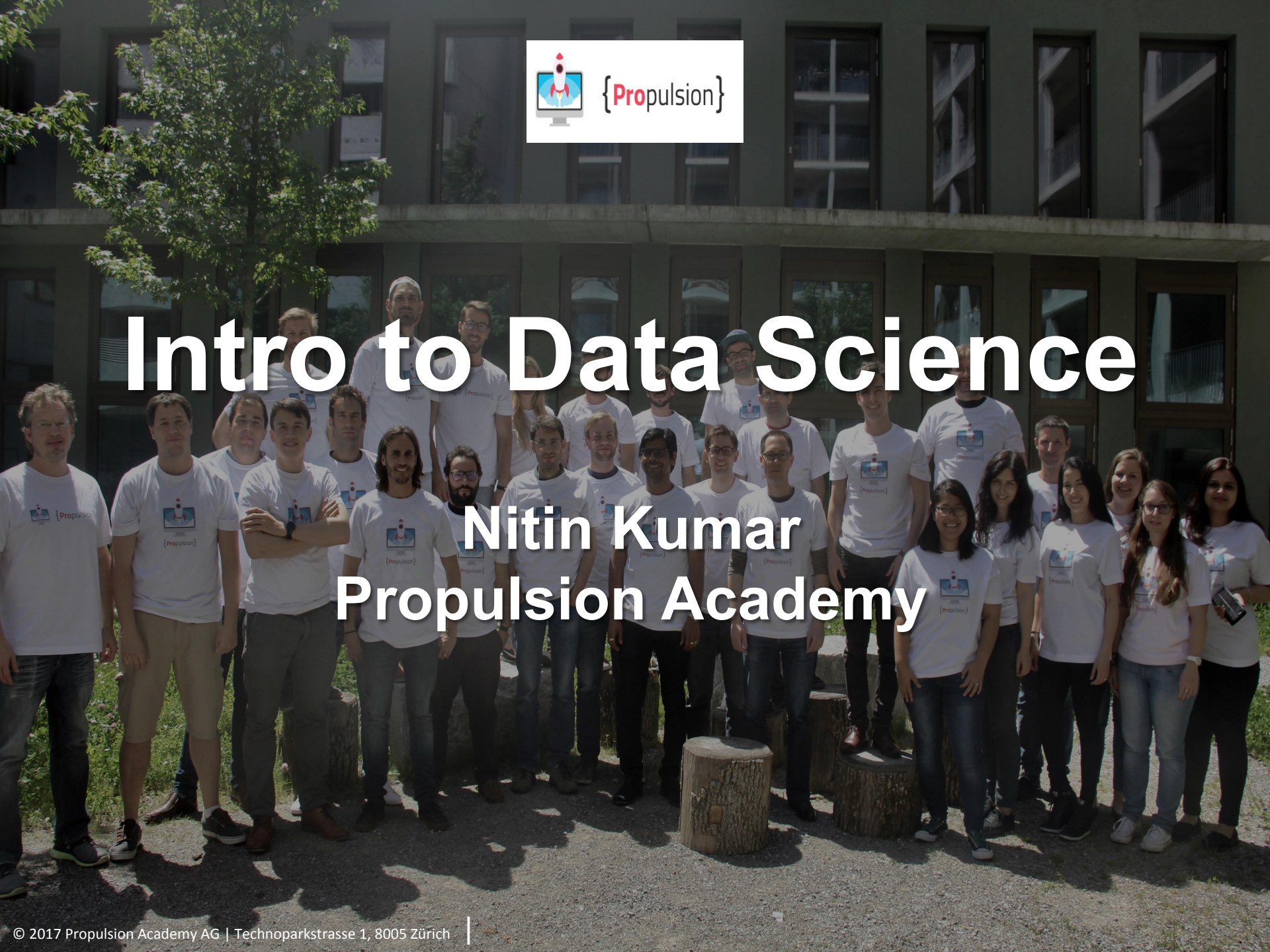


Intro to Data Science

Nitin Kumar
Propulsion Academy

Biodata

2003-2006
B.Sc.
Microbiology

2006-2008
M.Sc.
Biotechnology

2008-2013
PhD
Data mining/
Cancer genomics/
Biostatistics

2013-2015
Post-doc
Mathematical Biology/
Data Scientist

2015-2016
Quantitative data scientist
(Consultant)

2017 –
Program manager/
(Data science consultant)

