Big Data in Health

ARRIEL BENIS, PHD, HOLON INSTITUTE OF TECHNOLOGY, ISRAEL &

Prof. Stefan Darmoni, MD, PhD, Rouen University hospital & LIMICS INSERM U1142, Sorbonne Université, Paris, France

Introduction to Big Data, Big Data in Healthcare, and NoSQL

From Data to Big Data and Artificial Intelligence

- **Every minute of the day…**
- **A Revolution in Data Availability**
- **Data for… Anything!**
- **Data, Big Data, Artificial Intelligence… From Fiction to Reality !**

Data from a “Business” Perspective

- What’s a Business? A “Organization” and more…
- Paradigm shift Data as a Critical Organizational Resource
- From Data to Wisdom… or the Big Data Holy Grail - The DIKW model
- The DIKW model as a Business Intelligence Environment and Data Science

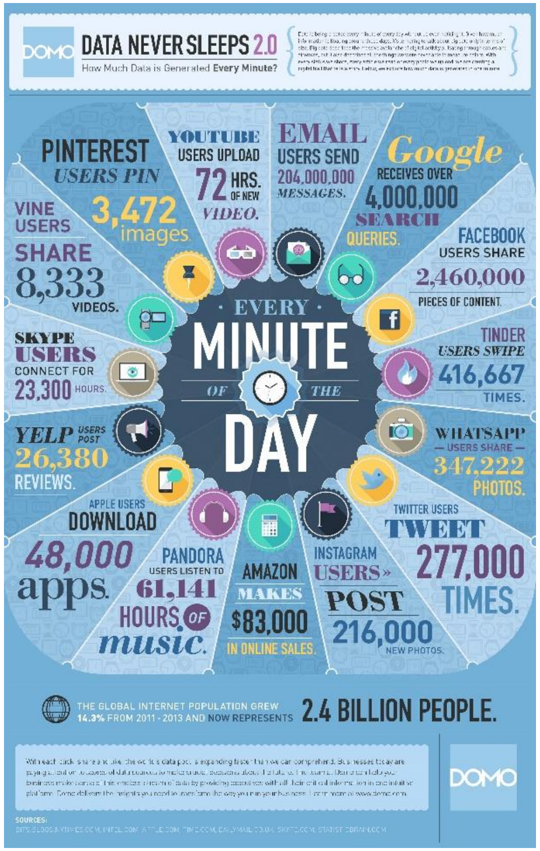
Big Data, definitions

- Types of Data / Big Data
- From the 3Vs to 10 Vs
- The Big Data Ecosystem is rich
- NoSQL

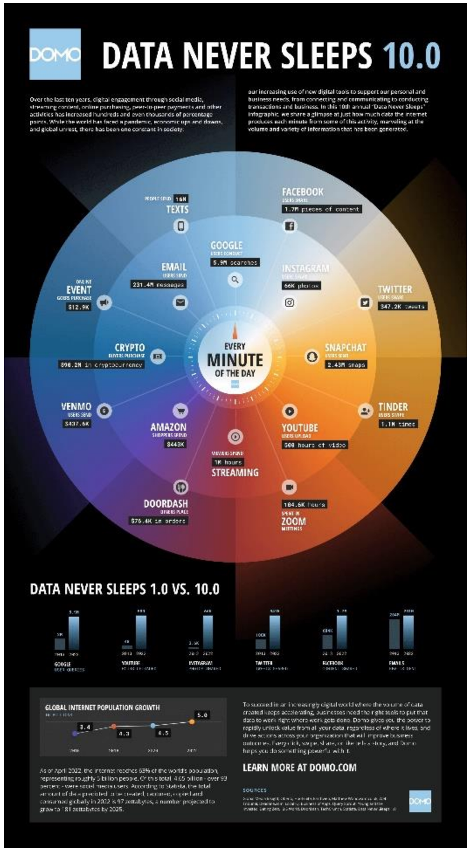
Big Data in Medicine

- Big Data and medical research
- Available Biobanks increasing
- Multi-sources for Multi-objectives
- Big Data and Machine Learning
- Big Health Data a competitive business

Every minute of the day…
A simple view for a BIG problem



2013



2022

From Data to Big Data and Artificial Intelligence

Who? Where? When? Why? What? How? How much?

Large-scale (BIG) Data is everywhere… and we need understand them!!!

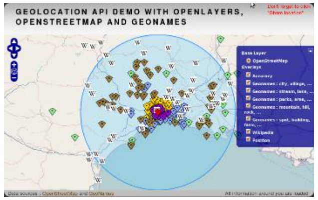
- **DATA DELUGE** (מבול) or **DATA TSUNAMI**
  Enormous data growth
  due to advances in data generation and collection technologies
- THE MANTRA
  GATHER WHATEVER DATA YOU CAN WHENEVER AND WHEREVER POSSIBLE.
- Expectations
  DATA will have **VALUE** either for the **PURPOSE COLLECTED** or for a **PURPOSE NOT ENVISIONED** *(the "surprising effect")*.

Cybersecurity

Search Engines
Social Media
Social Networks

eBusiness

Sensors



Geolocation

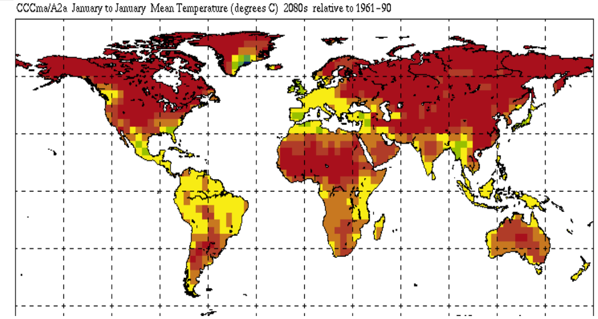
Health

From Data to Big Data and Artificial Intelligence

Who? Where? When? Why? What? How? How much?

Data are a Great opportunities to solve society's major problems

Improving **health care** and reducing costs



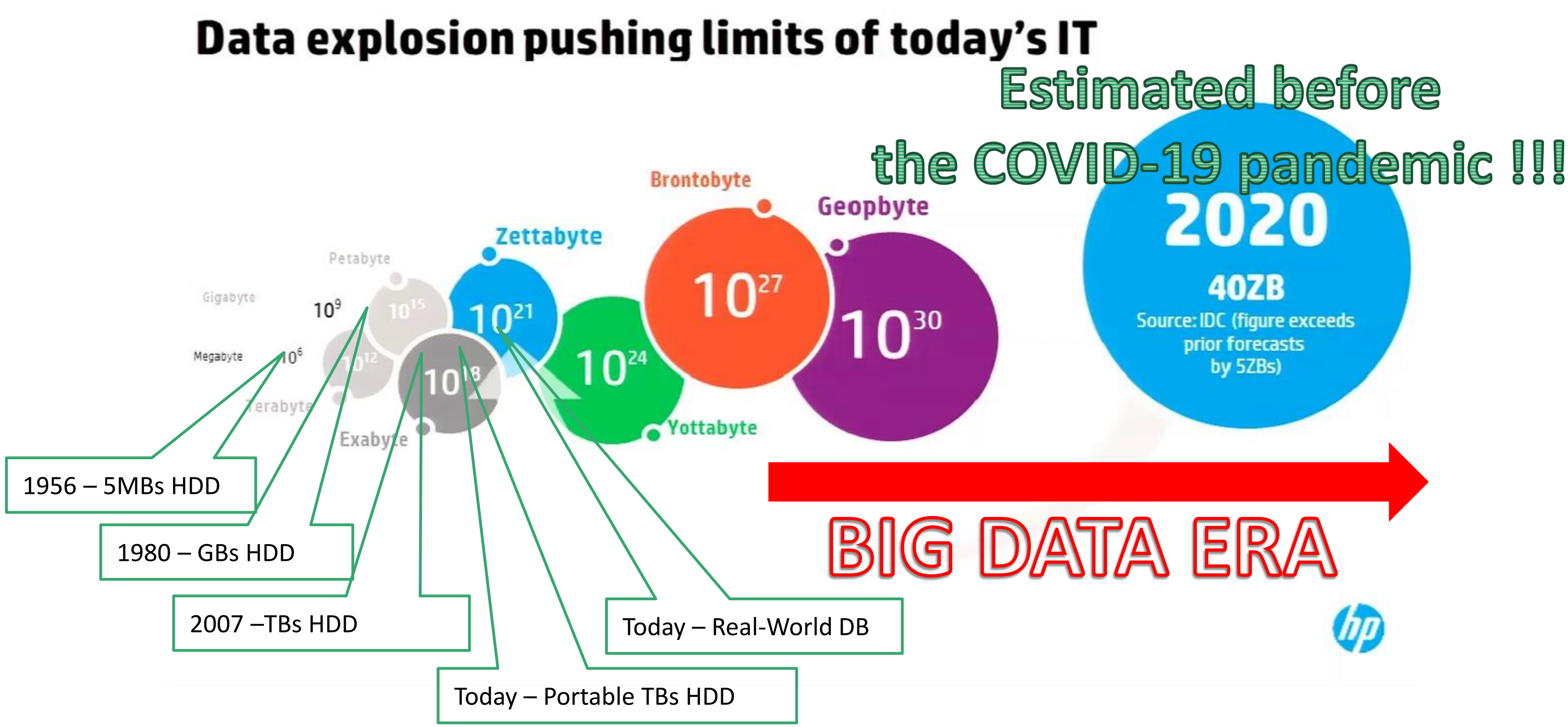
Predicting the impact of **climate change**

Finding alternative/ **green energy** sources

Reducing **hunger and poverty** by increasing **agriculture production**

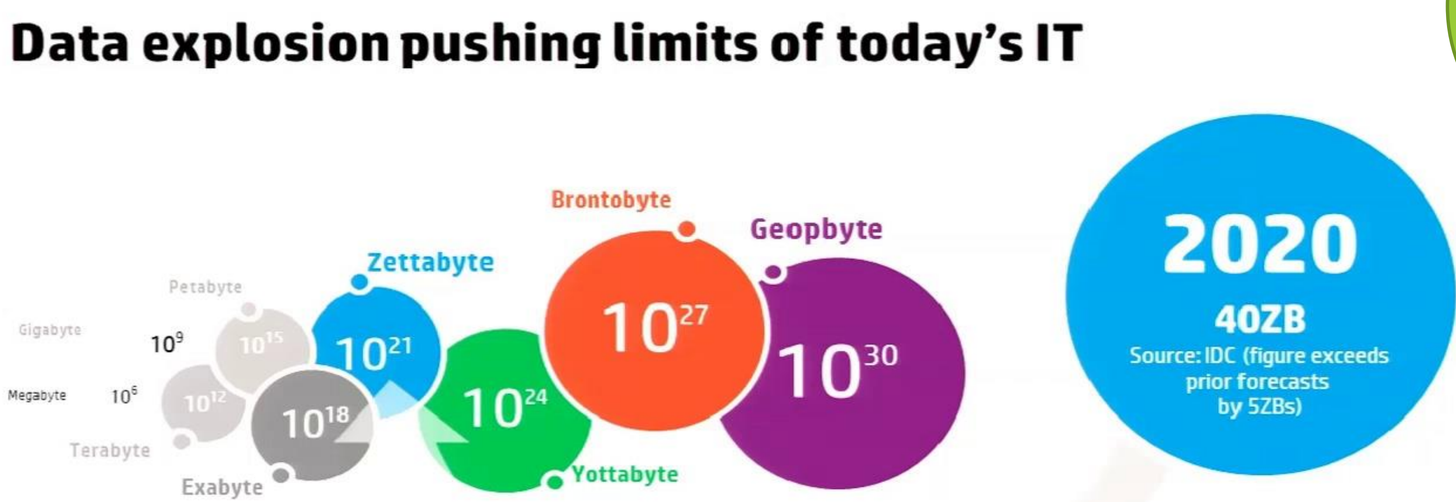
From Data to Big Data and Artificial Intelligence

