

# **O que é a ciência de dados** *(data science)*. Discussão do conceito

Luís Borges Gouveia

Universidade Fernando Pessoa

Versão 1.3, Outubro, 2015

# Nota prévia

- Esta apresentação tem por objetivo, proporcionar uma introdução sobre que é e como se enquadra a ciência de dados
- Para o efeito, são utilizados diversos gráficos e imagens retiradas na *World Wide Web* de diferentes atores associadas com a prática desta área e a quem é realizada referência.
- No entanto, a estrutura, sequência e o suporte das imagens, representam uma linha de pensamento que é independente da origem dessas mesmas imagens e que pretende ser ilustrada por elas e orientar o aprofundamento dos temas

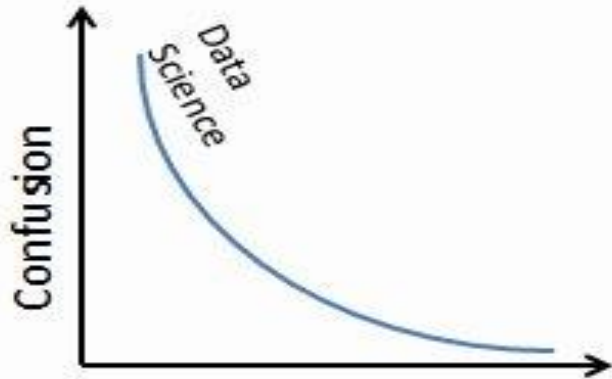
# *Data science* ou a ciência de dados

- Considera o estudo da origem da informação, o que representa e como pode ser transformada numa fonte valiosa para a criação de negócio e de estratégias para o contexto em análise
- A exploração de quantidades massivas de dados estruturados e não estruturados para identificar padrões que podem ajudar uma organização no controle de custos, aumento de eficiência, reconhecimento e descoberta de novos mercados e oportunidades e aumento de vantagem competitiva
- Transformação de dados disponíveis em informação, com recurso a técnicas de análise de dados, experiência, mas também inteligência e criatividade
- É a extração de conhecimento a partir de grandes conjuntos de dados, com recurso a métodos científicos

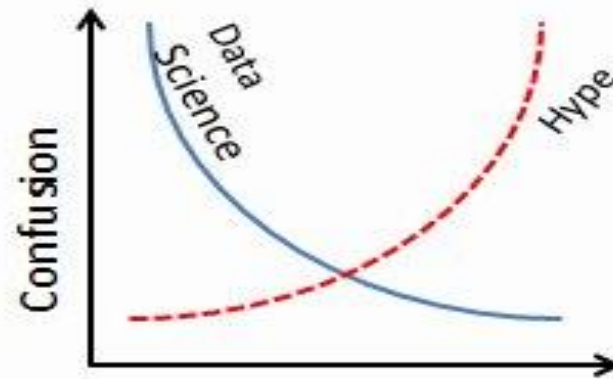
# Método científico?

- Colocar uma questão
- Colocar uma hipótese
- Traçar um plano para a comprovar
- Elaborar um contexto de observação
- Observar e experimentar
- Registrar os resultados
- Analisar os resultados
- Chegar a uma conclusão

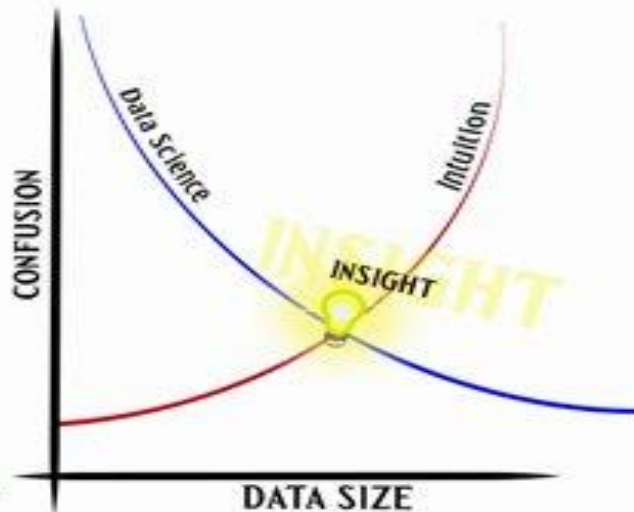
# Numa perspectiva mais operacional



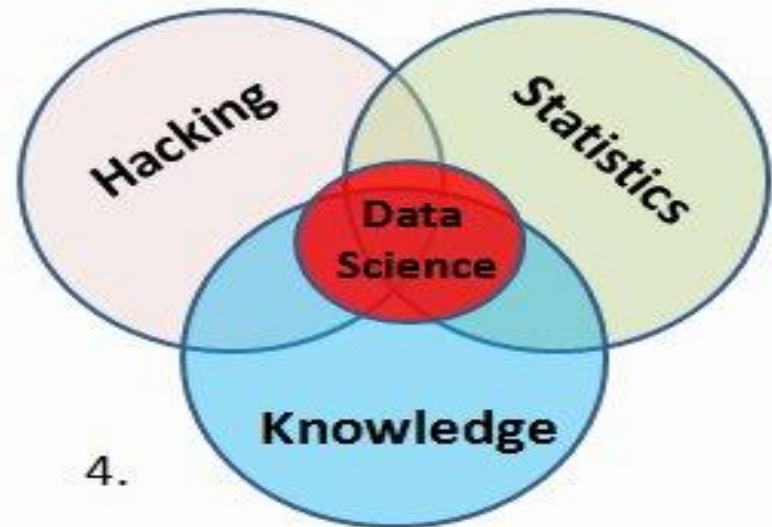
1. Data Size



2. Data Size



3.



4.

# O perfil multidisciplinar do profissional da ciência de dados

## MODERN DATA SCIENTIST

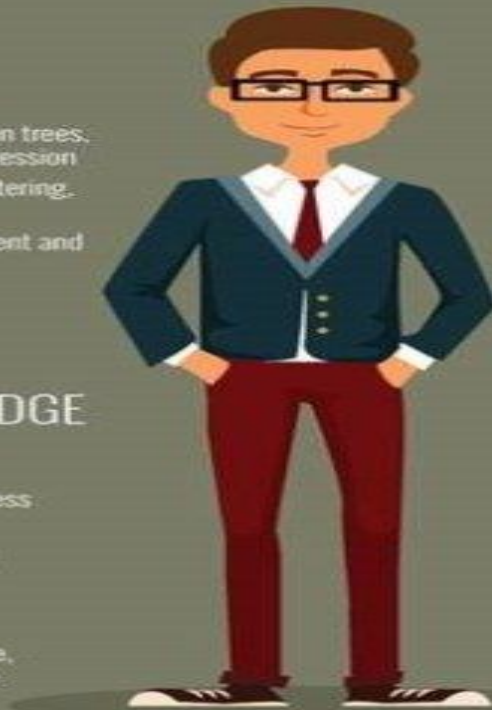
Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard.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

### DOMAIN KNOWLEDGE & SOFT SKILLS

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



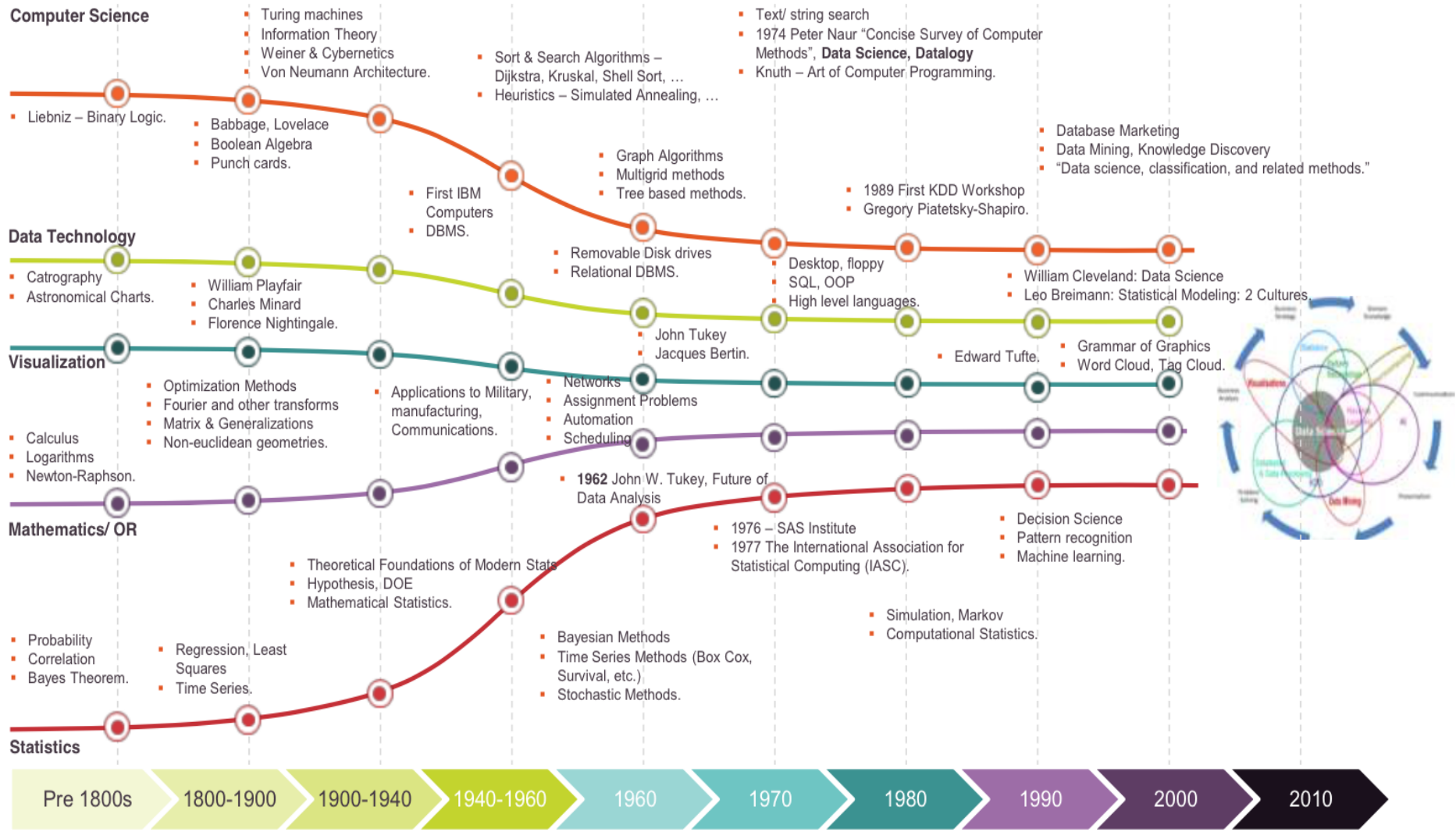
### PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package 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

