Big Data

BIG DATA

Data Analytics

From data to wisdow

What it can do for you?

Powered by HAL 9000 your Big Data solution















Big Data

BIG DATA

Data Analytics

From data to wisdow

What it can do for you?

Powered by HAL 9000 your Big Data solution

Who I am?



- **□** FAPESP IPT Scholarship
- **☐ Eng. Production POLI Mastering of science**
- ☐ FGV-SP MBA
- **☐** Computer Science UFMA Graduation
- ☐ Civil Engineering UEMA Graduation
- ☐ IT Market over the last 20 years
- ☐ Big Data HAL research group co-founder
- □ CDO at Intellibrand



















Big Data

BIG DATA

Data Analytics

From data to widows

What it can do for you?

Powered by HAL 9000 your Big Data solution









It is 1744, suddenly you discovered how many people had been born and dead over the last 50 years in a small city in Germany !!! What do you do with this data?



AGENDA



- I. Mankind big data history
- II. What is big data nowadays
- III. From data to widows
- IV. What is not big data
- V. Hands-on
- **VI. References**

Mankind big data history

- 1) Sapiens revolutions
- 2) Deterministic mindset
- 3) The world is not so easy
- 4) No, we can't control!!

1. Sapiens Revolutions

b Prehistoric humans (~2 million years) were no more important and impressive than other mammals. Human cultures began to take shape about 70,000 years ago.

Symbols empowered humans to connect around ideas that do not physically exist. These shared "myths" have enabled humans to take over the globe and make humankind overcome the forces of natural selection.



Agricultural Revolution

10,000 year ago

It made life worst at first – future **worries**: the weather, the crop yield this year Need to **write**, **measure**, **weight**

Cognitive Revolution

70,000 year ago



Communication

Talk about things we never **seen, touch or smell** – religions, myths, legends, fantasies **Collaboration** in large numbers







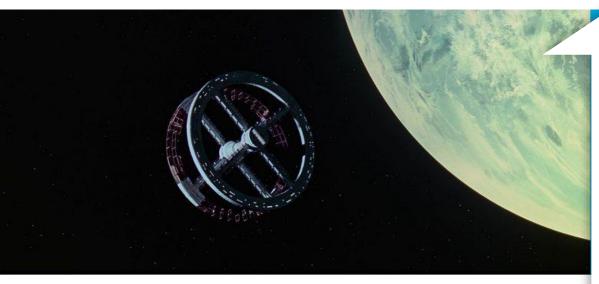




2. Deterministic Mindset

Laplace's Demon 1814 - scientific determinism – once you known location and momentum of atoms you

know the future based on the laws of classical mechanics



Scientific Revolution 500 years ago



Started in **Europe**

Admit our ignorance (ignoramus)

Special attention to observation and mathematics - data collection

Research method - turn data into knowledge

From Fuclid to Newton **deterministic** knowledge









3. The world is not so easy God does not play dice! YES, he does.





Not only does God play dice, but... he sometimes throws them where they cannot be seen."

@HuffPostU

The Nature of Space and Time, 1996, Professor Stephen Hawking and Roger Penrose

Scientific Revolution



Relativity and quantum mechanics as well as biology, economy and psychology do not fit determinism

STATISTICS













4. No, we can't control But we can sufficiently estimate

In 1744 Alexander Webster and Robert Wallace analyzed born & death numbers (1.238 babies born & 1.174 deaths) and concluded:

Years old person 1/100 chance to die In a certain year

ministers alive in a certain period of time

Years old person 1/39 chance to die In a certain year



In average 27 ministers dies per year

- **18** of them leaving a widow
- **5** without widow but with orphan children
- **2** with widow and orphan kids below 16 from previous engagement

Finally, they estimated **how long** till the widow die or get married again









4. No, we can't control But we can sufficiently estimate

In 1744 Alexander Webster and Robert Wallace analyzed born & death numbers (1.238 babies born & 1.174 deaths) and concluded:

Having these **information** they could estimate how much each minister have to pay in order to guarantee the future of your loved ones.



From data to knowledge to business

According to their accounts in 1765 the fund was supposed to have accumulated a total of £ 58.348 pounds. When this year arrived they had have £ 58.347.



3,500 employees £ 100 billion pounds









What is big data nowadays

- 1) Definitions
- 2) IBM 4 V's
- 3) Let's try to define Big Data
- 4) Mainstreamed representation
- 5) Big Data Analytics

1. Definitions

We live a data-driven world.

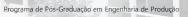
Data drives the modern organizations of the world.

