European Commission

Data Science and its role in Big Data analytics

THE CONTRACTOR IS ACTING UNDER A FRAMEWORK CONTRACT CONCLUDED WITH THE COMMISSION

Eurostat

Outline

1. Data Science, basic concepts
2. A short history
3. A new concept of Science?
4. Big Data as the new frontier of Data Science
5. Data, information, knowledge

European Commission

Data Science

Interdisciplinary

...[DS includes] mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products

Mout

The field of data science is emerging at the intersection of the fields of social science and statistics, information and computer science, and design

Berkeley School of Information

New kinds of Data

Extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD). "Unstructured data" can include emails, videos, photos, social media, and other user-generated content.

Wikipedia

First, the raw material, the "data" part of Data Science, is increasingly heterogeneous and unstructured. Second, computers interpret data automatically, making them active agents in the process of sense making.

Dhar

Data as product

...merely using data isn't really what we mean by "**data science**." A data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product. Data science enables the creation of data products

Loukadis (O'Reilly Media)

New methods for making-sense to data

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies

Rouse

At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.

New York University

Data Science landscape

European Commission

- Nanotechnologies
- Physics
- Robotics
- Mathematics
- Statistics
- Information theory
- Information technology
- AI

FIELDS

- Signal processing
- Probability models
- Machine learning
- Statistical learning
- Data mining
- Database
- Data engineering
- Pattern recognition
- Visualization
- Predictive analytics
- Uncertainty modeling
- Data warehousing
- Data compression
- Computer programming
- High Performance Computing

TECHINIQUES

Data Science

(WIKIPEDIA)

OBJECTS

APPROACHES

Methods that scale to Big Data are of particular interest in data science, although the discipline is not generally considered to be restricted to such data.

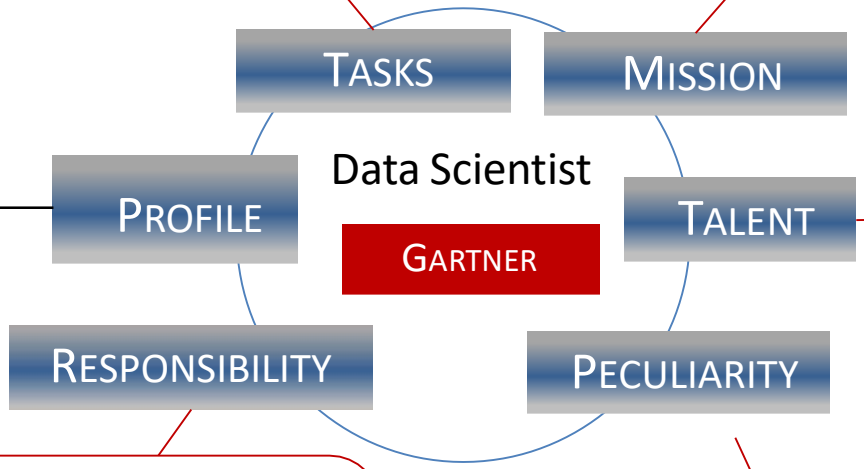
The development of machine learning, a branch of artificial intelligence used to uncover patterns in data from which predictive models can be developed, has enhanced the growth and importance of data science.

Who is a Data Scientist?

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.

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

The data scientist has emerged as a new role, distinct from — but ~~with similarities to —~~ 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

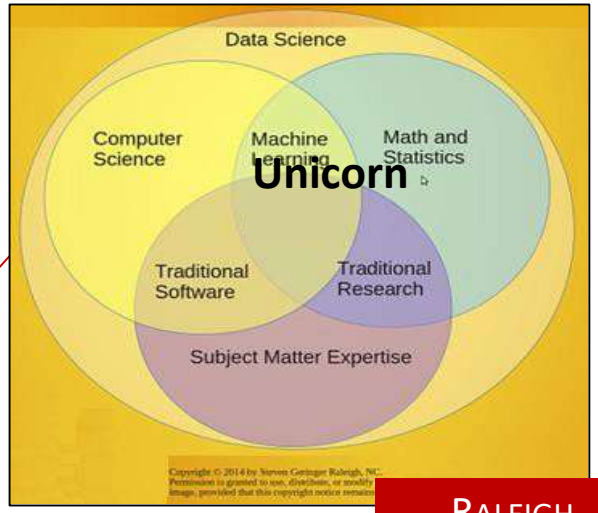
An individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining and visualization techniques.

