Data Gathering and Cleaning

Claudio Delrieux

Data Gathering and Cleaning, Part I

- We can roughly conceptualize the tasks involved in a data science (DS) project into five: gather, clean, explore, model, and adequately represent the required datasets.

Data Gathering and Cleaning, Part II

- The different types of databases you may encounter are SQL (Postgre, Oracle, etc.), which require ETL processes.
- Non-relational databases (NoSQL) like Mongo, Solr, Riak, Cassandra, among others require specific programming.
- Another way to obtain data is to scrape from the websites using web scraping tools.

Data Gathering and Cleaning, Part III

- Websites such as Facebook and Twitter allows users to connect to their web servers and access their data using their Web APIs.
- The most traditional way of obtaining data is directly from files, such as downloading from existing datasets which are stored mostly in CSV files.

Data Gathering and Cleaning, Part IV

- Regarding the second task, obtaining a clean dataset, this one is likely the most widely ranging, complex, and sometimes frustrating.
- As we saw in project 2 last week, there are a variety of library functions devoted to some of these tasks.

Data Gathering and Cleaning, Part V

- One of the most common tasks is *filtering*, which can be regarded as finding a good subset of observations.
- *Projection*, on the other hand, is selecting the adequate variables that best represent the intended model.

Data Gathering and Cleaning, Part VI

- There are much more complex tasks that involve reshaping the dataset, combining two or more datasets into one, data reconciliation, imputation of corrupt or missing data, etc.

The End

Data Exploration, Modeling, and Representation

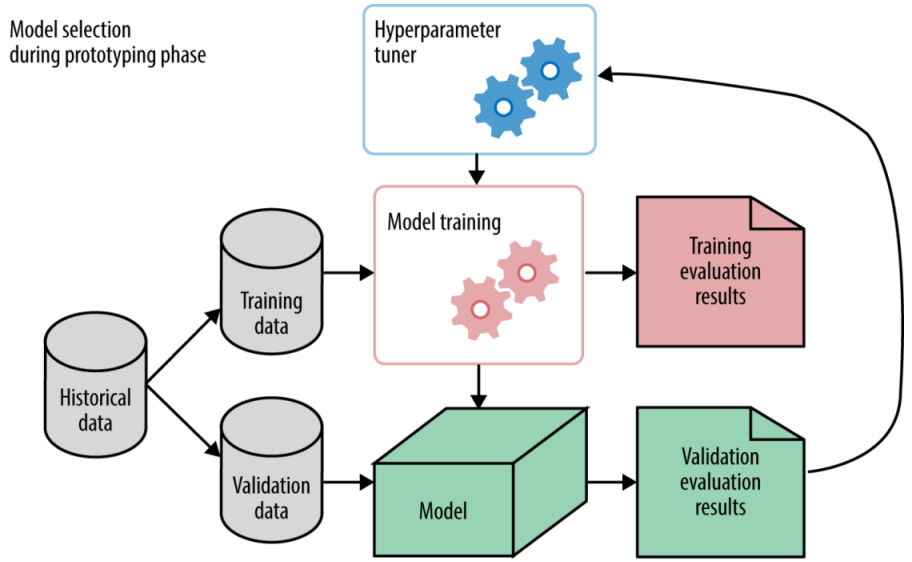
Claudio Delrieux

Data Exploration, Modeling, and Representation, Part I

- WRT data exploration consists of inspecting data and its properties. Different data types like numerical, categorical, ordinal, and nominal require different treatments.
- Then compute and visualize descriptive statistics to extract features and test significant variables or properties, find correlations, groups, or trends.

Data Exploration, Modeling, and Representation, Part II

- Building the data model is where most of the value is created (*"where the magic happens"*).
- However, a careful understanding of what is expected of a DS project is required before undergoing the model development.



Data Exploration, Modeling, and Representation, Part III

- As a general rule, we may classify the kind of required model by understanding the expected outcome.
- If we need to predict a category, group, or class of new data, then the most likely model is a *classifier*.
- A classifier is a supervised learning model that uses the known features of the training dataset to predict a categorical value.

Data Exploration, Modeling, and Representation, Part IV

- If we need to forecast a value, then the most likely model is a *regressor*.
- A *regressor* is a supervised learning model that uses the known values of the training dataset to learn a mathematical model that is able to predict the numerical value of a specific value.

Data Exploration, Modeling, and Representation, Part V

- If we need to group together records in our dataset (in not priorly known or understood groups), then the most likely model is *clustering*.
- *Clustering* is an unsupervised learning method that finds similarities and dissimilarities among records, using them to infer a partition of the dataset into groups.

Data Exploration, Modeling, and Representation, Part VI

- Sometimes, variations of these models are required. For instance, in *outlier detection,* we may use classifiers or clustering to build “normal” classes or clusters, and outliers would be data out of these classes or clusters.
- Also, other analysis models exist for specific purposes: time series, networks, and spatial analysis being the most prominent cases.

Data Exploration, Modeling, and Representation, Part VII

- Finally, model representation is required to help non-technical laypersons understand the meaning of the analysis.
- This is a key aspect that strongly depends on the deploy context (to whom and how are we communicating the analysis results).
- Thus it is difficult to generalize, but in most cases clever actionable dashboards depicting the KPIs are the most adequate solution.

The End

Examples in Corporate Activities

Claudio Delrieux

Examples in Corporate Activities, Part I

- A typical DS project arising in business is to prevent churning, to predict defaults, or similar customer-side prospective analysis.
- For this, the most important data source is the customer historical data, filtered (with no outliers) and class-balanced. This filtering is sometimes performed by some expert in the company, but other times is part of the analyst's task.

Examples in Corporate Activities, Part II

- As an example, suppose we are dealing with a telco, and our historical data looks something like this.

Client #	Income	Handset model	Plan type	Leftovers	Over-charges	Satisfaction	Churn?
342234	$67,000	I-phone	A	32 min/month	None	Good	No
243525	$88,000	Galaxy	A	None	8 min/month	Average	No
664342	$54,000	Lenovo	B	100 min/month	None	Good	Yes
121255	$59,000	Galaxy	B	None	None	Poor	Yes

Examples in Corporate Activities, Part III

- If a representative is contacting customers in risk of churning, also the data product needs to be understandable.
- In a situation like this, a decision tree-based model seems best (simple and transparent).
- This kind of model is sensitive to outliers, unbalanced classes, and care must be taken to avoid overfitting.

Examples in Corporate Activities, Part IV

- First an exploratory analysis is recommended over single variables (i.e., column-wise) to ensure the integrity and regularity of values.
- Second, if the amount of cases for the predicted variable (churn yes/no) is not properly balanced, then filter out randomly chosen cases of the most populous class.

Examples in Corporate Activities, Part V

- Third, while computing the decision tree, verify an adequate amount of cases per node, and also a reduced height.
- Usually more than one candidate model may arise under these constraints, each with different performance evaluation, so alternatives may have to be discussed with management.

Examples in Corporate Activities, Part VI

- Other common data products are related to targeted advertisement, and recommendation systems.
- These cases differ from the prior ones in that the context is dynamic (new products are to be advertised or recommended), and thus the models typically involve “on the fly,” unsupervised procedures.

Examples in Corporate Activities, Part VII

- In grocery stores, for instance, tables as the ones we showed before are infeasible (literally millions of rows and thousands of columns; these are typical big data contexts).
- Thus a “collaborative filtering” for profiling in advertisement or recommendation can be performed finding similitudes among different customers (similarity matching).

