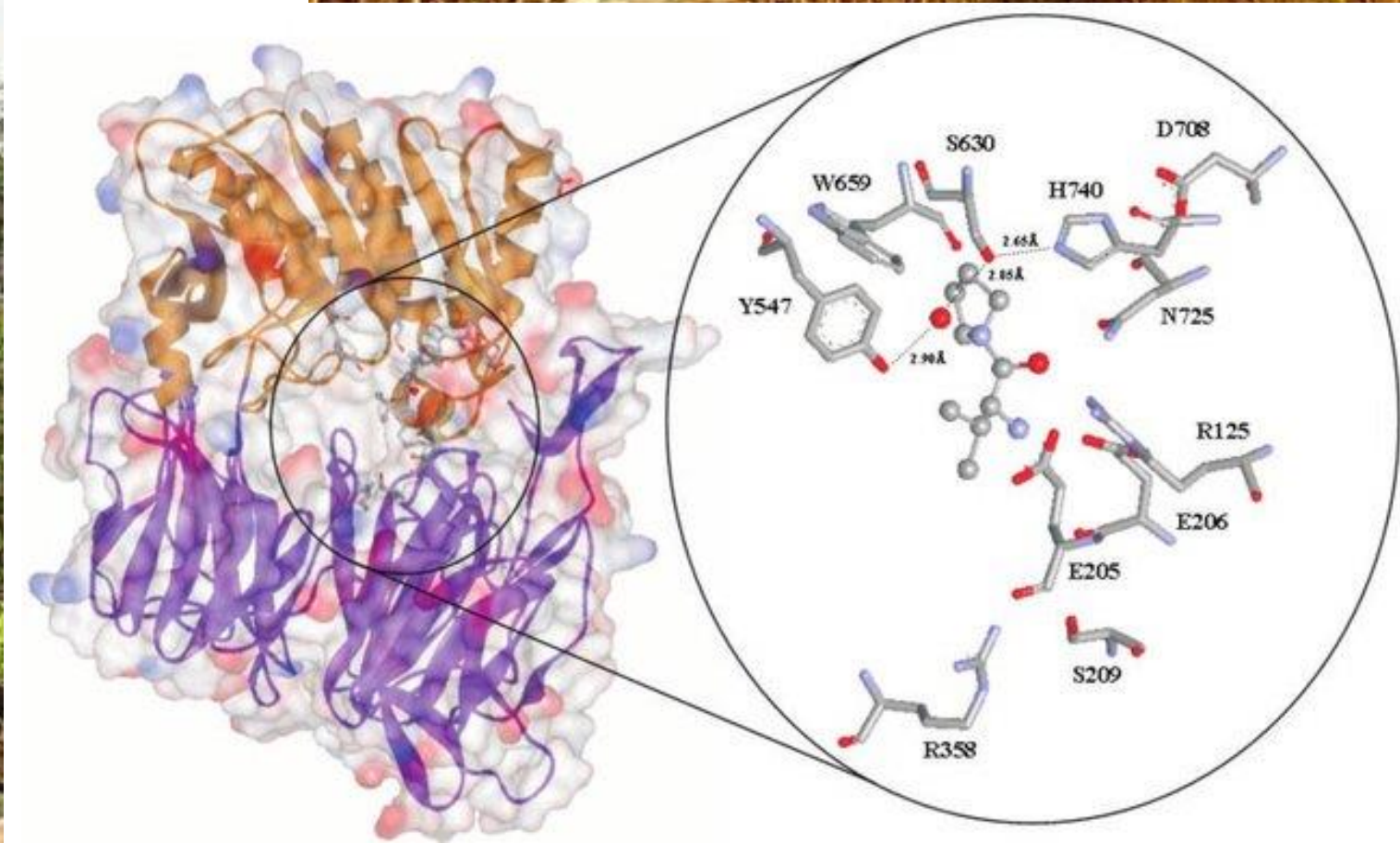
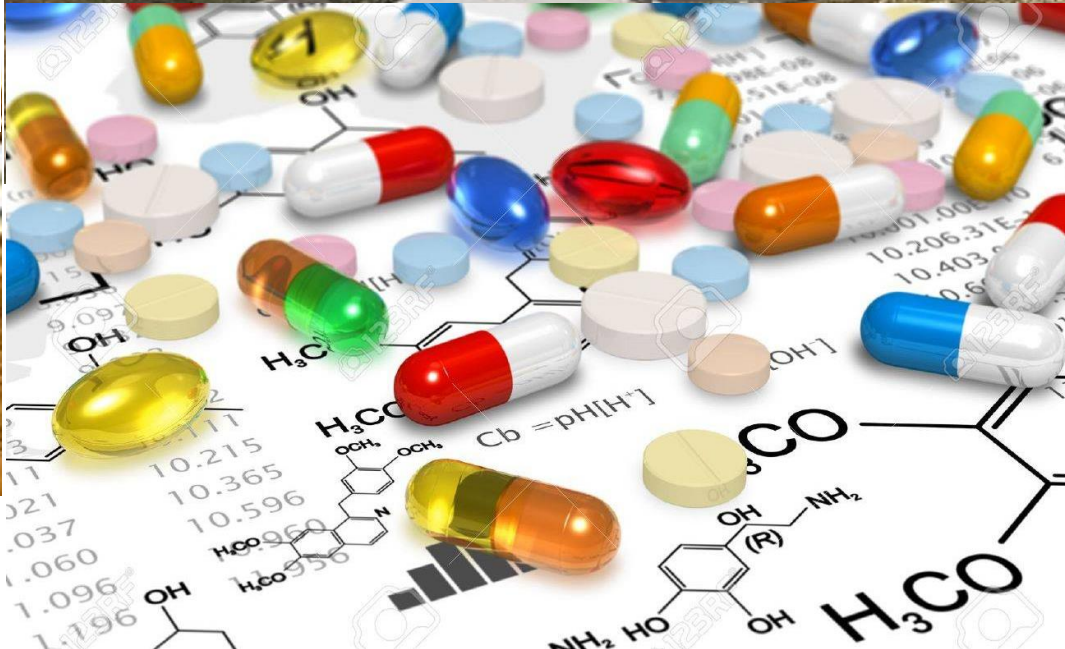
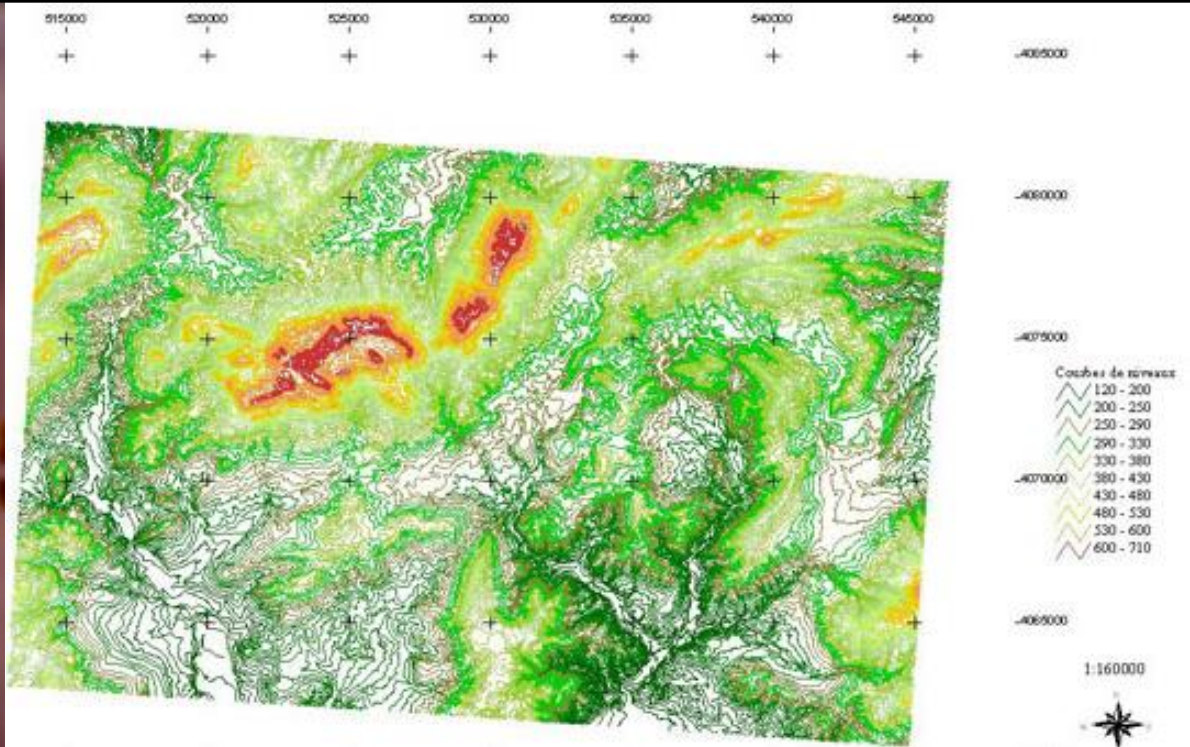
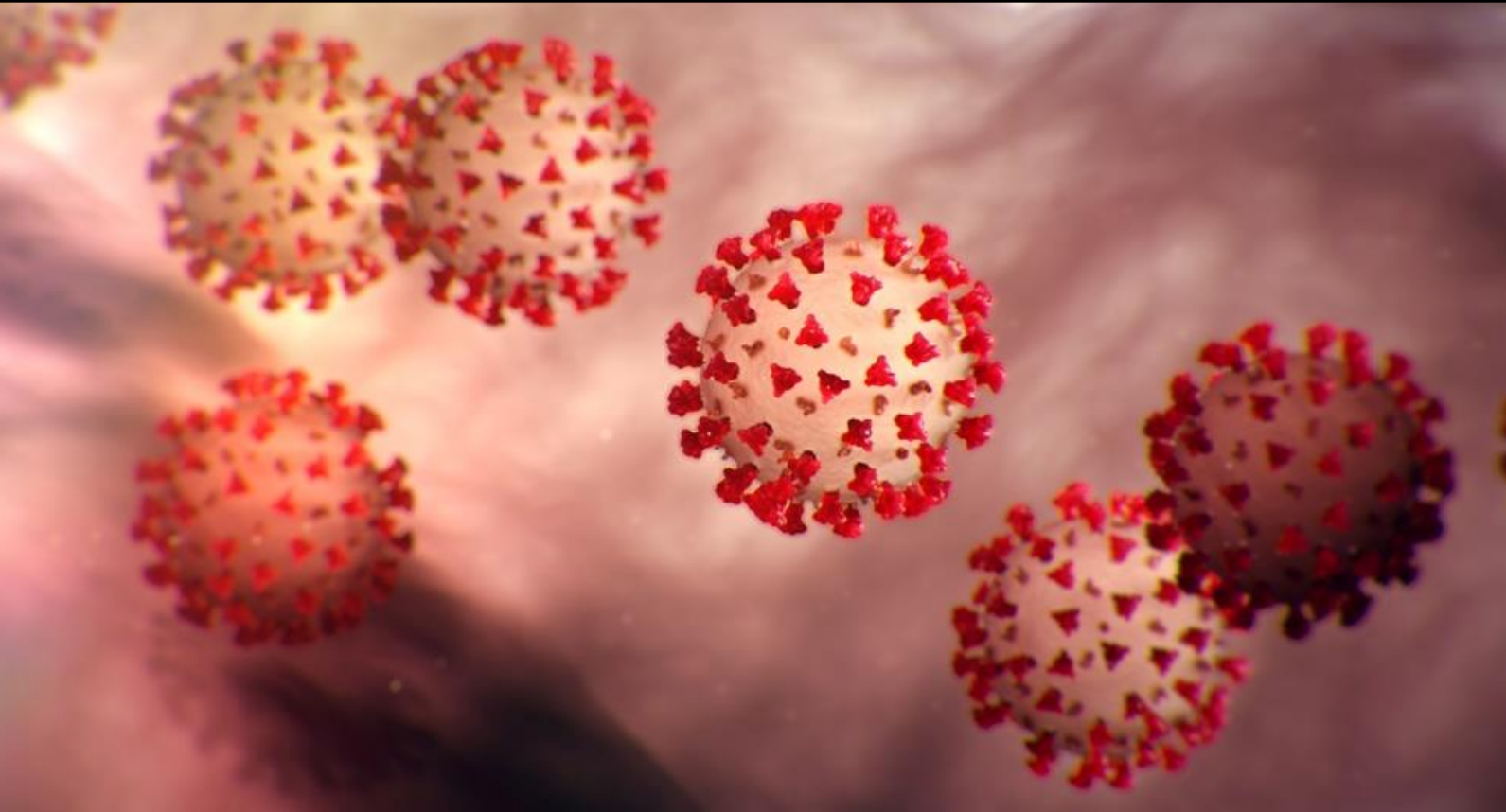


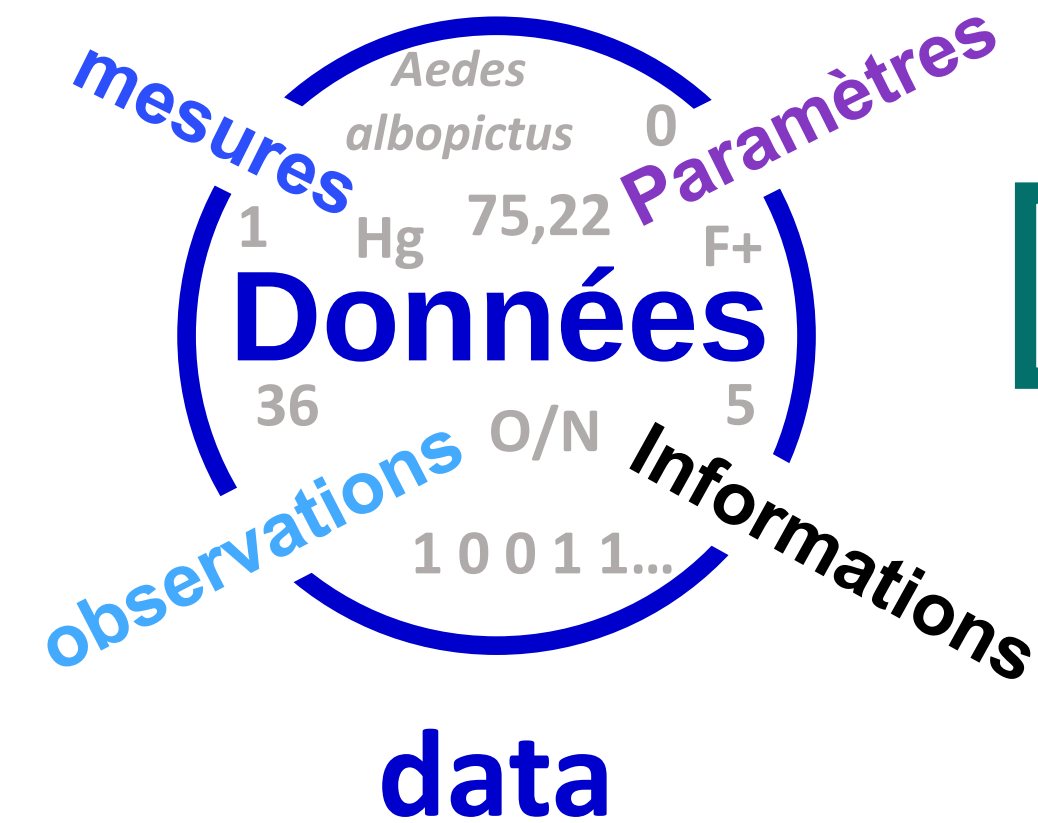


La Data Science

Qu'est-ce que c'est ? A quoi ça sert ? D'où ça vient ?

