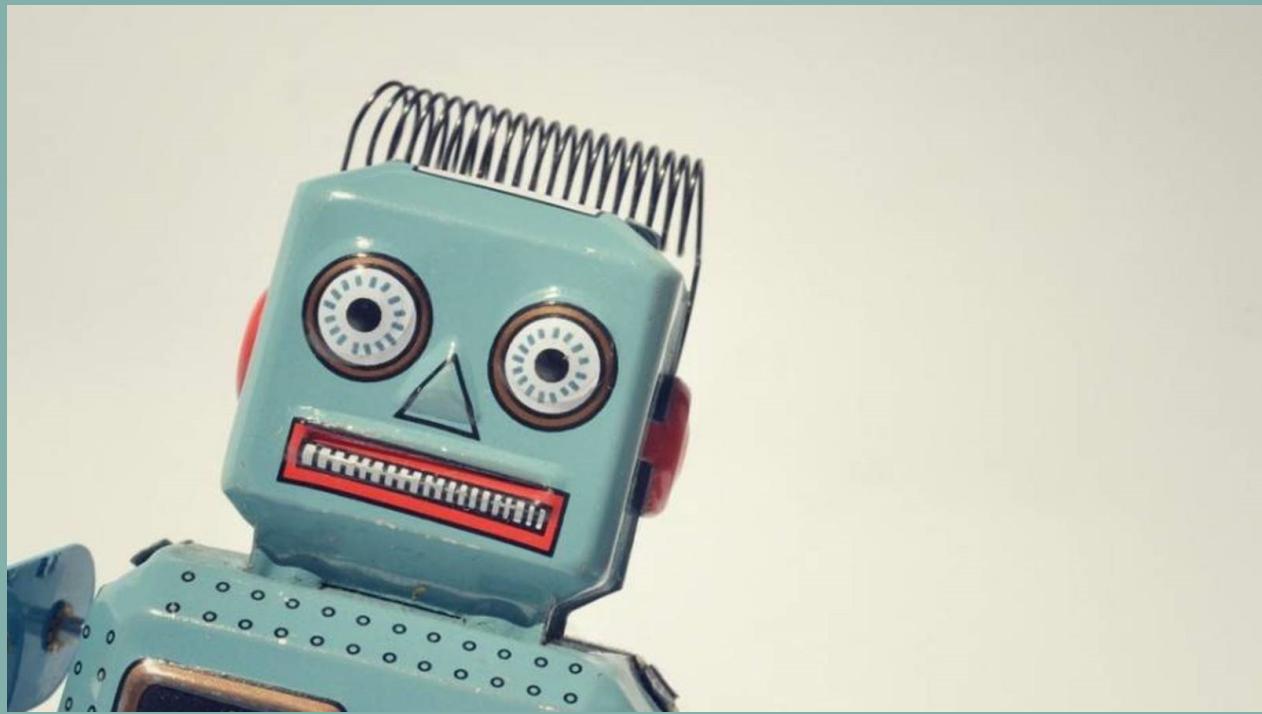


DIGITAL METHODS FOR ANALYSING TEXTS

//

01_Introduction to text mining

Ana Valdivia
Research Associate
King's College London



WHO AM I?//



Ana Valdivia

Research Associate at Security Flows (King's College London, UK)

PhD in Natural Language Processing (Universidad de Granada, Spain)

Research interests:

- *Fairness, Accountability, Transparency in AI*
- *Natural Language Processing*
- *Data Feminism*
- *Design Justice*
- *Hype in AI and Big Data*



WHO ARE YOU?//

Who are you?

What do you expect from this course?



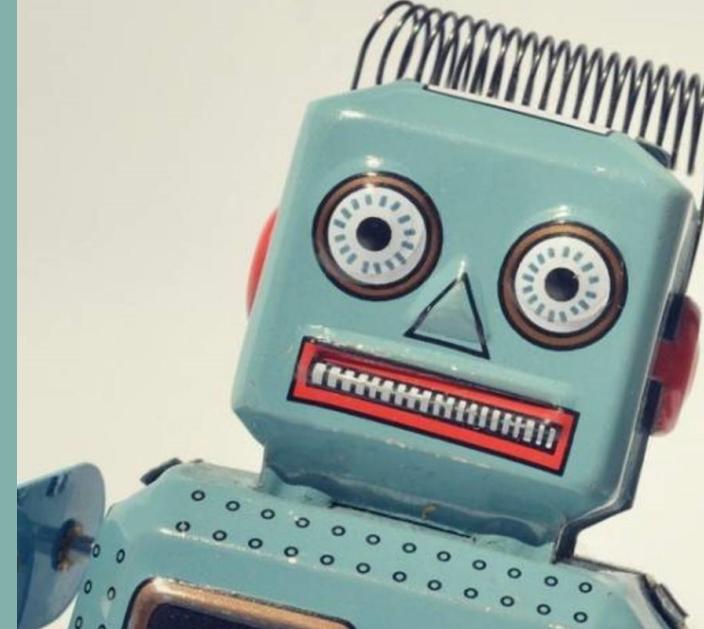
ABOUT THIS COURSE//

- Familiarises PhD students with the main **text mining** techniques in social science and develops basic skills in **digital methods**.
- You will learn how to approach and **manage text data**, **analyse texts**, and **visualize** this information.
- Every **Monday** and **Wednesday** from 10AM to 1PM.