The End

Examples in Government

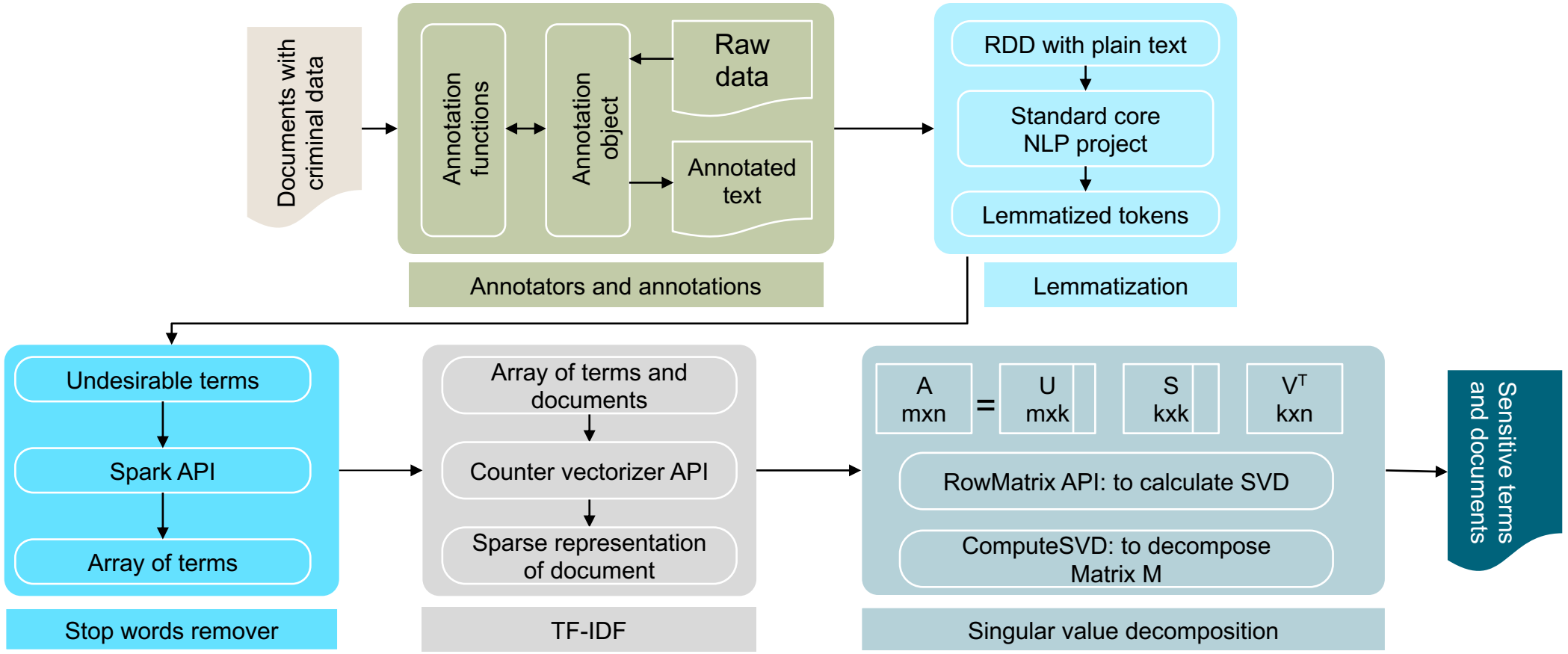
Claudio Delrieux

Examples in Government, Part I

- Terrorist activity detection and counteraction is key to break up threats at national and international levels.
- Information sources are diverse (text mining, emails, speech-to-text, social networks, and sometimes official documents from foreign countries), but analysis is mostly done on specific tagging tasks.

Examples in Government, Part II

- The goal is to detect sensitive information in this huge stream of text.



Examples in Government, Part III

- With the rise of face recognition systems and handheld MAC-address monitoring, CCTV-based surveillance is reaching unexpected (and perhaps undesired) capabilities.

Examples in Government, Part IV

- Facial recognition tools are currently helping identify and capture suspected criminals.
- Also, data analysis tools allow police to analyze motion and behavior data to identify jaywalking and track down violators.
- In China (and perhaps other countries), a vast database of information is kept on every citizen, including criminal and medical records, travel bookings, social media comments, and store visits.

The End

Examples in Science

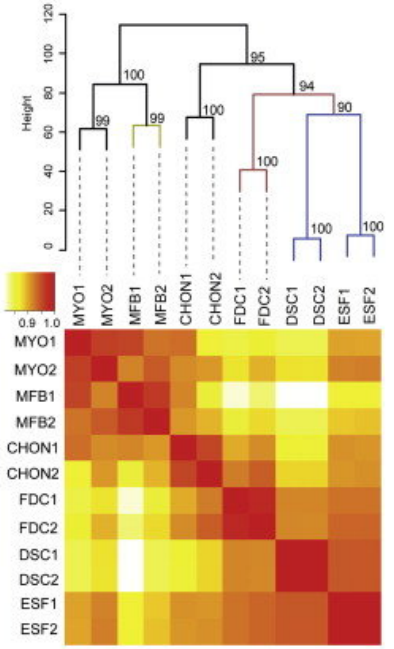
Claudio Delrieux

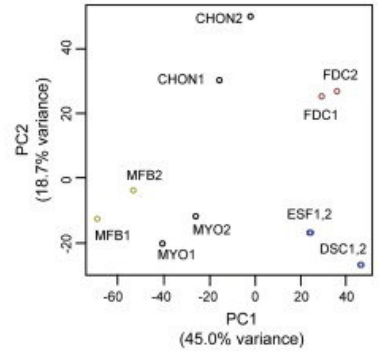
Examples in Science, Part I

- Scientists are several times required to model their data collections in a way such that makes sense for some specific purpose
- For instance, trying to organize the known living species is both useful, but puzzling
- A taxonomic tree is recognized as the right model, but its inference requires grouping together sets of species according to their features, for example, their genomic information

Examples in Science, Part II

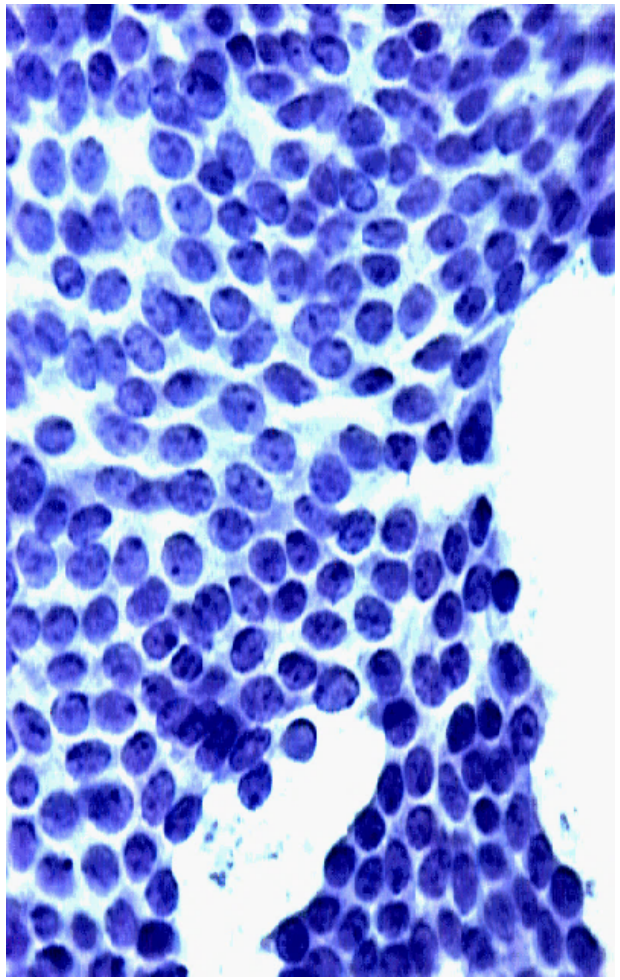
- The similitudes in specific RNA sequences of two given species can be put together in a covariance matrix (heat map)
- This allows one to perform a *hierarchical clustering* (above) based on the proximities in the principal components (below)





Examples in Science, Part III

- In many scientific and medical contexts, we have to deal with visual information that is plain to the expert human eye, but difficult to cope with using computer vision



