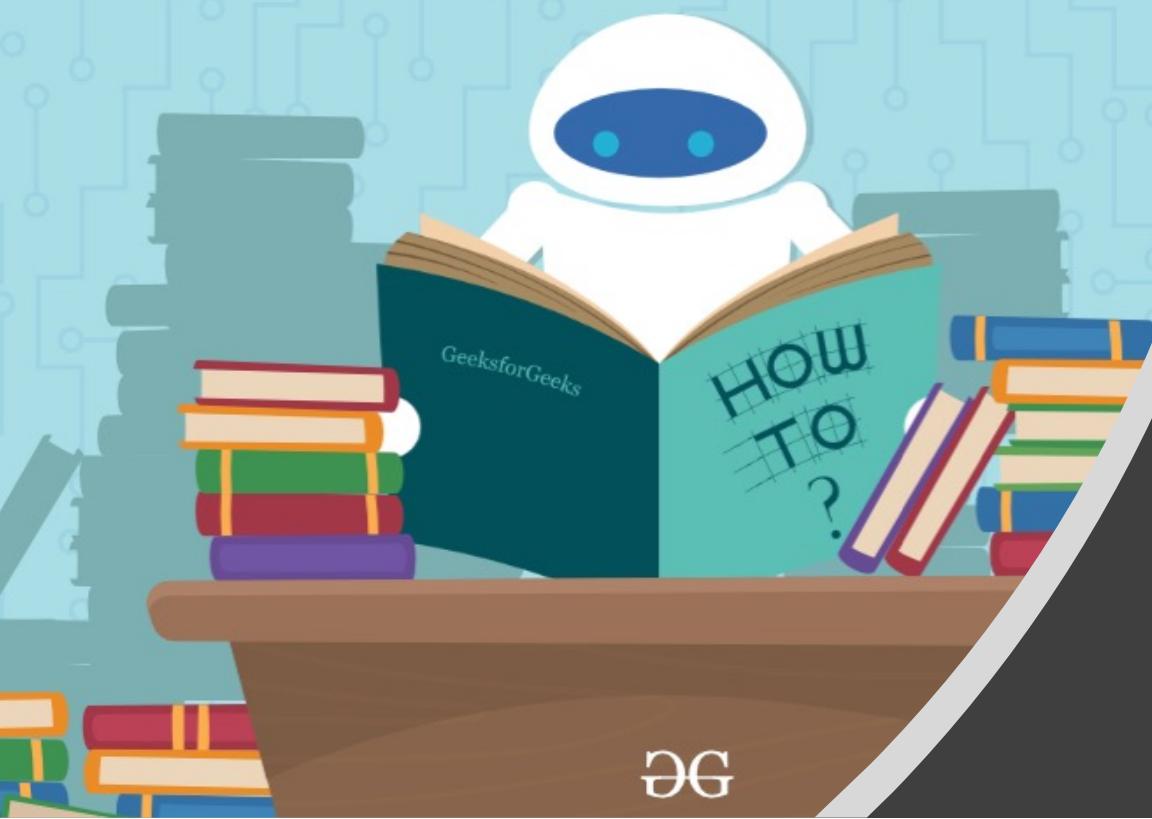


MACHINE LEARNING



Machine Learning for Economics

Marina Rizzi

28 June 2023

"Labor Economics" Course by Francesco
Devicienti

Overview of this lecture:

- What it is Machine Learning (and some common models used there)
- How do Economists use Machine Learning?
- Let's dive into some coding!

Aim of this lecture:

- Give an introductory view of what Machine learning is, and introduce to the main methods and terminology used
- Give examples of where Machine learning could be used in Economics
- Give a glimpse of the implementation of some codes and algorithms in Python

Pipeline:

-What is Machine Learning

-Machine Learning in Economics (with examples of papers)

-AI Instruments for Researchers

1) What is Machine Learning

Let's start with a question: what is Artificial Intelligence, in your opinion?

A Ansa

L'intelligenza artificiale per decifrare una lingua sconosciuta

Utilizzare l'intelligenza artificiale per aiutare a decifrare una lingua antica che ancora oggi resta sconosciuta.

23 ore fa



greenMe

L'intelligenza artificiale ci sta davvero aiutando a decifrare una lingua di 3.500 anni fa

Un gruppo di ricerca dell'Università di Bologna ha usato l'intelligenza artificiale per decifrare una lingua sconosciuta, il cipro-minoico.

2 ore fa



The Washington Post

Robots trained on AI exhibited racist and sexist behavior

Those virtual robots, which were programmed with a popular artificial intelligence algorithm, were sorting through billions of images and...

1 ora fa



India Education Diary

Massachusetts Institute of Technology: Artificial intelligence model finds potential drug molecules a thousand times faster

Massachusetts Institute of Technology: Artificial intelligence model finds potential drug molecules a thousand times faster.

4 ore fa

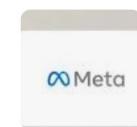


Tgcom24

Fake news, Zuckerberg arruola l'Intelligenza Artificiale anti-bufale

Meta, la società madre di Facebook e Instagram di Mark Zuckerberg, sta testando Sphere, uno strumento basato sull'Intelligenza Artificiale...

23 ore fa



F Forbes

Artificial Intelligence: Not A Panacea For Supply Chain Issues, But Extremely Helpful

Artificial intelligence is another area of technology investment that holds potential, and early results are promising.

1 giorno fa



City AM

From translations to chatbots: The future is knocking as UK firms rush to implement artificial intelligence

Many businesses are opting to take a low- or no-code approach to implementing artificial intelligence solutions into their operations, according to new research...

2 giorni fa

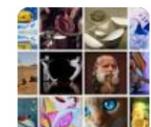


TechCrunch

Perceptron: AI that solves math problems, translates 200 languages, and draws kangaroos

Research in the field of machine learning and AI, now a key technology in ... artificial intelligence — and explain why they matter.

21 ore fa



Artificial Intelligence: (many definitions):

- ***“It is the study of how to train the computers so that computers can do things which at present human can do better.”***

-----> we want to add all the capabilities to a machine, that the human intelligence contains.

Machine Learning: often defined as a subfield of Artificial Intelligence (application of AI).

Artificial Intelligence



Any technique that enables computers to mimic human intelligence. It includes *machine learning*

Machine Learning



A subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes *deep learning*

Deep Learning



A subset of machine learning based on neural networks that permit a machine to train itself to perform a task.

What is Machine Learning?

- Often categorized as a subfield of Artificial Intelligence
- Concretely: building *mathematical model* to help understand data.
- "Learning": when we give to these models ***tunable parameters*** that can be adapted to observed data.
-----> Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.
- Learning without being explicitly coded (but, taking data as input, it is able to generate decisions or prediction as outputs).

More technical definition of Machine Learning:

“Machine Learning is said to learn from experience E w.r.t some class of task T and a performance measure P if learners performance at the task T as measured by P improves with experience E.”

Applications developed thanks to Machine Learning:

Virtual Personal Assistants, Product Recommendations, Self Driving Cars, Translators Softwares, Fraud Detection, Preventive Health Care

- Recent surge in the popularity of Machine Learning in a different variety of fields:
 - huge availability of data
 - more powerful computers

Machine Learning was important to solve a variety of problems that were difficult to be solved with "hard coding" (i.e. describing in a detailed way the action that the algorithm should have performed ----> Machine learning algorithms do not need a precise definition of what to do (i.e. how a chair looks like) but learn from the data the actions or the classifications they should perform)

- **Categories of Machine Learning Algorithms:**

- Supervised VS Unsupervised Algorithms (and also... Semi-supervised)
(one of the most important division)

