SYSTEMS ENGINEERING RESEARCH CENTER

DAU

Digital Readiness Series

Data analytics

Data and the world: State of Practice

By
Dr. Carlo Lipizzi
August 6, 2020

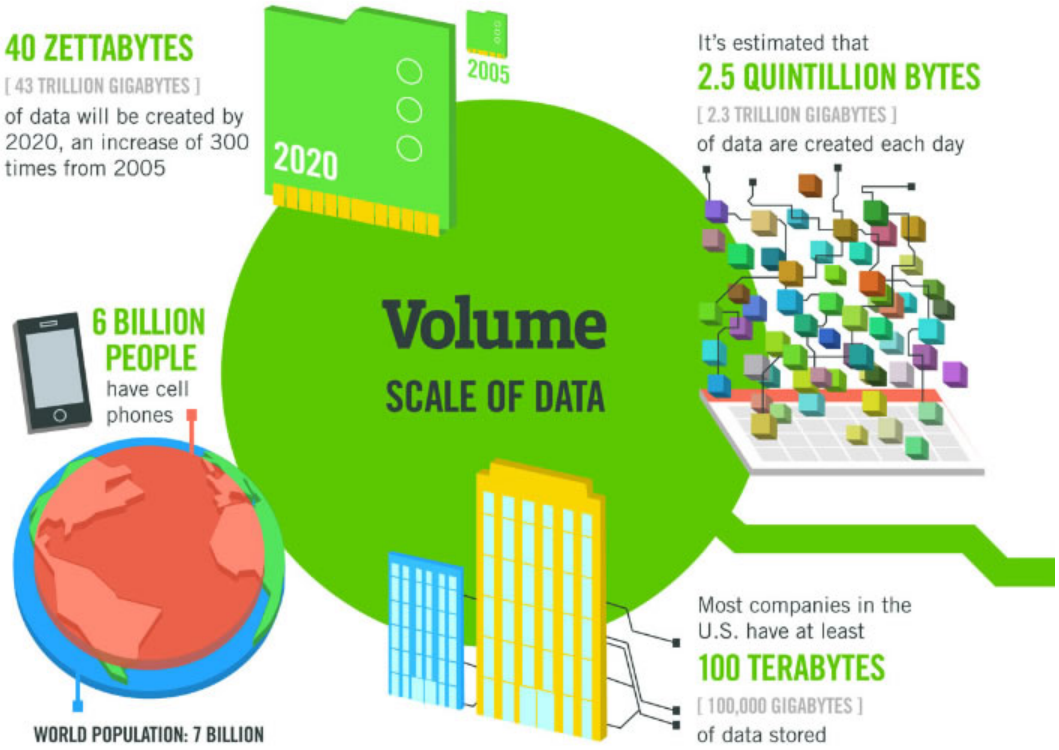
www.sercuarc.org

SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

Data growth



- 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

SYSTEMS ENGINEERING RESEARCH CENTER

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

SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

The relevance of the Medium: The Medium is the message?

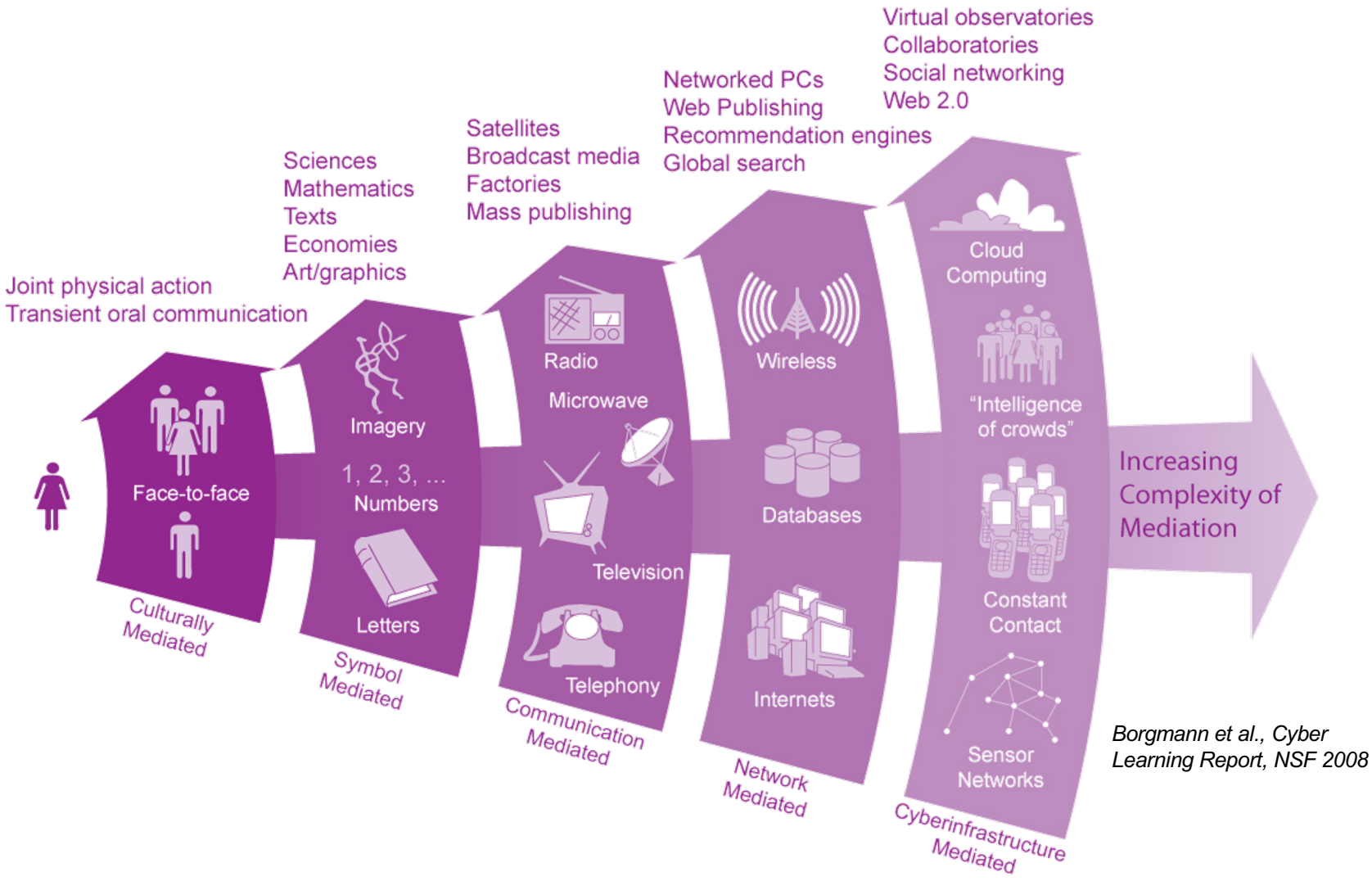
Marshall McLuhan (1911 – 1980)

"…technological media are staples or natural resources, exactly as are coal and cotton and oil"

"… it is the medium that shapes and controls the scale and form of human association and action. The content or uses of such media are as diverse as they are ineffectual in shaping the form of human association"

SYSTEMS ENGINEERING RESEARCH CENTER

The complexity of Learning Mediation



Borgmann et al., Cyber Learning Report, NSF 2008

Mediated Learning Experience refers to the way in which stimuli experienced in the environment are transformed by a mediating agent, usually a parent, teacher, sibling, or other decisive element in the life of the learner

The “Printing” society

Focused on

Religion

Education

Industry

Thought

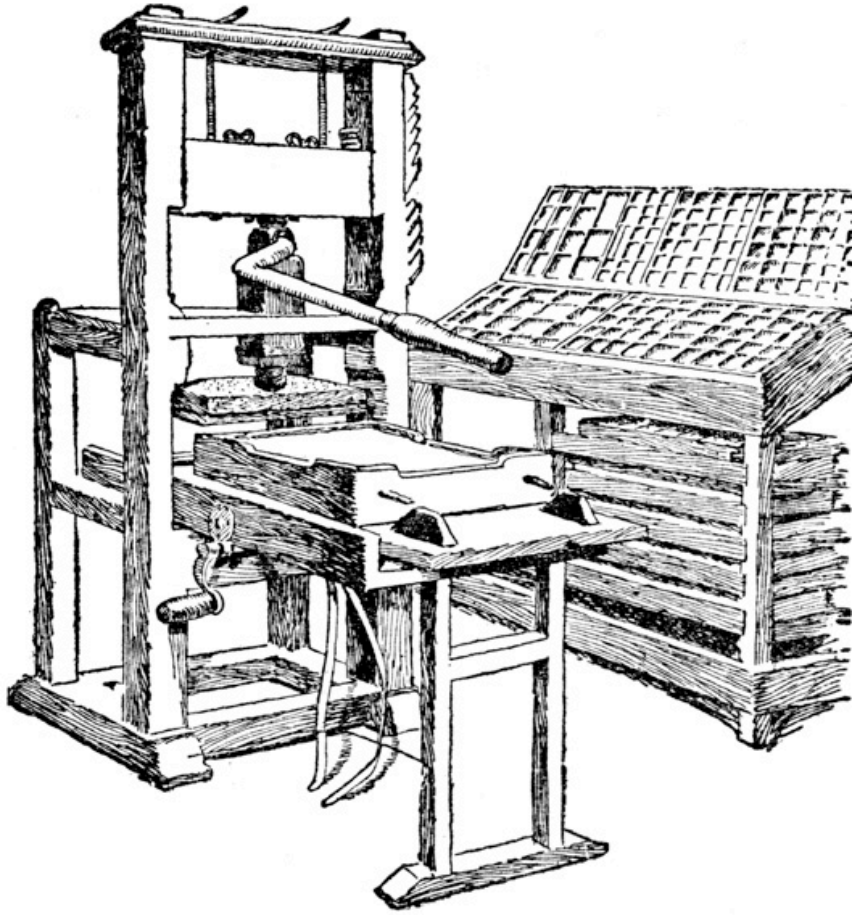
Conflict

Ideas

Community

Organization

Truth



Johannes Gutenberg 1398 – 1468

The “Mass Media” society

Focused on

Leisure time
Education
Knowledge of the other
Politics
Global Connections
Speed of Life
Mean World Effect
Minimizing Empathy

“At an accelerating pace throughout the century, the electronic transmission of news and entertainment changed virtually all features of American Life” (Robert Putnam, Bowling Alone)