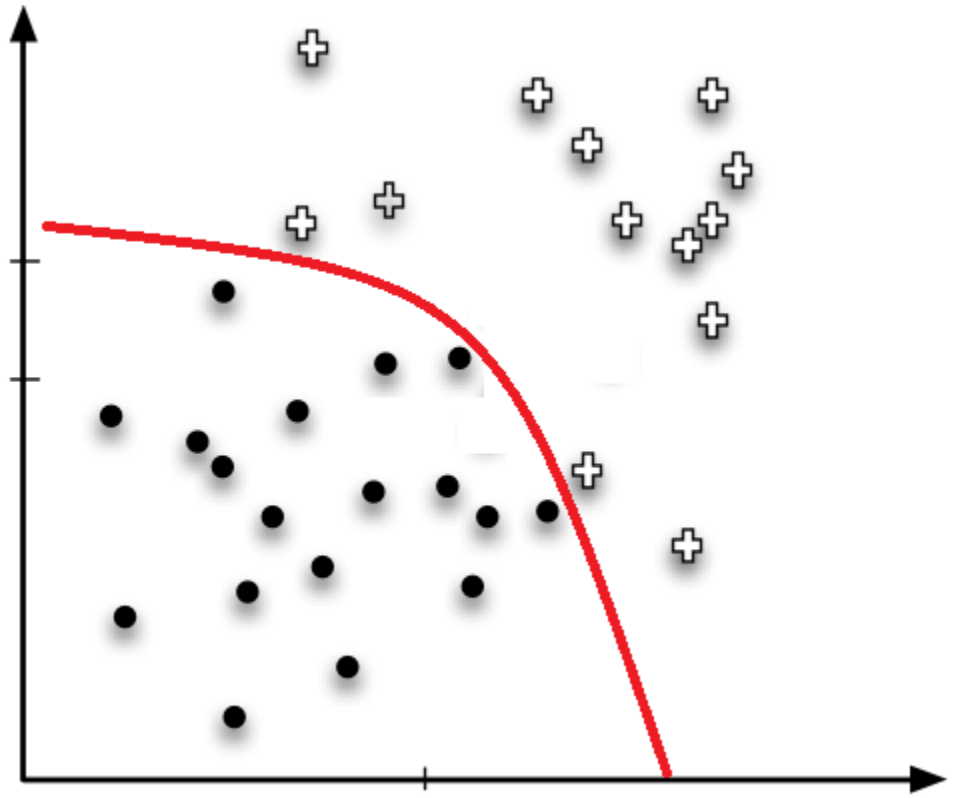
(*Cell image from the Wisconsin Breast Cancer dataset*)

Examples in Science, Part IV

- Typical image analytics provide parameters as:
  - **Radius:** *mean of distances from center to points on the perimeter*
  - **Texture:** *standard deviation of grayscale values*
  - **Perimeter:** *perimeter of the mass*
  - **Area:** *area of the mass*
  - **Smoothness:** *local variation in radii lengths*
  - **Compactness:** *computed as: perimeter^2/area*
  - **Concavity:** *severity of concave portions of the contour*
  - **Concave points:** *number of concave portions of the contour*
  - **Symmetry:** *a measure of the nuclei symmetry*
  - **Fractal dimension:** *“coastline approximation”*
- With this set of parameters, the purpose is to infer malignant vs. nonmalignant cells

Examples in Science, Part V

- A typical representation of the parameter space may lead to something as shown, where a separating function has to be inferred using *multivariate nonlinear regression*



The End

Classification Worked-Out Example

Claudio Delrieux

Classification Worked-Out Example, Part I

- Using SciKit-Learn iris dataset, we will recap different aspects of building a classifier.
- We want to recognize species of irises. The data consists of measurements of three different species of irises.

Setosa Iris

Versicolor Iris

Virginica Iris

Classification Worked-Out Example, Part II

- The features in the dataset are petal and sepal width and length (in cm), and the target class to predict (0, 1, 2).

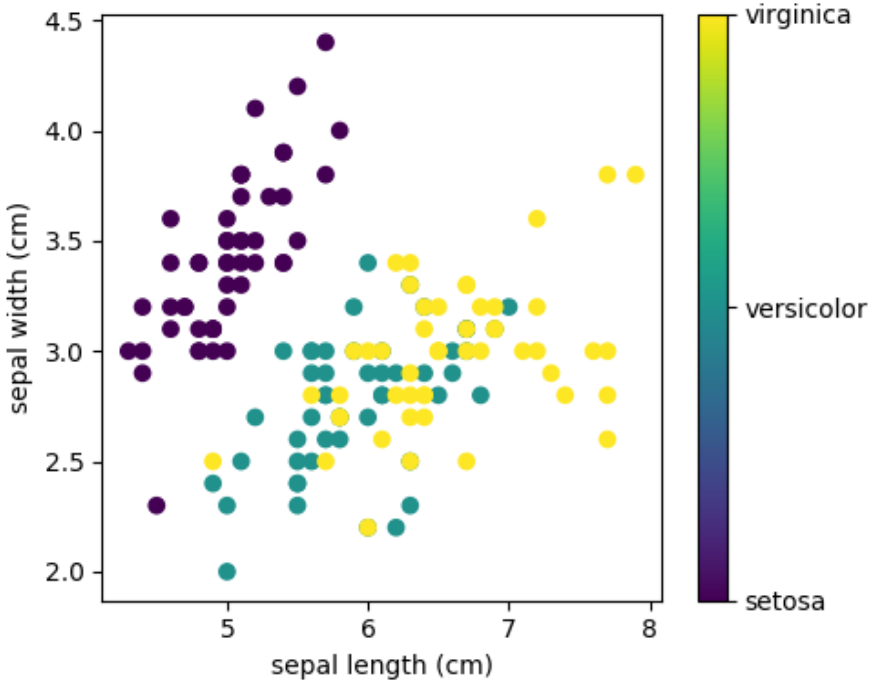
```
from sklearn.datasets import load_iris
iris = load_iris()
>>> print(iris.data.shape)
(150,4)
>>> print(iris.data[0])
[5.1  3.5  1.4  0.2]
```

Classification Worked-Out Example, Part III

- We can represent a projection of two of the features in a scatterplot.

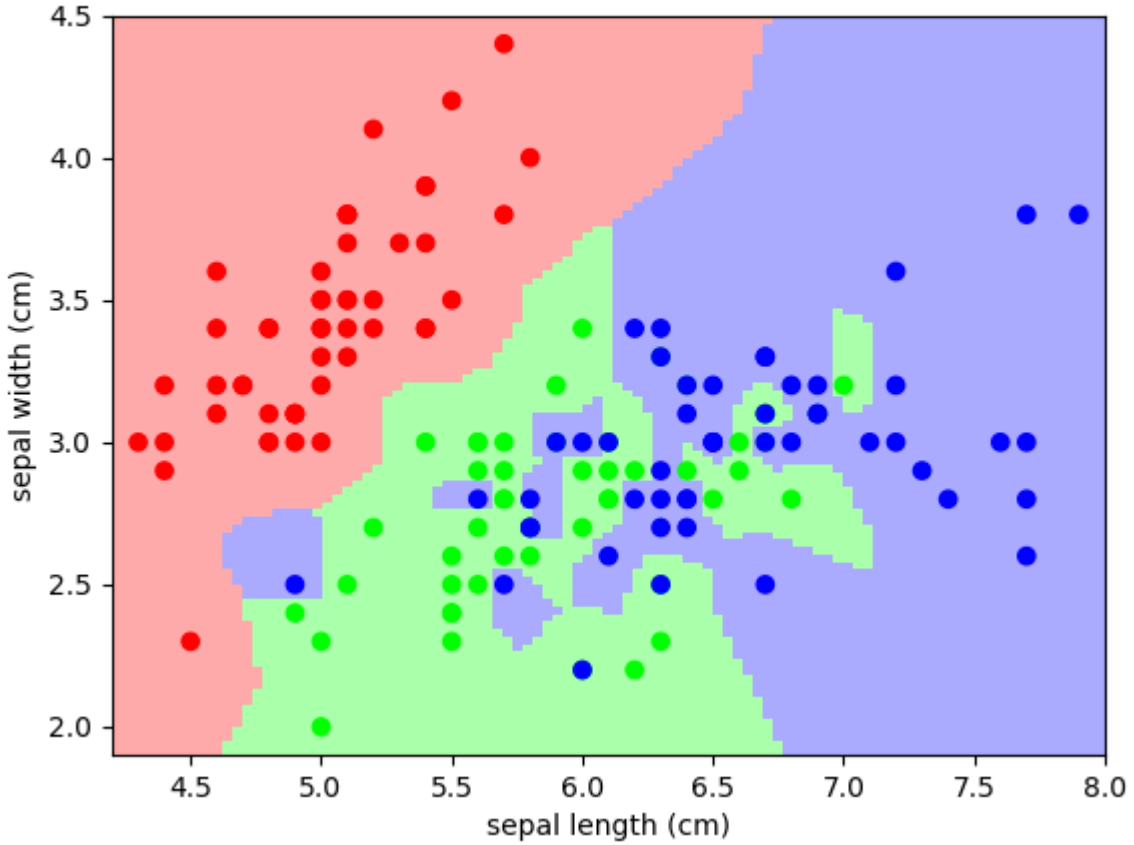
```
plt.figure(figsize=(5, 4))
plt.scatter(iris.data[:,x_index],
            iris.data[:, y_index],
            c=iris.target)
...

plt.show()
```



