

Data Science 101

Janis Keuper



INSTITUTE FOR MACHINE
LEARNING AND ANALYTICS



- What is “Data Science ?”
- Key Skills for Data Scientists
- Data Science Work Flow
- Python for Data Science
- Case Study I: predicting supermarket sales



Simple Question: What is Data Science?



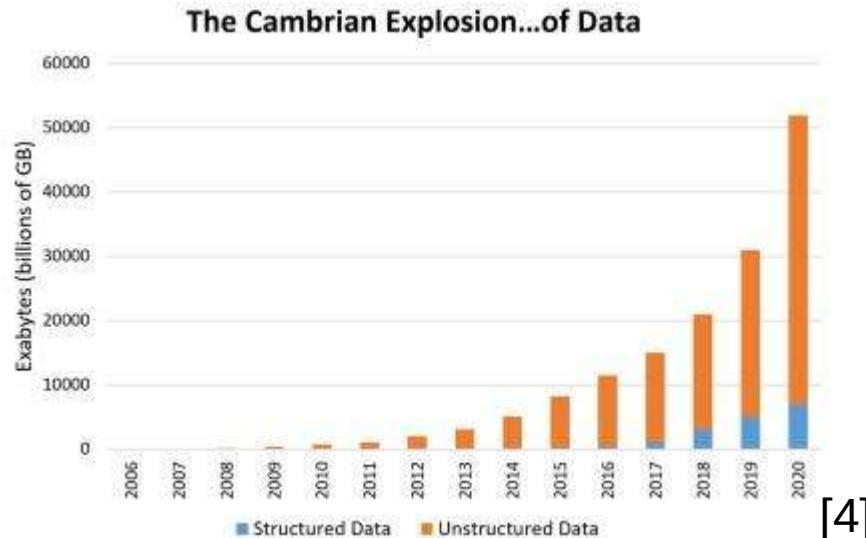
Simple Question: What is Data Science?

→ many different answers

(all right in some way)

What is Data Science?

Historic motivation



[4]

Multiplying Factor	SI Prefix	Scientific Notation	Name
1 000 000 000 000 000 000 000 000	Yotta (Y)	10^{24}	1 septillion
1 000 000 000 000 000 000 000 000	Zetta (Z)	10^{21}	1 sextillion
1 000 000 000 000 000 000 000	Exa (E)	10^{18}	1 quintillion
1 000 000 000 000 000 000	Peta (P)	10^{15}	1 quadrillion
1 000 000 000 000	Tera (T)	10^{12}	1 trillion
1 000 000 000	Giga (G)	10^9	1 billion
1 000 000	Mega (M)	10^6	1 million
1 000	kilo (k)	10^3	1 thousand
0.001	milli (m)	10^{-3}	1 thousandth
0.000 001	micro (u)	10^{-6}	1 millionth
0.000 000 001	nano (n)	10^{-9}	1 billionth
0.000 000 000 001	pico (p)	10^{-12}	1 trillionth
0.000 000 000 000 001	femto (f)	10^{-15}	1 quadrillionth
0.000 000 000 000 000 001	atto (a)	10^{-18}	1 quintillionth
0.000 000 000 000 000 000 001	zepto (z)	10^{-21}	1 sextillionth
0.000 000 000 000 000 000 000 001	yocto (y)	10^{-24}	1 septillionth

[4]

Historic motivation

Leads to many new questions:

- *Where is this data coming from?*
- *What are we going to do with it?*
- *Who is going to do it?*

Historic motivation

Leads to many new questions:

- *Where is this data coming from?*
 - *Internet of Things (IoT: sensors every where)*
 - *Industry 4.0 (connecting everything)*
 - *Big Data (storing all the information you can in the hope generate gain)*
 - *Social Media (and other internet services)*
- *What are we going to do with it?*
- *Who is going to do it?*

Historic motivation

Leads to many new questions:

- *Where is this data coming from?*
 - *Internet of Things (IoT: sensors every where)*
 - *Industry 4.0 (connecting everything)*
 - *Big Data (storing all the information you can in the hope generate gain)*
 - *Social Media (and other internet services)*
- *What are we going to do with it?*
 - *Analyze the past*
 - *Predict the future (at least try to)*
 - *Machine Learning (data driven computing)*
 - *AI (what ever that is)*

