Big Data Analytics Trends
storage, network science and graph stores

Genoveva Vargas-Solar
Senior Scientist, French Council of Scientific Research, LIG-LAFMIA
genoveva.vargas@imag.fr

http://vargas-solar.com/big-linked-data-keystone/

Big is not a matter of size …

it is a matter or **representativity** & **consumption capacity**

HOW BIG IS YOUR DATA REALLY

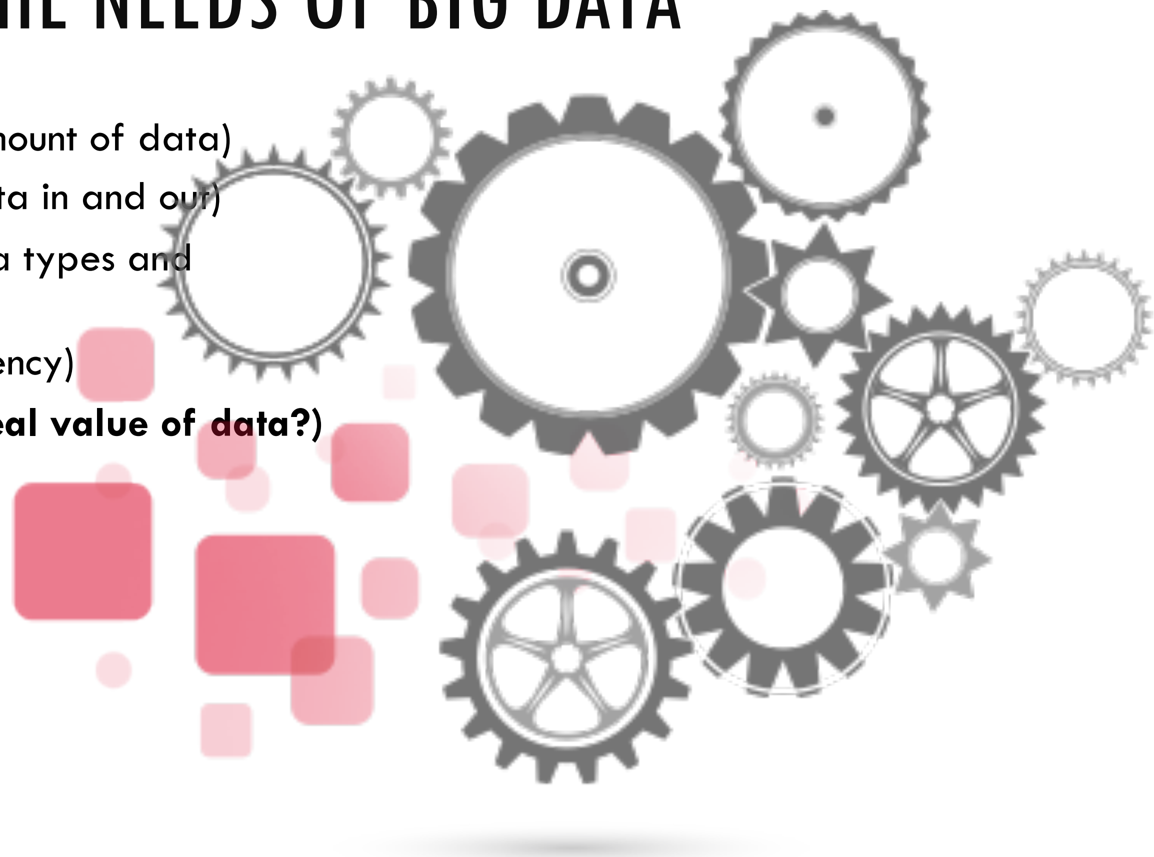
Unit	Size	
Byte (B)	8 bits	One grain of rice
Kilobyte (KB)	2^{10} bytes	A cup of rice
Megabyte (MB)	2^{20} bytes	8 bags of rice
Gigabyte (GB)	2^{30} bytes	3 container lorries
Terabyte (TB)	2^{40} bytes	2 container ships
Petabyte (PB)	2^{50} bytes	Covers Manhattan
Exabyte (EB)	2^{60} bytes	Covers the UK (3 times)
Zettabyte (ZB)	2^{70} bytes	Fills the Pacific ocean

David Wellman

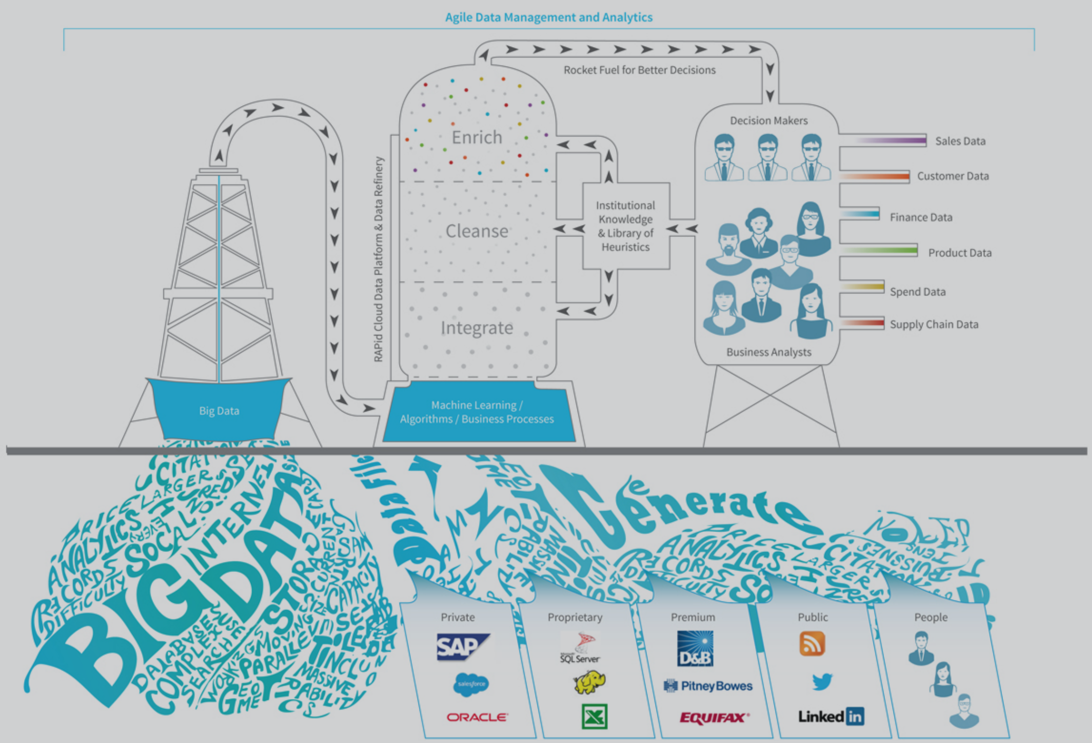
Collection of data sets **so large** and **complex** *that it becomes* **difficult** to **process** *using* **on-hand database management** *tools or* **traditional** *data* **processing** applications