Classification Worked-Out Example, Part IV

- This plot shows that the dataset is not separable.
- A further exploration of the feature space using K-NN (N=1) shows that the classes are indeed mixed.



Classification Worked-Out Example, Part V

- Several methods for estimators are available.
- For example, model = KNeighborsClassifier(n_neighbors=1)

model.fit(): fits training data

model.predict(): predicts the label of new data

model.predict_proba(): returns the probability that a new observation has each categorical label

model.score(): implements a score method

Classification Worked-Out Example, Part VI

- Apart from K-NN, there are other estimators.
  - SVM (and kernel-SVM)
  - Decision trees (and random forests)
  - Bayesian classifier
  - Logistic regression
  - Ada boost (and ensemble classifier)
  - Linear (and quadratic) discriminant analysis

The End

Clustering Worked-Out Example

Claudio Delrieux

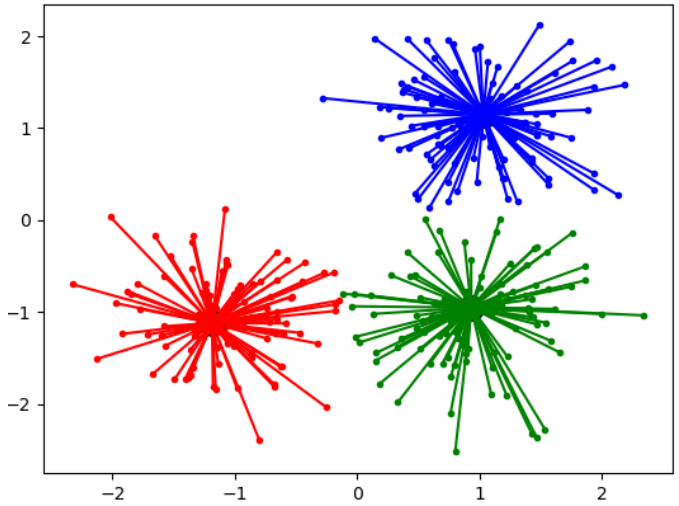
Clustering Worked-Out Example, Part I

- In SciKit-Learn we have the **make_blobs** function that allows us to create synthetic datasets (isotropic Gaussian blobs).

```
from sklearn.cluster import AffinityPropagation
from sklearn import metrics
from sklearn.datasets import make_blobs
centers = [[1, 1], [-1, -1], [1, -1]]
X, labels_true = make_blobs(n_samples=300,
  centers=centers,cluster_std=0.5,random_state=0)
```

Clustering Worked-Out Example, Part II

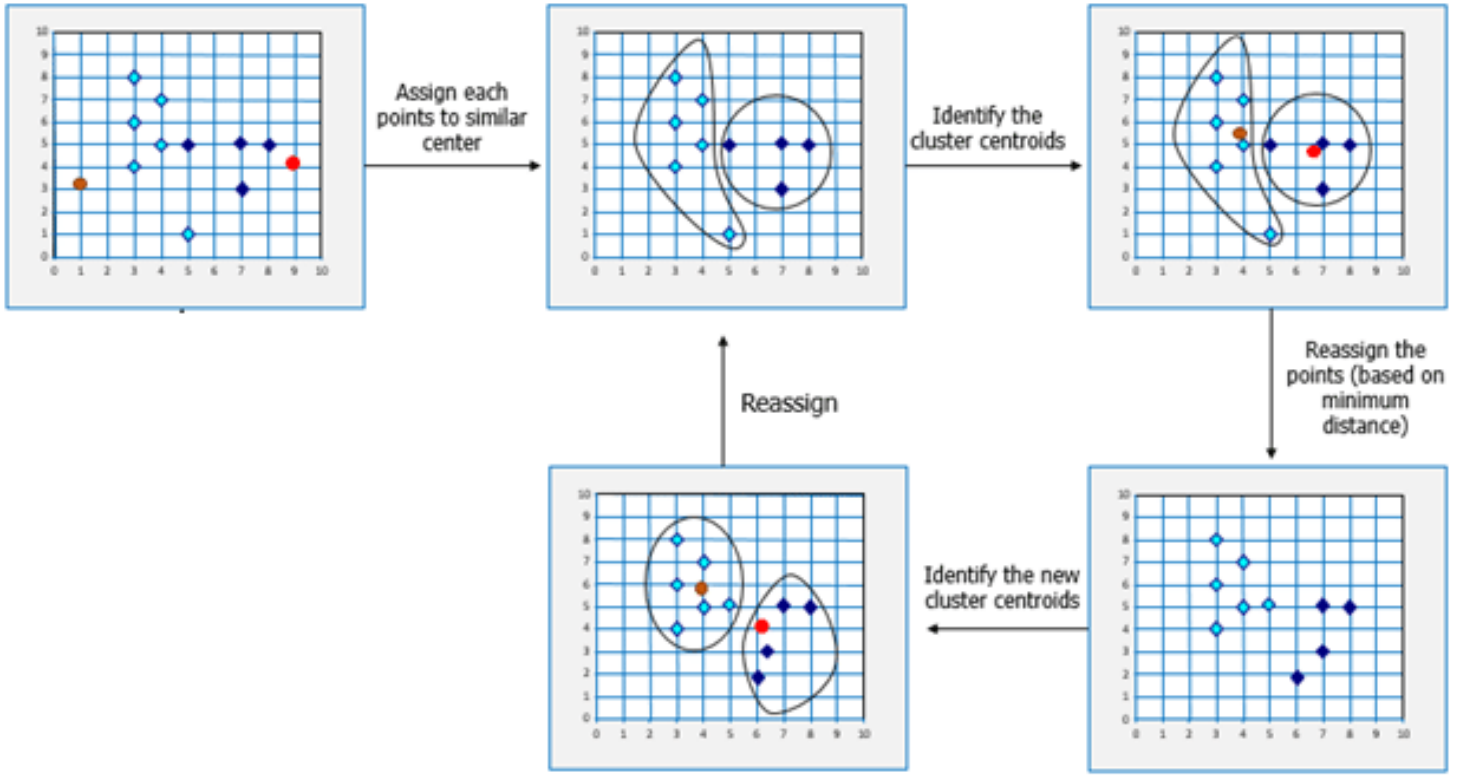
- With this we run one of the clustering methods, *affinity propagation*, that expands the cluster by sending messages among pairs of points.




```
af = AffinityPropagation(preference=-50).fit(X)
cluster_centers_indices=af.cluster_centers_indices
labels = af.labels_
n_clusters_ = len(cluster_centers_indices)
```

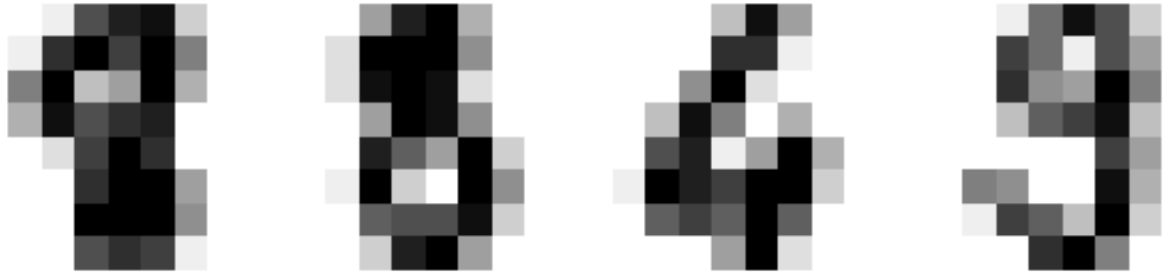
Clustering Worked-Out Example, Part III

- We have also the classic K-means algorithm that iterates between cluster assignation and centroid recomputing.



