ML Fundamentals

Instituto Balseiro

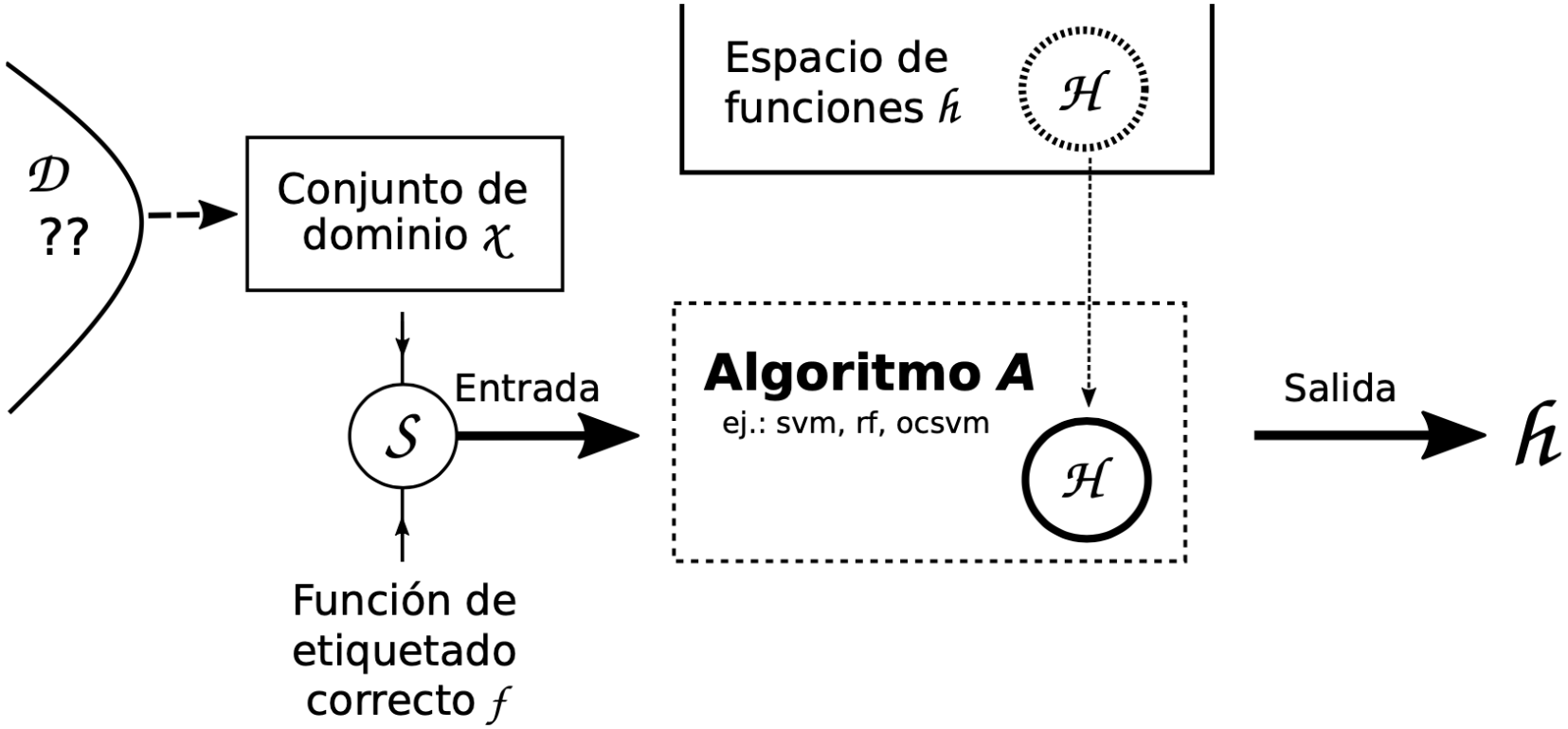
Instituto Balseiro
26/08/2022

Announcements

- Inscripciones: **hasta el 02/09** ⚠️
- Vocacionales: ya cerró. Acerbo y Garrido hablar después conmigo
- Oyentes: confirmar participación
- en Discussions: Cuento corto de Miguel y GPT-3 - notificaciones de github 👈🚦
- Práctica: Hoy entra la G2, y el lunes la P1
  - **P1 se entrega el 09/09.**

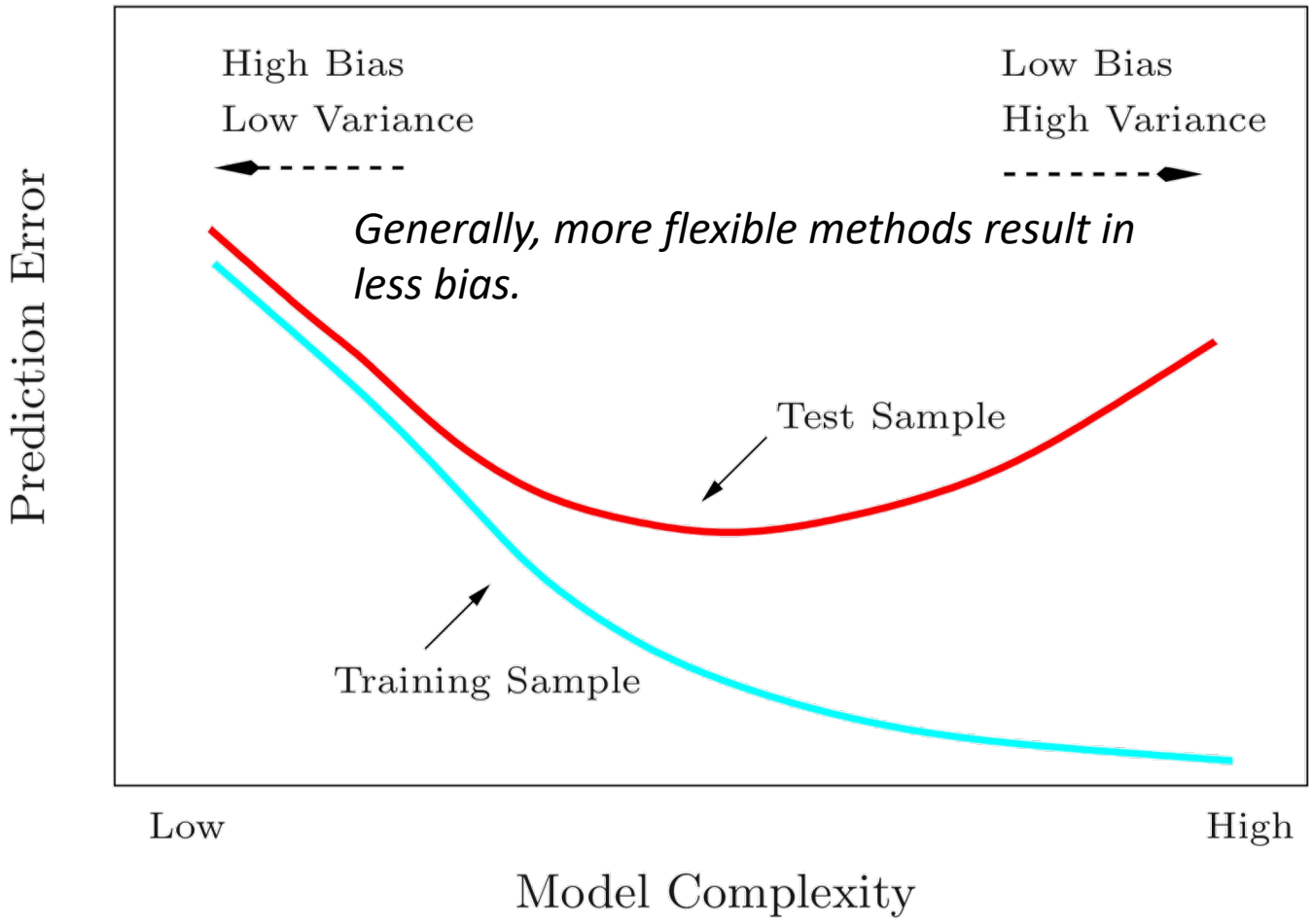
Last lecture review

Supervised Learning conceptual model



Técnicas de Aprendizaje Automático Aplicadas a Simulaciones Numéricas de Colisiones de Material Granular Poroso, Daniela N. Rim, Seminario de investigación y/o desarrollo, FCEN, UNCUyo, 2019. Basado en 2014 - Shalev-Shwartz - Understanding machine learning= From theory to algorithms.

The Bias–Variance Tradeoff



$$\mathrm{E}\left[\left(y-\hat{f}\left(x\right)\right)^{2}\right]=\left(\mathrm{Bias}\left[\hat{f}\left(x\right)\right]\right)^{2}+\mathrm{Var}\left[\hat{f}\left(x\right)\right]+\sigma^{2}$$

where

$$\mathrm{Bias}\left[\hat{f}\left(x\right)\right]=\mathrm{E}\left[\hat{f}\left(x\right)\right]-\mathrm{E}\left[f(x)\right]$$

and

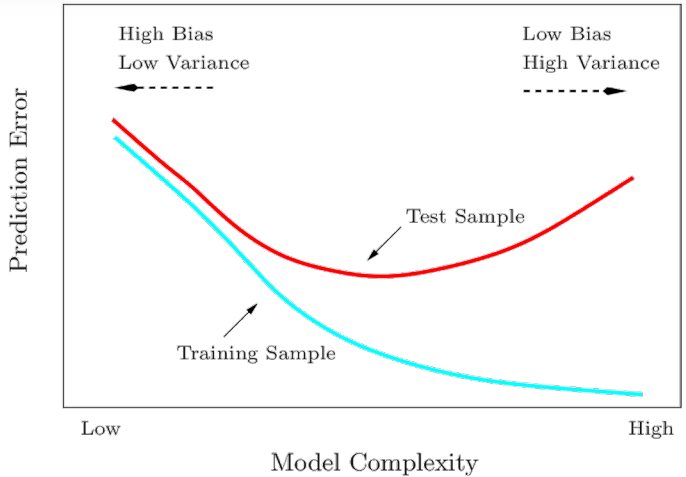
$$\mathrm{Var}\left[\hat{f}\left(x\right)\right]=\mathrm{E}[\hat{f}\left(x\right)^{2}]-\mathrm{E}[\hat{f}\left(x\right)]^{2}.$$

Towards a good model

- Model training: estimating the parameters of the model that lower a given loss function for that particular training data. *Optimization error/Loss/Cost.*

- Model selection: estimating the performance of different models in order to choose the best one: *validation*.

- Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data. *Test.*



Lecture 4
Data

ML Fundamentals – Lecture 4

- Data limitations
- Features
- Modeling process and Feature Engineering
- Data cleaning
- Exploratory Data Analysis (EDA)
- Data preprocessing
- Categorical data
- Feature selection

Data limitations

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features

Insufficient Quantity of Training Data

- Depending on the algorithm, one might need a huge amount of observations to train properly.
- In general, there is a tradeoff between the algorithm and the amount of data available.
- There are techniques to 'recycle' models (pre-trained models/transfer learning).

Nonrepresentative Training Data

- A model will be inaccurate if it was trained with data different from the data it's trying to predict.
- Irreducible noise (important with scarce data)
- Sampling bias
- Etc.

