Machine Learning in Python

Rickard Sjögren, Scientist @ Corporate Research, Sartorius AG

UmeJUG, March 27, 2019

What is machine learning?

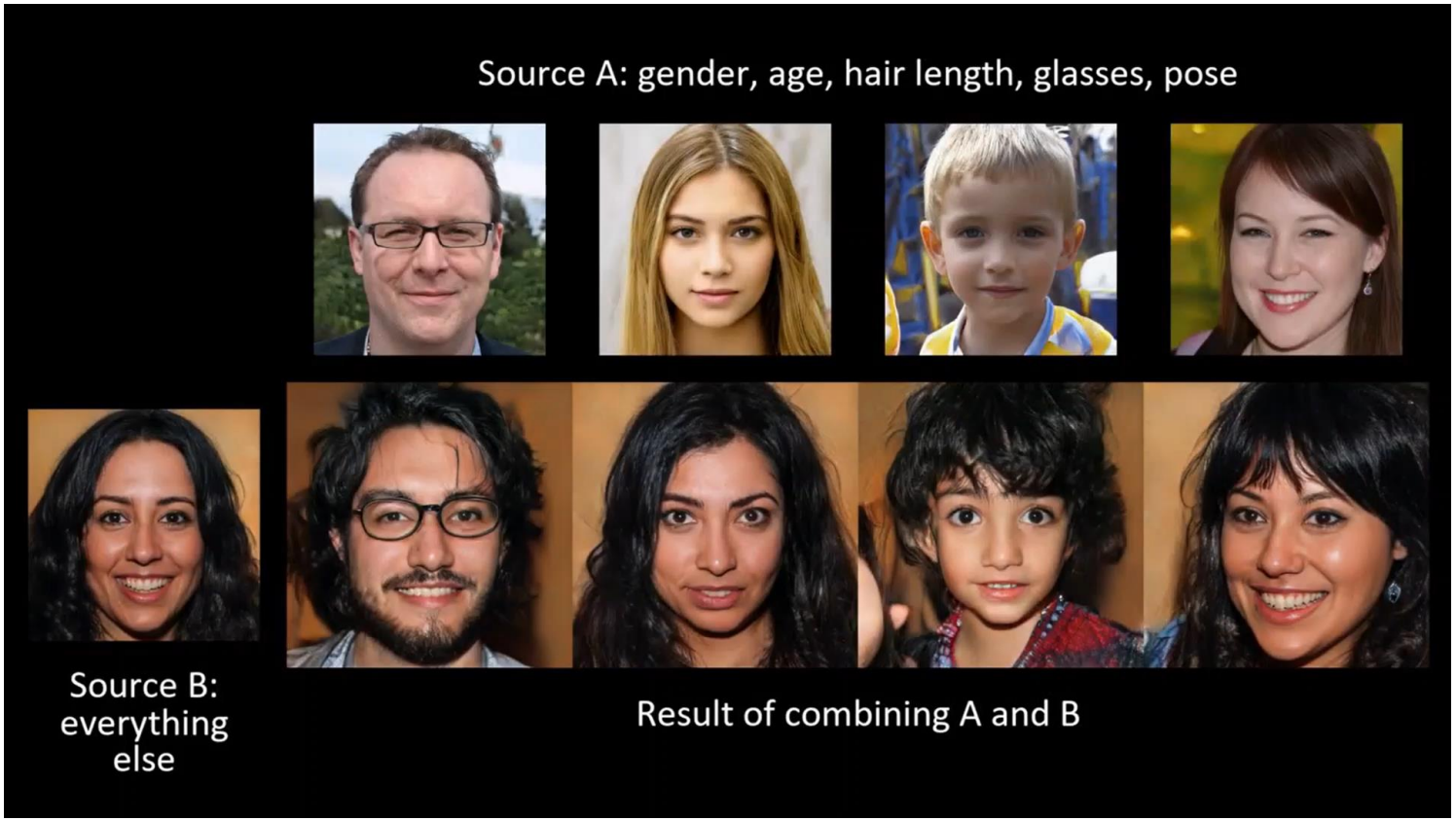
"Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead”

- https://en.wikipedia.org/wiki/Machine_learning

What is machine learning?

How it's branded

What is machine learning?



What you can do

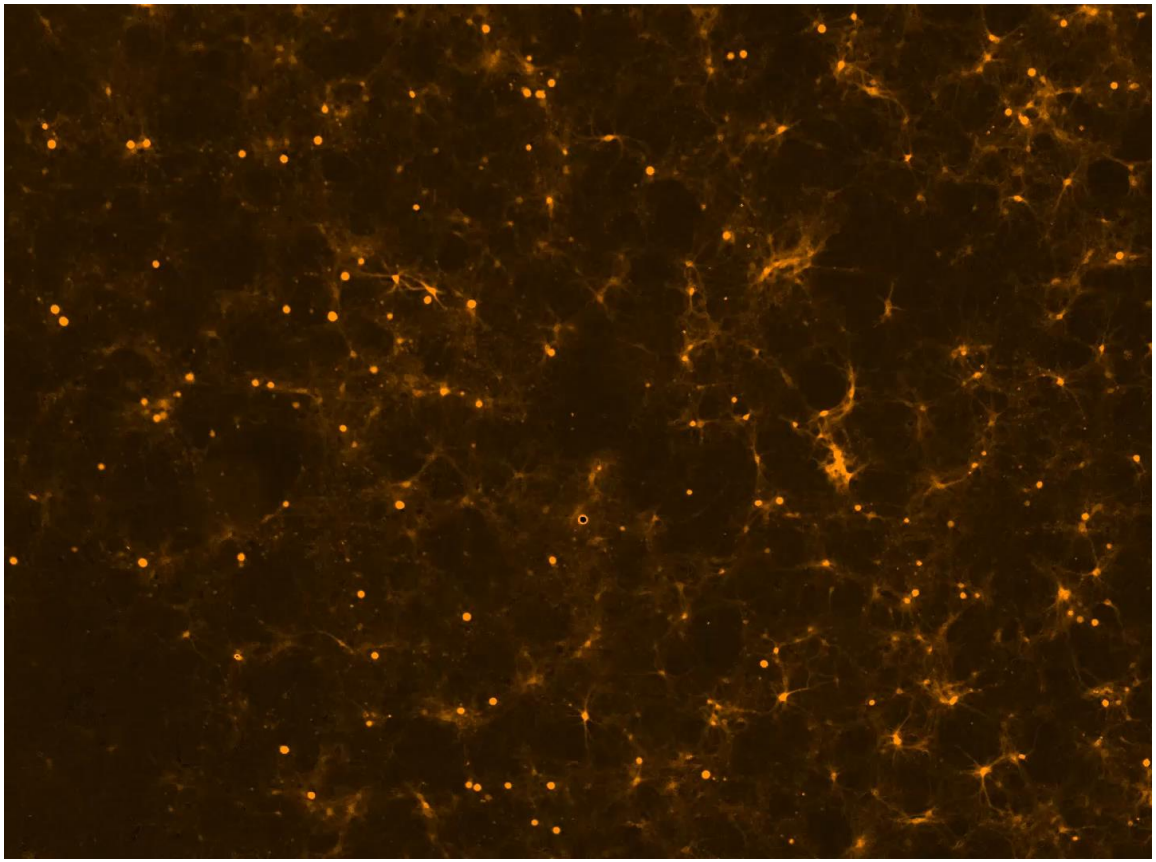
Karras, Tero, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." *arXiv preprint arXiv:1812.04948* (2018).