La DATA SCIENCE qu'est-ce que c'est, à quoi ça sert ?

Produire des **connaissances** en faisant parler les **données**



data

- Les data dans « data science »
- Terrain, labos, entreprises,...
- Informations diverses, archives
- Satellites, réseaux, Web

[ordinateurs]

Data Science

[outils & disciplines]

**Produire
Connaissances**

Ça marche!

Connaissances

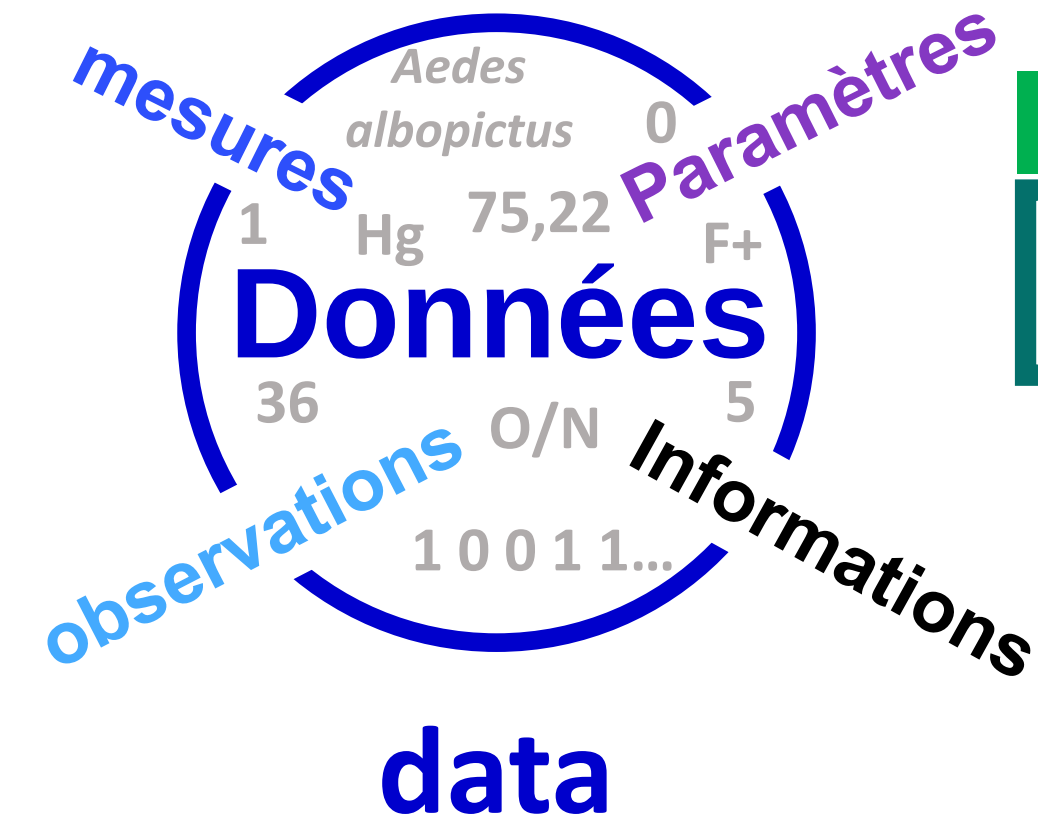
- **Décisions**
- **Actions**
- **Contrôles**
- **Prévisions**

*Pilotage voiture autonome,
Développement médicament,
Fidélisation clientèle,...*

▷ Approche multidisciplinaire dont les applications dépendent de la puissance des ordinateurs utilisés

La DATA SCIENCE qu'est-ce que c'est, à quoi ça sert ?

Produire des **connaissances** en faisant parler les **données**



- Les data dans « data science »
- Terrain, labos, entreprises,...
- Informations diverses, archives
- Satellites, réseaux, Web

ordinateurs / serveurs

Data Science

- ▶ Automatisation
- ▶ Bases de données
- ▶ Informatique
- ▶ Mathématiques
- ▶ Statistique
- ▶ Algorithmique
- ▶ Machine learning (*apprentissage*)
- ▶ Visualisation et cartographie
- ▶ Communication

Connaissances

→ **Décisions**

→ **Actions**

→ **Contrôles**

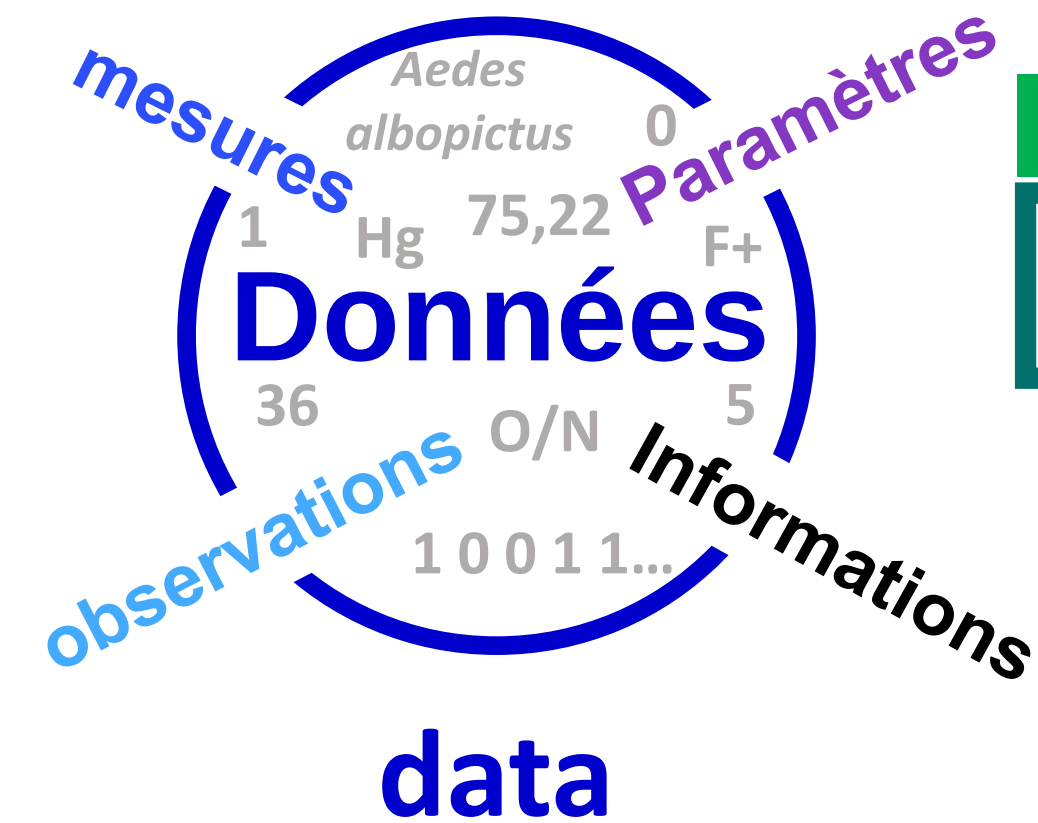
→ **Prévisions**

*Pilotage voiture autonome,
Développement médicament,
Fidélisation clientèle,...*

▷ Gros volume de données / Importante puissance de calcul / Espace de stockage conséquent

La DATA SCIENCE qu'est-ce que c'est, à quoi ça sert ?

Produire des **connaissances** en faisant parler les **données**



- Les data dans « data science »
- Terrain, labos, entreprises,...
- Informations diverses, archives
- Satellites, réseaux, Web

ordinateurs / serveurs

Data Science



Energie

+

Eau

!!!

Connaissances

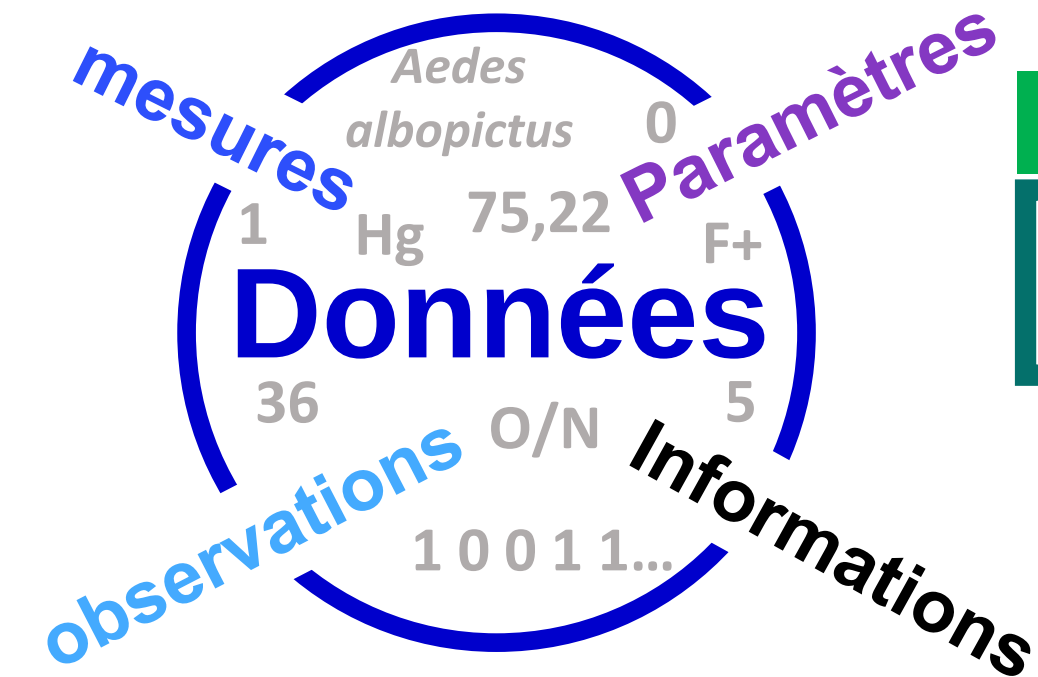
- **Décisions**
- **Actions**
- **Contrôles**
- **Prévisions**

*Pilotage voiture autonome,
Développement médicament,
Fidélisation clientèle,...*

▷ Gros volume de données / Importante puissance de calcul / Espace de stockage conséquent

La DATA SCIENCE qu'est-ce que c'est, à quoi ça sert ?

Produire des **connaissances** en faisant parler les **données**



ordinateurs / serveurs

Data Science

- ▶ Bases de données
- ▶ Informatique
- ▶ Mathématiques
- ▶ Statistique
- ▶ Algorithmique
- ▶ Machine learning (*apprentissage*)
- ▶ Visualisation et cartographie

Connaissances

→ **Décisions**

→ **Actions**

→ **Contrôles**

→ **Prévisions**

modèles ± fiables ± précis

DATA = OR BLANC !

- Les data dans « data science »
- Terrain, labos, entreprises,...
- Informations diverses, archives
- Satellites, réseaux, Web

précision et fiabilité

Energie

+

Eau

!!!

▷ Gros volume de données / Importante puissance de calcul / Espace de stockage conséquent

Définition

Discipline (en fait plutôt une démarche, une philosophie) visant à **l'extraction de connaissances** à partir d'ensembles de données et s'appuyant sur des outils **statistiques**, mathématiques et **informatiques** (notamment de **visualisation**).

Peut être considérée comme une **science des données numériques**.

Permet faire parler et de valoriser les données

Métiers (au moins pour l'instant), il existe une myriade de métiers associés à l'étiquette « *data scientist* » : data engineer, data analyst)

« Le Geek, l'Analyste et le Communicateur ». Une fable ?

Valorisation des données

L'élément central (et pas qu'en termes de dollars)

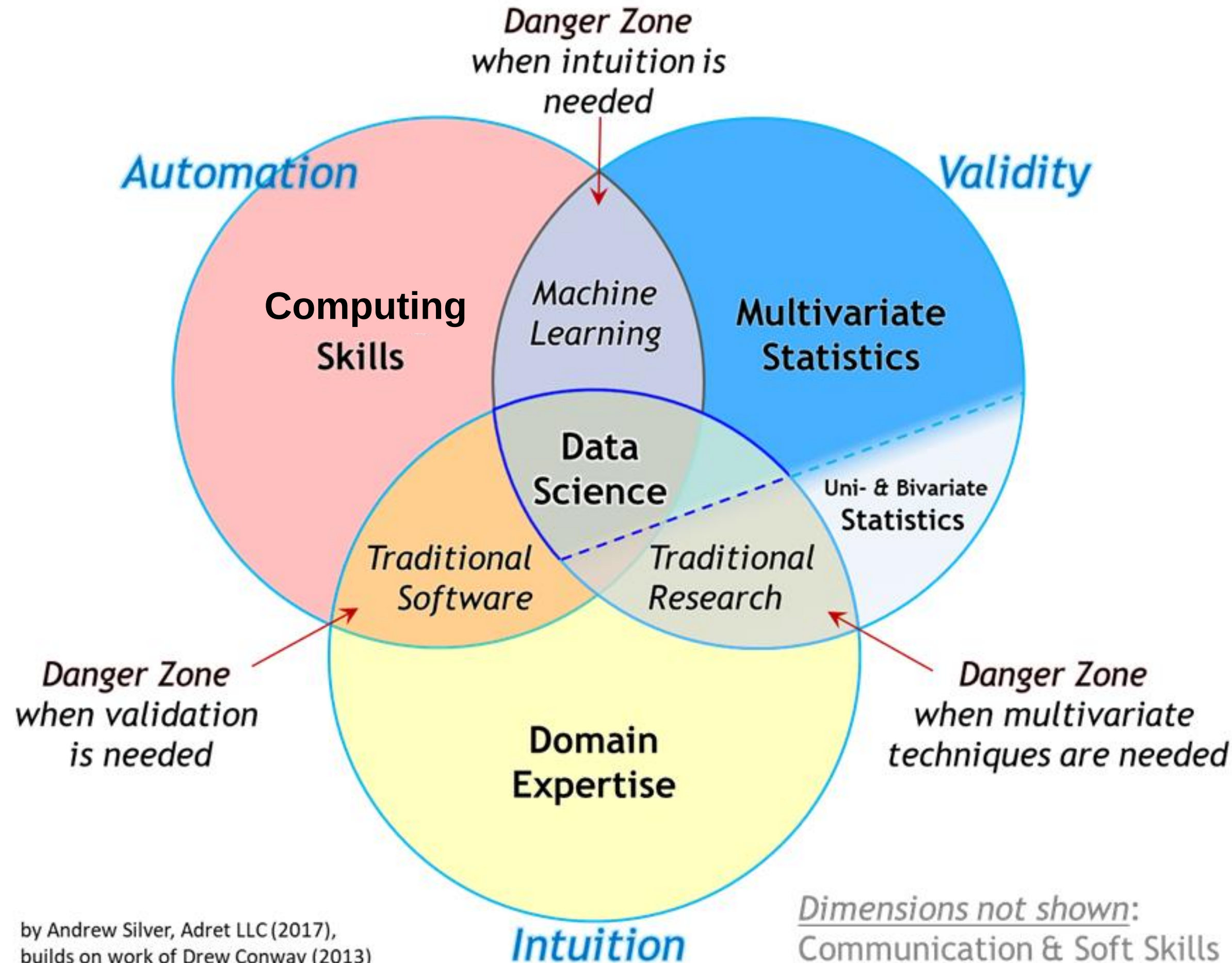
Il va donc falloir apprendre à faire parler les données

(préparation, gestions des valeurs manquantes, statistique descriptive, modélisation, statistique inférentielle, apprentissage...)