Poor-Quality Data

- Usually dealt with when data cleaning, but sometimes irremediable.
  - Errors
  - Outliers
  - Other noise (upper/lower, encoding, etc.)
  - Missing data or NA with no clue on its meaning

Irrelevant Features

- *Garbage in, garbage out*
- Sometimes not trivial to identify which

What's ML?

A little deeper...

- ML = ***representation*** + evaluation + optimization

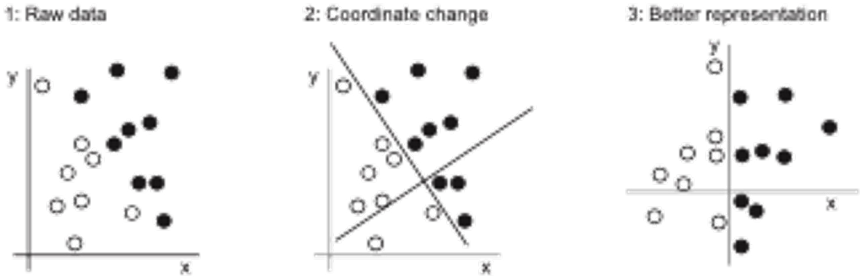
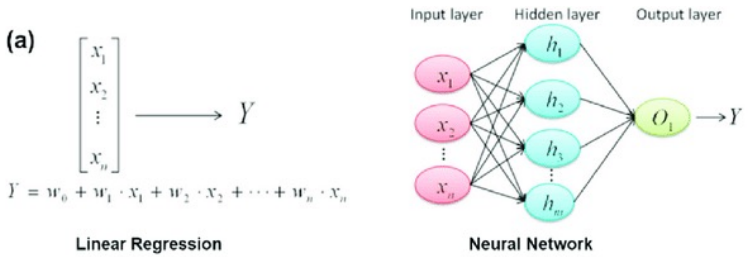


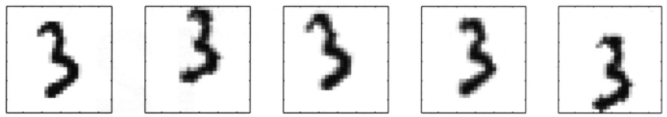
Figure 1.4 Coordinate change



2018 - Chollet - Deep Learning with Python

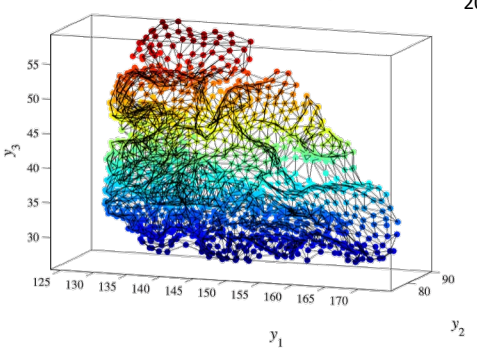
feature representation

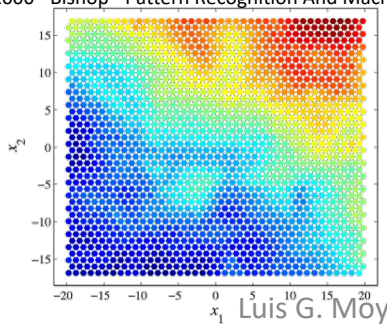
learner space



2006 - Bishop - Pattern Recognition And Machine Learning

manifold learning





2007 - Lee - Nonlinear Dimensionality Reduction

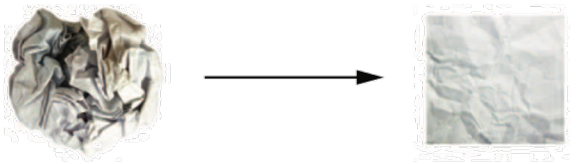


Figure 2.9 Uncrumpling a complicated manifold of data

Features

- Variables that are discriminant, available and (mostly) informative.

Features									Target
Cust. ID	State	Acct length	Area code	Int'l plan	Voicemail plan	Total messages	Total mins.	Total calls	Churned?
502	FL	124	561	No	Yes	28	251.4	104	False
1007	OR	48	503	No	No	0	190.4	92	False
1789	WI	63	608	No	Yes	34	152.2	119	False
2568	KY	58	606	No	No	0	247.2	116	True

[Feature engineering] is an art like engineering is an art, like programming is an art, like medicine is an art. There are well defined procedures that are methodical, provable and understood.

Feature Engineering

aka data wrangling or data munging

In the words of Pedro Domingos:[84]

. . . some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.

In the words of Jason Brownlee:[51]

[Feature engineering] is an art like engineering is an art, like programming is an art, like medicine is an art. There are well defined procedures that are methodical, provable and understood.

In the words of Andrew Ng:[337]

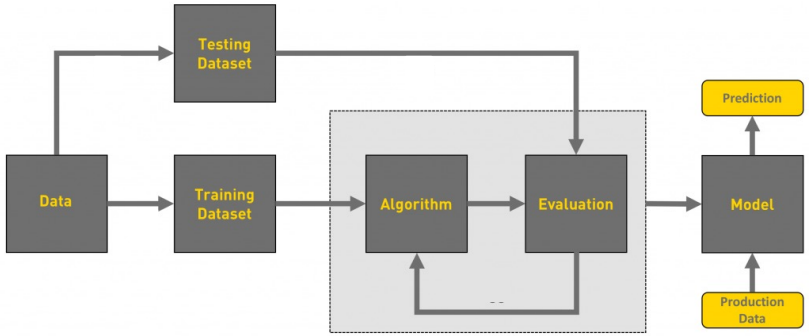
Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.

In the words of Yoshua Bengio:[337]

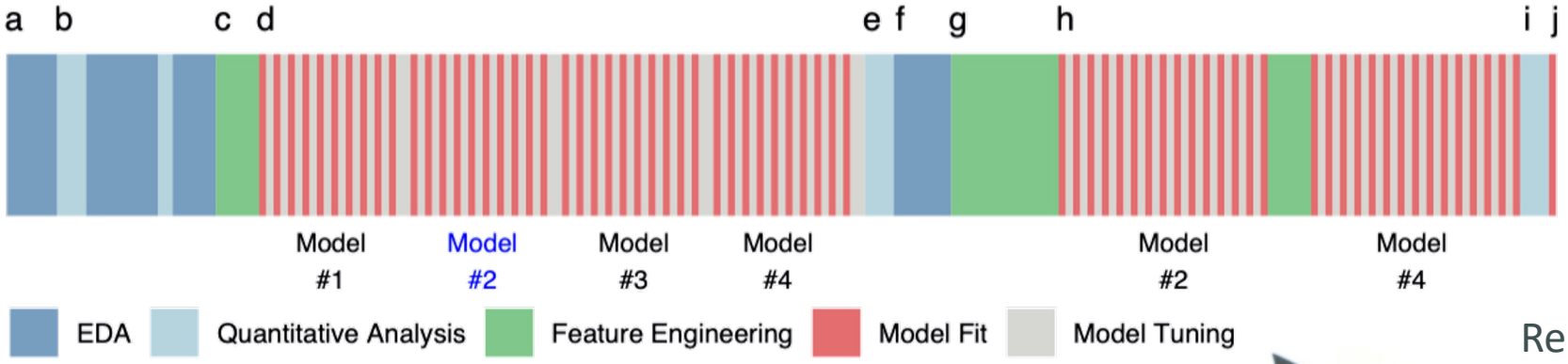
Good input features are essential for successful ML. Feature engineering is close to 90% of effort in industrial ML.

Model != Modeling Process

1. iterative
2. heuristic/creative
3. documented



A typical modeling process, comprising several ML cycles:



2019 - Book - Kuhn - Feature Engineering and Selection=A Practical Approach for Predictive Models, Fig 1.4, p9

Reproducible Research

Feature Engineering

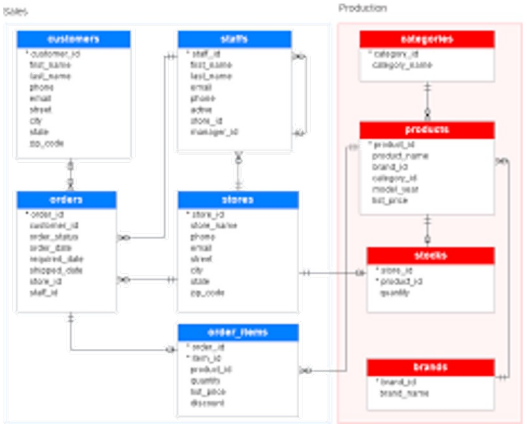
aka data wrangling or data munging

- **EDA**: exploratory analysis of the raw data.
  - identifying good features and expanding them
- **Error analysis**: which features are hurting performance more than helping?
  - identifying redundant/uninformative features and dropping them.

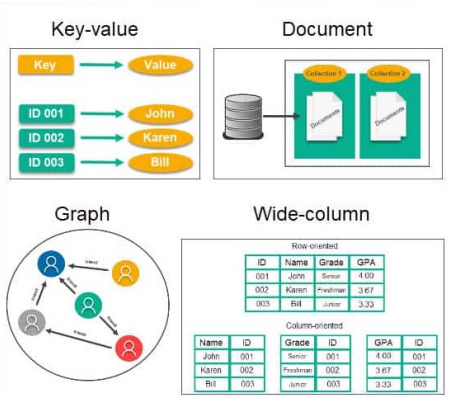
Domain Knowledge

Feature types

Type	Description / *Note*	Example
Binary	One of two values	Did the customer leave a tip? Did the turbine make noise? Did the student turn in the exam before the end of the allocated time? Does the person like cats?
	simplest	Did the licence plate have a match in the DMV database?
Categorical	One among many values	Colour of the car, Brand of the car, Credit card type, Number of bedrooms (*also a number*)
	values known beforehand	Type of mattress (*also mattress surface area*)
Discrete	Can be mapped to the integers	Number of pizzas eaten, Times visited the gas pump, Times the turbine got serviced last year, Steps walked last week (*compare: distance walked, which is continuous*), Number of people living in the house.
	order is important	
Continuous	Can be mapped to a real number	Engine temperature, Latitude, Longitude, Speech length, Colour intensity at centre of image, Distance walked, Temperature at centre of dish (*as measured by an IR camera*)
	representation issues	
Complex	Records, lists, sets	Date (year, month, day; note that it can be expressed as number of days from a fixed date), Product name (*it might include brand, product type, colour, size and variant*), Location (latitude and longitude), Complaint (a sequence of characters), Countries visited (it is a set of categories)
	challenging	



Structured: Spreadsheets, csv, arrays, dataframes, lists, SQL



Unstructured:
twitter, JSON, MongoDB, NoSQL

Dataset types

- Structured vs unstructured data
- File types (txt, CSV, SQLite, JSON, BigQuery)
- Sizes (Small, Medium, Large, Big data)
- Licenses (Creative Commons, GPL, Other Database, Other)

Missing values

blancks, NaNs, NAs, <special holder>, etc.