What is machine learning?

What you can do

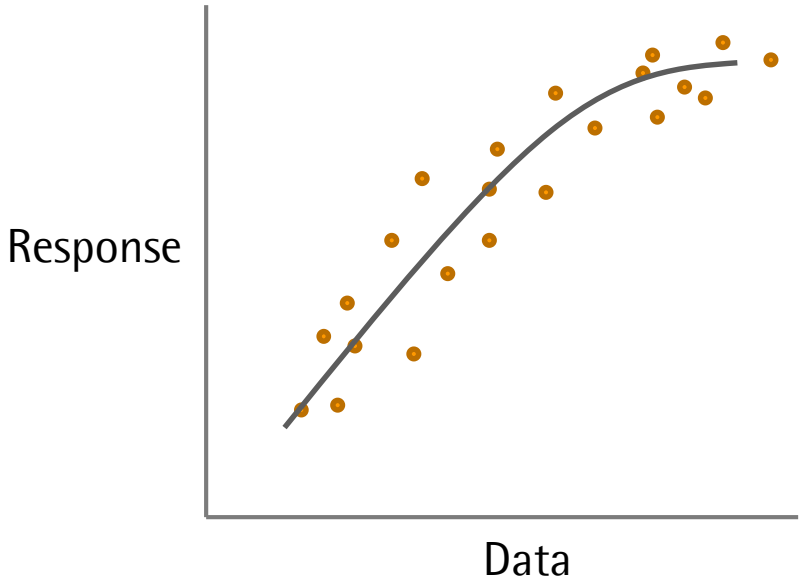
Peng, X. B., Kanazawa, A., Malik, J., Abbeel, P., & Levine, S. (2018, December). SFV: reinforcement learning of physical skills from videos. In *SIGGRAPH Asia 2018 Technical Papers* (p. 178). ACM.

What is machine learning?



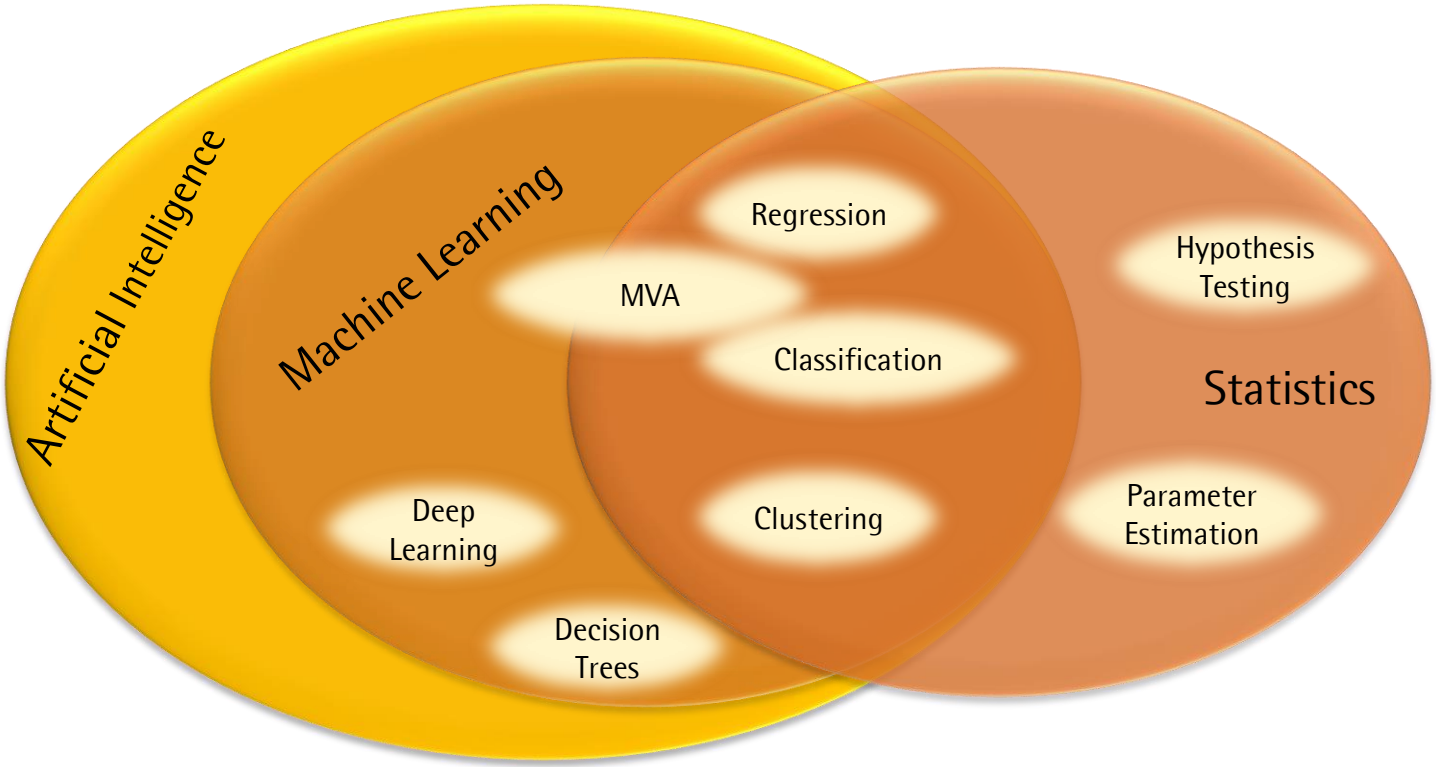
What you can do

What is machine learning?