Session date	Lecture	Lecture topic	Seminar topic
12 April	1	Introduction to text mining	Import text data
14 April	2	Analysing text	Methods for text preprocessing
19 April	3	Analysing words	Methods for word analysis
21 April	4	Topic modelling	Methods for analysing topics
26 April	5	Ethics NLP + live coding	Biases and real example analysis
28 April	6	Text mining in the real world	Analysing your own text

INTRODUCTION

//



INTRODUCTION//

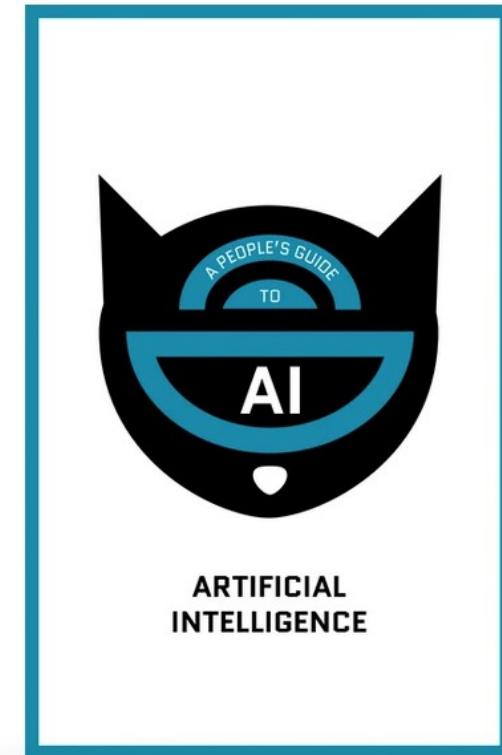


What is AI?

When you hear the words “**Artificial Intelligence**”, what are the first three things that come to your mind?

1. _____
2. _____
3. _____

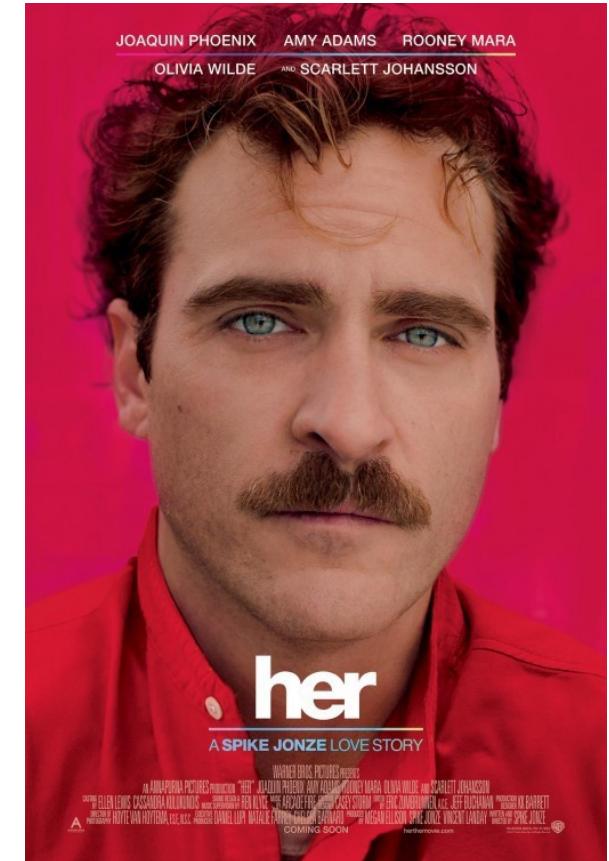
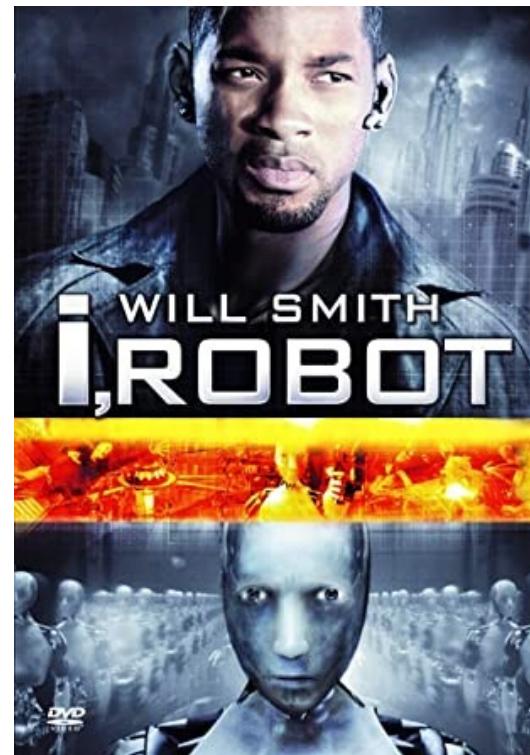
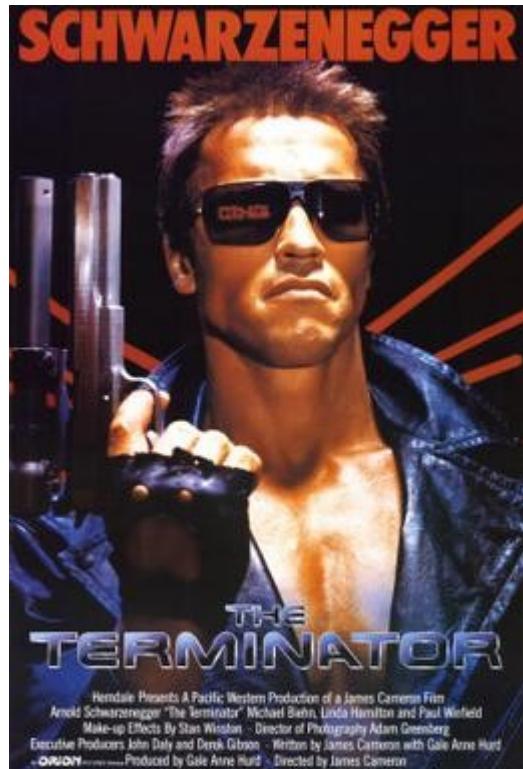
Write here



INTRODUCTION//



What is AI?



