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

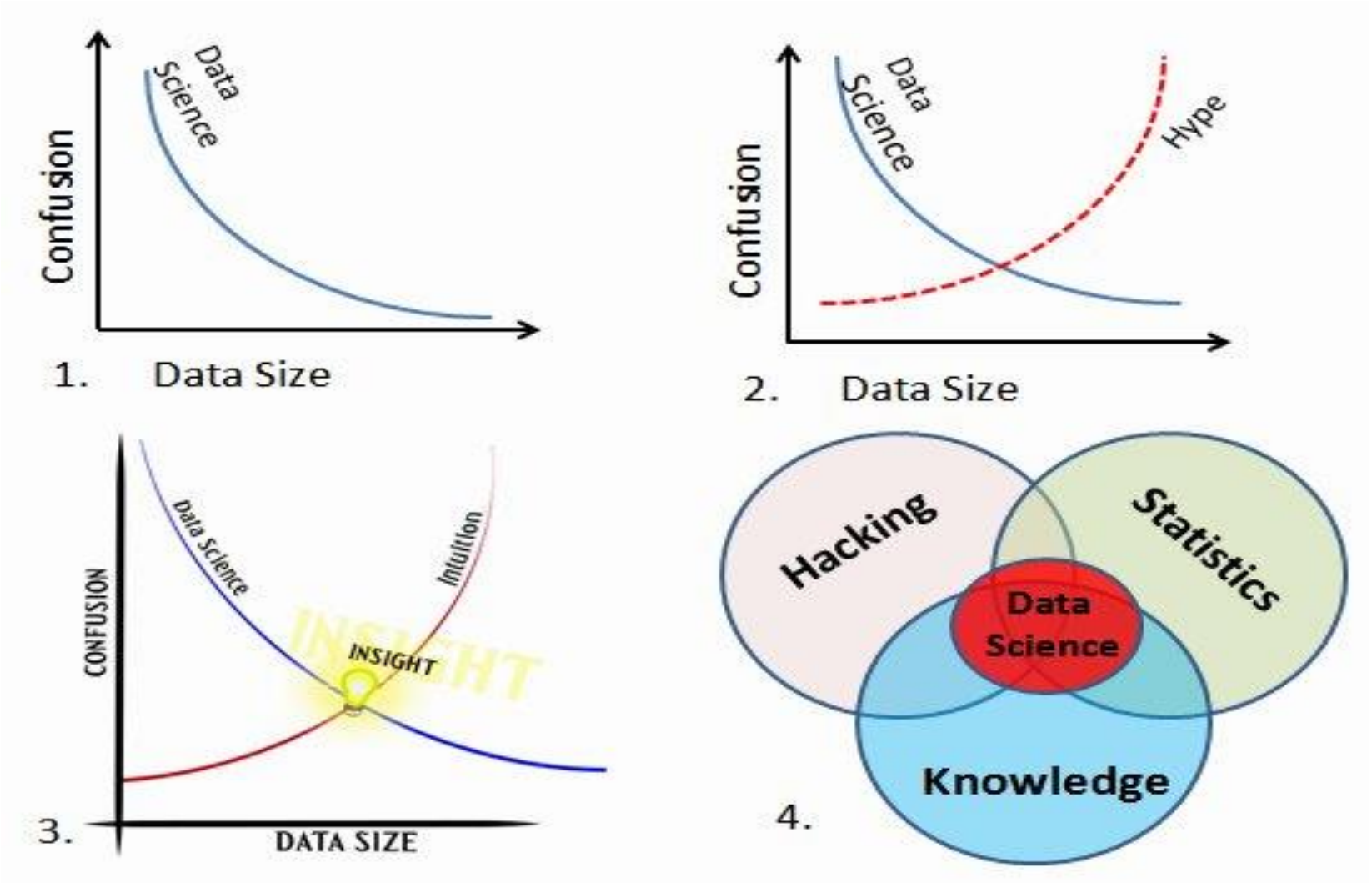
Data science ou a ciência de dados

- 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?

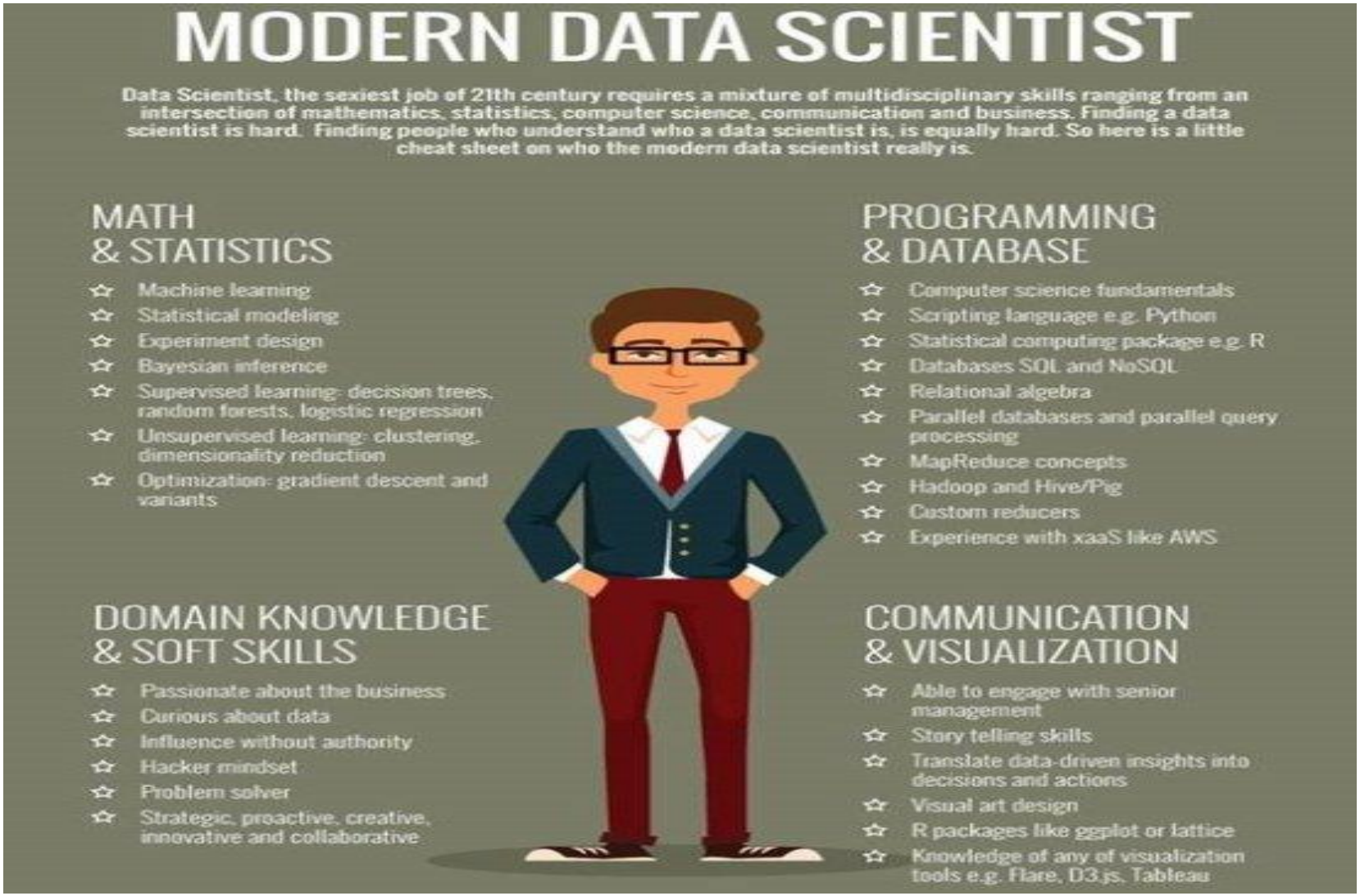
- Colocar uma questão
- Colocar uma hipótese
- Traçar um plano para a comprovar
- Elaborar um contexto de observação
- Observar e experimentar
- Registar os resultados
- Analisar os resultados
- Chegar a uma conclusão

Numa perspetiva mais operacional



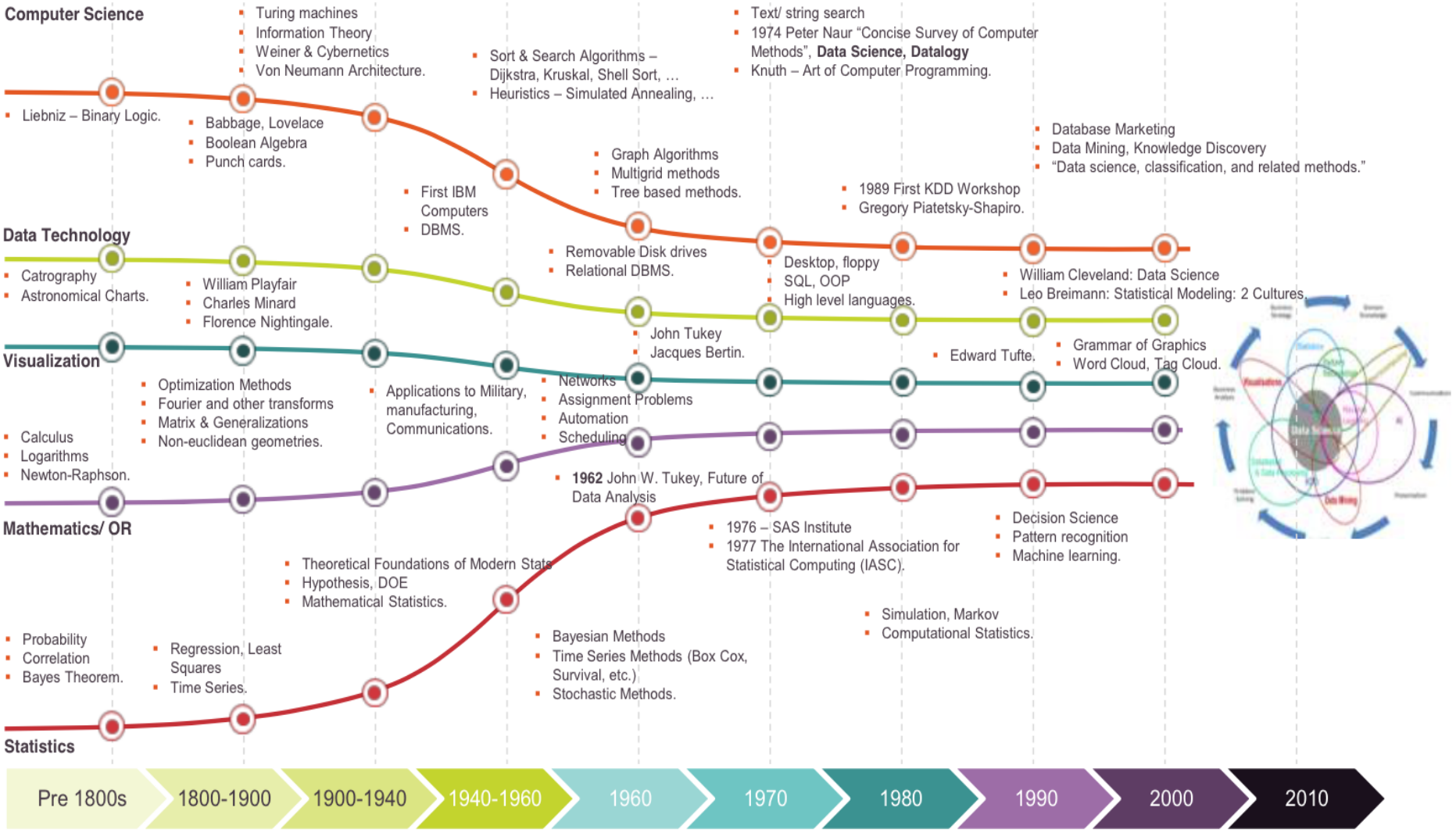
http://www.kdnuggets.com/2013/04/data-scientist-hat.html

