Digital Readiness Series

Data analytics

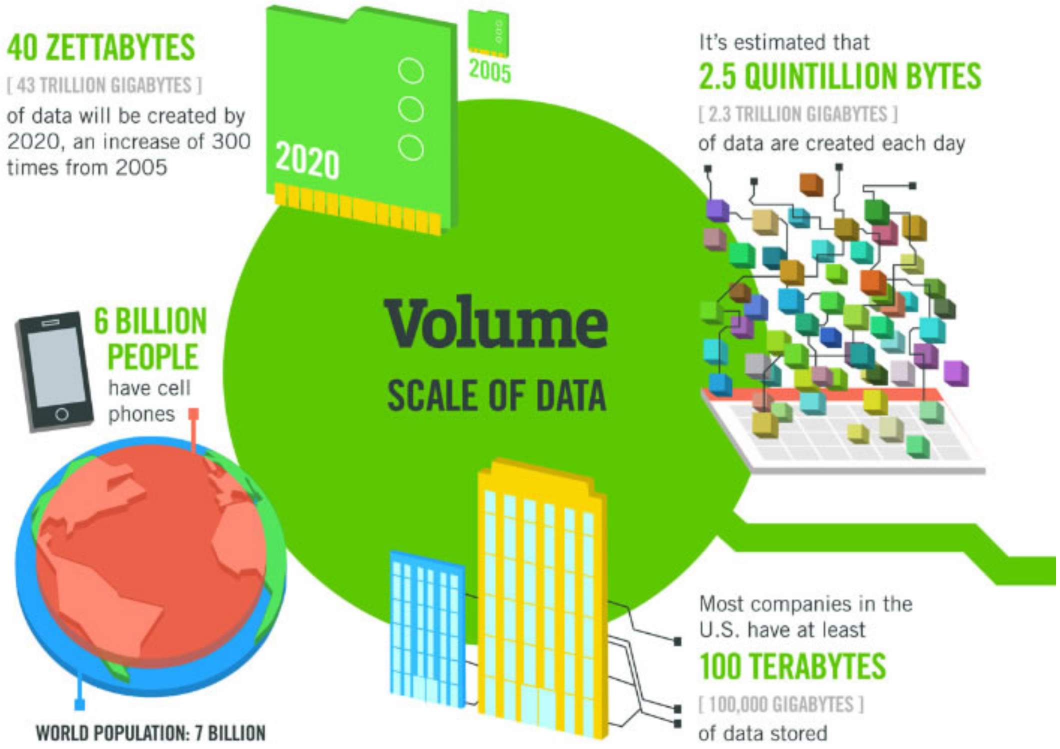
The world of Data

By
Dr. Carlo Lipizzi

July 30, 2020

www.sercuarc.org

Data growth

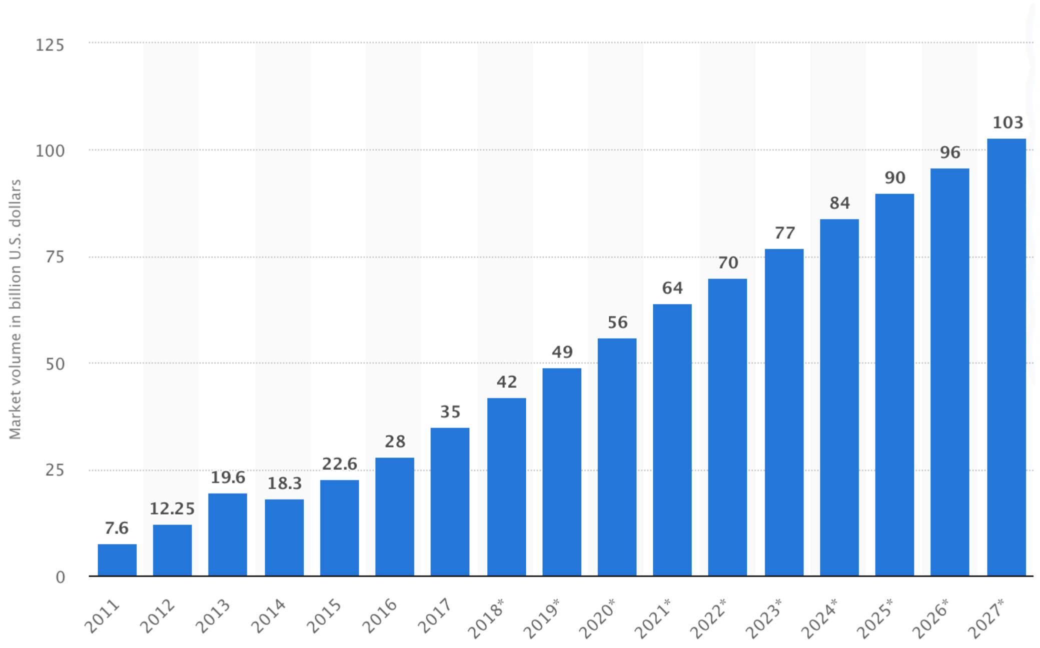

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

- The digital tools we are using every day are creating data from everything we do at an unprecedented rate: every day, 2.5 quintillion (1018) bytes of data are created and 90% of the data in the world today was created within the past two years.
- Data piles up quickly in business applications, and compound annual data growth threatens to bury today's application infrastructure. A senior executive at a major bank remarked, "There are only 3 things certain in life: death, taxes, and data growth" [from Wired]
- Because so much of the population is generating it, Big Data can provide potentially useful information for our lives and businesses
- Mining the Big Data requires a combination of tools, ability to represent knowledge and domain-specific expertise

The “datafication”

- It is happening as result of the digital transformation process that is creating a new kind of economy based on the “datafication” of virtually any aspect of human social, political and economic activity as a result of the information generated by the digitally connected individuals, companies, institutions and machines

The Data Market



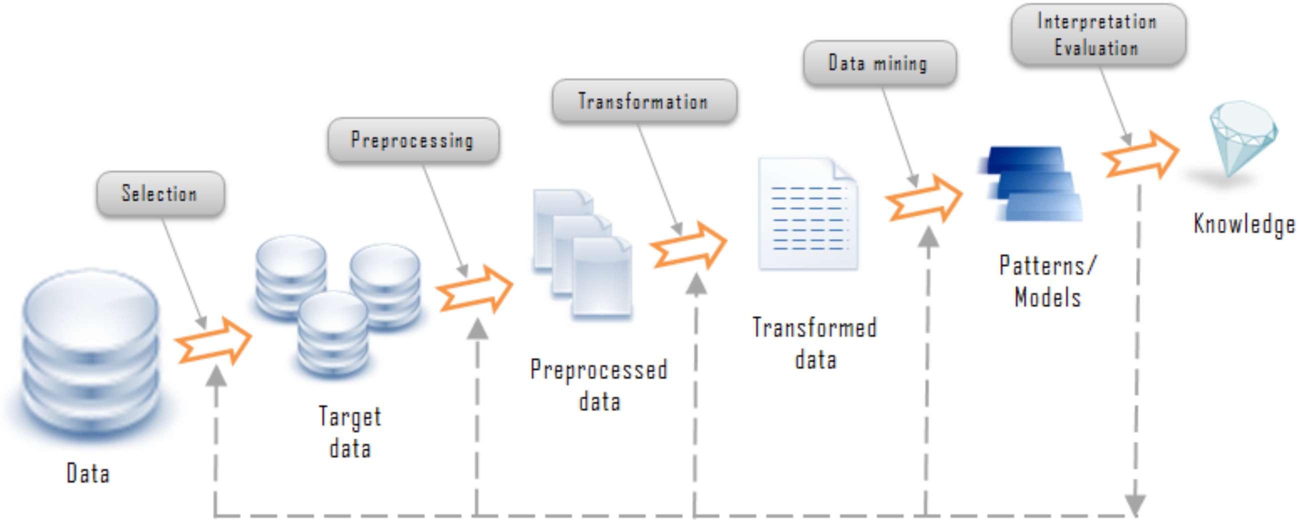
© Statista

- The global big data and business analytics market was valued at 169 billion U.S. dollars in 2018 and is expected to grow to 274 billion U.S. dollars in 2022
- As of November 2018, 45 percent of professionals in the market research industry reportedly used big data analytics as a research method

What is Data Science?

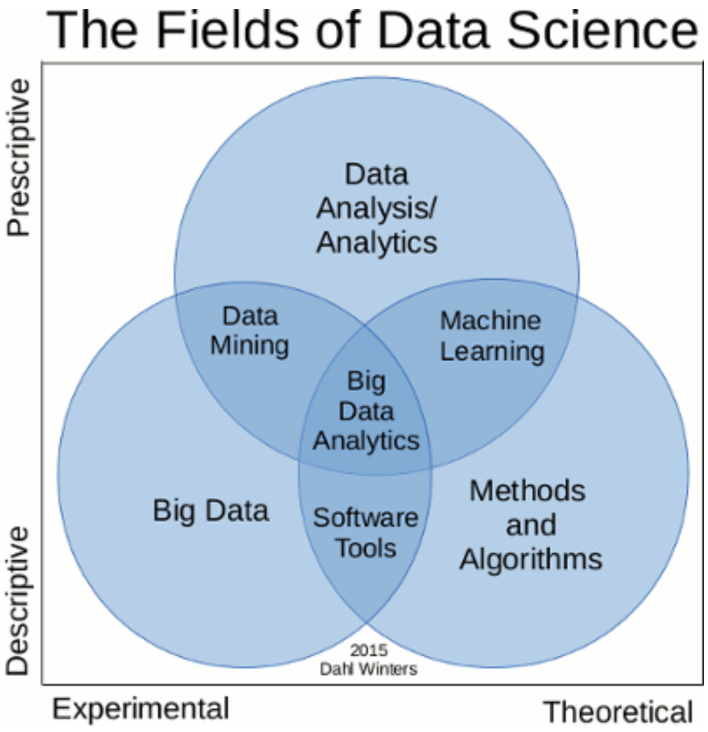
What is Data Science?