3-Vs Definition

- Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making and process automation." ("Gartner IT Glossary, n.d.")
- "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." (TechAmerica Foundation's Federal Big Data Commission, 2012)

Clearly, **size** is the characteristic that comes to mind considering the question "what is big data?" However, other characteristics of big data have emerged. Laney (2012) suggested the three V's are the dimensions of challenges in data management. Gandomi & Haider(2014)







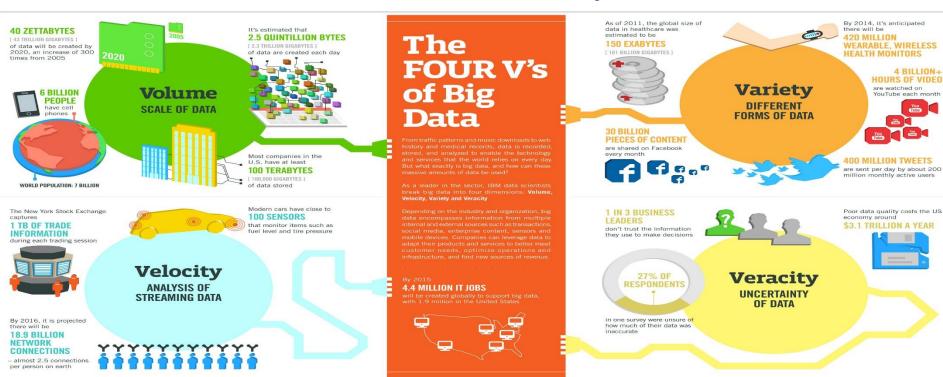






2. IBM 4 V's

Oracle 5th V (Value) + SAS 6th and 7th (Variability and Visualization)



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS







TRM



3. Let's try to define Big Data

[◦] More definitions of big data – academic viewpoint

In general, big data is perceived as a **source of innovative products, services, and business** opportunities (Davenport et al., 2012;)

Moreover, big data is believed to **result in more efficient and effective operations** by, for example, **optimizing** supply chain flows; setting the **most profitable price** for products and services; **selecting the right** people for certain tasks and jobs; **minimizing errors** and quality problems, and **improving** customer relationships (Chen et al., 2012).

Additionally, further economic and social value can be **gained** from big data through **enhanced decision making** (Sharma et al., 2014) and more informed strategizing (Constantiou and Kallinikos, 2015).

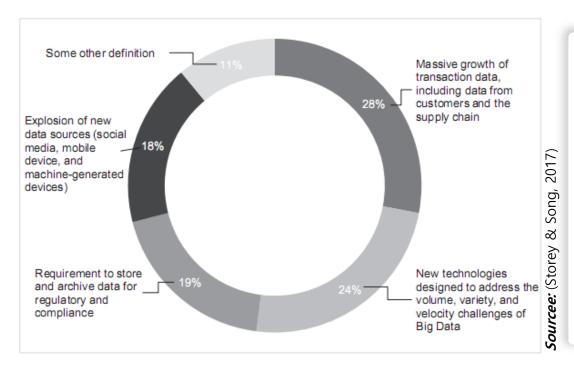






3. Let's try to define Big Data

⁶ More definitions of big data – practitioner viewpoint



Conclusion

Thus, both the academic and practitioner-oriented literatures are characterized by a strong focus on the opportunities that big data provides for organizations.





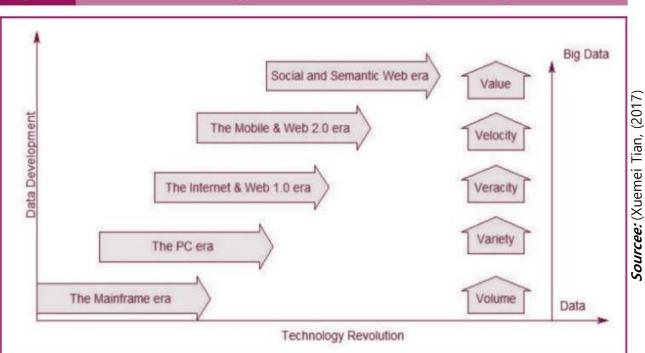




3. Let's try to define Big Data

⁶ More definitions of big data – academic viewpoint

Figure 1 Waves in the technology revolution and the emergence of big data



Davenport (2014) predicted that the term was facing a relatively short life span.

Mayer-Schonberger and Cukier (2013), an authoritative source, concurred that there was no rigorous definition of big data.

Chart discussion

Big data - (R)evolution or Challenges







4. Mainstreamed representation

A classical and mainstreamed Big Data representation – Magic Funnel



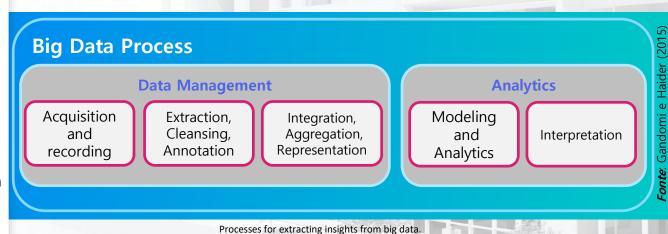


5. Big Data Analytics

⁶Big data are worthless. Its potential value is unlocked only when leveraged to drive decision making.

• From data to insights

- Organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights.
- There are five stages to extract insights from big data.
- These five stages form the two main sub-processes: data management and analytics.