PassengerId	Survived	Pclass	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Female	26	0	0	STON/02. 3101282	7.925		S
4	1	1	Female	35	1	0	113803	53.1	C123	S
5	0	3	Male	35	0	0	373450	8.05		S
6	0	3	Male		0	0	330877	8.4583		Q

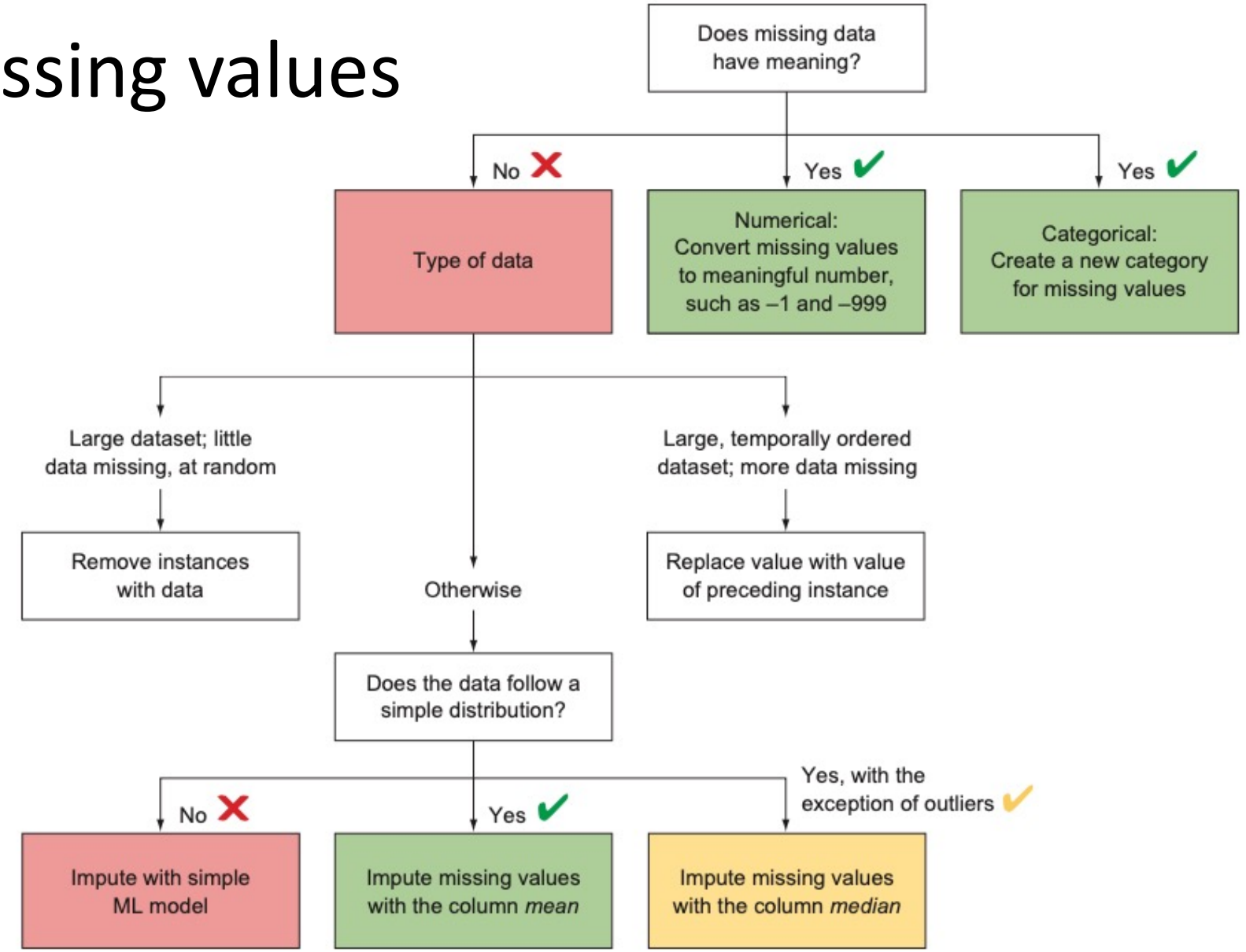
Missing values

- Causes: structural, random (MCAR, MAR), specific
- Is the NA informative?
- Should NAs stay? Or should they go?
- Imputation

https://scikit-learn.org/stable/modules/impute.html#impute

2019 - Kuhn Sec. 8.2

Missing values



Henrik Brink, Joseph Richards, Mark Fetherolf - Real-World Machine Learning-Manning Publications (2016)
Josse et al. *On the consistency of supervised learning with missing values*. 2019. ⟨hal-02024202v2⟩

EDA: Exploratory data analysis

- Load and inspect data
- Serialization formats
  - JSON
  - HDF5
  - Pickle
    - most objects!
    - easy to use!
    - … but be careful

```
import numpy as np
import pandas as pd
import pickle
```

```
>>> pdDf = pd.read_csv('100 Sales Record.csv')
>>> pdDf.head()
```

```
with open('test.pkl','wb') as f:
    pickle.dump(pdDf, f)
```

```
with open("test.pkl", "rb") as f:
    d4 = pickle.load(f)

>>> d4.head()
```

Warning: The `pickle` module **is not secure**. Only unpickle data you trust.

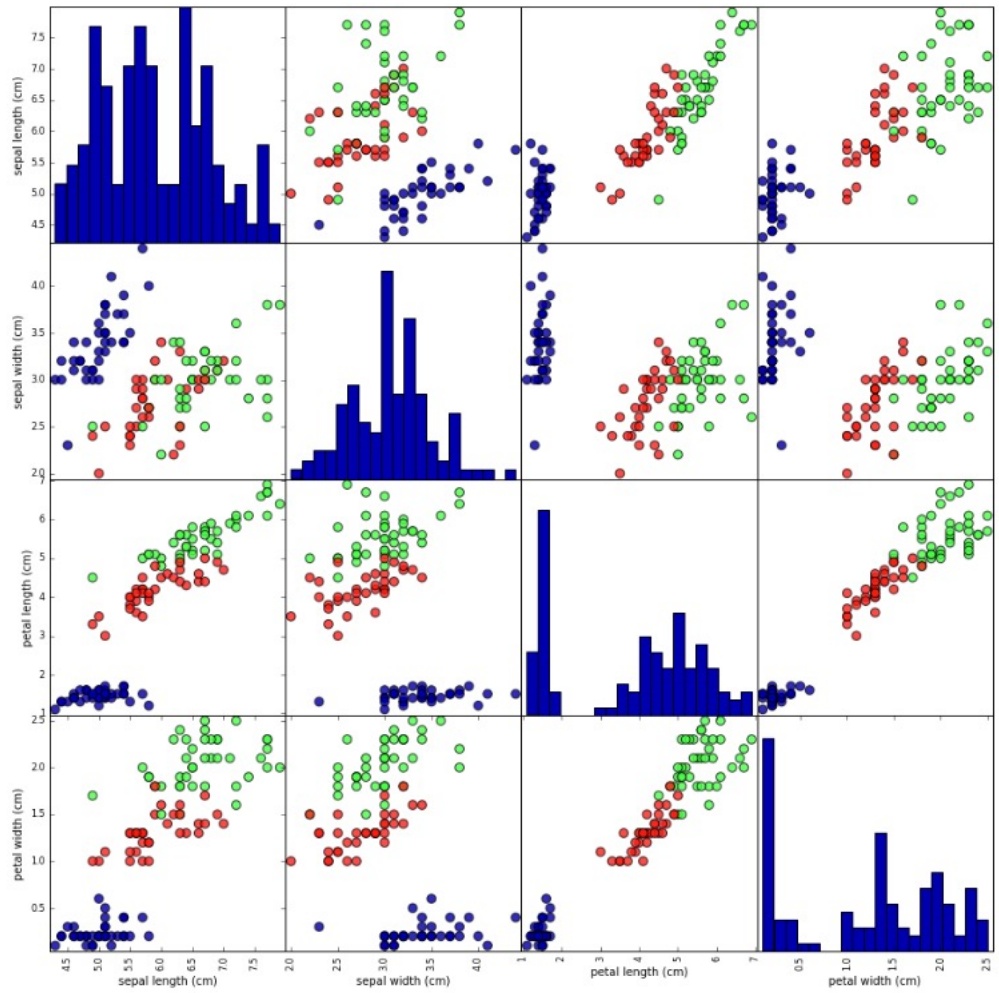
It is possible to construct malicious pickle data which will **execute arbitrary code during unpickling**. Never unpickle data that could have come from an untrusted source, or that could have been tampered with.

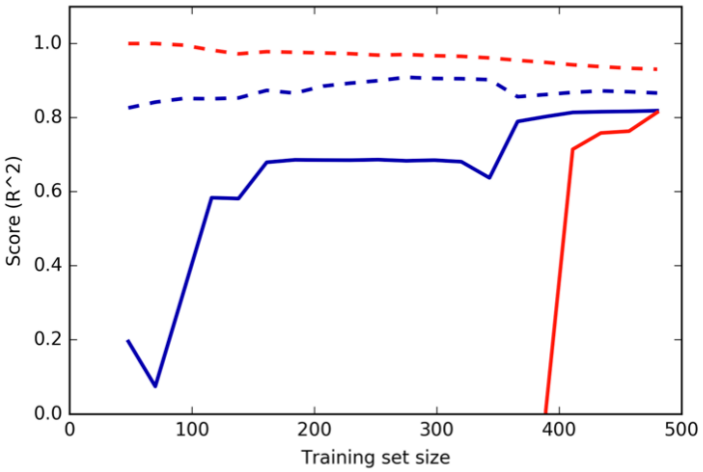
Consider signing data with `hmac` if you need to ensure that it has not been tampered with.

