**Pascal RIGOLET**

Bat 112 Institut Curie – Paris-Saclay

06 20 65 77 70

pascal.rigolet@u-psud.fr

**BADE**

(**BA**se de **D**onnées pour l'**E**nvironnement)

**Apprendre à créer et gérer une Base de Données**

⇨ **Présentiel + Distanciel (+ Comodal?)**

**Mesures Covid19 à respecter**

**Connexion windows**

User : prénom.nom@u-psud.fr
(ou : prénom.nom@universite-paris-saclay.fr)

Password : celui de votre messagerie

# BASES ET BANQUES DE DONNEES
## Objectifs de l'enseignement

- **Apprendre à créer et gérer une Base de Données**
- **Nous disposons de 8 séances pour tout faire**

  ⇨ **L'ACCENT SERA MIS SUR LA PRATIQUE**
- **Le site web de l'enseignement sera bientôt opérationnel**
- **Contrôle des connaissances :** vous rendrez **un projet** **en 3 parties**

  **1 -** Création d'une **base de données** personnelle [coef 3]

  (application : Environnement/Pollution Chimique/Géologie/Ecologie)

  **2 – Rapport de 12 pages** sur cette base de données [coef 2]

  *(objectifs ; de la création à l'utilisation ; format pdf)*

  **3 – Oral de présentation du projet** [coef 2]

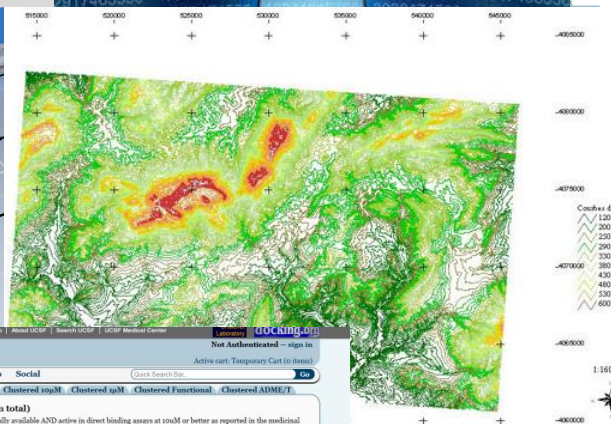
  *(dont démonstration de l'utilisation de la base de données + diaporama)*

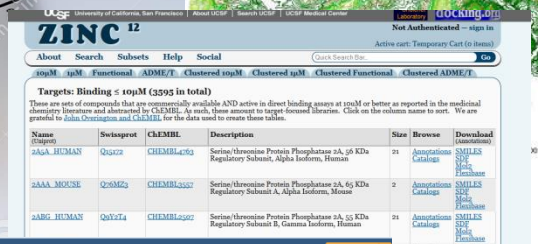
- On commencera les projets en classe

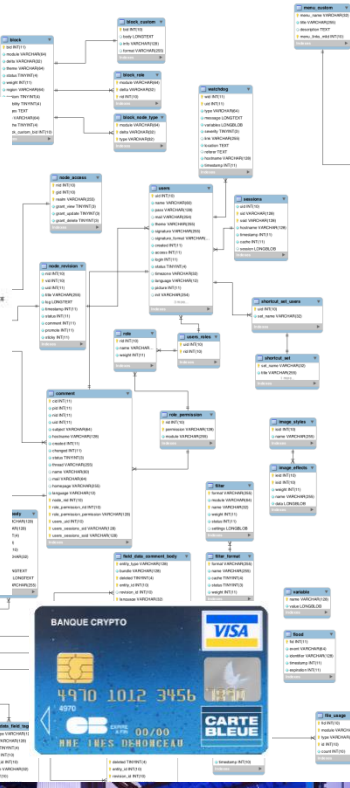
# Système de Gestion de Base de Données (SGBD)
# A – Introduction

- **Qu'est-ce qu'on entend généralement par BASE DE DONNEES ?**
- **Petit historique**
- **Limites, fiabilité et stockage des données**
- **Vers la structuration des données**
- **Quelques exemple de bases connues**
- **Une base de données est généralement associée à des services**
- **Quelques mots de vocabulaire**
- **Codage ASCII**