### 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

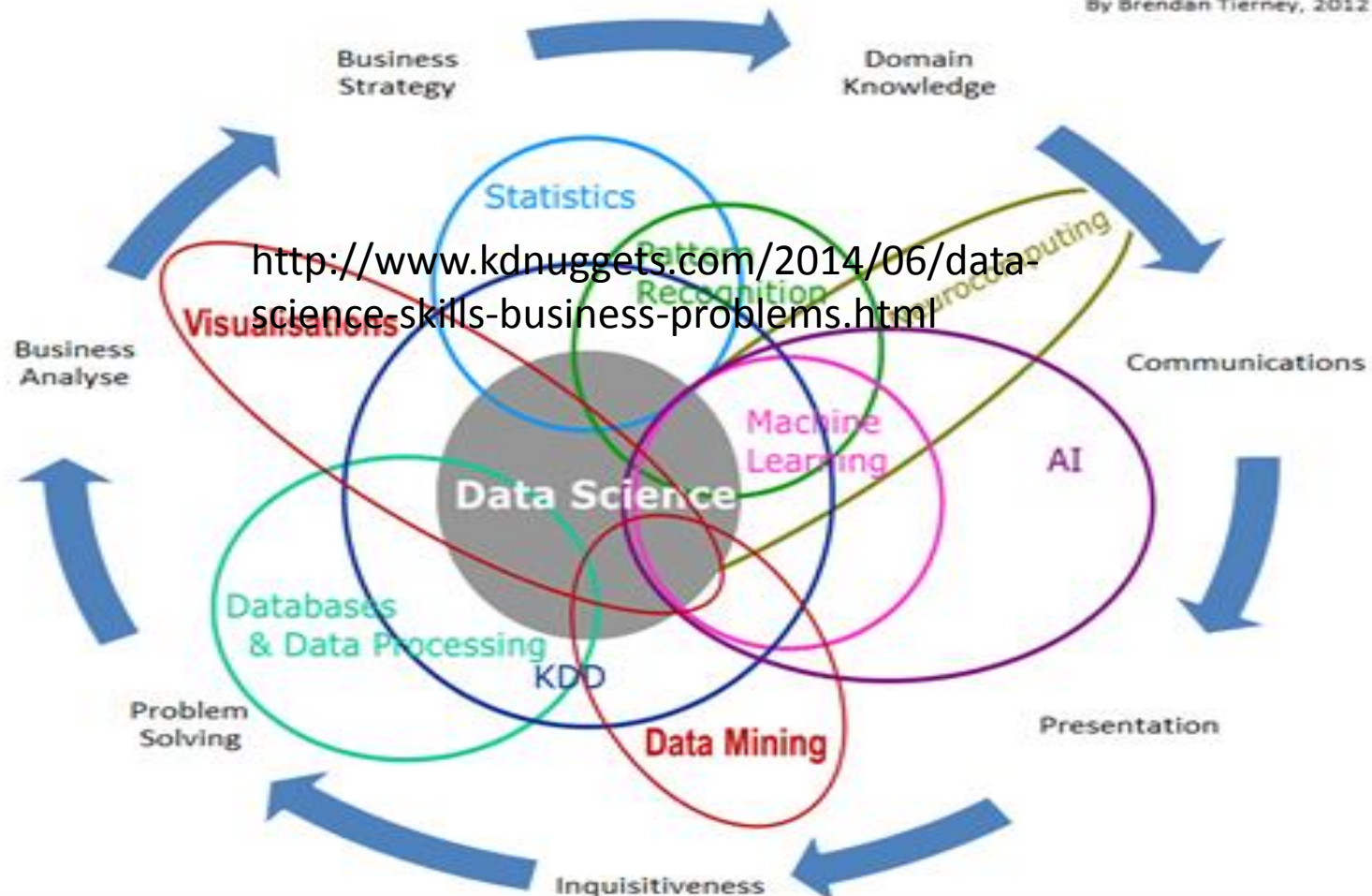
# A história e contribuições para a formação da ciência de dados



# Multidisciplinaridade da ciência de dados

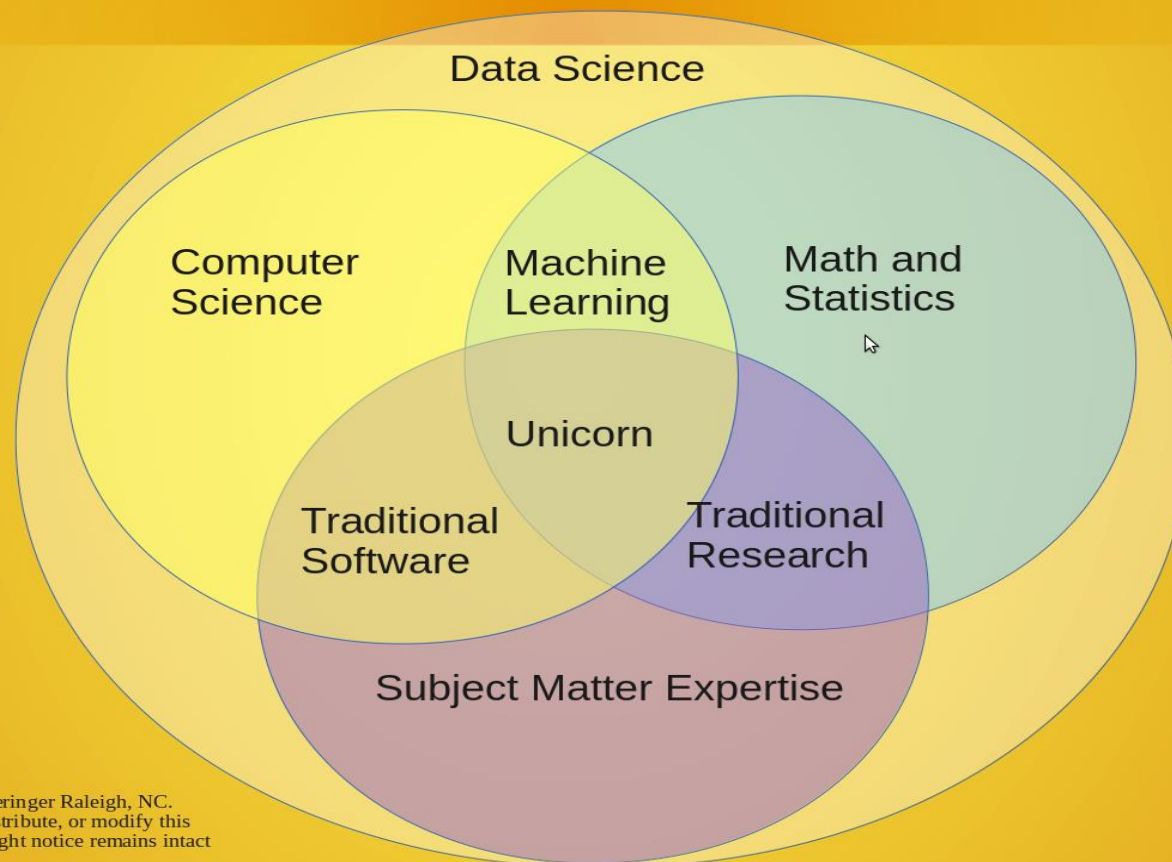
## Data Science Is Multidisciplinary

By Brendan Tierney, 2012



# A ciência de dados face às ciências da computação

## Data Science Venn Diagram v2.0



Copyright © 2014 by Steven Geringer Raleigh, NC.  
Permission is granted to use, distribute, or modify this  
image, provided that this copyright notice remains intact

# Peso económico e relevância da análise de dados

## Exhibit 15

### Sectors differ in their ability to use and obtain value from big data analytics

QUALITATIVE

#### Big data ease of capture

Reflects ability to own or access data and analytics

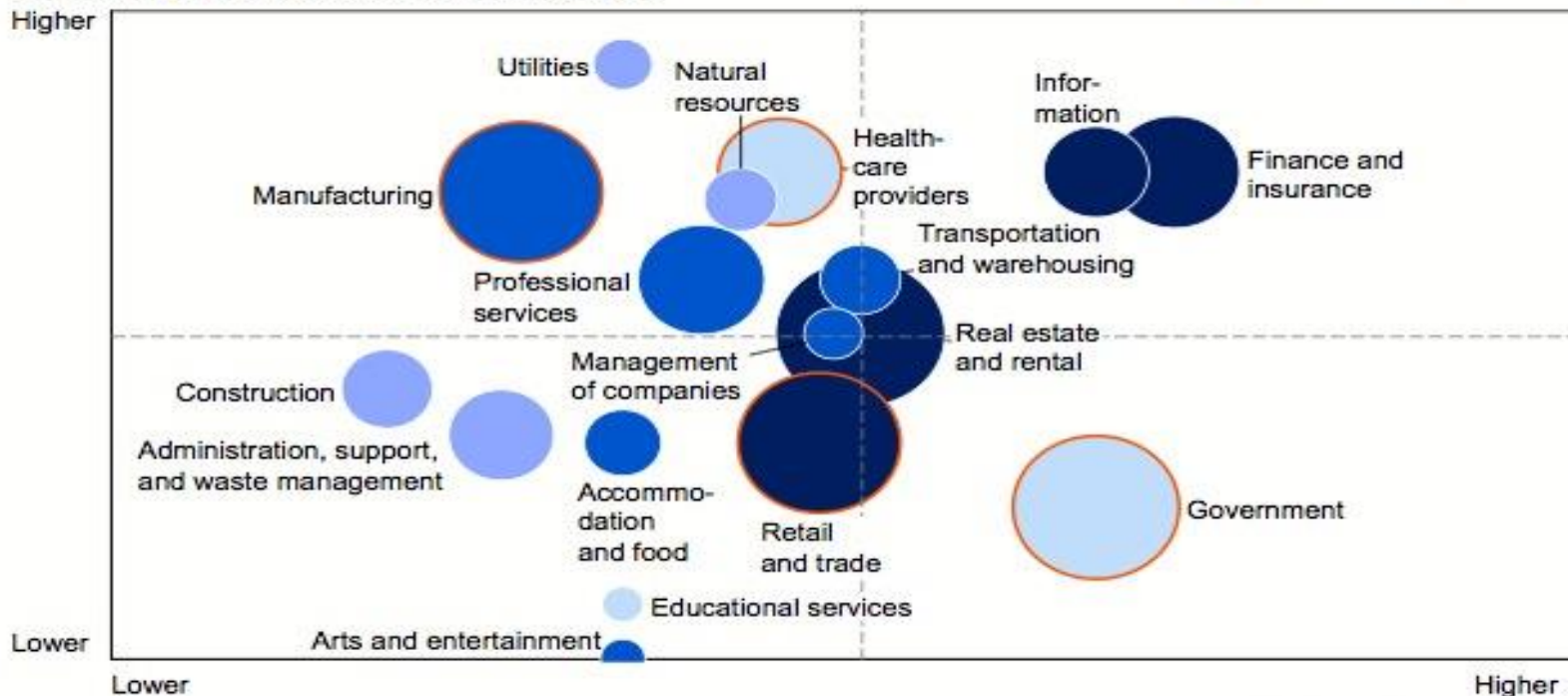
Higher

○ Bubble size = GDP

○ Sectors studied in this report

Competitive intensity to adopt big data

● Highest    ● Moderate  
● High      ● Low



Lower

Lower

Higher

**Big data value potential**

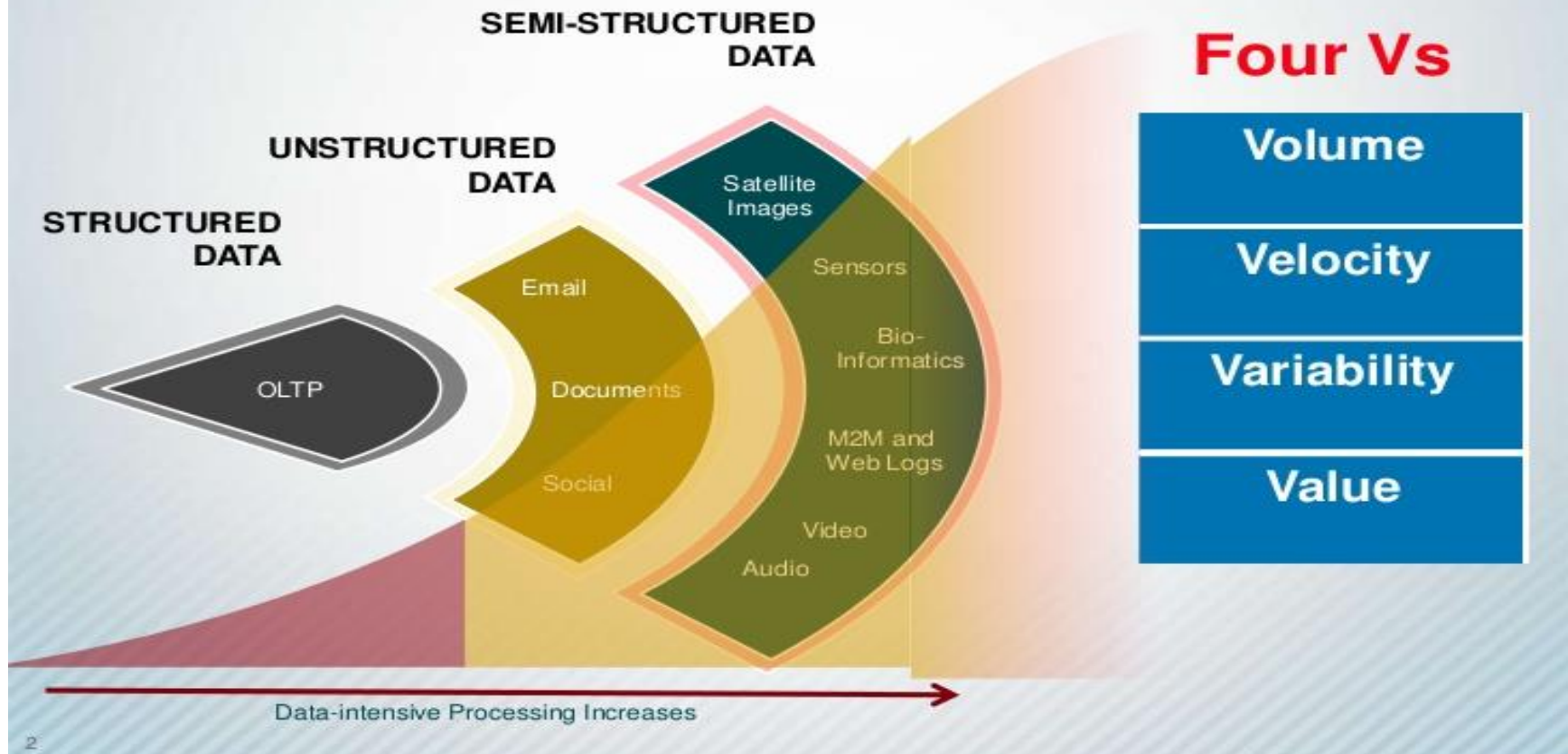
Reflects value of data and/or competitive advantage achieved

