web

14. Larousse É. Encyclopédie Larousse en ligne - Internet : structure du réseau [Internet]. [cité 20 mars 2019]. Disponible sur: http://www.larousse.fr/encyclopedie/images/Internet__structure_du_r%C3%A9seau/1314077

15. ITU-T Recommendation database [Internet]. ITU. [cité 4 juin 2021]. Disponible sur: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11559&lang=fr

16. Dv D. Medical Internet of Things and Big Data in Healthcare. Healthc Inform Res. 31 juill 2016;22(3):156-63.

17. The Internet of Things: Sizing up the opportunity | McKinsey [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.mckinsey.com/industries/semiconductors/our-insights/the-internet-of-things-sizing-up-the-opportunity

18. Zwolenski M, Weatherill L. The Digital Universe: Rich Data and the Increasing Value of the Internet of Things. J Telecommun Digit Econ. 30 sept 2014;2(3):9-9.

19. Managing Large Data Volumes from Scientific Facilities [Internet]. [cité 3 juin 2021]. Disponible sur: https://ercim-news.ercim.eu/en89/special/managing-large-data-volumes-from-scientific-facilities

20. Big Data et Machine Learning - 3e éd.. Les concepts et les outils de la data science - Pirmin Lemberger,Marc Batty,Médéric Morel,Jean-Luc Raffaëlli [Internet]. [cité 21 mai 2021]. Disponible sur: https://www.decitre.fr/ebooks/big-data-et-machine-learning-3e-ed-les-concepts-et-les-outils-de-la-data-science-9782100803422_9782100803422_9.html

21. Jean-Laurent Philippe. Calculer plus vite, plus haut, plus fort. Pour Sci Hors-sér. (98).

22. comment SMB the first to. Intel kills off Nervana's NNP-T chip in favor of Habana processors [Internet]. [cité 25 mai 2021]. Disponible sur: https://www.datacenterdynamics.com/en/news/intel-kills-nervanas-nnp-t-chip-favor-habana-processors/

23. Hadoop: From Experiment To Leading Big Data Platform [Internet]. InformationWeek. [cité 3 juin 2021]. Disponible sur: https://www.informationweek.com/big-data/news/software-platforms/hadoop-from-experiment-to-leading-big-data-platform/240157176

24. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6. USA: USENIX Association; 2004. p. 10. (OSDI'04).

25. Buyya R, Broberg J, Goscinski AM. Cloud Computing: Principles and Paradigms. John Wiley & Sons; 2010. 747 p.

26. Gregorio R. IoT Semantic Interoperability: Research Challenges, Best Practices, Recommendations and Next Steps [Internet]. EGM. [cité 30 mai 2021]. Disponible sur: https://www.egm.io/iot-semantic-interoperability-research-challenges-best-practices-recommendations-and-next-steps

27. Gros plan sur big data [Internet]. ISO. [cité 30 mai 2021]. Disponible sur: https://www.iso.org/cms/render/live/fr/sites/isoorg/contents/news/2014/03/Ref1821.html

28. Esante [Internet]. InterHop. 2021 [cité 4 juin 2021]. Disponible sur: https://interhop.org/projets/esante

29. Gouvernement Français. COMMUNIQUE DE PRESSE : Le Gouvernement annonce sa stratégie nationale pour le Cloud. 2021 mai.

30. #LeBrief. L'ensemble des services de Google était en panne pendant près d'une heure hier [Internet]. Next INpact. 2020 [cité 25 mai 2021]. Disponible sur: https://www.nextinpact.com/lebrief/45076/lensemble-services-google-ont-ete-en-panne-pendant-pres-dune-heure-hier

31. Hermann V. Amazon S3 : c'est une erreur humaine qui a provoqué la panne de plusieurs heures [Internet]. Next INpact. 2017 [cité 25 mai 2021]. Disponible sur: https://www.nextinpact.com/article/25768/103533-amazon-s3-cest-erreur-humaine-qui-a-provoque-panne-plusieurs-heures

32. Tout sur les systèmes d'information - Jean-François Pillou - Librairie Eyrolles [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.eyrolles.com/Informatique/Livre/tout-sur-les-systemes-d-information-9782100502769/

