UPC Universitat Politècnica de Catalunya BarcelonaTech | IDEAI Intelligent Data Science and Artificial Intelligence Research Center

Data Science

K. Gibert

Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, (IDEAI-UPC) Barcelona

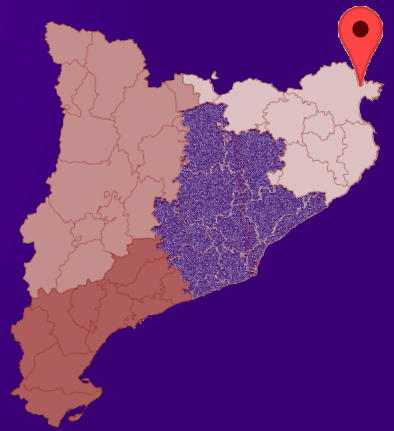
Dean of Illustrious Professional College of Informatics Engineering of Catalonia

President and founder of donesCOEINF

https://www.eio.upc.edu/en/homepages/karina,

karina.gibert@upc.edu, karina.gibert, @karinagibertk

Figuerenca



Setembre de 2024, IDEAI-UPC

IDEAI UPC

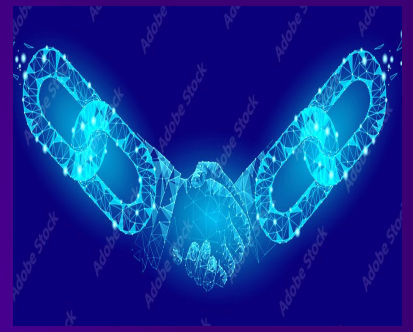
Outline

- **Introduction**
- **Data Science**
- **Brief History**
- **Concept**
- **Added value**
- **New professional profile**
- **KDD underlying process**
- **What to do with data**

Eras of Society

MASUDA, Yoneji. *The information society as post-industrial society*. World Future Society, 1981.
(ONU, 2005) Secretary-General's Roadmap for digital cooperation United Nations june 2020 disponible a https://www.un.org/en/content/digital-cooperation-roadmap/
HLP-DC (2019). The Age of Digital Interdependence - Report of the United Nations Secretary-General's High-level Panel on Digital Cooperation.

Digital Transformation



IoT Wereables 5G Cloud comp HPC Blockchain

Digital Society
NU 2018

Disclosing Complexity

Fro m 1956

IDEAI UPC

Introduction

- Knowledge Society *[United Nations, 2005]*
  - Great need of getting knowledge from
    - **Data**
    - **Organizations**
    - **Natural, industrial or artificial phenomena**
  - Support complex decision making processes
- Enormous quantities of data to analyze
  - Boom Internet late 1990s

    WWW [Tim Berners-Lee, CERN,1990]; 1995 www free&global
  - New technologies
  - Exponentially increasing
- Classical data analysis is poor
  - Too much data
  - Phenomena too complex
- New approaches required

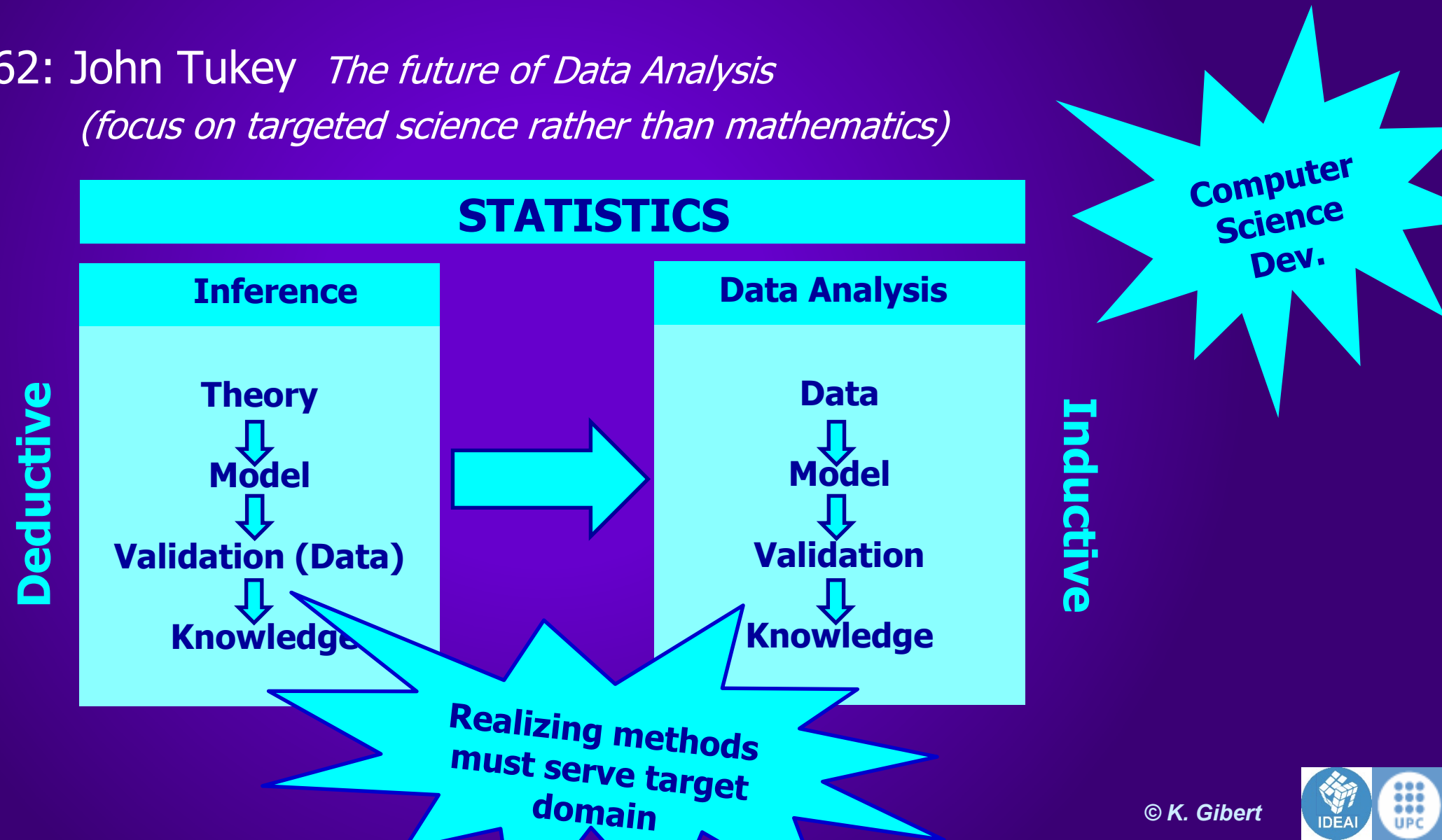
Data Science
Brief History

[Gibert, EMSO 2018]

- Availability of data
- New data-centered paradigm
- Learn from more complex phenomenon
- Methodological challenges
- Multidisciplinariety awareness
- Added value awareness
- Data Science boom
- The Fact gap
- New profesional profile and shortage awareness