Safer serialization formats such as `json` may be more appropriate if you are processing untrusted data. See Comparison with json.

https://docs.python.org/3/library/pickle.html

EDA: Exploratory data analysis





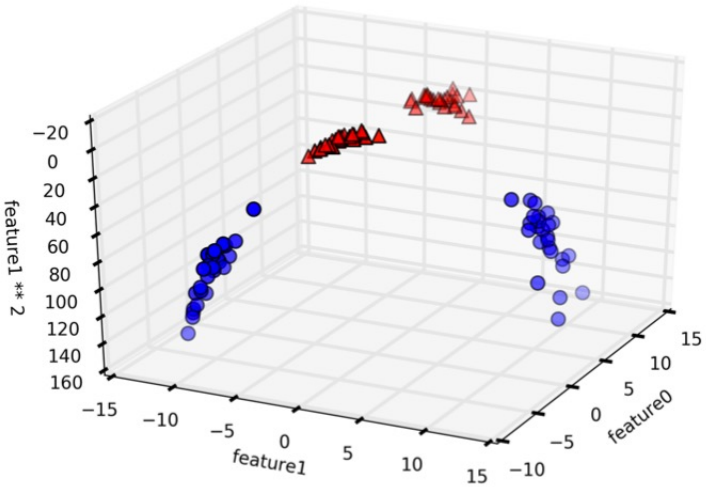
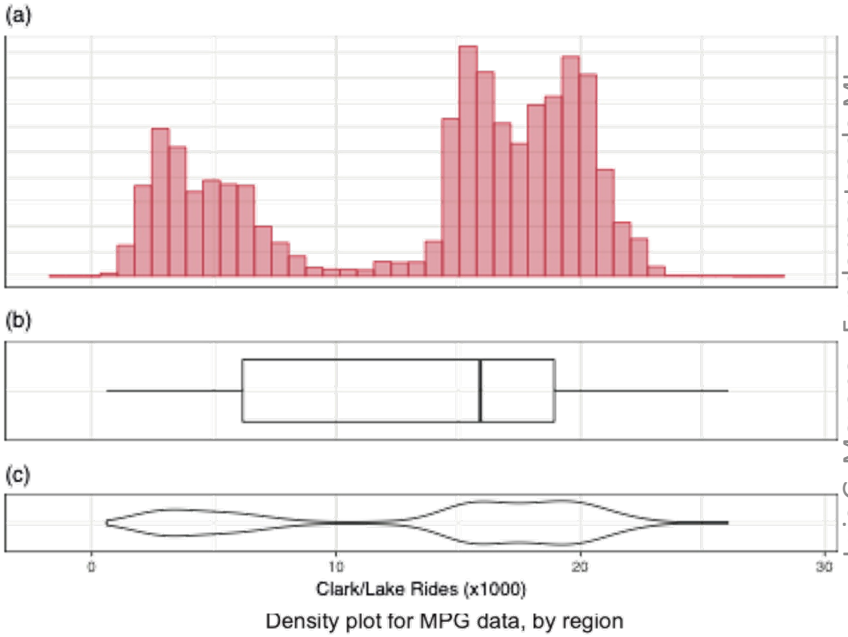
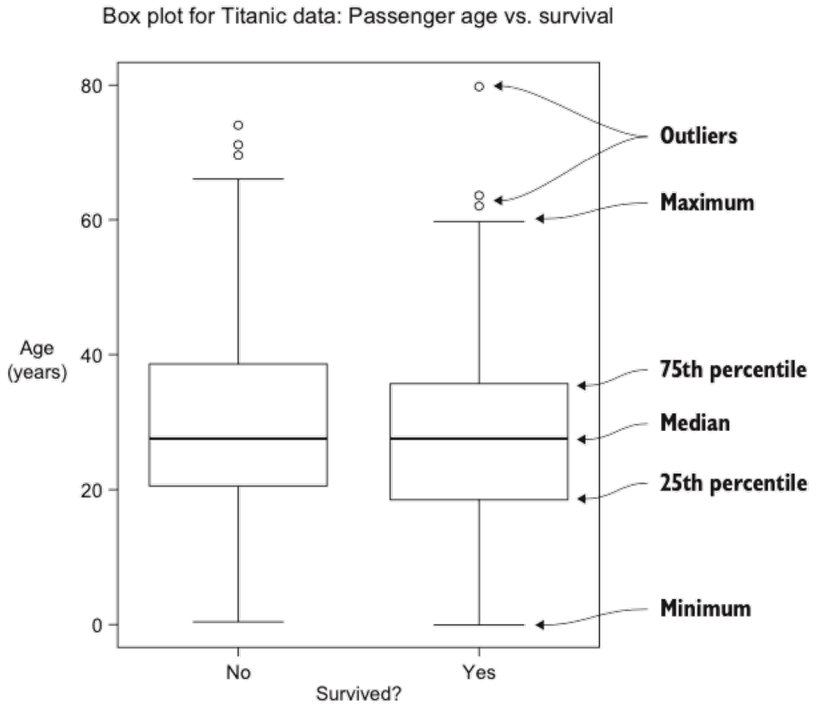


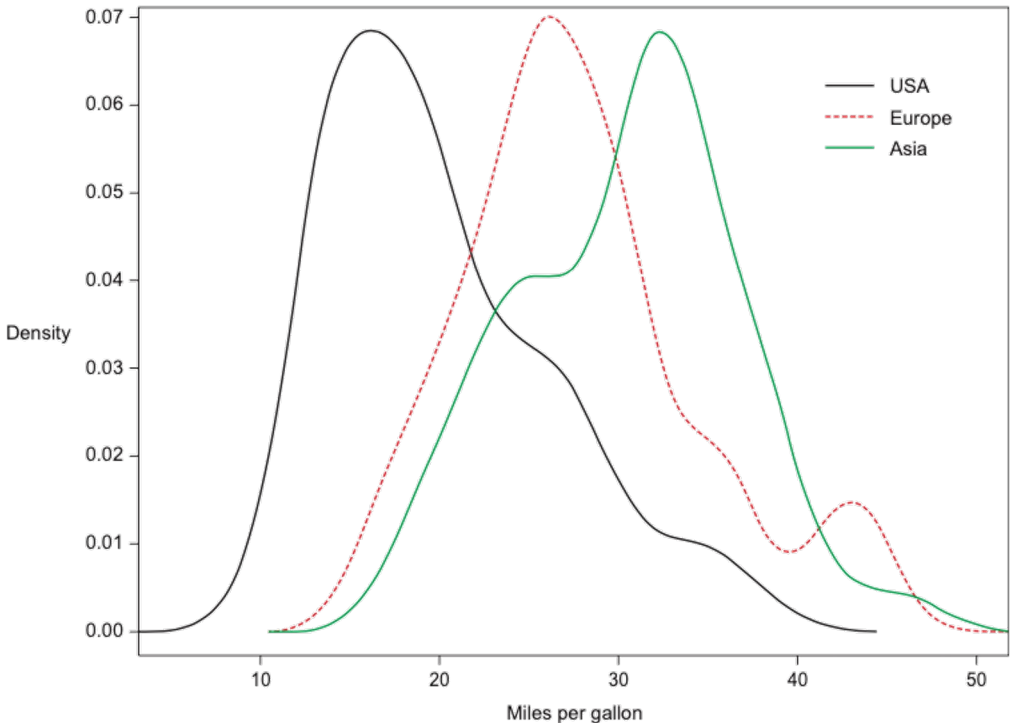
Figure 1-3. Pair plot of the Iris dataset, colored by class label

EDA: Comparing data

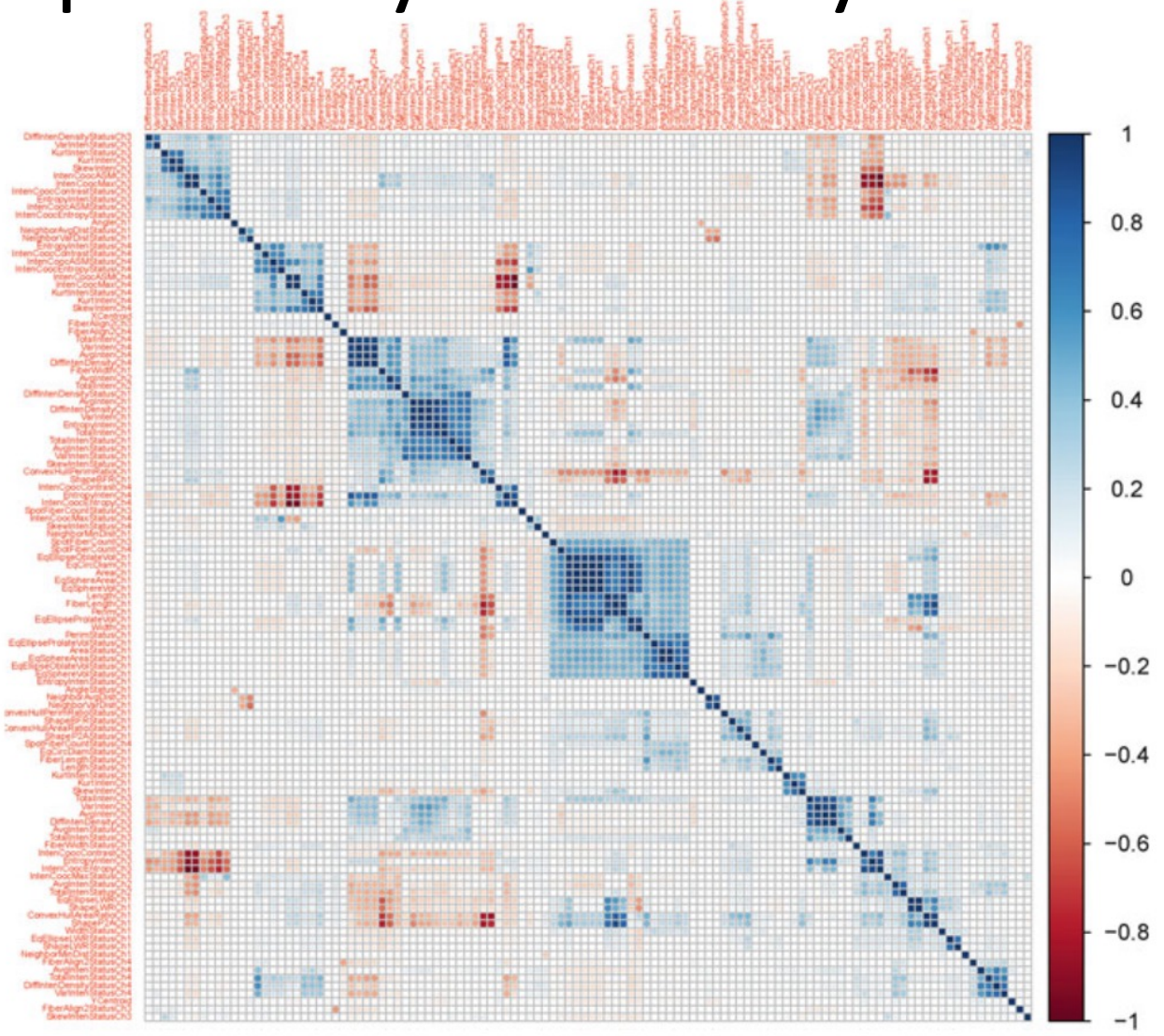


Density plot for MPG data, by region





EDA: Exploratory data analysis

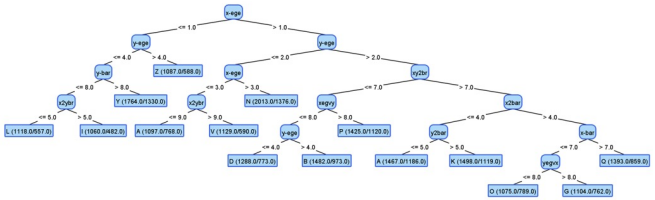


(Multi)collinearity

- Redundant predictors frequently add more complexity to the model than information they provide to the model.
- In some models, may lead to unstable behavior, numerical errors or poor interpretability.
- Other models (e.g. trees) deal well with this.
- For prediction, it's generally **not** an issue.

No-free-lunch theorem

- Always check your assumptions before relying on a model or search algorithm.
- There is no “super algorithm” that will work perfectly for all datasets.

