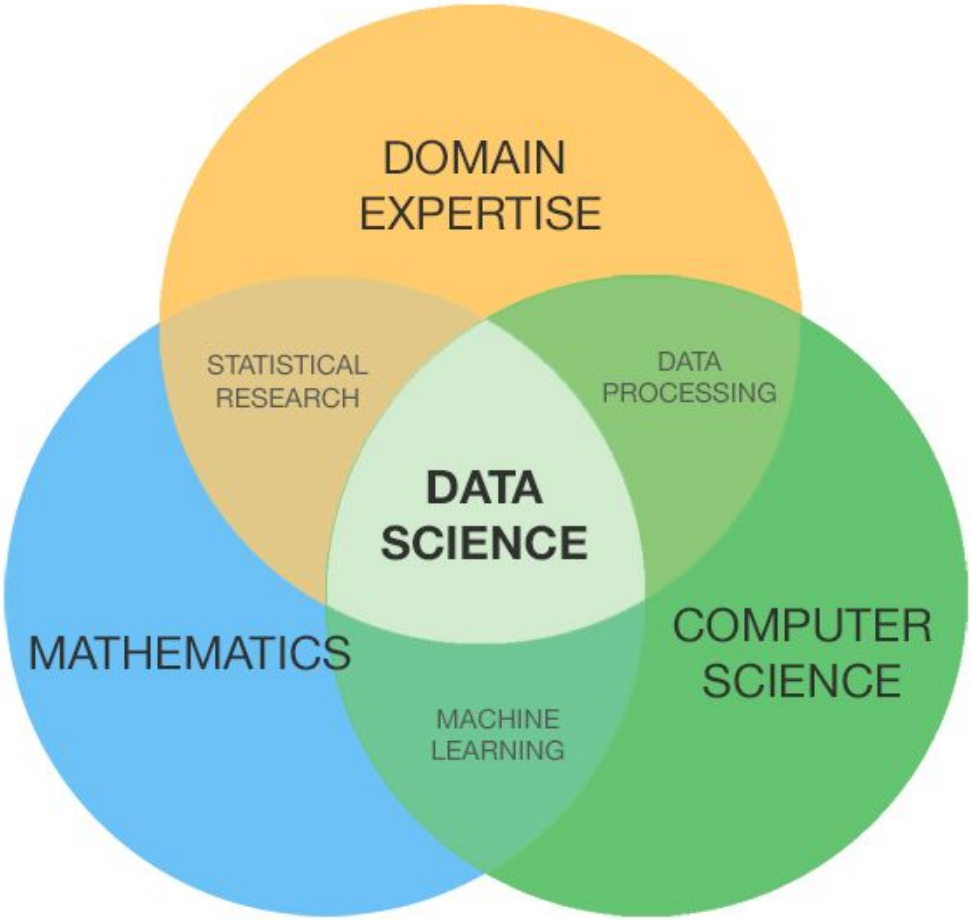
Data Science

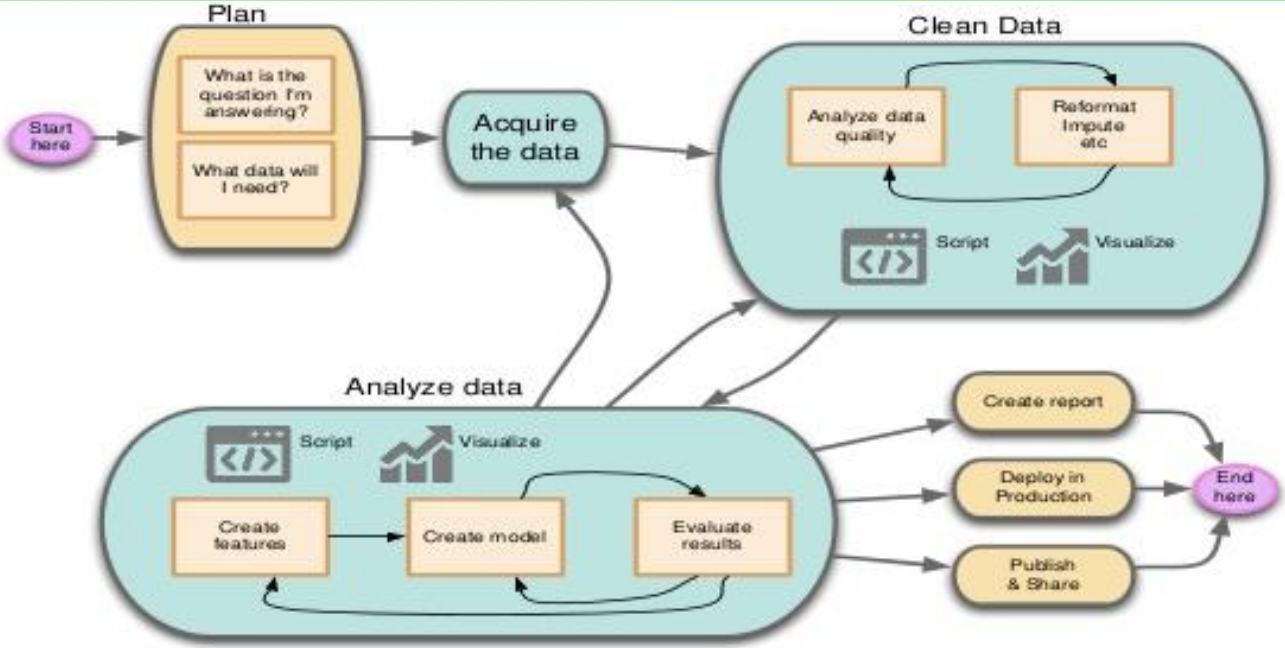
Overview, tools, applications

M. Glowacki - Axiomato



Source: Palmer, Shelly. Data Science for the C-Suite.
New York: Digital Living Press, 2015. Print.

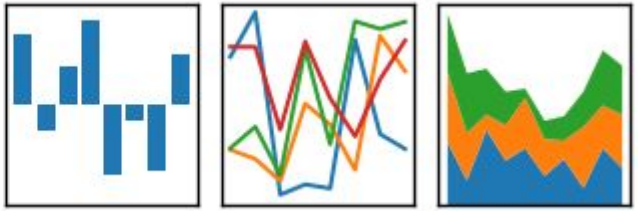
The Data Science Workflow…



EDA

pandas

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



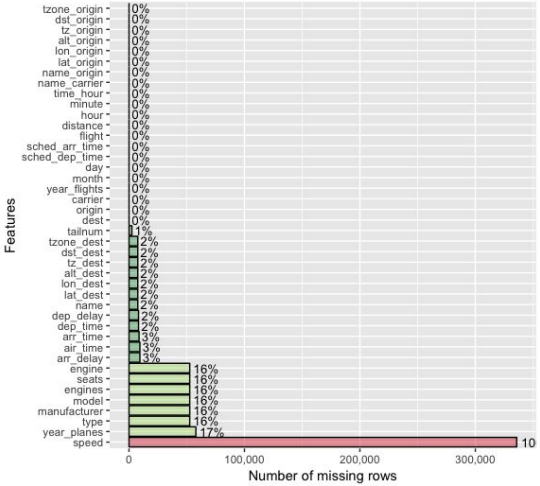
- Python library
- Fast in-memory data wrangling and cleaning
- DataFrame concept (different types, NaN)
- Basic statistics
- Charting (wrapper around matplotlib.pyplot) - scatterplot, histogram, boxplot, barplot, piechart