- Supervised: classification or regression

- Unsupervised: clustering, dimensionality reduction,

- Reinforcement Learning

Supervised VS Unsupervised Machine Learning

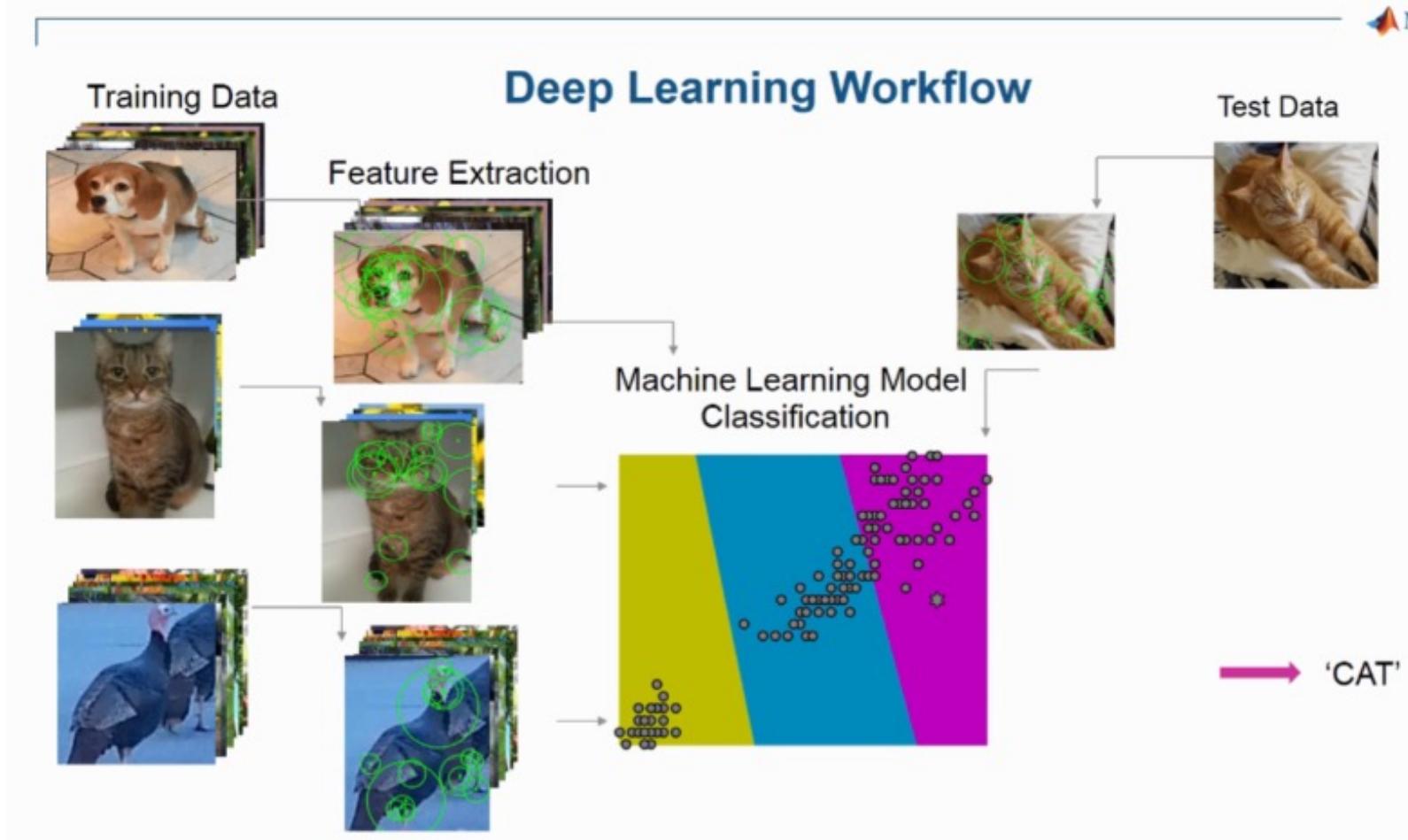


Supervised Machine Learning

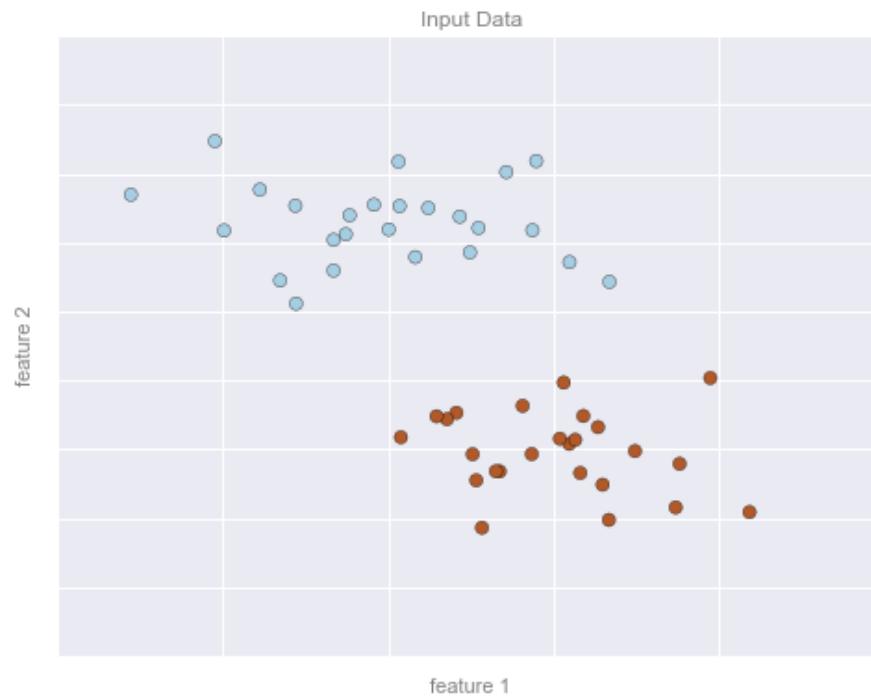
- Modeling the relationship between features of the data and some associated labels
- Trained model ----> usable to apply labels to new, unknown data
- **Classification** (discrete labels) and **regression** (continuous labels)

Formally:

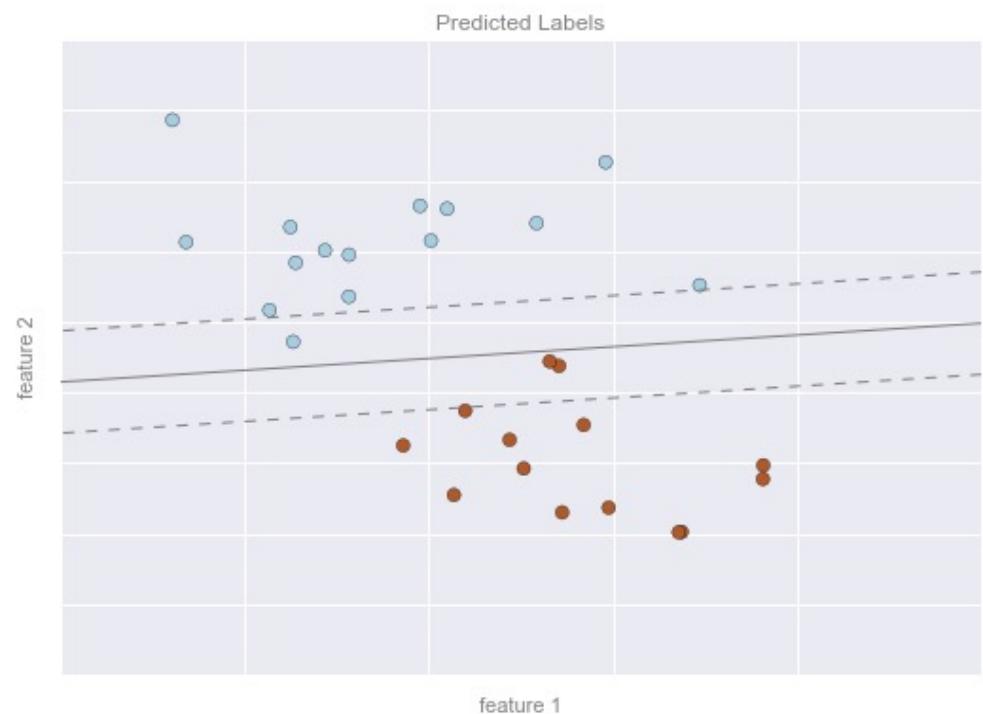
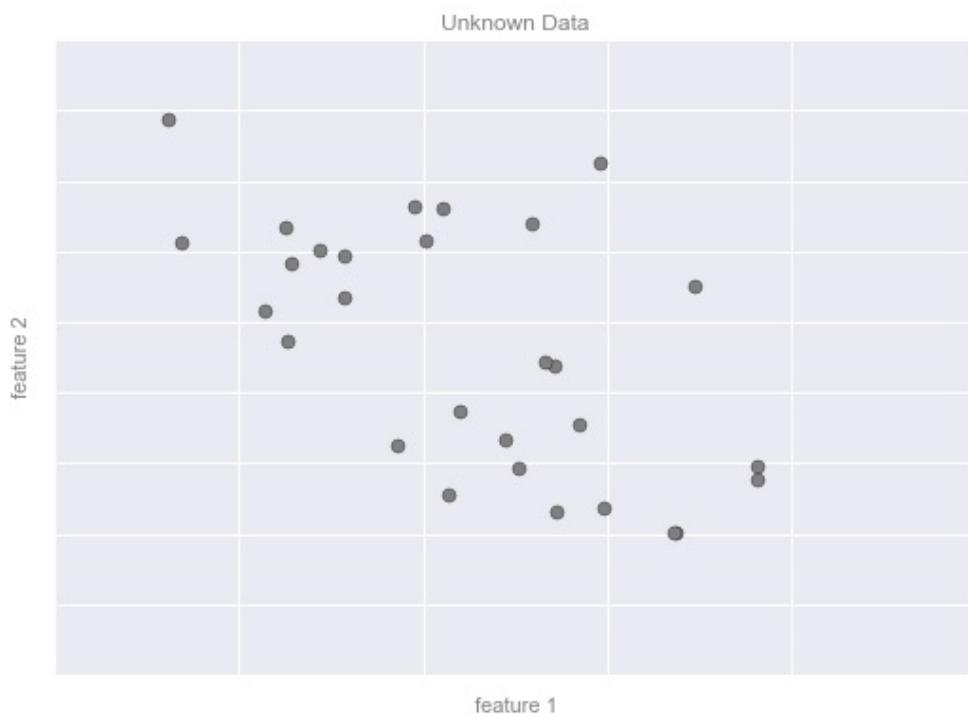
- Features set X ; Label set Y
- Unknown target function $f : X \rightarrow Y$
- Learning algorithm A uses the training set to select a function that approximates f (usually, there is a loss function that has to be minimized)



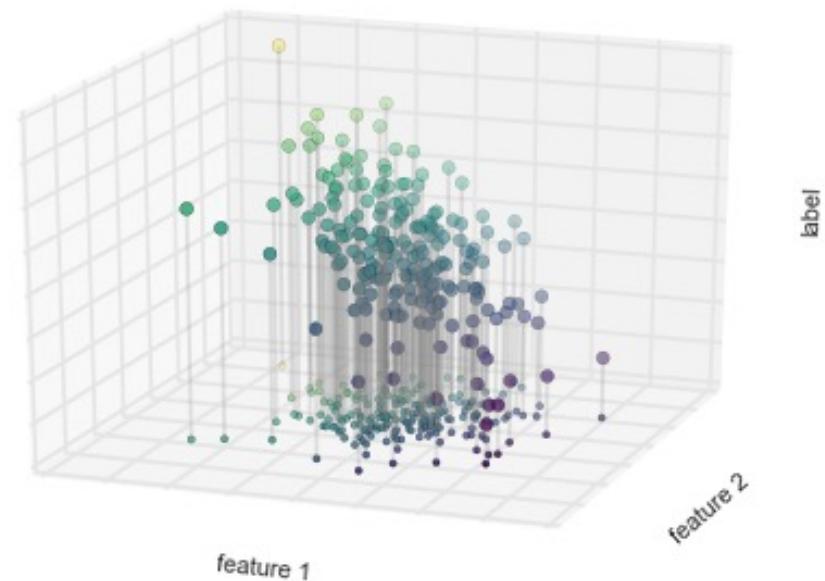
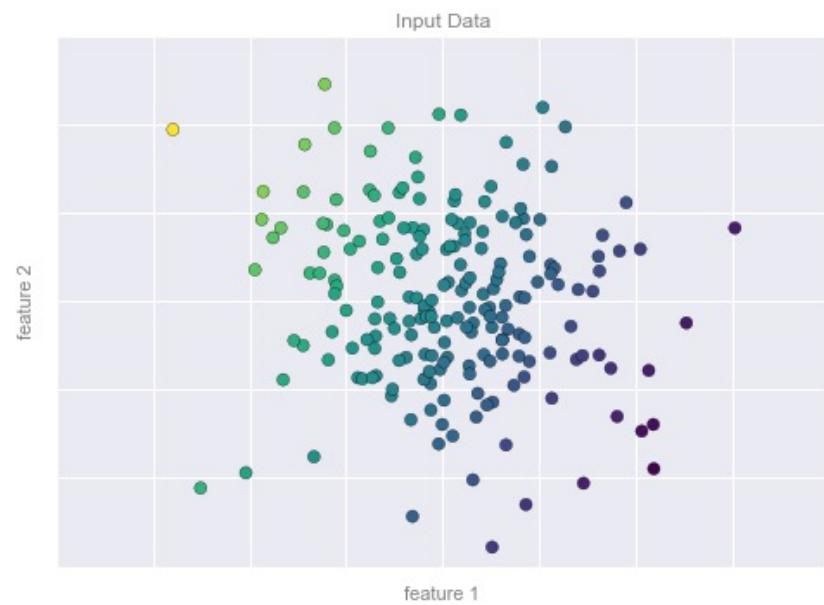
Classification