SOURCE: US Bureau of Economic Analysis; McKinsey Global Institute analysis

# Uma enorme massa de dados e os desafios associados

BIG DATA IS NOT JUST ABOUT SIZE

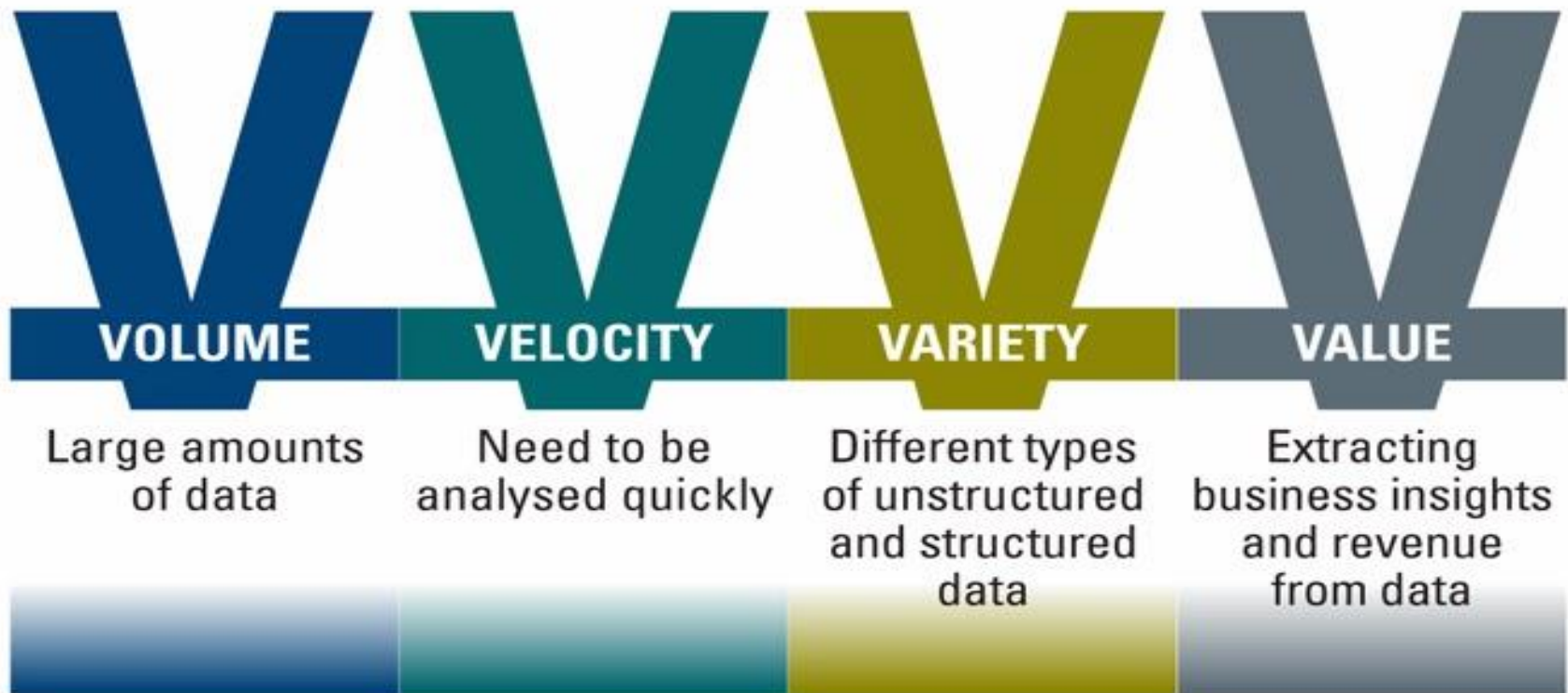
HITACHI  
Inspire the Next



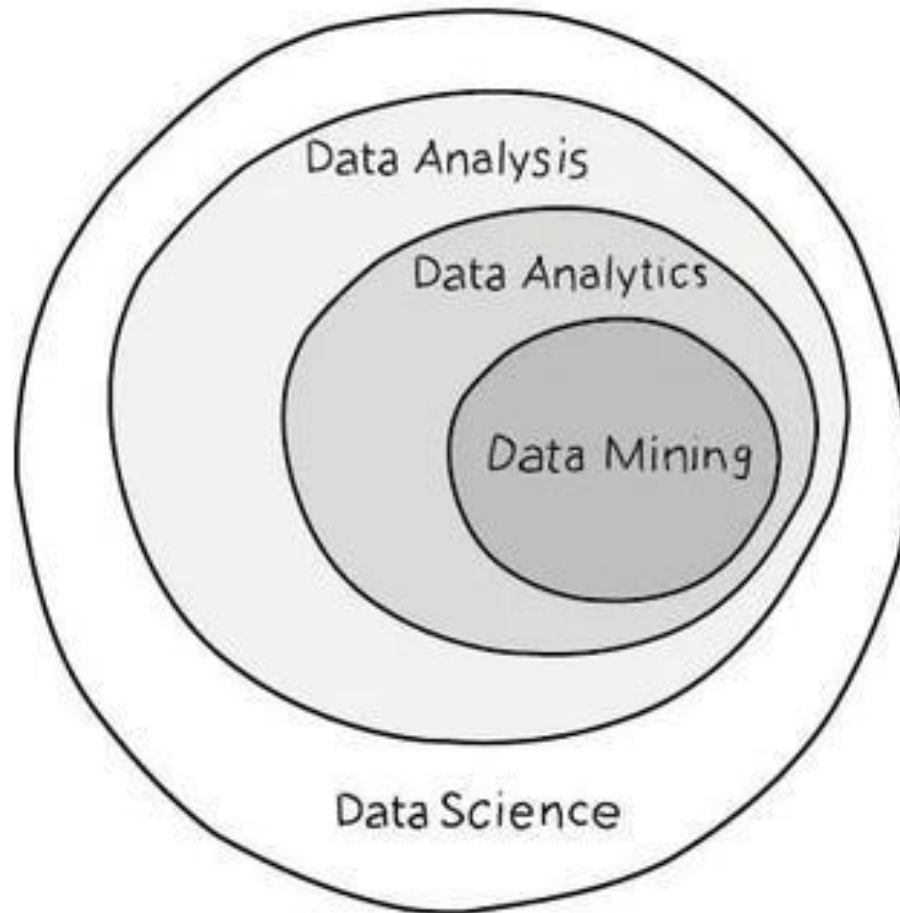
Os dados massivos são mais do que a sua quantidade (como extrair valor, em tempo útil, de um grande volume de dados)

## Big Data: The four Vs

Volume, Velocity, Variety and Value



# A relação entre algumas das disciplinas que emergem do potencial de dados digitais

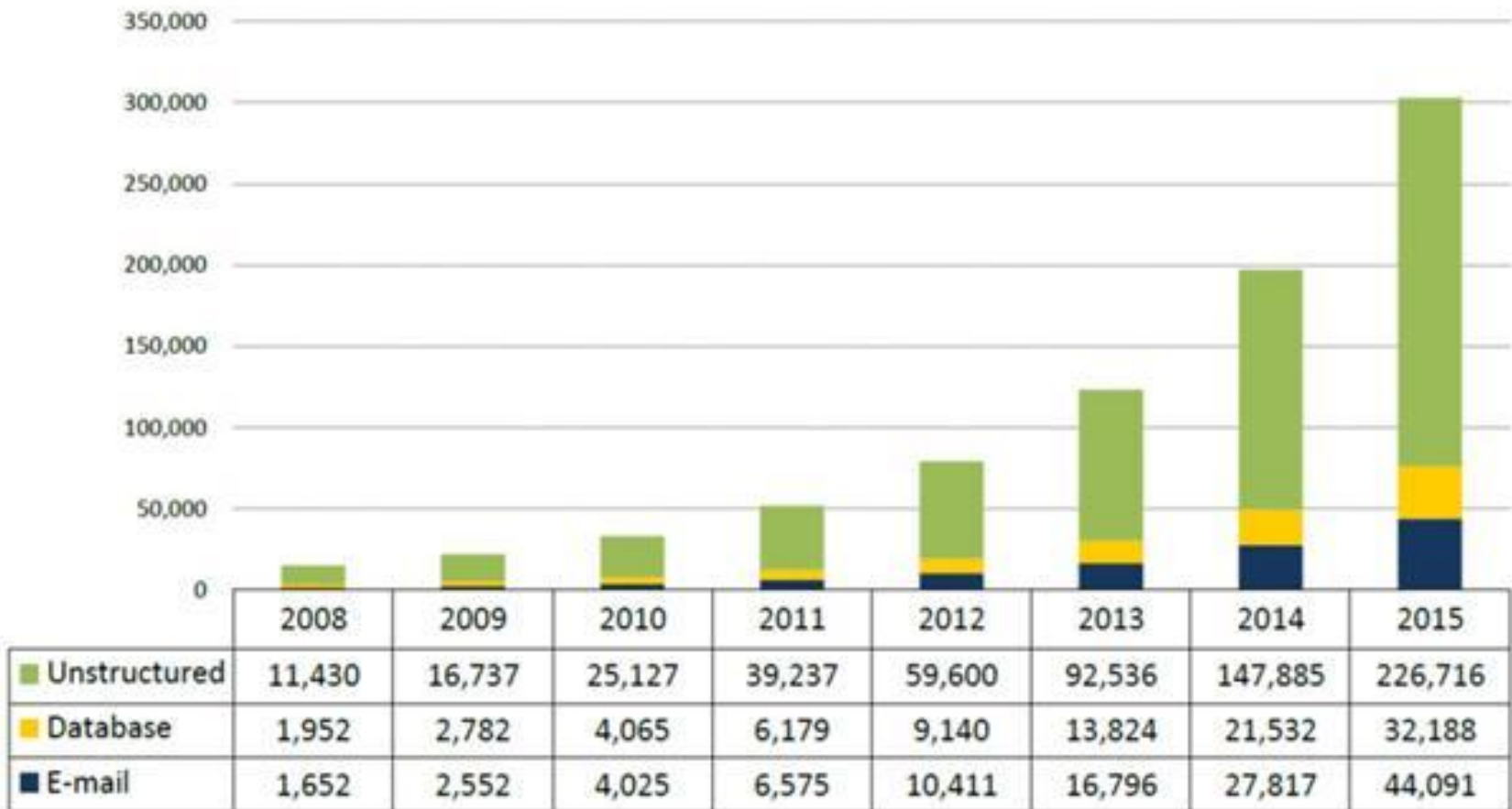


# Diferenciar a inteligência de negócios, da análise de dados e da ciência de dados

DISCIPLINE	TECHNOLOGIES	SKILLS	FOCUS
<b>BUSINESS INTELLIGENCE</b>	<ul style="list-style-type: none"> <li>• ETL Tools / SQL</li> <li>• RDBMS</li> <li>• Reporting</li> <li>• Visualization</li> </ul>	<ul style="list-style-type: none"> <li>• Programming</li> <li>• Data Analysis</li> <li>• Data Modeling</li> <li>• Report Development</li> <li>• Basic Statistics</li> <li>• Technical Architecture</li> <li>• Business Analysis &amp; Strategy</li> <li>• Presentation</li> </ul>	<ul style="list-style-type: none"> <li>• Information Delivery and Reporting</li> <li>• Data Visualization</li> <li>• Descriptive Statistics</li> <li>• Data Integration and Consolidation</li> </ul>
<b>DATA ANALYSIS</b>	<ul style="list-style-type: none"> <li>• Data Modeling Software</li> <li>• Diagramming Software</li> <li>• Documentation Software</li> <li>• SQL</li> <li>• Data Profiling Software</li> </ul>	<ul style="list-style-type: none"> <li>• Data Modeling</li> <li>• Business Analysis</li> <li>• Data Manipulation</li> <li>• Basic statistics</li> </ul>	<ul style="list-style-type: none"> <li>• Business Rules</li> <li>• Data Definitions and Lineage</li> <li>• Data Entity Relationships</li> <li>• Data Attributes</li> <li>• Data Structures</li> <li>• Sources and Targets of Data</li> <li>• Data Quality</li> </ul>
<b>DATA SCIENCE</b>	<ul style="list-style-type: none"> <li>• Statistics Software</li> <li>• Columnar Data</li> <li>• Map-Reduce</li> <li>• NoSQL</li> <li>• Programming Languages</li> <li>• Graphing/Charting Software</li> </ul>	<ul style="list-style-type: none"> <li>• Advanced Statistics</li> <li>• Programming</li> <li>• Business Analysis</li> <li>• Modern Data Management Technologies and Architectures</li> </ul>	<ul style="list-style-type: none"> <li>• Predictive Modeling</li> <li>• Advanced Statistical Analysis</li> <li>• Data Mining</li> <li>• Unstructured Data Management</li> <li>• Large data volumes</li> <li>• Research</li> </ul>