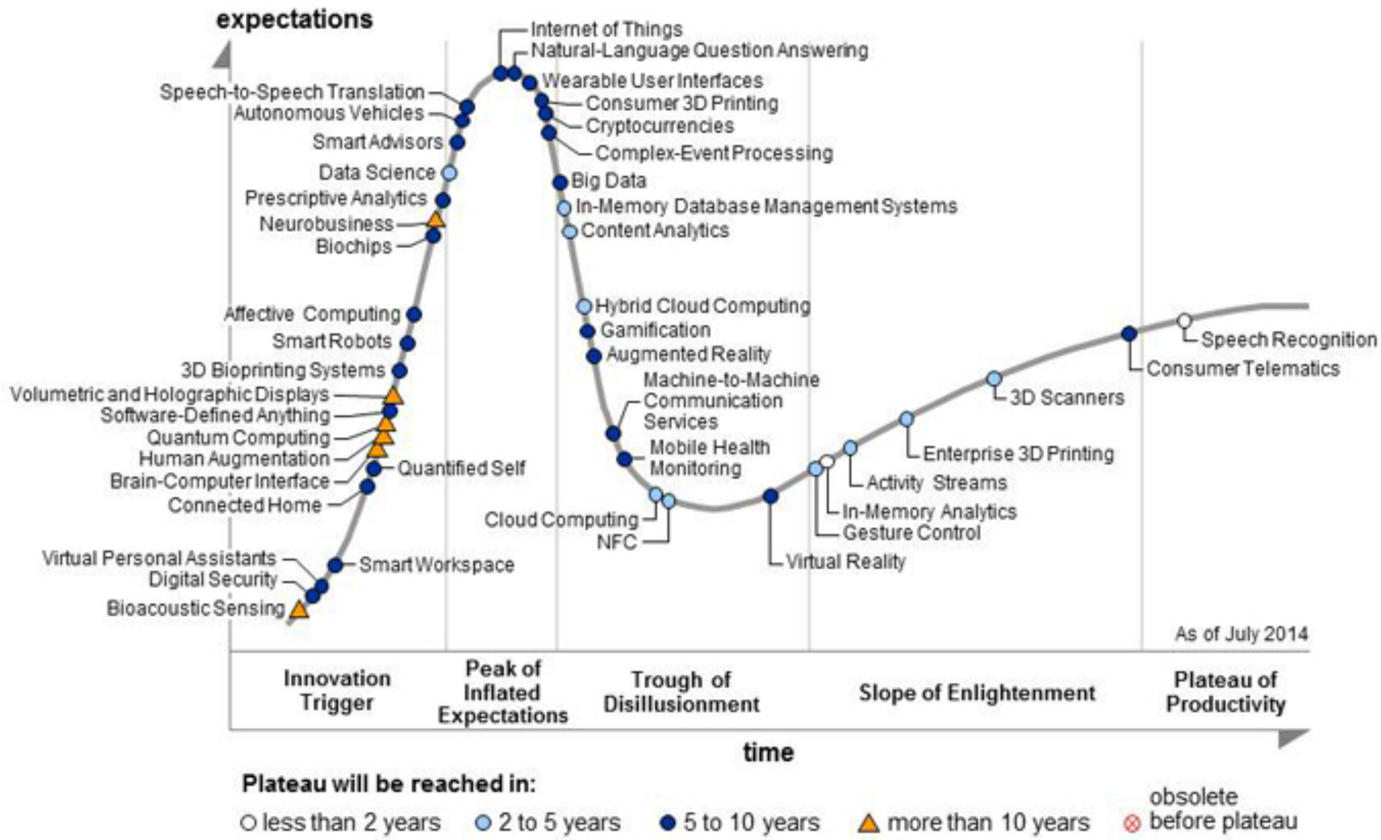
THE V'S & THE NEEDS OF BIG DATA

- increasing **volume** (amount of data)
- **Velocity** (speed of data in and out)
- **Variety** (range of data types and sources)
- **Veracity** (data consistency)
- **Value (which is the real value of data?)**