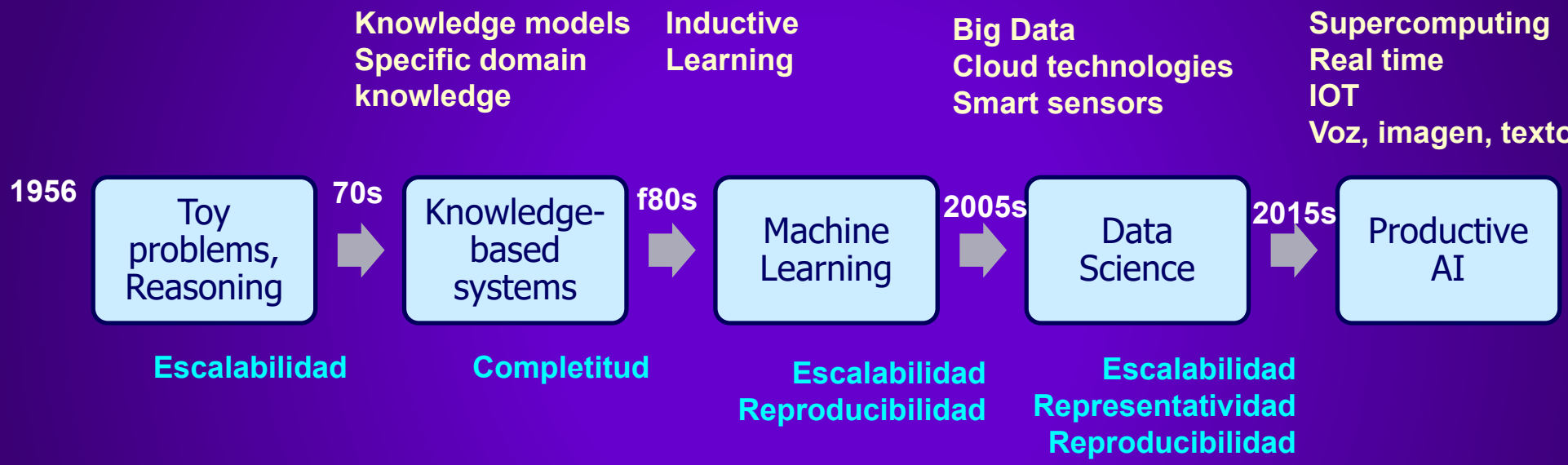
Data Analysis, Data Mining, Data Science

- 1962: John Tukey *The future of Data Analysis*
  (focus on targeted science rather than mathematics)



Artificial Intelligence

Evolution



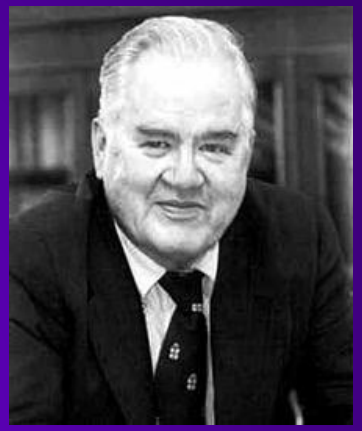
Spain:

- 1st thesis en IA 19981 (UPC, UCM)
- 1st research groups 85 (UPC)
- 1st companies 1987 (AIs SA, BCN)

Data Analysis

"Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics"

[John Tukey, 1962]



U. Princeton

IEEE Medal of Honour 1982

"more emphasis needs to be placed on using data to suggest hypotheses to test […] Exploratory Data Analysis and Confirmatory Data Analysis can—and should—proceed side by side"

[John Tukey, 1977] Exploratory Data Analysis

Data Analysis, Data Mining, Data Science

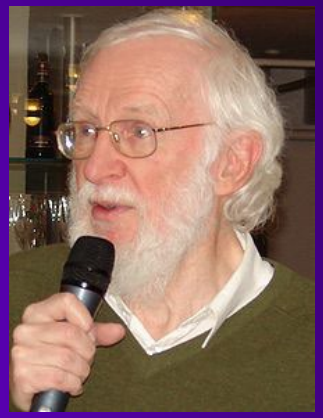
- 1962: John Tukey *The future of Data Analysis*
  (focus on targeted science rather than mathematics)
- 1974: Peter Naur UoC *coins the term*
  (data processing to get new knowledge for decision support)

Data science is the science of dealing with data (1974)

Data Science

"Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences"

[Peter Naur, 1974]



U. Copenhagen

25/10/1928-2/1/2016

Turing Prize 2005 (ALGOL 60)

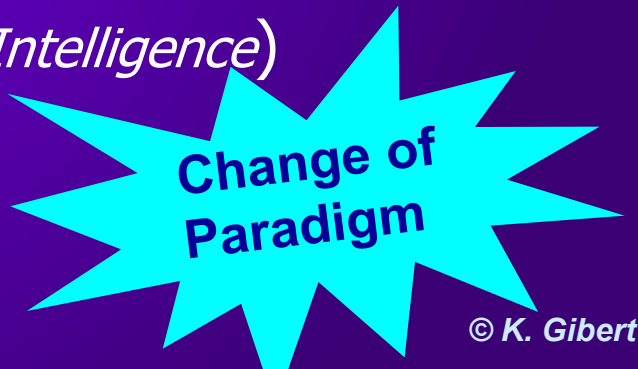
"In good data processing new, so far unknown date, may be used directly by humans to guide their actions"

[Peter Naur, 1974]

IDEAI UPC

Data Analysis, Data Mining, Data Science

- 1962: John Tukey *The future of Data Analysis*
  (focus on targeted science rather than mathematics)
- 1974: Peter Naur UoC *coins the term*
  (data processing to get new knowledge for decision support)
- 1977: International Association for Statistical Computing
  multidisciplinar (Stats+CS+K) & knowledge production approach
- 80s: AI: machine learning
  focus on data rather than experts as main source for knowledge
- 1985: AI&Stats Society
  (*links Statistics, Computing and Artificial Intelligence*)



Data Analysis, Data Mining, Data Science

- 1962: John Tukey *The future of Data Analysis*
  (focus on targeted science rather than mathematics)
- 1974: Peter Naur UoC *coins the term*
  (data processing to get new knowledge for decision support)
- 1977: International Association for Statistical Computing
  multidisciplinar (Stats+CS+K) & knowledge production approach
- 80s: AI: machine learning
  focus on data rather tan experts as main source for knowledge
- 1985: AI&Stats Society
  (*links Statistics, Computing and Artificial Intelligence*)

Artificial Intelligence and Statistics

Interdisciplinar research field

- Starting:
  - 1985: Douglas Fisher and Bill Gale (AI&Stats Society)
  - 1986: First Int'l Conference on AI & Stats
- Main goals:
  - Promote communication between AI and Statistics communities

→ *"We feel that there is great potential for development at the intersection of Artificial Intelligence, Computational Science and Statistics"*

Cheeseman and Oldford 94.

- Improve research in problems common to both (Data Mining and Knowledge Discovery, ...)

Data Analysis, Data Mining, Data Science

- 1962: John Tukey *The future of Data Analysis*
  (focus on targeted science rather than mathematics)
- 1974: Peter Naur UoC *coins the term*
  (data processing to get new knowledge for decision support)
- 1977: International Association for Statistical Computing
  multidisciplinar and knowledge production approach
- 80s: AI: machine learning
  focus on data rather tan experts as main source for knowledge
- 1985: AI&Stats Society
  (*links Statistics, Computing and Artificial Intelligence*)
- 1989: Knowledge Discovery from Databases (IJCAI)

Data Mining and Knowledge Discovery

- Interdisciplinary problem

  "Non trivial identifying of valid, novel, potentially useful, ultimately understandable patterns in data"

  [Fayyad 96]

Chief Data Officer & Group Managing Director
Barclays Bank
Chairmann Oasis-500
Yahoo Chief Data Officer&EVP (2004-2008)

- Starting:
  - 1989: First Int'l Workshop on KDD in IJCAI
  - 1994: First proceedings
  - August 1995: First Int'l Conference on KDD *(4000 submissions!!)*
  - 1996: First State of the art (Fayyad et al.)
  - 1997: Data Mining & Knowledge Discovery journal launch