# O exponencial crescimento dos dados, ocorre em várias categorias

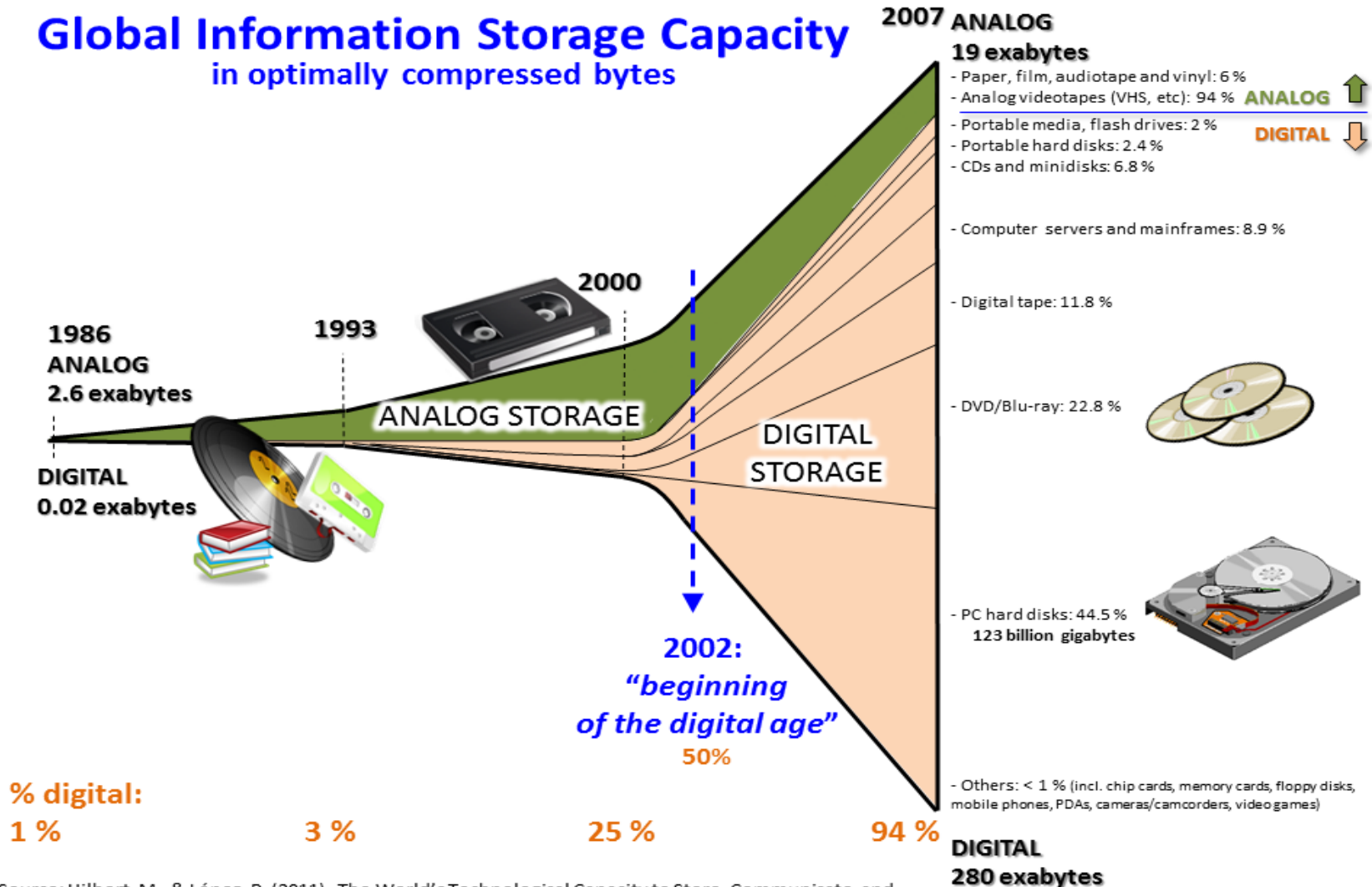
Total Archived Capacity, by Content Type, Worldwide, 2008-2015 (Petabytes)



Source: Enterprise Strategy Group, 2010.

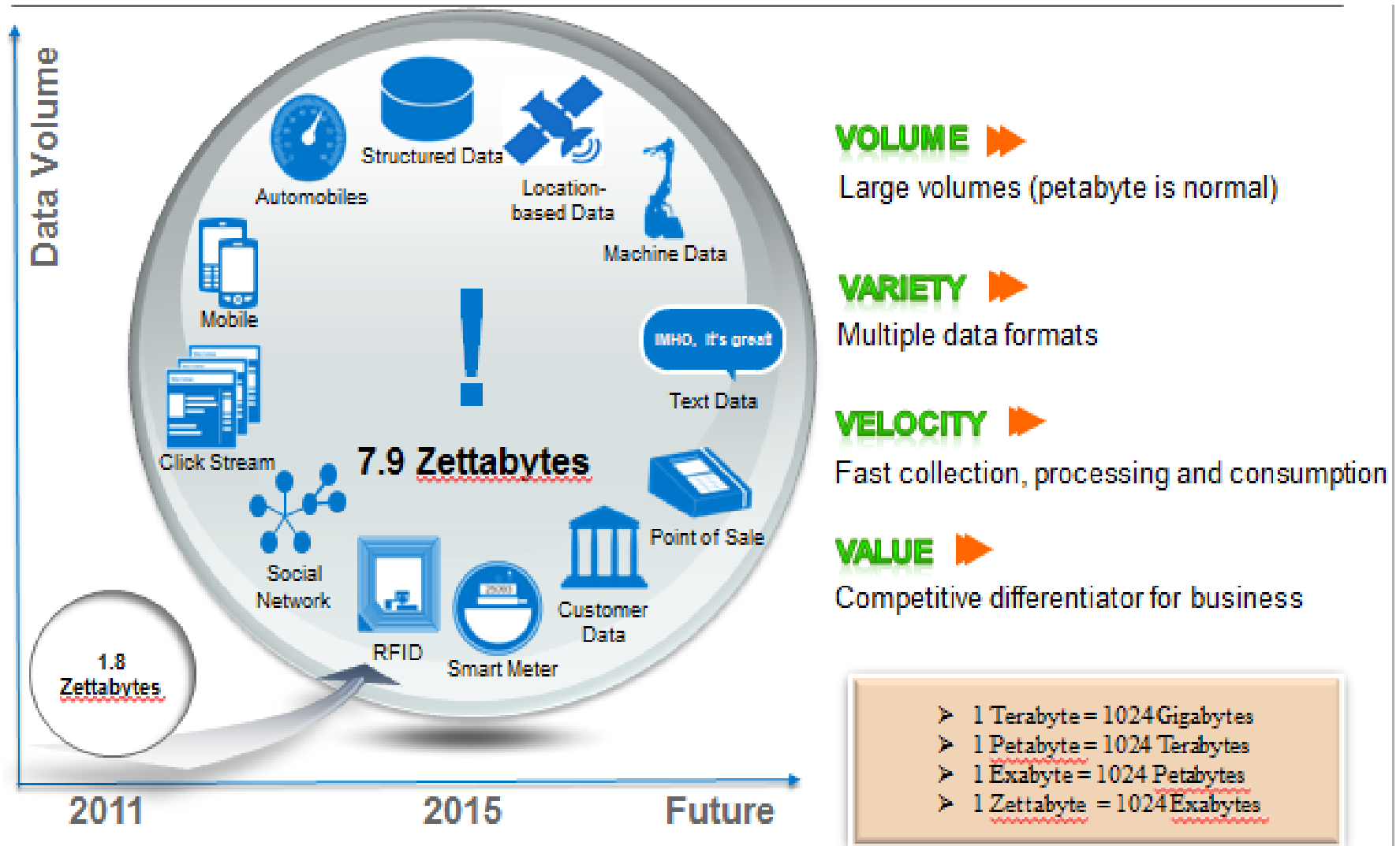
# Apesar de tudo, ainda existe informação analógica a considerar

## Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

A ciência de dados é aplicada a diferentes setores (por exemplo, RFID, em 2015, representa quase 23% do total)

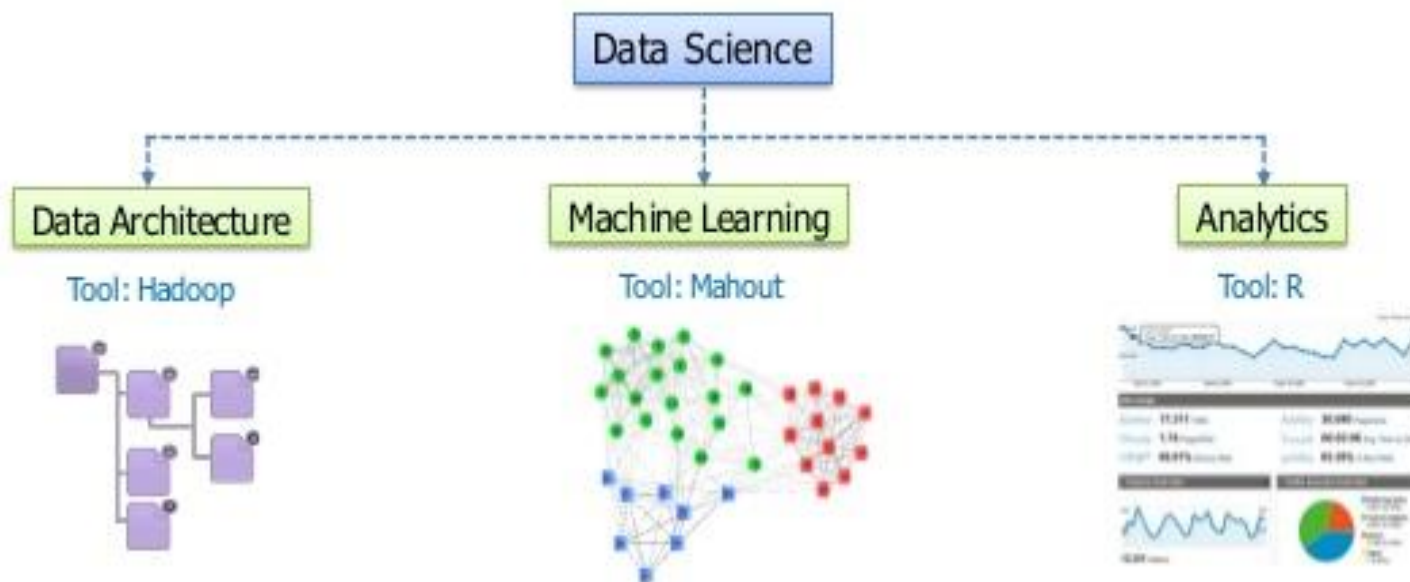


# O maior crescimento é o de dados não estruturados (dentro e fora da empresa)

## Exhibit 1 Examples of Unstructured Data

Corporate Enterprise Systems	Internet
<ul style="list-style-type: none"><li>• Corporate and personal email.</li><li>• Text and instant messages.</li><li>• Customer call logs, correspondence, and sales rep comments.</li><li>• Voice mail and phone logs.</li><li>• Documents, presentations, spreadsheets, and reports.</li><li>• Encrypted files and messages.</li><li>• Images and audio and video files.</li><li>• Calendars and contacts.</li><li>• Cellphone and vehicle location data.</li><li>• Internet histories, email-file transmission, and system access logs.</li></ul>	<ul style="list-style-type: none"><li>• Social network profiles such as Facebook, LinkedIn, Twitter, and Google Plus.</li><li>• Social influence sites such as expert blogs, user forums, industry-specific sites, and consumer review sites (Amazon, ZDNet, and Zillow).</li><li>• Activity-generated data such as location information.</li><li>• Government, publisher, and aggregator databases, wikis.</li><li>• Newspapers, articles, and white papers.</li><li>• Websites.</li></ul>