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.

D. Laney, L. Kart

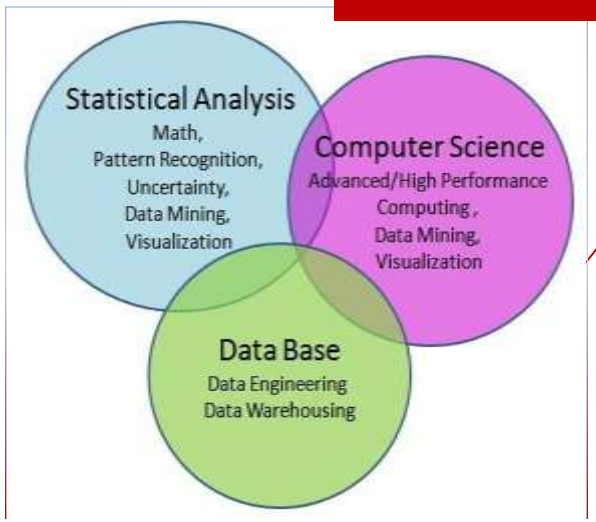
Venn diagram

CONWAY

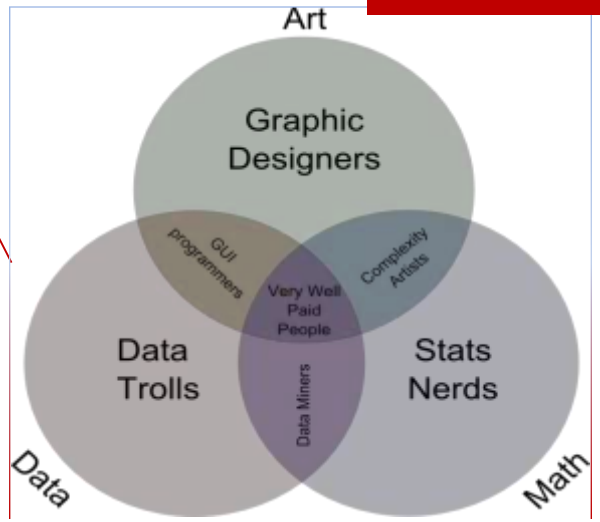


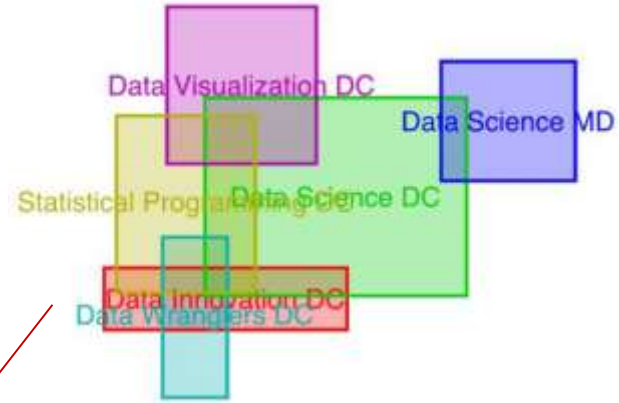
