



UNIVERSIDAD DE CHILE

# Minería de Datos

Welcome to the Machine Learning class

---

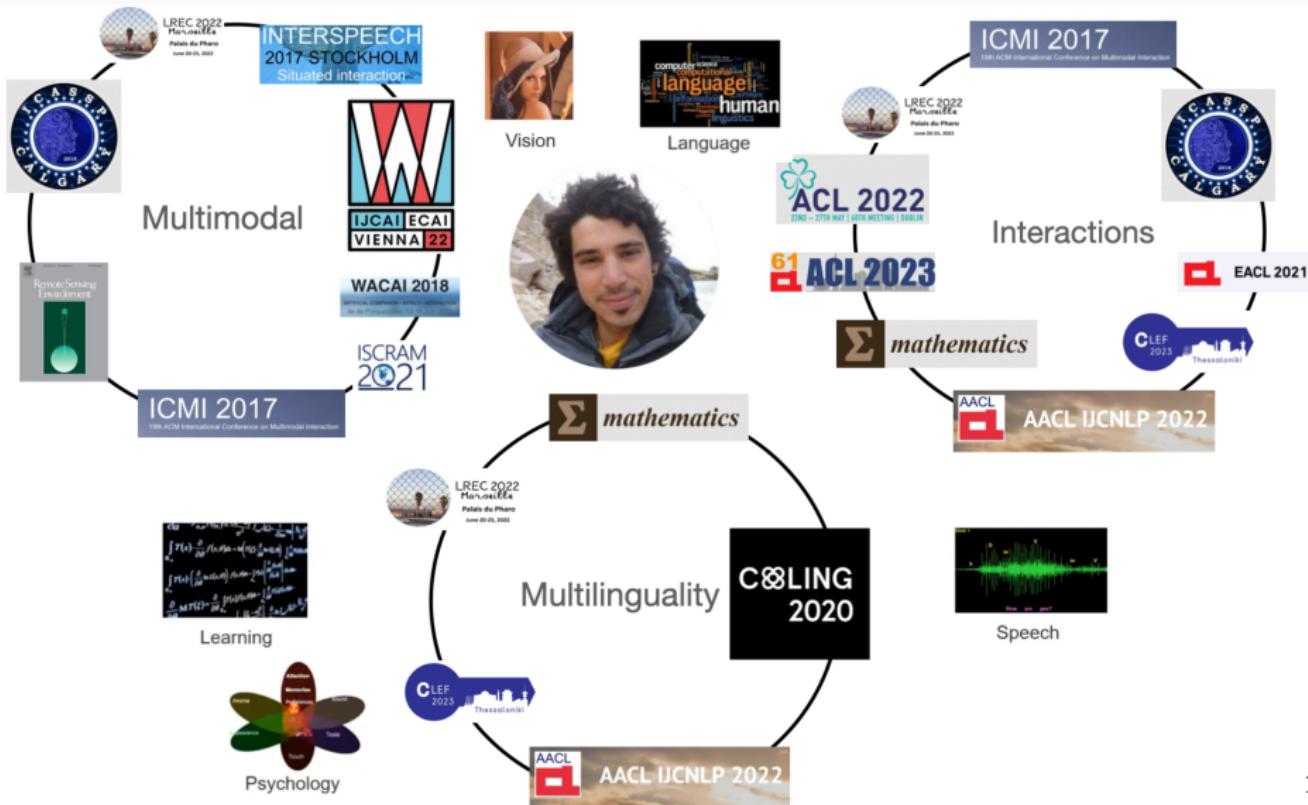
Valentin Barriere

Universidad de Chile – DCC

CC5205, Fall 2025

# Introduction

# Who am I?



## Definition

Data mining aims at the extraction of knowledge from large amounts of data using automatic or semi-automatic methods.