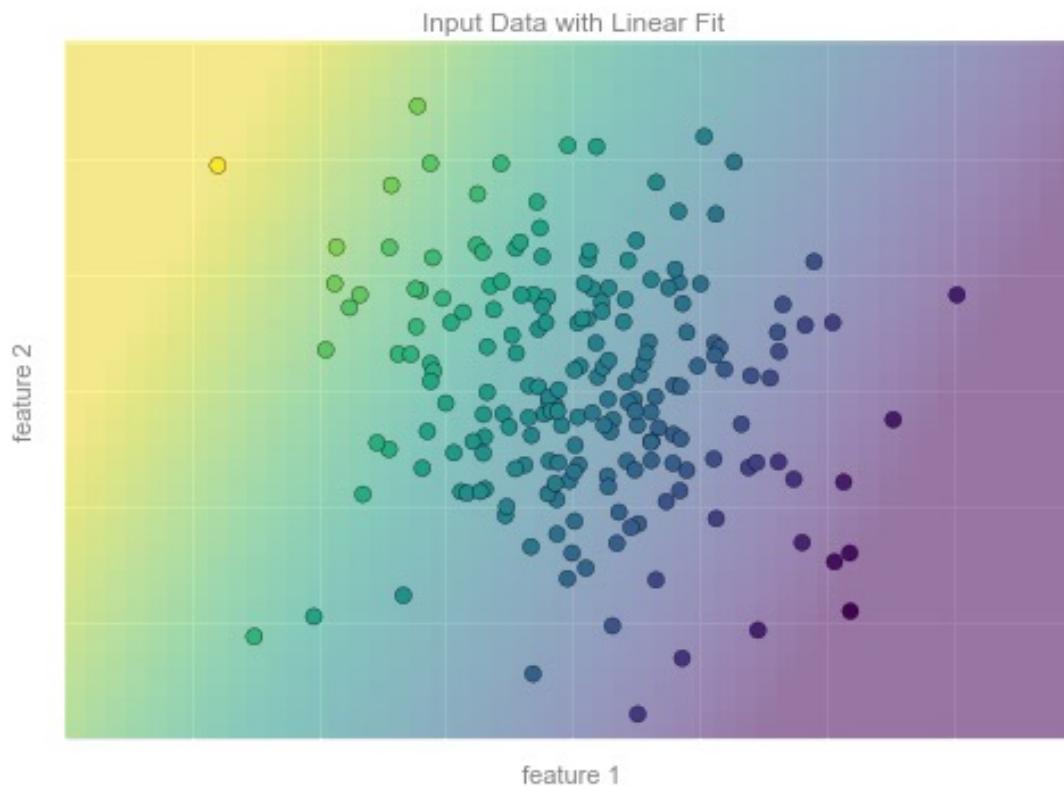
Classification



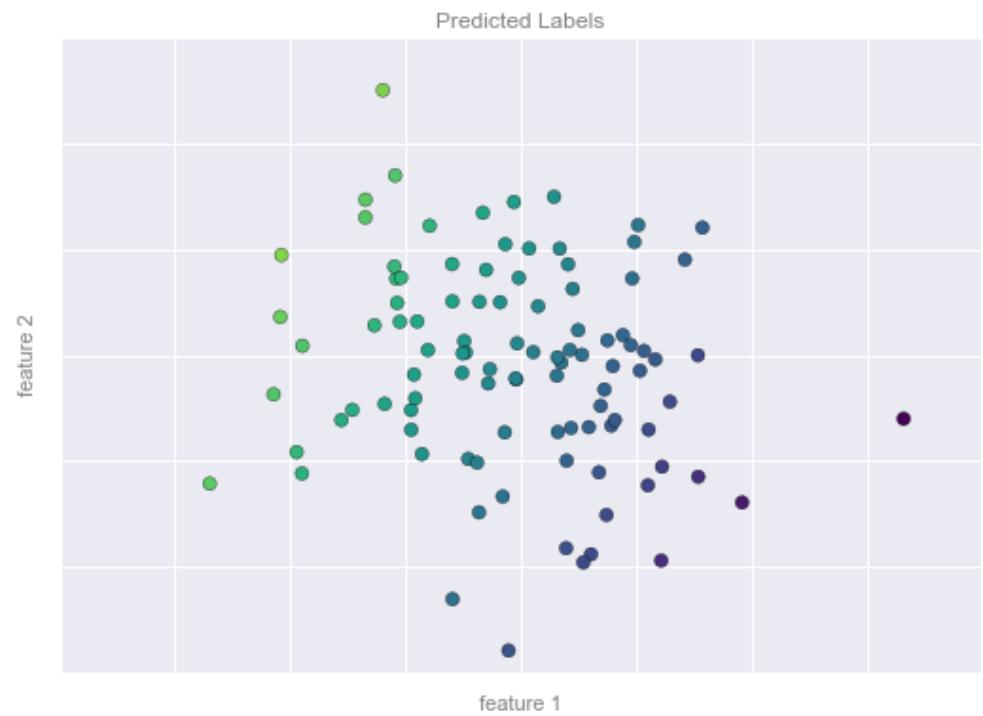
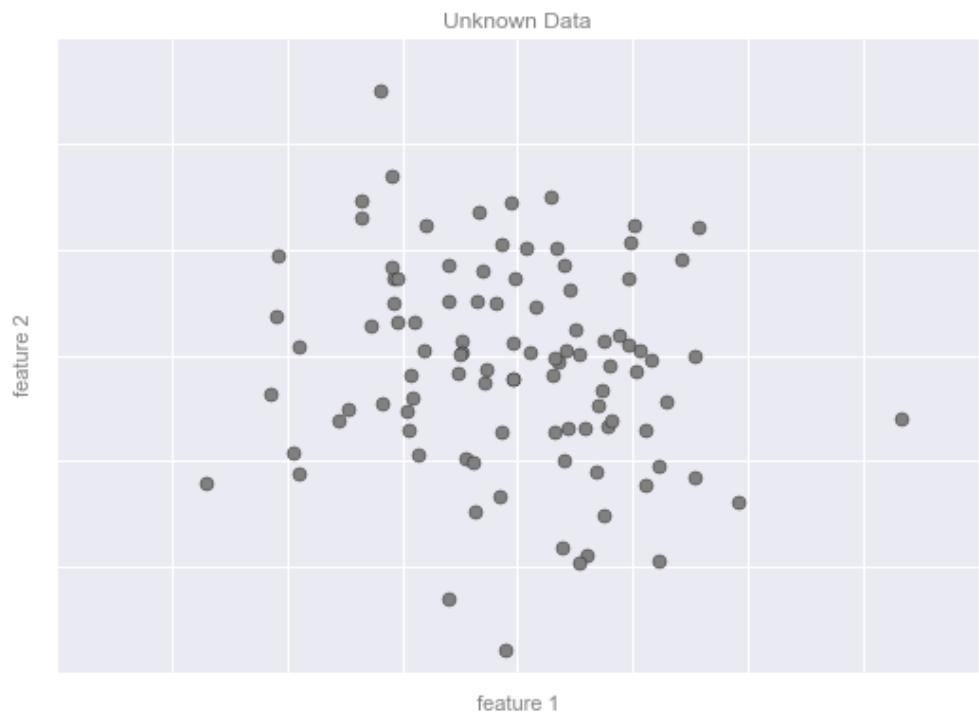
Regression