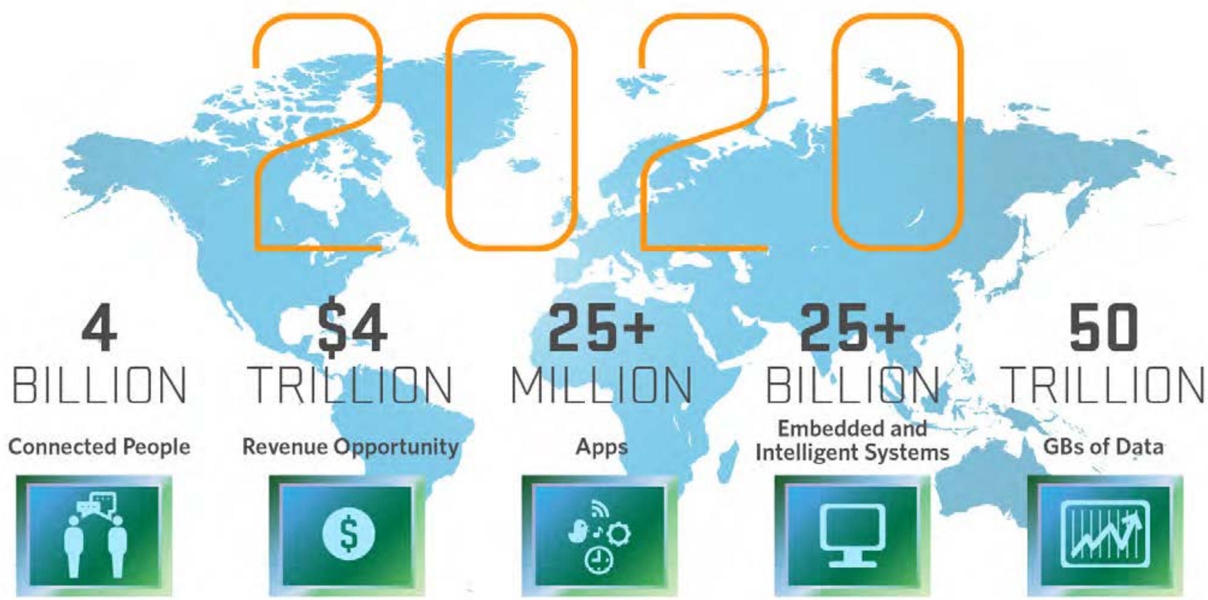
BIG DATA PROCESSING AT GLANCE





http://www.gartner.com/newsroom/id/2819918

INTERNET OF THINGS



Source: Mario Morales, IDC

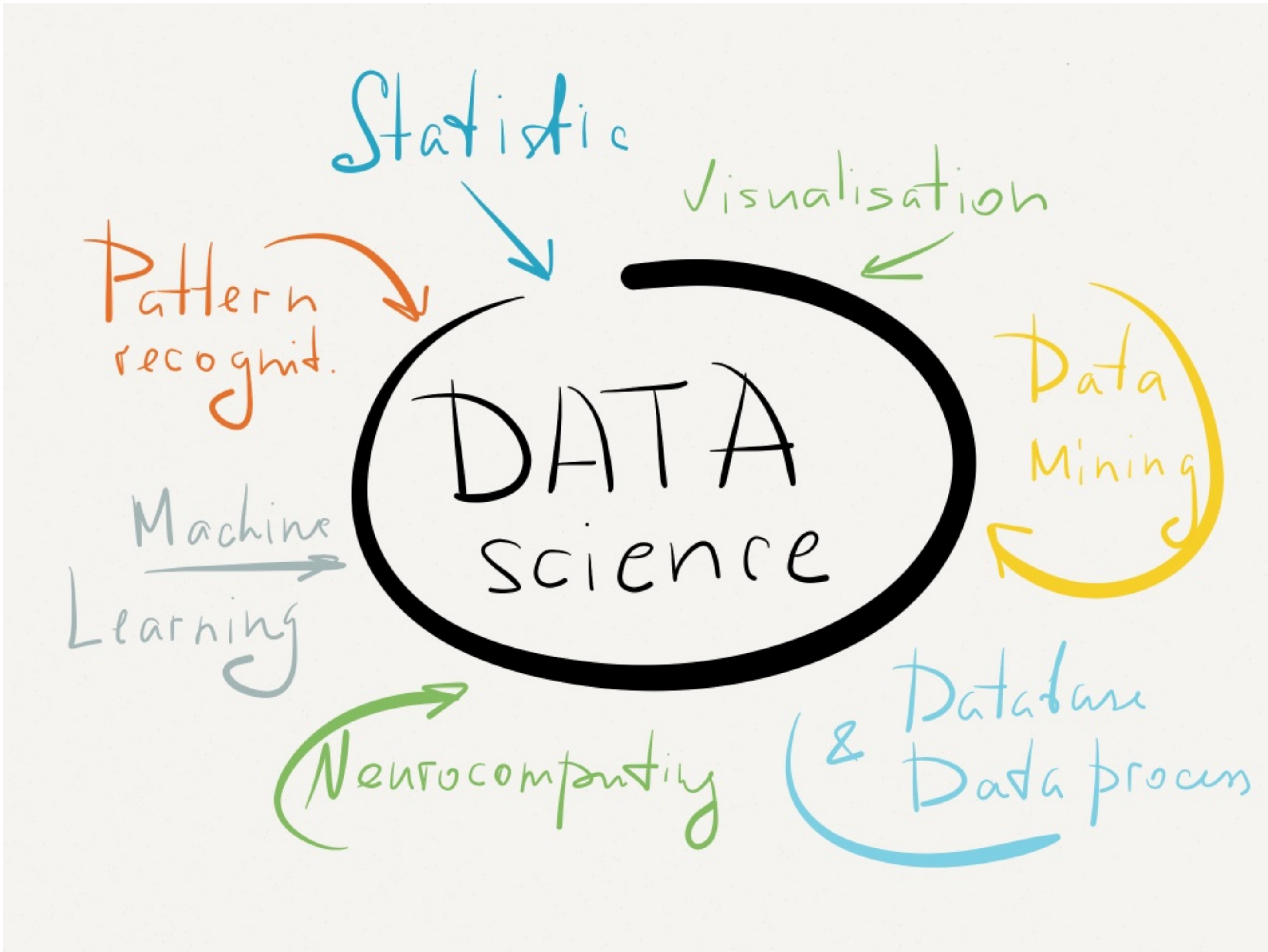
BIG DATA AT A BRONTO SCALE

1 bit	Binary digit
8 bits	1 byte
[illegible]	[illegible]
1000 Petabytes	1Exabyte
1000 Exabyte	1 Zettabyte
1000 Zettabytes	1 Yottabyte

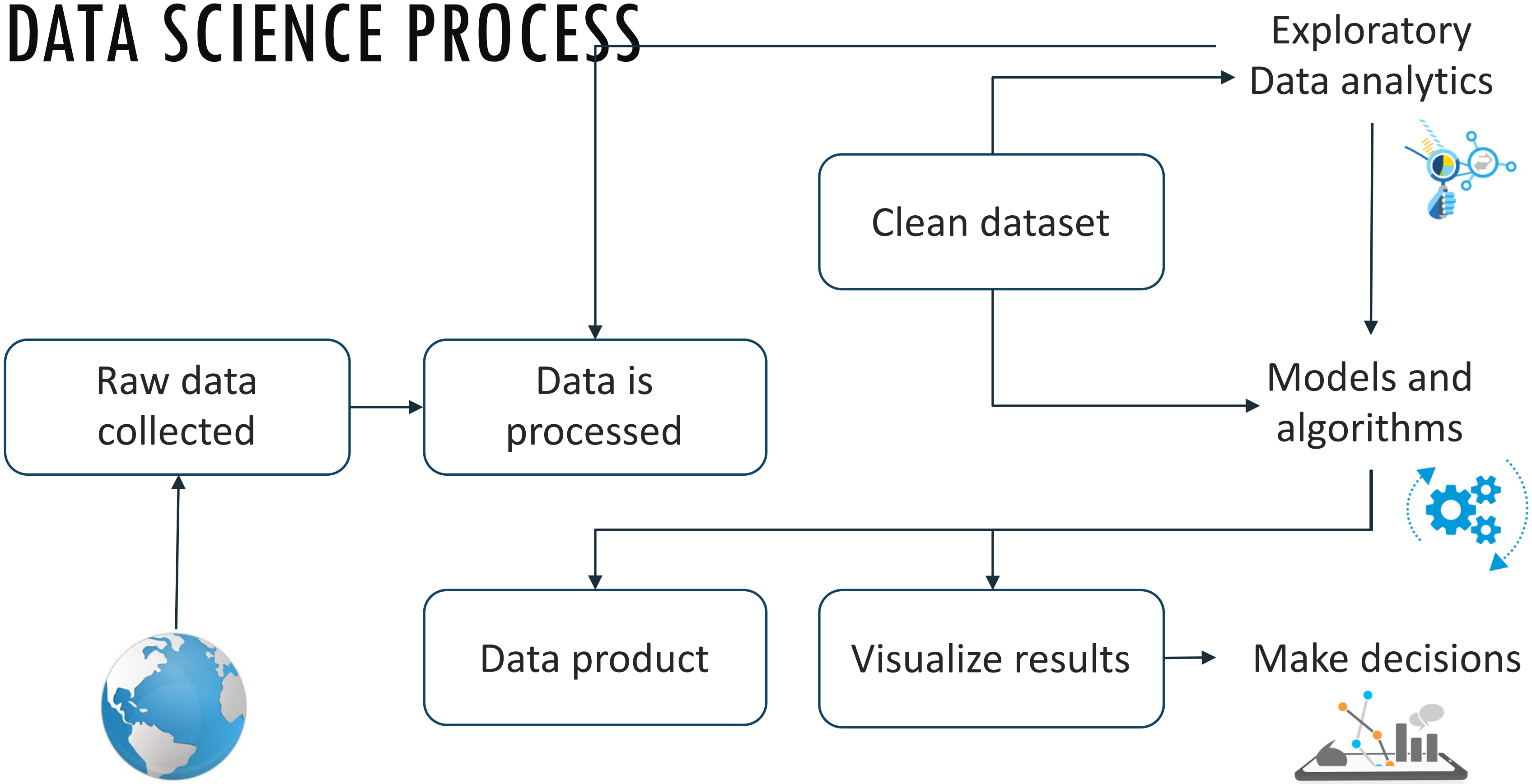
We will no longer have the luxury of dealing with just "big" data

http://spectrum.ieee.org/computing/software/beyond-just-big-data


You are here
Map of the Internet, The Opte Project, www.opte.org


Statistic
Visualisation
Pattern recognit.
DATA science
Data Mining
Machine Learning
Neurocomputing
& Database Data process

DATA SCIENCE PROCESS



What about analytics ?

PRINCIPLE

Given lots of data

Discover patterns and models that are:

- **Valid:** hold on new data with some certainty
- **Useful:** should be possible to act on the item
- **Unexpected:** non-obvious to the system
- **Understandable:** humans should be able to interpret the pattern

NEW TYPES OF HUGE DATA COLLECTIONS

Thick data: combines both quantitative and qualitative analysis,

Long data: extends back in time hundreds or thousands of years

Hot data: used constantly, meaning it must be easily and quickly accessible

Cold data: used relatively infrequently, so it can be less readily available

http://spectrum.ieee.org/computing/software/beyond-just-big-data

DATA COLLECTIONS

Different sizes, evolution in structure, completeness, production conditions & content, access policies modification …