Clustering Worked-Out Example, Part IV

- We can use the “digits” dataset and see if the samples cluster together in a reduced space. For dimensionality reduction, we will use the two main principal components.



Clustering Worked-Out Example, Part V

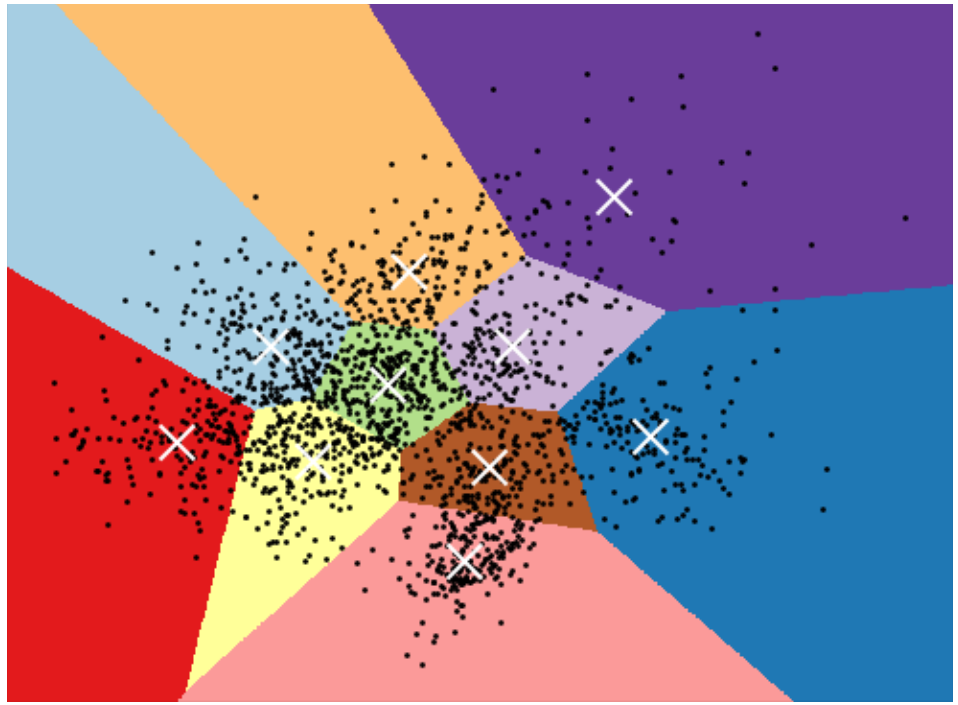
```
from sklearn.cluster import KMeans
from sklearn.datasets import load_digits
from sklearn.decomposition import PCA

X_digits, y_digits = load_digits(return_X_y=True)
data = scale(X_digits)
n_digits = len(np.unique(y_digits))
reduced_data = PCA(n_components=2).fit_transform(data)

kmeans = KMeans(init='k-means++', n_clusters=n_digits, n_init=10)
kmeans.fit(reduced_data)
```

Clustering Worked-Out Example, Part VI

- The white crosses indicate the ten cluster centroids, and the colored regions are the regions in PCA closer to each of these centroids.



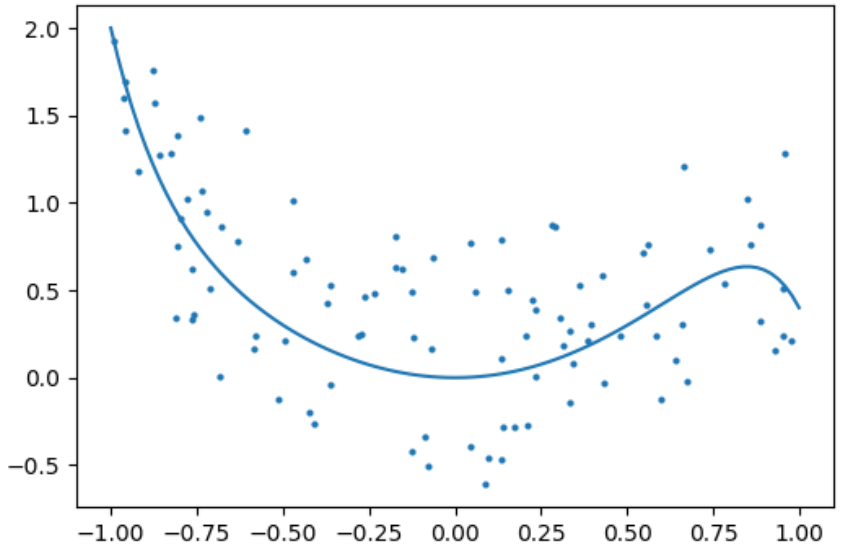
The End

Regression Worked-Out Example

Claudio Delrieux

Regression Worked-Out Example, Part I

We have this synthetic dataset (actually the points, not the curve).




```
rng = np.random.RandomState(0)
x = 2*rng.rand(100) - 1
f=lambda t:1.2 * t**2 + .1 * t**3 - .4 * t **5 - .5 * t ** 9
y = f(x) + .4 * rng.normal(size=100)
x_test = np.linspace(-1, 1, 100)
```

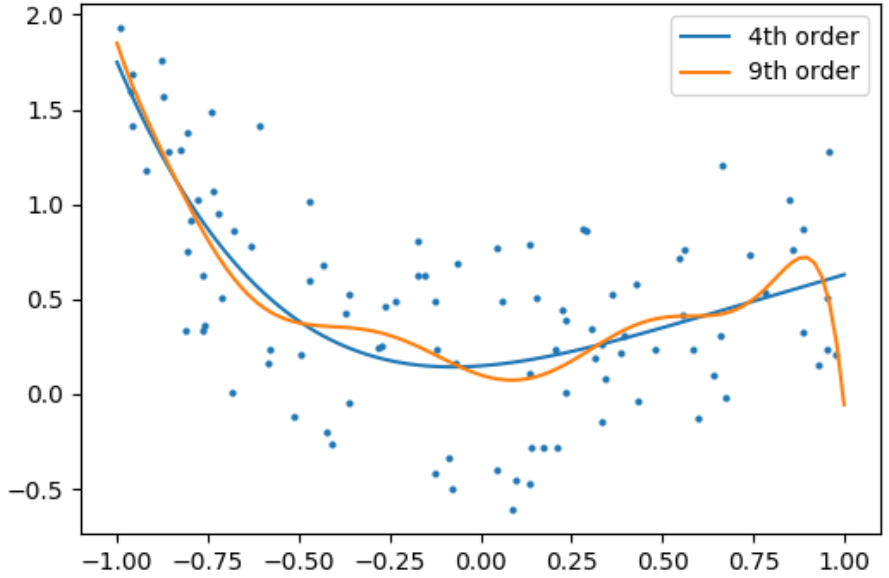
Regression Worked-Out Example, Part II

With this, a polynomial regression of the desired order can be performed.

```
X = np.array([x**i for i in range(5)]).T
X_test = np.array([x_test**i for i in range(5)]).T
regr = linear_model.LinearRegression()
regr.fit(X, y)
plt.plot(x_test, regr.predict(X_test), label='4thorder')
```