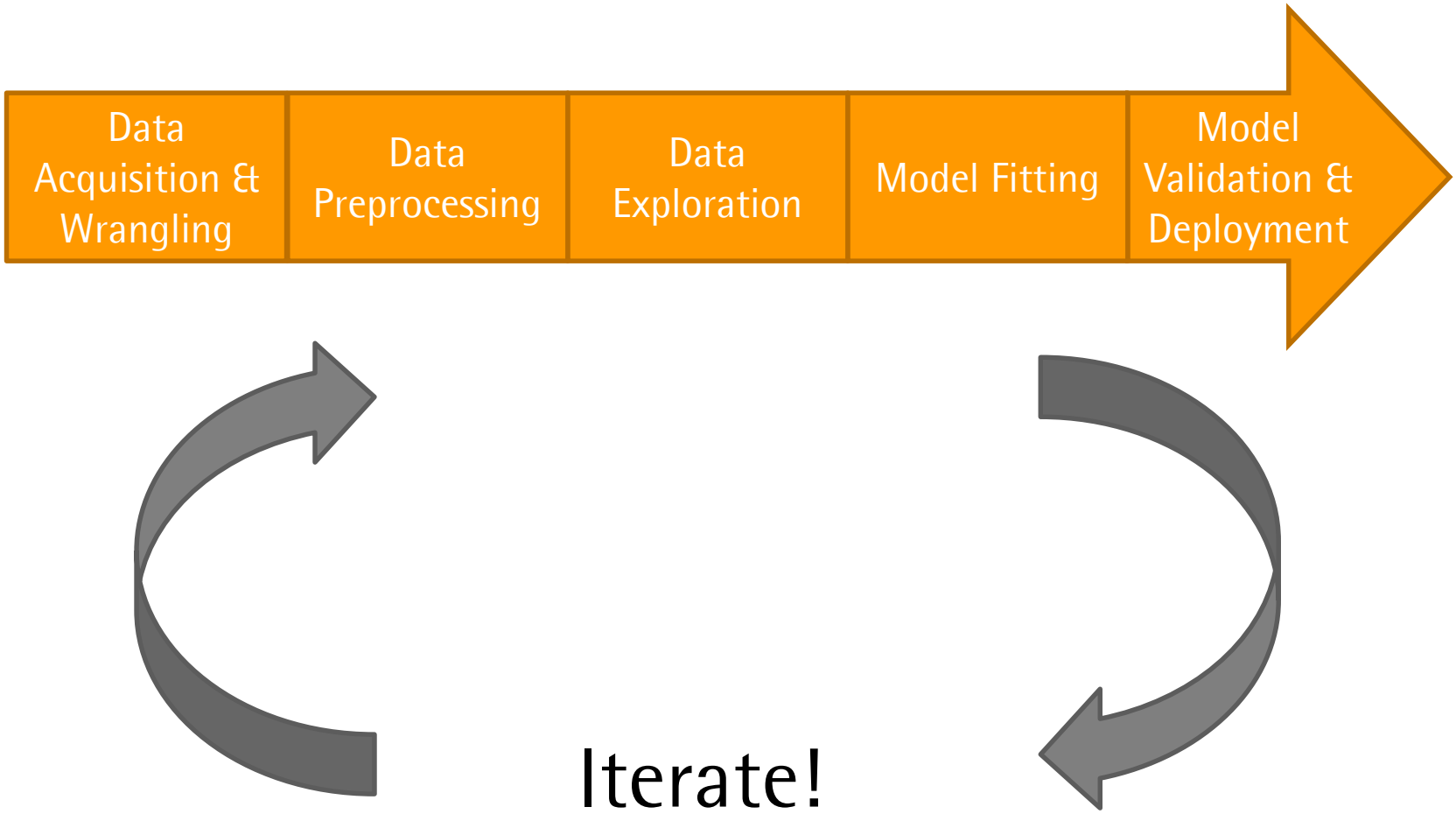


What you typically do

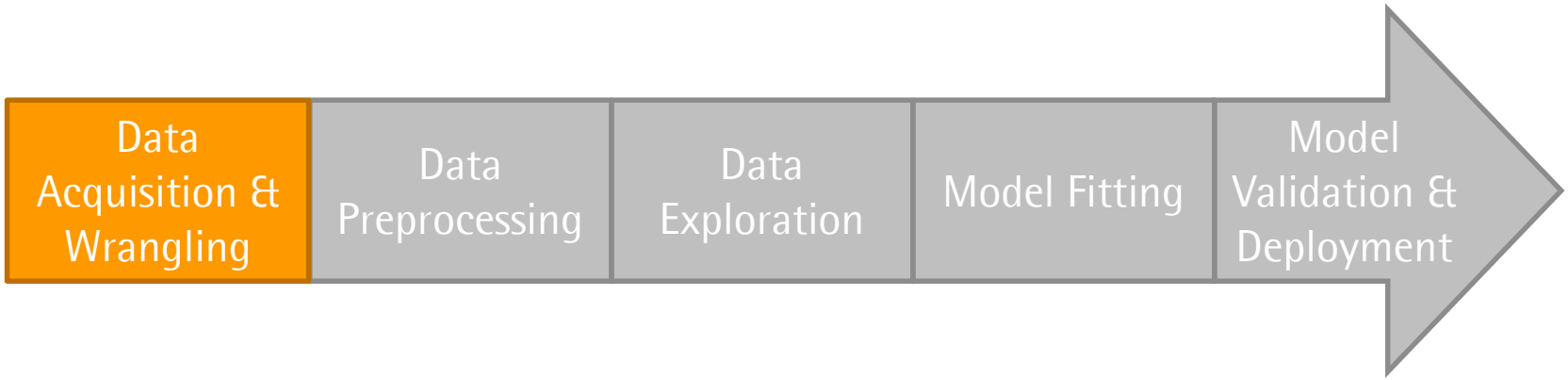
What is machine learning?