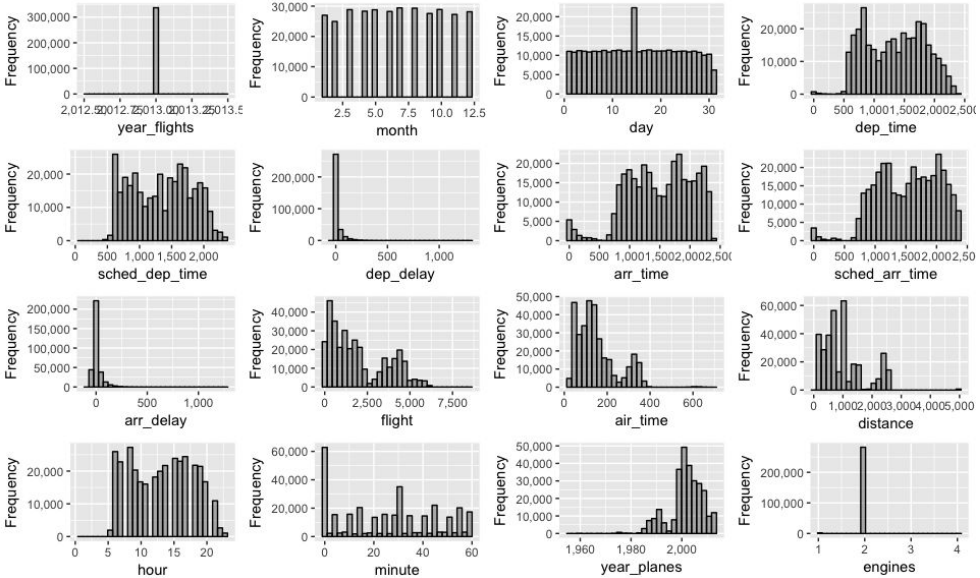
* For “real arrays” it is better to use Numpy.

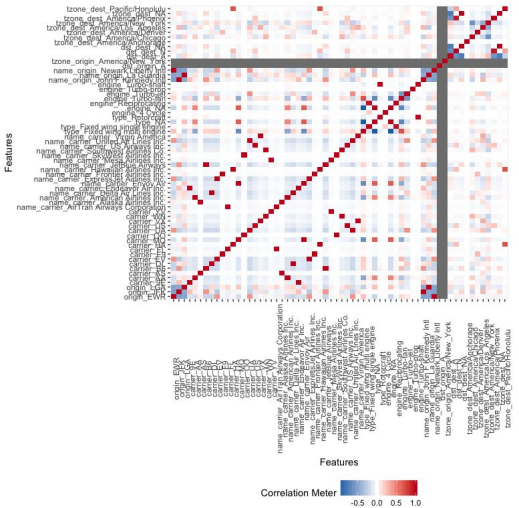
** For internal analysis I also use Trifacta Wrangler and various R packages: DataExplorer, Caret, LargeVis

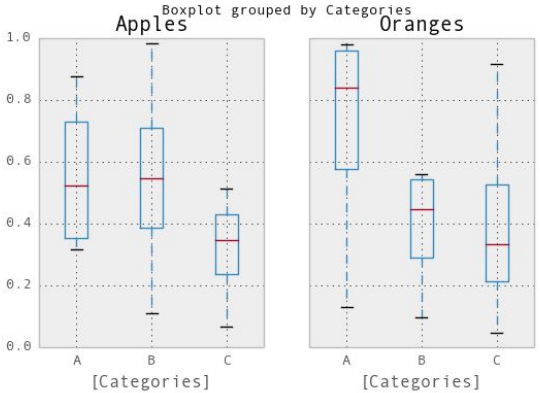
Data preparation (cleaning and feature engineering)