Historic motivation

Leads to many new questions:

- **Where is this data coming from?**
 - *Internet of Things* (IoT: sensors every where)
 - *Industry 4.0* (connecting everything)
 - *Big Data* (storing all the information you can in the hope generate gain)
 - Social Media (and other internet services)
- **What are we going to do with it?**
 - Analyze the past
 - Predict the future (at least try to)
 - *Machine Learning* (data driven computing)
 - *AI* (what ever that is)

Biggest Buzz-Words in IT

Historic motivation

Leads to many new questions:

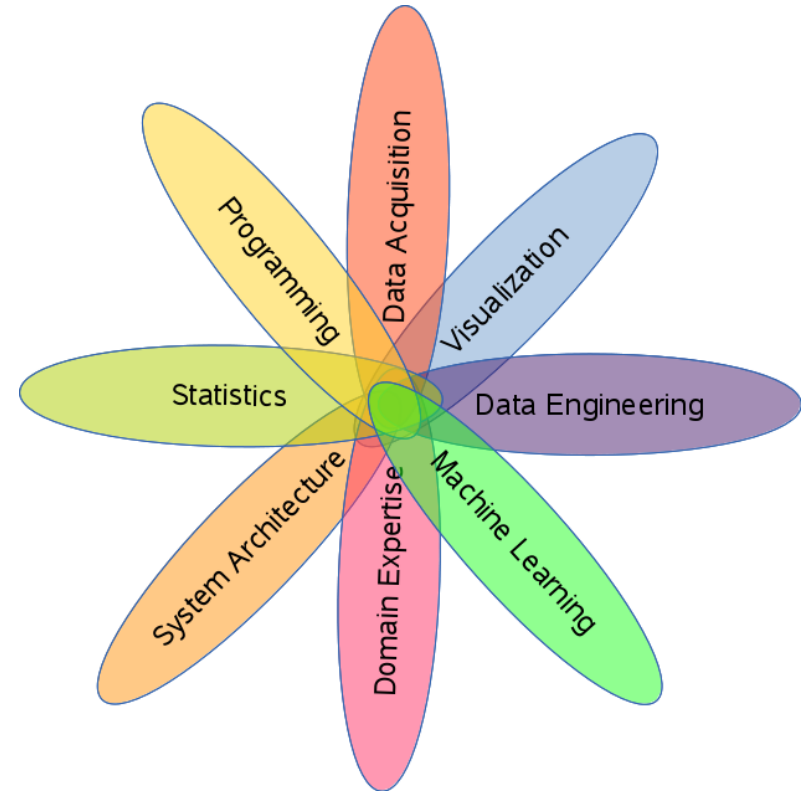
- *Where is this data coming from?*
- *What are we going to do with it?*
- *Who is going to do it? → **Data Science !***

What is Data Science?

Motivation by tasks:

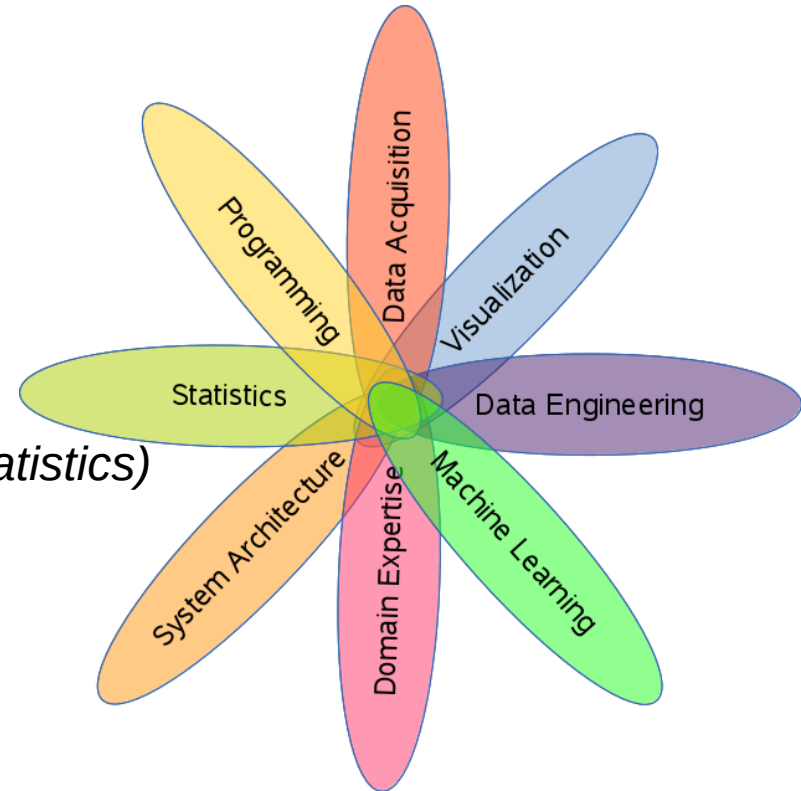
What is expected of Data Science?

