Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

DATA SCIENCE

NDBI048

Introduction

https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

OUTLINE

- Motivation
- What is data science?
- Who is data scientist?
- Techniques and technologies
- Use cases
- Map of related courses
- Organization of the course

DATA

- Data affect life and work
  - Data leaks, cybercrimes, …
- Why it is so important now?
- History:
  - 19th century – industrial age (mechanical inventions) → goods, pollution
  - 20th century – informational age (machines gathering/storing data) → amounts of data
  - 21st century – HUGE amounts of data

FACEBOOK BY THE NUMBERS:
STATS, DEMOGRAPHICS & FUN FACTS
(LAST UPDATE: APRIL 2020)

- 2.5 billion monthly active users (2.989 billion in 2023)
- 5 billion comments are left on Facebook pages monthly
- 55 million status updates are made every day
- Every 60 seconds
  - 317,000 status updates
  - 147,000 photos uploaded
  - 54,000 shared links

https://www.omnicoreagency.com/facebook-statistics/

http://www.ibmbigdatahub.com/

WHAT IS BIG DATA?

Mobile devices
(tracking all objects all the time)

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Sensor technology and networks
(measuring all kinds of data)

Gartner: *“**Big Data**” is high **v**olume, high **v**elocity, and/or high **v**ariety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.*

IBM: *Depending on the industry and organization, **Big Data** encompasses information from internal and external sources such as transactions, social media, enterprise content, sensors, and mobile devices.*
Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

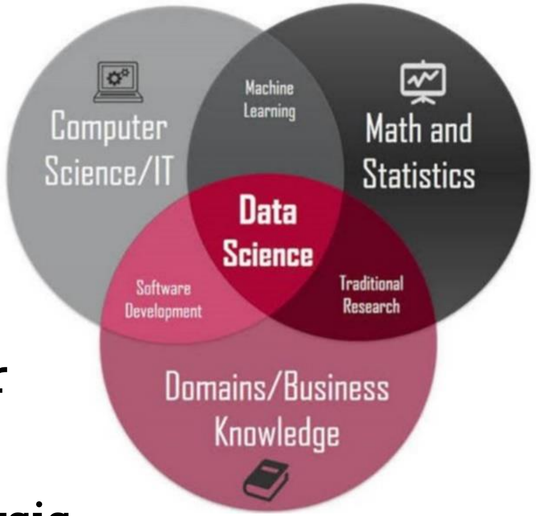
WHAT IS DATA SCIENCE?

- Data science is the art and science of acquiring knowledge through data
- Take data → use it → acquire knowledge → use it to:
  - Make decisions
  - Predict the future
  - Understand the past/present
  - Create new industries/products
- Different forms of data
  - On-line/off-line, real-time/past-time, organized/unorganized, missing/incomplete/wrong, different scales, …
- Data is cleaned → relationships are revealed