SYSTEMS ENGINEERING RESEARCH CENTER

The “Always-on Internet” society

Focused on

Sharing

Cooperation

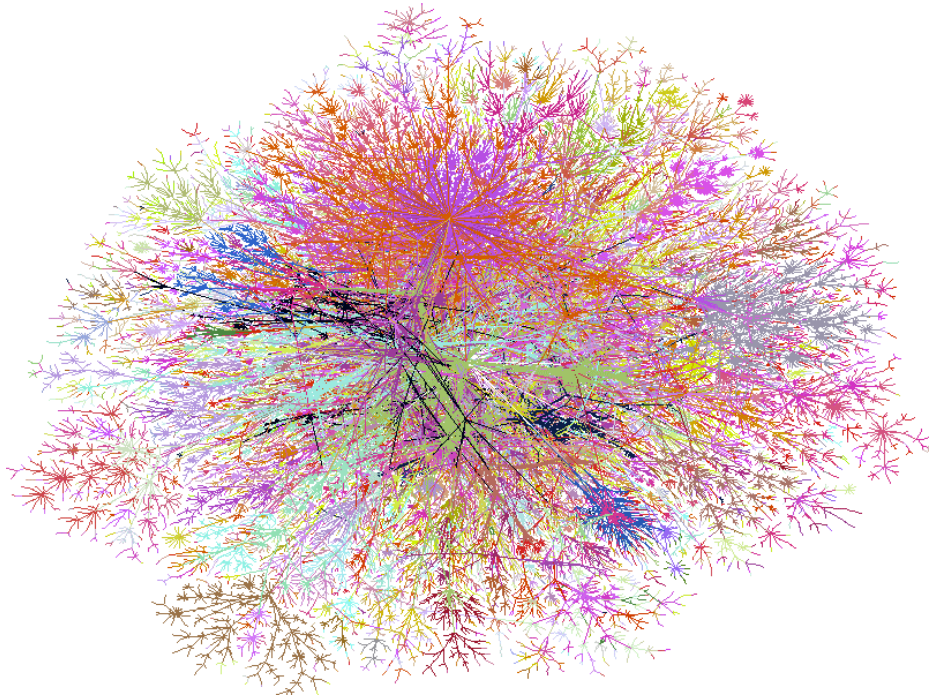
Collective Actions

Reduced Privacy

Source uncertainty

Information overload

“Living in the cloud”



SYSTEMS ENGINEERING RESEARCH CENTER

"Living in the cloud" effects: the overstimulation

- The average attention span in 2015 - 8 seconds
- The average attention span in 2000 - 12 seconds
- The average attention span of a goldfish - 9 seconds
- Percent of teens who forget major details of close friends and relatives - 25 %
- Percent of people who forget their own birthdays from time to time - 7 %
- Average number of times per hour an office worker checks their email inbox - 30
- Average length watched of a single internet video - 2.7 minutes

SYSTEMS ENGINEERING RESEARCH CENTER

How we communicate in "living in the cloud" times

- Percent of page views that last less than 4 seconds: 17 %
- Percent of page views that lasted more than 10 minutes: 4 %
- Percent of words read on web pages with 111 words or less: 49 %
- Percent of words read on an average (593 words) web page: 28 %
- Users spend only 4.4 seconds more for each additional 100 words

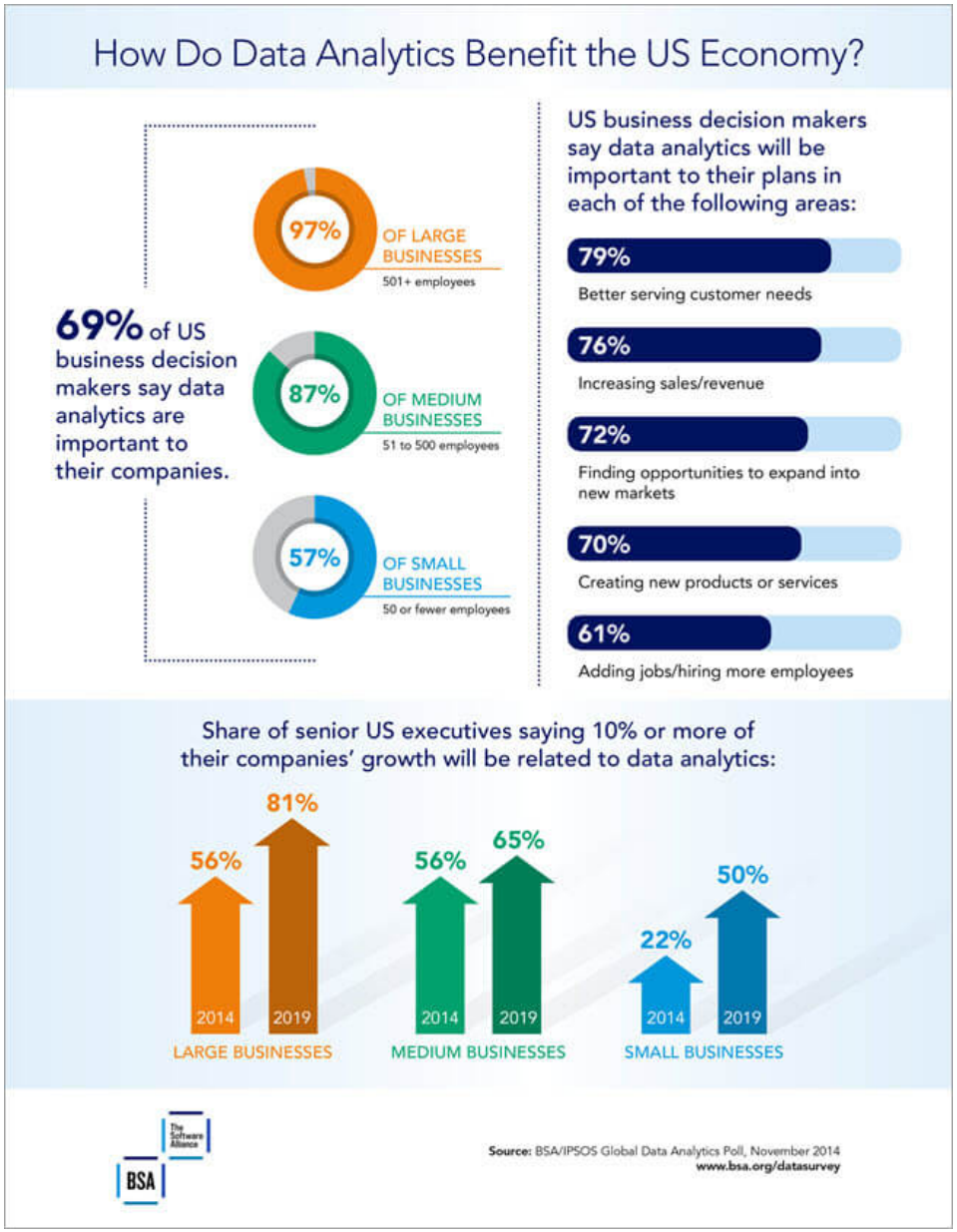
Source: Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer: "Not Quite the Average: An Empirical Study of Web Use," in the ACM Transactions on the Web, vol. 2, no. 1 (February 2008).

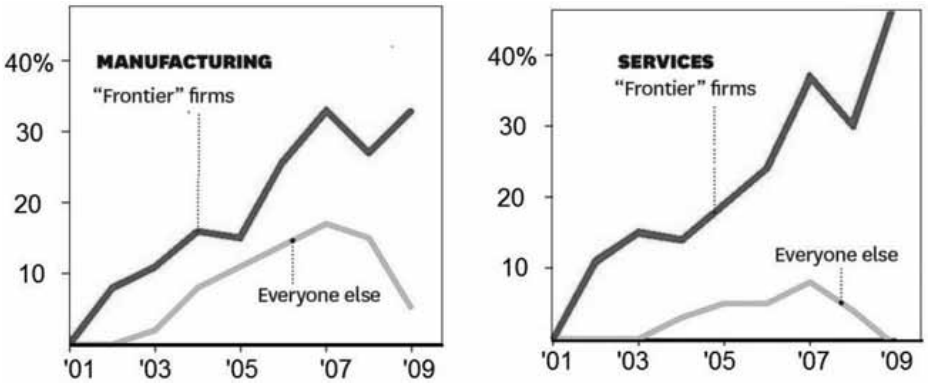
SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

Data and the Economy



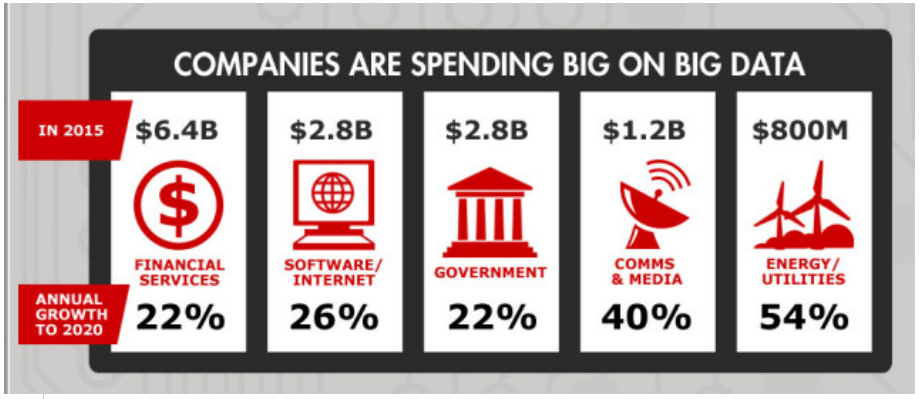


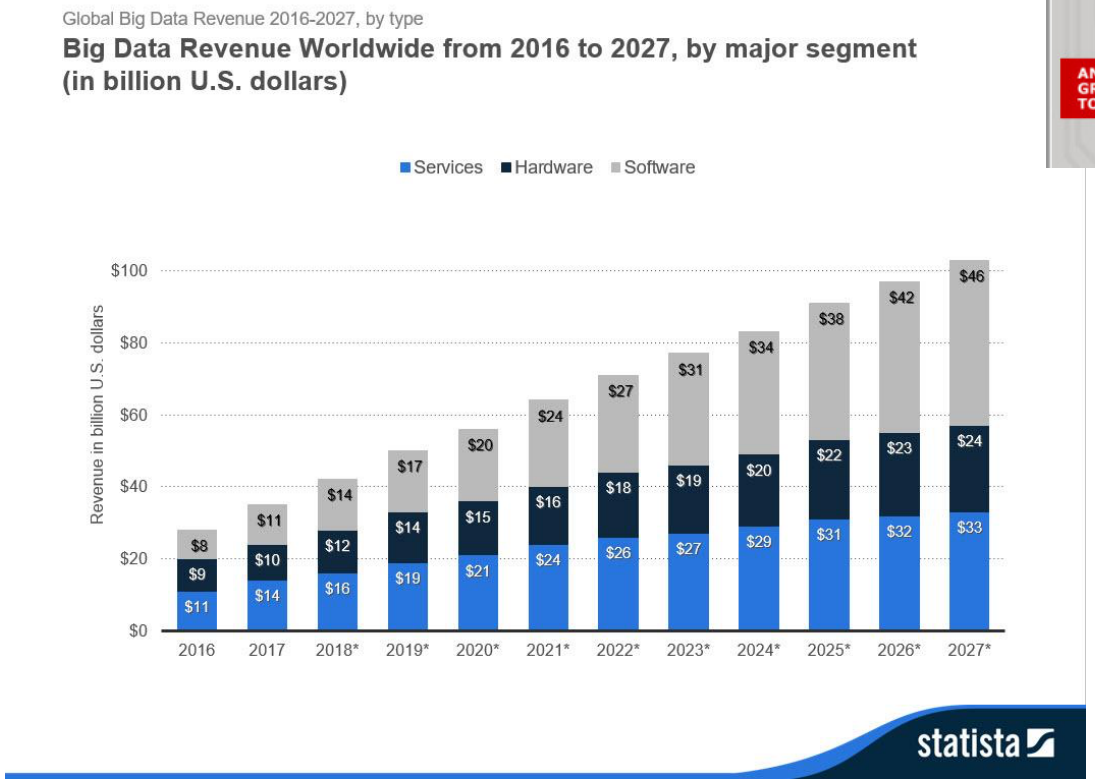
Source: Organization for Economic Co-operation and Development (OECD)