What are typical tasks / applications?



Typical Data Science Tasks:

- *Data Acquisition*
- *Data cleaning*
- *Storing data and making it accessible*
- *Selecting and pre-processing data*
- *Structuring and describing data (Descriptive Statistics)*
- *Exploration and Inference (Learning)*
- *Modeling and Abstraction*
- *Verification*
- *Visualization and Reporting*



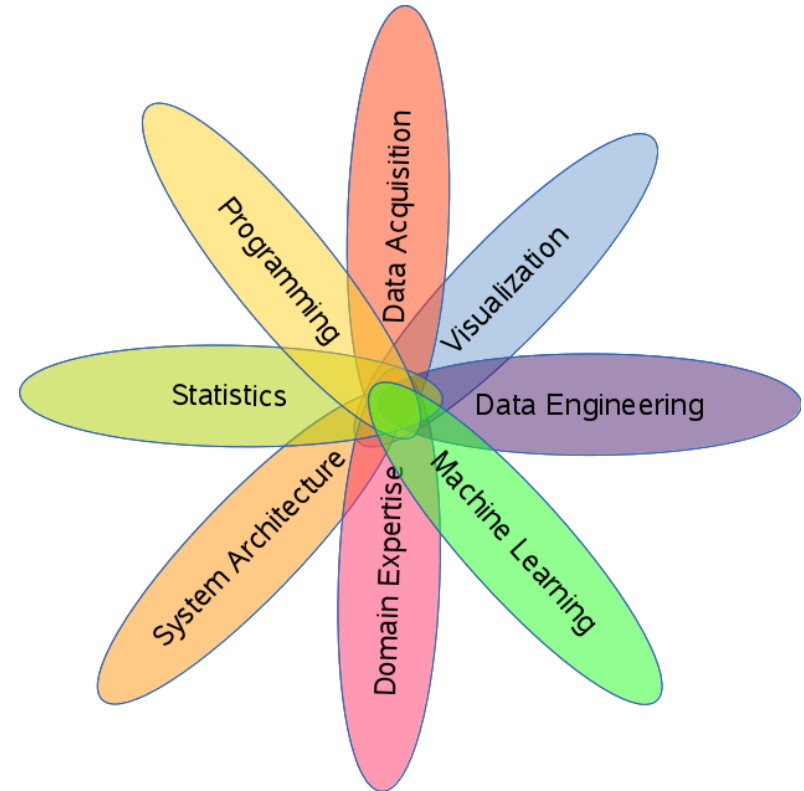
What is Data Science?

Motivation by tasks:

What is expected of Data Science?

What are typical tasks / applications?

→ **Data Science as mesh-up of many tasks and skills**



Data Science – Science of data?