Wikipedia: Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

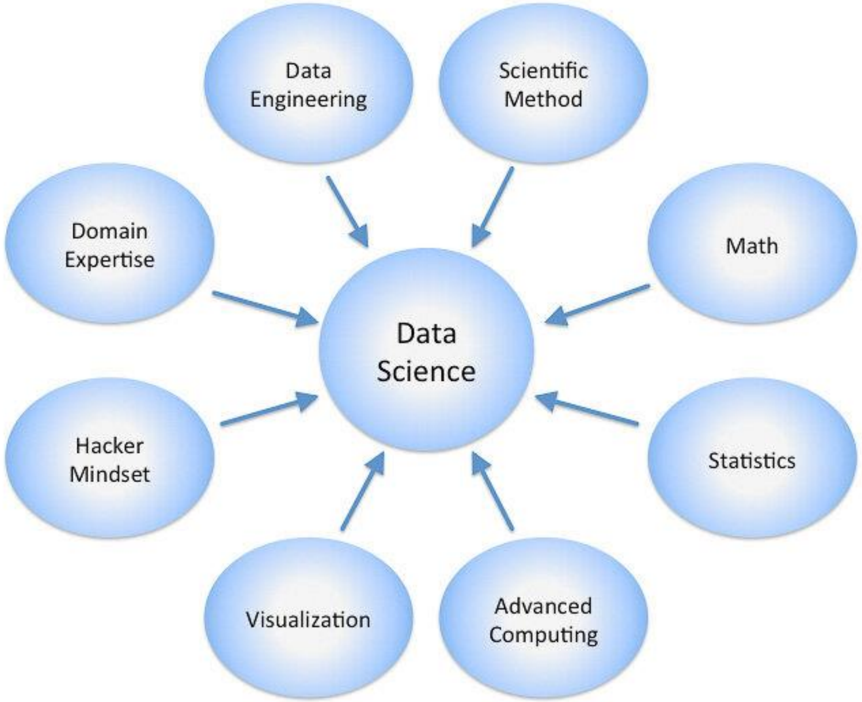
https://www.kdnuggets.com/2020/08/top-10-lists-data-science.html

WHAT IS DATA SCIENCE?

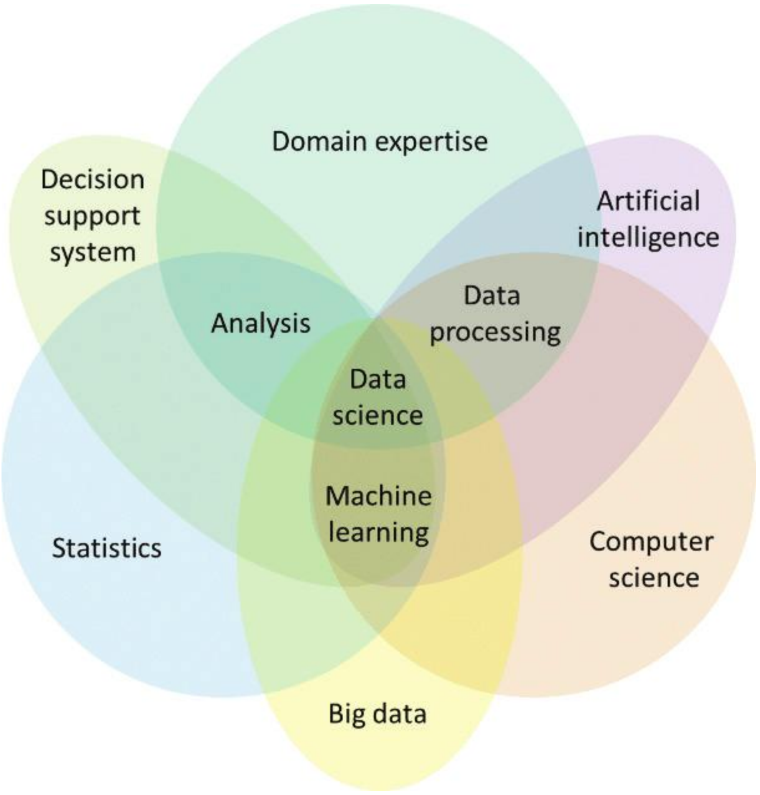


- **Computer science**: usage of code to create outcomes on the computer
  - Design and implement complex algorithms
- **Math and statistics**: usage of equations and formulas to perform analysis
  - Theorize and evaluate algorithms, tweak the existing procedures to fit specific situations
    - Formalization of relationships between variables
  - Statistics, probability, creation of models
    - Model = formal relationship between data simulating a real-world phenomenon
- **Domain knowledge**: understanding the problem domain
  - Medicine, finance, social science, …
    - Not just the one you know – data scientist can adapt
  - Apply concepts and results in a meaningful and effective way
- Difficult for a single person to master all three → teams of data scientists