Integrate analytics

management dashboards

and operational systems.

Perform evaluation against

metrics:

Communicate results and recommendations

procedures into

Monitor performance:

Identify parts that need to be

improved

What is the question to solve and

1. Business

Understanding

5. Model

Building

8. Review and

Monitoring

6. Evaluation

7. Deployment

Build models

Perform analysis and iterate

Programa de Pós-Graduação em Engenharia de Produção





and workflow?

Select key variables and determine

correlation between them



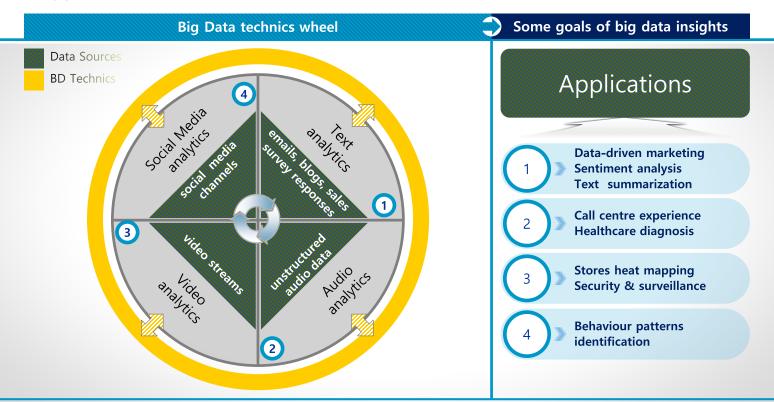


Song, 2017

ಹ

Source: (Storey &

5. Big Data Analytics Technics











||| From data to wisdom

^oWe live a data-driven world.

Data drives the modern organizations of the world.

3 fundamental keys

- In data science there are 'Data', 'Information' and 'Knowledge', often times the lines are blurred about what is each one of them.
- Data are facts of the world (financial transactions, age, temperature, etc..).
- Information appears when we work with those numbers and we can find value and meaning, helping us to make informed decisions.
- Knowledge is when data and the information turn into a set of rules to assist the decisions.
 In fact, we cannot store knowledge because it implies theoretical or practical understanding of a subject.

It is fundamental to understand what type of problems need to be addressed within that specific industry and identify ways in which data can be leveraged to drive solutions and otherwise unobtainable business insights.



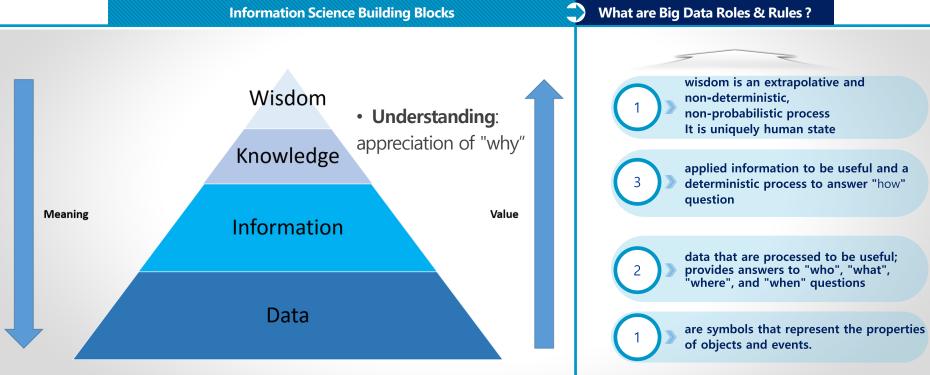








^bBig Data as part of the building blocks











^oBig Data as part of the building blocks

Diferences

- Data "19770524", "Caio Azevedo", "tennis" (raw groups of symbols).
- **Information** 1977-05-24 is a Date, Caio is a Person, tennis is a sport (<u>data</u> in context).
- Knowledge 1977-05-24 is Caio's date of birth, and he likes tennis (linking and transforming information).
- Understanding Caio's birthday is on May 24th. If he likes tennis then he probably plays it.
- Wisdom Buy Caio a racket for his birthday and he'll be very happy.