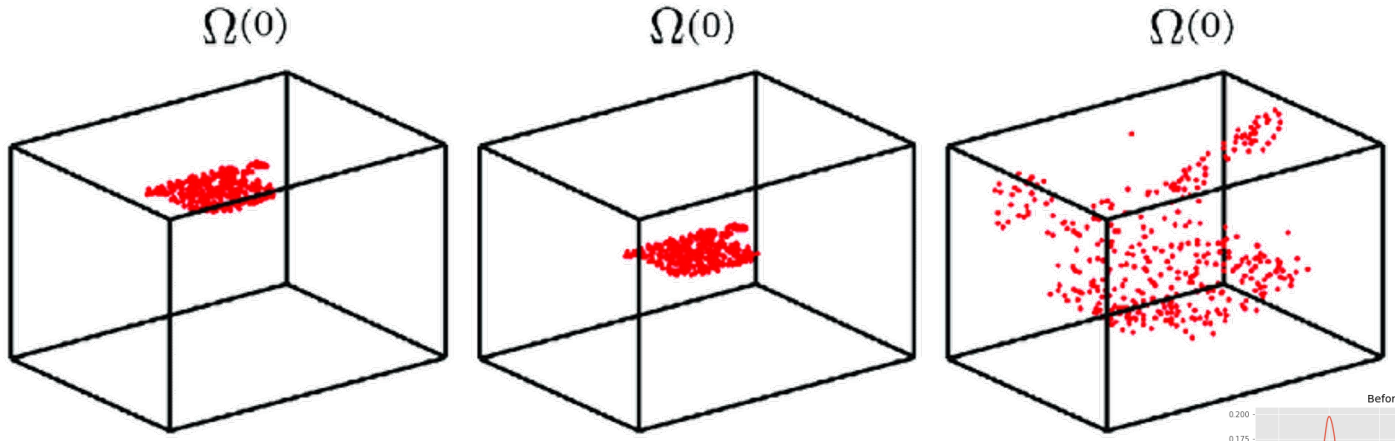


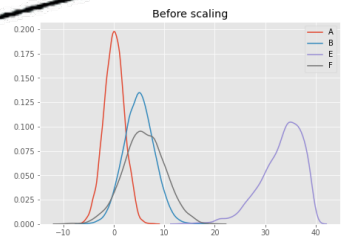
Wolpert, D.H., Macready, W.G. (1995-1997)

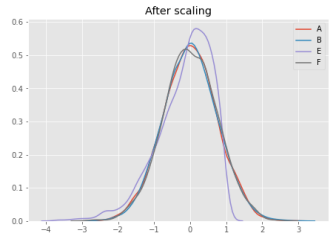
Data pre-processing

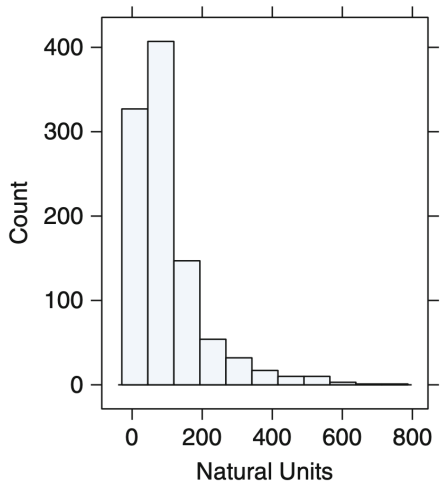
- Normalization: centering and scaling
  - [0, 1]?
  - http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html
  - Whitening (decorrelation, identity covariance)
- dimensionality reduction (PCA, t-SNE, word2vec)
- outlier removal
- near-zero variance (bad, but check target correlation)
- collinear features

