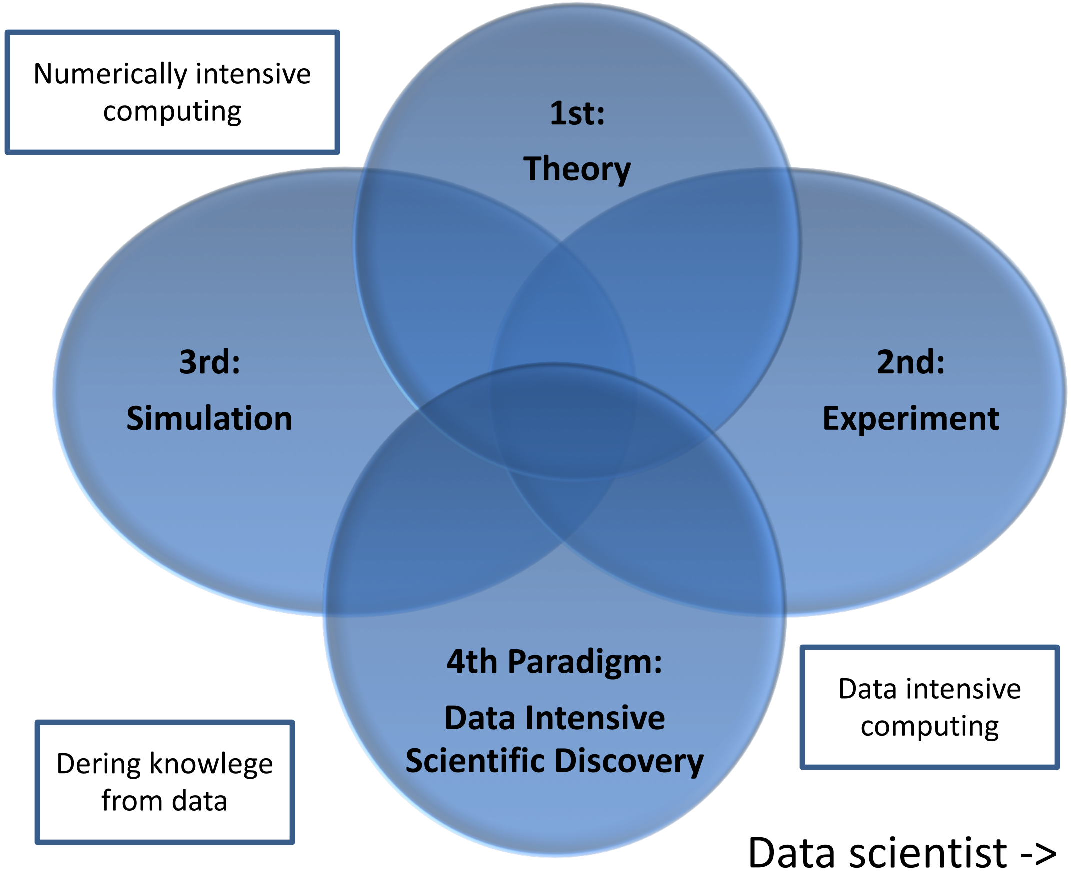
A Short Introduction to Data Science

Marian Bubak

Sano Centre of Computational Medicine
http://sano.science
and
AGH University of Science and Technology
Krakow, Poland
bubak@agh.edu.pl
http://dice.cyfronet.pl/

Theory, experiment, simulation, and data

Numerically intensive computing



Data intensive computing

Dering knowlege from data

Data scientist -> EDISON Project

Three laws

- **1971 – Moors's law – transistors;** being the doubling of microprocessing power roughly every two years.

- **1995 – Metcalfe's law – network volume**; which states that the value of a telecommunications network is proportional to the square of the number of connected users of the system (n^2).

