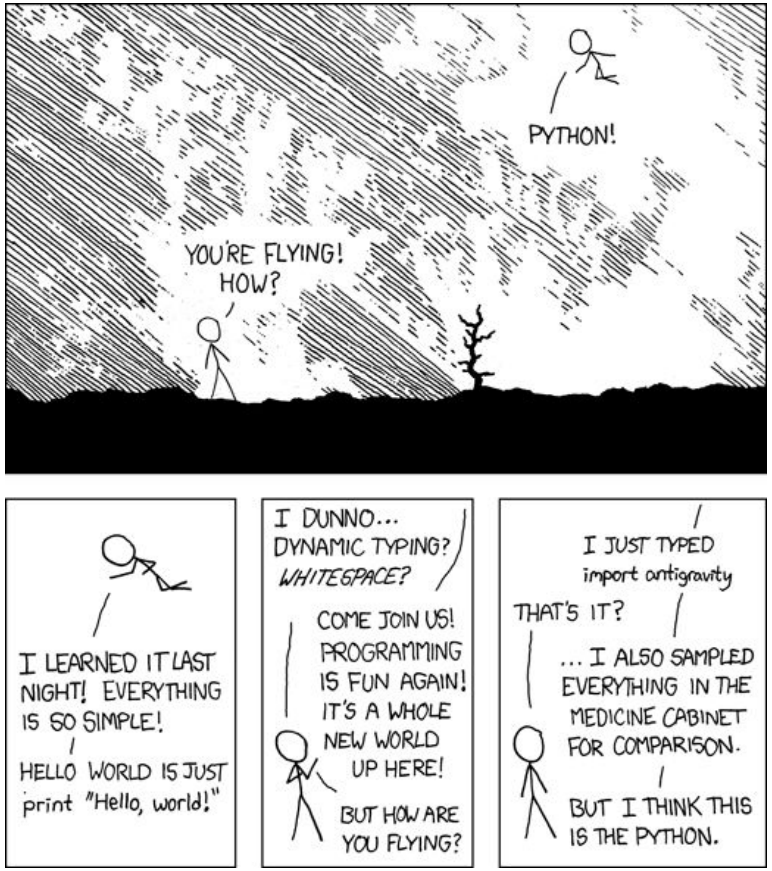
UMM
UNIVERSITÄTSMEDIZIN
MANNHEIM

Medizinische Fakultät Mannheim
der Universität Heidelberg

Universitätsklinikum Mannheim

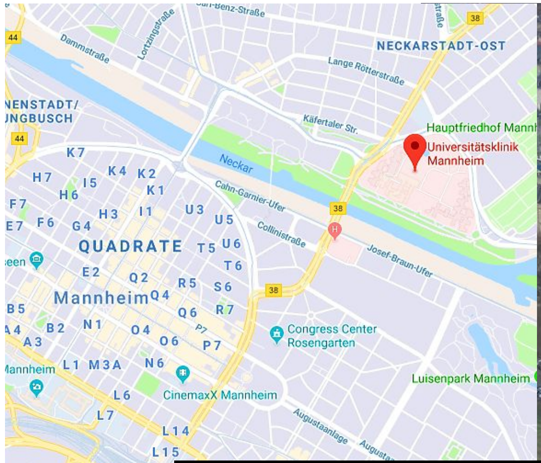
Python for Big Data Analysis


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

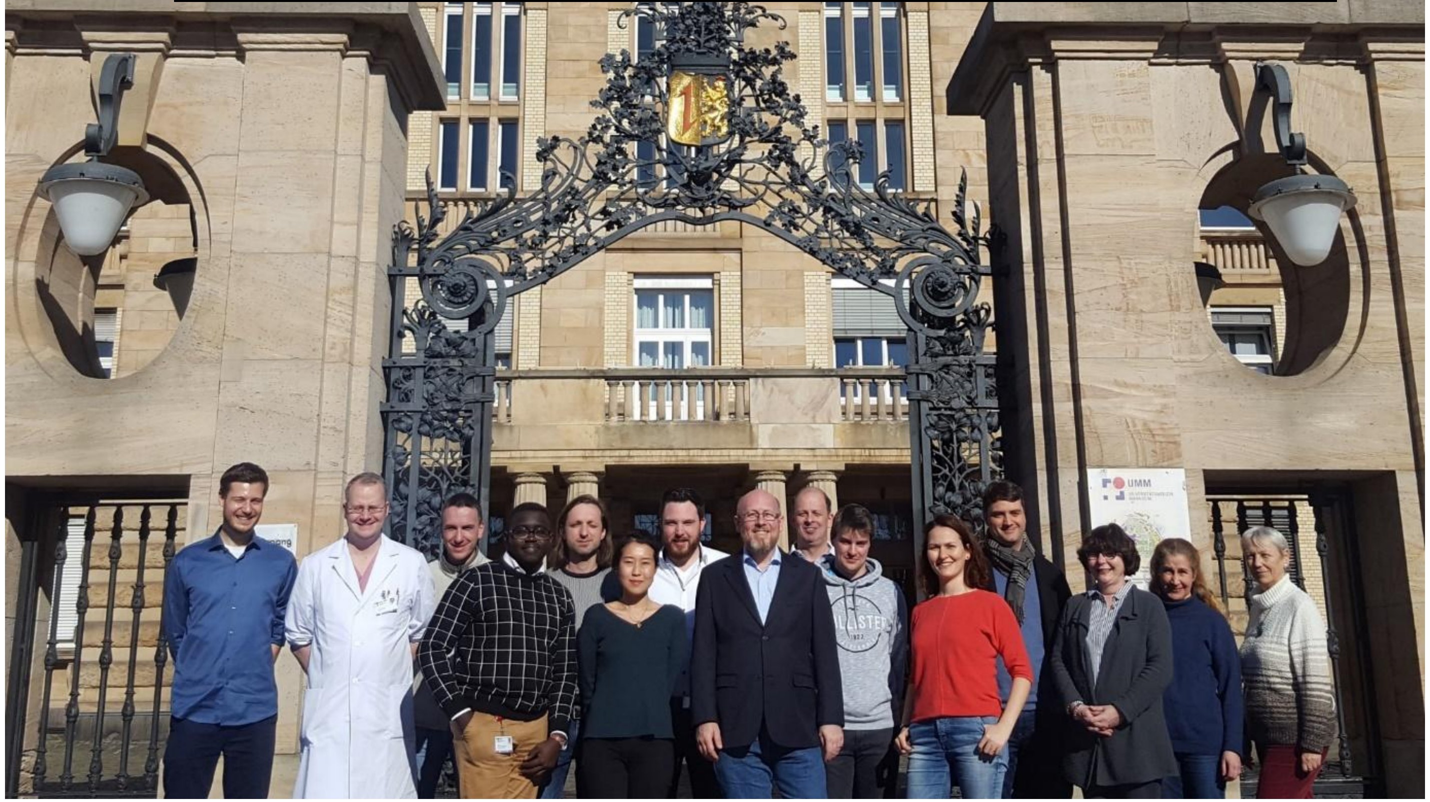
Kim Hee
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health

About us

- **Kim Hee - PhD candidate since 2016**
  - Computer engineering background
  - Research interest: microbiology image data analysis with consideration of scalability
  - Responsibility: ETL for laboratory data