RALEIGH

MOUT



ERICKSON





5

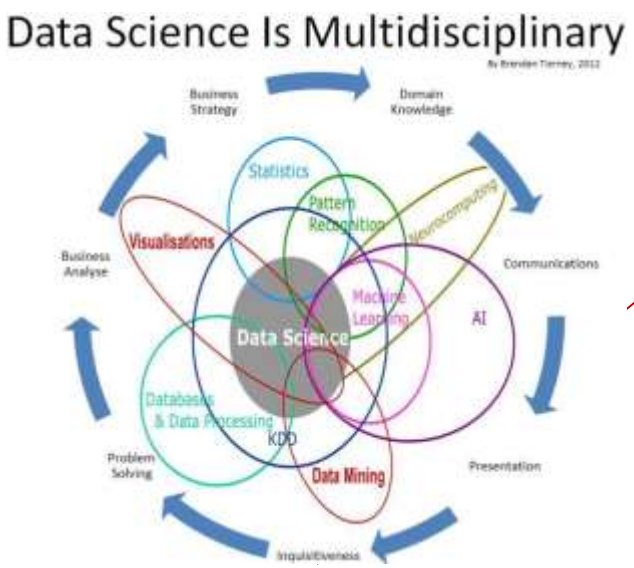
4

6

Venn diagram

>7

7



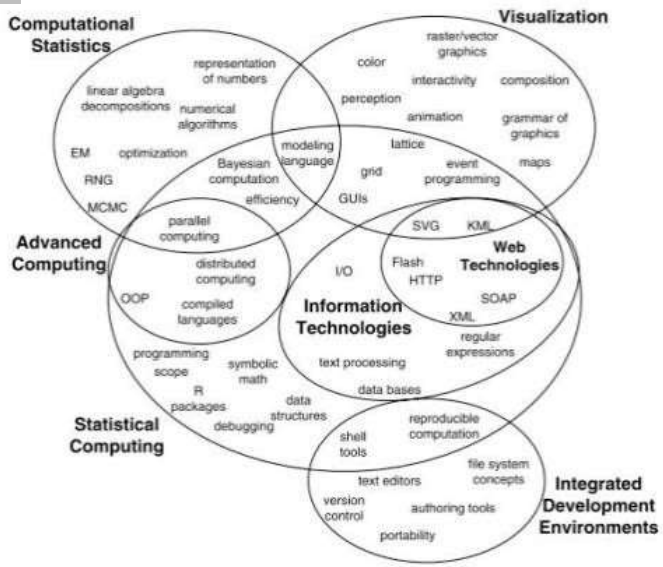
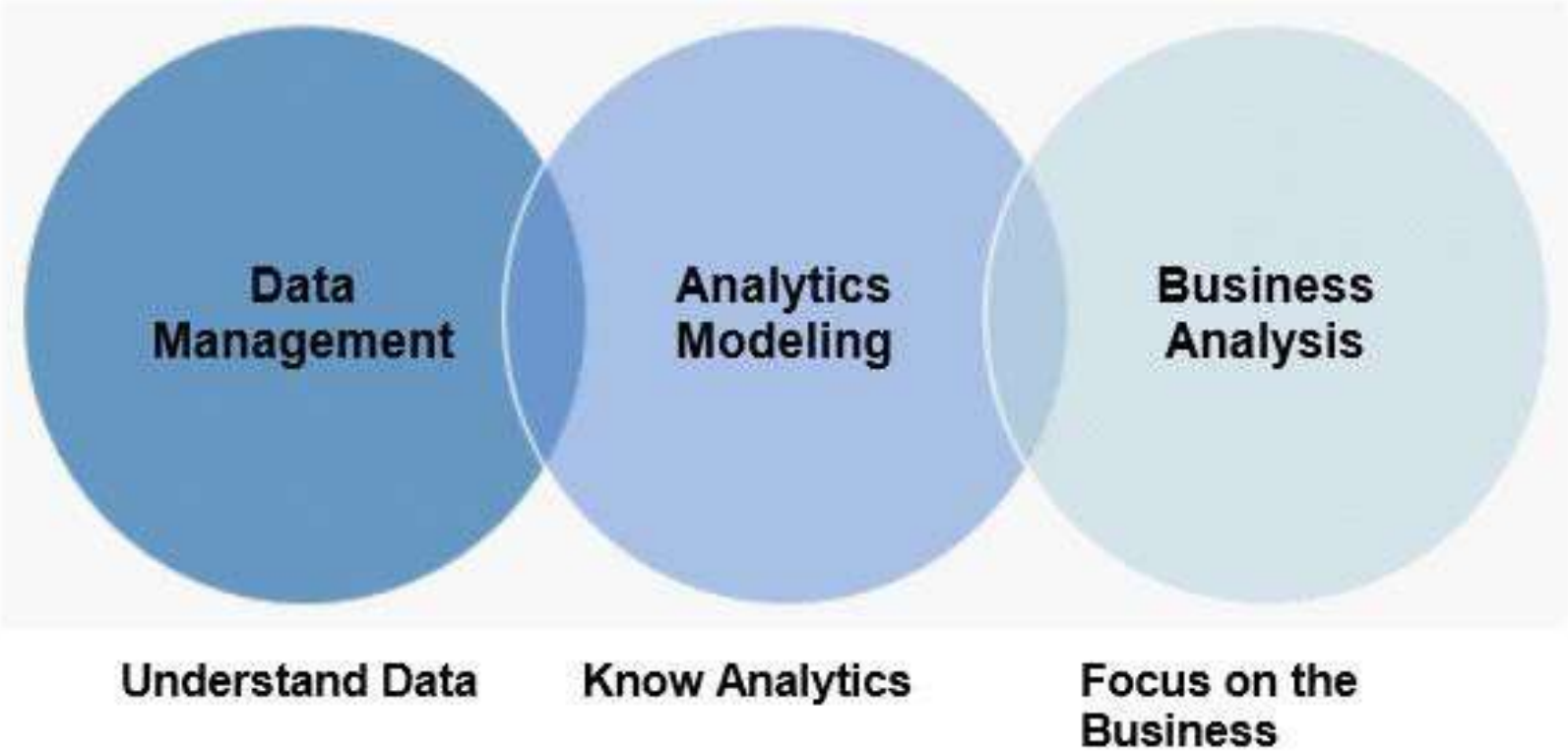