Les données : de l'or blanc !

A vous de jouer! Saurez-vous relever ce défi?

Data Science – Les zones dangereuses (diagramme de Venn)



by Andrew Silver, Adret LLC (2017),
builds on work of Drew Conway (2013)

Le cheminement des données

Acquisition

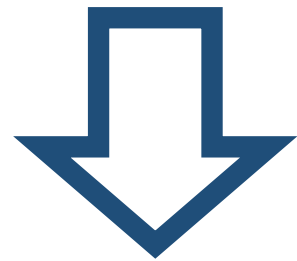
(acquérir les données de façon intelligente, plan d'expérience, campagne de collecte, fiabilité des data)



Filtrer, organiser, importer

Stockage

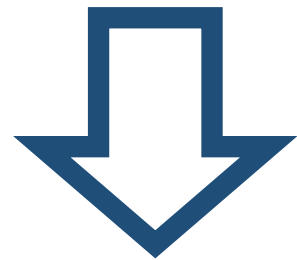
(organisation, formatage, labels, rôle central des **bases de données**)



Nettoyer, décrire, regrouper, interroger, extraire

Traitement - Analyse

(codage, recodage, scoring, **analyse statistique**, modélisation, prédictions, **machine learning**, ...)



Faire parler les données selon un objectif, un but précis

Communication

(conclusions, nouvelles connaissances, décisions, ordres à machines, prévention,)

Objectif

L'objectif de l'enseignement est de former les étudiants à **la manipulation, au stockage, au traitement et à l'analyse des données biologiques** acquises au laboratoire, sur le terrain ou en entreprise et à leur **représentation graphique**.

Il s'agira d'apprendre à exploiter au mieux les données biologiques dans le **contexte du *Big Data***,

Constituant une véritable initiation aux **data sciences en biologie**, il sera axé sur la réalisation de

bases de données, le **traitement** et **l'analyse des données** et le **machine learning**)

(+ communication)

Un plus pour votre CV

Acquisition d'une compétence très demandée (recherche / ingénierie / entreprise)



Compétences acquises

- À l'issue de cette formation, vous serez **capables de prendre en charge un travail de data sciences en biologie**
- l'expertise acquise ira de la création à l'extraction d'informations d'une base de données puis à l'exploitation de ces informations à l'aide d'un traitement statistique complet associé aux techniques d'apprentissage (machine learning) et à la communication des résultats sous forme de présentation synthétique (comprenant la rédaction d'un rapport écrit de qualité professionnelle)
- vous serez ainsi **armés pour répondre aux nouveaux challenges de l'analyse des données en biologie** et on pourra vous confier **la responsabilité d'un projet de data sciences dans le cadre de votre métier**, dans le secteur privé comme dans le secteur publique (notamment Recherche Scientifique).

Le deuxième effet KiSSCool (*second effet bénéfique*)

- **Redonner le sens pratique** à des étudiants gavés de théorie
(« je connais mais ne sais pas l'appliquer »); c'est pourquoi cette UE est très axée sur la pratique.
- Savoir **se débrouiller seul** (repérer la situation, passer à la pratique, travailler vite et bien).
Etre capable d'analyser ses propres données expérimentales, de production ou de marketing
(*contexte stage, labo, bureau d'étude, entreprise, ...*)
- Parler un **langage de base en statistique, machine learning, base de données, etc..**
(mais éviter les excès : « *l'inférence de l'inféron* ») permettant de faire appel aux spécialistes lorsque cela est nécessaire
- Ne pas confondre data sciences, apprentissage supervisé, bioinformatique, biostatistique et autre spécialisations
(attention aux escrocs!)
- Essayer de **mieux décrypter le monde dans lequel nous vivons**
- Prendre des **responsabilités** (et des risques mesurés)
- Apprendre à **se faire confiance** (confiance en soi tout d'abord et confiance en son équipe)
- En faire profiter le CV

- ⇒ Cours/TD/**Ateliers Pratiques/Projets** [40 h]
- ⇒ Développement du Projet choisi à chaque séance (*sauf première séance*)
- ⇒ Salles informatiques du Bât 625 et 336
- ⇒ **L'accent** est mis sur **la pratique**
- ⇒ Un **site web** où vous retrouverez cours, défis, jeux (équipes), corrigés, auto-évaluations, **avancement des projets**, forum, liaison avec l'enseignant
- ⇒ (vidéos des cours et de présentation de l'enseignement)

Responsable : Pascal RIGOLET (pascal.rigolet@universite-paris-saclay.fr)

Plans des séances, photos de tableaux, diaporamas et vidéos de l'enseignement

Séance 1 - Introduction aux bases de données

- Support de la séance (fichier pdf)
- Séance 2A - Les informations dans les bases de données sont destinées à être exploitées (Séance distancielle remplaçant en partie le cours annulé du mercredi 16 septembre)
- Vidéo à visionner >>> Les data sciences - Qu'est-ce que c'est ? A quoi ça sert? (vidéo mp4)
- Vidéo à visionner >>> Types des bases de données (vidéo mp4)
- Fiche résumé sur les types de variables (fichier pdf)
- Le rôle central des bases de données dans le traitement et l'analyse des données (fichier pdf)

Accueil | Objectifs | Contact | Liens utiles | Offres de stage

Menu

- Objectifs
- > Espace L2S4
- Les défis du L2
- Je vous recommande
- Supports de Cours
- Projets
- Contrôle Connaissances
- Examen à distance
- Data Sciences
- Bases de Données
- Une base sur le web
- Web Design
- Machine Learning
- Sondages
- Forum
- Video Teaching
- Exercices
- Tables
- Pour approfondir
- Divers
- Autoformation Stat
- Fichiers de Données
- Autoformation-Databases

Data Sciences

Séances et documents - Année 2020

Une étroite complicité entre l'enseignant et les étudiants est la clé du succès de cette formation. Je vous souhaite beaucoup de plaisir dans cette UE.

Espace Numérique de Travail (ENT) - Enseignement Numérique (EN)

'a rien sans rien!). Cela vous initie à l'enseignement numérique qui devient incontournable à notre époque e

[dernière mise à jour le 07/10/2020]

PARTIE DU SITE EN COURS DE DEVELOPPEMENT - PROCHAINEMENT OPERATIONNEL

Les mots de la semaine :

data sciences ; outils de visualisation ; arbres de décision ; apprentissage

Présentation de l'option TAB

- Présentation de l'enseignement (fichier pdf)
- Présentation vidéo de l'UE (fichier mp4) [durée : 5 min. 15]

Plans des séances, photos de tableaux et diaporamas projetés en cours

> Chaque séance est couplée à une ou plusieurs formations numériques à distance de quelques minutes

Séance 1 - Introduction : Qu'est-ce que les data sciences?

Vous avez entendu parler de Data Sciences mais vous vous demandez encore ce que cela peut bien être et ce qui se cache derrière ce concept. C'est vrai qu'avec la foison d'informations qui sont diffusées par les médias ou disponibles sur internet on a un peu de mal à faire la part des choses concernant cette approche multidisciplinaire émergente. On trouve un peu tout et n'importe quoi sur ce domaine. Cette section devrait vous aider à y voir plus clair et tenter de vous expliquer tout cela de la façon la plus concise et la plus simple qui soit. Ce sera l'occasion de revisiter les types de variables utilisées dans le traitement des données, notamment en statistique puis de découvrir comment sont codées et structurées les informations dans le chemin qui part des data et conduit jusqu'au Big Data.

- Les data sciences - Qu'est-ce que c'est ? A quoi ça sert? (vidéo mp4 de 4 minutes)
- Types des Variables : tout ce qu'il faut savoir (vidéo mp4 de 4 minutes)
- Fiche résumé sur les types de variables (fichier pdf)
- Des data au Big Data - Codage et structuration d'un tableau (fichier pdf)
- Le diaporame sur l'analyse des données et les données (fichier pdf)

Séance 2 - Le/la data scientist, un métier

Plutôt que d'aller chercher la perle rare qui maîtrise les data sciences regroupent un ensemble de compétences (techniques, analytiques, de communication, etc.). Nous allons voir comment se présente un data scientist (ou data analyst) et comment il travaille.

- Comment devenir Data Scientist? (vidéo mp4 de 4 minutes)

- Travail sur des **jeux de données réels**
- Thématiques : Environnement, Epidémiologie, Biomédical, Pollution, Agronomie, Aide au diagnostic...
- Utilisation de logiciels dédiés
- **Apprendre à raisonner** avant d'appliquer le modèle adapté à la problématique.
- Travail organisé en **groupes de 3 à 4 en salle**
- **Préparation et avancement des projets en classe**
- **Jeux** individuels et par équipe
- Apprentissage d'une démarche
- Vous servira dans vos **stages** (*ainsi que dans la **suite de votre cursus***)

Contrôle des connaissances

- **Réalisation d'un projet de data sciences sur une thématique choisie par le groupe d'étudiants**

[fichier de données + résultats de l'analyse numérique + rapport écrit (12 à 15 pages environ) + présentation orale]

Le contrôle des connaissances portera essentiellement sur un **projet** qui mettra l'accent non pas sur des calculs mais sur des stratégies de traitement des données et sur l'analyse de résultats présentés sous forme de graphiques et de tableaux.

Ce projet sera démarré en classe (ateliers).

- **Contrôle continu** : défis et jeux savants

- **Première session** : $NF = 0,70 \cdot PROJET + 0,30 \cdot CC$

NF : Note Finale CC : Contrôle Continu

- **Deuxième session** : $NF = 0,70 \cdot PROJET + 0,30 \cdot CC[REP]$

Enseignant intervenant

Pascal RIGOLET (responsable du module) : pascal.rigolet@universite-paris-saclay.fr



Un des slogans de Science Po :

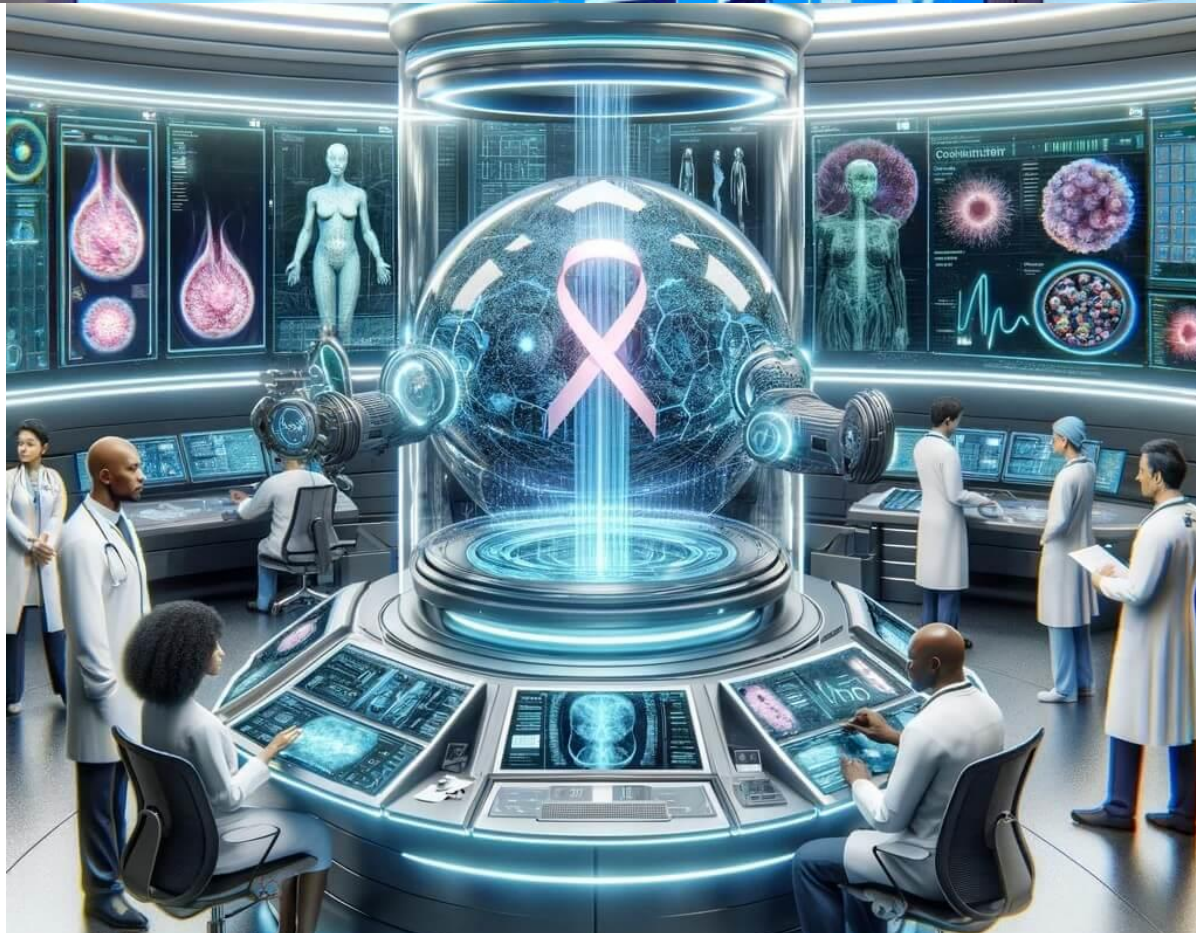
« Sciences Po Grenoble propose, en cinq ans, un parcours intellectuel ambitieux destiné à **former les décideurs de demain** »

Cherchez l'erreur!