# Base de Données





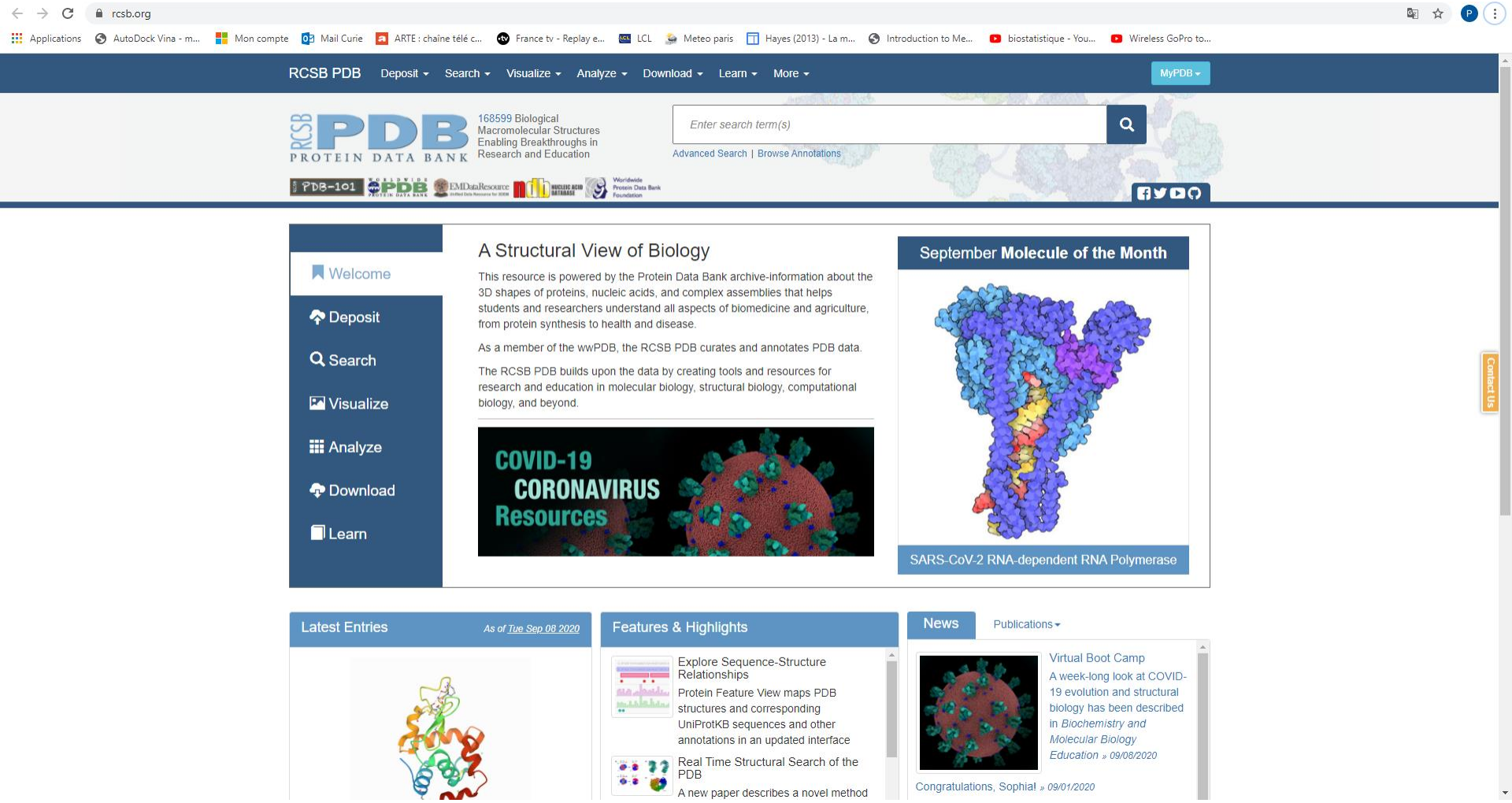




**Big Data :** tous les deux jours, l'humanité produit autant d'information que ce qu'elle a généré depuis l'aube de la civilisation jusqu'en 2003. Plus de 90% des données disponibles ont été produites ces 2 dernières années; ce volume double tous les deux ans d'où l'enjeu considérable que constituent le **traitement**, **l'analyse**, le **stockage** et le **décryptage** de ces « mégadonnées » (« données massives »)