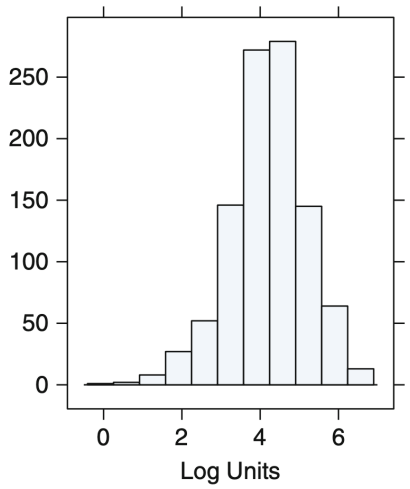
EDA: Exploratory data analysis

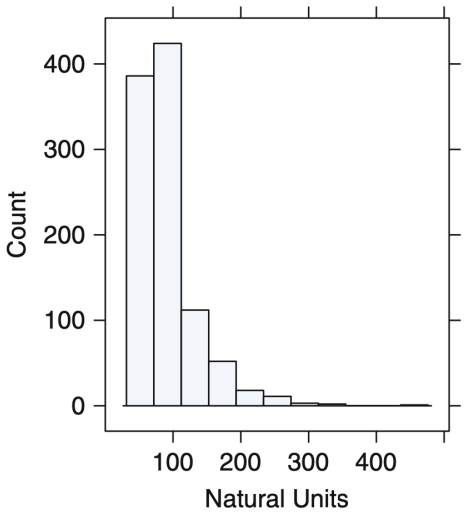


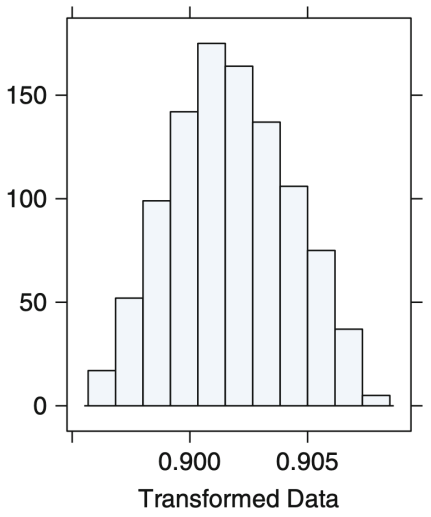




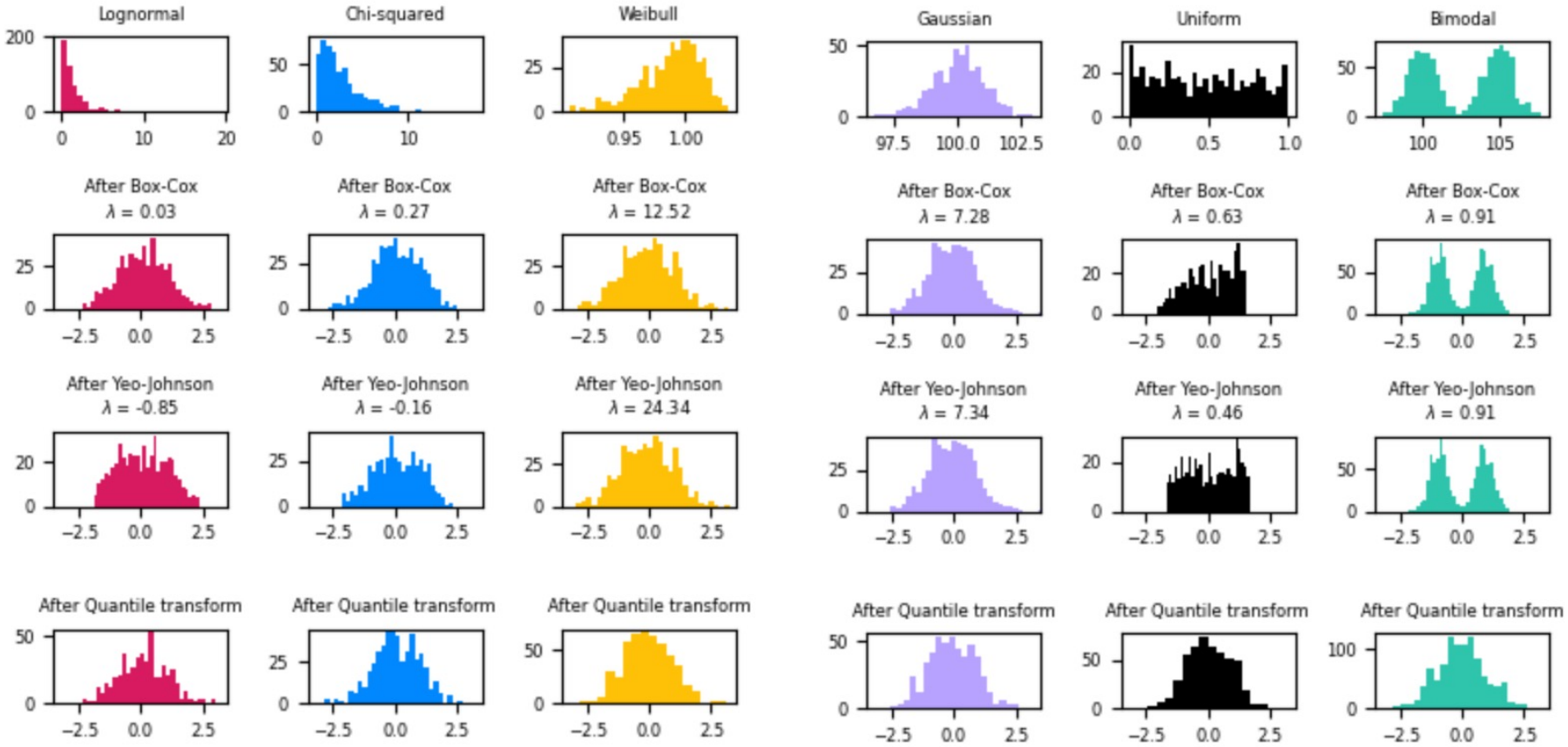




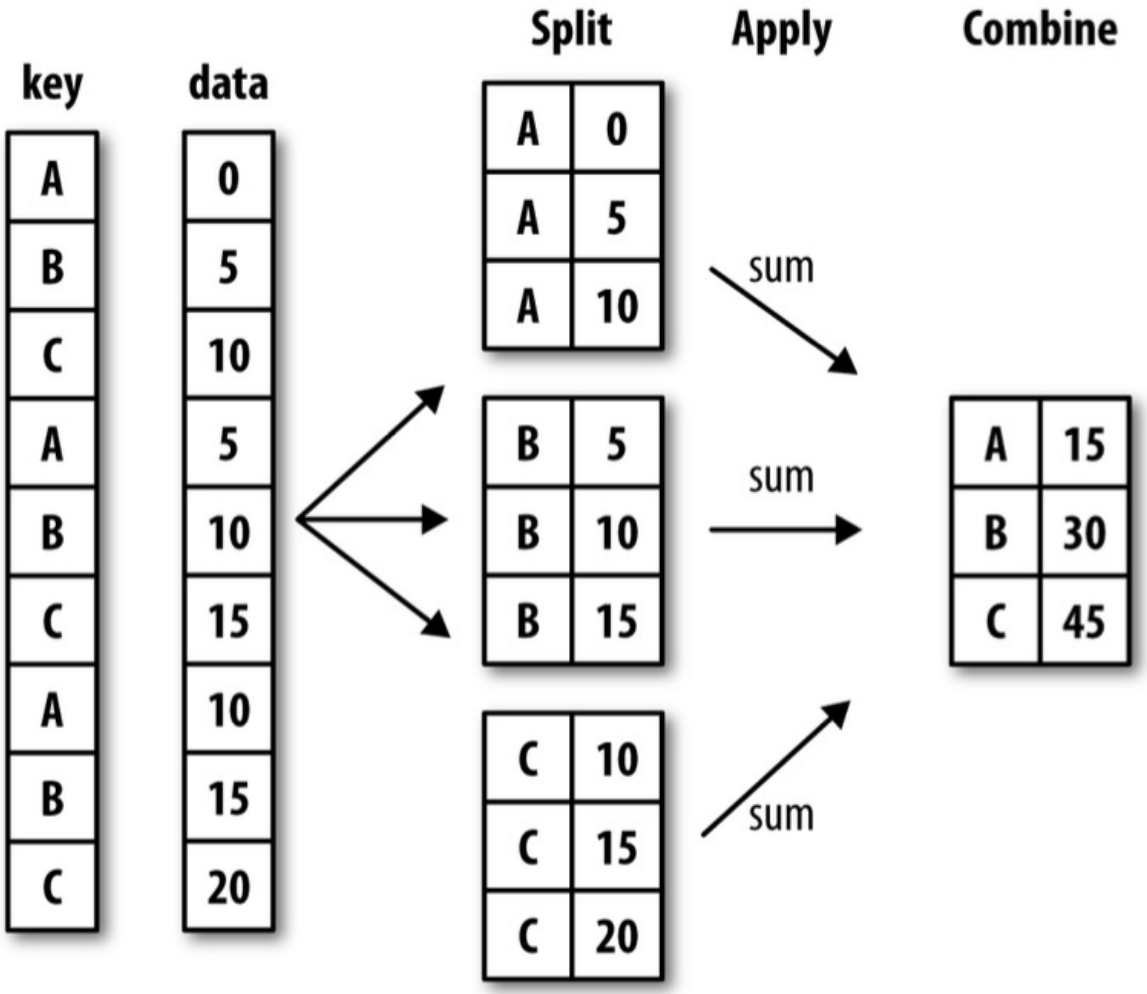




Box-Cox, Yeo-Johnson, Quantile



Group data (Split + Apply + Combine)



Other transformations

- pivoting
- decomposing
- discretizing
- kernel transformations

Original Data

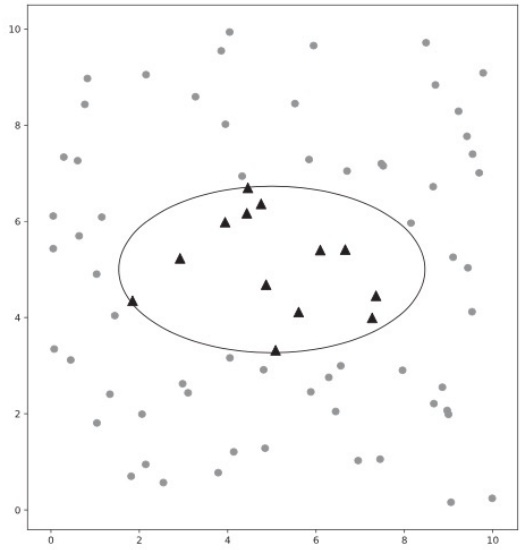
ID	Activity	Target
1	read	n
1	save	n
1	search	y
2	save	n
2	read	n
3	save	n
4	search	n
5	read	y
5	search	y

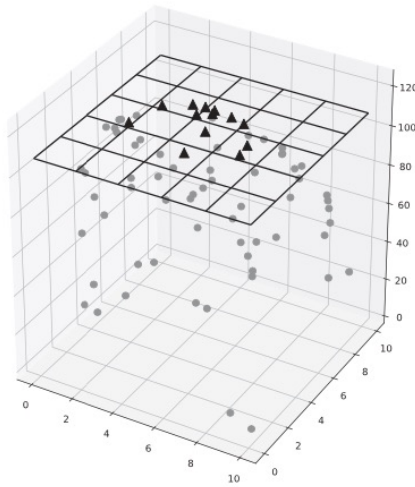
→

Data after Pivot

ID	Read?	Save?	Search?	Target
1	y	y	y	y
2	y	y	n	n
3	n	y	n	n
4	n	n	y	n
5	y	n	y	y

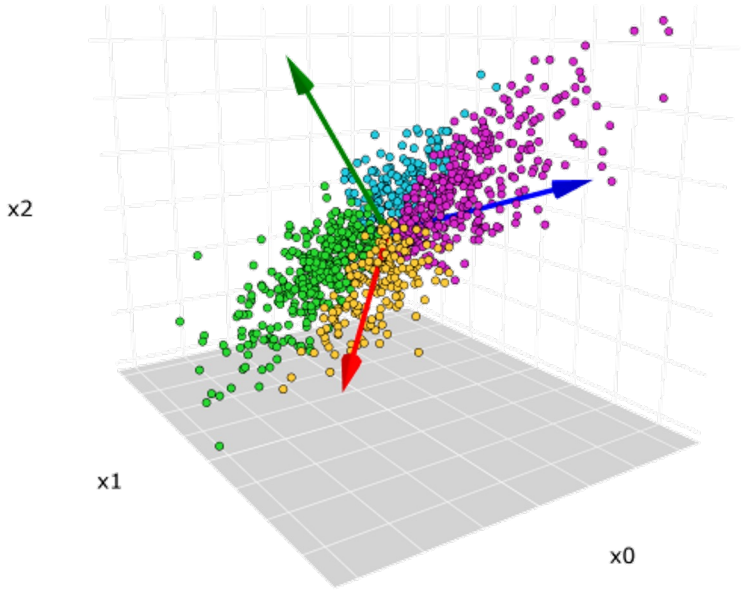
$$K_{\text{rbf}}(\vec{x}, \vec{y}) = \exp\left(-\gamma \left\| \vec{x} - \vec{y} \right\|^2\right)$$