INTRODUCTION//



What is AI?

We need to **demystify** Artificial Intelligence.

INTRODUCTION//



What is AI?

We need to **demystify** Artificial Intelligence.

INTRODUCTION//



What is AI?

What is intelligence?

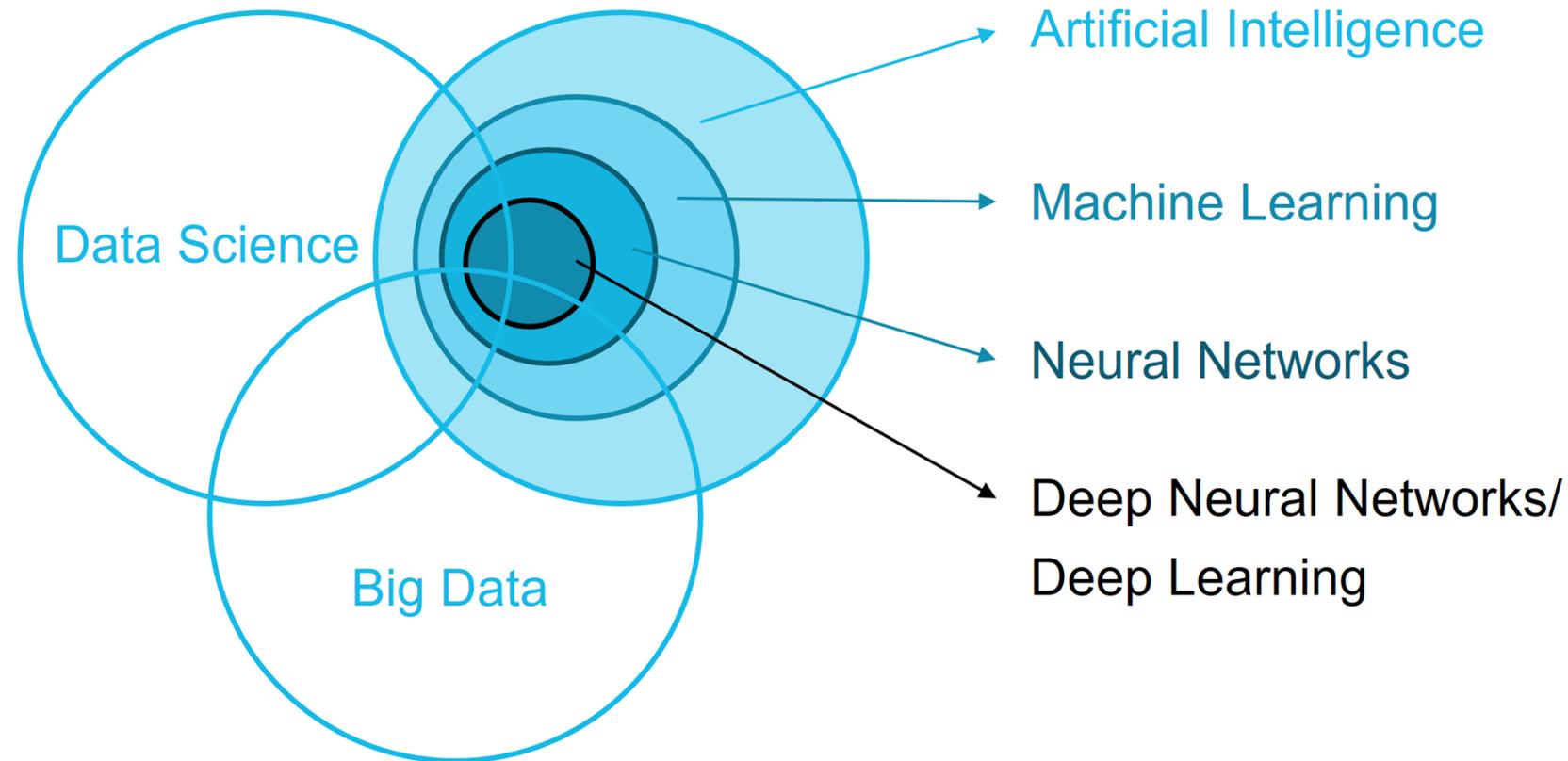
“Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by *perceiving their environment* through **data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or **processing the information**, derived from this data and **deciding the best action(s)** to take to achieve the given goal.”**

```
3 require File.expand_path('../config/environment', __FILE__)
4 # Prevent database truncation if the database needs cleaning.
5 abort("The Rails environment is running in production mode!") if Rails.env.production?
6 require 'spec_helper'
7 require 'rspec/rails'
8
9 require 'capybara/rspec'
10 require 'capybara/rails'
11
12 Capybara.javascript_driver = :webkit
13 Category.delete_all
14 Shoulda::Matchers.configure do |config|
15   config.integrate do |with|
16     with.test_framework :rspec
17     with.library :rails
18   end
19 end
20
21 # Add additional requires below this line to include more spec helpers
22
23 # Requires supporting files with custom matchers and helper methods
24 # in spec/support/ and its subdirectories
25 # run as spec files by default
26 # in _spec.rb will need to be specified
27 # run twice. It is recommended
28 # end with _spec.rb
29
30 # -- --
31
32 # No results found for 'rspec/gf'
```

European Comission.

