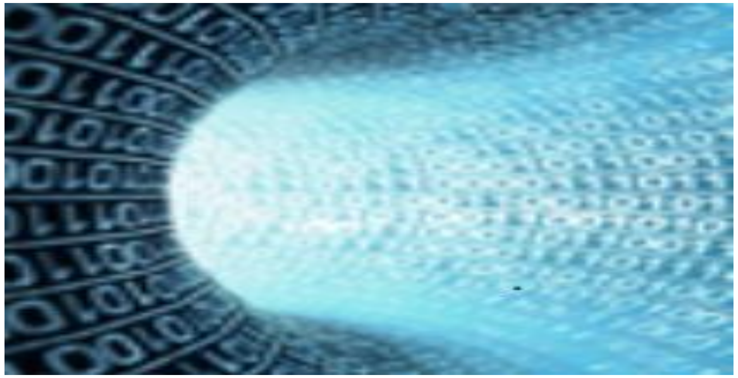
Пловдивски университет „Паисий Хилендарски“
Факултет по математика и информатика

Data Science and Big Data
Наука за данните и Големите данни

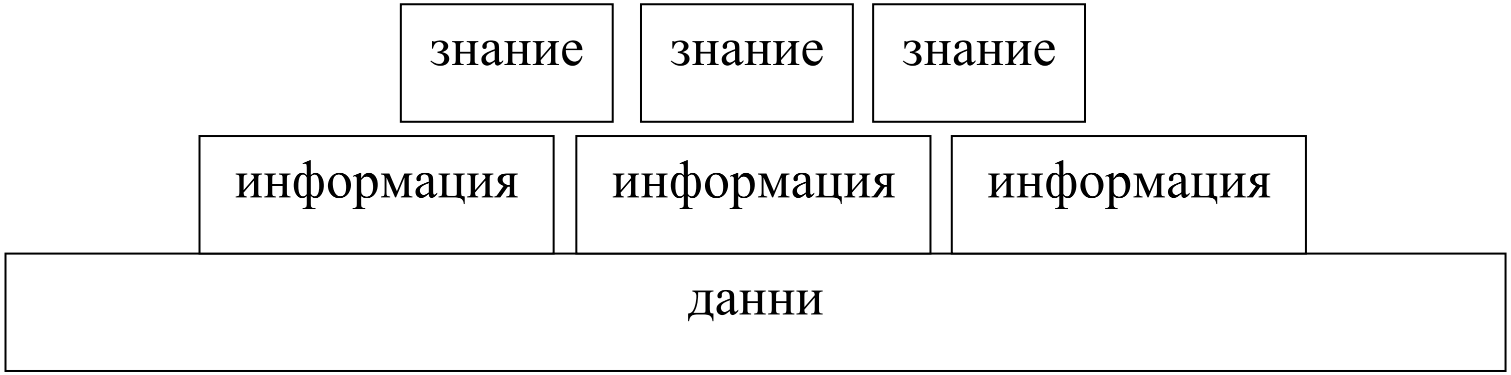


Проф. д.м.н. Снежана Гочева-Илиева
(По материали от Интернет)

28.02.2018

Данни

Терминът **данни** означава неструктурирани факти за обекти, които се съхраняват в «суров вид», без да се използват. Когато тези данни се обработват с някаква цел (за намаляване неопределеността на някакъв обекта(и), те се превръщат в информация. Данните често са възприемани като най-ниското ниво на абстракция, от което информацията и знанието произхождат.