- **Today – Watson's law – data and knowledge;** (not actually a law yet, this is hopeful IBM postulation and suggestion at this stage), which is the use of and application of AI in business, smart cities, consumer applications and life in general.

https://www.forbes.com/sites/adrianbridgwater/2018/03/20/ibm-ceo-rometty-proposes-watsons-law-ai-in-everything/#d4491224d087

The Fourth Paradigm – Data Intensive Scientific Discovery

- Talk by Jim Gray o the NRC-CSTB1 in Mountain View, CA, on January 11, 2007 http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

- The Fourth Paradigm – Data Intensive Scientific Discovery, Eds. Tony Hey, Stewart Tansley, and Kristin Tolle, Microsoft, 2009 https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf

- Albert-Laszlo Barabasi – The network science http://barabasi.com/

- Brian P. Schmidt - https://www.mso.anu.edu.au/~brian/ https://journals.aps.org/rmp/pdf/10.1103/RevModPhys.84.1151

- http://dice.cyfronet.pl/ ; http://sano.science

Big Data

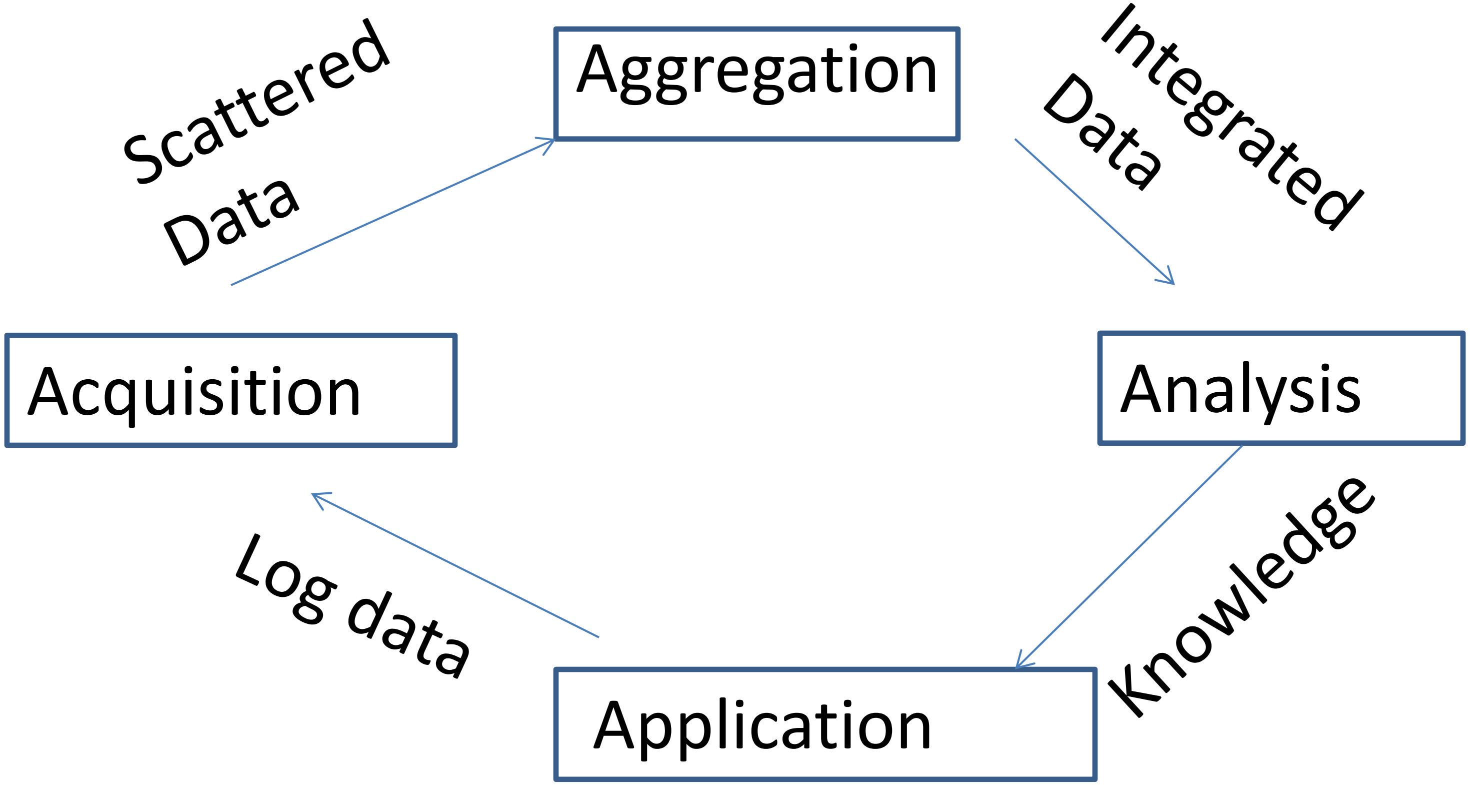
- Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization (Gartner, 2012)
- Complicated (intelligent) analysis of data may make a small data “appear” to be “big”
- Any data that exceeds our current capability of processing can be regarded as “big”