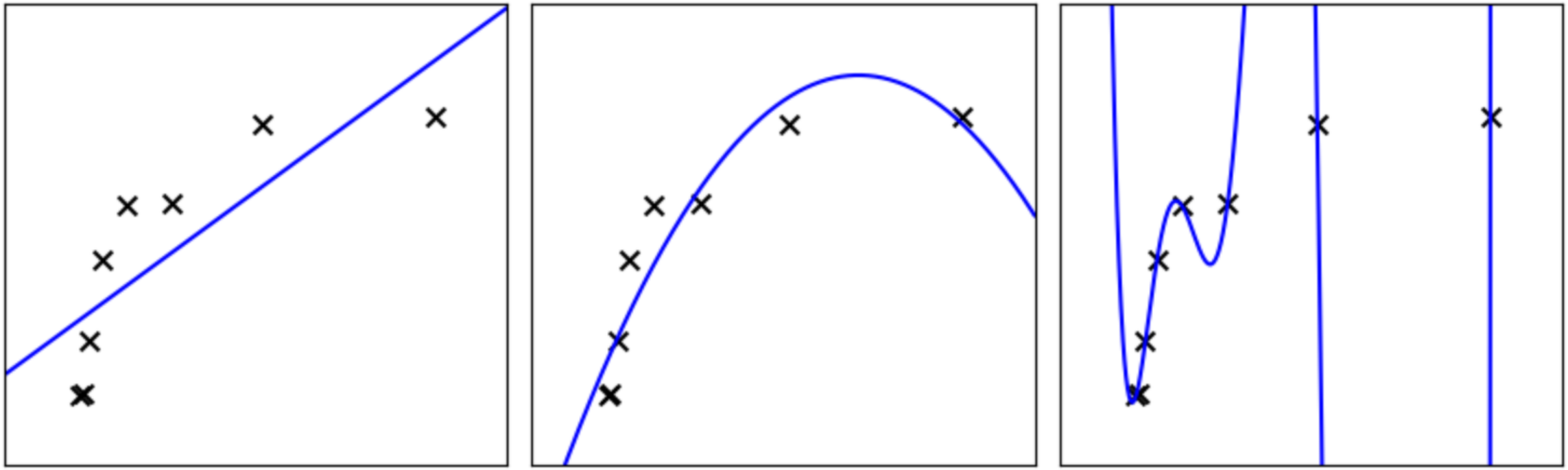
Regression Worked-Out Example, Part III

Repeating the code above for n = 5 and n = 9, we can have different fits. A higher order fit (even though it is the right order) appears to be sensitive to noise, while a lower order fit appears to adjust data better.



Regression Worked-Out Example, Part IV

In this, it is important to recall the concepts of underfitting (high bias) and overfitting (high variance), and try to establish a right equilibrium in the resulting model.



Regression Worked-Out Example, Part V

Apart from linear regression, SciKitLearn makes available other data products.

- LASSO, Ridge, ElasticNet (L1 and L2 regularization)
- RANSAC (dealing with outliers)
- Decision tree, random forest
- SVR, kernel SVR
- KNN regressor
- LARS

The End

Other Data Analysis Techniques

Claudio Delrieux

Other Data Analysis Techniques, Part I

- Apart from classification, clustering and regression, which are salient in the DS toolset, other analysis techniques may be required for specific purposes.
- Among them we can mention outlier detection, time series analysis, network analysis, dimension reduction, filtering, and spatial analysis (even though the list is longer).

Other Data Analysis Techniques, Part II

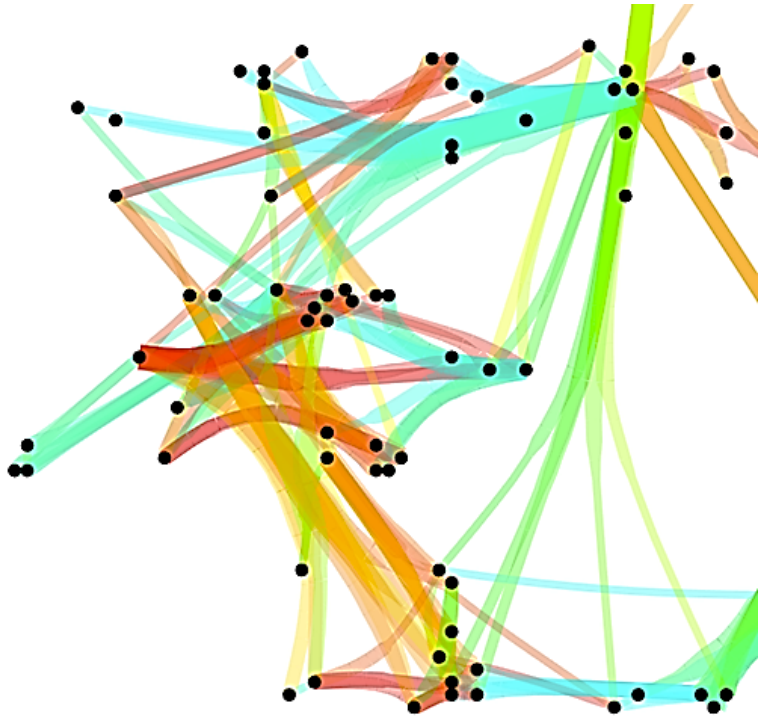
- **Outlier detection** is concerned with detecting atypical data (offline or online) according to some criteria.
- Outliers can be cases that don’t fit to any known category (a variant of classifiers), or that cannot be reasonably grouped together with other cases (a variant of clustering), or anomalous values that cannot be regressed within ordinary parameters.

Other Data Analysis Techniques, Part III

- **Time series** are unique in the sense that **time** is a specific independent variable that imposes particular features to the dataset. Time series are also analyzed in the context of **signal processing**, but with different purposes. In DS the interest is mainly focused on establishing trends, periodic patterns, and fluctuation analysis.

Other Data Analysis Techniques, Part IV

- **Network analysis** is unique in the sense that it involves connectivity relations that provide hierarchical and topological properties to the structure. Typical analysis goals are path (and shortest path) finding, hub and clique identification, and spanning tree computation.



Other Data Analysis Techniques, Part V

- **Dimension reduction** is not *per se* an analysis technique. The main purpose is to eliminate superfluous variables or to find alternative, more compact, representations of the dataset. In these reduced representations, ulterior analysis tasks are much easier since there is no superfluous information obscuring the relevant factors.

Other Data Analysis Techniques, Part VI

- **Filtering** also is not an analysis technique, but can be used for analysis purposes. In collaborative filtering, for instance, elements in the dataset are filtered according to a specific item, which implicitly performs a grouping or clustering of the part of the dataset that is similar in some sense to that specific item.

Other Data Analysis Techniques, Part VII

- **Spatial analysis** is unique in the sense that specific geographical, geometrical, or political division properties are relevant. Classification, clustering, and regression with spatial data require taking into account weighing factors, and the statistical models differ from the standard ones.