O perfil multidisciplinar do profissional da ciência de dados



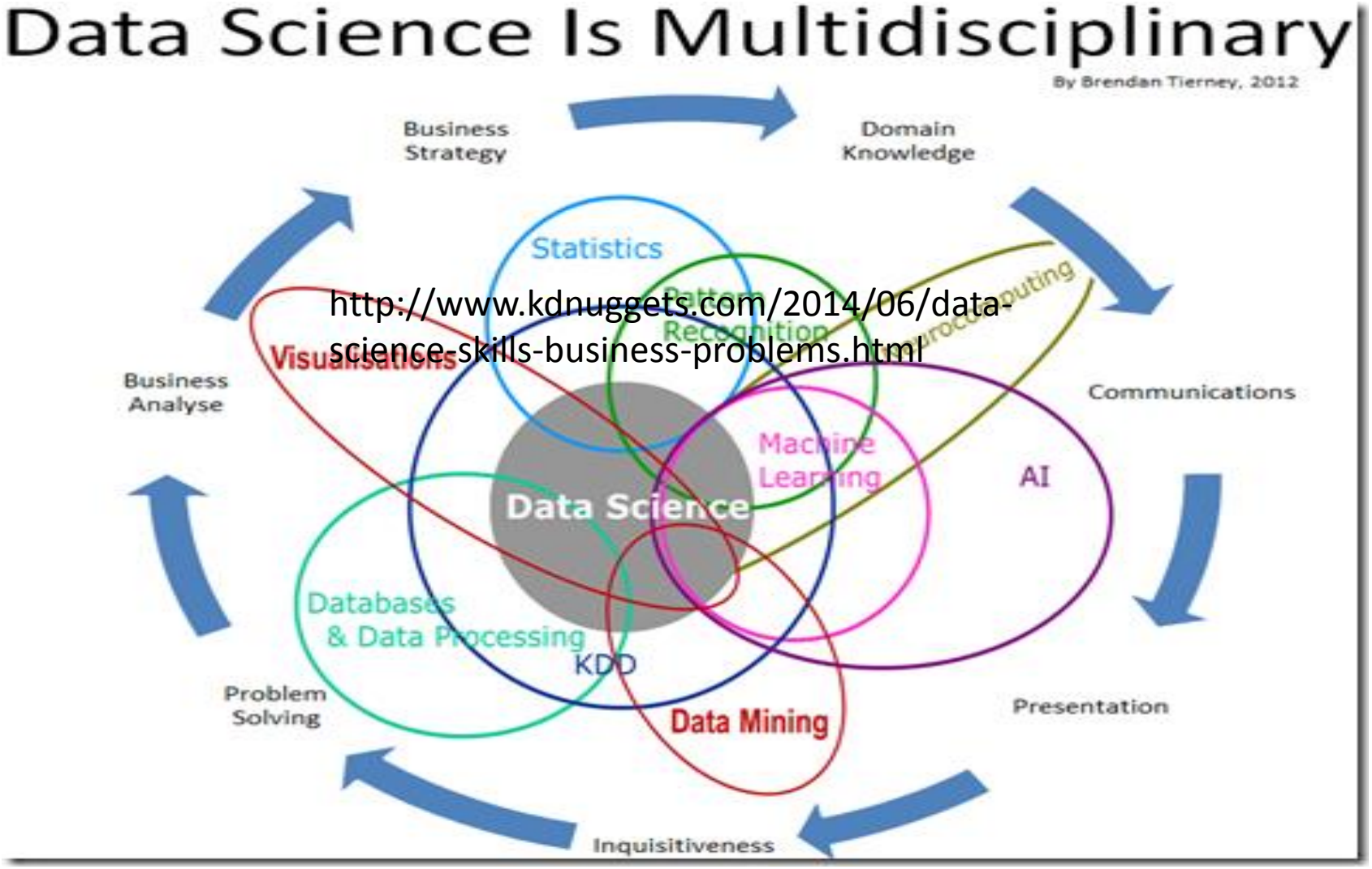
http://pt.slideshare.net/AdamRabinovitch/opus-data-scientist

A história e contribuições para a formação da ciência de dados



http://www.datasciencecentral.com/profiles/blogs/two-great-visualizations-about-data-science

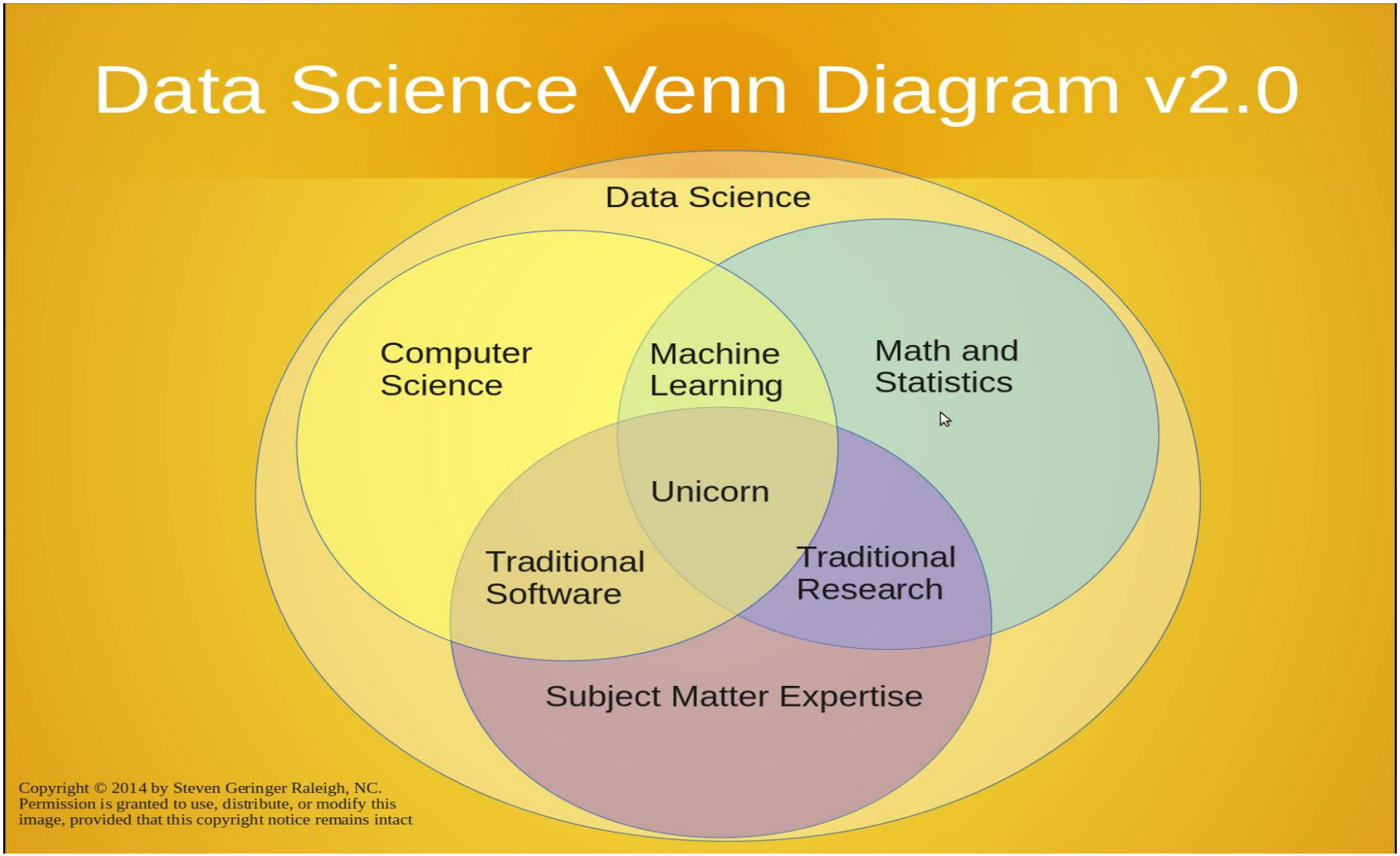
Multidisciplinaridade da ciência de dados



http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html

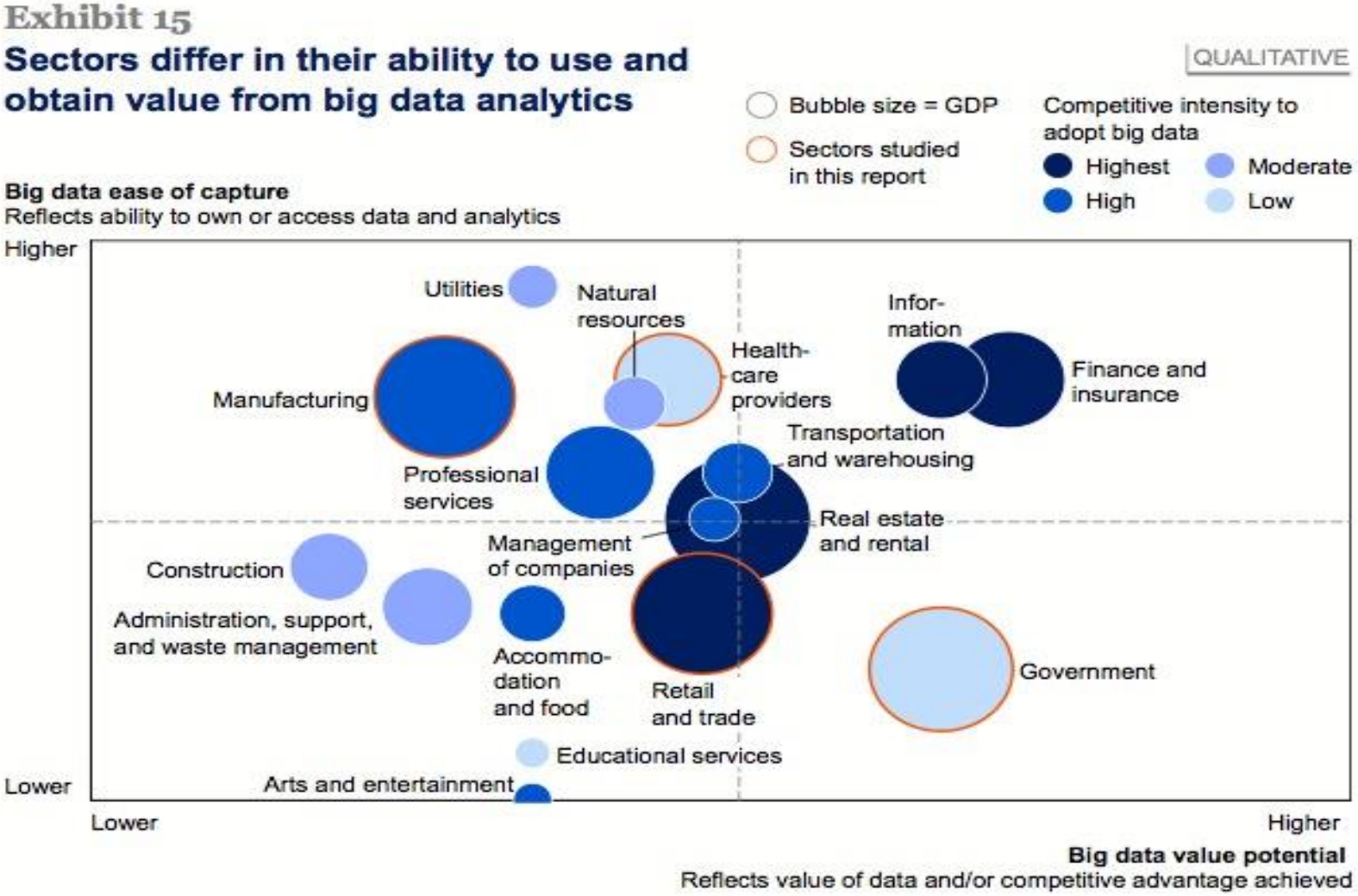
http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html