  - Big Data Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1 and related links

Big Data is important - examples

- Science
  - Large Synoptic Survey Telescope will create 140 TB every 5 day
  - Biomedical computation e.g. decoding human genome, personalized medicine
  - Social science
- Business
  - Facebook: 40 x 10^9 photos from users
  - Walmart: > 10^6 customer transactions every hour, imported into databases estimated to contain 2.5 PB
  - Falcon Credit Card Fraud Detection System protects 2.1 x 10^9 active accounts world-wide

Life cycle of Data, 4 A



Pedro Domingos, *The Master Algorithm*

https://homes.cs.washington.edu/~pedrod/

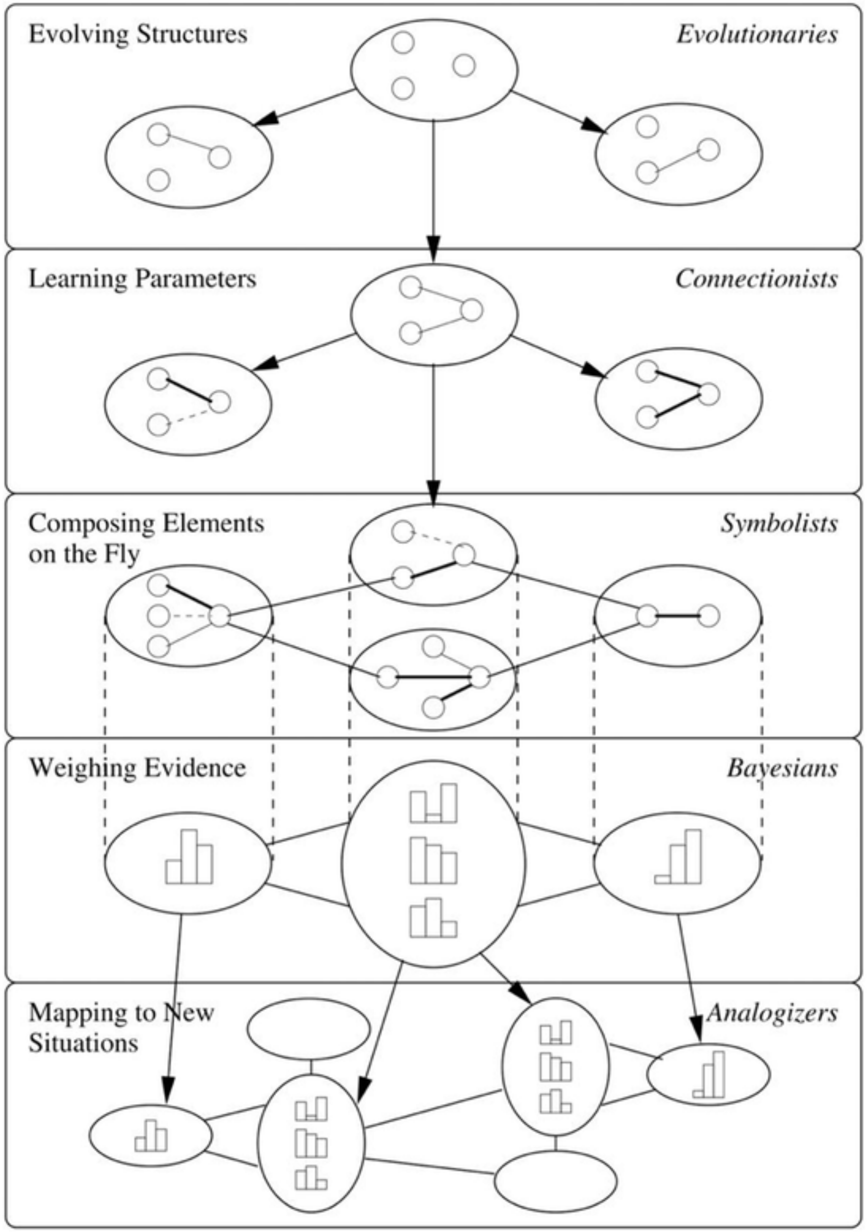
Central hypothesis: All knowledge - past, present and future - can be derived from data by a single, universal learning algorithm.

