Big Data em Saúde: Desafios e Perspectivas

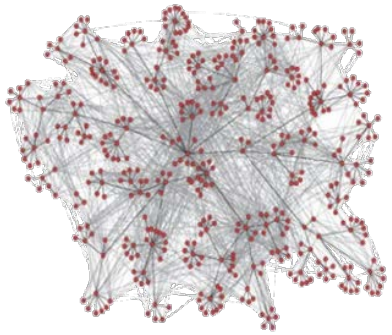
Agma J. Machado Traina
agma@icmc.usp.br

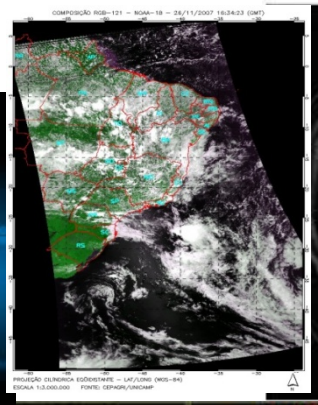
Databases and Images Research Group
Computer Science Department
University of São Paulo (USP) at São Carlos
Brazil

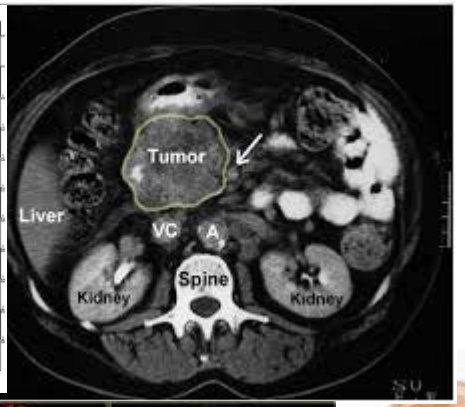
Introduction

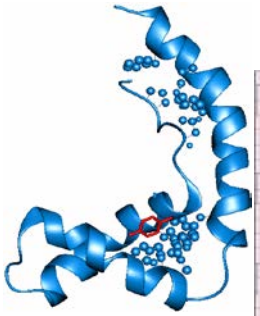
Current Scenario...