Machine Learning Process

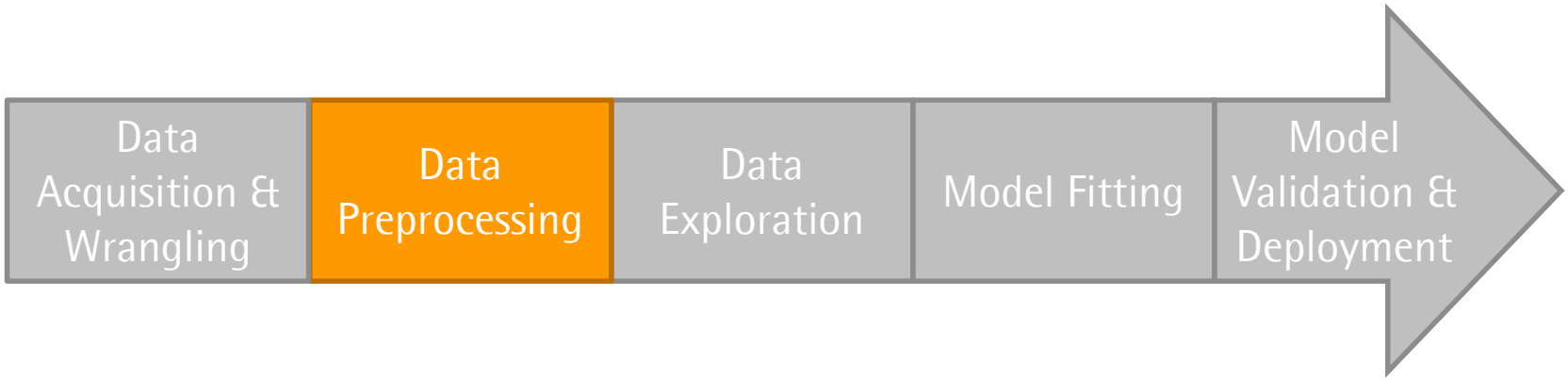


Machine Learning Process



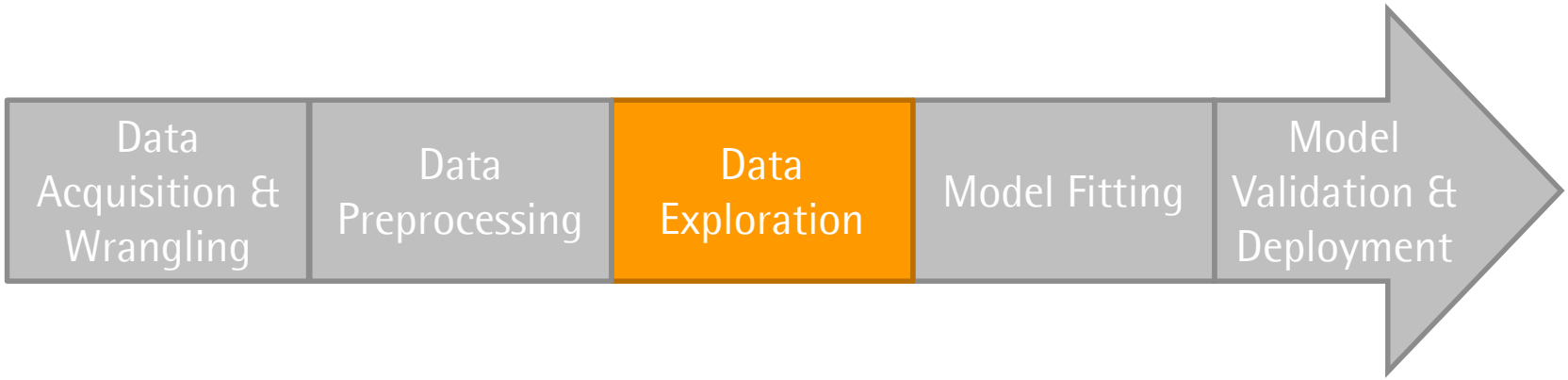
- Collect data and make it analyzable
  - Connect and query databases
  - Scrape and clean web-based text sources
- Time-consuming and often very manual
- Highly domain-specific

Machine Learning Process



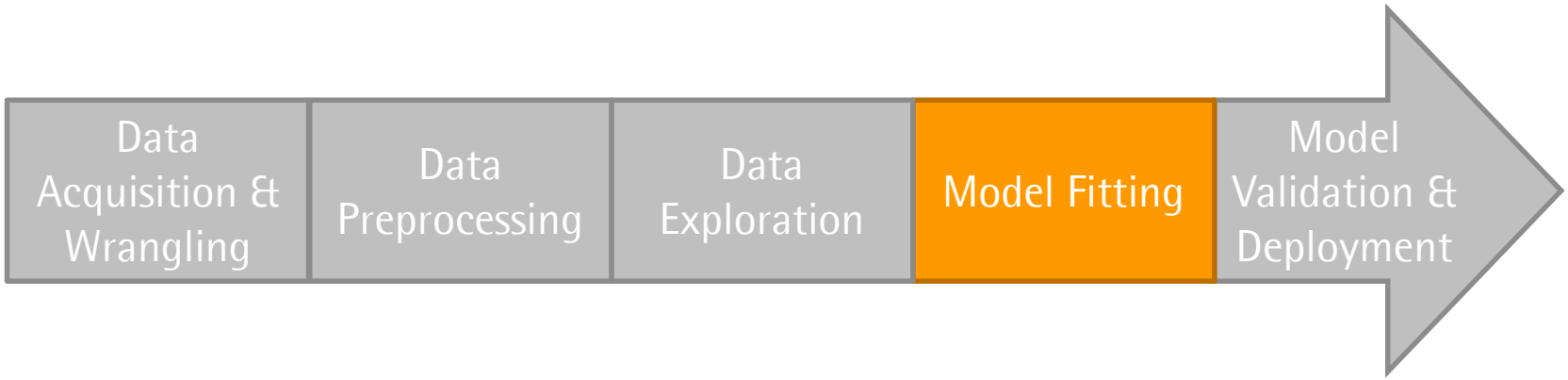
- "All you need is a table"
  - Transformation of unstructured data into structured
- Feature engineering
  - Find useful table representation