33. The CIA Triad — Confidentiality, Integrity, and Availability Explained [Internet]. freeCodeCamp.org. 2020 [cité 4 juin 2021]. Disponible sur: https://www.freecodecamp.org/news/the-cia-triad-confidentiality-integrity-and-availability-explained/

34. Epitech. Big data et cybersécurité : quels sont les risques ? [Internet]. Futura. [cité 4 juin 2021]. Disponible sur: https://www.futura-sciences.com/tech/dossiers/internet-big-data-boom-donnees-numeriques-1936/page/4/

35. The Paris Agreement | UNFCCC [Internet]. [cité 16 mai 2021]. Disponible sur: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

36. François Schneider. L'Effet Rebond. L'Ecologiste Ed Fr Ecol. 2003;4(11):45.

37. Microsoft finds underwater datacenters are reliable, practical and use energy sustainably [Internet]. Innovation Stories. 2020 [cité 26 oct 2020]. Disponible sur: https://news.microsoft.com/innovation-stories/project-natick-underwater-datacenter/

38. Fredericia, Danemark – Centres de données – Google [Internet]. Centres de données Google. [cité 26 oct 2020]. Disponible sur: https://www.google.com/intl/fr/about/datacenters/locations/fredericia/

39. EUR-Lex - 32016R0679 - EN - EUR-Lex [Internet]. [cité 25 mai 2021]. Disponible sur: https://eur-lex.europa.eu/eli/reg/2016/679/oj/fra

40. CNIL, CNOM. Guide pratique sur la protection des données personnelles. juin 2018;40.

41. Le Cloud Act, une nouvelle loi qui renforce l'ingérence des autorités américaines sur les opérateurs de Cloud des US [Internet]. Les Echos. 2018 [cité 4 juin 2021]. Disponible sur: https://www.lesechos.fr/idees-debats/cercle/le-cloud-act-une-nouvelle-loi-qui-renforce-lingerence-des-autorites-americaines-sur-les-operateurs-de-cloud-des-us-131174

42. Rees M. Health Data Hub : le Conseil d'État exige des correctifs face au risque de surveillance américaine [Internet]. 2020 [cité 26 oct 2020]. Disponible sur:

https://www.nextinpact/article/44169/health-data-hub-conseil-detat-exige-correctifs-face-au-risque-surveillance-americaine

43. La CNIL, c'est quoi ? | Besoin d'aide | CNIL [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.cnil.fr/fr/cnil-direct/question/la-cnil-cest-quoi

44. Comprendre le RGPD | CNIL [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.cnil.fr/fr/comprendre-le-rgpd

45. Qu'est-ce le Privacy By Design ? [Internet]. Données & RGPD. 2019 [cité 4 juin 2021]. Disponible sur: https://donnees-rgpd.fr/definitions/privacy-by-design/

46. Les stratégies des GAFAM et des BigTech dans la santé [Internet]. Les Echos Études. [cité 4 juin 2021]. Disponible sur: https://www.lesechos-etudes.fr/etude/gafam-bigtech-sante/

47. Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery [Internet]. 2009 [cité 4 juin 2021]. Disponible sur: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

48. Epitech. Open data : l'open science ou la science ouverte [Internet]. Futura. [cité 4 juin 2021]. Disponible sur: https://www.futura-sciences.com/tech/dossiers/internet-big-data-boom-donnees-numeriques-1936/page/5/

49. Comment le Big data révolutionne la recherche en santé [Internet]. Institut Pasteur. 2018 [cité 4 juin 2021]. Disponible sur: https://www.pasteur.fr/fr/journal-recherche/dossiers/comment-big-data-revolutionne-recherche-sante

50. Arthur Samuel [Internet]. [cité 20 oct 2019]. Disponible sur: http://infolab.stanford.edu/pub/voy/museum/samuel.html

51. Dornbusch P. News sur les échecs [Internet]. [cité 2 juin 2021]. Disponible sur: http://www.chess-and-strategy.com/

52. AlphaGo: The story so far [Internet]. Deepmind. [cité 26 mai 2021]. Disponible sur: /research/case-studies/alphago-the-story-so-far