The “Real-World” facing to a Revolution in Data Availability



https://wwwyoutube.com/watch?v=2D8oji5EKbM

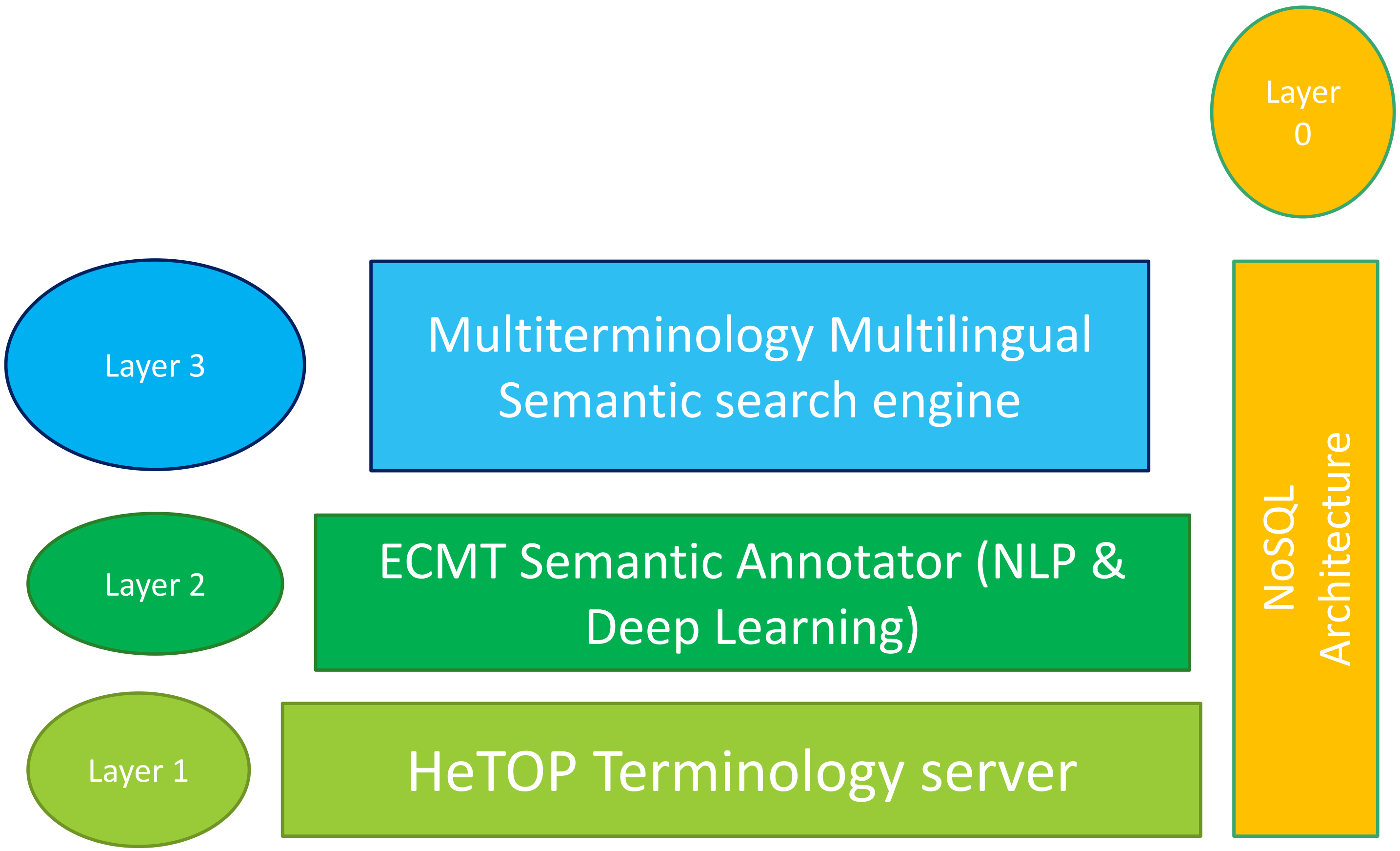
The “Real-World” facing to a Revolution in Data Availability



Googol
10^{100}

Number of atoms in the universe
10^{78-86}

Semantic Clinical Data Warehouse (CDW) in Rouen University Hospital, France



HeTOP Terminology server

Layer 1

- URL: www.hetop.eu
- Multi Terminology & cross lingual → matrix navigation (among languages & among terminologies
- 100 termino-ontologies included in 55 languages
- 2 M different concepts in English; 0.7 M in French
  - ≈ 165 K concepts in French in UMLS (2018AB) vs. ≈ 630 K in HeTOP (x3.6); the most advance terminology server in France >> tool of the French (National) Digital Health Agency; relations are tricky to manage
- **Over 100 million RDF triplets (2014) → big data +++**

Order of magnitude = 10^8

ECMT Semantic Annotator (NLP & Deep Learning)

Layer 2

- Based on HeTOP; 50 chosen KOS out of 75 (no interface terminologies)
- $21.4 * 10^6$ health documents
- Processing time: 30 hours (two servers 1 To; one with 196 cores and the second with 144 cores => computer sobriety
- **$5.2 * 10^9$ medical concepts extracted; $2.6 * 10^9$ after filtering**

Real example of big data in health (RWD: real world data)

Siefridt C, Grosjean J, Lefebvre T, Rollin L, Darmoni S, Schuers M. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. Int J Med Inform. 2020 Jan;133:104009. doi: 10.1016/j.ijmedinf.2019.104009. Epub 2019 Nov 1.

Neveol & coll. Clinical Natural Language Processing in languages other than English: opportunities and challenges. Journal of Biomedical Semantics 2018 9:12

Data for… Anything!
The World Economic Forum viewpoint

https://www.youtube.com/watch?v=eVSfJhssXUA

From Data to Big Data and Artificial Intelligence

Tang Yu, An AI-Powered Robot, Named CEO Of A Chinese Company

Tang Yu will handle the organisational and operational aspects for the company, which is worth nearly $10 billion.

World News | Edited by Nikhil Pandey | Updated: September 08, 2022 1:49 pm IST

An AI-powered robot is the new CEO of a Chinese company. (Representational Photo)

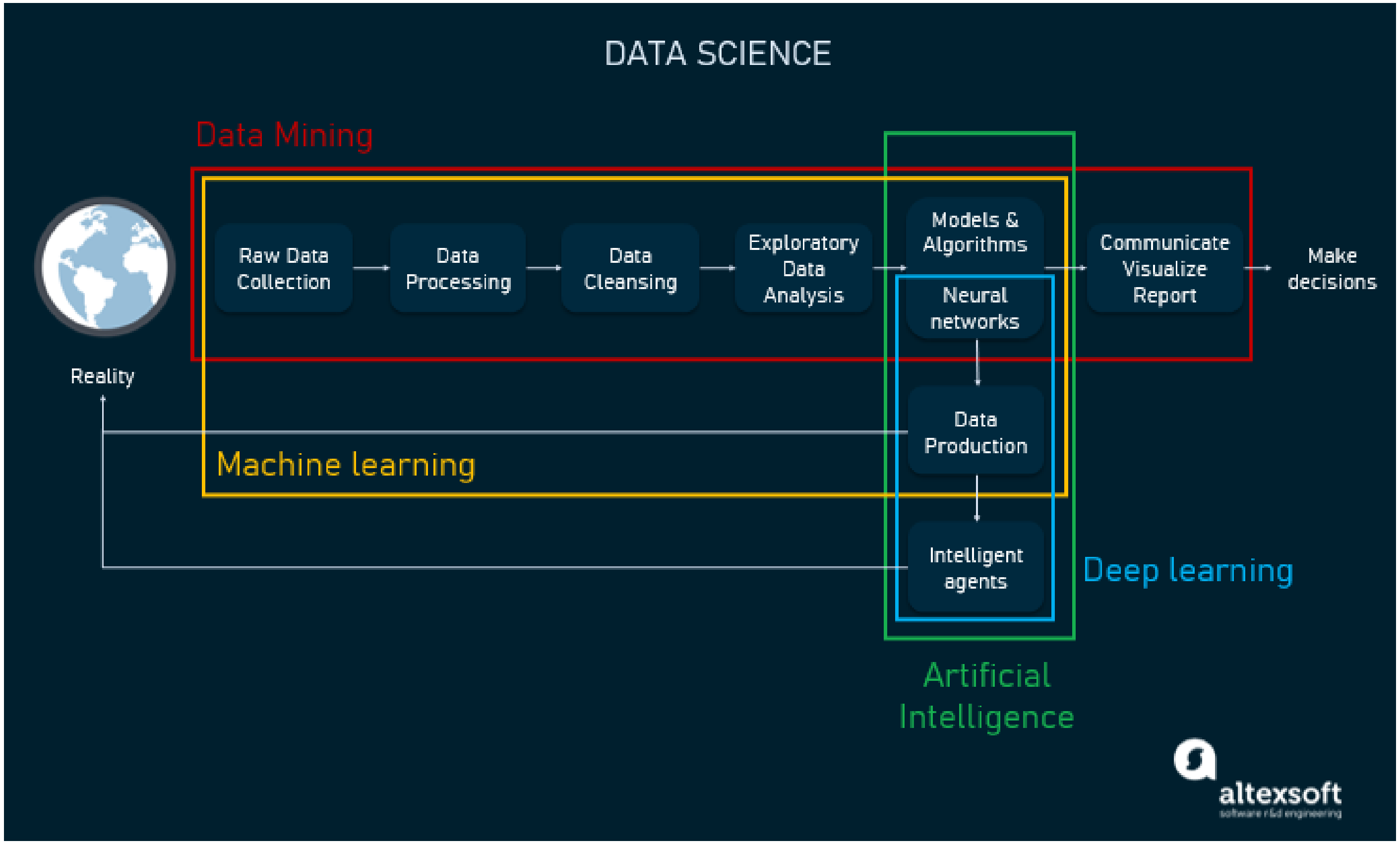
In several science fiction movies, robots are seen ruling the planet and taking humans as their slaves. But many people don't take the claims made in these movies too seriously, and their forecasts are wrong many times. However, a recent move by a Chinese company has shocked the business world as well as social media. The metaverse firm has chosen an AI-powered virtual humanoid robot as its chief executive officer (CEO). The official announcement was made last week.

According to a report in UK-based *The Metro*, Tang Yu, the humanoid robot, will be leading the operations at China's NetDragon Websoft, making her the first

2022 !

Not 2026….

From Reality-to-Reality AI *via Fiction*
From Data to AI



Introduction to Big Data, Big Data in Healthcare, and NoSQL

From Data to Big Data and Artificial Intelligence
- Every minute of the day…
- A Revolution in Data Availability
- Data for… Anything!
- Data, Big Data, Artificial Intelligence… From Fiction to Reality !

