CDS6214

Data Science Fundamentals

Lecture 1
Introduction and Overview

MMU
MULTIMEDIA UNIVERSITY

big data

data science

Big Data for real-world insights

- Big Data in the Hospitality Industry
  https://www.youtube.com/watch?v=mK1stwMHb7Y
- How Big Data Could Transform The Health Care Industry
  https://www.youtube.com/watch?v=_mXrZEIpNMw
- Big Data in Telefónica: Dynamic Insights
  https://www.youtube.com/watch?v=APDjX3cZ7Ps
- How video analytics can improve retail customer experience
  https://www.youtube.com/watch?v=o0kIIGC4Fuw

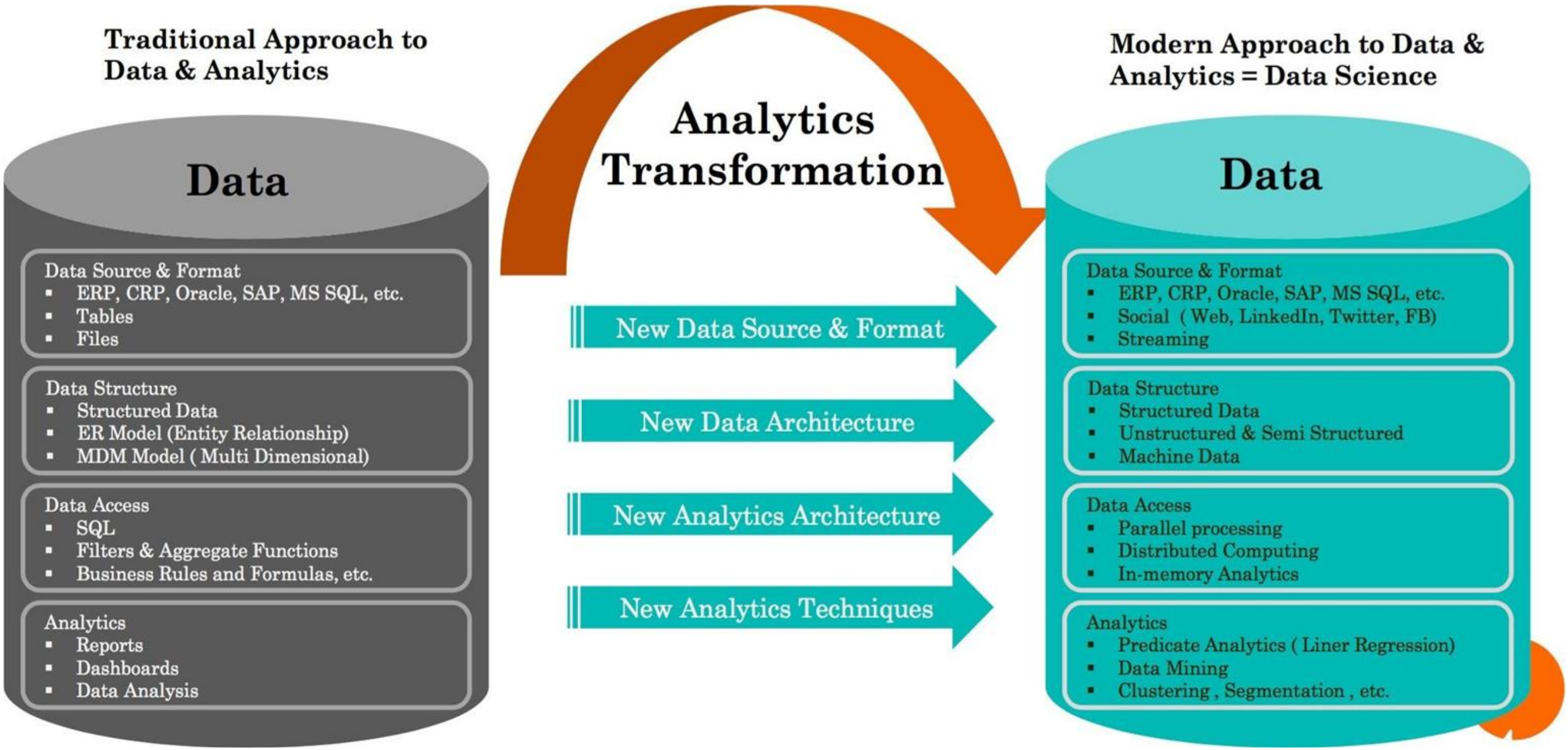
Big Data

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

C.B.B.D.Manyika,“BigData:TheNextFrontierforInnovation,Competition,andProductivity,” McKinsey Global Institute, 2011

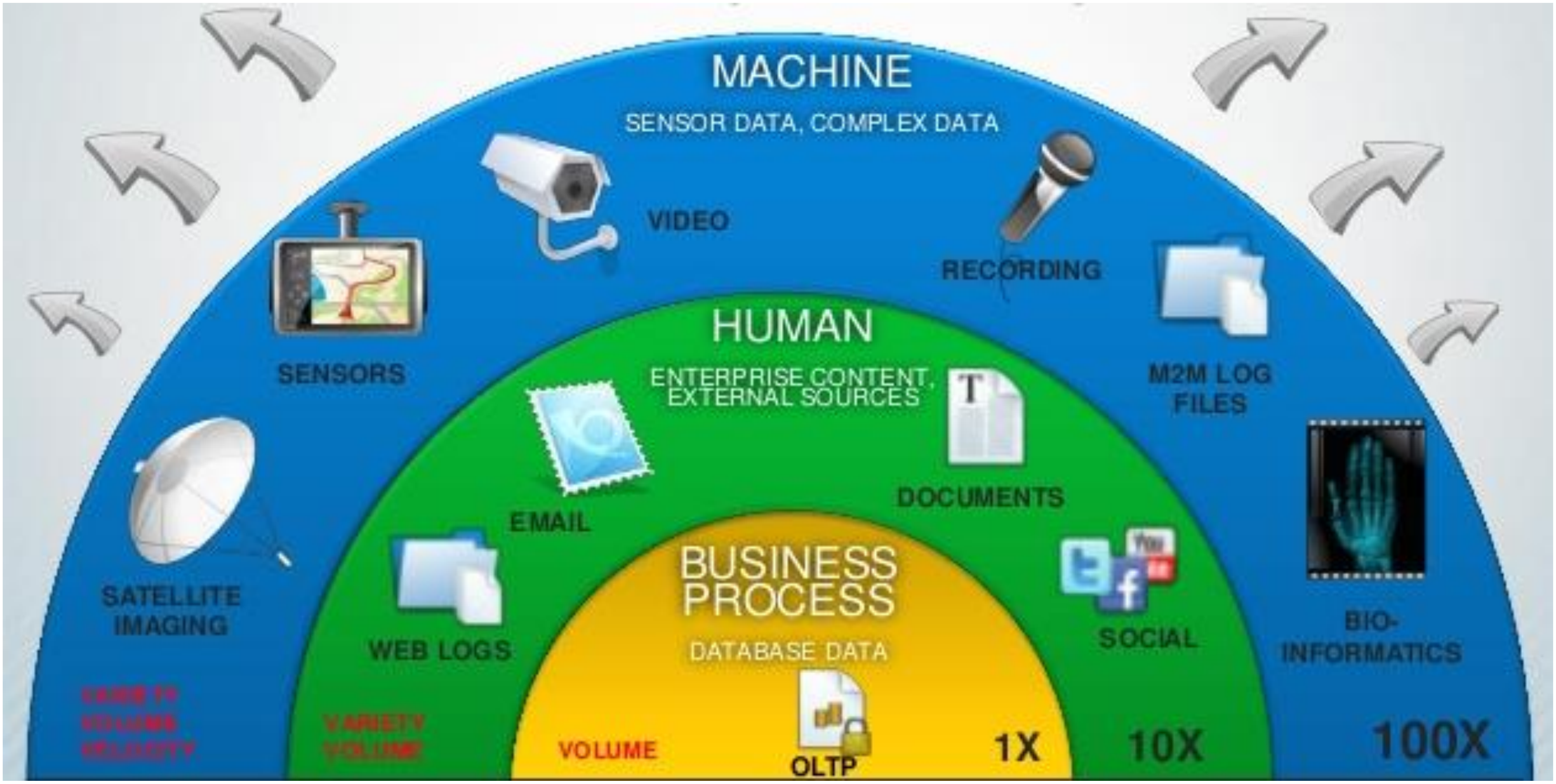
http://www.biznetsoftware.com/big-data-can-biznet-help-manage/

Big Data (2)



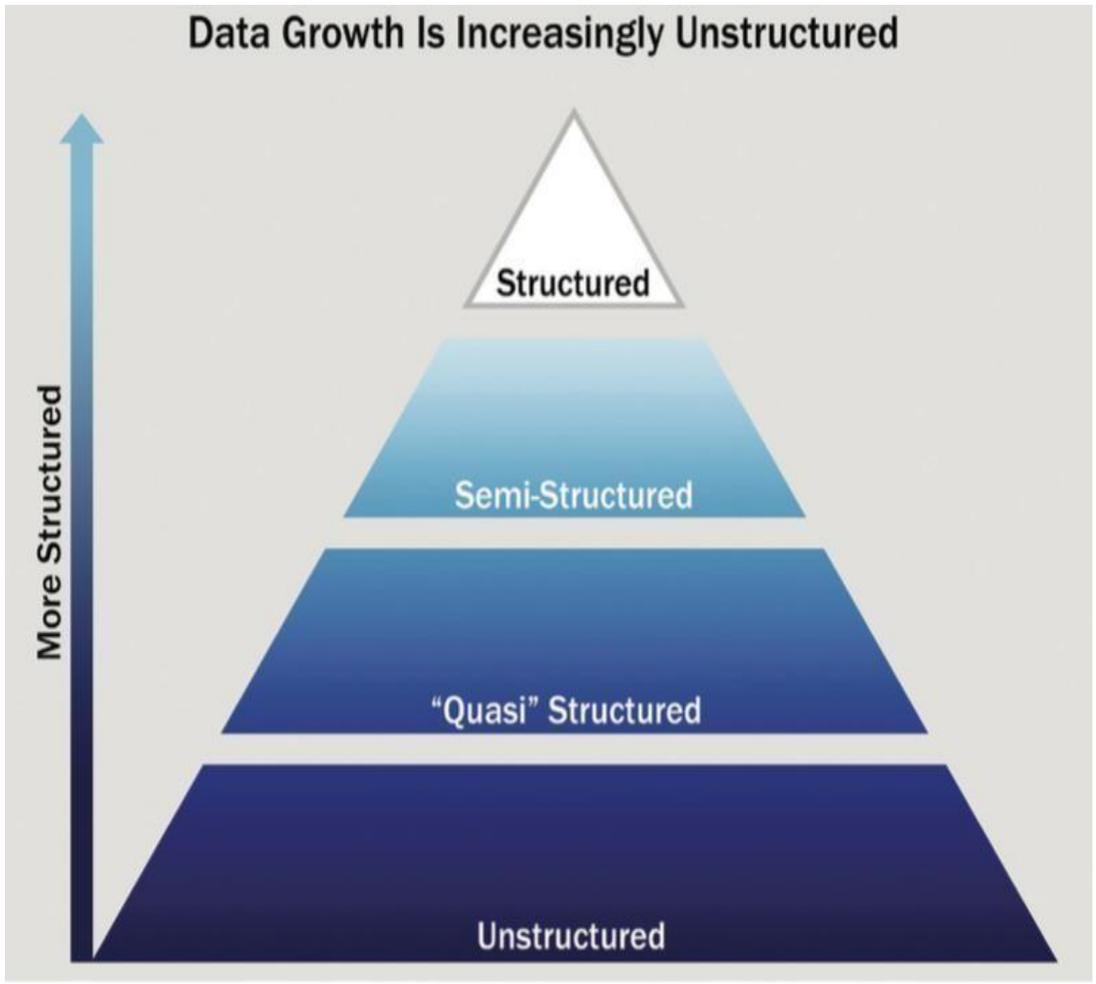
http://www.slideshare.net/petekua/mdec-data-science-vs-data-scientist-v2-04212015

Big Data Sources



http://www.slideshare.net/bjorna/big-data-in-oil-and-gas

Data Structures



- **Structured data:**

Data containing a defined data type, format, and structure(e.g. online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spread- sheets).