53. TURING AM. I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind. 1 oct 1950;LIX(236):433-60.

54. 1956 : Et l'intelligence artificielle devint science | Les Echos [Internet]. [cité 22 oct 2019]. Disponible sur: https://www.lesechos.fr/2017/08/1956-et-lintelligence-artificielle-devint-science-1116632

55. Yann Le Cun. Quand la machine apprend. Odile Jacob. 2019.

56. Théorie des probabilités : définition et explications [Internet]. Techno-Science.net. [cité 4 mars 2021]. Disponible sur: https://www.techno-science.net/definition/6387.html

57. Algèbre linéaire - Définition et Explications [Internet]. Techno-Science.net. [cité 4 mars 2021]. Disponible sur: https://www.techno-science.net/glossaire-definition/Algebre-lineaire.html

58. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Red Hook, NY, USA: Curran Associates Inc.; 2012. p. 1097-105. (NIPS'12).

59. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and Improving the Image Quality of StyleGAN. ArXiv191204958 Cs Eess Stat [Internet]. 23 mars 2020 [cité 2 juin 2021]; Disponible sur: http://arxiv.org/abs/1912.04958

60. Lancement de l'initiative scikit-learn, bibliothèque logicielle de référence en machine learning | Inria [Internet]. [cité 5 mars 2021]. Disponible sur: https://www.inria.fr/fr/lancement-de-linitiative-scikit-learn

61. Arute F, Arya K, Babbush R, Bacon D, Bardin JC, Barends R, et al. Quantum supremacy using a programmable superconducting processor. Nature. oct 2019;574(7779):505-10.

62. Quand l'ordinateur quantique de Google se fait damer le pion par un PC classique. Le Monde.fr [Internet]. 7 déc 2020 [cité 25 mai 2021]; Disponible sur: https://www.lemonde.fr/sciences/article/2020/12/07/quand-l-ordinateur-quantique-de-google-se-fait-damer-le-pion-par-un-pc-classique_6062533_1650684.html

63. Zhou Y, Stoudenmire EM, Waintal X. What Limits the Simulation of Quantum Computers? Phys Rev X. 23 nov 2020;10(4):041038.

64. Du pixel à la vision par ordinateur : introduction à l'analyse... [Internet]. [cité 2 juin 2021]. Disponible sur: https://www.edimark.fr/correspondances-onco-theranostic/pixel-a-vision-par-ordinateur-introduction-a-analyse-images

65. Geitgey A. Machine Learning is Fun! Part 3: Deep Learning and Convolutional Neural Networks [Internet]. Medium. 2020 [cité 3 juin 2021]. Disponible sur: https://medium.com/@ageitgey/machine-learning-is-fun-part-3-deep-learning-and-convolutional-neural-networks-f40359318721

66. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. ArXiv150601497 Cs [Internet]. 6 janv 2016 [cité 3 juin 2021]; Disponible sur: http://arxiv.org/abs/1506.01497

67. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. nov 1998;86(11):2278-324.

68. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. ArXiv190308297 Cs Stat [Internet]. 19 mars 2019 [cité 4 juin 2021]; Disponible sur: http://arxiv.org/abs/1903.08297

69. Using AI to help find answers to common skin conditions [Internet]. Google. 2021 [cité 4 juin 2021]. Disponible sur: https://blog.google/technology/health/ai-dermatology-preview-io-2021/

70. A deep learning system for differential diagnosis of skin diseases | Nature Medicine [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.nature.com/articles/s41591-020-0842-3

71. Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, et al. Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care

Physicians and Nurse Practitioners in Teledermatology Practices. JAMA Netw Open. 28 avr 2021;4(4):e217249.