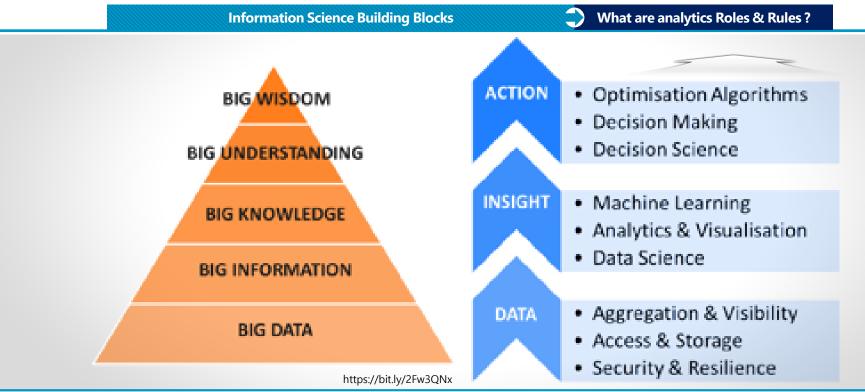








^oBig Data as part of the building blocks



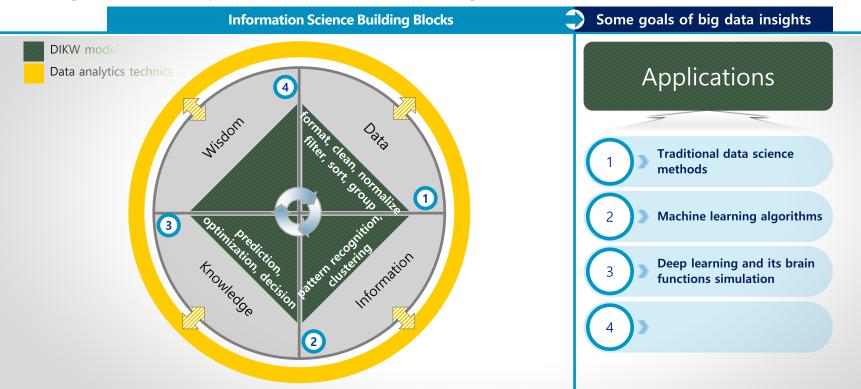








Big Data as part of the building blocks











| What is not big data

- 1) Big Data Buzzwords
- 2) Business Intelligence
- 3) Artificial Intelligence
- 4) Machine Learning
- 5) Deep Learning

DA USP

1. Big Data Analytics Big data buzzwords



IV. What is not Big Data

2. Big Data Analytics ^b Business Intelligence



Diferences

- Business analytics reports-oriented
- Requires an experienced analyst to define database requests and data profile understanding
- Make use of data mining, DW and Dashboards
- BI tells us what was and what is
- Business insights are mathematics models oriented
- The specialist complements model analysis
- IA tells us what to do

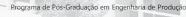
What they have in common?

Big Data as Raw Material

Business Intelligence



Inteligência Artificial





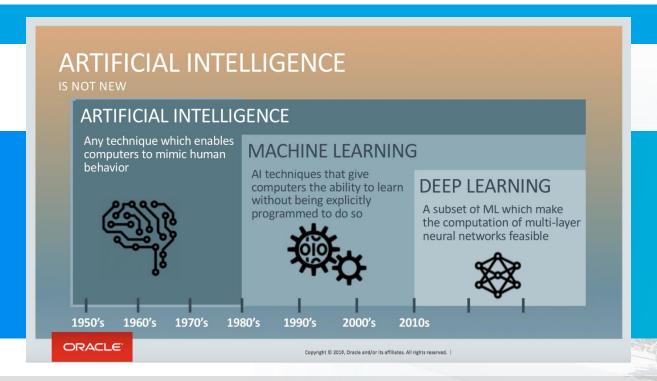






IV. What is not Big Data

3. Big Data Analytics Artificial Intelligence



Artificial Intelligence

- It is a process simulation of human intelligence performed by machines or computer systems
- It is supported on math, statistics and probability, logic and philosophy, linguistics, neuroscience and decision theory.
- Robotic, machine learning, deep learning, computational vision and natural language processing are subsets of IA.





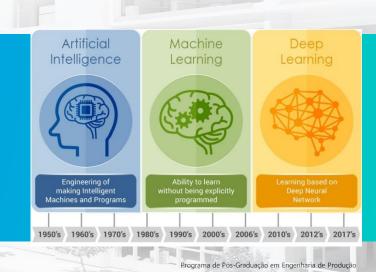




4. Big Data Analytics Machine Learning

Machine Learning

- The science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.
- Different types of ML algorithms, grouped by either learning style (i.e. supervised, unsupervised or semisupervised learning) or by similarity in form or function (i.e. classification, regression, decision tree, clustering, deep learning, etc.).

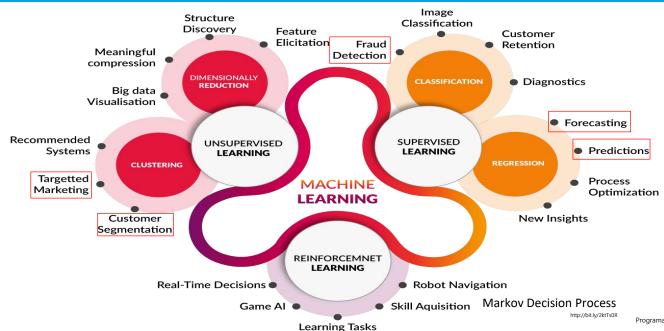








4. Big Data Analytics Machine Learning



https://bit.ly/2PzoHFa https://bit.ly/2WlRe0Z https://bit.ly/2OexcoQ









5. Big Data AnalyticsDeep Learning



Artificial Intelligence

Machine Learning

Deep Learning

The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to yest amounts of data.