Machine Learning Process



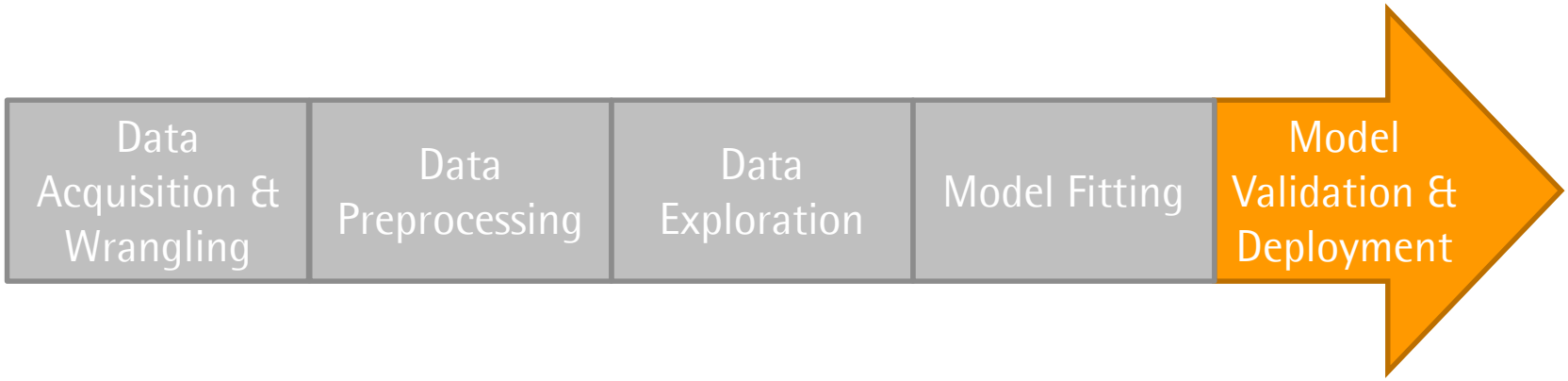
- Get to know your data
- Explore and visualize general trends
- Detect outliers and abnormal distributions

Machine Learning Process



- What most people think of as machine learning
- Algorithms that explain your relationship of interest
  - Support vector machines, Random Forests, Linear regression, Deep learning, Principal Component Analysis…

Machine Learning Process

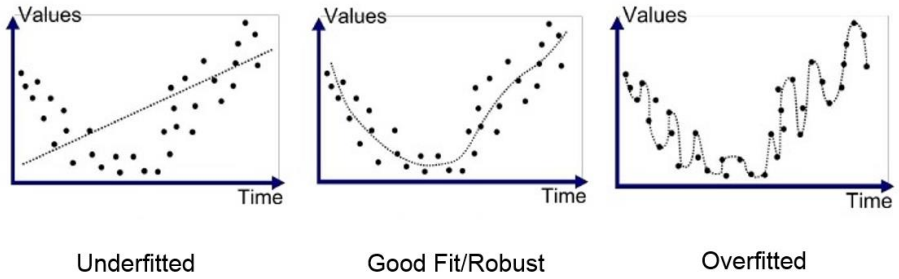


- Make sure that your model does what it's supposed to
- Cross-validation to find suitable model parameters
- Test-set evaluation to report performance to stakeholders

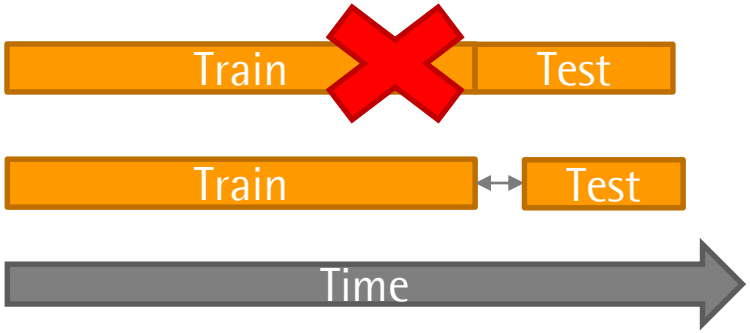
Why validation is so damn important

- You need an honest evaluation of what your model know
- A bit confusing nomenclature
  - Validation-set typically used to find good model parameters to balance over/under-fitting
  - Test-set used to asses real performance
- Beware of data leakage!
  - Preprocessing fitted for train- and test-data at the same time
  - Correlation between close time points in time series
  - Data from same person present in both training and test-set