Regression

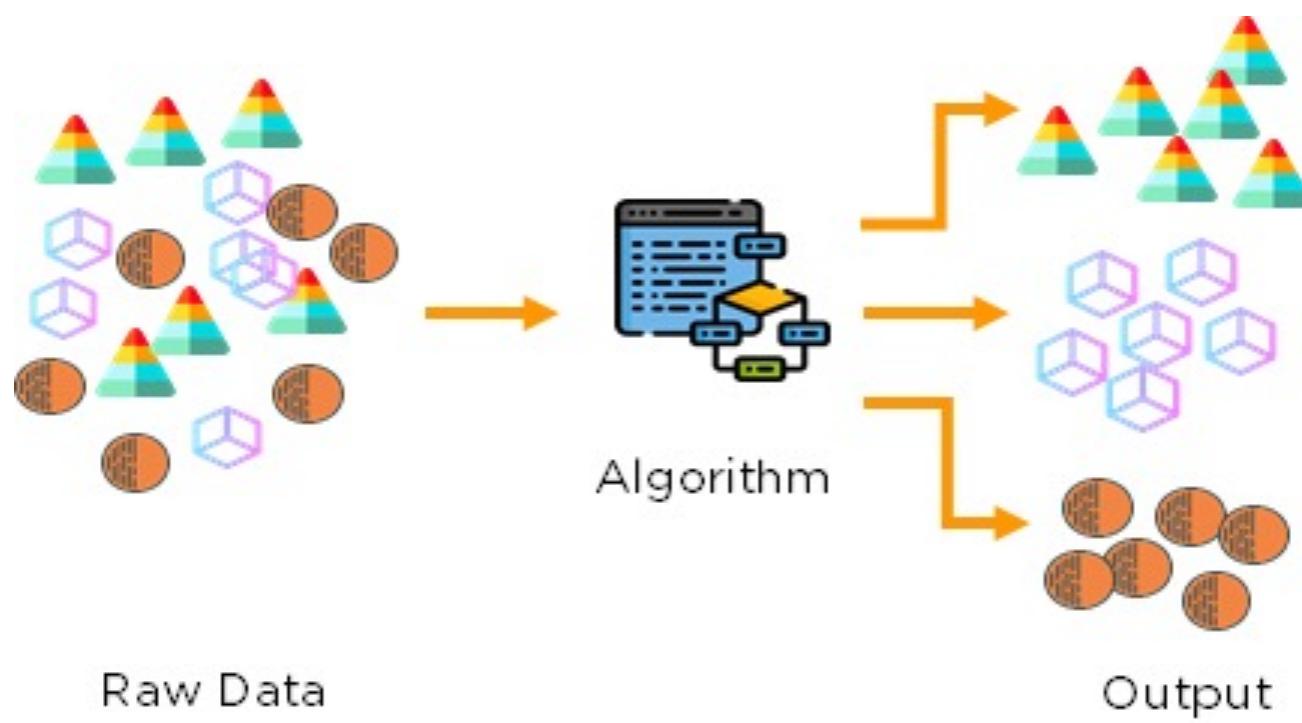


Regression

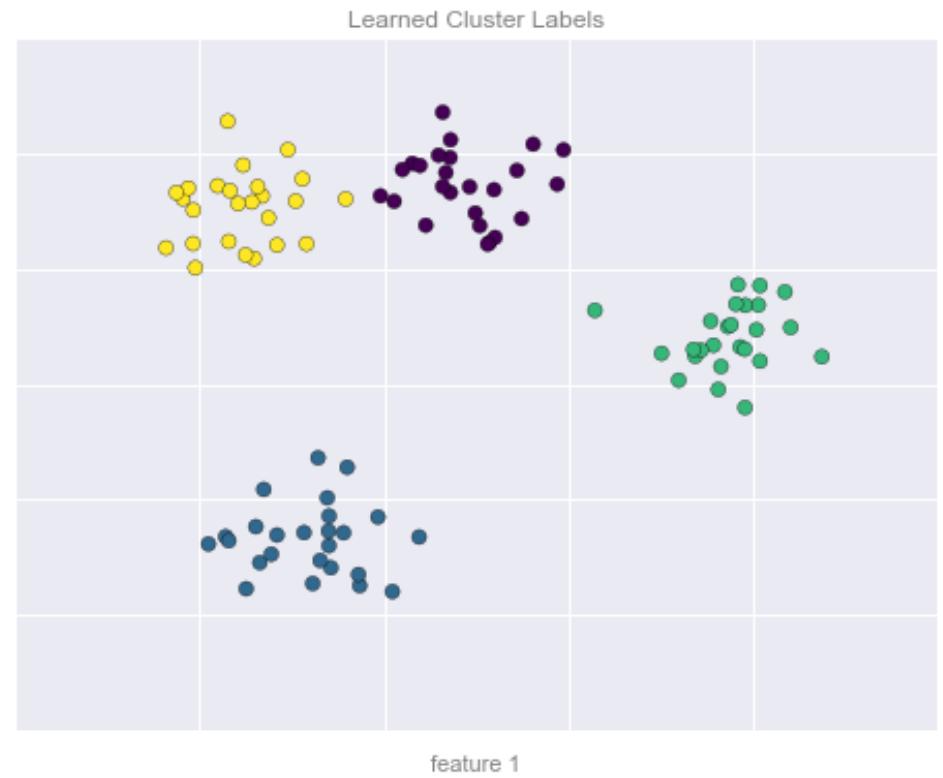
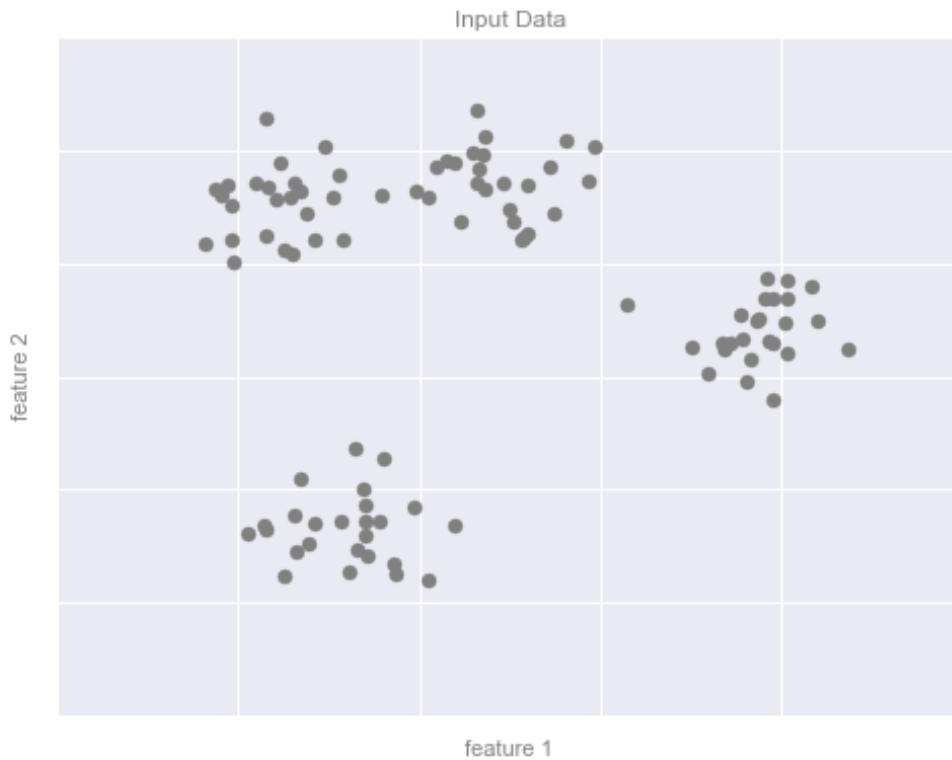


Unsupervised Machine Learning

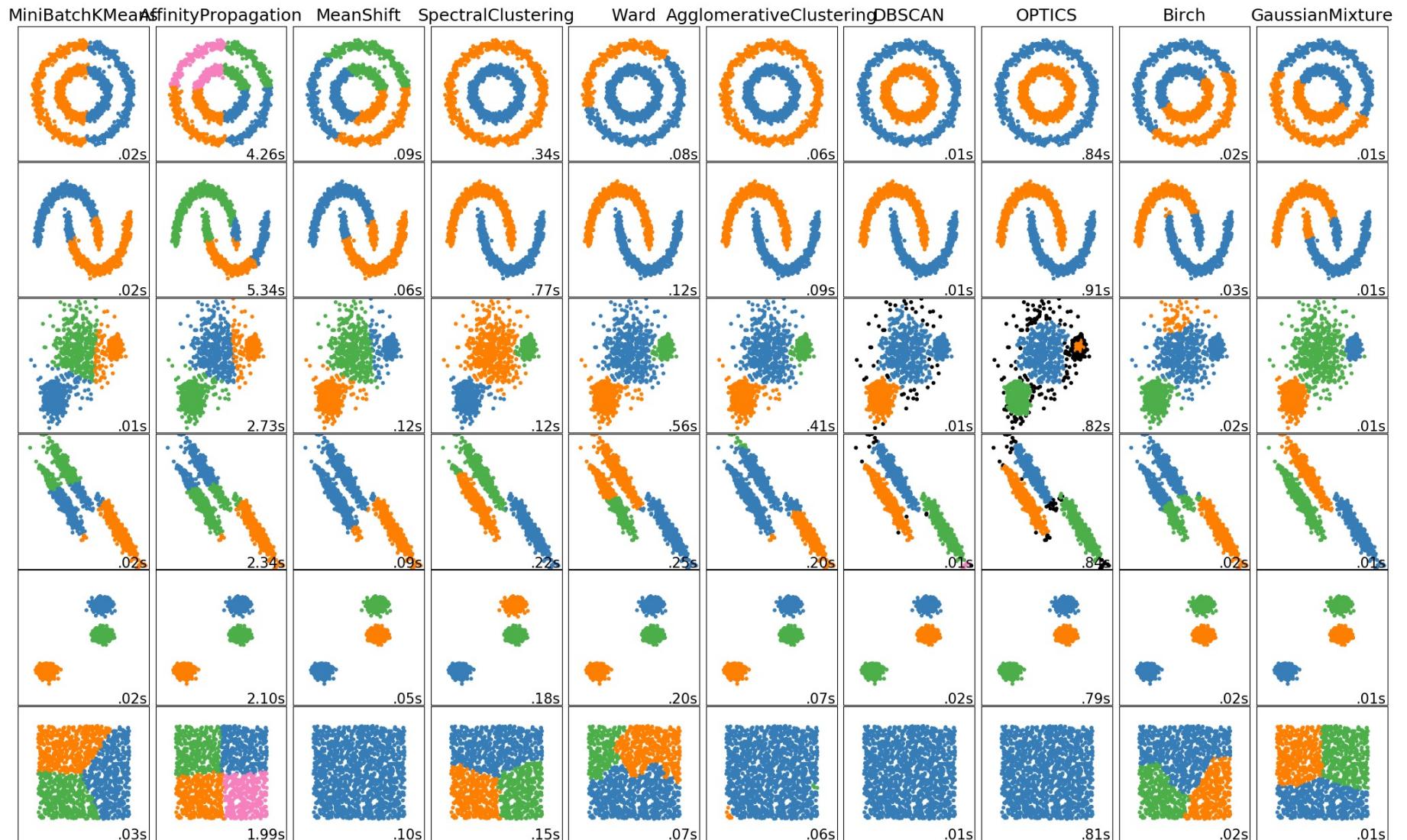
- Modeling the features of a dataset without reference to any label
 - "Let the dataset speak for itself"
- Given an input Set X: find patterns, classify input into categories



Clustering



Clustering



Some Machine Learning Methods:

Supervised Machine Learning

- Logistic Regression
- Random Forest
- Support Vector Machine
- Naive Bayes

Unsupervised Machine Learning:

- K-Means (for clustering)

for Topic Modeling:

- LDA Algorithm

Logistic Regression

- The probabilities describing the possible outcomes of a single trial are modeled using a **logistic function**:

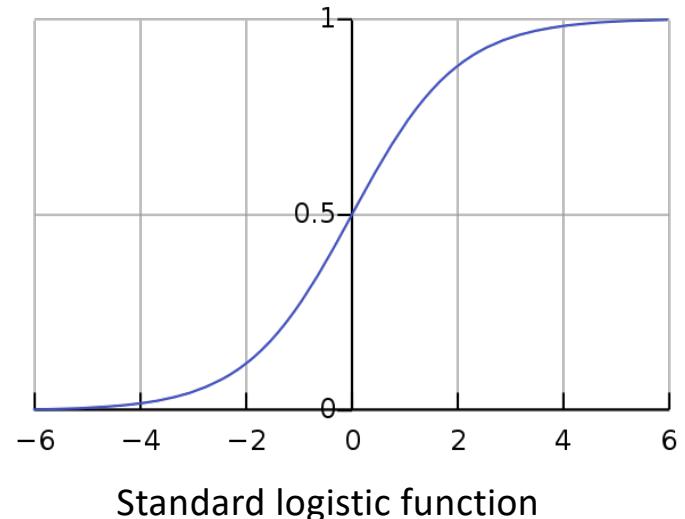
$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

where x_0 is the x value of the sigmoid's midpoint

L is the curve's maximum value

k is the logistic growth rate or steepness of the curve

(The standard logistic function, with L=1, k=1 and $x_0=0$ is called the sigmoid)



Regularized logistic regression:

- Add a term in your minimization problem that give a "cost" to the number of coefficient that are different from zero (or give a cost the higher the size of the coefficients)

(Lasso, Ridge Regressions, etc)

When you implement logistic regression in scikit-learn, the default is to apply the L2 regularization (i.e. a Ridge regression)

Regularized logistic regression is used also to interpret the most predictive words for a category, calculating the **marginal effects** for each word and listing the most predictive words for each category.