# Componentes (e ferramentas) da ciência de dados



# A visualização de dados é relevante

## A PERIODIC TABLE OF VISUALIZATION METHODS

>☀< <b>C</b> continuum											<b>Data Visualization</b> <i>Visual representations of quantitative data in schematic form (either with or without axes)</i>					<b>Strategy Visualization</b> <i>The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.</i>					>☀< <b>G</b> graphic facilitation						
>☀< <b>Tb</b> table	>☀< <b>Ca</b> cartesian coordinates											<b>Information Visualization</b> <i>The use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with it</i>					<b>Metaphor Visualization</b> <i>Visual Metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed</i>					>☀< <b>Me</b> meeting trace	>☀< <b>Mm</b> metro map	>☀< <b>Tm</b> temple	>☀< <b>St</b> story template	>☀< <b>Tr</b> tree	>☀< <b>Ct</b> cartoon
>☀< <b>Pi</b> pie chart	>☀< <b>L</b> line chart											<b>Concept Visualization</b> <i>Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.</i>					<b>Compound Visualization</b> <i>The complementary use of different graphic representation formats in one single schema or frame</i>					>☀< <b>Co</b> communication diagram	>☀< <b>Fp</b> flight plan	>☀< <b>Cs</b> concept skeleton	>☀< <b>Br</b> bridge	>☀< <b>Fu</b> funnel	>☀< <b>Ri</b> rich picture
>☀< <b>B</b> bar chart	>☀< <b>Ac</b> area chart	>☀< <b>R</b> radar chart cobweb	>☀< <b>Pa</b> parallel coordinates	>☀< <b>Hy</b> hyperbolic tree	>☀< <b>Cy</b> cycle diagram	>☀< <b>T</b> timeline	>☀< <b>Ve</b> venn diagram	>☀< <b>Mi</b> mindmap	>☀< <b>Sq</b> square of oppositions	>☀< <b>Cc</b> concentric circles	>☀< <b>Ar</b> argument slide	>☀< <b>Sw</b> swim lane diagram	>☀< <b>Gc</b> gant chart	>☀< <b>Pm</b> perspectives diagram	>☀< <b>D</b> dilemma diagram	>☀< <b>Pr</b> parameter ruler	>☀< <b>Kn</b> knowledge map										
>☀< <b>Hi</b> histogram	>☀< <b>Sc</b> scatterplot	>☀< <b>Sa</b> sankey diagram	>☀< <b>In</b> information lense	>☀< <b>E</b> entity relationship diagram	>☀< <b>Pt</b> petri net	>☀< <b>Fl</b> flow chart	>☀< <b>Cl</b> clustering	>☀< <b>Lc</b> layer chart	>☀< <b>Py</b> minto pyramid technique	>☀< <b>Ce</b> cause-effect chains	>☀< <b>Tl</b> toulmin map	>☀< <b>Dt</b> decision tree	>☀< <b>Cp</b> cpm critical path method	>☀< <b>Cf</b> concept fan	>☀< <b>Co</b> concept map	>☀< <b>Ic</b> iceberg	>☀< <b>Lm</b> learning map										
>☀< <b>Tk</b> tukey box plot	>☀< <b>Sp</b> spectrogram	>☀< <b>Da</b> data map	>☀< <b>Tp</b> treemap	>☀< <b>Cn</b> cone tree	>☀< <b>Sy</b> system dyn./ simulation	>☀< <b>Df</b> data flow diagram	>☀< <b>Se</b> semantic network	>☀< <b>So</b> soft system modeling	>☀< <b>Sn</b> synergy map	>☀< <b>Fo</b> force field diagram	>☀< <b>Ib</b> ibis argumentation map	>☀< <b>Pr</b> process event chains	>☀< <b>Pe</b> pert chart	>☀< <b>Ev</b> evocative knowledge map	>☀< <b>V</b> Vee diagram	>☀< <b>Hh</b> heaven 'n' hell chart	>☀< <b>I</b> infomural										

**Cy** Process Visualization

Note: Depending on your location and connection speed it can take some time to load a pop-up picture.

version 1.5

© Ralph Lengler & Martin J. Eppler; www.visual-literacy.org

**Hy** Structure Visualization

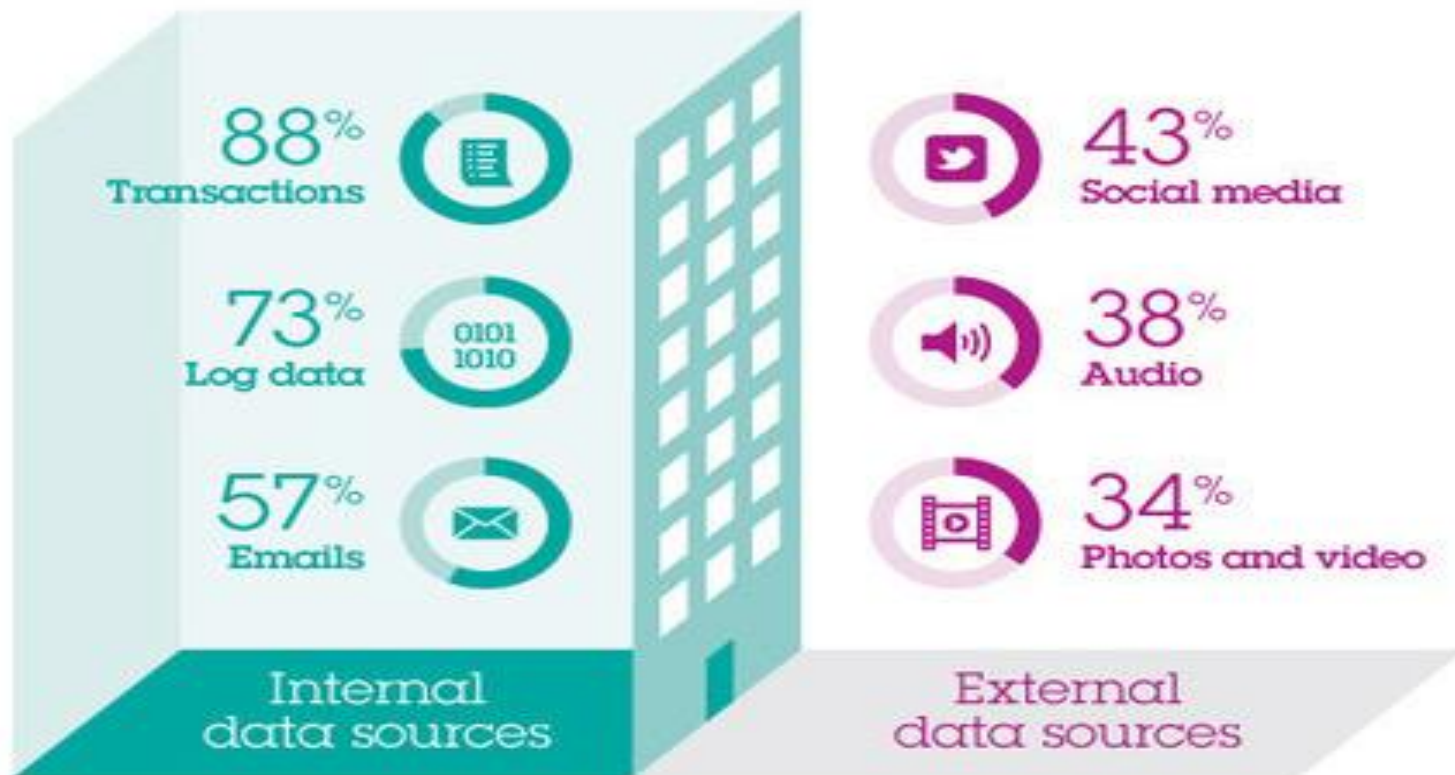
- ☀ **Overview**
- ☐ **Detail**
- ☉ **Detail AND Overview**
- < > **Divergent thinking**

>☀< <b>Su</b> supply demand curve	>☀< <b>Pc</b> performance charting	>☀< <b>St</b> strategy map	>☀< <b>Oc</b> organisation chart	<☐> <b>Ho</b> house of quality	>☀< <b>Fd</b> feedback diagram	☐ <b>Ft</b> failure tree	>☀< <b>Mq</b> magic quadrant	>☀< <b>Ld</b> life-cycle diagram	>☀< <b>Po</b> porter's five forces	<☐> <b>S</b> s-cycle	>☀< <b>Sm</b> stakeholder map	☉ <b>Is</b> ishikawa diagram	☀ <b>Tc</b> technology roadmap
☀ <b>Ed</b> edgeworth	>☀< <b>Pf</b> portfolio	☀ <b>Sg</b> strategic	>☀< <b>Mz</b> mintzberg's	<☐> <b>Z</b> zwickly's	<☐> <b>Ad</b> affinity	☐ <b>De</b> decision	>☀< <b>Bm</b> bcg matrix	>☀< <b>Stc</b> strategy	>☀< <b>Vc</b> value chain	<☐> <b>Hy</b> hype-cycle	>☀< <b>Sr</b> stakeholder	>☀< <b>Ta</b> tabs	<☐> <b>Sd</b> spray

# As origens dos dados e o seu tipo

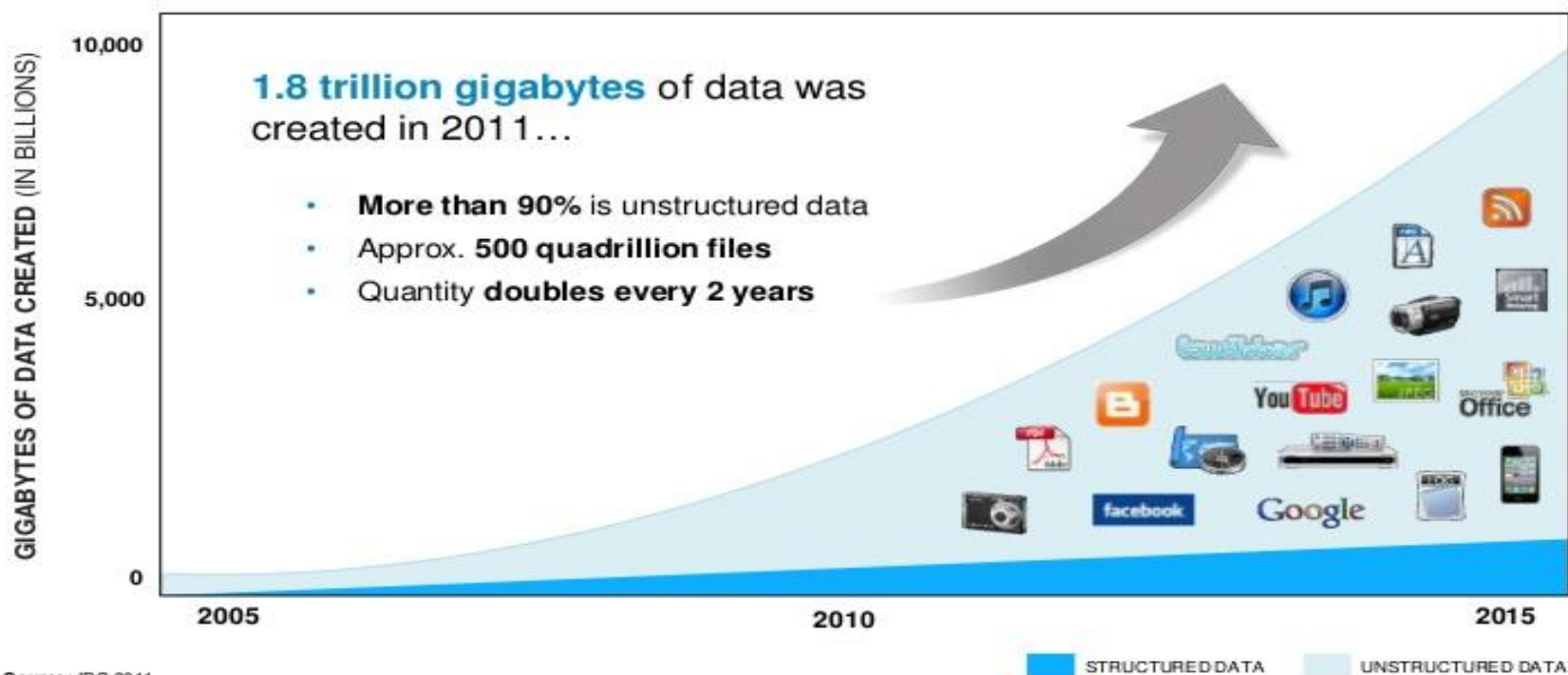
## Where does big data come from?