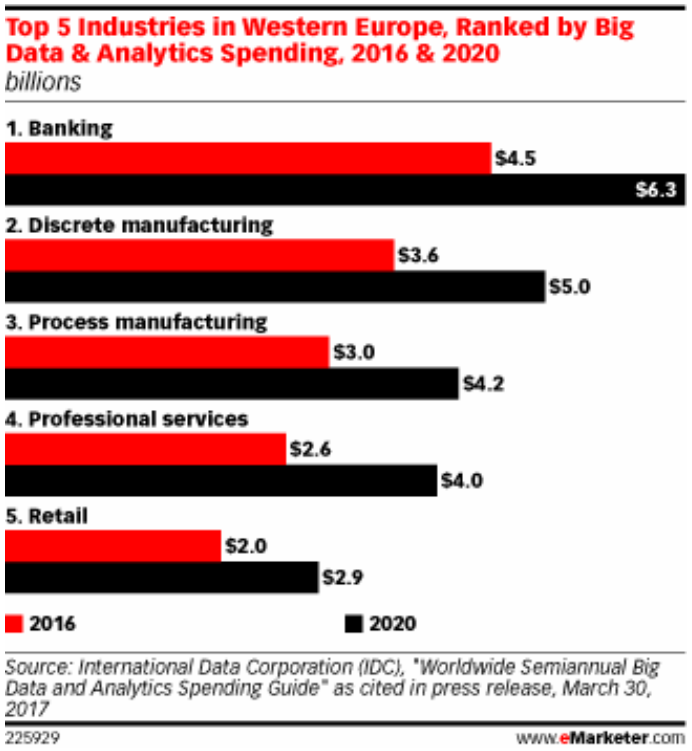
- Leading firms are more productive, more profitable, more innovative, and pay higher wages
- One reason often cited in the academic literature is that superstar firms are succeeding in large part due to leveraging on data

SYSTEMS ENGINEERING RESEARCH CENTER

How companies leverage on Data

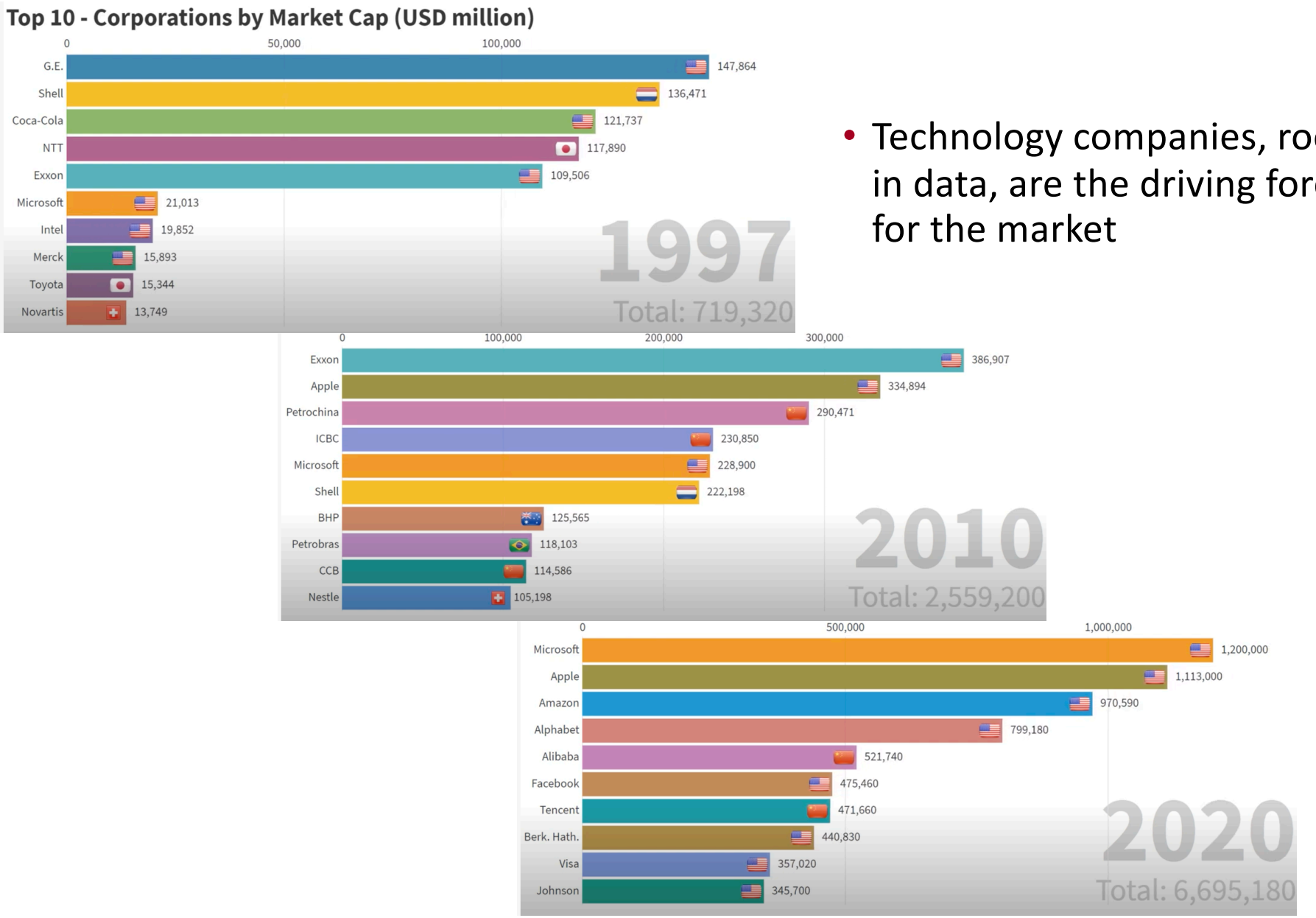






SYSTEMS ENGINEERING RESEARCH CENTER

The rise of giant tech companies

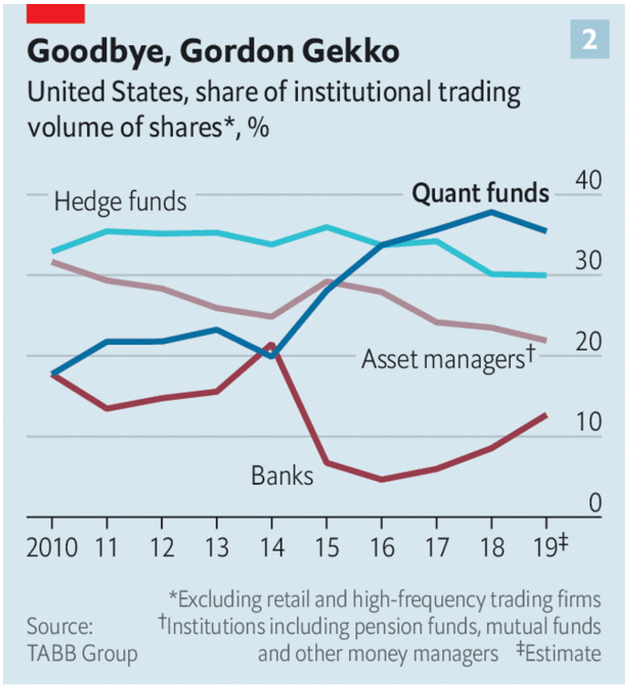


- Technology companies, rooted in data, are the driving force for the market

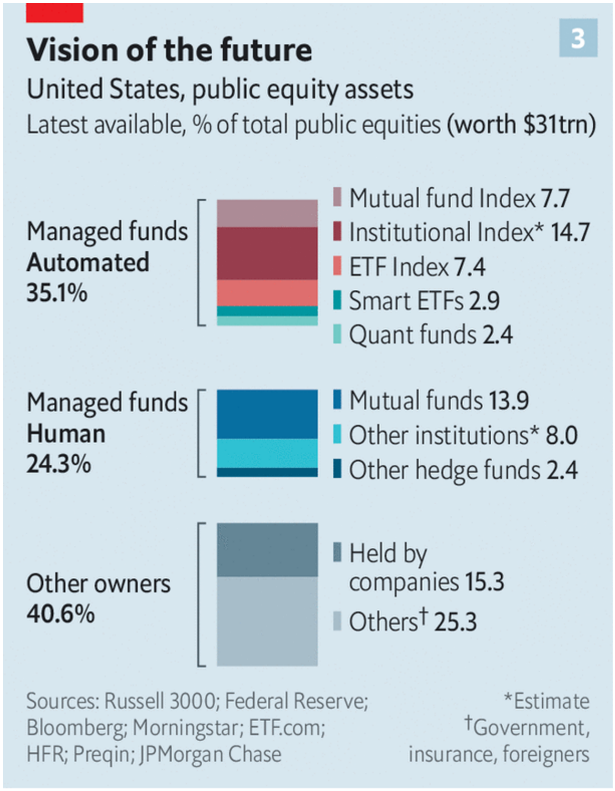
Source: YouTube - https://youtu.be/fgEHSzZm6gg

SYSTEMS ENGINEERING RESEARCH CENTER

Finance and the Algorithms



The Economist

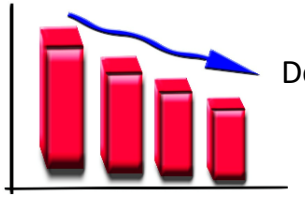


The Economist

- In early ’70s it was electronic execution; in mid ‘70s first index fund
- In the ‘80s and ‘90s quantitative hedge funds and exchange-traded funds. Quant funds program algorithms choose stocks based on factors determined by data analytics driven by economic theories
- According to Deutsche Bank, 90% of equity-futures trades and 80% of cash-equity trades are executed by algorithms without any human input