Random Forest

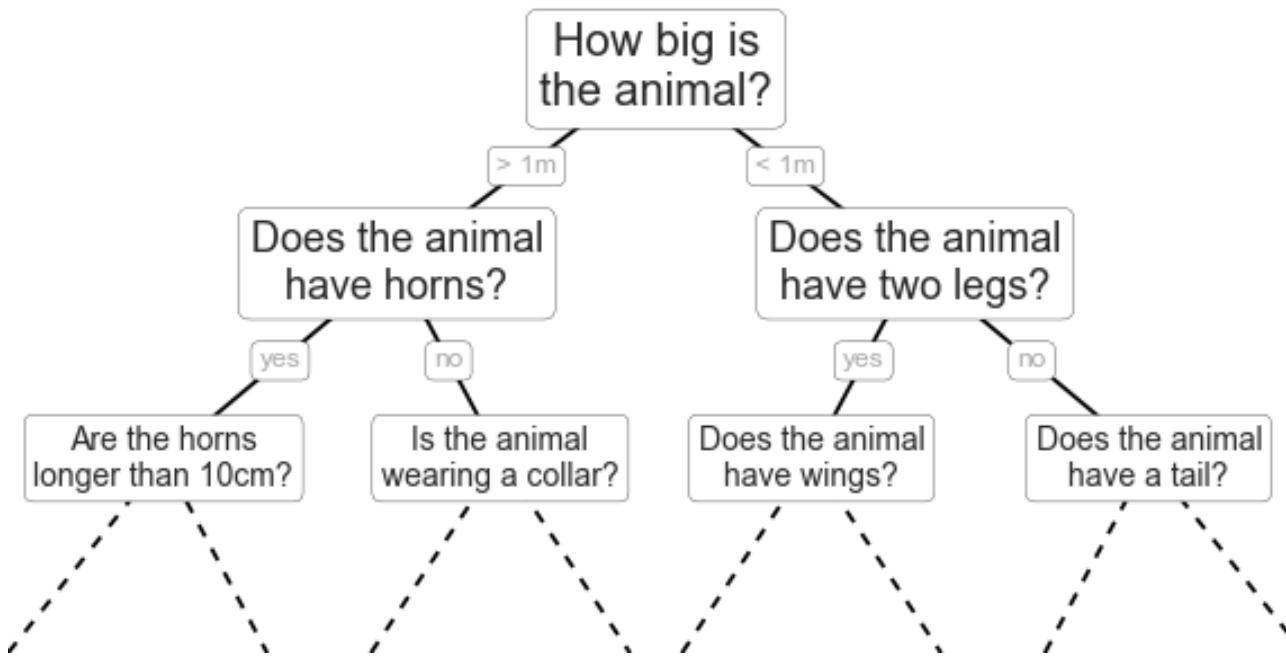
- Ensemble learner built on decision trees

Ensemble methods: they relies on aggregating the result of an ensemble of simpler estimators

What are Decision trees?

- Methods to classify data, based on "asking a series of questions" with a binary outcome
- Very intuitive (and interpretable) way to classify and label data

Decision Trees



- Decision trees tend to overfit the data
- **Random forest:** ensemble of decision trees (each tree in the ensemble is built from a sample drawn with replacement (bootstrap sample) from the training set)

Other important concepts in Supervised Machine Learning:

-Hyperparameter: the parameter of your model that you can choose at the beginning (i.e. the λ parameter in the LASSO regression that determine the weight of your cost term)

----> When you choose your optimal model within a class of models, you are comparing different models with different hyperparameters, and you will choose the one that obtain the best performances

(different methods for choosing hyperparameters, i.e. Grid Search, etc)

-Cross validation: resampling method – uses different portions of the data to test and train a model on different iteration

(i.e. you don't want your results to depend on the particular test set you chose)

K-fold

4-fold validation ($k=4$)



K-means

- Used for clustering
- Searches for a pre-determined number of clusters within an unlabeled multidimensional dataset

Optimal clustering:

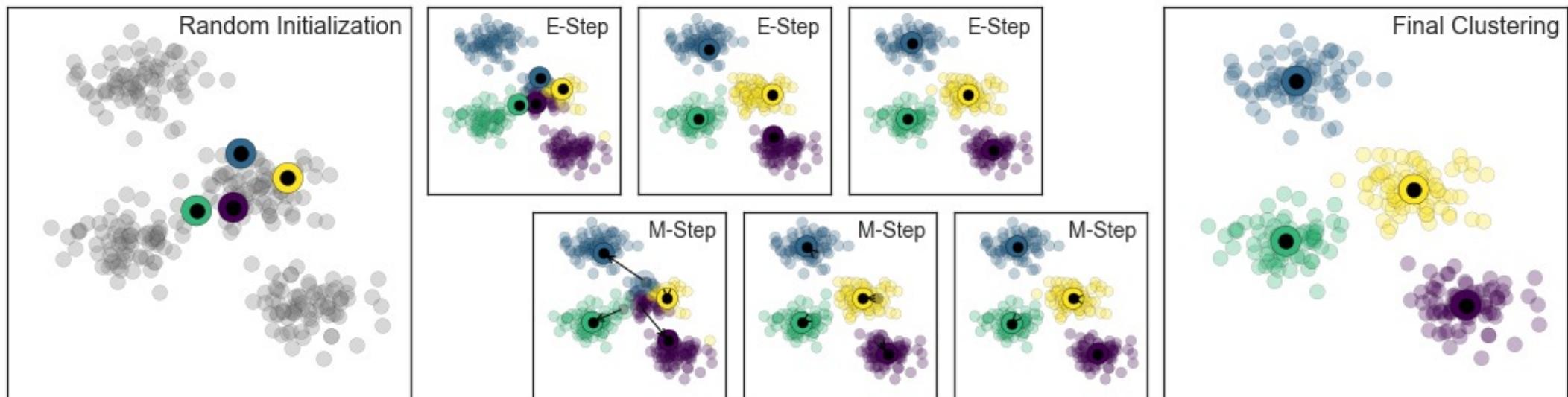
- The "cluster center" is the arithmetic mean of all the points belonging to the cluster
- Each point is closer to its own cluster center than to other cluster centers.

K-means

How does the algorithm work?

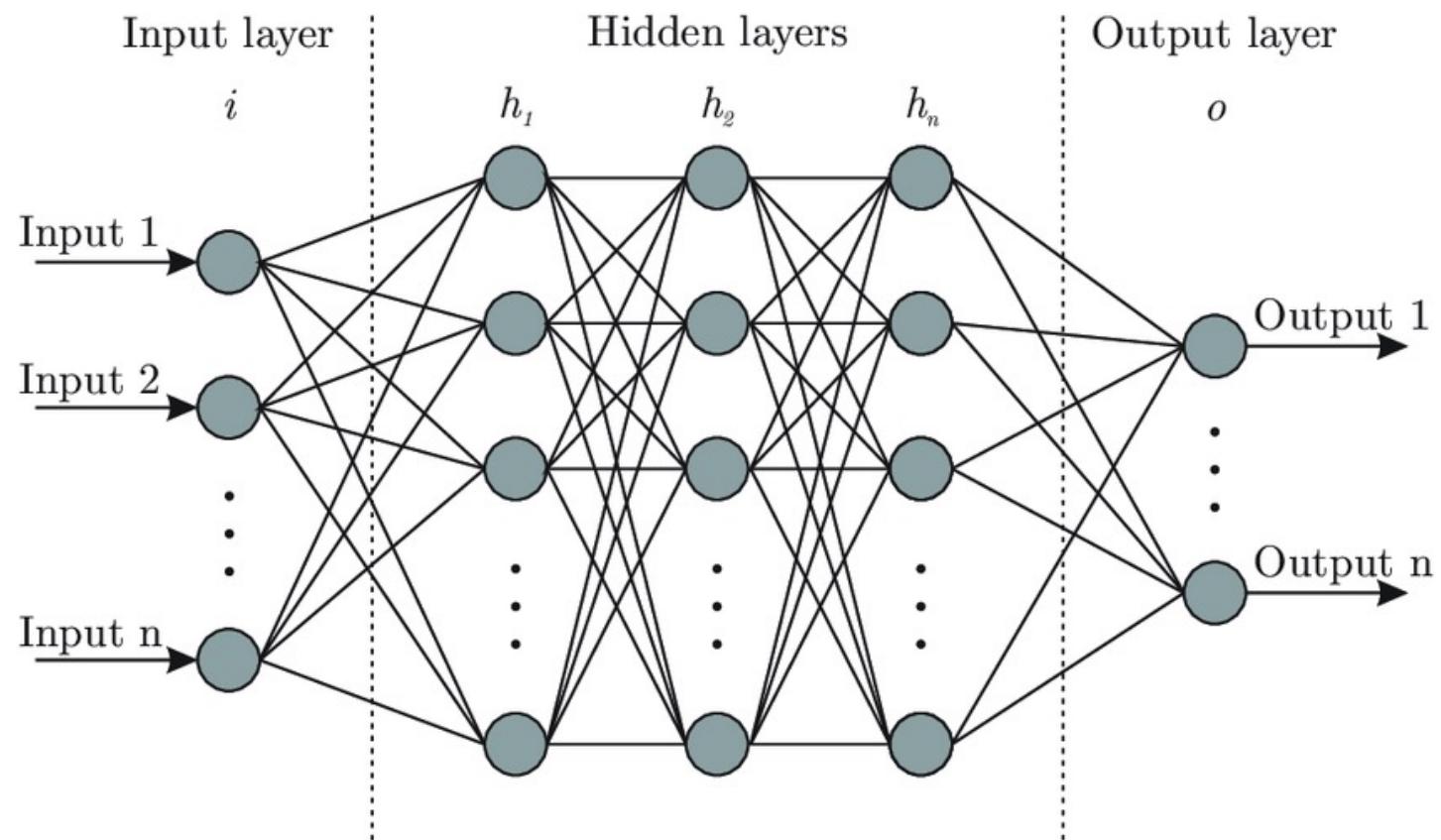
Expectation-maximization algorithm:

- 1) Guess some cluster centers
- 2) Repeat until converged
 - a) E-Step: assign points to the nearest cluster center
 - b) M-Step: set the cluster centers to the mean



A fast (very fast....) glimpse of Deep Learning

- Part of machine learning methods, based on artificial neural networks
- Inspired by how the brain works
- The frontier, for the majority of tasks in which machine learning is employed
- Applications: computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection, board game programs, ...