Data from a "Business" Perspective
- **What's a Business? A "Organization" and more…**
- **Paradigm shift Data as a Critical Organizational Resource**
- **From Data to Wisdom… or the Big Data Holy Grail - The DIKW model**
- **The DIKW model as a Business Intelligence Environment and Data Science**

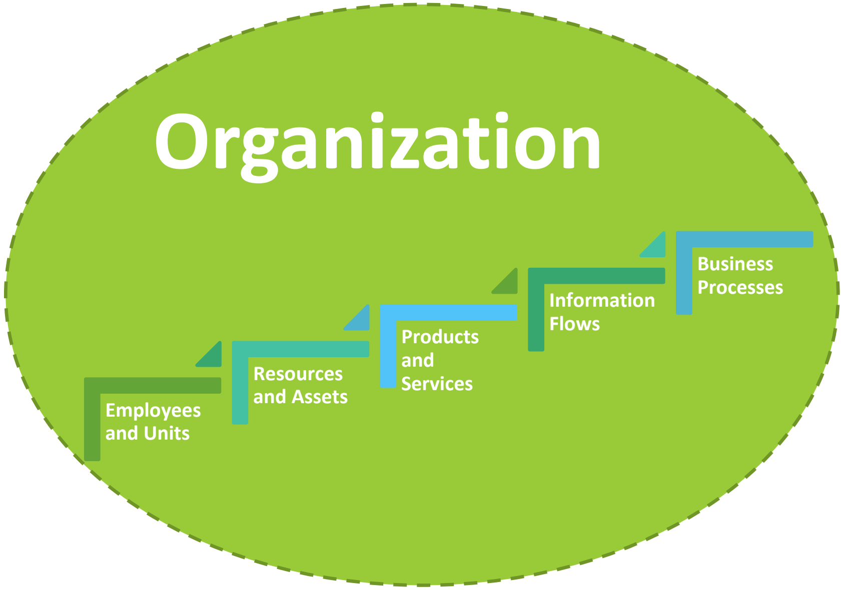
Big Data, definitions
- Types of Data / Big Data
- From the 3Vs to 10 Vs
- The Big Data Ecosystem is rich
- NoSQL

Big Data in Medicine
- Big Data and medical research
- Available Biobanks increasing
- Multi-sources for Multi-objectives
- Big Data and Machine Learning
- Big Health Data a competitive business

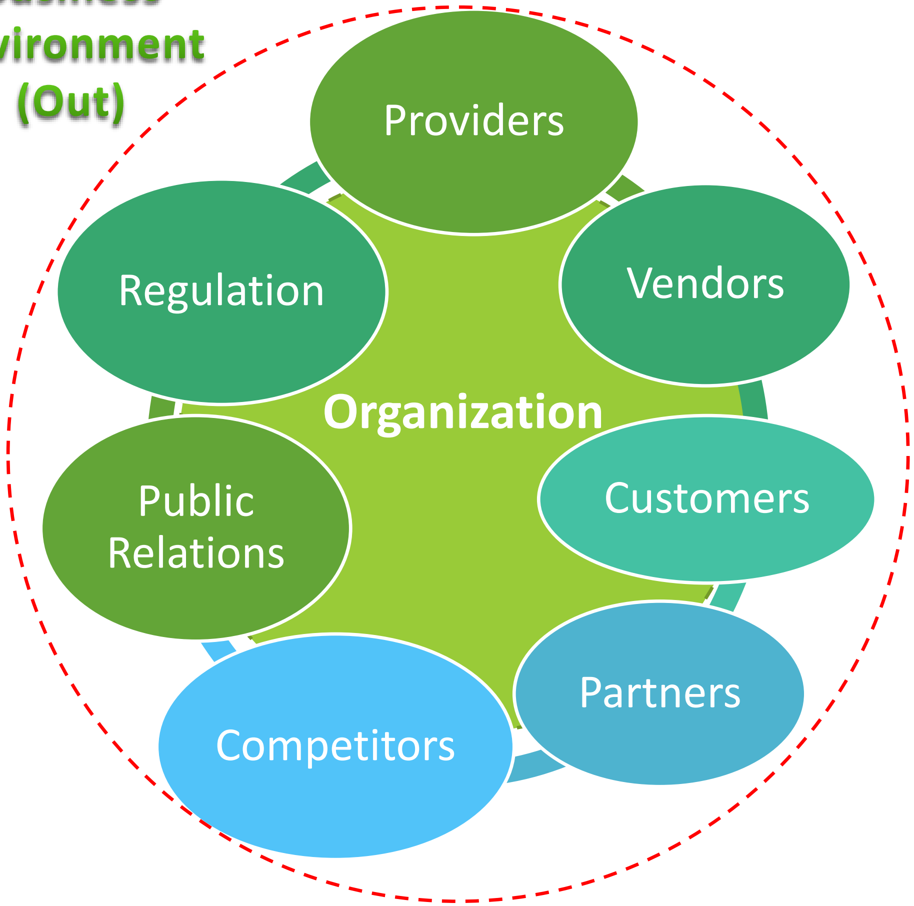
Data from a “Business” Perspective

What’s a Business? A “Organization” and more…

STRATEGIC OBJECTIVES / MISSIONS DESIRED OUTCOMES



Business Environment (Out)



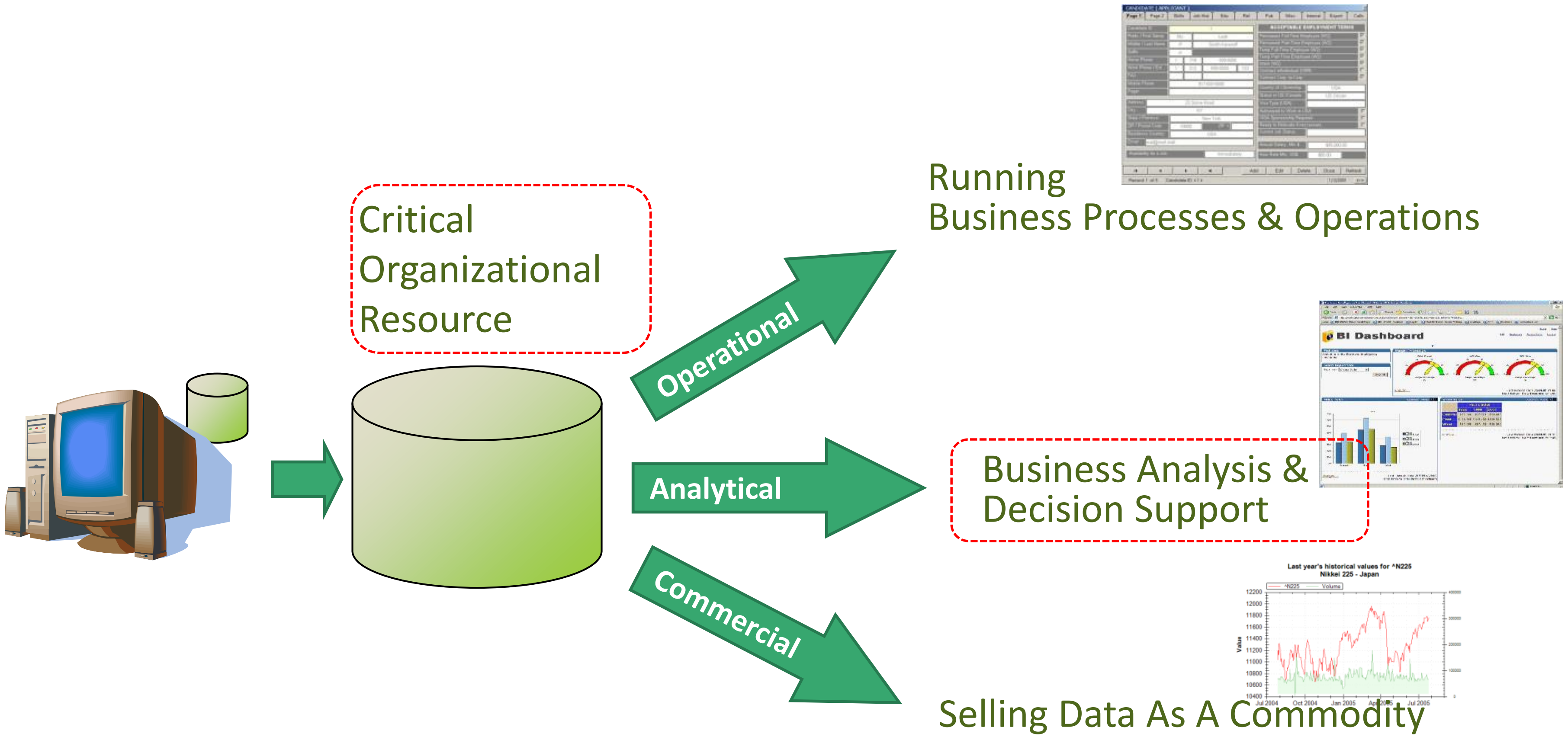
What’s a Business?
A “Organization” and more… (2)

1. An **organized group** of **objects** *(humans, machines, etc.)* **sharing** a set of **desired outcomes**
   - Private or Public sector business
   - Government agency
   - Non-Governmental Organization (NGO) and Non-Profit Organization
   - Community (e.g., Online, Offline)
2. A **goal-driven entity** with certain value **measures for success**
3. An **organization aims at forming** the **best fit between the business environment and the organization’s mission, objective and desired outcomes**
   - Fit is achieved
     - by managing organizational **resources**
     - through an ongoing process of **information exchange**

Data from a "Business" Perspective

Paradigm shift

Data as a Critical Organizational Resource

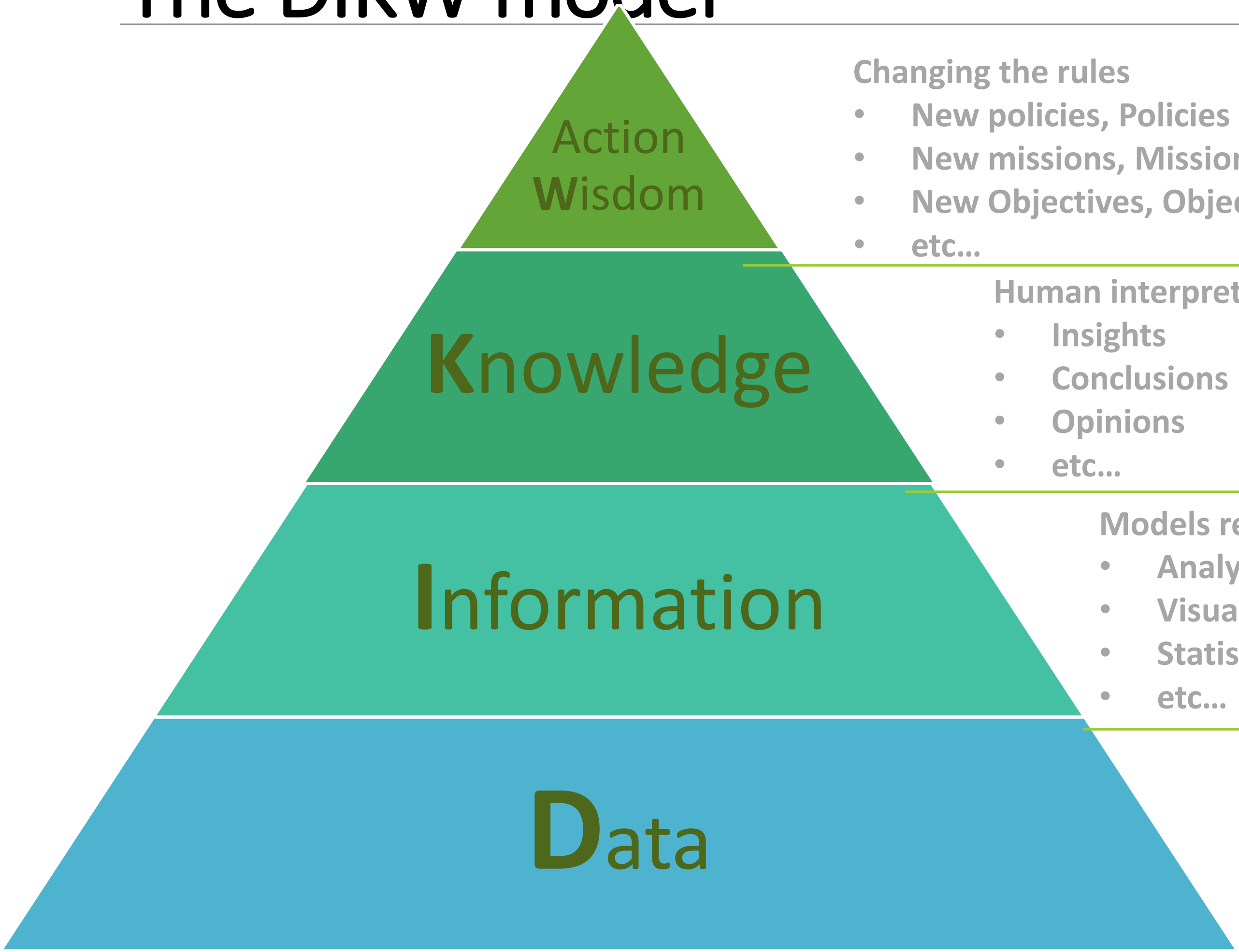


DIFFERENT FORMS OF DATA USE

Data from a "Business" Perspective

From Data to Wisdom… or the Big Data Holy Grail

The DIKW model



Changing the rules
- New policies, Policies updates
- New missions, Missions updates
- New Objectives, Objectives updates
- etc…

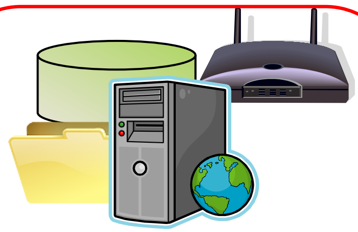
Human interpretation
- Insights
- Conclusions
- Opinions
- etc…

Models reflecting real-world behavior
- Analytical models
- Visualization
- Statistical analysis
- etc…

Data Resources
- Databases
- Tables
- Records
- etc.