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



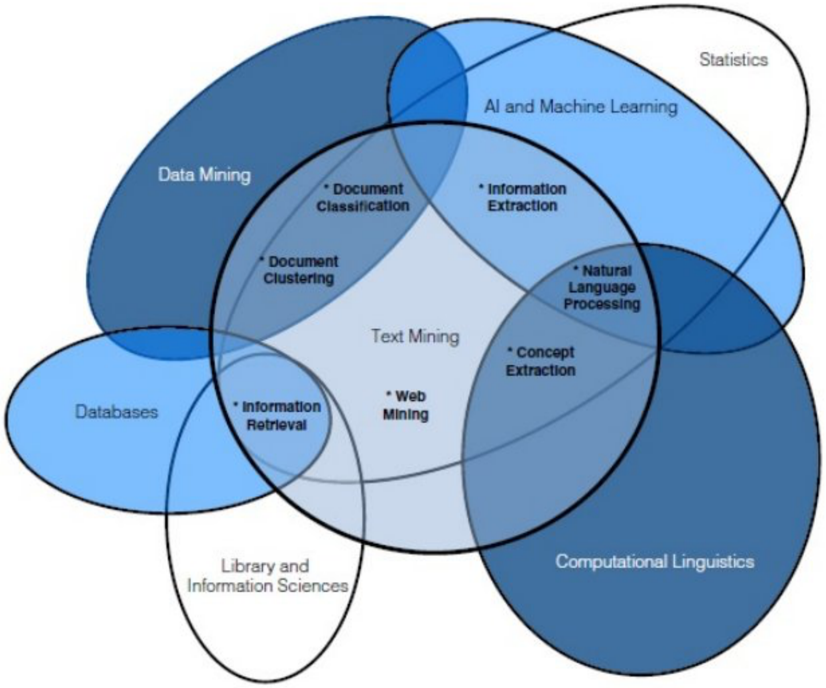
What is Data Science?

- Is the science of analyzing raw data in order to make conclusions about that information



- Data-centered systems: Models and systems in Data Science are developed bottom up, extracting them from the data
- A Data scientist exists at the intersection of information technology, statistics and business. The primary goal of a data scientist is to increase efficiency and improve performance by discovering patterns in data and defining courses of action

Data Science Components



--- Focus on data complexity ---

- **Data engineering**: collecting and organizing data
- **Data exploration**: how to work with data
- **Data mining**: extracting knowledge from data
- **Data visualization**: representing metrics in an intuitive way
- **Data-driven systems**: Bottom-up machine learning
- **Natural Text Processing**: text is data

Why Data Science now?

- More intense competition
- Recognition of the value in data sources
- Availability of quality data
- The exponential increase in data processing and storage capabilities and decrease in cost

Data Science/Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Source: Pearson Education, Inc.

Why Data Science?

- **Making better informed decisions**

  e.g. strategies, recommendations

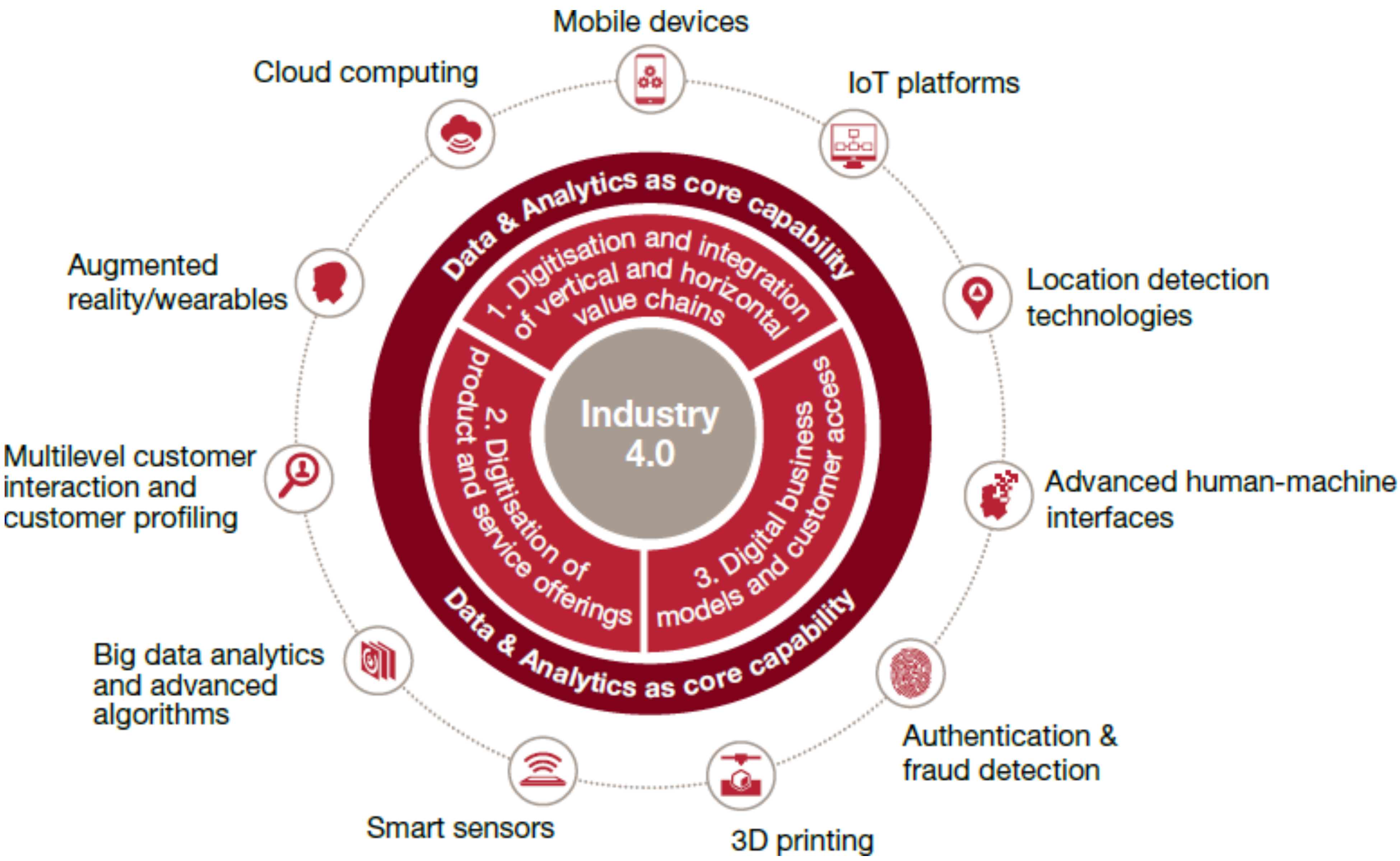
- **Discovering hidden insights**

  e.g. anomalies forensics, patterns, trends

- **Automating business processes**

  e.g. complex events, translation

Business Environment

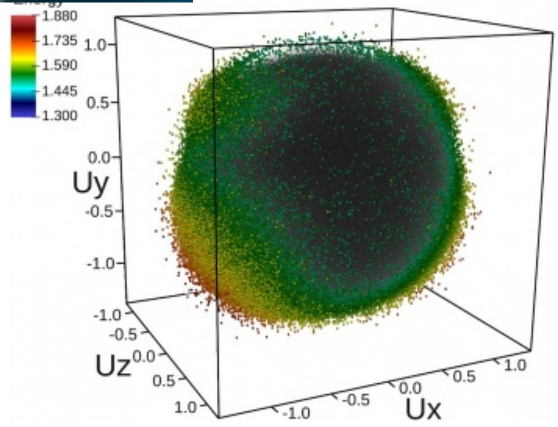

Mobile devices
IoT platforms
Location detection technologies
Advanced human-machine interfaces
Authentication & fraud detection
3D printing
Smart sensors
Big data analytics and advanced algorithms
Multilevel customer interaction and customer profiling
Augmented reality/wearables
Cloud computing
Data & Analytics as core capability
1. Digitisation and integration of vertical and horizontal value chains
2. Digitisation of product and service offerings
3. Digital business models and customer access
Industry 4.0
Data & Analytics as core capability

Why Mine Data?

- Large amount of data collected and warehoused
  - —Web data, e-commerce
  - —Purchases at stores
  - —Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - —Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Mine Data?

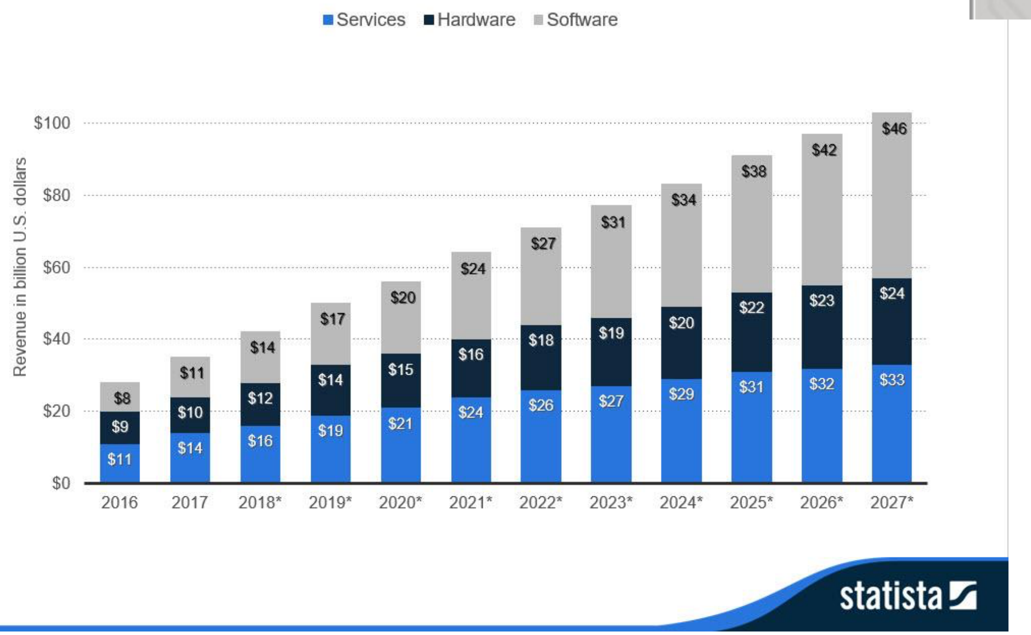
- Data collected and stored at very high speeds
  - —remote sensors on a satellite
  - —telescopes scanning the skies
  - —microarrays generating gene expression data
  - —scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists in
  - —classifying and segmenting data
  - —Hypothesis Formation