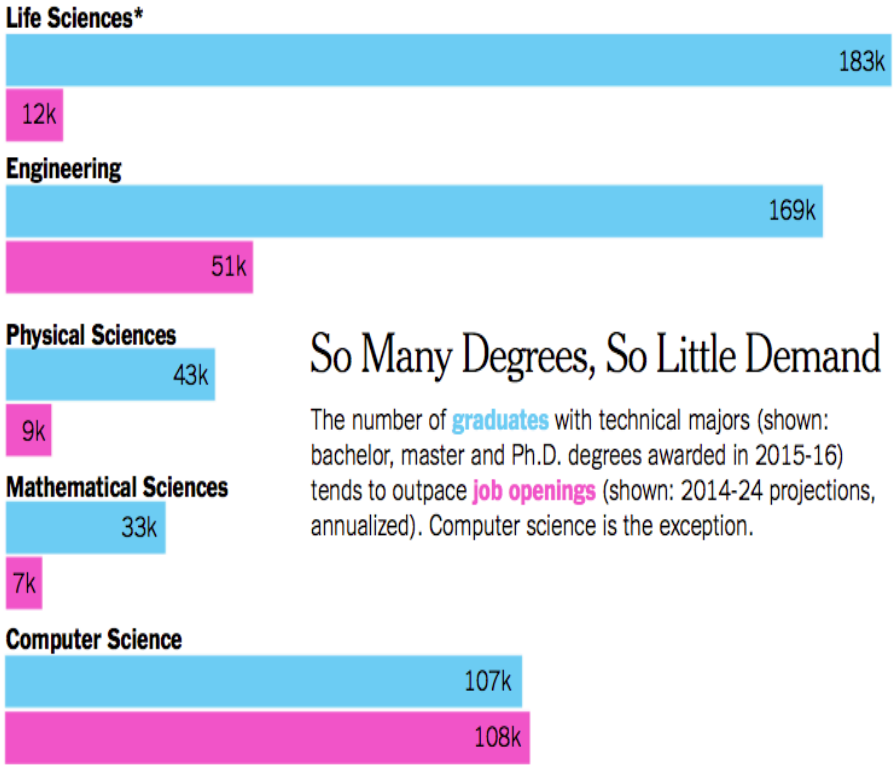
Table of contents

1. Why?
2. What?
3. Who?
4. Where?
5. How?
6. Propulsion

Why are we talking about it?

Why are we talking about it?

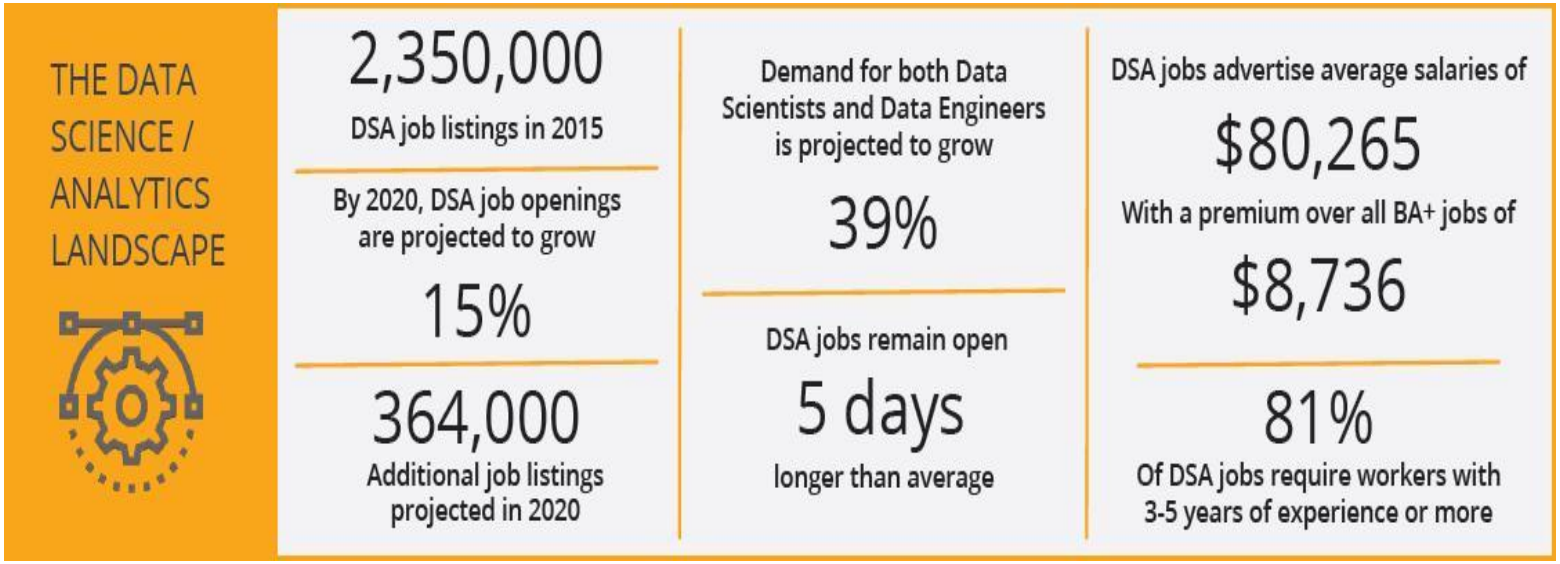
Graduate degrees in US vs market demand



https://www.nytimes.com/2017/11/01/education/edlife/stem-jobs-industry-careers.html

Why are we talking about it?

Data science and analytics jobs in US



https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF

What is data science?

{Propulsion}

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

When Harvard Business Review called it "The Sexiest Job of the 21st Century"[6] the term became a buzzword, and is now often applied to business analytics,[7] or even arbitrary use of data, or used as a sexed-up term for statistics.[8] While many university programs now offer a data science degree, there exists no consensus on a definition or curriculum contents.[7] Because of the current popularity of this term, there are many "advocacy efforts" surrounding it.[9]

https://en.wikipedia.org/wiki/Data_science#cite_note-Harvard-6

What is Data Science?