A ciência de dados face às ciências da computação

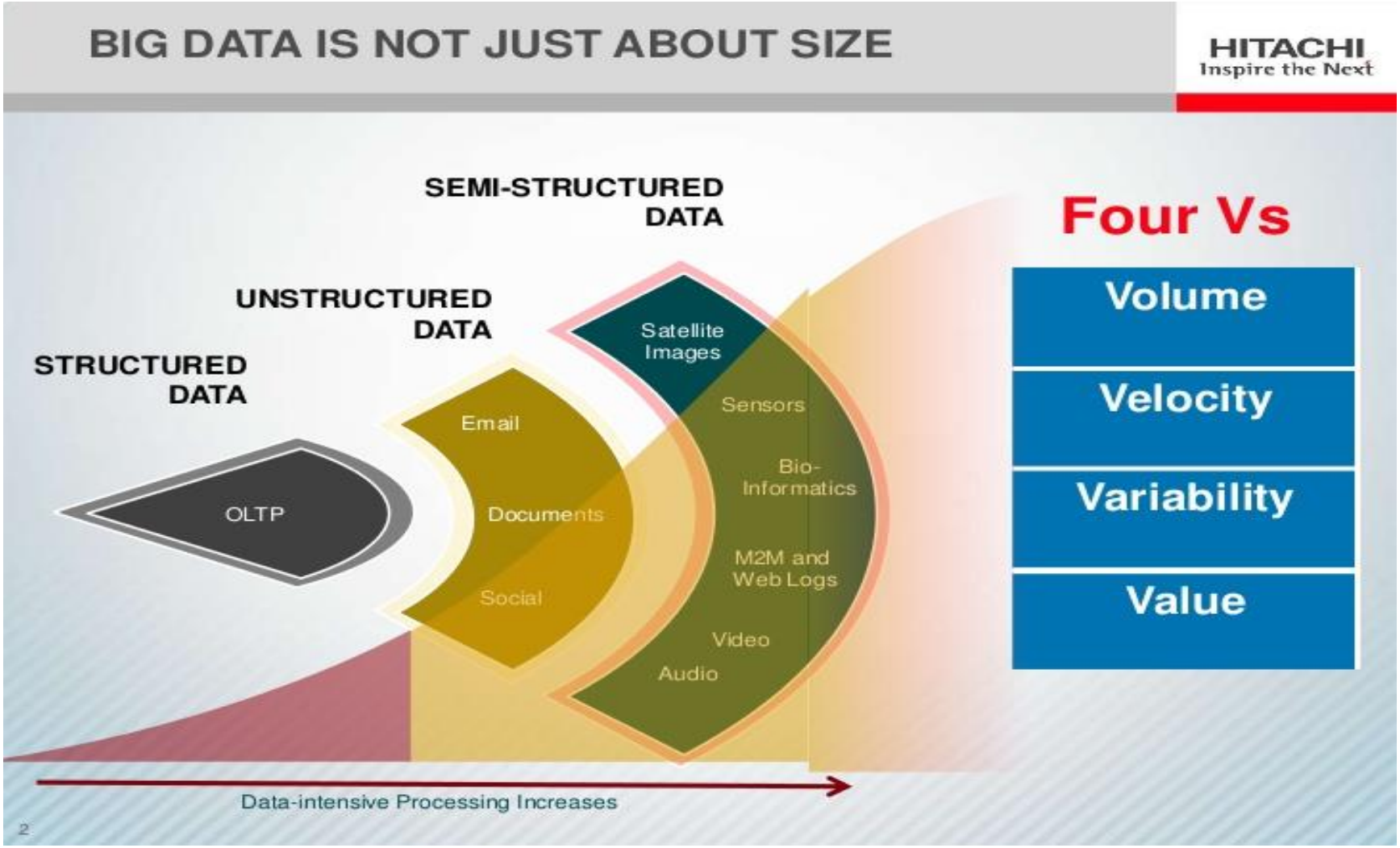


http://robjhyndman.com/hyndsight/am-i-a-data-scientist/

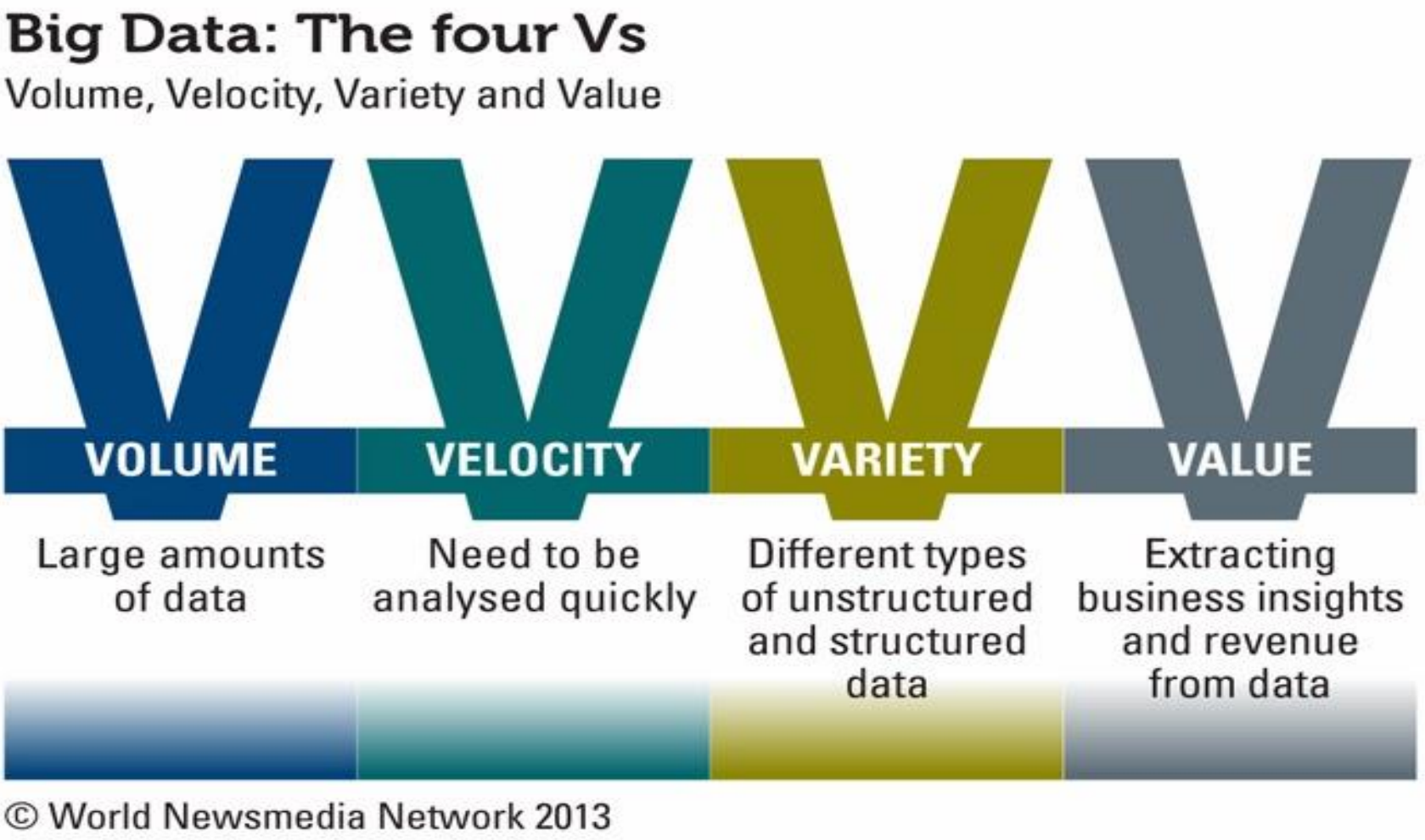
Peso económico e relevância da análise de dados



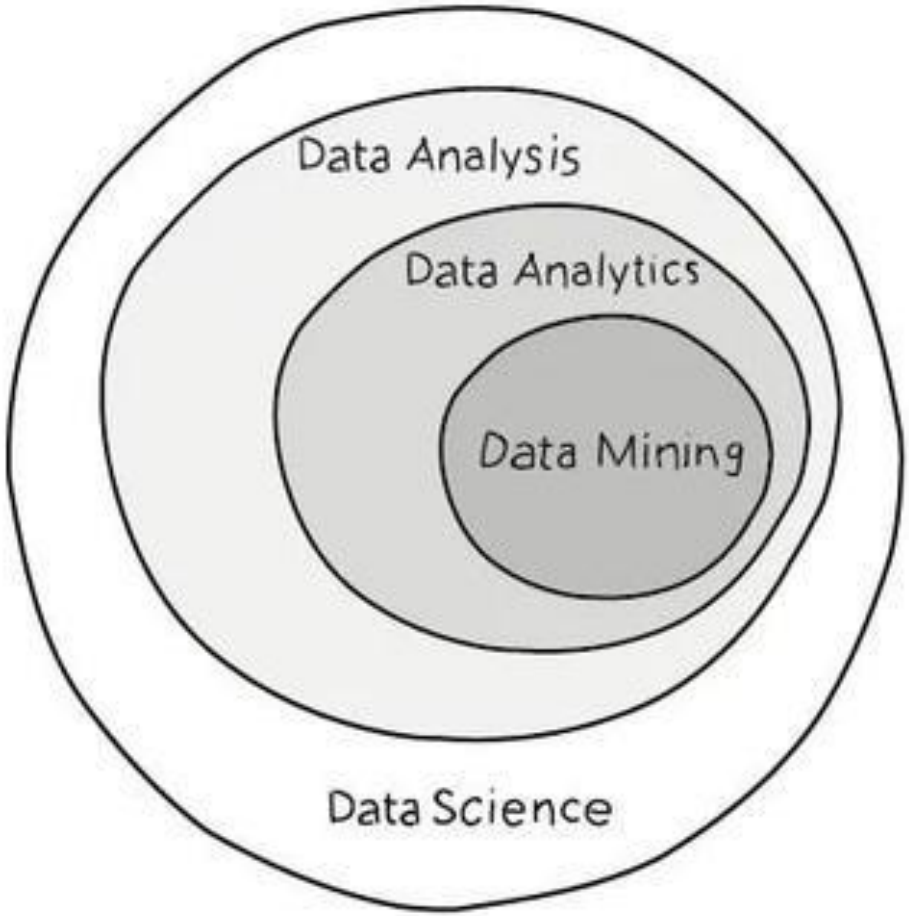
Uma enorme massa de dados e os desafios associados



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



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



http://www.datasciencecentral.com/profiles/blogs/two-great-visualizations-about-data-science

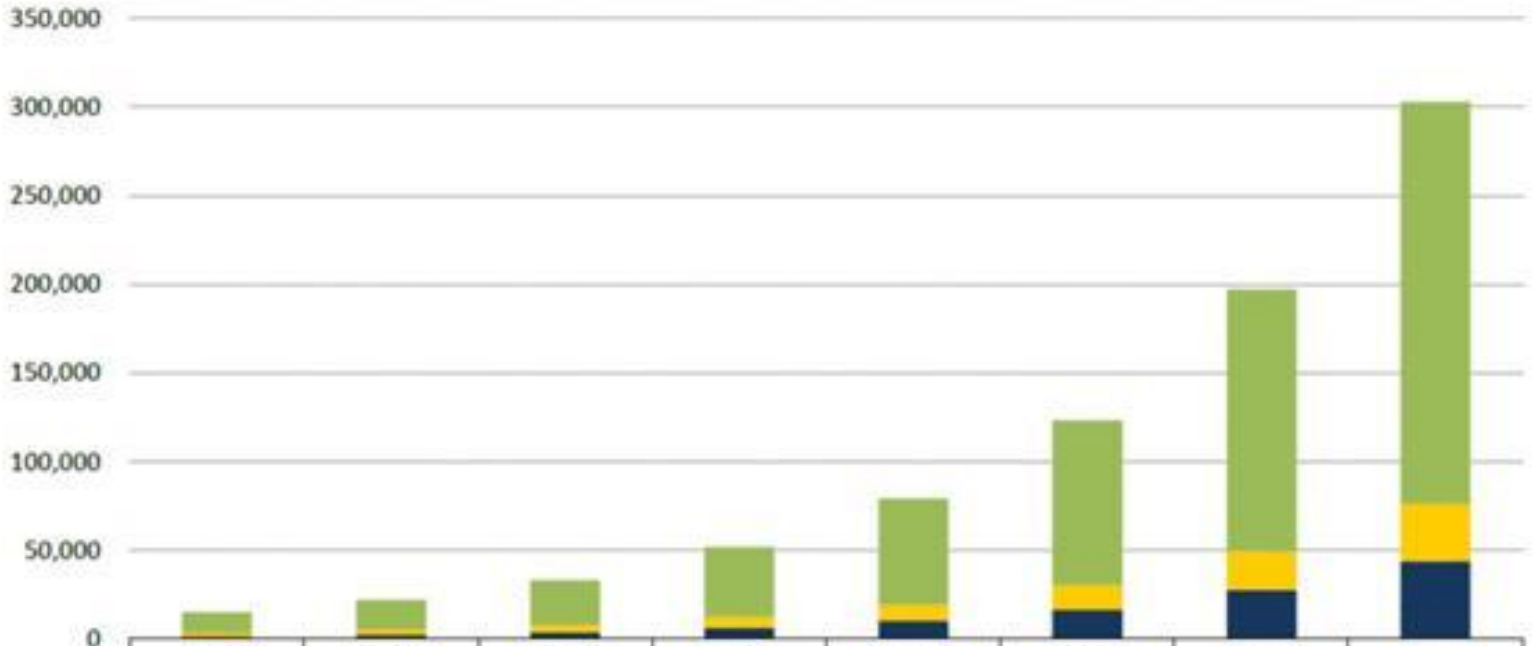
Diferenciar a inteligência de negócios, da análise de dados e da ciência de dados

DISCIPLINE	TECHNOLOGIES	SKILLS	FOCUS
BUSINESS INTELLIGENCE	• ETL Tools / SQL • RDBMS • Reporting • Visualization	• Programming • Data Analysis • Data Modeling • Report Development • Basic Statistics • Technical Architecture • Business Analysis & Strategy • Presentation	• Information Delivery and Reporting • Data Visualization • Descriptive Statistics • Data Integration and Consolidation
DATA ANALYSIS	• Data Modeling Software • Diagraming Software • Documentation Software • SQL • Data Profiling Software	• Data Modeling • Business Analysis • Data Manipulation • Basic statistics	• Business Rules • Data Definitions and Lineage • Data Entity Relationships • Data Attributes • Data Structures • Sources and Targets of Data • Data Quality
DATA SCIENCE	• Statistics Software • Columnar Data • Map-Reduce • NoSQL • Programming Languages • Graphing/Charting Software	• Advanced Statistics • Programming • Business Analysis • Modern Data Management Technologies and Architectures	• Predictive Modeling • Advanced Statistical Analysis • Data Mining • Unstructured Data Management • Large data volumes • Research

http://www.ediscoveryreadingroom.com/?p=2735

O exponencial crescimento dos dados, ocorre em várias categorias

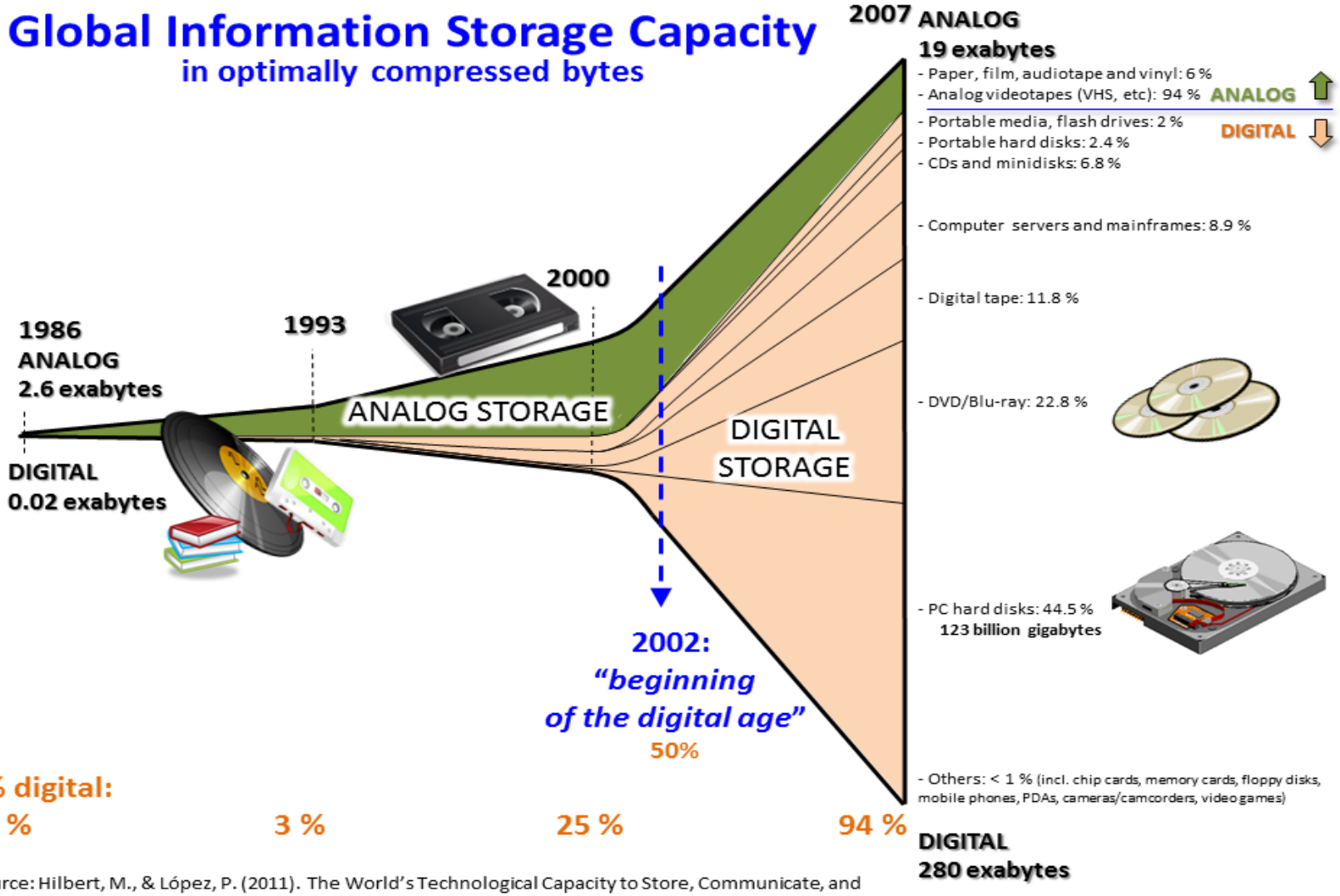
Total Archived Capacity, by Content Type, Worldwide, 2008-2015 (Petabytes)



