



INTRODUCTION TO BIG DATA

Fábio Nogueira

BRIEF CURRICULUM

Prof. Fábio Nogueira

Doctorate in Computer Science (UFPE)

Graduated in Electrical Engineering (UFCG)

Analyst at the Central Bank of Brazil

fabio.nogueira.souza@gmail.com

Areas:

- Distributed Systems
- Middleware
- Component and Service Orientation
- Autonomic Computing

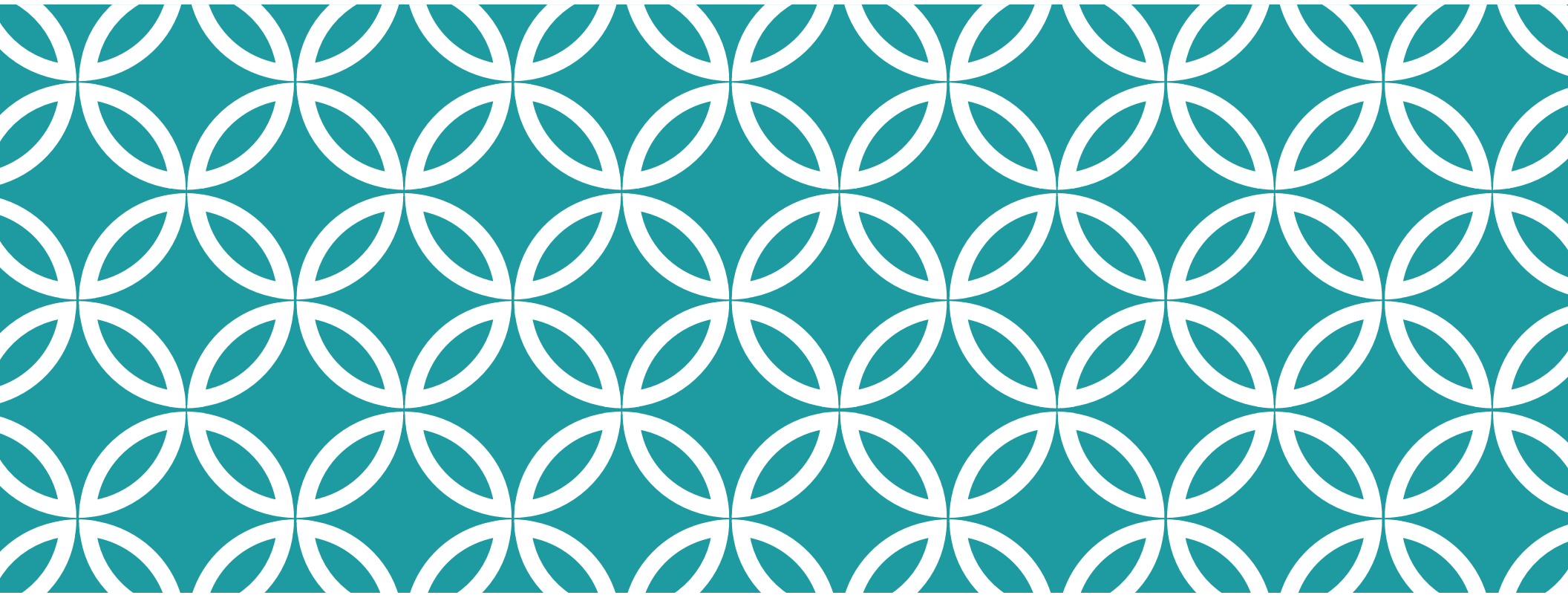
AGENDA

Big Data

Data Analytics in Big Data

Apache Hadoop

Apache Spark



BIG DATA |

CONTEXT

“The data-driven world will be **always on, always tracking, always monitoring, always listening and **always watching** – **because it will be always learning.**”**

Data Age 2025: The Digitalization of the World

CONTEXT

NSA scandal: what data is being monitored and how does it work?

Everything you need to know about data gathering from internet companies by the US National Security Agency

● **NSA has direct access to Google, Facebook and Apple**

What is the scandal?

The US's National Security Agency (NSA), its wiretapping agency, has been monitoring communications between the US and foreign nationals over the internet for a number of years, under a project called Prism. Some of the biggest internet companies, from Apple to Google to Yahoo, are involved. The US government confirmed the existence of the scheme and its application on Thursday night.

What data is being monitored?

Potentially, everything. The PowerPoint slide about Prism says it can collect "email, chat (video, voice), videos, photos, stored data, VoIP [internet phone calls], file transfers, video conferencing, notifications of target activity - logins etc, online social networking details" and another category called "special requests".