Data collections' releases

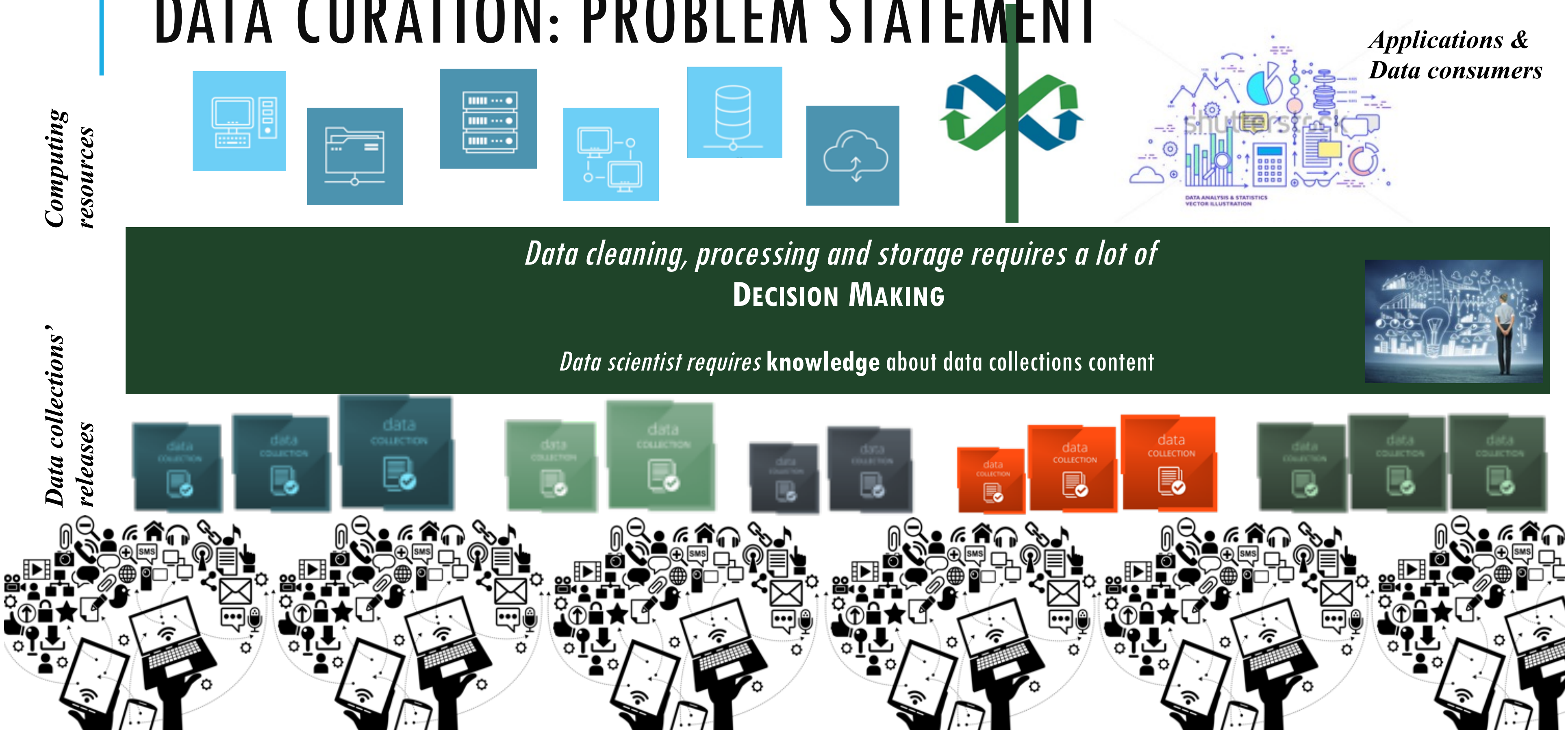
DATA COLLECTIONS

NOT MANAGEABLE NEITHER EXPLOITABLE AS SUCH

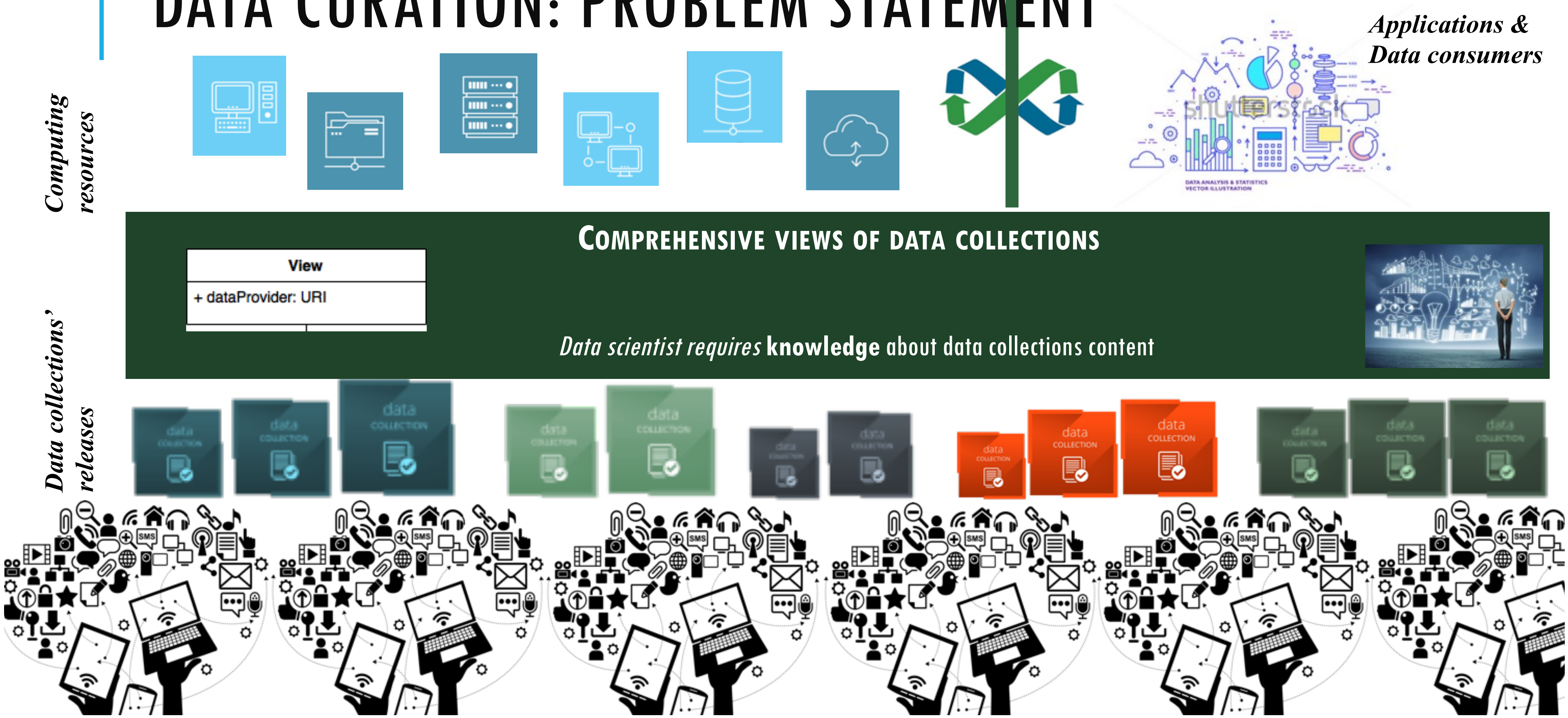
RAW DATA:
heterogeneous (*variety*), huge (*volume*), incomplete, unprecise, missing, contradictory (*veracity*), continuous releases produced at different rates (*velocity*), proprietary, critical, private (*value*)

Data collections' releases

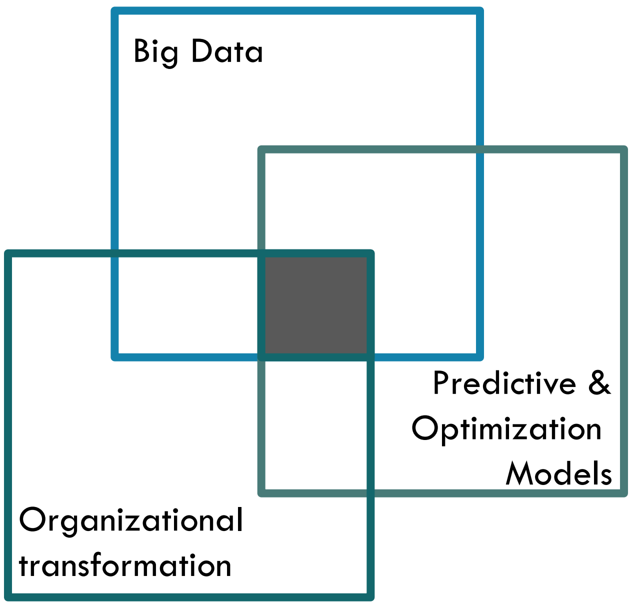
DATA CURATION: PROBLEM STATEMENT



DATA CURATION: PROBLEM STATEMENT



CAPTURING VALUE FROM ADVANCED ANALYTICS



Based on three guiding principles

Decision backwards

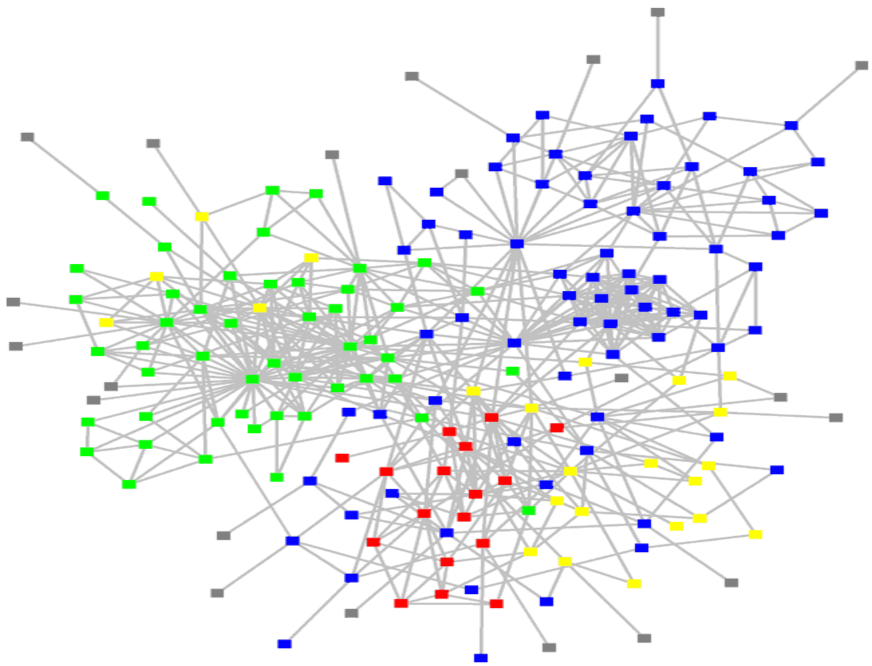
Step by step

Test and learn

The “Social Graph” behind Facebook

Keith Shepherd's "Sunday Best”. http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