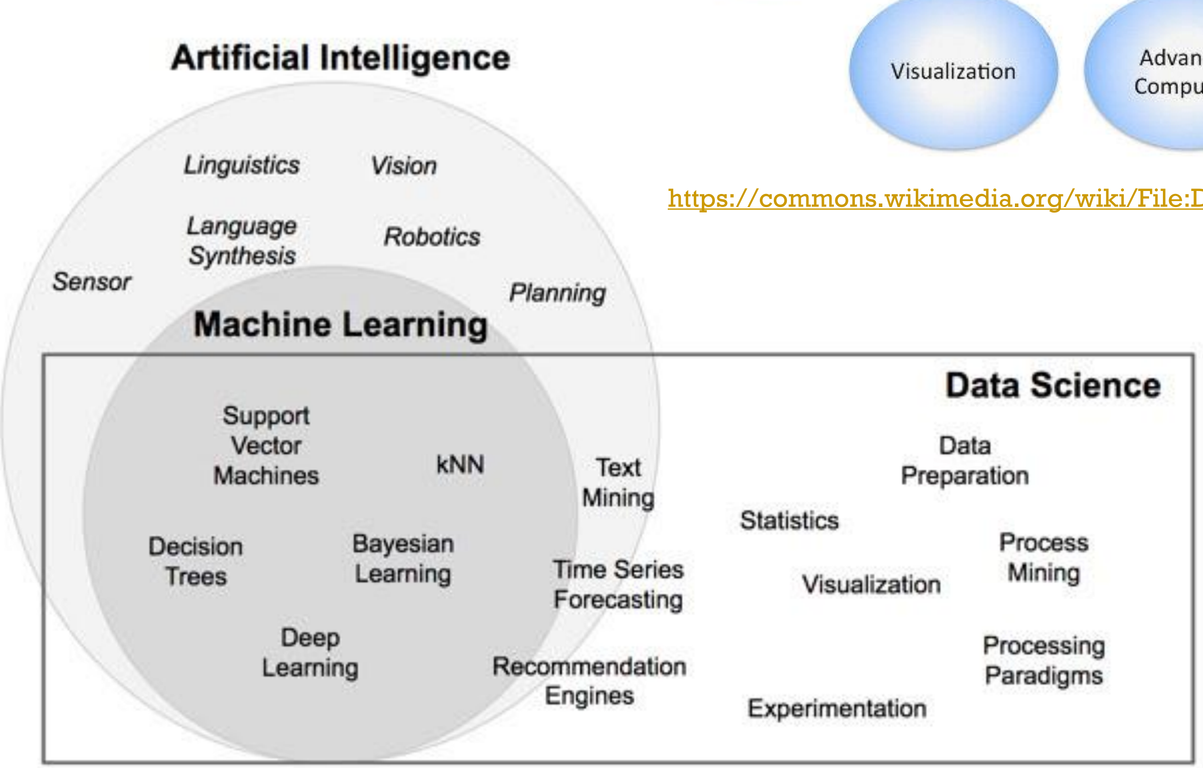
WHAT IS DATA SCIENCE?



