FOReSiGHT

SAPIENZA
UNIVERSITÀ DI ROMA

Data science

Outline

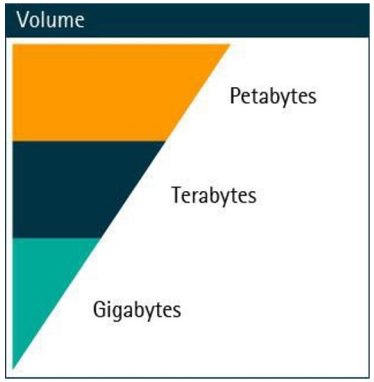
- Big Data & Challenges
- What is Data Science
- Data Science & Academia
- Data Science & Others
- Case Studies
- Essential points
- Conclusion

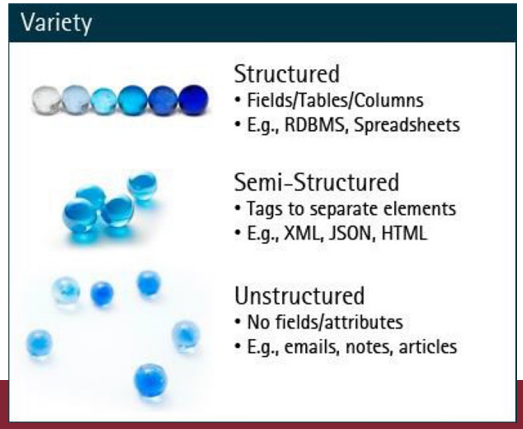
Data All Around

- **Lots of data is being collected and warehoused**
  - Scientific Experiments
  - Internet of Things
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network
  - ……many more!

Big Data

- Big Data are data sets so large or so complex that traditional methods of storing, accessing, and analyzing their breakdown are too expensive. However, there is a lot of potential value hidden in this data, so organizations are eager to harness it to drive innovation and competitive advantage.
- Big Data technologies and approaches are used to drive value out of data rich environments in ways that traditional analytics tools and methods cannot.





What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling
- Data Driven
  - Predictive Analytics
  - Deep Learning

https://youtu.be/X3paOmcrTiQ
https://youtu.be/pzo13OPXZS4

Big Data & Data Science

- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- Employment of data scientists is projected to grow 36 percent from 2021 to 2031, much faster than the average for all occupations
  https://www.bls.gov/ooh/math/data-scientists.htm
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"
  - Plans for Data Science Stream at AUST
  - RDA-CODATA School of Research Data Science

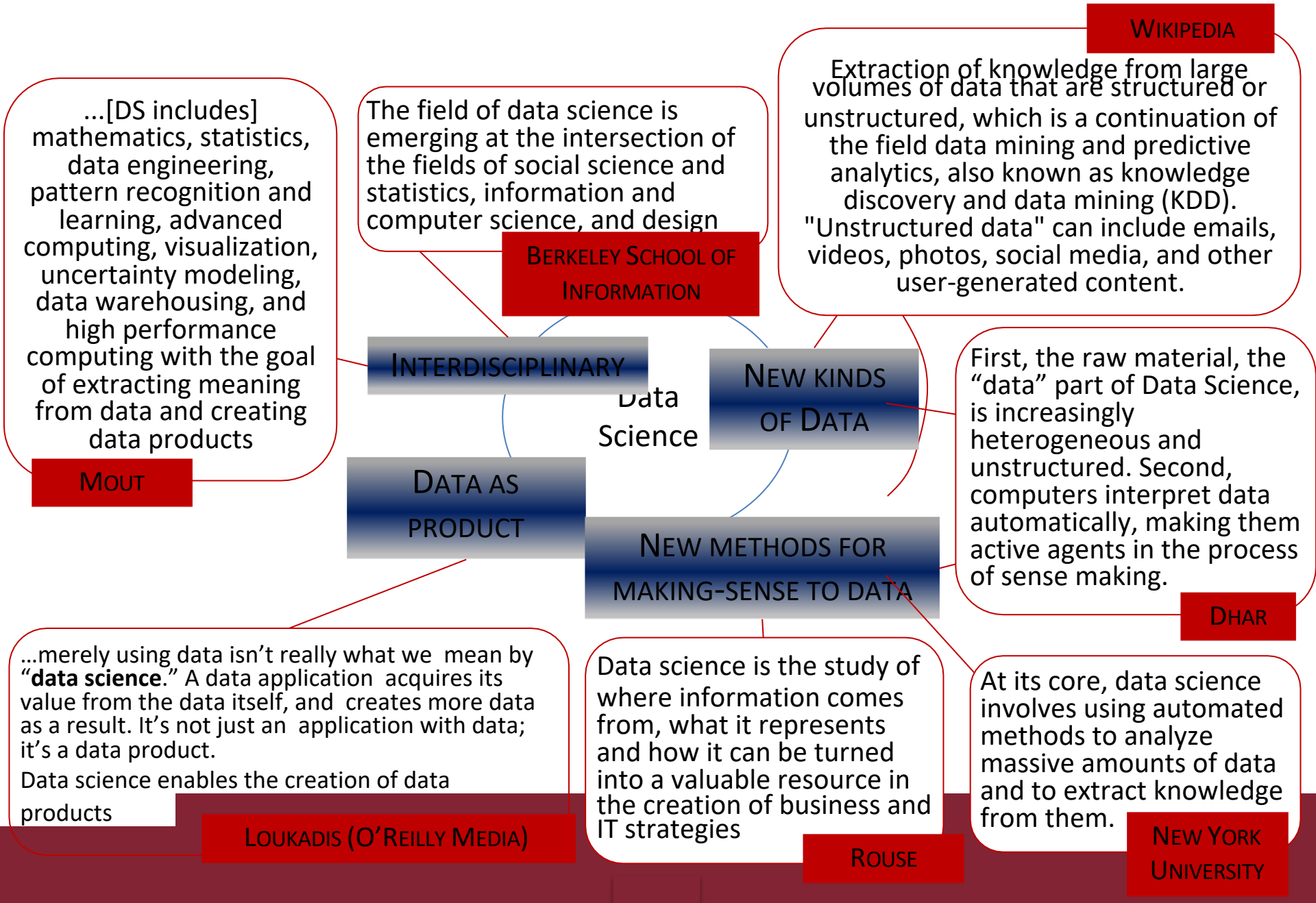
What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

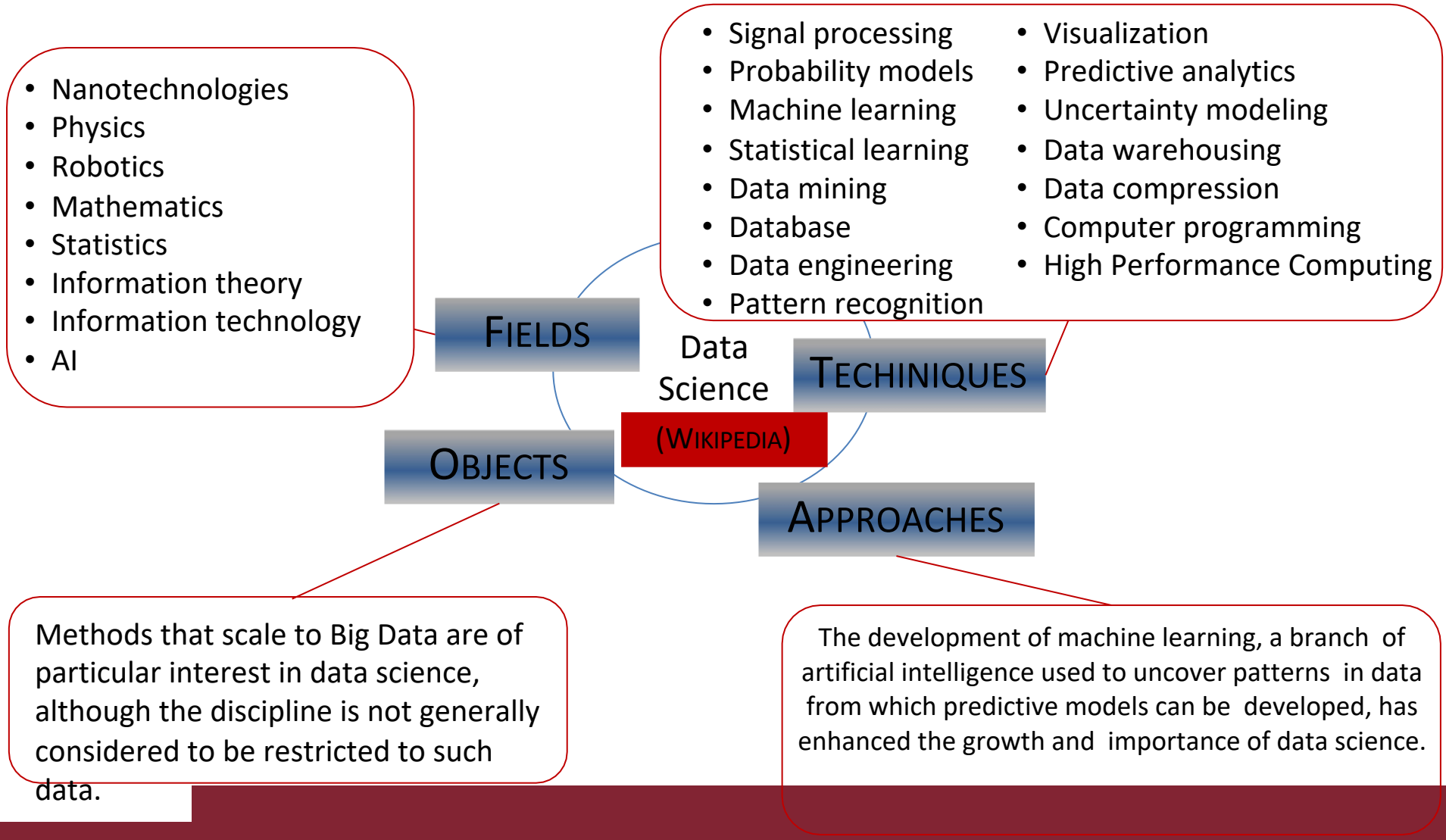
What is Data Science?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Definitions