Figure 3. Core Data Scientist Skills



Source: Gartner (March 2012)

European Commission

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 IaaS 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

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*);

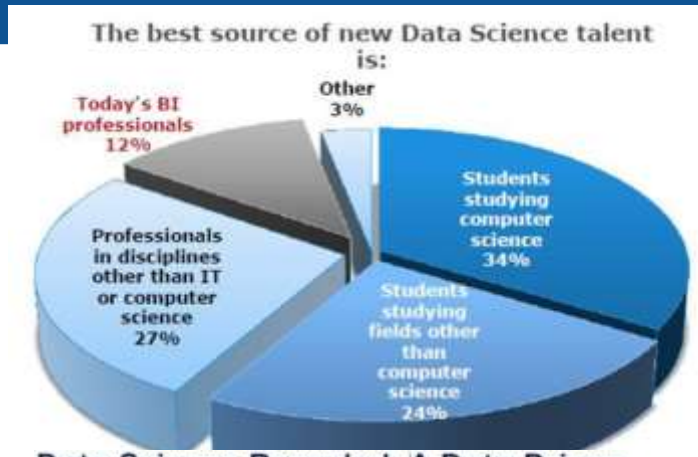
(c) To stand beside other disciplines, offering the systematic treatment of its own conceptual foundations (*new theories*).

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

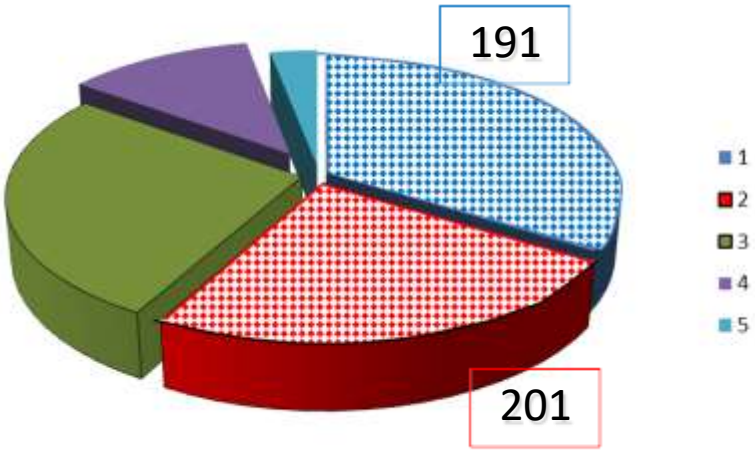
→ to be anyone's own area of specialisation → **Transdisciplinary** (like cybernetics or semiotics) or **interdisciplinary** (like biochemistry or cognitive science)?