- **Semi-structured data:**

Textual data files with a discernible pattern that enables parsing (e.g. XML data files that are self-describing and defined by an XML schema).

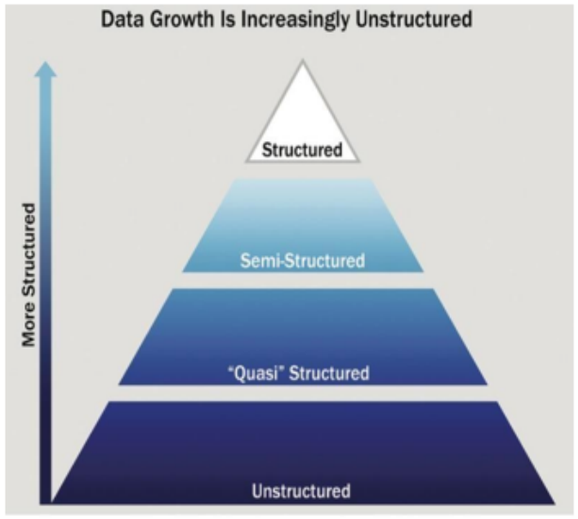
- **Quasi-structured data:**

Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).

- **Unstructured data:**

Data that has no inherent structure, which may include text documents, PDFs, images, and video.

Data Structures



- **Structured data:**
Data containing a defined data type, format, and structure(e.g. online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spread- sheets).

- **Semi-structured data:**
Textual data files with a discernible pattern that enables parsing (e.g. XML data files that are self-describing and defined by an XML schema).

- **Quasi-structured data:**
Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).

- **Unstructured data:**
Data that has no inherent structure, which may include text documents, PDFs, images, and video.

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

```
<SampleXML>
  <Colors>
    <Color1>White</Color1>
    <Color2>Blue</Color2>
    <Color3>Black</Color3>
    <Color4 Special="Light">Green</Color4>
    <Color5>Red</Color5>
  </Colors>
  <Fruits>
    <Fruits1>Apple</Fruits1>
    <Fruits2>Pineapple</Fruits2>
    <Fruits3>Grapes</Fruits3>
    <Fruits4>Melon</Fruits4>
  </Fruits>
</SampleXML>
```

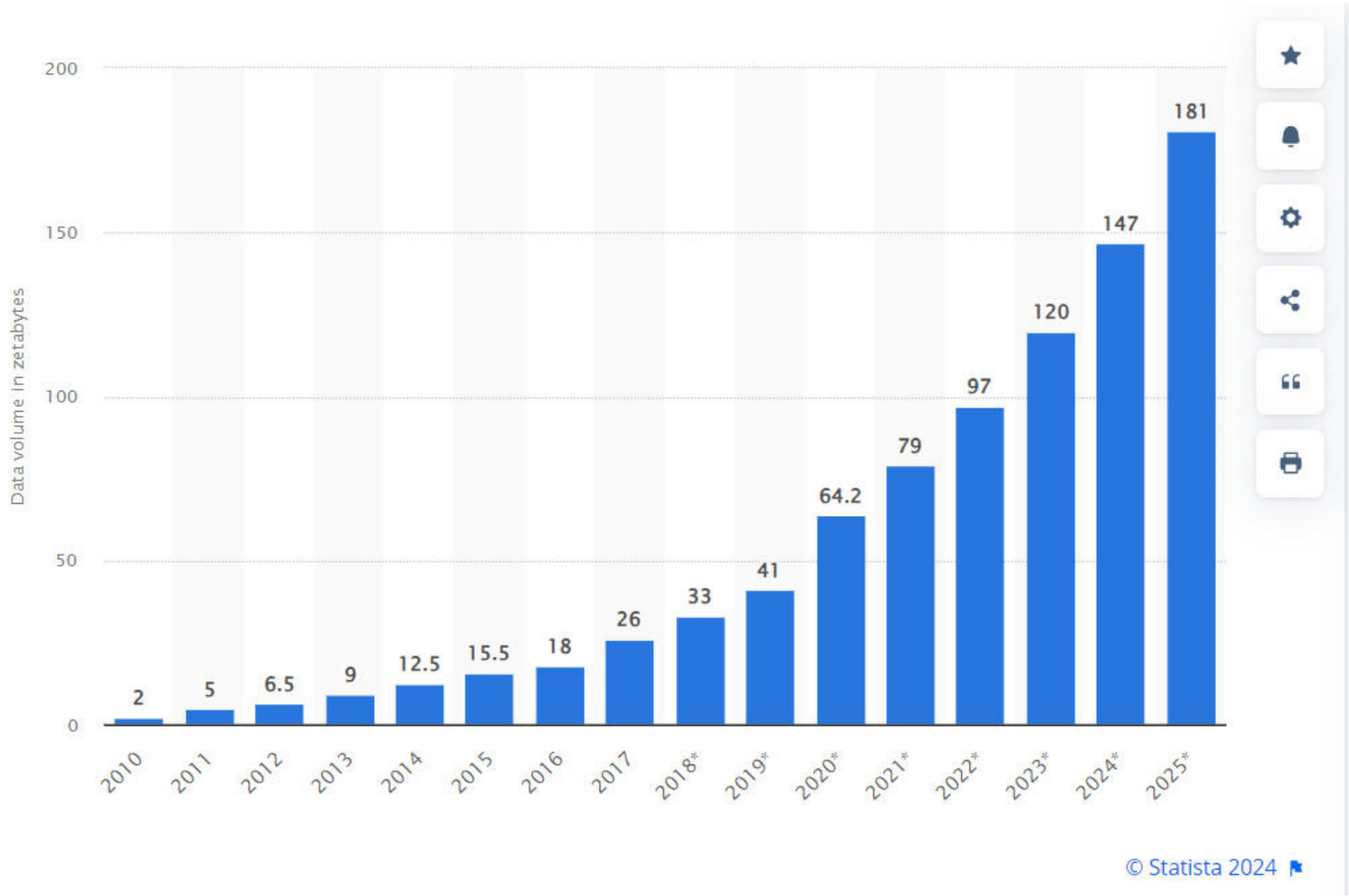
```
unix time ;IP address     ; session ID                              ; page request; referee

1074589200;193.179.144.2 ;1993441e8a0a4d7a4407ed9554b64ed1;/dp/?id=124   ;www.google.cz;
1074589201;194.213.35.234;3995b2c0599f1782e2b40582823b1c94;/dp/?id=182   ;
1074589202;194.138.39.56 ;2fd3213f2edaf82b27562d28a2a747aa;/              ;www.seznam.cz;
1074589233;193.179.144.2 ;1993441e8a0a4d7a4407ed9554b64ed1;/dp/?id=148   ;/dp/?id=124;
1074589245;193.179.144.2 ;1993441e8a0a4d7a4407ed9554b64ed1;/sb/           ;/dp/?id=148;
1074589248;194.138.39.56 ;2fd3213f2edaf82b27562d28a2a747aa;/contacts/     ; /;
1074589290;193.179.144.2 ;1993441e8a0a4d7a4407ed9554b64ed1;/sb/           ;/sb/;
```

Big Data Growth

Key enablers for the growth of “Big Data” are:

- Increase of storage capacities
- Increase of processing power
- Availability of data

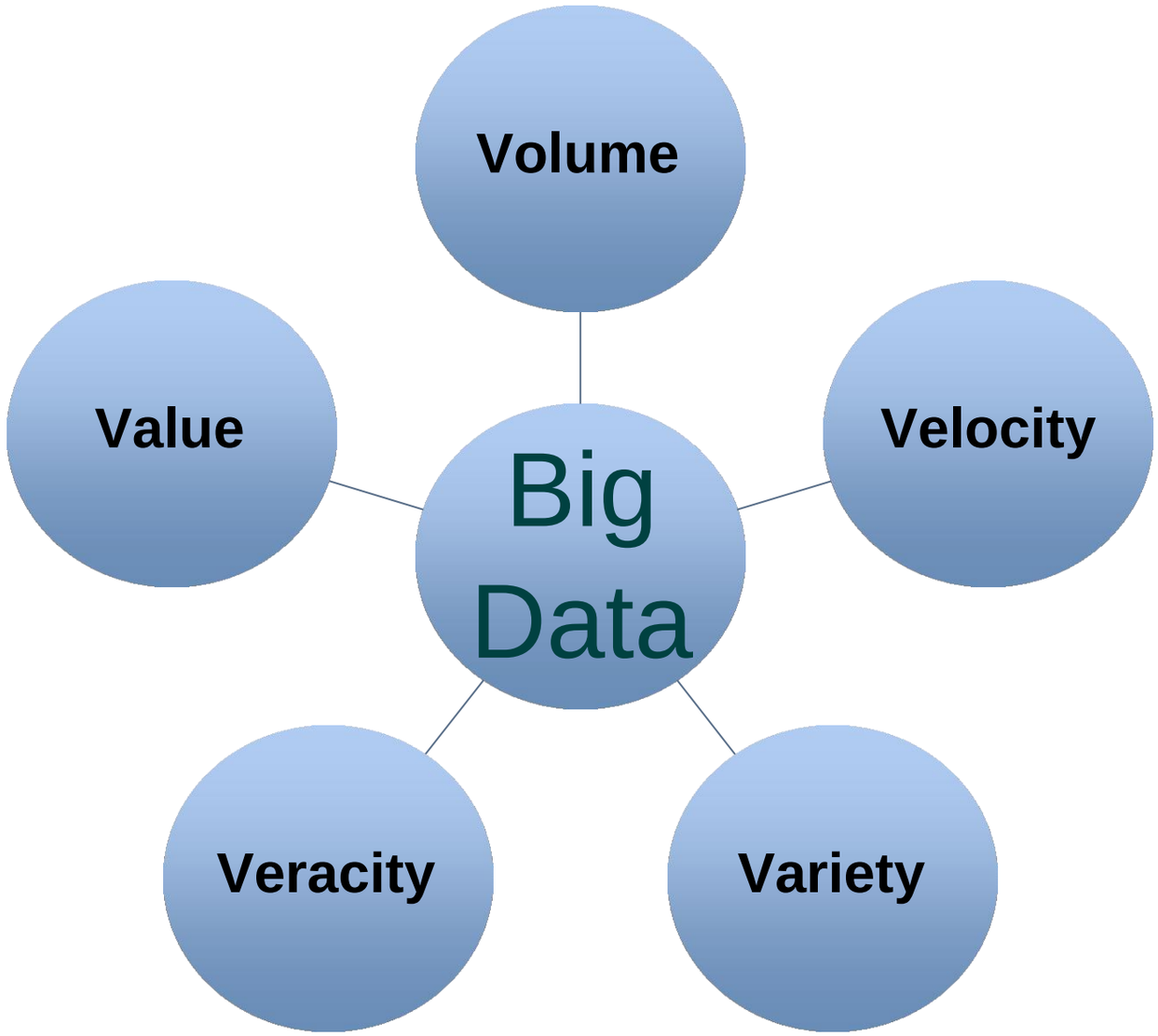


Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025(in zettabytes)