- All ML methods have implicit assumptions
- Make the assumptions explicit (Hume, "no free lunch")
- Evidence for a Master Algorithm: neuroscience, evolution, physics, statistics, computer science
- Machine learning versus knowledge engineering (Minsky, Chomsky, Fodor)

Machine learning allows computers to program themselves

- Give it the input and the desired output, out comes a program
- Just add data
- Simple methods allows to write complex programs

Five tribes of machine learning



Evolutionaries - nature's learning algorithm
•Evolutionary algorithms, crossover
•Can learn structure, wide hypothesis space
•Needs a way to ‘fill' the structure

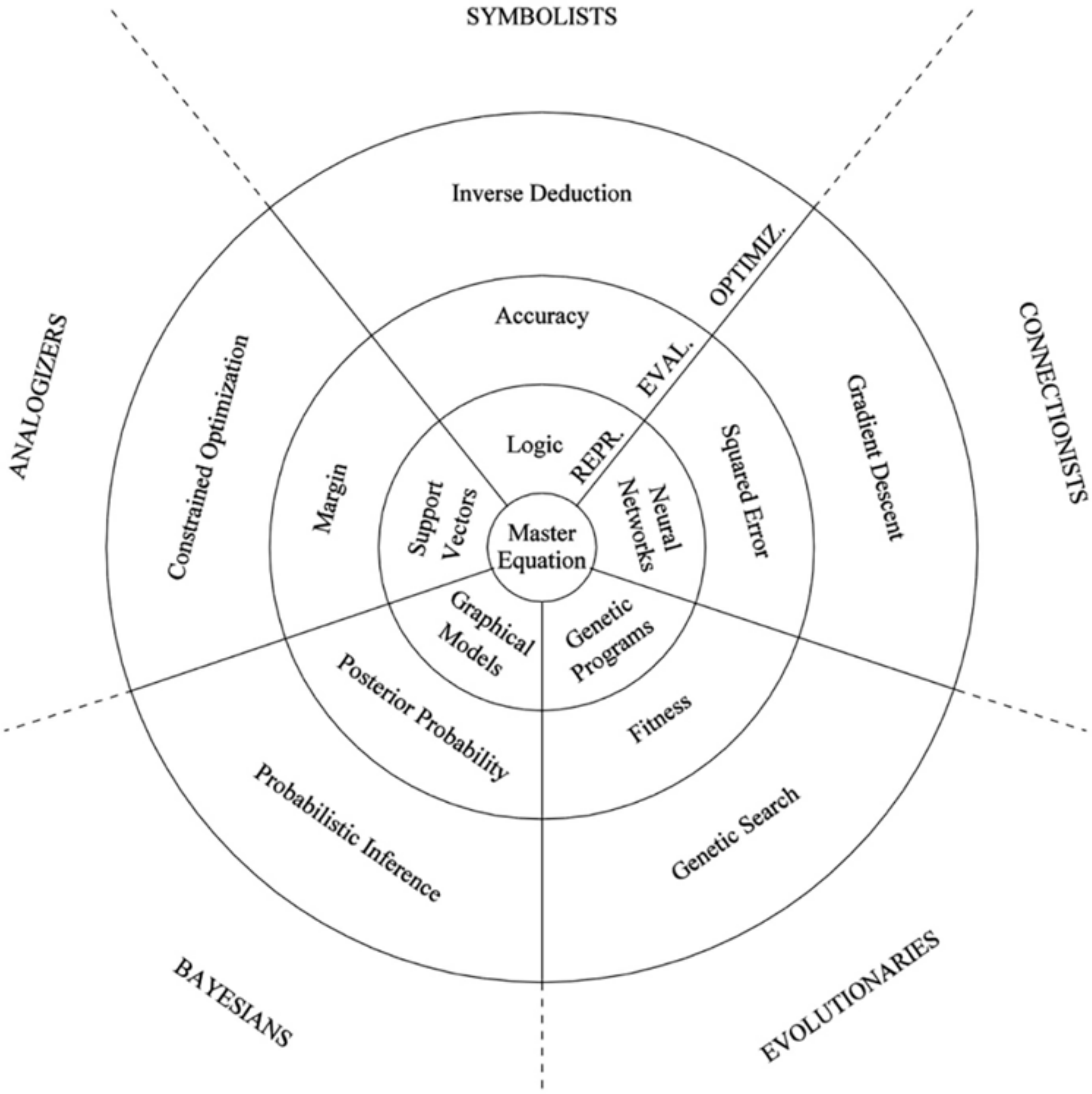
Connectionists - reverse engineer the brain
•Hebbs rule: neurons that fire together, wire together
•Neural networks, back propagation
•Good on signal processing
•Hard to add reasoning/explanations