Data Science landscape



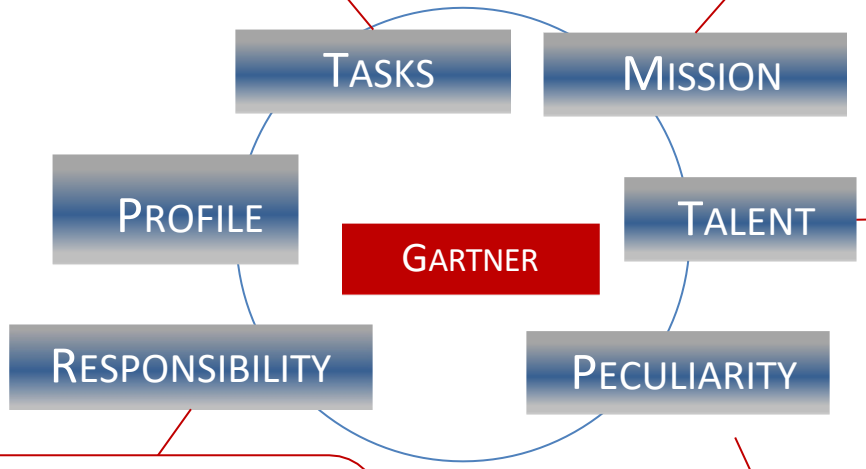
Who is a Data Scientist?

A data scientist may or may not have specialized industry knowledge to aid in modeling business problems and with understanding and preparing data.

In addition to advanced analytic skills, this individual is also proficient at integrating and preparing large, varied datasets, architecting specialized database and computing environments, and communicating results

The data scientist has emerged as a new role, distinct from — but

those of business intelligence (BI) analysts and statisticians



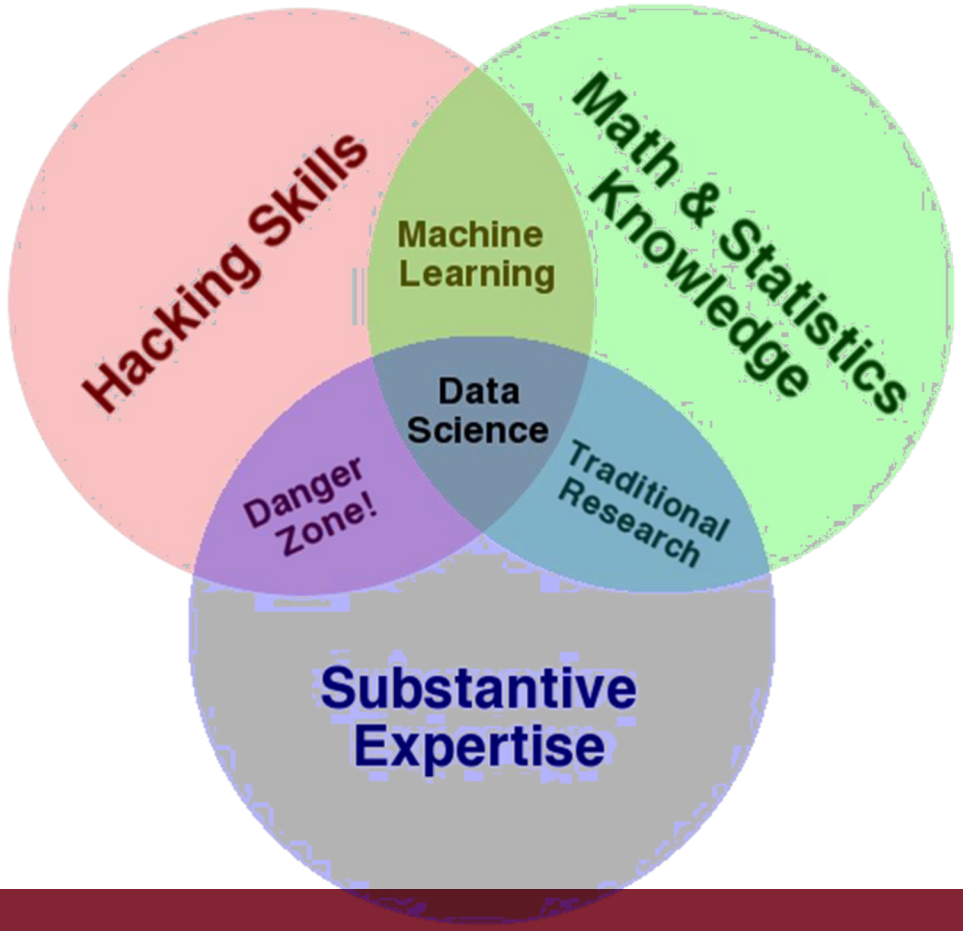
Creating value from data requires a range of talents: from data integration and preparation, to architecting specialized computing/database environments, to data mining and intelligent algorithms

An individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use o statistical, algorithmic, mining and visualization techniques

Data scientists can be invaluable in generating insights, especially from "big data;" but their unique combination of technical and business skills, together with their heightened demand, makes them difficult to find or cultivate.

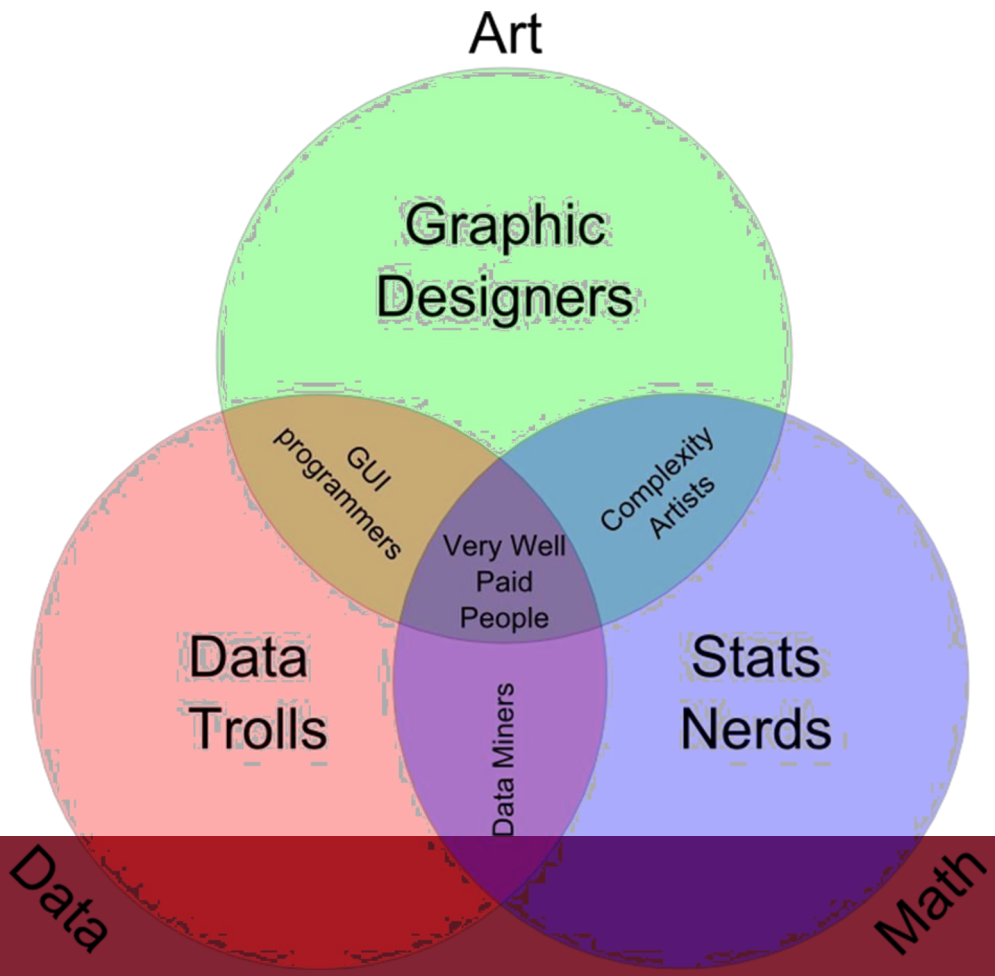
What is Data Science?