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
E-mail	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

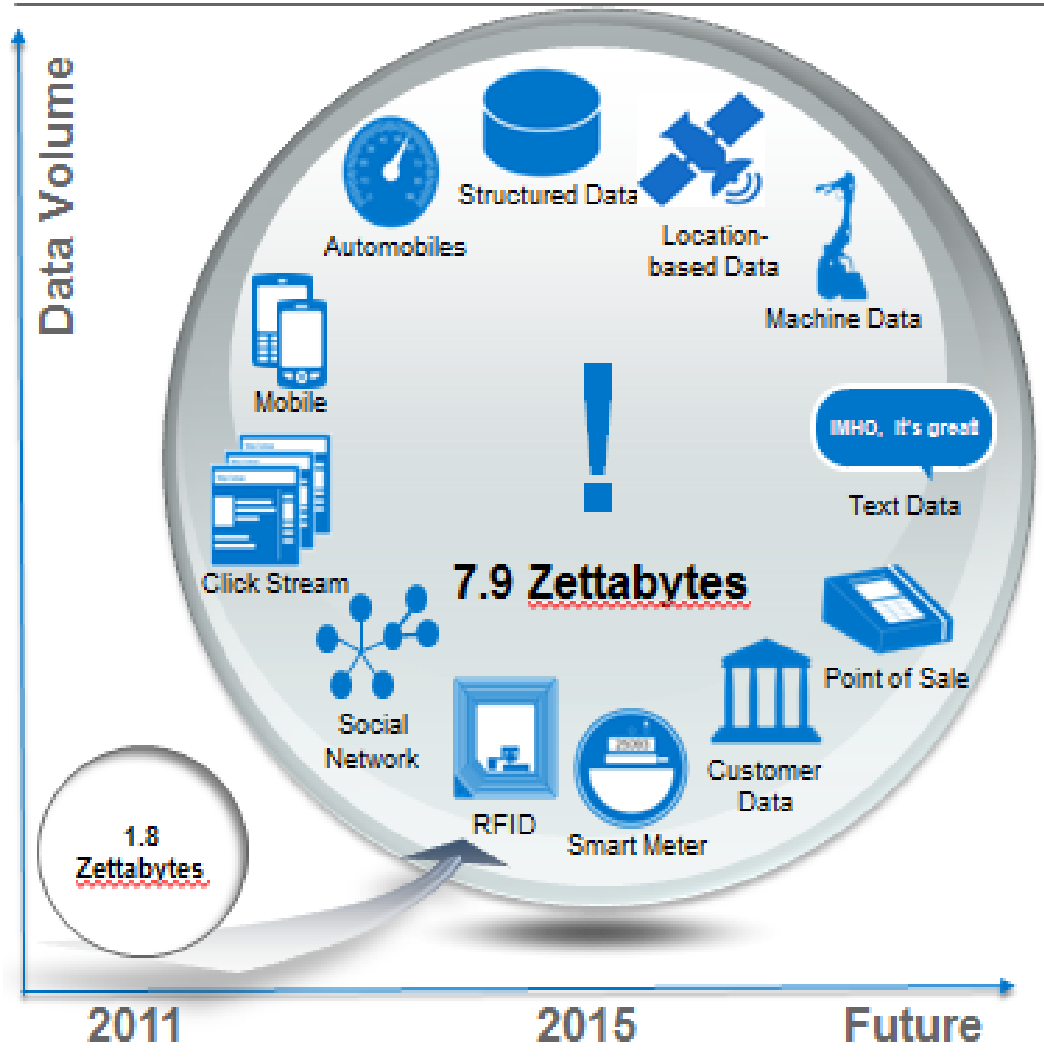
Source: Enterprise Strategy Group, 2010.

Apesar de tudo, ainda existe informação analógica a considerar



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)



VOLUME
Large volumes (petabyte is normal)

VARIETY
Multiple data formats

VELOCITY
Fast collection, processing and consumption

VALUE
Competitive differentiator for business

- 1 Terabyte = 1024 Gigabytes
- 1 Petabyte = 1024 Terabytes
- 1 Exabyte = 1024 Petabytes
- 1 Zettabyte = 1024 Exabytes

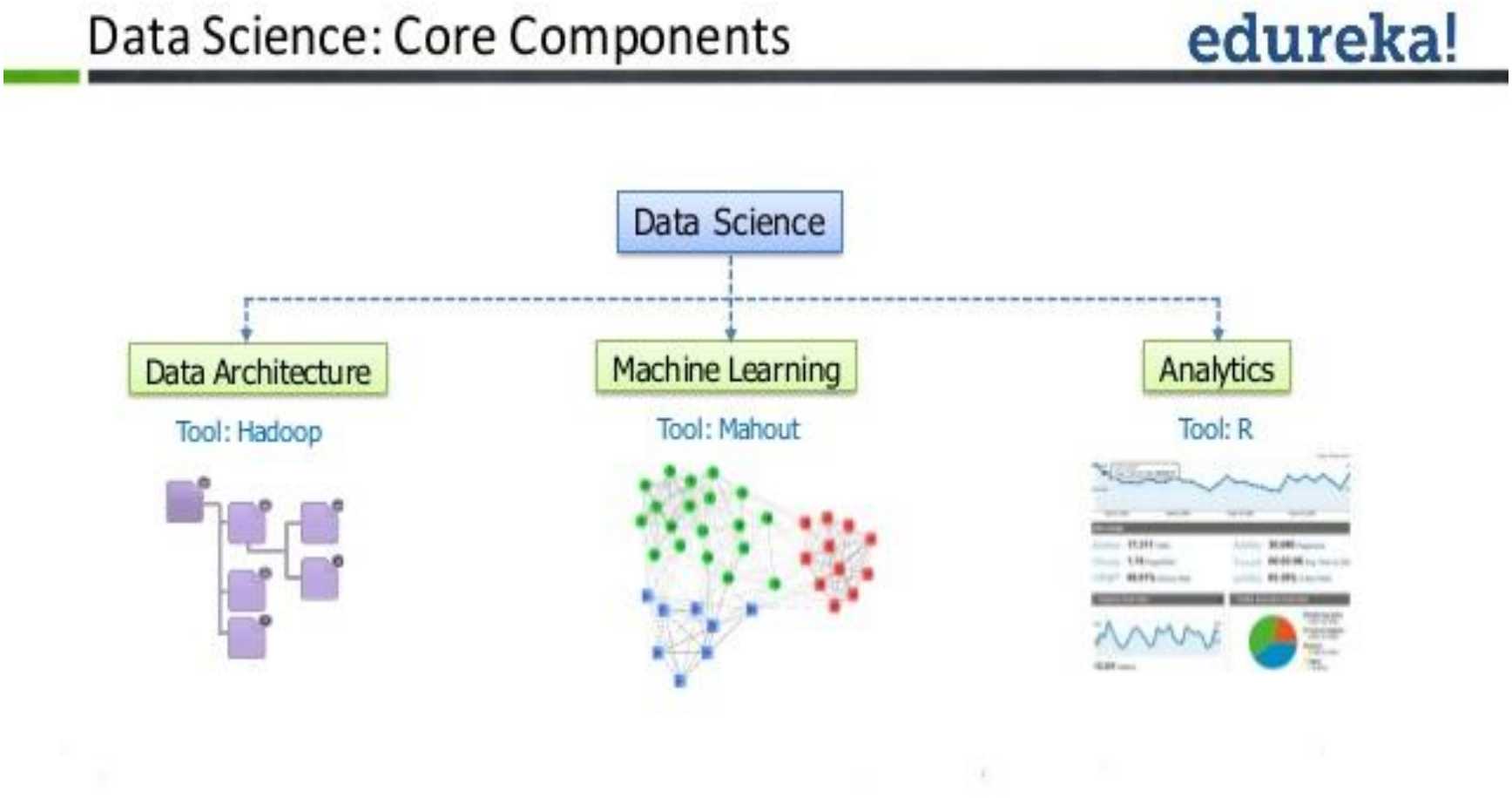
O maior crescimento é o de dados não estruturados (dentro e fora da empresa)

Exhibit 1 Examples of Unstructured Data

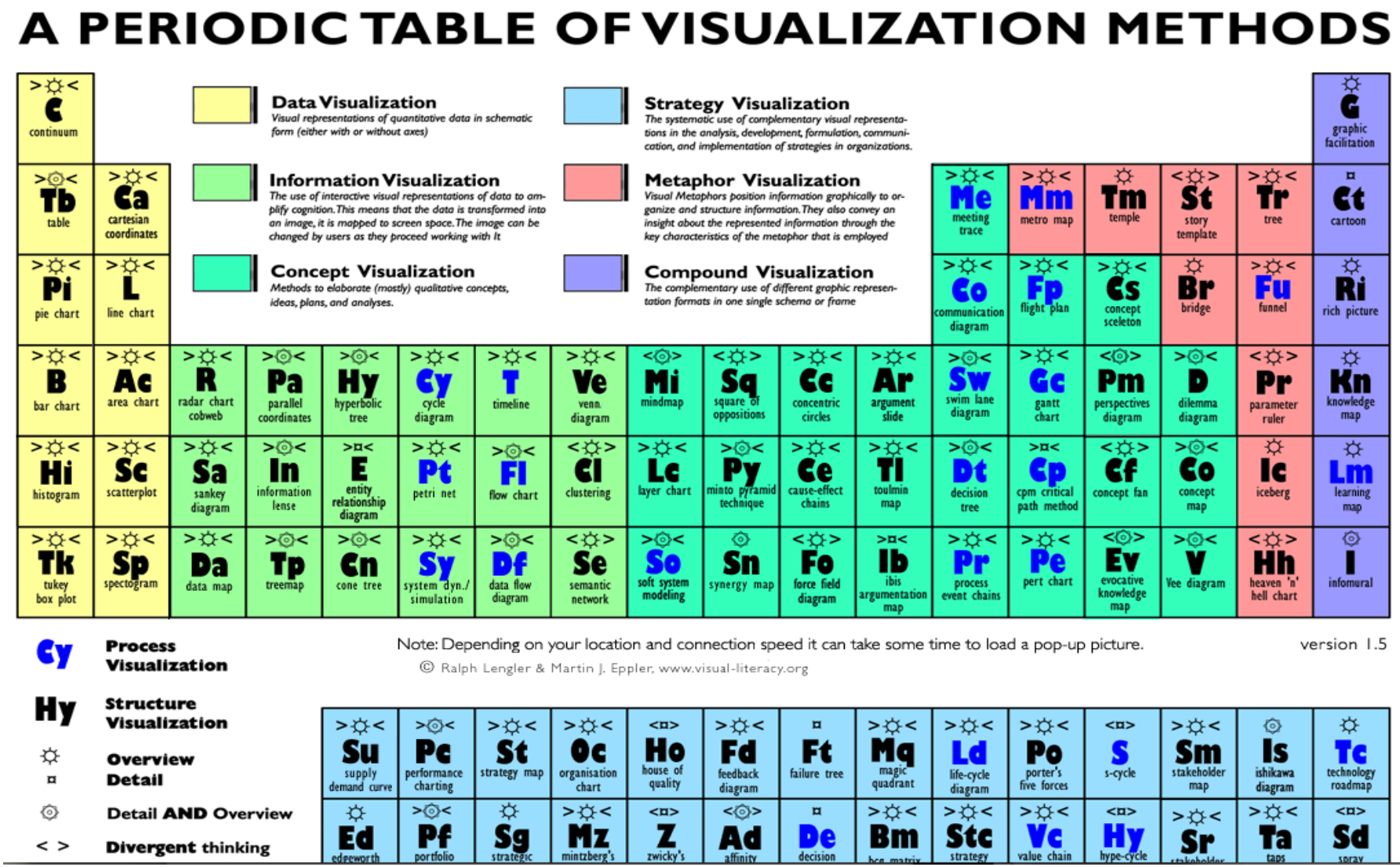
Corporate Enterprise Systems	Internet
• Corporate and personal email. • Text and instant messages. • Customer call logs, correspondence, and sales rep comments. • Voice mail and phone logs. • Documents, presentations, spreadsheets, and reports. • Encrypted files and messages. • Images and audio and video files. • Calendars and contacts. • Cellphone and vehicle location data. • Internet histories, email-file transmission, and system access logs.	• Social network profiles such as Facebook, LinkedIn, Twitter, and Google Plus. • Social influence sites such as expert blogs, user forums, industry-specific sites, and consumer review sites (Amazon, ZDNet, and Zillow). • Activity-generated data such as location information. • Government, publisher, and aggregator databases, wikis. • Newspapers, articles, and white papers. • Websites.