Symbolists - using reasoning, rule based
•Logic, decision trees, inverse deduction
•Easy to add knowledge
•Impossible to code everything in rules

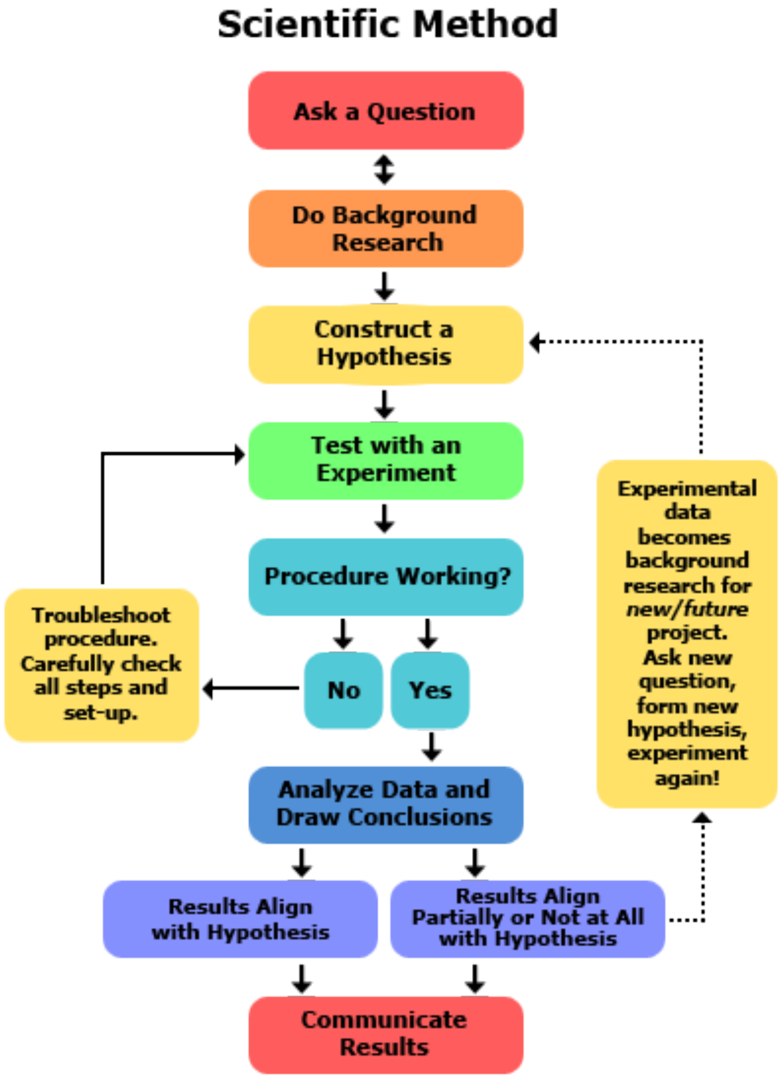
Bayesians - probabilities and statistics
•Bayesian networks, Kalman filter, Markov networks
•Bayes theorem, Probabilistic reasoning from first principles
•Hard to do unite logic and probability

