How to start your journey as a Data Scientist?

PARVANEH SHAFIEI

WOMAN IN DATA SCIENCE – TURIN 26TH FEBRUARY

Who I am

- Past Web & Software developer
- Master: computer science in Polimi

aizoon® AUSTRALIA EUROPE USA TECHNOLOGY CONSULTING

REPLY

EY Building a better working world

R Ladies Milan

Where are Rladies now?

+70 CHAPTERS

+20 COUNTRIES

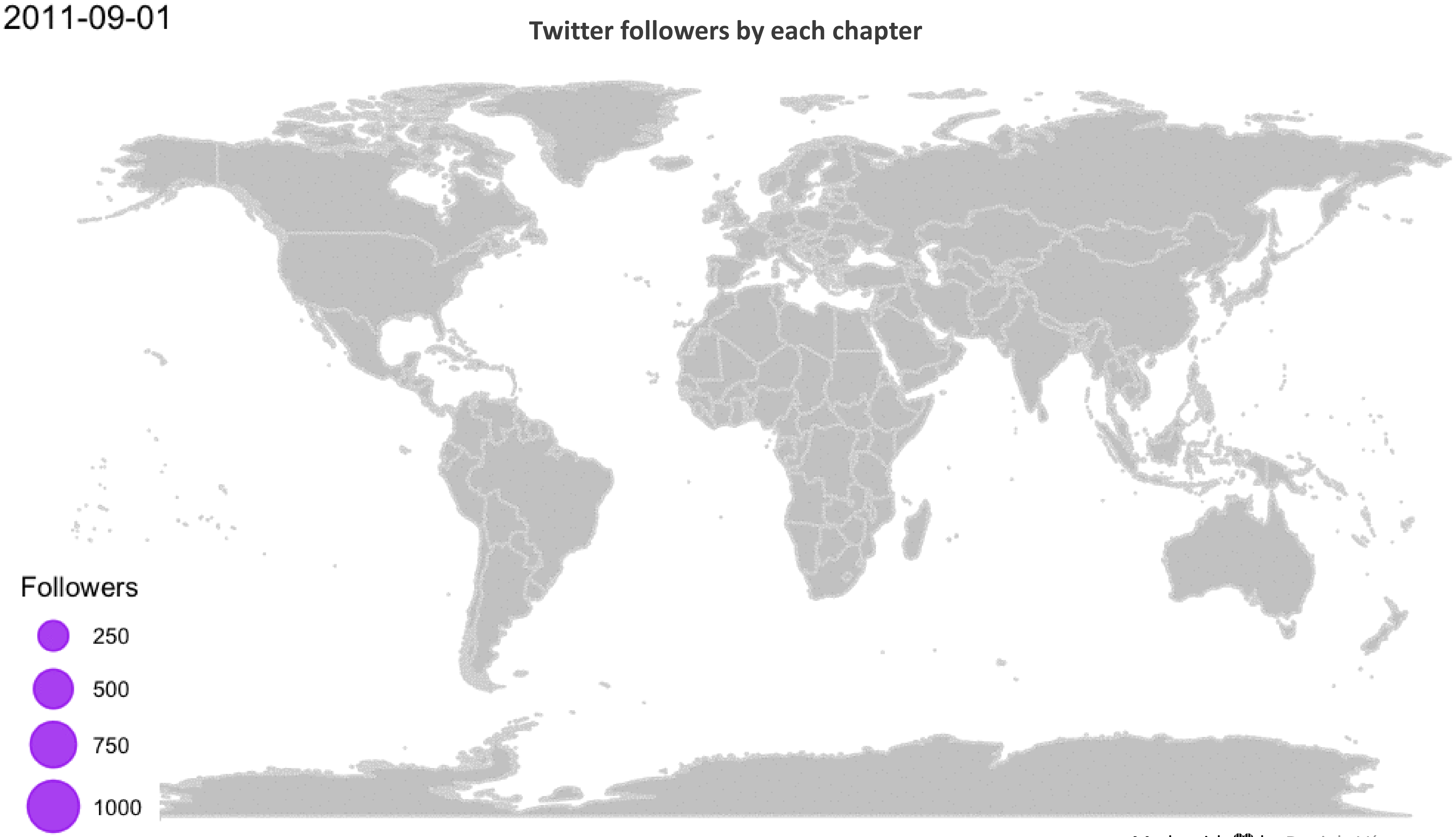
- Founded July 2017
- Near 250 Rladiers

14% R package developers are female

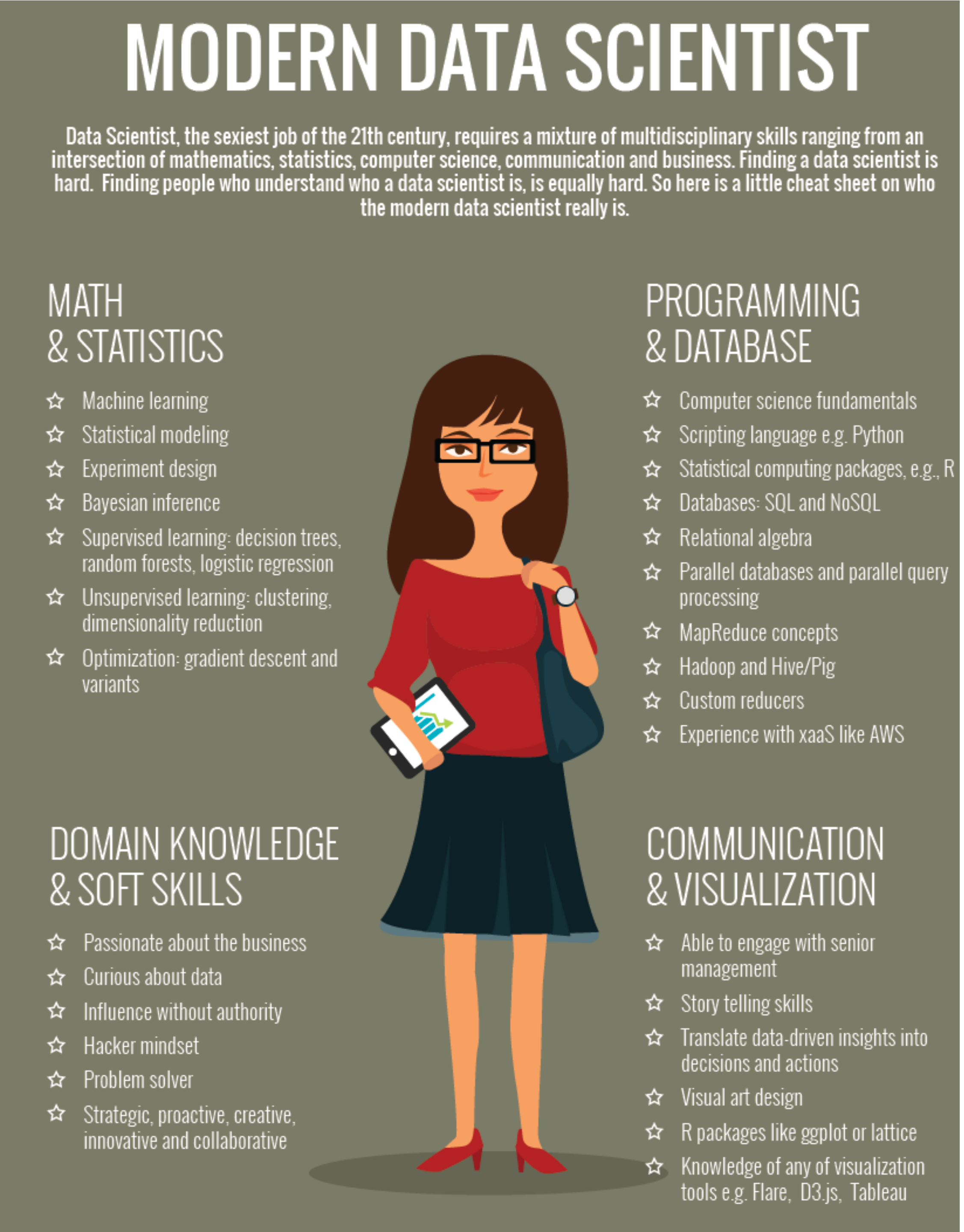
26% Only make up data professionals

20% Of all tech startups across the world are founded by women

Twitter followers by each chapter



Made with ♥ by Daniela Vázquez



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

”A data scientist is better at statistic than a software engineer and is better at software engineering than a statistician“

Who ia Data scientist?