{Propulsion}

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data scienc with data.[3] It er omputer science, in

Turing awa en) and asserted th

When Harv siness analytics,[7] degree,

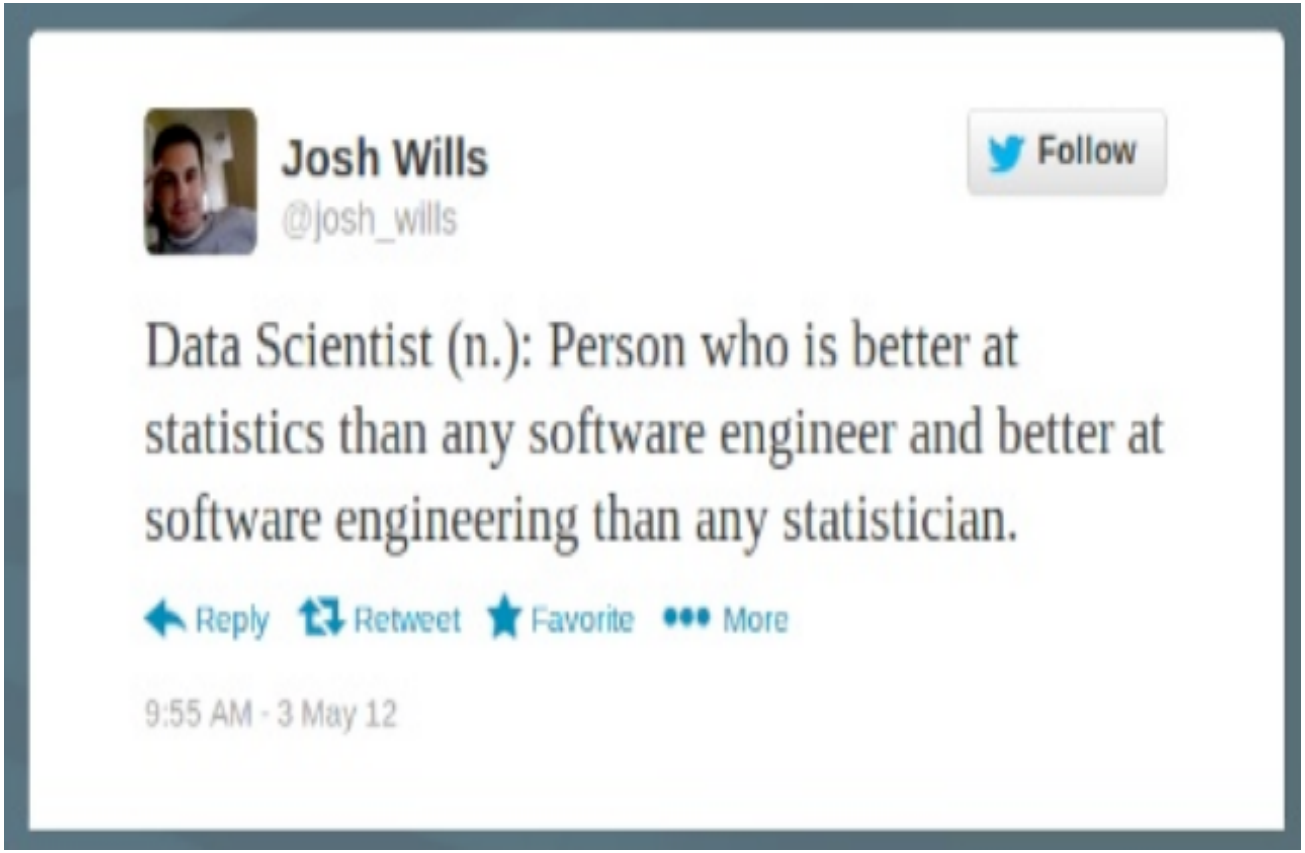
there exists no consensus on a definition or curriculum contents.[7] Because of the current popularity of this term, there are many "advocacy efforts" surrounding it.[9]