https://www.statista.com/statistics/871513/worldwide-data-created/

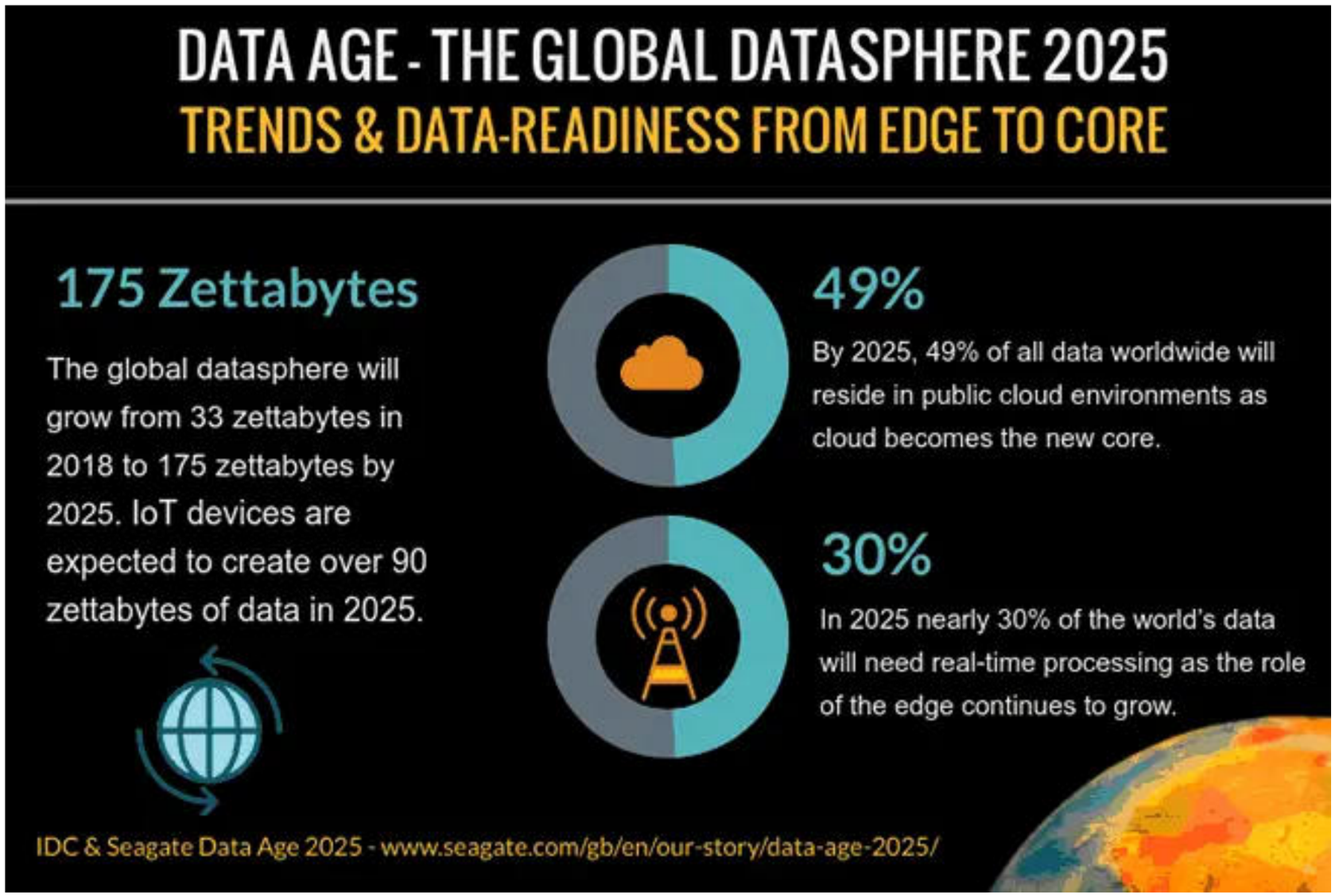
Characterization of Big Data

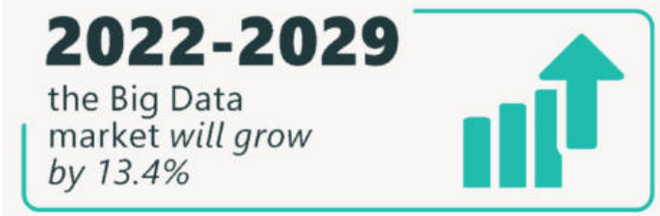
There are 5Vs often used to describe the characteristics of Big Data



BD Characterization: Volume (1)

Volume: the vast amount of data generated each second (Scale of data)





BD Characterization: Volume (2)

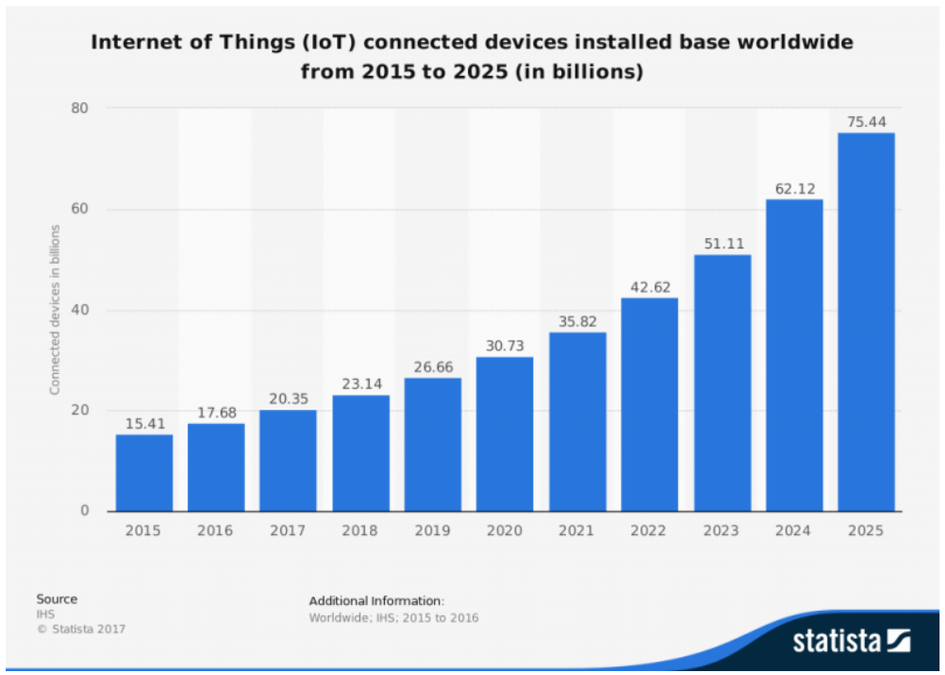
- 1 Gigabyte (GB) = 1,000,000,000 byte
- 1 Terabyte (TB) = 1,000 Gigabyte (GB)
- 1 Petabyte (PB) = 1,000,000 Gigabyte (GB)
- 1 Exabyte (EB) = 1,000,000,000 Gigabyte (GB)
- 1 Zettabyte (ZB) = 1,000,000,000,000 (GB)

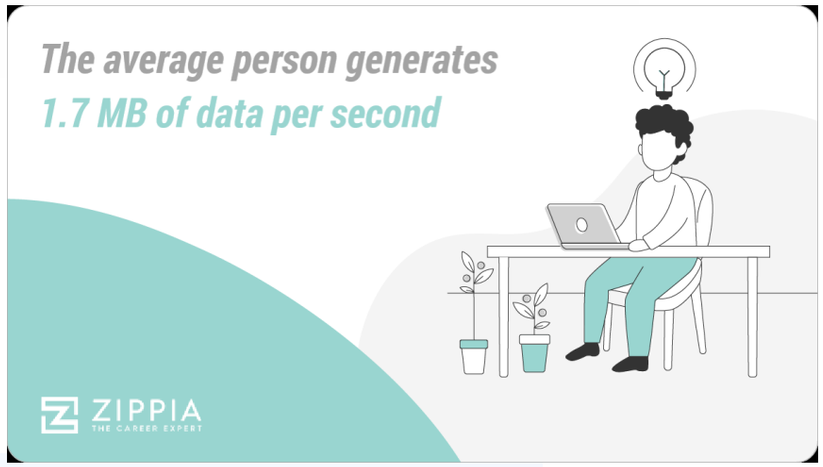
Number in words	*Number in figures*	*Number in standard form*	*Number written as a decimal*
One thousand	1,000	10^3	
Ten thousand	10,000	10^4	0.01 million
One hundred thousand	100,000	10^5	0.1 million
One million	1,000,000	10^6	
Ten million	10,000,000	10^7	0.01 billion
One hundred million	100,000,000	10^8	0.1 billion
One billion	1,000,000,000	10^9	
Ten billion	10,000,000,000	10^{10}	0.01 trillion
One hundred billion	100,000,000,000	10^{11}	0.1 trillion
One trillion	1,000,000,000,000	10^{12}	
One quadrillion	1,000,000,000,000,000	10^{15}	

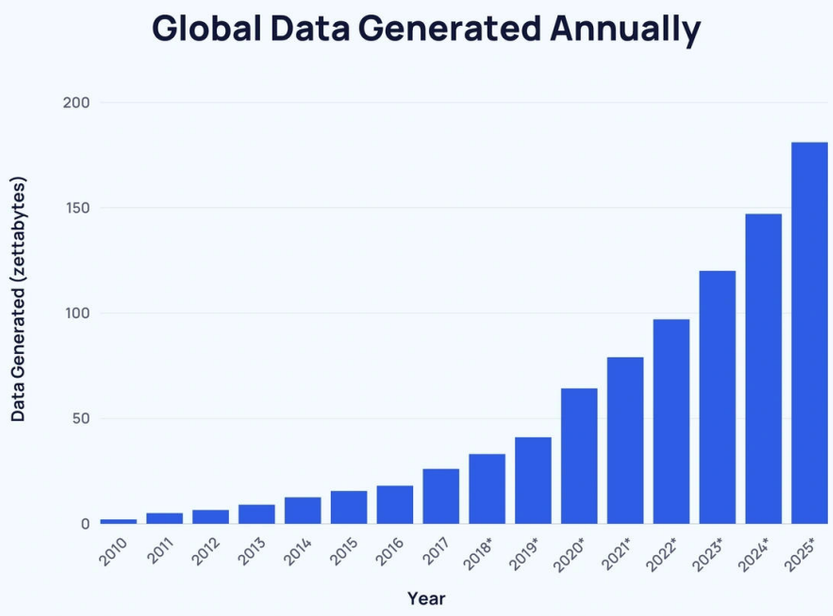
BD Characterization: Velocity

Velocity: the speed at which data is generated

- 463 ZB of data will be created every day by 2025, which will be worth $229.4 billion.