- **Passionate** about data
- **Thinker**
- A **listener** to understand business problem
- A good communicator and **storyteller**
- Have **technical skills** for coding and analyzing data
- **Obsess** with solving problems

Data scientists

- Use analytics & technical skills to extract, analyze and model data

R julia tableau python Spark MATLAB Azure Machine Learning Qlik

Statistician

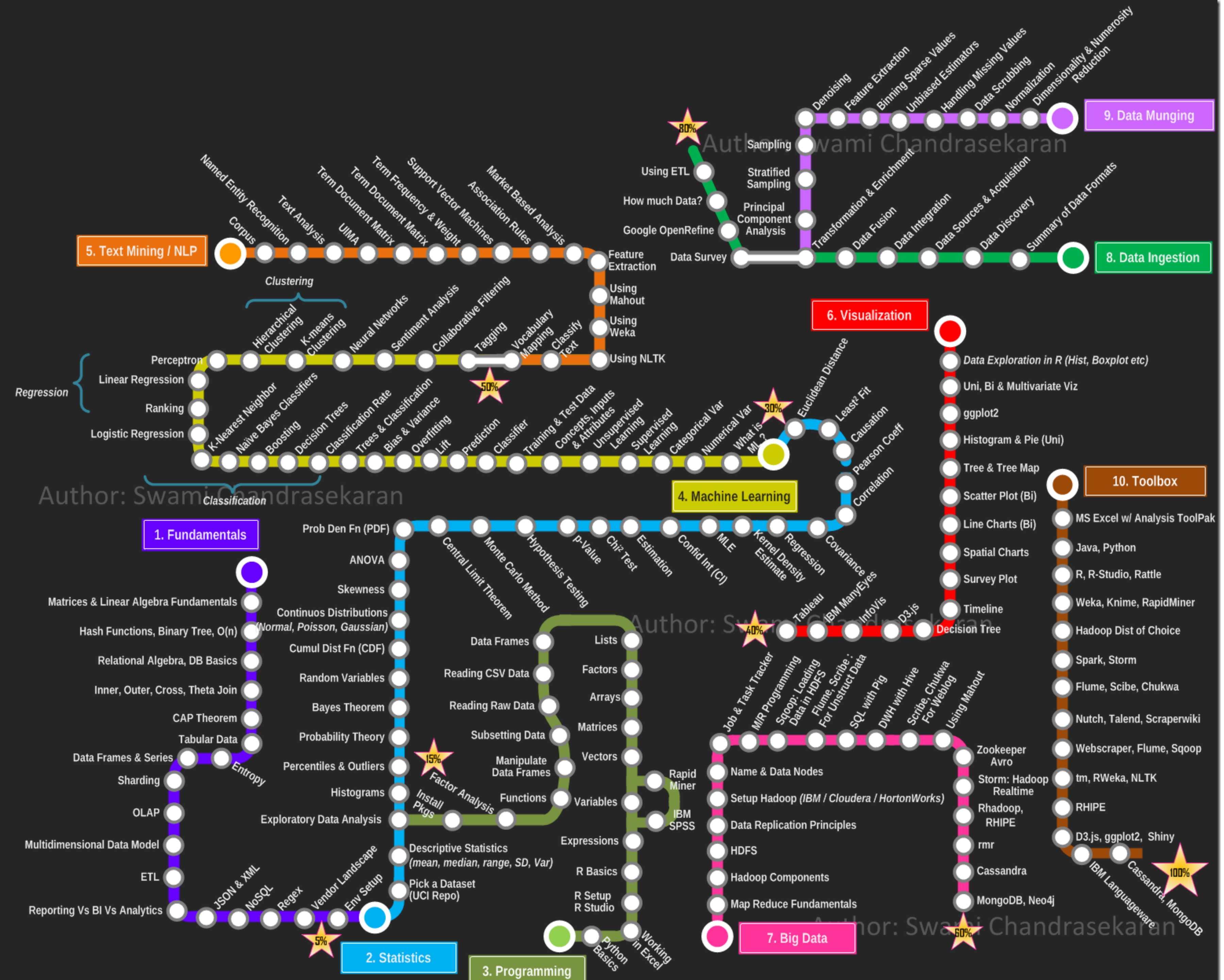
- Understand statistic and apply it on real problems

R MATLAB sas

Data engineer

- Responsible for architecture of data
- Ensure the flow of data within servers & applications

Spark mongoDB cassandra cloudera MapReduce



1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%

2. Statistics
Pick a Dataset (UCI Repo)
Descriptive Statistics (mean, median, range, SD, Var)
Exploratory Data Analysis
Histograms
Percentiles & Outliers
Probability Theory
Bayes Theorem
Random Variables
Cumul Dist Fn (CDF)
Continuos Distributions (Normal, Poisson, Gaussian)
Skewness
ANOVA
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least² Fit
Euclidean Distance
30%

3. Programming
Python Basics
Working in Excel
R Setup
R Studio
R Basics
Expressions
Variables
IBM SPSS
Rapid Miner
Vectors
Matrices
Arrays
Factors
Lists
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Factor Analysis
Install Pkgs
15%

4. Machine Learning
What is ML?
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Classification Rate
Decision Trees
Boosting
Naive Bayes Classifiers
K-Nearest Neighbor
Classification
Logistic Regression
Ranking
Linear Regression
Regression
Perceptron
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
50%

5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Classify Text
Vocabulary Mapping

6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau
40%

7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%

8. Data Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%

9. Data Munging
Dimensionality & Numerosity Reduction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis

10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%

Be ready for the journey!

Learning in data science is not a linear path!

- Data science is **not about just one specific skills**
- You must know all things and **be expert in one** of them
- It is a **fast evolving field**
- It does **not need** to be **expert in the domain**

Where to start learning or add more skill sets to your toolbox?