- Some definitions link computational, statistical, and substantive expertise.

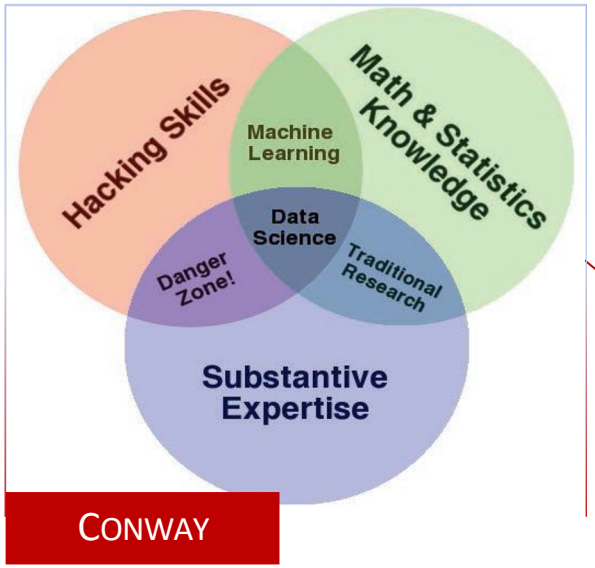


What is Data Science?

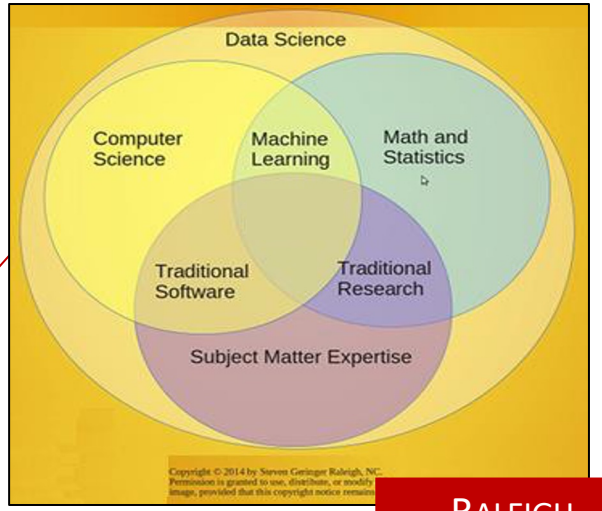
- Other definitions focus more on technical skills alone.