знание знание знание

информация информация информация

данни

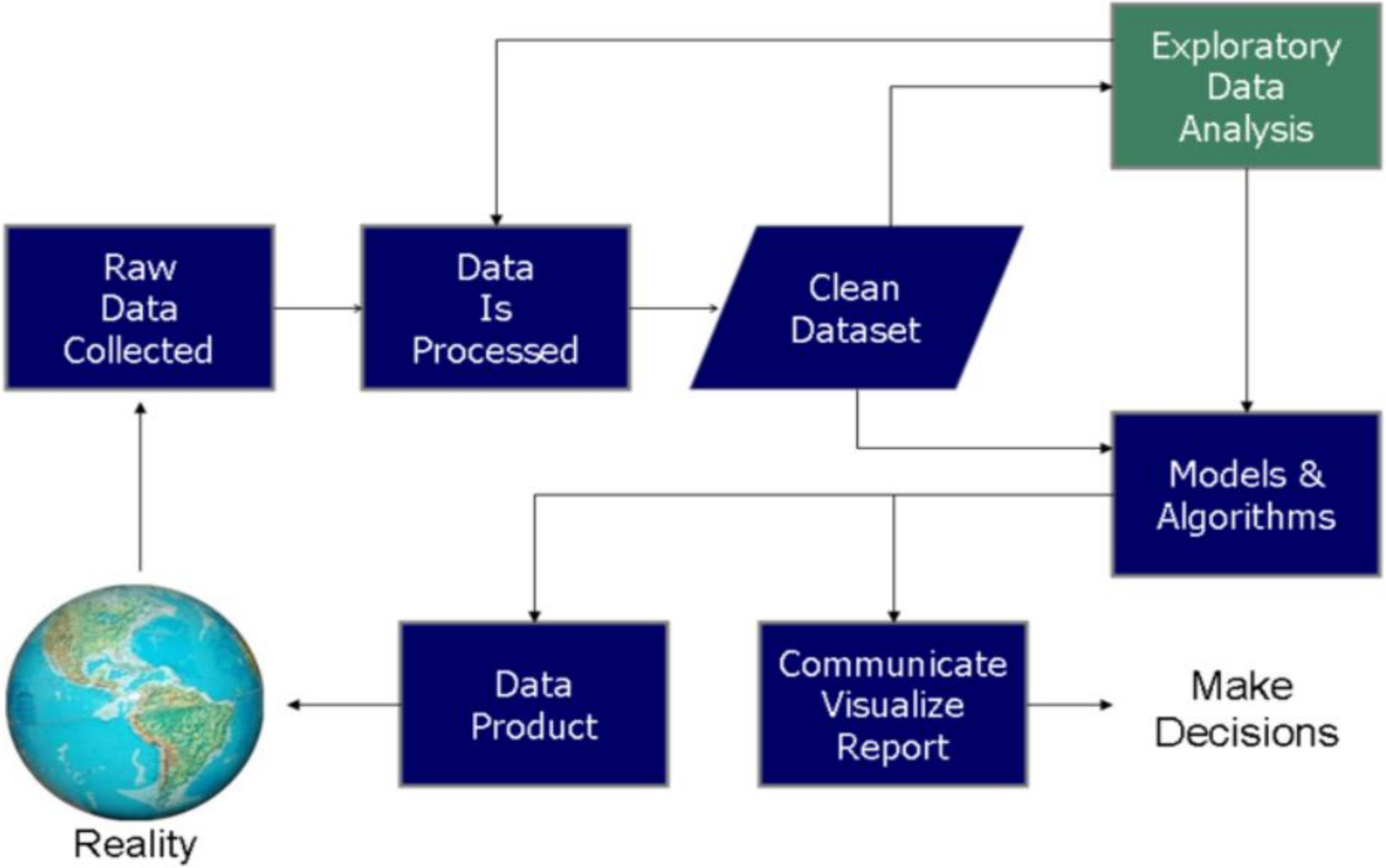
Какво е наука за данните?

Data science, also known as **data**-driven **science**, is an interdisciplinary field of **scientific** methods, processes, algorithms and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured, similar to **data** mining.

Data science is a "concept to unify **statistics, data analysis, machine learning** and their **related methods**" in order to "understand and analyze actual phenomena with data. It employs techniques and theories from the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, uncertainty quantification, computational science, data mining, databases, and visualization.

2010- Възникна нова професия – **data scientist** (учен по данните)