https://en.wikipedia.org/wiki/Clever_Hans

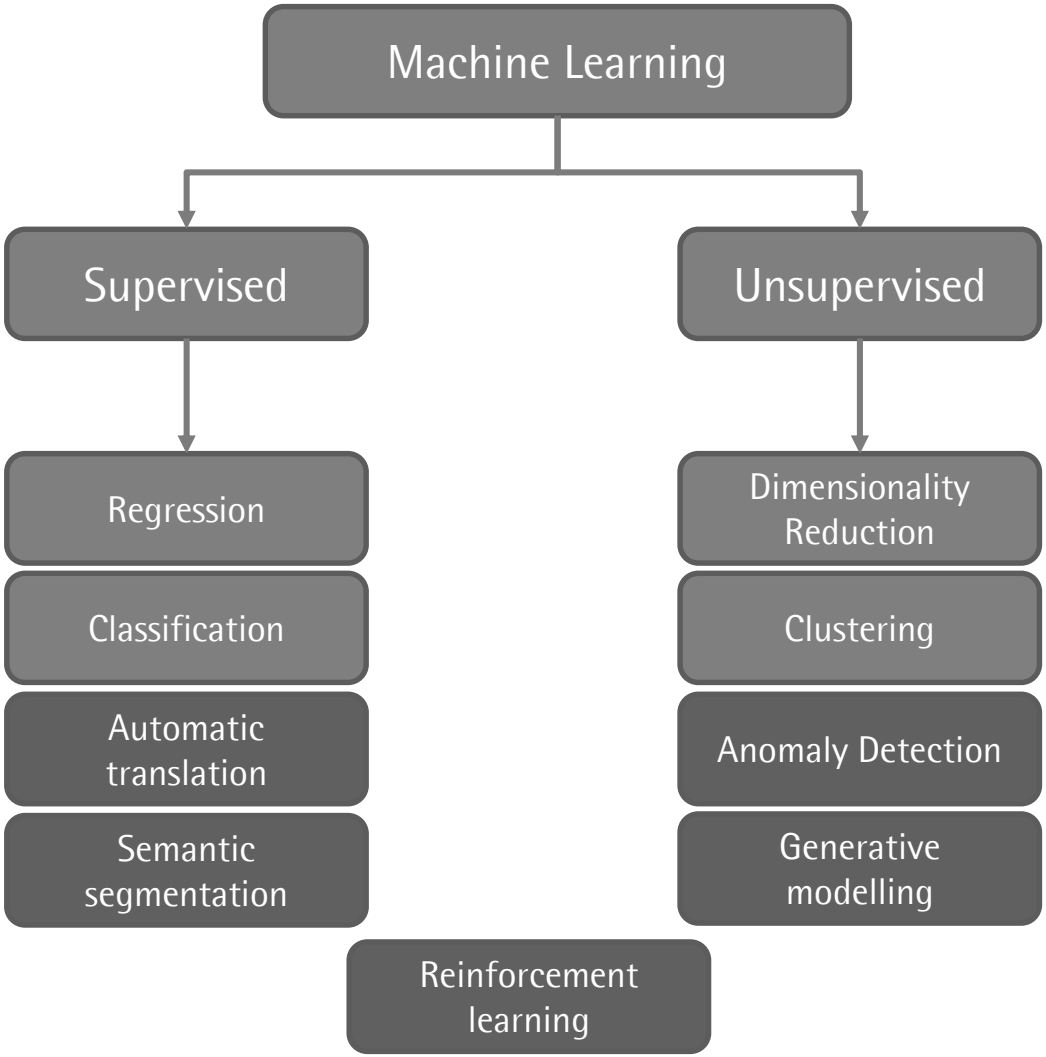


https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

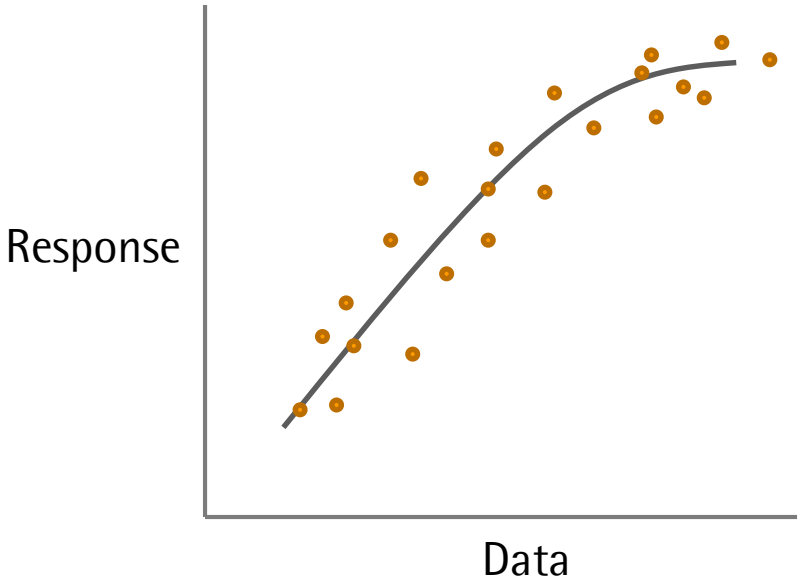


Machine Learning Methods

Machine Learning Models

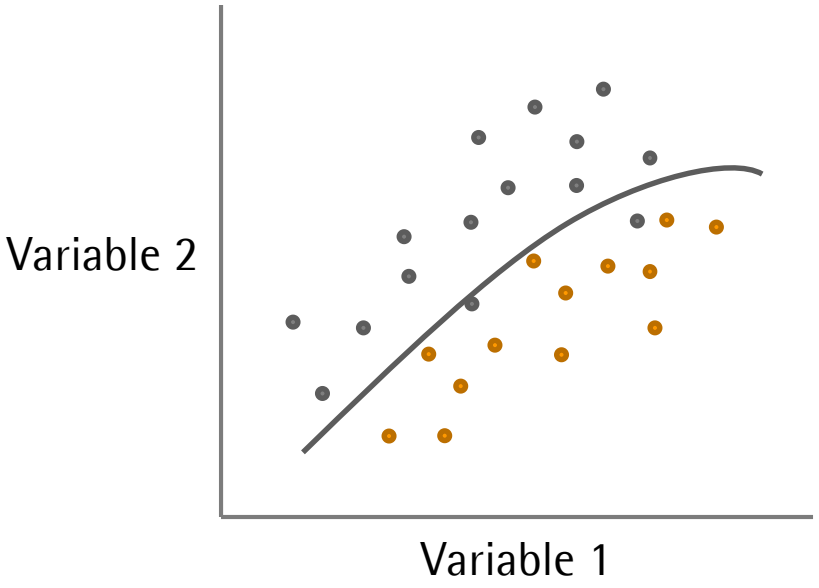


Supervised Machine Learning - Regression



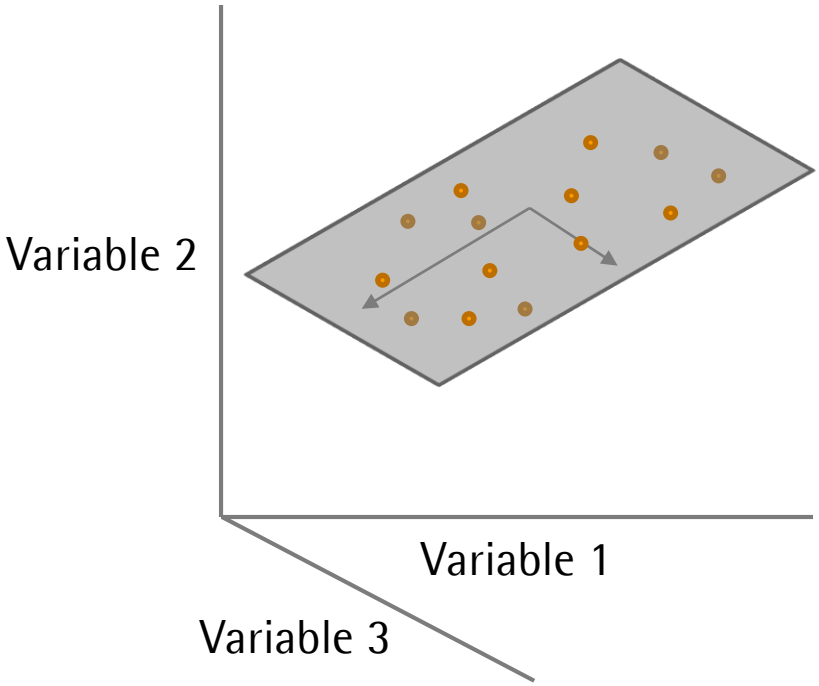
- Describe numerical relationship between variables
  - $Y \approx f(X, \beta)$
- Popular methods include
  - Linear regression
  - Partial Least Squares Regression (PLS)
  - Ridge or LASSO-regression
- Applications include
  - Business demand forecasting
  - Prediction of pharmaceutical binding affinity
  - Estimated future process yield

Supervised Machine Learning - Classification



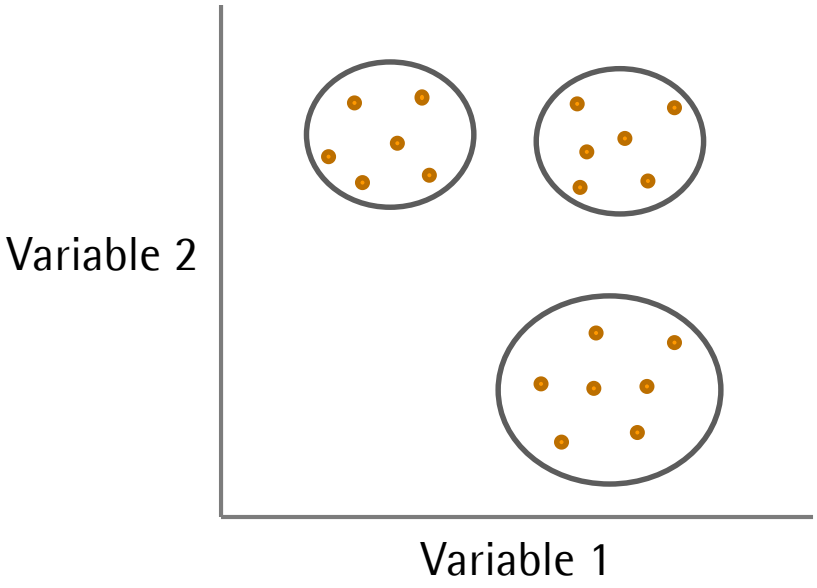
- Describe decision boundaries between classes
  - "Put data into buckets"
- Popular methods include
  - Support vector classifiers
  - Logistic regression
  - Random forests
- Applications include
  - Click stream analysis
  - Clinical case/control discrimination
  - Optical Character Recognition