It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

https://en.wikipedia.org/wiki/Data_science#cite_note-Harvard-6

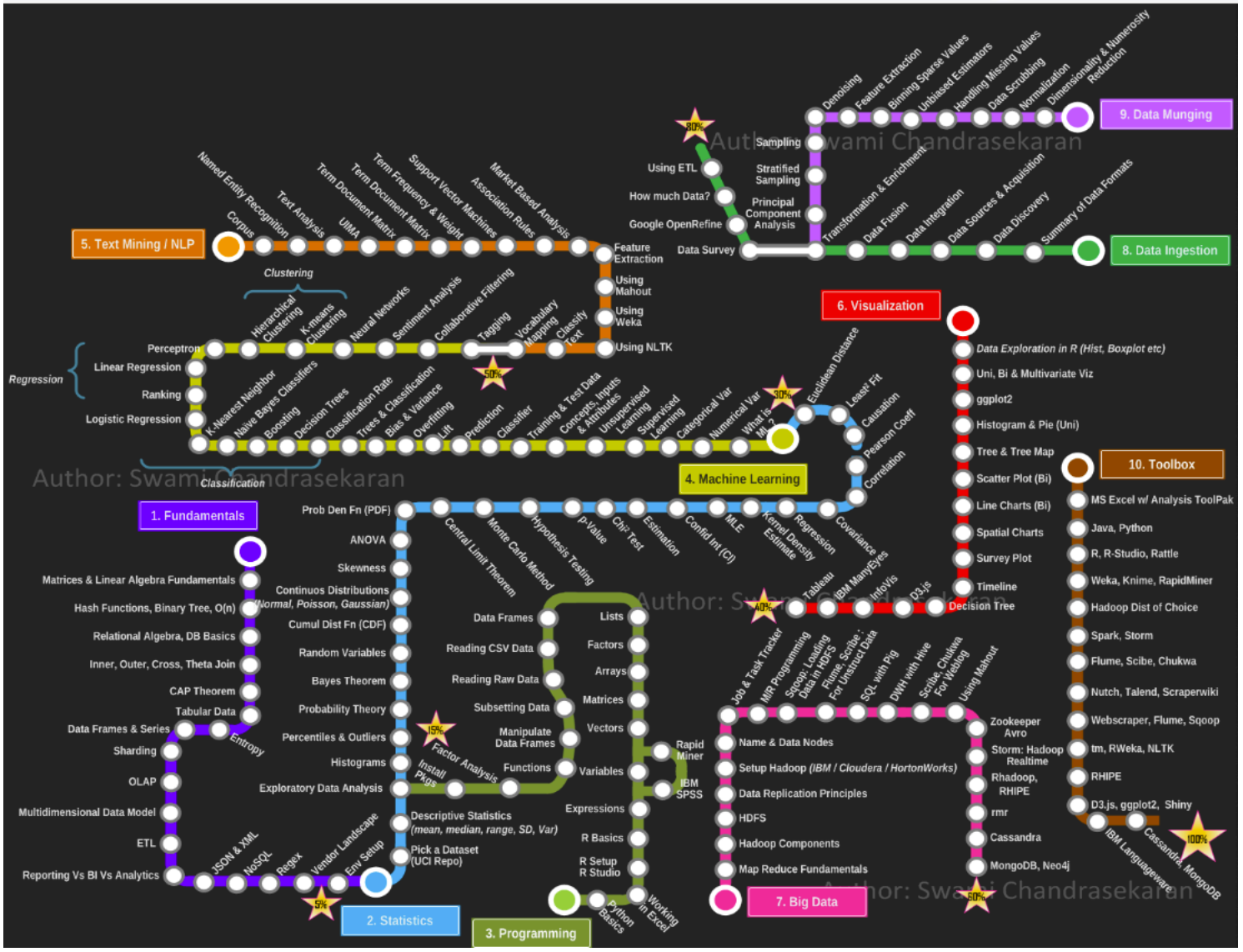
Who is a data scientist?

{Propulsion}



https://en.wikipedia.org/wiki/Data_science#cite_note-Harvard-6

Roadmap to become a data scientist



http://nirvacana.com/thoughts/becoming-a-data-scientist/

What does a data scientist do?

- Looks for the data they need
- Explores data
- Finds out why it looks like the way it does
- Cleans up data garbage
- Creates reference datasets
- Tries to understand the problem from different angles

80% of total project time

- Finally – Makes an attempt to write really amazing and sophisticated algorithm that can change the world

20% of total project time

Where is data science needed?

Targeted advertisement

Fraud detection

Hotel recommendation system

- Everywhere!

Churn analysis

Lung cancer diagnostics

Burglary risk estimation

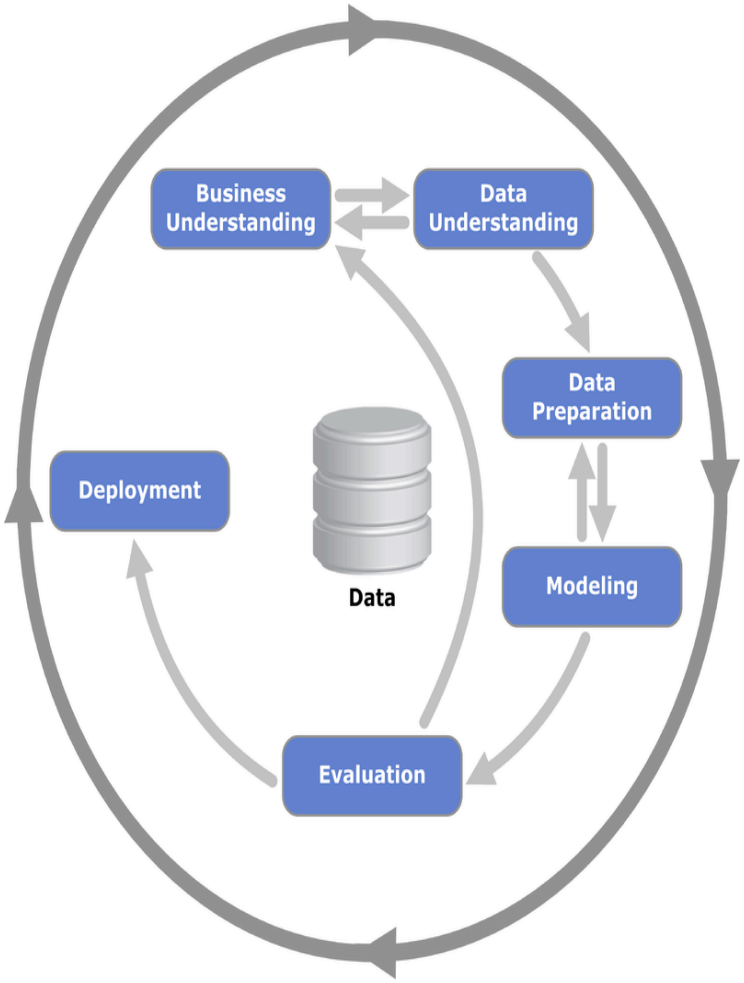
Swiss job market

Table of contents

1. Why?
2. What?
3. Who?
4. Where?
5. How?
6. Propulsion

Are there known standards for data project cycle?