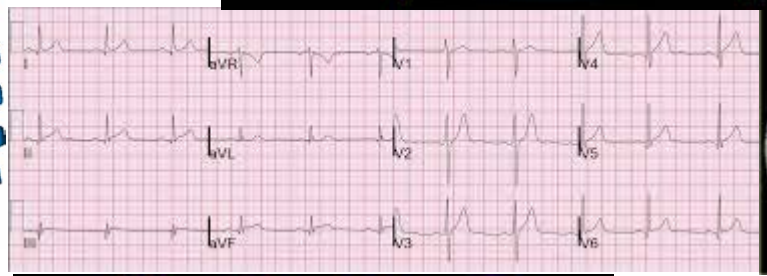
Complex big data everywhere

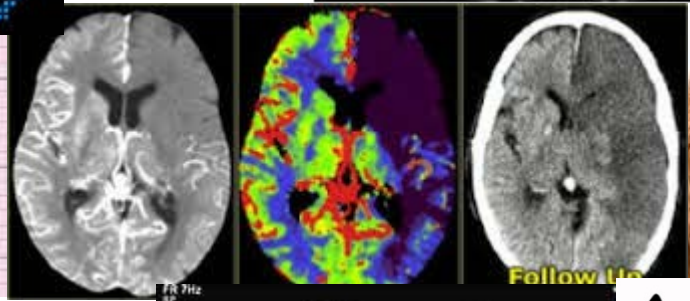


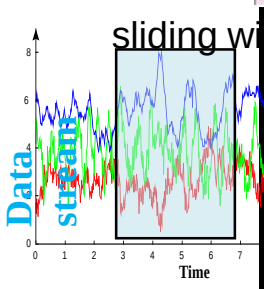


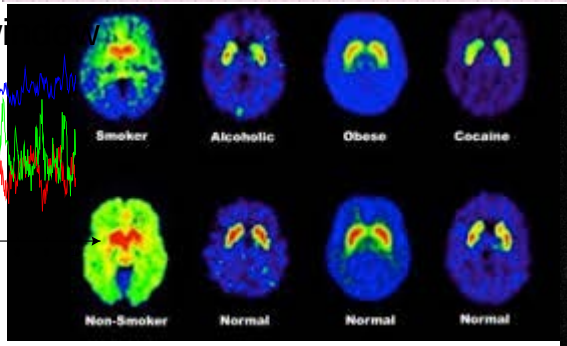


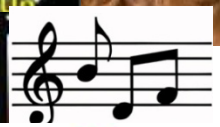






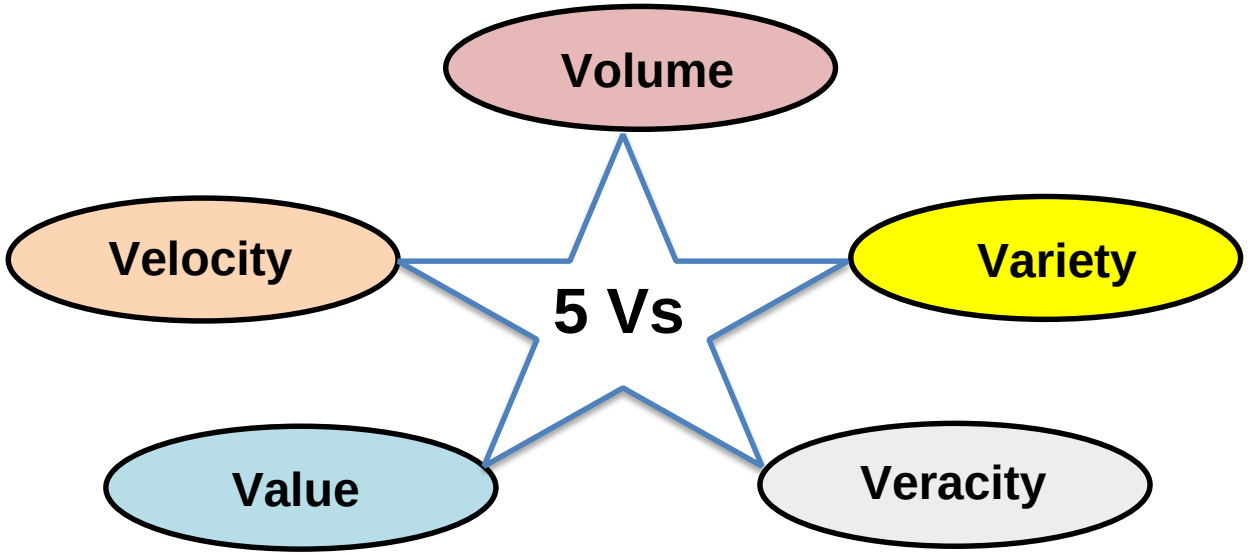






Current Scenario... Big Data

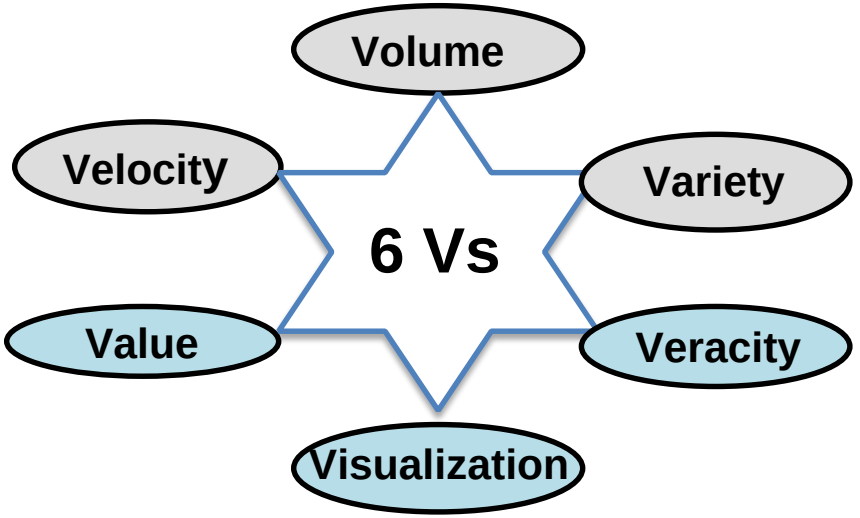
150 exabytes or 10^{18} bytes of new healthcare data generated yearly in USA, growing 48% annually [1]



[1] Nature Medicine | VOL 25 | Jan 2019 | 24–29 | www.nature.com/naturemedicine
Stanford Health. Harnessing the power of data in health. *Stanford Medicine 2017 Health Trends Report* (2017).