SYSTEMS ENGINEERING RESEARCH CENTER

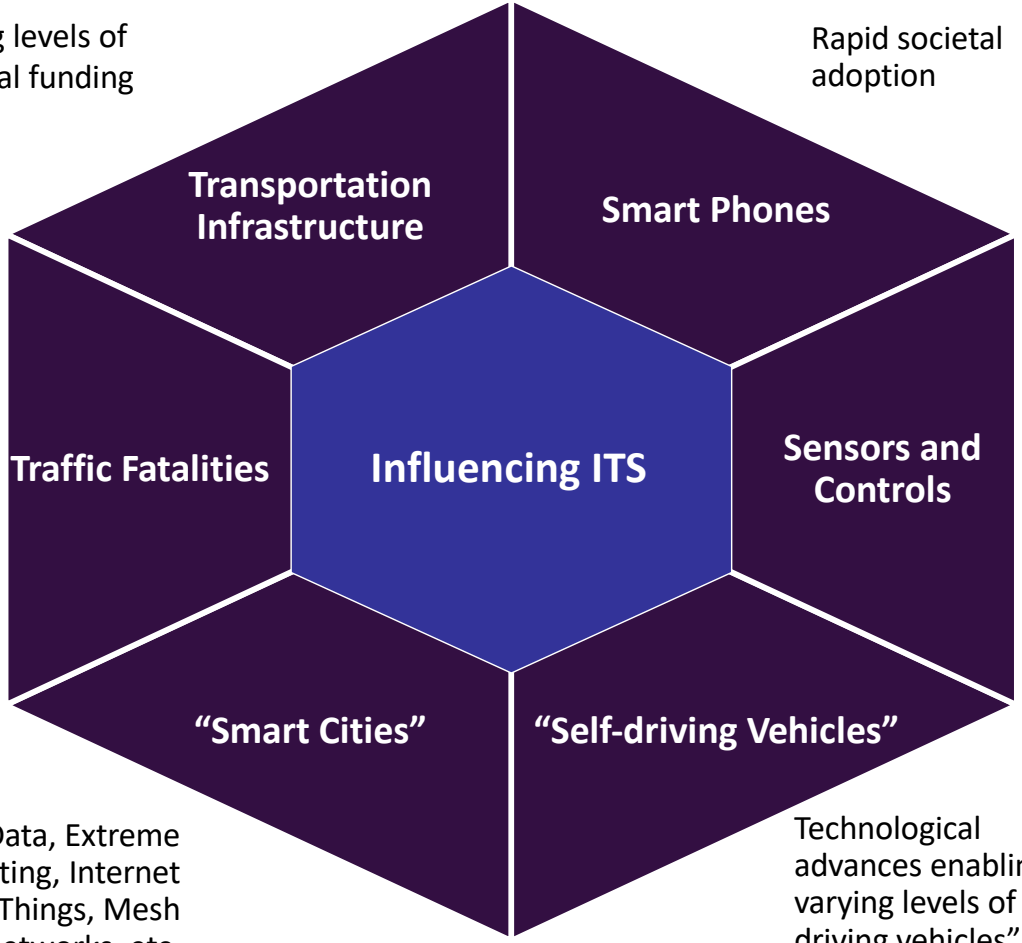
Data and Transportation - ITS



Declining levels of federal funding

Rapid societal adoption

Distracted drivers, bicyclists and pedestrians are increasing traffic fatalities



Rapid developments for automotive use

Big Data, Extreme Computing, Internet of Things, Mesh Networks, etc.

Technological advances enabling varying levels of "self-driving vehicles"

Data and Transportation - ITS

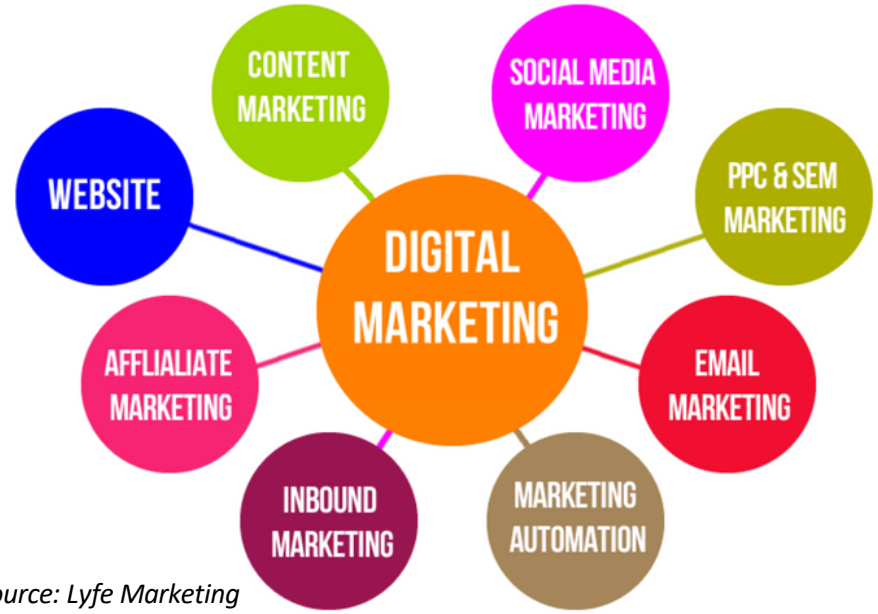
Public and Private Sector Roles (2016-Future)

	Public Sector	Private Sector
Infra—structure	Focus will be on repairing and maintaining physical infrastructure (including ordinary traffic engineering and signage as well as expanding **intelligent** traffic signal systems to improve traffic flow)	Transportation Networks will become elements of Smart Cities, and the private sector will own much of the **data**
Vehicles	Will promote advances in **collision avoidance systems** to reduce V2V and "Vehicle-to-Pedestrian" (V2P) collisions	Vehicles will become Internet Protocol (IP) nodes, collecting **data** for Smart Cities companies
Travelers	Step up efforts to reduce collisions resulting from distracted driver and distracted pedestrians/bicycles	Travelers will make trip choices based on user-optimized constraints, subject to **algorithms** used by private companies providing crowd-sourced traffic information and navigation services

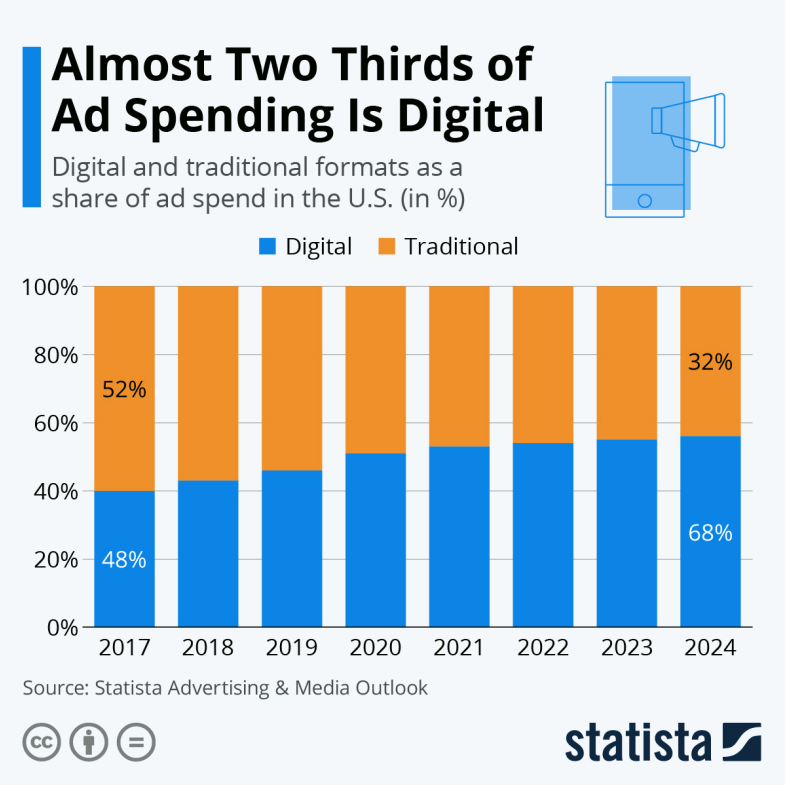
Data and Marketing

Source: Lyfe Marketing

Data and Marketing

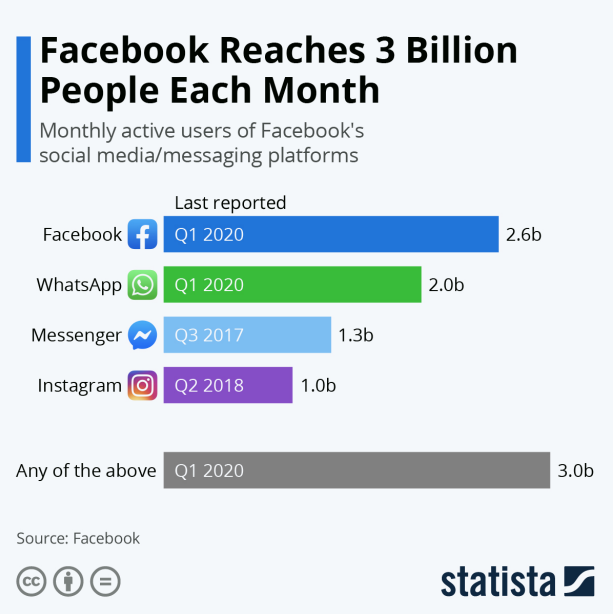


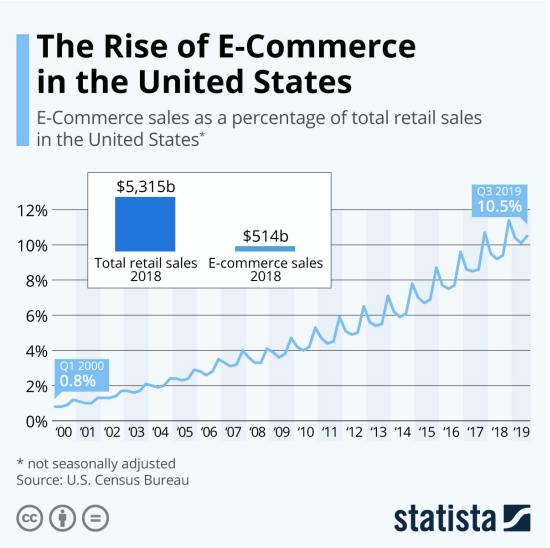
Source: Lyfe Marketing



- 60% of marketers across various industries have already shifted their efforts towards digital marketing
- 94% of B2B marketers are actively using LinkedIn for marketing
- Mobile will be accounting for over 70% of digital ad spend by 2019
- 90% B2C businesses report social media as being the most effective content marketing tactic

Data and Marketing





- Almost 60% of U.S. adults use Facebook on a regular basis
- More than 80% of shoppers/buyers do their research online before investing in a product/service

SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

SYSTEMS ENGINEERING RESEARCH CENTER

Top data trends happening now

Top 10 Data and Analytics Trends That Will Change Your Business

Scaling Business Impact	• Smarter, faster, more responsible AI • Decline of the dashboard • Decision intelligence • X analytics
Transforming Deployment	• Augmented data management: Metadata is the new black • Cloud is a given • Data and analytics worlds collide
Increasing Data and Analytics Value	• Data marketplaces/exchanges • Practical blockchain (for data and analytics) • Relationships form the foundation of data and analytics value