https://commons.wikimedia.org/wiki/File:DataScienceDisciplines.png

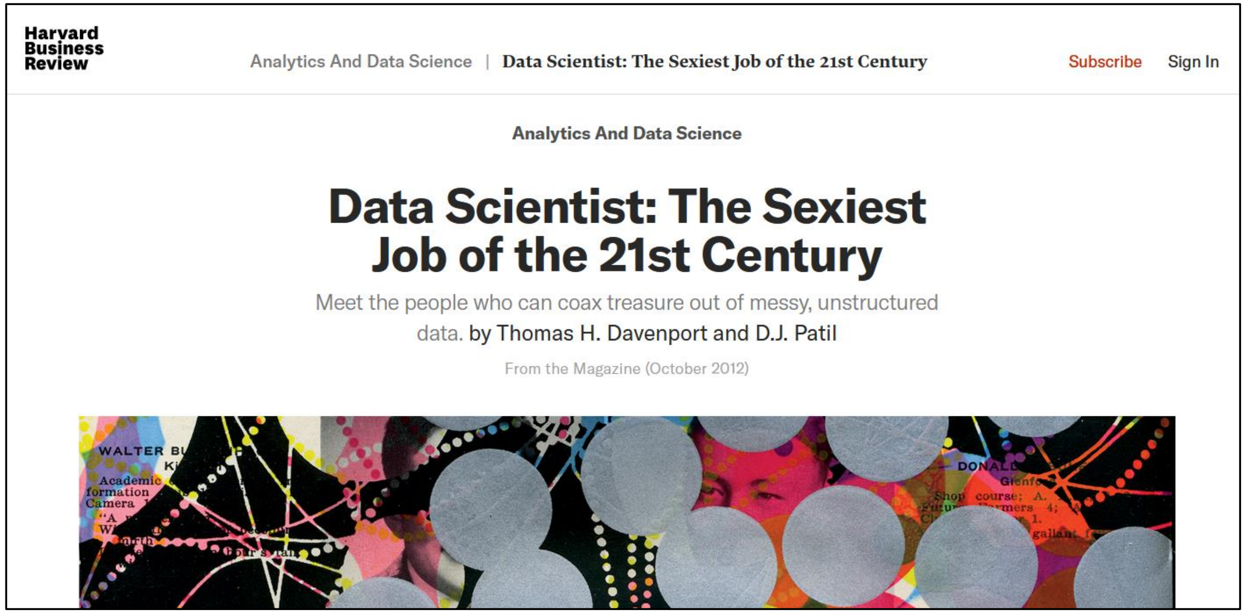


Palmer S. Data Science for the C-Suite. New York: Digital Living Press; 2015



http://www.introdatascience.com/course-slides.html

WHO IS DATA SCIENTIST?



Harvard Business Review

Analytics And Data Science | Data Scientist: The Sexiest Job of the 21st Century

Subscribe Sign In

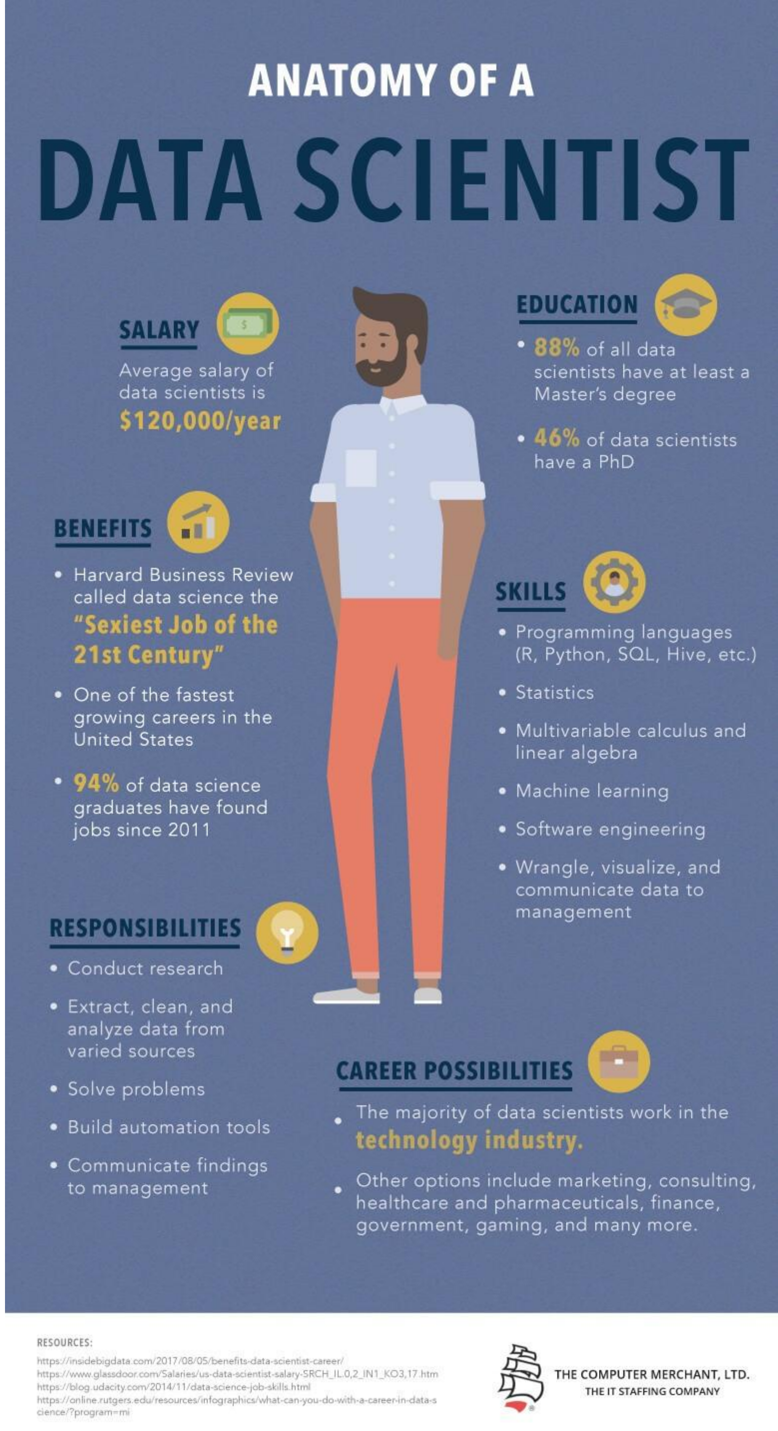
Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