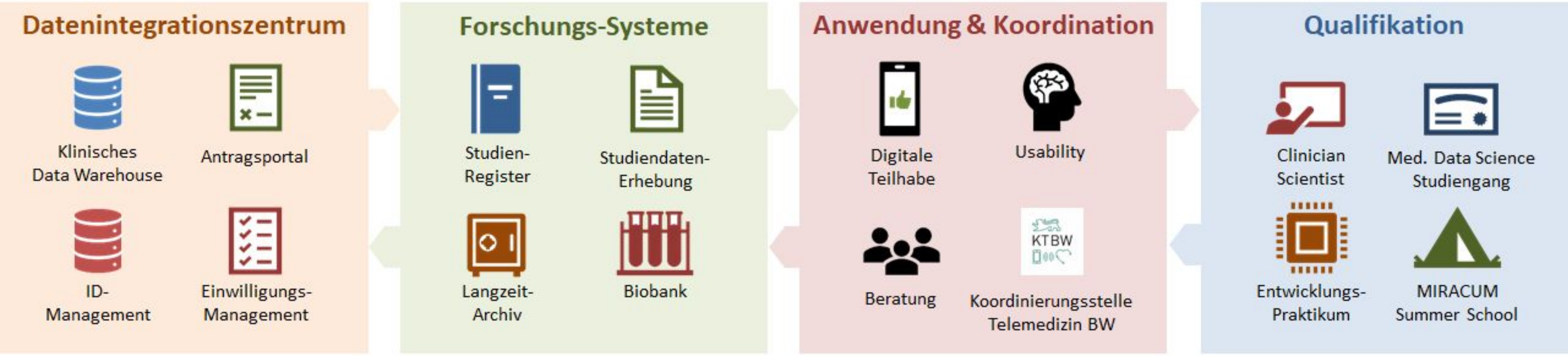


Heinrich-Lanz-Zentrum für Digitale Gesundheit



Heinrich-Lanz-Center (HLZ) for Digital Health

- **Our lab is currently active in the following areas:**



Datenintegrationszentrum

Klinisches Data Warehouse

Antragsportal

ID-Management

Einwilligungs-Management

Forschungs-Systeme

Studien-Register

Studiendaten-Erhebung

Langzeit-Archiv

Biobank

Anwendung & Koordination

Digitale Teilhabe

Usability

Beratung

Koordinierungsstelle Telemedizin BW

Qualifikation

Clinician Scientist

Med. Data Science Studiengang

Entwicklungs-Praktikum

MIRACUM Summer School

Versorgungs-IT

MEDIZIN INFORMATIK INITIATIVE

KOMPETENZZENTRUM FORSCHUNGSDATEN EIN DIENST VON URZ

Metropolregion Rhein-Neckar

TMF

hochschule mannheim

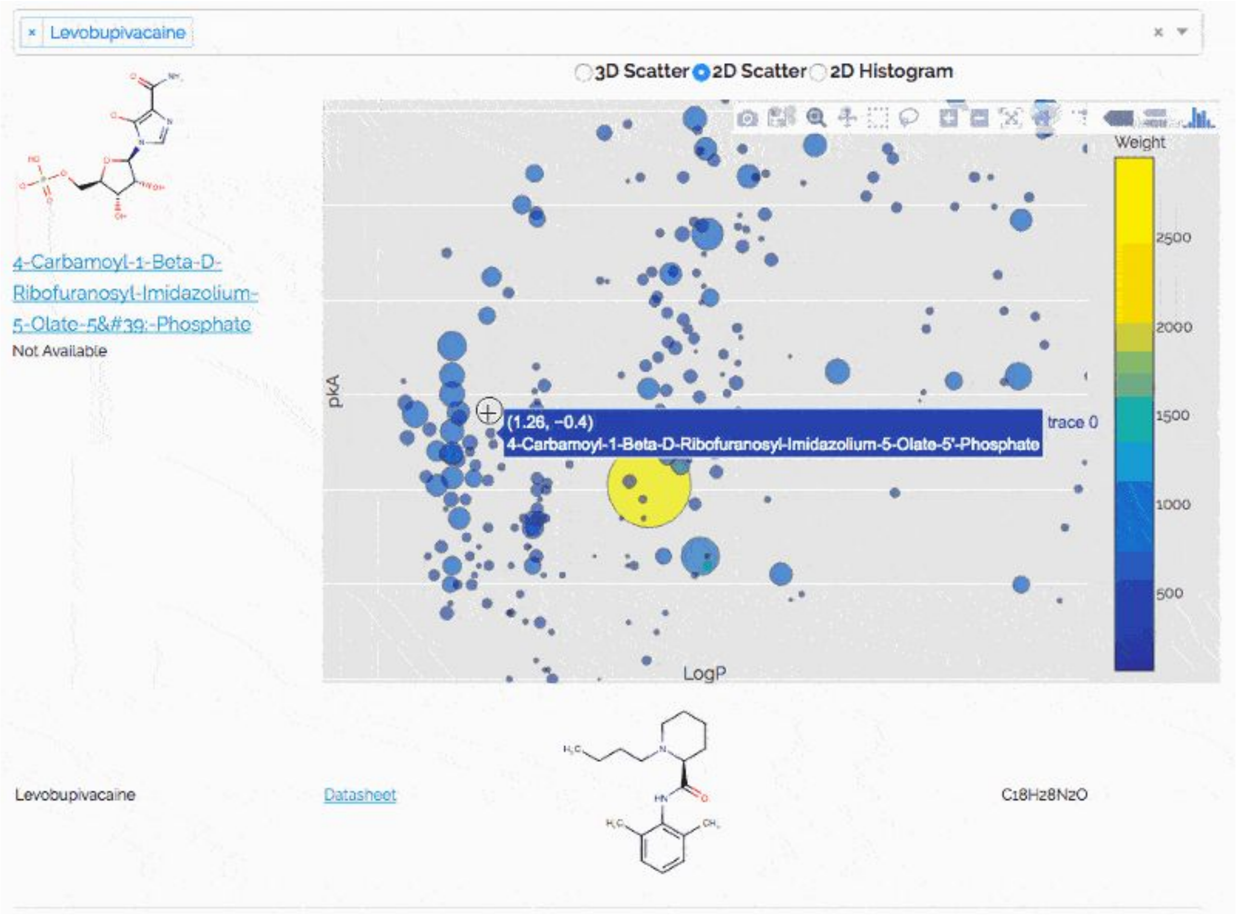
miracum

Challenges in digital healthcare domain

- **Data silos across and within the healthcare providers**
- **Premature status of interoperability and data standards**
- **Data quality is low (ex: completeness, consistency, accuracy, timeliness and more)**
- **In-house database management system is not ready for data analysis**
- **Not many successful healthcare applications for patients**
- **Not enough study programs**
- **Legal issues**
- **and more**

In-house database management system is for neither cyclic data analyse nor visualization

- In-house database is built for the billing purpose
- Data analysis with a conventional database management system (CDBMS) is possible, but not optimal because too many I/Os occur and memory is not properly utilized for data analysis
- Data visualization with SQL is not possible


× Levobupivacaine
○3D Scatter ●2D Scatter ○2D Histogram
4-Carbamoyl-1-Beta-D-Ribofuranosyl-Imidazolium-5-Olate-5&#39;-Phosphate
Not Available
pKA
LogP
Weight
2500
2000
1500
1000
500
(1.26, −0.4)
4-Carbamoyl-1-Beta-D-Ribofuranosyl-Imidazolium-5-Olate-5'-Phosphate
trace 0
Levobupivacaine
Datasheet
C18H28N2O

Tutorial organization

Time	Topic
09:00 - 09:30	**Set the scene**
09:30 - 10:45	Data analysis and Data visualization
10:45 - 11:15	Break
11:15 - 12:15	Machine Learning
12:15 - 13:00	Python at Scale (PySpark)

Agenda

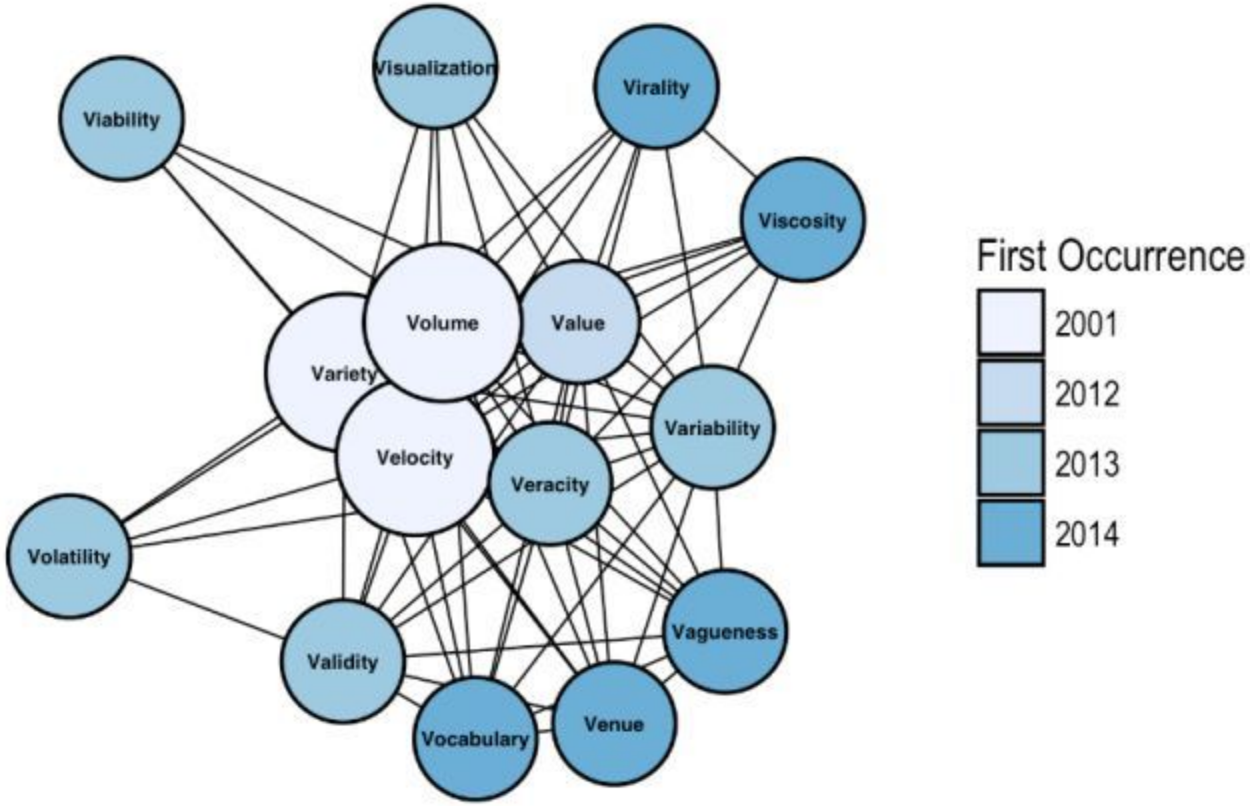
- Big Data and Data Science
- Why Python for Data Science?
- Hands-on: Jupyter Notebook Launch

What is Big Data?

- Big data is NOT defined in terms of data size
- Big data refers to datasets that a conventional database software is not capable to capture, store, manage, process, analyze, and understand
- Those datasets are large, fast, complex, and not structured
- 3Vs (volume, variety and velocity) are three properties of big data introduced by D. Laney, “3D data management: Controlling data volume, velocity and variety,” Appl. Deliv. Strateg. File, vol. 949, 2001.
    - **Volume** represents the ever - growing amount of data and challenges the current stage of storage systems.
    - **Velocity** describes how quickly the data is retrieved, stored and processed.
    - **Variety** describes the multitude of data sources like sensors, smart devices and social media often in unstructured data formats.