72. Intelligence artificielle et discrimination : tout savoir sur les biais de l'IA [Internet]. Formation Data Science | DataScientest.com. 2020 [cité 4 juin 2021]. Disponible sur: https://datascientest.com/intelligence-artificielle-biais-ia

73. Google. Building a More Equitable Camera | Google [Internet]. 2021 [cité 4 juin 2021]. Disponible sur: https://www.youtube.com/watch?v=2DXY9cR7vN4&t=3s

74. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters [Internet]. 10 oct 2018 [cité 4 juin 2021]; Disponible sur: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

75. Les algorithmes reproduisent les stéréotypes sexistes de notre société [Internet]. Les Echos Start. 2019 [cité 4 juin 2021]. Disponible sur: https://start.lesechos.fr/innovations-startups/tech-futur/les-algorithmes-reproduisent-les-stereotypes-sexistes-de-notre-societe-1175518

76. AI for humanity [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.aiforhumanity.fr

77. Sandrine Rigaud. Cash Investigation. Au secours, mon patron est un algorithme. 2019.

78. Transformation numérique de l'industrie française : quels impacts sur les métiers, l'emploi et la formation ? [Internet]. [cité 3 juin 2021]. Disponible sur: https://syntec-numerique.fr/actu-informatique/transformation-numerique-industrie-francaise-quels-impacts-sur-metiers-emploi

79. Managing the black box of artificial intelligence (AI) [Internet]. Deloitte United States. [cité 25 mai 2021]. Disponible sur: https://www2.deloitte.com/us/en/pages/advisory/articles/black-box-artificial-intelligence.html

80. CDS Coalition. Voluntary Industry Guidelines for the Design of Medium Risk Clinical Decision Support Software to Assure the Central Role of Healthcare Professionals in Clinical Decision-Making [Internet]. 2017. Disponible sur: http://cdscoalition.org/wp-content/uploads/2017/08/CDS-3060-Guidelines-Final-2.pdf

81. MÉDECINE PRÉDICTIVE [Internet]. SFMPP. [cité 3 nov 2019]. Disponible sur: https://www.sfmpp.org/medecine-predictive/

82. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. mars 2011;8(3):184-7.

83. Marie-Odile Safon. La e-santé : télésanté, santé numérique ou santé connectée. 2019;377.

84. Dépistage de la Covid-19 : saisie dans SI-DEP des tests antigéniques [Internet]. [cité 23 avr 2021]. Disponible sur: https://www.ameli.fr/medecin/actualites/depistage-de-la-covid-19-saisie-dans-si-dep-des-tests-antigeniques

85. WHO Global Observatory for eHealth, World Health Organization. MHealth: new horizons for health through mobile technologies. [Internet]. Geneva: World Health Organization; 2011 [cité 25 mai 2021]. Disponible sur: http://www.who.int/goe/publications/goe_mhealth_web.pdf

86. Hausse de 85% de l'audience des applications de santé [Internet]. MMAF. 2020 [cité 4 juin 2021]. Disponible sur: https://www.mobilemarketing.fr/hausse-de-85-des-applications-de-sante/

87. OMS | Prévention des maladies chroniques: un investissement vital [Internet]. WHO. World Health Organization; [cité 25 mai 2021]. Disponible sur: http://www.who.int/chp/chronic_disease_report/part1/fr/

88. Dépakine : Sanofi mis en examen pour « homicides involontaires » [Internet]. Les Echos. 2020 [cité 25 mai 2021]. Disponible sur: https://www.lesechos.fr/industrie-services/pharmacie-sante/sanofi-mis-en-examen-pour-homicides-involontaires-1228189

89. Uem JMT van, Maier KS, Hucker S, Scheck O, Hobert MA, Santos AT, et al. Twelve-week sensor assessment in Parkinson's disease: Impact on quality of life. Mov Disord. 2016;31(9):1337-8.

90. Microsoft and Adaptive Biotechnologies announce partnership using AI to decode immune system; diagnose, treat disease [Internet]. The Official Microsoft Blog. 2018 [cité 3 nov 2019]. Disponible sur: https://blogs.microsoft.com/blog/2018/01/04/microsoft-adaptive-biotechnologies-announce-partnership-using-ai-decode-immune-system-diagnose-treat-disease/

91. Adaptive Biotechnologies and Microsoft launch groundbreaking ImmuneCODE database to share populationwide immune response to the COVID-19 virus [Internet]. Stories. 2020 [cité 28 mai 2021]. Disponible sur: https://news.microsoft.com/2020/06/11/adaptive-biotechnologies-and-microsoft-launch-groundbreaking-immunecode-database-to-share-populationwide-immune-response-to-the-covid-19-virus/

92. Loupy A, Aubert O, Orandi BJ, Naesens M, Bouatou Y, Raynaud M, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. BMJ. 17 sept 2019;366:l4923.