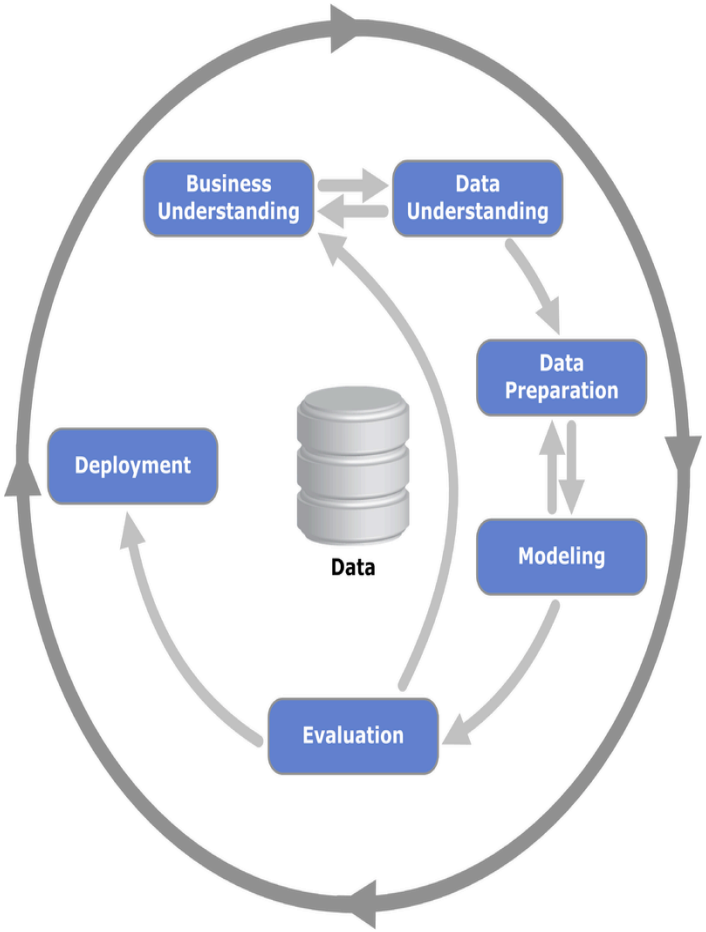
- Cross-industry standard process for data mining (CRISP-DM)



https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

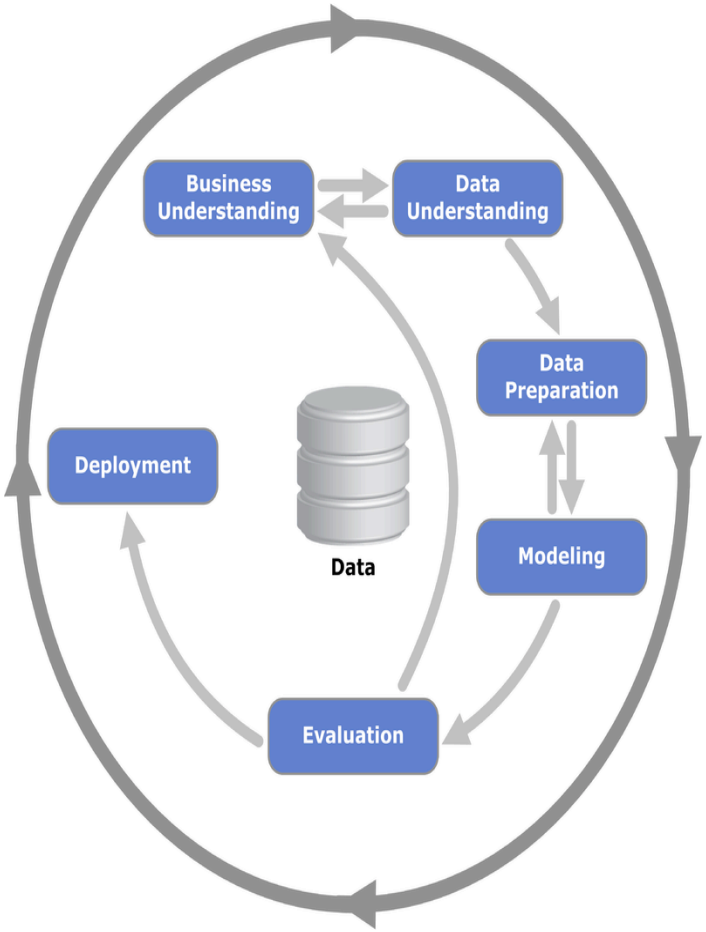
Business understanding

- Business objectives – background, frame the problem, detection criteria?
- Assess situation – what data do we require, what assumptions are we making, costs and benefits?
- Goals of the the project
- Project plan and decision model
- Sketch a plan