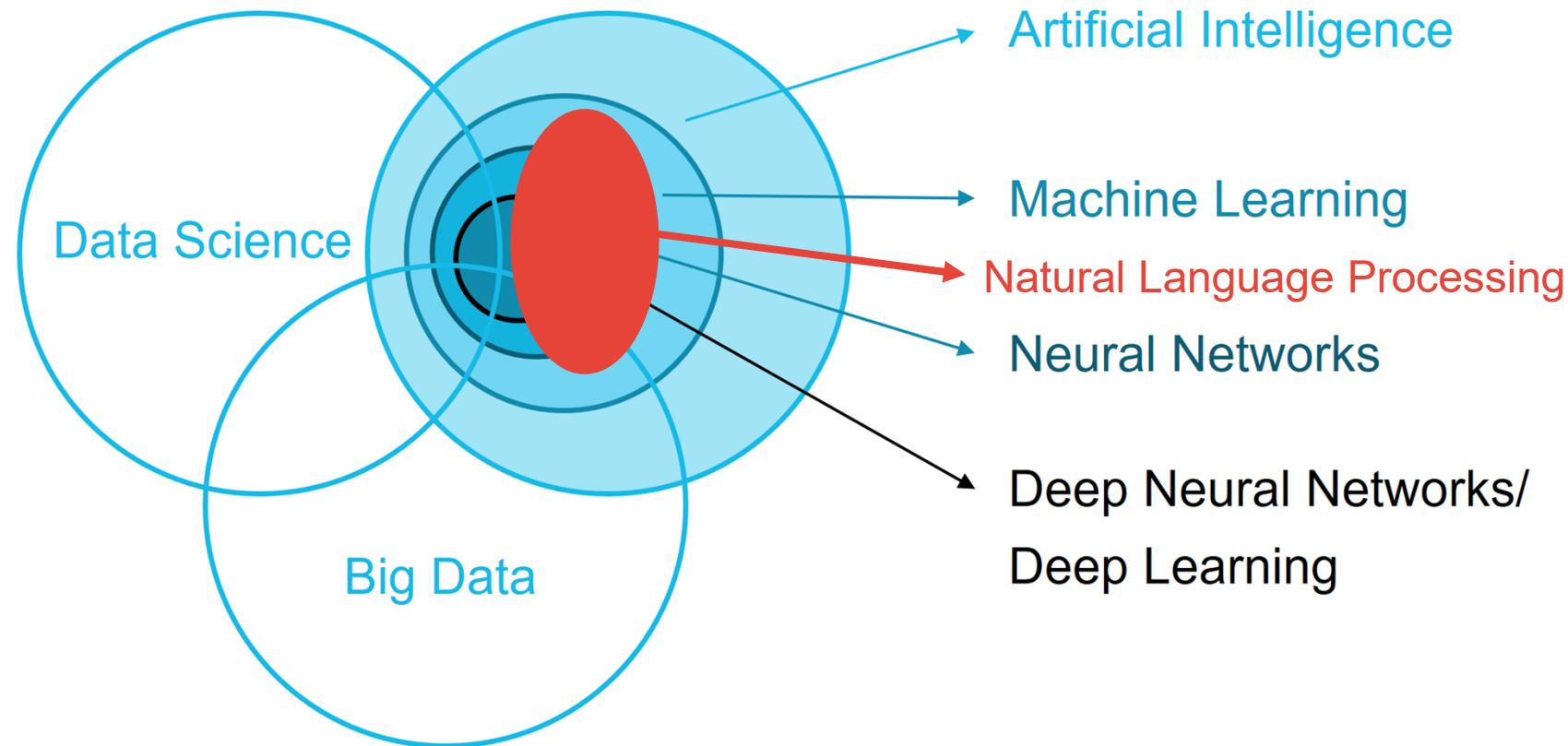
INTRODUCTION//

What is AI?



INTRODUCTION//

What is AI?



INTRODUCTION//



What is Natural Language Processing?

Digital Methods to Analyse Text = Natural Language Processing

INTRODUCTION//



What is Natural Language Processing?

“Human language appears to be unique phenomenon, without significant analogue in the animal world.”

Noam Chomsky.





INTRODUCTION//

What is Natural Language Processing?

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that is focused on enabling computers to understand, process and generate human languages, to get computers closer to a human-level understanding of language.

INTRODUCTION//



What is Natural Language Processing?



Data



Computation
performance



Natural Language
Processing

INTRODUCTION//



Could you list three NLP-based applications?