Online courses

DataCamp

coursera

edX

UDACITY

MITOPENCOURSEWARE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

fast.ai

Datasets for practices

kaggle

UCI
Machine Learning Repository

MLCOMP

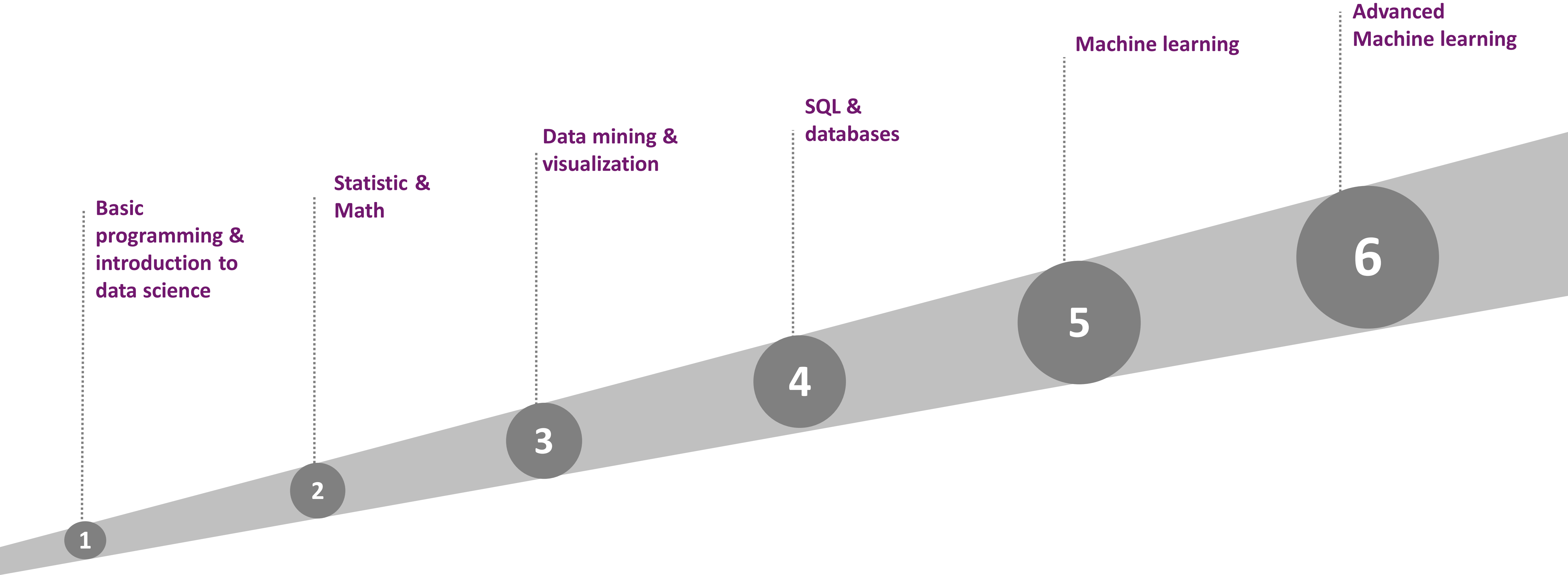
UCI KDD Archive

Google Trends Datastore

CrowdFlower

DATA.GOV.UK Beta
Opening up Government

There is a starting point but no end road!



1. Basic programming & introduction to data science
2. Statistic & Math
3. Data mining & visualization
4. SQL & databases
5. Machine learning
6. Advanced Machine learning

Beginners steps

Basic programming & introduction to data science

- R
- Python

- Benefits & potential of the languages
- Answer possible questions in the field
- How to code and use the basics such as libraries, functions,..
- How to load & read & manipulate data

DataCamp

Introduction to R

Intro to Python for Data Science

coursera

R Programming

Introduction to Data Science in Python

edX

Introduction to Python: Absolute Beginner

Introduction to R for Data Science

Beginners steps

Statistic & Math

- Learn fundamental concepts of statistics such as p-values, variance, correlation, statistical hypothesis..
- Evaluate various types of data and how to interpret their structure
- How to apply various statistical methods on the data

DataCamp

A Hands-on Introduction to Statistics with R

Intro to Statistics with R: Introduction

coursera

Statistical Inference

Introduction to Probability and Data

edX

Statistics and R

Introductory Statistics : Basic Ideas and Instruments for Statistical Inference

Beginners steps

Data mining & visualization

- Handle anomalies in data such as missing values, outliers,..
- Explore correlations among variables
- Apply feature engineering on the data
- Create graph & visualization to demonstrate findings in the data

DataCamp

Data Visualization with ggplot2

Introduction to Data Visualization with Python

coursera

Exploratory Data Analysis

Data Management and Visualization

edX

Analyzing and Visualizing Data with Power BI

Data Analysis: Visualization and Dashboard Design

Beginners steps

SQL & databases

- Understand how relational databases are working
- How to interact with databases for fetching, saving and manipulation of data

DataCamp

Intro to SQL for Data Science

Introduction to Databases in Python

coursera

Using Databases with Python

Managing Big Data with MySQL

edX

Querying Data with Transact-SQL

Beginners steps

Machine learning

- Learn various algorithms such as regression, random forest, classification tree, etc. and their concepts and understand where to use them
- How to apply predictive modeling on set of data
- Learn how to train various model and what are the metrics of trained models and how to compare them

DataCamp

Machine Learning Toolbox

Introduction to Machine Learning

coursera

Machine Learning

Practical Machine Learning

edX

Applied Machine Learning

Principles of Machine Learning

Intermediate & advanced steps

Advanced machine learning

- Learn how to manipulate unstructured data
- Learn and understand advanced topics such as deep learning, social network analysis, text mining, time series processing,..

DataCamp

Text Mining: Bag of Words

Manipulating Time Series Data in R with xts & zoo

coursera

Deep Learning Specialization

Practical Time Series Analysis

edX

Deep Learning Explained

Graph Algorithms

Being specialist in one field?