It proposes to use a set of algorithms [...] to build models from the data, that is, to find interesting structures or patterns according to predefined criteria, and extract as much knowledge as possible from them.<sup>1</sup>

---

<sup>1</sup>[https://fr.wikipedia.org/wiki/Exploration\\_de\\_données](https://fr.wikipedia.org/wiki/Exploration_de_données)

# Data

Different types of data:

- Structured data:
  - Social data: Age, Salary, Skin color, Place of residence
  - Metric data: *Likes* on a post, Time spent on a page, Number of mutual friends
- Unstructured data:
  - Text: Sentence, Paragraph, Document
  - Sound: Song, Speech
  - Image: Photo, Video

## Different types of Mining:

- Data exploration: Detect simple values, biases
- Classification/Regression task: Use data to characterize new data **by class or with a value** in a supervised manner
- Clustering task: Group data into classes in an unsupervised manner
- Dimensionality reduction: Develop common structures for compressed data representations

# Outline : Applications

## Applications

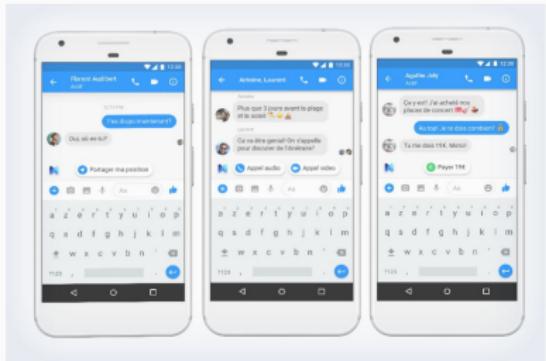
Meaning of the Terms

Prerequisites

Data Exploration Practical:  
MovieLens  
Overview

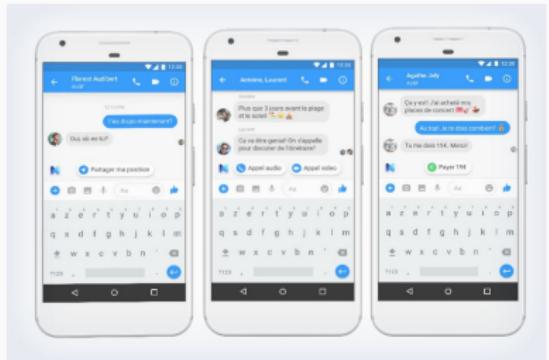
# Applications (I/II)

- Event detection in text



# Applications (I/II)

- Event detection in text

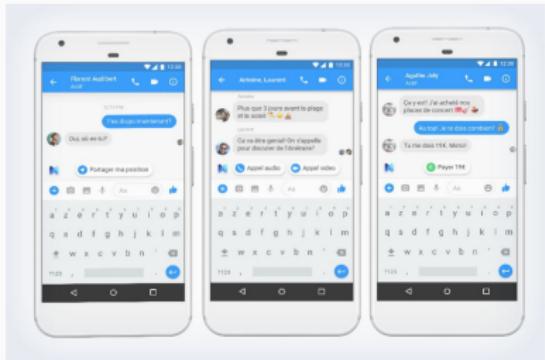


- Automatic processing of user opinions



# Applications (I/II)

- Event detection in text



- Automatic processing of user opinions



Les connaissez-vous ?



**Marina Dunion**  
Digital Marketing @Air France  
& Co-Founder @Flexifly  
● Teddy Viraye-  
Chevalier et 3 autres  
relations



**Salvatore Anzalone**  
Post-Doc at ISIR, University  
Pierre et Marie Curie, Paris  
● Thomas Janssoone et  
2 autres relations

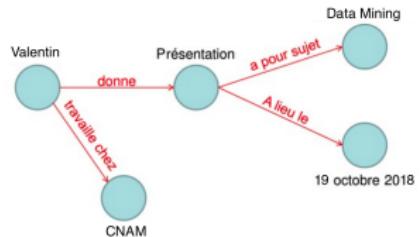


**Halla Olafsdottir**  
Medical Solutions Project  
Manager | Chef de Projet  
● Télécom ParisTech

- Recommending items to a user

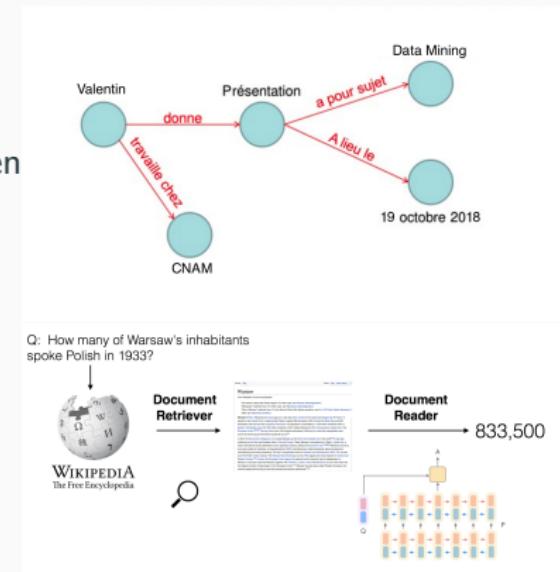
## Applications (II/II)