The inflation of Big Data Vs (2017)

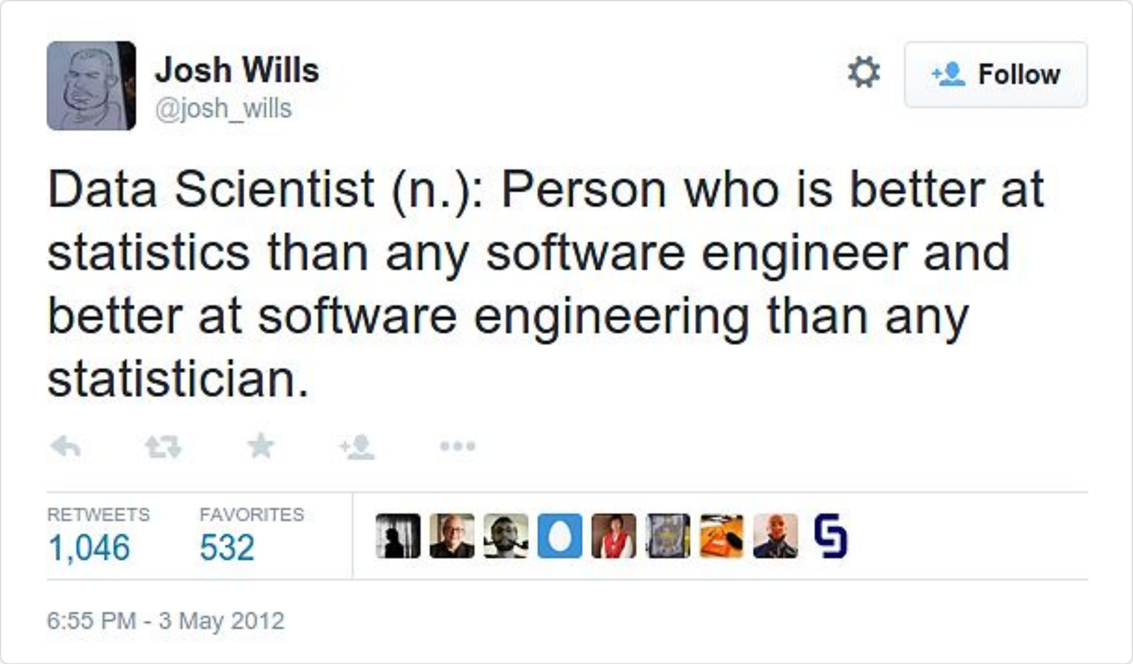
- These 42 Vs are orbiting the original three



42 V's of Big Data and Data Science visualized by Tom Shafer (2017)
https://www.elderresearch.com/blog/42-v-of-big-data

What is Data Science?

- An informal definition of data scientist by Josh Wills:



Josh Wills
@josh_wills

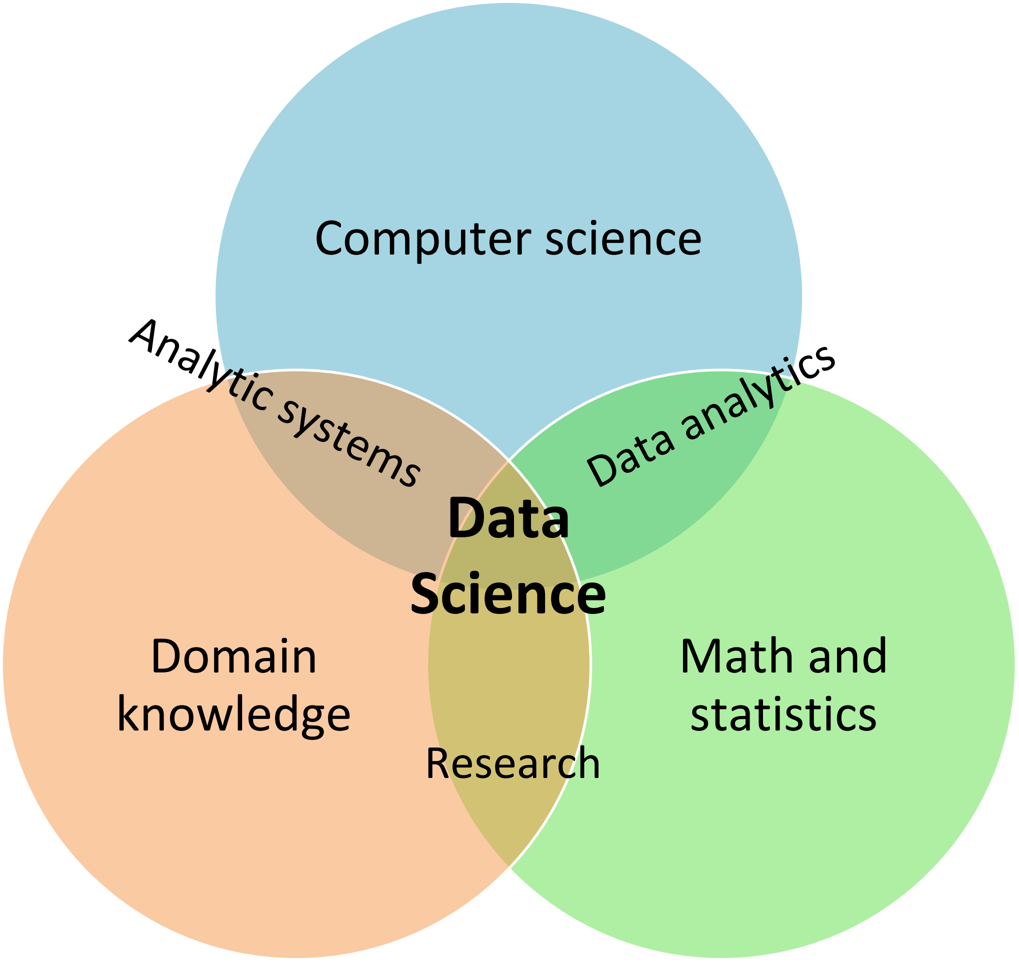
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 1,046
FAVORITES 532

6:55 PM - 3 May 2012

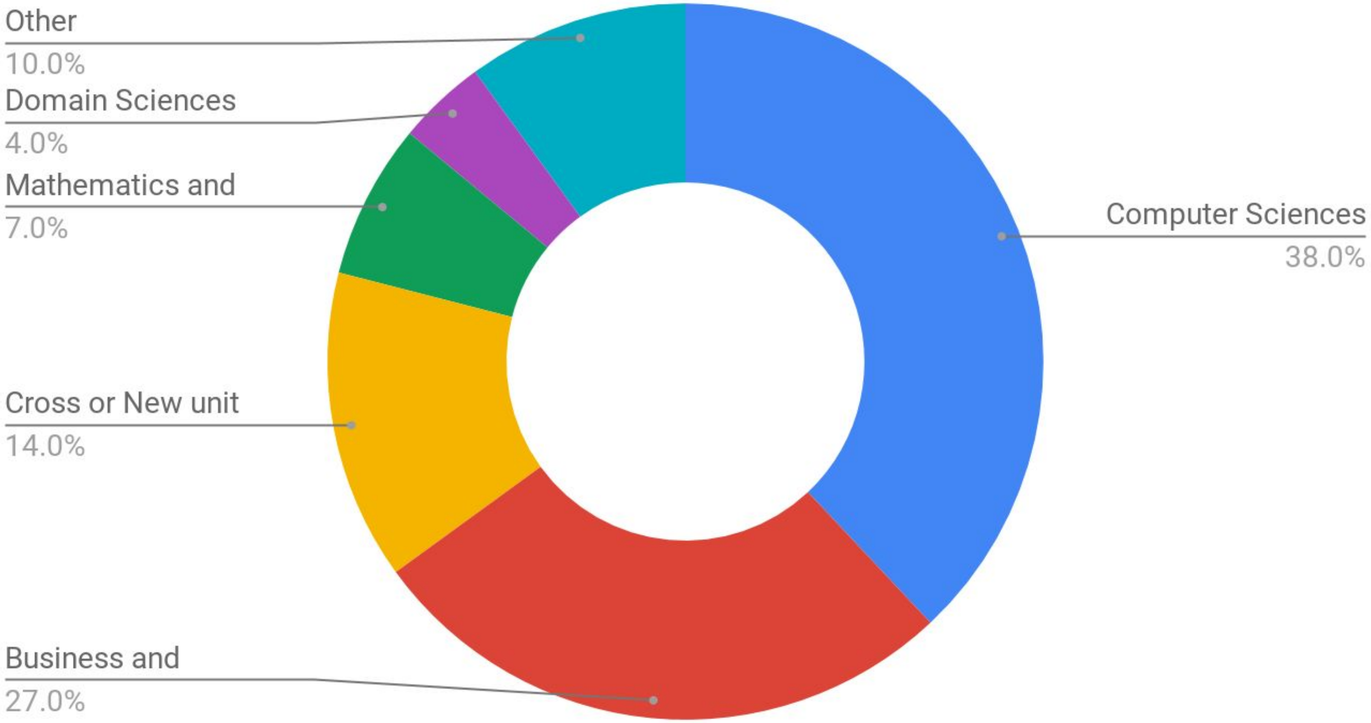
What is Data Science?