Other topics

- Depends on the field, business and type of problems

DataCamp

Building Web Applications in R with Shiny

Network Analysis in R

Building Chatbots in Python

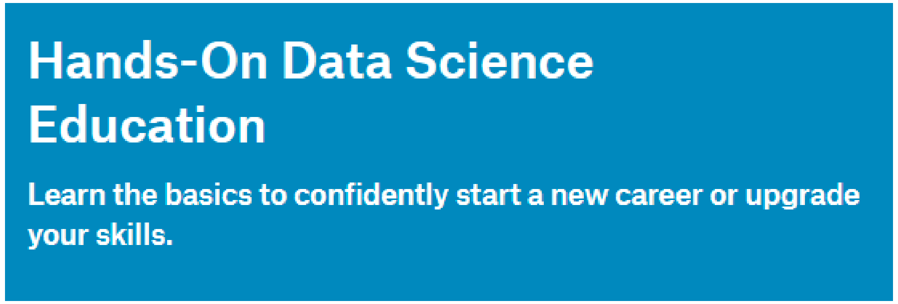
Credit Risk Modeling in R

coursera

Bayesian Methods for Machine Learning

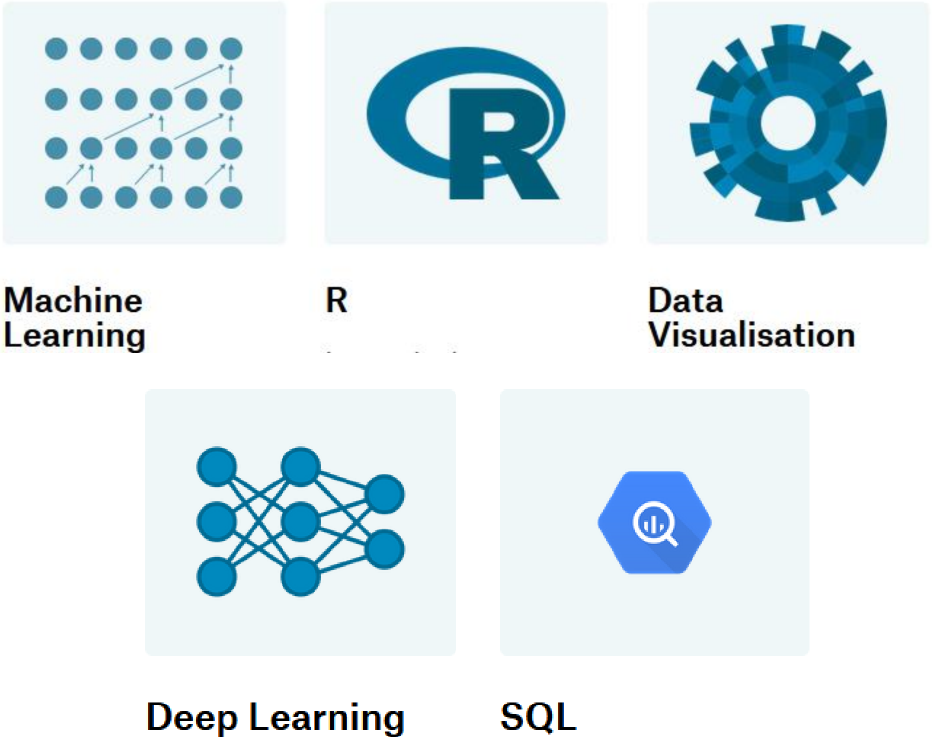
Practical Reinforcement Learning

Kaggle learning



Hands-On Data Science Education

Learn the basics to confidently start a new career or upgrade your skills.



Machine Learning

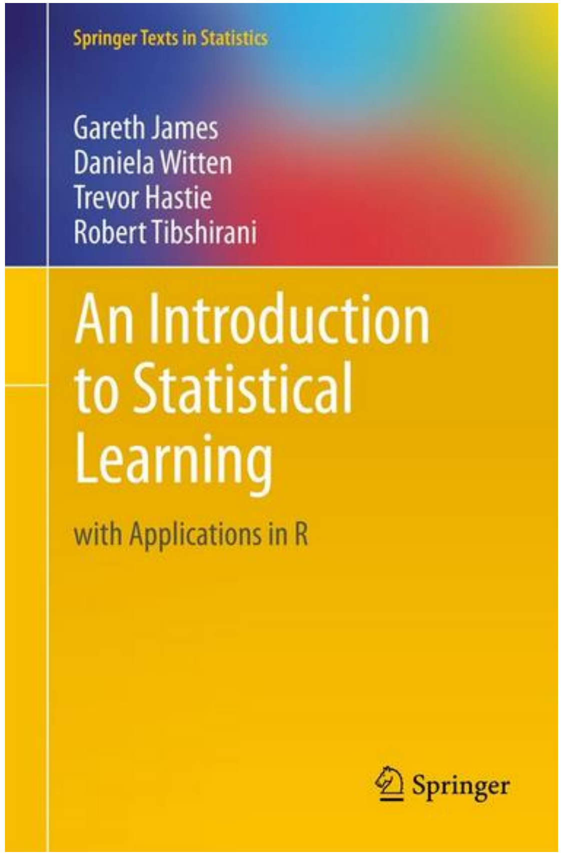
R

Data Visualisation

Deep Learning

SQL

If you have to read just one single book / or watch just 15 hours videos



Springer Texts in Statistics

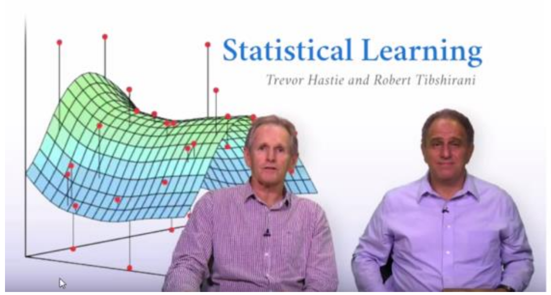
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

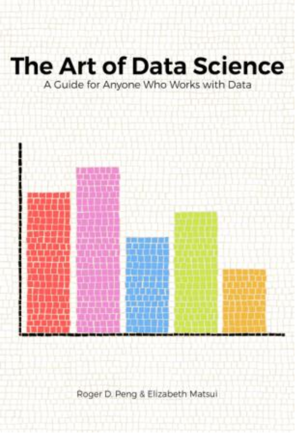
with Applications in R

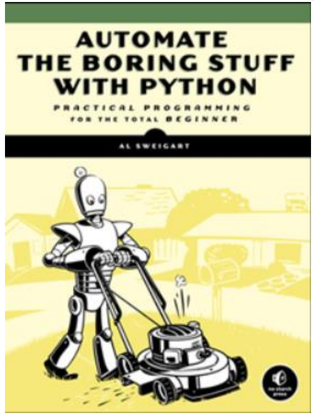
Springer

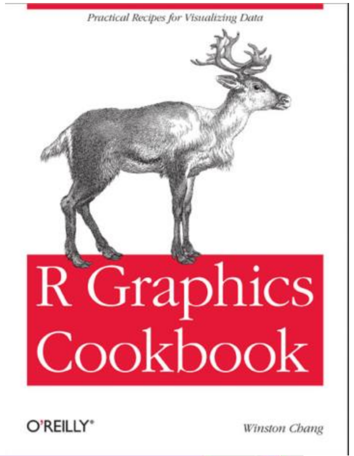
S Statistical Learning

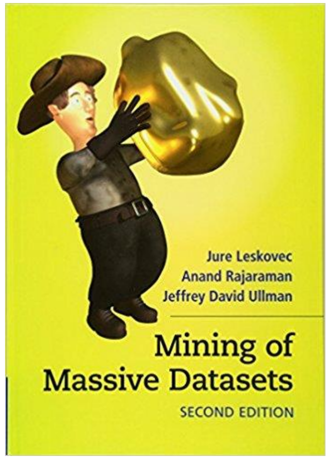


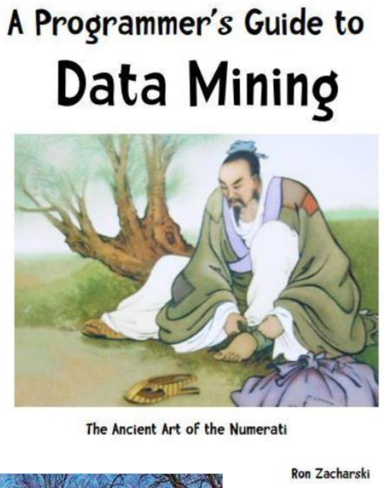
Free books

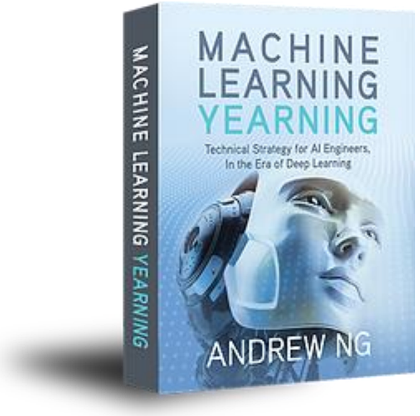

The Art of Data Science
A Guide for Anyone Who Works with Data
Roger D. Peng & Elizabeth Matsui