1. _____

2. _____

3. _____

Write here

INTRODUCTION//

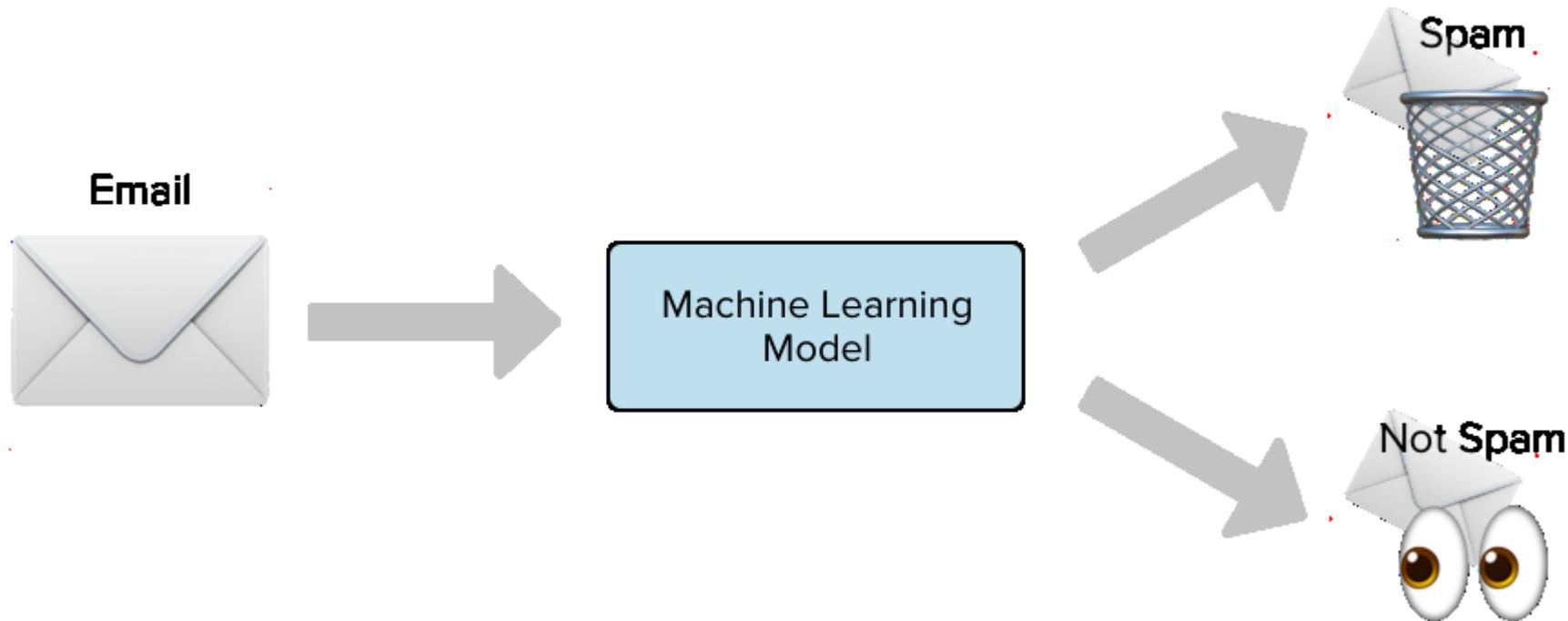


How do you think spam detection works?

INTRODUCTION//



How do you think spam detection works?





INTRODUCTION//

What is Natural Language Processing?

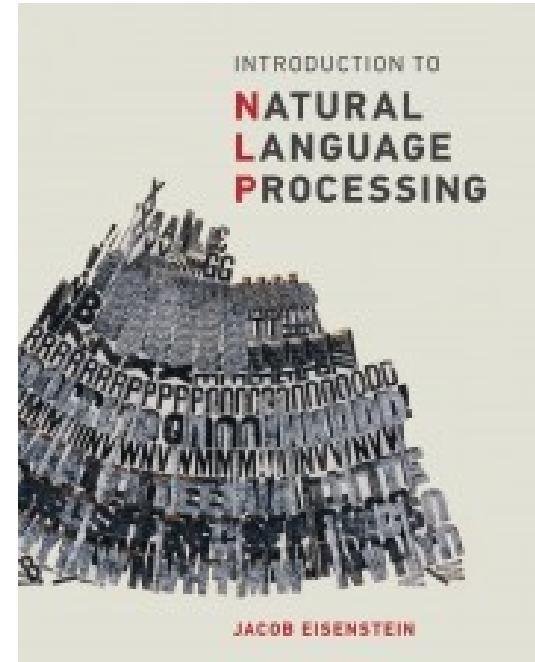
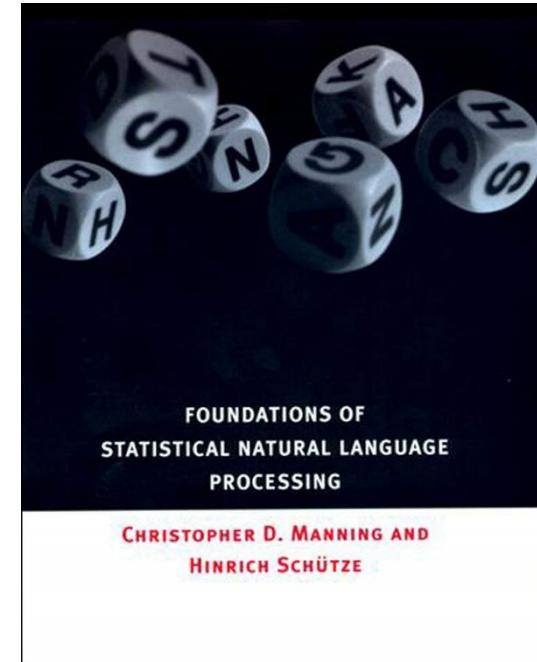
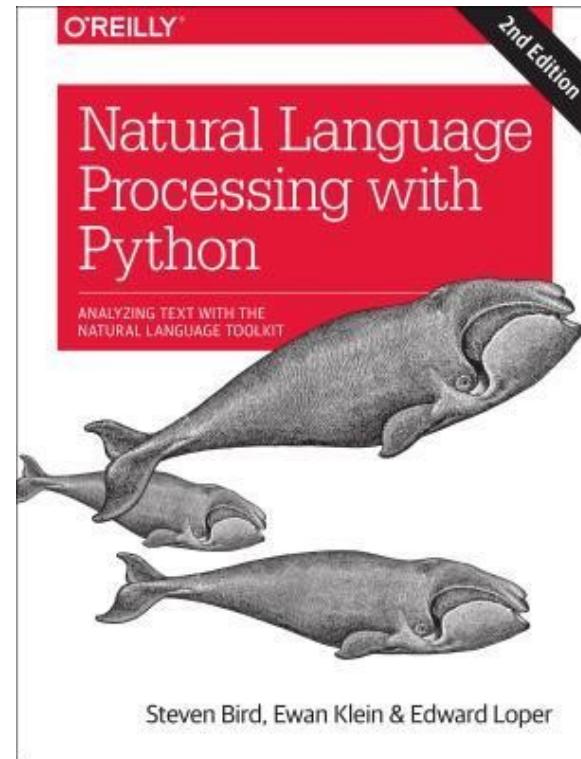
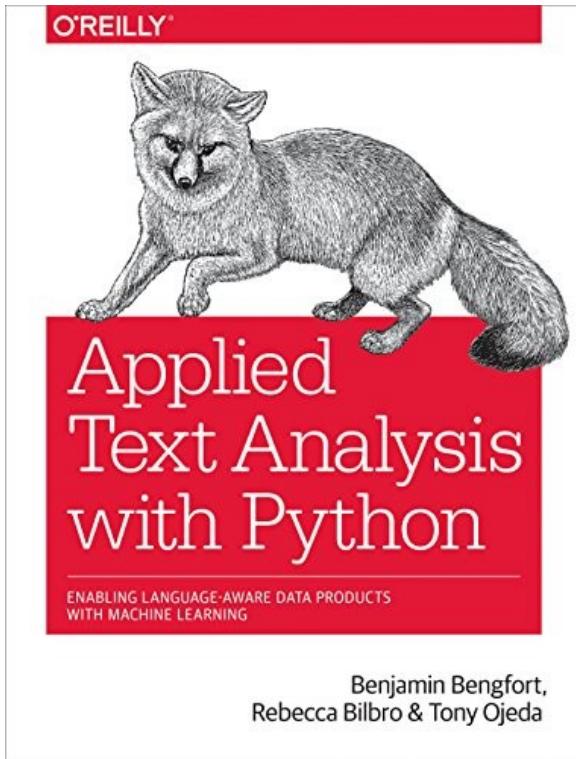
We will learn **basic skills** to deal with text in different layers:

- Text,
- words,
- relations, and
- topics.

INTRODUCTION//

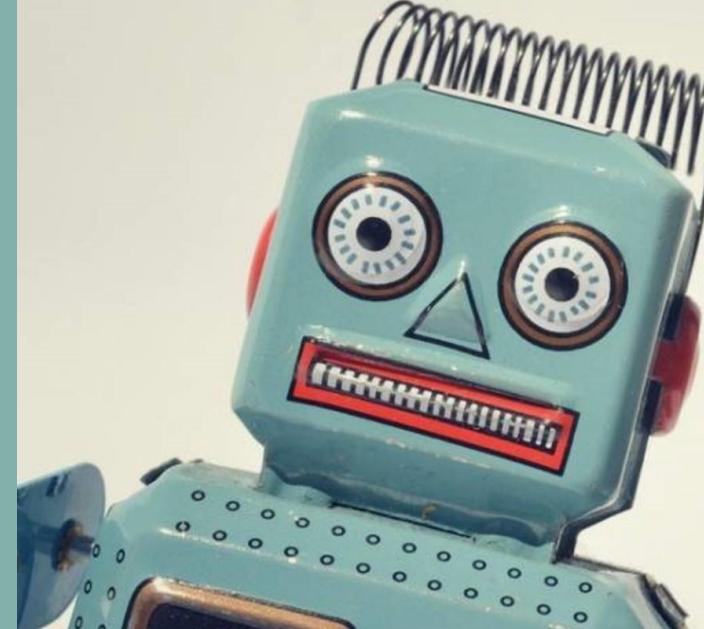


Books



HISTORY OF NLP

//

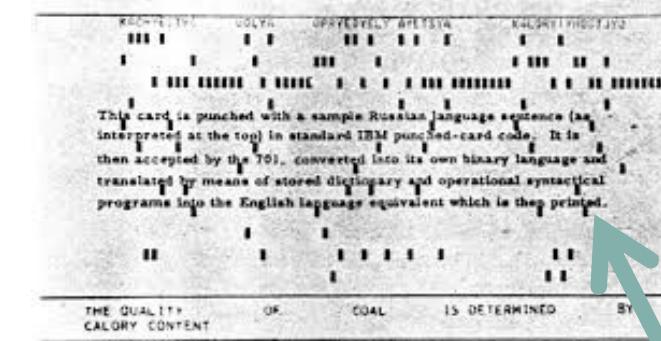


HISTORY OF NLP//

The history of NLP (and AI) is military history:

- Computing Machinery and Intelligence by Alan Turing (1950).
- The Georgetown-IBM System, a machine translation system (1954).
- PARRY, a chatbot to simulate a person with paranoid schizophrenia (1972).

Example of a punched card with a Russian sentence and English translation (Georgetown-IBM System)



To read more about the Georgetown-IBM System, click [here](#).

HISTORY OF NLP//



It has rapidly evolved thanks to gamers.



A **graphics processing unit (GPU)** is a specialized electronic circuit.



HISTORY OF NLP//

Where are we now?