Data Analysis, Data Mining, Data Science

- 1994: Companies have many data underconsumed *(Berry 94)*
- 1994: Bussiness Week *cover story on Marketing Databases*
  (bussiness impact of KDD)
- 1996: Vth Conf. Int'l Federation of Classification Societies
  (Kobe) *Data Science in title of conf & two DS special sessions*
- 1999: Knowledge @Warton: *Data Mining huge companies 'rewards*
  John Zahavi: *Scalability and Complexity*
- 2001: L. Breiman: *Statistical Modeling: The two cultures*
  (critics model-based statistical approach)
- 2005: T. Davenport: *Competing on analytics*
  (companies compete with strategic information)

Data Science
A Bussiness
Opportunity

IDEAI UPC

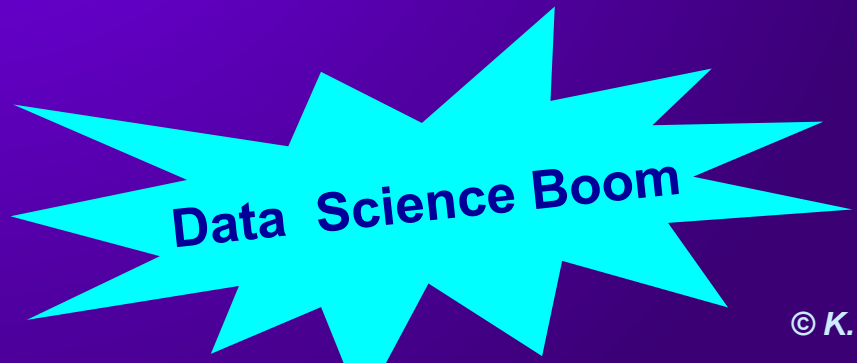
Data Analysis, Data Mining, Data Science

- 2002: Commitee on Data for Science and Technology
- (CODATA, Int'l Council for Science) *lead Data Science development*
- Ap 2002: Data Science Journal launch *(CODATA)*
- 2003: Journal of Data Science launch
- 2009: National Science and Technology Council

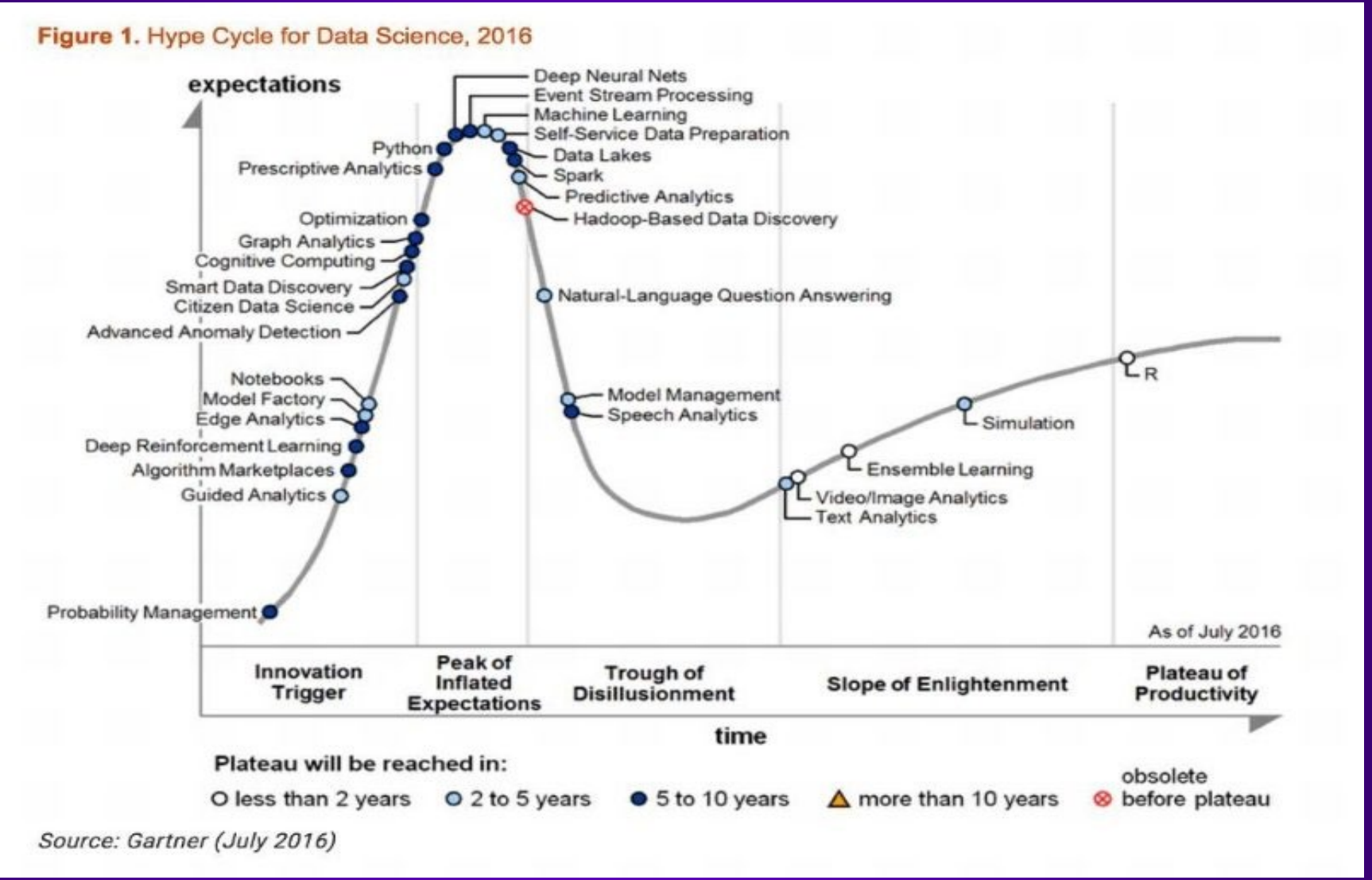
  (Commitee on Science, working group in Digital Data),

  Data Science key for the success of scientific enterprises
- 2012: Davenport and Patil:

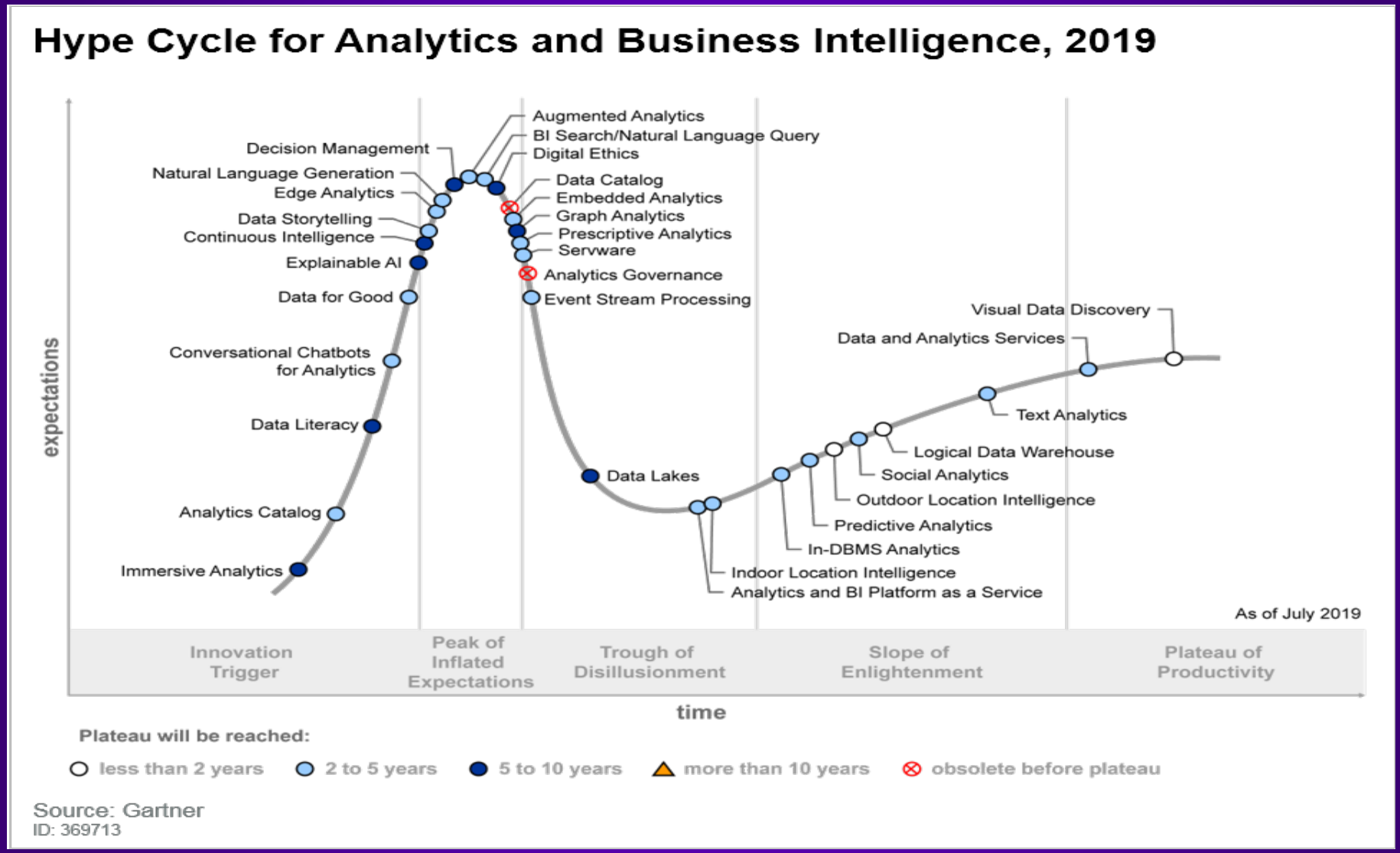
  Data Scientists: The sexiest job of the 21st century



Gartner

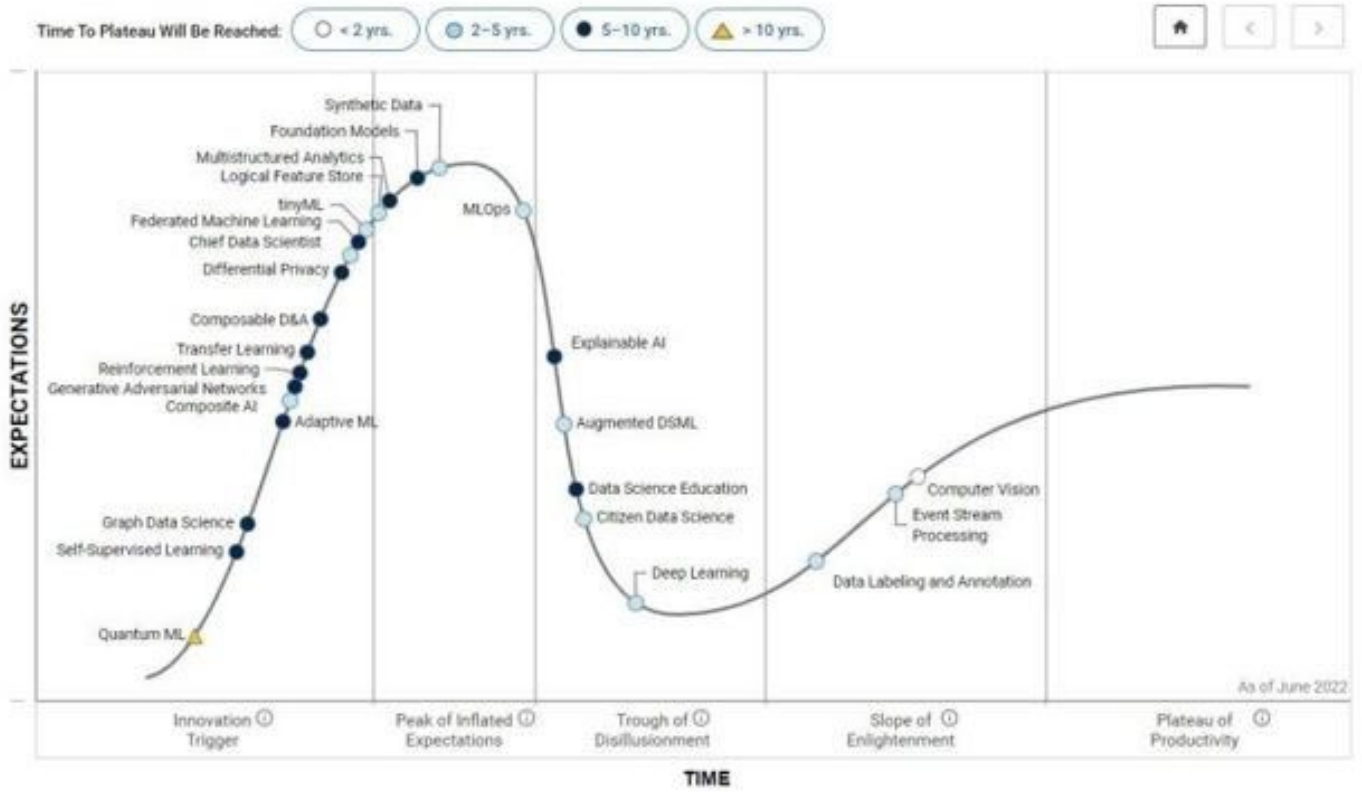


Gartner



Gartner

Gartner, “Hype Cycle for Data Science and Machine Learning, 2022”, Farhan Choudhary, Peter Krensky, June 29, 2022