A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning

Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

Deep Learning

- The subset of ML composed of algorithms that permit software train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.
- Discovers complex structures in large data sets using back-propagation algorithm to indicate how a system must change its internal parameters used to calculate the representation at each layer of the representation in the previous layer.











5. Big Data AnalyticsDeep Learning (how to)

Input	output
0	0
1	2
2	4
3	6
4	8

output = 2 x input

Input (x)	Actual Output (<i>W</i>)	Output (y)
0	0	0
1	3	2
2	6	4
3	9	6
4	12	8

Random initialization

y = **W**.**x** (*weights*)

Ex: y=3.x. y=0,5.x or y=5.x.

Х	Υ	<i>W</i> =3	rmse(3)	W=3.0001	rmse	
0	0	0	0	0	0	
1	2	3	1 (3-2)^ 2	3.0001	1.0002	
2	4	6	4	6.0002	4.0008	
3	6	9	9	9.0003	9.0018	
4	8	12	16	12.004	16.0032	
Tot	al		30	-	30.006	

rmse: root mean square error δ W=0.0001 (differentiation) Error rate **=0.006/0.0001=60x**

https://omar-florez.github.io/scratch_mlp/ https://medium.com/datathings/neural-networks-and-backpropagation-explained-in-a-simple-way-f540a3611f5e

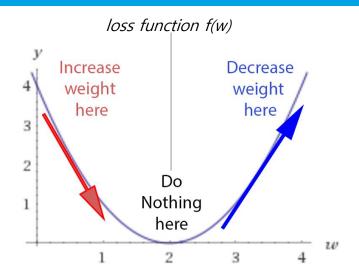






IV. What is not Big Data

5. Big Data Analytics Deep Learning (how to)



If we initialize randomly the network, we are putting any random point on this curve (let's say w=3). The learning process is actually saying this:

- Let's check the derivative.
- If it is positive, meaning the error increases if we increase the weights, then we should decrease the weight.
- If it's negative, meaning the error decreases if we increase the weights, then we should increase the weight.
- If it's 0, we do nothing, we reach our stable point.

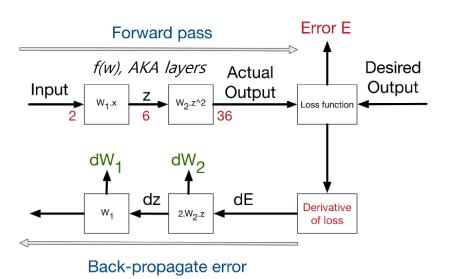
https://omar-florez.github.io/scratch_mlp/ https://medium.com/datathings/neural-networks-and-backpropagation-explained-in-a-simple-way-f540a3611f5e





□□□□□□□ IV. What is not Big Data

5. Big Data AnalyticsDeep Learning (how to)



Weight update

The learning rate is introduced as a constant (usually very small), in order to force the weight to get updated very smoothly and slowly (to avoid big steps and chaotic behavior).

New weight = old weight-Derivative Rate (dW) * learning rate

In order to validate this equation:

- dW> 0, an increase in weight will increase the error, thus the new weight should be smaller.
- dW < 0, it means that an increase in weight will decrease the error, thus we need to increase the weights.
- dW=0, it means that we are in a stable minimum. Thus, no update on the weights is needed -> we reached a stable state.



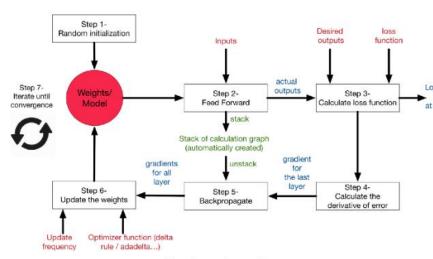






□□□□□□□ IV. What is not Big Data

5. Big Data AnalyticsDeep Learning (how to)



Neural networks step-by-step

How many iterations are needed to converge?

- This depends on how **strong the learning rate** we are applying. High learning rate means faster learning, but with higher chance of instability.
- Loss (error) It depends as well on the **meta-parameters** of the network metric (how many layers, how complex the non-linear functions are). The more it has variables the more it takes time to converge, but the higher precision it can reach.
 - It depends on the **optimization method** use, some weight updates rule are proven to be faster than others.
 - It depends on the **random initialization** of the network. Maybe with some luck you will initialize the network with **W=1.99** and you are only one step away from the optimal solution.
 - It depends on the **quality** of the training set. If the input and output has no correlation between each other, the neural network will not do magic and can't learn a random correlation.