STRUCTURE OF AN ORGANIZATION

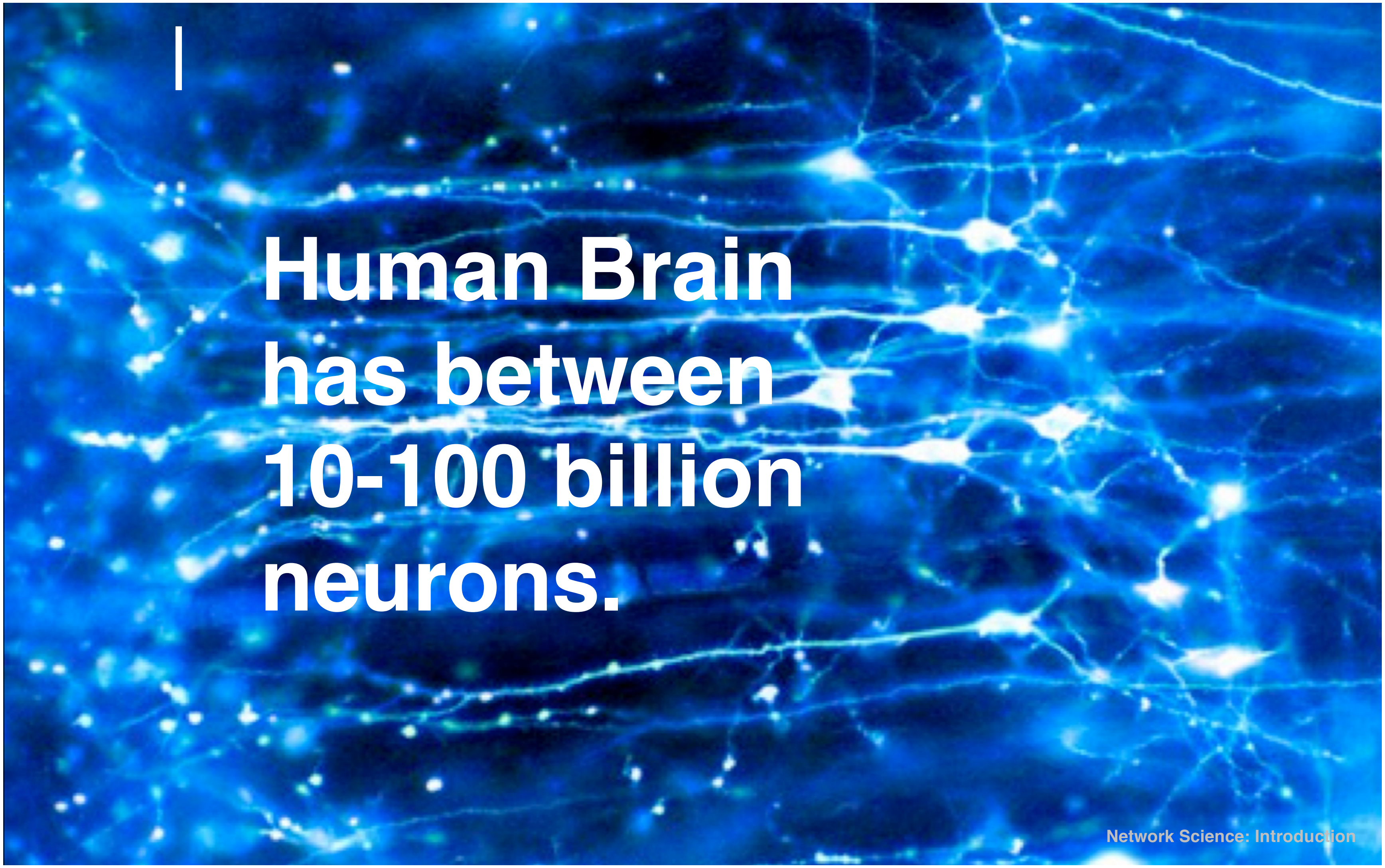


www.orgnet.com

: departments

: consultants

: external experts


Human Brain
has between
10-100 billion
neurons.
Network Science: Introduction

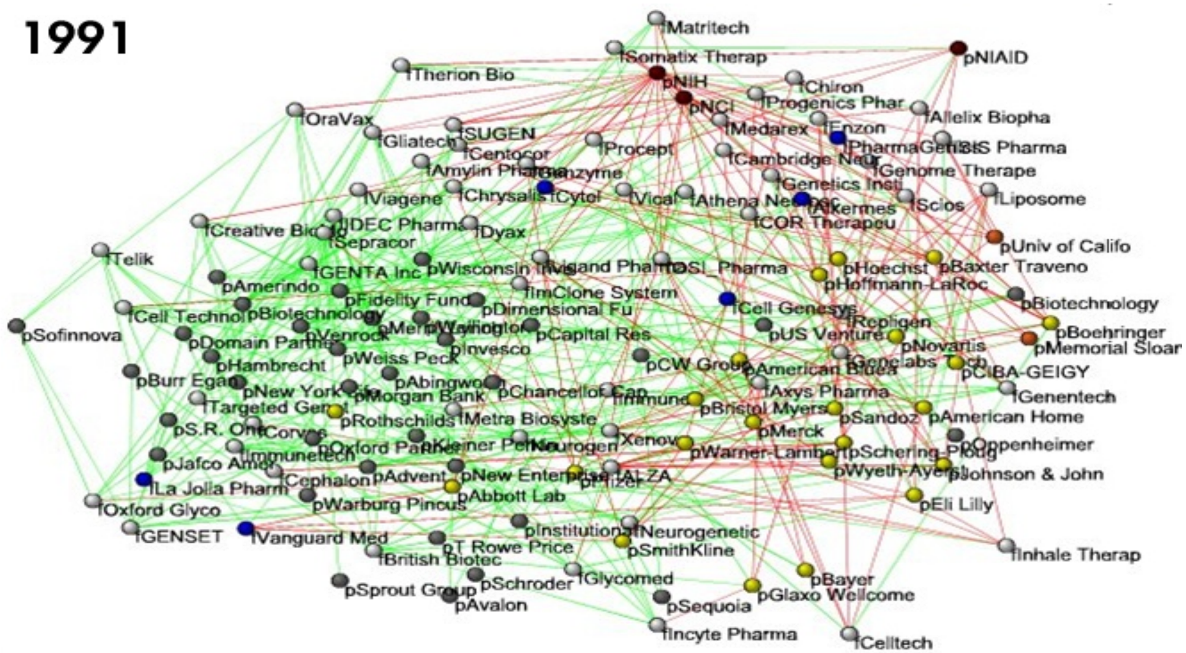
BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