Introduction: Big Data challenges

- Development of **new**, **scalable** and **expandable** big data infrastructure (*volume*, *velocity*, *variety*),
- **analytical methods** (*veracity* and *value* and
- **visualization** techniques to support understanding data/information/knowledge gathered yielding better decisions and outcomes.

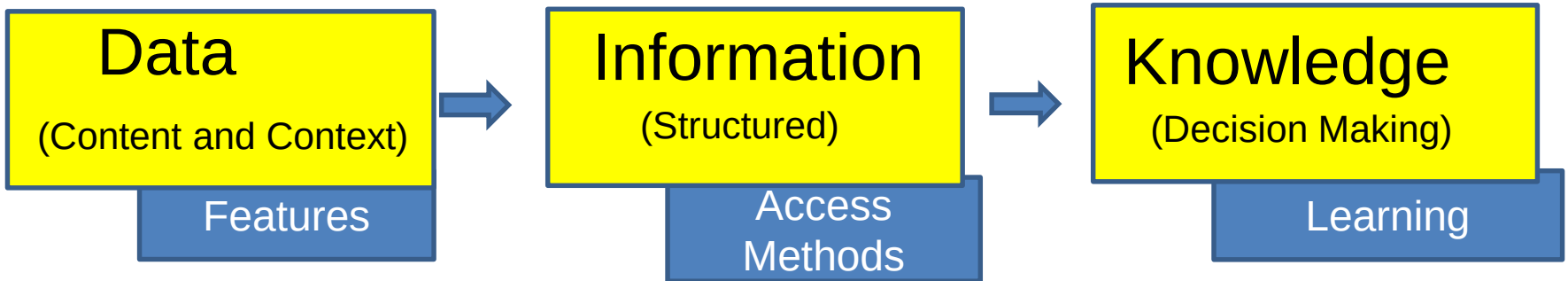


Current Scenario... Big Data

A great need for developing **automated systems** that could help users to **retrieve complex data** from the databases, employing their inherent **content** supported by their **context**

Thus, for each type of data it is important:

- Extract the relevant **features** that **best describe** them
- Get **dependable** data
- **Index** the data for **fast retrieval/processing** (**scalability**)
- Process queries on their content (**Similarity queries**)



DIKW Pyramid

garbage in ► garbage out

DIKW Pyramid

Considering the *DIKW* Pyramid:

- **How** are my data?
- **What** the data can **provide**?
- **When** the data is useful?
- **Who** owns the data? Who benefits from the data?

Ethical issues

Ethical issues: some studies

- **United Kingdom** survey (April 2018): 65% have reservations on allowing personal data to improve healthcare and is unfavorable to AI replacing doctors and nurses [1]
- **Germany**: 83% medical students are motivated with AI, but skeptical regarding diagnosing ... [2]
- **USA:** ~50% healthcare organizations decision-makers are not confident that AI will improve medicine, but ~ 50% of them have reservations [3]:
  - produce fatal errors,
  - not work properly,
  - not meet currently hyped expectations

1. Fenech M, Strukelj N, Buston O. *Ethical, social and polictical challenges of artificial intelligence in health.* 2018 April. http://futureadvocacy.com/wp-content/uploads/2018/04/1804_26_FA_ETHICS_08-DIGITAL.pdf.
2. Pinto dos Santos D, Giese D, Brodehl S, Chon SH, StaabWet al. *Medical students' attitude towards artificial intelligence: a multicentre survey*. Eur Radiol 2018 Jul 6. https://doi.org/10.1007/s00330-018-5601-1 PMID: 29980928
3. Intel Corporation. *Overcoming barriers in AI adoption in healthcare*. 2018 April https://newsroom.intel.com/wp-content/uploads/sites/11/2018/07/healthcare-iot-infographic.pdf.