The End

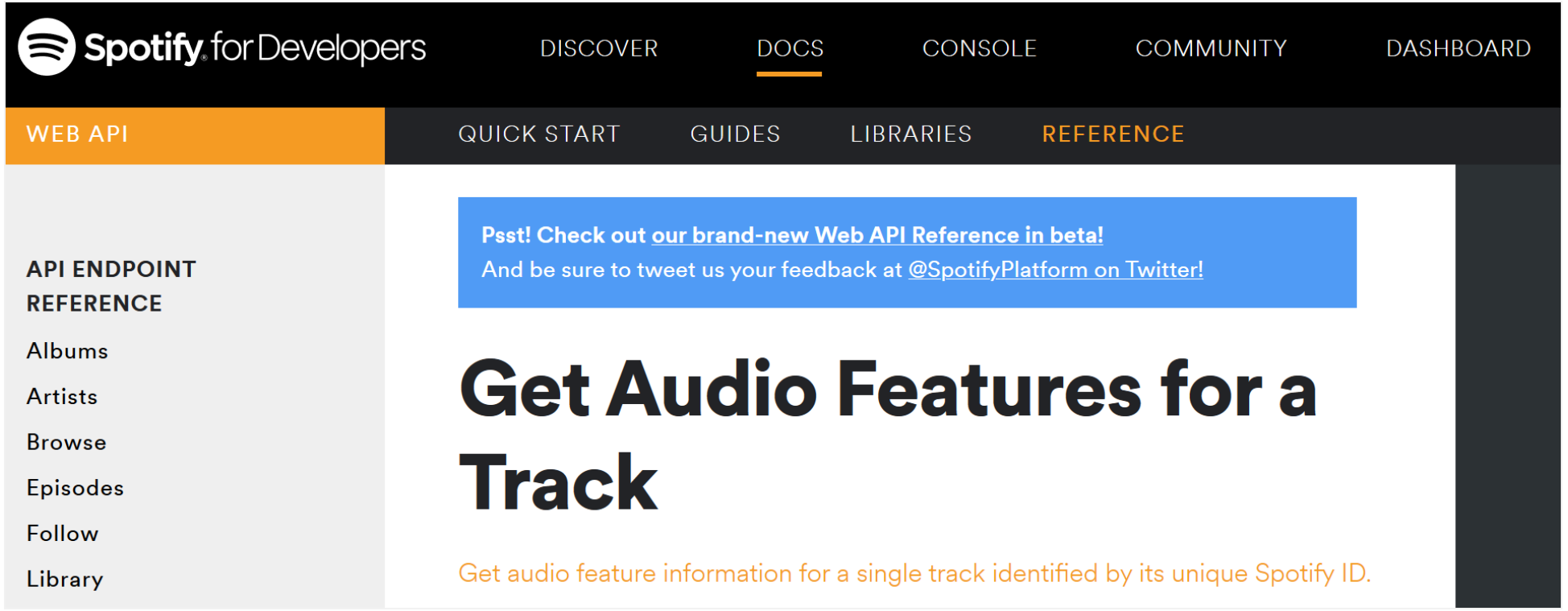
Cases for Mid-Term Project

Classifiers

Claudio Delrieux

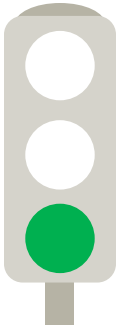
Cases for Mid-Term Project: Classifiers, Part I

- **The Beat of the Music:** The goal of this project is to predict the liking of songs through their musical attributes.



Cases for Mid-Term Project: Classifiers, Part II

- This dataset is composed by 2017 songs with attributes from Spotify's API. Each song has been labeled by someone who used "1" if she likes the song or "0" if she doesn't like it. Each row represents a song. There are 16 columns, 13 of which are song attributes, one column for song name, one for artist, and a column called "target," which is the aforementioned label for the song.



Cases for Mid-Term Project: Classifiers, Part III

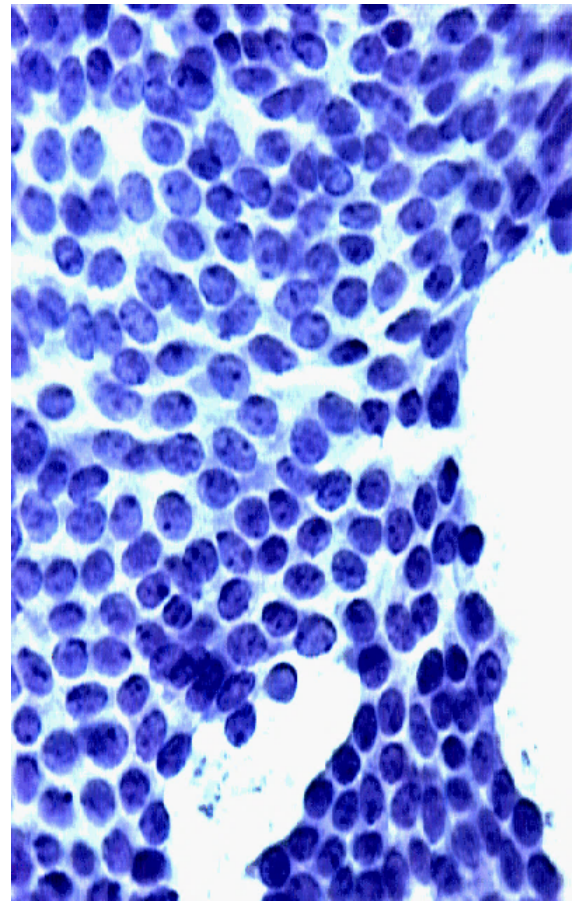
- **COVID-19 Detection in Chest X-Ray Images:** The goal of this project is to build a model that identifies COVID-19 pneumonia in chest X-ray images. There are 606 X-ray images organized into three folders (COVID-19, normal, pneumonia).

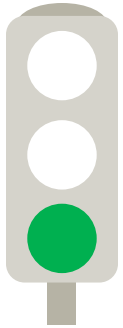
Cases for Mid-Term Project: Classifiers, Part IV

- **Classifying Fake News:** Your goal is to determine if an article is fake news or not. There are two files, one with news labeled as fake, and another with news labeled as true. You have four columns on each file: title, text, subject, date.

Cases for Mid-Term Project: Classifiers, Part V

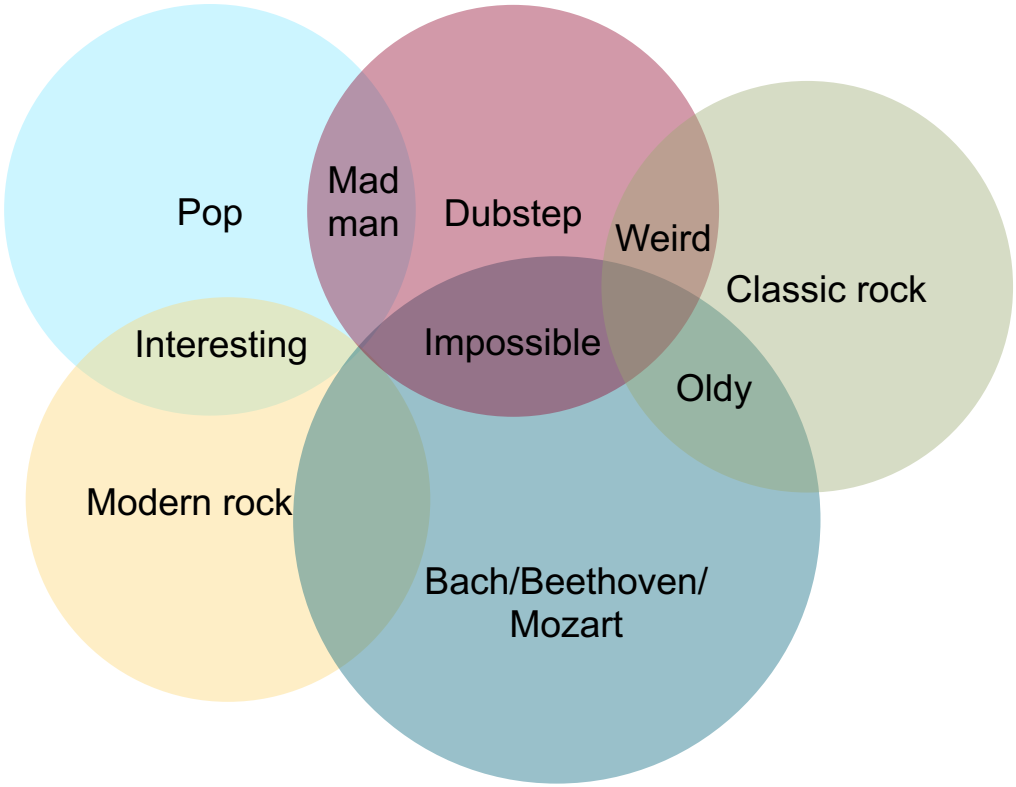
- **Breast Cancer Diagnostic:** This is the Winsconsin Breast Cancer dataset mentioned before. The goal is to build a model able to predict breast cancer tissues as malignant or benign.

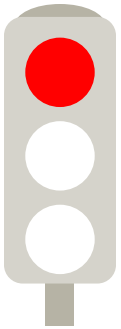




Cases for Mid-Term Project: Classifiers, Part VI

- **Music Genre Classification:** The dataset consists of 1,000 audio tracks each 30 seconds long, with 10 genres (100 tracks each). Extract audio features and create a model able to classify tracks by genre using these features.
- A notebook useful to open and explore audio tracks is provided.