**Ceux qui maîtriseront les data,
contrôleront beaucoup de choses, ils décideront pour les autres**

**Je me bats pour que vous ayez un pouvoir de décision
en tant que biologistes (santé, écologie et environnement)**

Le traitement et le stockage de l'information ne datent pas d'hier !

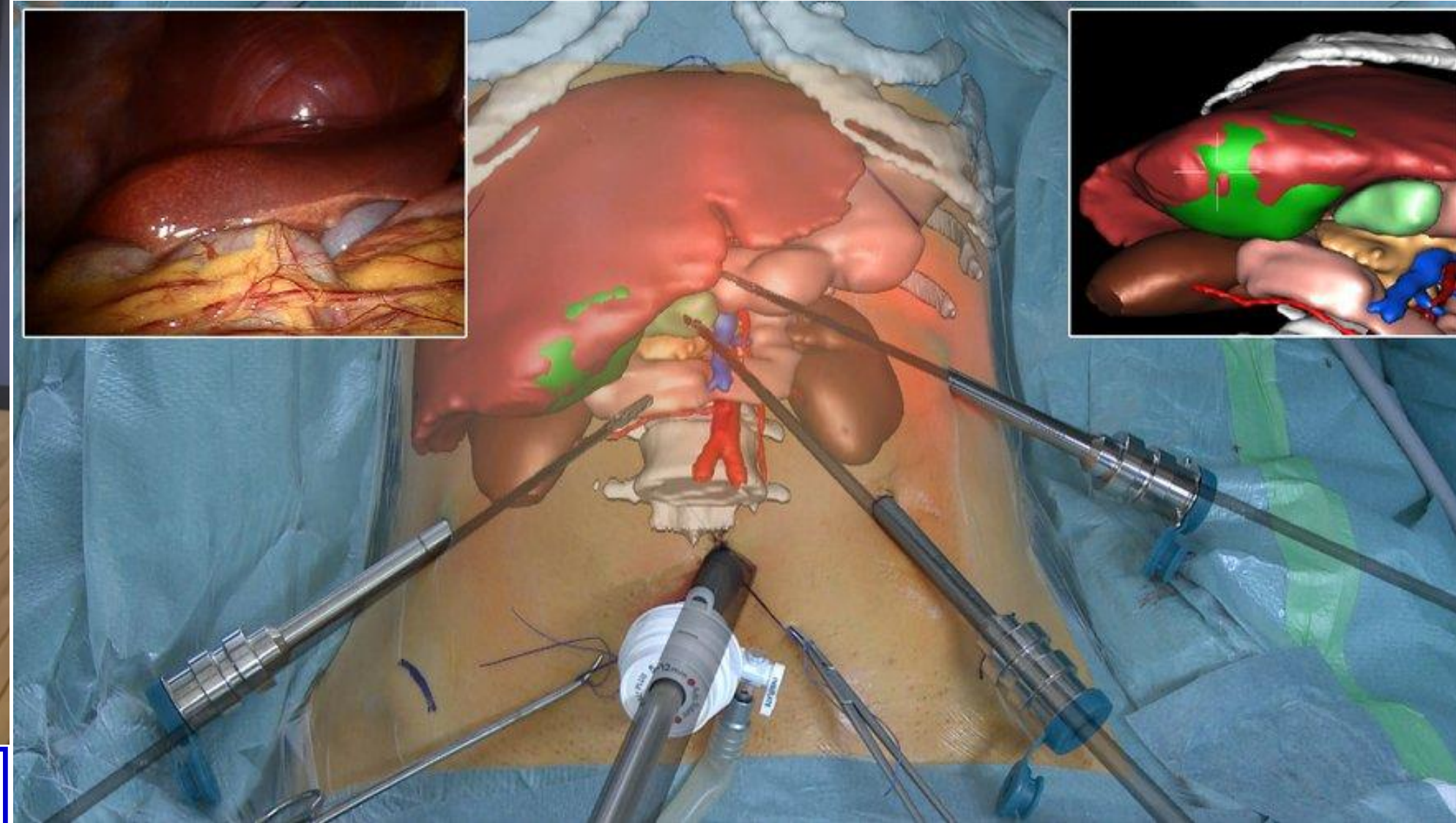


Le métavers, une révolution à laquelle il faut se préparer en biologie



Également gestion de l'environnement et biodiversité, santé (étude en temps réel du corps humain en complément des autres diagnostics, scanner, radio, etc...) et à une échelle très précise (mm ou μm), assistance à opération en chirurgie humaine (c'est déjà le cas; projection de la partie à enlever sur l'organe ou la région à traiter pour enlever ce qu'il faut avec une très grande précision), etc...

Un métavers est un monde virtuel fictif.



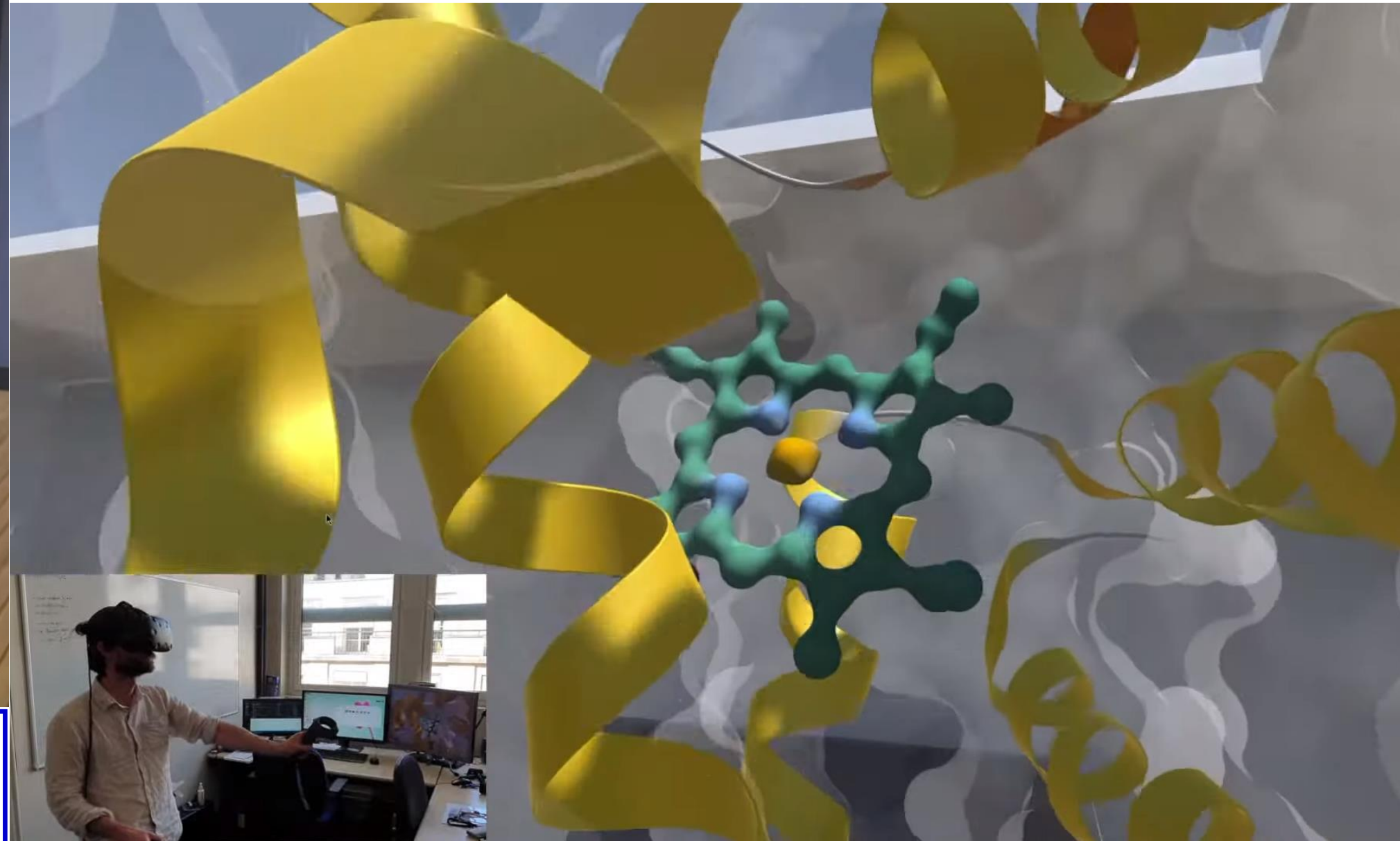
Un métavers est un monde virtuel fictif. Le terme est régulièrement utilisé pour décrire une future version d'Internet où des espaces virtuels, persistants et partagés sont accessibles via interaction 3D. Le métavers déjà en route en biologie (structurale). C'est peut-être vous qui en construirez les applications à venir...

Le métavers, une révolution à laquelle il faut se préparer en biologie



Également gestion de l'environnement et biodiversité, santé (étude en temps réel du corps humain en complément des autres diagnostics, scanner, radio, etc...) et à une échelle très précise (mm ou μm), assistance à opération en chirurgie humaine (c'est déjà le cas; projection de la partie à enlever sur l'organe ou la région à traiter pour enlever ce qu'il faut avec une très grande précision), etc...

Un métavers est un monde virtuel fictif.



Un métavers est un monde virtuel fictif. Le terme est régulièrement utilisé pour décrire une future version d'Internet où des espaces virtuels, persistants et partagés sont accessibles via interaction 3D. Le métavers déjà en route en biologie (structurale). C'est peut-être vous qui en construirez les applications à venir...

A close-up photograph of two large red ants on a forest floor. The ant in the foreground is facing left, with its head and antennae clearly visible. The ant in the background is facing right. The ground is covered with dry leaves and twigs.

<https://lemag.ird.fr>

La chasse aux virus dans les forêts tropicales, un travail de fourmis...

24.03.2023

SANTÉ

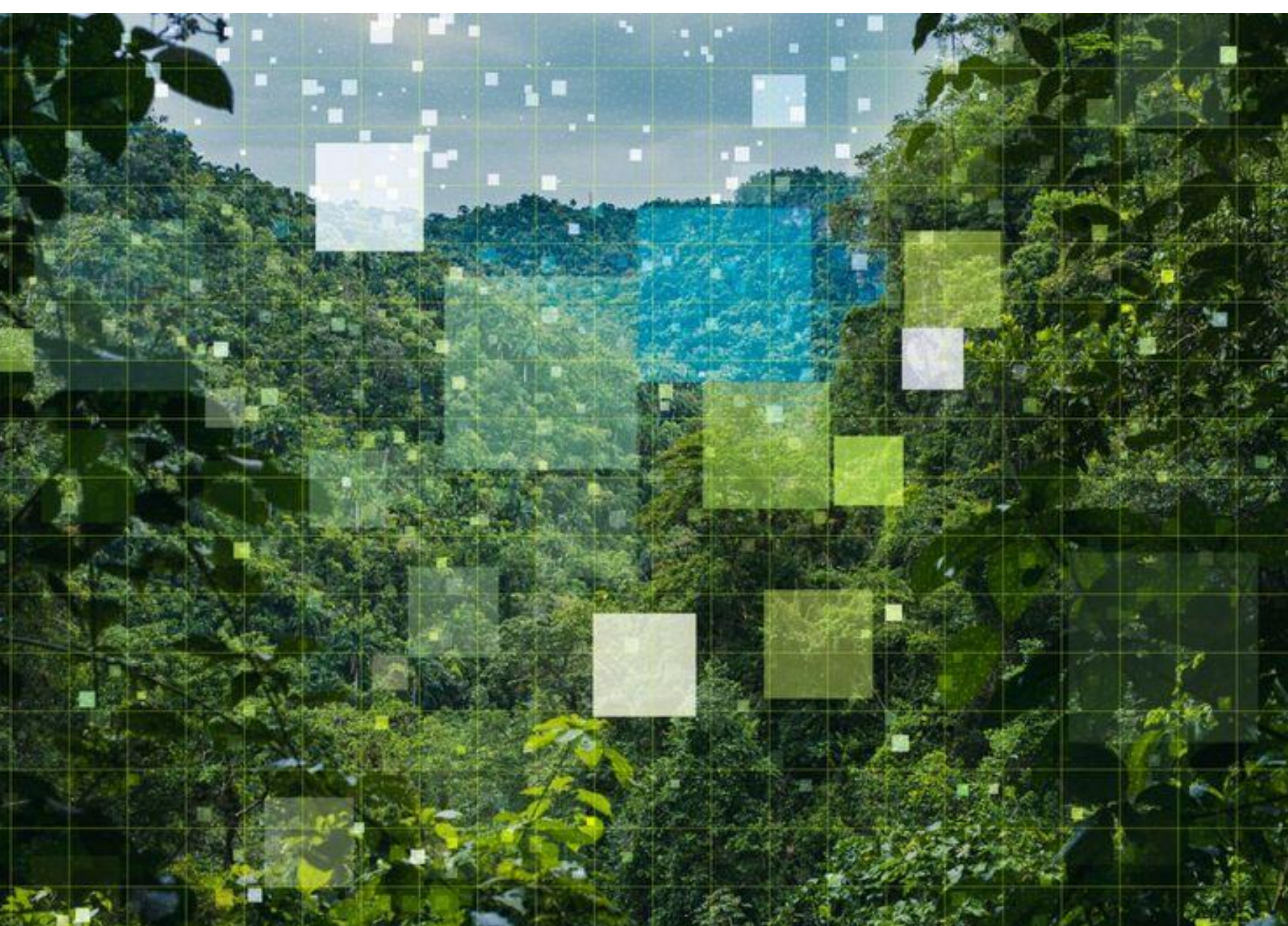
VIVANT

Objectif : identifier les virus avec lesquels les fourmis ont été en contact.

Récupérer l'ensemble du matériel génétique (ADN et ARN) au sein d'un échantillon puis analyser la portion correspondant au génome viral total (*l'échantillon contient également du matériel génétique provenant des animaux ou plantes que la fourmi a consommés*).