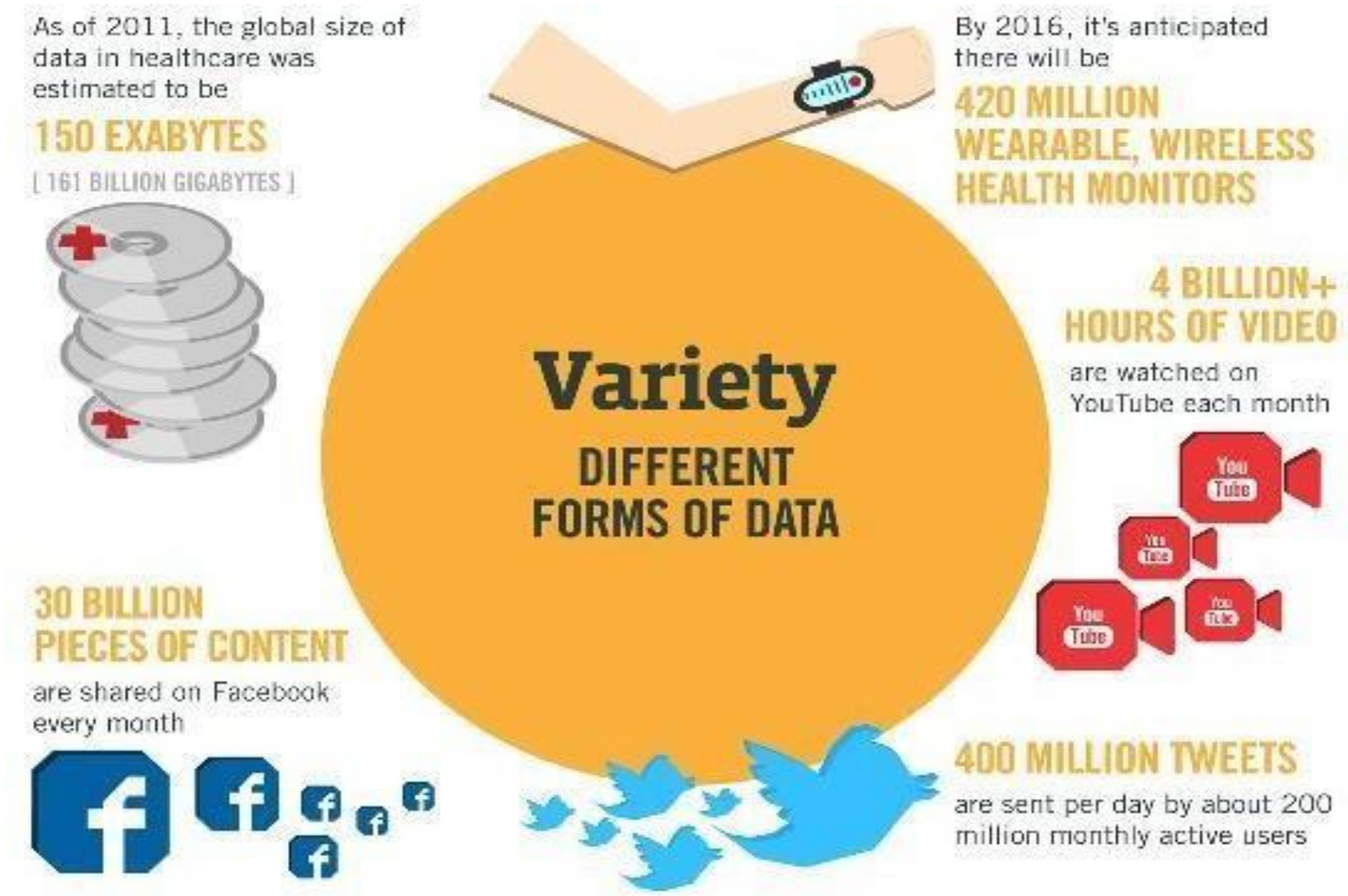


https://www.g2.com/articles/big-data-statistics

Data We Create Online in 60 Seconds

BD Characterization: Variety

Variety: the different types of data available (diversity)



3 Important Statistics About How Much Data Is Created Every Day

FinancesOnline
REVIEWS FOR BUSINESS

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista

3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION
searches on Google by the end of 2021

1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

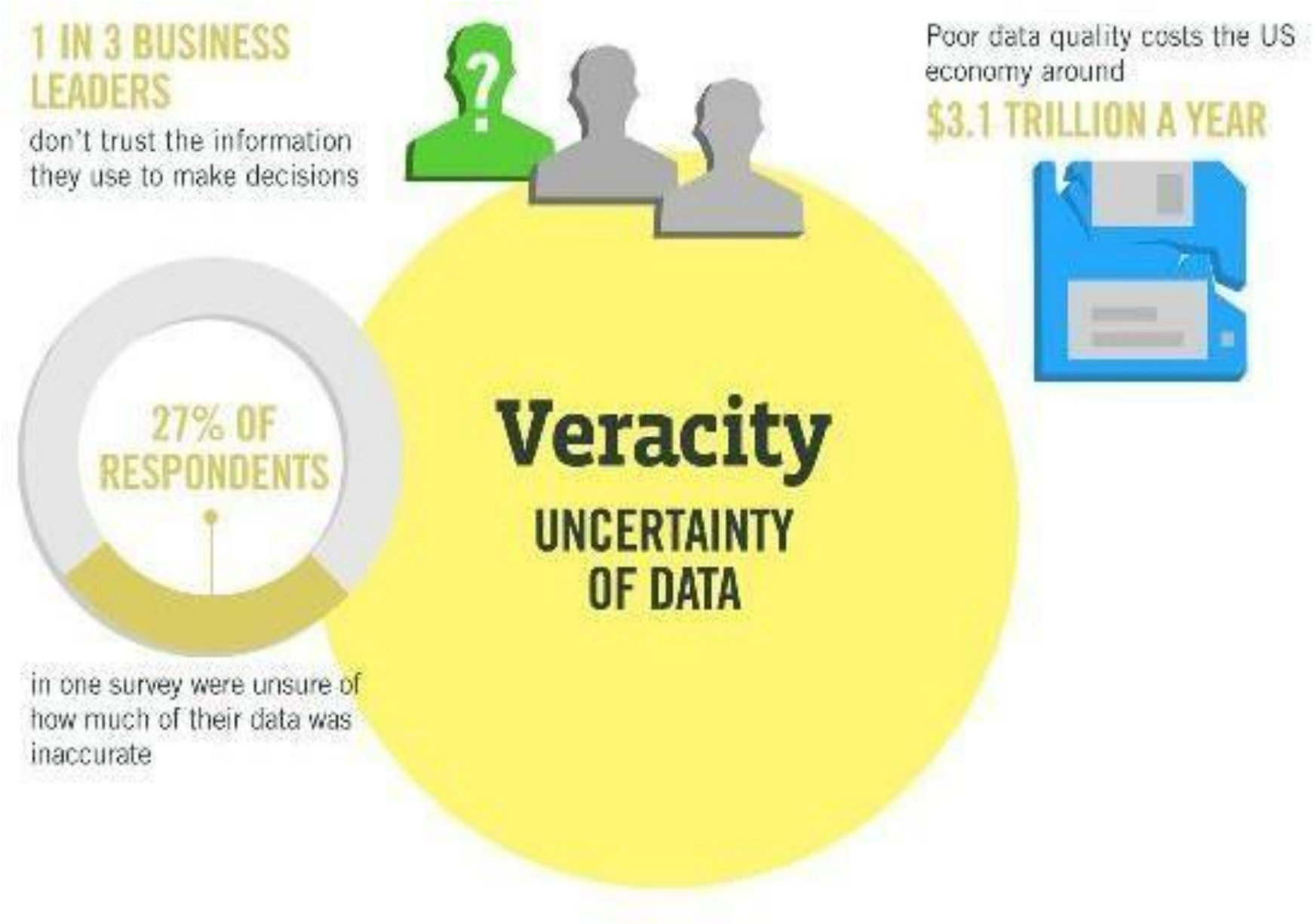
278,108 PETABYTES
global IP data per month by the end of 2021

230,000
new malware versions created every day

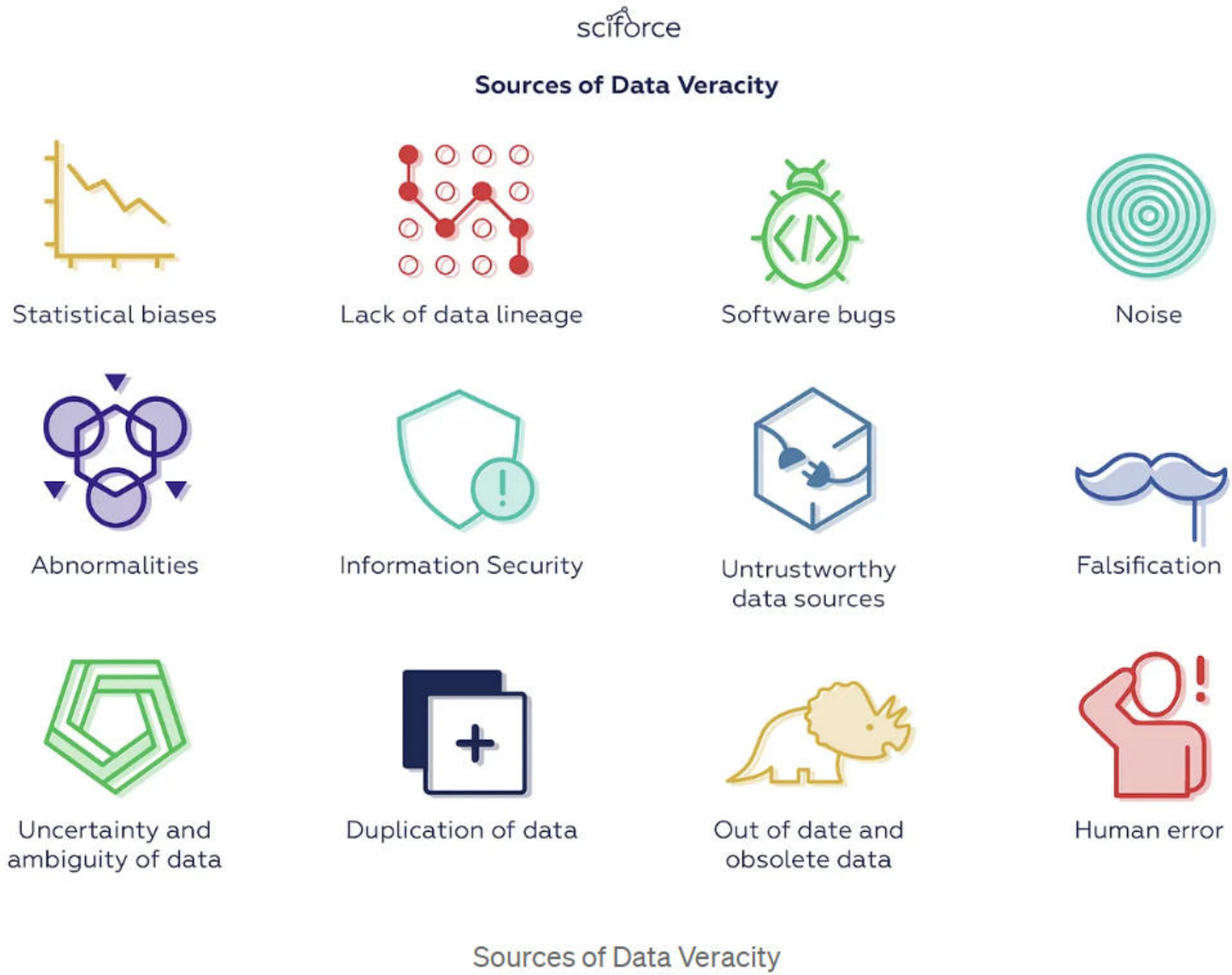
82%
share of video in total global internet traffic at the end of 2021