Data from a “Business” Perspective

The DIKW model as a **Business Intelligence Environment**



- Databases
- Files
- Data feeds
- Internet
- etc.

Source
BIG DATA

Extract, Transform, Load

ETL

Acquisition

Data WareHouse

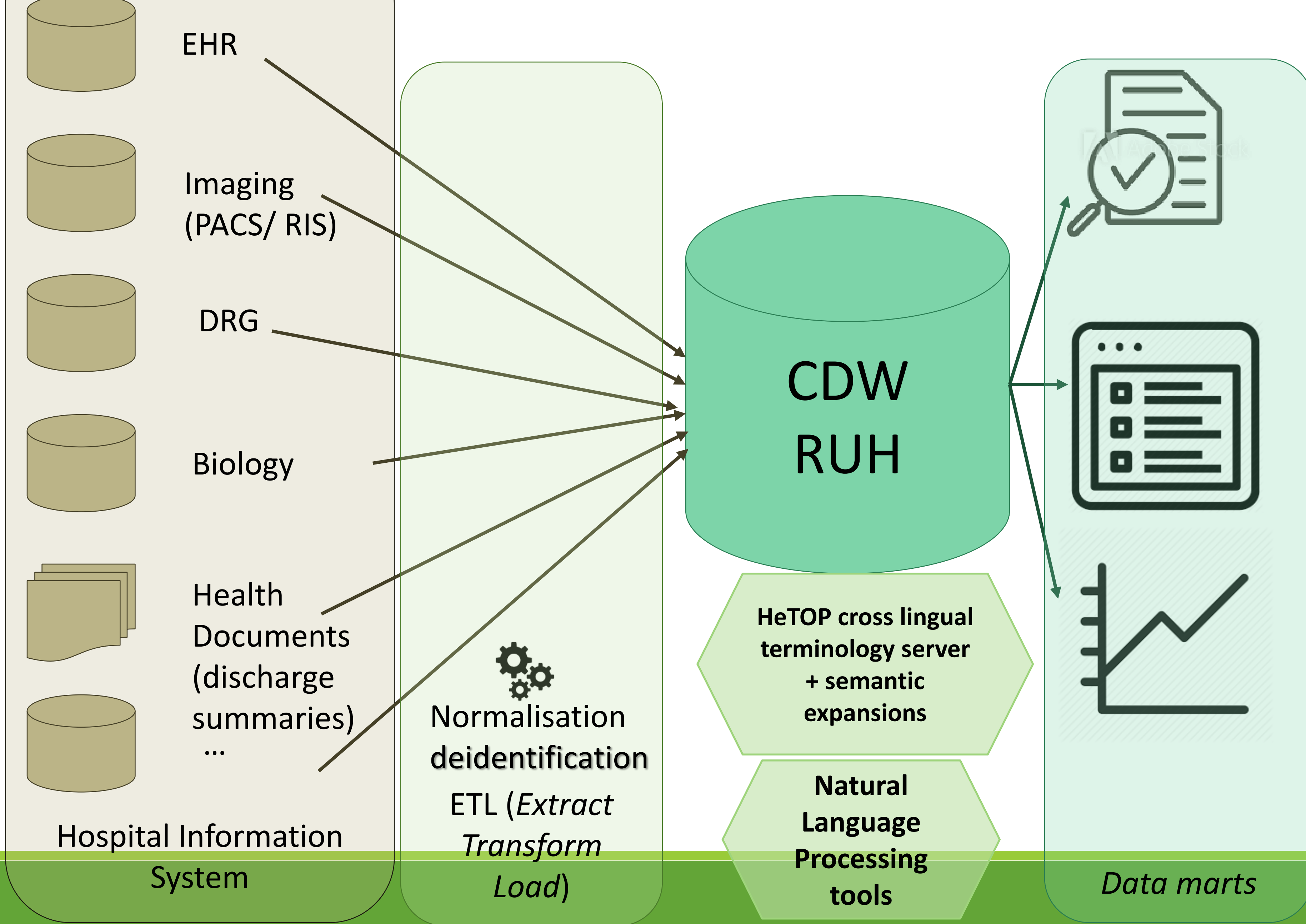
DWH

Infrastructure
BIG DATA

- Reporting
- OLAP
- Dashboards
- Data Mining
- Etc.

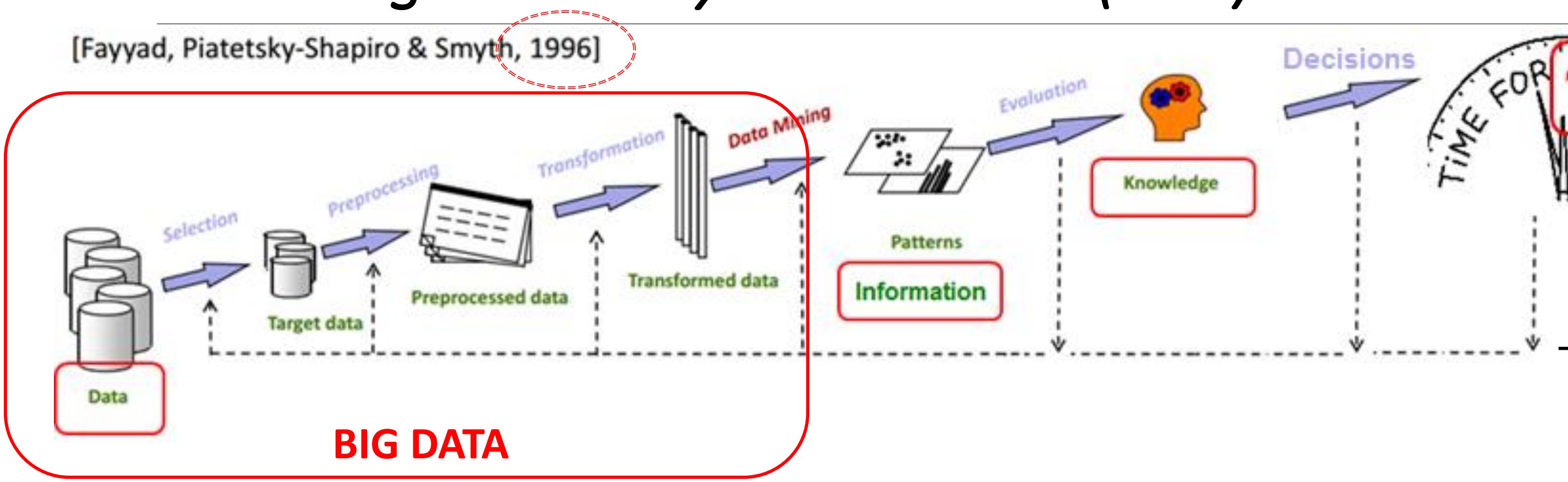
Applications

Model of CDW in RUH



DIKW/Big Data from a **Data Science perspective**

Knowledge Discovery in Databases (KDD)



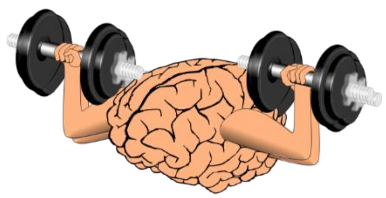
Continuous improvements, changes
Elasticity of time between each step !
Added value at each step !

In science we call it ***"KDD"*** *Knowledge Discovery in Databases*	*In Business we call it* ***"BI", Analytics***
In the past we called it ***"Data", "Databases"***	*Today, we call it* ***"Big Data"***

Exercice

Business… but all we discussed about is relevant for the “Healthcare and Medicine” field.

5W2H* ?



* Who, What, When, Where, Why, How, How much

Introduction to Big Data, Big Data in Healthcare, and NoSQL

From Data to Big Data and Artificial Intelligence
- Every minute of the day…
- A Revolution in Data Availability
- Data for… Anything!
- Data, Big Data, Artificial Intelligence… From Fiction to Reality !

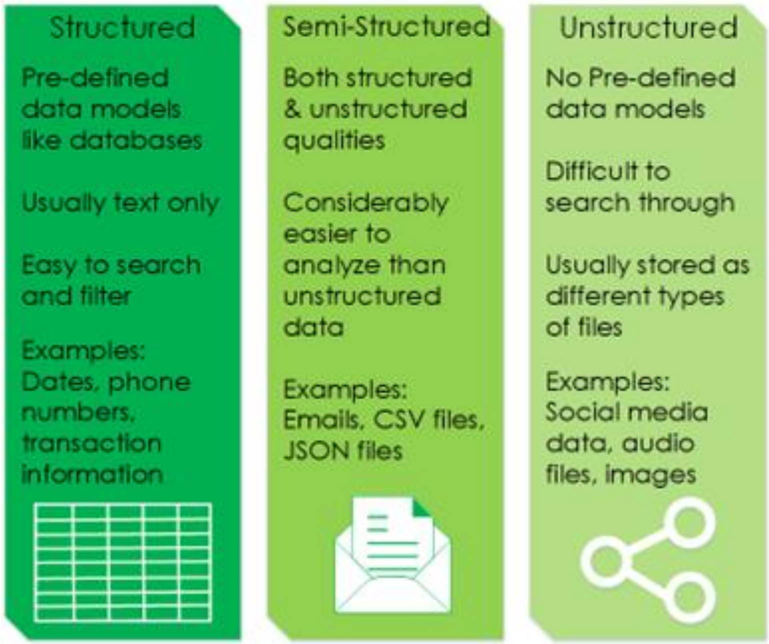
Data from a "Business" Perspective
- What's a Business? A "Organization" and more…
- Paradigm shift Data as a Critical Organizational Resource
- From Data to Wisdom… or the Big Data Holy Grail - The DIKW model
- The DIKW model as a Business Intelligence Environment and Data Science

Big Data, definitions
- **Types of Data / Big Data**
- **From the 3Vs to 10 Vs**
- **The Big Data Ecosystem is rich**
- **NoSQL**

Big Data in Medicine
- Big Data and medical research
- Available Biobanks increasing
- Multi-sources for Multi-objectives
- Big Data and Machine Learning
- Big Health Data a competitive business

Big Data, definitions

Types of Big Data



https://mdaca.io/2021/05/whats-the-big-data/

Structured:
Biology, DRGs data, CPOE data (prescription)

Semi-Structured:

Unstructured:
Text from health documents

Quality of data

1. Quality of data is varying according to the type of data
   - Structured
   - Semi-structured
   - Non structured (80% of the data in France; a little bit less in AngloSaxon countries)
2. The better quality lays in structured data
   - BUT even structured data may generate errors
   - In biology, some errors exists from automats (e.g. kaliemia K+ in blood); supervision by a human biologist, as kaelimia is of utmost importance for the patient +++
3. Least quality for unstructured data
   - Need of a semantic annotator to extract medical concepts
   - Precision (false positive) / Recall (false negative) is varying according to the context

Criteria to measure the quality of a documentary system (information science)

	Relevant	Non Relevant	
Transmitted Documents	A	B	A+B
Non Transmitted Documents	C	D	C+D
	A+C	B+D	

Recall = A/A+C ; Silence = 1-Rappel = C/A+C = false negatives

Precision = A/A+B ; Noise = 1 – Précision = B/A+B = fals positives

Recall = sensitivity (biostatistics)

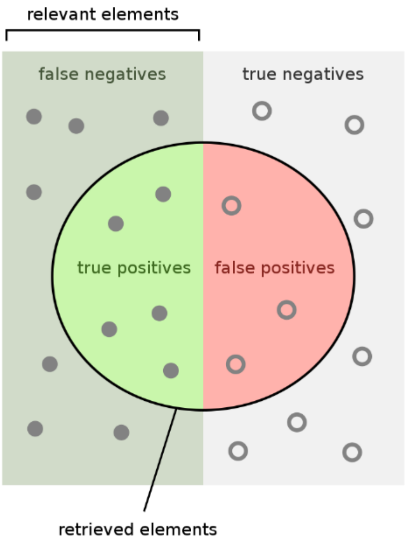
Precision = positive predictive value (PPV) (biostatistics)

Criteria to measure the quality of a documentary system (information science)

Recall = A/A+C ; Silence = 1-Rappel = C/A+C = false negatives
Precision = A/A+B ; Noise = 1 – Précision = B/A+B = fals positives
Recall = sensitivity (biostatistics)
Precision = positive predictive value (PPV) (biostatistics)