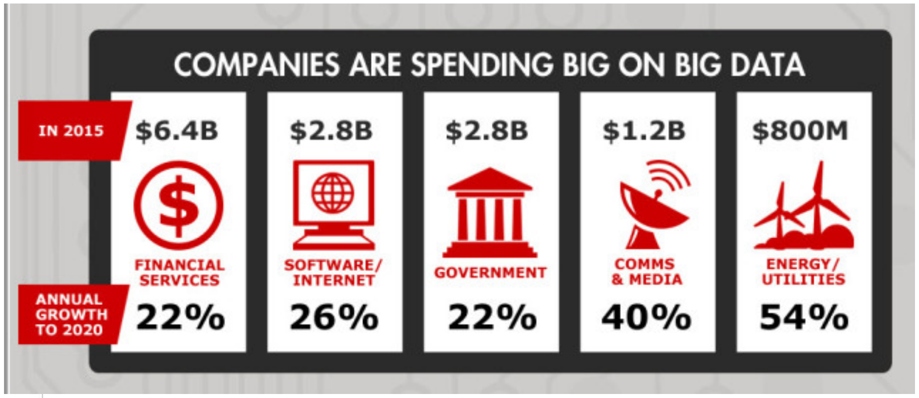


How companies leverage on Data

Global Big Data Revenue 2016-2027, by type

Big Data Revenue Worldwide from 2016 to 2027, by major segment (in billion U.S. dollars)



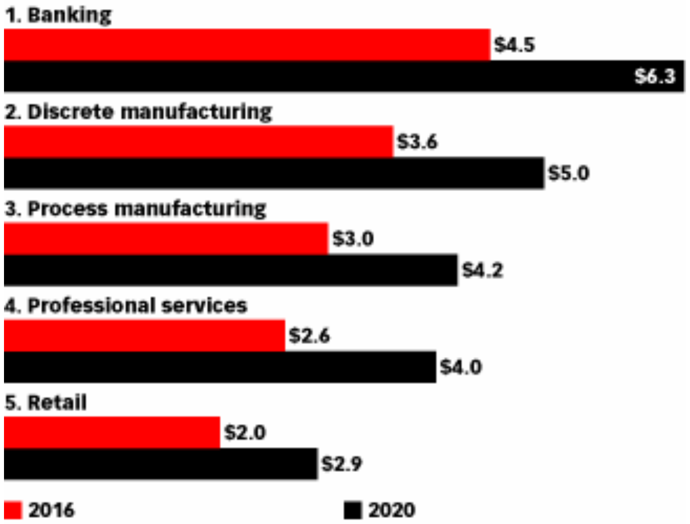


COMPANIES ARE SPENDING BIG ON BIG DATA

	FINANCIAL SERVICES	SOFTWARE/ INTERNET	GOVERNMENT	COMMS & MEDIA	ENERGY/ UTILITIES
IN 2015	$6.4B	$2.8B	$2.8B	$1.2B	$800M
ANNUAL GROWTH TO 2020	22%	26%	22%	40%	54%

Top 5 Industries in Western Europe, Ranked by Big Data & Analytics Spending, 2016 & 2020

billions



Source: International Data Corporation (IDC), "Worldwide Semiannual Big Data and Analytics Spending Guide" as cited in press release, March 30, 2017

225929 www.eMarketer.com

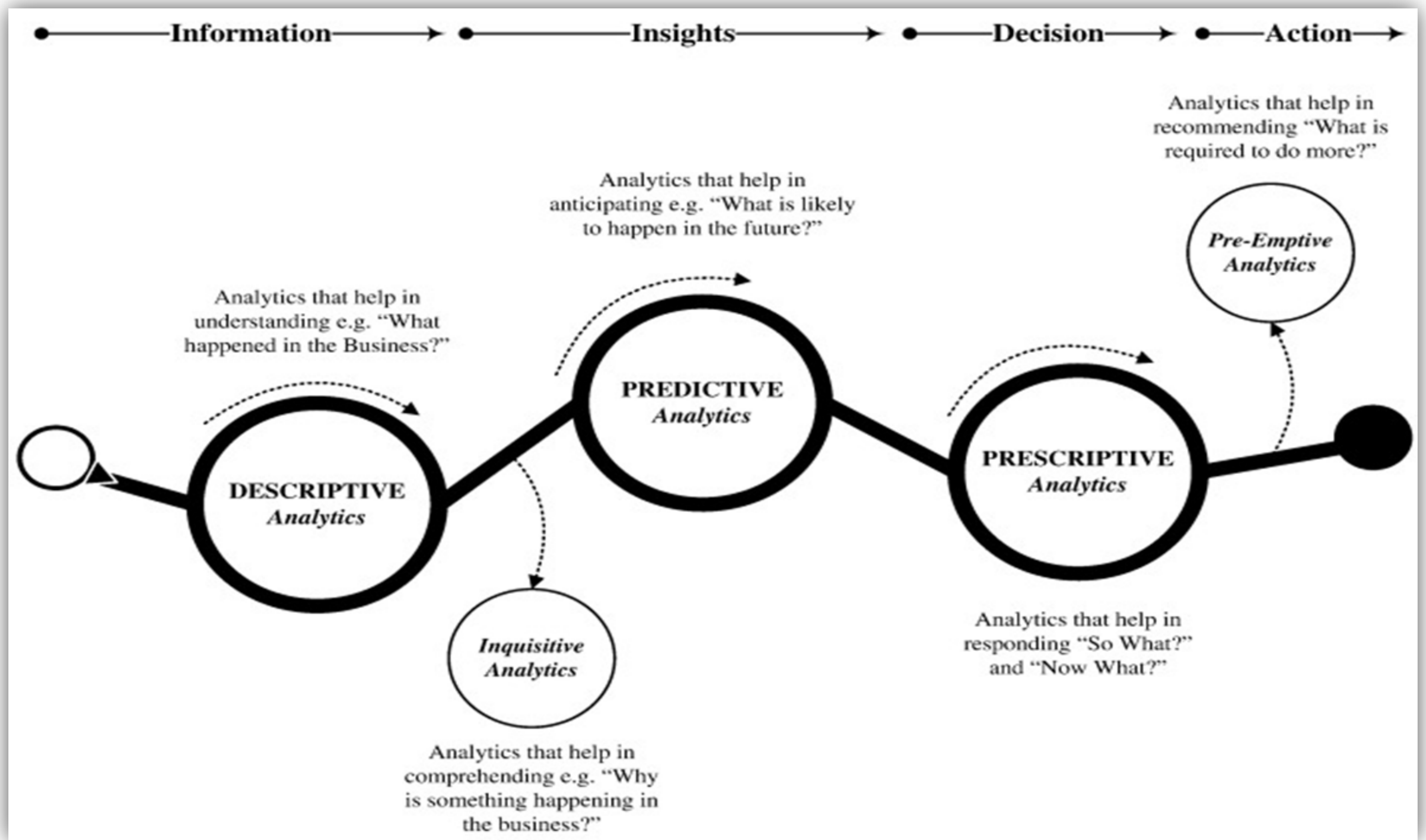
Identifying Insurance Fraud

- Business need
  - Save and make money by reducing fraudulent auto insurance claims
- Data & Analytics
  - Predictive analytics against years of historical claims and coverage data
  - Text mining adjuster reports for hidden clues, e.g. missing facts, inconsistencies, changed stories
- Results
  - Improved success rate in pursuing fraudulent claims from 50% to 88%; reduced fraudulent claim investigation time by 95%
  - Marketing to individuals with low propensity for fraud

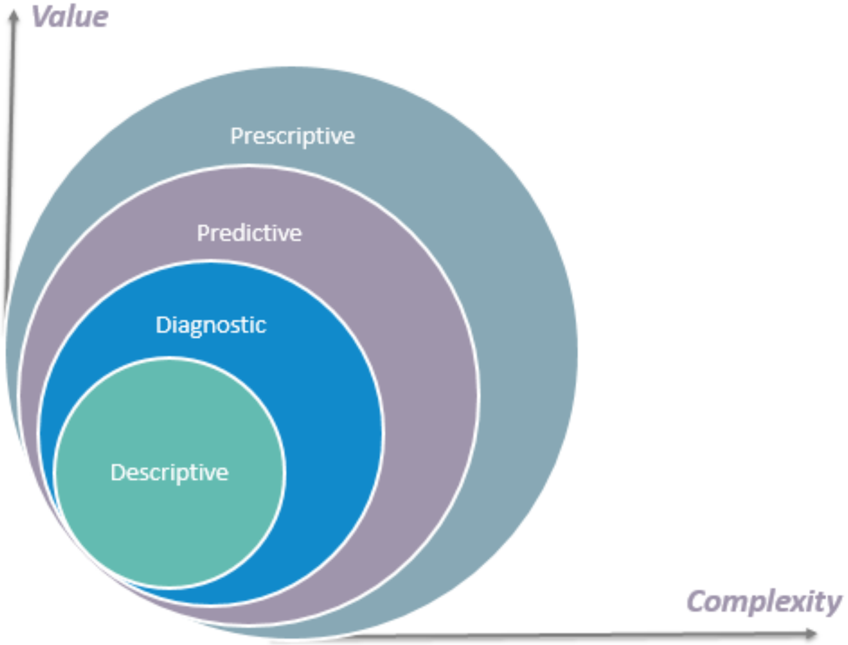
Improving Corporate Image

- Business need
  - Improve reputation, brand and buzz by tapping social media
- Data & Analytics
  - Continually scanning Twitter for mentions of their business
  - Integrating tweeters with their robust customer management system
- Results
  - Detected tweet from a top customer lamenting late flight—no time to dine at Morton’s
  - Tuxedo-clad waiter waiting for him when he landed with a bag containing his favorite steak, prepared the way he normally likes it