AUTOMATE THE BORING STUFF WITH PYTHON
PRACTICAL PROGRAMMING FOR THE TOTAL BEGINNER
AL SWEIGART

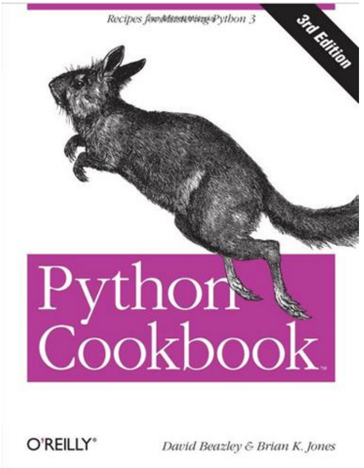

Practical Recipes for Visualizing Data
R Graphics Cookbook
O'REILLY
Winston Chang

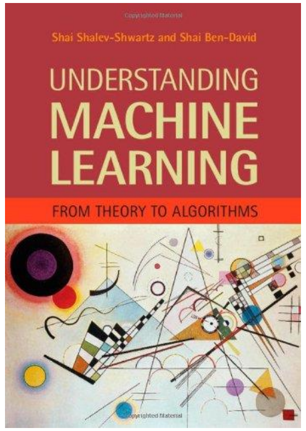

Jure Leskovec
Anand Rajaraman
Jeffrey David Ullman
Mining of Massive Datasets
SECOND EDITION

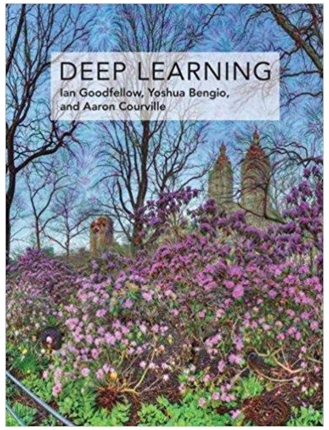

A Programmer's Guide to
Data Mining
The Ancient Art of the Numerati
Ron Zacharski


MACHINE LEARNING YEARNING
Technical Strategy for AI Engineers, In the Era of Deep Learning
ANDREW NG
MACHINE LEARNING YEARNING

Probability and Statistics for Programmers
Think Stats
O'REILLY
Allen B. Downey


Recipes for Mastering Python 3
3rd Edition
Python Cookbook
O'REILLY
David Beazley & Brian K. Jones


Shai Shalev-Shwartz and Shai Ben-David
UNDERSTANDING MACHINE LEARNING
FROM THEORY TO ALGORITHMS


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Websites

KDnuggets

R-bloggers

No Free Hunch

BANANA DATA

Revolutions

Daily news about using open source R for big data analysis, predictive modeling, data science, and visualization since 2008

Analytics Vidhya

Learn everything about analytics

Data Science Central

FastML

simplystats

visualising data

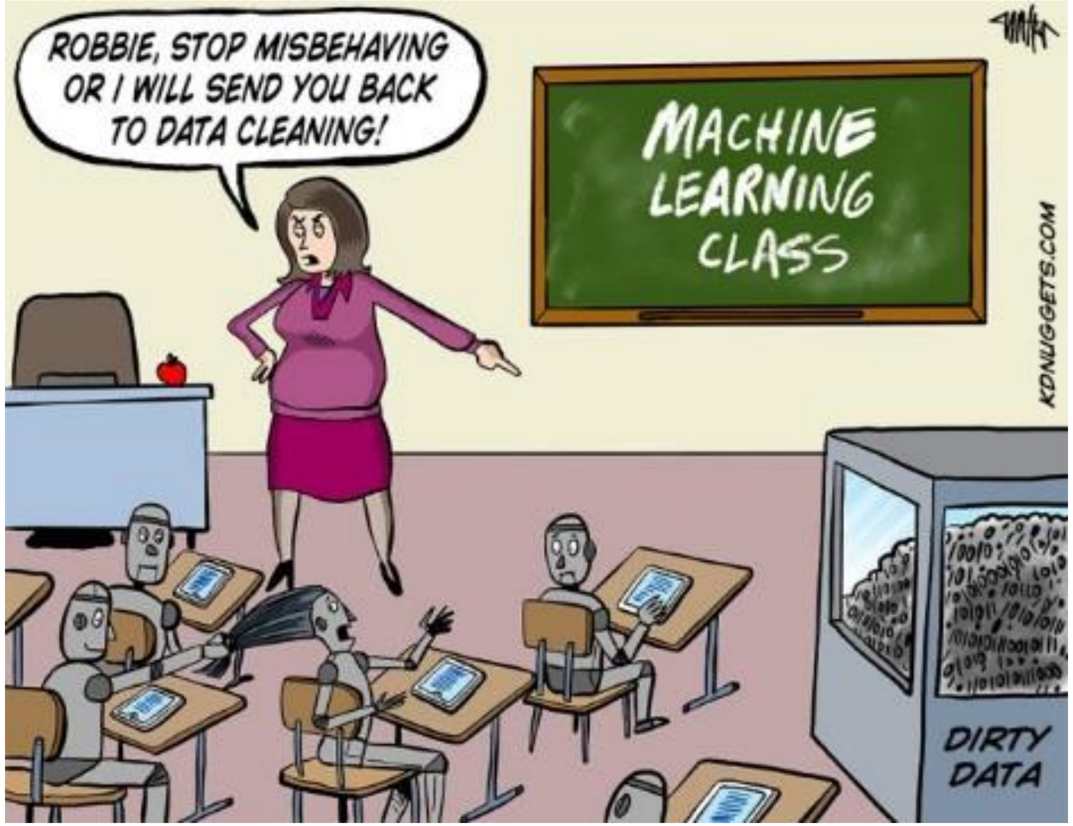
information is beautiful

ŷhat

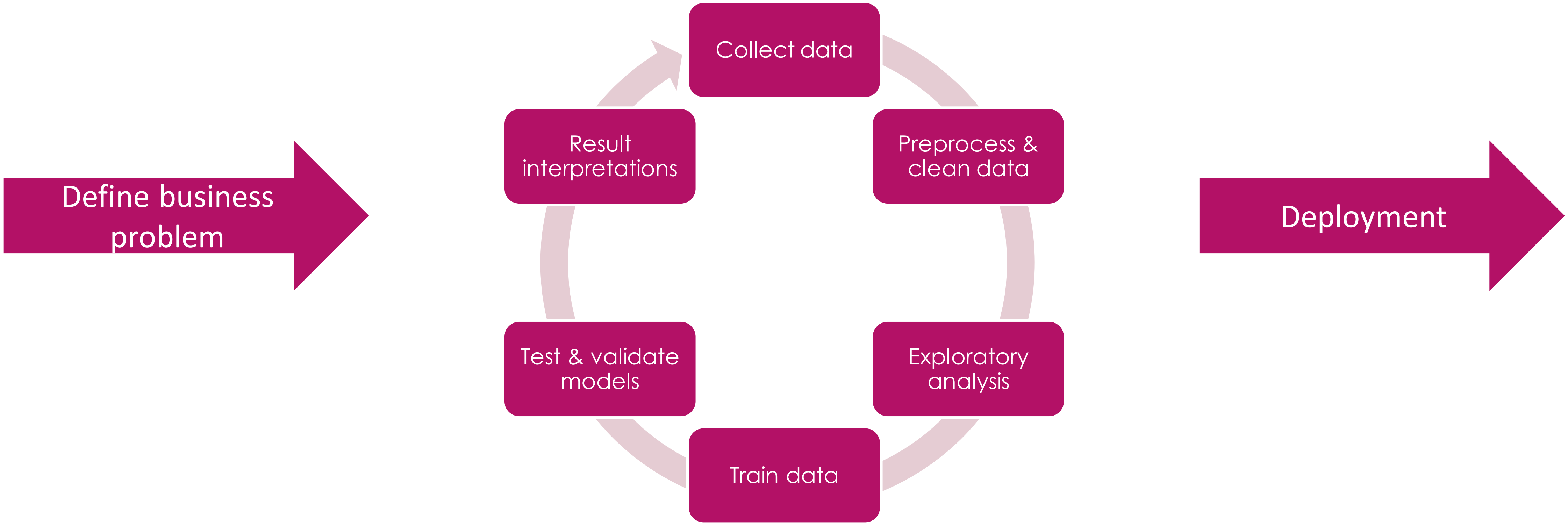
You don’t have any experience?