- Data science is a multidisciplinary domain
- Data science aims to extract knowledge from data with a set of tools, scientific methods and scalable systems



Data science educational programs in EU (2016)

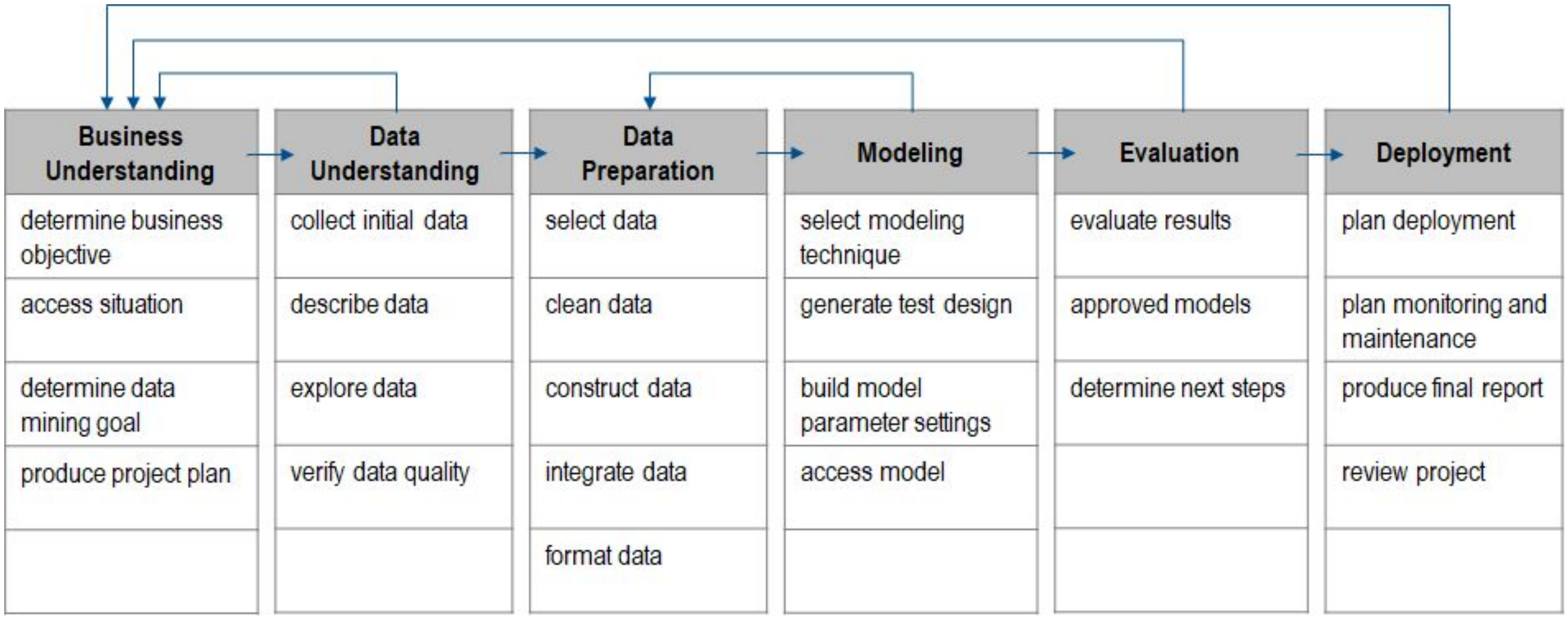
- Data Science is hard to teach due to the nature of the multidisciplinary
- ...and they are teaching different things...



EDISION survey result in 2016: Data Science educational programs providers
(https://cordis.europa.eu/project/rcn/198292/results/en)

CRISP-DM: the most used methodology for a data science project

- CRISP-DM is an open standard process model that describes common approaches used by data mining experts.
- developed by DamilerChrysler, SPSS, NCR and OHRA and released in 1999



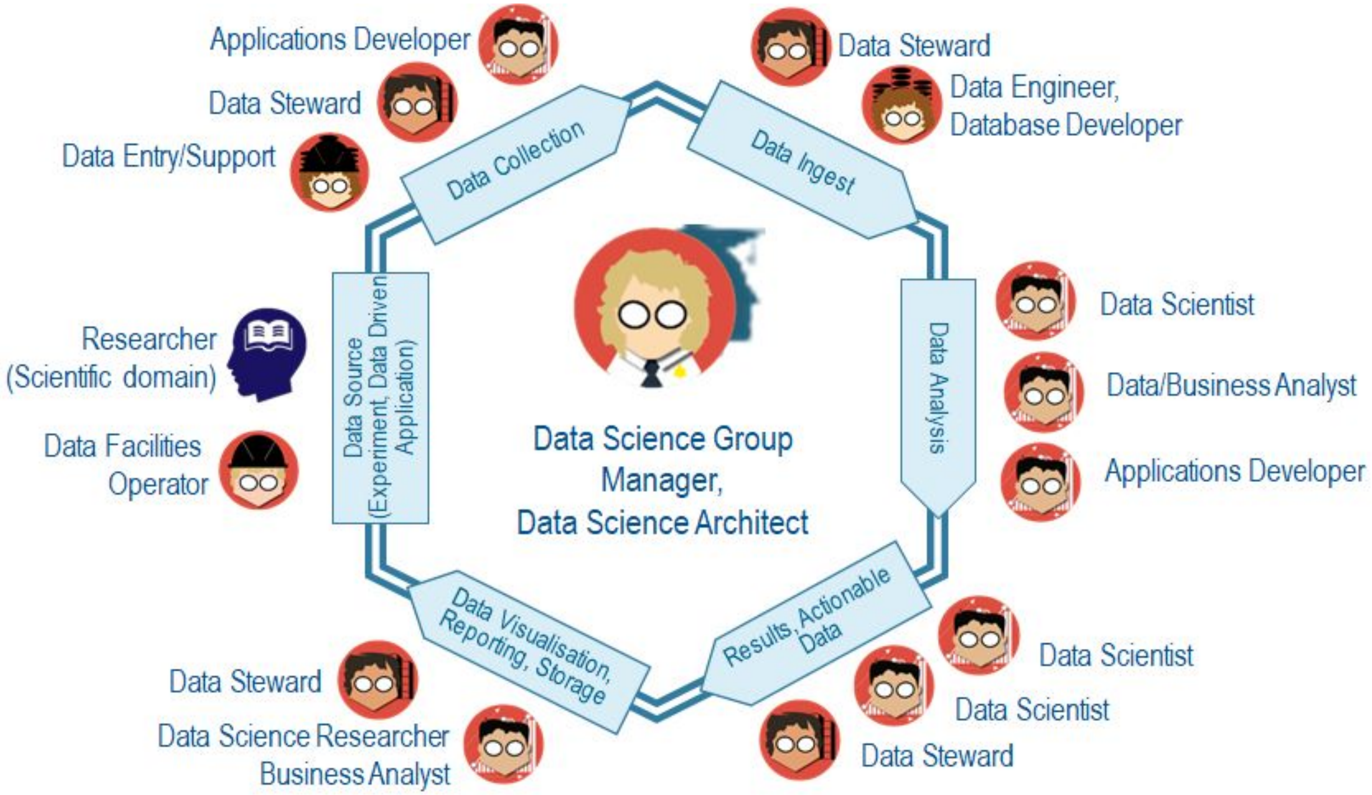
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
determine business objective	collect initial data	select data	select modeling technique	evaluate results	plan deployment
access situation	describe data	clean data	generate test design	approved models	plan monitoring and maintenance
determine data mining goal	explore data	construct data	build model parameter settings	determine next steps	produce final report
produce project plan	verify data quality	integrate data	access model		review project
		format data			

which steps are unique compared to a research project?

Data science team needs not only a data scientist

https://cordis.europa.eu/project/rcn/198292/results/en
https://zenodo.org/record/167597#.XW0rSOgvNyE

Profile Group	Profile Title
Data Science Managers	Data Science (group) Manager
	Data Science Infrastructure Manager
	Research Infrastructure Manager
Data Science Professionals	Data Scientist
	Data Science Researcher
	Data Science Architect
	Data Science Programmer/Engineer
	Data Analyst
	Business Analyst
Data handling Professionals	Data Stewards
	Digital data curator
	Digital Librarians
	Data Archivists
Database Professionals	Large scale database designer
	Large scale database admin
	Scientific database administrator
Technicians and associate profession	Big Data facilities Operator
	Large scale data storage operator
	Scientific database operator



Applications Developer
Data Steward
Data Entry/Support
Data Steward
Data Engineer, Database Developer
Data Collection
Data Ingest
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect
Data Analysis
Data Scientist
Data/Business Analyst
Applications Developer
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Steward
Data Science Researcher
Business Analyst
Data Scientist
Data Scientist
Data Steward