ANATOMY OF A

DATA SCIENTIST

SALARY

Average salary of data scientists is **$120,000/year**

EDUCATION

- **88%** of all data scientists have at least a Master's degree
- **46%** of data scientists have a PhD

BENEFITS

- Harvard Business Review called data science the **"Sexiest Job of the 21st Century"**
- One of the fastest growing careers in the United States
- **94%** of data science graduates have found jobs since 2011

SKILLS

- Programming languages (R, Python, SQL, Hive, etc.)
- Statistics
- Multivariable calculus and linear algebra
- Machine learning
- Software engineering
- Wrangle, visualize, and communicate data to management

RESPONSIBILITIES

- Conduct research
- Extract, clean, and analyze data from varied sources
- Solve problems
- Build automation tools
- Communicate findings to management

CAREER POSSIBILITIES

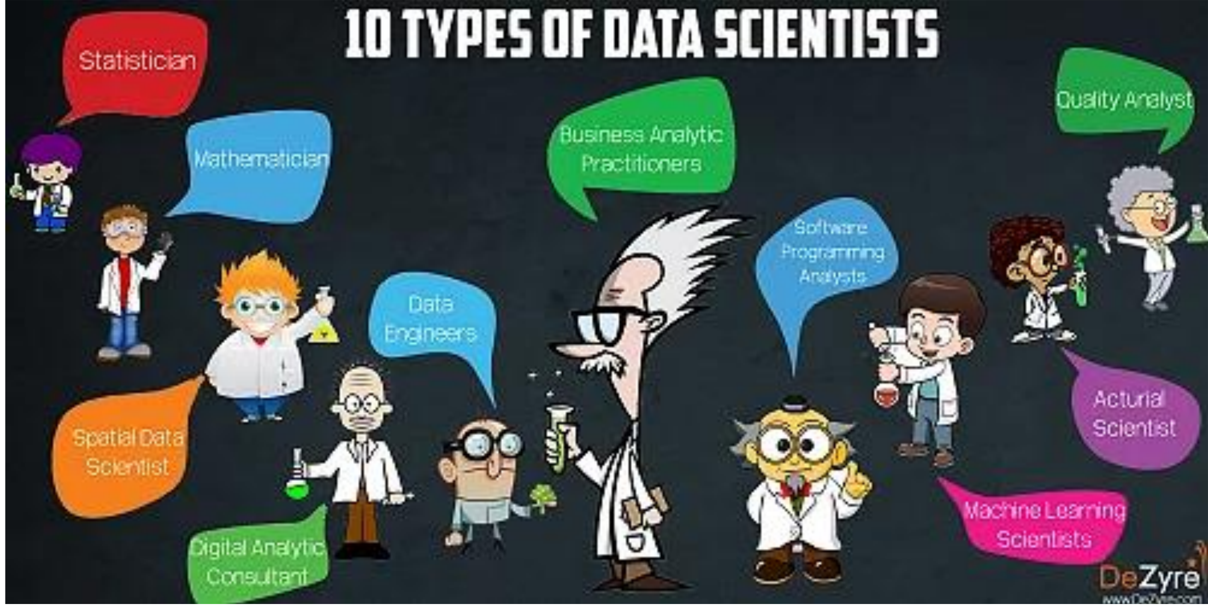
- The majority of data scientists work in the **technology industry.**
- Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.

RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
https://online.rutgers.edu/resources/infographics/what-can-you-do-with-a-career-in-data-science/?program=mi

THE COMPUTER MERCHANT, LTD.
THE IT STAFFING COMPANY

https://elearninginfographics.com/anatomy-data-scientist-infographic/

10 TYPES OF DATA SCIENTISTS


10 TYPES OF DATA SCIENTISTS
Statistician
Mathematician
Business Analytic Practitioners
Quality Analyst
Software Programming Analysts
Data Engineers
Acturial Scientist
Spatial Data Scientist
Digital Analytic Consultant
Machine Learning Scientists
DeZyre
www.DeZyre.com

- 400 different designations

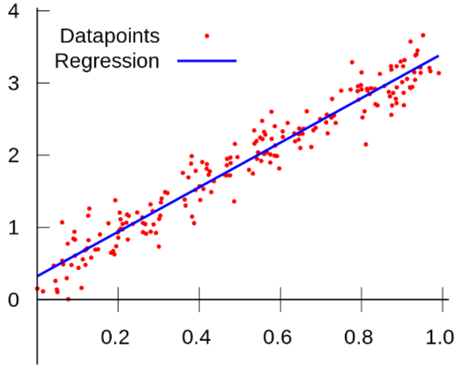
1. Data Scientist as Statistician
    - Traditional
    - Hypothesis testing, confidence intervals, analysis of variance, data visualization and quantitative research
2. Data Scientist as Mathematician
    - Deep knowledge of operations research and applied mathematics
3. Data Scientists Vs Data Engineers
    - Data engineer - responsibility to design, build and manage the information captured by an organization
    - Work closely together
4. Data Scientist as Machine Learning Scientist
    - AI, neural networks
5. Data Scientist as Actuarial Scientist
    - Apply mathematical and statistical models to BFSI (Banking, Financial Services and Insurance)

https://www.projectpro.io/article/10-different-types-of-data-scientists/179

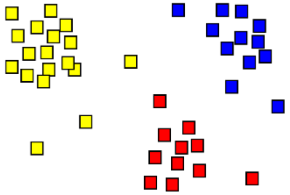
10 TYPES OF DATA SCIENTISTS

6. Data Scientist as Business Analytic Practitioner
   - „Sit“ between front end decision making teams and the back end analysts
   - ROI (Return On Investment) analysis, ROI optimization, dashboards design, performance metrics determination, high level database design
7. Data Scientist as Software Programming Analyst
   - Programming skills to automate routine big data related tasks to reduce computing time
   - Database and ETL (Extract Transform Learn) tools
8. Spatial Data Scientist
   - Special handling of spatial data (e.g., GPS coordinates)
9. Data Scientist as Digital Analytic Consultant
   - Configuring websites to collect data and direct it to analytics tools (e.g., Google Analytics), visualizing (filtering], processing and designing dashboards
10. Data Scientist as Quality Analyst
    - Large data sets to be analyzed to maintain quality control and meet minimum performance standards