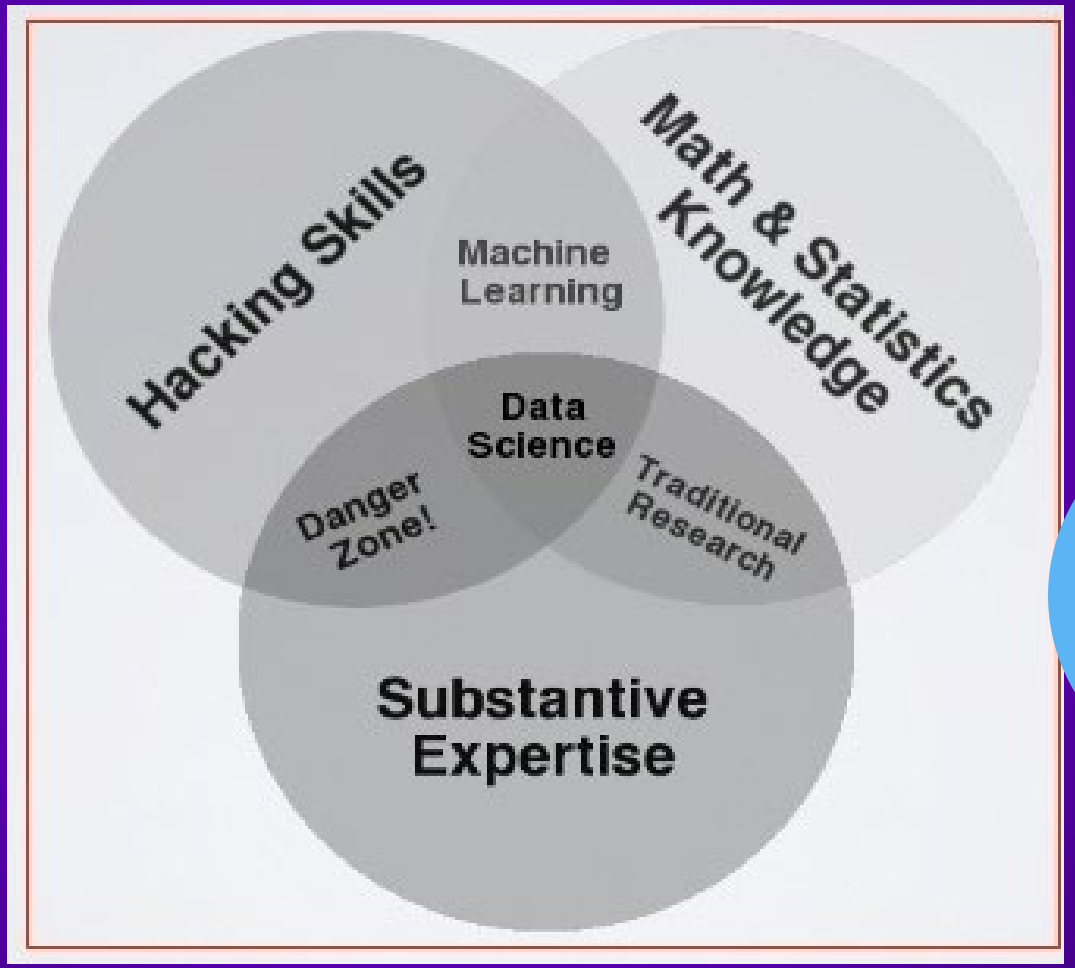


Gartner, “Hype Cycle for Data Science and Machine Learning, 2022”, Farhan Choudhary, Peter Krensky, June 29, 2022

Data Science concept

- 1974: P. Naur: *DS: science of dealing with data*
- 1997: Jeff Wu: *Statistics=Data science??*
- 2001: Cleveland: *Call for establishing a new discipline to enlarge the field of Statistics[…] Because […]implies substantial changes […] call Data Science (Stats+CS+DM)*
- 2010: Drew Conway @NYU The Data Science Venn Diagramm
  (CS+Stats+Maths+Expertise)

The Data Science Ven Diagramm Drew Conway 2010



Multidisciplinar team

Introduction

The annals of Artificial Intelligence

- Darmouth Summer School on AI, McCarthy 1956
  McCarty, Minsky, Shannon, More, Samuel…

Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

"Empirical Science. It studies the mechanisms for intelligent actions"
[McCarthy 56]

"The science of making machines do what Men do as intelligent beings"
[Minsky 68]

"AI can have two purposes. One is to use the power of computers to augment human thinking, just as we use motors to augment human or horse power […] The other is to use a computer […] to understand how Humans think […]. In a humanoid way. If you test your programs not merely by what they can accomplish, but how they accomplish it, […] you're using Artificial Intelligence to understand the human mind"
[Simon 94]

McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E., (2006) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904
https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf; **http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf**
Minsky (1968) Semantic Information Processing. MIT press

Data Science concept

- 1974: P. Naur: *DS: science of dealing with data*
- 1997: Jeff Wu: *Statistics=Data science??*
- 2001: Cleveland: *Call for establishing a new discipline to enlarge the field of Statistics[…] Because […]implies substantial changes […] call Data Science (Stats+CS+DM)*
- 2010: Drew Conway @NYU The Data Science Venn Diagramm *(CS+Stats+Maths+Expertise)*

  Mike Loukadis @ o'Reilly: *What is Data Science?*
- 2013: Somohano: *Discover what we don't know from data*

Data Science
Carlos Somohano 2013

- *Discovering what we don't know from data*
- *Getting predictive, actionable insight from data*
- *Creating Data Products with business impact*
- *Communicating relevant business from data*
- *Building confidence in decisions that drive business value*

[Carlos Somohano, 2013]

Data Science London Founder

Non-profit organization for DS since Feb 2012

IDEAI UPC

Data Science concept

- 1974: P. Naur: *DS: science of dealing with data*
- 1997: Jeff Wu: *Statistics=Data science??*
- 2001: Cleveland: *Call for establishing a new discipline to enlarge the field of Statistics[…] Because […]implies substantial changes […] call Data Science (Stats+CS+DM)*
- 2010: Drew Conway @NYU The Data Science Venn Diagramm *(CS+Stats+Maths+Expertise)*

  Mike Loukadis @ o'Reilly: *What is Data Science?*
- 2013: Somohano: *Discover what we don't know from data*
- 2013: Mattmann: *Data Mining+Algorithm+Data Management*
- 2017: Lauro: *process to transform data into actionable knowledge for predictions and support and validate decisions. CS (language), Stats (logics), domain expertise catalytic element to make transformation*

Data Science concept

- 2018: Gibert, Horsburg, Athanasiadis, Holmes *[ENVSOFT, 2018]*

Data science : emergent multidisciplinary field combining

- *Data analysis*
- *Data processing*
- *Domain expertise*

To transform data into understandable and actionable knowledge

Relevant for informed decision making (bridging the Fact Gap)

- *involves intensive consumption of available and required data*
- *Copes with data heterogeneity*
- *BigData is a tool, not the focus, but domain complexity*

Gibert, K., ,J. Horsburgh, I. Athanasiadis, G. Holmes (2018) "Environmental Data Science." Environmental Modelling & Software 106: 4-12. DOI:10.1016/j.envsoft.2018.04.005

Data Science

[Gibert, EMSO 2018]

Added value

Data Science *[Gibert, EMSO 2018]*

- 50s: informed decision making *(Luhn 1958)*

  Expert-based
- 2010: nascent data-centered economy *(cukier2010)*
- 2004: The Fact Gap *(Hammond 2004)*

- Data Science new decision making-paradigm
  - Data-driven decisions
  - Added value of organizations is information (coming from data)

Data Mining and Knowledge Discovery

- Banca d'Italia [1995]: *Built a KDD system for*
  - Daily update of the whole set of movements
  - Decide what and how to analyze
  - Select relevant results
  - Produce a daily 2-pages synthesis (natural language)

Daily support to main boss decision making

Data Mining and Knowledge Discovery

- Banca d'Italia
  - *Built a KDD system for*

  Daily support decision making of the main boss
- Technological problems
  - Millions of movements per day
  - Time to transmit to the central server?
  - Time to update the database?
  - How to select and retrieve proper data to analyze from DB?
  - How to validate results and verify technical assumptions?
- Methodological problems
  - Which is important to analyze today?
  - Which is the proper Data Mining technique?
  - Which are relevant results?
  - How to express results for the main boss?

Data Mining and Knowledge Discovery

- Big Supermarket chains *(Wal-Mart, EEUU, 1992 [Kelly 1996])*
  - Daily update the datawarehouse with costumer's bill contents
    (*20milions daily transactions [Babcock 1994]*)
  - Decide what/how to analyze: Habits *(Market Basket analysis [Brin 1997])*
  - Select relevant results
    - What is buyed more
    - Main associations between products

30% of transactions containing beer also contain diapers??????
2% of transactions contain both of these items [Agrawaal 1996]

  - Analyze the pattern in depth

Data Mining and Knowledge Discovery

- Big Supermarket chains *(Wal-Mart, EEUU, 1992 [Kelly 1996])*
  - Daily update the datawarehouse with costumer's bill contents
    (*20milions daily transactions [Babcock 1994]*)
  - Decide what/how to analyze: Habits *(Market Basket analysis [Brin 1997])*
  - Select relevant results
    - What is buyed more
    - Main associations between products

of transactions containing beer also contain diapers???????
of transactions contain both of these items [Agrawaal 1994]

  - Analyze the pattern in depth
    - *Friday between 5 and 7 pm, Young customers, Males*
  - Understanding the pattern

Just-married with small kid cannot meet friends in pub on Friday night for party
Helps wife with the shopping (required things…. Diapers for the kid);
Beer is his personal reward to spend Friday at home

What is the usefullness of elicitating this knowledge?