Source: Gartner

SYSTEMS ENGINEERING RESEARCH CENTER

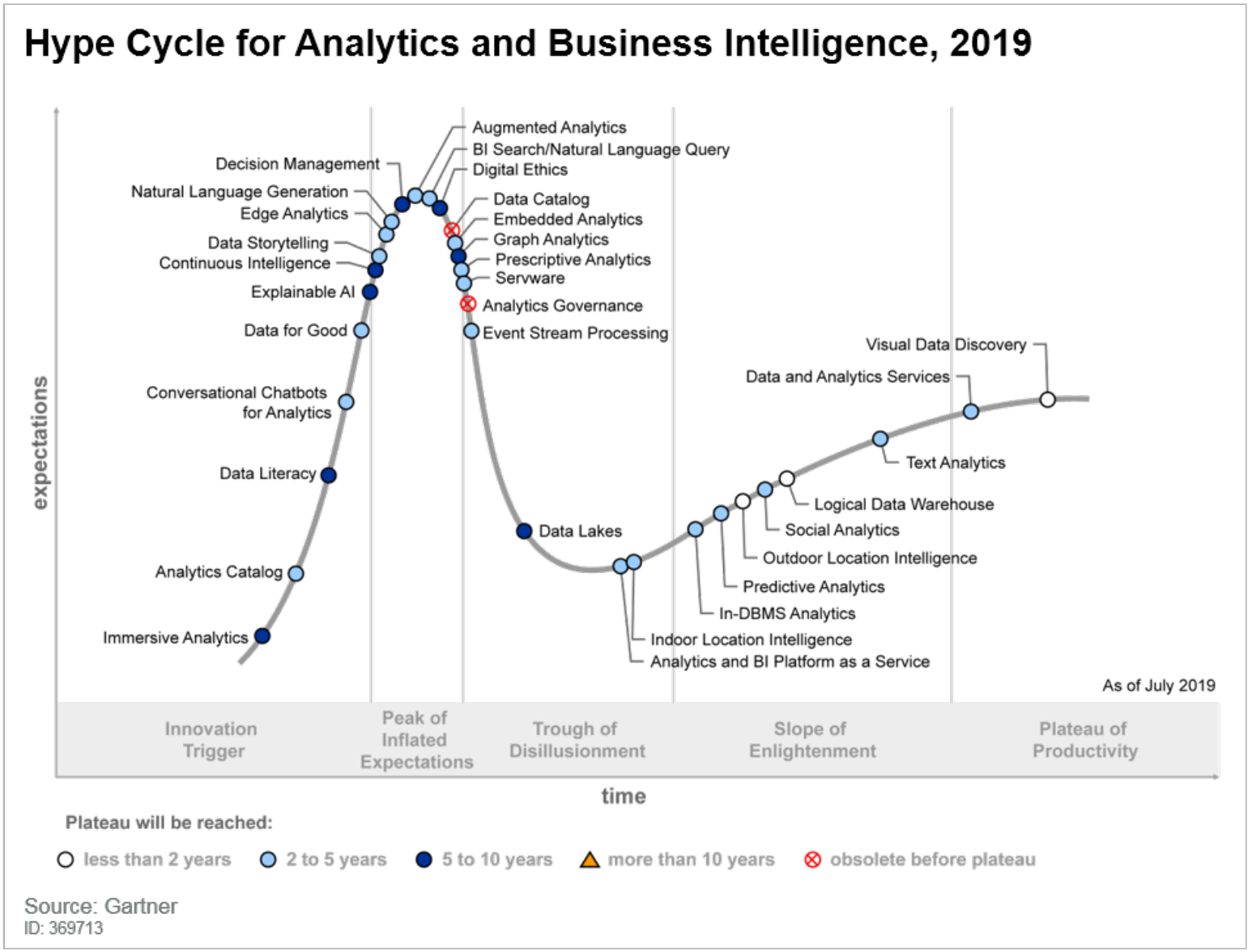
Top data trends happening now

Key contributing factors already happening

- *Decline of the Dashboard* - Data stories will be the most widespread way of consuming analytics, and stories will be automatically generated using augmented analytics techniques
- *X Analytics* (e.g., text analytics, video analytics, audio analytics, etc.) - AI-supported content analytics for video, audio, vibration, text, emotion will trigger major innovations and transformations
- *Augmented Data Management*: Metadata Is "the New Black" - Organizations will utilize active metadata, machine learning and data fabrics to dynamically connect, optimizing and automating most of the data management processes
- *Data Marketplaces and Exchanges* - Large organizations will be either sellers or buyers of data via formal online data marketplaces
- *Relationships Form the Foundation of Data and Analytics Value* - Graph technologies will facilitate rapid contextualization for decision making

Source: processed Gartner info

Analytics and Business Intelligence



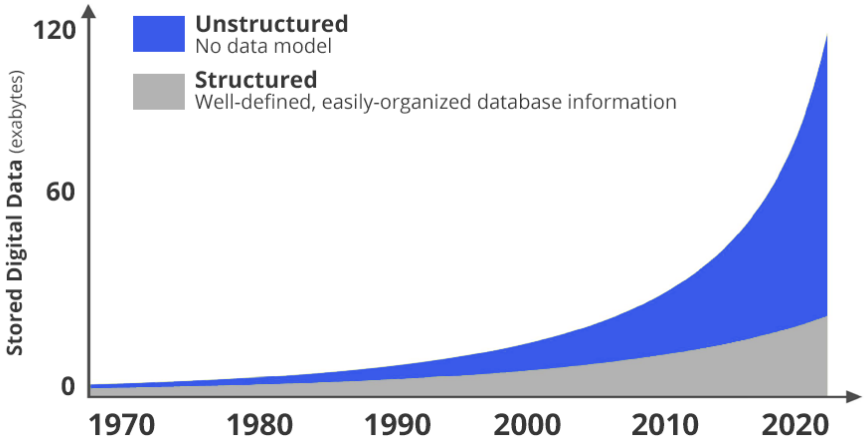
SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

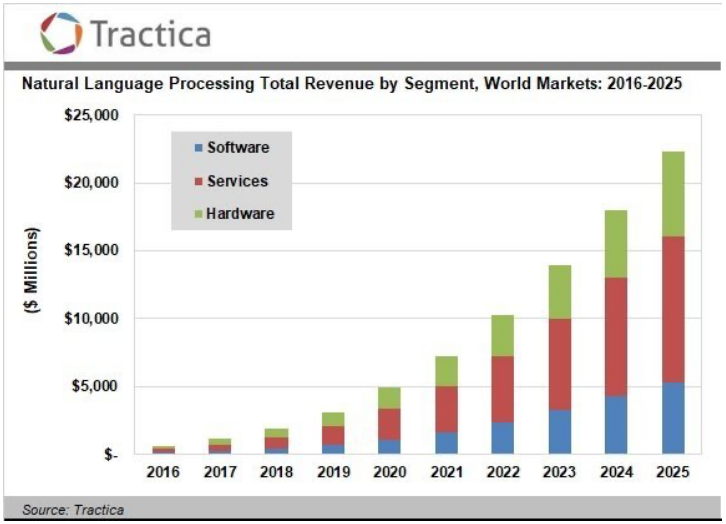
- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

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



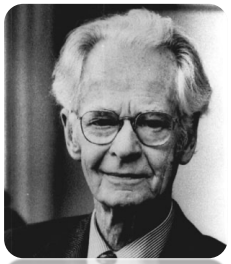
Source: m-files.com



- 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)

SYSTEMS ENGINEERING RESEARCH CENTER

How we acquire Language



B.F. Skinner

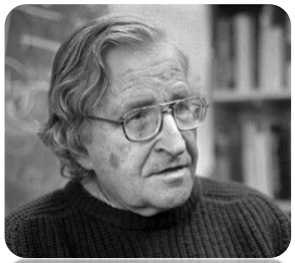
Behaviorist theory

infants learn language from other human role models through a process involving **imitation**, **rewards**, and **practice**

J. Piaget

Constructivist theory

Language is acquired within the context of the child's broader intellectual development. **Language is not an independent system**, but part of our general cognitive makeup



N. Chomsky

Nativist theory

Children are born equipped with an **innate template for language**, and this blueprint aids the child in the task of constructing a grammar for their language

SYSTEMS ENGINEERING RESEARCH CENTER

Language as part of human evolution

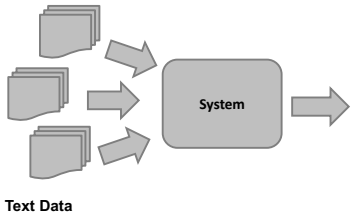
Neanderthal australopithecine Homo erectus

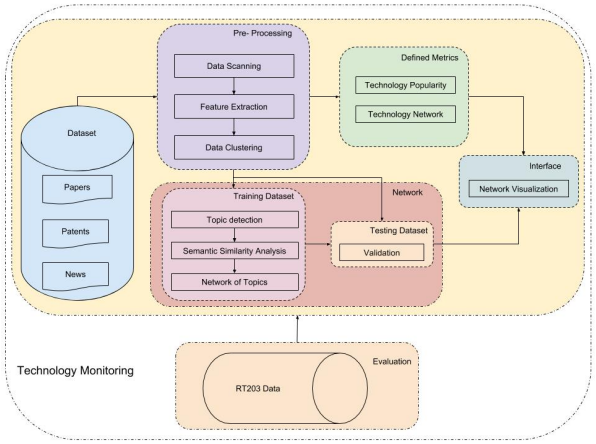
SYSTEMS ENGINEERING RESEARCH CENTER

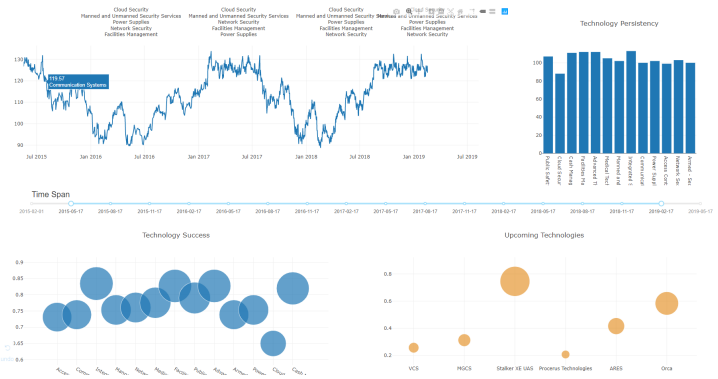
Implementing NLP

- Language is constantly changing, and NLP has to follow the changes, going from processing based on predefined structures (taxonomies/ontologies, syntax) to based on structures deducted from the text itself
- Language has a double bias: one from the originator, one from the recipient (listener/reader). We may not be able to do much on the first, we need to address the second
- If we want to process large amount of text, we need to create a numeric/computational layer out it
- To properly use language in our analyses, we need to
  - Extract a computational structure from the text
  - Create different structures/"knowledge bases" for the different points of view
  - Create metrics that can properly represent the type of information we want to extract