Problems of Machine Learning Algorithms:

- Interpretability
- They made persist bias that could exist in the dataset that is used for training
- Privacy

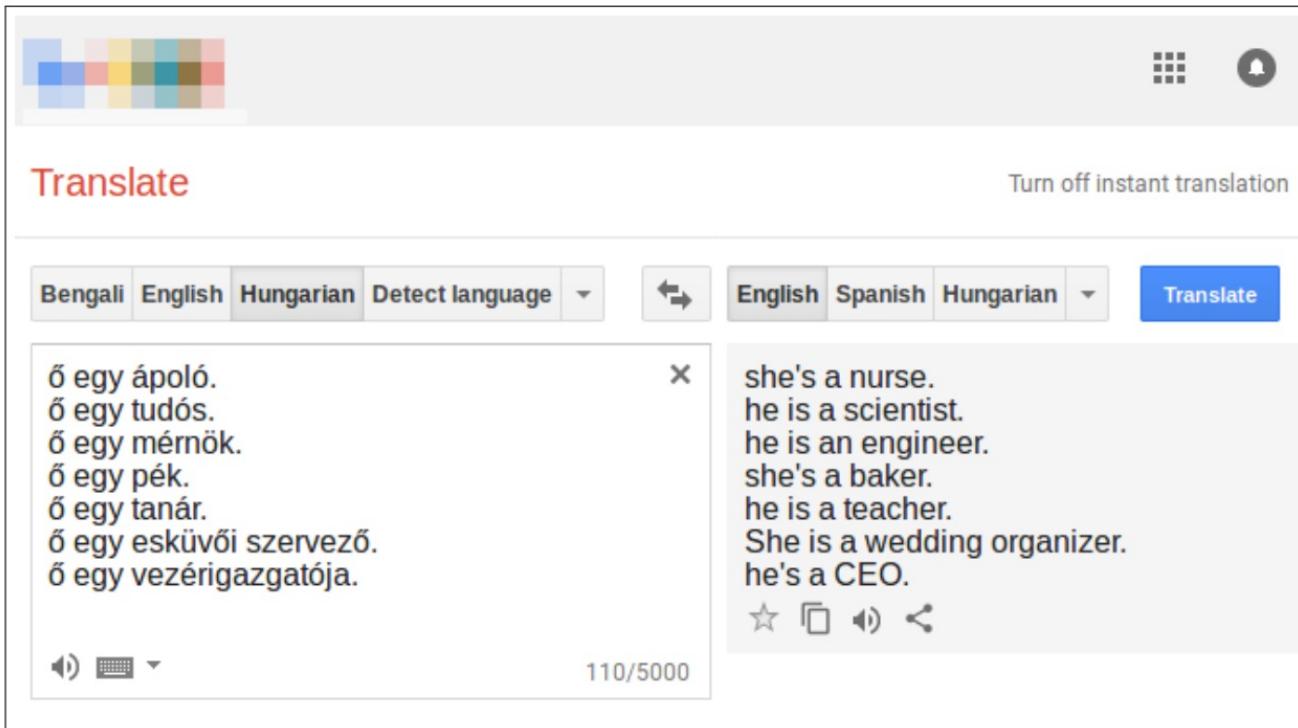


Figure 1: Translating sentences from a gender neutral language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in machine translation. This screenshot from Google Translate shows how occupations from traditionally male-dominated fields [40] such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female.

From the paper "Assessing Gender Bias in Machine Translation – A Case Study with Google Translate" (Prates, Avelar and Lamb, 2019). Link here: <https://arxiv.org/pdf/1809.02208.pdf>

**2) How do Economists use
Machine Learning?**

How do economists use Machine Learning algorithms?

Mainly, **two ways** in which algorithms are used in economic analysis and research:

- Machine Learning and Econometrics:** to obtain (and use) econometrics methods that can give more information than usual econometrics models (i.e. random forest for estimation of heterogeneous treatment effects)
- "**Text-as-data papers**": To obtain information from "unstructured data" (i.e. videos, images, texts, etc) and use these information in their analysis (i.e. "Text-as-data" type of use)

I will start from describing you how I use Machine Learning in my research....

"Regulation of Social Media and the Evolution of Content: A cross-platform analysis" - Marina Rizzi

Research Question: how online platform regulation affect abusive content, in the platform itself and in other complementary platforms

Specifically: Banning of racist hate speech in Twitter. How this policy affect content in Twitter (the regulated platform) and in Parler (a less regulated platform)?

Some coding examples for SUPERVISED Machine Learning (i.e. How I apply Supervised Machine Learning to my research)

What did I need in my research?

I was in need to understand whether a post can be considered hate speech (and in particular racist hate speech) or not

What did I find online (luckily)? A pre-labelled dataset!! ☺

Pre-labelled dataset (the "Measuring Hate Speech Dataset")

Text	Hatespeech Dummy	Target Race	Target Women	Target Sexuality	Target Religion	Target Gender	Target Disability	Target Origin
White people are trash.	1	1	0	0	0	0	0	0
We are already heavily taxed, you can't expect the hard working kiwi taxpayer to do both, first we fix our homeless children & those in poverty, then once o...	0	0	0	0	0	0	0	0
Queer rights are human rights	0	0	0	1	0	0	0	0

- I train an algorithm on the **pre-labelled** dataset

```
df['text_clean02'] = df['text'].apply(lambda x: clean_text02(x))

X = df['text_clean02']
y = df['hate_speech_dumm']

Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, random_state=1)

vectorizer = CountVectorizer()
Xtrain_counts = vectorizer.fit_transform(Xtrain)

tfidf_transformer = TfidfTransformer()
Xtrain_tfidf = tfidf_transformer.fit_transform(Xtrain_counts)

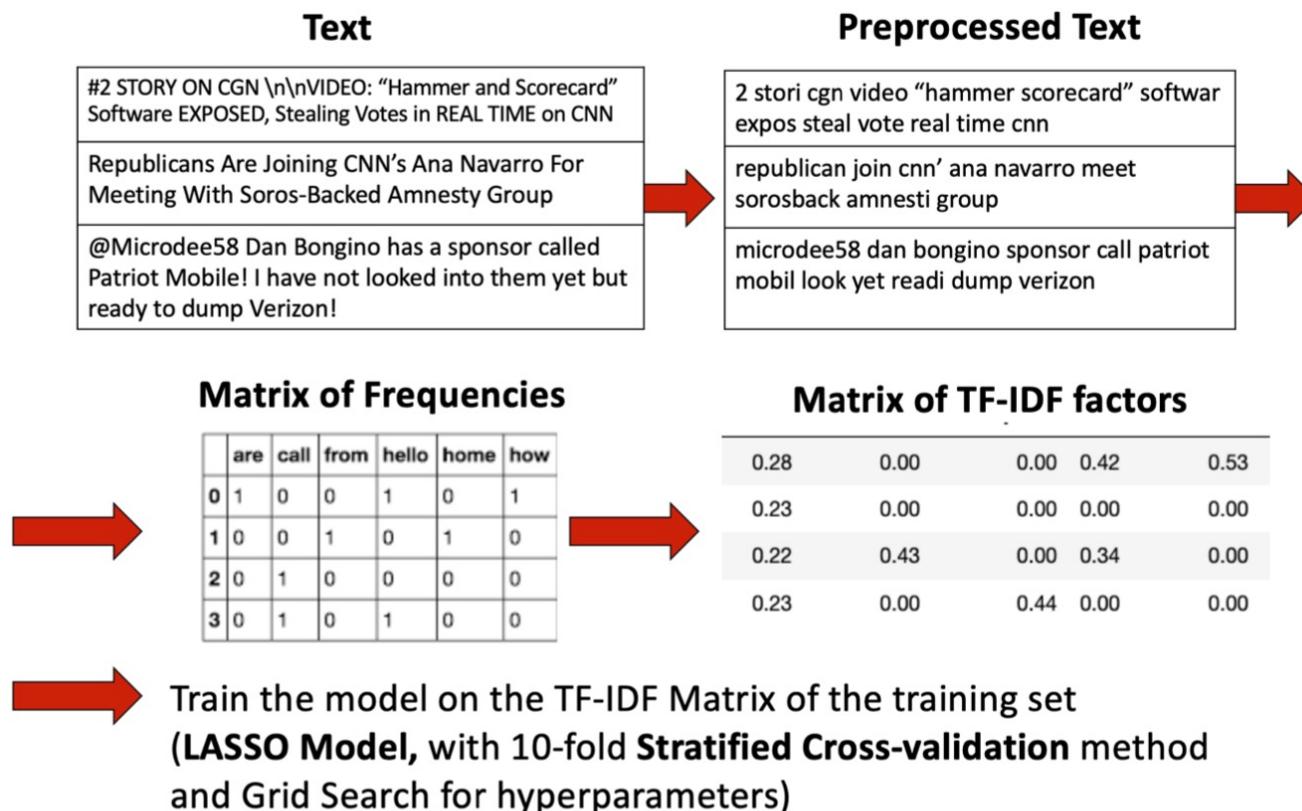
Xtest_counts = vectorizer.transform(Xtest)
Xtest_tfidf = tfidf_transformer.fit_transform(Xtest_counts)

model = LogisticRegressionCV(Cs=20, solver='liblinear', penalty='l1', cv=10)

model.fit(Xtrain_tfidf, ytrain)
y_model = model.predict(Xtest_tfidf)
```

Quick visualization of the pipeline I have to implement to create a supervised machine learning algorithm from text data

Measuring Hate Speech dataset (pre-labelled)



I apply the algorithm on the dataset I want to analyze:

Predicted with Text Classifiers

Post	Platform	User	Created at	likes	retweets	[...]	Hate Speech	Hate Speech toward Ethnic groups	Talking about Ethnic groups
Be sure to tune in tomorrow!	Twitter	Governor Kaylvey	2020-11-02	34	8		0	0	0
I had a chat with the Mexican lady of the library yesterday	Twitter	JenniferC	2020-12-28	3	0		0	0	1
Get a life you animal f**, shut up!	Twitter	ny2356	2020-11-18	2	0		1	0	0

I obtain the panel dataset I needed for my analysis!