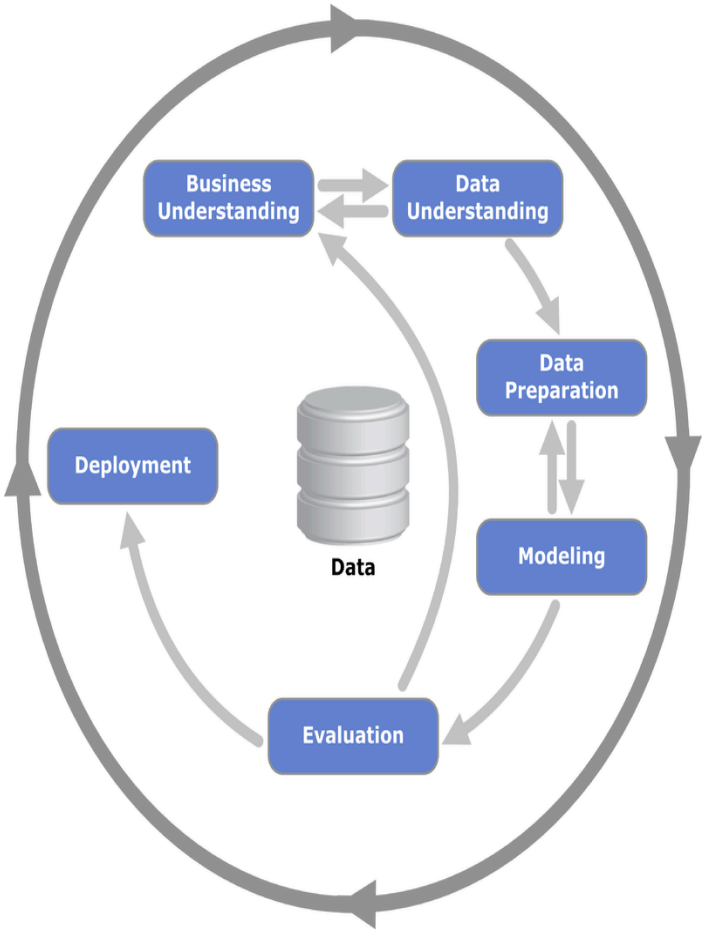
Data understanding

- Collect initial data (database, files, documents), make reports
- Data description and exploration (variables, distributions, correlations, clustering, PCA, …..)
- Data quality issues – how may data formats, spelling mistakes, empty fields



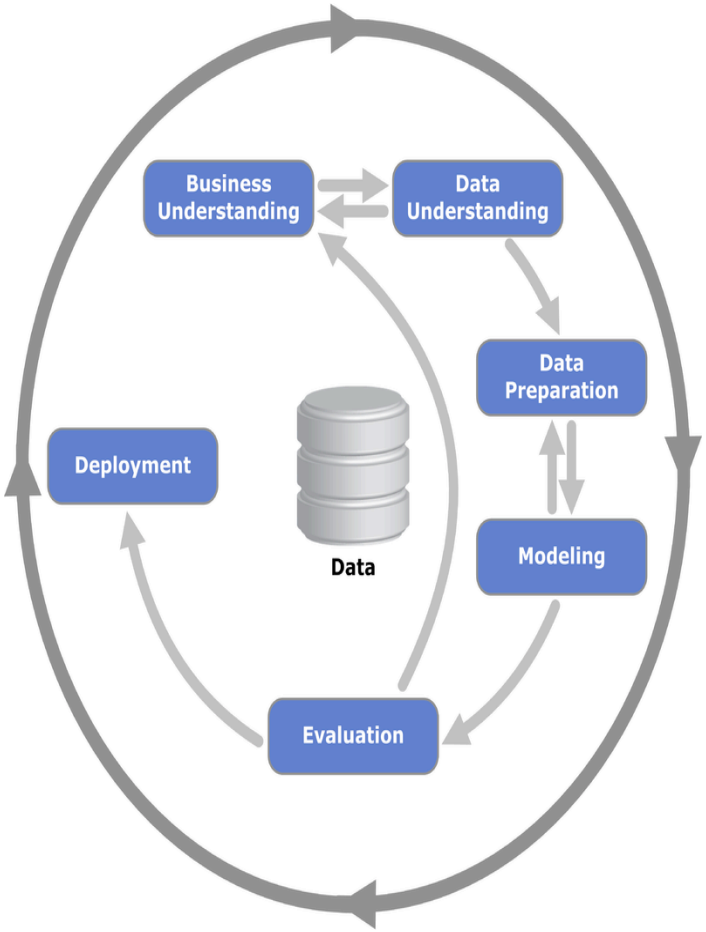
Data preparation

- Select data – what to include, what not and why?
- Data cleaning – remove empty rows/columns, correct for spellings, split address into individual fields
- Merge data – Single data table?
- Feature Engineering
- Define final data