Data Science Process



Raw Data Collected

Data Is Processed

Clean Dataset

Exploratory Data Analysis

Models & Algorithms

Data Product

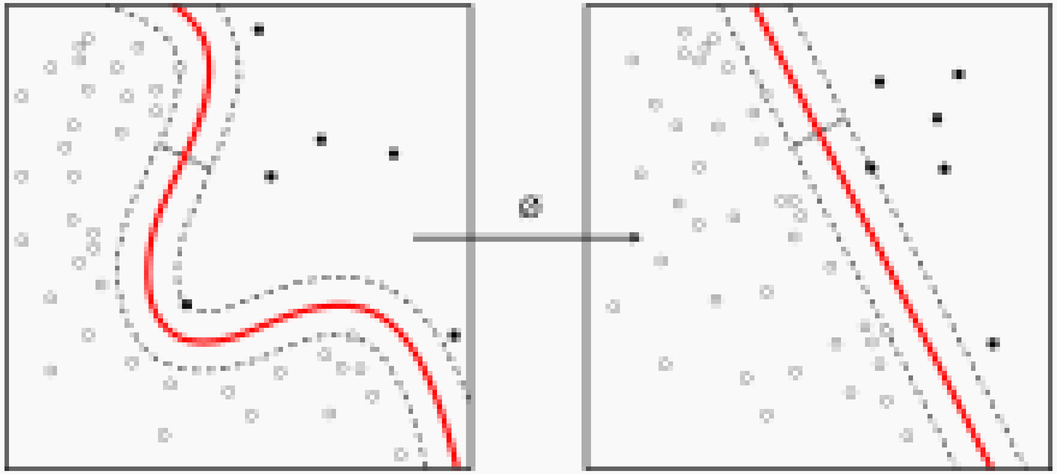
Communicate Visualize Report

Make Decisions

Reality

Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013

Machine learning and data mining



Problems

Classification

- Clustering
- Regression
- Anomaly detection
- Association rules
- Reinforcement learning
- Structured prediction
- Feature engineering

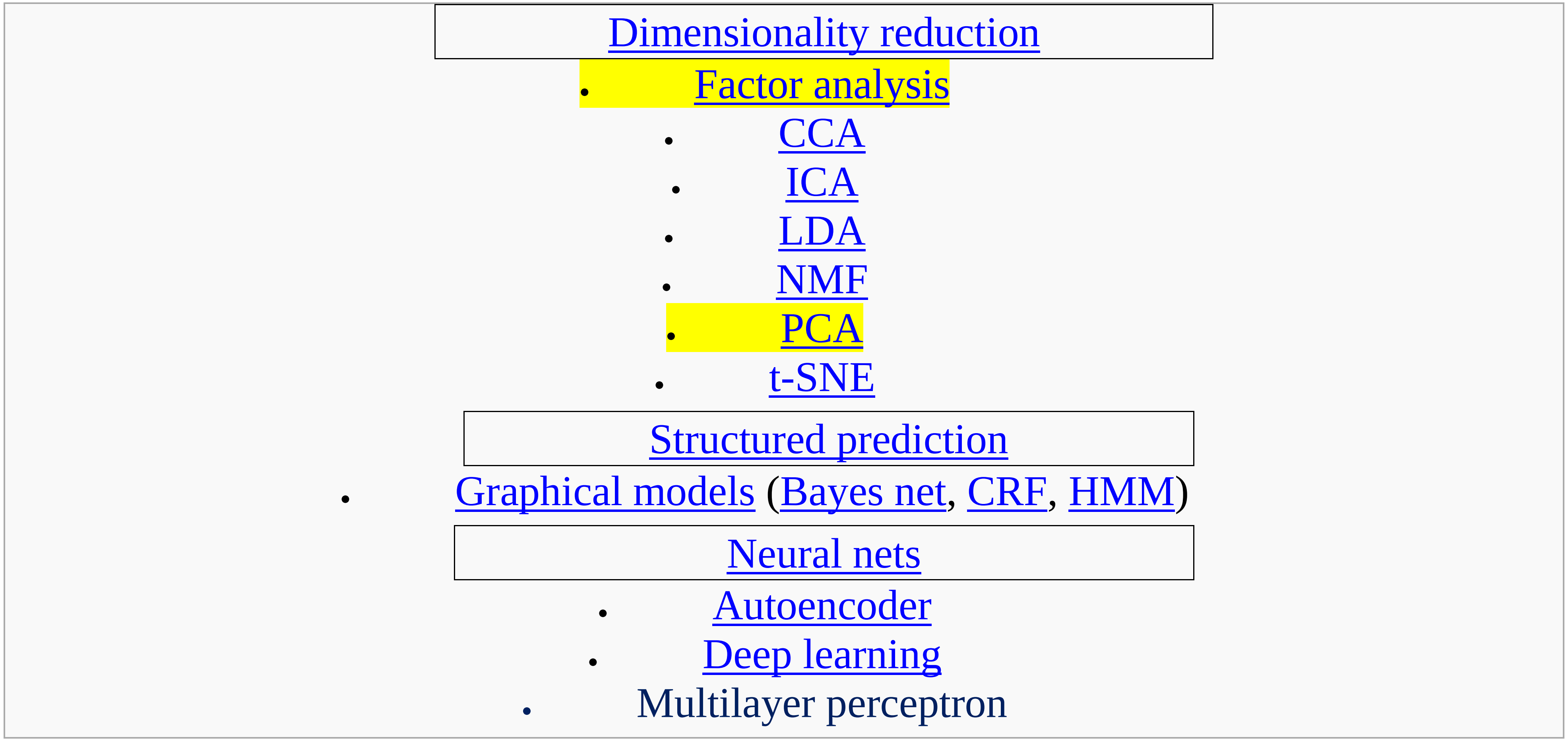
- Feature learning
- Online learning
- Grammar induction