Date	User	Upvotes	Reposts	N. of Posts	N. of Hate Speech posts	N. of Hate Speech Posts toward Minorities	N. of Hate Speech Posts toward Women
2020-11-01	007cigarjoe	14	3	55	3	1	0	
2020-11-02	007cigarjoe	63	21	60	4	2	0	
2020-11-01	AdamianVahe	0	0	2	0	0	0	
2020-11-04	Aewilson	2	0	1	0	0	0	
2020-11-01	Rexomus	3	0	7	1	0	1	
2020-11-06	Rexomus	4	2	10	2	1	1	

Difference-in-Differences regression:

$$y_{i,p,t} = \alpha + \beta_1 Post_t * Parler_p + \mu_i + \delta_t + \varepsilon_{i,p,t}$$

$y_{i,t}$: outcome variable (i.e. % of hate speech posts) for user i in time t in platform p

$Post_t$: dummy = 1 if day t is after Twitter's regulation

$Parler_p$: dummy = 1 if I am considering user accounts in Parler

μ_i individual fixed effects and δ_t daily fixed effects

Table: Spillover effects for users that have accounts in both platforms

VARIABLES	(1) Hate Speech	(2) Racist Hate Speech	(3) Hate Speech based on Sexuality	(4) Hate Speech based on Origin
Post Treatment	-0.00279 (0.00267)	-0.00380*** (0.00112)	-0.000728 (0.000567)	-0.000256 (0.00126)
Parler	0.0160*** (0.00478)	0.00258 (0.00176)	0.00106 (0.00115)	0.0102*** (0.00233)
Post Treatment * Parler	0.0145*** (0.00490)	0.00569*** (0.00218)	0.00149 (0.00134)	0.00260 (0.00263)
Observations	35,290	35,290	35,290	35,290
R-squared	0.112	0.051	0.049	0.079
Individual Fixed Effects	Yes	Yes	Yes	Yes

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Other examples of Economics papers that uses supervised machine learning approaches:

- Ash, E., Galletta, S., & Giommoni, T. (2021). **A machine learning approach to analyze and support anti-corruption policy.** R&R at American Economic Journal: Economic Policy
- Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). **Text classification of ideological direction in judicial opinions.** *International Review of Law and Economics*, 62, 105903.
- Ash, E., Guillot, M., & Han, L. (2021, November). **Machine Extraction of Tax Laws from Legislative Texts.** In *Proceedings of the Natural Legal Language Processing Workshop 2021* (pp. 76-85).

Text-as-data Papers

(i.e. papers that use text to obtain some variables for their research question).

We are going to have a brief overview of the following papers:

- Ash, E., Chen, D. L., & Ornaghi, A. (2021). **Gender attitudes in the judiciary: Evidence from US circuit courts.** (R&R at *American Economic Journal: Applied Economics*)
- Cagé, J., Hervé, N., & Viaud, M. L. (2020). **The production of information in an online world: Is copy right?.** *The Review of Economic Studies*, 87(5)
- Sockin, J. (2021). **Show Me the Amenity: Are Higher-Paying Firms Better All Around?.** Available at SSRN.
- Hansen, S., Ramdas, T., Sadun, R., & Fuller, J. (2021). **The demand for executive skills** (No. w28959). National Bureau of Economic Research.
- Wu, A. (2017) **Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum**

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

Research question: do gender attitudes influence interactions with female judges in US Circuit Courts?

The paper: the authors propose a judge-specific measure of gender attitudes based on use of gender-stereotyped language in the judge's authored opinions. It exploits the quasi-random assignment of judges to cases.

Data: The authored opinions of 139 judges in U.S. (total corpus of over 14 million sentences), and outcomes: reversals of district court decisions, opinion assignment, and citations + decisions in gender-related cases.

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

Identification strategy: Differences-in-differences. Exploit the quasi-random assignment of judges to cases (in addition on conditioning on judges' characteristics)

Results: higher-slant judges vote more conservatively in gender-related cases. They interact differently with female colleagues: they are more likely to reverse lower-court decisions if the lower-court judge is a woman than a man, are less likely to assign opinions to female judges, and cite fewer female-authored opinions.

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

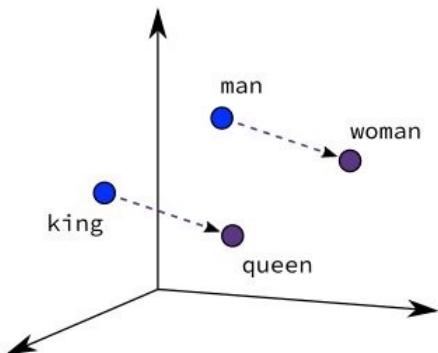
How does this paper exploit Machine Learning?

They construct a judge-specific measure of gender attitudes based on use of gender-stereotyped language in the judge's authored opinions.

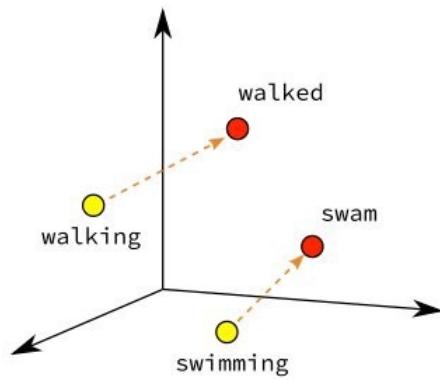
----> They use NLP (natural language processing) to develop a measure of *gender slant* based on how strongly judges associate men with careers and women with families in the opinions they write

----> In particular: *word embedding* ----> an algorithm that distributes words in a vector space, based on their co-occurrence in a corpus, and should represent their semantic meaning and relationships)

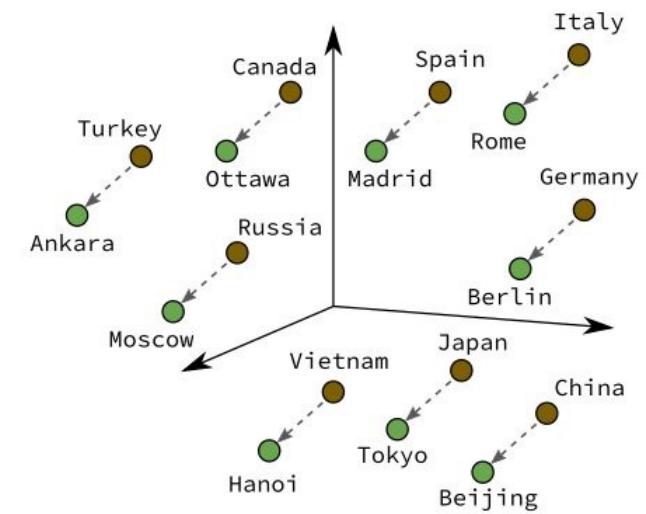
Word Embedding



Male-Female



Verb Tense



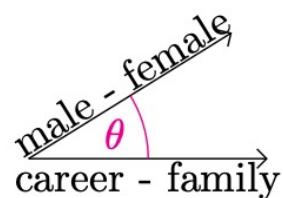
Country-Capital

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

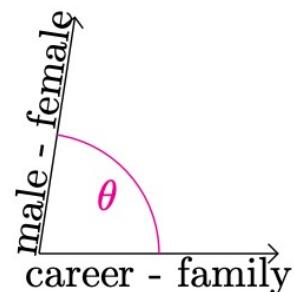
Cosine similarity ($\cos(\theta)$) between male – female (gender dimension) and career - family (career-family dimension)

Figure 2: Measuring Gender Attitudes using Cosine Similarity

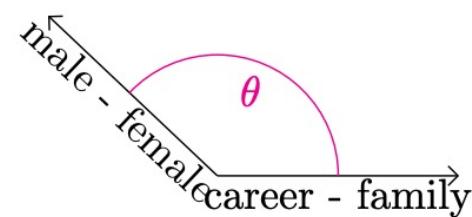
(a) Slant ≈ 1



(b) Slant ≈ 0



(c) Slant ≈ -1



Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

- For each judge: their corpus of authored opinions
 - > They train an embedding for each judge
 - > they obtain a specific gender slant measure for each judge (139 circuit judges)
- (validation checks** - female and younger judges have lower gender slant)

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

Validation checks

Table 2: Correlates of Gender Slant

Dependent Variable	Gender Slant					
	(1)	(2)	(3)	(4)	(5)	(6)
Democrat	-0.027 (0.172)				-0.003 (0.178)	0.083 (0.269)
Female		-0.502* (0.288)			-0.592*** (0.202)	-0.713** (0.276)
Minority			-0.098 (0.329)		-0.164 (0.194)	0.453 (0.283)
Born in 1920s				-0.069 (0.191)	0.080 (0.208)	0.152 (0.299)
Born in 1930s					-0.765*** (0.203)	-0.740*** (0.234)
Born after 1940					-0.537** (0.229)	-0.558** (0.258)
Daughter						-0.490* (0.275)
Observations	139	139	139	139	139	98
Outcome Mean	0.000	0.000	0.000	0.000	0.000	-0.085
Adjusted R2	-0.007	0.020	-0.007	0.087	0.440	0.529
Circuit FE					X	X
Additional Controls					X	X
Number of Children FE						X

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

$$\text{Conservative Vote}_{jcit} = \beta \text{Gender Slant}_j + X_j' \gamma + \delta_{ct} + \varepsilon_{ictj}$$

ConservativeVote_{jcit}: is an indicator variable equal to 1 if judge j of circuit c voted conservatively (against expanding women's rights) in case i during year t

One of the results: judges with higher slant are more likely to vote conservatively in gender-related cases.