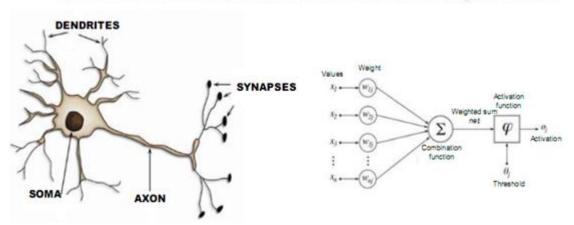


5. Big Data Analytics

Deep Learning (Math)



An artificial neuron is a mathematical model of a biological neuron



1. The feed forward algorithm...

$$n_l = S\left[\sum_{l=1} (w_l i_{l-1})\right]$$

Where **n** is a **neuron** on *layer I*, and **w** is the **weight** value on layer I, and **i** is the value on **I-1** layer. All input values are set as the first layer of neurons. Then, each neuron on the following layer s takes the sum of all the neurons on the previous layer multiplied by the weights that connect them to the relevant neuron on that following layer. This summed value is then activated.

https://medium.com/@ODSC/5-essential-neural-network-algorithms-9336093fdf56



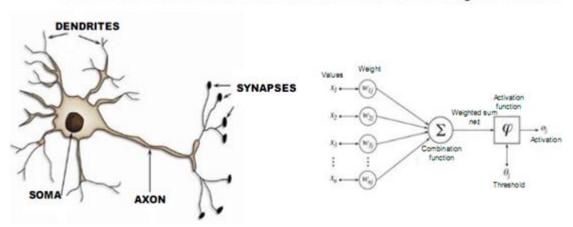






5. Big Data AnalyticsDeep Learning (Math)

An artificial neuron is a mathematical model of a biological neuron



2. A common activation algorithm: Sigmoid...

$$S(t) = \frac{1}{1+e^{-t}}$$

Normalize input values to a proportional value bet ween 0 and 1 converting a value to a probability, which reflects a neuron's weight or co nfidence. This introduces nonlinearity to a model, allowing it to pick up on observations with greater insight.

https://medium.com/@ODSC/5-essential-neural-network-algorithms-9336093fdf56



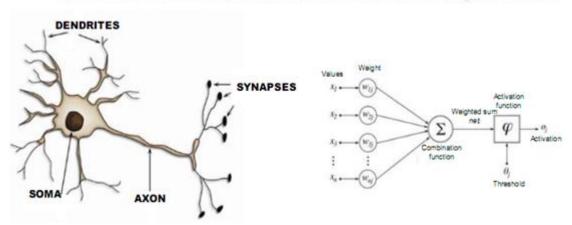






5. Big Data AnalyticsDeep Learning (Math)

An artificial neuron is a mathematical model of a biological neuron



3. The cost function...

$$(E = \frac{1}{2}(n_L - t)^2)$$

The squared cost function lets you find the error by calculating the difference between the ou tput values and target values. The target/desired values could be a binary vector for classification.

https://medium.com/@ODSC/5-essential-neural-network-algorithms-9336093fdf56





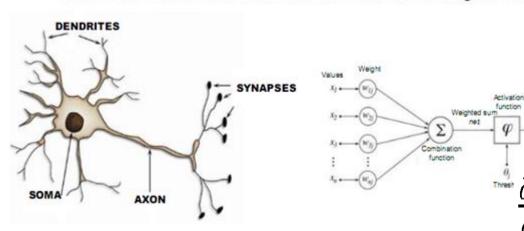




□□□□□□□ IV. What is not Big Data

5. Big Data AnalyticsDeep Learning (Math)

An artificial neuron is a mathematical model of a biological neuron



4. The back propagation...

The error from the cost function is then passed back by being multiplied by the derivative of the sigmo id function S'.

$$\delta_L = (\Delta E_{n_L} * S'(n_L))$$

Recursive accumulation contributed to the error (unique neuron). Past weight values must be transposed to fit the following layer

$$\frac{\partial E}{\partial n} = \delta_L = [T(w_{l+1}) * \delta_{l+1} * n_L (1 - n_L)]$$

https://medium.com/@ODSC/5-essential-neural-network-algorithms-9336093fdf56



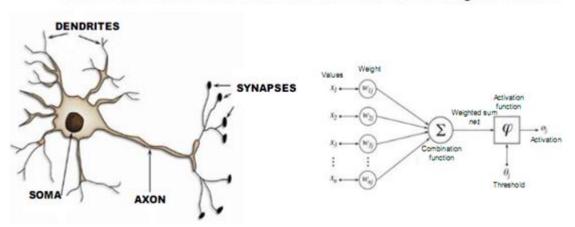






5. Big Data AnalyticsDeep Learning (Math)

An artificial neuron is a mathematical model of a biological neuron



5. Applying the learning rate/weight updating...

The change now needs to be used to adapt the weight value. The eta represents the learning rate:

$$w = w - \left(\eta * \frac{\partial E}{\partial w}\right)$$

https://medium.com/@ODSC/5-essential-neural-network-algorithms-9336093fdf56









6. Big Data Analytics⁶ Applications

Medical

- •Skin cancer identification
- FDA approval for death prediction
- Radiology
- Predict disease from patient

Agriculture

- Identify Plant Pests
- Create more efficient seeds
- Monitor crops i real time
- •Identify soil defects & nutrients