- Detection of relationships between entities in a text



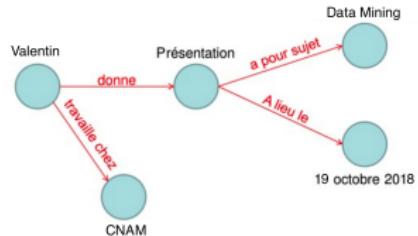
# Applications (II/II)

- Detection of relationships between entities in a text
- Question answering

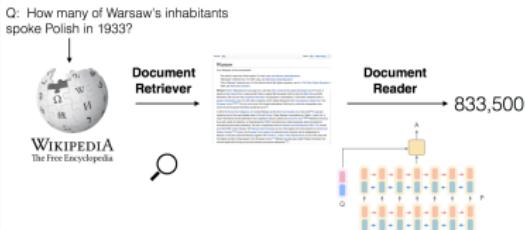


# Applications (II/II)

- Detection of relationships between entities in a text



- Question answering



- IE module for a conversational agent



# Outline : Meaning of the Terms

Applications

**Meaning of the Terms**

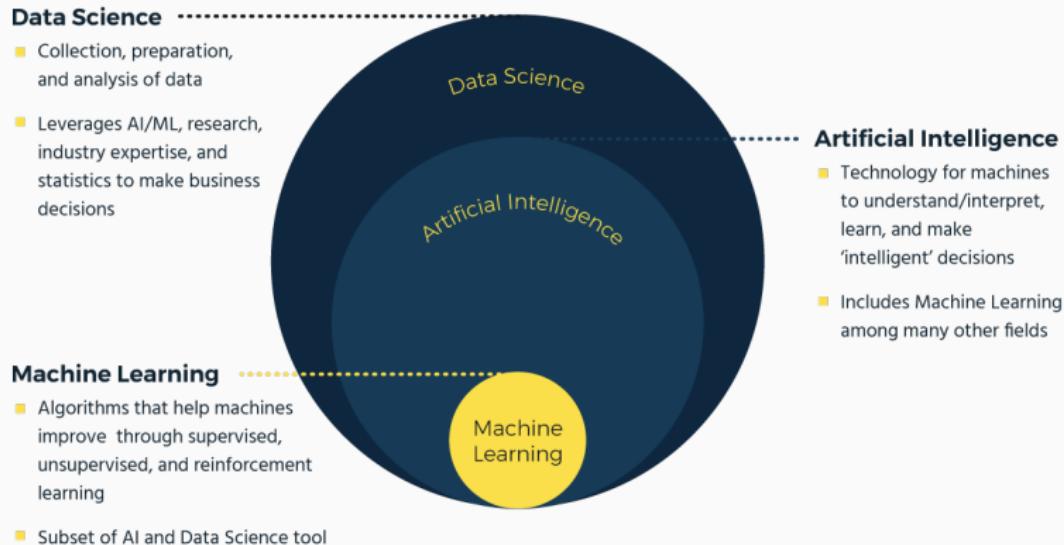
Prerequisites

Data Exploration Practical:

MovieLens

Overview

## AI vs. Data Science vs. Machine Learning



**Figure 1:** Differences between fields

## In summary

Data Science focuses on analyzing data to extract knowledge, Machine Learning uses algorithms to make predictions and decisions based on data, and Artificial Intelligence refers to the development of systems that can perform intelligent tasks autonomously.

Overly simplistic definition:

- Data mining generates understanding.
- Machine learning generates predictions.
- Artificial intelligence generates actions.

# Example on a Music Platform

## Data Scientist

Collects and analyzes user data from music platforms to identify patterns and musical preferences.

## Machine Learner

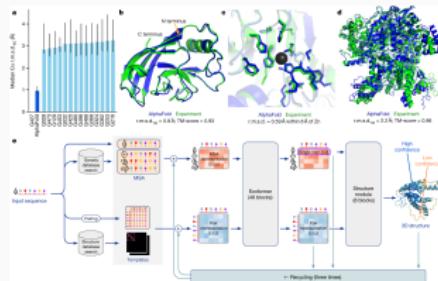
Develops and optimizes a music recommendation model using machine learning algorithms to predict user preferences.