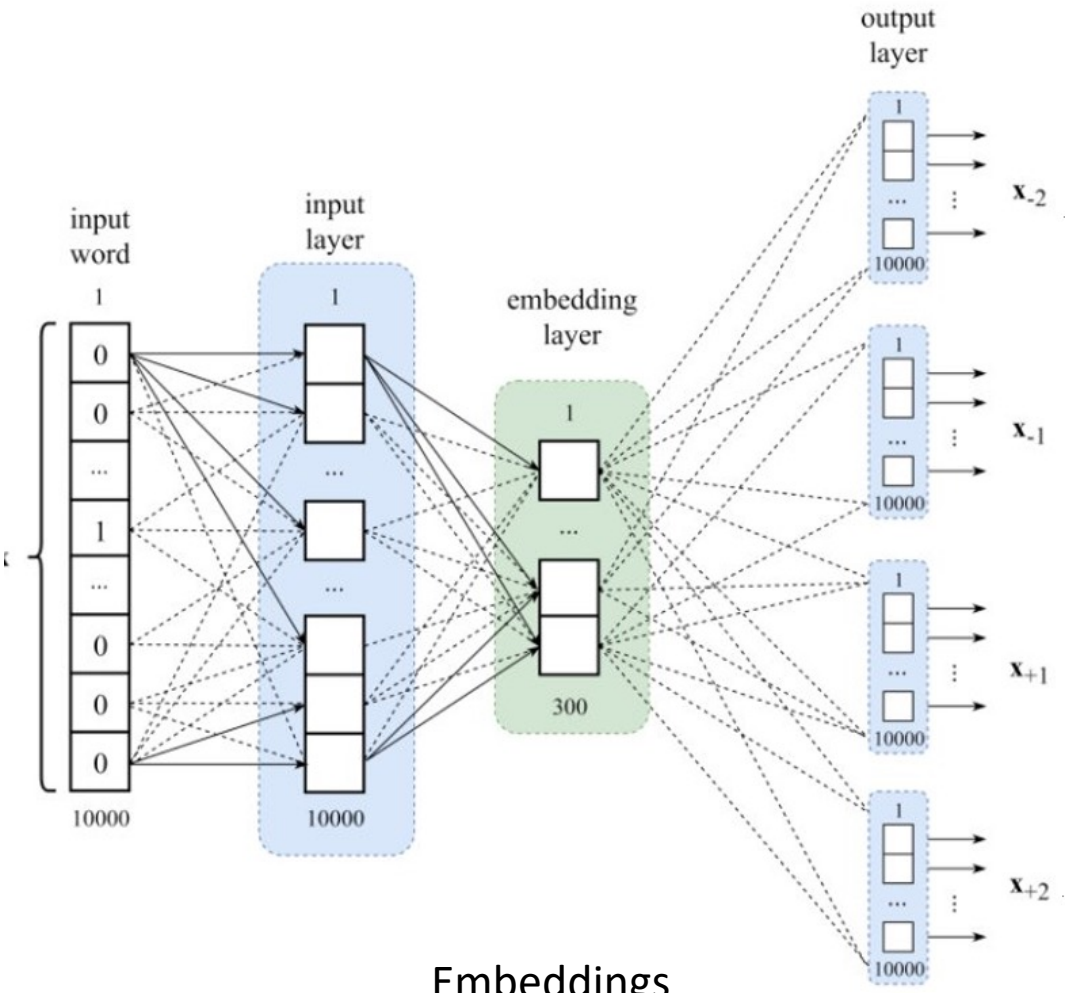


Learning representations

Dimensionality reduction and embeddings



PCA



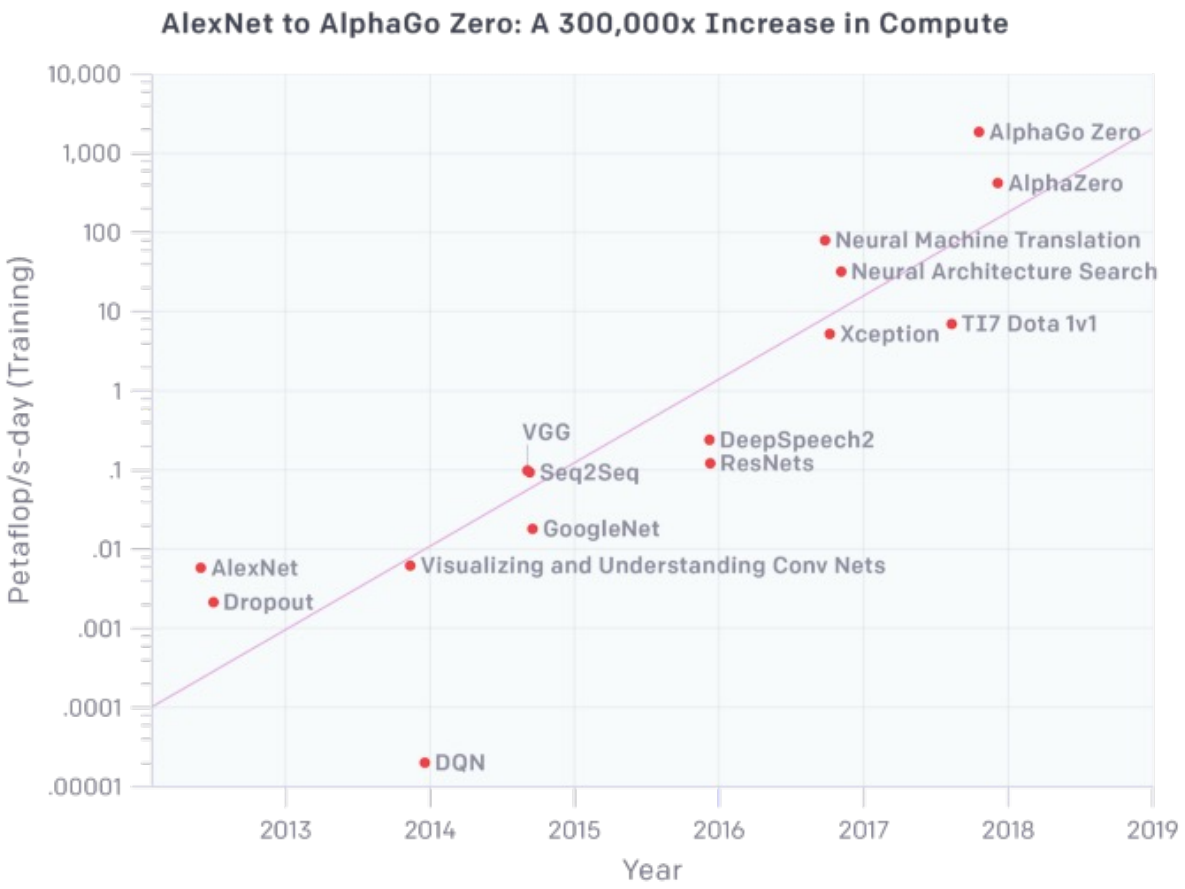
Embeddings

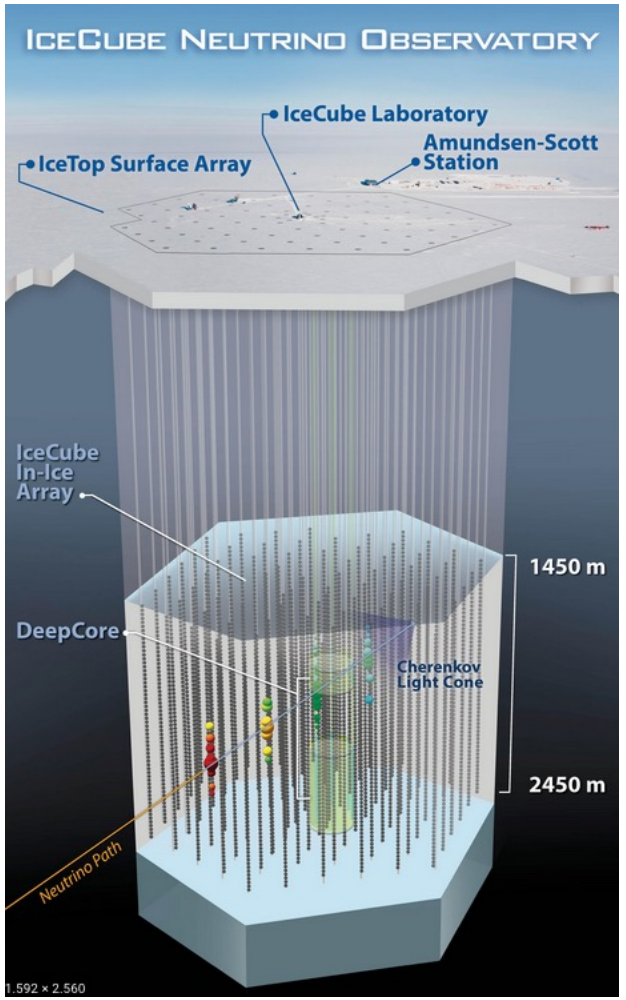
Deep Learning doesn't need FE

- DL offers the possibility to eliminate FE by trading large amounts of training data for the painstaking process of feature improvement.
  2020 Duboue p127

- Pero, pero, pero... Good features solve problems using fewer resources.

Big Data


AlexNet to AlphaGo Zero: A 300,000x Increase in Compute
Petaflop/s-day (Training)
10,000
1,000
100
10
1
.1
.01
.001
.0001
.00001
2013
2014
2015
2016
2017
2018
2019
Year
AlphaGo Zero
AlphaZero
Neural Machine Translation
Neural Architecture Search
Xception
TI7 Dota 1v1
VGG
Seq2Seq
DeepSpeech2
ResNets
GoogleNet
AlexNet
Dropout
Visualizing and Understanding Conv Nets
DQN


IceCube Neutrino Observatory
IceCube Laboratory
Amundsen-Scott Station
IceTop Surface Array
IceCube In-Ice Array
1450 m
DeepCore
Cherenkov Light Cone
2450 m
Neutrino Path
1.592 × 2.560

Categorical features

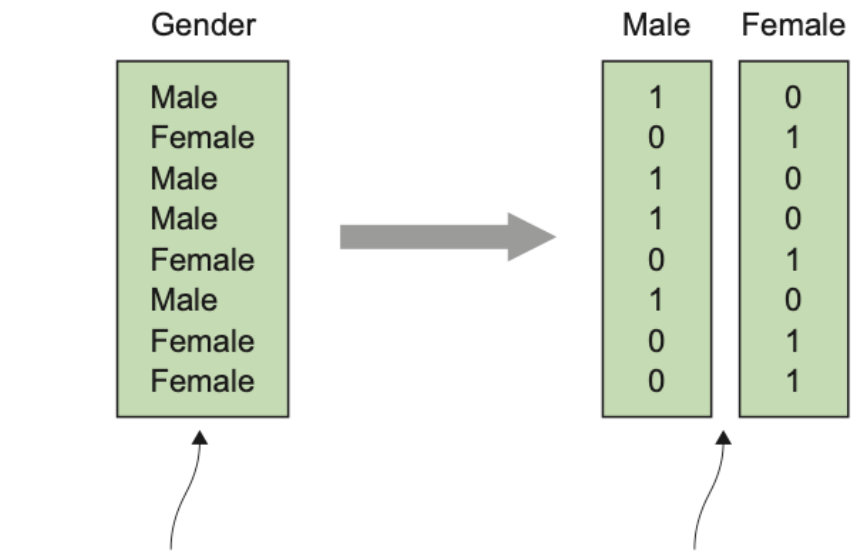
Person	Name	Age	Income	Marital status
1	Jane Doe	24	81,200	Single
2	John Smith	41	121,000	Married

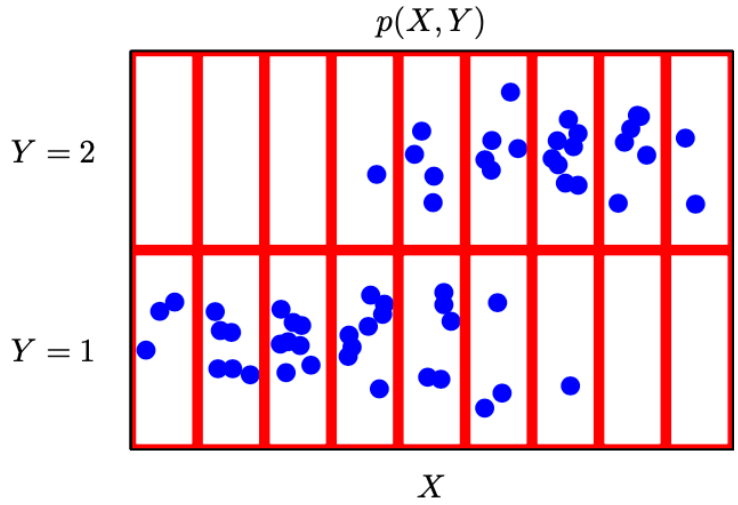
Categorical features

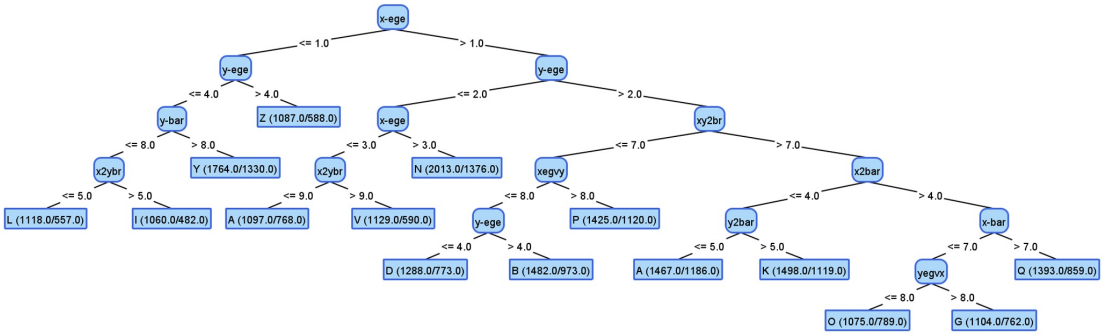
PassengerId	Survived	Pclass	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Female	26	0	0	STON/02. 3101282	7.925		S
4	1	1	Female	35	1	0	113803	53.1	C123	S
5	0	3	Male	35	0	0	373450	8.05		S
6	0	3	Male		0	0	330877	8.4583		Q