http://www.journalofaccountancy.com/issues/2014/jan/20126972.html

Componentes (e ferramentas) da ciência de dados

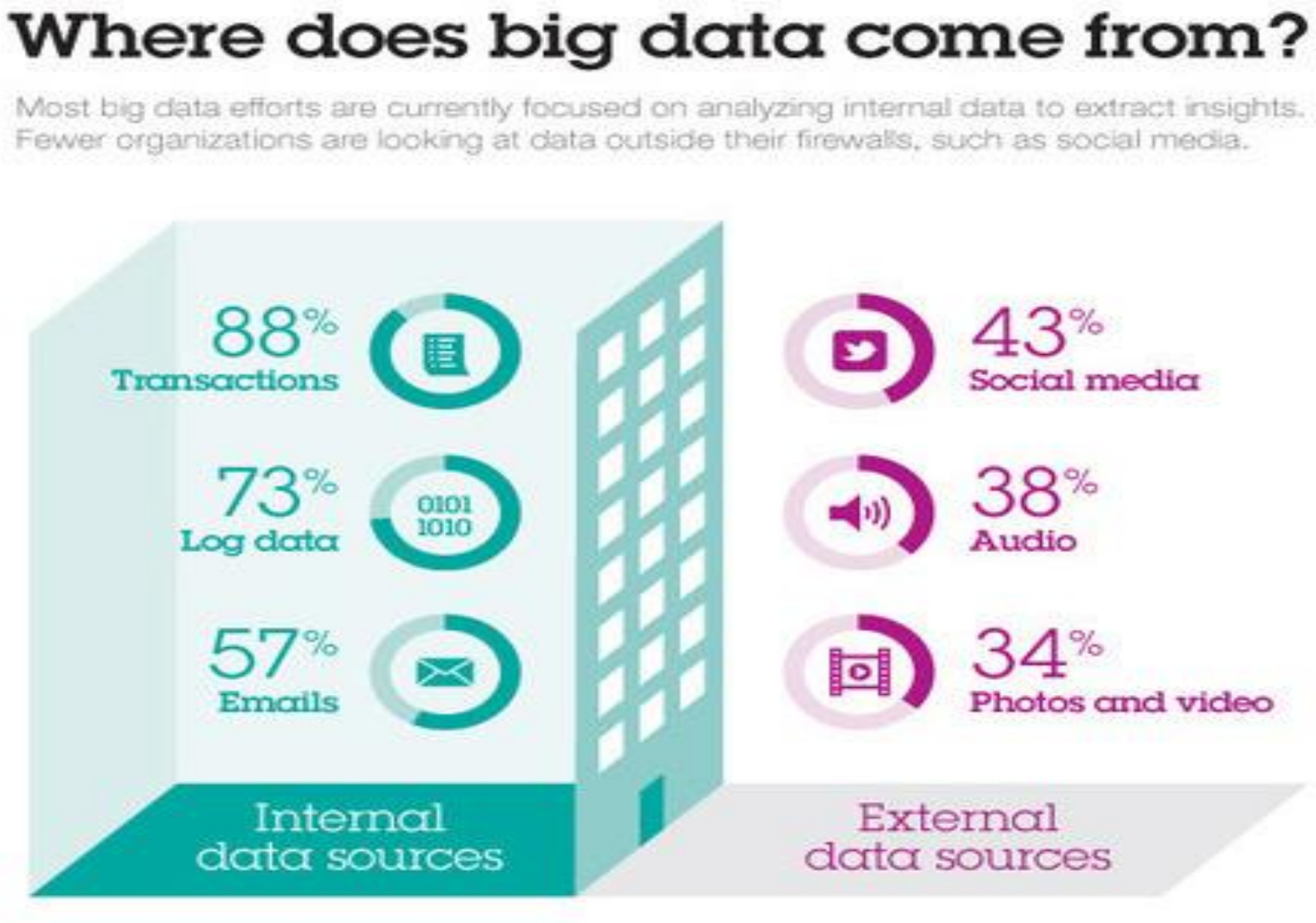


A visualização de dados é relevante



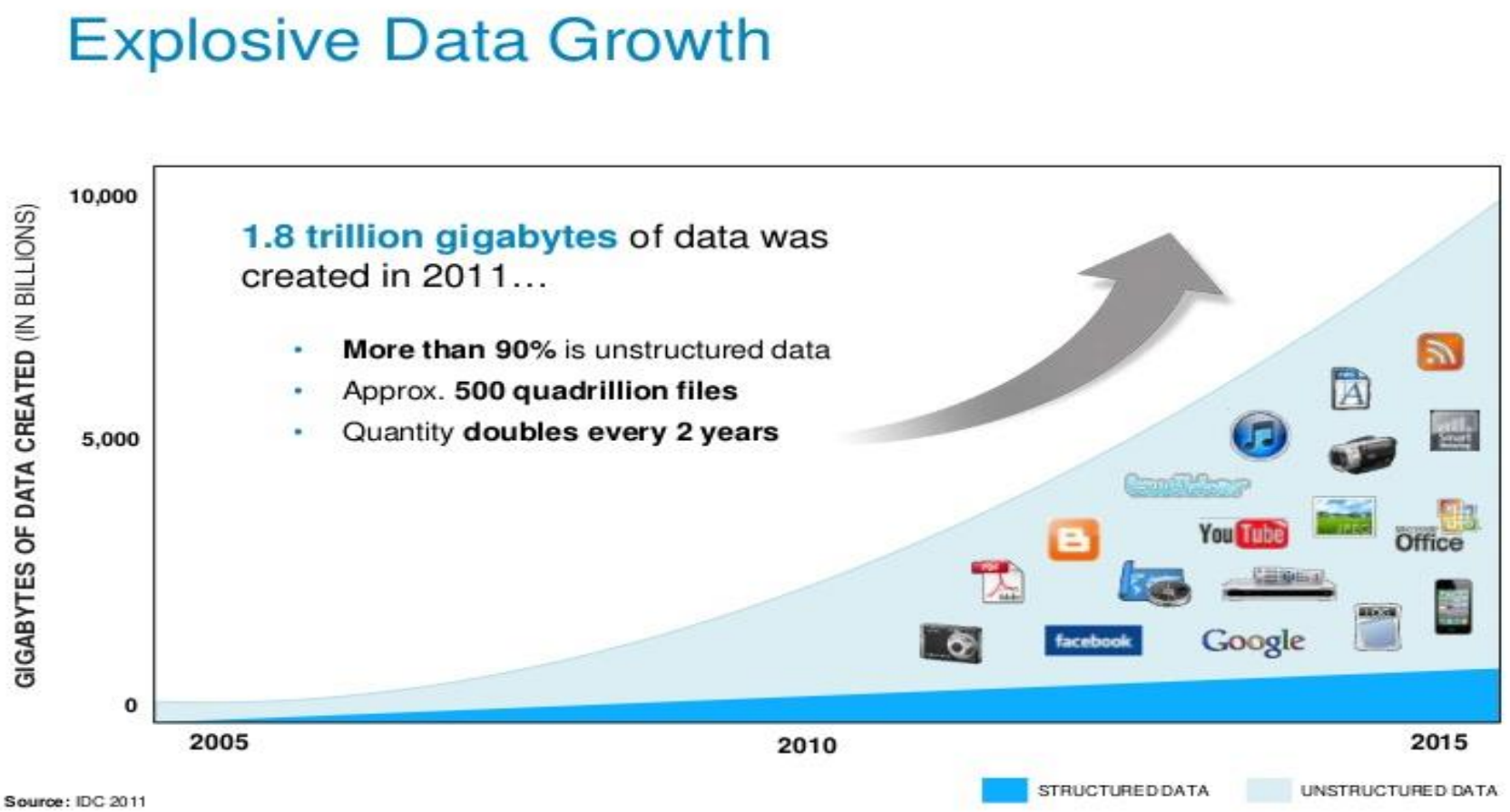
http://www.datasciencecentral.com/profiles/blogs/two-periodic-tables-for-data-scientists

As origens dos dados e o seu tipo



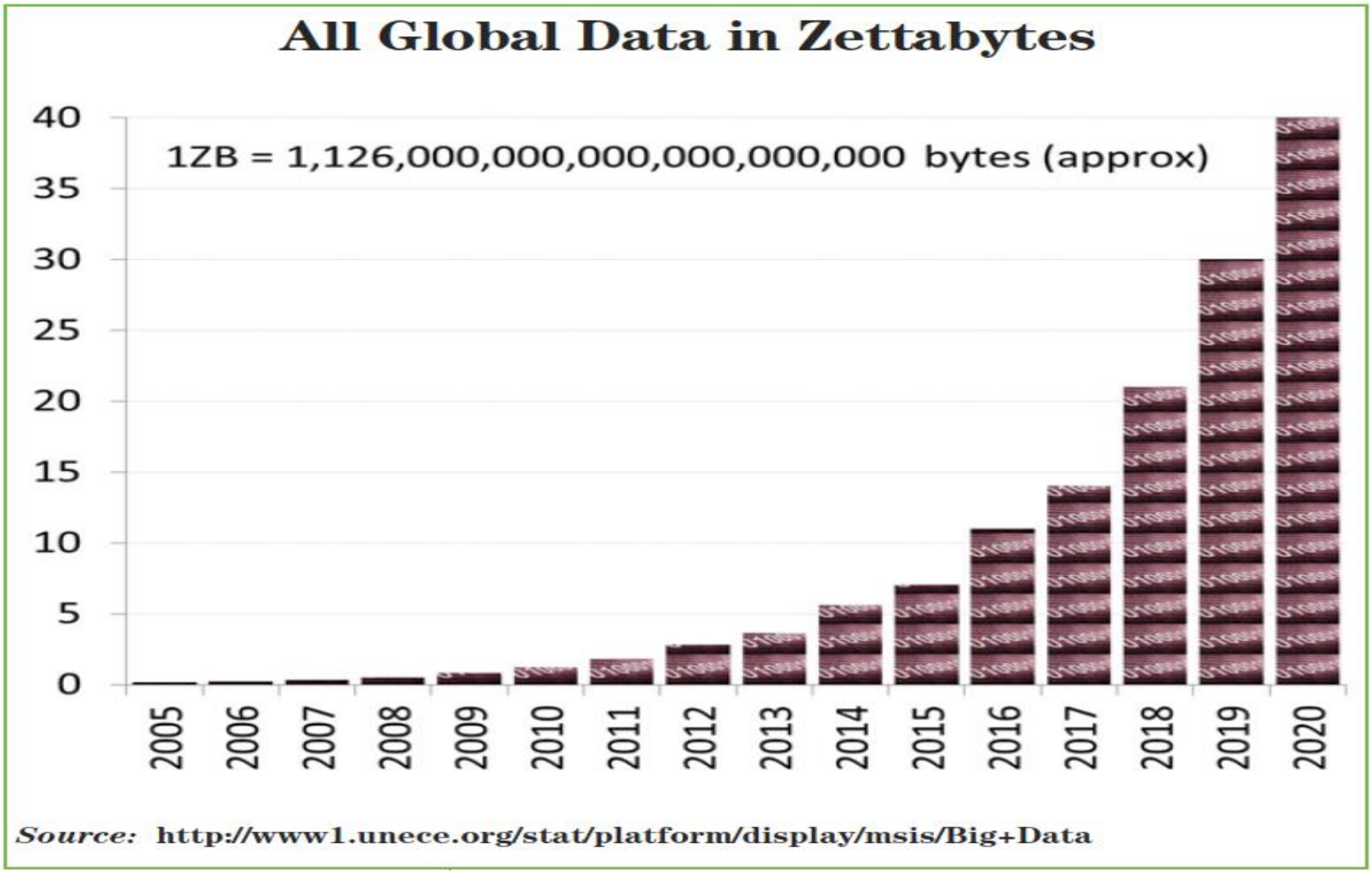
http://www.ibmbigdatahub.com/infographic/where-does-big-data-come