NLP at work: monitoring technology

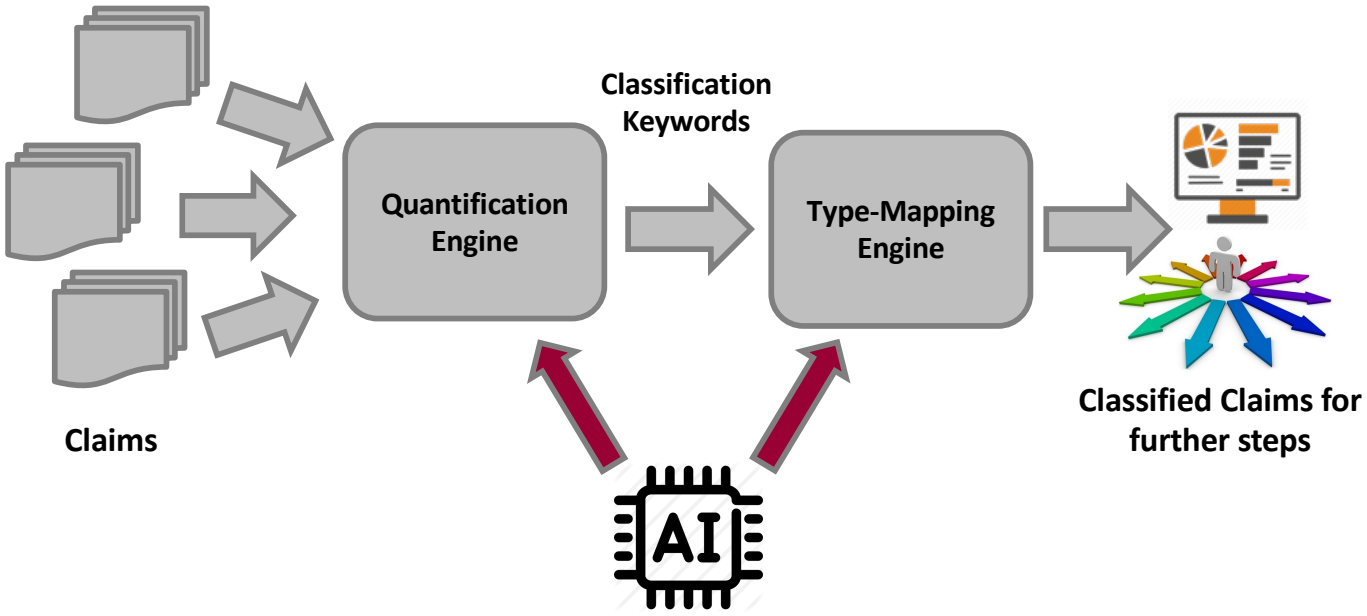






A radar screen for coming and “future” technologies, along with a technology taxonomy generator

NLP For Insurance



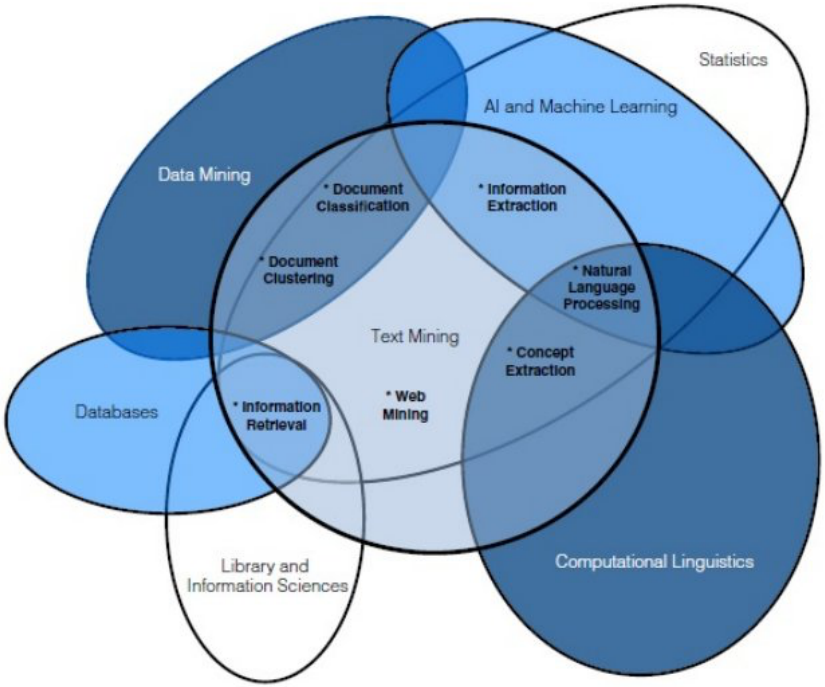
- This is a preprocessing/triage system taking Claims as input and process them for the rule system
- Claims have distinctive characteristics/keywords, that we use as benchmarks for the triage
- We compare – using a computational representation of the expert's knowledge – words in the claim with the benchmarks
- From the comparison we extract a number we use for the triage

SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

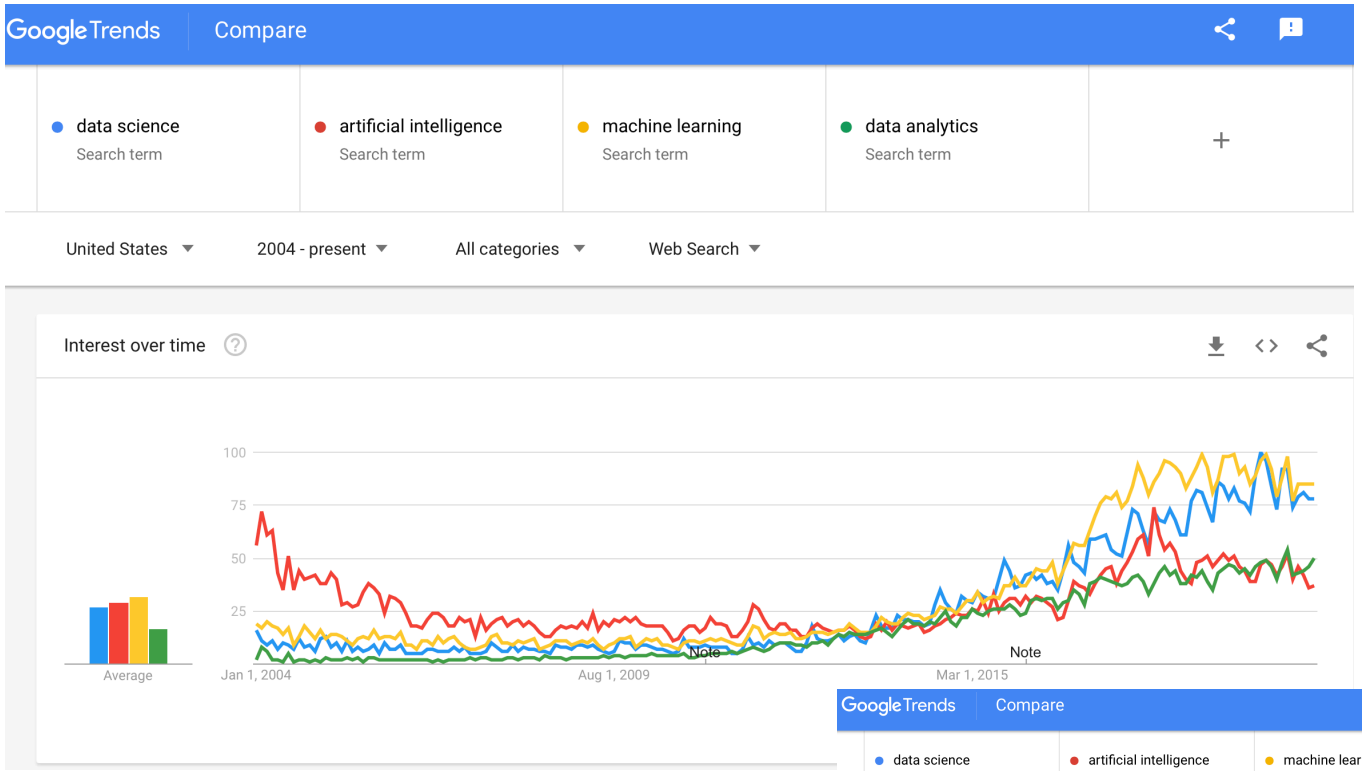
Data Science Components

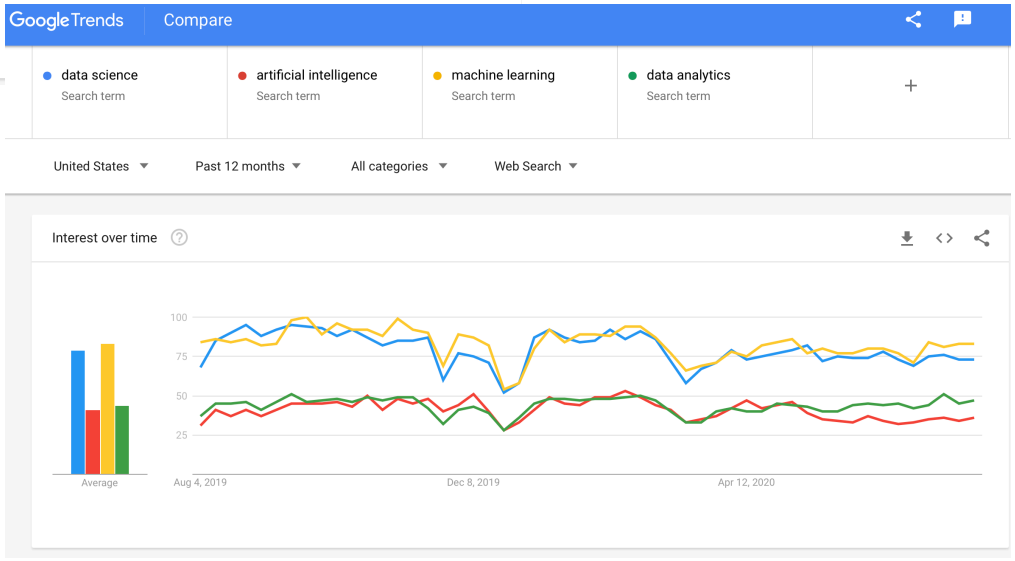


--- Focus on data complexity ---

- **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

The overlapping world of Analytics and AI/ML



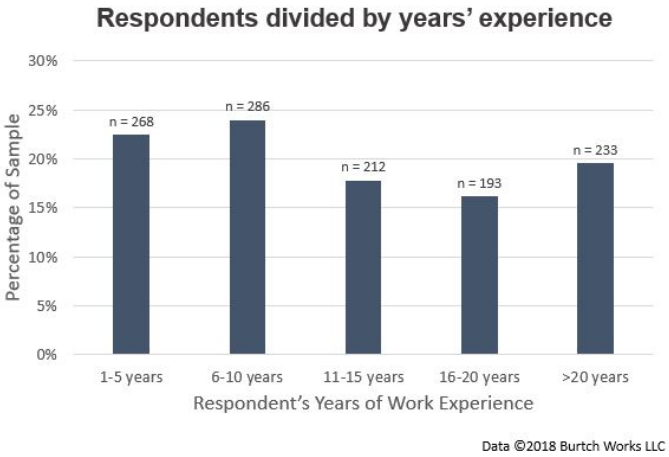


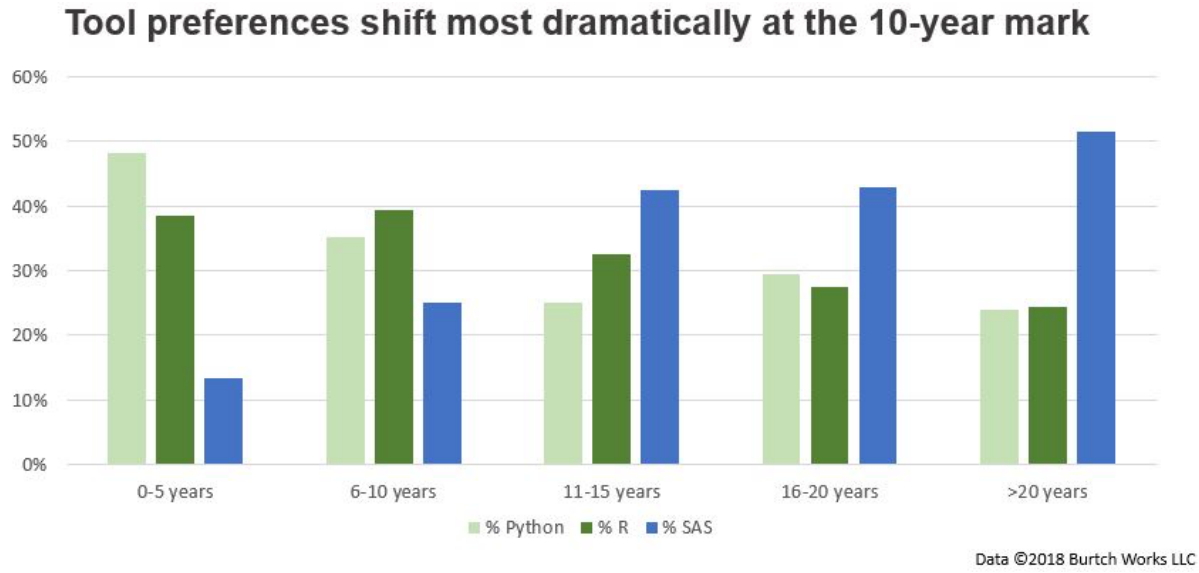
SYSTEMS ENGINEERING RESEARCH CENTER

Tools for Data Science

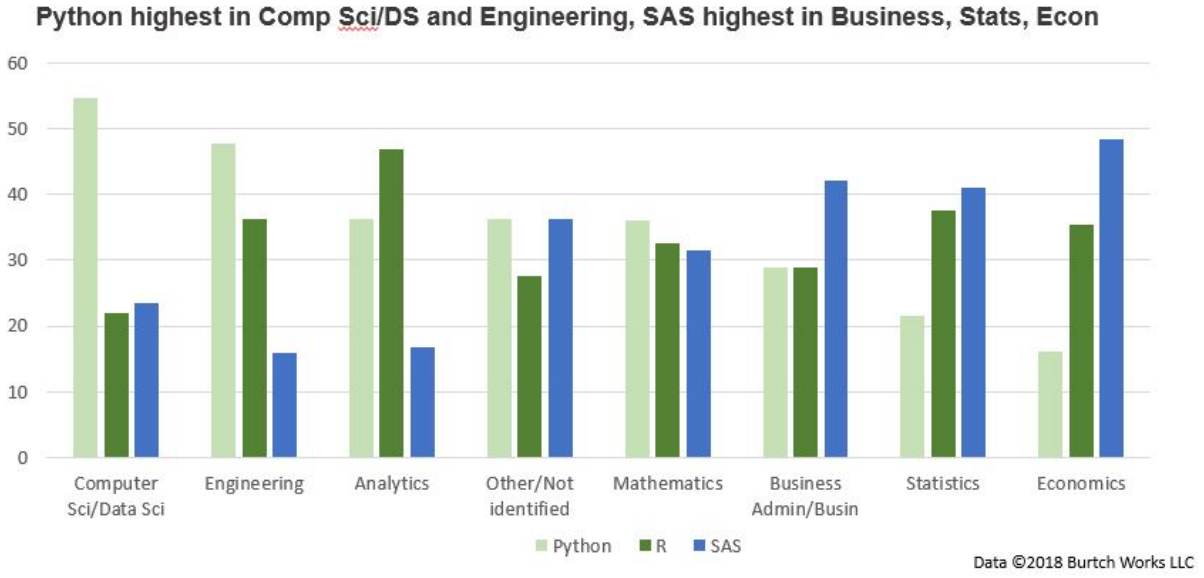
- Data Science is part science part craftmanship
- Accomplishments in Data Science occur via an interaction between the data scientist and the data
- The use of intermediaries in this relationship reduce the effectiveness of the process
- Depending on the size of the project, the direct relationship can be with the actual system or with a fully functional prototype, to be scale up to the final system
- Three main categories of tools:
  - Programming languages – Python and R in particular
  - UI-based tools – Knime, Rattle, Alterix, Rapidminer
  - Commercial ”statistical” tools – SAS, SPSS
- The majority of systems running in operational environments are based programming languages

Tools for Data Science





The comparisons are made using a specific commercial tool, SAS, but similar results would appear for other tools in the same category



Tools for Data Science

- Uses simple and intuitive GUI
- Easy node configuration and execution
- Open Source
- Many relevant examples
- Useful help – node description
- Good for beginners
- KNIME allows users to:
  - visually create data flows
  - selectively execute analysis steps
  - inspect results
- Integration of various Python, R, Perl, Java snippets
- Portability – PMML, XML

Using UI-based tools: demo

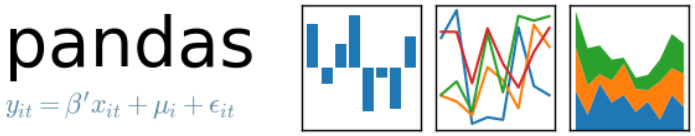
Python as a Data Science tool

- Python is a multi-paradigm, high-level, interpreted, programming language
- With more than 40,000 libraries adding specific functionalities, it can be optimized for a wide variety of domains
- High-level: easy to write, closer to human language than to machine language
- Interpreted: the program is executed directly by the interpreter, instead of being translated in machine language
- These characteristics granted it popularity in text-mining, data analysis, scientific simulations, web-scraping, and many other scripting tasks
- Its flexibility comes with the cost of a lower performance with respect to other languages (C, C++, Java). For this reason it cannot be used for applications such as high-frequency trading

Python as a Data Science tool

- Guido Van Rossum created the first version of Python in 1989
- The language is named after Monty Python. This reflects in the use of example variable and function names such as spam, eggs, bacon, and sausage

Python Libraries for Data Science



Pandas:

- adds data structures and tools designed to work with table-like data
- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

SciKit-Learn:

- provides data science/machine learning algorithms: classification, regression, clustering, model validation etc.
- built on NumPy, SciPy and matplotlib

Using Python: demo

SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

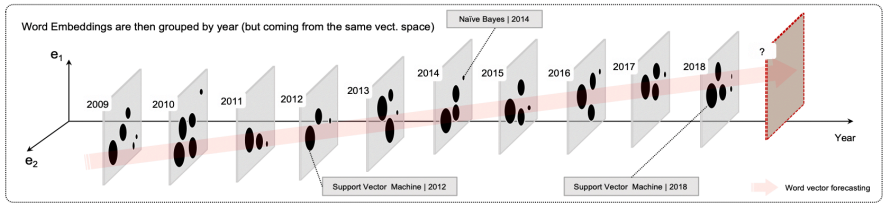
- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

Analytics @ Stevens/SSE/SERC

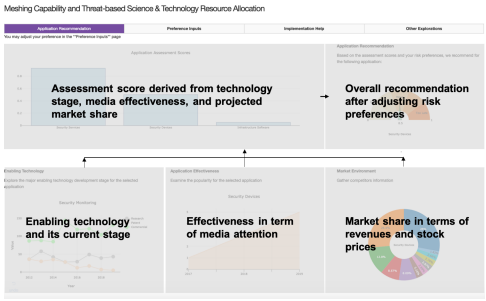
Dr. Carlo Lipizzi

NLPlab - We develop **Natural Language Processing** & **Machine Learning** solutions [https://nlplab.sercuarc.org]. ~25 people, 3 DoD-sponsored research projects

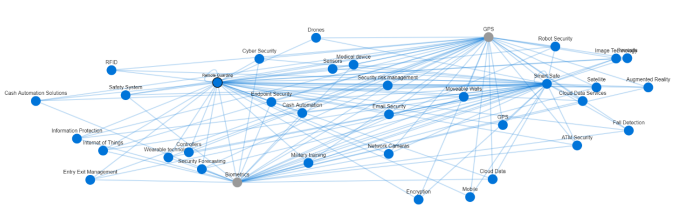
Predicting new technologies



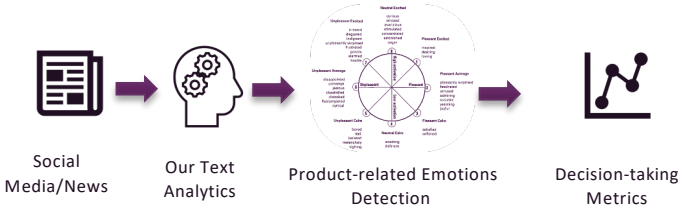
DSS - Extracting risk elements from text to take competitive decision



Technologies/application relationships – "kill chain"