- Do the action!
- Apply what you learn on some use cases from Kaggle or any other data set that you can
- Do the competitions in Kaggle, share the code in blogs, GitHub account, LinkedIn,...
- You can find internship, asking companies directly
- Ask mentors to follow you in the way

SMILE! OUR AI WILL TAKE A PICTURE OF YOUR FACE AND WILL KNOW EXACTLY WHAT YOU WOULD LIKE FOR DINNER TONIGHT.
KDNUGGETS.COM


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Ok, I’m all set, but how can I tackle a real problem?



Define business problem

Collect data

Preprocess & clean data

Exploratory analysis

Train data

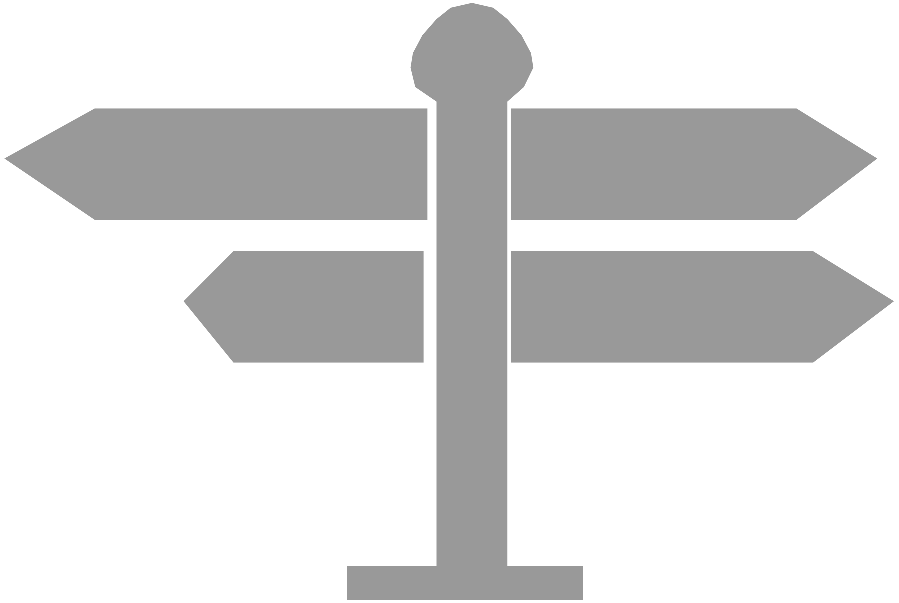
Test & validate models

Result interpretations

Deployment

Define the business problem

- Listen to business's need
- You must be a good listener

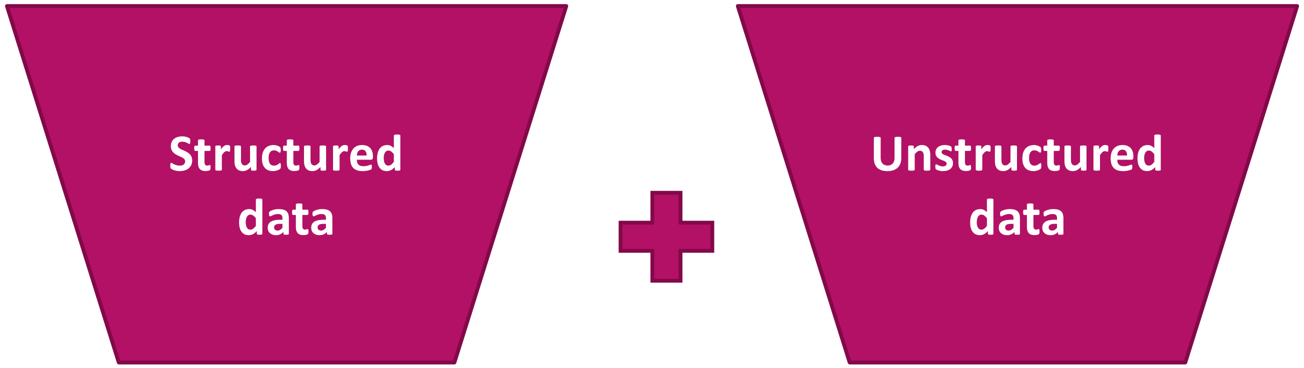


- The business you are facing do not know their problem yet? That is even fine too!
- A lot of time, they don't have any idea where to start

Collect data

Data are in a structured format such as database, csv file,...

Data are in a unstructured format such as text, images, audio and video



Structured data

+

Unstructured data

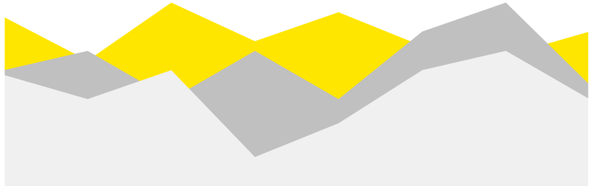
Near 80% of data in organization are unstructured

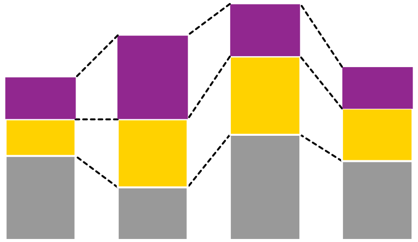
Preprocessing & cleaning data

- **Never trust** your data
- **Check consistency & structure** of the data to be sure about its quality
- Fill **missing values**
- **Be aware of noisy data** and treat them well!
- Check **outliers** & remove them
- Apply **data transformation** such as normalization or aggregation