<https://bit.ly/2N8CwIM>

DESCUBRA A REDE IT TRENDS

COMPUTERWORLD
FROM IDG

CarreiraNegóciosSegurançaInovaçãoPlataformasSegurançaEstrat

Home > Segurança

Facebook pagará US\$ 644 mil ao Reino Unido em caso da Cambridge Analytica

Valor da multa é o teto máximo que o governo do Grã-Bretanha pode aplicar em caso de violação da lei de proteção de dados

Da Redação
01/11/2019 às 16h30

Em março de 2018, veio a público que a consultoria política havia tido acesso a dados pessoais de 87 milhões de usuários da rede social, em um dos maiores escândalos de privacidade que atingiu o Facebook. Nos meses que sucederam, **Mark Zuckerberg, CEO do Facebook**, enfrentou questionamentos de legisladores dos EUA e da União Europeia sobre como a Cambridge Analytica obteve acesso a tais dados.

<https://bit.ly/2PFUFiX>

WHAT IS BIG DATA?

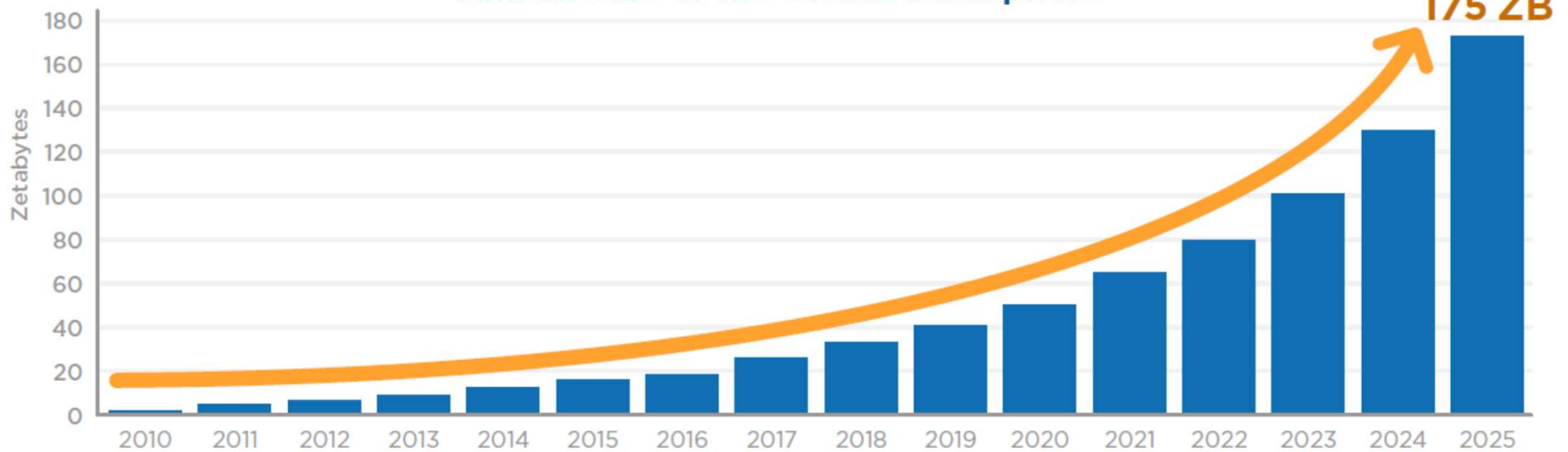
Laney's 3Vs model [1]

- Volume
- Velocity
- Variety

VOLUME

DIGITAL AGE 2025

Annual Size of the Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

<https://bit.ly/2pAgRjN>

VOLUME

Measuring data

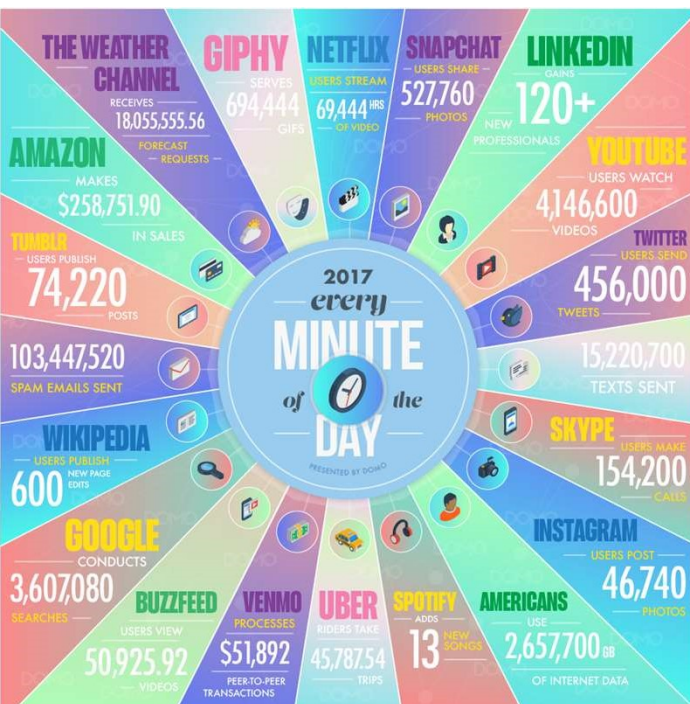
Abbreviation	Unit	Value	Size (in bytes)
b	bit	0 or 1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	1,000 bytes	1,000 bytes
MB	megabyte	1,000 ² bytes	1,000,000 bytes
GB	gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB	terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB	petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB	exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