Unicorn



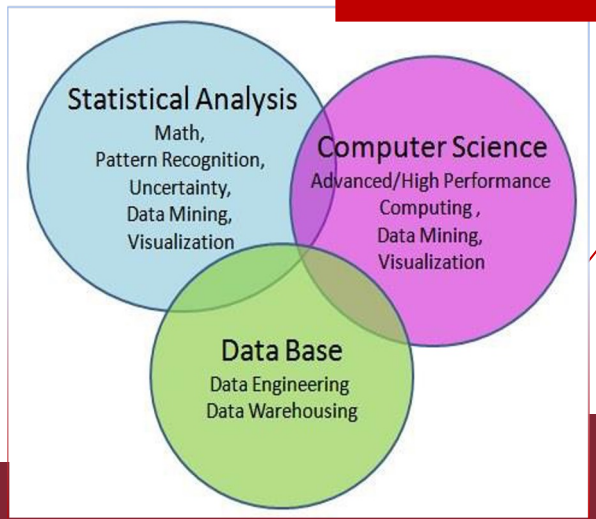
CONWAY



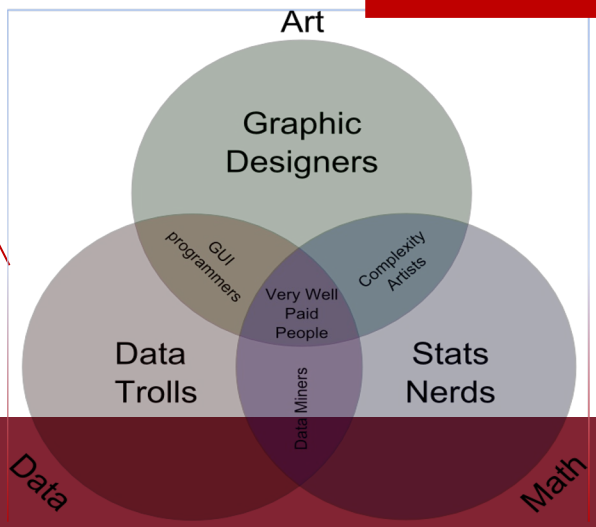
RALEIGH

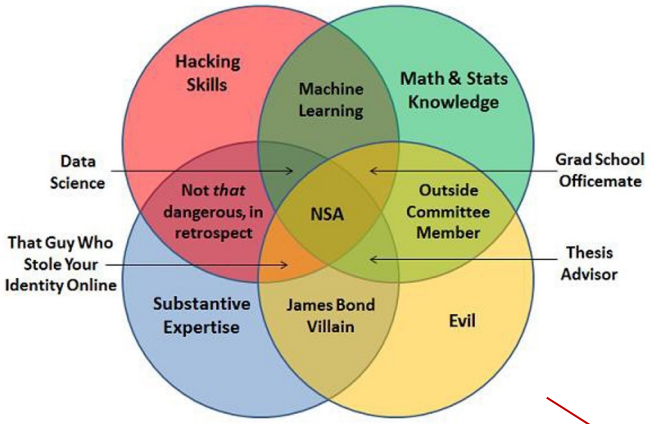
Venn diagram

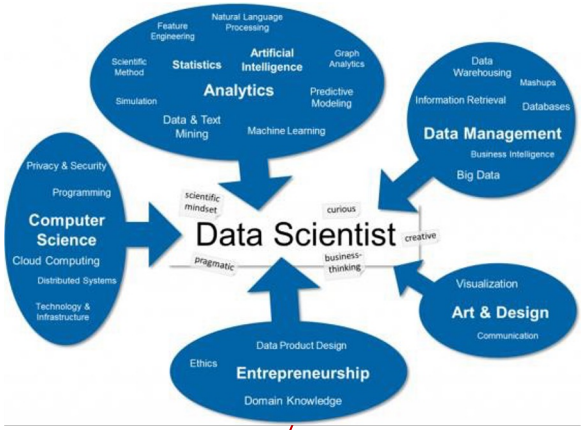
MOUT



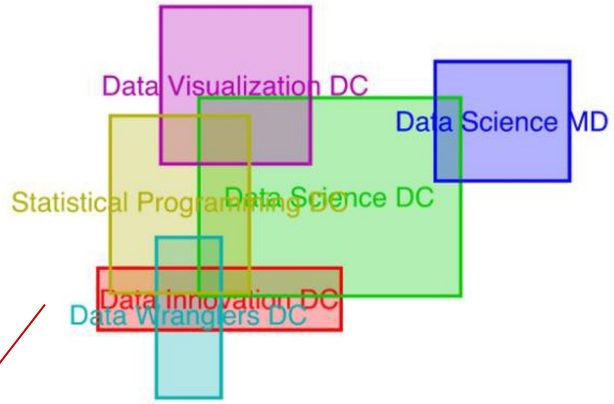
ERICKSON

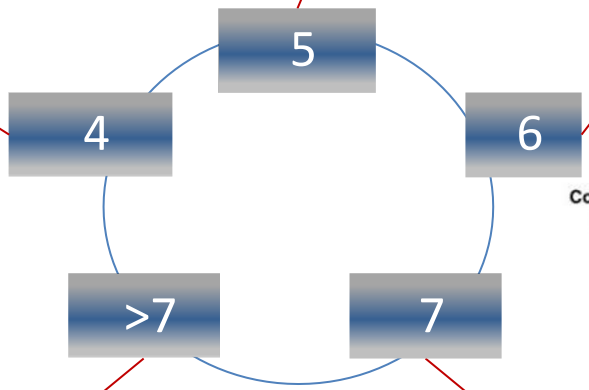




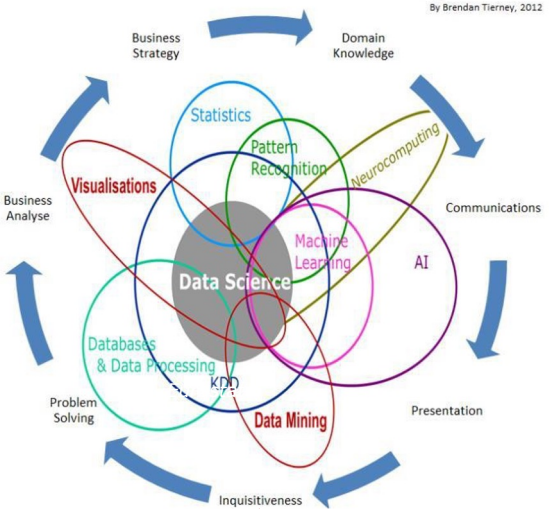


https://youtu.be/5Zg-C8AAIGg





Data Science Is Multidisciplinary



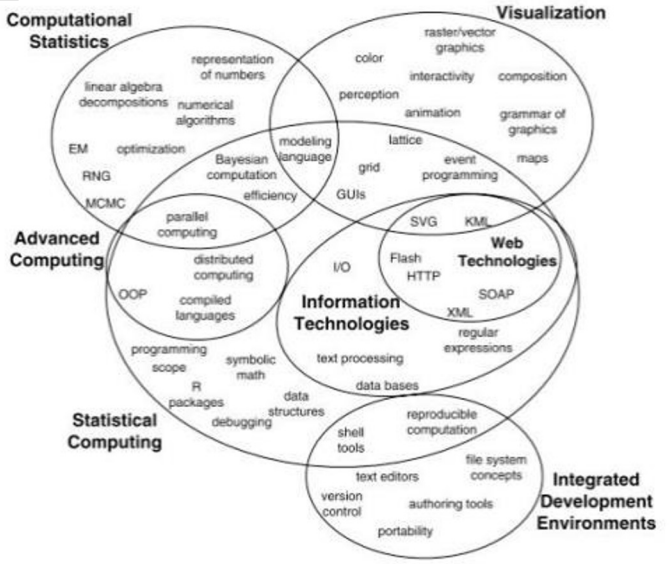
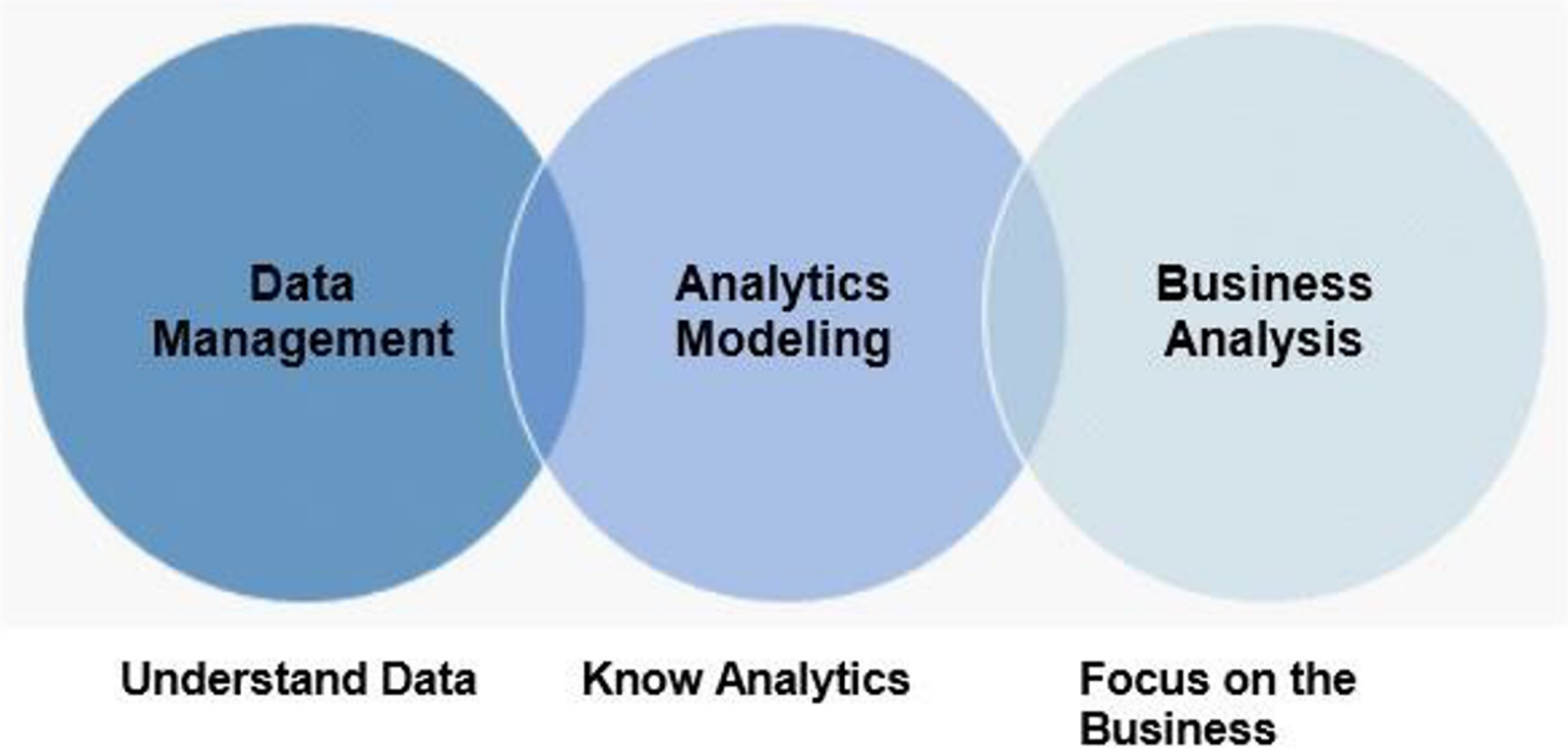


Figure 3. Core Data Scientist Skills



Source: Gartner (March 2012)

MODERN DATA SCIENTIST

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

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

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

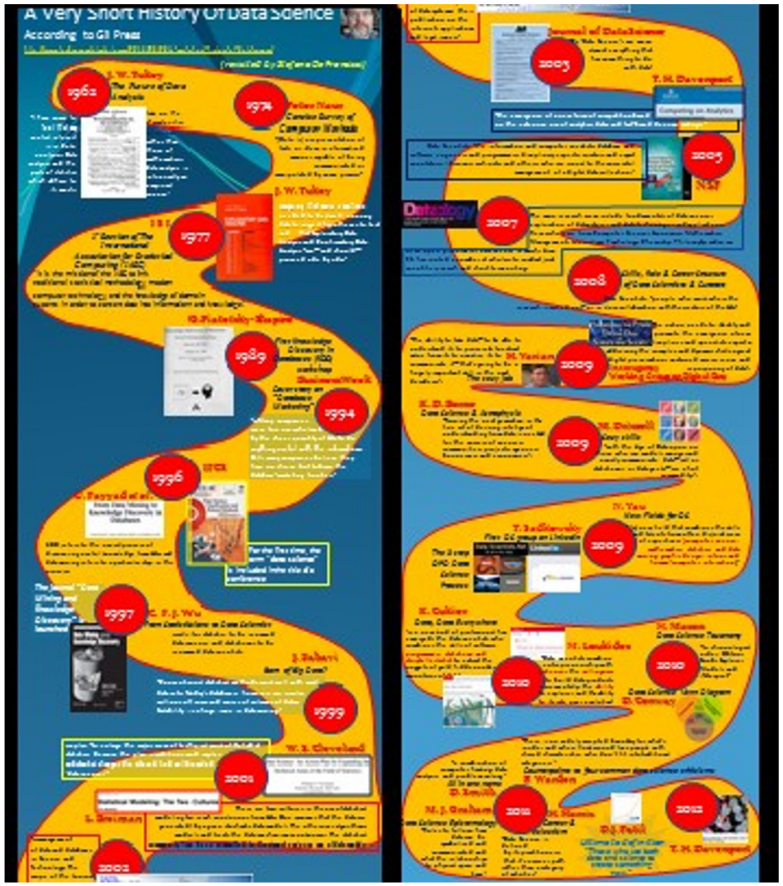
COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

Short History of Data Science (Loosely based on Gil Press version)



http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science

1962

J. W. Tukey

The Future of Data Analysis

"I have come to feel that my central interest is in *data analysis*... Data analysis, and the parts of statistics which adhere to it, must...


THE ANNALS
of
MATHEMATICAL
STATISTICS

Contents

take on the characteristics of science rather than those of mathematics... data analysis is intrinsically an empirical science"

1974

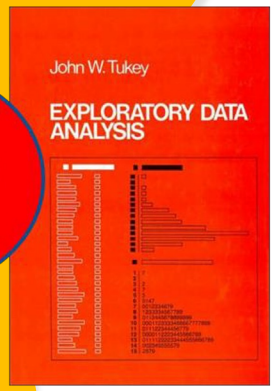
Peter Naur

Concise Survey of Computer Methods

"[Data is] **a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."**

J. W. Tukey

...arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis "can—and should—proceed side by side."