BD Characterization: Veracity

Veracity: the trustworthiness of data



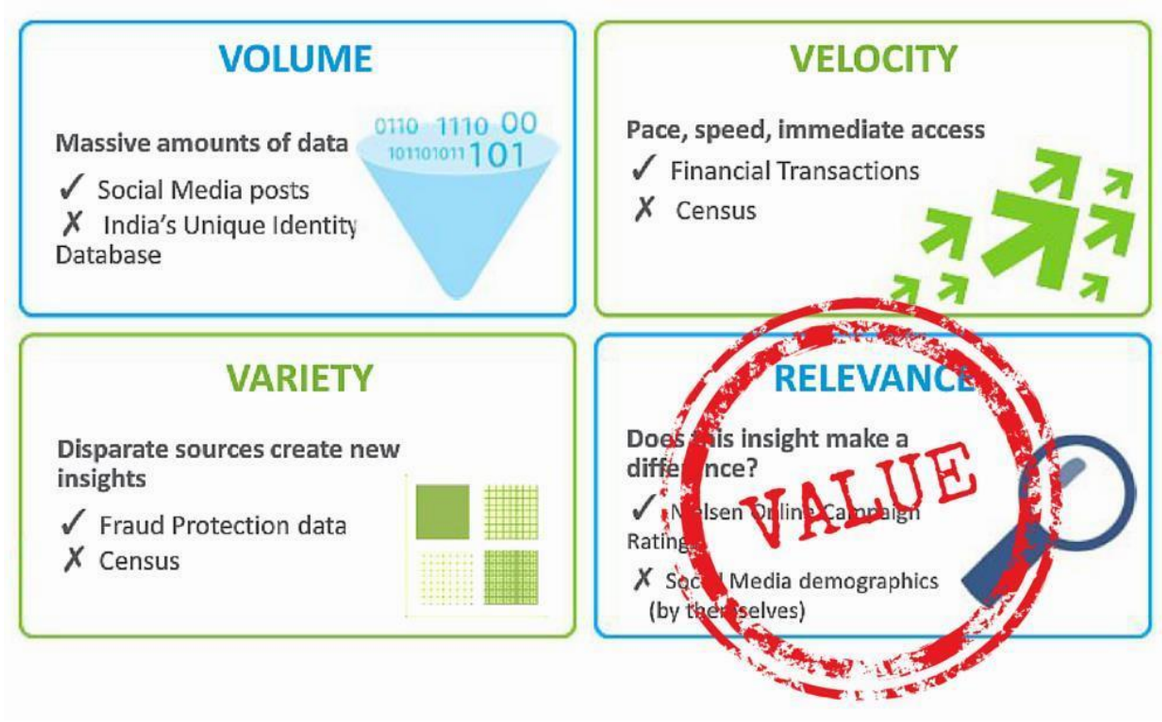
BD Characterization: Veracity

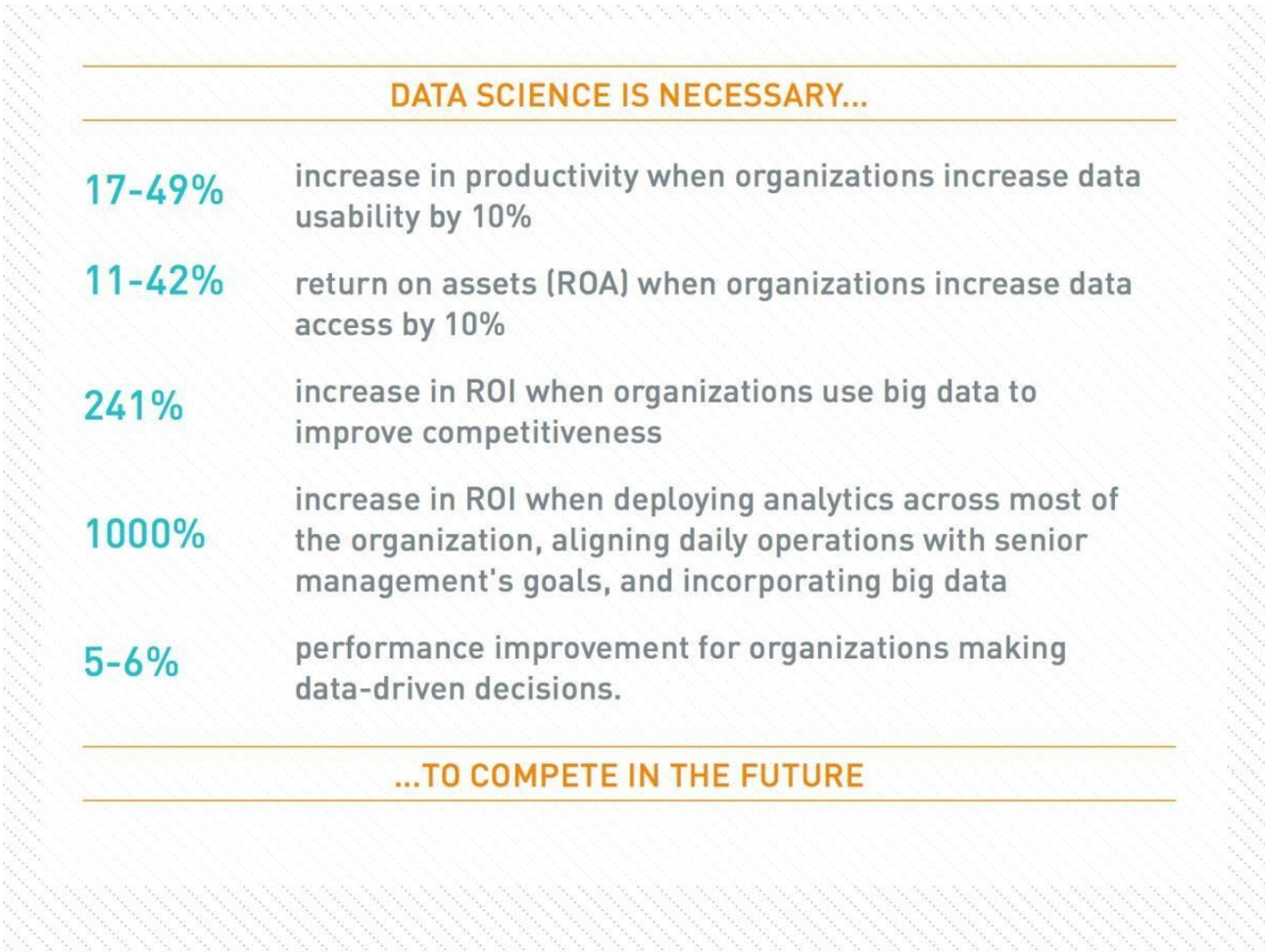


https://medium.com/sciforce/data-veracity-a-new-key-to-big-data-38e110391c7d

BD Characterization: Value

Value: the meaningfulness of data, creation of actionable insights, data monetization etc.





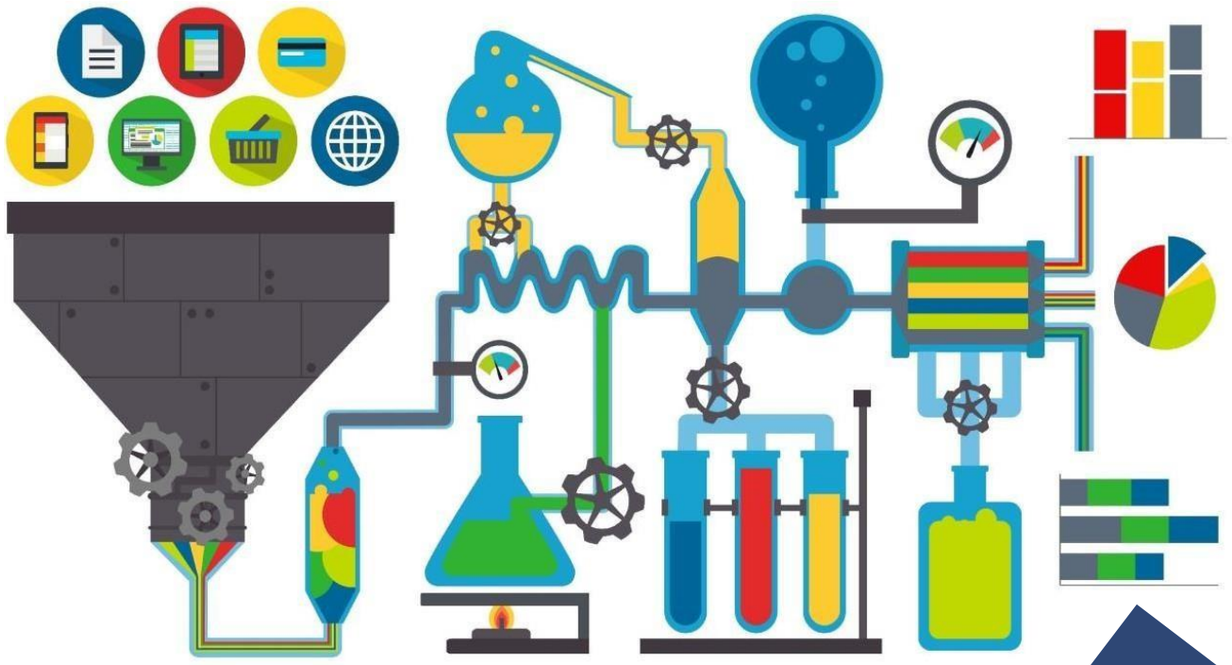
Source: Booz Allen Hamilton

Insights from Big Data

Big Data has been leveraged to create actionable insights in various domains:

- **Facebook**: analyses location information to make it easier to find friends to connect to,identify global migration patterns, to determine where the different football team fanbases live.
- **Target:** predictively models data to determine which of its customers are pregnant, to focus baby-related marketing to them
- **Tesco PLC:** collected refrigerator-related data points to monitor it's performance, towards proactive maintenance (servicing of machines) to cut down on energy costs.
- **Macy's Inc.**: adjusts pricing of it's items in near-real time by monitoring demand and inventory.
- **Siemens:** leveraged on sensor-data analytics and predictive maintenance to reduce train failures.
- **Google Flu Trends:** aggregates Google search queries to predict outbreaks of flu.