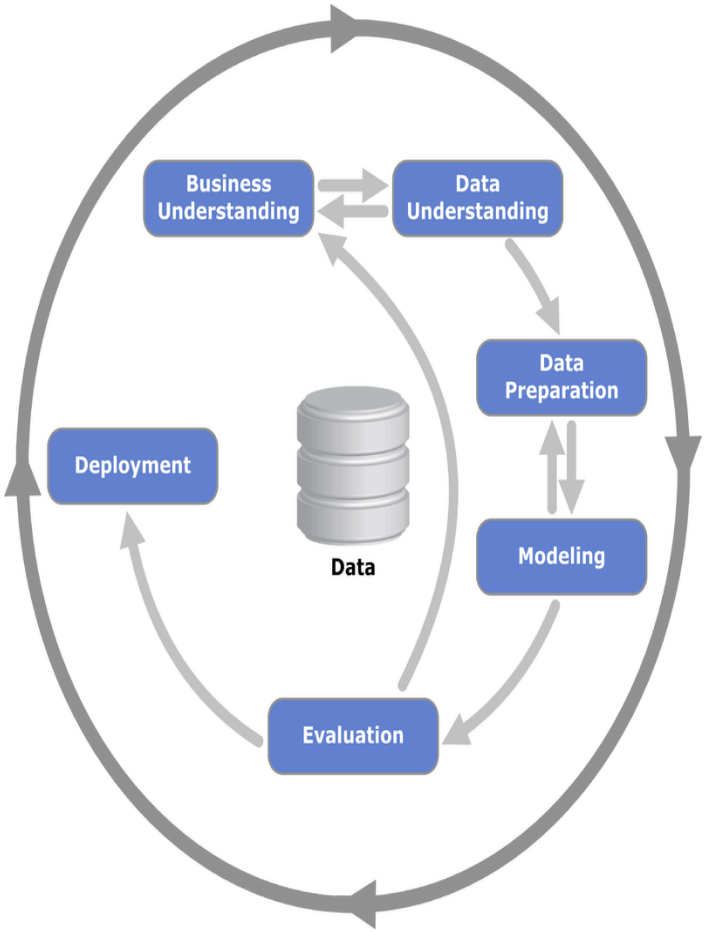
Modeling

- Generate test design - Reference dataset
- Technique selection – regression, SVM, PCA, clustering, AI, Ensemble
- Modeling assumptions, description and Parameter optimization
- Assessment criteria – Mean squared error, ROC, precision, recall
- Model revision



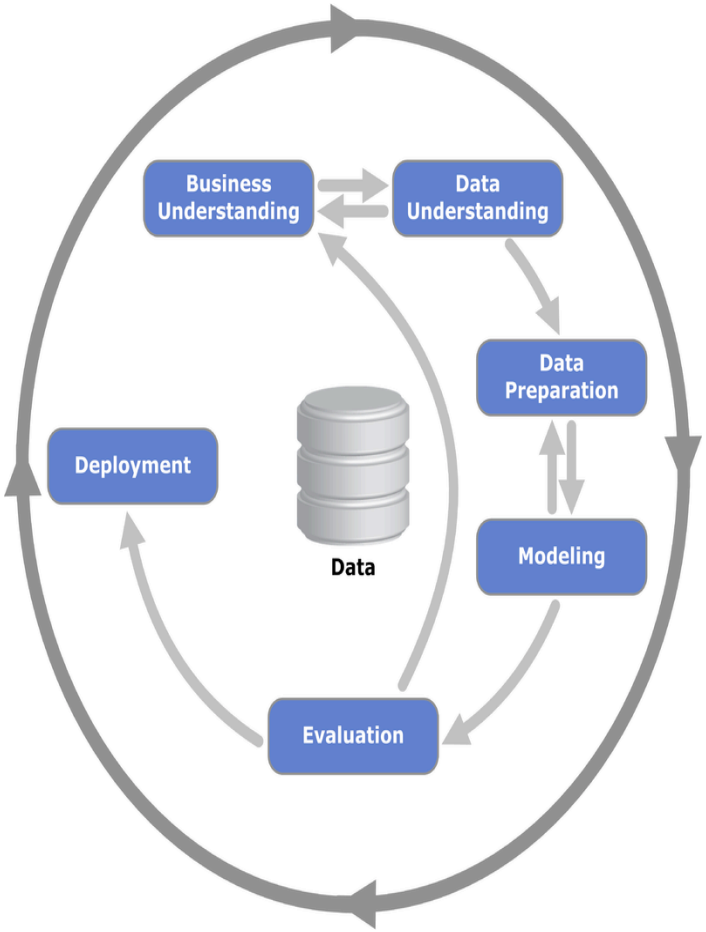
Evaluation

- Results assessment –business success criteria, approved models
- Process review – check data flow, check code, check logging, process-timing
- Define next steps



Deployment

- Make a deployment plan
- Plan monitoring and maintenance
- Final reports
- Review project and make documentation