**Chaque année**, l'Humanité produit plus de 2000 **Exabit** d'informations nouvelles ($10^{21}$ bits, 1000 milliards de milliards de bits), soit 2 Zettaoctets et a déjà stocké **en 2020 plusieurs dizaines de Zettaoctets** d'information ($10^{22}$ bits,10000 milliards de milliards), environ 40 milliards de téraoctets… Cela a un coût ⇨ **Matière + Energie + Eau**

# PDB


rcsb.org
Applications
AutoDock Vina - m...
Mon compte
Mail Curie
ARTE : chaîne télé c...
France tv - Replay e...
LCL
Meteo paris
Hayes (2013) - La m...
Introduction to Me...
biostatistique - You...
Wireless GoPro to...
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
168599 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search term(s)
Advanced Search | Browse Annotations
PDB-101
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
A Structural View of Biology
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.
COVID-19 CORONAVIRUS Resources
September Molecule of the Month
SARS-CoV-2 RNA-dependent RNA Polymerase
Contact Us
Latest Entries
As of Tue Sep 08 2020
Features & Highlights
Explore Sequence-Structure Relationships
Protein Feature View maps PDB structures and corresponding UniProtKB sequences and other annotations in an updated interface
Real Time Structural Search of the PDB
A new paper describes a novel method
News
Publications
Virtual Boot Camp
A week-long look at COVID-19 evolution and structural biology has been described in Biochemistry and Molecular Biology Education » 09/08/2020
Congratulations, Sophia! » 09/01/2020

# Système de Gestion de Base de Données (SGBD)
# E –Les données dans la base

- Quantité et qualité des données
- Performance de la base
- Exploitabilité des données
- Le moindre détail peut avoir de l’importance
- Meilleur compromis

# *Qu'est-ce qu'une base de données?*

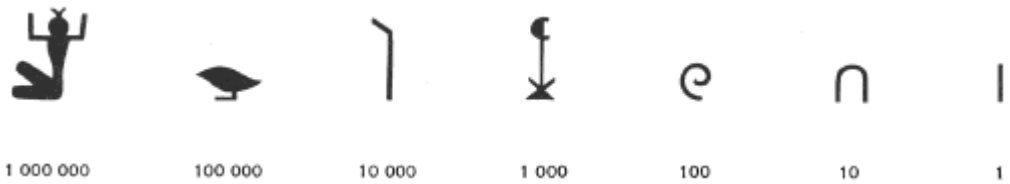
# Système de Gestion de Base de Données (SGBD)
# E – Quelques définitions

- **Une base de données c'est essentiellement une collection structurée d'informations non nécessairement du même type (format)**
- **Une base de données est usuellement localisée en un seul lieu et un seul support (dupliqué en fait) qui est généralement informatique (numérique)**
- **Pièce centrale des dispositifs informatique qui servent à la collecte, au stockage et à l'utilisation des informations**
- **SGBDR : logiciel moteur qui pilote la base et en permet la manipulation et l'exploitation (interrogation)**
- **Toutes les secondes le volume d'information ne cesse d'augmenter contribuant au Big Data. Mais sans analyse et sans base de données (contribuant à préparer les données à leur analyse), le Big Data n'est rien.**