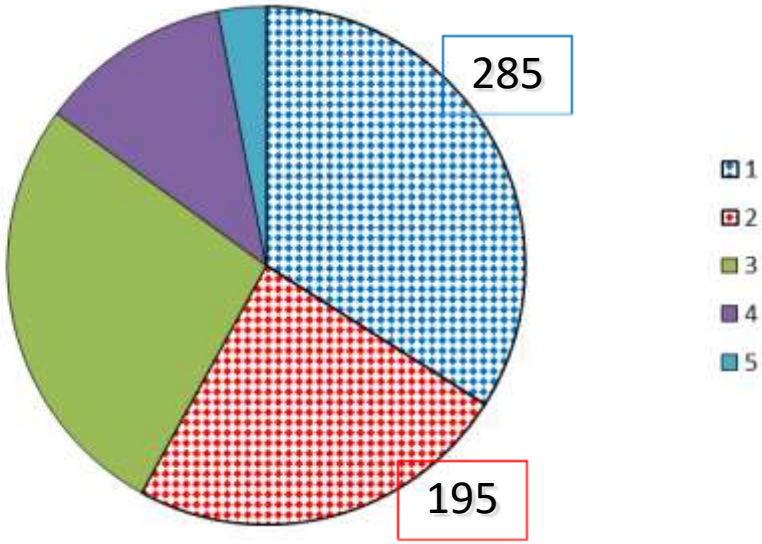
L. Floridi



Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field

http://www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf



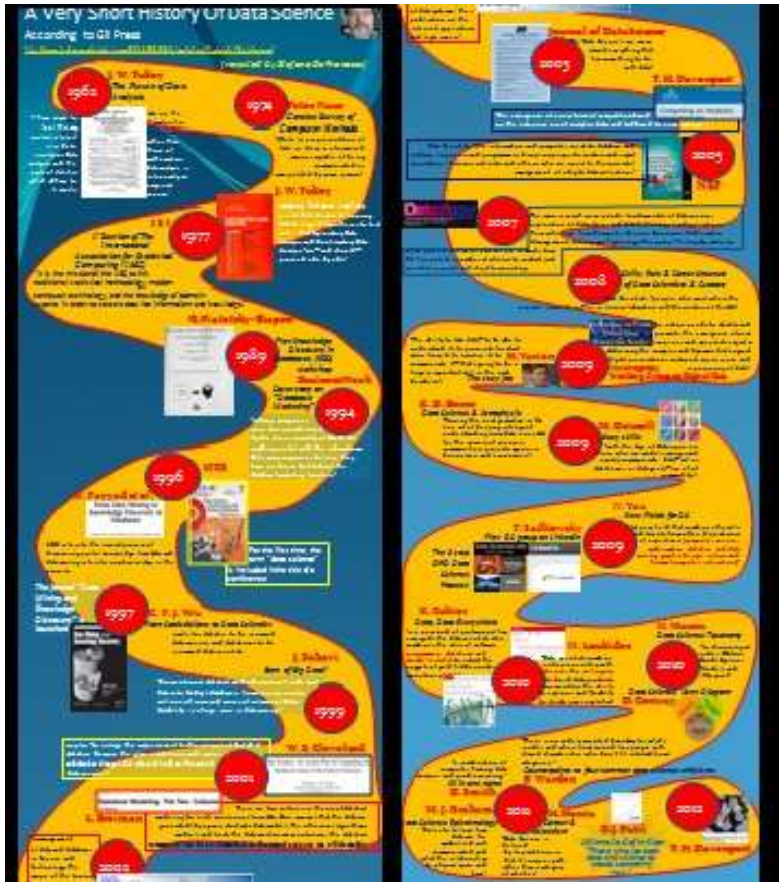


$$\text{Lie factor} = \frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}} = \begin{cases} = 1 : \text{Truth} \\ \neq 1 : \text{Lie} \end{cases}$$

$$\text{where size of effect} = \frac{|\text{second value - first value}|}{\text{first value}}$$

	second value	first value	value
Size of effect shown in graphic	191	201	0,050
Size of effect in data	285	195	0,462
Lie factor			**0,108**

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

European Commission

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."

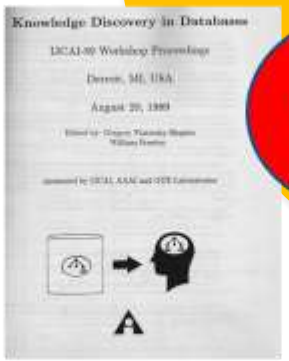
1977

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



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

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

U. Fayyad *et al.*

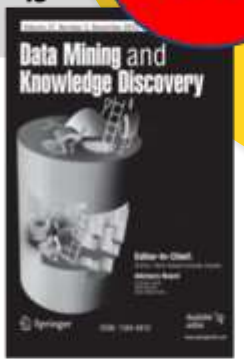
From Data Mining to Knowledge Discovery in Databases

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

European Commission

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

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

2001

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.