F-measure or F-score is the harmonic mean of precision & recall

F-measure = 2 (R * P) / (R + P)



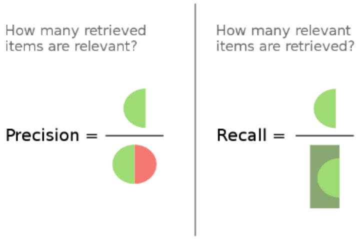


Table 7. System performance for ICD10 coding on the **French raw** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

	ALL				**EXTERNAL**			
	Team	P	R	F	**Team**	P	R	F
Official runs	SIBM-run1	**.857**	**.689**	**.764**	SIBM-run1	**.567**	**.431**	**.490**
	LITL-run2	.666	.414	.510	LIRMM-run1	.443	.367	.401
	LIRMM-run1	.541	.480	.509	LIRMM-run2	.443	.367	.401
	LIRMM-run2	.540	.480	.508	LITL-run2	.560	.283	.376
	LITL-run1	.651	.404	.499	LITL-run1	.538	.277	.365
	TUC-MI-run2	.044	.026	.033	TUC-MI-run2	.010	.004	.005
	TUC-MI-run1	.025	.015	.019	TUC-MI-run1	.006	.005	.005
	average	.475	.358	.406	**average**	.367	.247	.292
	median	.541	.414	.508	**median**	.443	.283	.376
Non-official	LIMSI-run2	.872	.784	.825	LIMSI-run2	.700	.594	.643
	LIMSI-run1	.883	.760	.817	LIMSI-run1	.709	.559	.625
	TUC-MI-run1-corrected	.883	.539	.669	TUC-MI-run1-corrected	.780	.290	.423
	TUC-MI-run2-corrected	.882	.536	.667	TUC-MI-run2-corrected	.767	.283	.414
	UNIPD-run1	.629	.468	.537	UNIPD-run2	.350	.381	.365
	UNIPD-run2	.518	.384	.441	UNIPD-run1	.362	.251	.296
	Mondeca-run1	.375	.131	.194	Mondeca-run1	.335	.228	.271
	Frequency baseline	.339	.237	.279	Frequency baseline	.381	.110	.170

Several formal evaluation of a semantic annotator

Same tool with different P/R according to the context

Corpora	P/R first iteration	P/R second iteration
CDW Rouen	0.36/0.63	0.62/0.68
LiSSa, bibliographic database (scientific articles)	0.72/0.85	0.91/0.87
Documents in General Practice	0.80/0.84	

Big Data, definitions

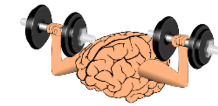
From the 3Vs to 10 Vs

Volume / נפח	•More than 90% of ever generated data where produced these last years
Velocity / מהירות	•Speed at which data is being generated, produced, created, or refreshed
Variety / שוני	•Structured, Semi-Structured, Unstructured data
Variability / השתנות	•Inconsistencies in the data resulting from multiple disparate data types and sources
Veracity / אמינות	•Confidence or trust in the data… dropping when Volume, Velocity, Variety, Variably increase
Validity / תקפות	•Accuracy and correctness of the data is for its intended use
Vulnerability / פגיעות	•Security concerns
Volatility / נדיפות – אי-יציבות	•How old does data need before being considered irrelevant, historic, or not useful any longer?
Visualization /ויזואליזציה	•Challenges due to technical (in-memory, scalability, processing time) and human (perception)
Value	•Driving business value from data

Big Data, definitions

From the 3Vs to 10 Vs

Examples in health and medicine



Volume	• …
Velocity	• ..
Variety	• …
Variability	• ..
Veracity	• …
Validity	• …
Vulnerability	• …
Volatility	• ….
Visualization	• …
Value	• …

CDW Rouen - volumetry

Patients	2 millions
Stays (hospitalisations, séances, consultations)	22.3 millions
Documents (reports, notes, letters, manual prescriptions...)	21.4 millions > **1G** medical concepts
CPOE	1.8 millions
Lab tests (hematology, biochemistry...)	176 millions
Medical devices	116,000
Diagnostics (ICD-10)	11.6 millions
Procedures (imaging, surgery...) (CCAM)	10.7 millions

Big Data, definitions

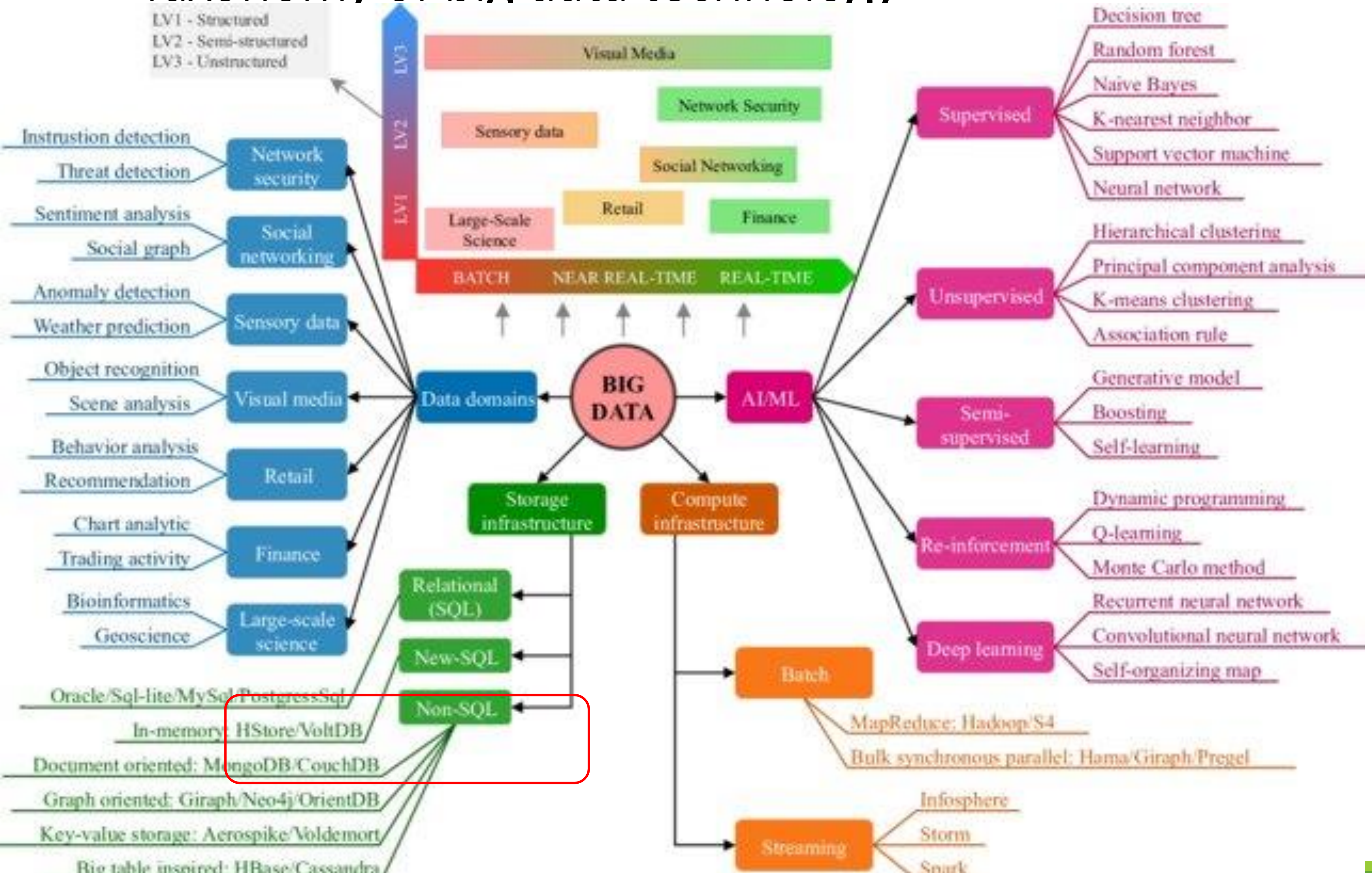
The Big Data Ecosystem is rich

The Big Data technology stack is evolving rapidly

Big Data, definitions

The Big Data Ecosystem is rich (2)

Taxonomy of big data technology



NoSQL – a critical component of the Big Data Ecosystems

All in the NoSQL Family

NoSQL databases are geared toward managing large sets of varied and frequently updated data, often in distributed systems or the cloud. They avoid the rigid schemas associated with relational databases. But the architectures themselves vary and are separated into four primary classifications, although types are blending over time.

Document databases

Store data elements in document-like structures that encode information in formats such as JSON.

+

Common uses include content management and monitoring web and mobile applications.

+

EXAMPLES
Couchbase Server, CouchDB, MarkLogic, MongoDB

Graph databases

Emphasize connections between data elements, storing related "nodes" in graphs to accelerate querying.

+

Common uses include recommendation engines and geospatial applications.

+

EXAMPLES
AllegroGraph, Amazon Neptune, ArangoDB, IBM Db2 Graph, Neo4j

Key-value stores

Use a simple data model that pairs a unique key and its associated value in storing data elements.

+

Common uses include storing clickstream data and application logs.

+

EXAMPLES
Aerospike, Amazon DynamoDB, Azure Table Storage, Redis, Riak

Wide-column stores

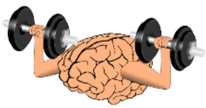
Also called table-style databases, they store data across tables that can have very large numbers of columns.

+

Common uses include internet search and other large-scale web applications.

+

EXAMPLES
Accumulo, Cassandra, Google Cloud Bigtable, HBase, ScyllaDB



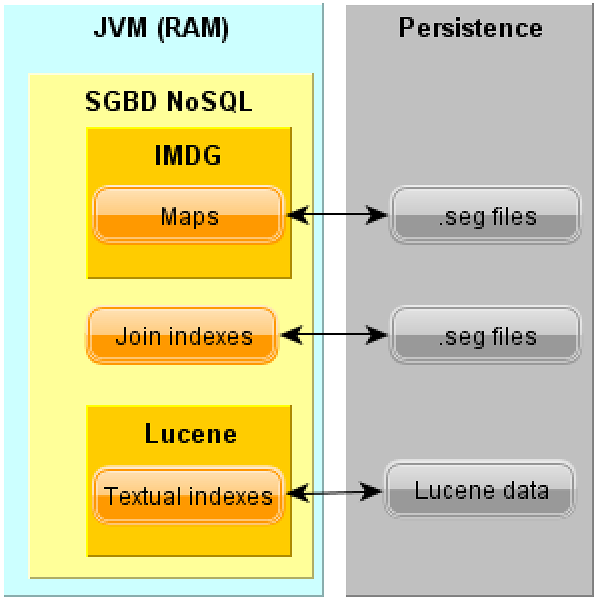
Why can it be important in healthcare research ?