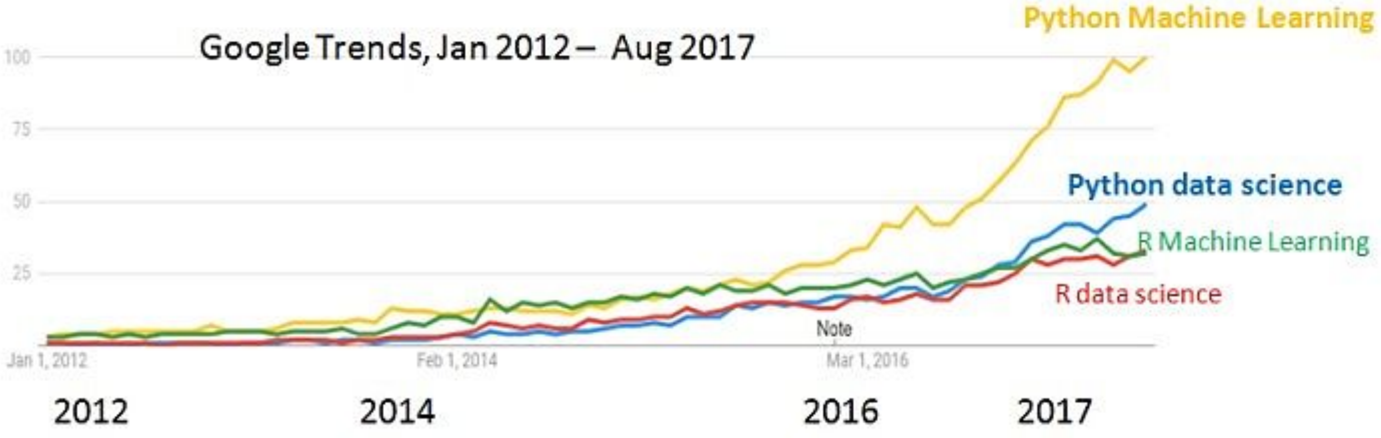
Agenda

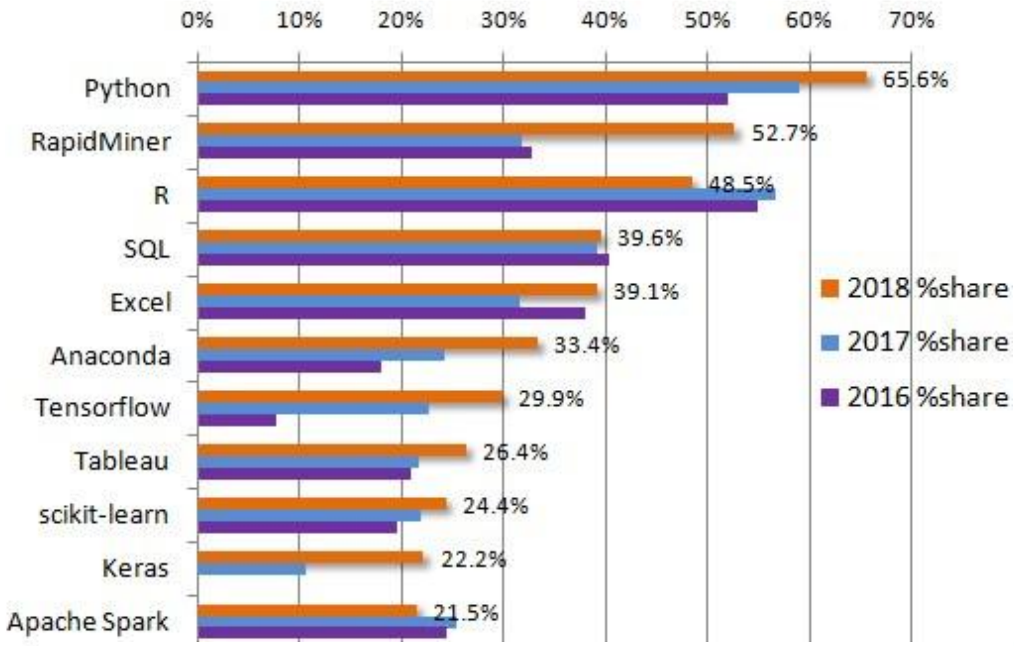
- Big Data and Data Science
- Why Python for Data Science?
- Hands-on: Jupyter Notebook Launch

Python is the most popular language among data scientists

> “ Python remains the big kahuna,
> but specialist languages hold their own ”

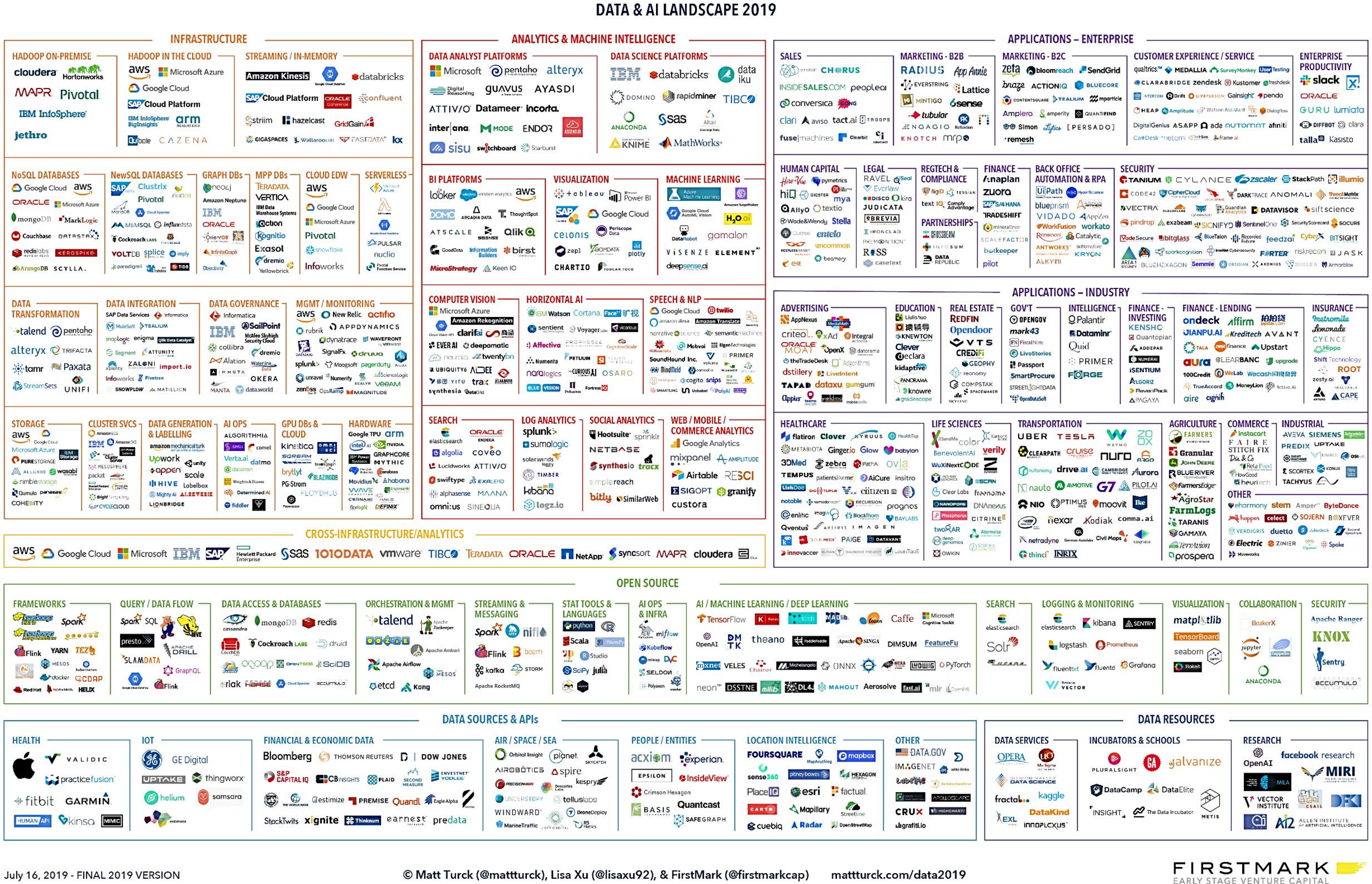
Stephen Cass (2019), *The Top Programming Languages 2019*, IEEE Spectrum



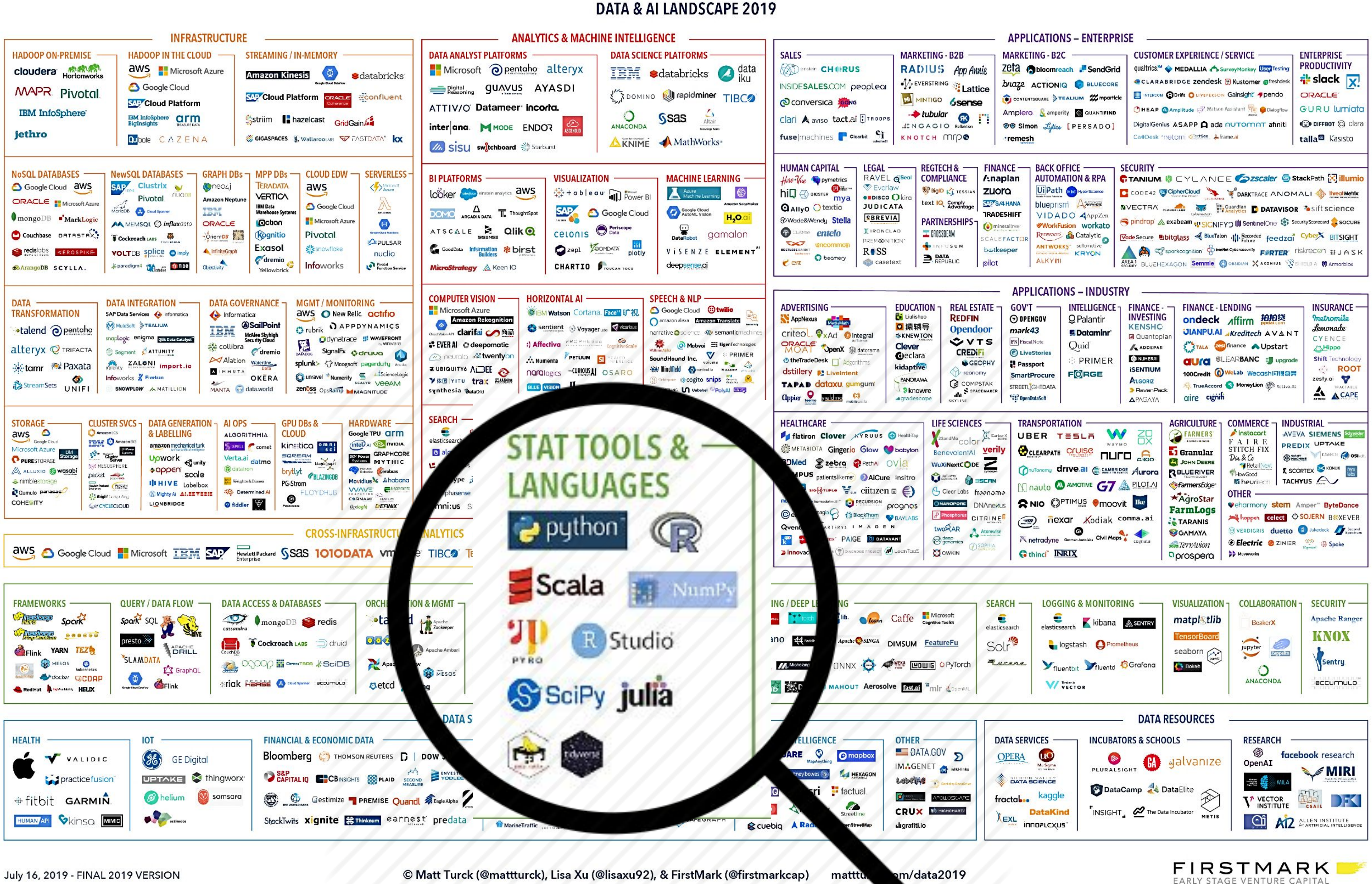


Source: KDnuggets Software Poll had over 2,300 voters in 2018 (https://goo.gl/WNZZPD) 2,900 voters in 2017 (https://goo.gl/qKp9Yr) and 2,895 voters in 2016 (https://goo.gl/23SXb2)

Python is highly compatible language in big data world



Python is highly compatible language in big data world



Why is the compatibility matter?