Most big data efforts are currently focused on analyzing internal data to extract insights. Fewer organizations are looking at data outside their firewalls, such as social media.



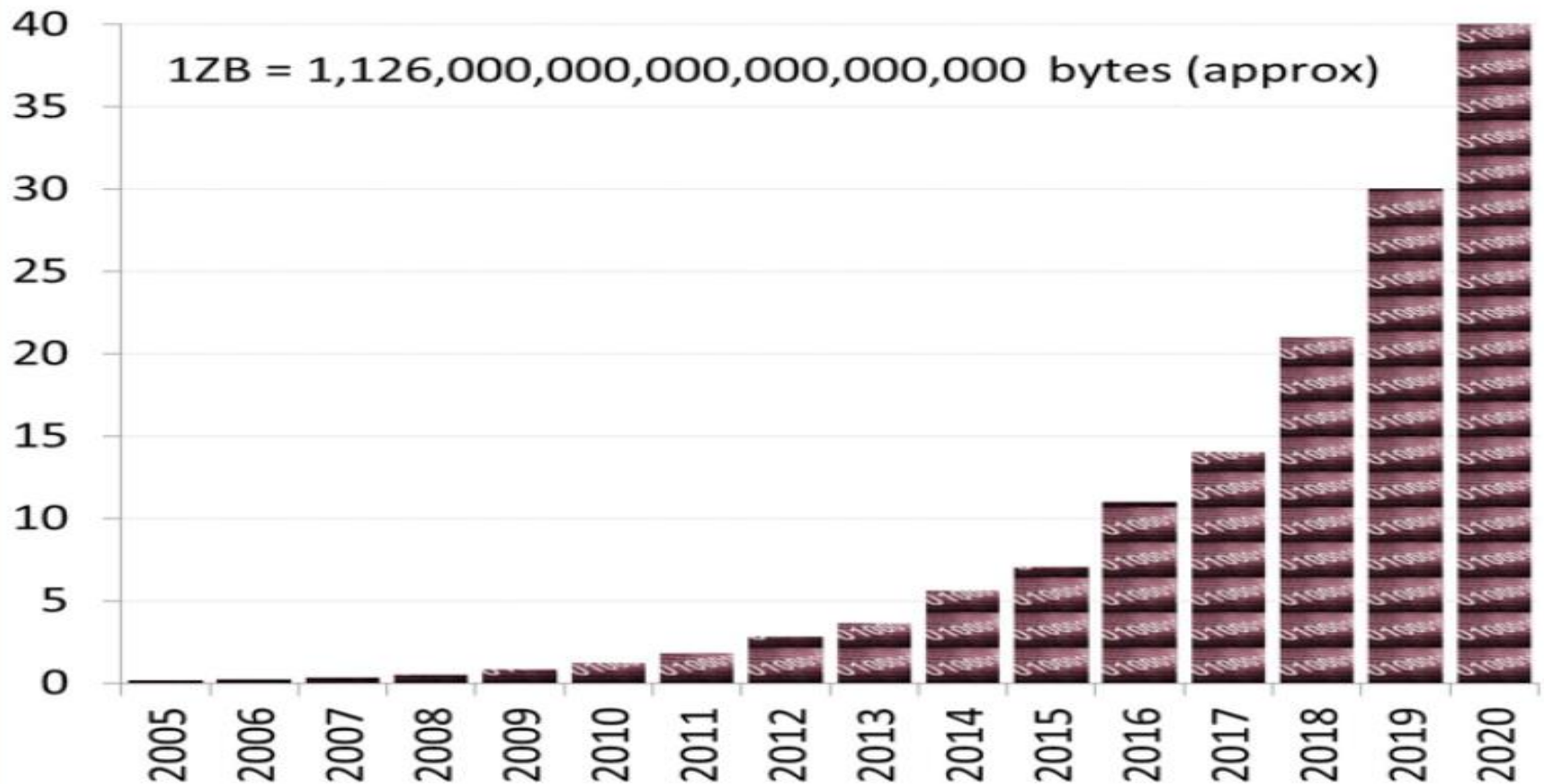
# A explosão de dados é essencialmente a explosão dos dados não estruturados

## Explosive Data Growth



# O que se pretende dizer por quantidades massivas de dados?

## All Global Data in Zettabytes



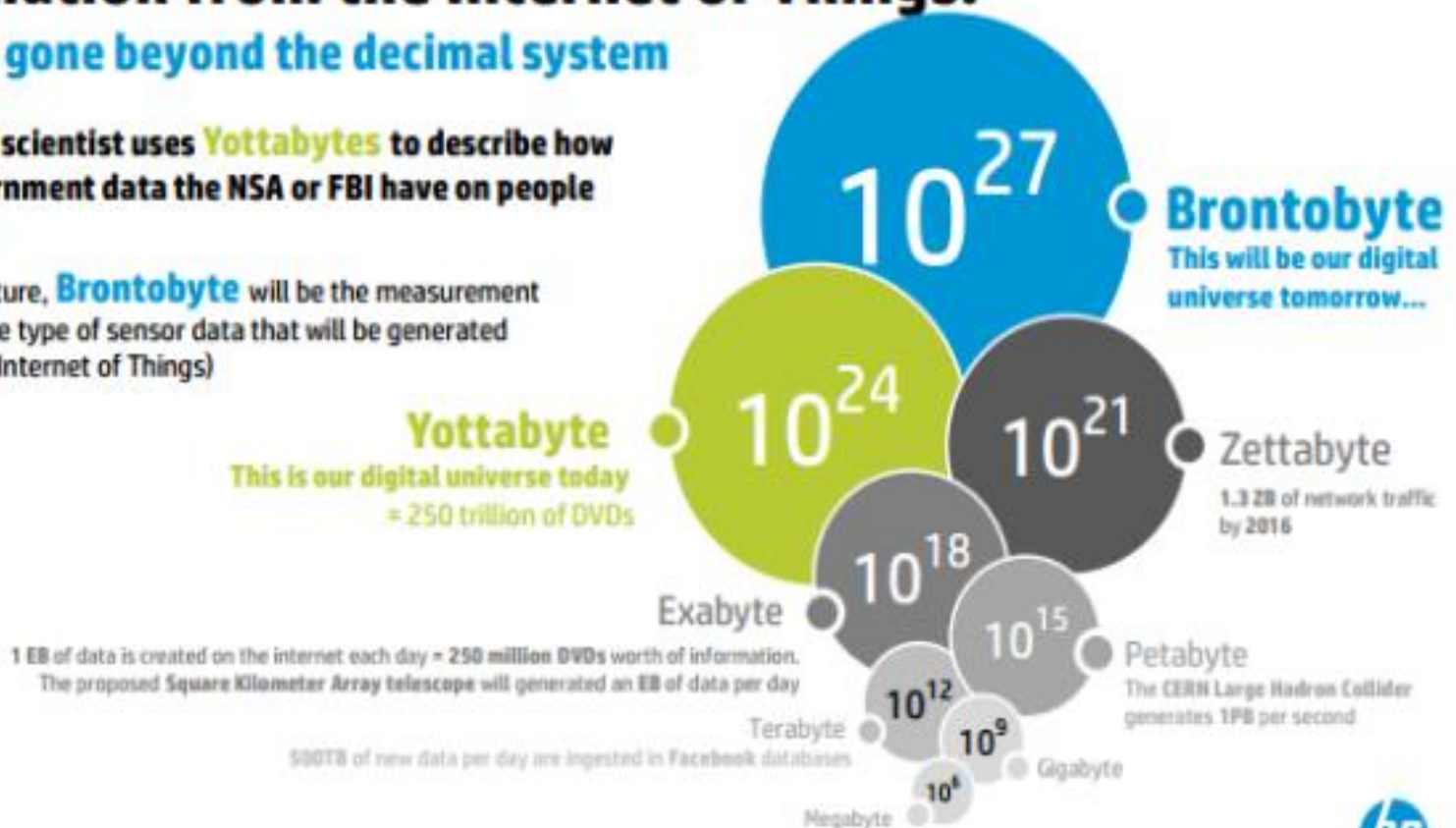
Source: <http://www1.unece.org/stat/platform/display/msis/Big+Data>

# Até onde vai escalar o crescimento de dados?

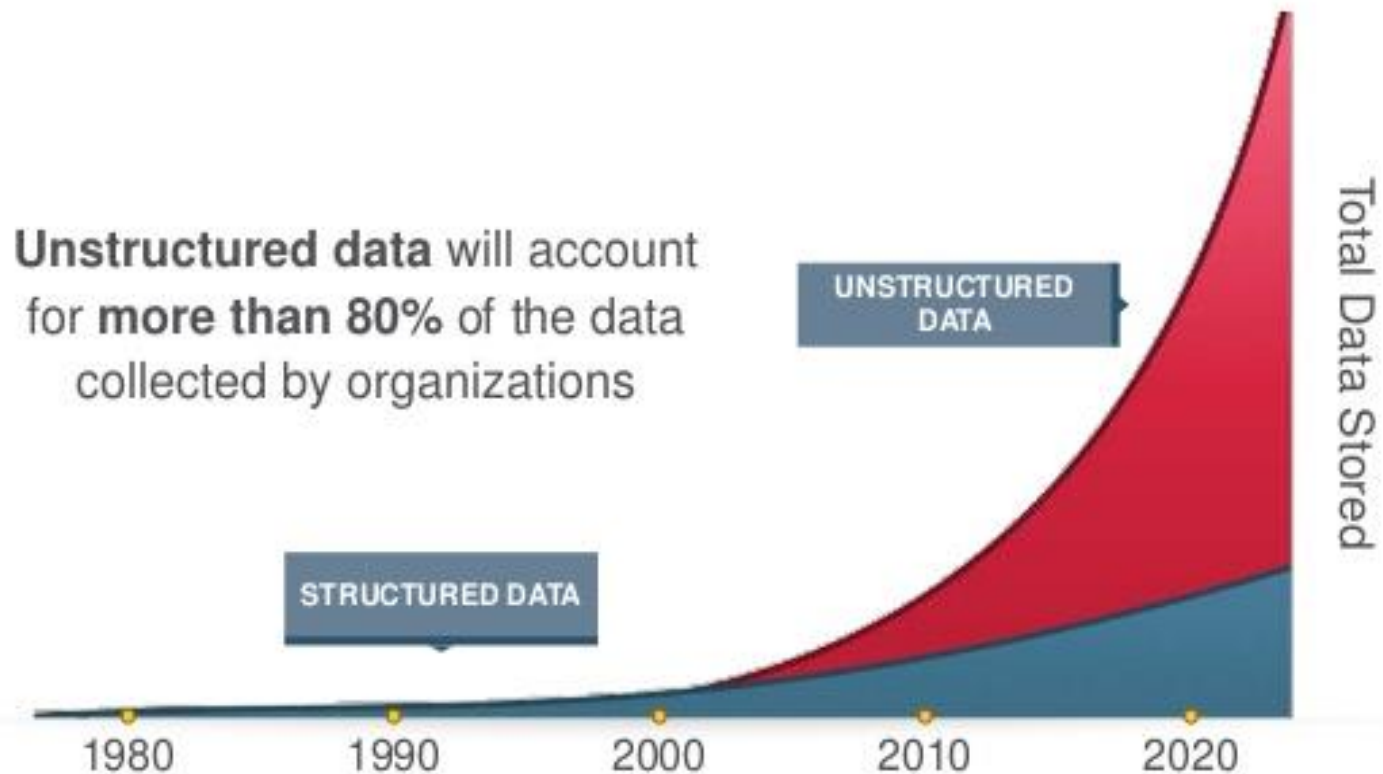
## Information from the Internet of Things: We have gone beyond the decimal system