93. SANTÉ 2030 : une analyse prospective de l'innovation en santé [Internet]. [cité 4 juin 2021]. Disponible sur: https://www.leem.org/publication/sante-2030-une-analyse-prospective-de-linnovation-en-sante

94. Jonveaux P. Médecine prédictive. La lettre de L'hépato-Gastroentérologue. III(6):3.

95. Comité Consultatif National d'Ethique. PROBLEMES ETHIQUES POSES PAR DES DEMARCHES DE PREDICTION FONDEES SUR LA DETECTION DE TROUBLES PRECOCES DU COMPORTEMENT CHEZ L'ENFANT [Internet]. 2007. Report No.: 95. Disponible sur: https://www.ccne-ethique.fr/sites/default/files/publications/avis095.pdf

96. Cancer [Internet]. [cité 26 janv 2021]. Disponible sur: https://www.who.int/fr/news-room/fact-sheets/detail/cancer

97. Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012. Etude à partir des registres des cancers du réseau Francim - Partie 1 : tumeurs solides [Internet]. [cité 22 août 2019]. Disponible sur: https://www.santepubliquefrance.fr/docs/estimation-nationale-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-entre-1980-et-2012.-etude-a-partir-des-registres-des-cancers-du-reseau

98. Song M, Wu K, Meyerhardt JA, Ogino S, Wang M, Fuchs CS, et al. Fiber Intake and Survival After Colorectal Cancer Diagnosis. JAMA Oncol. 1 janv 2018;4(1):71.

99. Futura. Côlon [Internet]. Futura. [cité 22 août 2019]. Disponible sur: https://www.futura-sciences.com/sante/definitions/biologie-colon-6869/

100. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, et al. Chimeric Antigen Receptor T Cells for Sustained Remissions in Leukemia. N Engl J Med. 16 oct 2014;371(16):1507-17.

101. SPF. Evaluation épidémiologique du programme de dépistage organisé du cancer colorectal en France. Résultats 2009-2010 [Internet]. [cité 20 févr 2021]. Disponible sur: /maladies-et-traumatismes/cancers/cancer-du-colon-rectum/evaluation-epidemiologique-du-programme-de-depistage-organise-du-cancer-colorectal-en-france.-resultats-2009-2010

102. VIDAL - Cancer colorectal - Prise en charge [Internet]. [cité 26 août 2019]. Disponible sur: https://www.vidal.fr/recommandations/3506/cancer_colorectal/prise_en_charge/

103. Stades du cancer colorectal - Cancer du côlon [Internet]. [cité 22 août 2019]. Disponible sur: https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-colon/Stades-du-cancer-colorectal

104. Mlecnik B, Bindea G, Kirilovsky A, Angell HK, Obenauf AC, Tosolini M, et al. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. Sci Transl Med. 24 févr 2016;8(327):327ra26.

105. Saito T, Nishikawa H, Wada H, Nagano Y, Sugiyama D, Atarashi K, et al. Two FOXP3 + CD4 + T cell subpopulations distinctly control the prognosis of colorectal cancers. Nat Med. juin 2016;22(6):679-84.

106. Recommandations Cancer colorectal [Internet]. VIDAL. [cité 25 janv 2021]. Disponible sur: https://www.vidal.fr/

107. Institut National du Cancer. Evaluation médico-économique du dépistage du cancer colorectal / Revue de la littérature «Études de coûts ». 2019 avr.

108. Canard JM. Le coût de la coloscopie de prévention et du CCR. 2015 déc 5; Clinique du Trocadéro, Paris.

109. STIVARGA 40 mg cp pellic [Internet]. VIDAL. [cité 25 janv 2021]. Disponible sur: https://www.vidal.fr/

110. Loriot M-A, Ciccolini J, Thomas F, Barin-Le-Guellec C, Royer B, Milano G, et al. Dépistage du déficit en dihydropyrimidine déshydrogénase (DPD) et sécurisation des chimiothérapies à base de fluoropyrimidines : mise au point et recommandations nationales du GPCO-Unicancer et du RNPGx. Bull Cancer (Paris). 1 avr 2018;105(4):397-407.

111. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou F-S, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. The Lancet. 26 mai 2018;391(10135):2128-39.

112. About Digital Pathology [Internet]. [cité 12 mars 2020]. Disponible sur: https://digitalpathologyassociation.org/about-digital-pathology

113. François-Xavier Frenois. Pathologie digitale : fondamentaux technologiques. Corresp En Onco-Théranostic. mars 2019;VIII(1):8-20.

114. IPSOGEN : une société spécialisée dans la conception de Biopuces [Internet]. [cité 29 oct 2019]. Disponible sur: https://www.gazettelabo.fr/archives/prives/2000/52IPSOGEN.html

115. HalioDx [Internet]. Marseille Immunopole. [cité 29 oct 2019]. Disponible sur: https://marseille-immunopole.org/entreprises/halio-dx/

116. https://www.haliodx.com, 2018 halioDx. HalioDx will develop new assays for QIAGEN in the frame of a multi–year and multi–project agreement [Internet]. HalioDx. [cité 29 oct 2019]. Disponible sur: https://www.haliodx.com/company/news/detail/News/haliodx-will-develop-new-assays-for-qiagen-in-the-frame-of-a-multi-year-and-multi-project-agreem/

117. HalioDx et Keen Eye s'associent pour enrichir Immunoscore d'une composante « IA » [Internet]. [cité 4 juill 2019]. Disponible sur: https://www.gazettelabo.fr/breves/8192HalioDx-KeenEye-enrichir-Immunoscore-composante-IA.html

118. Multimodal Neurons in Artificial Neural Networks [Internet]. OpenAI. 2021 [cité 9 mai 2021]. Disponible sur: https://openai.com/blog/multimodal-neurons/

119. MedImmune completes acquisition of Definiens - AstraZeneca [Internet]. [cité 4 mai 2021]. Disponible sur: https://www.astrazeneca.com/media-centre/press-releases/2014/medimmune-definiens-aquisition-tissue-phenomics-26112014.html

120. Davenport TH, Patil DJ. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review [Internet]. 1 oct 2012 [cité 17 oct 2019];(October 2012). Disponible sur: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

121. Big Data : la feuille de route entre en action [Internet]. [cité 17 oct 2019]. Disponible sur: http://www.economie.gouv.fr/big-data-feuille-route-en-action

122. TROUVE A. MATH - M2 MVA Mathématiques / Vision / Apprentissage [Internet]. [cité 17 oct 2019]. Disponible sur: http://math.ens-paris-saclay.fr/version-francaise/formations/master-mva/

123. Masson E. Toxicité sévère à la capécitabine liée à un déficit enzymatique en dihydropyrimidine déshydrogénase (DPD) [Internet]. EM-Consulte. [cité 8 mars 2021]. Disponible sur: https://www.em-consulte.com/article/248018

124. Data Science & Artificial Intelligence: Unlocking new science insights [Internet]. [cité 8 mars 2021]. Disponible sur: https://www.astrazeneca.com/r-d/data-science-and-ai.html

125. Médicaments du futur : Iktos explore l'espace chimique grâce à l'IA [Internet]. Les Echos. 2020 [cité 8 mars 2021]. Disponible sur: https://www.lesechos.fr/industrie-services/pharmacie-sante/medicaments-du-futur-iktos-explore-lespace-chimique-grace-a-lia-1182596

# TABLE DES ILLUSTRATIONS

## SERMENT DE GALIEN

***Je jure, en présence de mes maîtres de la Faculté, des conseillers de l'Ordre des pharmaciens et de mes condisciples :***

- ***D'honorer ceux qui m'ont instruit dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle à leur enseignement.***

- ***D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de l'honneur, de la probité et du désintéressement.***

- ***De ne jamais oublier ma responsabilité et mes devoirs envers le malade et sa dignité humaine, de respecter le secret professionnel.***

- ***En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser des actes criminels.***

***Que les hommes m'accordent leur estime si je suis fidèle à mes promesses.***

***Que je sois couvert d'opprobre, méprisé de mes confrères, si j'y manque.***