Ethical issues: Principles

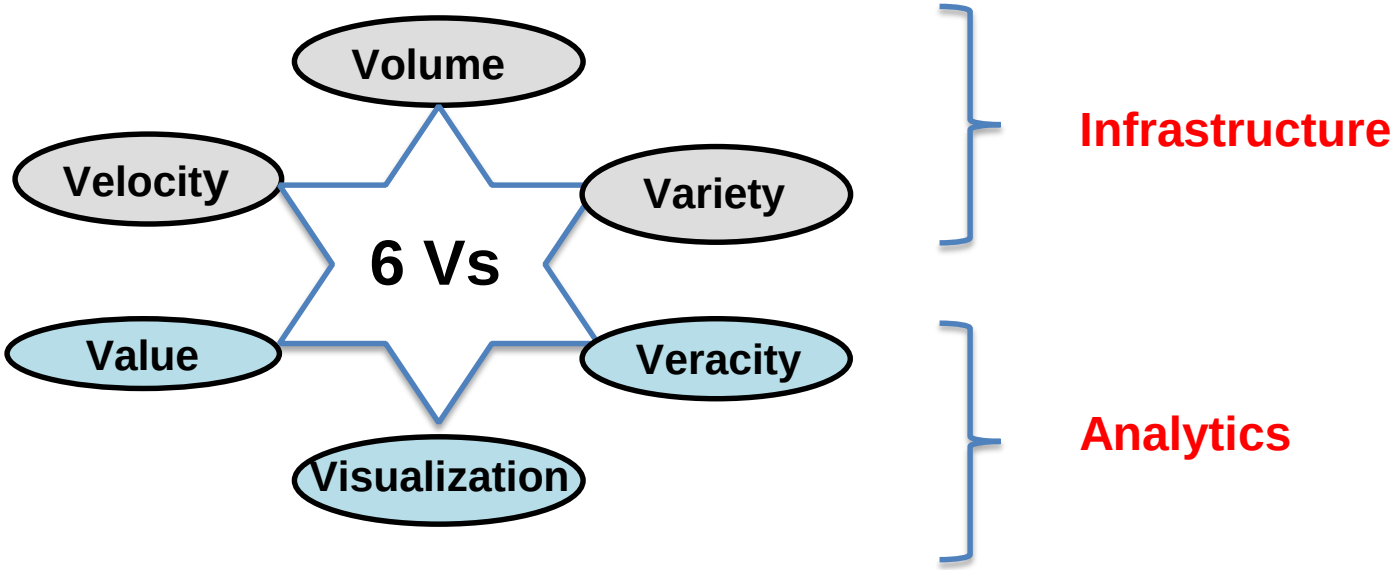
- transparency
- fairness
- non-maleficence
- liability
- privacy

❑ European General Data Protection Regulation (GDPR): Europe
❑ Health Insurance Portability and Accountability Act (HIPAA): USA
...

Explainable AI (XAI) vs. Black Box

Big Data still has meaningfull challenges !

- Development of **new**, **scalable** and **expandable** big data infrastructure (*volume*, *velocity*, *variety*),
- **analytical methods** (*veracity*, *value* *and* *visualization*)



Complex data with Missing values

- Big data applications must deal with:
  - Large number of data elements (i.e., cardinality)
  - High dimensionality (i.e., number of attributes)
  - Complexity of the features that describe the attributes
  - Non-dimensional data (e.g. DNA sequences)

Missing data can occur due to:

- **Preventable errors** or **mistakes** (e.g. failing to appear for a medical exam,…etc).
- Problems **outside of control** (e.g. failure of the equipment, low battery,…etc).
- Privacy or security reason.
- **Legitimate** (e.g. a survey question that does not apply to the respondent).

	A_1	A2	A3	A4	A5
Obj_1					
Obj_2					
Obj_3					
Obj_4					
Obj_5					

	A_1	A2	A3	A4	A5
Obj_1		?		?	
Obj_2	?	?			
Obj_3	?	?	?	?	?
Obj_4		?			
Obj_5		?			?

Safia Brinis, Caetano Traina Jr., Agma Traina - *Hollow-tree: a Metric Access Method for Data with Missing Values*, Journal of Intelligent Information Systems, pp. 1-26. DOI: 10.1007/s10844-019-00567-8

Conclusions

Complex big data bring new interesting challenges:

- regarding BD infrastructure and analytics as well

- Access methods to query processing (**high-dimensional** and **adimensional data)**,

- **Scalable approaches** to deal with **missing data,**

- New mechanisms for **organizing the data,**

- To ease the problem of **“garbage in garbage out”**

- A **closer relationship with related fields** to gather/convey to the users the knowledge & wisdom needed and desired.

Many opportunities to work/research!

Thanks!

- To all the members of the **Databases and Images Group (GBdI)** ICMC-USP/São Carlos and **MiVisBD** research

- To the Ciência e Inovação Digital em Saúde FAPESP organizers

- To **you all** for attending to this panel

ACKNOWLEDGMENT