L'IA au service des agriculteurs



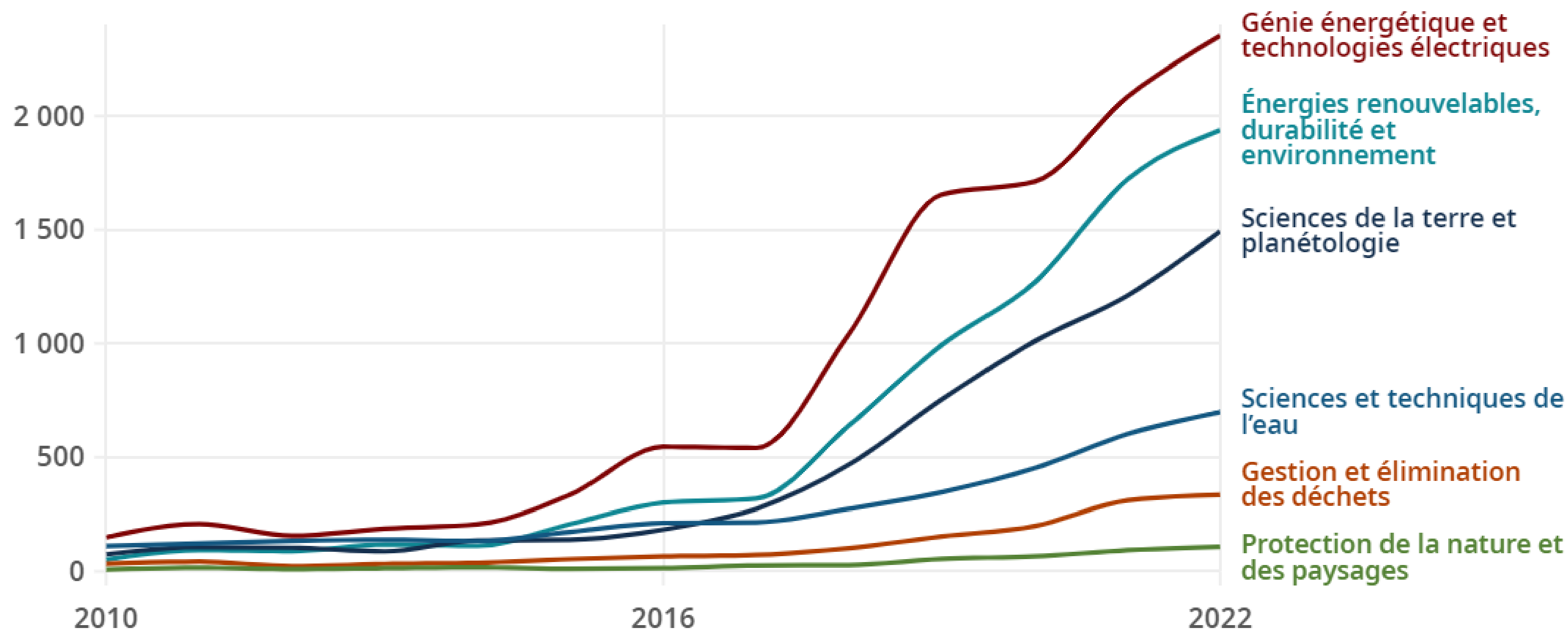


De véritables Salles de Contrôle



Publications scientifiques sur l'IA et l'environnement par grand thème

Nombre de publications, pays de l'OCDE



Le stockage et le traitement de l'information ne datent pas d'hier !

Le stockage et le traitement de l'information ne datent pas d'hier !

Décomptes et registres (Babylone, Haute Egypte, antiquité, moyen âge). De tous temps les humains ont eu besoin de compter, dénombrer, prédire, échanger, rapporter, archiver



1	𐎶	11	𐎶𐎵	21	𐎶𐎵𐎶	31	𐎶𐎵𐎶𐎵	41	𐎶𐎵𐎶𐎵𐎶	51	𐎶𐎵𐎶𐎵𐎶𐎵
2	𐎶𐎵	12	𐎶𐎵𐎶𐎵	22	𐎶𐎵𐎶𐎵𐎶	32	𐎶𐎵𐎶𐎵𐎶𐎵	42	𐎶𐎵𐎶𐎵𐎶𐎵𐎶	52	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵
3	𐎶𐎵𐎶	13	𐎶𐎵𐎶𐎵𐎶	23	𐎶𐎵𐎶𐎵𐎶𐎵	33	𐎶𐎵𐎶𐎵𐎶𐎵𐎶	43	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵	53	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵𐎵
4	𐎶𐎵𐎶𐎵	14	𐎶𐎵𐎶𐎵𐎶𐎵	24	𐎶𐎵𐎶𐎵𐎶𐎵𐎶	34	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵	44	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵𐎵	54	𐎶𐎵𐎶𐎵𐎶𐎵𐎶𐎵𐎵𐎵
5	𐎶𐎵𐎶𐎵𐎵	15	𐎶𐎵𐎶𐎵𐎵𐎵	25	𐎶𐎵𐎶𐎵𐎵𐎵𐎵	35	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵	45	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵	55	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵
6	𐎶𐎵𐎶𐎵𐎵𐎵	16	𐎶𐎵𐎶𐎵𐎵𐎵𐎵	26	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵	36	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵	46	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵	56	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵
7	𐎶𐎵𐎶𐎵𐎵𐎵𐎵	17	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵	27	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵	37	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵	47	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	57	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵
8	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵	18	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵	28	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵	38	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	48	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	58	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵
9	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵	19	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵	29	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	39	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	49	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵	59	𐎶𐎵𐎶𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵𐎵
10	𐎶	20	𐎶𐎵	30	𐎶𐎵𐎵	40	𐎶𐎵𐎵𐎵	50	𐎶𐎵𐎵𐎵𐎵		

Babylone, écriture cunéiforme

Système d'écriture mis au point en Basse Mésopotamie vers 3400 avant notre ère et qui s'est répandu ensuite dans tout le Proche-Orient ancien

Dates approximatives :

	— 3300	— 2800	— 2400	— 1800	— 700
L'étoile (signe du ciel et du dieu)					
La parcelle de terre (signe de la terre)					
La silhouette humaine (signe de l'homme)					
Le triangle pubien (signe de la femme)					
La femme + le schéma du massif montagneux (la femme étrangère → l'esclave)					
L'oiseau					
Le poisson					
La tête de vache (signe de la vache)					
L'épi d'orge (signe de l'orge, le grain)					



A peu près au même moment (vers 3200 ans avant notre ère) apparaissent les Hiéroglyphes en Egypte ancienne

1 000 000	100 000	10 000	1 000	100	10	1

Quelques chiffres et nombres, haute Egypte

Les Pionniers

Le premier programme informatique pour une machine mécanique



Ada Lovelace (1815-1852)

Pionnière de l'informatique, première programmeuse.

En 1842 elle a réalisé le **premier programme** informatique (théorique) pour l'ancêtre de l'ordinateur : la machine analytique de Charles Babbage (inexistante).

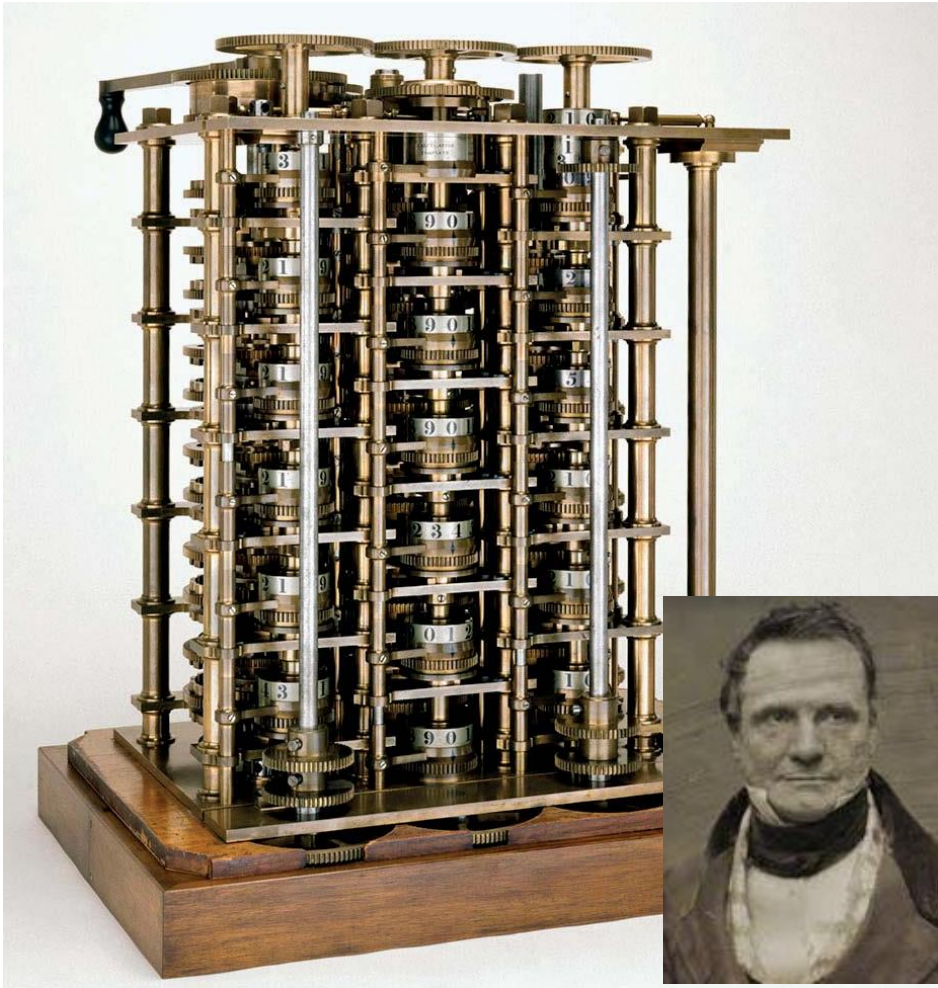


Diagram for the computation by the Engine of the Numbers of Bernoulli. See Note G. (page 722 et seq.)

Number of Operation.	Nature of Operation.	Variables acted upon.	Variables receiving results.	Indication of change in the value on any Variable.	Statement of Results.	Data.										Working Variables.										Result Variables.				
						V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	V_{16}	V_{17}	V_{18}	V_{19}	V_{20}	V_{21}	V_{22}	V_{23}	V_{24}	
						0 0 0 1	0 0 1 2	0 0 4 0	0 0 4 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0		
						1	2	n																						
1	$\times V_2 \times V_3$	V_2, V_3	V_4	$V_4 = V_2 \times V_3$	$= 2n$	...	2	n	2n	2n	2n	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	$- V_4 - V_5$	V_4, V_5	V_6	$V_6 = V_4 - V_5$	$= 2n - 1$	...	1	...	2n - 1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	$+ V_6 + V_7$	V_6, V_7	V_8	$V_8 = V_6 + V_7$	$= 2n + 1$	...	1	...	...	...	2n + 1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	$- V_8 - V_9$	V_8, V_9	V_{10}	$V_{10} = V_8 - V_9$	$= 2n - 1$	...	...	...	0	0	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	$+ V_{10} + V_{11}$	V_{10}, V_{11}	V_{12}	$V_{12} = V_{10} + V_{11}$	$= \frac{1}{2} \cdot \frac{2n-1}{2n+1}$	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6	$- V_{12} - V_{13}$	V_{12}, V_{13}	V_{14}	$V_{14} = V_{12} - V_{13}$	$= \frac{1}{2} \cdot \frac{2n-1}{2n+1} = A_0$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7	$- V_{14} - V_{15}$	V_{14}, V_{15}	V_{16}	$V_{16} = V_{14} - V_{15}$	$= n - 1 (= 2)$	...	1	...	n	...	...	...	...	...	...	n - 1	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8	$+ V_{16} + V_{17}$	V_{16}, V_{17}	V_{18}	$V_{18} = V_{16} + V_{17}$	$= 2 + 0 = 2$	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9	$- V_{18} - V_{19}$	V_{18}, V_{19}	V_{20}	$V_{20} = V_{18} - V_{19}$	$= \frac{2n}{2} = A_1$	...	...	...	...	...	2n	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10	$\times V_{20} \times V_{21}$	V_{20}, V_{21}	V_{22}	$V_{22} = V_{20} \times V_{21}$	$= B_1 \cdot \frac{2n}{2} = B_1 A_1$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11	$+ V_{22} + V_{23}$	V_{22}, V_{23}	V_{24}	$V_{24} = V_{22} + V_{23}$	$= \frac{1}{2} \cdot \frac{2n-1}{2n+1} + B_1 \cdot \frac{2n}{2}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12	$- V_{24} - V_{25}$	V_{24}, V_{25}	V_{26}	$V_{26} = V_{24} - V_{25}$	$= n - 2 (= 2)$	...	1	...	...	...	...	...	...	...	...	n - 2	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13	$- V_{26} - V_{27}$	V_{26}, V_{27}	V_{28}	$V_{28} = V_{26} - V_{27}$	$= 2n - 1$	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14	$+ V_{28} + V_{29}$	V_{28}, V_{29}	V_{30}	$V_{30} = V_{28} + V_{29}$	$= 2 + 1 = 3$	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
15	$- V_{30} - V_{31}$	V_{30}, V_{31}	V_{32}	$V_{32} = V_{30} - V_{31}$	$= \frac{2n-1}{3}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16	$\times V_{32} \times V_{33}$	V_{32}, V_{33}	V_{34}	$V_{34} = V_{32} \times V_{33}$	$= \frac{2n-1}{2} \cdot \frac{2n-1}{3}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17	$- V_{34} - V_{35}$	V_{34}, V_{35}	V_{36}	$V_{36} = V_{34} - V_{35}$	$= 2n - 2$	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18	$+ V_{36} + V_{37}$	V_{36}, V_{37}	V_{38}	$V_{38} = V_{36} + V_{37}$	$= 3 + 1 = 4$	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		

- Entrée/Sorties
- Instructions
- Fichiers
- Format
- Algorithme
- Programme
- Données
- Variables

Entrées



Sorties

De nombreuses autres femmes ont par la suite contribué au développement de l'informatique

Le tout premier ordinateur, le Zuse Z3 (Berlin, 1941)



Le Zuse Z3 était un calculateur à relais électromécaniques conçu par l'ingénieur allemand Konrad Zuse. On le programmait au moyen de rubans perforés et les nombres entrés en base décimale étaient convertis en binaire à virgule flottante. Zuse fut le premier programmeur d'une machine réelle. L'institut de recherche aéronautique allemand l'utilisait pour réaliser des analyses statistiques sur les vibrations des ailes (bombes volantes). Le Z3 était complet au sens de Turing.

Le mot « ordinateur » été créé par IBM en 1953 (machines automatiques répondant à la fameuse architecture de Von Neumann).
L'informatique, désignant le traitement automatique de l'information, combine les mots Information et automatique.