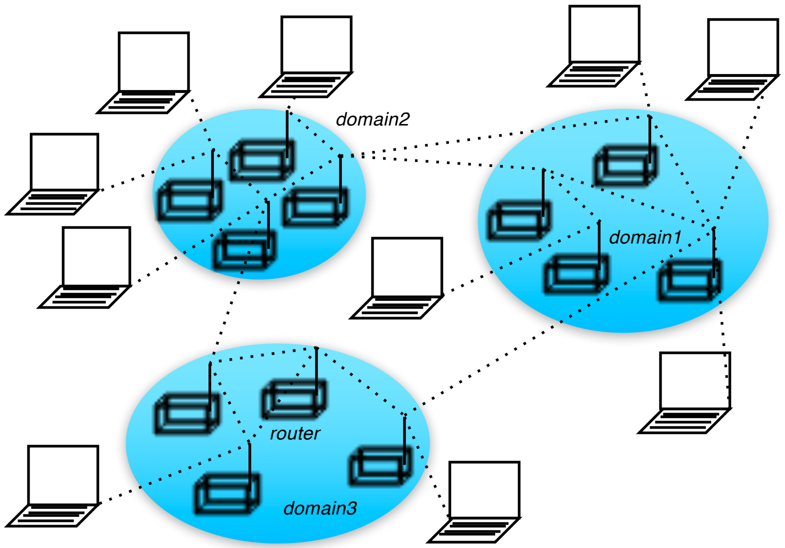
Links:

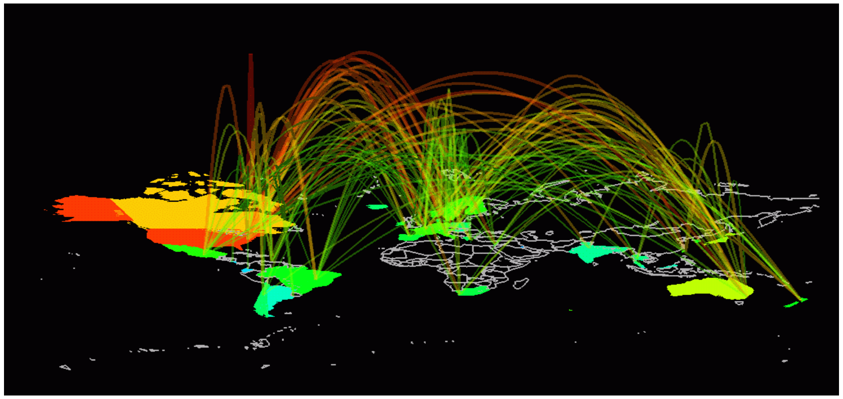
- Collaborations
- Financial
- R&D

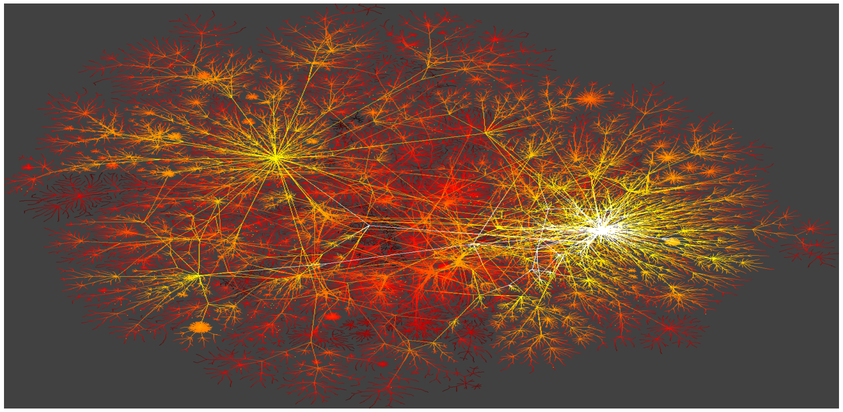


http://ecclectic.ss.uci.edu/~drwhite/Movie

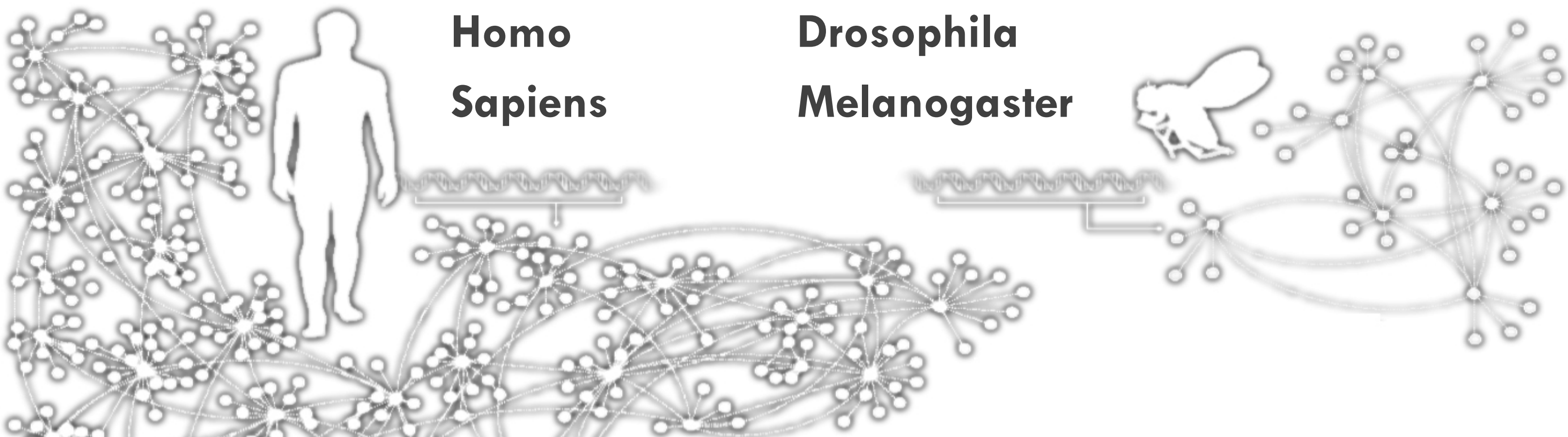
INTERNET

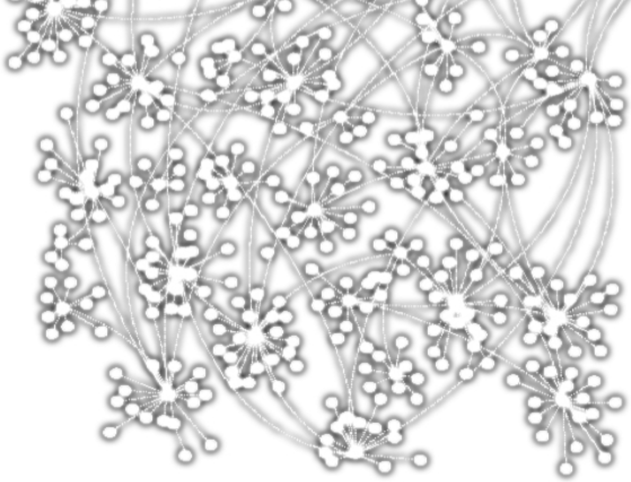






HUMAN GENES



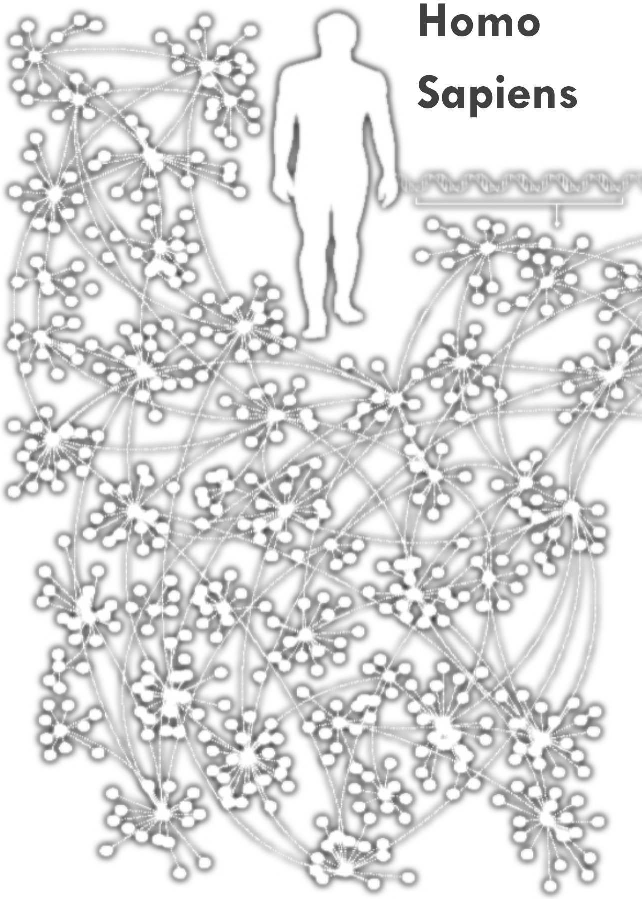


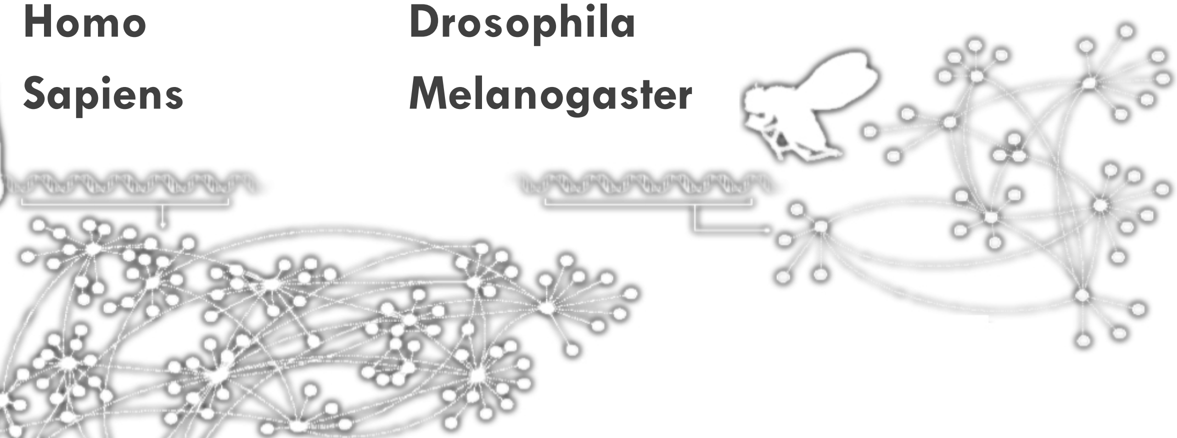
In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them.

HUMAN GENES

Homo Sapiens

Drosophila Melanogaster





Complex systems

Made of many non-identical **elements** connected by diverse **interactions.**

NETWORK

ECONOMIC IMPACT

Google

Market Cap(2010 Jan 1):

$189 billion

Cisco Systems

networking gear Market cap

(Jan 1, 2919):

$112 billion

Facebook

market cap:

$50 billion

www.bizjournals.com/austin/news/2010/11/15/facebooks... - Cached

Data was not
stored

Beginning of the use of BDs
& basic reports

Great variety of visual
resources to analyse data

DATA CONTAINS VALUE & KNOWLEDGE

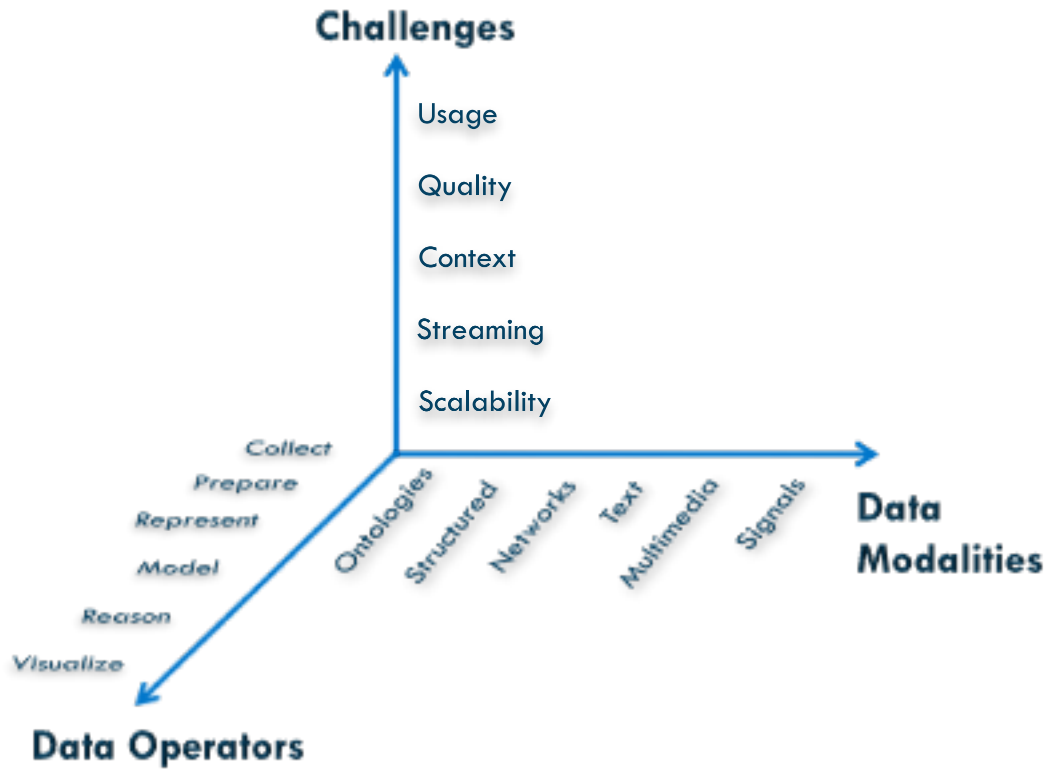
KNOWLEDGE EXTRACTION

Data needs to be

- Stored ← this class
- Managed
- ANALYZED ← this class

Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

WHAT MATTERS WHEN DEALING WITH DATA?



DATA MINING: CULTURES

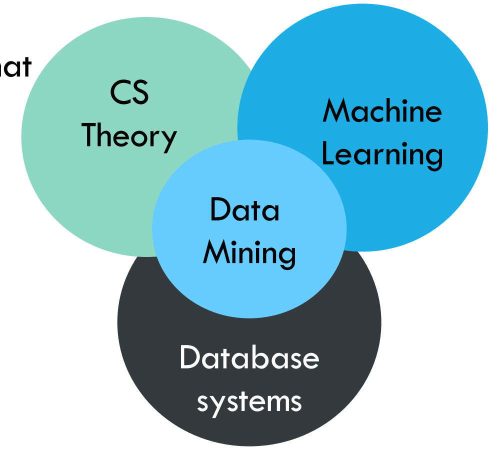
Data mining overlaps with:

- **Databases:** Large-scale data, simple queries
- **Machine learning:** Small data, Complex models
- **CS Theory:** (Randomized) Algorithms

Different cultures:

- To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
  - Result is the query answer
- To a ML person, data-mining is the **inference of models**
  - Result is the parameters of the model

In this class we will do both!



DATA MINING TASKS

Descriptive methods

- Find human-interpretable patterns that describe the data
  - **Example:** Clustering

Predictive methods

- Use some variables to predict unknown or future values of other variables
  - **Example:** Recommender systems