Babcock, C. (1994) Parallel processing mines retail data. Computer world 6
Agrawaal R, Srikant, R (1994): Fast algorithms for mining association rules in large databases . In procs 20th Int'l Conf. VLDB, Santiago, Chile pp 487-499

Data Mining and Knowledge Discovery

- What is the value of identifying that

  Young new fathers buy diapers and beer on Friday

  - Acquiring strategic information
  - Capacity of planning actions

    Wal-Mart moved the beer next to the diapers and beer sales went up
  - Capacity of becoming PROACTIVE

    What about moving snacks (peanuts and pretzels) next to diapers?

- Support decision making through informed-decision

  Buying department

  Marketing department

Important economic implications

- From then on: apparently misplaced things in stores

Data Mining and Knowledge Discovery

- From then on: apparently misplaced things in stores

The added value of Big Data

1994: Companies have many data underconsumed *(Berry 94)*

- 2016: Caffo: *DS hype flame out when is about data rather than Scienc*
- 2017: Baeza-Yates: *Big data or right data*

- In-silico discoveries from bigdata

Data Science

[Gibert, EMSO 2018]

Data scientists

A new profile

Data Scientist profile

- Multidisciplinariety
  +
- High-category goals

→ New skills

- 2001: WS Cleveland Bell Labs *limits of Data analysis. Call for involvement of academics in DS CV*
- 2005: National Science Foundation *defines CV for Data Scientists*
- 2008: JISC: *The skills, role and career structure of Data Scientists& Curators*
- 2017: ACM: *guidelines to integrate Data Science into degree programmes*

Integral approach
Call for Data Science CV

Data Analysis, Data Mining, Data Science

- 2009: Hal Varian, *Google's chief economist:*
  scarce ability of understand data and extract value
  Jeff Hammebacher @Facebook: *What does a Data Scientist do?*
- 2010: Hilary Masson & Chris Wiggins @ Dataist
- 2011: DJ Patil @Linkedin: *data scientist vs data analyst*
  First Data Scientist Chief in the White House USA
- 2012: Josh Wills: *A Data Scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician*
- 2012: Davenport and Patil:
  Data Scientists: The sexiest job of the 21st century
- 2013: Carlos Somohano @DataScience London
  Vincent Granville @DSC *Horizontal vs. Vertical Data Scientist*

Data Science

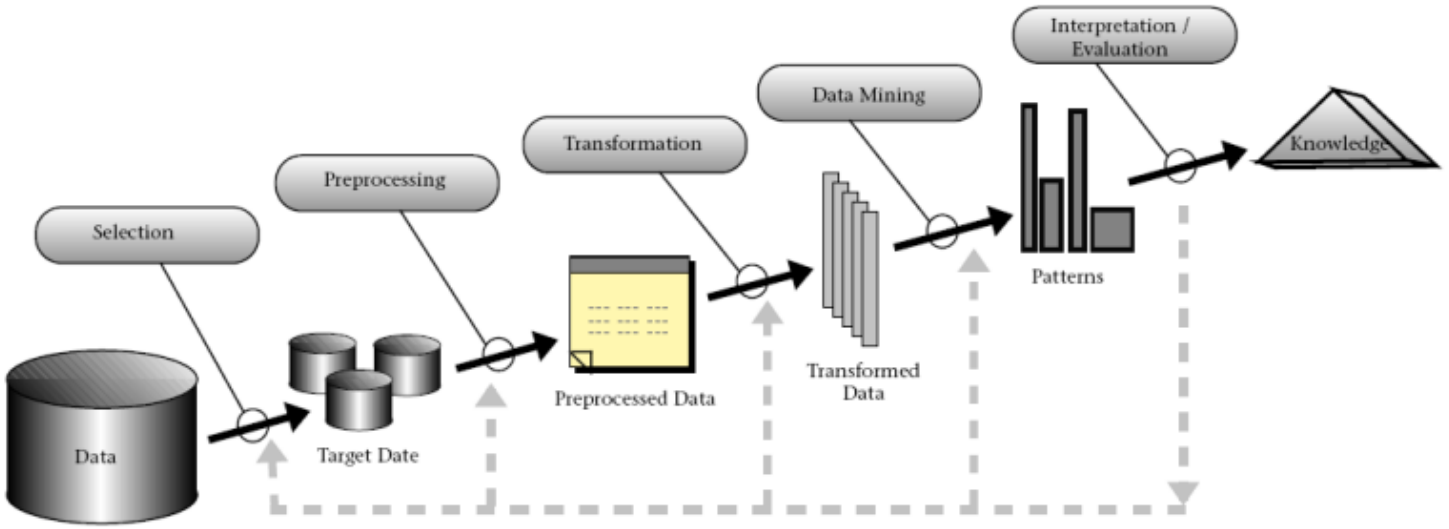
[Gibert, EMSO 2018]

Underlying KDD process

IDEAI UPC

Data Mining and Knowledge Discovery

Knowledge Discovery System [Fayy96]:

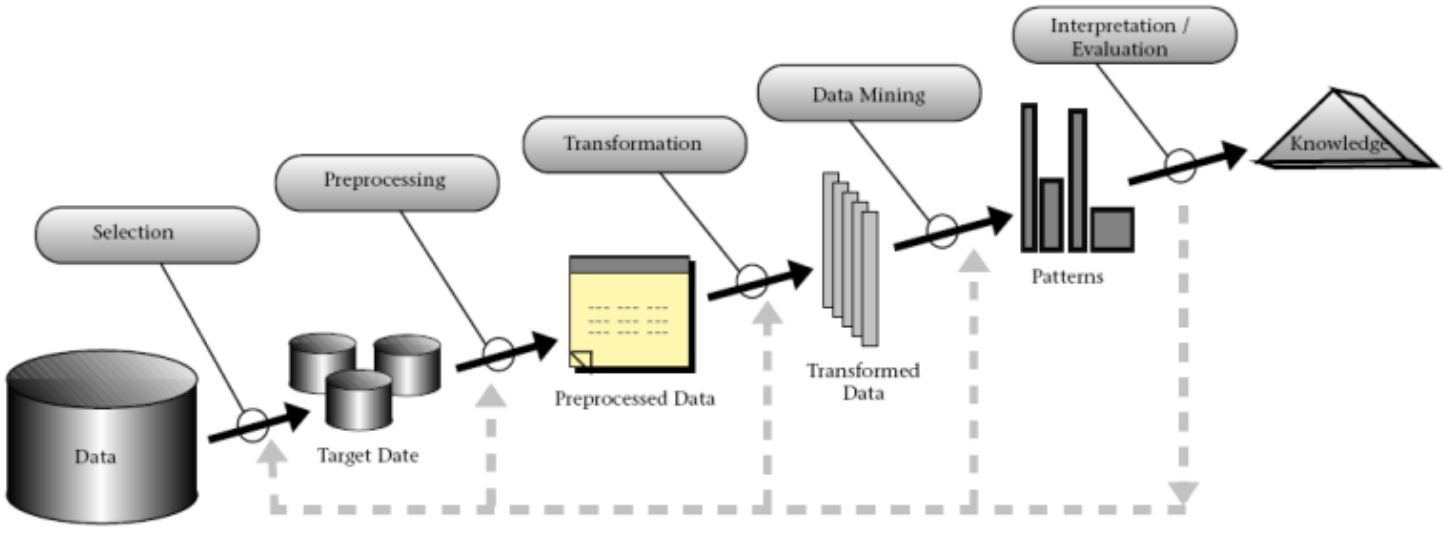


- Problem definition
- Data collection
- Data cleaning and preprocessing
- Dimensionality reduction
- DM technique choice
- Data mining
- Interpretation and discovered knowledge production