Today data scientist uses **Yottabytes** to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



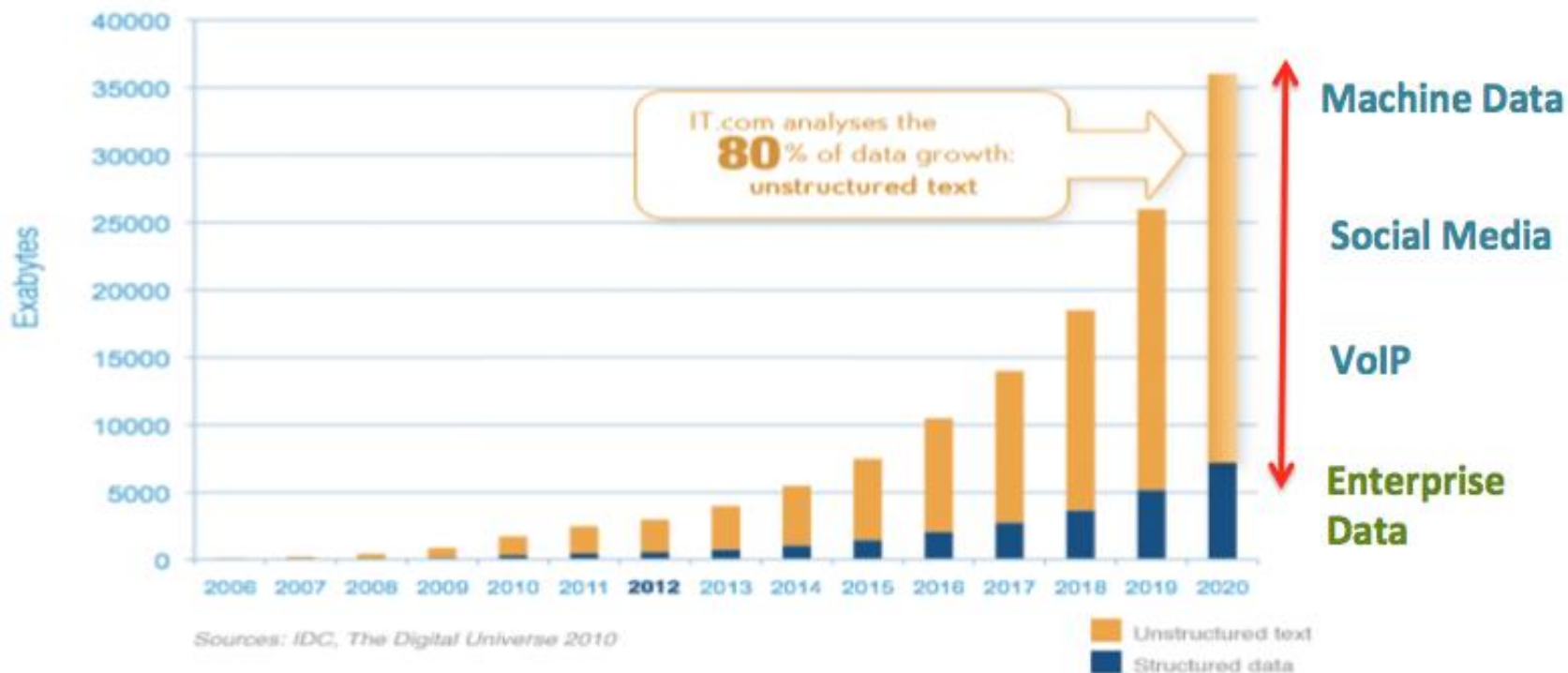
O crescimento dos dados não estruturados vai mudar o paradigma da sua gestão, nas empresas



# Nem todos os dados não estruturados tem a mesma origem

- Organizations are redefining data strategies due to the requirements of the evolving Enterprise Data Warehouse (EDW).

Worldwide Corporate Data Growth



# Dados estruturados versus dados não estruturados (origem)

## *Structured Data*



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

## *Unstructured Data*



# De que estamos a falar, quando falamos de dados não estruturados?

## STRUCTURED VS. UNSTRUCTURED DATA

### Structured Data

High Degree of organization, such as a relational database

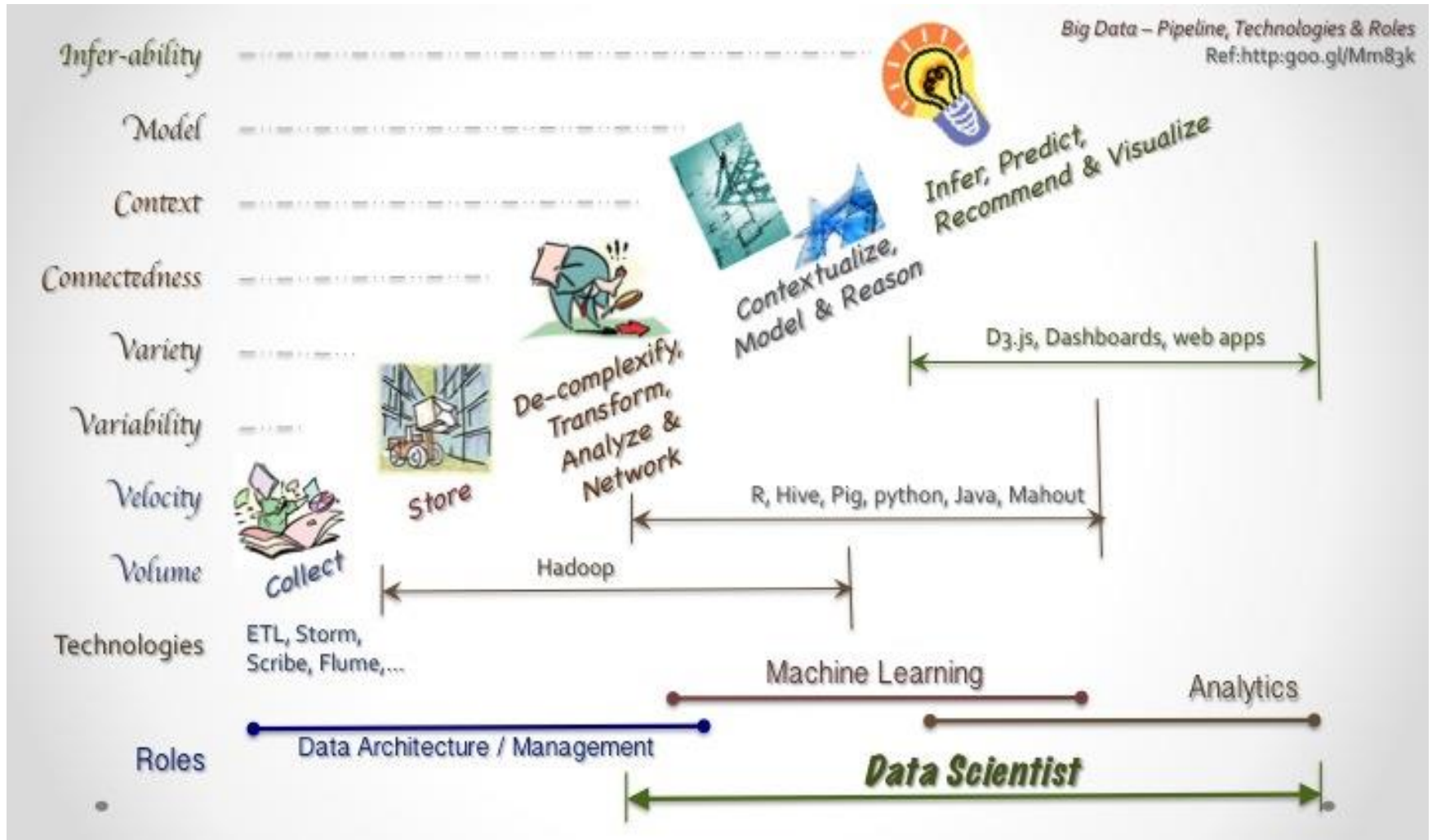
Column	Value
Patient	Joe Brown
Date of Birth	02/13/1972
Date Admitted	02/05/2014

### Unstructured Data

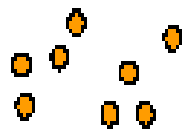
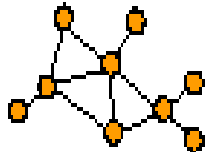

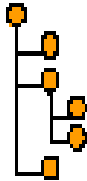
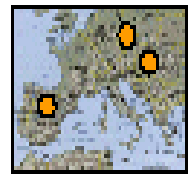

Information that is difficult to organize using traditional mechanisms

"The patient came in complaining of chest pain, shortness of breath, and lingering headaches...smokes 2 packs a day... family history of heart disease...has been experiencing similar symptoms for the past 12 hours...."

# Papeis, competências e atividades no contexto da ciência de dados



# Relacionar datos por recurso a modelos apropiados

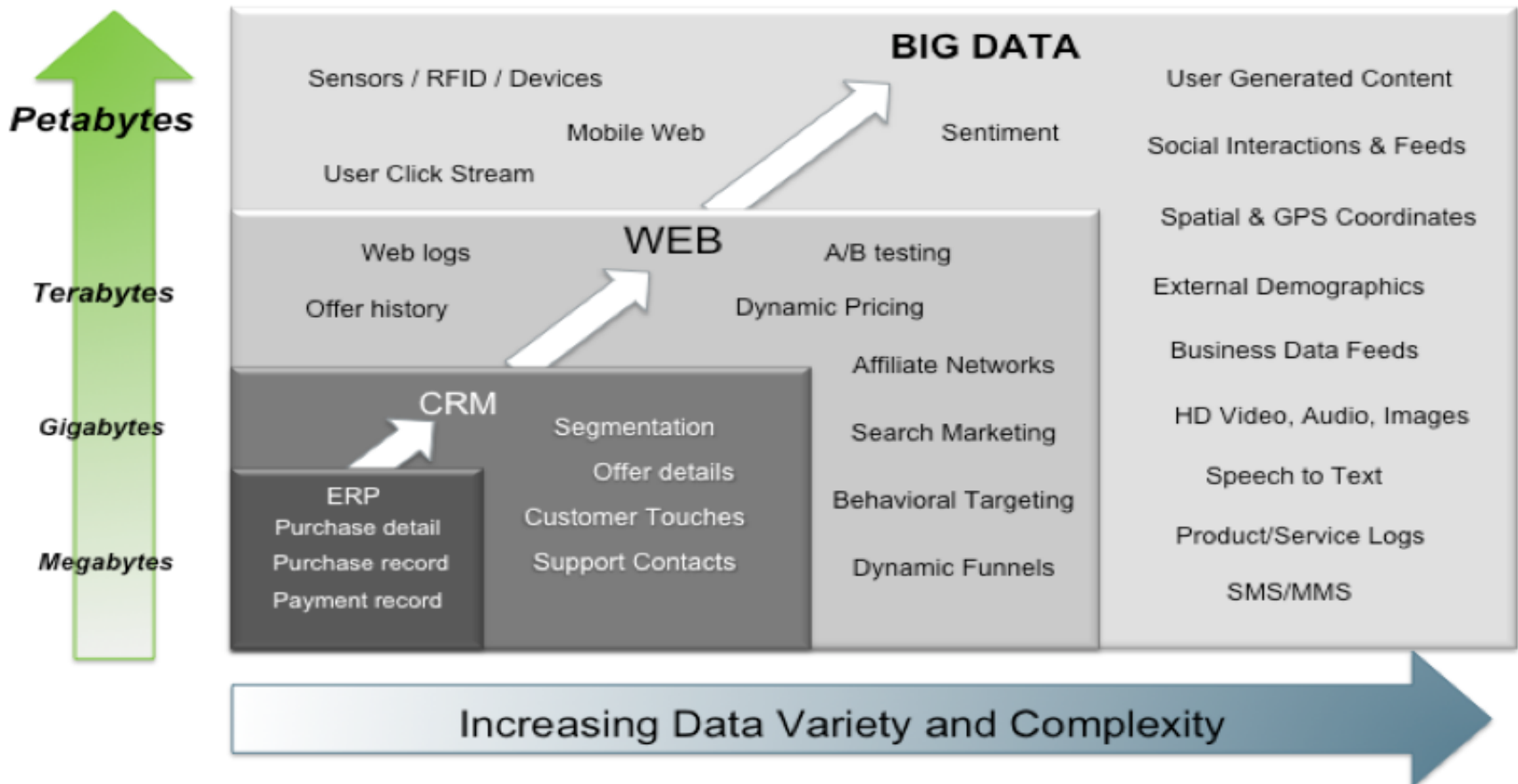
Relationship Type:	General Similarity	Explicit Reference	Field/Value Co-occurrence	Parent/Child	Spatial	Temporal
<b>Model Type:</b>	<i>Vector-space</i> 	<i>Network</i> 	<i>Multidimensional Index</i> 	<i>Hierarchical</i> 	<i>Spatial</i> 	<i>Ordinal Index</i> 
<b>Examples:</b>	<i>Reports, articles, DB records</i>	<i>References &amp; citations, hyperlinks</i>	<i>DB records, document metadata</i>	<i>File paths, taxonomies, IP addresses</i>	<i>Geolocations, CAD models</i>	<i>Event descriptions</i>