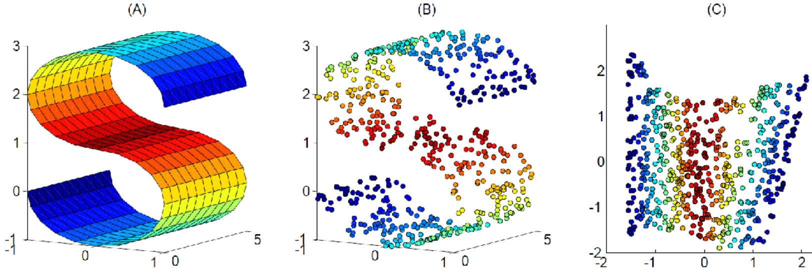
COMMON DATA SCIENCE TECHNIQUES



- Linear regression
  - Modelling the relationship between dependent (response) and independent (predictor) variables
- Logistic regression
  - Model the probability of a certain event (e.g., win/lose, healthy/sick)



- Decision trees
  - Classification and data fitting
- Support-vector machine
  - Supervised learning that analyze data for classification and regression analysis
- Cluster analysis
  - Group data together
- Dimensionality reduction
  - Reduce the complexity of data computation so that it can be performed more quickly.
- Machine learning
  - Inferencing patterns from data
- Naive Bayes classifiers
  - Classification by applying the Bayes' theorem



COMMON DATA SCIENCE TECHNOLOGIES

- **Programming Languages**: R, Python, Scala, Julia, SQL, Java, …
- **Data Modeling and Visualization Tools**: Scikit-learn, Pandas, Tableau, TensorFlow, Numpy, Mat plotlib, Shiny, D3, ggplot2, …
  - Support for statistics, data visualization, algorithms, data modeling, data analysis, …
- **Database Tools**: NoSQL, NewSQL, and relational databases
  - E.g., MySQL, Redshift, Hadoop, HBase, MongoDB, Cassandra, …
- **Big Data Tools**: Hadoop, Spark, Pig, Drill, Hive, Presto, …
  - Analyze data and provide a framework for processing and distributing large data.

DATA SCIENCE USE CASES (TELECOM)

- **Fraud detection**
  - Illegal access, authorization, theft or fake profiles, cloning, behavioral fraud, …
  - Apply machine learning algorithms to an immense amount of customer and operator data to spot the characteristics of normal traffic
  - Algorithms define the anomalies and with the help of data visualization techniques present them as alerts to the analysts
- **Predictive analytics**
  - Knowledge of customer preferences = better understanding of the customer
  - Uses historical data to build forecasts
- **Customer segmentation**
  - Segmentation of the market and targeting the content respectively
  - Telecommunication: customer value segmentation, customer behaviour segmentation, customer lifecycle segmentation, and customer migration segmentation