NoSQL Architecture IN Rouen CDW

Layer 0

Java

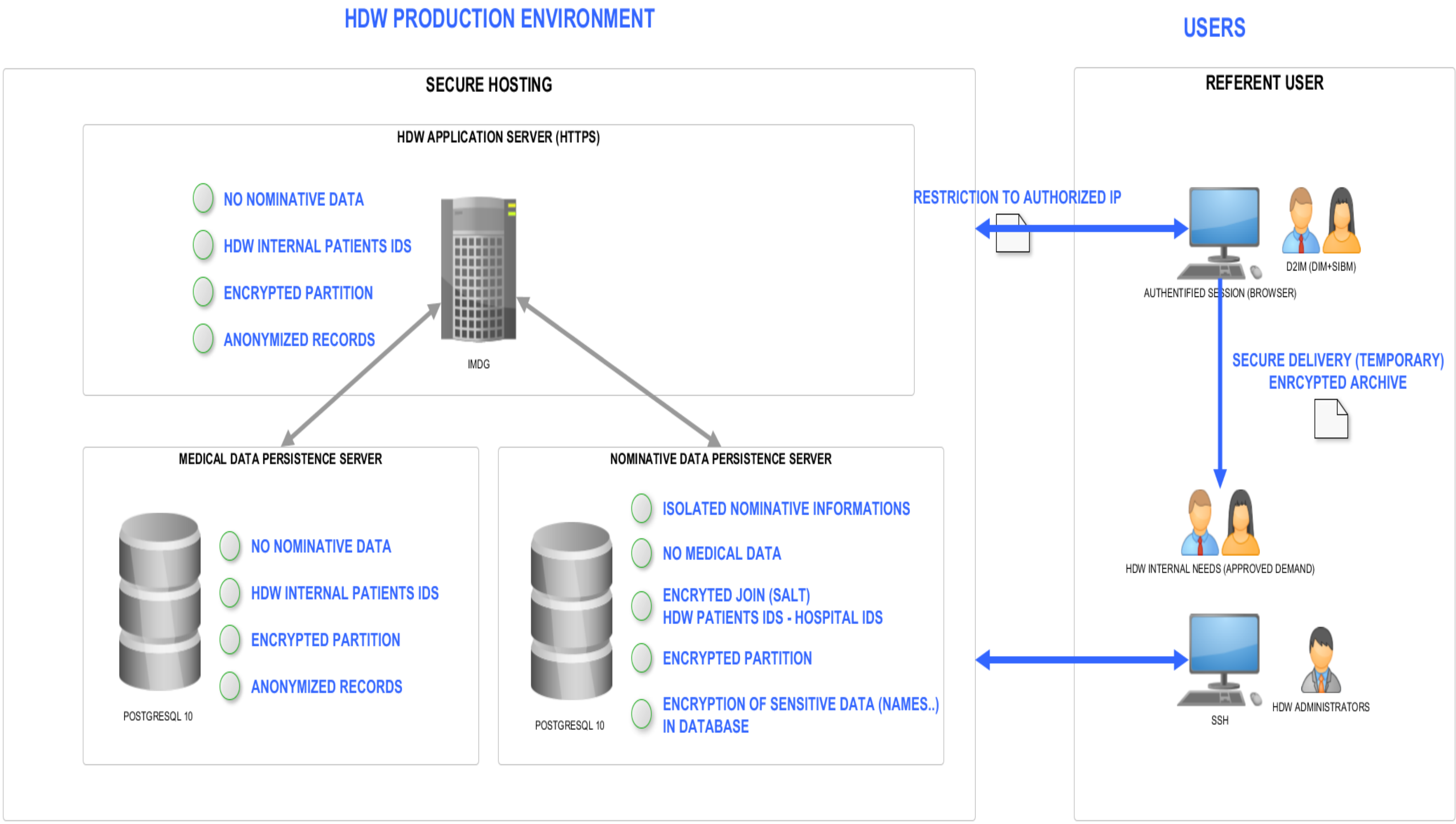
Custom NoSQL (In Memory Data Grid –IMDG-): Key-Value store, join indexes (replacing SQL joins) & Lucene (NLP)



Three powerful servers:

- 1 To RAM & 196 cores (prod)
- 1 To RAM & 144 cores (preprod) * 2
- Semantic Annotator + deep learning

HEALTH DATA WAREHOUSE - SECURE ARCHITECTURE PROPOSAL



Acess Policy to EDSaN

https://edsan.chu-rouen.fr/edsan/acces-aux-donnees/procedures/

EDSaN has been credited by the French CNIL, respect to European GPDR (nov 2018)

1. Demand sent to Department of Clinical Research (commission of qualification)
2. Transmission to the Scientific and Ethical Committee
3. « Physical » appointment in the DDH to create the EDSaN queries
   - Ethical & Juridical aspects: EDSaN provides access to all RUH health since 2000
   - Scientific aspects : bias of data, bias of tools...
4. Access to queried and specific data in a encrypted environment (nominative access)
5. Potential exports (realized after validation by a specific committee)

Introduction to Big Data, Big Data in Healthcare, and NoSQL

From Data to Big Data and Artificial Intelligence
- Every minute of the day…
- A Revolution in Data Availability
- Data for… Anything!
- Data, Big Data, Artificial Intelligence… From Fiction to Reality !

Data from a "Business" Perspective
- What's a Business? A "Organization" and more…
- Paradigm shift Data as a Critical Organizational Resource
- From Data to Wisdom… or the Big Data Holy Grail - The DIKW model
- The DIKW model as a Business Intelligence Environment and Data Science

Big Data, definitions
- Types of Data / Big Data
- From the 3Vs to 10 Vs
- The Big Data Ecosystem is rich
- NoSQL

Big Data in Medicine
- **Big Data and medical research**
- **Available Biobanks increasing**
- **Multi-sources for Multi-objectives**
- **Big Data and Machine Learning**
- **Big Health Data a competitive business**

Big Data in Medicine

Big Data and medical research

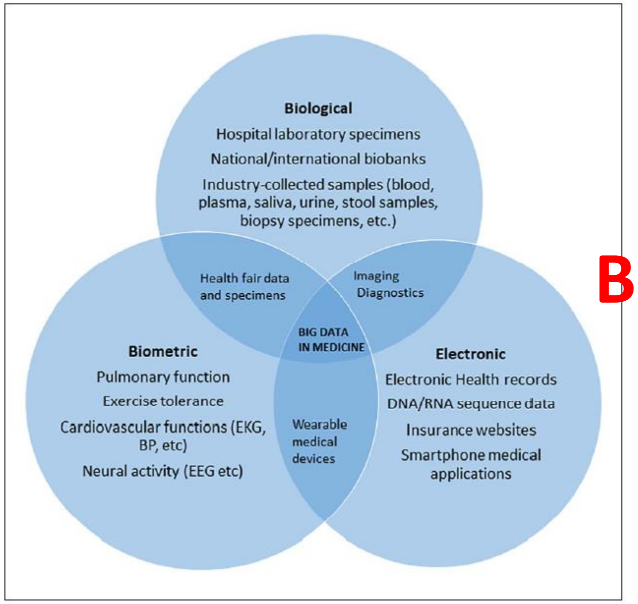
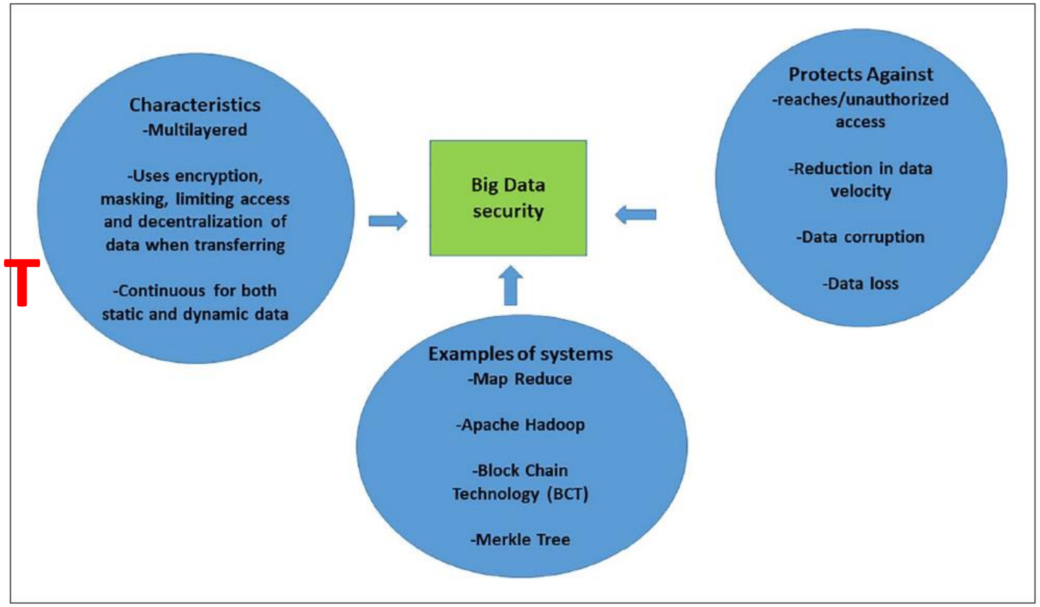


Figure.1. Big data in medicine.

BUT

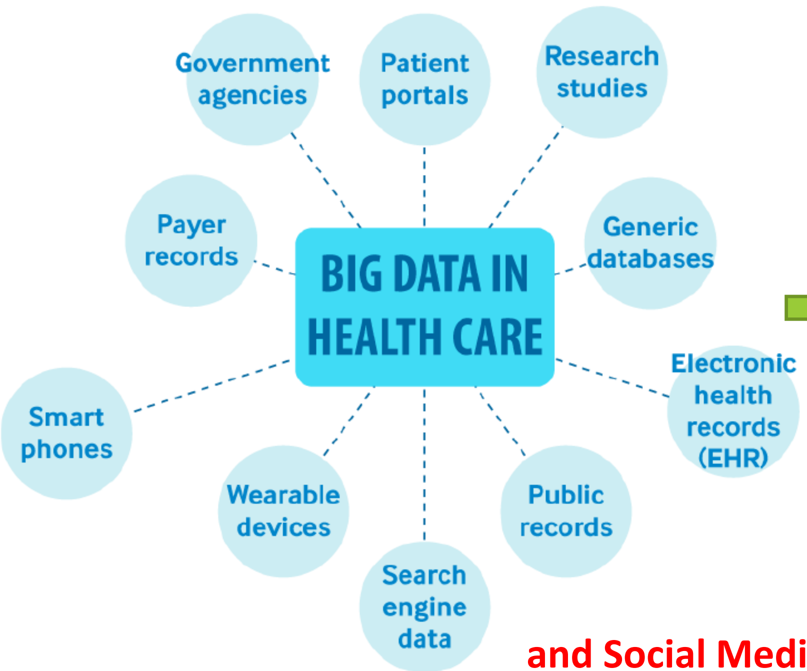


Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med. 2020 Jun 25;8:2050312120934839. doi: 10.1177/2050312120934839. PMID: 32637104; PMCID: PMC7323266.

Big Data in Healthcare

Summary

Sources of Big Data in Health Care



and Social Media

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Applications for Big Data in Healthcare

Diagnostics
Data mining and analysis to identify causes of illness

Preventative medicine
Predictive analytics and data analysis of genetic, lifestyle, and social circumstances to prevent disease

Precision medicine
Leveraging aggregate data to drive hyper-personalized care

Medical research
Data-driven medical and pharmacological research to cure disease and discover new treatments and medicines

Reduction of adverse medication events
Harnessing of big data to spot medication errors and flag potential adverse reactions

Cost reduction
Identificaton of value that drives better patient outcomes for longterm savings

Population health
Monitor big data to identify disease trends and health strategies based on demographics, geography, and socio-economics

HDW Objectives in general

To optimize DRGs by semiautomatic detection of atypical profiles between coding and HDW data (business €/$/£/₪)

To Improve clinical research thanks to feasibility studies prior to clinical trials & optimization of inclusions

Detection specific patients profiles

- e.g. patients frequently admitted to the emergency department

To create and maintain epidemiological cohorts and registries

HDW Objectives in general

To detect specific adverse events

- Vigilances, iatrogenic, cross infections

To create and follow-up quality indicators

To Develop and assess computer aided decision support systems (CDASS)

Access to epidemiology surveillance tools at various level (individual or collective)

Tools to improve clinical practice

- Dashboard to provide feedback to individual or collective practice

HDW Objectives for pharmaceutical companies

To improve clinical research thanks to feasibility studies prior to clinical trials & optimization of inclusions

- How many patients has the disease A (incidence) in 20xy?