# As diferenças entre inteligência de negócios e a ciência de dados

	<b>Business Intelligence</b>	<b>Data Science</b>
<b>Perspective</b>	Looking backwards	Looking forwards
<b>Actions</b>	Slice and Dice	Interact
<b>Expertise</b>	Business User	Data Scientist
<b>Data</b>	Warehoused, Siloed	Distributed, real-time
<b>Scope</b>	Unlimited	Specific business question
<b>Questions</b>	What happened?	What will happen? What if?
<b>Output</b>	Table	Answer
<b>Applicability</b>	Historic, possible confounding factors	Future, correcting for influences
<b>Tools</b>	SAP, Cognos, Microstrategy, SAS	Revolution R Enterprise QlikView, Tableau, Jaspersoft
<b>Hot or not?</b>	So 1997	Transformational

# Lidar com dados massivos exige novos tipos de respostas pelas empresas

Big Data = Transactions + Interactions + Observations



**Source:** Contents of above graphic created in partnership with Teradata, Inc.

# Descobrir dados para inteligência de negócios e modelos para a ciência de dados

## Data science is very different from traditional business intelligence (BI)

Both approaches are necessary for the contemporary enterprise

### BI

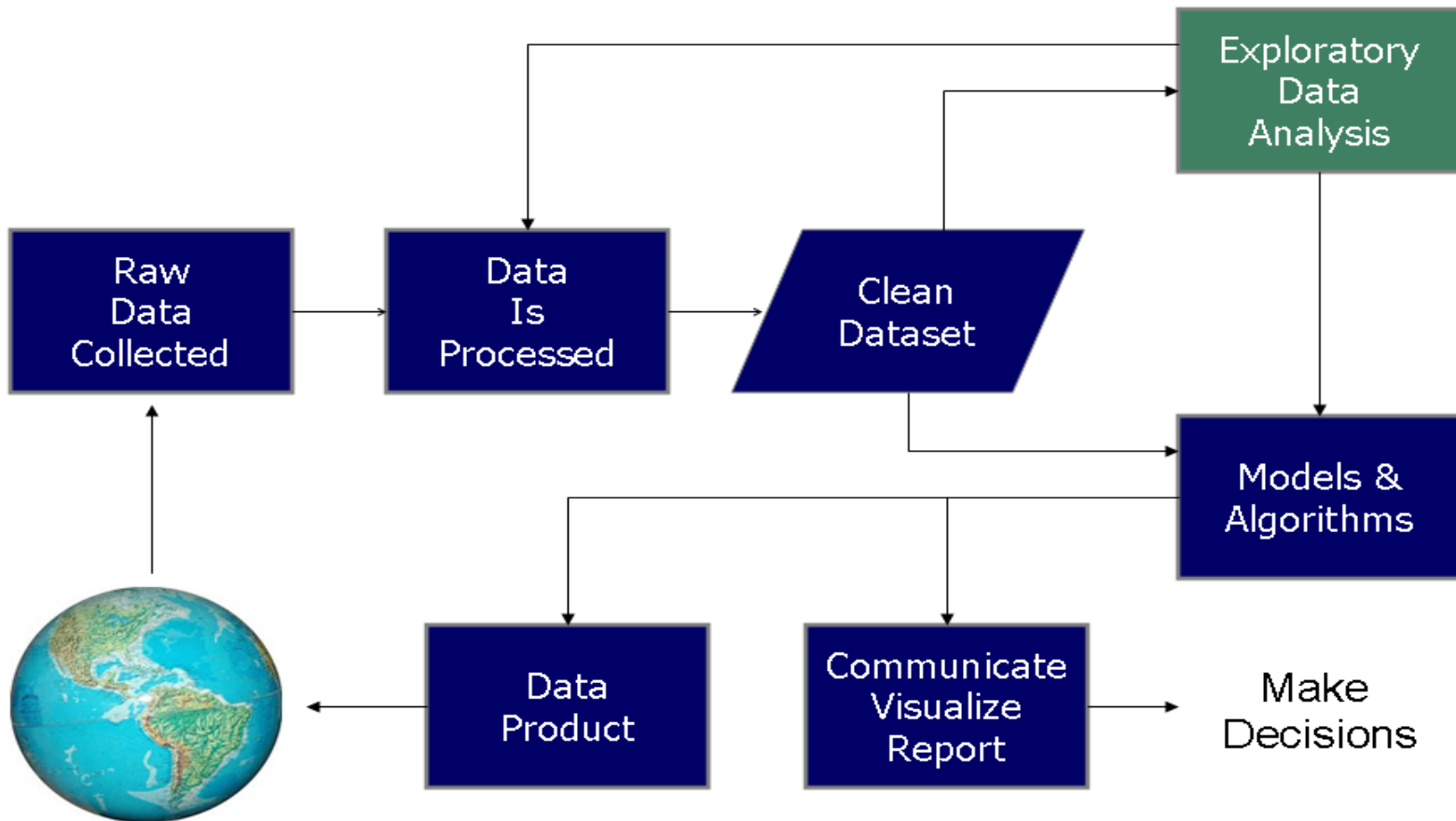
- Choose a business outcome to improve
- Decide what data will be relevant
- Create a data model
- Design reports, dashboards, and/or visualize

### Data Science

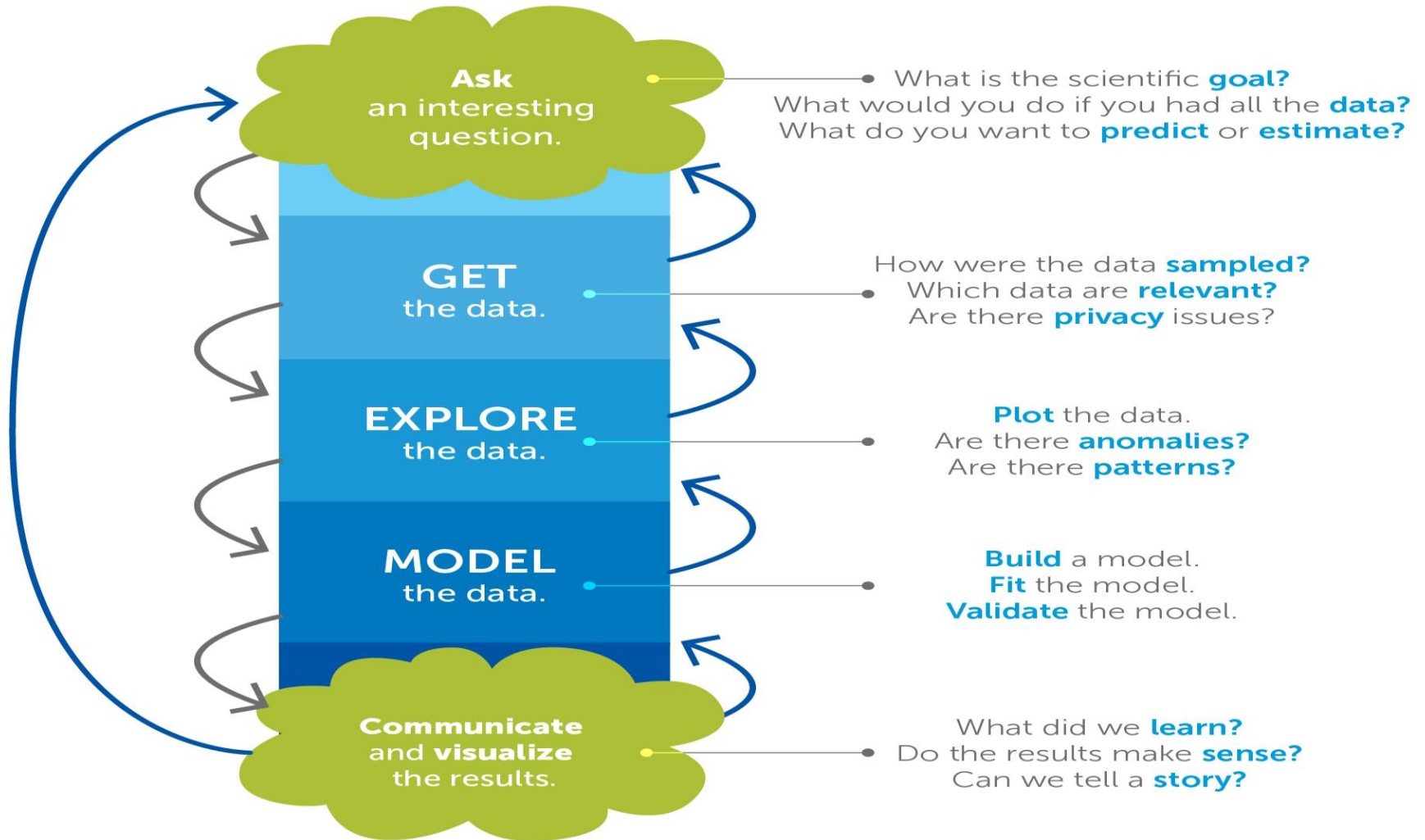
- Choose business outcome to improve
- Assemble all possible data
- Run algorithms on the data to find a model
- Evaluate the model
- Operationalize the model

# O processo de produção da ciência de dados (criar modelos e tomar decisões)

## Data Science Process



# The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course <http://cs109.org/>.

# Existe uma multiplicidade de ferramentas associadas com cada um dos processos de ciência de dados



# Comentários finais

- Área de grande potencial
  - Quer em dimensão de negócio, quer em empregabilidade
- Existe um enorme leque de aplicação
  - onde quer que exista a possibilidade de obter dados em grande quantidade ou de grande complexidade, em formato digital
- A ênfase deve ser na ciência e não nos dados
  - implicando o uso das diferentes técnicas de um modo ordenado
- Os profissionais de ciência de dados são especialistas de análise de dados que possuem competências técnicas para resolver problemas complexos e a curiosidade de explorar quais os problemas que devem ser resolvidos
  - existe uma dimensão de criatividade aplicada que é componente essencial do trabalho em ciência de dados

# Tarefas típicas de profissionais de ciência de dados

- Recolher grandes quantidades de dados não tratadas para transformar em dados úteis
- Resolver problemas relacionados com negócio ou contextos bem definidos, com recurso a técnicas orientadas a dados
- Trabalhar com uma variedade de linguagens de programação
- Dominar conceitos estatísticos, incluindo distribuições e testes estatísticos
- Dominar e acompanhar o estado de arte de técnicas analíticas como aprendizagem automática, *deep learning* e análise de texto
- Comunicar com equipas técnicas e de gestão
- Descobrir critérios e ordem em padrões de dados, bem como identificar tendências que podem contribuir para a eficácia do negócio ou do contexto em estudo

# Técnicas e ferramentas para a ciência de dados

- Visualização de dados: a apresentação de dados de forma gráfica de modo a ser mais facilmente entendida
- Aprendizagem automática: um ramo da inteligência artificial baseado em algoritmos matemáticos e na automação
- *Deep learning*: uma área da investigação em aprendizagem automática que usa os dados para modelar abstrações complexas
- Reconhecimento de padrões: tecnologia que reconhece padrões em dados
- Preparação de dados: o processo de conversão dos dados em bruto num formato que possa ser mais facilmente tratado ou consumido
- Análise de texto: o processo de examinar dados não estruturados de forma a extrair aspetos relevantes sobre o negócio ou o contexto em estudo

# Oportunidade e valor