https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features

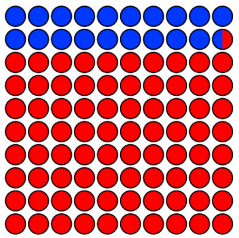
One-hot enconding, aka dummy variables

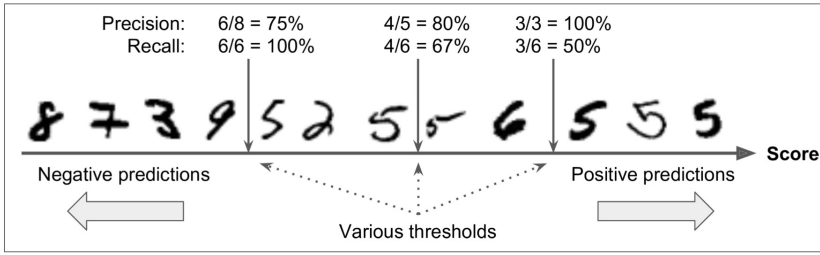






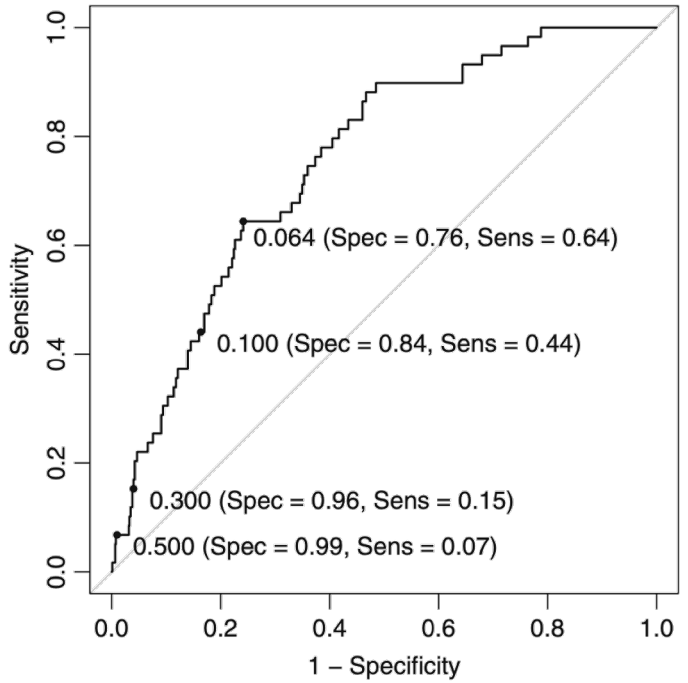
Severe Class Imbalance

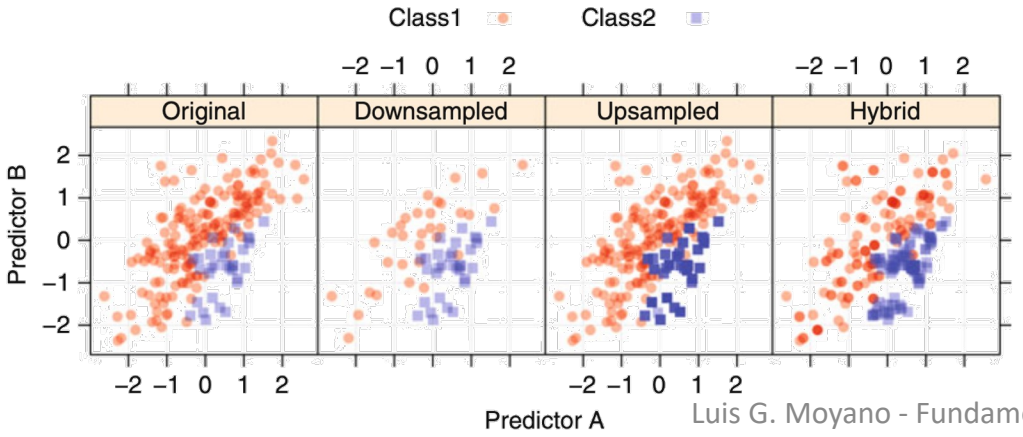




1. Alternate Cutoffs
2. Sampling methods
   - SMOTE
   - ROSE
3. Cost-sensitive training
   - error/class penalty

	0.50 Cutoff		0.064 Cutoff	
	Insurance	no insurance	Insurance	no insurance
Insurance	11	19	71	441
Noinsurance	105	1827	45	1,405



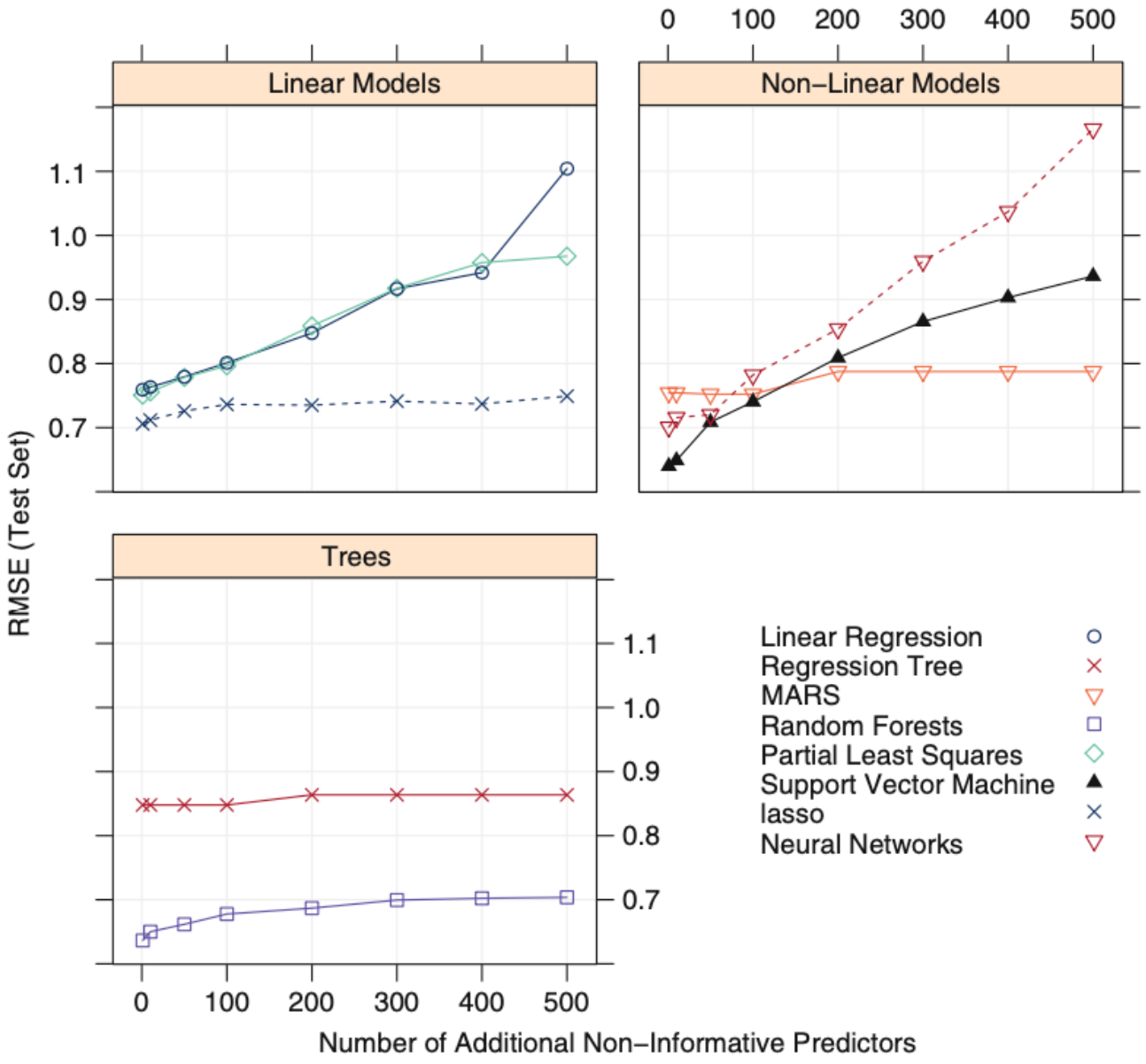


2018 - Book - Kuhn - Applied Predictive Modeling - Chapter 16

Feature engineering workflow (#ponele)

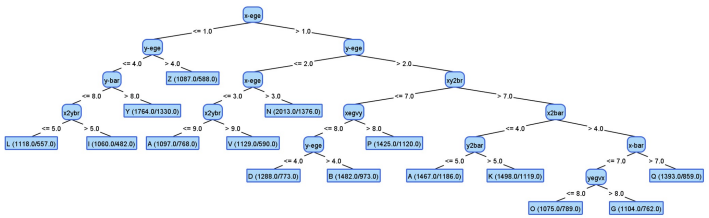
- Include all the features that you suspect to be predictive of the target variable.
  Fit an ML model. If the accuracy of the model is sufficient, stop.
- Otherwise, expand the feature set by including other features that are less obviously related to the target.
  Fit an ML model. If the accuracy of the model is sufficient, stop.
- Otherwise, starting from the expanded feature set, run an ML *feature selection algorithm* to choose the best, most predictive subset of your expanded feature set.

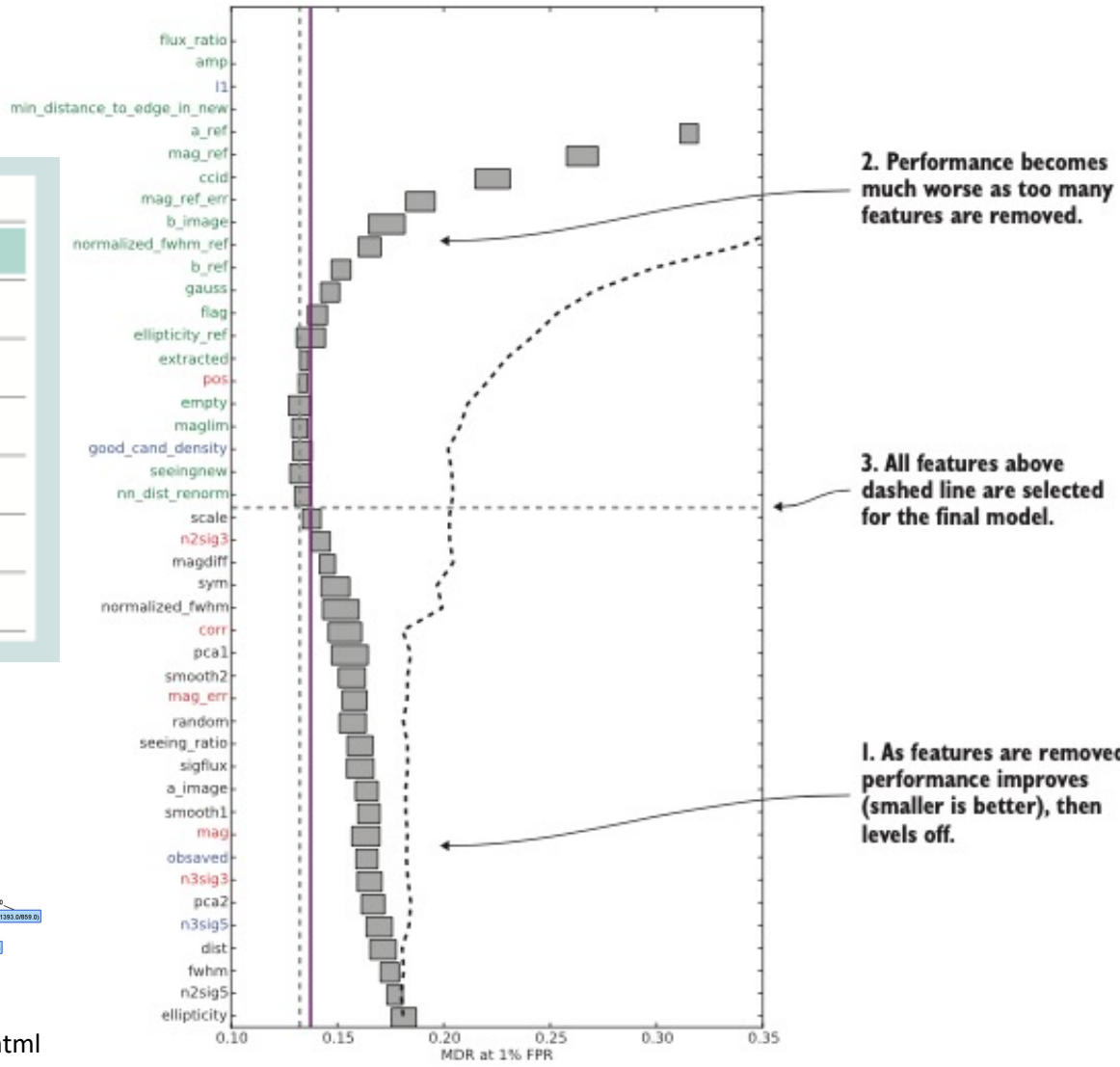
Non-informative Predictors



Feature selection

Feature	Importance
birthyear	
timezone	
datetime_week_of_year	
datetime_day_of_year	
lat	
datetime_hour_of_day	
lng	





https://scikit-learn.org/stable/modules/feature_selection.html

Lecture 4 – Data – Summary

- Data limitations
- Features
- Modeling process and Feature Engineering
- Data cleaning
- Exploratory Data Analysis (EDA)
- Data preprocessing
- Categorical data
- Feature selection

Next: **Linear algorithms**