Detection specific patients profiles

- e.g. patients frequently admitted to the emergency department

To create and maintain epidemiological cohorts and registries

To detect specific adverse events

- Vigilances, iatrogenic, cross infections

To create and follow-up quality indicators

To Develop and assess computer aided decision support systems (CDASS)

Access to modelization tools for physicians & researchers

CDW State of the art

In the world:

i2b2 (Harvard MS): over 150 University Hospitals are using I2B2 (SQL tecnology)

…

In France :

ehop (Rennes UH); over 15 UH in France are using Ehop

Dr Warehouse (Necker, Foch – Paris)

ConSoRe Continuum Soins Recherche specialised in cancer hospitald

EDSaN (Rouen UH); will be used in CDW in general practice

CDW state of the art

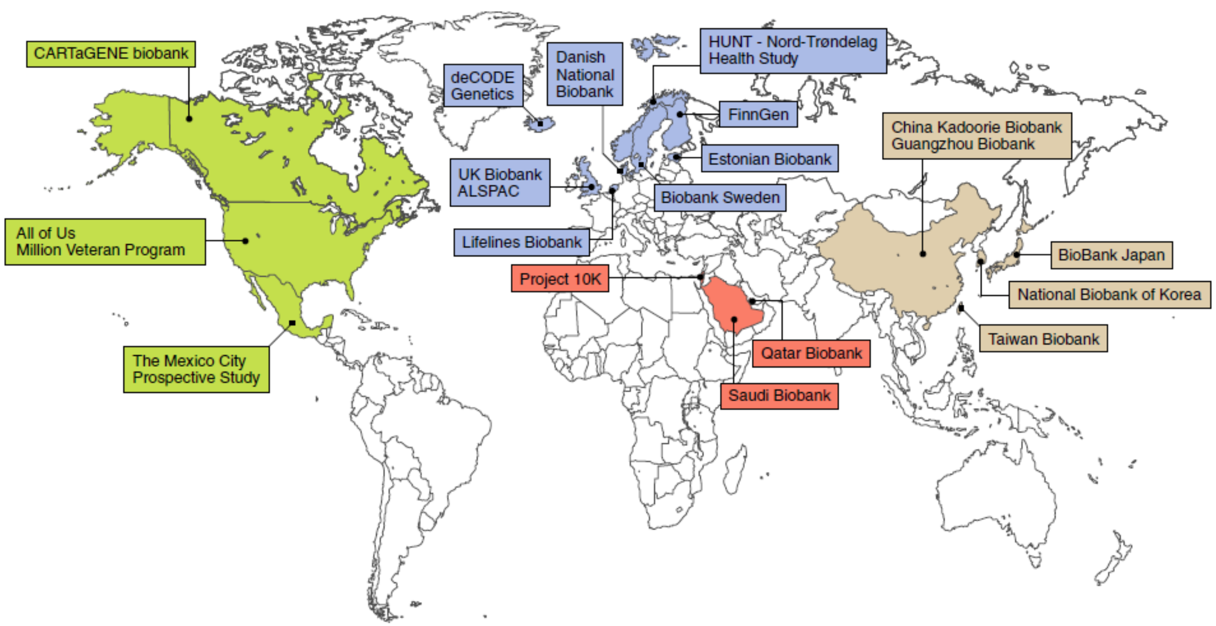
« CDW » at the national level; in fact, direct use of national EHR, Electronic Health Record

Existing in some « small » countries (in terms of population)

- Complexity of digital health not linear with population
- e.g. Israel (e.g. Clalit), Taiwan, Singapore, Denmark
- Main publications about COVID-19

At the HMO level in the US

Available Biobanks increasing and providing more big health data



Location	Biobank	N (goal)
Canada	CARTaGENE biobank[119]	43,000
USA	All of Us[33] Million Veteran Program[49]	1,000,000 > 600,000
Mexico	The Mexico City Prospective Study[52]	150,000
Iceland	deCODE Genetics	500,000
UK	UK Biobank[38] Avon Longitudinal Study of Parents and Children (ALSPAC)[20]	500,000 > 15,000
Netherlands	Lifelines Biobank[120]	> 167,000
Denmark	Danish National Biobank[121]	
Norway	HUNT - Nord-Trøndelag Health Study[122]	125,000
Sweden	Biobank Sweden	
Finland	FinnGen	500,000
Estonia	Estonian Biobank[123]	52,000
Israel	Project 10K	10,000
Saudi Arabia	Saudi Biobank	200,000
Qatar	Qatar Biobank[124]	60,000
China	China Kadoorie Biobank[51] Guangzhou Biobank[125]	> 500,000 30,000
Japan	BioBank Japan[126]	200,000
Korea	National Biobank of Korea[127]	500,000
Taiwan	Taiwan Biobank[128]	200,000

Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med. 2020 Jan;26(1):29-38. doi: 10.1038/s41591-019-0727-5. Epub 2020 Jan 13. PMID: 31932803.

Multi-sources for Multi-objectives optimization

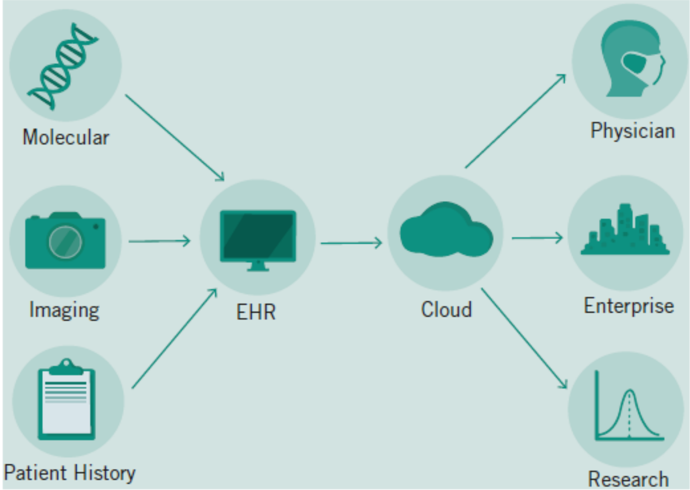


Fig. 2 General model of care envisioned. Here, various heterogeneous data types are fed into a centralized EHR system that will be uploaded to a secure digital cloud where it can be de-identified and used by research and enterprise, but primarily by physicians and patients.

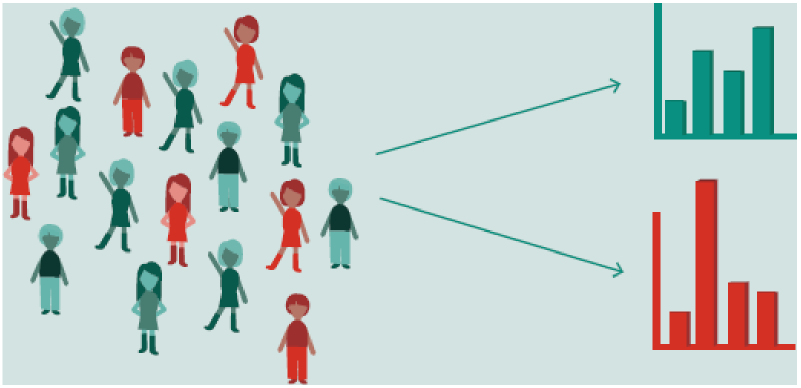


Fig. 6 Minority representation in current large-scale experiments integrating across a variety of factors is often lacking. The "All of Us" study will meet this need by specifically aiming to recruit a diverse pool of participants to develop disease models that generalize to every citizen, not just the majority (Denny et al. 2019). Future global Big Data generation projects should learn from this example in order to guarantee equality of care for all patients.

Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity (Edinb). 2020 Apr;124(4):525-534. doi: 10.1038/s41437-020-0303-2. Epub 2020 Mar 5. PMID: 32139886; PMCID: PMC7080757.

Big Data and Machine Learning for a personalized medicine

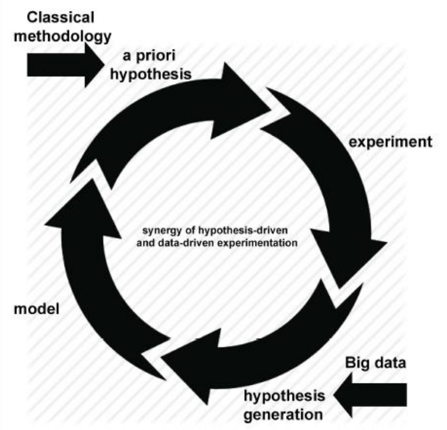


FIGURE 1 | The synergistic cycle of hypothesis-driven and data-driven experimentation.

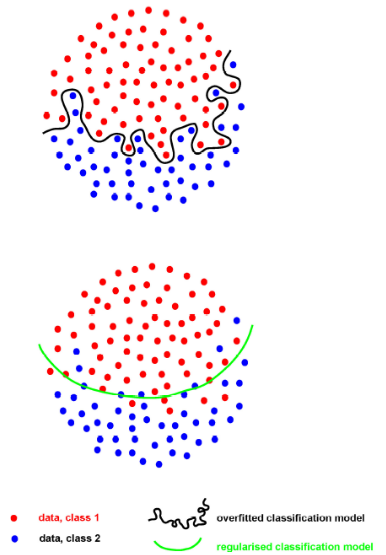


FIGURE 2 | The concept of overfitting and model regularization.

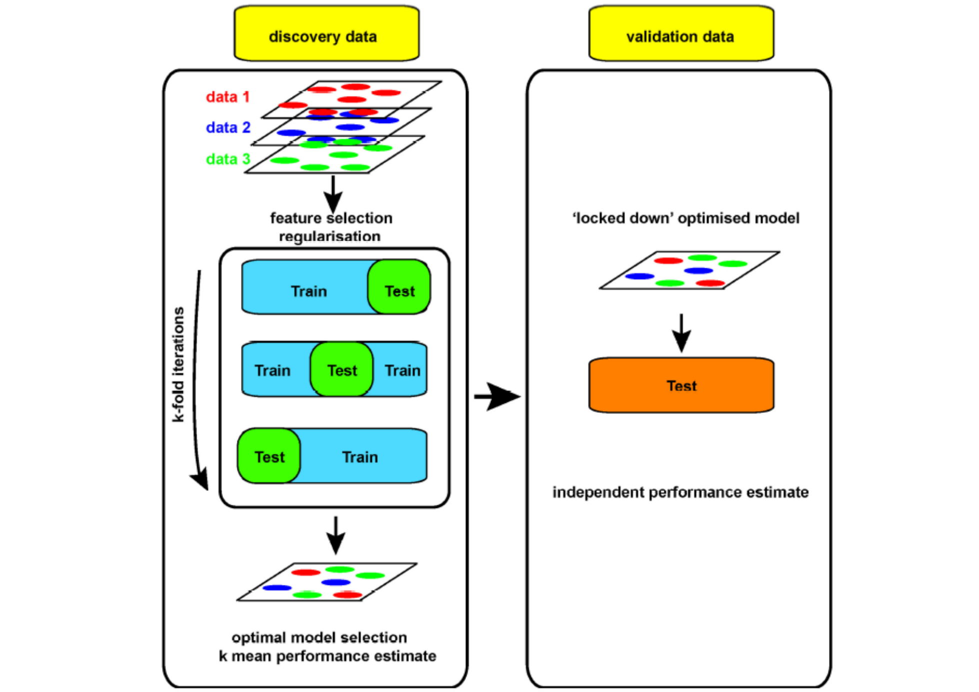


FIGURE 3 | Model performance evaluation.

Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, Spreafico R, Hafler DA and McKinney EF (2019) From Big Data to Precision Medicine. Front. Med. 6:34. doi: 10.3389/fmed.2019.00034

Big Data in Medicine

Big Health Data a competitive business

Table 2 List of some of big companies which provide services on big data analysis in healthcare sector

Company	Description	Web link
IBM Watson Health	Provides services on sharing clinical and health related data among hospital, researchers, and provider for advance researches	https://www.ibm.com/watson/health/index-1.html
MedeAnalytics	Provides performance management solutions, health systems and plans, and health analytics along with long track record facility of patient data	https://medeanalytics.com/
Health Fidelity	Provides management solution for risks assessment in workflows of healthcare organization and methods for optimization and adjustment	https://healthfidelity.com/
Roam Analytics	Provides platforms for digging into big unstructured healthcare data for getting meaningful information	https://roamanalytics.com/
Flatiron Health	Provides applications for organizing and improving oncology data for better cancer treatment	https://flatiron.com/
Enlitic	Provides deep learning using large-scale data sets from clinical tests for healthcare diagnosis	https://www.enlitic.com/
Digital Reasoning Systems	Provides cognitive computing services and data analytic solutions for processing and organizing unstructured data into meaningful data	https://digitalreasoning.com/
Ayasdi	Provides AI accommodated platform for clinical variations, population health, risk management and other healthcare analytics	https://www.ayasdi.com/
Linguamatics	Provides text mining platform for digging important information from unstructured healthcare data	https://www.linguamatics.com/
Apixio	Provides cognitive computing platform for analyzing clinical data and pdf health records to generate deep information	https://www.apixio.com/
Roam Analytics	Provides natural language processing infrastructure for modern healthcare systems	https://roamanalytics.com/
Lumiata	Provides services for analytics and risk management for efficient outcomes in healthcare	https://www.lumiata.com
OptumHealth	Provides healthcare analytics, improve modern health system's infrastructure and comprehensive and innovative solutions for the healthcare industry	https://www.optum.com/

Fig. 4 Illustration of application of "Intelligent Application Suite" provided by AYASDI for various analyses such as clinical variation, population health, and risk management in healthcare sector

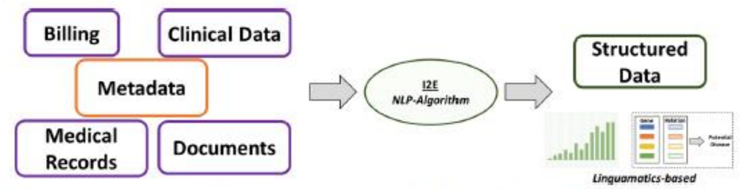


Fig. 5 Schematic representation for the working principle of NLP-based AI system used in massive data retention and analysis in Linguamatics

Dash, S., Shakyawar, S.K., Sharma, M. *et al.* Big data in healthcare: management, analysis and future prospects. *J Big Data* **6**, 54 (2019). https://doi.org/10.1186/s40537-019-0217-0

Big Data in Medicine

Prof. David Salt, University of Aberdeen Scootland, UK

Why is Big Data important for biology?

https://www.youtube.com/watch?v=K-qOxElpLk8

Deep learning in RUH, France

Medical word embeddings (Word2Vec, FastText,Glove)

- PhD Emeric Dynomant (OMICX)
- Two different corpus
  - 12 M health documents from HSDW
  - 180 K abstracts from LiSSa, French bibliographic database

Doc2Vec, Patient2Vec (in progress)

- PhD Mikaël Dusenne, MD
  - Hybrid semantic annotator (“old” NLP + deep NLP)
  - Doc2Vec2DRGs, using an other tool !!! ELMO?

Réservation Cinémas Pathé | Medical Embedding - Result | darmoni - PubMed - NCBI | Clinical Natural Language P | LIMICS | licence pour logiciel - Tradu

https://cispro.chu-rouen.fr/winter/query/

Medical word embeddings querying page

12M ▾ | endocardite | Search

GloVe	infectieuse, myocardite, eto, native, streptocoque, bovis, bactériémie, faecalis, _endocardite
FastText (CBOW)	proprio_septive, myopericardite, septo_optique, endo_aortique, endocartite, endoculaire, acrodermite, rhino_septale, salmonellose, épidermolyse
FastText (Skip-Gram)	endocartite, endocardique, proctologique, extancilline, septo_basale, prolongements, recanalisée, précentrale, dantrolene, podoscopique
Word2Vec (Skip-Gram)	bovis, sanguinis, eto, gordonii, gallolyticus, aorto_mitrale, mutans, infectieuse, streptoccoque, salivarius
Word2Vec (CBOW)	endocartite, _endocardite, native, bovis, médiastinite, myocardite, mutans, gallolyticus, myopéricardite, tamponnade

Home

Dynomant E, Lelong R, Dahamna B, Massonaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ.
Word embedding for French natural language in healthcare: a comparative study.
JMIR Med Inform. 2019 Jul;7(3):e12310. DOI : 10.2196/12310

Réservation Cinémas Pathé | Medical Embedding - Result | darmoni - PubMed - NCBI | Clinical Natural Language P | LIMICS | licence pour logiciel - Tradu

https://cispro.chu-rouen.fr/winter/query/

Medical word embeddings querying page

LiSSa ▾ endocardite Search

Word2Vec (Skip-Gram)	endocardites, bactériémie, infectieuse, valvulopathie, septicémie, mycotique, spondylodiscite, médiastinite, valvulaire, bioprothèse

Home

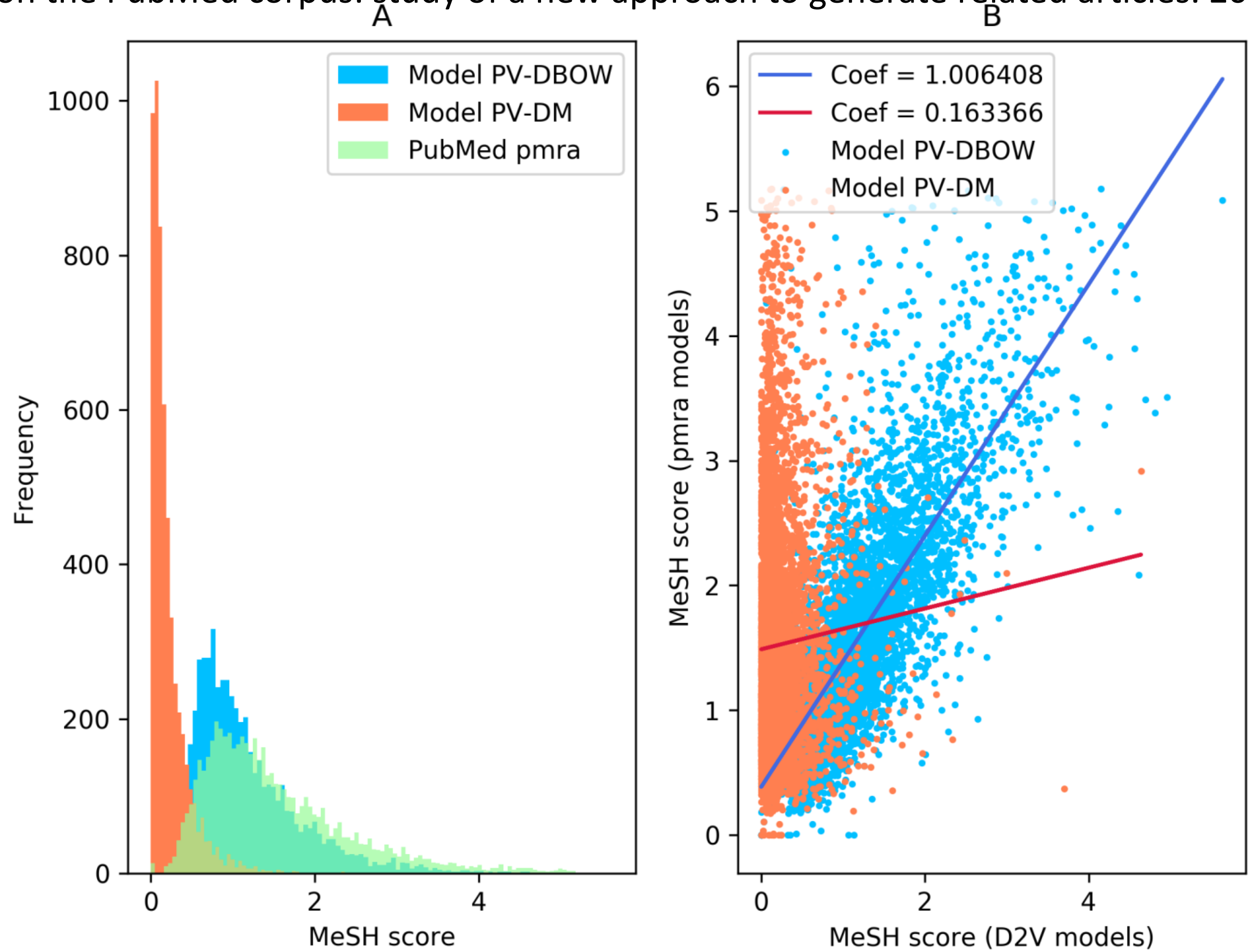
Wordembeddings in two different contexts

QUERY: “facebook”

LiSSa corpus (300k)	internet, twitter, web, blog, e_learning, blogs, internautes, tic, game, …
RUH documents (12M)	reproches, injures, messages, insultes, rumeurs, ex_conjointe, menaces, insultant, …

Doc2Vec2PubMed

Dynomant E, Darmoni SJ, Lejeune E, Kerdelhué G, Leroy JP, Lequertier V, Canu S, Grosjean J. Doc2Vec on the PubMed corpus: study of a new approach to generate related articles. 2019 Nov.



Perspectives
Towards « One Health »

1. Clinical
   - Hospital
   - General Pratice
   - National (EHR Israël, Denmark, Singapore, Taiwan; France & US: DRGs)
2. Omics
   - Genomics, Transcriptomics, Metabolomics, Nutriomics…
3. Environment
4. IoT & medical devices
5. Social Media, email…
6. Veterinary medicine
7. …

Introduction to Big Data, Big Data in Healthcare, and NoSQL

From Data to Big Data and Artificial Intelligence
- Every minute of the day…
- A Revolution in Data Availability
- Data for… Anything!
- Data, Big Data, Artificial Intelligence… From Fiction to Reality !

Data from a “Business” Perspective
- What’s a Business? A “Organization” and more…
- Paradigm shift Data as a Critical Organizational Resource
- From Data to Wisdom… or the Big Data Holy Grail - The DIKW model
- The DIKW model as a Business Intelligence Environment and Data Science

Big Data, definitions
- Types of Data / Big Data
- From the 3Vs to 10 Vs
- The Big Data Ecosystem is rich
- NoSQL

Big Data in Medicine
- Big Data and medical research
- Available Biobanks increasing
- Multi-sources for Multi-objectives
- Big Data and Machine Learning
- Big Health Data a competitive business

Publications

https://edsan.chu-rouen.fr/edsan/communication/publications/

Romain Lelong; Badisse Dahamna; Romain Leguillon; julien Grosjean; Catherine Letord; SJ Darmoni & Lina F. Soualmia. Assisting Data Retrieval with a Drug Knowledge Graph. **ICIMTH2021**, 2021.

Stéfan J. Darmoni. IA au sein d'un entrepôt de données de santé à Rouen. **Bulletin de l'AfIA** , 04, Number 112, Pages 18-20, 2021.

Pressat-Laffouilhère T, Balayé P, Dahamna B, Lelong R, Billey K, Darmoni SJ, Grosjean J. Evaluation of Doc'EDS: A French Semantic Search Tool to Query Health Documents from A Clinical Data Warehouse. BMC Med Inform Decis Mak. 2020 Sep. DOI : 10.21203/rs.3.rs-59497/v1

Dynomant E, Lelong R, Dahamna B, Massonaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ. Word embedding for French natural language in healthcare: a comparative study. JMIR Med Inform. 2019 Jul;7(3):e12310. DOI : 10.2196/12310

Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials. JMIR Med Inform. 2019;30. DOI : 10.2196/13917

Siefridt C, Grosjean J, Lefebvre T, Rollin L, Darmoni SJ, Schuers M. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. Int J Med Inform. 2019. DOI : 10.1016/j.ijmedinf.2019.104009

Grosjean J, Letord C, Charlet J, Aimé X, Danès L, Rio J, Zana I, Darmoni SJ, Duclos C. Un modèle sémantique d'identification du médicament en France. Atelier IA & Santé; 2019 Juil 1; Toulouse, France.

Dynomant E, Darmoni SJ, Lejeune E, Kerdelhué G, Leroy JP, Lequertier V, Canu S, Grosjean J. Doc2Vec on the PubMed corpus: study of a new approach to generate related articles. 2019 Nov.

Lelong R. Accès sémantique aux données massives et hétérogènes en santé. Normandie Université. 2019 Juin.

Ndangang M, Grosjean J, Lelong R, Dahamna B, Kergourlay I, Griffon N, Darmoni SJ. Terminology Coverage from Semantic Annotated Health Documents. Stud Health Technol Inform. 2018;255:20-4. DOI : 10.3233/978-1-61499-921-8-20

Lelong R, Soualmia LF, Sakji S, Dahamna B, Darmoni SJ. NoSQL technology in order to support Semantic Health Search Engine. MIE 2018: Medial Informatics Europe; 2018 Apr 24-26; Gothenburg, Sweden.

Lelong R, Soualmia LF, Dahamna B, Griffon N, Darmoni SJ. Querying EHRs with a Semantic and Entity-Oriented Query Language. Stud Health Technol Inform. 2017;235:121-5. DOI : 10.3233/978-1-61499-753-5-121

Cabot C. Recherche d'information clinomique au sein du Dossier Patient Informatisé : modélisation, implantation et évaluation. Normandie Université, 2017.