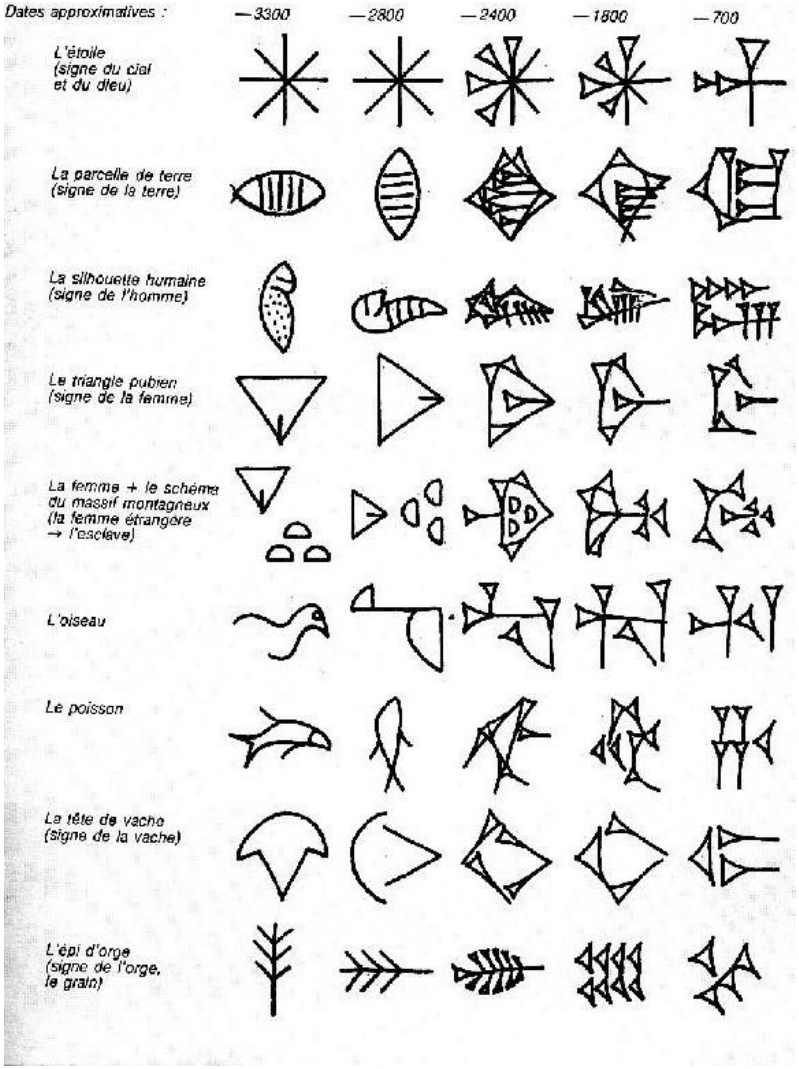
# Système de Gestion de Base de Données (SGBD)
# A-2 – Quelques exemples de bases de données identifiables

- Décomptes et registres (Babylone, Haute Egypte, antiquité,  moyen âge)

*Babylone, écriture cunéiforme*




*Les chiffres et nombres, haute Egypte*


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 schème 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)