Unsupervised Machine Learning – Dimensionality reduction



- Find descriptive representation of high-dimensional data
- Popular methods include
  - Principal Component Analysis (PCA)
  - Non-negative Matrix Factorization (NMF)
  - Uniform Manifold Approximation and Projection (UMAP)
- Applications include
  - Exploratory analysis of complex data
  - Data compression
  - Feature engineering

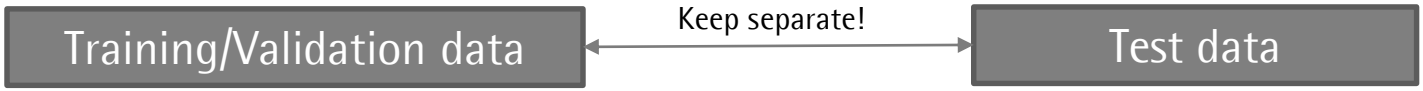
Unsupervised Machine Learning - Clustering

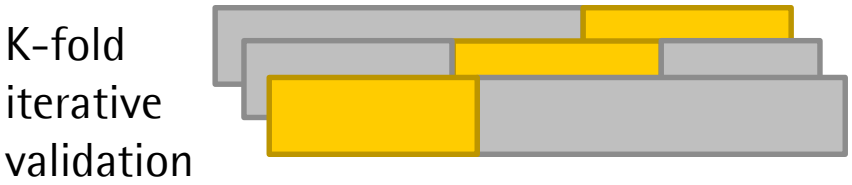


- Find groups in data
  - "Create buckets"
- Methods include
  - K-Means clustering
  - DBSCAN
  - Hierarchical /Agglomerative Clustering
- Applications include
  - Exploratory analysis
  - Customer segmentation
  - Cancer subtype detection

Model Validation

- How to choose train-val-test split?
  - If you don't know anything, randomize
  - Stratify based on prior knowledge to reduce risk of bias
  - Fixed or k-fold training/validation depends on size of data
- Double-check for dependencies between training- and test-data
- Use as much data in test set as you can afford