Terminologycal ambiguity
Data Mining vs KDD

Data Mining and Knowledge Discovery

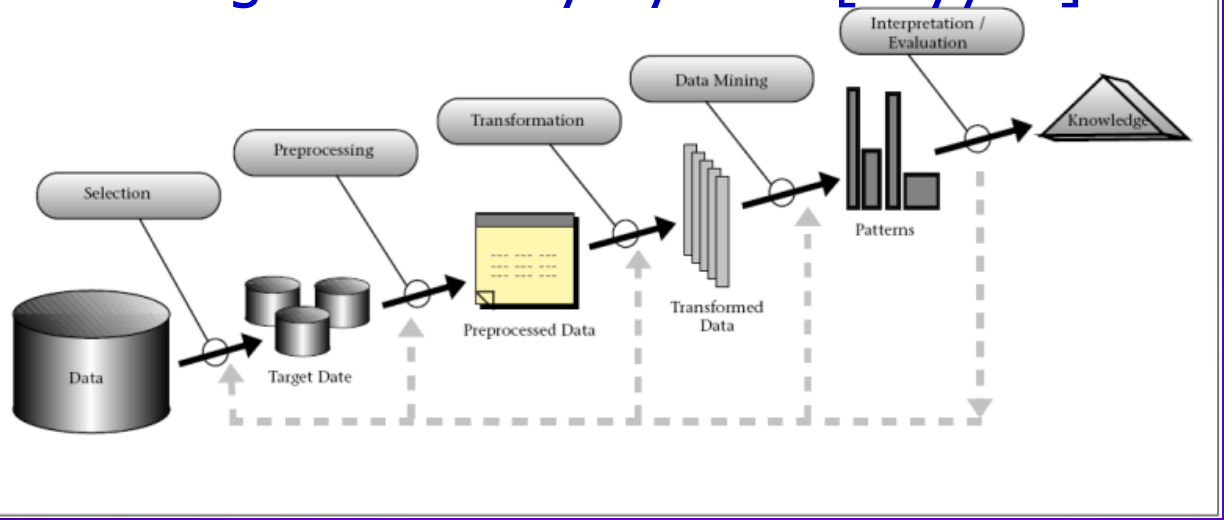
Knowledge Discovery System [Fayy96]:



- Very ambicious goals

Data Mining and Knowledge Discovery

Knowledge Discovery System [Fayy96]:



- Very ambicious goals
- No complete system on yet
  - Connection to DataWarehouses
  - Tools to assist preprocessing
  - Collection of data mining techniques *(AMD, NN, IR, AssR, Reg…)*
  - Some help on reporting phase
  - Manual process management and knowledge production

Data Mining and Knowledge Discovery

- New paradigm proposed by Fayyad

"Most previous work on KDD has focussed on [...] data mining step. However, the other steps are of considerable importance for the successful application of KDD in practice""

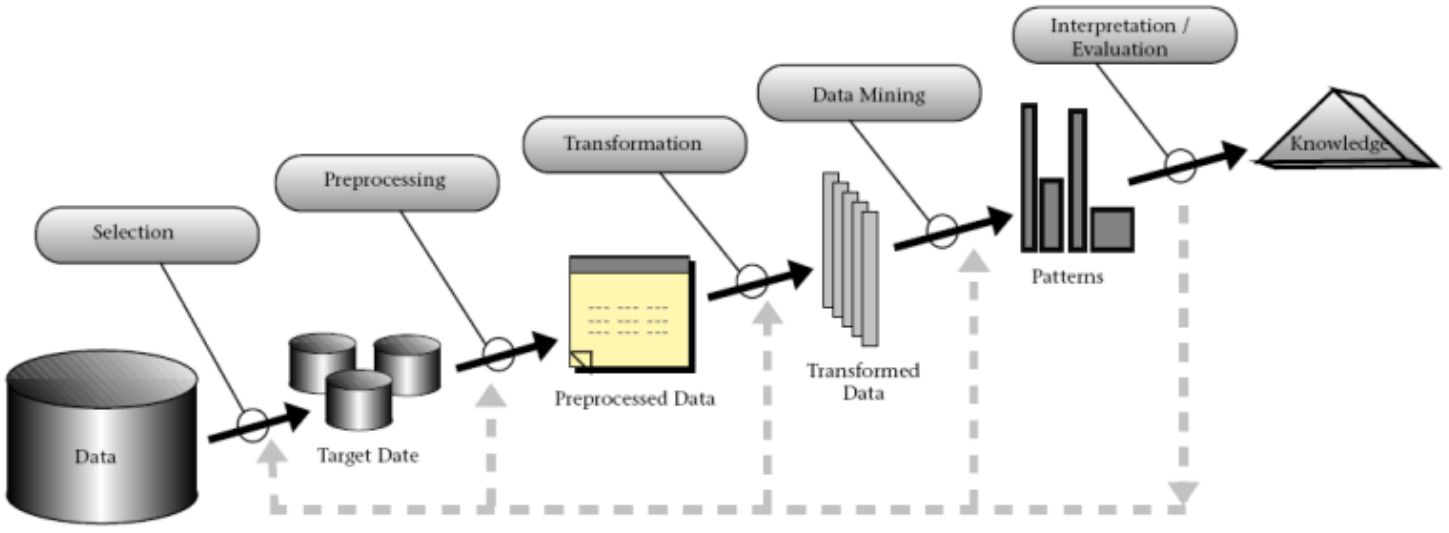
[Fayyad 96]

- Include prior and posterior analysis in KDD
  - Requires Great efforts in real applications
    - Specially in medical systems (uncertainty, imprecise, multi-scaled,..)
  - Time consuming, difficult (no standard methodology stablished)
  - Expert interaction required
  - Domain-dependent?

  - After good prior analysis, proper data mining easy

Data Mining and Knowledge Discovery

Knowledge Discovery System [Fayy96]:



- Wide scope approach

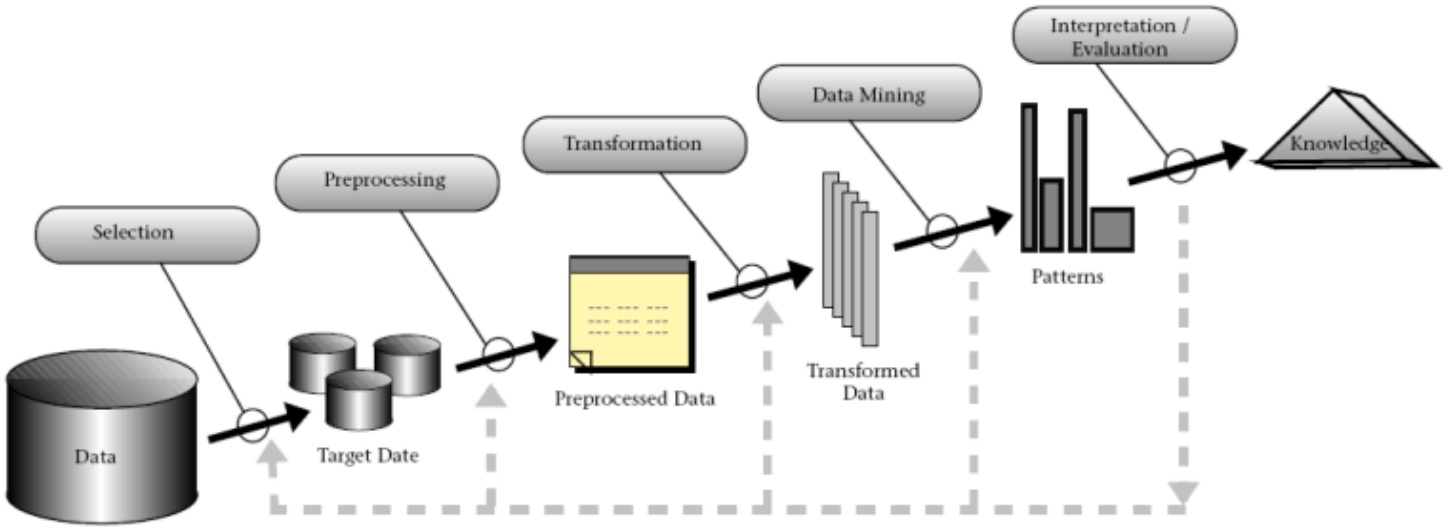
- Also interesting to better understand very complex small datasets

Multidisciplinariety

Combination or hybridation of techniques

Data Mining and Knowledge Discovery

Knowledge Discovery System [Fayy96]:



- **Preprocessing (see paper)**

Gibert, K., Sànchez–Marrè, M., & Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications*, *29*(6), 627-663.