- **Watson** (IBM), question-answering computer system (**2006**).
- **Virtual assistants**, Siri, Amazon Alexa, Cortana, Google Assistant (**2011-2016**).
- **GPT-3** (OpenAI), autoregressive language model that uses deep learning to produce human-like text (**2020**).

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



See this [video](#).

HISTORY OF NLP//



Are you impressed by the **performance** of this chatbot?

Have you had good **experience** when **interacting** with chatbots?

HISTORY OF NLP//



NLP (and AI) is far from being
general intelligent.

HISTORY OF NLP//

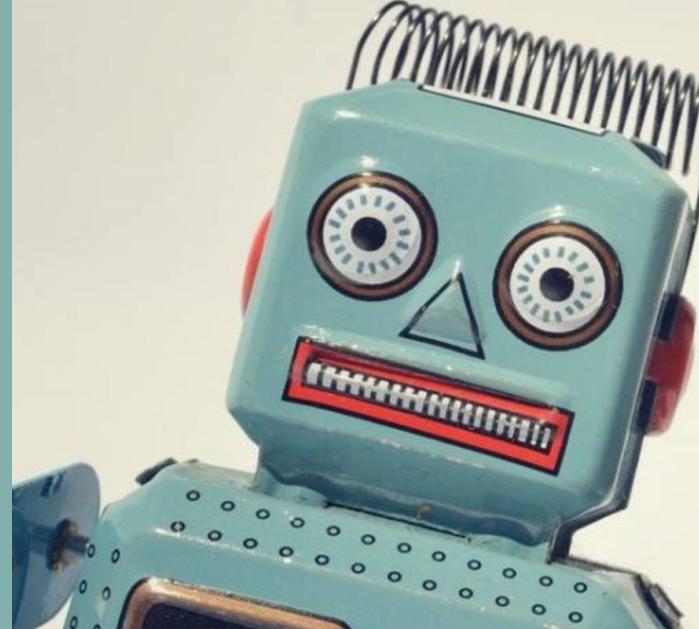


However, it is good at analysing large amount of data and identifying patterns on it.

Which is super difficult for us, the human beings.

LANGUAGE AS DATA

//

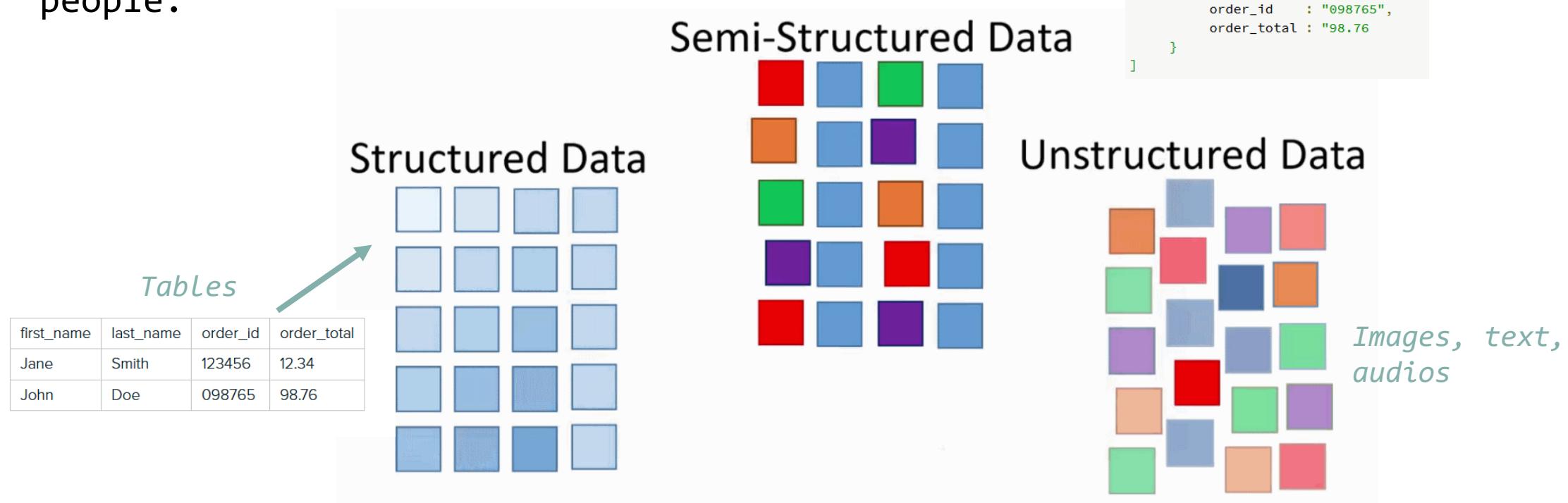


LANGUAGE AS DATA//



Unstructured data

Language is **unstructured data** that has been produced by people to be understood by other people.



LANGUAGE AS DATA//



Unstructured data

Language is **unstructured data** that has been produced by people to be understood by other people.

Myanmar / Aung San Suu Kyi
faces four charges as junta cracks down on dissent

Pay gap / Women earning two-thirds less than men in top finance roles

Labour / Alleged leakers of antisemitism report should not be named, rules judge



Title	Keywords	Date	Text
Aung San Suu...	Myanmar	01/03/2021	...
Women earning two-thirds...	Pay gap	01/03/2021	...
Alleged leakers of...	Labour	01/03/2021	...

LANGUAGE AS DATA//



Which other columns we could include?