Definition of Data:



Wikipedia:

Data (*/ˈdɜrtə/ DAY-tə, /ˈdætə/ DA-tə, or /ˈdɑ:tə/ DAH-tə*) is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

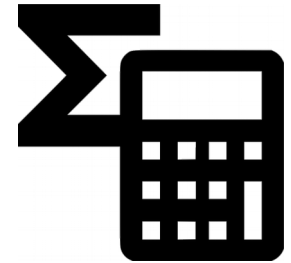
Scientific perspectives at data:

- **Looking at Data like a Statistician, you see the world in terms of**
 - Variables (sets)
 - (joint) Distributions
 - Densities
 - Correlations
 - Likelihoods
 - Hypothesis



Scientific perspectives at data:

- **Looking at Data like a Mathematician, data forms**
 - Structures (Algebras, Sets, Groups, Graphs ...)
 - Spaces (i.e. high-dimensional Euclidean space)
 - Change (i.e. expressed in differential equations)
 - in practice: a lot of numerical optimization
- **Looking at Data like a Computer Scientist [Programmer]**
 - you think in data structures
 - algorithms and their efficiency
 - levels of abstraction
 - use cases (reusability)
 - parallelization



Scientific perspectives at data:

- **Looking at Data like a Data Engineer, you see**

- Databases
- Storage
- System Architectures



- **Looking at Data like a Hacker, you think about**

- how to get the data (from devices, web databases, streams ...)
- how to extract the data (from binaries, proprietary formats ...)
- how to use data in unconventional ways



Scientific perspectives at data:

- **Looking at Data like a Visual Artist, you think about**

- how to visualize the key information hidden in the data
- how to depict change
- how to setting data into relation
- how to present complex problems



- **Looking at Data as a Domain Expert**

- Domain knowledge is important to set data in the correct meta-relation
 - i.e. helps to ask the "right" questions
 - helps to avoid trivial miss-interpretations (due to missing meta data/information)
- But, domain knowledge also might be misleading and preventing data-driven discoveries



Just be:

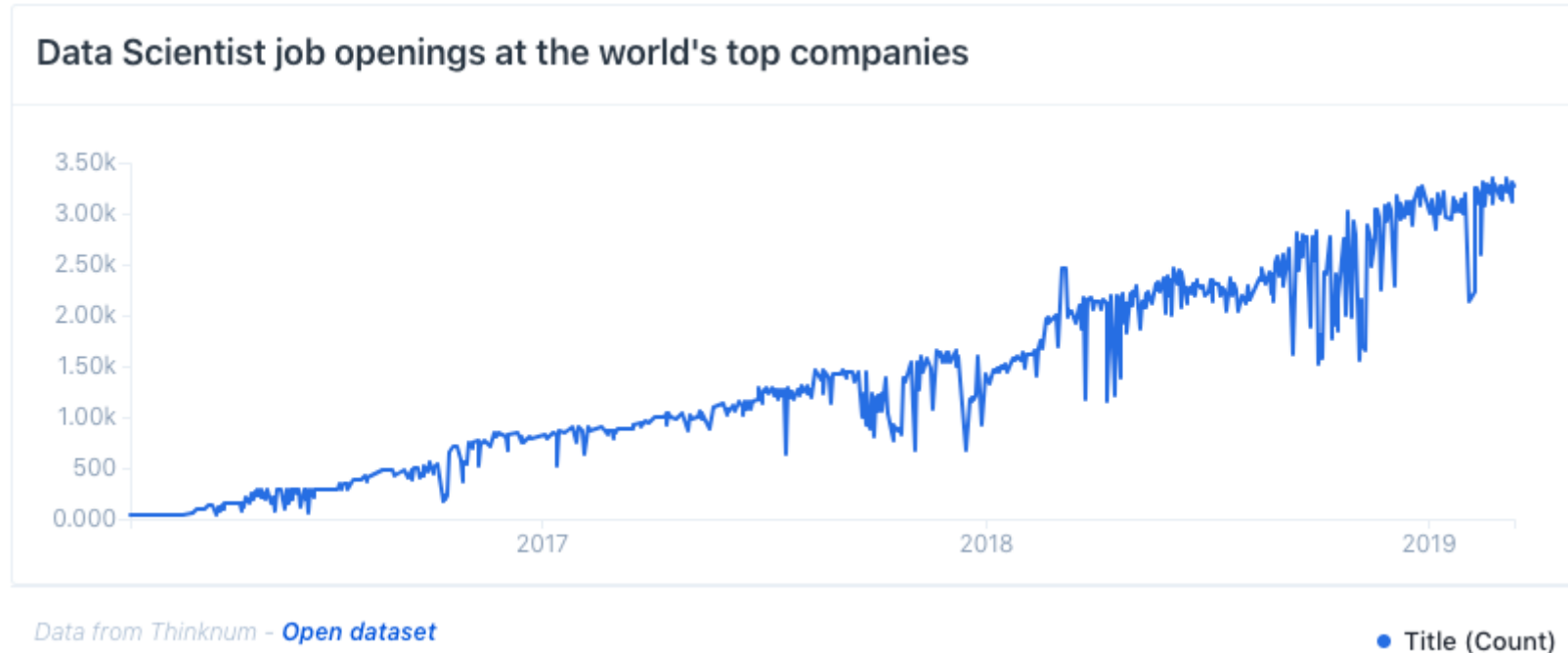
- A Statistician
- Mathematician
- Programmer
- Engineer
- Hacker
- Artist
- And Expert in many domains :-)