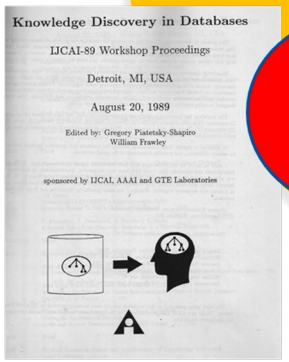
1977

I S I

1° Section of The International Association for Statistical Computing (IASC)

"It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

G.Piatetsky-Shapiro


Knowledge Discovery in Databases

IJCAI-89 Workshop Proceedings

Detroit, MI, USA

August 20, 1989

Edited by: Gregory Piatetsky-Shapiro
William Frawley

sponsored by IJCAI, AAAI and GTE Laboratories

1989

First Knowledge Discovery in Databases (KDD) workshop

BusinessWeek

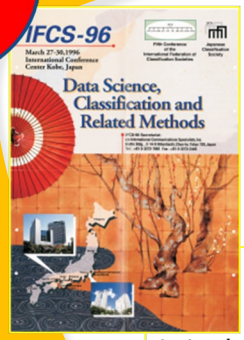
Cover story on "Database Marketing"

1994

"…Many companies were too overwhelmed by the sheer quantity of data to do anything useful with the information… Still, many companies believe they have no choice but to brave the database-marketing frontier."

1996

IFCS


IFCS-96

March 27-30,1996
International Conference
Center Kobe, Japan

Data Science, Classification and Related Methods

U. Fayyad *et al.*

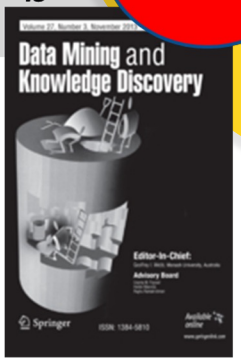
From Data Mining to Knowledge Discovery in Databases

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

For the first time, the term "**data science**" is included in the title of a conference

The journal "Data Mining and Knowledge Discovery" is launched



1997

C. F. J. Wu

From Statisticians to Data Scientist

...calls for statistics to be renamed data science and statisticians to be renamed data scientists

J. Zahavi

Born of Big Data?

"Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data... Scalability is a huge issue in data mining."

1999

...a plan "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'"

W. S. Cleveland

2001

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

L. Breiman

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.

"…management of data and databases in Science and Technology. The scope of the Journal includes descriptions of data systems, their publication on the internet, applications and legal issues."

2002

DATA SCIENCE Journal — CODATA

2003

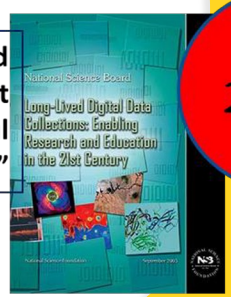
Journal of Data Science

"By "Data Science", we mean almost everything that has something to do with data"

T. H. Davenport

"the emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making… "

Data Scientists: "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."



2005

NSF

Dataology
上海市数据科学重点实验室(复旦大学)
Shanghai Key Laboratory of Data Science, Fudan University

2007

The main research areas include fundamentals of data science, exploration of datanature, and data technologies and applications. Researchers are from Computer Science, Economics, Mathematics, Management, Journalism, Psychology, Chemistry, Philosophy, and so on.

As an open platform for data science research, Area 96 has invited a number of scholars to conduct joint scientific research and short term visiting.

2008

Skills, Role & Career Structure of Data Scientists & Curators

Data Scientists: "people who work where the research is carried out–or in close collaboration with the creators of the data"

"The nation needs to identify and promote the emergence of new disciplines and specialists expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data".

Interagency Working Group on Digital Data

"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…".

H. Varian

The sexy job

2009

K. D. Borne

Data Science & Astrophysic

"Training the next generation in the fine art of deriving intelligent understanding from data is needed for the success of sciences, communities, projects, agencies, businesses, and economies. ".

2009

M. Driscoll

Sexy skills

"with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity.".

N. Yau

New Fields for DS

" [a] new field that combines the skills and talents from often disjoint areas of expertise… [computer science; mathematics, statistics, and data mining; graphic design; infovis and human-computer interaction]"

2009

T. Sadkowsky

First DS group on LinkedIn

The 3 step OPD Data Science Process

K. Cukier

Data, Data Everywhere

"… a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data"

2010

M. Loukides

"Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution"

H. Mason

Data Science Taxonomy

2010

"In chronological order: Obtain, Scrub, Explore, Model, and iNterpret"

Data Science Venn Diagram

D. Conway

"a combination of computer hacking, data analysis, and problem solving "

All in one name

D. Smith

M. J. Graham

Data Science Epistemology

"Rules to follow. how data can be symbolized and communicated and what the relationships to physical space and time"

2011

"There is no widely accepted boundary for what's inside. and when I look around I see people with shared characteristics who don't fit into traditional categories. "

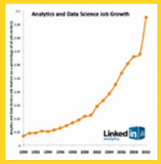
Counterpoints to four common data science criticisms

P. Warden

H. Harris

Carreer & eclectism

"Data Science is defined by its practitioners, that it's a career path rather than a category of activities"



D.J. Patil

Ultimate definition

"Those who use both data and science to create something new."

2012

Data Scientist: The Sexiest Job of the 21st Century

T. H. Davenport

Steps to a Metaphisics of Data Science

- How does the Data Science in the context of the Knowledge Organization?
- What are its relations with other fields of scientific knowledge?
- Can DS be explained as part of the philosophy of science?

	Data	Information	Knowledge
Scientific context	Data Science	Information Science	Knowledge Science
Philosophical context	Philosophy of Data	Philosophy of Information	Philosophy of Knowledge (Epistemology, Gnoseology)

https://podcasts.ox.ac.uk/luciano-floridi-ethics-open-data

Is Data Science a maturity science?

Types of domain dealt by an intellectual enterprises:

(a) topics (facts, data, problems, phenomena, observations, and the like)

(b) methods (techniques, approaches, and so on)

(c) theories (hypotheses, explanations, and so forth)

Feature of a new discipline:

(a) To represent an autonomous field (*unique topics*)

(b) To provide an innovative approach to both traditional and new philosophical topics (*original methodologies*);

(c) To stand beside other disciplines, offering the systematic treatment of its own conceptual foundations (*new theories*).

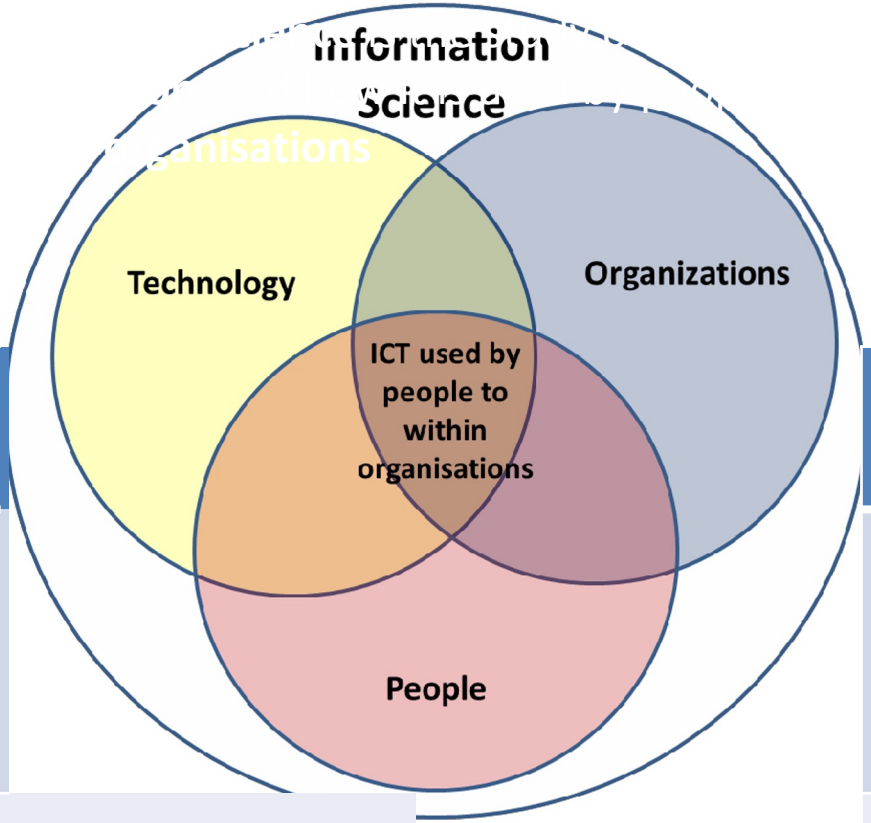
If a discipline attempts to innovate in more than one of these domains simultaneously is premature, as detaches itself too abruptly from the normal and continuous thread of evolution of its general field (Stent 1972).

As everyone's concern is nobody's business → crossroad of
- technical matters
- theoretical issues
- applied problems
- conceptual analyses

→ to be anyone's own area of specialisation → **Transdisciplinary** (like cybernetics or semiotics) or **interdisciplinary** (like biochemistry or cognitive science)?