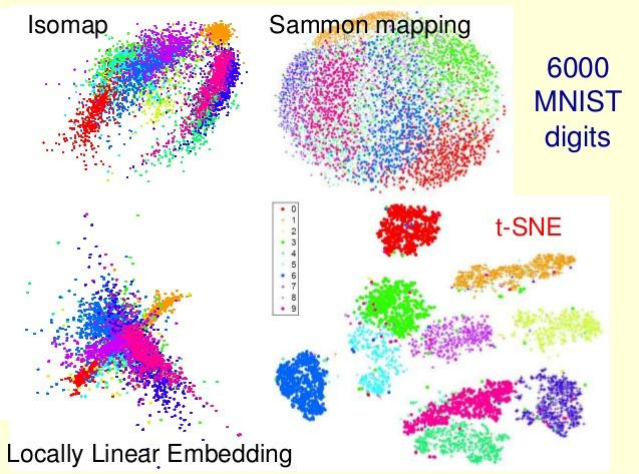
- Near zero variance
- Missing values - imputation strategy
- Outliers removal, winsorization, ...
- Categorical data inspection and one-hot encoding
- Unbalanced datasets (imbalanced-learn, metric selection)
- Centering and scaling, skewness; MinMax, Root, Log scaling
- Correlation analysis
- Dimensionality reduction (e.g. PCA)
- Feature generation (binning, interactions: ratios, multipliers, sums etc.)
- Feature selection
- NLP for text


Features
Number of missing rows
Group
Good
OK
Remove


Frequency
year_flights
month
day
dep_time
sched_dep_time
dep_delay
arr_time
sched_arr_time
arr_delay
flight
air_time
distance
hour
minute
year_planes
engines


Features
Correlation Meter


Boxplot grouped by Categories
Apples
Oranges
[Categories]


Isomap
Sammon mapping
6000
MNIST
digits
t-SNE
Locally Linear Embedding

ML tasks

1. **Supervised**
   - **classification**
   - **regression**
2. Unsupervised (clustering, dimensionality reduction,...)
3. Other: semi-supervised learning, online learning, reinforcement learning, ranking learning, one-shot learning, structured prediction, transfer learning...

Comparing Classification & Regression

property	supervised classification	regression
output type	discrete (class labels)	continuous (number)
what are you trying to find?	decision boundary	"best fit line"
evaluation	accuracy	"sum of squared error" — or — r^2 ("r squared")

ML metrics

Classification:

- Confusion matrix and derivatives: ACC, F1, precision, recall
- AUC, log-loss
- ...

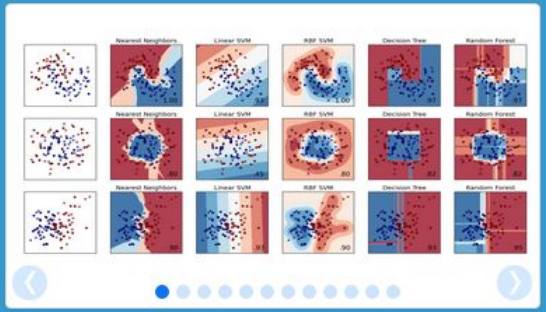
Regression:

- Mean squared error
- Median absolute error
- ...

White-box vs black-box - Model interpretability

- GAM - Generalized Additive Models
- Linearity, montonicity
- Feature importance
- Tree interpreters (hot-topic)
- Complex decision tree is also a kind of black-box

- Python
- Open-source
- Single machine, in-memory, no GPU support
- Small, medium sized datasets
- Full-flow
- Classic ML algorithms (supervised, unsupervised …)
- Very selective algo inclusion rule(3 years since publication, +200 citations, …)
- Unified API (popular in other ml libs in python e.g. lightning)
- Lack of p, confidence interval,... for lin regression - use StatsModels if you like Python



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *Isomap*, *non-negative matrix factorization*. — *Examples*

Model selection

Comparing, validating and choosing parameters and models.