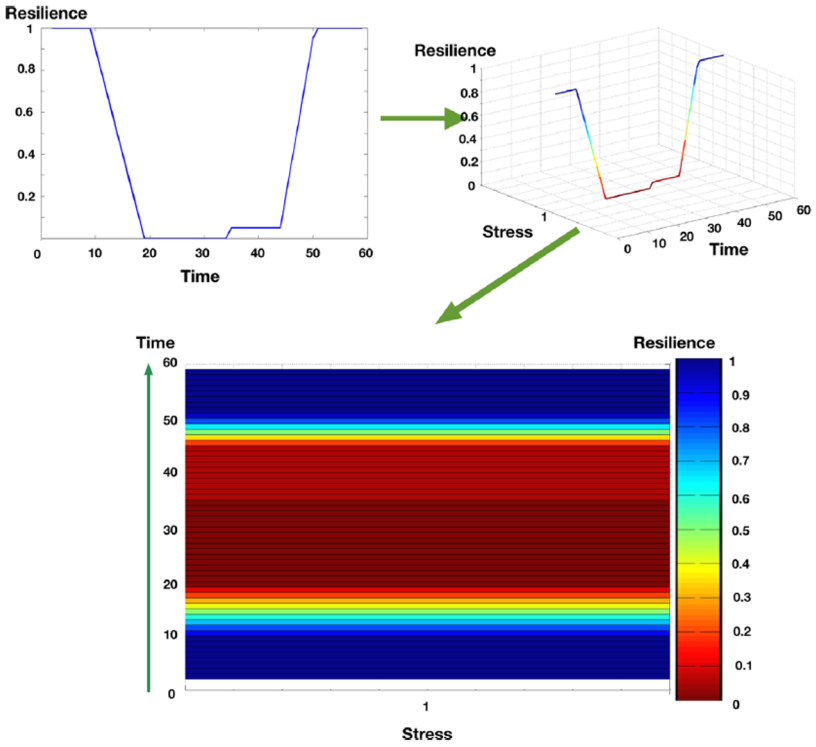


Detecting and measuring emotional reactions to products



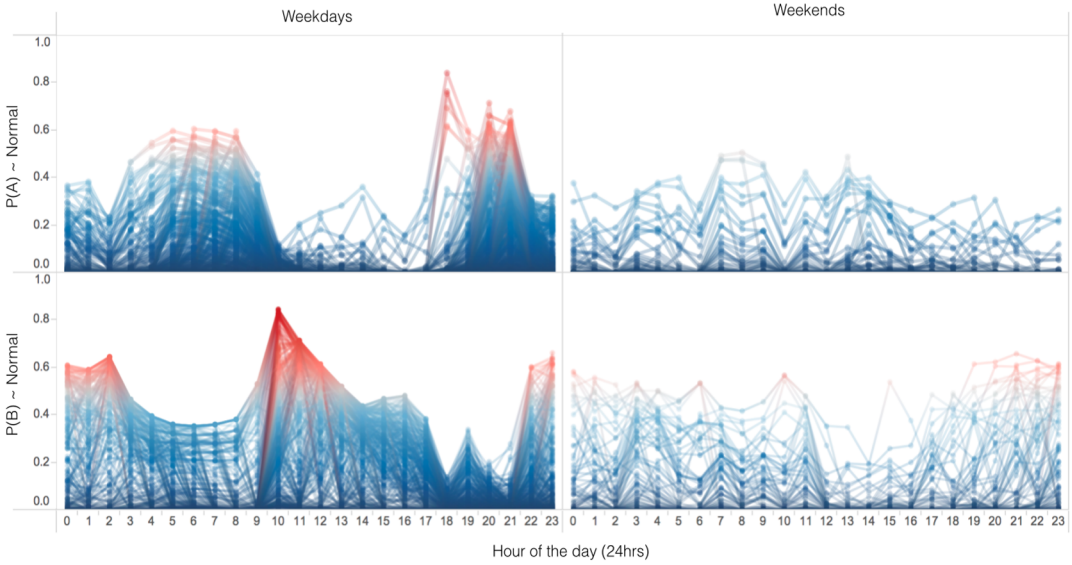
Analytics @ Stevens/SSE/SERC

Resilience Analytics for Real-Time Decision-Making and Disaster Preparedness

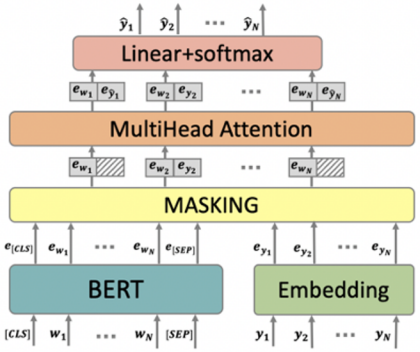


Urban Logistics and Multi-Objective Optimization to Drive Improvements in Policy-Making

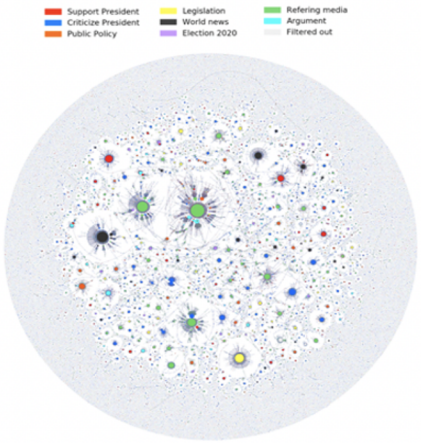
Dr. Jose Ramirez-Marquez



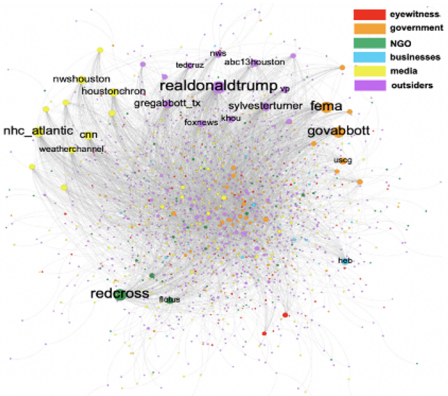
Analytics @ Stevens/SSE/SERC



Deep learning architecture design



conversation dynamics networks



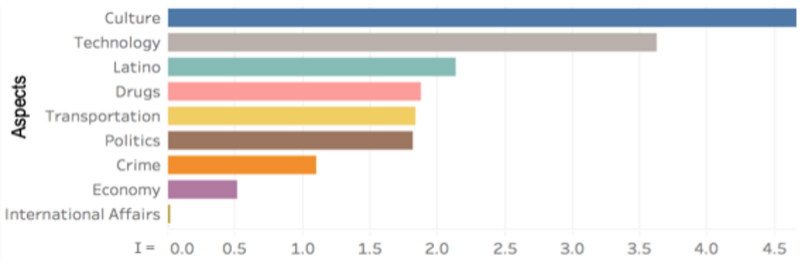
Disaster-motivated network analysis

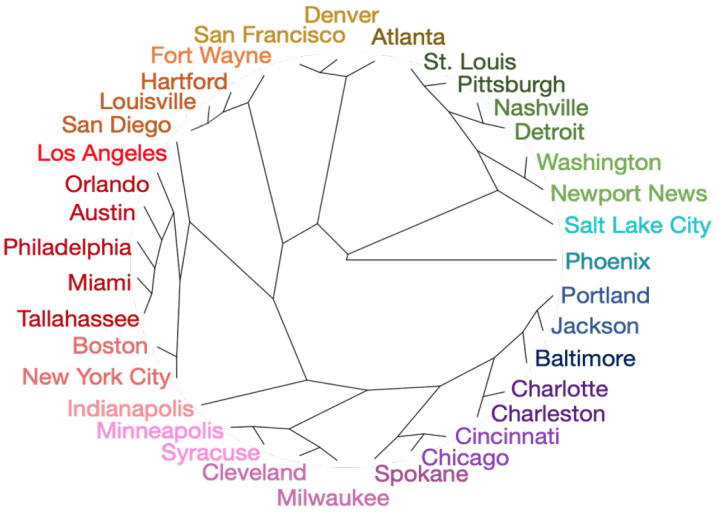
Multi-lingual AI

Pouria Babvey

Analytics @ Stevens/SSE/SERC

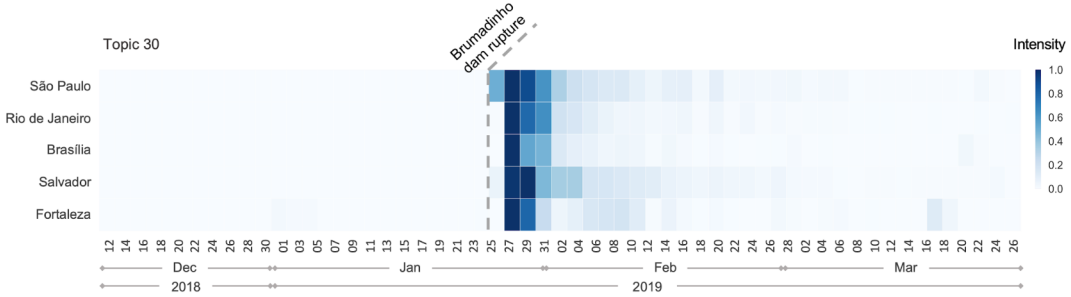
City Identity Profiles and Identity Similarity Using Online News Data

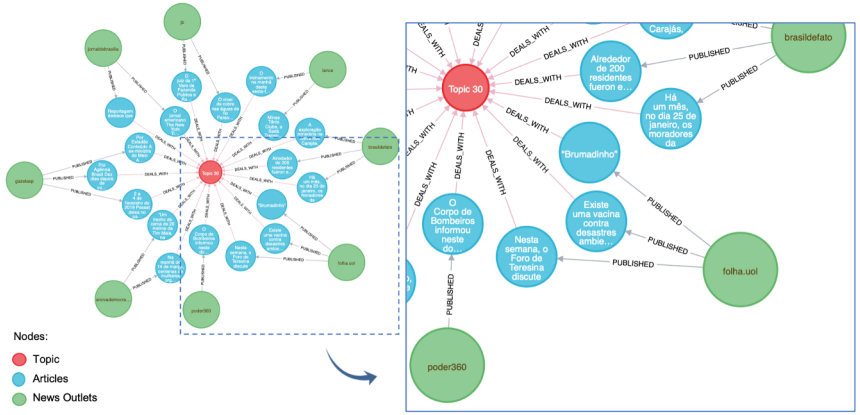




Fernanda Capela

Impact of Disruptive News Events Using Topic Resilience and Network Analysis





SYSTEMS ENGINEERING RESEARCH CENTER

Agenda

- The datatification
- Data changing Society
- Data changing Businesses
- Top data trends happening now
- The world of Natural Language
- Data science is hands-on: tools and methods
- Putting things together
- What's coming

Next Week - *Data for the upcoming world: Horizon scanning*

- Future cannot be predicted, but in science there is a high level of consistency over time. Data Science today is a steppingstone for an even more informed and complex way of living and doing business, with a continuous integration of sources and media, creating semantic synergies, pushing the boundaries of convenience, value and privacy.
- In this seminar, we scan the major trends in Data Science, starting from the current emerging trends, extrapolating scenarios and presenting live examples of emerging applications

SYSTEMS ENGINEERING RESEARCH CENTER

DAU

STEVENS INSTITUTE of TECHNOLOGY

THE INNOVATION UNIVERSITY®

Thank you!

School of Systems&Enterprises

Dr. Carlo Lipizzi

clipizzi@stevens.edu

NLP LAB