L. Floridi

Beyond Data Science?



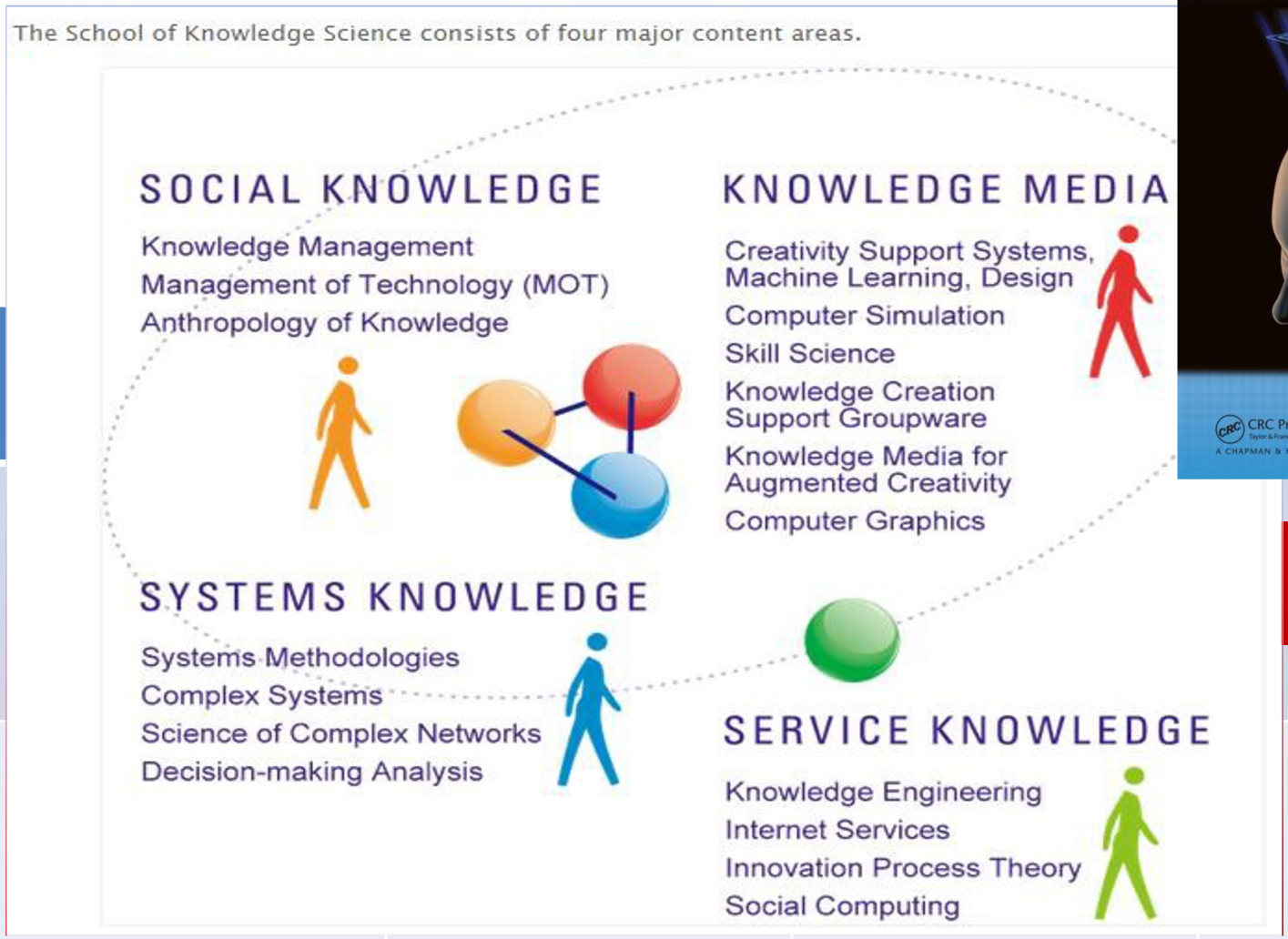
Information Science sits at the intersection of technology, people, and organizations.

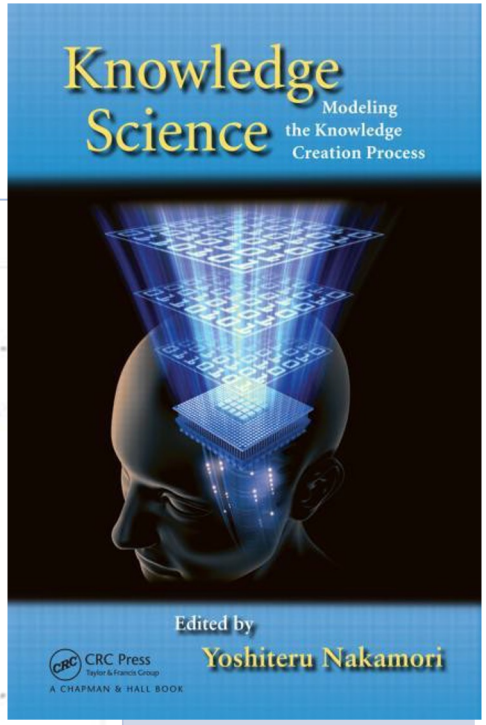
It is a distinct discipline and has a focus on Information and Communication Technologies (ICT) used by people to manage information within organisations.

Information	Knowledge
Information Science	

http://infosci.otago.ac.nz/what-is-information-science/

Beyond Data Science?





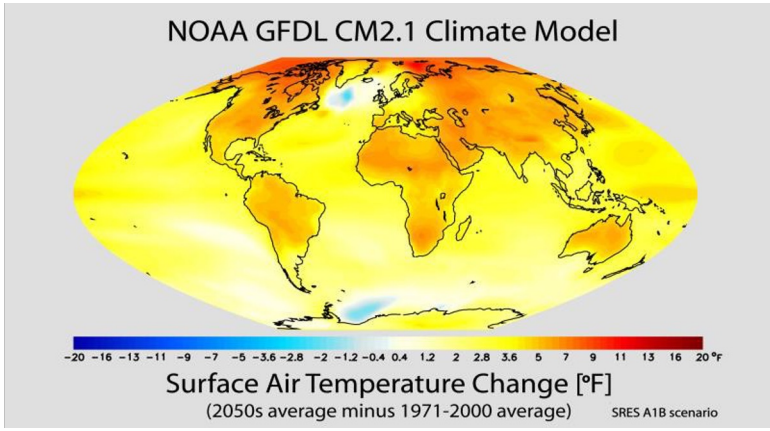
Knowledge Science

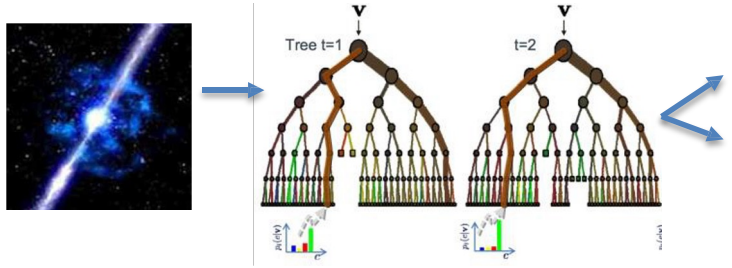
https://www.crcpress.com/Knowledge-Science-Modeling-the-Knowledge-Creation-Process/Nakamori/9781439838365

Data Science Vs Analysis Vs Software Delivery

Component	Traditional Analysis	Traditional Software Delivery	Data Science
Tools	SAS, R, Excel, SQL, in-house tools	Java, source control, Linux, continuous integration, unit testing, bug reports and project management	R, Java, scientific Python libraries, Excel, SQL, Hadoop, Hive, Pig, Mahout and other machine learning libraries, github for source control and issue management
Analytical Methods	Regressions, classifications, measuring prediction accuracy and coverage/error, sampling	N/A	Classification, clustering, similarity detection, recommenders, unsupervised and supervised learning, small- and large-scale computations, measuring prediction accuracy and coverage/error
Team Structure	Statisticians, Mathematicians, Scientists	Developers, Project Managers, Systems Engineers	Mathematicians, Statisticians, Scientists, Developers, Systems Engineers
Time Frame	Either: • Usually on-going research and discovery within a team in the organization Or: • Specific project to determine answers	Regular software release cycle, continuous delivery, etc.	Either: • Discovery/learning phase leading to product development Or: • On-going research and product invention/improvement