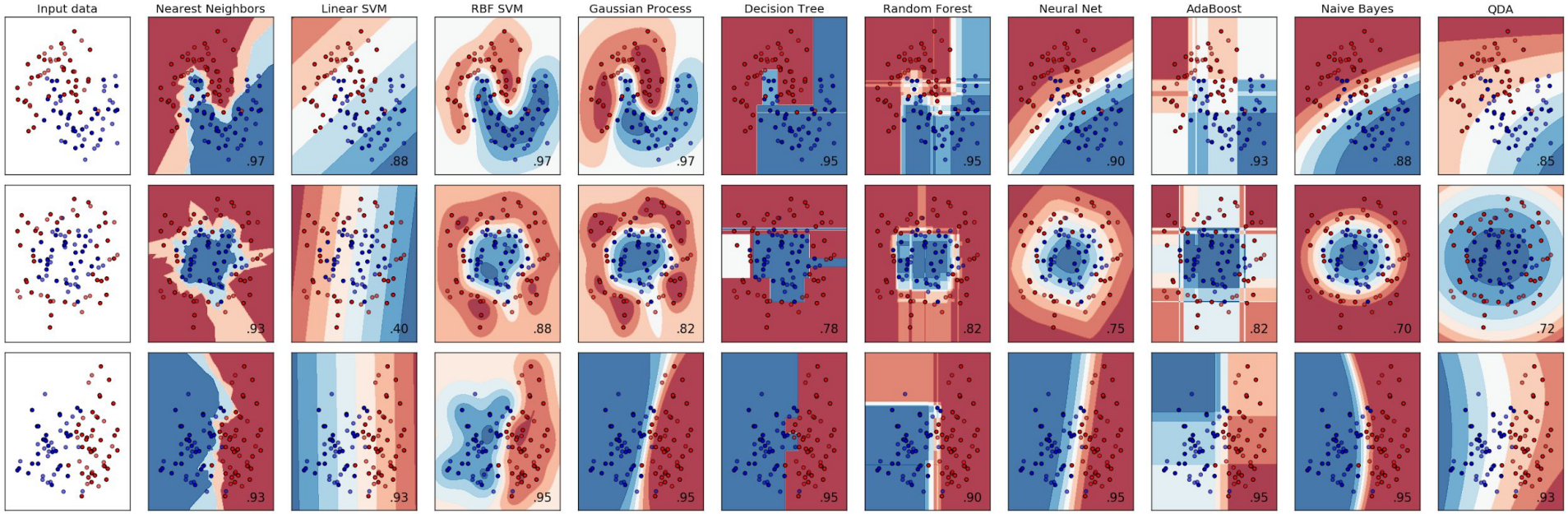
Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*. — *Examples*