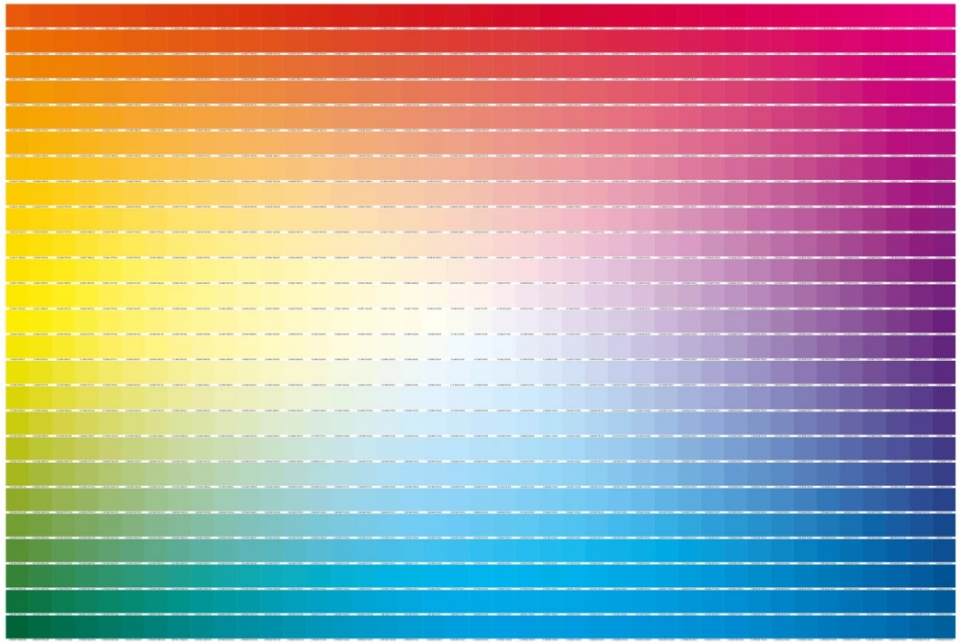


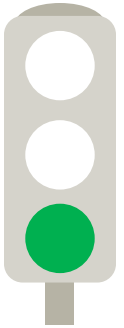
Cases for Mid-Term Project: Classifiers, Part VII

- **Detecting Credit Card Fraud:** The datasets contain transactions made by European cardholders during two days, where we have 492 frauds out of 284,807 labeled transactions. The goal is to build a model able to recognize which transactions are fraudulent.

Cases for Mid-Term Project: Classifiers, Part VIII

- **Rainbow Guru:** The dataset contains a CSV file that has 865 color names with their corresponding RGB (red, green, and blue) values of the color. It also has the hexadecimal value of the color. The goal is to build a model able to predict color names from their hexadecimal/RGB code.





Cases for Mid-Term Project: Classifiers, Part IX

- **Recognizing Traffic Signs:** The goal is to build a model capable of determining the type of traffic sign that is displayed in an image captured under different real-life conditions. Provided are labeled images already split in three files (tran/val/test) and a predefined list of signs to recognize.

The End

Cases for Mid-Term Project

Clustering

Claudio Delrieux

Cases for Mid-Term Project: Clustering, Part I

Trends in Beer Preferences: Your goal is to group similar beers together according to reviews, and to recommend one in each group.

Cases for Mid-Term Project: Clustering, Part II

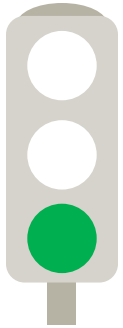
This dataset contains 1.5M reviews of beer scrapped from BeerAdvocates. Try to answer using the data: If you had to pick five beers to recommend, which would you pick? Why? Which of the factors (taste, aroma, appearance, palate) are most important in determining the overall quality of a beer? If I usually enjoy IPAs, which beer should I try?

Cases for Mid-Term Project: Clustering, Part III

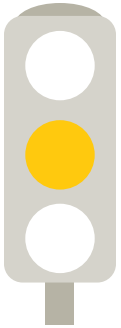
Wholesale Customer Segmentation: The goal is to identify and describe the various customer segments hidden in the data based on their annual spending on diverse product categories.

Cases for Mid-Term Project: Clustering, Part IV

Violent Crime Rates by U.S. State: This data set contains statistics for violent crimes in the U.S. states during 1973, and the percent of population living in urban areas. The goal is to identify hierarchies of U.S. states according to violent crime rates.



Cases for Mid-Term Project: Clustering, Part V



Chatbot: Build a model that predicts answers using predefined patterns and responses. You are provided with a file that contains these patterns. Words and classes files are provided as extra help.

You: Hello, how are you?

Bot: Hi there, how can I help?

You: What can you do?

Bot: I can guide you through adverse drug reaction list, blood pressure tracking, hospitals and pharmacies.

The End

Cases for Mid-Term Project

Regression

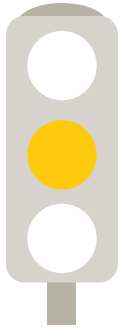
Claudio Delrieux

Cases for Mid-Term Project: Regression, Part I

Boston Housing Prices: The dataset contains 506 observations and 14 variables. The goal is to understand the variables driving the prices of homes and to predict prices from the attributes.

Cases for Mid-Term Project: Regression, Part II

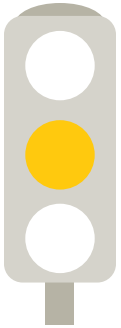
Predicting NBA Salaries: The goal is to build a model that predicts salaries of NBA players. We have two files. The first contains players and their corresponding stats. The second provides seasonal information about their salaries from 1985 to 2018.



Cases for Mid-Term Project: Regression, Part III

January Flight Delay Prediction: The dataset contains flights from January 1, 2019 till January 31, 2019 (around 400,000 rows and 21 features). The goal is to predict flight delays.

Cases for Mid-Term Project: Regression, Part IV

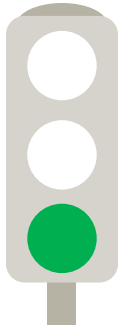


Predicting IMDB Rating for Movies: The goal of this project is to build a model that predicts the IMDB rating score based on attributes as duration, actors, or even titles of 5,000 movies (28 attributes in all).

IMDb

Cases for Mid-Term Project: Regression, Part V

Spotify Top Songs: The dataset contains daily ranking of the 200 most listened songs in 53 countries in 2017 and 2018. The goal of this project is to understand which features make a song popular.



Spotify

Cases for Mid-Term Project: Regression, Part VI

Wine Reviews: The goal is to regress wine prices from attributes in 300K reviews scraped from Wine Enthusiast during 2017. Which are positively reviewed? Does this correlate with price?

The End

Wrap-Up

Claudio Delrieux

Wrap-Up, Part I

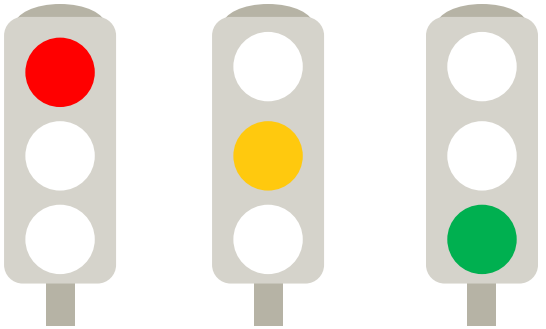
- Next, we will present some of the tools and libraries that may be needed for your mid-term project.
- After that, there will be time for consultation regarding the project preparation.
- Finally, you will present your mid-term projects according to the following guidelines.

Wrap-Up, Part II

- Mid-term project presentation will consist of:
  - Class presentation (four minutes plus two minutes for questions)
  - A documented netbook containing the complete project
  - A three to four-page report including dataset description, data preparation and analysis, results and discussion

Wrap-Up, Part III

- Mid-term projects have a “traffic light” indicating “easy,” “intermediate,” and “difficult,” and therefore the expected presentation and grading will not be the same in each case.



Wrap-Up

The End