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

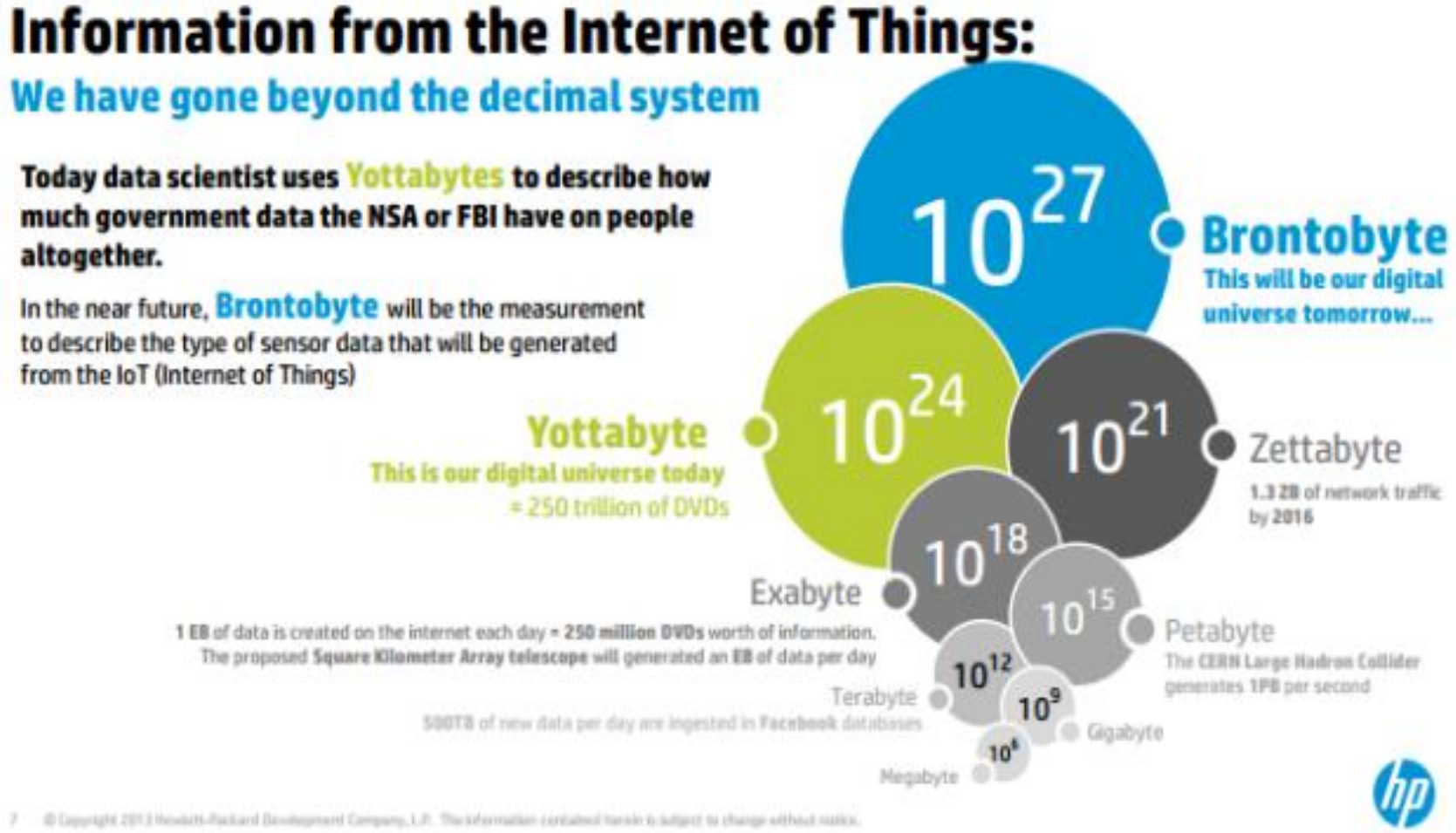


cloudera intel pentaho Big Data Solutions DELL

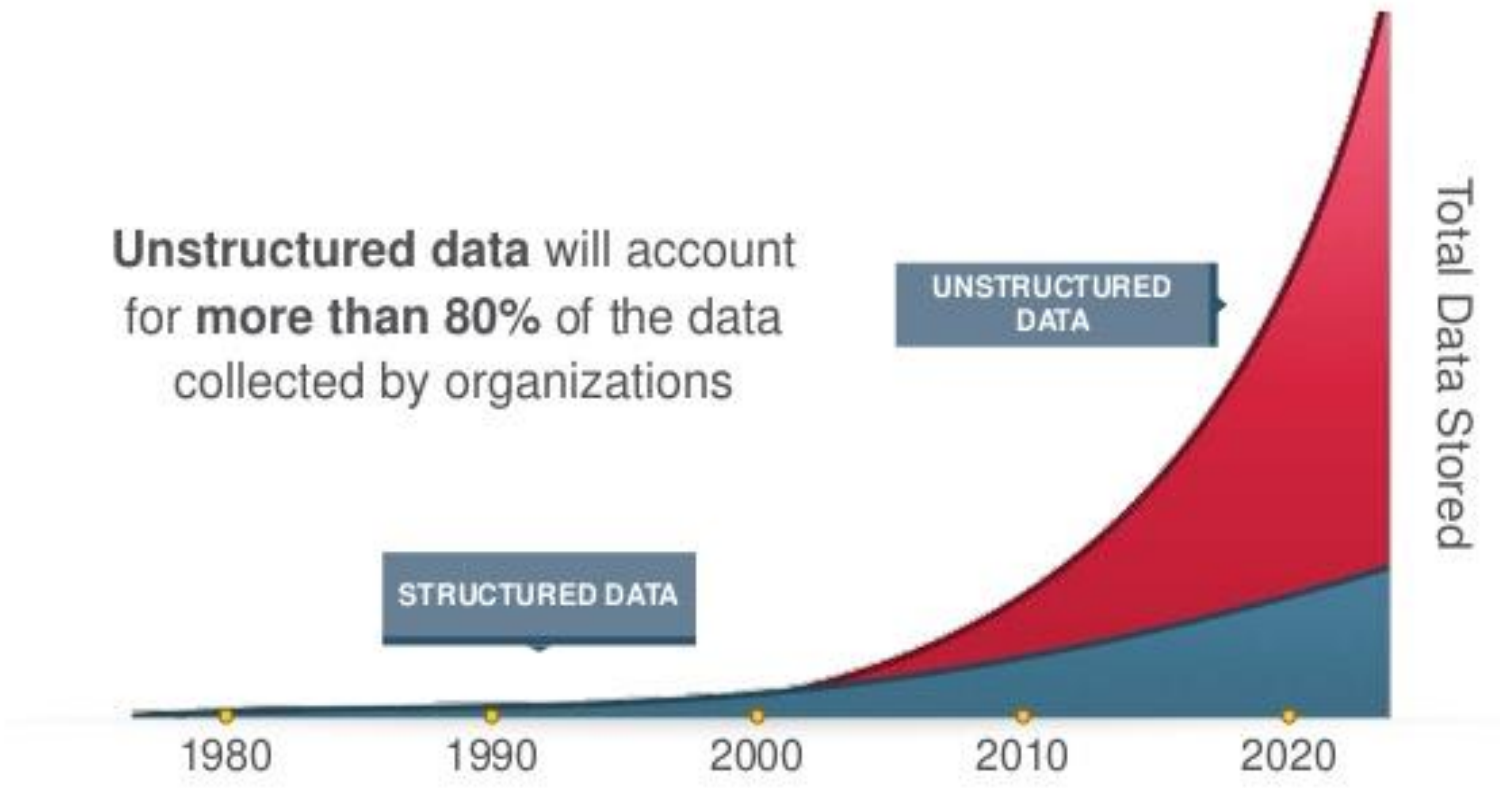
O que se pretende dizer por quantidades massivas de dados?



Até onde vai escalar o crescimento de dados?



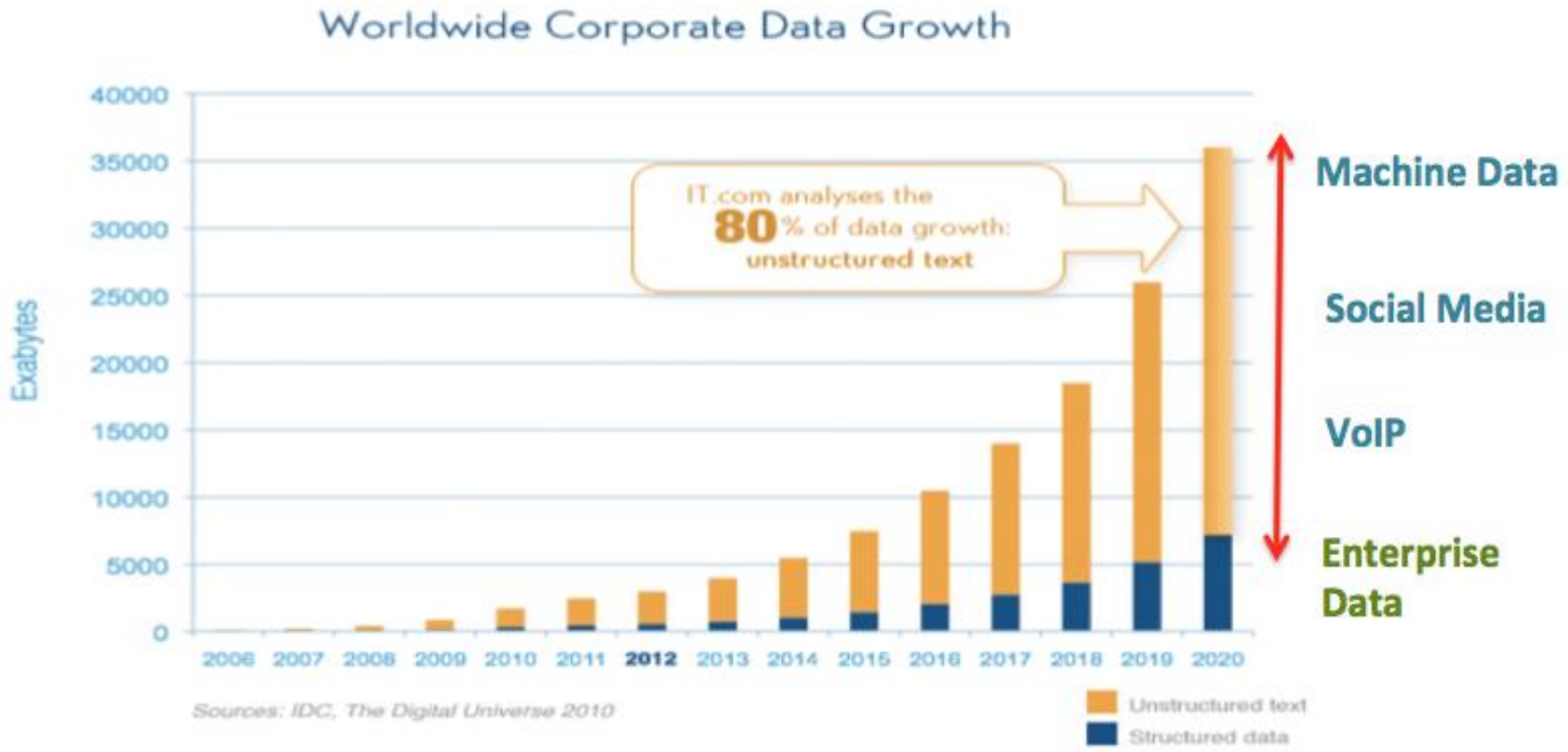
O crescimento dos dados não estruturados vai mudar o paradigma da sua gestão, nas empresas



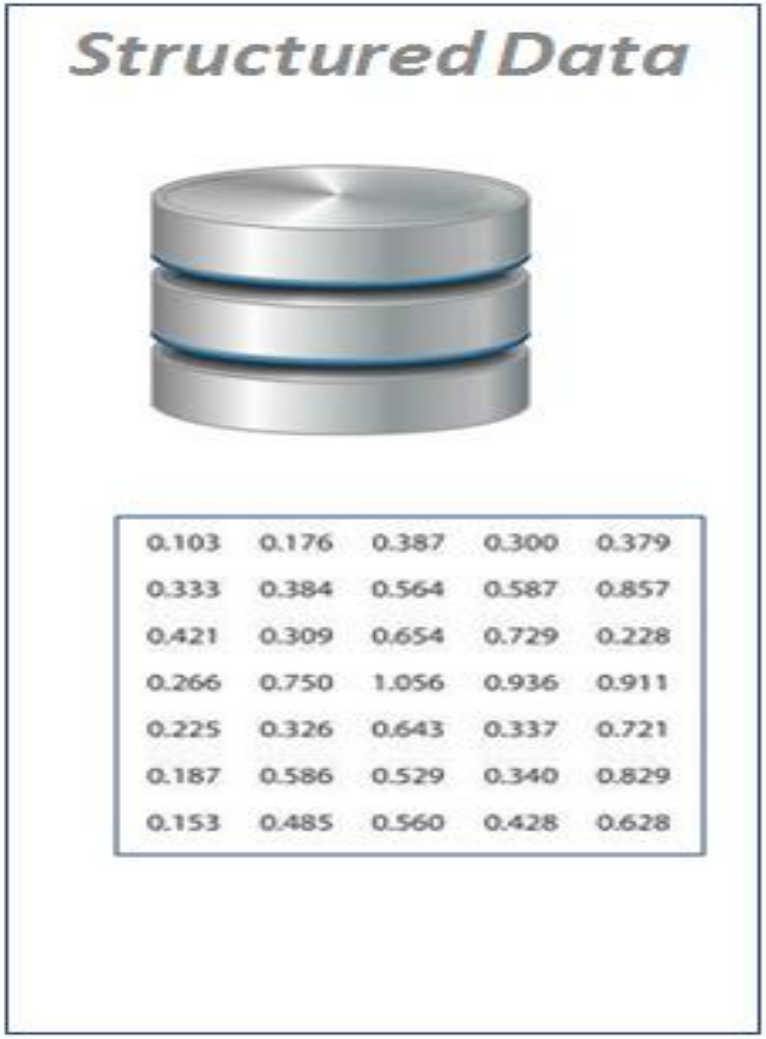
Source: Human-Computer Interaction & Knowledge Discovery in Complex Unstructured, Big Data

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).