European Commission

"...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"

HARNESSING THE POWER of DIGITAL DATA for SCIENCE AND SOCIETY

"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

2009

"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

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

European Commission

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

European Commission

"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

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

All in one name

D. Smith

M. J. Graham

Data Science Epistemology

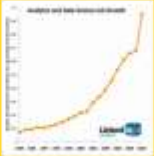
"Rules to follow. how data can be symbolized and communicated and what the relationships to physical space and time"

2011

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

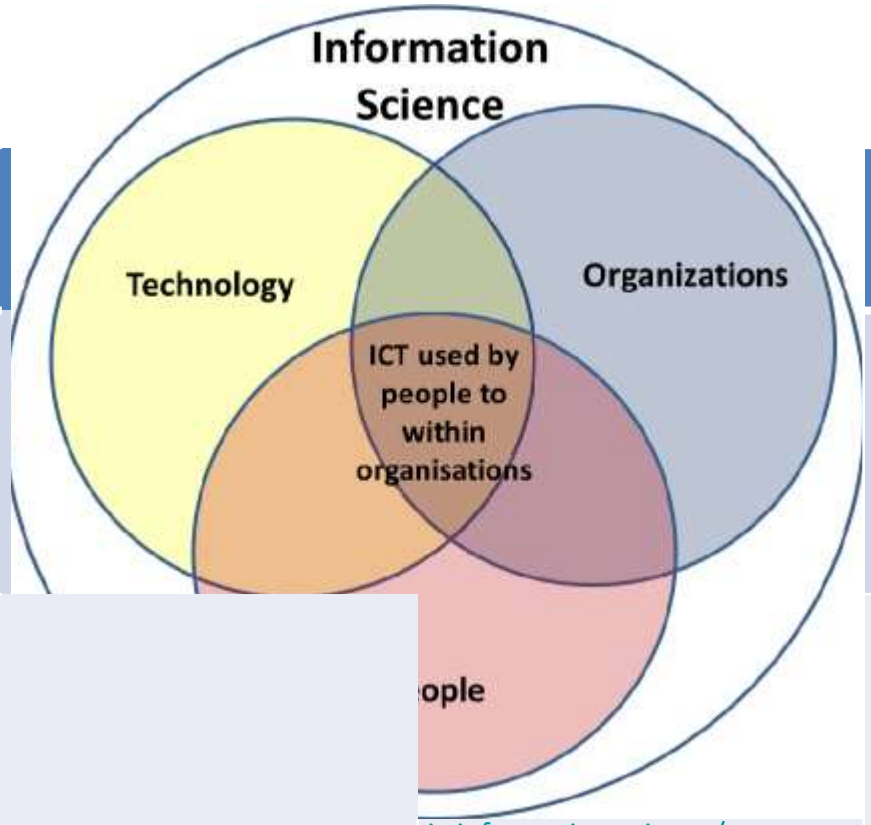
Steps to a Metaphisics of Data Science

- 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)

Beyond Data Science?

Information Science is the study of **information** and how it is used by people within **organisations**