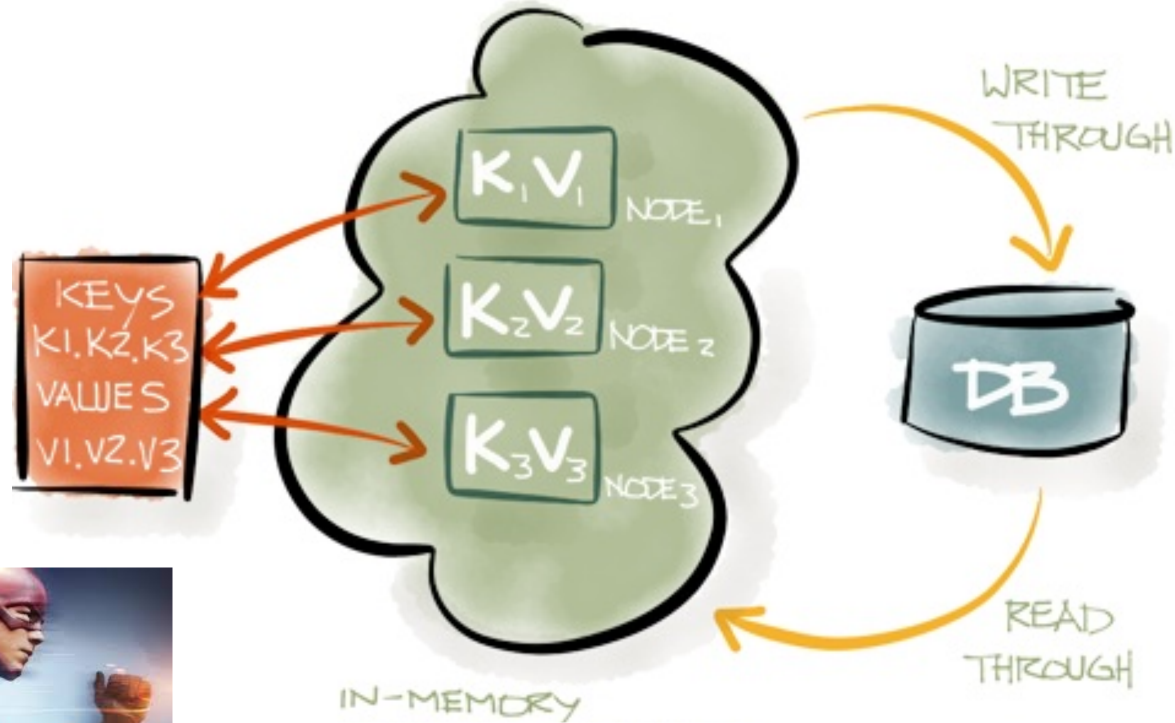
HOW IT ALL FITS TOGETHER

High dim. data	Graph data	Infinite data	Machine learning	Apps
Locality sensitive hashing	PageRank, SimRank	Filtering data streams	SVM	Recommender systems
Clustering	Community Detection	Web advertising	Decision Trees	Association Rules
Dimensionality reduction	Spam Detection	Queries on streams	Perceptron, kNN	Duplicate document detection

Data management guided by the RUM conjecture

(Read, Update, Memory (or storage) overhead)

DEALING WITH DATA FOR DATA SCIENCE TASKS



DEALING WITH DATA FOR DATA SCIENCE TASKS

Persistence

Crystal
DNA

Magnetic
Solid state

Paper/card

Permanence

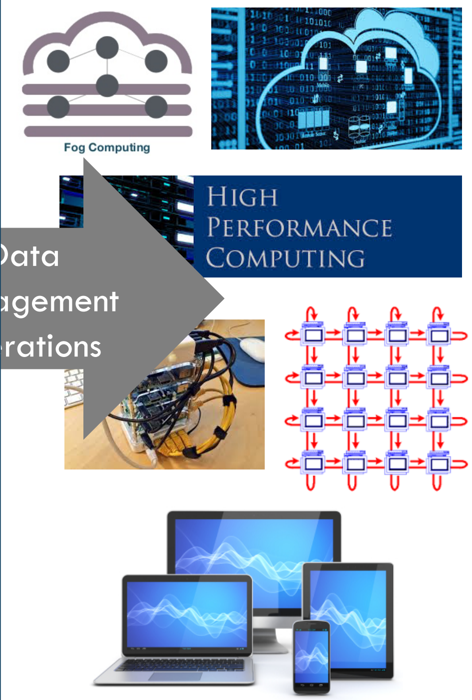
For how long will data survive time?

Functional architecture

The next generation of data management systems

Data management operations

Deployment architecture



CHALLENGES AND OBJECTIVE

How to combine, deploy, and deliver DBMS functionalities:

- **Compliant** to application/user requirements
- **Optimizing** the consumption of computing resources in the presence of **greedy** data processing tasks
- Delivered according to **Service Level Agreement (SLA)** contracts
- Deployed in **elastic** and distributed **platforms**

Final remarks & Lecture program

FINAL REMARKS

Data collections

- **New scales:** bronto scale due to emerging IoT
- **New types**: thick, long hot, cold
- New **quality measures**: QoS, QoE, SLA

Data **processing** & **analytics**

- Complex jobs, stream analytics are still open issues
- Economic cost model & business models (Big Data value & pay-as-U-go)

CONTENT

Big Data Analytics Trends

- Big data and beyond the mirror
- Big Data analytics, Data mining, Data science
- Cooking data: the big picture

Data management at scale: all you need for cooking data

- High performance execution environments
- Data as service tools: distributed storage, data access API, more complex data processing, declarative languages
- New data analytics stacks

Modeling & Predictive analytics

- Clustering at different scales
- Network science
- Graph analytics

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

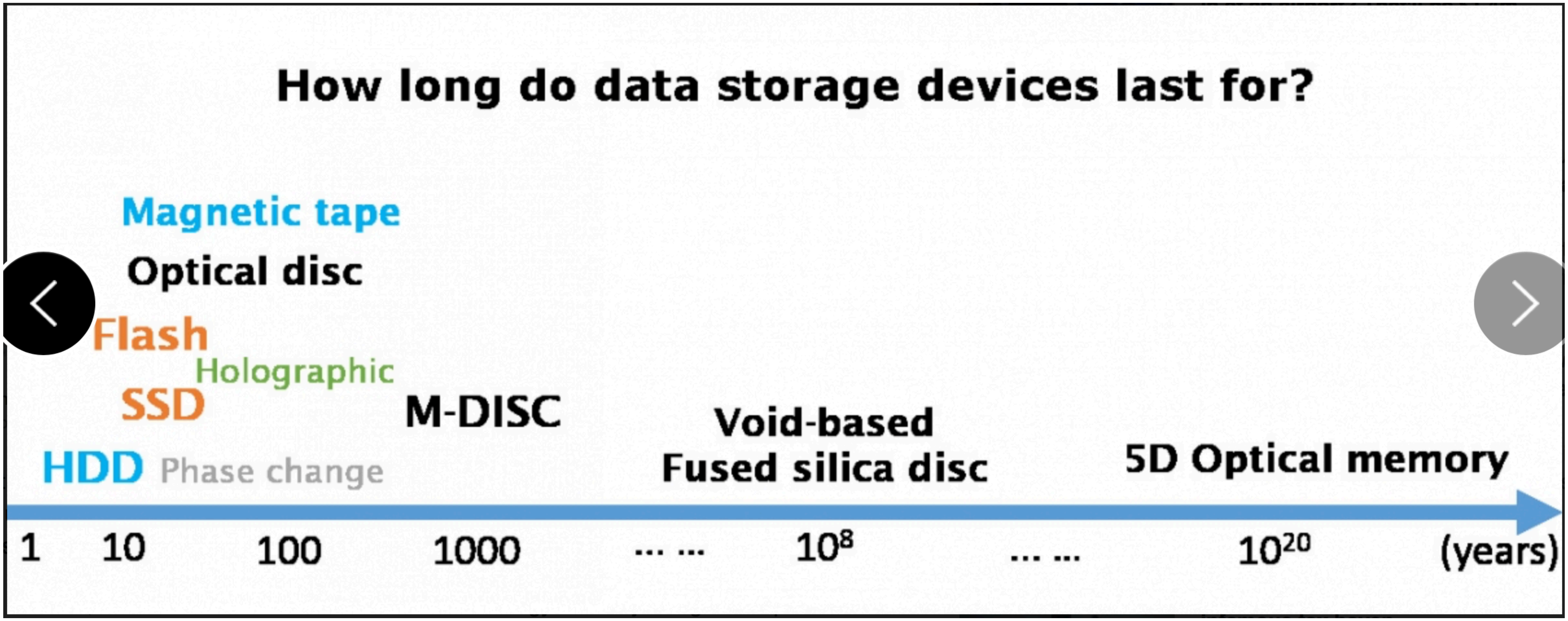
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Genoveva Vargas-Solar

CR1, CNRS, LIG-LAFMIA

Genoveva.Vargas@imag.fr

http://vargas-solar.com/big-linked-data-keystone/


How long do data storage devices last for?
Magnetic tape
Optical disc
Flash
Holographic
SSD
M-DISC
Void-based
Fused silica disc
HDD
Phase change
5D Optical memory
1
10
100
1000
... ...
10^8
... ...
10^20
(years)