How to churn out Insights from Big Data?



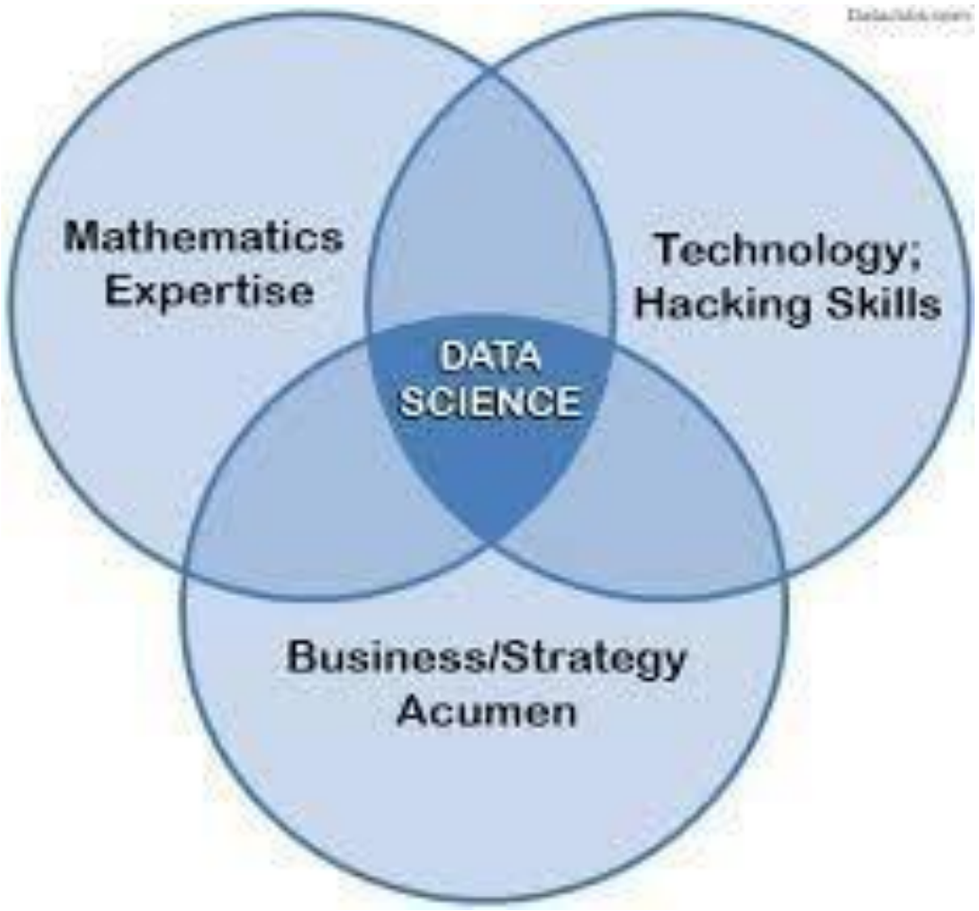
http://blog.onapproach.com/big-data-insights-found-in-unexpected-places

(MachineLearning,Artificial Intelligence and Data) Landscape

https://annageller.medium.com/2023-state-of-data-infrastructure-key-trends-from-matt-turcks-mad-landscape-7dc24e14a815

What is Data Science?

Data Science enables us to gain insights from data through the use of technology, statistics and business acumen.



http://vitalflux.com/big-data-is-not-just-about-hadoop-stack-implementation/

What is Data Science?

Data Science is about drawing useful conclusions from large and diverse data sets through **exploration**, **prediction**, and **inference**.

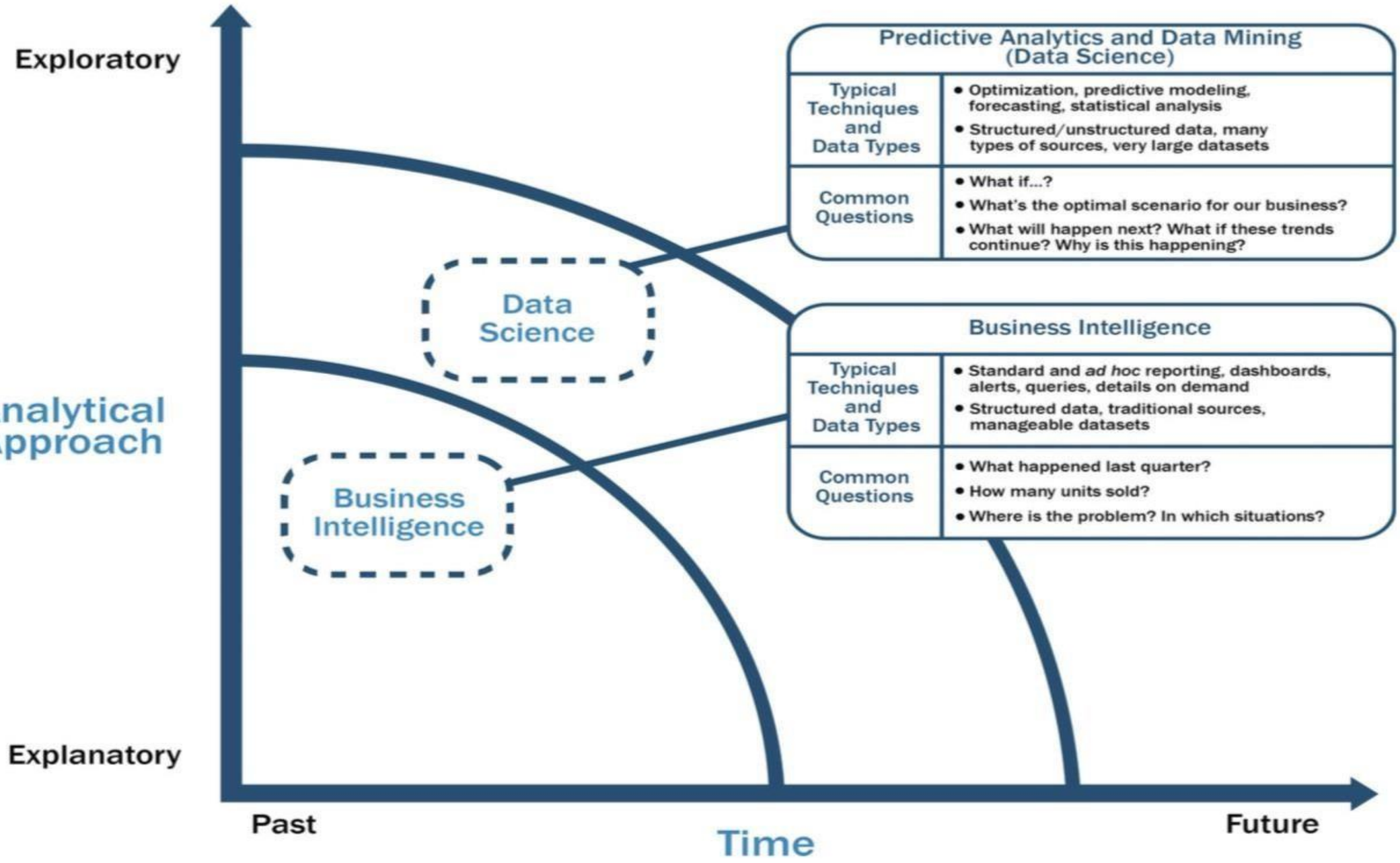
Identifying patterns in information
(visualizations, descriptive stats)

Using information to make informed decisions
(machine learning, optimization)

Quantifying the degree of certainty
(statistical tests, models)

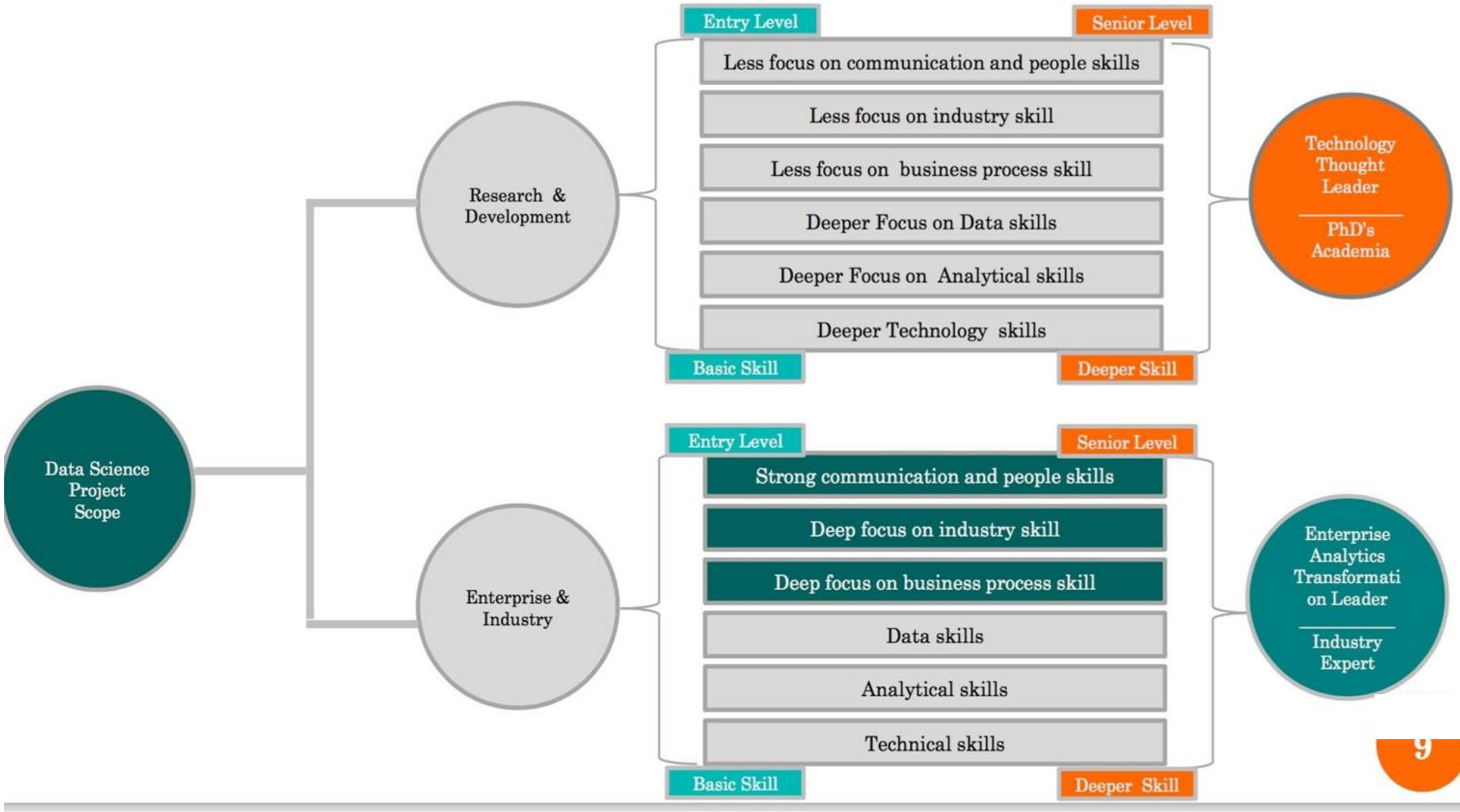
(Adhikari & DeNero, **Computational and Inferential Thinking)**