$1 million solution has never been used

- Netflix held the Netflix Prize open competition [1] for the best algorithm to predict user ratings for films.
- On September 21, 2009 Netflix awarded the $1M Grand Prize to team “BellKor’s Pragmatic Chaos”.
- The solution improved the recommendation algorithm by 10%. **But Netflix never implemented that solution itself.**
- According to the Netflix blog post [2]:

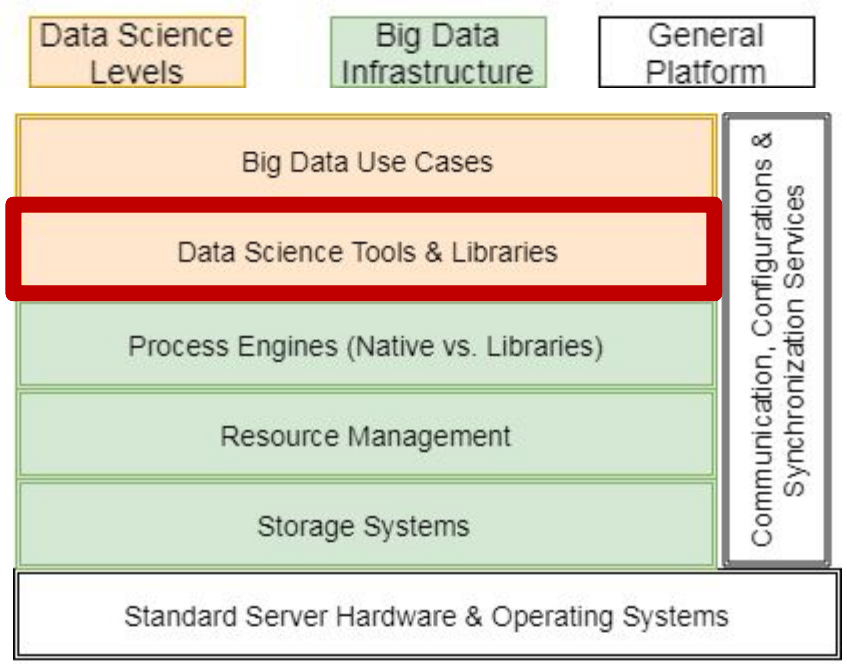
> “*We evaluated some of the new methods offline but* ***the additional accuracy gains that we measured did not seem to justify the engineering effort*** *needed to bring them into a production environment.*”

[1] https://www.kaggle.com/netflix-inc/netflix-prize-data

[2] https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

Compatibility Evaluation in Big Data stack (1/3)
among data science layer

- Python is the first-class citizen in the deep learning/AI world



Data Science Levels

Big Data Infrastructure

General Platform

Big Data Use Cases

Data Science Tools & Libraries

Process Engines (Native vs. Libraries)

Resource Management

Storage Systems

Communication, Configurations & Synchronization Services

Standard Server Hardware & Operating Systems

Ivanov, T., & Singhal, R.. (2018). *ABench: Big Data Architecture Stack Benchmark*. Paper presented at the Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, Berlin, Germany, April 09 - 13, 2018.

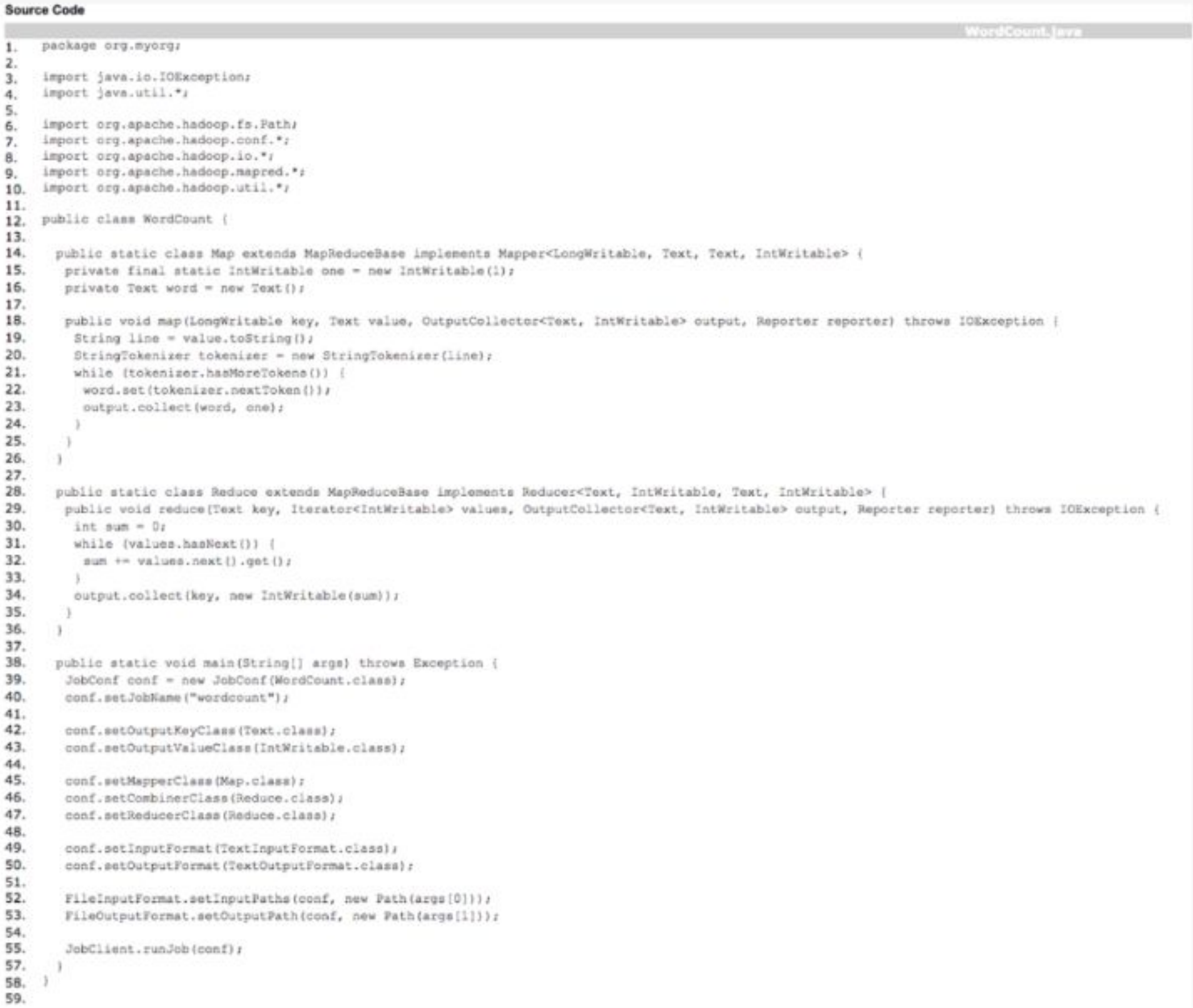
	Python	R	Other APIs
TensorFlow	Yes	No	c++, Java, Go, Swift
Keras	Yes	Yes	No
MXNet	Yes	Yes	c++, Scala, Julia, Perl
PyTorch	Yes	No	No
CNTK	Yes	No	c++

compatibility table (https://goo.gl/na3DzY)

Compatibility Evaluation in Big Data stack (1/3)

among data science layer

Java MapReduce

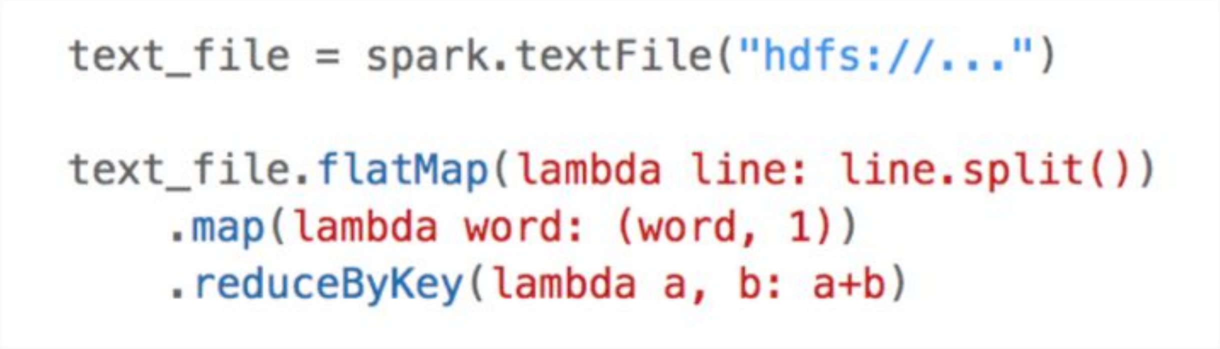


Source Code

WordCount.java

```java
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();

   public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);
    while (tokenizer.hasMoreTokens()) {
     word.set(tokenizer.nextToken());
     output.collect(word, one);
    }
   }
  }

  public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
   public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) {
     sum += values.next().get();
    }
    output.collect(key, new IntWritable(sum));
   }
  }

  public static void main(String[] args) throws Exception {
   JobConf conf = new JobConf(WordCount.class);
   conf.setJobName("wordcount");

   conf.setOutputKeyClass(Text.class);
   conf.setOutputValueClass(IntWritable.class);

   conf.setMapperClass(Map.class);
   conf.setCombinerClass(Reduce.class);
   conf.setReducerClass(Reduce.class);

   conf.setInputFormat(TextInputFormat.class);
   conf.setOutputFormat(TextOutputFormat.class);

   FileInputFormat.setInputPaths(conf, new Path(args[0]));
   FileOutputFormat.setOutputPath(conf, new Path(args[1]));

   JobClient.runJob(conf);
  }
}