Dados estruturados versus dados não estruturados (origem)



https://www.laserfiche.com/ecmblog/4-ways-to-manage-unstructured-data-with-ecm/

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

STRUCTURED VS. UNSTRUCTURED DATA

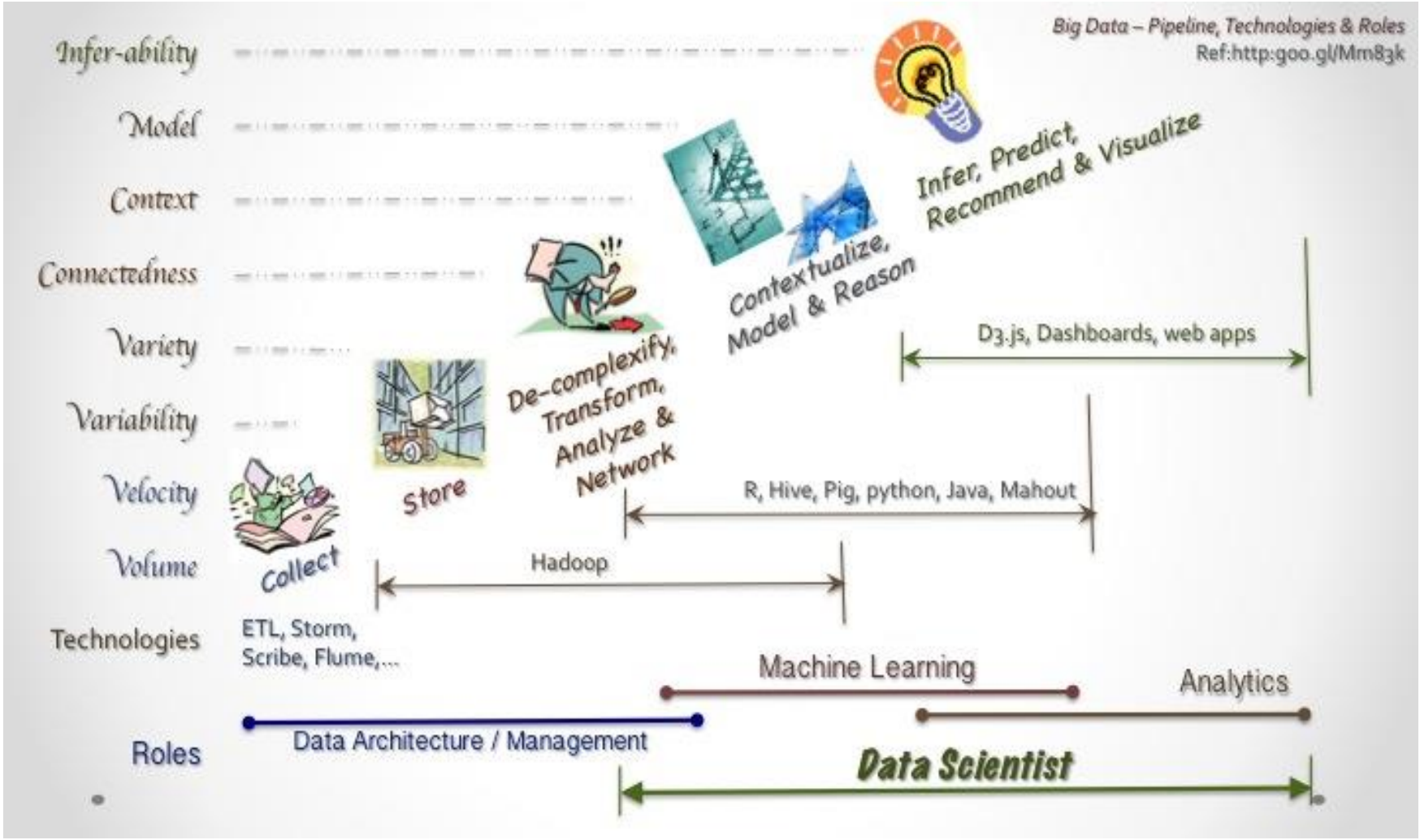
Structured Data
High Degree of organization, such as a relational database

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

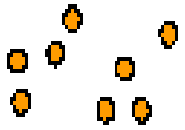
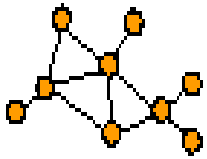
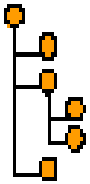
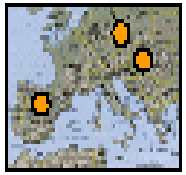
"The patient came in complaining of chest pain, shortness of breath, and lingering headaches…smokes 2 packs a day… family history of heart disease…has been experiencing similar symptoms for the past 12 hours…."

Papeis, competências e atividades no contexto da ciência de dados



https://doubleclix.wordpress.com/2012/12/15/what-or-who-is-a-data-scientist/

Relacionar dados por recurso a modelos apropriados

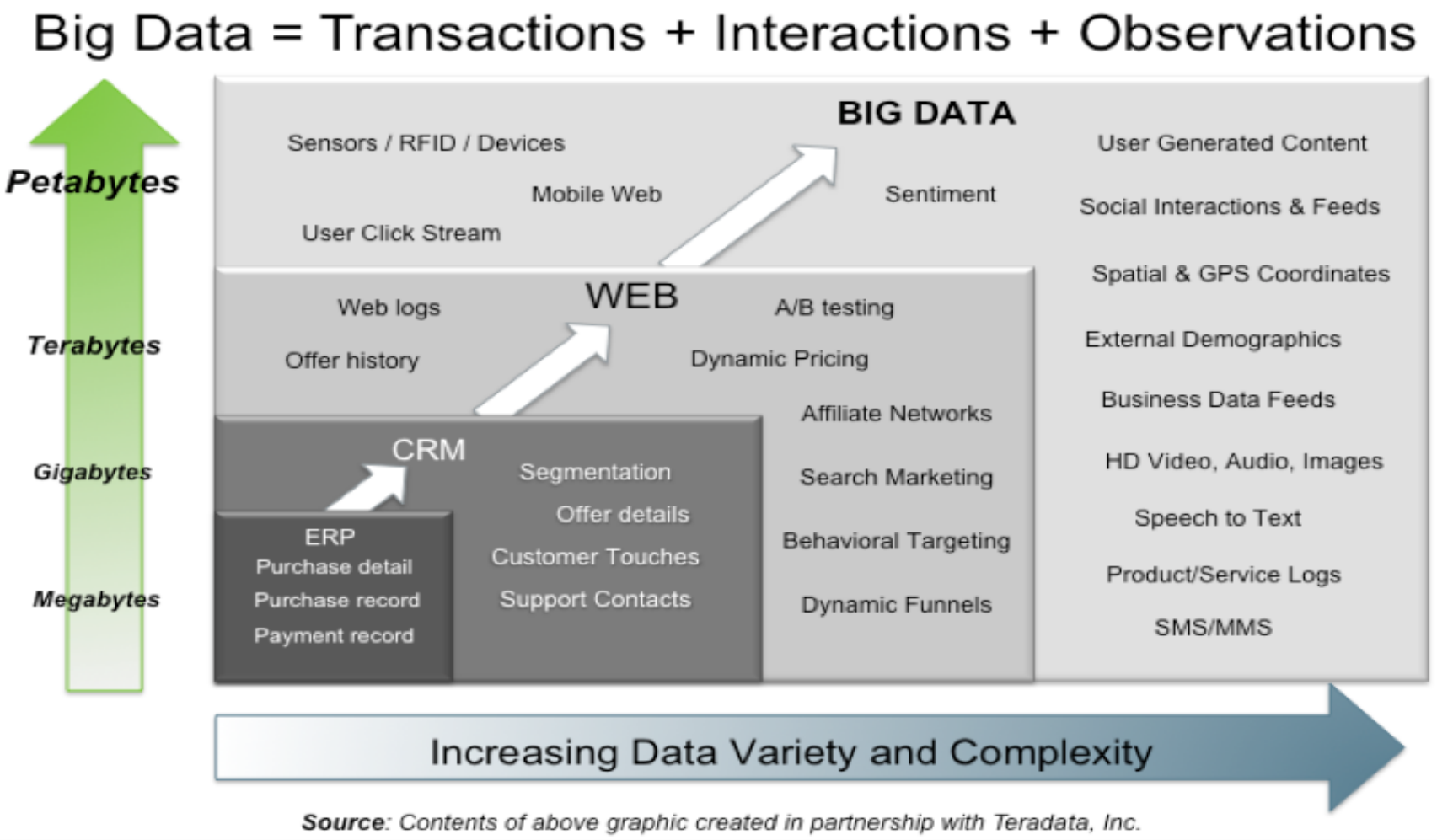
Relationship Type:	General Similarity	Explicit Reference	Field/Value Co-occurrence	Parent/Child	Spatial	Temporal
Model Type:	*Vector-space* 	*Network* 	*Multidimensional Index* 	*Hierarchical* 	*Spatial* 	*Ordinal Index* 
Examples:	*Reports, articles, DB records*	*References & citations, hyperlinks*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

As diferenças entre inteligência de negócios e a ciência de dados

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

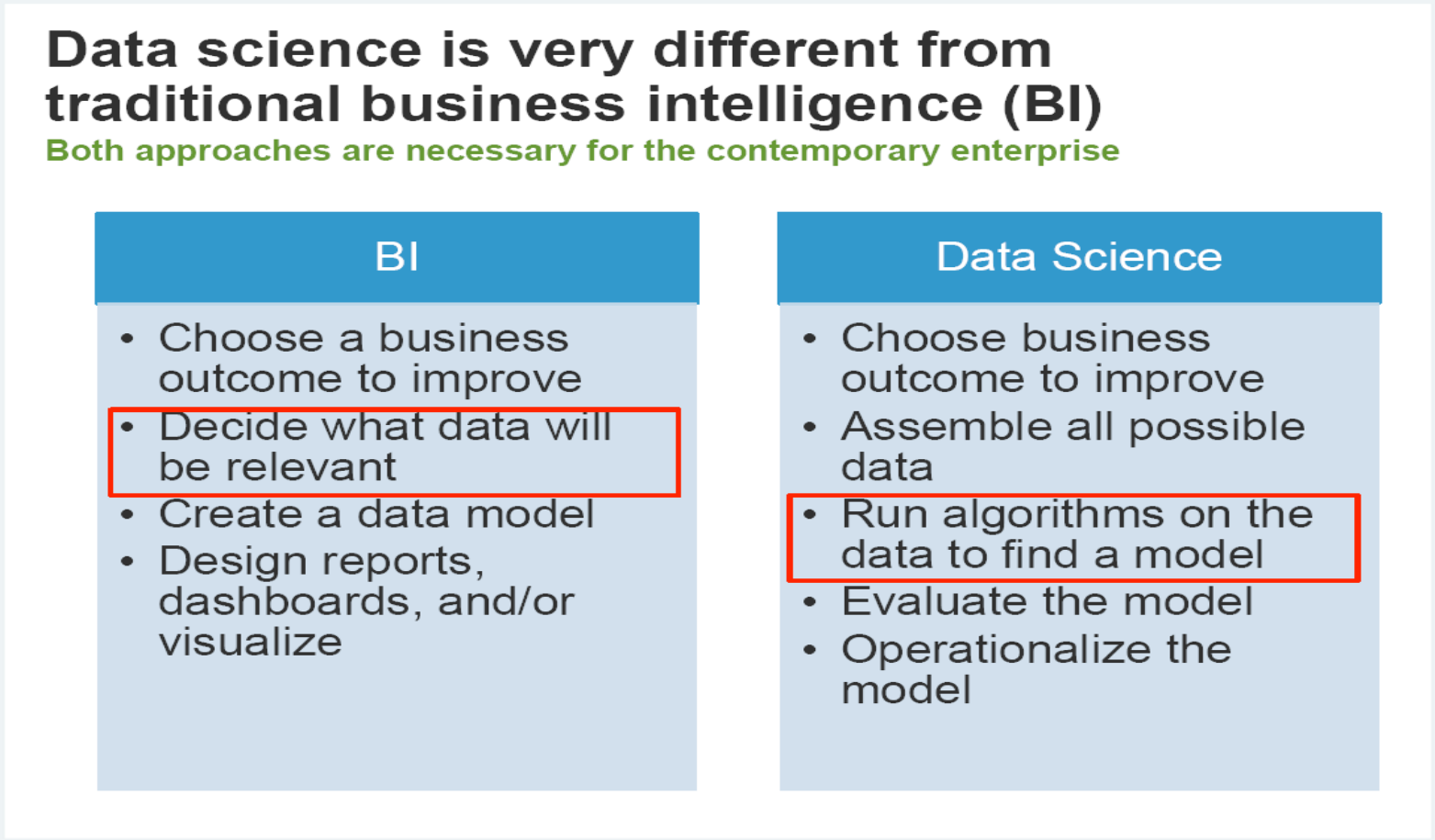
http://blog.revolutionanalytics.com/2013/05/statistics-vs-data-science-vs-bi.html

Lidar com dados massivos exige novos tipos de respostas pelas empresas



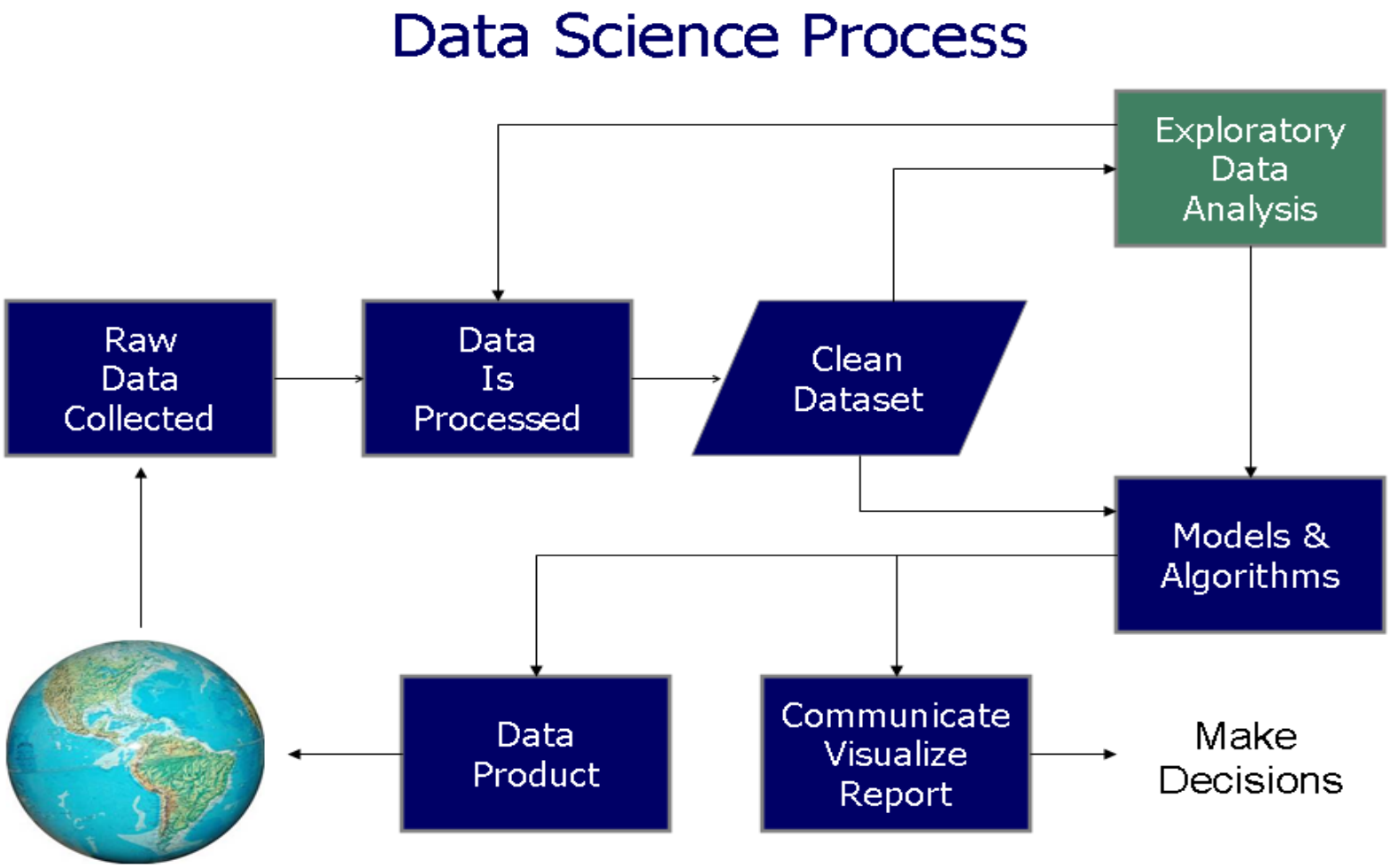
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