What to do with Data Science

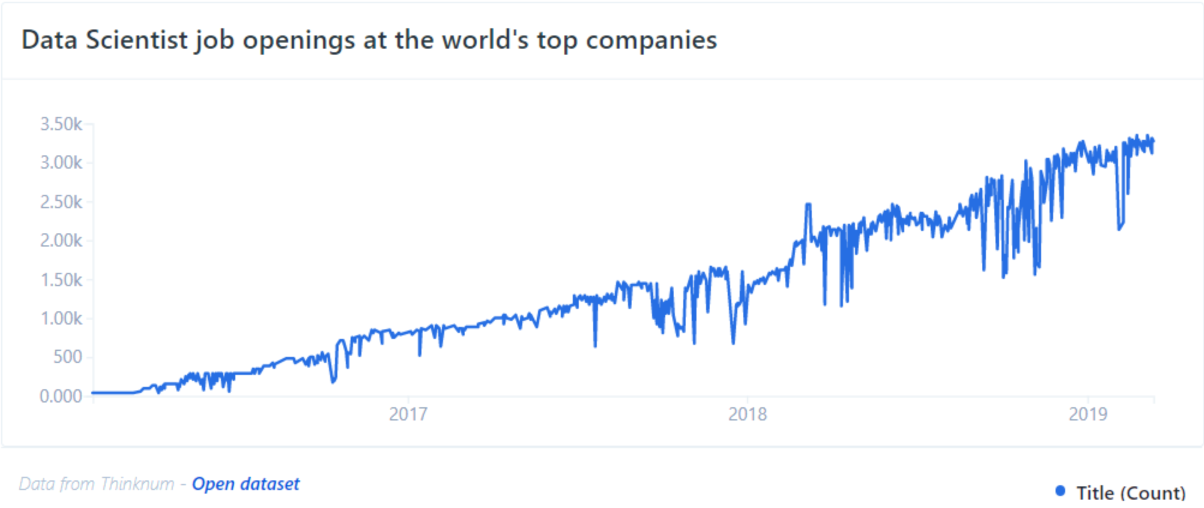

Information
Insights
Decision
Action
Analytics that help in recommending "What is required to do more?"
Analytics that help in anticipating e.g. "What is likely to happen in the future?"
Pre-Emptive Analytics
Analytics that help in understanding e.g. "What happened in the Business?"
PREDICTIVE Analytics
PRESCRIPTIVE Analytics
DESCRIPTIVE Analytics
Inquisitive Analytics
Analytics that help in responding "So What?" and "Now What?"
Analytics that help in comprehending e.g. "Why is something happening in the business?"

Categories



- **Descriptive analytics** describes what has happened over a given period of time
- **Diagnostic analytics** focuses on why something happened
- **Predictive analytics** moves to what is likely going to happen in the near term
- **Prescriptive analytics** suggests a course of action