Konrad Zuse

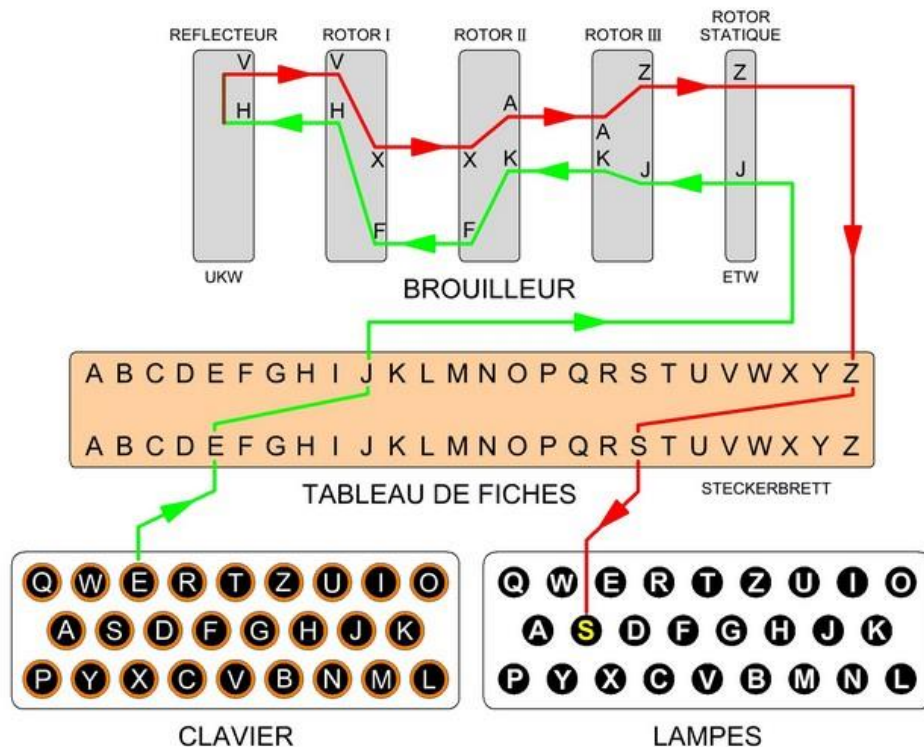
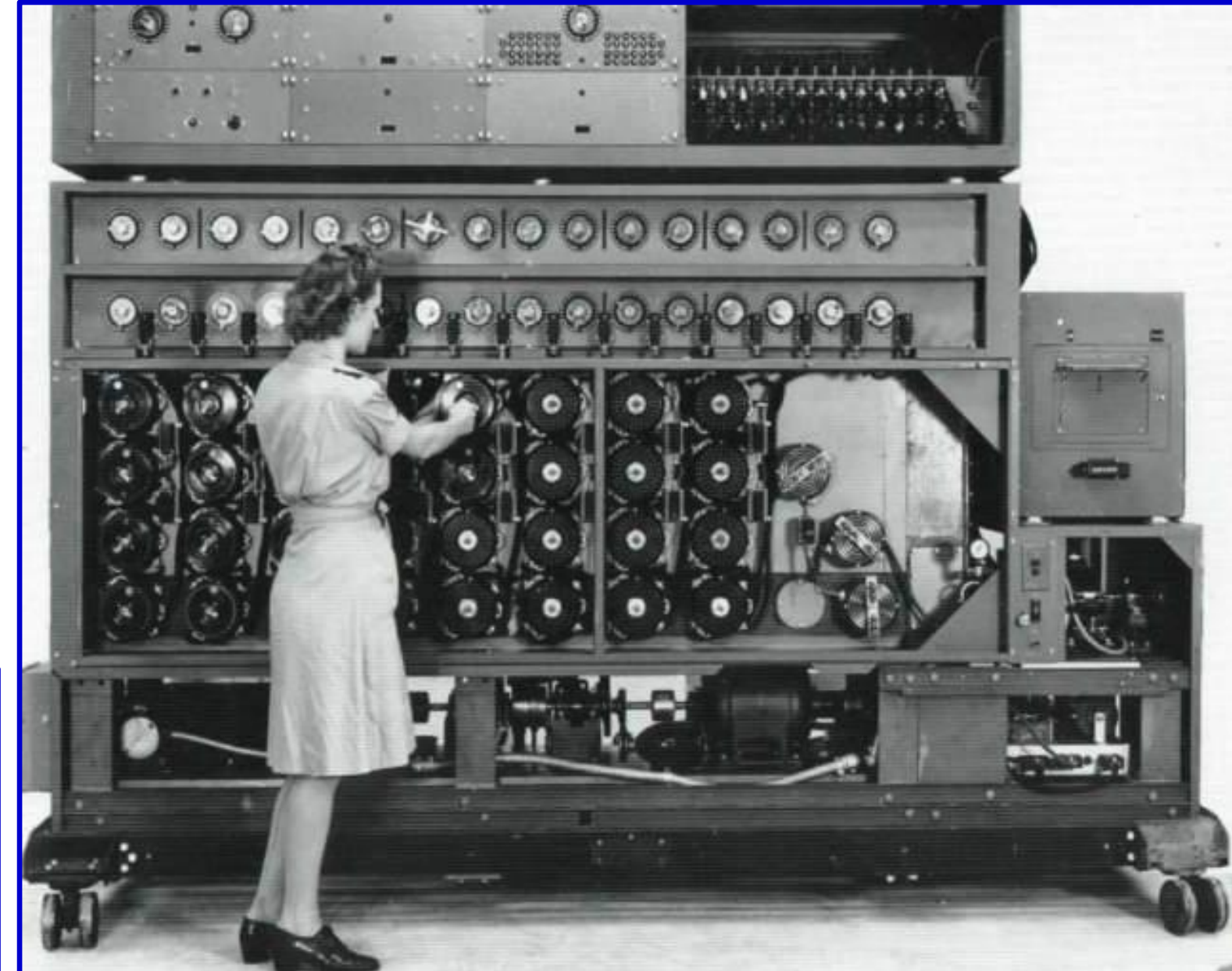
Le premier combat de titans informatique - Machines électromécaniques



Enigma, machine de cryptage simple mais aux nombre de combinaisons de chiffrage vertigineux, fut utilisée par l'Allemagne nazie et notamment ses armées pour coder ses messages. Pendant la 2^{ème} guerre mondiale, les cryptanalystes britanniques, dont **Alan Turing**, continuèrent les travaux du mathématicien polonais **Marian Rejewski** et furent capables de déchiffrer les messages Enigma en perfectionnant les « **bombes électromécaniques** » inventées par Rejewski.



Alan Turing
père du 1^{er} ordinateur
digne de ce nom :
la Bombe (1940-45)



Le premier ordinateur binaire (1943)

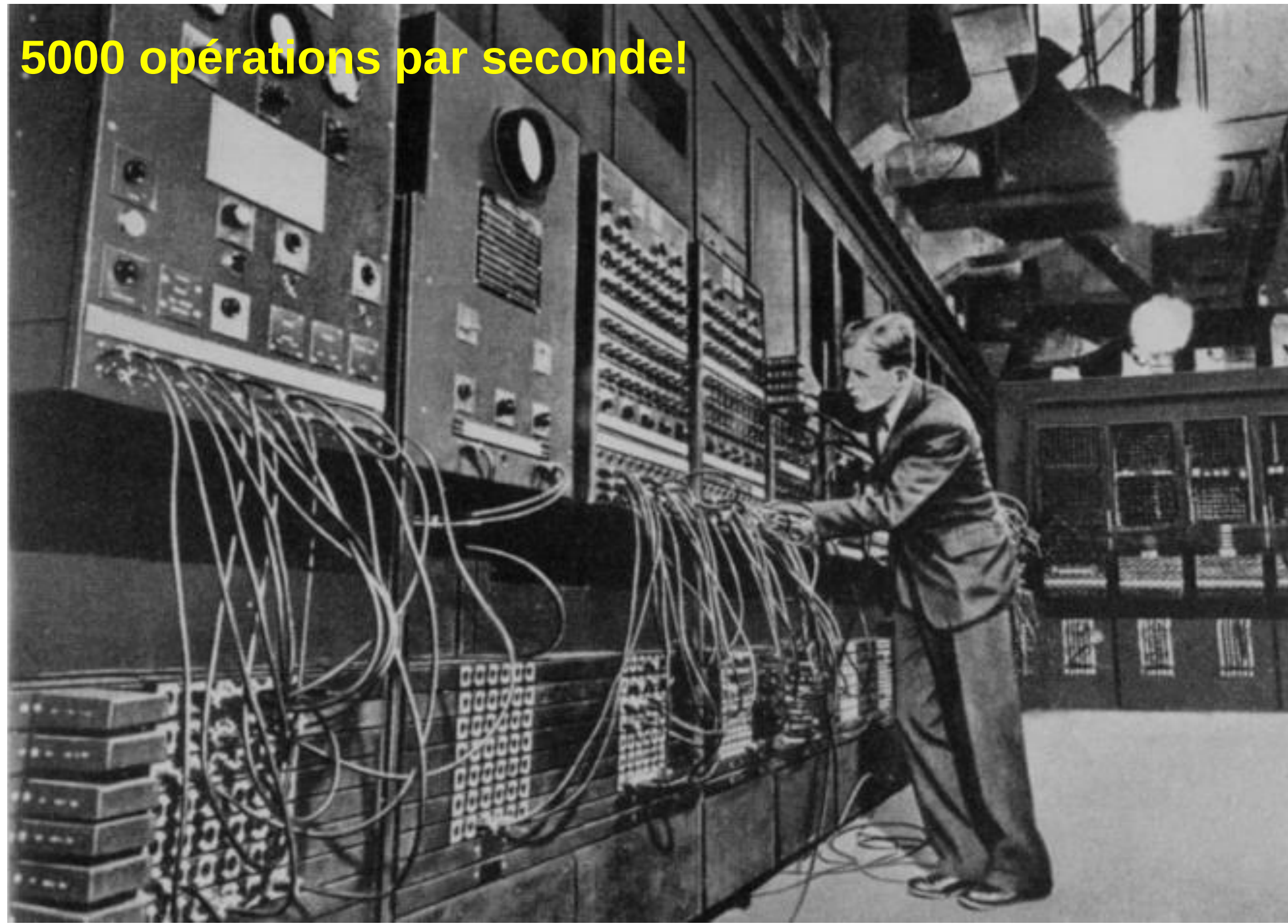
Les COLOSSUS MARK I (suivit du MARK II) ont contribué à casser le code Lorenz (binaire) et au lancement surprise du débarquement



Le premier ordinateur numérique électronique programmable (1946)

L'ENIAC acronyme de *Electronic Numerical Integrator And Computer*

« Le Giganteste Cerveau » pesait 30 tonnes et était destiné à l'armée américaine (balistique



Naissance de l'Intelligence Artificielle

VOL. LIX. No. 236.]

[October, 1950]

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

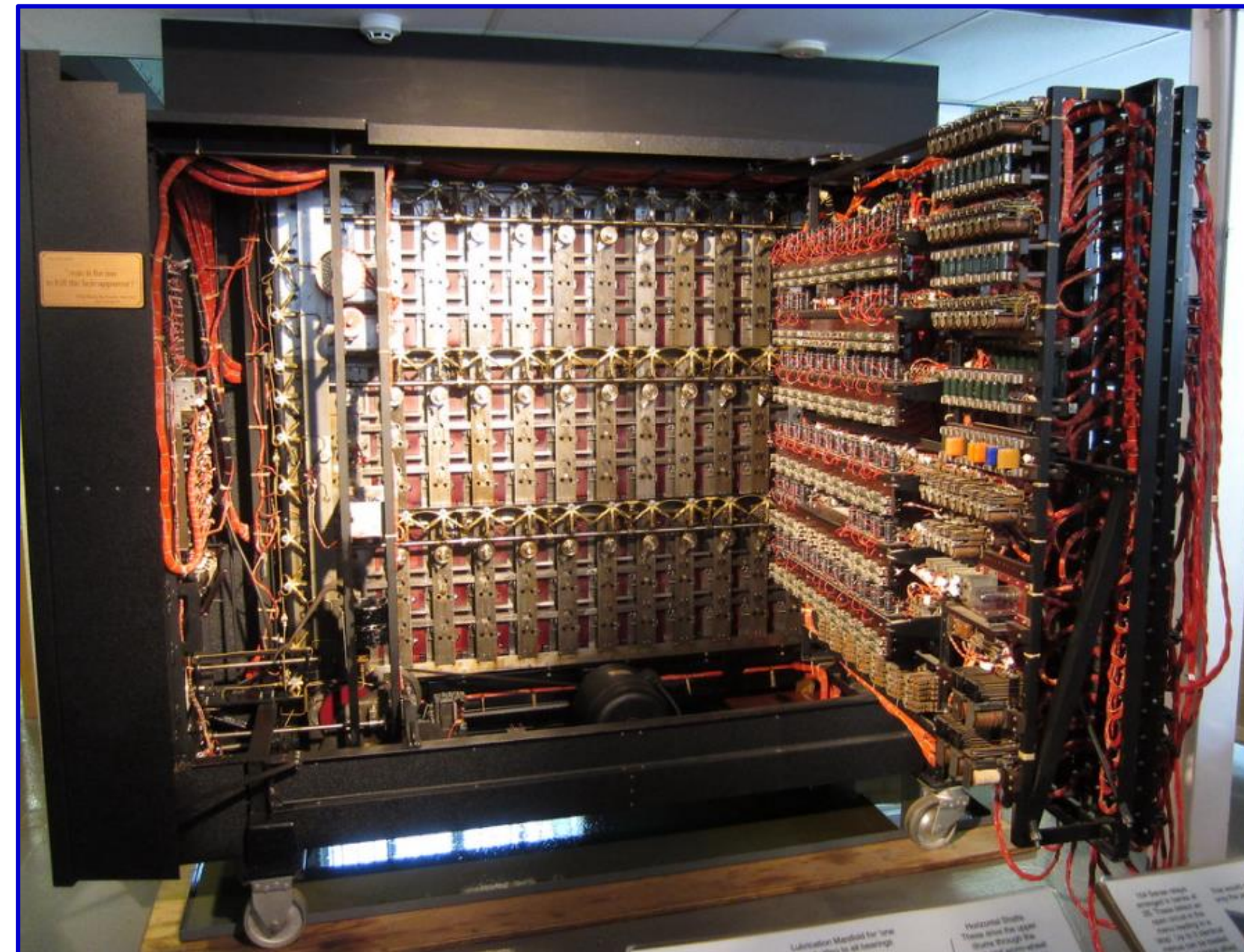
The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is

Le test de Turing

En octobre 1950, le mathématicien britannique Alan Turing signe l'un de ses articles les plus célèbres, sous le titre "Machines de calcul et intelligence" débutant par : « **Je propose de réfléchir à la question : les machines peuvent-elles penser ?** »



Alan Turing
père du 1^{er} ordinateur
digne de ce nom :
la Bombe (1940-45)



Naissance de l'Intelligence Artificielle

VOL. LIX. No. 236.]