```

Spark's Python API



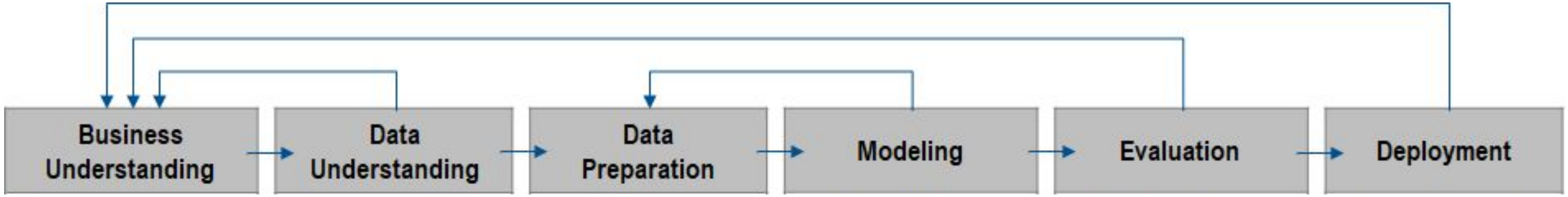
```python
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

source: https://generalassemb.ly/data/data-science/spark

Compatibility Evaluation in Big Data stack (1/3)

among data science layer

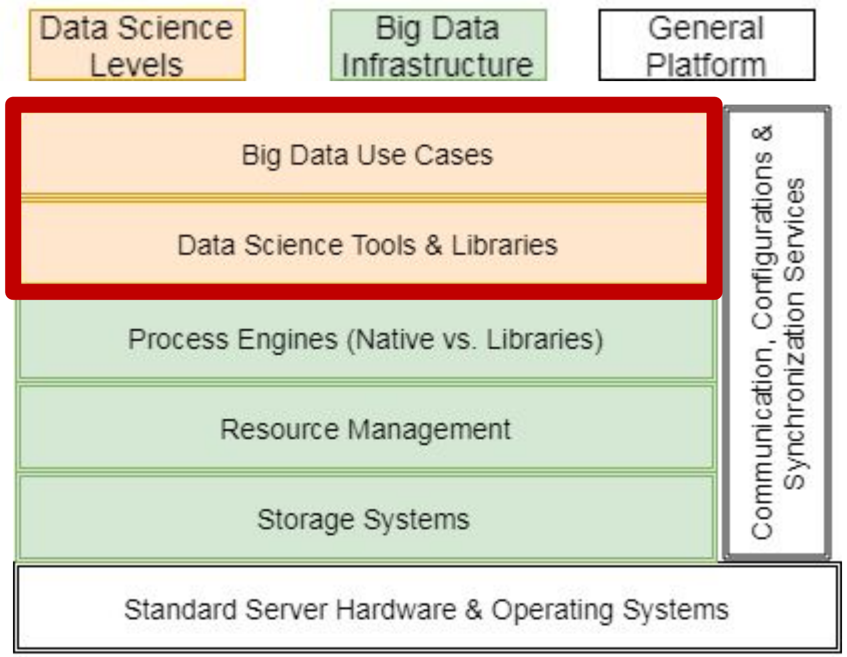
- **The requirements for a data science project cannot be done up-front.**
- **With data science, you learn as you go, not before you go.** You must be learned through experimentation, trial and error, and iteration.



CRISP-DM methodology

Compatibility Evaluation in Big Data stack (2/3)

application layer vs. data science layer


Data Science Levels
Big Data Infrastructure
General Platform
Big Data Use Cases
Data Science Tools & Libraries
Process Engines (Native vs. Libraries)
Resource Management
Storage Systems
Communication, Configurations & Synchronization Services
Standard Server Hardware & Operating Systems

Ivanov, T., & Singhal, R.. (2018). *ABench: Big Data Architecture Stack Benchmark*. Paper presented at the Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, Berlin, Germany, April 09 - 13, 2018.

Java
JS
php
PROGRAMMING REPUBLIC OF PERL
GO
Java
R
julia
APACHE Spark
K
scikit learn
SQL

Popular web programming languages
https://t.ly/VAYlP

> “ costs scale super-linearly with the number of people involved. ”

Eric Colson (2019). *Why Data Science Teams Need Generalists, Not Specialists.* Harvard Business Review

> “ The number of relationships (r) grows as a function number of members (n) per this equation: r = (n^2-n) / 2. And, each relationship bares coordination costs ”

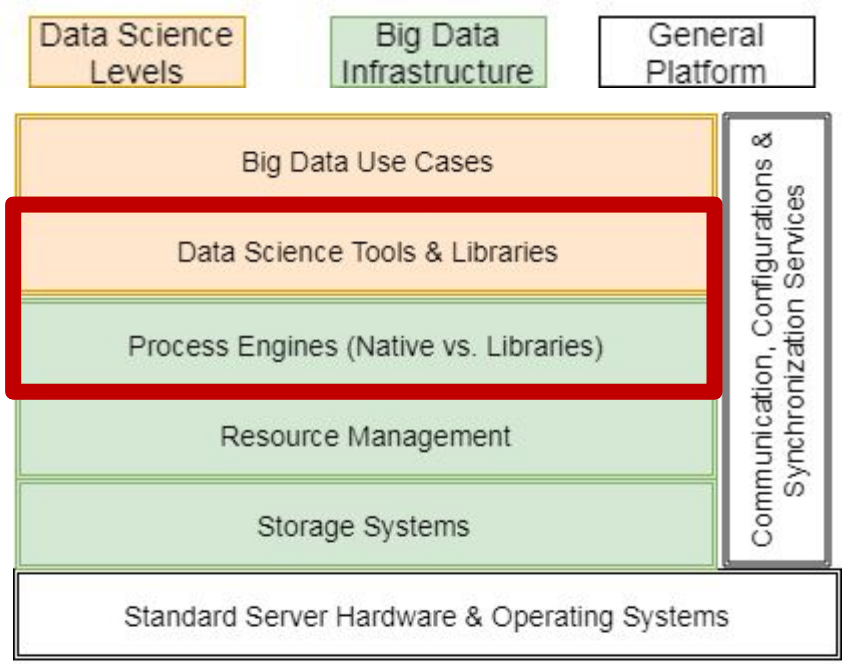
Hackman, J. R., & Hackman, R. J. (2002). *Leading teams: Setting the stage for great performances*. Harvard Business Press.

UMM UNIVERSITÄTSMEDIZIN MANNHEIM

Medizinische Fakultät Mannheim der Universität Heidelberg
Universitätsklinikum Mannheim

Compatibility Evaluation in Big Data stack (3/3)
data science layer vs. process engine layer

- Two layers of data science tools and process engines have a high compatibility


Data Science Levels
Big Data Infrastructure
General Platform
Big Data Use Cases
Data Science Tools & Libraries
Process Engines (Native vs. Libraries)
Resource Management
Storage Systems
Communication, Configurations & Synchronization Services