## Artificial Intelligence

Implements a social agent that can interact with the user, to enhance the personalization of music recommendations and provide a more accurate and contextualized experience.

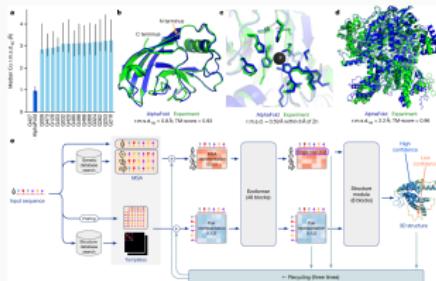
## **Significance of the Work: Why Do It?**

- Scientific advancement



# Significance of the Work: Why Do It?

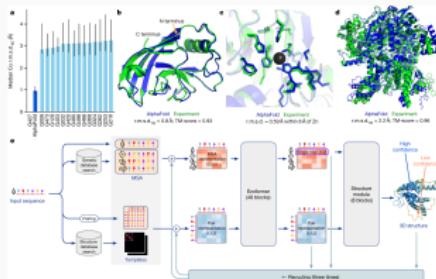
- Scientific advancement



- Prevention and management of natural disasters

# Significance of the Work: Why Do It?

- Scientific advancement



- Prevention and management of natural disasters

- Impact on public health



**Open Chronic**

Améliorer la prise en charge des malades chroniques

Santé    Promotion 3

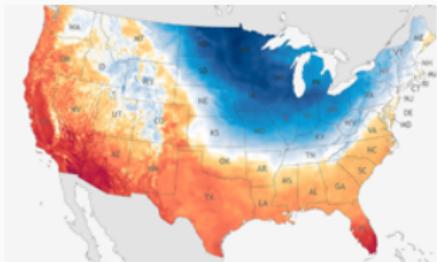
Ministère de la santé, Direction de la recherche, des études, de  
et des statistiques

Paris

Data science

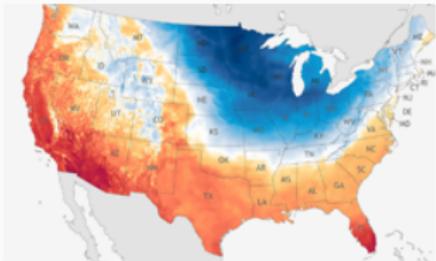
## **Significance of the Work: Why Do It?**

- Environmental sustainability

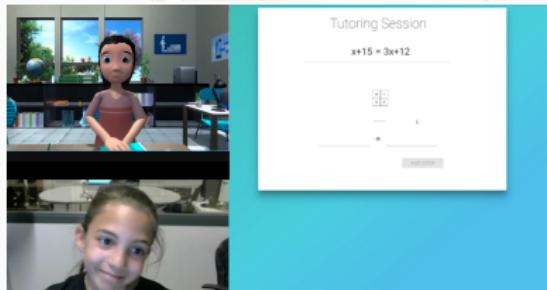


# Significance of the Work: Why Do It?

- Environmental sustainability

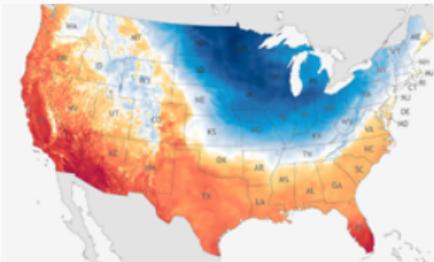


- Boost to education and research

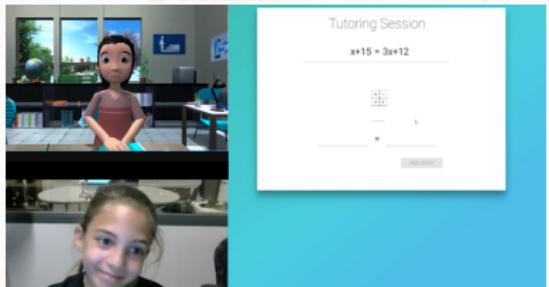


# Significance of the Work: Why Do It?

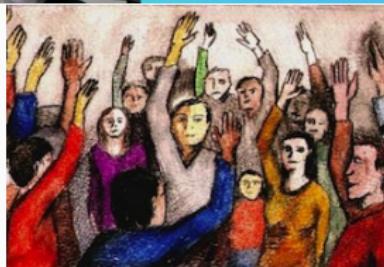
- Environmental sustainability



- Boost to education and research



- Participatory democracy



# Outline : Prerequisites

Applications

Meaning of the Terms

Prerequisites

Data Exploration Practical:

MovieLens

Overview

# The Program

## Theoretical Parts

Foundations in statistics and linear algebra: General overview of statistical learning, mathematical foundations of different models, intuitive approach

## Practical Parts

Python basics: Use of data manipulation tools, use of a DL library, use of an ML library, sentiment analysis, ranking based on wine preferences, ...<sup>2</sup>

---

<sup>2</sup>Non-contractual for this course

# Material

- Computer
- Jupyter Notebook and Anaconda:  
<https://www.anaconda.com/download/>
- The notebooks and cheatsheets available online:

## Python For Data Science Cheat Sheet

### NumPy Basics

Learn Python for Data Science interactively at [www.DataCamp.com](http://www.DataCamp.com)

### NumPy

The NumPy library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Use the following import convention:

```
>>> import numpy as np
```

### NumPy Arrays



### Creating Arrays

More... [View on GitHub](#)

## Python For Data Science Cheat Sheet

### Pandas Basics

Learn Python for Data Science interactively at [www.DataCamp.com](http://www.DataCamp.com)

### Pandas

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.

pandas

Use the following import convention:

```
>>> import pandas as pd
```

### Pandas Data Structures

#### Series

A one-dimensional labeled array capable of holding any data type



## Python For Data Science Cheat Sheet

### Matplotlib

Learn Python interactively at [www.DataCamp.com](http://www.DataCamp.com)

### Matplotlib

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

### 1 Prepare The Data

Also see Lists & NumPy

```
>>> # Import numpy as np
>>> x = np.linspace(0, 10, 10)
>>> y = np.sin(x)
>>> U = -1 + x**2
>>> V = 1 + y - 4*x
```

### 2D Data & Images

```
>>> data = 2 * np.random.random((10, 10))
>>> data2 = 3 * np.random.random((10, 10))
>>> data3 = 4 * np.random.random((10, 10))
>>> U = -1 + data**2
>>> V = 1 + data2 + data3
```

### As

### Ar

### C

### 4 C

### Colours

### 5 DM

### IM

### Markers

### M

### Plots

### Sc

### St

### T

### W

### X

### Y

### Z

### z

## Python For Data Science Cheat Sheet

### Scikit-Learn

Learn Python for data science interactively at [www.DataCamp.com](http://www.DataCamp.com)

### Scikit-learn

Scikit-learn is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.

### A Basic Example

```
>>> from sklearn.model_selection import train_test_split
>>> from sklearn.metrics import accuracy_score
>>> from sklearn.datasets import load_iris
>>> X, y = iris.data[:, :-1], iris.target
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier(n_neighbors=5)
>>> knn.fit(X_train, y_train)
>>> y_pred = knn.predict(X_test)
>>> accuracy_score(y_test, y_pred)
```

## Cred

### Sup

Lines  
>>> f  
>>> l

Supp  
>>> t

Nain  
>>> E

>>> g

KNN  
>>> E

>>> K

Unsi  
>>> S

Princ  
>>> P

KIM  
>>> F

>>> R

Mod  
>>> M

## Python For Data Science Cheat Sheet

### Keras

Learn Python for data science interactively at [www.DataCamp.com](http://www.DataCamp.com)

### Keras

Keras is a powerful and easy-to-use deep learning library for Theano and TensorFlow that provides a high-level neural networks API to develop and evaluate deep learning models.

### A Basic Example

```
>>> import numpy as np
>>> from keras.models import Sequential
>>> from keras.layers import Dense
>>> data = np.random.random((1000, 10))
>>> labels = np.random.randint(0, 2, size=(1000, 1))
>>> model = Sequential()
>>> model.add(Dense(32, activation='relu',
>>>                 input_dim=10))
>>> model.add(Dense(1, activation='sigmoid'))
>>> model.compile(optimizer='rmsprop',
>>>                 loss='binary_crossentropy',
>>>                 metrics=['accuracy'])
>>> model.fit(data, labels, epochs=10, batch_size=32)
```

## Mod

### Sequ

>>> En

>>> ms

>>> mo

>>> Mult

Binary G

>>> fr

mo

>>> mo

## Python For Data Science Cheat Sheet

### Jupyter Notebook

Learn More Python for Data Science at [www.DataCamp.com](http://www.DataCamp.com)

### Saving/Loading Notebooks

Create new notebook

Make a copy of the current notebook

Name a copy...