Contrast: Scientific Computing

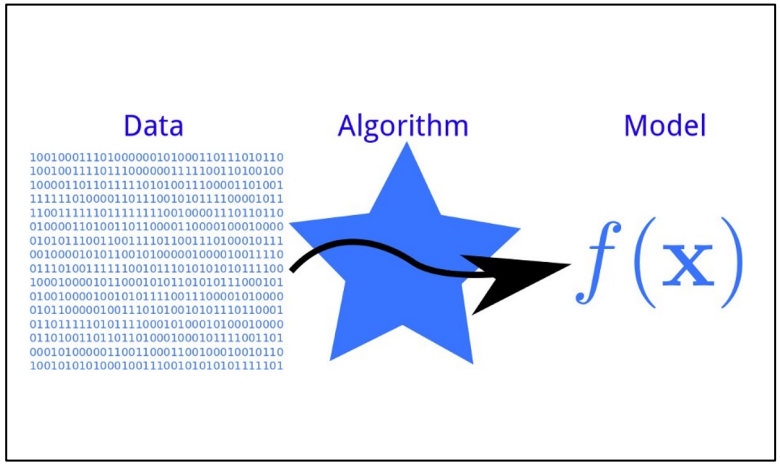


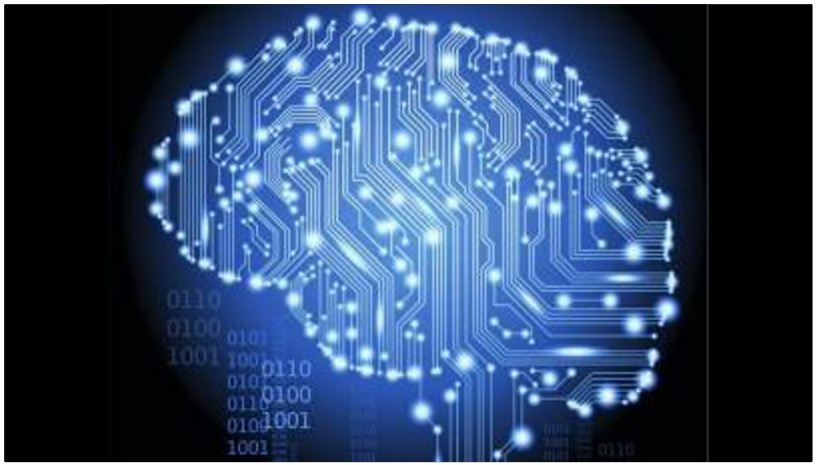


Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach
General inference engine replaces model
Structure not related to problem
Statistical models handle true randomness, and **un-modeled complexity**.
Run on cheaper computer Clusters (EC2)

Contrast: Machine Learning





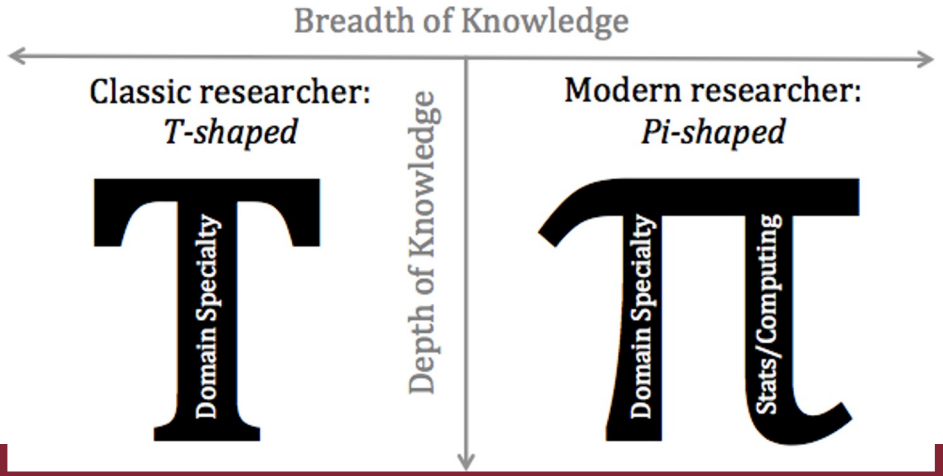
Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper ☺	Take action!

Contrast: Data Engineering

	Data Science	Data Engineering
Approach	Scientific (Exploration)	Engineering (Development)
Problems	Unbounded	Bounded
Path to Solution	Iterative, exploratory, nonlinear	Mostly linear
Education	More is better (PhD's common)	BS and/or self-trained
Presentation Skills	Important	Not as important
Research Experience	Important	Not as important
Programming Skills	Not as important	Important
Data Skills	Important	Important

Data Science & Academia

- In the words of Alex Szalay, these sorts of researchers must be "Pi-shaped" as opposed to the more traditional "T-shaped" researcher. In Szalay's view, a classic PhD program generates T-shaped researchers: scientists with wide-but-shallow general knowledge, but deep skill and expertise in one particular area. The new breed of scientific researchers, the data scientists, must be Pi-shaped: that is, they maintain the same wide breadth, but push deeper both in their own subject area and in the statistical or computational methods that help drive modern research:



https://youtu.be/FsSrzmRawUg

Data Science & Academia

- In a post by Jake Vanderplas in 2014 related to SciFoo discussion on: ***Academia and Data Science***, the following questions below were discussed.
- I encourage you to develop your own thoughts on them and come up with your assessment

  - Where does Data Science fit within the current structure of the university & research institutions?
  - What is it that academic data scientists want from their career? How can academia offer that?
  - What drivers might shift academia toward recognizing & rewarding data scientists in domain fields?
  - Recognizing that graduates will go on to work in both academia and industry, how do we best prepare them for success in both worlds?

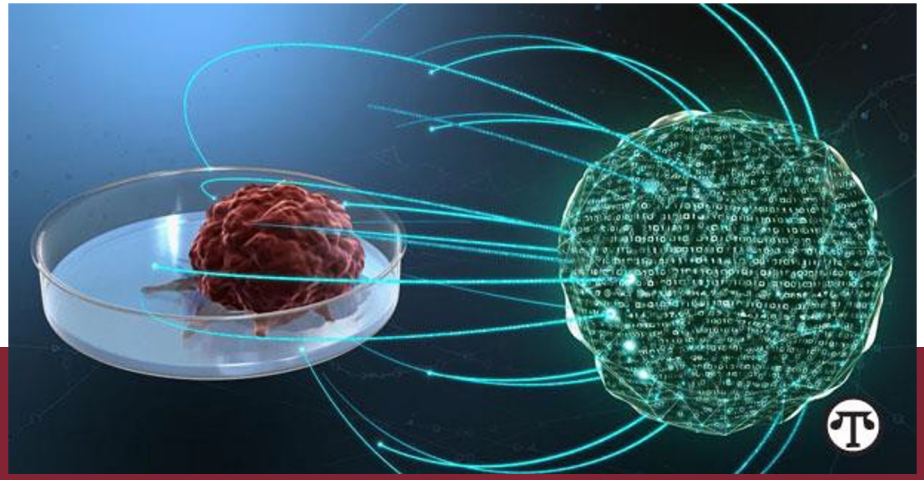
Data Science Applications

	Business	Health Care	Urban Leaving
Summary	From car design to insurance to pizza delivery, businesses are using data science to optimize their operations and better meet their customers' expectations.	Tomorrow's healthcare may look more efficient thanks to things like electronic health records. It also may look a lot more effective. Reduced readmissions, better care, and earlier detection are on the horizon.	For the first time in human history, more people live in cities than in suburban or rural areas. An emerging field called "urban informatics" combines data science with the unique challenges facing the world's growing cities
What is happening?	Two-Way Street for the Ford Focus Electric Car	Reducing Hospital Readmissions	Taking on Megacity Traffic
	Better Fraud Detection Boosts Customer Satisfaction	Better Point-of-Care Decisions	Fighting Crime with Data "predictive policing"
	E-Commerce Insights: Domino's Secret Sauce		
What is possible	Using Social Data to Select Successful Retail Locations .	Medical Exams by Bathroom Mirrors	Instrumenting cities

Contrast: Computational Sciences

Data Science: Case Study Cancer Research

- Cancer is an incredibly complex disease; a single tumor can have more than **100 billion cells**, and each cell can acquire mutations individually. The disease is always changing, evolving, and adapting.
- Employ the power of big data analytics and high-performance computing.
- Leverage sophisticated pattern and machine learning algorithms to identify patterns that are potentially linked to cancer
- Huge amount of data processing and recognition



Data Science: Case Study Health Care

Stanford MEDICINE | News Center

FRONT PAGE ALL NEWS TOPICS MULTIMEDIA

Email Share 73 Tweet

Search News... Go

Stanford Medicine, Google team up to harness power of data science for health care

Stanford Medicine will use the power, security and scale of Google Cloud Platform to support precision health and more efficient patient care.

AUG 8 2016

Stanford Medicine and Google are working together to transform patient care and medical research through data science.

The new collaboration combines Stanford Medicine's excellence in health-care research and clinical work with Google's expertise in cloud technology and data science. Stanford's forthcoming Clinical Genomics Service, which puts genomic sequencing into the hands of clinicians to help diagnose disease, will be built using Google Genomics, a service that applies the same technologies that power Google Search and Maps to securely store, process, explore and share genomic data sets.