Standard Server Hardware & Operating Systems

Ivanov, T., & Singhal, R.. (2018). *ABench: Big Data Architecture Stack Benchmark*. Paper presented at the Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, Berlin, Germany, April 09 - 13, 2018.

	Java	Scala	Python	R	Other APIs
Hadoop	Yes	No	Yes	No	c/c++, ruby, groovy, Perl
Spark	Yes	Yes	Yes	Yes	
Flink	Yes	Yes	Yes	Yes	
Kafka	Yes	Yes	Yes	Yes	c/c++, ruby, groovy, Go, .NET, and more

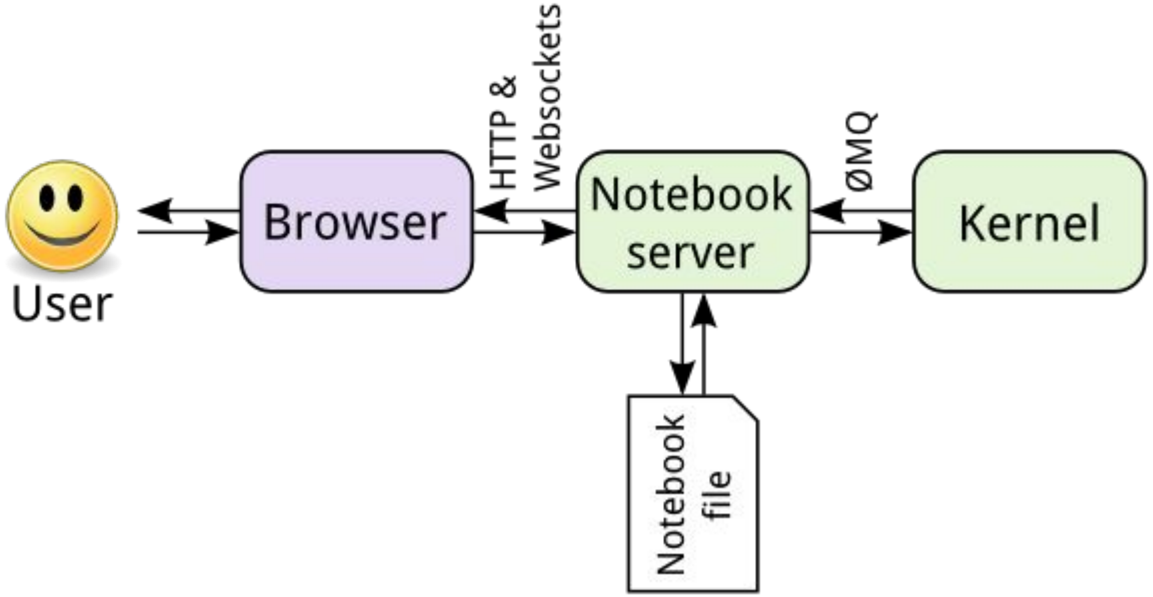
https://data-flair.training/blogs/hadoop-vs-spark-vs-flink/
https://www.confluent.io/blog/12-programming-languages-walk-into-a-kafka-cluster

Agenda

- Big Data and Data Science
- Why Python for Data Science?
- Data Mining Methodology
- Hands-on: Jupyter Notebook Launch

Jupyter Notebook

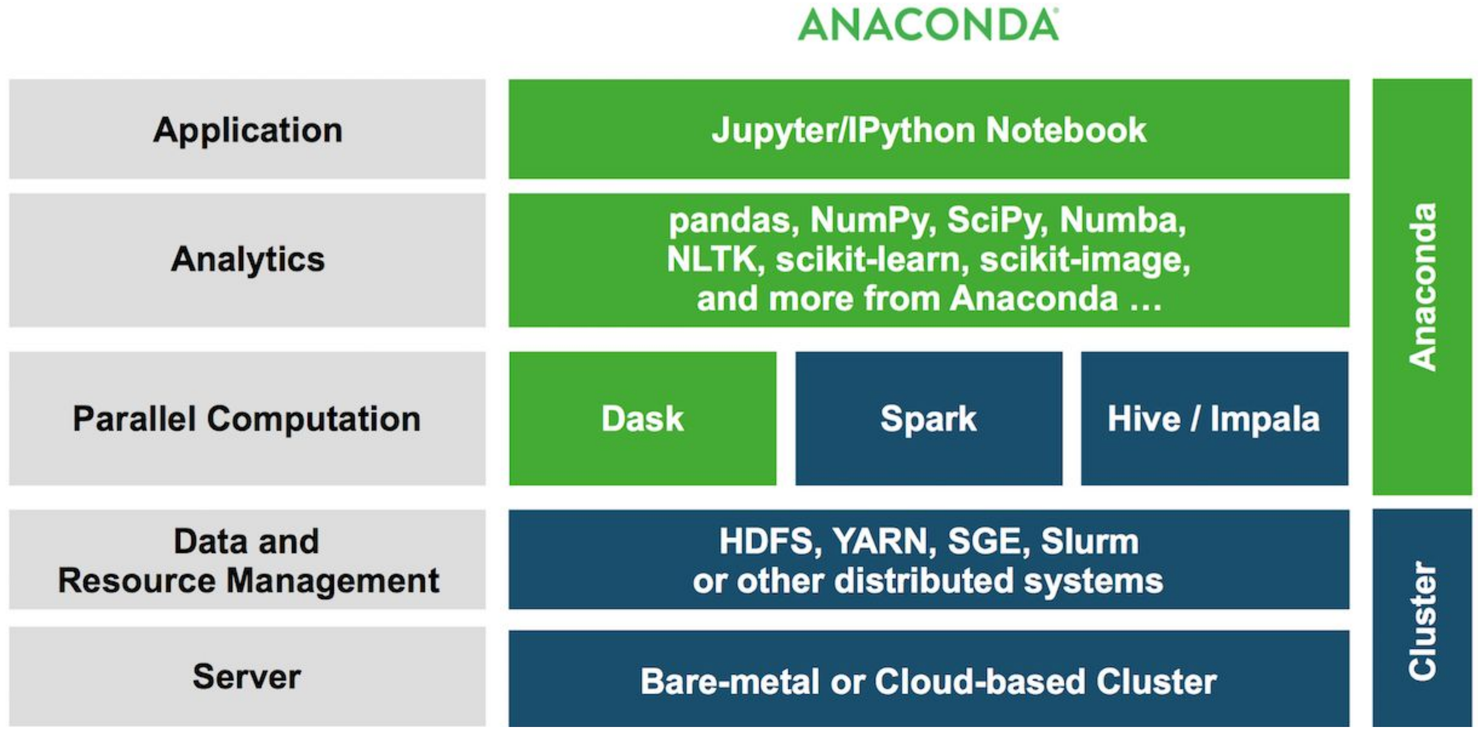
- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text
- Jupyter supports over 40 programming languages, including **Ju**lia, **Pyt**hon, **R**, and Scala. (https://github.com/jupyter/jupyter/wiki/Jupyter-kernels)



https://jupyter.readthedocs.io/en/latest/architecture/how_jupyter_ipython_work.html

Install Jupyter Notebook using Anaconda

- Anaconda is a open-source distribution including Jupyter Notebook, Python, and popular libraries for data science project



ANACONDA

Application — Jupyter/IPython Notebook

Analytics — pandas, NumPy, SciPy, Numba, NLTK, scikit-learn, scikit-image, and more from Anaconda …

Parallel Computation — Dask | Spark | Hive / Impala

Data and Resource Management — HDFS, YARN, SGE, Slurm or other distributed systems

Server — Bare-metal or Cloud-based Cluster

Anaconda

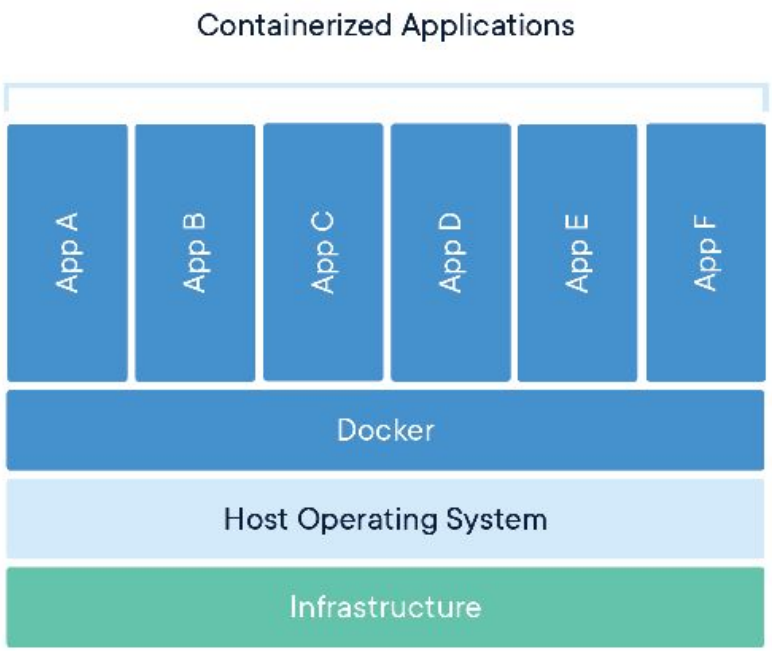
Cluster

https://www.anaconda.com

docker pull Jupyter Notebook

https://hub.docker.com/u/jupyter

- Docker allows you to decouple applications from infrastructure
  (i.e. it is no longer needed to match the development environment to the production environment)
- Docker
  - Docker is an open platform that performs OS-level virtualization
  - Unlike system level virtualization, kernel is the same for all users



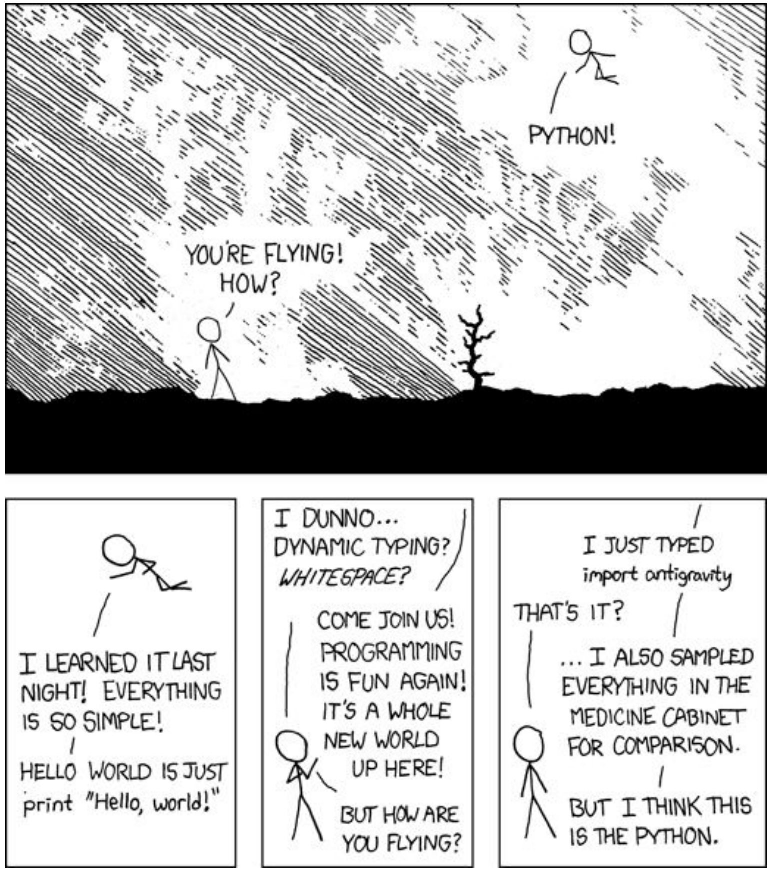
https://docs.docker.com/engine/docker-overview

Jupyter Notebook Launch

1. Open a browser and go to 129.206.5.27:PORT
2. Open a browser and go to 129.206.7.29:PORT

THANK YOU

Kim Hee
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.