On top of that:



- Very good communication skills
 - Presenting your work
 - Getting information from domain experts
- Be a quick learner
 - Adaption to new domains
 - Fast moving field (e.g. machine learning)

Job Opportunities Data Science



<https://media.thinknum.com/articles/massive-increase-in-demand-for-data-science-jobs-in-2019/>

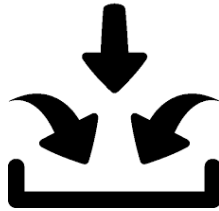
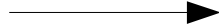


Start: Hypothesis
Idea
Problem setting

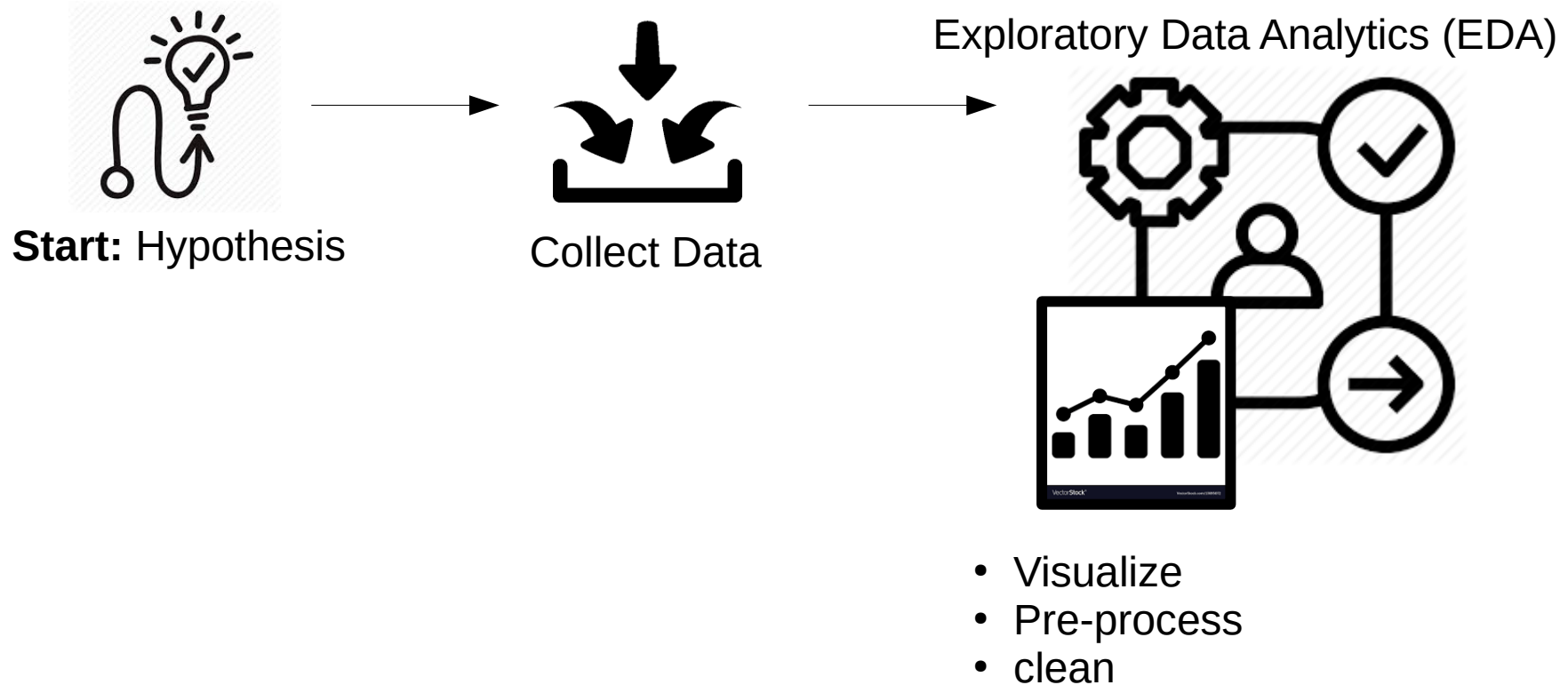
Need: Basic understanding of the
problem → Domain knowledge