Lloyd Minor, dean of the School of Medicine, says the collaboration with Google marks a "milestone for the future of patient care and research."
Glenn Matsumura

Stanford Medicine includes the Stanford School of Medicine, Stanford Health Care and Stanford Children's Health. Together, Stanford Medicine and Google will build cloud-based applications for exploring massive health-care data sets, a move that could transform patient care and medical research.

"Stanford Medicine and Google are committing to major investments in preventing and curing diseases that afflict ordinary people worldwide. We're proud to be setting this milestone for the future of patient care and research," said Lloyd Minor, MD, dean of the School of Medicine.

The agreement — considered key to Stanford Health Care's development of the Clinical Genomics Service — makes Google Inc. a formal business associate of Stanford Medicine. As such, Google and Stanford will both comply with the Health Insurance Portability and Accountability Act, a federal law that regulates the privacy and security of medical information. HIPAA requires that Stanford Medicine patient data stored on Google Cloud Platform servers stay private. Patient information will be encrypted, both in transit and on servers, and kept on servers in the United States.

LEADING IN PRECISION HEALTH

Stanford Medicine is leading the biomedical revolution in precision health, defining and developing the next generation of care that is proactive, predictive and precise.

Learn more →

A LEGACY OF INNOVATION

Stanford Medicine's unrivaled atmosphere of breakthrough thinking and interdisciplinary collaboration has fueled a long history of achievements.

View timeline →

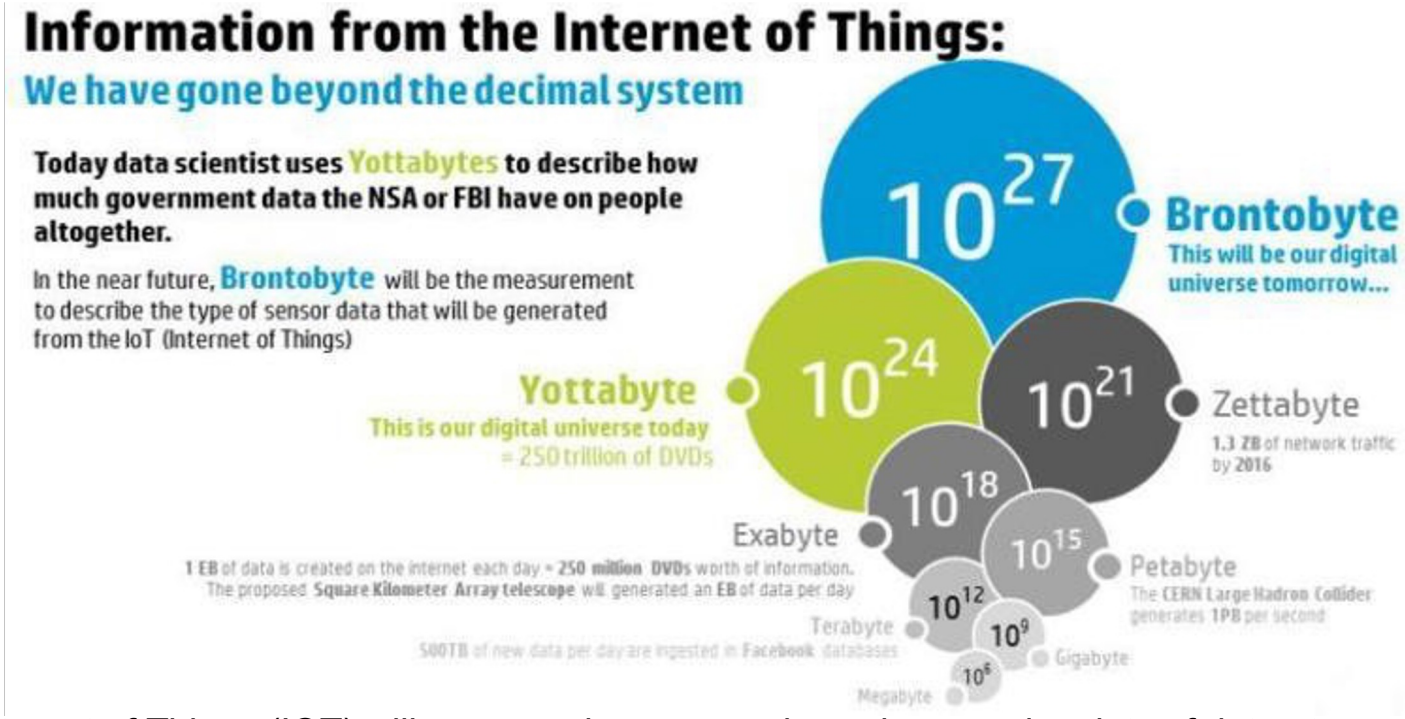
http://med.stanford.edu/news/all-news/2016/08/stanford-medicine-google-team-up-to-harness-power-of-data-science.html

Data Science: Case Study Elections

- The Obama campaigns in 2008 and 2012 are credited for their successful use of social media and data mining.
- Micro-targeting in 2012
  - http://www.theatlantic.com/politics/archive/2012/04/the-creepiness-factor-how-obama-and-romney-are-getting-to-know-you/255499/
  - http://www.mediabizbloggers.com/group-m/How-Data-and-Micro-Targeting-Won-the-2012-Election-for-Obama---Antony-Young-Mindshare-North-America.html
- Micro-profiles built from multiple sources accessed by aps, real-time updating data based on door-to-door visits, focused media buys, e-mails and Facebook messages highly targeted.
- 1 million people installed the Obama Facebook app that gave access to info on “friends”.

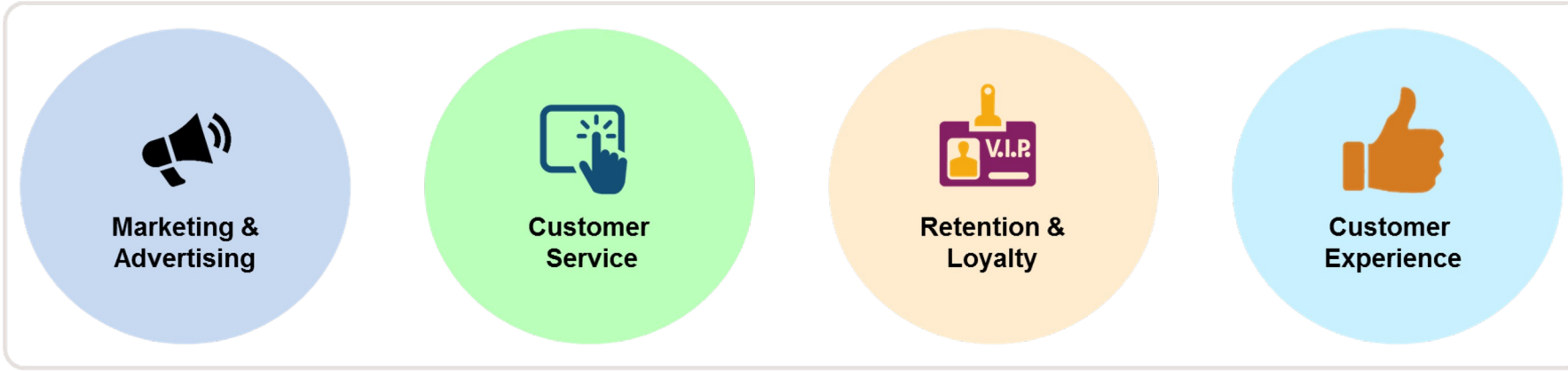
Data Science: Case Study Internet of Things (IoT)

- The Internet of Things is rapidly growing. It is predicted that more than 25 billion devices will be connected by 2020.



- The Internet of Things (IOT) will soon produce a massive volume and variety of data at unprecedented velocity. If "Big Data" is the product of the IOT, "Data Science" is it's soul.

Data Science: Case Study
Customer Analytics



Leveraging customer data to move ever closer to the elusive goal of truly personalized marketing: the right offer, at the right time, in the right location and context, to the right person.

By capturing and analyzing the data from customer touch points within an organization, companies can identify customer pain points and issues proactively and update their customer service FAQs or other communications with existing customers.

Using customer data and analytics, these companies deploy and refine predictive models that help them retain customers with proactive approaches. Investments, in terms of offers and upgrades, can be made at the right time to increase the likelihood of retaining desirable customers.

The experience that customers have with companies matters a great deal. Other recent research has highlighted the critical connection between experience and company financial performance.

Essential Points

- **Big Data has given rise to Data Science**
- **Data science is rooted in solid foundations of mathematics and statistics, computer science, and domain knowledge**
- **Sexy profession – Data Scientists ☺**
- **Not every thing with data or science is Data Science!**
- **The use cases for Data Science are compelling**

Conclusion

- In this section you have learned
  - What Big Data Challenges are
  - What exactly is Data Science and what do Data
  - Scientists do
  - Data Science contrasted with other disciplines
  - Case Study & Use Cases

Questions?


Thank
You!