Data Scientists: demand & competencies

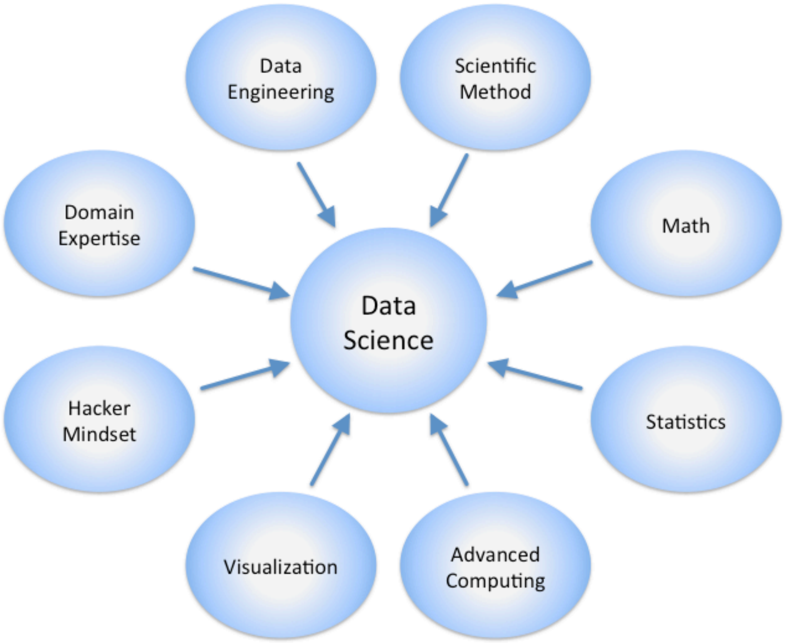


Data Scientist job openings at the world's top companies

Data from Thinknum - *Open dataset*

A steady growing demand

A wide set of competencies and direct/first hand experience



Data Science

Data Engineering

Scientific Method

Domain Expertise

Math

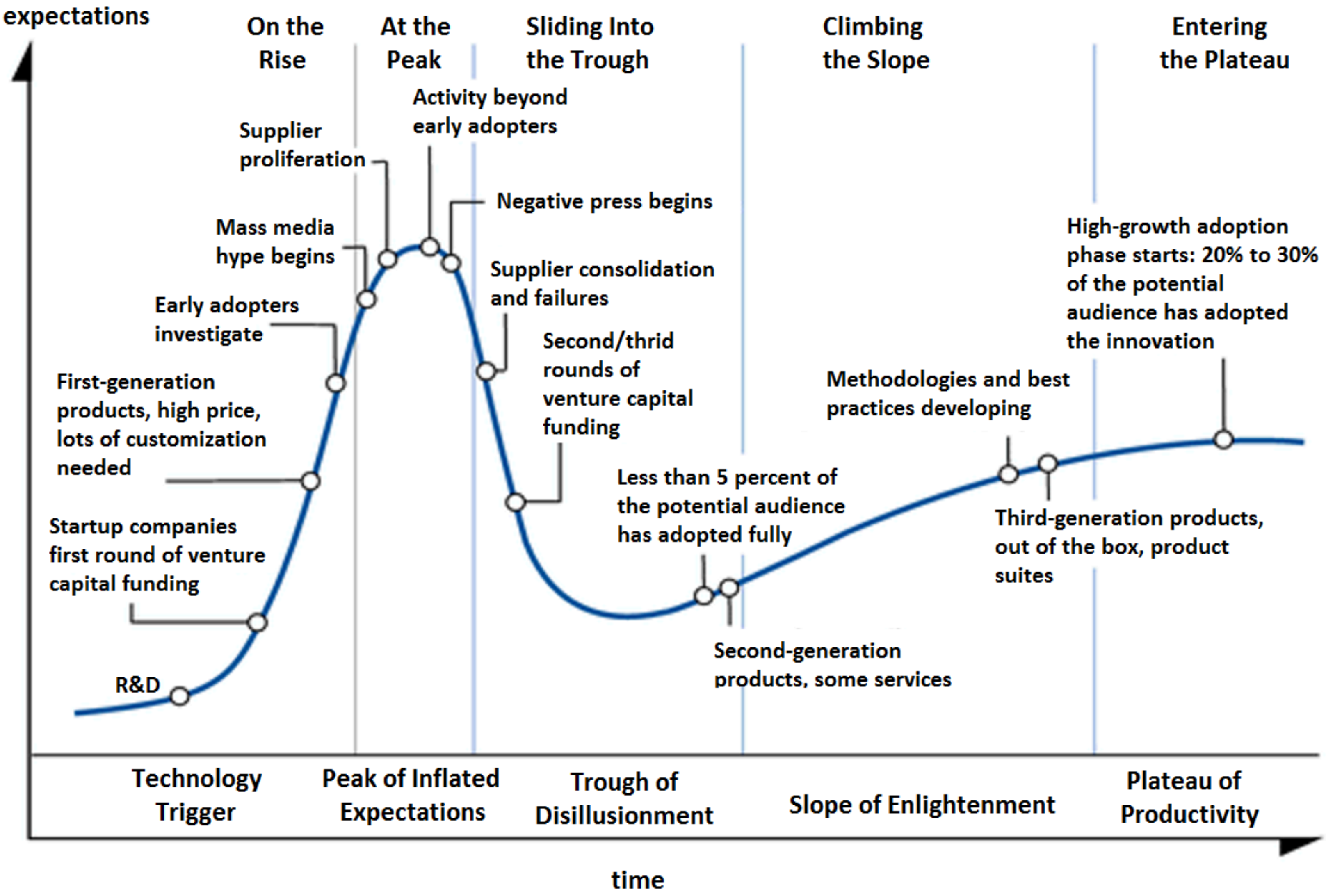
Hacker Mindset

Statistics

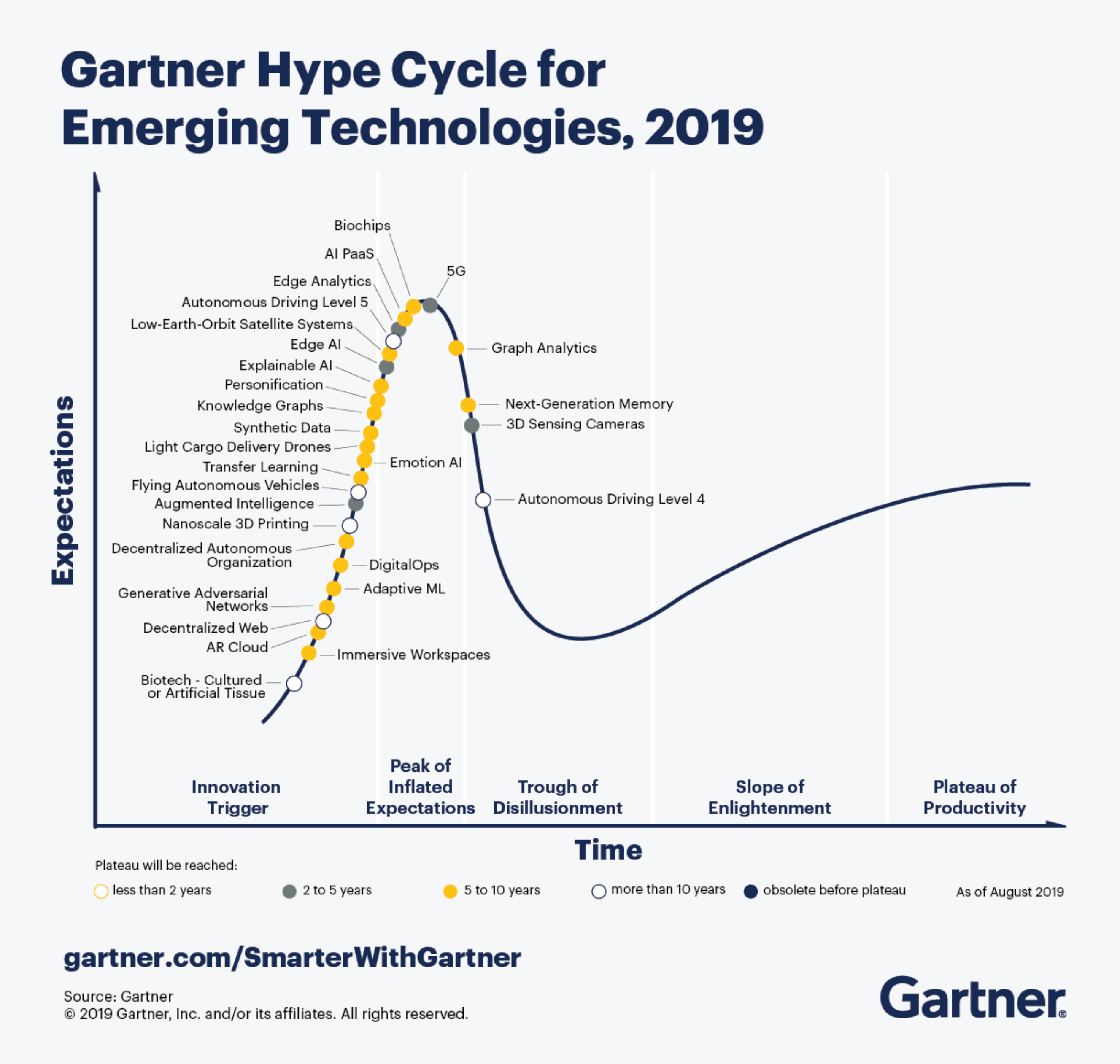
Visualization

Advanced Computing

Technology evolution: using Gartner model


expectations
On the Rise
At the Peak
Sliding Into the Trough
Climbing the Slope
Entering the Plateau
Activity beyond early adopters
Supplier proliferation
Negative press begins
Mass media hype begins
Supplier consolidation and failures
High-growth adoption phase starts: 20% to 30% of the potential audience has adopted the innovation
Early adopters investigate
Second/thrid rounds of venture capital funding
Methodologies and best practices developing
First-generation products, high price, lots of customization needed
Less than 5 percent of the potential audience has adopted fully
Third-generation products, out of the box, product suites
Startup companies first round of venture capital funding
Second-generation products, some services
R&D
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

Emerging Technologies

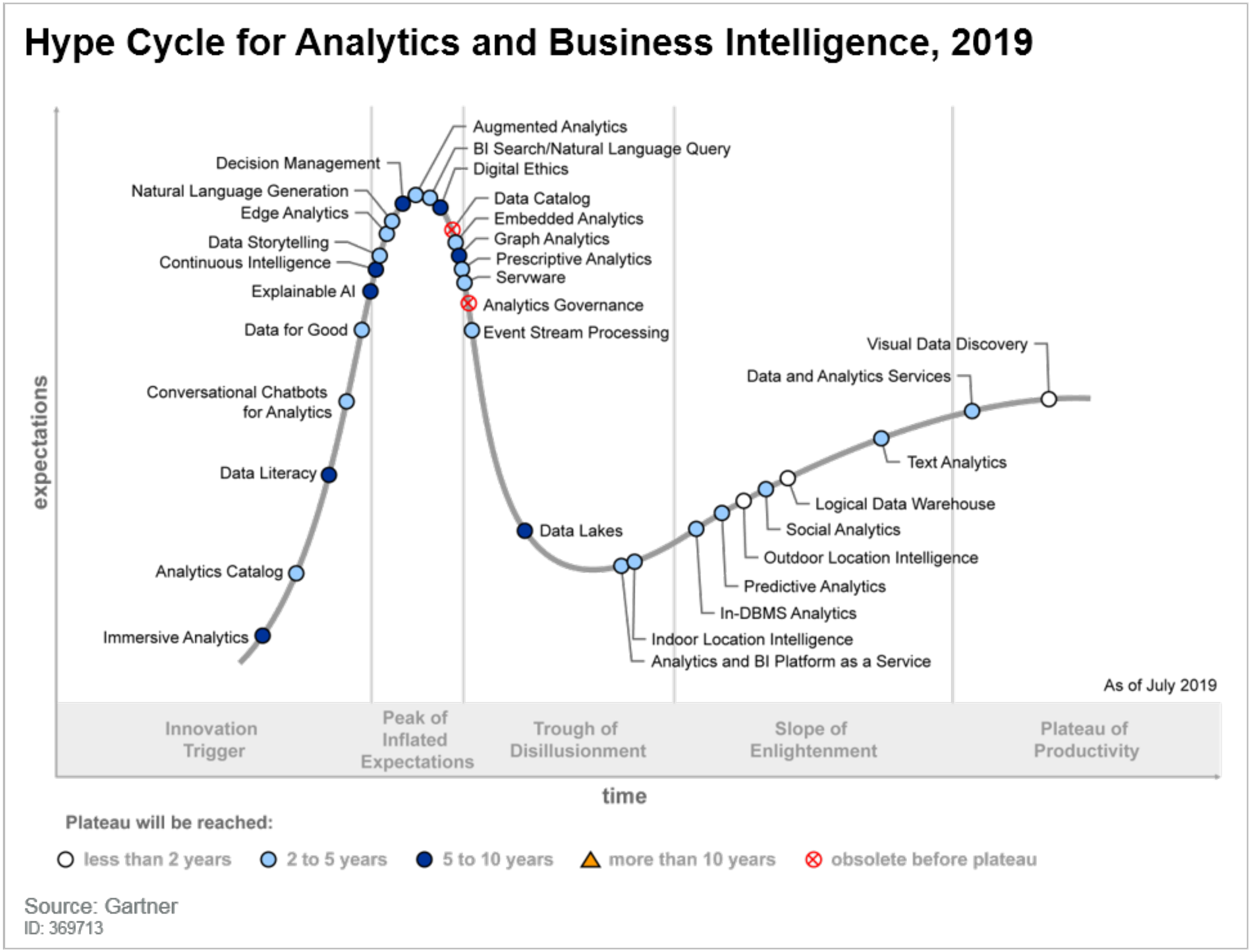


Gartner Hype Cycle for Emerging Technologies, 2019

gartner.com/SmarterWithGartner

Source: Gartner
© 2019 Gartner, Inc. and/or its affiliates. All rights reserved.

Analytics and Business Intelligence



Hype Cycle for Analytics and Business Intelligence, 2019

Source: Gartner
ID: 369713

Challenges of Data Science

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Data Science vs. Statistical Analysis

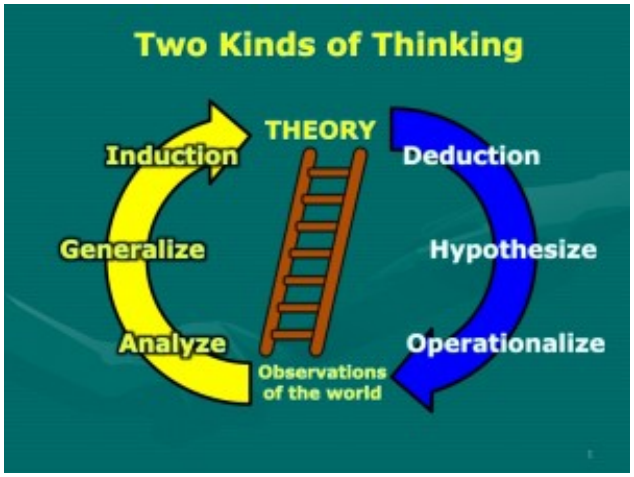
- **Statistical Analysis:**
  - Ill-suited for Nominal and Structured Data Types
  - Interpretation of results is difficult and daunting
  - Expert user guidance is critical
  - **Deductive**

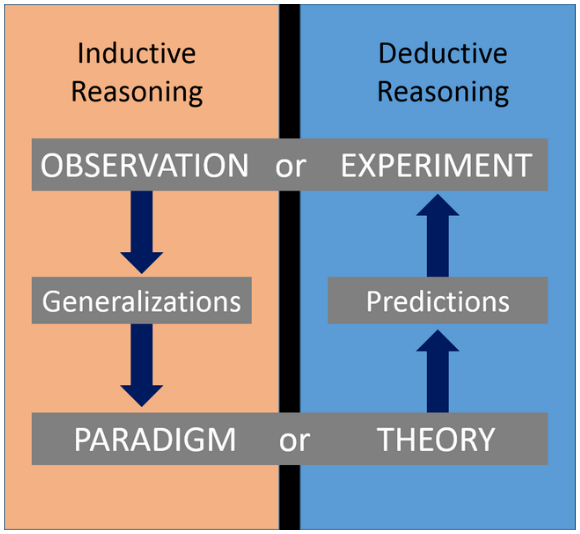
- **Data Science:**
  - Need of large Datasets
  - Efficiency & Scalability of Algorithms is important
  - Real World/ Pre-existing Data - not user generated
  - Even in large datasets, quality is relevant
  - Data may not be static
  - **Inductive**

Data Science and Induction Principle

Induction vs Deduction

- **Deductive reasoning** links premises to reach conclusions:
  — All men are mortal. Socrates is a man. Therefore, Socrates is mortal
- **Induction reasoning** adds information. Premises seek to supply strong evidence for (not absolute proof of) the truth of the conclusion:
  — I have a bag of many coins, and I've pulled 10 at random and they've all been pennies, therefore this is probably a bag full of pennies





Images from Daniel Miessler's newsletter

The Problems with Induction

- From true facts, we may induce false models
- The Black Swan
  - European swans are all white
  - Induce: ”Swans are white” as a general rule
  - Discover Australia and black Swans...
  - Problem: the set of examples is not random and representative
- Another example: distinguish US tanks from Iraqi tanks
  - Method: Database of pictures split in train set and test set; Classification model built on train set
  - Result: Good predictive accuracy on test set; Bad score on independent pictures
  - Why did it go wrong: other distinguishing features in the pictures (hangar versus desert)
- The risk of false model can be reduced by
  - A combination of quantity and quality of data
  - A thorough data preparation
  - Appropriate/advanced algorithms
  - A proper knowledge of the domain for testing the results/reiterate the process

Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - Web

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - — Examples: eye color of a person, temperature, etc.
  - — Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - — Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

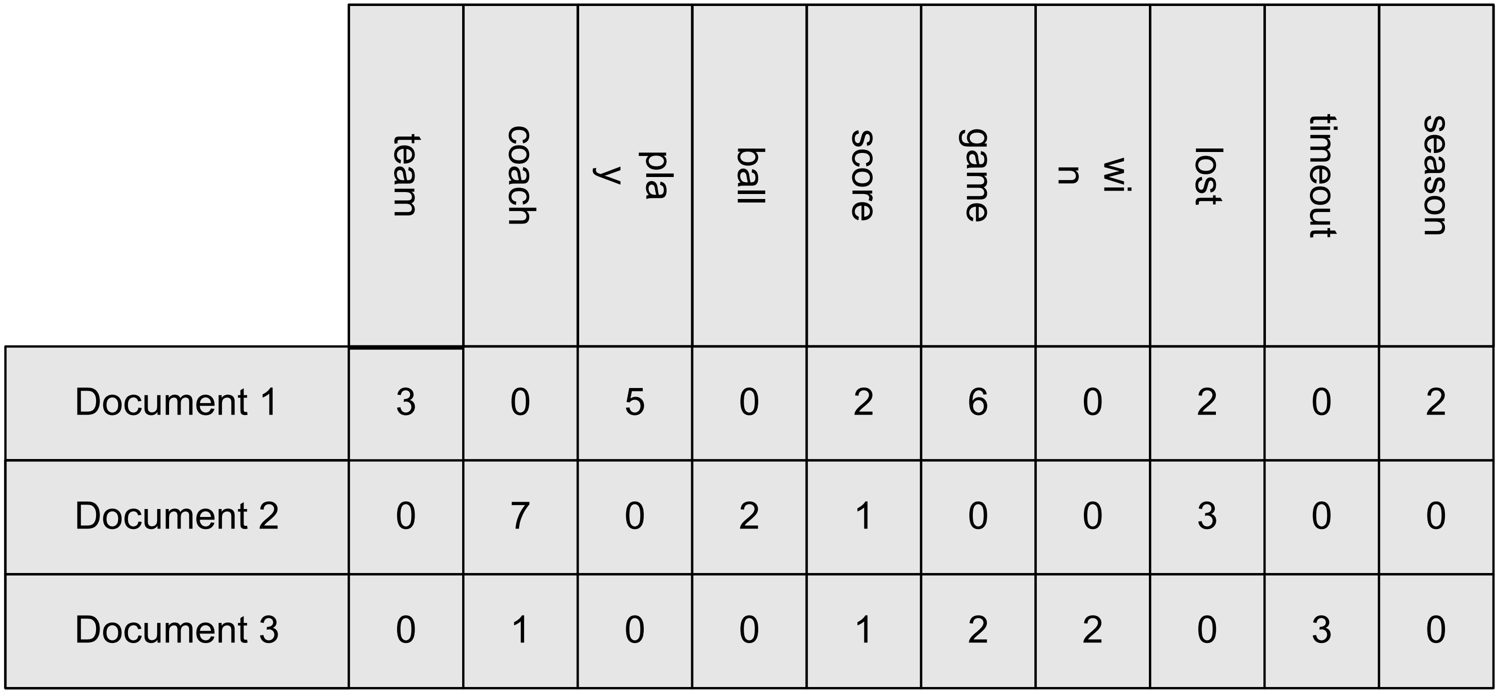
A Dataset

	X_1	X_2	X_3	Y
x_1				
x_2				
x_3				
x_4				
x_5				
x_6				

- **Rows**: Data/points/instances/examples/samples/records
- **Columns**: Features/attributes/dimensions/independent variables/covariates/predictors
- Variables: Target/outcome/response/label/dependent-independent

Document Data

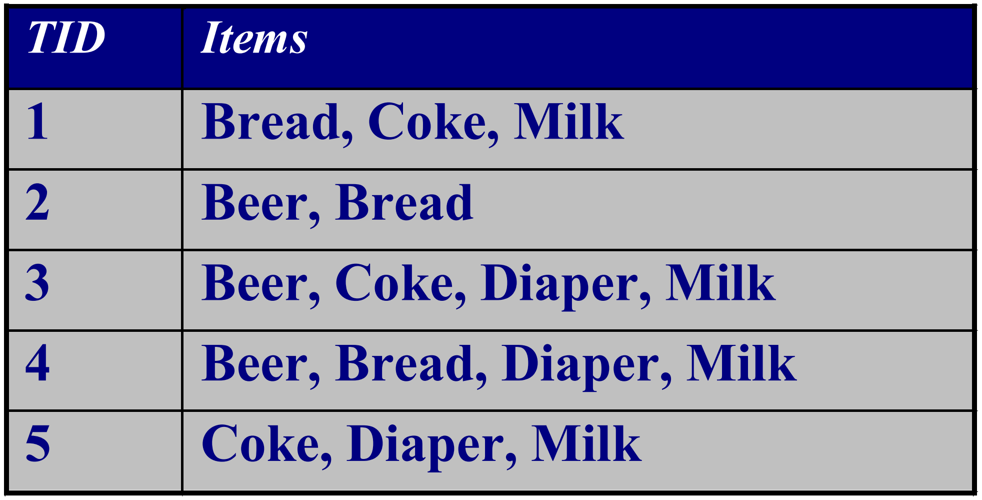
- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document



	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

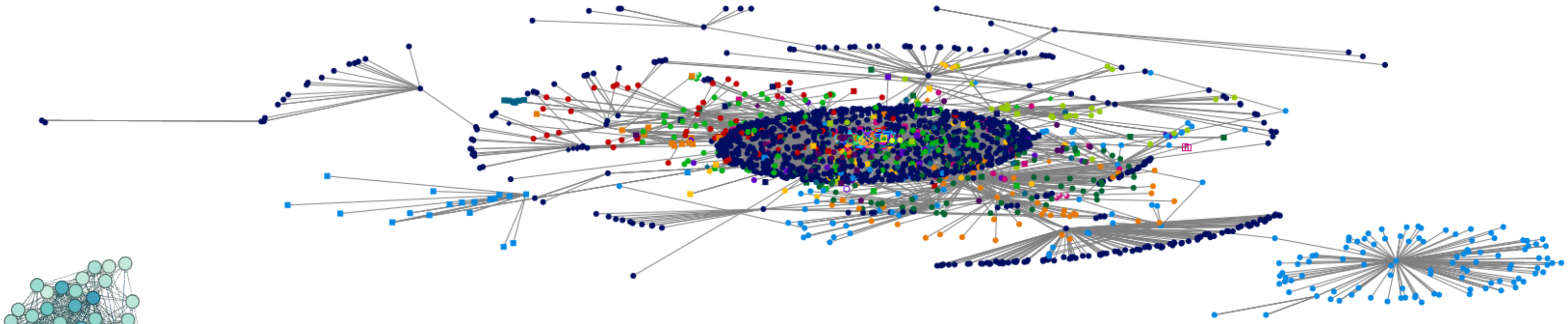
- A special type of record data, where
  - ○ each record (transaction) involves a set of items.
    - — For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items

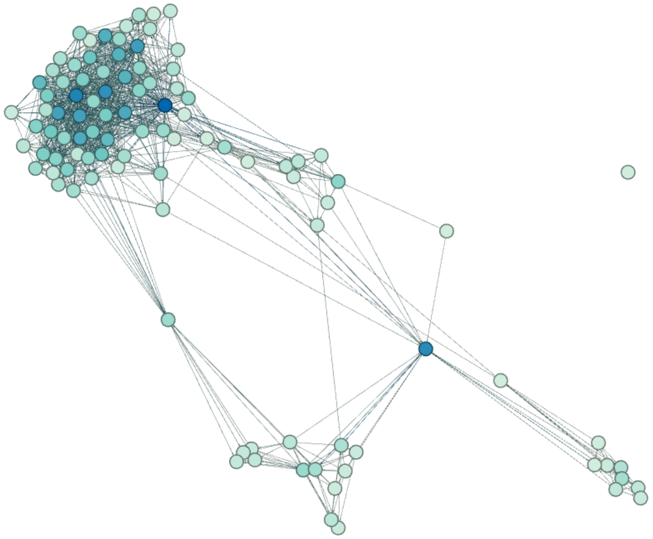


TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data

- This is graph for the 2nd presidential debate (Oct. 16-2012), showing who the most influencers are in the conversation. Nodes/Vertices are users
- It reflects the "popularity" of the users. For example, the big aggregation is around a user ("politifact"), with 910 re-tweets, with the second largest being "Eat_A_Brick" with 115 (south-east of politifact)