- **Data Mining Methods Conceptual Map (DMMCM) (see paper)**

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. Environmental modelling & software, 110, 3-27.

Data Mining and Knowledge Discovery

A balanced perspective cannot be acquired by studying disciplines in pieces; the consilience among them must be pursued. Such unification will be difficult to achieve. But I think it is inevitable. Intellectually it rings true, and it gratifies impulses that arise from the admirable side of human nature. To the extent that the gaps between the great branches of learning can be narrowed, diversity and depth of knowledge will increase.

[E.O. Wilson 1998]

Biologist, Harvard U, EEUU

Twice Pulitzer; Times (1995) 25 most influencial people in America

Data Mining and Knowledge Discovery

The Elephant and the blind Men (Ancient India)

[Puchala 1971]

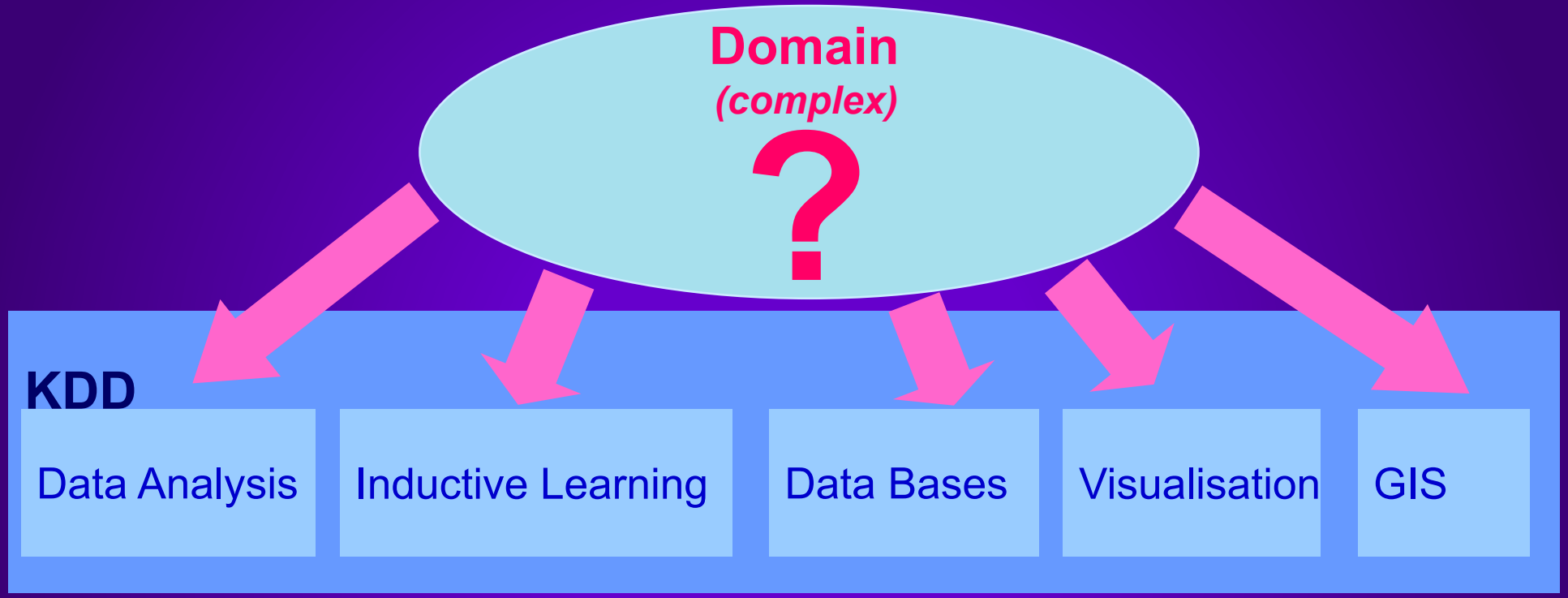
An elephant came to a small town (had ever seen one)
Ancient council (5 blind men) went to feel the elephant with their hands.
Later on, they sat down and began to discuss their experiences.

- One who touched the trunk and feld like a thick tree branch.
- Another who touched the tail feld like a snake or rope.
- Another who touched the leg, feld like a pillar.
- Another who touched the ear, said like a huge fan
- Another who touched the side, said like a wall.

Consilience
Multidisciplinariety

All had different partial views of the same reality
Putting all partial views together, the complete view could emmerge

Data Mining and Knowledge Discovery



- Prediction
- Description
- Summary
- Overview

Discovered knowledge base
Domain model

- Validity
- Utility
- Simplicity/complexity
- Novelty

KDD uses

Decision Support

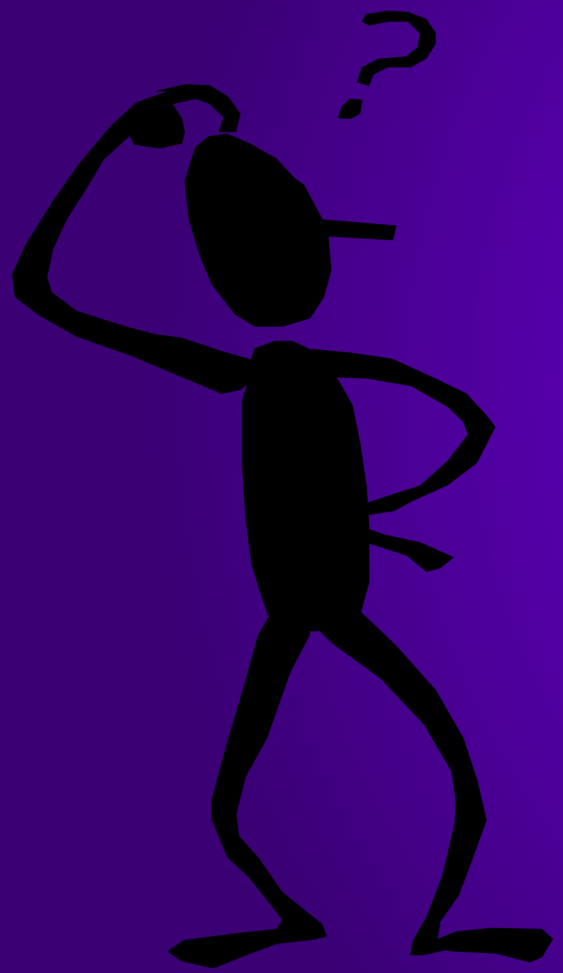
- Improving complex decision making
- Intelligent Decision Support Systems

Bussiness intelligence

Domains

- Marketing
- Bussiness
- Research

Medical, industrial, environmental applications



Any question?...

K. Gibert

https://www.eio.upc.edu/en/homepages/karina

karina.gibert@upc.edu @Karinagibertk

KEMLG-@-IDEAI: Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, Barcelona

Dean of Illustrious Professional College of Informatics Engineering of Cataloonia (COEINF)

Founder and president of donesCOEINF