Title	Keywords	Date	Text
Aung San Suu...	Myanmar	01/03/2021	...
Women earning two-thirds...	Pay gap	01/03/2021	...
Alleged leakers of...	Labour	01/03/2021	...

LANGUAGE AS DATA//



Building a Custom Corpus

Corpora are collections of related documents that contain natural language.

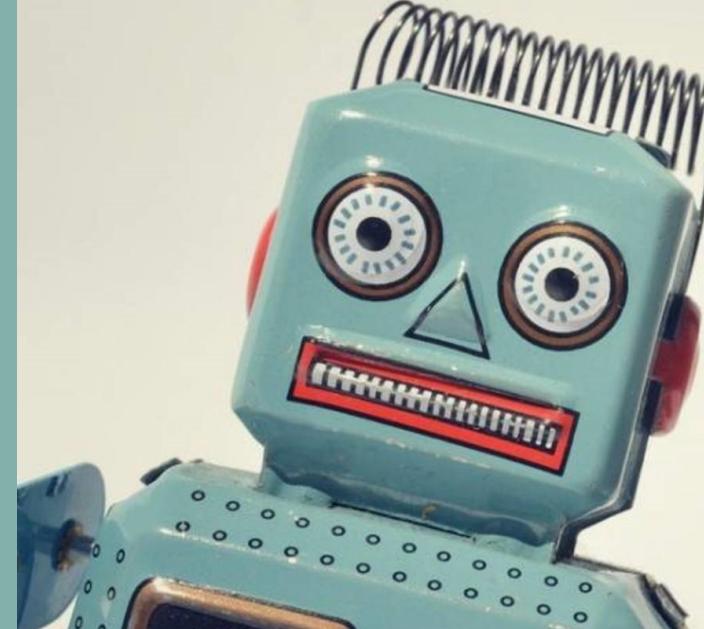
A **corpus** can be large or small, though generally they consist of dozens or even hundreds of gigabytes of data inside of thousands of documents.

- **Annotated:** Supervised learning algorithms.
- **Unannotated:** Unsupervised learning algorithms.



JUPYTER NOTEBOOKS

//



JUPYTER NOTEBOOKS//



Python is a **programming language** that lets you work quickly and integrate systems more effectively...

It contains relevant **NLP libraries** to analyse text (NLTK, Spacy and Gensim)

JUPYTER NOTEBOOKS//

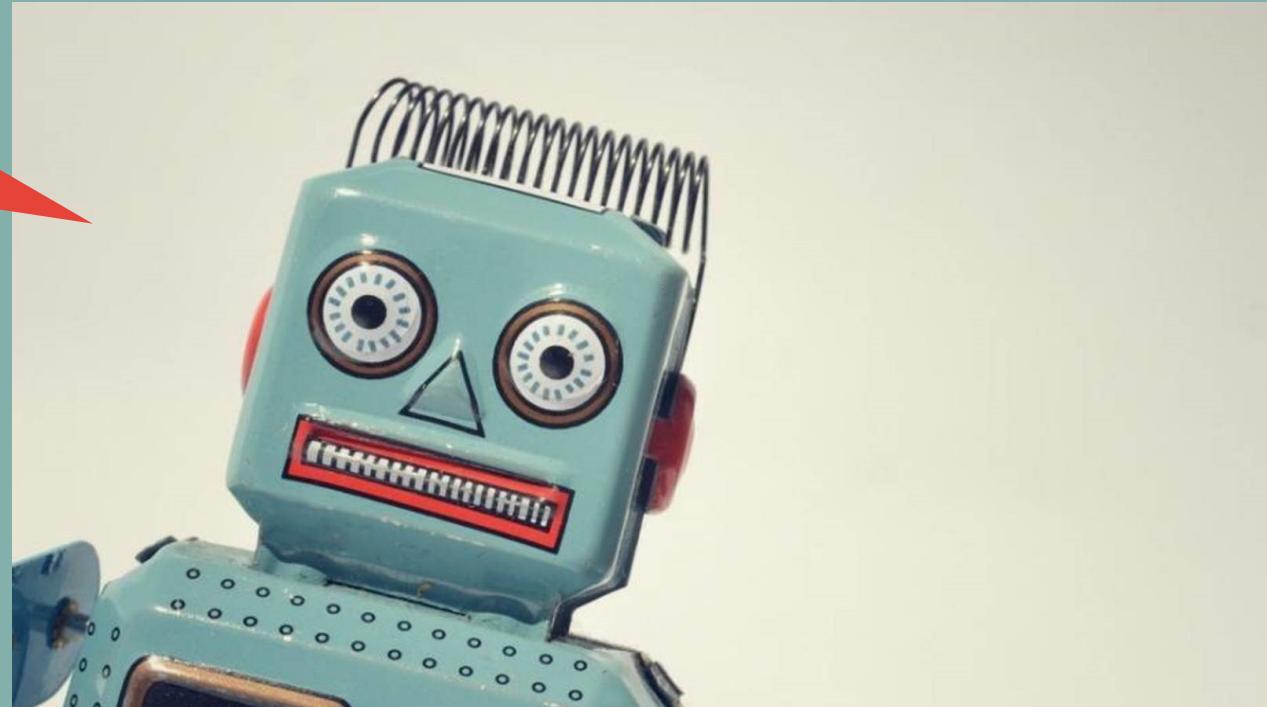


The Jupyter Notebook is a web application that allows to create documents that contain live code cells, visualizations, explanatory text and equations:

- Interactive
- Support multiple languages (Python)
- Extensible
- And it's open source!



LET'S CODE!



**WE'LL BE BACK IN 15
MIN...**