Back to data scientist

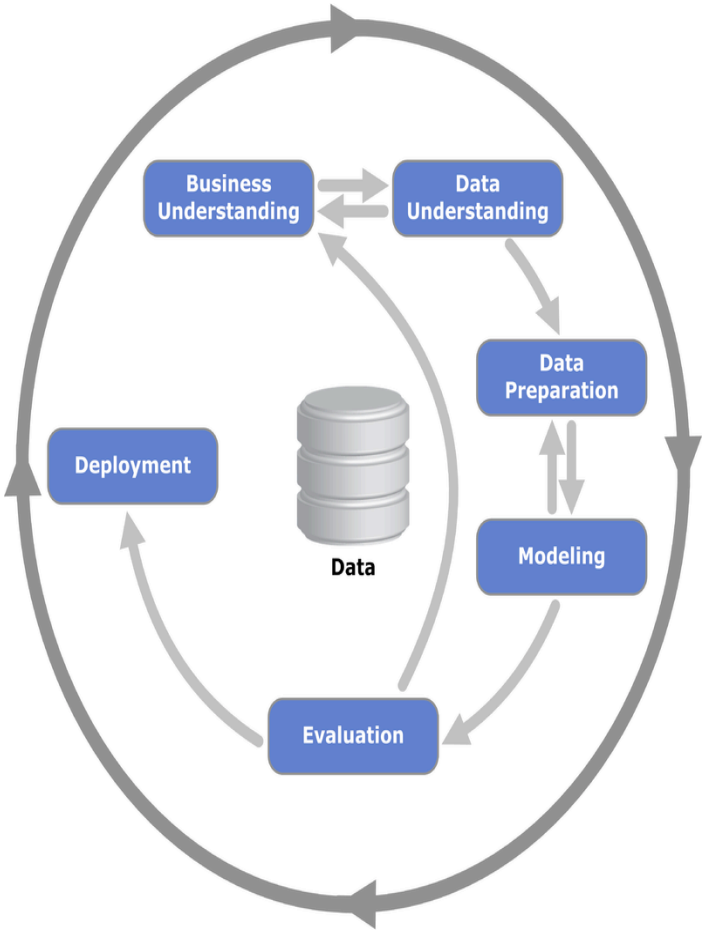


Table of contents

1. Why?
2. What?
3. Who?
4. Where?
5. How?
6. Propulsion

Propulsion academy

- Full and part-time Full-Stack Development and Data Science bootcamps
- Hans-on and project-oriented approach
- Curricula representing actual industry needs
- Incubator-like atmosphere

Data science tools and technologies

{Propulsion}

Primary programming languages used throughout the field.

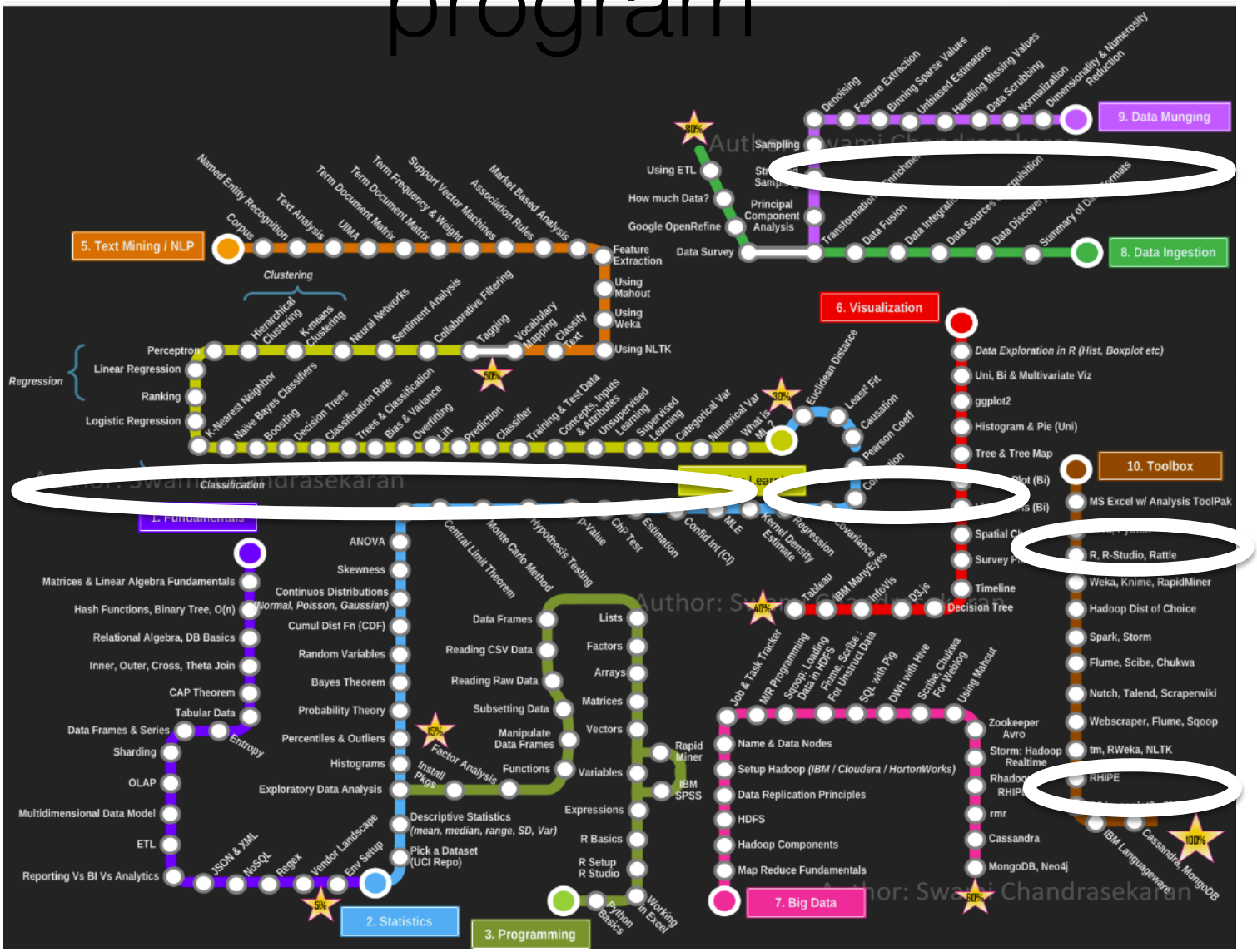
Open-standard file format to read and write data.

Python libraries for data processing, matrices, tables …

Web application to share documents and live code.

Main library for machine learning tools.

Core concepts covered in program



Where are our students today?

{Propulsion}

Next data science program

- 15th January, 2018

- Thanks!