VELOCITY

DATA NEVER SLEEPS REPORT

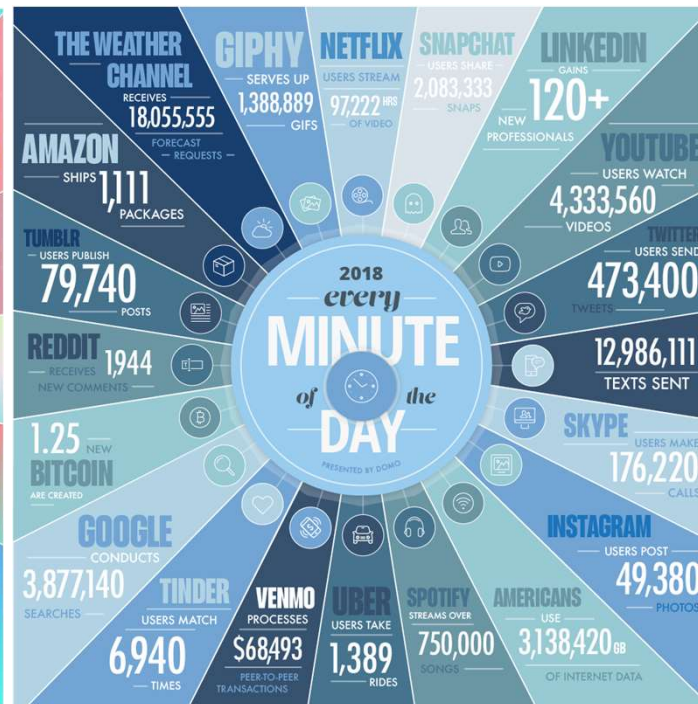
2017

<http://bit.ly/2lw0mD9>



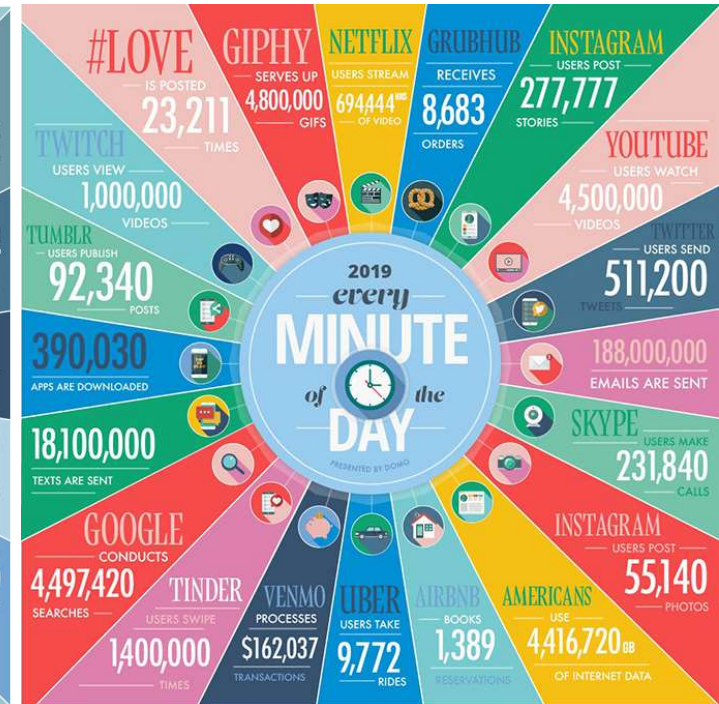
2018

<http://bit.ly/2ISP37Z>



2019

<http://bit.ly/2IVgDI2>



VARIETY

Data formats

Non-aligned data structures

Inconsistent data semantics



<https://bit.ly/36lDvgi>

4VS MODEL

“...big data technologies describe a new generation of technologies and architectures, designed to economically extract **value** from very large **volumes** of a wide **variety** of data, by enabling the high-**velocity** capture, discovery, and/or analysis”

- Extracting Value from Chaos (IDC)

5VS MODEL

“...dealing effectively with *Big Data* requires one to create **value** against the **volume**, **variety** and **veracity** of data while it is still in motion (**velocity**), not just after it is at rest”

- Understanding big data: Analytics for enterprise class hadoop and streaming data

“TO INFINITY AND BEYOND!”