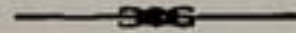
[October, 1950]

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY



I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is

Le test de Turing

En octobre 1950, le mathématicien britannique Alan Turing signe l'un de ses articles les plus célèbres, sous le titre "Machines de calcul et intelligence" débutant par : « **Je propose de réfléchir à la question : les machines peuvent-elles penser ?** »



Alan Turing

père du 1^{er} ordinateur
digne de ce nom :

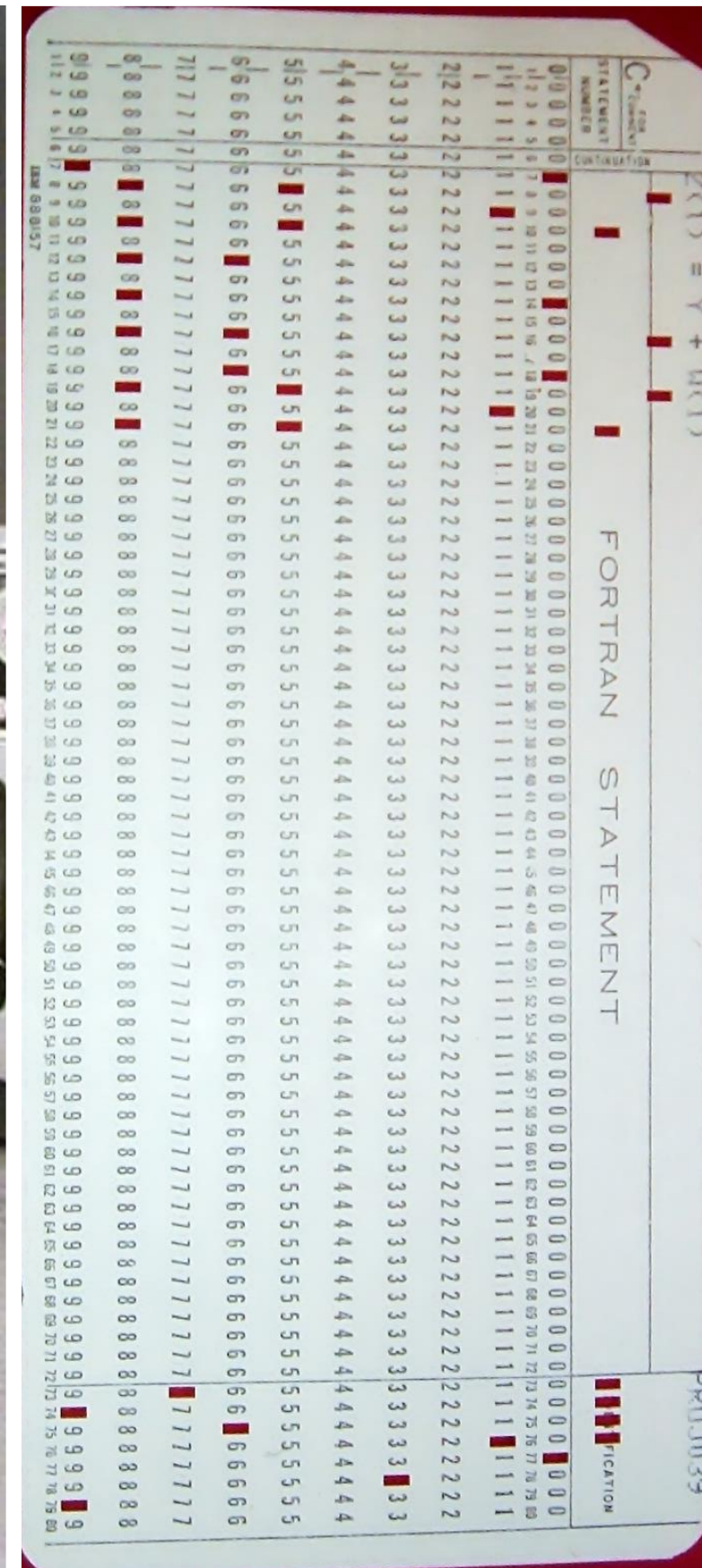
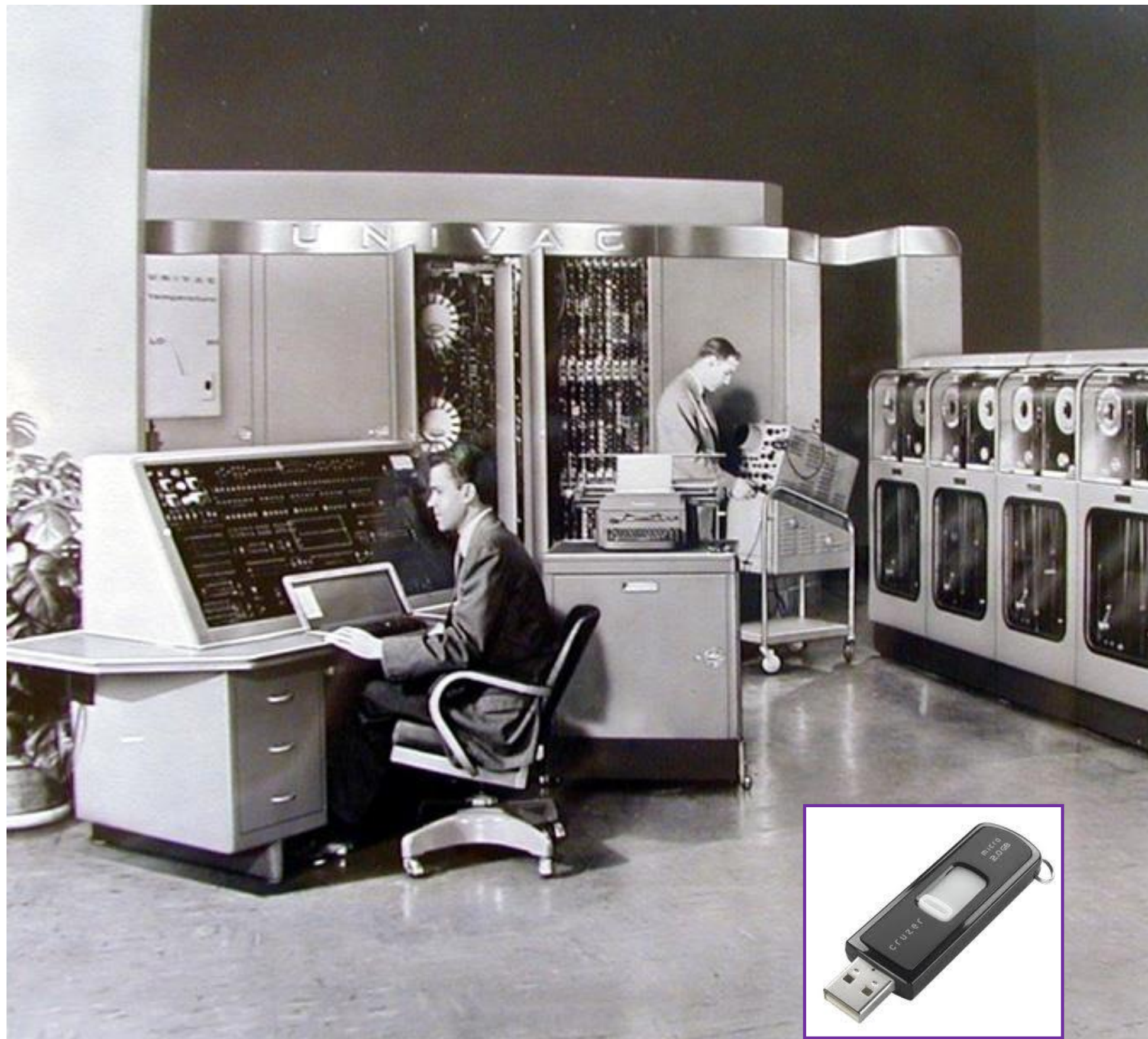
la Bombe (1940-45)

L'expression **Intelligence artificielle** apparaît pour la première fois en **1956**.

Plusieurs chercheurs américains, dont John McCarthy et Marvin Minsky se sont réunis à l'université de Dartmouth aux Etats-Unis. Leur projet est alors de développer l'intelligence artificielle. Le terme reprenait le projet cybernéticien de marier l'étude du cerveau avec celle des machines.

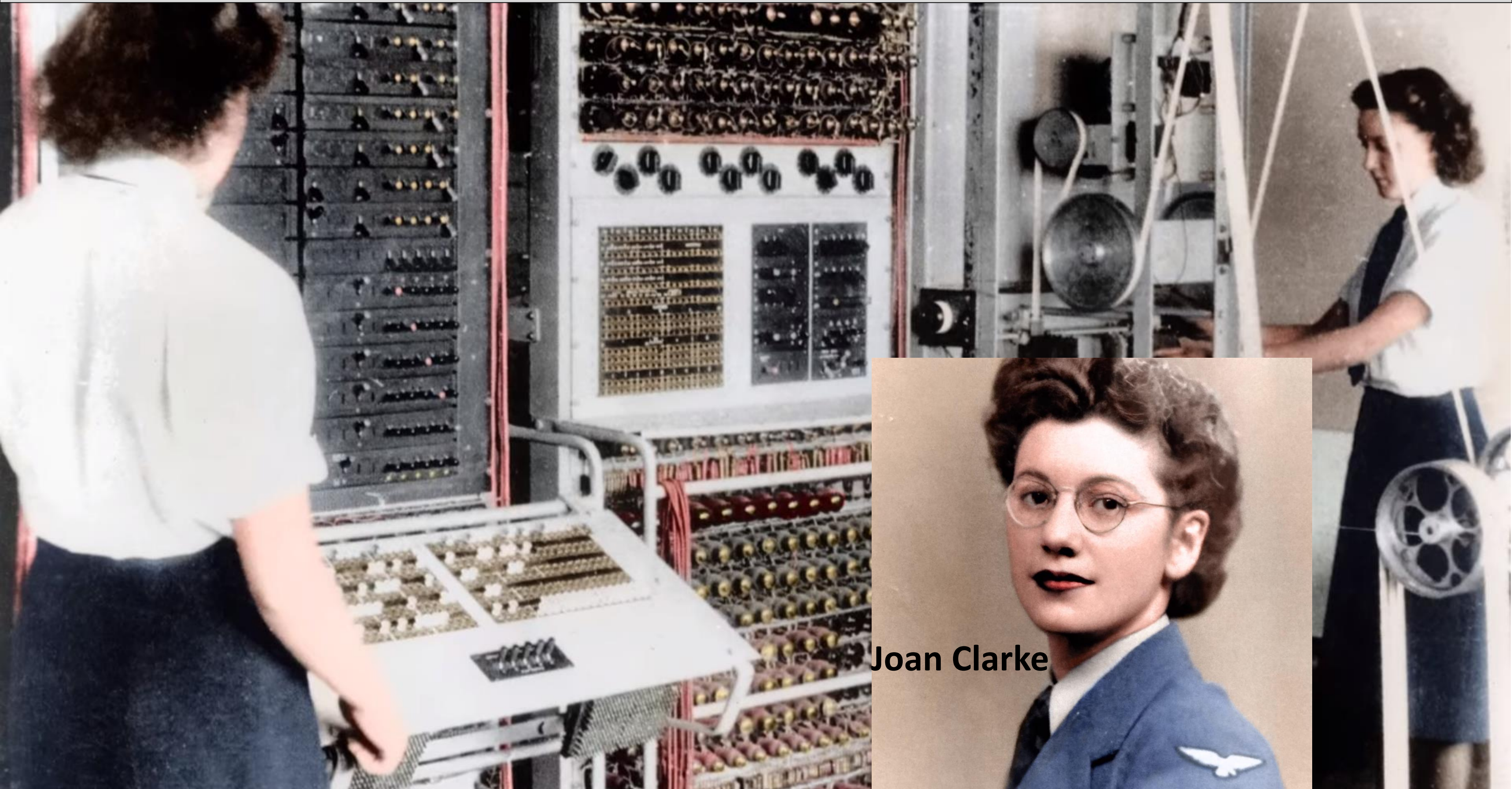
Les premiers gros ordinateurs et leur utilisation civile

L'Univac (1951)



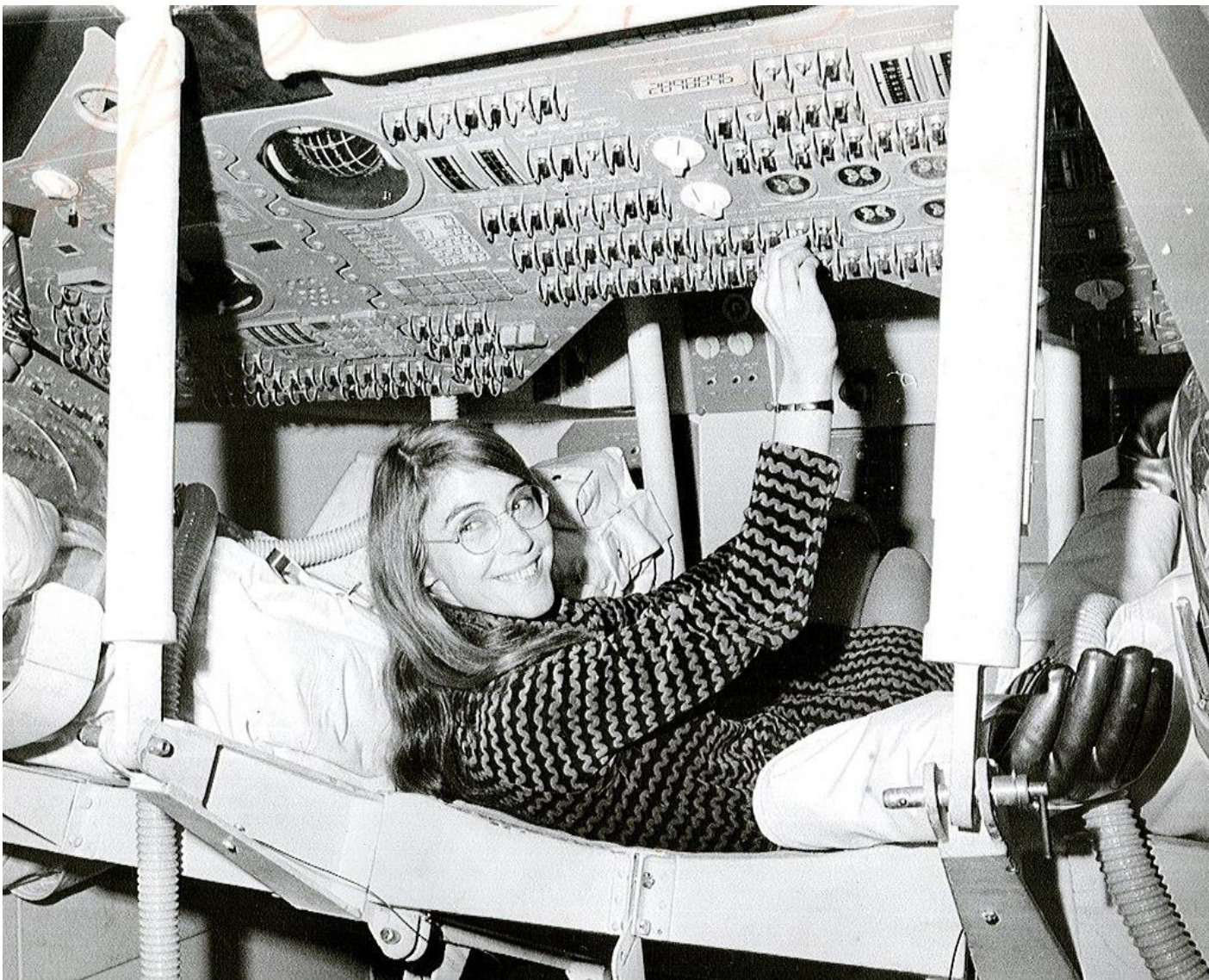
Carte perforée (instruction)

L'apport des femmes au développement de l'informatique



Joan Clarke

L'apport des femmes à l'informatique de programmation



Margareth Hamilton (née en 1936)

Programme spatial Apollo



Mary Jackson (1921-2005)

Informaticienne et ingénieure aérospatiale (NASA).