Supervised learning, (**classification** • **regression**)

- Decision trees
- Ensembles (Bagging, Boosting, Random forest)
- *k*-NN
- Linear regression
- Naive Bayes
- Neural networks
- Logistic regression
- Perceptron
- Relevance vector machine (RVM)
- Support vector machine (SVM)

Clustering

- BIRCH
- CURE
- Hierarchical
- *k*-means
- Expectation–maximization (EM)



Dimensionality reduction

- Factor analysis
- CCA
- ICA
- LDA
- NMF
- PCA
- t-SNE

Structured prediction

- Graphical models (Bayes net, CRF, HMM)

Neural nets

- Autoencoder
- Deep learning
- Multilayer perceptron

От: Wikipedia

Забележка. В жълт фон са темите, които преподаваме във ФМИ на специалност БИТ 1 и 2 курс, и част от тях – и в някои избираеми.

Applications of Data Science

- **Internet search**: Search engines make use of data science algorithms to deliver best results for search queries in a fraction of seconds.
- **Digital Advertisements**: The entire digital marketing spectrum uses the data science algorithms - from display banners to digital billboards.
- **Recommender systems**: A lot of companies use this system to promote their products and suggestions in accordance with the user’s demands and relevance of information. The recommendations are based on the user’s previous search results.

https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article

Big Data

Big Data refers to humongous volumes of data that cannot be processed effectively with the traditional applications that exist. The processing of Big Data begins with the raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.

The definition of Big Data, given by Gartner is, "**Big data is high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation**".