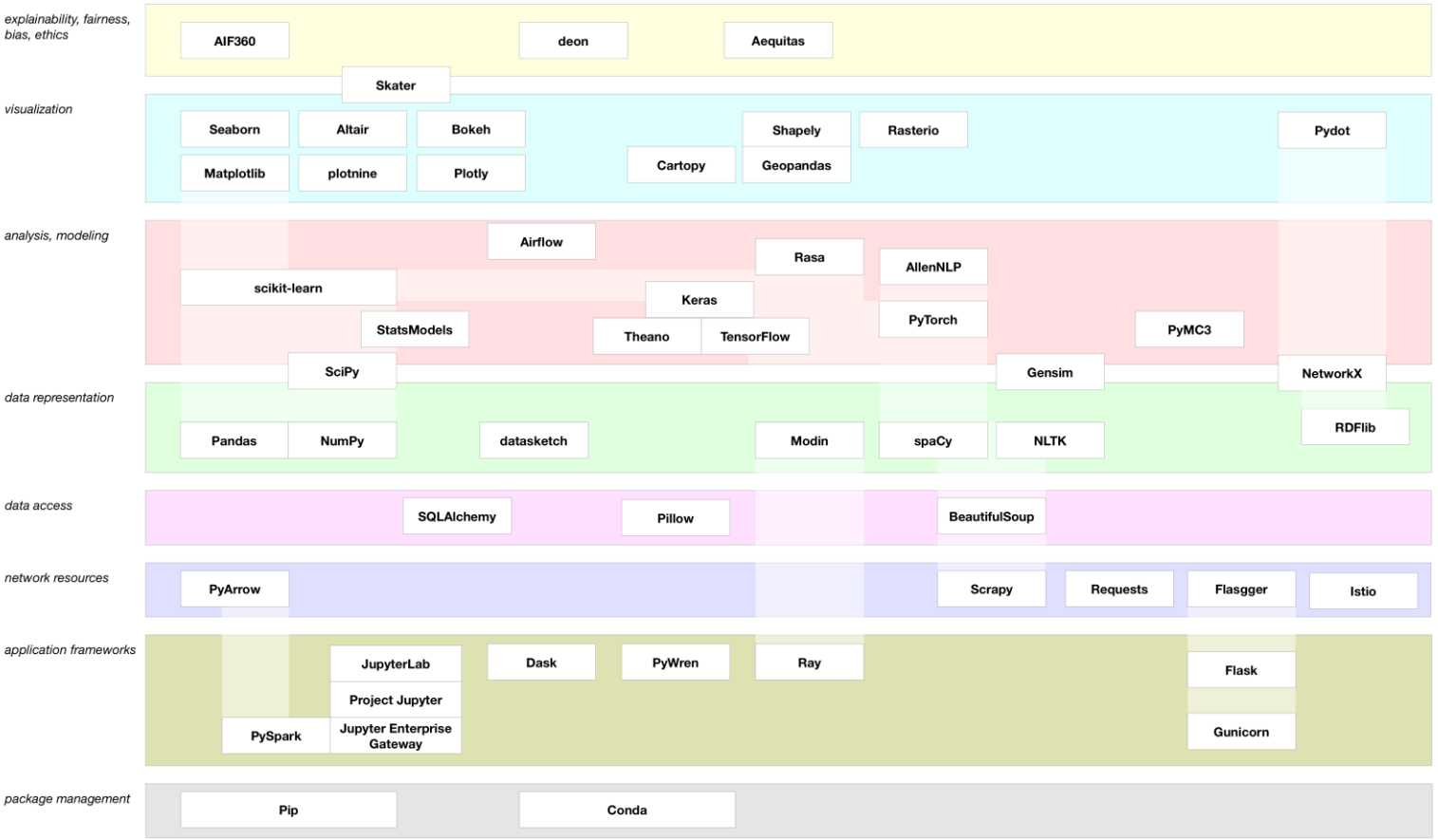




sartorius

Machine Learning in Python

Why Python?



https://community.ibm.com/community/user/datascience/blogs/paco-nathan/2019/03/12/a-landscape-diagram-for-python-data

NumPy

- Fundamental package for scientific computing.
  - N-dimensional array object
  - broadcasting functions
  - tools for integrating C/C++ and Fortran code
  - Linear algebra, Fourier transform, and random number functionality

- Most numerical Python libraries built are on top of NumPy

```
import numpy as np
```

http://www.numpy.org/

Pandas

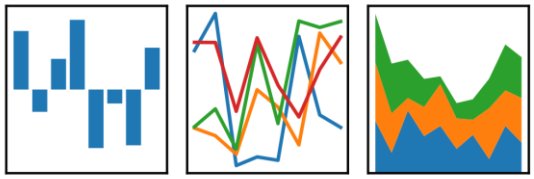
- Data structures for data analysis
  - DataFrame and Series
  - Semantic indexes
  - Groupby, resampling, joins
  - IO (csv, Excel, HDF5, parquet)

- NumPy N-dimensional array as backend

```
import pandas as pd
my_df = pd.DataFrame(my_array)
```

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



https://pandas.pydata.org/

Scikit-Learn

- Comprehensive library of machine learning methods
  - Models and algorithms
  - Preprocessing tools
  - Validation algorithms, model metrics, train-val-test splitting
  - Excellent documentation

- Also based on NumPy N-dimensional array

```
from sklearn.decomposition import PCA
my_pca_model = PCA(n_components=2).fit(my_array)
```

https://scikit-learn.org/

Some resources

- Scikit-Learn user guide
  - Comprehensive explanations of many algorithms
  - https://scikit-learn.org/stable/user_guide.html
- Machine Learning Mastery
  - Plenty of easy-to-follow tutorials
  - https://machinelearningmastery.com/
- Machine Learning on Coursera by Andrew Ng
  - Classic course, still very good
  - https://www.coursera.org/learn/machine-learning?

Today

https://github.com/Umetrics/machine-learning-workshop

Introduction to Pandas missing unfortunately

https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html