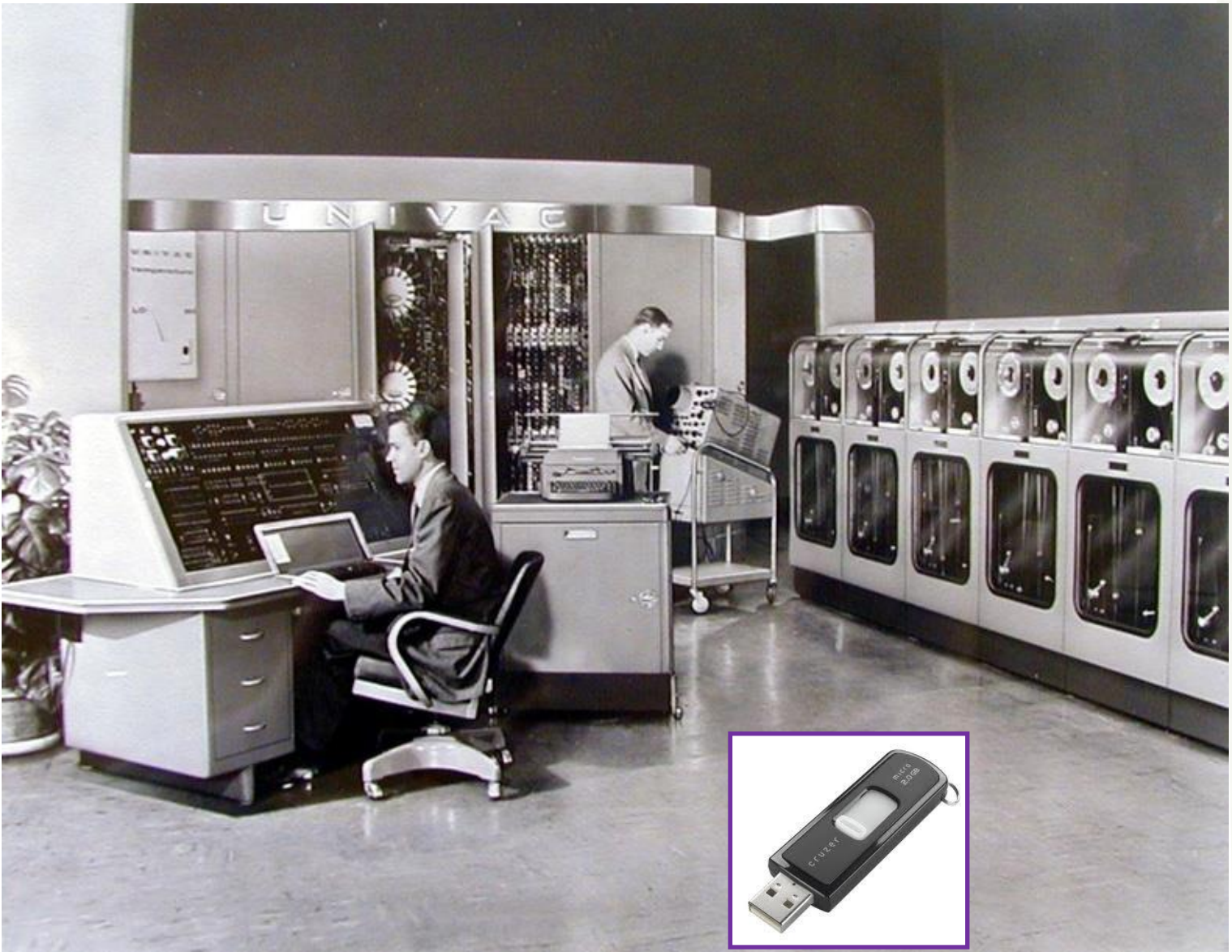

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59

# Système de Gestion de Base de Données (SGBD)
# A-2 – Quelques exemples de bases de données identifiables

- Décomptes et registres (Babylone, Haute Egypte, antiquité,  moyen âge)
- Archives dans l'histoire
- Coordonnées bancaires (dès le 19ème siècle)
- Compagnies de téléphone
- Clients/fournisseurs
- Horaires des train et réservations de billets
- Fiches de police – Carte d'identité, Passeport, …
- Sécurité Sociale (vers 1946 avec fiches support papier, fichiers boîtes) puis ordinateurs et fichiers (N° INSEE)
- Premier ordinateurs commandés par les USA pour applications civiles : UNIVAC recensement de 1950
- Loi informatique et liberté de 1976
- Google, Yahoo et autres « moteurs de recherches »
- Face book (et ses dangers)
- Apogée  et tous les fichiers gérant l'inscription et les notes des étudiants
- Edwige (un projet avorté?)
- **Chaque français en moyenne est enregistré dans une centaine de DB**

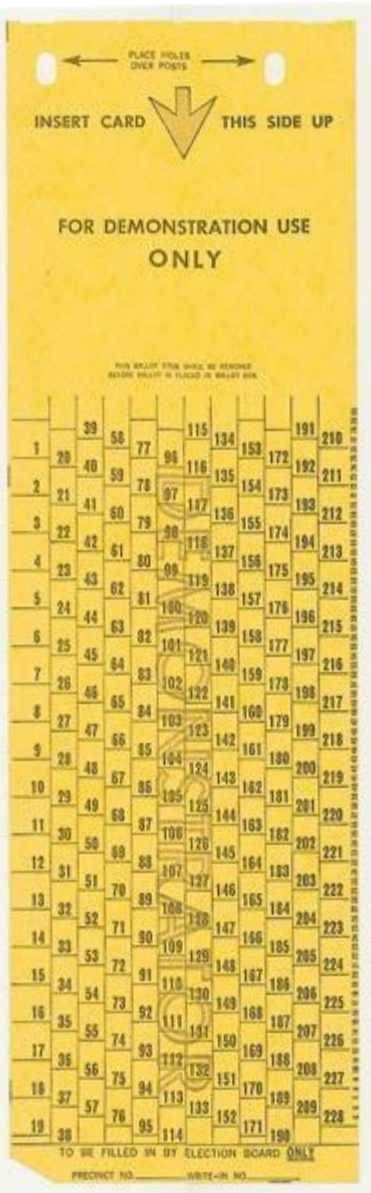
# Système de Gestion de Base de Données (SGBD)
# A-2 – Quelques exemples de bases de données identifiables