Big data and the blind men

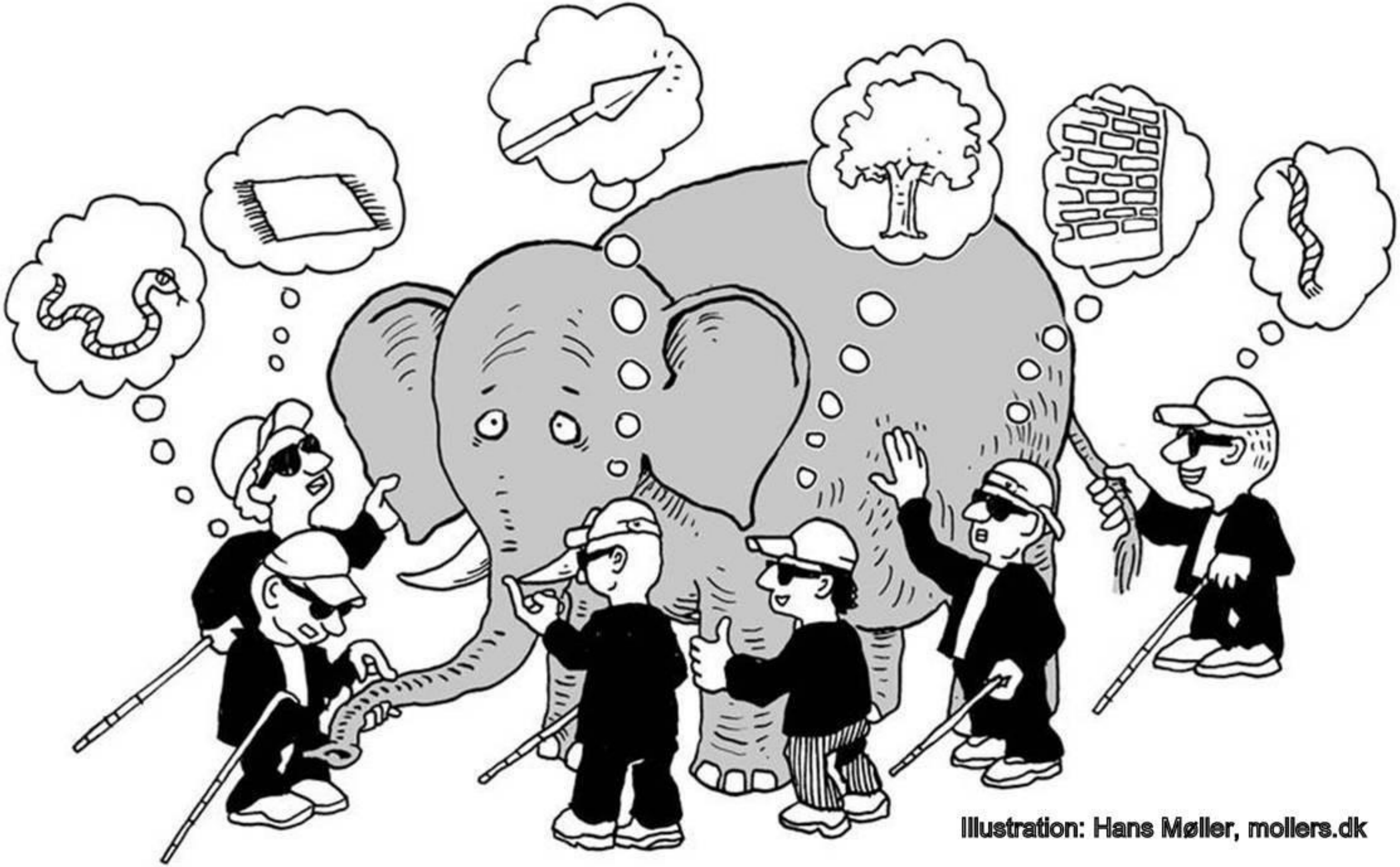


Illustration: Hans Møller, mollers.dk

Applications of Big Data

- **Big Data for financial services (за банки и др. финансови услуги)**: Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks use big data for their financial services. The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems which can be solved by big data.
- **Big Data in communications (за комуникации)**
- **Big Data for Retail**: **(Търговия на дребно)**

The skills you require

To become a Data Scientist

- **Education: 88% have a Master’s Degree and 46% have PhDs**
- In-depth **knowledge of SAS and/or R - statistical programming languages**: For Data Science, R is generally preferred.
- **Python coding**: Python is the most common coding language that is used in data science along with Java, Perl, C/C++.
- **Hadoop platform**: Although not always a requirement, knowing the Hadoop platform is still preferred for the field.
- **SQL database/coding**: Though NoSQL and Hadoop have become a major part of the Data Science background, it is still preferred if you can write and execute complex queries in SQL.
- **Working with unstructured data**: It is most important that a Data Scientist is able to work with unstructured data be it on social media, video feeds, or audio.

To become a Big Data professional

- **Analytical skills**: The ability to make sense of the piles of data that you get. With analytical abilities, you will be able to determine which data is relevant to your solution, more like problem-solving.
- **Creativity**: You need to have the ability to create new methods to gather, interpret, and analyze a data strategy.
- **Mathematics and statistical skills**: Good, old-fashioned "number crunching". This is extremely necessary, be it in data science, data analytics, or big data.
- **Computer science**: Computers are the workhorses behind every data strategy. Programmers will have a constant need to come up with algorithms to process data into insights.
- **Business skills**

https://databricks.com/blog/2018/01/17/matei-zaharias-5-predictions-about-ai-in-2018.html

Databricks' Chief Technologist, Matei Zaharia:

Five key predictions for 2018

1. **Data will be the central competitive advantage.**
2. **AI will find new use cases, starting with verticals.**
3. **Data scientists will continue to grow in number.**
4. **Deep learning frameworks will start to converge and move up in abstraction.**
5. **The cloud will enable new data application architectures.**

Моето мнение за бъдещето:

Crowdsourcing paradigm – използвай тълпите!

www.shutterstock.com · 718442440

Примери и приложения на краудсорсинга

- **Колективни дейности – Wikipedia, freelancer, facebook…**
- **Събиране на данни за определена област – ресторанти, хотели – Clickworkers.com …**
- **Маркетнг, продажби, нови продукти …**
- **Решаване на трудни научни проблеми – AmazonTurk.com**

Използвай тълпите:

- Създай работа на всички според способностите им!
- Слоновете идват на тълпи!

www.shutterstock.com · 715430131

www.shutterstock.com · 496098325

Благодаря за вниманието!