Pharma

- Design drugs
- Bioinformatics
- Predict the chemical reactions between candidate compounds and target molecules
- •Identify one or more genes responsible for a disease

Autonomous Vehicles

- Map raw pixels from camera directly to steering commands
- Drive in unstructured conditions
- •Car and lane detection
- Motion control & planning
- Optimization of AV traffic

Data Centers

- Data center security
- Reduce electricity usage
- Server optimization

Source: https://semiengineering.com/deep-learning-spreads/









V Hands-on

- 1) Finding employees address list
- 2) "Bolsa Família" Case
- 3) People Analytics profile mapping & evaluation
- 4) Predictions Flight passengers and Titanic deaths















1. Real life situations Big Data HAL



Data analytics cycle

Scope definition



Activities

▶ Raw Data collection

 Business team identify under which scenarios are going to be submitted to analysis and/or pattern identification

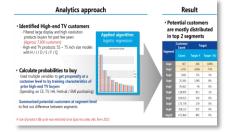
Data preparation



▶ Meaningful Dataset

- Raw extraction from datasources
- Dataset creation considering relevant parameters

HAL analytics



▶ Patters identification

- Dataset submission to HAL algorithms
- Results analysis
- Patterns identification

Result

Data Result interpretation to Decision make



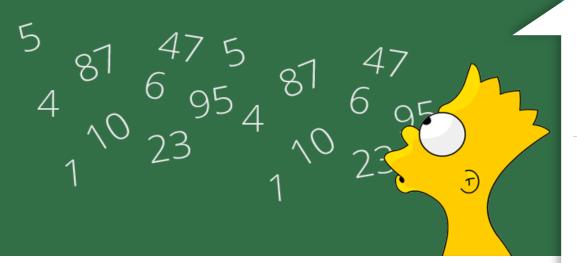






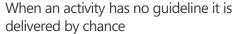
Auditing







Randomness Random Investigation



Professional experience empiric process



Auditor has fundamental rule being the main guide of the auditing process

Programa de Pós-Graduação em Engenharia de Produção

V. Hands-on











Auditing











V. Hands-on

^bAuditing

"Bolsa Familia" Auditing











V. Hands-on

People Analytics

"Betting only on data and algorithms is very reductionist. They will not sol ve all the problems but will make the skills that have more nuances more valuable such as creativity, problem solving, compassion and empathy for the other." - Susan David author of "Emotional Agility" about humans adva ntages over the machines



Team-members commitment

https://bit.lv/2EceZ5Q









1. Real life situations People Analytics





Mapping your team-members





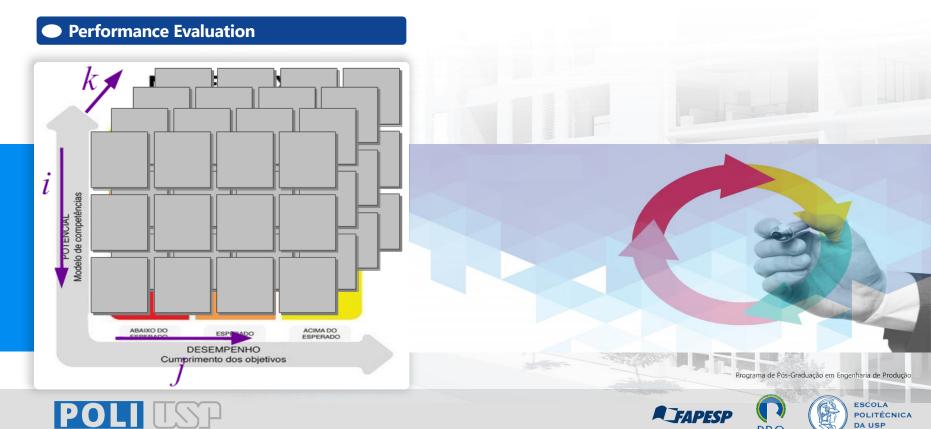






V. Hands-on

1. Real life situations • People Analytics



Classification

Suppliers Classification

ID	Fornecedor	Pontuais	Atrasados	% de atrasos	Média dias atraso	Desvio Padrão (atraso)
1	OLIGAM INDUSTRIA E COMERCIO EIRELI - ME	6	84	93	34	30
2	EXPRESSA DISTRIBUIDORA DE MEDICAMENTOS LTDA	49	128	72	30	35
3	Johnson & Johnson	136	309	69	18	31
4	CM HOSPITALAR	21	62	75	25	20
5	COLOPLAST DO BRASIL Itda	26	52	67	26	28
6	SCHARLAB BRASIL MATERIAL PARA LABORAT	30	58	66	24	32
7	E.J.A Drogaria e Perfumaria Itda - epp	56	63	53	29	27
8	DUPATRI HOSP. COM. IMPORTA	126	257	67	15	20
9	CIRÚRGICA SANTA CRUZ	0	24	100	39	24
10	VIX COM	32	43	57	27	16
11	AGLON COMERCIO E REPRESENTA	25	58	70	18	20
12	Accord Farmaceutica Ltda	11	25	69	42	36
13	SANOFI-AVENTIS FARMAC	74	62	46	28	33
14	CRISMED COMERCIAL HOSPITALAR LTDA.	133	178	57	14	13
15	FARMA VISION DISTRIBUIDORA DE MEDICAMENTOS LTDA -	85	80	48	22	22
16	COMERCIAL CIRURGICA RIOCLARENSE LTDA	275	279	50	15	18
17	ANBIOTON IMPORTADORA LTDA	23	30	57	29	20