Save current notebook and record checkpoint

Preview of the printed notebook

Close notebook & stop running any scripts

Close and save...

Trusted Notebooks

Share and edit

### Open an existing notebook

Rename notebook

Revert notebook to a previous checkpoint

Download notebook as...

- Python notebook

- HTML

- Markdown

- LaTeX

- PDF

### Writing Code And Text

## Python For Data Science Cheat Sheet

### Pandas Basics

Learn Python for Data Science interactively at [www.DataCamp.com](http://www.DataCamp.com)

### Pandas

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.

pandas

Use the following import convention:

```
>>> import pandas as pd
```

### Pandas Data Structures

#### Series

A one-dimensional labeled array capable of holding any data type



# Outline : Data Exploration Practical: MovieLens

Applications

Meaning of the Terms

Prerequisites

Data Exploration Practical:  
MovieLens

Overview

## Data Exploration: MovieLens

---

A simple study of a movie review dataset

- 3 million ratings
- Social descriptors: age, gender, ...
- First basic approach to data mining

movielens

# Introduction to pandas

Pandas



- Python library for data manipulation:  
<https://pandas.pydata.org/>
- Enables performing operations and visualizations
- Easy to use

# Outline : Overview

---

Applications

Meaning of the Terms

Prerequisites

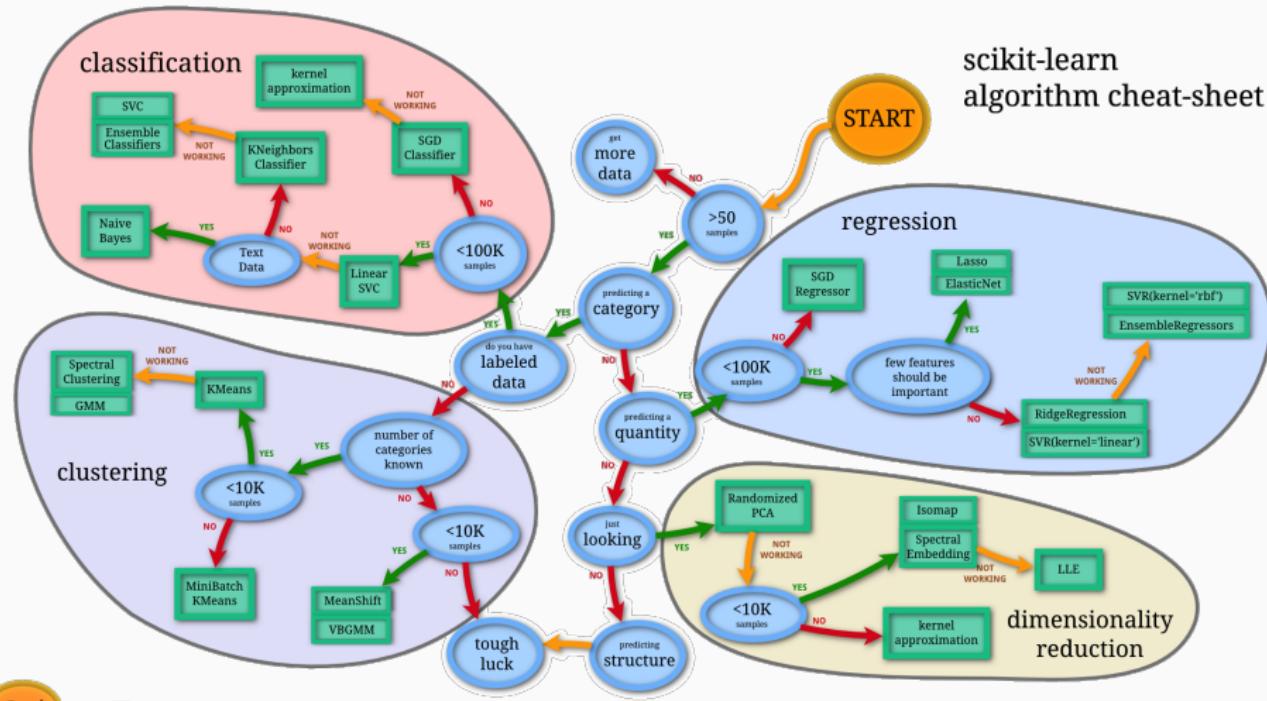
Data Exploration Practical:  
MovieLens  
**Overview**

# The Different Methods

---

- Classification: predict a given class
- Regression: predict a value
- Clustering: group elements into clusters (undetermined classes)
- Dimensionality reduction: reduce the data representation space
- Anomaly detection

## The Different Methods



The scikit-learn logo consists of the word "scikit" above the word "learn", where the "e" in "learn" is stylized as a lowercase "l". Both words are written in a white, lowercase, sans-serif font.