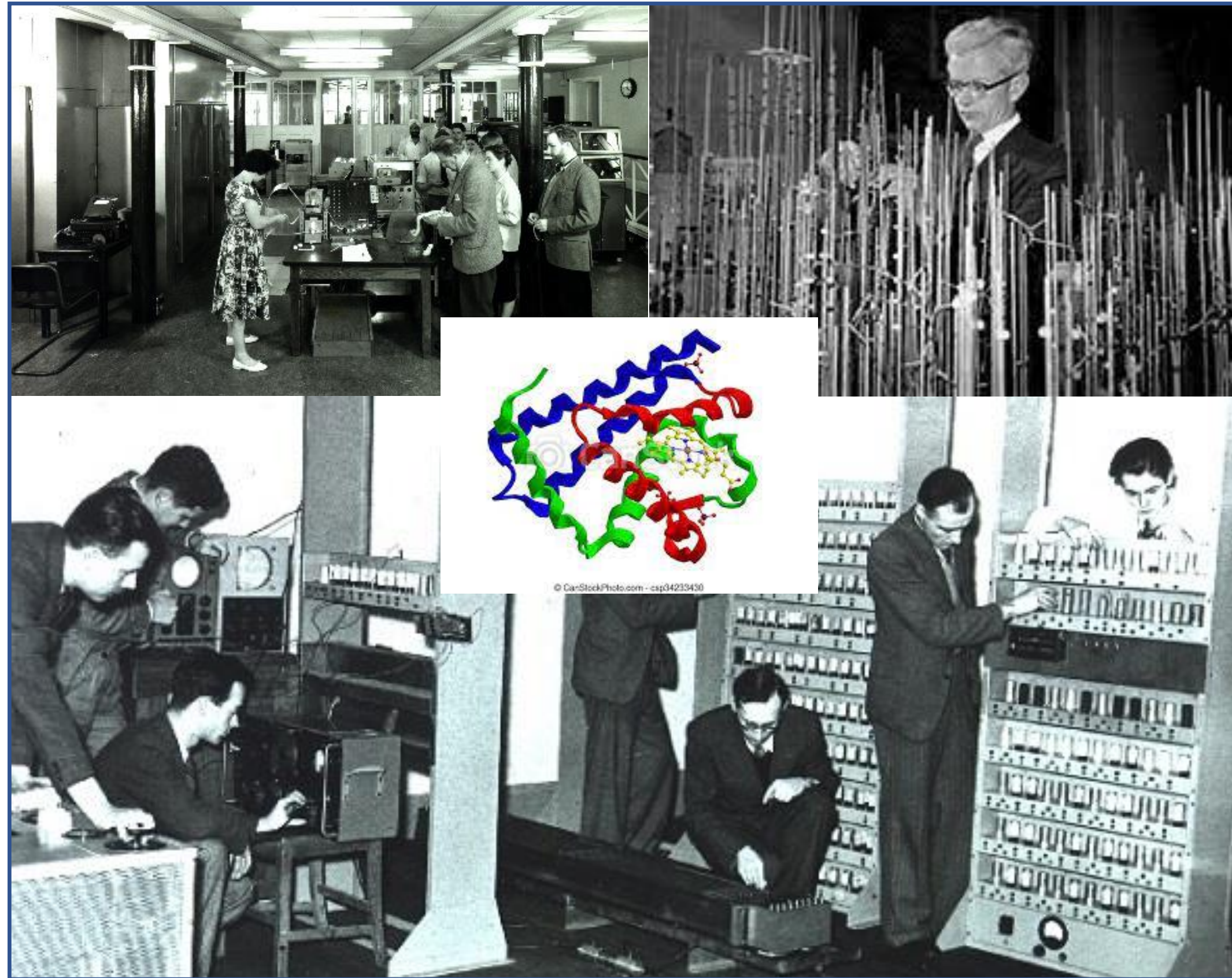
Le film « Les figures de l'ombre » retrace les destins de Katherine Johnson, Mary Jackson et Dorothy Vaughan qui ont permis de grandes avancées pour la NASA

Au fait, quelle fut la première contribution de l'informatique et de l'univers computationnel à la biologie?

La résolution de la S3D de la mioglobine

Et toute la chaîne de programmes pour résoudre graphiquement la structure cristallographique des protéines mais ça c'est une autre histoire, 1965, environ 30 ans avant le séquençage du génome humain (EDSAC 2 (Electronic Delay Storage Automatic Calculator du Mathematical Laboratory à Cambridge)).

Aujourd'hui les applications des data sciences à la biologie sont innombrables et en faire un historique ferait de vous des experts, une vraie branche de l'histoire des sciences à développer



Margaret Dayhoff (dans les années 1960-1980)

Pionnière dans l'application des méthodes mathématiques et informatiques à la biochimie, elle a en particulier mis en place les premières grandes bases de données de gènes et de protéines et les outils pour les interroger.

Elle est également connue pour le code à une lettre des acides aminés, facilitant les calculs et pour la fameuse matrice de substitution qui a contribué à la comparaison et la prédiction de la structure 2D et 3D des protéines.

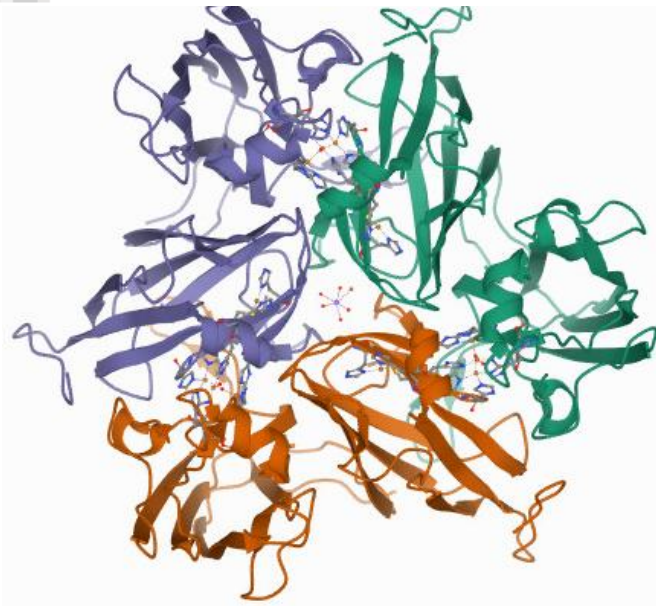
Tout prenait du temps à cette époque mais on faisait déjà du bon boulot.

Chapeau madame Dayhoff pour cette belle contribution!
(Belle est son second prénom)!



```
1 -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD-----LSHGS 53 P69905 HBA_HUMAN
1 MVHLTPEEKSAVTALWGKVNV--DEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAMGN 58 P68871 HBB_HUMAN
1 MVHLTPEEKTAIVNALWGKVNV--DAVGGEALGRLLVVYPWTQRFFESFGDLSSPDVAMGN 58 P02042 HBD_HUMAN
  : ** : * : * : * : * : * : * : * : * : * : * : * : * :
54 AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLKRVDFVNFKLLSHCLLVTLAAH 113 P69905 HBA_HUMAN
59 PKVKAHGKKVLGAFSDGLAHLDNLKGTFTAILSELHCCKLHVDPENFRLLGNVLVCVLAHH 118 P68871 HBB_HUMAN
59 PKVKAHGKKVLGAFSDGLAHLDNLKGTFTSQSELHCCKLHVDPENFRLLGNVLVCVLAARN 118 P02042 HBD_HUMAN
  : ** : * : * : * : * : * : * : * : * : * : * : * : * :
114 LPAEFTPAVHASLDKFLASVSTVLTISKYR 142 P69905 HBA_HUMAN
119 FGKEFTFPVQAAYQKVVAGVANALAHKYH 147 P68871 HBB_HUMAN
119 FGKEFTPQMQAAYQKVVAGVANALAHKYH 147 P02042 HBD_HUMAN
  : ** : * : * : * : * : * : * : * : * : * : * : * : * :
```

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901



Vous et la Data Science

Du travail, des contraintes mais ça vaut le coup!

Avec un peu de travail, vous participerez, vous la génération Z, à un monde nouveau et mieux maîtrisé, notamment en Biologie : écologie, santé, environnement, Dans un contexte humaniste

Etre biologiste?

« On » vous l’a maintes fois répété : être biologiste, c’est faire de la bio et de la bio et encore de la bio et rien d’autre! ...

Ni le temps ou l’occasion de faire autre chose...

En clair les **biologistes** seraient des **pouvoyeurs de données** mais n’auraient ni la compétence, ni le droit (droit en tant que spécialistes) d’analyser leurs propres données; ce rôle étant confié à d’autres (les « gentils » mathématiciens, modélisateurs et autres statisticiens ou informaticiens).

Les robots commencent à faire des tâches qu’il y a peu techniciens, ingénieurs et chercheurs réalisaient seuls ou en équipe.

Les robots et autres automates et ordinateurs seront bientôt ... vos collègues!

Soyez acteurs de ce nouveau monde

Ne vous faites pas manipuler! Donnez-vous des armes pour ne plus être à la solde de gens plus ou moins bien intentionnés.

Apprenez à faire parler les données pour contribuer à un monde plus humaniste.

Vous et la Data Science

Vous pouvez créer votre propre entreprise ou votre propre emploi

Ne pas en avoir peur.

Le monde change,

entrer dans une grande entreprise, assister à des réunions stériles et porter cravate ou tailleur est un modèle dépassé.

Vous ne pourrez pas dire : « je ne savais pas! »

A vous de jouer!

Saurez-vous relever ce défi?

La balle est dans votre camp.

Ce que vous allez apprendre

Ce que vous allez apprendre

- **Formater et nettoyer les données**
- **Manipuler les données** d'un jeu de données (data set)
- **Créer base de données** (Access, MySQL)
- **Analyse statistique pratique** (régression linéaire multiple, MCO [OLS] en anglais, régression logistique et risk ratios, Analyse factorielle, R (interfacé))
- **Machine learning**, apprentissage supervisé et non supervisé;
Si vous aimez les animaux : Python, Anaconda, Pandas, Spider, Jupiter..
- **Communication** : comment présenter les résultats en restant honnête
- Du vocabulaire spécifique bien sûr
- **Comment gérer et conduire un projet** (cahier des charges, mile stones, méga instructions, méthode agile, chef de projet,...)
- Mais ce sera à vous de choisir ce pourquoi vous vous sentez motivés et de de pousser plus loin en autoformation (étant donné la pression sur les masters bio "pas de place pour du hors bio" mais rassurez-vous dans d'autres départements ou facs ou bien dans d'autres pays, on a compris qu'il y avait de la place pour ce hors bio et qu'il fallait se dépêcher d'en faire avant d'être à la ramasse)
- Vous êtes 52 à avoir choisi cette UE, c'est un effectif assez important
- En conséquence, cet effectif va nous demander un peu de **discipline** pour **honorer le cdc** et arriver au terme de cet apprentissage

Projet Personnel d'Etudes et d'Insertion (PPEI)

⇒ une formation obligatoire à la conduite de projet

Les Projets

Le but des UEs «**Projet Personnel d'Etudes et d'Insertion** » est de vous apprendre à construire et concrétiser un **projet en groupe** dans une thématique donnée (dans votre domaine scientifique de prédilection).

Dans cette UE « **Data Science** » vous réaliserez vos projets par groupes de 3 ou 4 (4 étant l'idéal).

Attention, cela demande un réel travail partagé, une bonne **coordination**, un bon **esprit d'équipe** et une bonne **organisation**!

Vous avez jusqu'à **début mai** (date à définir en fonction de vos examens) pour **réaliser et présenter par écrit et oralement votre projet**.

Cela va aller plus vite que vous ne l'imaginez.

Les Projets, un travail d'équipe

Projets : les 4 compétences

- > Base de données,
- > Analyse statistique (avec R),
- > Machine learning (en programmation python ou R),
- > Communication

Et

- > le/la chef(fe) de projet

(crédité de 1,5 points de bonus)

mais attention il s'agit d'une charge et d'une responsabilité importantes

Les rôles utiles des postit, du tableau blanc (paper board) et du story telling!



Les équipes-projets

Votre **équipe-projet** contiendra **4 membres**, chacun des membres se spécialisant plus sur une des 4 expertises/compétences (**Bases** de Données et SQL, **Analyse Statistique** et R, **Machine Learning** et Python, **Communication**, design & style écrit et oral).

L'un(e) de ces 4 membres assurera la fonction de **chef(fe) de projet** (créditée de points bonus, conditionnés au résultats obtenus).

Bref, vous réaliserez un **vrai travail d'équipe** comme il est de coutume de le faire en *data sciences* en réunissant des **compétences**, des **aspirations** et des **talents complémentaires** qui servent un **projet commun**, le votre.

Les équipes-projets

La thématique sera empruntée, au choix, **dans n'importe quel domaine de la biologie** qui intéresse les membres de l'équipe, projet biologique de "biodatasciences".

Votre **enseignant** sera à la fois **animateur**, supporter et **membre mobile de chaque équipe**. Il est là pour vous remonter le moral au cas où.

Vous aurez tous les mêmes enseignements mais vous pourrez décider de vous orienter vers un des aspects qui vous séduit le plus dans les data sciences. On choisira (peut-être) des jetons couleurs virtuels (je vous expliquerai) pour déterminer ce choix.

Les Projets

Vous en sortirez **enrichis**, formés à la conduite de projet et pourrez valoriser cette expérience sur **votre CV**

[YouTube ou un Podcast dédié (pour les **3 meilleurs** projets)].

Je réfléchis à un **prix pour récompenser les 5 meilleurs projets**.

Proposition de création de startup (en accompagnement total) pour les étudiants qui seront intéressés.

Une courte **vidéo** explicative sera bientôt réalisée à propos de vos projets.

Les mains dans le cambouis en salle informatique

On commence dès aujourd'hui

La première séance en salle informatique

Bâtiment 625 (hbar) sur le plateau
(bâtiment d'enseignement de la physique)

Salle 306



Nous allons travailler en **Salle 306** sur un jeu de données au format CSV
et manipuler R et Python (JupyterLab, Pandas)
Si j'en ai le temps, je vous présenterai l'ENT (site web)