Predictive



Titanic Predictive



https://www.kaggle.com/francksylla/titanic-machine-learning-from-disaster



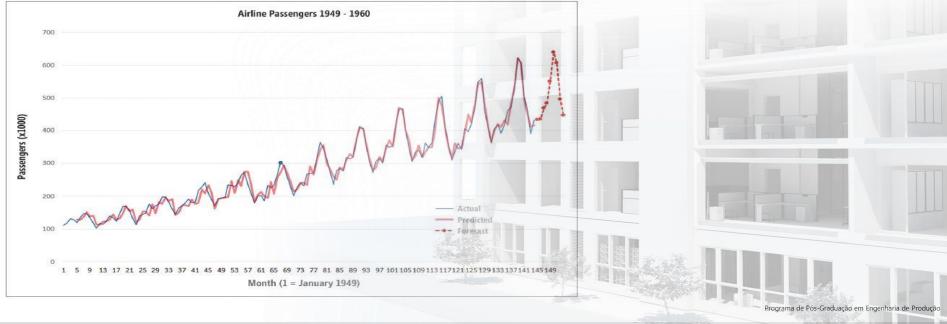






• Predictive

Airline passengers prediction









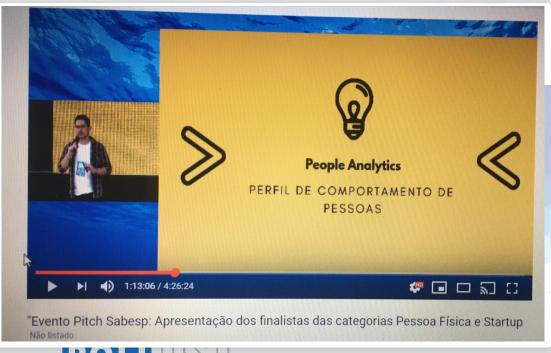


V. Hands-on



1. Real life situations People Analytics & Auditing

People behavior for fraud identification







1. References

VI. References

Harari Y. (2015). Sapiens a brief history of humankind. Harper

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data conce pts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Xuemei Tian, (2017) "Big data and knowledge management: a case of déjà vu or back to the future?", Journal of Knowledge Management, Vol. 21 Issue: 1, pp.113-131.

Storey, V. C., & Song, I. Y. (2017). Big data technologies and manag ement: What conceptual modeling can do. *Data & Knowledge Engineering*, *108*, 50-67.

Wen, J., Li S., Lin Z., Hu Y., Huang C. (2011). Systematic literature review of machine learning based software development effort estimation models

Sammut C., Webb G. (2017) Encyclopedia of Machine Learning and Data Mining, 2nd Edition.

ACKOFF, R. L., Ackoff's Best. New York: John Wiley & Sons, pp 170 – 172, 1999.

nttp://damielhulme.blogspot.com/2012/10/in-beginning-bigbuzz-was-bigdata.html nttps://www.quest.com/community/quest/database-management/b/database-management-blog/posts/data-informa

http://datascienceacademy.com.br/blog/10-carreiras-em-big-data-e-data-science/

https://www.datamation.com/big-data/big-data-technologies.html

nttp://datascienceacademy.com.br/blog/cientista-de-dados-por-onde-comecar-em-8-passo

http://www.bigdatabusiness.com.br/cientista-de-dados-que-profissao-e-essa-2/

https://ec.europa.eu/digital-single-market/en/what-big-data-can-do-you

https://emerj.com/ai-glossary-terms/what-is-machine-learning/

https://www.orgilly.com/ideas/machine-learning-a-quick-and-simple-definition

http://www.bigdatabosiness.com.br/o-que-o-big-data-muda-na-realidade-dos-profissionais-de-marketing/

https://extra.globo.com/noticias/educacao/profissoes-de-sucesso/profissionals-de-big-data-estao-entre-os-cinco-contratados-no-brasil-22098050.html

https://selecthub.com/business-intelligence/business-intelligence-vs-business-analytics/

https://blog.eduonix.com/artificial-intelligence-vs-business-intelligence









Thanks

caio.aze@usp.br

From data to widows

What it can do for you?

Powered by HAL 9000 your Big Data solution

POLIUST





INSTITUTO DE **PESQUISAS TECNOLÓGICAS**