Vision

Verification

Validation

Vocabulary

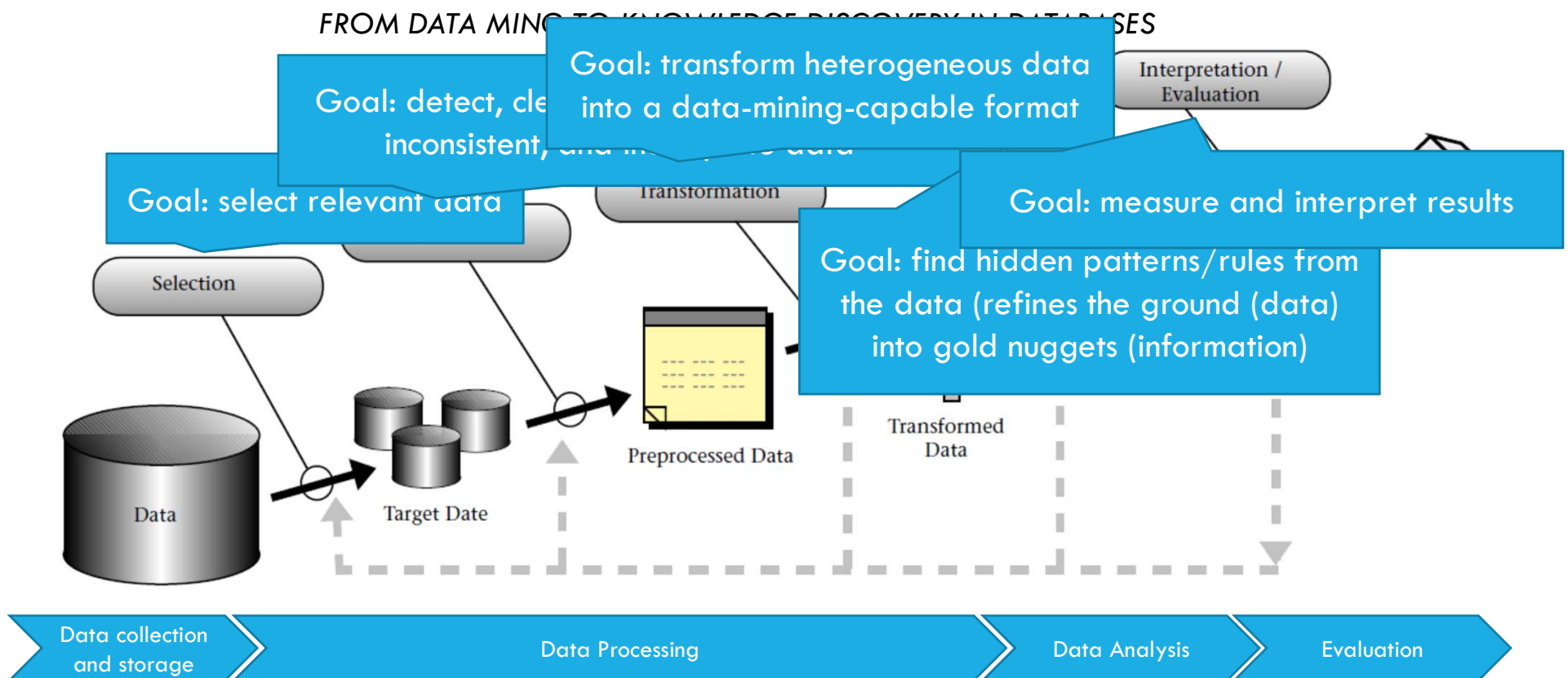
Vagueness

...

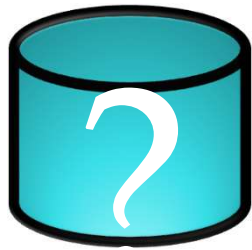


BIG DATA ⇒ BIG CHALLENGES

KNOWLEDGE DISCOVERY IN DATABASES (KDD)



BUT, BIG DATA = BIG CHALLENGES



DATA STORAGE



Scale Up



DATA PROCESSING



DATA ANALISYS



QUESTIONS???