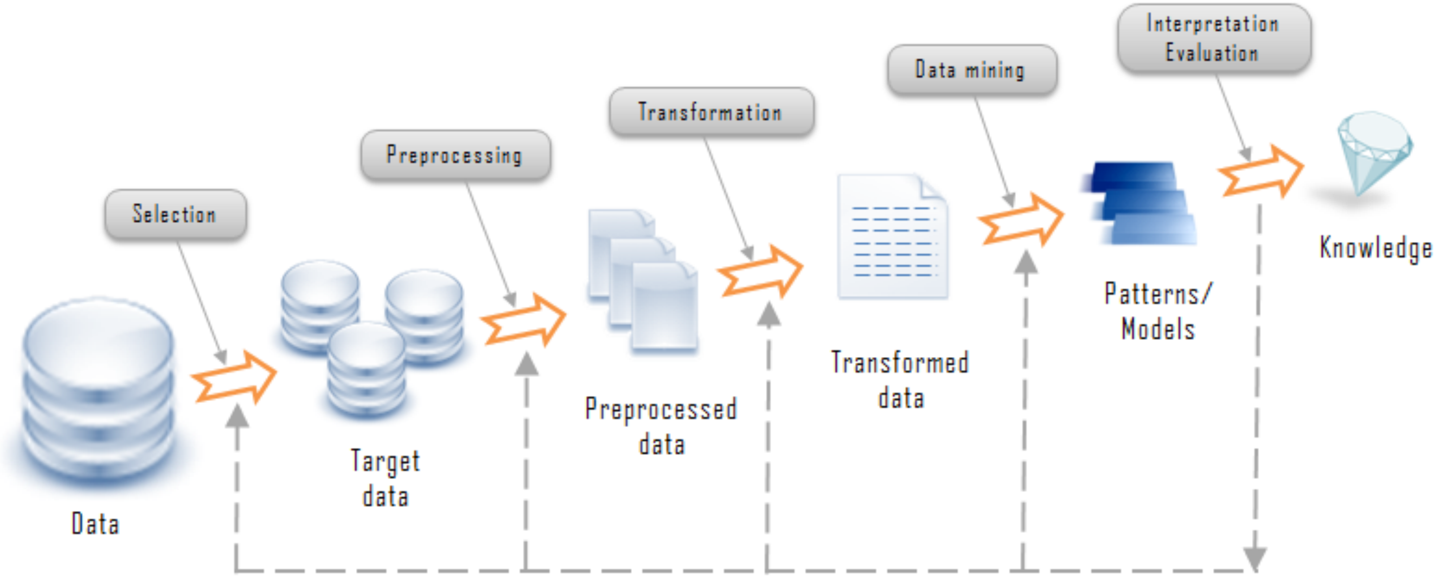




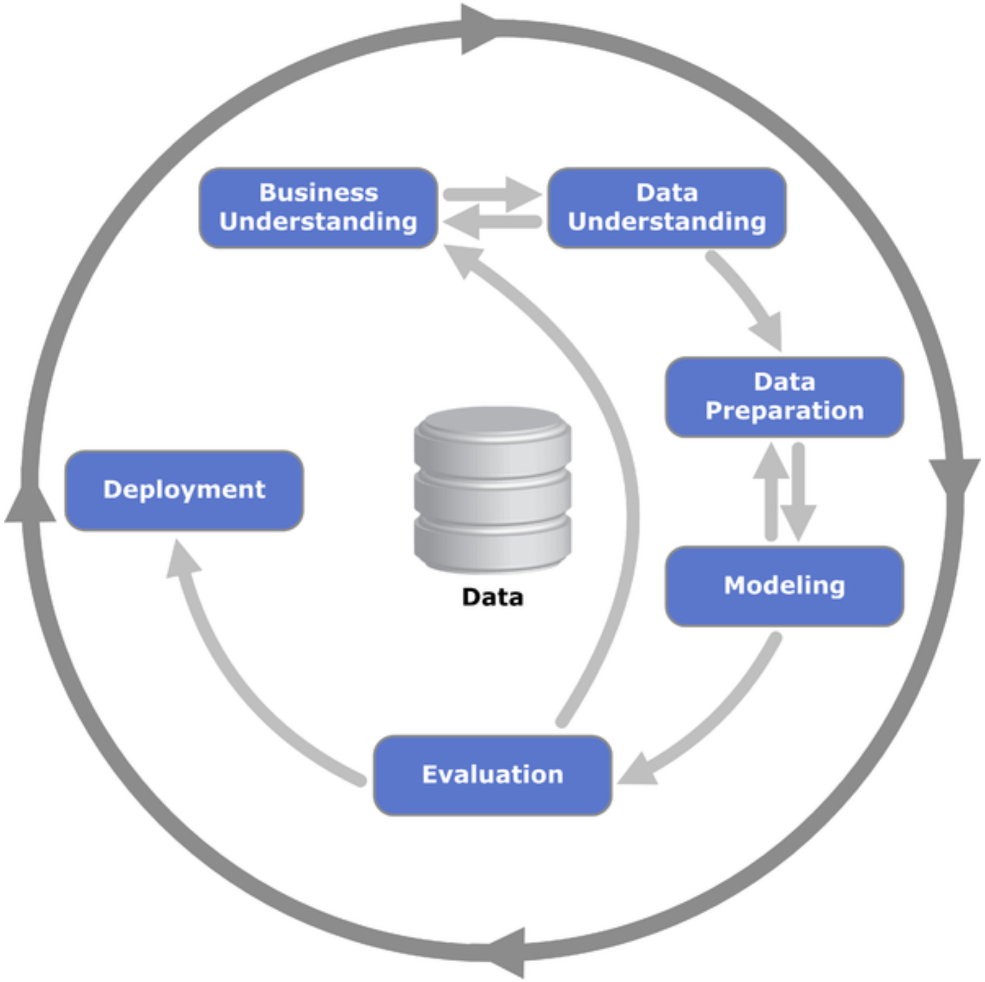
- This is my Facebook network, consisting of 313 nodes (people I'm connected with) and 1323 edges (links between my "friends")
- The "islands" represent in a sense areas of interest for me. The largest is my professional network on the top left

Data Mining Methodologies

- Several non formal methodologies available. Two more formally defined are:
  - SEMMA. It is a list of sequential steps developed by SAS Institute Inc
  - CRISP-DM. Polls conducted in multiple years show that it is the leading methodology used by data miners *[Gregory Piatetsky-Shapiro – WDD Nuggets]*



The phases of CRISP-DM



- A de facto industry standard for data mining
- Created between 1997-1999 by DaimlerChrysler, SPSS and NCR
- Acronym stands for Cross-Industry Standard Process for Data Mining
- Consists of 6 phases, intended as a cyclical process
- Not all phases are necessary in every analysis
- Somewhat similar to SAS Institute's SEMMA

5 Main Learning Tasks

What:

- Classification: predict a discrete target variable
- Clustering: predict clusters
- Regression: predict a continuous target variable
- Density estimation: predict the distribution
- Dimensionality reduction: predict new features

How:

- Supervised learning: We're predicting a target variable for which we get to see examples (regression, classification)
- Unsupervised learning: We're predicting a target variable for which we never get to see examples (density estimation, clustering, dimensionality reduction)

Data Mining Tasks

Prediction Tasks

- Use some variables to predict unknown or future values of other variables

Description Tasks

- Find human-interpretable patterns that describe the data

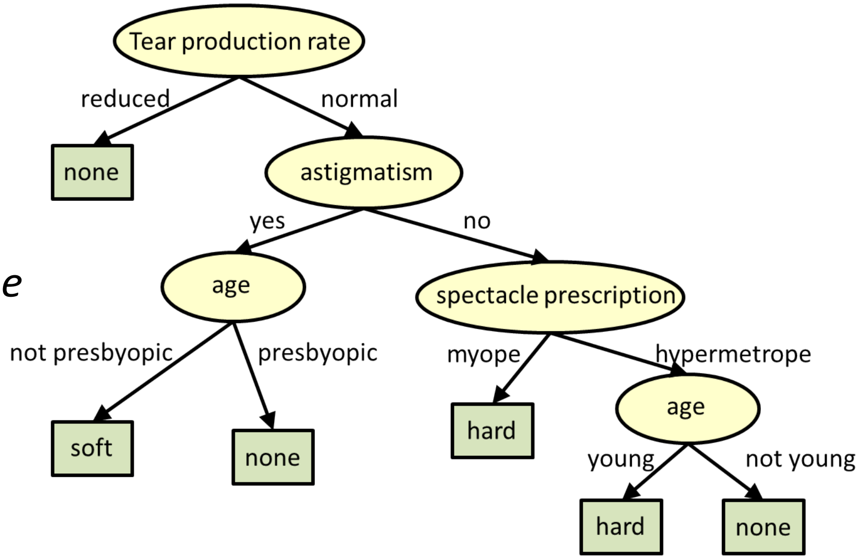
Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification

- Given a collection of records (training set). Each record contains a set of attributes, one of the attributes is the class
- Find a model for class attribute as a function of the values of other attributes
- Goal: previously unseen records should be assigned a class as accurately as possible
    - — A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

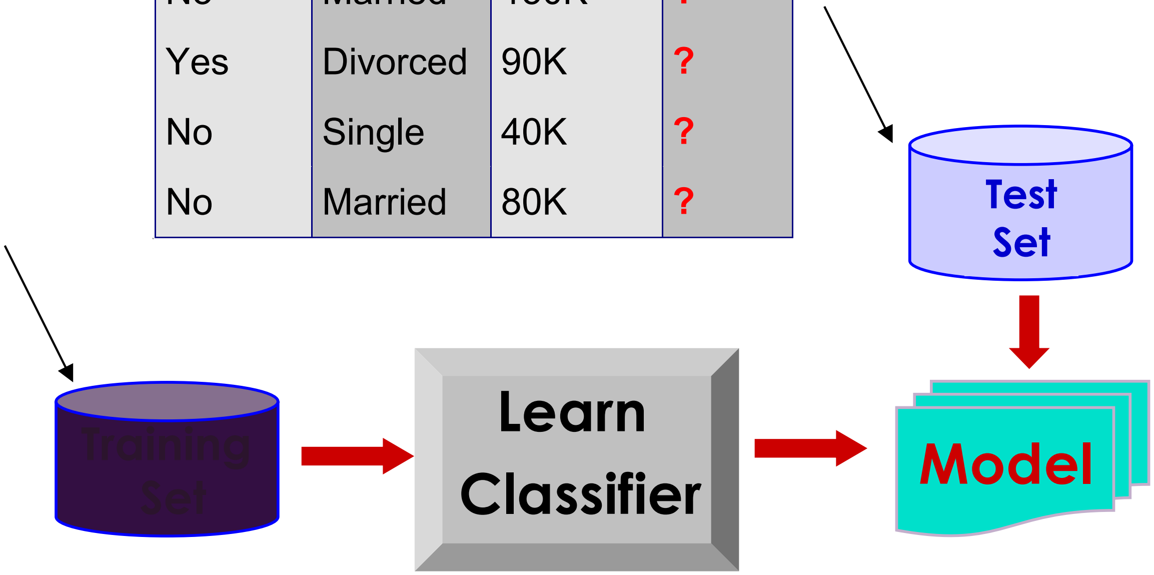
Example of algorithm: Decision Tree



Classification Example

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product
  - Approach:
    - Use the data for a similar product introduced before
    - We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute
    - Learn a model for the class of the transactions
    - Use this model to detect fraud by observing credit card transactions on an account