BI vs. Data Science



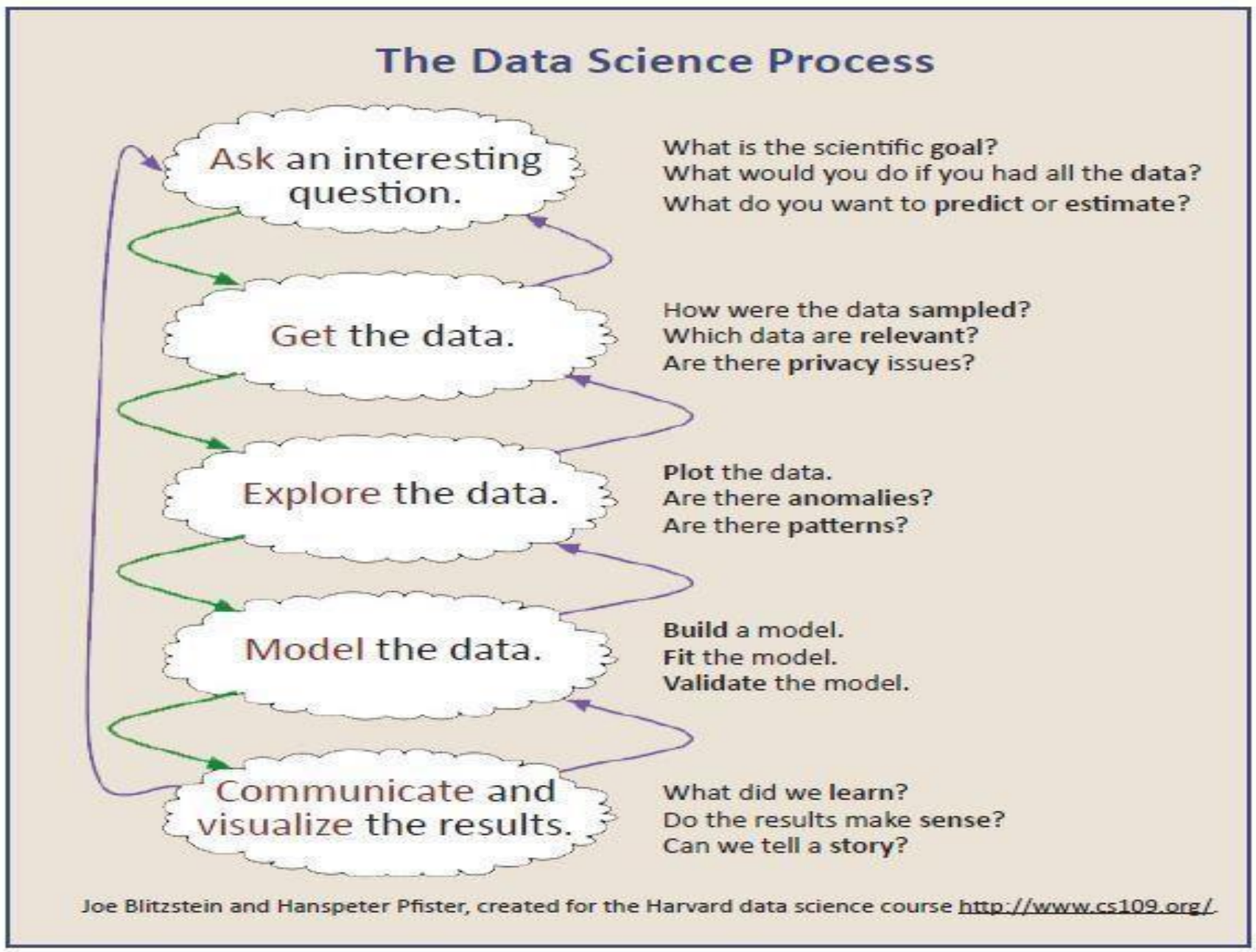
EMC Education Services. (2015). Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.

Data Science in Academia vs. the Industry



http://www.slideshare.net/petekua/mdec-data-science-vs-data-scientist-v2-04212015

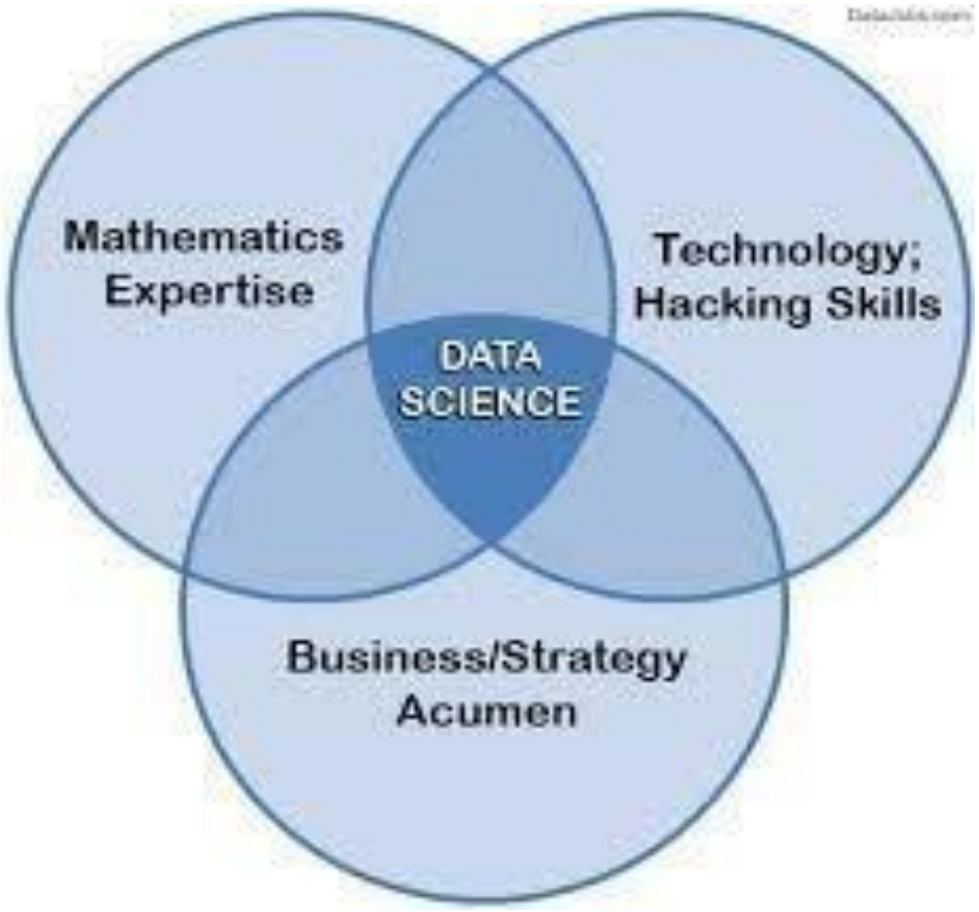
Data Science Process



https://insidebigdata.com/2014/11/12/ask-data-scientist-data-science-process/

A revisited slide: What is Data Science?

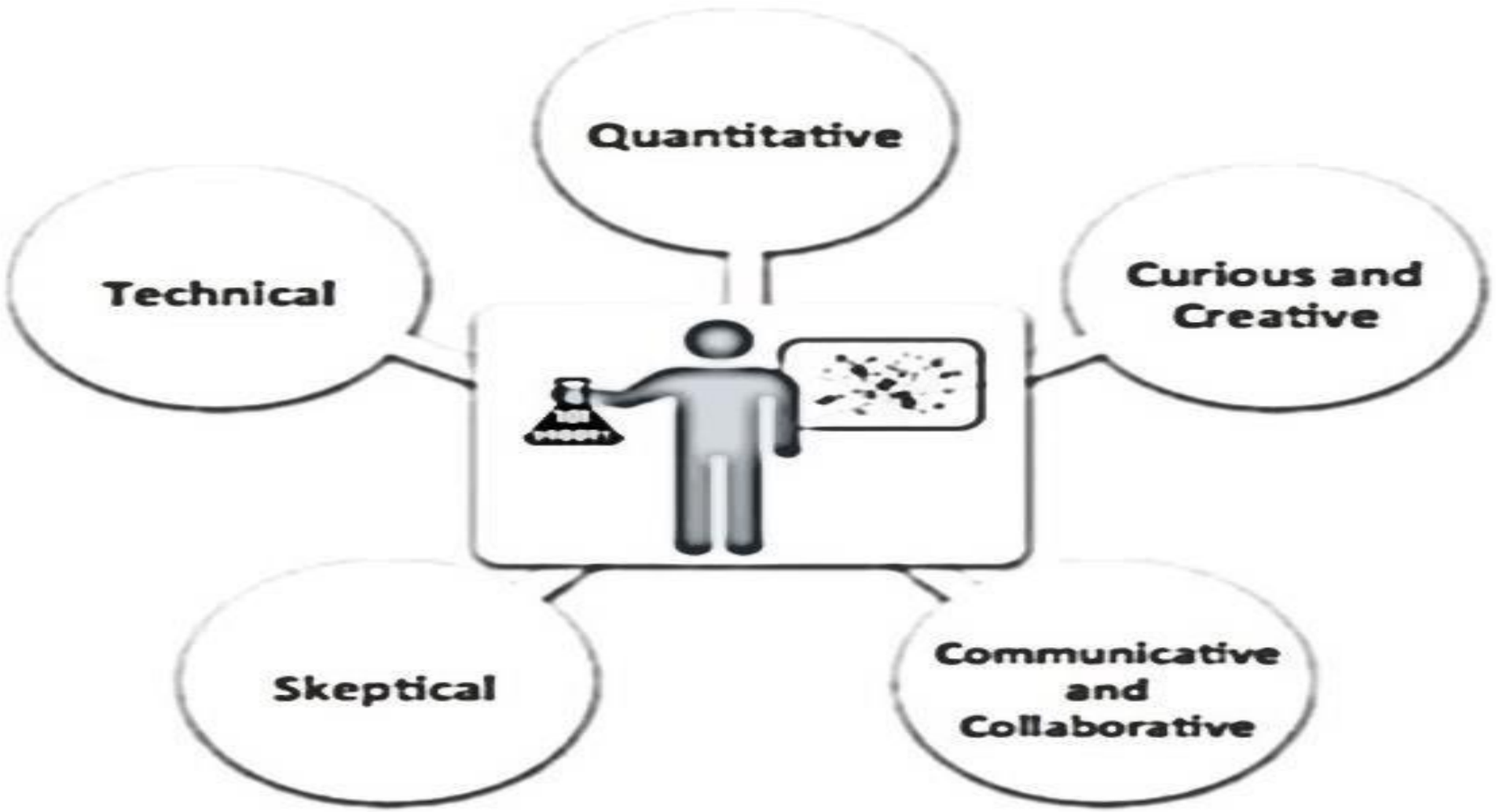
Data Science enables us to gain insights from data through the use of technology, statistics and business acumen.