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



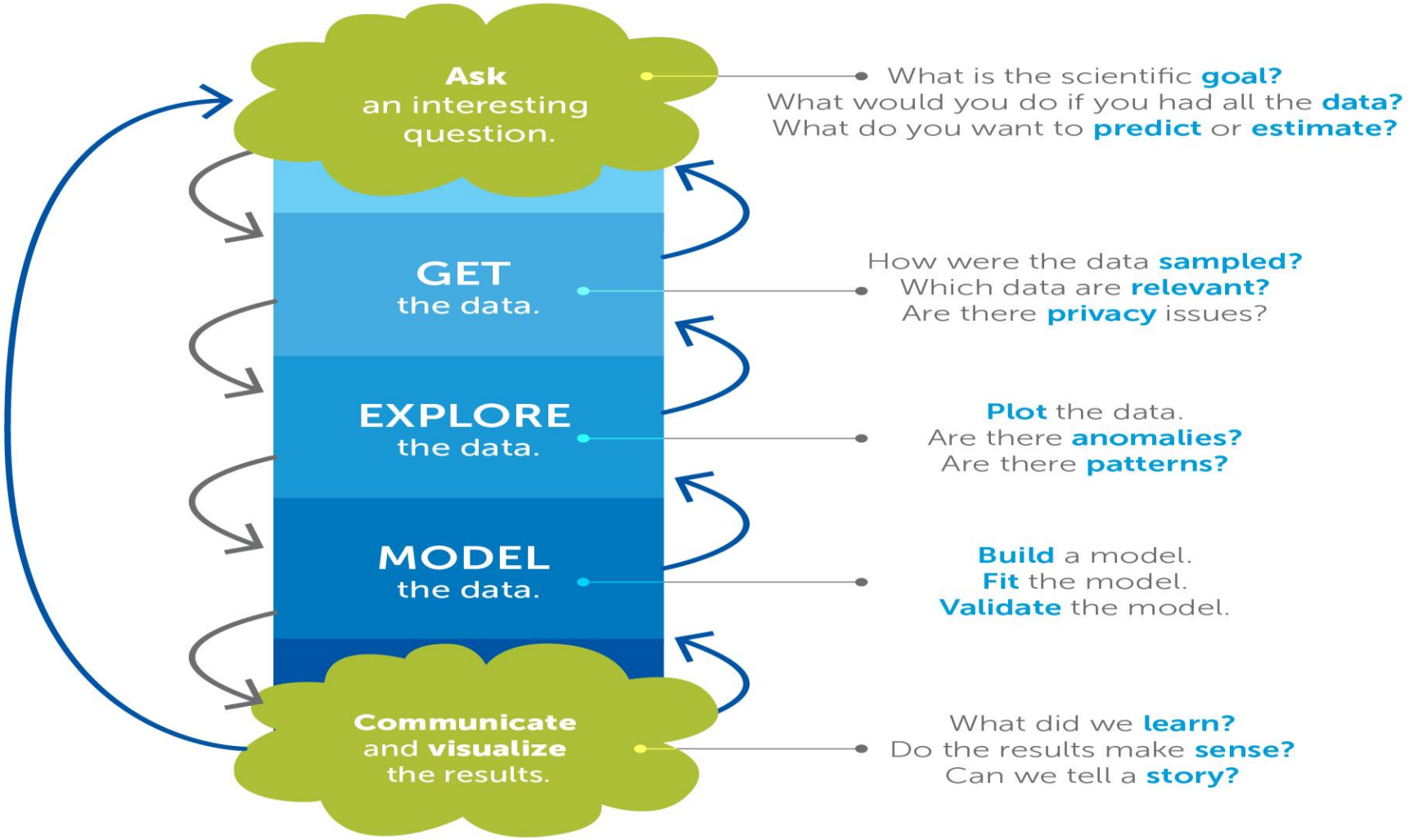
http://semanticommunity.info/Big_Data_Technology_Symposium

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



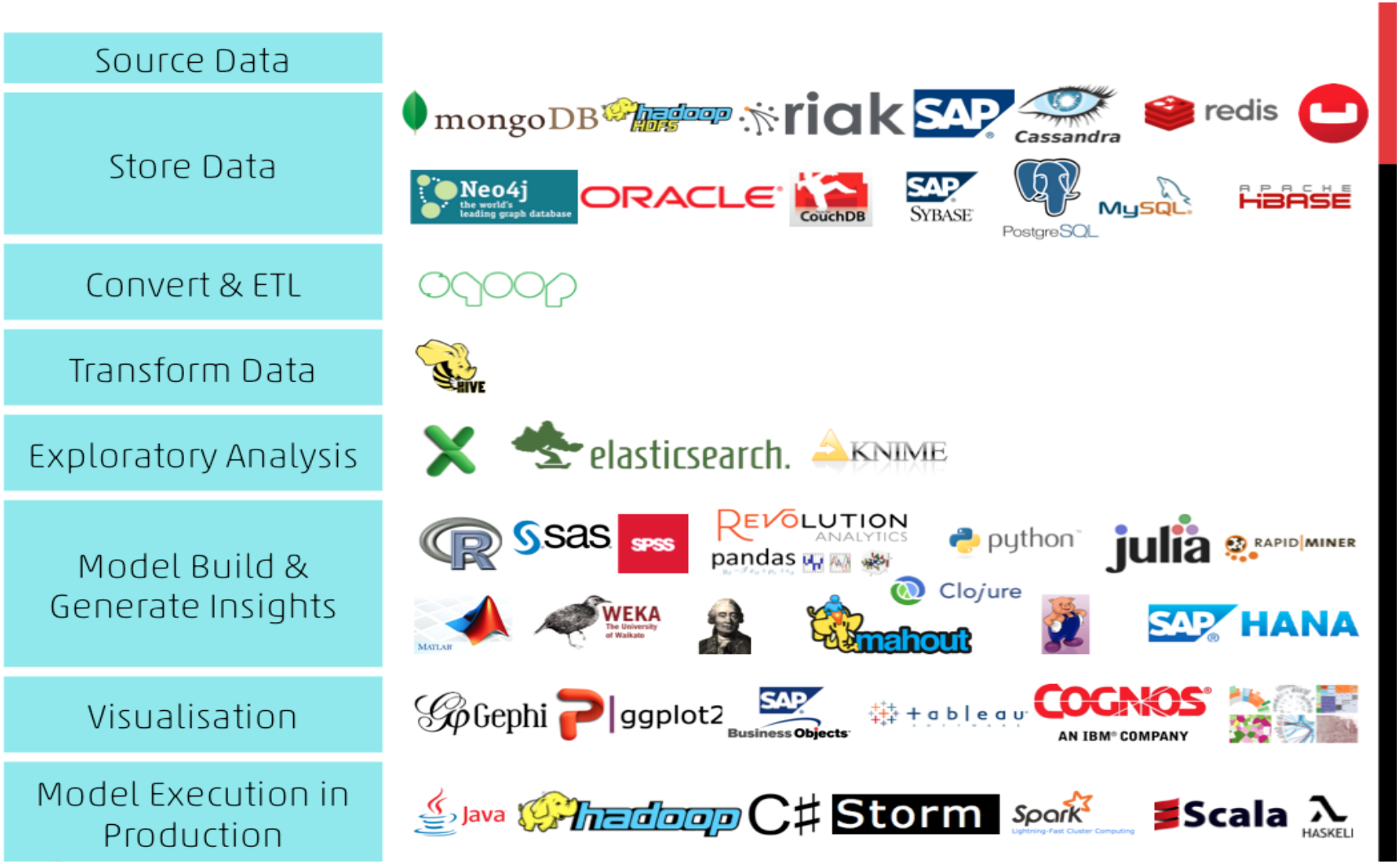
https://en.wikipedia.org/wiki/Data_science#/media/File:Data_visualization_process_v1.png

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



https://becomingadatascientist.wordpress.com/2013/07/26/choosing-a-data-science-technology-stack-w-survey/

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 enfase 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

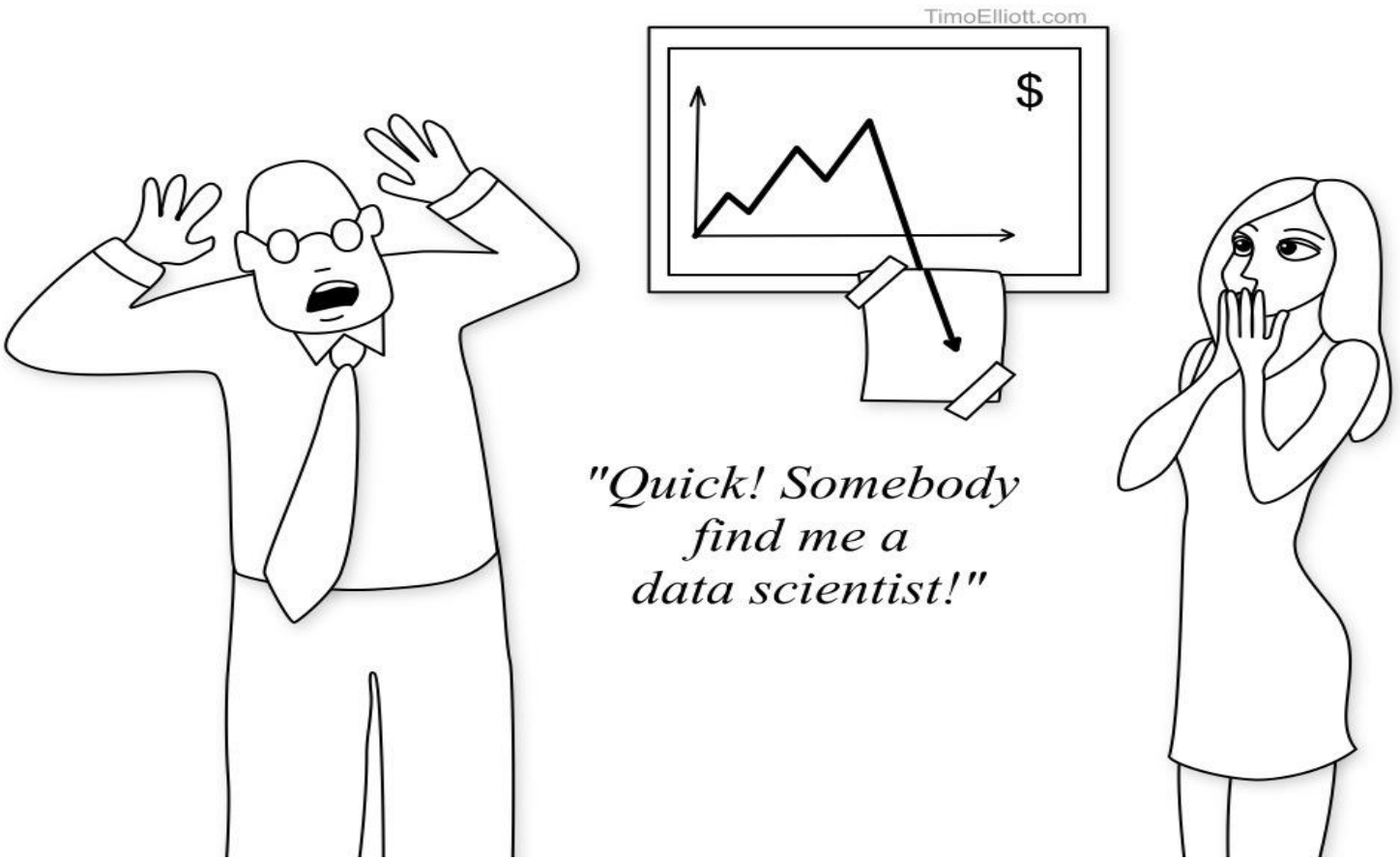
Tarefas típicas de profissionais de ciência de dados

- 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



http://timoelliott.com/blog/2013/08/what-is-a-data-scientist-and-what-do-they-do-cartoons.html