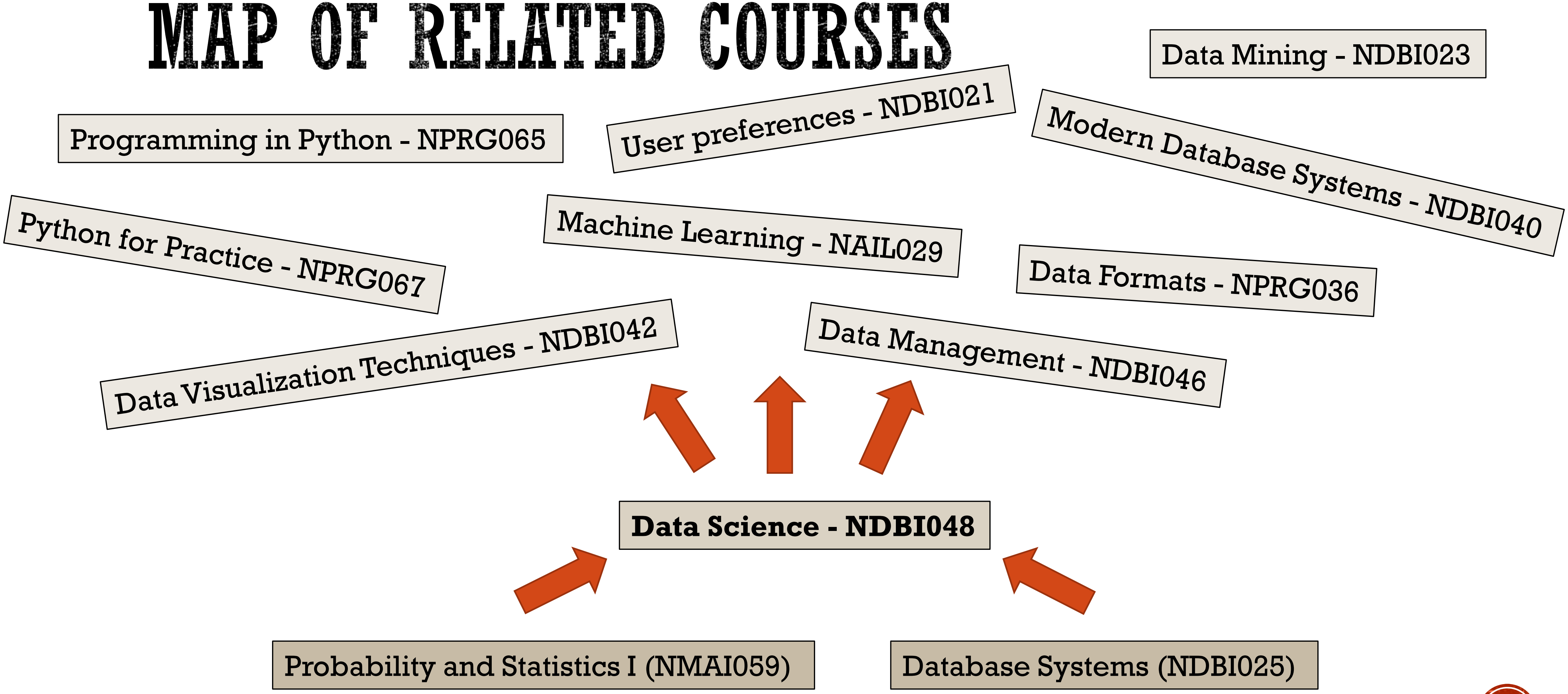
https://www.kdnuggets.com/2019/02/top-10-data-science-use-cases-telecom.html

DATA SCIENCE USE CASES

- **Customer churn prevention**
  - Insights concerning customers feelings about the services → immediate addressing the satisfaction-related issues and churn prevention
- **Lifetime value prediction**
  - Customers tend to search for better and cheaper services
  - Lifetime value = all the future profits and revenues generated by a customer
- **Network management and optimization**
- **Customer sentiment analysis**
- **Recommendation engines**
- **Price optimization**
- **…**

https://www.kdnuggets.com/2019/02/top-10-data-science-use-cases-telecom.html

MAP OF RELATED COURSES



Data Mining - NDBI023

Programming in Python - NPRG065

User preferences - NDBI021

Modern Database Systems - NDBI040

Python for Practice - NPRG067

Machine Learning - NAIL029

Data Formats - NPRG036

Data Visualization Techniques - NDBI042

Data Management - NDBI046

Data Science - NDBI048

Probability and Statistics I (NMAI059)

Database Systems (NDBI025)

ORGANIZATION OF THE COURSE

https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

REFERENCES

- Sinan Ozdemir: Principles of Data Science
- Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta: Practical Data Science Cookbook
- Frank Kane: Hands-On Data Science and Python Machine Learning