## L'Univac (1951)









## *Les limites du monde connu …*

**Gogol** (« Googol ») **: $10^{100}$** ⇨ grand nombre versus infini

- Très supérieur au nombre de particules dans l'Univers connu (environ $10^{80}$)

  et pourtant simple à écrire (1 suivi de 100 zéros)

  mais très inférieur au nombre de cubes de Planck (côté de la longueur de Planck : $1{,}616 \times 10^{-35}$ m) dans l'univers observable.
- Mathématicien Edward Kasner (« *Mathematics and the Imagination* » - 1938)
- Un peu de légende geek (son neveu de 9 ans aurait trouvé ce nom)
- Origine du nom explicitement revendiqué par les fondateurs de Google
- $10^{gogol}$ : « gogolplex »
- Une certaine écriture de l'infini?...
- Si unité = mètre : infiniment petit (connu) : **$10^{-32}$** ; infiniment grand (connu) : **$10^{32}$**
- On ne peut pas représenter l'infini dans une base de donnée : monde concret/connu
- Génome humain : 23000 gènes ; 3,2 milliards de pdb ; 750 Mo (haploïde)
- Diamètre de l'univers observable : $9.2 \times 10^{10}$ années-lumière

# Système de Gestion de Base de Données (SGBD)
# E – Fiabilité des données

***Hélas, les données ne sont ni fiables, ni éternelles…***

- Utilité d'une base de données (il faut qu'elle serve à quelque chose)
- Sécurité des données
- Contenu = vérité ?
- Stockage fichiers en plusieurs exemplaires (non nécessairement identiques)
- Versions ; date mise à jour
- Stockage : altération avec le temps
- Autres problèmes …

# Système de Gestion de Base de Données (SGBD)
# E – Fiabilité des données

***Support de l'information : un problème à moyen et long terme…***

- **Conservation des données**

  Papier : ~ 1000 ans
  Pierre : ~ 10000 ans
  Bambou & papyrus : ~ plusieurs milliers d'années
  Peaux : plusieurs centaines d'années
  Grottes : > 35000 ans (Chauvet)
  Toiles, peintures, mosaïques : 1000 à 3000 ans
  Argile (écriture cunéiforme) : > 4500 ans
  Film, photos, ciné : ~ 100 ans (demandent ensuite restauration)
  CD : 5 à 10 ans maxi
  Clé USB : 0 à 20 ans maxi
  Disquette : 15 ans
  Bandes magnétiques : 100 ans et 1000 ans (nouvelle génération)
  Disque dur : 5 à 10 maxi
  Cloud : ??? (dépendant du reste; supports changés régulièrement)
  ADN : 300000 ans
  Quartz (gravure) : > 1 milliard d'années!!

- Produire, stocker et exploiter **les data consommeront** d'ici un siècle à un siècle et demi environ toute la matière et toutes les **ressources de la planète !!!**

# Système de Gestion de Base de Données (SGBD)
# E – Quelques définitions

**Une base de données (relationnelle) c'est essentiellement une collection (structurée) d'informations non nécessairement du même type (format)**

*(Il est vrai qu'il existe des bases de données non structurées pour le big data mais qu'il faut passer par une structuration des données pour pouvoir les traiter, les analyser et les exploiter)*

⇨ **Il va donc falloir apprendre à stoker et organiser l'information en vue de son exploitation : il existe des méthodes pour cela; cela ne s'improvise pas.**

⇨ **Pas de place pour le bazar…**

⇨ **mais la logique et le bon sens ont tous leurs droits!**