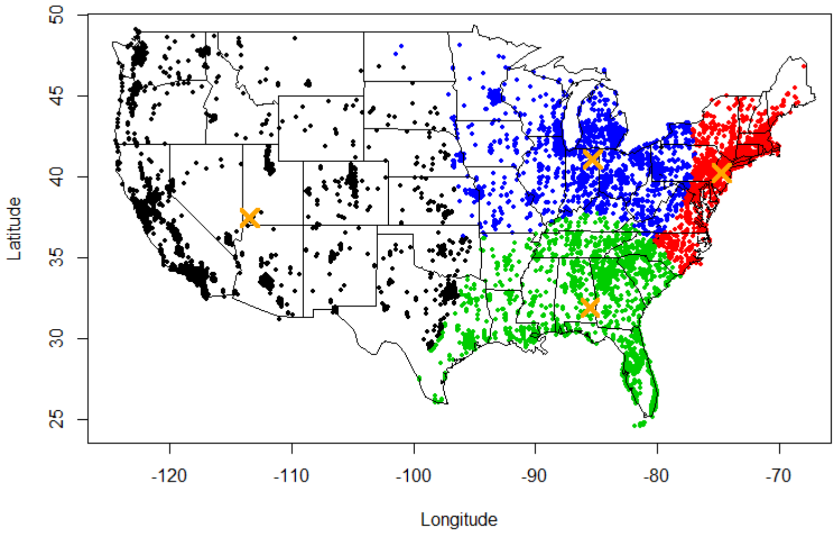
Classification: Application 3

- Customer Churn
  - Goal: To predict whether a customer is likely to be lost to a competitor
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal
    - Find a model for loyalty

Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another
- Similarity Measures:
  - Euclidean Distance if attributes are continuous
  - Problem-specific Measures

Example of algorithm: k-means on geo-location



Clustering: Application 1

- Market Segmentation
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information
    - Find clusters of similar customers
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters

Clustering: Application 2

- Document Clustering
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents

Association Rule Discovery

- Given a set of records each of which contains some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips

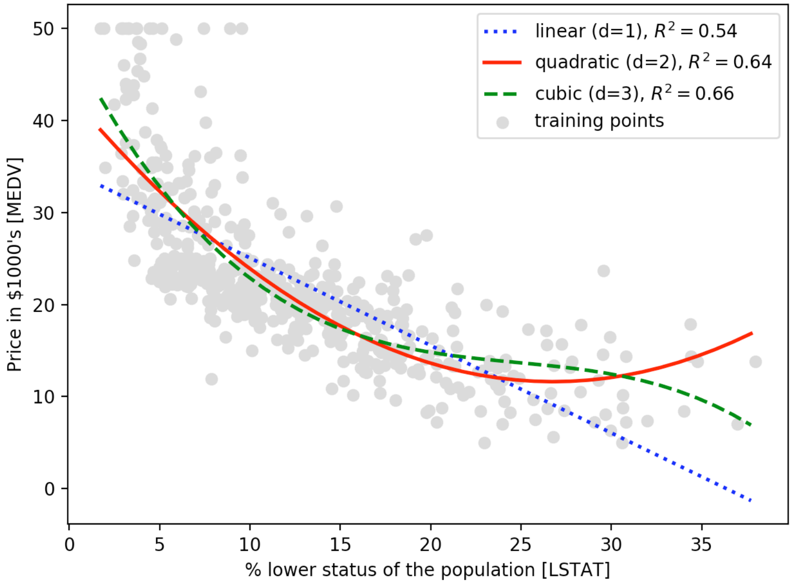
Association Rule Discovery: Application 2

- Supermarket shelf management
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Examples:
  - — Predicting sales amounts of new product based on advertising expenditure
  - — Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - — Time series prediction of stock market indices

Different forms of Regression



Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:

Credit Card Fraud Detection

SCAM ALERT

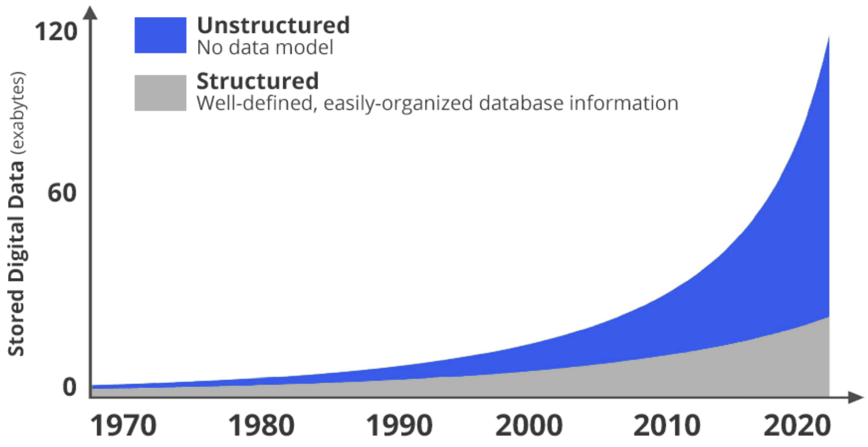
TARGET

Target announced millions of Credit Cards stolen - details here.

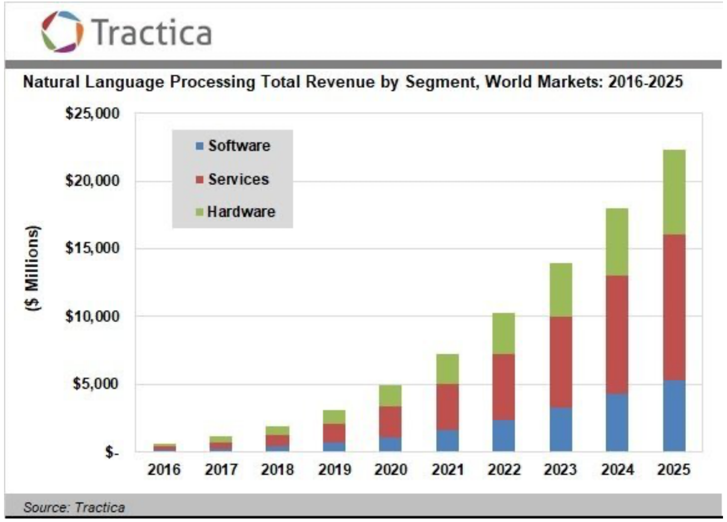
Network Intrusion Detection

Natural Language as source of Data

- 85-90 percent of all corporate data is in some kind of unstructured form, such as text and multimedia *[Gartner, 2019]*
- Tapping into these information sources is a need to stay competitive



Source: m-files.com



Source: Tractica

- Examples of application of **Natural Language Processing**: insurance (claim processing); law (court orders); academic research (research articles); finance (reports analysis); medicine (discharge summaries); technology (patent files); marketing (customer comments)

Challenges in Natural Language Processing

- Semantic ambiguity and context sensitivity
  - —automobile = car = vehicle = Toyota
  - —Apple (the company) or apple (the fruit)
- Syntactic/formal ambiguity
  - —Misspelling
  - —Different words for the same concept (e.g.: street; st.)
- Implicit knowledge
  - —We talk about things giving for granted common or specific knowledge

Implementing NLP

- Language is changing constantly, and NLP is following the changes, going from processing based on predefined structures (taxonomies/ontologies, syntax) to structures deducted from the text itself

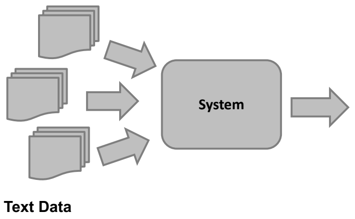
Limitations of the traditional/deductive approach

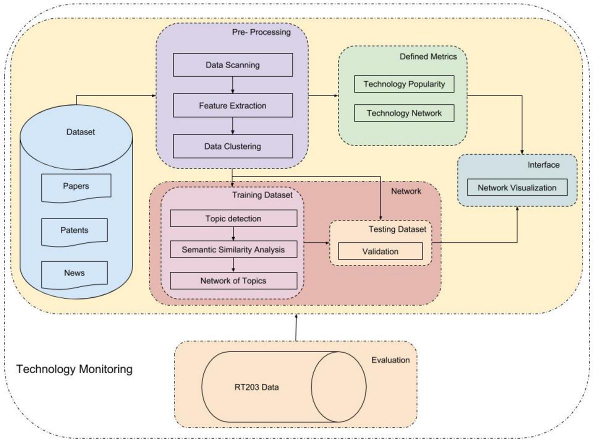
- Today, language is more fragmented, has less structure, has more jargons
- Different points of view may provide different interpretations

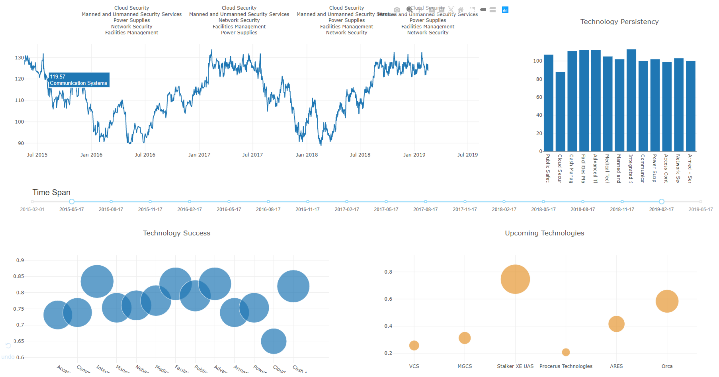
Machine Learning/inductive approach

- Extracting a numerical structure from text
- Different structures for different points of view
- Different structures automatically extracted over time

NLP at work: monitoring technology







A radar screen for coming and "future" technologies, along with a technology taxonomy generator

Next Week - *Data and the world: State of Practice*

- Data science is becoming an integral part of everything we do each day
- In this seminar, we describe how data driven systems are changing the way we live and work
- We will present what is the state of the practice in terms of approaches and key applications

STEVENS INSTITUTE of TECHNOLOGY

1870

THE INNOVATION UNIVERSITY®

Thank you!

Dr. Carlo Lipizzi
clipizzi@stevens.edu