http://vitalflux.com/big-data-is-not-just-about-hadoop-stack-implementation/

Data Scientists

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician." - *Josh Wills*



http://www.mastersindatascience.org/careers/data-scientist/

Data Scientists: Technical Skills

- Math (e.g. linear algebra, calculus and probability)
- Statistics (e.g. hypothesis testing and summary statistics)
- Machine learning tools and techniques (e.g. k-nearest neighbors, random forests, ensemble methods, etc.) for data mining
- Software engineering skills (e.g. distributed computing, algorithms and data structures)
- Data cleaning
- Data mining
- Data visualization
- R or Python
- SQL databases and database querying languages
- Unstructured databases
- Big data platforms like Spark, Hadoop, Hive & Pig

Data Scientists: Business Skills

- Analytical Problem-Solving
- Effective Communication
- Intellectual Curiosity
- Industry Knowledge

Responsibilities of a Data Scientist

- Conduct undirected research and frame open-ended industry questions
- Extract huge volumes of data from multiple internal and external sources
- Thoroughly clean and prune data to discard irrelevant information
- Explore and examine data from a variety of angles to determine hidden weaknesses, trends and/or opportunities
- Devise data-driven solutions to the most pressing challenges
- Employ sophisticated analytics programs, machine learning and statistical methods to prepare data for use in predictive and prescriptive modeling
- Invent new algorithms to solve problems and build new tools to automate work
- Communicate predictions and findings to management and IT departments through effective data visualizations and reports
- Recommend cost-effective changes to existing procedures and strategies

Analyst vs. Engineer vs. Scientist

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:

Excel, Tableau, SQL

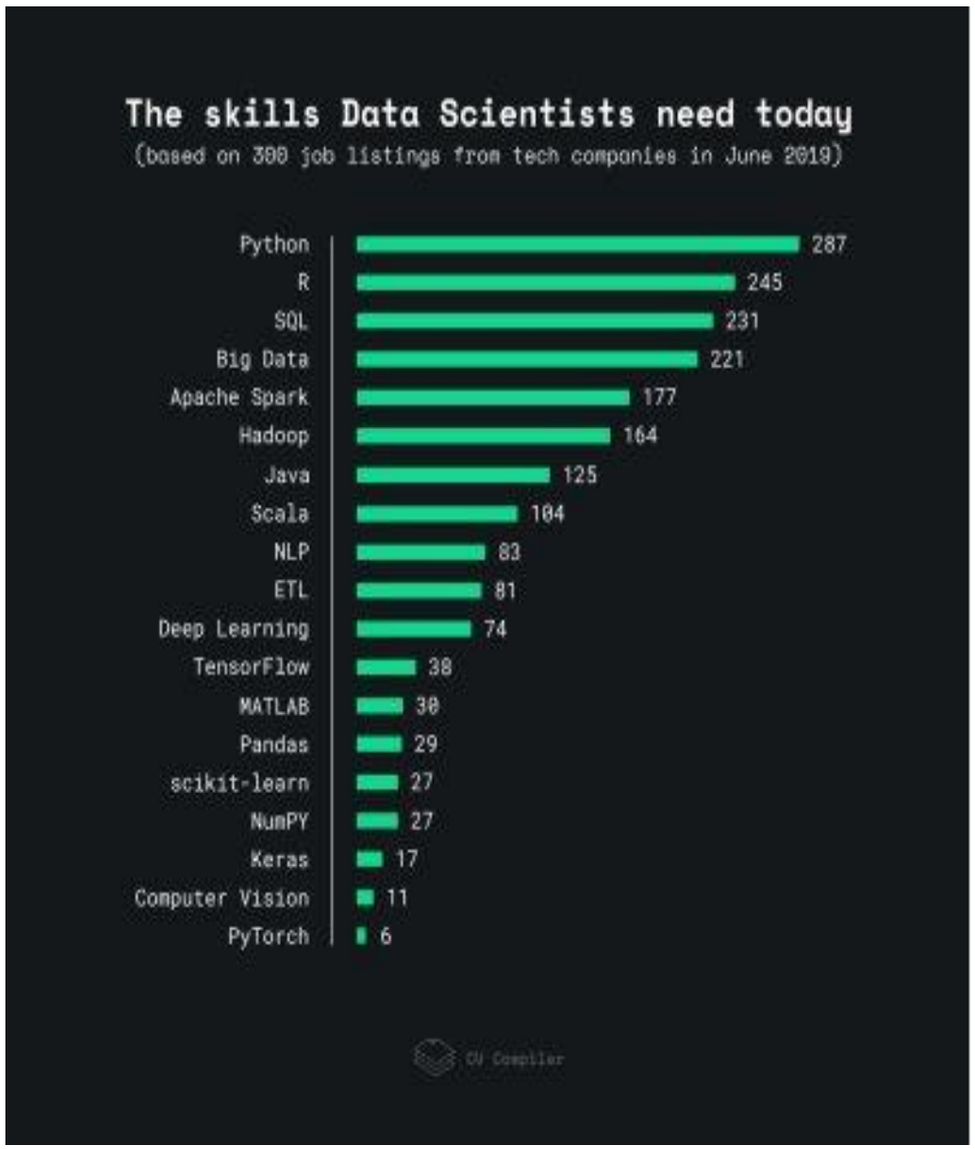
https://medium.com/@vegi/data-scientist-vs-data-analyst-vs-data-engineer-using-word-cloud-902ab83d0879

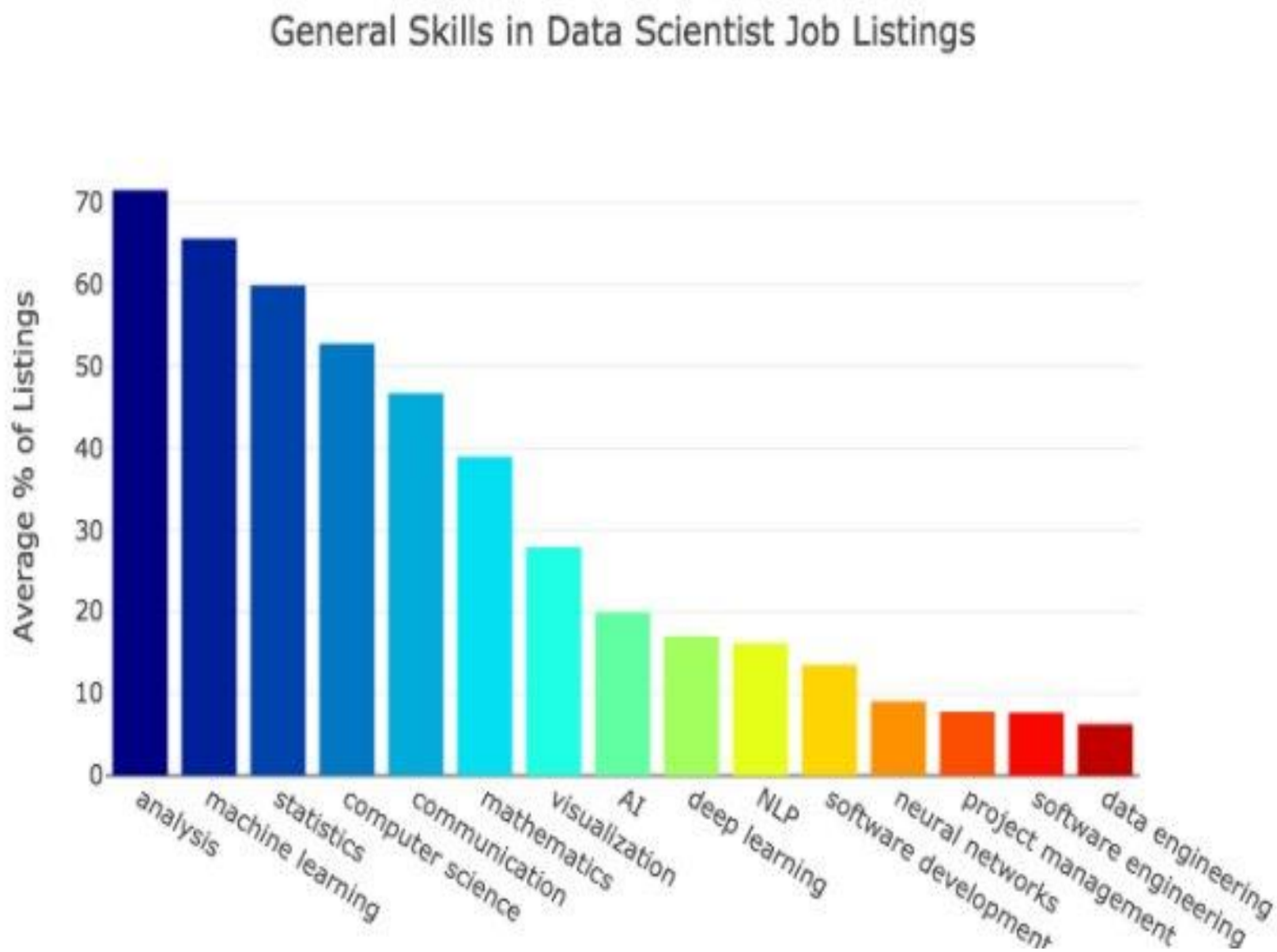
Programming Tools/Languages for Data Science

Worldwide, Mar 2023 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	27.91 %	-0.6 %
2		Java	16.58 %	-1.6 %
3		JavaScript	9.67 %	+0.6 %
4		C/C++	6.93 %	-0.5 %
5		C#	6.88 %	-0.5 %
6		PHP	5.19 %	-0.6 %
7		R	4.23 %	-0.2 %
8	↑	TypeScript	2.81 %	+0.6 %
9	↑	Swift	2.28 %	+0.2 %
10	↓↓	Objective-C	2.26 %	+0.0 %

Skills Data Scientists Need Today





proschool
An ims Initiative

DATA SCIENTIST
MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

PROGRAMMING & DATABASE

- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

DOMAIN KNOWLEDGE & SOFT SKILLS

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking

COMMUNICATION & VISUALIZATION

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design

Challenges in Data Science

- Validity of Assumptions
- Making ad-hoc explanations of data patterns
- Over-generalizing
- Communication
- Validation of models, data pipeline integrity
- Using statistical tests correctly
- Prototype ⇒ Production transitions

Textbook/References

- No specific textbook to use, but a few major references:
  - EMC Education Services (Editor). (2015). Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.
  - **NEW!** Adhikari, A. & DeNero J. (2019). Computational and Inferential Thinking. Online Book. https://www.inferentialthinking.com/chapters/intro.html
  - O'Neil, C. & Schutt, R. (2013). Doing Data Science Straight Talk from the Frontline. O'Reilly Media.

End of Lecture 1