Exploratory analysis

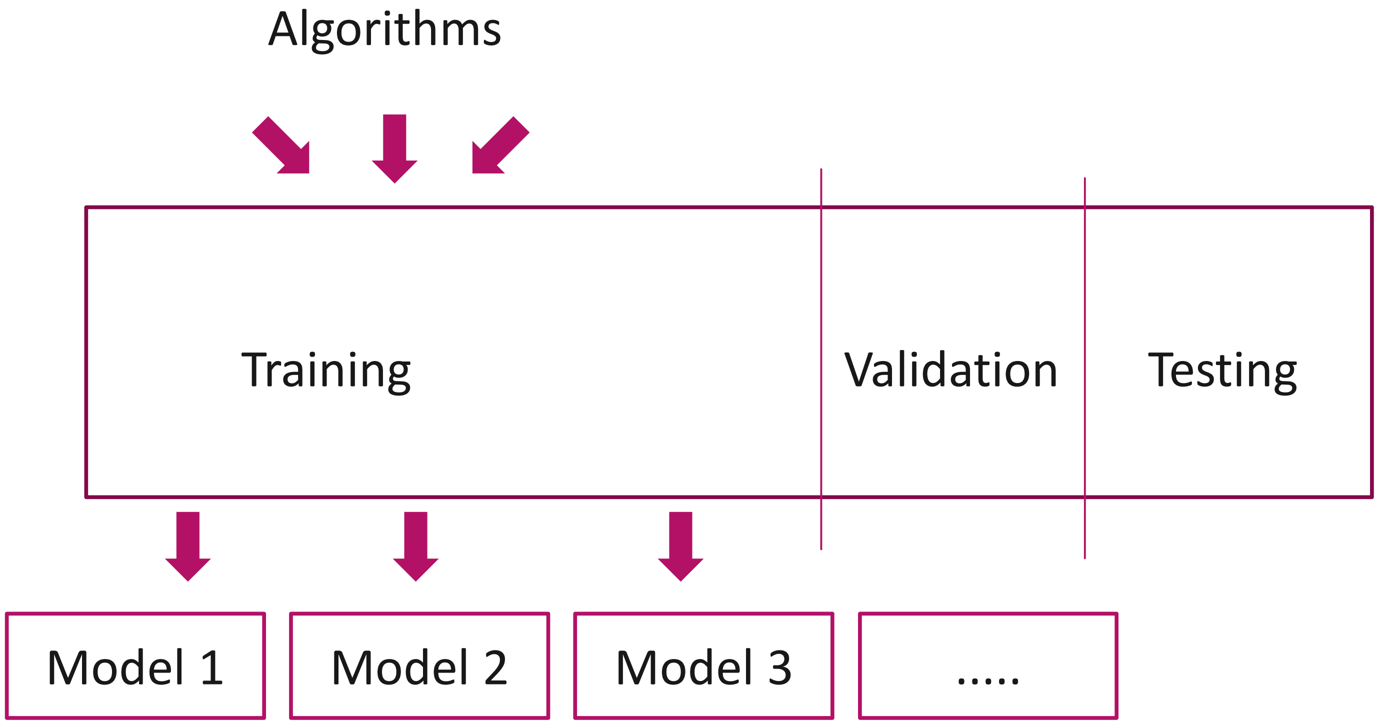
- **Understand and summarize the** data
- **Explore** for insights & hidden patterns
- Investigate the correlations among various variables
- Create some **hypothesis** based on the findings
- **Understand which model & technique to** use for analyzing the data
- **Feature engineering** can happen in this stage too





Think about the questions that you are going to answer

Model training, testing & validation



Algorithms

Training

Validation

Testing

Model 1

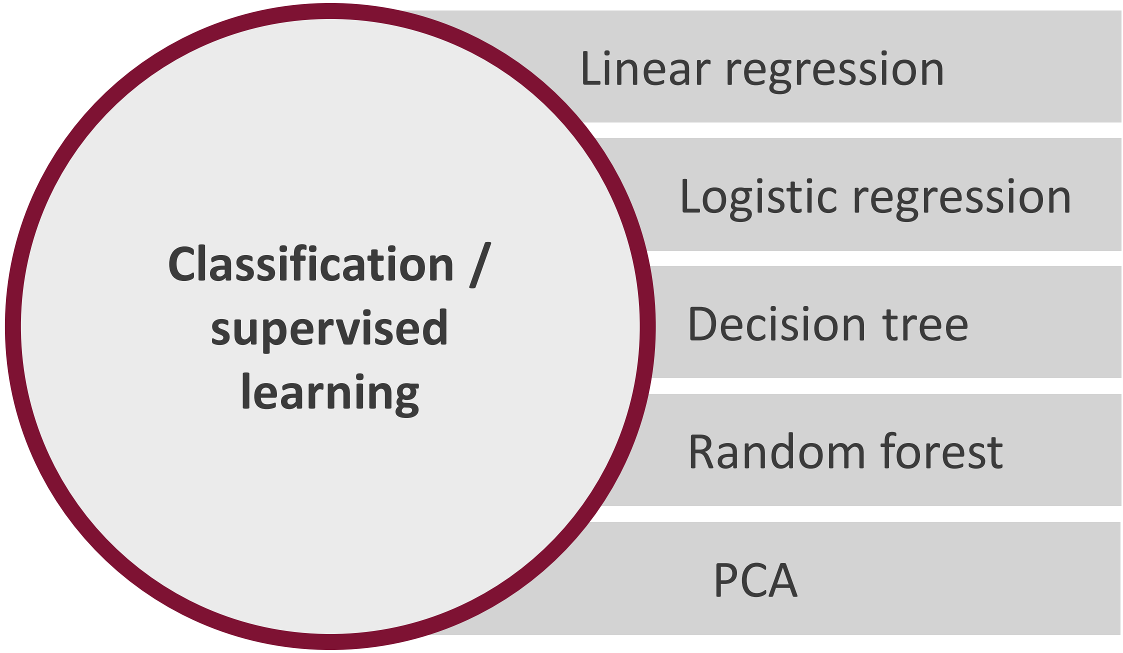
Model 2

Model 3

.....

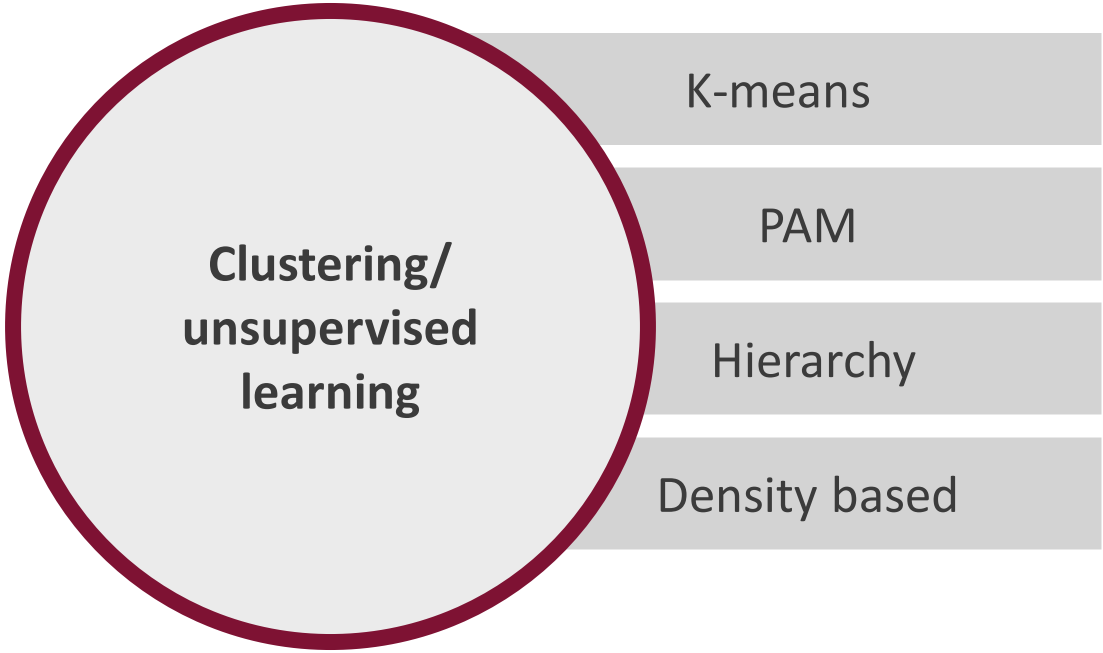
60%, 20%, 20%

Algorithms?