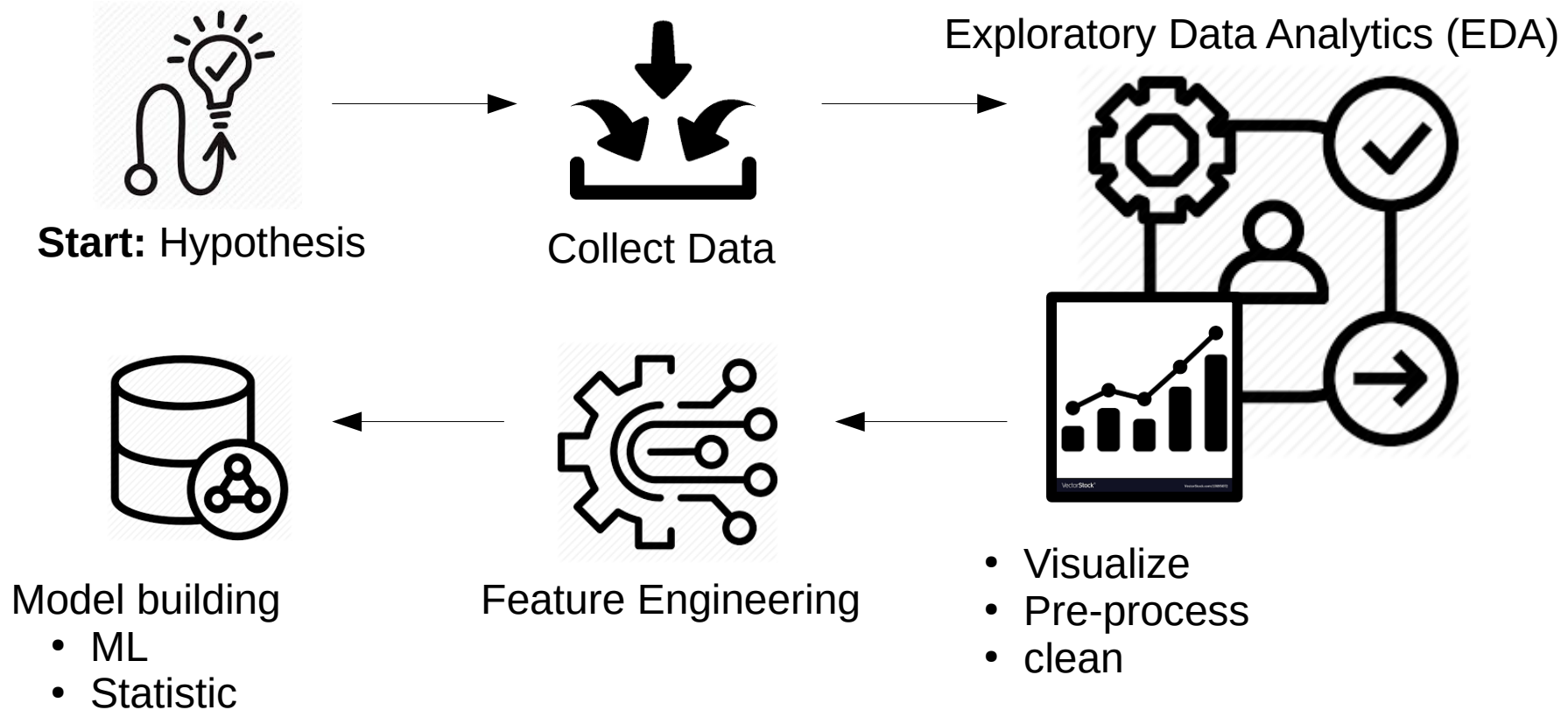


Start: Hypothesis



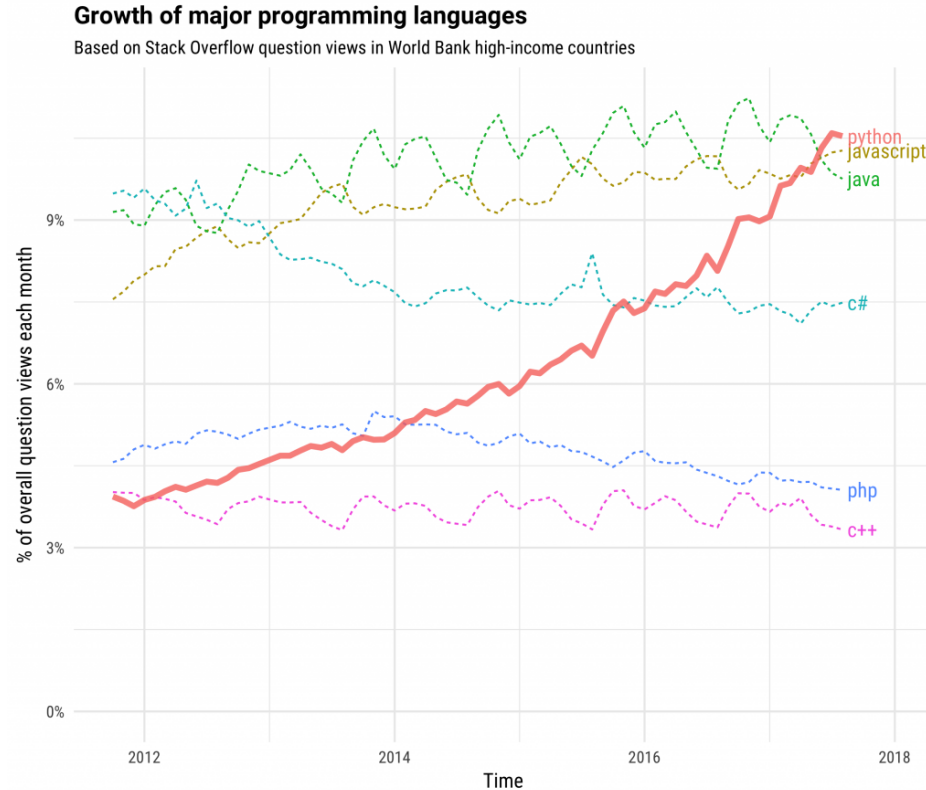
Collect Data





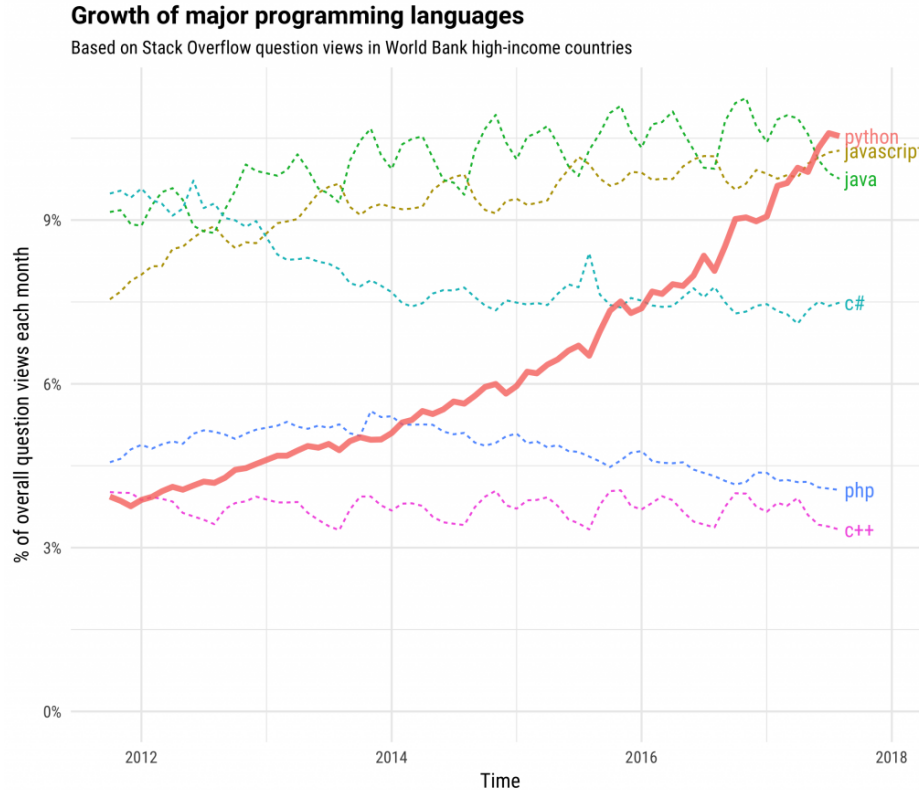


- We will use Python (more exactly tools with python interface) for this course
- Python is the most dominant language used in Data Science, Data Analytics and Machine Learning today
- Alternatives: R, Julia and Matlab



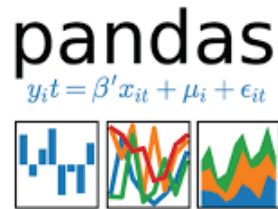
[2]

- Triggered by ML + Data Science
- Python pushes convergence of technologies:
 - HPC
 - Cloud
 - Big-Data



[2]

- Very low entry barrier (easy to learn)
- Quite universal in programming approaches
- Easy to interface existing high performance libs
- Huge Community
- **Libraries !!!**
- **Python as „glue code“ for rapid development and markup**





<https://www.anaconda.com/distribution/>

Install on you local computer