## Examples: Classification

---

- Speech emotion recognition:
- Animal species classification from images:
- Object detection in medical images:

## Examples: Classification

---

- Speech emotion recognition: **is the person angry or happy?**
- Animal species classification from images:
- Object detection in medical images:

## Examples: Classification

---

- Speech emotion recognition: **is the person angry or happy?**
- Animal species classification from images: **is it a cat or a puma?**
- Object detection in medical images:

## Examples: Classification

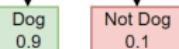
---

- Speech emotion recognition: **is the person angry or happy?**
- Animal species classification from images: **is it a cat or a puma?**
- Object detection in medical images: **is it a tumor?**

# Examples: Classification

- Speech emotion recognition: **is the person angry or happy?**
- Animal species classification from images: **is it a cat or a puma?**
- Object detection in medical images: **is it a tumor?**

Binary Classification



Multiclass Classification



Multilabel Classification



## Examples: Regression

---

- Speech emotion recognition:
- Damage assessment from post-earthquake images:
- Detection of Alzheimer's severity in voice:

## Examples: Regression

---

- Speech emotion recognition: **what is the intensity of the anger?**
- Damage assessment from post-earthquake images:
- Detection of Alzheimer's severity in voice:

## Examples: Regression

---

- Speech emotion recognition: **what is the intensity of the anger?**
- Damage assessment from post-earthquake images: **can the ambulance use the bridge?**
- Detection of Alzheimer's severity in voice:

## Examples: Regression

---

- Speech emotion recognition: **what is the intensity of the anger?**
- Damage assessment from post-earthquake images: **can the ambulance use the bridge?**
- Detection of Alzheimer's severity in voice: **how advanced is the condition?**

## Examples: Regression

- Speech emotion recognition: **what is the intensity of the anger?**
- Damage assessment from post-earthquake images: **can the ambulance use the bridge?**
- Detection of Alzheimer's severity in voice: **how advanced is the condition?**

### Age Prediction via Regression



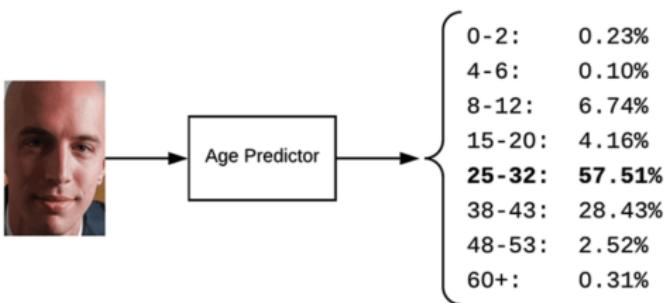
# Examples: Regression

- Speech emotion recognition: **what is the intensity of the anger?**
- Damage assessment from post-earthquake images: **can the ambulance use the bridge?**
- Detection of Alzheimer's severity in voice: **how advanced is the condition?**

## Age Prediction via Regression



## Age Prediction via Classification



## Examples: Clustering

- Topic mining in political forums:
- Misinformation classification on social networks:
- Customer segmentation:

## Examples: Clustering

- Topic mining in political forums: **what are citizens concerned about?**
- Misinformation classification on social networks:
- Customer segmentation:

## Examples: Clustering

- Topic mining in political forums: **what are citizens concerned about?**
- Misinformation classification on social networks: **we group these news items that seem unusual**
- Customer segmentation:

## Examples: Clustering

- Topic mining in political forums: **what are citizens concerned about?**
- Misinformation classification on social networks: **we group these news items that seem unusual**
- Customer segmentation: **people who like beers**

## Examples: Clustering

- Topic mining in political forums: **what are citizens concerned about?**
- Misinformation classification on social networks: **we group these news items that seem unusual**
- Customer segmentation: **people who like beers**



**Questions?**

## References i