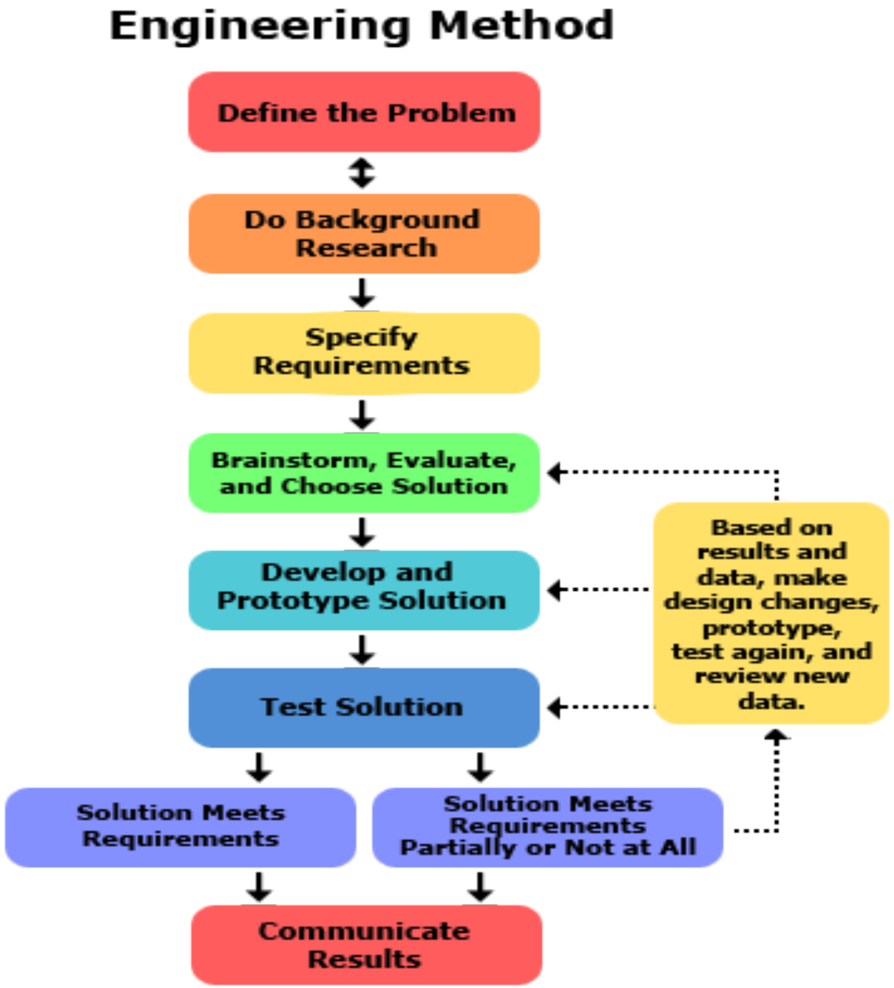
Analogizers - you are what you resemble
•(dis)similarity based
•kNN, SVM
•Analogy is powerful
•Hard to do rules and structure

Combining ensembles



Engineering and scientific methods


Scientific Method
Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Troubleshoot procedure. Carefully check all steps and set-up.
Analyze Data and Draw Conclusions
Results Align with Hypothesis
Results Align Partially or Not at All with Hypothesis
Experimental data becomes background research for *new/future* project. Ask new question, form new hypothesis, experiment again!
Communicate Results