- Easy installation on Linux/Mac/Win
 - See Howto in Moodle
 - User space install (no admin needed)
- Conda Python package management
 - Simple install/update/sharing of software
- All open source!



NOTE:

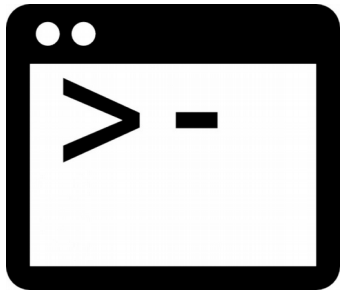
There are two versions of Python:

Python2 and Python3

Support for Python2 ended 31.12.19

→ get started with Python3!





Interactive Python Introduction

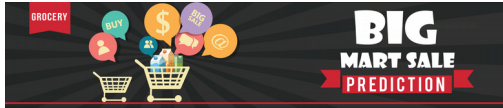
→ Lab session Friday



Case Study I: predicting sales

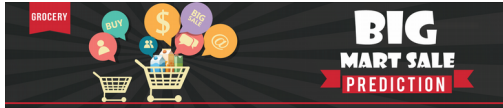
→ a Data Science competition by





Problem Formulation

Bigmart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a **model that can predict the sales, per product, for each store**. BigMart has collected sales data from the year 2013, for **1559 products across 10 stores in different cities**. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business.

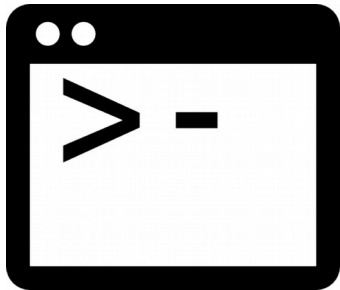


Problem Formulation

Bigmart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a **model that can predict the sales, per product, for each store**. BigMart has collected sales data from the year 2013, for **1559 products across 10 stores in different cities**. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business.

Discussion:

- What are the goals ?
- What data is supplied ?
- What expert knowledge is needed ?



Case Study I: predicting sales

→ demo on Colab

[1] free icons taken from <https://www.flaticon.com>

[2] <https://stackoverflow.blog/2017/09/06/incredible-growth-python/>

[3] <https://datahack.analyticsvidhya.com/contest/practice-problem-bigmart-sales-prediction/>

[4] https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462#