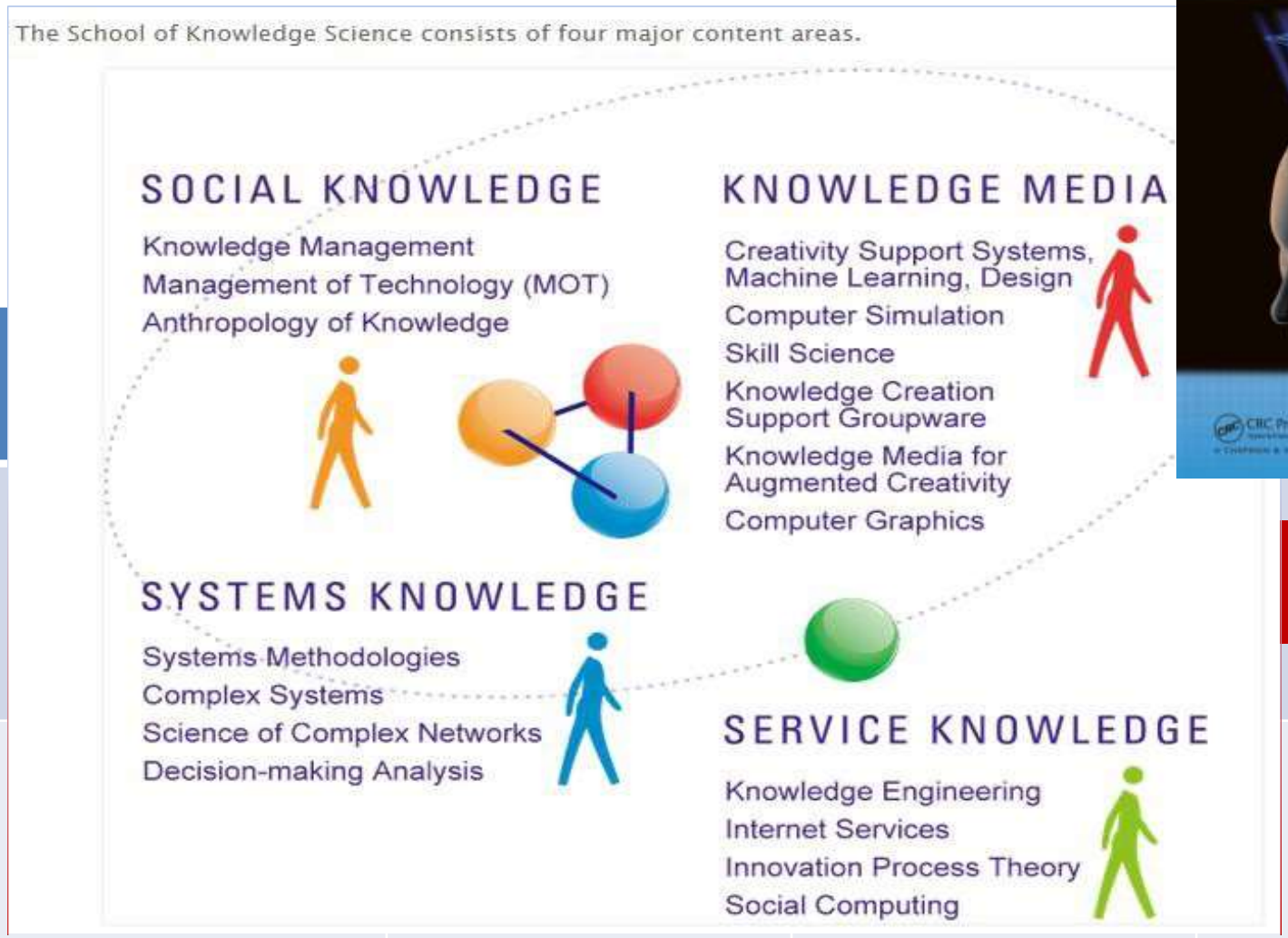


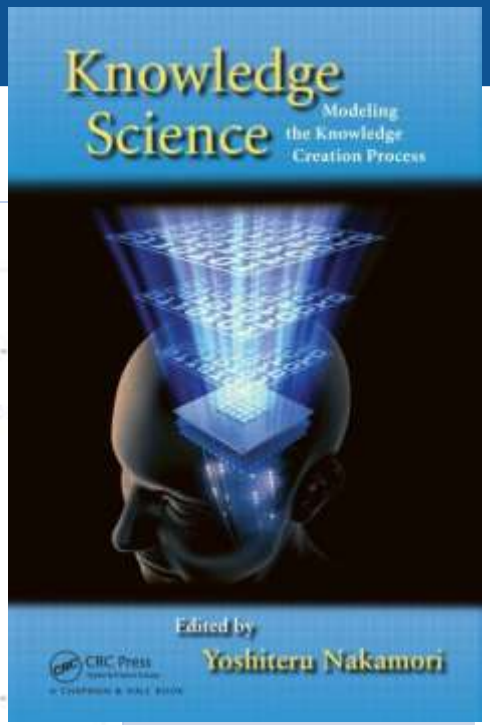
http://infosci.otago.ac.nz/what -is-information-science/

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	

Beyond Data Science?





Knowledge Science

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