Classification / supervised learning

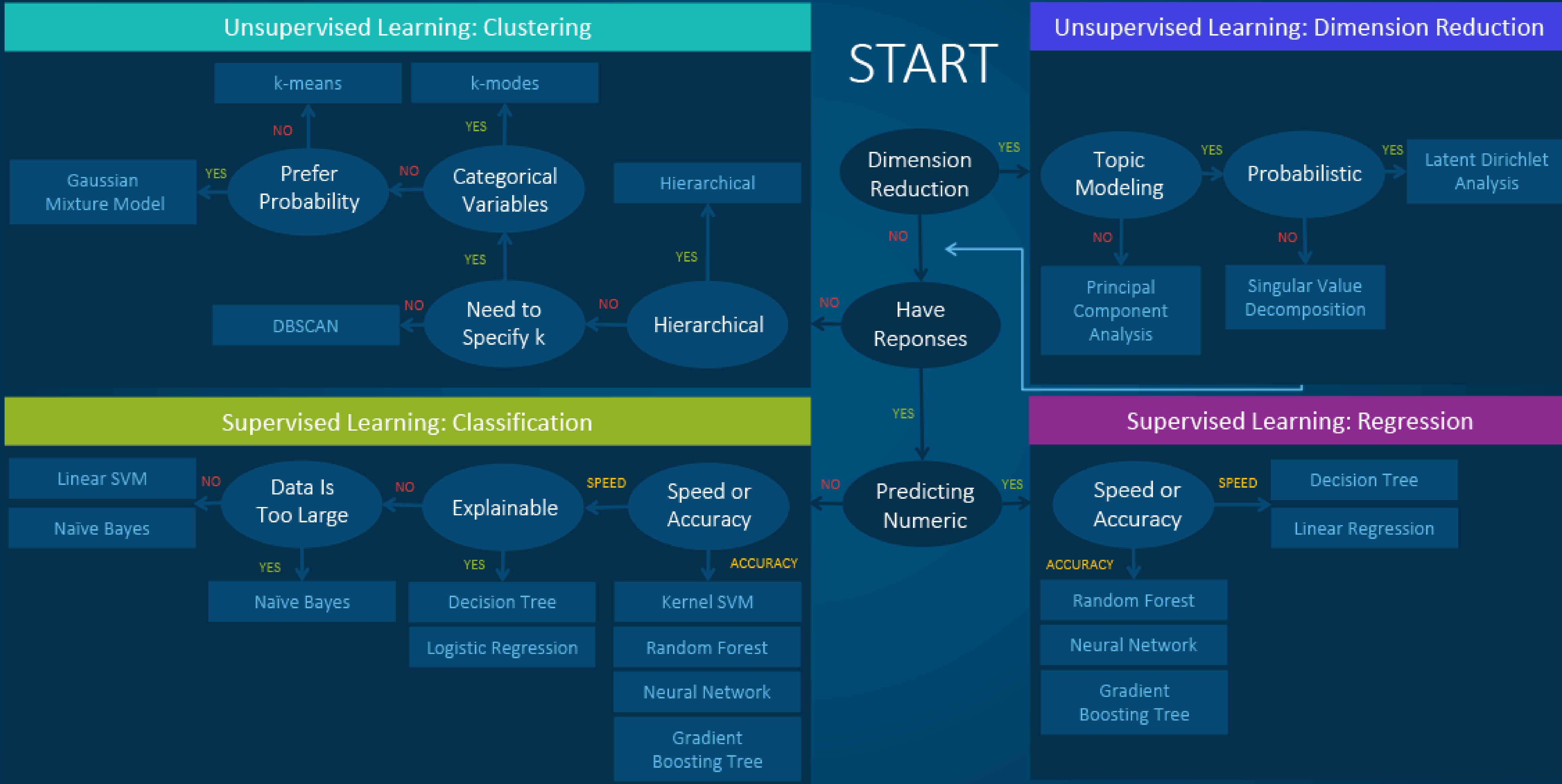
- Linear regression
- Logistic regression
- Decision tree
- Random forest
- PCA



Clustering/ unsupervised learning

- K-means
- PAM
- Hierarchy
- Density based

Machine Learning Algorithms Cheat Sheet



START

Dimension Reduction

YES

NO

Have Reponses

NO

YES

Predicting Numeric

NO

YES

Unsupervised Learning: Clustering

k-means

k-modes

NO

YES

Gaussian Mixture Model

YES

Prefer Probability

NO

Categorical Variables

Hierarchical

YES

YES

DBSCAN

NO

Need to Specify k

NO

Hierarchical

Unsupervised Learning: Dimension Reduction

Topic Modeling

YES

Probabilistic

YES

Latent Dirichlet Analysis

NO

NO

Principal Component Analysis

Singular Value Decomposition

Supervised Learning: Classification

Linear SVM

Naïve Bayes

NO

Data Is Too Large

NO

Explainable

SPEED

Speed or Accuracy

YES

YES

ACCURACY

Naïve Bayes

Decision Tree

Logistic Regression

Kernel SVM

Random Forest

Neural Network

Gradient Boosting Tree

Supervised Learning: Regression

Speed or Accuracy

SPEED

Decision Tree

Linear Regression

ACCURACY

Random Forest

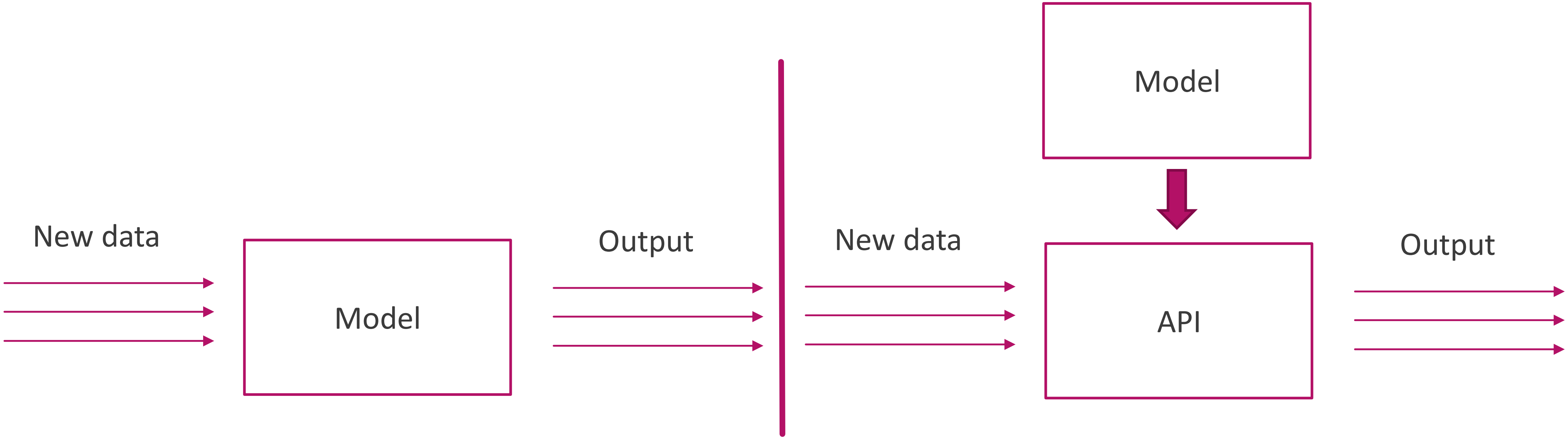
Neural Network

Gradient Boosting Tree

Deployment

Batch

Real-time streaming



"More data beats better models. Better data beats more data." — Riley Newman

“Enjoy your journey!”