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*. — *Examples*

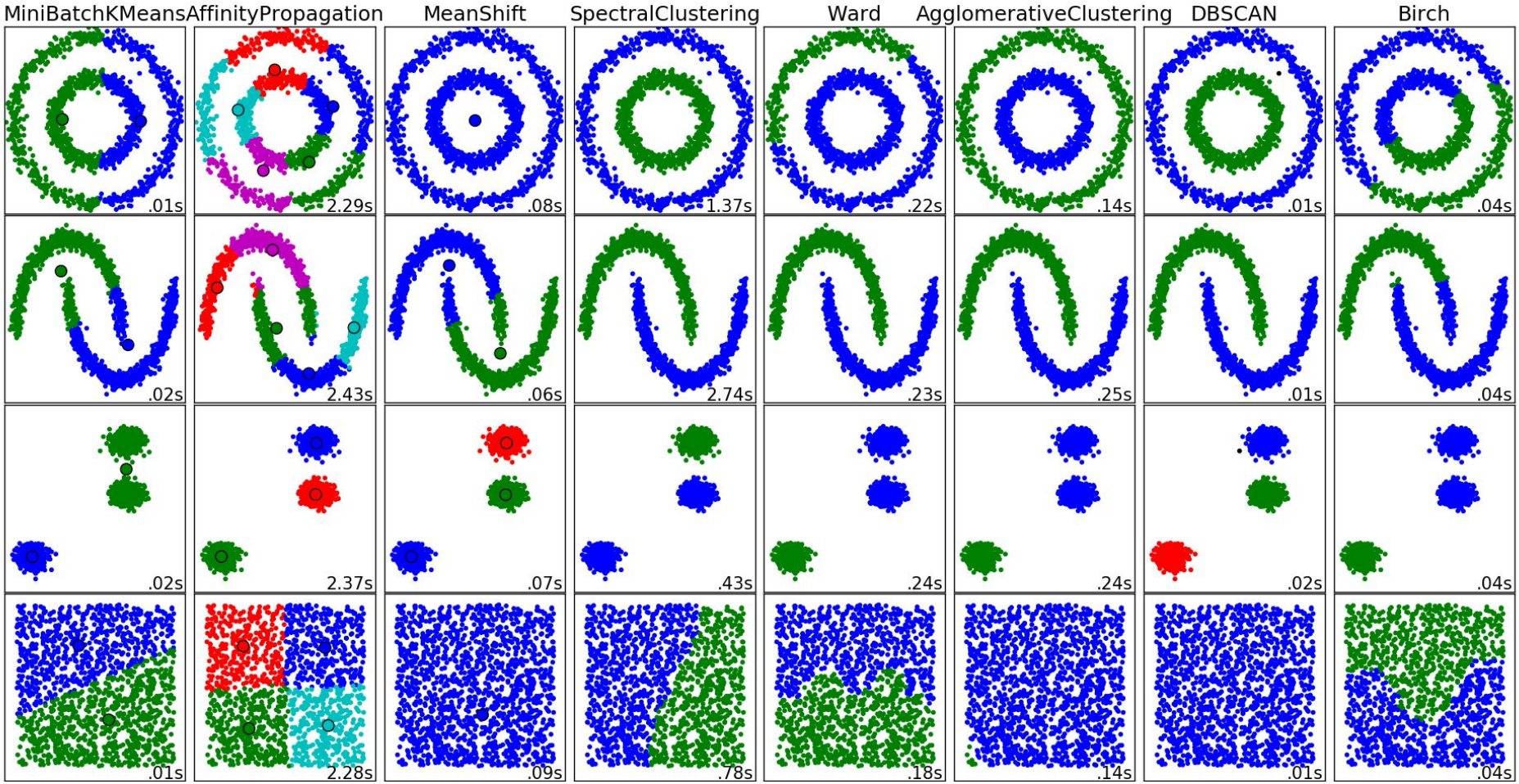
Visual comparison of classifiers



No Free Lunch

If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

David Wolpert and William Macready. No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation, 1:67, 1997.


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.29s
.08s
1.37s
.22s
.14s
.01s
.04s
.02s
2.43s
.06s
2.74s
.23s
.25s
.01s
.04s
.02s
2.37s
.07s
.43s
.24s
.24s
.02s
.04s
.01s
2.28s
.09s
.78s
.18s
.14s
.01s
.04s

Simple and consistent API

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier()
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
```

Xgboost, LightGBM

- Gradient boosting trees
- Open-source
- C++ with wrappers
- CPU and hybrid CPU/GPU (bleeding-edge)
- Scalable, distributed
- Classifiction, regression
- State of the art results for structured data
- Harder to tune (a lot of knobs) than RF
- Easy to overfit

Deep learning

- Keras with TF backend
- High-level NN API
- Open-source
- Python
- Grid-search across parameters and architectures…
- Input data need to be pre-processed
- Great for image classification etc.
- For structured data - part of ensemble, ROI

H2O

- Java internally
- Open-source
- Python/R interface, H2O Flow (GUI)
- Distributed, big-data
- A few but fast algorithms
- Spark by Sparkling Water
- AutoML

Spark MLib

- Distributed machine learning library
- **Scala**, Python, R
- More algos but not so good in benchmarks (Szilard)

Time series

- Keras based mcfly
- Facebook prophet
- pyFlux
- R packages e.g. Hyndman forecast
- Classic TS vs ML approach

Factorization machines

- fastFM and other implementations
- Typically C++ with python wrappers
- Recommender engines
- Great for modelling interactions with sparse and categorical data

R

- Similar to scikit-learn
- No unified API (caveat: caret package)
- More into descriptive than predictive modelling
- Domain specific packages (medicine, genomics etc.)
- Interesting packages (POV): DataExplorer, FastKNN, LargeVis, rotationForest...

Model tuning

- Manual
- **Grid-search**
- Random-search
- Bayes-search
- Multi-armed bandit approach (hyperband)
- ...

NUMERAI

A hedge fund built by a global community
of anonymous data scientists