Engineering Method
Define the Problem
Do Background Research
Specify Requirements
Brainstorm, Evaluate, and Choose Solution
Develop and Prototype Solution
Test Solution
Based on results and data, make design changes, prototype, test again, and review new data.
Solution Meets Requirements
Solution Meets Requirements Partially or Not at All
Communicate Results

The grand aim of science is to cover the greatest number of experimental facts by logical deduction from the smallest number of hypotheses or axioms. - Albert Einstein

Civilization advances by extending the number of important operations we can perform without thinking about them.
- Alfred North Whitehead

Basic steps of a simulation study

- **Problem definition** - goals of the study, what needs to be solved
- **Project planning** - work packages with a responsible parties; time and resources
- **System definition** - system components to be modeled and the performance measures to be analyzed
- **Model formulation** - understanding how the actual system behaves
- **Numerical model**
- **Numerical libraries an packages**
- **Input data collection & analysis**
- **Model translation, implementation** - the model is translated into a programming language
- **Verification & validation**
- **Experimentation & analysis** - alternative models, simulations, comparison with the real system
- **Documentation**

Lista przedmiotów

- Wprowadzenie do Data Science
- Statystyka
- Bazy danych
- Programowanie w języku Python
- Ekstrakcja danych ze źródeł internetowych
- Hurtownie danych
- Analiza dużych zbiorów danych w środowisku Spark
- Uczenie maszynowe
- Eksploracja danych
- Analiza danych tekstowych
- Sieci społeczne
- Analiza danych przestrzennych
- Wizualizacja dużych zbiorów danych
- Prawne aspekty analizy danych
- Seminarium - Projekty dyplomowy
- Projekt dyplomowy

https://informatyka.podyplomowe.agh.edu.pl/ds-detailed-ramowe-tresci

Literatura

- Big Data Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1 and related links
- EDISON: building the data science profession; EU Project
- Marcin Szeliga, Data Science i uczenie maszynowe, PWN, 2017
- Adam Zagdański, Artur Suchwałko, Analiza i prognozowanie szeregów czasowych, Praktyczne wprowadzenie na podstawie środowiska R, PWN, 2016

The final statement

Richard Feynman (The Feynman Lectures on Physics, Volume 3, Feynman's Epilogue):

"the powers of instruction are of very little efficacy except in those happy circumstances in which they are practically superfluous"

Wisława Szymborska: „ **Może to wszystko**”

Może to wszystko
dzieje się w laboratorium?
Pod jedną lampą w dzień
i miliardami w nocy?

Może jesteśmy pokolenia próbne?
Przesypywani z naczynia w naczynie,
potrząsani w retortach,
obserwowani czymś więcej niż okiem,
każdy z osobna
brany na koniec w szczypczyki?

Może inaczej:
żadnych interwencji?
Zmiany zachodzą same
zgodnie z planem?
Igła wykresu rysuje pomału
przewidziane zygzaki?

Może jak dotąd nic w nas ciekawego?
Monitory kontrolne włączane są rzadko?
Tylko gdy wojna i to raczej duża,
niektóre wzloty ponad grudkę Ziemi,
czy pokaźne wędrówki z punktu A do B?

Może przeciwnie:
gustują tam wyłącznie w epizodach?
Oto mała dziewczynka na wielkim ekranie
przyszywa sobie guzik do rękawa.

Czujniki pogwizdują,
personel się zbiega.
Ach cóż to za istotka
z bijącym w środku serduszkiem!
Jaka wdzięczna powaga
w przewlekaniu nitki!
Ktoś woła w uniesieniu:
Zawiadomić Szefa,
niech przyjdzie i sam popatrzy!

„Dekada Literacka” nr 6, s. 1 (KiP), 1992

New Scientist, 31 August 2016:

Could you be living inside a simulation created by a more advanced intelligence?
Where does your unerring belief that you are not come from?