Gender attitudes in the judiciary: Evidence from US circuit courts (Ash, Chen and Ornaghi, 2021)

Table 3: Effect of Gender Slant on Decisions in Gender-Related Cases

Dependent Variable	Conservative Vote				
	(1)	(2)	(3)	(4)	(5)
Gender Slant	0.041*** (0.016)	0.041*** (0.013)	0.041*** (0.012)	0.050*** (0.014)	0.046*** (0.012)
Democrat		-0.144*** (0.025)	-0.141*** (0.025)	-0.135*** (0.023)	-0.148*** (0.025)
Female		-0.031 (0.032)	-0.042 (0.032)	-0.017 (0.025)	-0.034 (0.034)
Observations	3086	3086	3086	3086	3086
Clusters	113	113	113	113	113
Outcome Mean	0.606	0.606	0.606	0.606	0.606
Circuit-Year FE	X	X	X	X	X
Additional Controls		X	X	X	X
Year of Appointment			X		
Exposure FE				X	
No Gender-Related Cases					X

Show Me the Amenity: Are Higher-Paying Firms Better All Around?

(Jason Sockin, 2021)

Research Question: Do higher-paying firms offer more favorable work, or compensate for less favorable work? (estimation of joint distribution of wages, amenities, and job satisfaction across firms)

Data: job reviews/reports from Glassdoors

Identification strategy: AKM (and other analysis)

Results: 1) High-paying firms are high-satisfaction firms because they offer better amenities 2) workers, especially high-earners, are willing to pay for job satisfaction 3) incorporating non-wage amenities nearly doubles the variance in total compensation across firms

Glassdoor review

4.0 ★★★★☆ ✓

Current Employee, more than 5 years

Very Supportive Culture

28 Jun 2022 - Senior National Account Manager in London, England

✓ Recommend CEO Approval ✓ Business Outlook

Pros

Great culture with a real emphasis on work life balance. Genuinely care about their employees and dedicate a lot of time to personal development

Cons

Massive company and progression can be slower than at other businesses. Minimum 2 years expected in each role before progression

[Continue reading](#)

Show Me the Amenity: Are Higher-Paying Firms Better All Around?

(Jason Sockin, 2021)

How does he use Machine Learning?

Semi-supervised Topic Modeling (Anchored Correlation Explanation – CorEx model – of Gallagher et al. (2017) on workers' free response description (pros and cons) in Glassdoor

- 1) The researcher chose 48 amenities (that he thinks are important in these descriptions)
- 2) He gives to the algorithm some topic-specific "anchor words", to help guide topic to convergence (and improve interpretability of topics) (the anchor words still refer to the 48 amenities he identifies)
- 3) The algorithm create the clusters (according to its objective)

Since last year we were uncertain on how this algorithm worked....

I tried to simulate an example using the "Corex" Topic Modeling Python package

	text	
0	i like dogs	
1	i like cats	
2	my cat destroyed everything inside my living room	
3	my dog like to sleep in the kitchen	
4	elephants like to eat a lot of fruits	
5	bananas are yellow	
6	an apple can have different colours	
7	i don't like so much to eat an apple alone	
8	i like pear	

No Anchors

```
[ ] from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
vectorizer = CountVectorizer()
Xcounts = vectorizer.fit_transform(df['text'])
Xcounts
```



```
<9x35 sparse matrix of type '<class 'numpy.int64'>'  
      with 46 stored elements in Compressed Sparse Row format>
```

	alone	an	apple	are	bananas	can	cat	cats	colours	destroyed	...	much	my	of	pear	room	sleep	so	the	to	yellow
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	1	...	0	2	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	1	0	1	1	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	1	0
5	0	0	0	1	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
6	0	1	1	0	0	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
7	1	1	1	0	0	0	0	0	0	0	...	1	0	0	0	0	0	1	0	1	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0

9 rows × 35 columns

```

##### Without anchors:
import numpy as np
import scipy.sparse as ss
from corextopic import corextopic as ct

# Train the CorEx topic model
topic_model = ct.Corex(n_hidden=2) # Define the number of latent (hidden) topics to use.
probabilities = topic_model.fit(Xcounts, words=columns_names)
probabilities02 = topic_model.fit_transform(Xcounts, words=columns_names)
#####, docs=docs)

prob_topics = pd.DataFrame(probabilities02)
prob_topics

```

	0	1	
0	False	True	
1	False	True	
2	False	False	
3	True	True	
4	True	True	
5	False	False	
6	True	False	

	text	text_tokenized	0	1
0	i like dogs	[i, like, dogs]	False	True
1	i like cats	[i, like, cats]	False	True
2	my cat destroyed everything inside my living room	[my, cat, destroyed, everything, inside, my, l...]	False	False
3	my dog like to sleep in the kitchen	[my, dog, like, to, sleep, in, the, kitchen]	True	True
4	elephants like to eat a lot of fruits	[elephants, like, to, eat, a, lot, of, fruits]	True	True
5	bananas are yellow	[bananas, are, yellow]	False	False
6	an apple can have different colours	[an, apple, can, have, different, colours]	True	False
7	i don't like so much to eat an apple alone	[i, don't, like, so, much, to, eat, an, apple,...]	True	True
8	i like pear	[i, like, pear]	False	True

With anchors

```
[ ] topic_model02 = ct.Corex(n_hidden=2)
topic_model02.fit(Xcounts, words=columns_names, anchors=[['dogs','cats'], ['apple','bananas', 'fruits']], anchor_strength=2)

<corextopic.corextopic.Corex at 0x7fe2146e1c30>
```

```
▶ prob_topics02 = pd.DataFrame(topic_model02.labels)
prob_topics02
```

	0	1
0	False	True
1	False	True
2	True	False
3	False	True
4	False	True
5	False	False
6	True	False
7	True	True
8	False	True

	text	text_tokenized	0	1	0	1
0	i like dogs	[i, like, dogs]	False	True	False	True
1	i like cats	[i, like, cats]	False	True	False	True
2	my cat destroyed everything inside my living room	[my, cat, destroyed, everything, inside, my, l...]	False	False	True	False
3	my dog like to sleep in the kitchen	[my, dog, like, to, sleep, in, the, kitchen]	True	True	False	True
4	elephants like to eat a lot of fruits	[elephants, like, to, eat, a, lot, of, fruits]	True	True	False	True
5	bananas are yellow	[bananas, are, yellow]	False	False	False	False
6	an apple can have different colours	[an, apple, can, have, different, colours]	True	False	True	False
7	i don't like so much to eat an apple alone	[i, don't, like, so, much, to, eat, an, apple,...]	True	True	True	True
8	i like pear	[i, like, pear]	False	True	False	True

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

The paper: it examines whether people in academia portray and judge women and men differently in everyday “conversations” that take place online.

How? The author combines methods from text mining, machine learning and econometrics to study the existence and extent of gender stereotyping on Economics Job Market Rumors forum

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

Results:

- Through a topic analysis, the author finds that the discourse tends to become significantly less academic or professional oriented, and more about personal information and physical appearance when women are mentioned.
- The words with the strongest predictive power on gender, selected by the Lasso-logistic model, provide a direct look into the gender stereotyping language on this forum.
- A panel data analysis reveals the state dependence between the content of posts within a thread. In particular, if women are mentioned previously in a thread, the topic is likely to shift from academic to personal.

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

Data: posts from the EJMR forum (2014-2016)

Table 2: Summary of the text data

	All	Gender Related	Female (Level 1)	Male (Level 1)
Number of Posts	1,143,416	237,863	56,171	181,692
Number of Threads	131,913	74,679	22,708	66,796

Notes: "Level 1" uses the most inclusive set of gender classifiers to identify gender related posts. Duplicate observations have been resolved by the Lasso-Logistic model in Section 2.2.

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

Topic Analysis: she manually classified the top 10.000 words into 15 categories.

Table 4: Categories of Words for Topic Analysis

Category	No. Words	Examples
<i>i. Academic/Professional</i>		
Economics	177	“economics”, “macro”, “empirical”, “QJE”, “Keynesian”
Academic-General	1,515	“research”, “papers”, “tenure”, “teaching”, “professor”
Professional	138	“career”, “interview”, “payrolls”, “placement”, “recruit”
<i>ii. Personal/Physical</i>		
Personal Information	118	“family”, “married”, “kids”, “relationship”, “lifestyle”
Physical Attributes	125	“beautiful”, “handsome”, “attractive”, “body”, “fat”
Gender related	86	“gender”, “femine”, “masculine”, “sexist”, “sexual”

Notes: “Gender related” category under *Personal/Physical* are not used as gender classifiers. The complete list of 15 categories can be found at Appendix I.

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

She uses two benchmark models to estimate the gender differences in topics. The first model looks at the effects of gender on the sum of word frequencies in each topic, while the second uses an indicator $1[Topic_i > 0]$ of whether any word from a given topic occurs:

$$(i) : Topic_i = \gamma_0 + \gamma_1 Female_i + e_i$$

$$(ii) : 1[Topic_i > 0] = \theta_0 + \theta_1 Female_i + u_i$$

$Topic \in \{Academic/Professional, Personal/Physical\}$

Table 5: Academic/Professional

	Level 1		Level 2		Level 3		Level 4	
	counts	1[<i>counts</i> > 0]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Female</i> _i	-1.756 (0.035)	-0.132 (0.002)	-1.893 (0.043)	-0.143 (0.003)	-2.084 (0.053)	-0.164 (0.003)	-2.241 (0.055)	-0.165 (0.003)
Constant	4.067 (0.017)	0.706 (0.001)	4.444 (0.022)	0.742 (0.001)	4.676 (0.024)	0.758 (0.001)	4.918 (0.027)	0.764 (0.002)
<i>N</i>	233,433	233,433	168,293	168,293	142,131	142,131	111,535	111,535
R ²	0.011	0.014	0.011	0.018	0.011	0.022	0.015	0.026
Adj. R ²	0.011	0.014	0.011	0.018	0.011	0.022	0.015	0.026
F Stat.	2,575.068	3,373.472	1,931.830	3,145.749	1,548.165	3,153.588	1,657.292	2,923.917

Notes: Standard errors are in parentheses. Restrict to posts with ≥ 3 and ≤ 252 words, roughly 98% of each sample. “Level 1” to “Level 4” refer to increasingly restrictive levels of gender classifiers to identify gender-related posts (see Figure 1). The odd numbered regressions estimate model (i), while the even numbered ones estimate model (ii).

Table 6: Personal/Physical

	Level 1		Level 2		Level 3		Level 4	
	counts	$1[counts > 0]$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Female_i$	0.883 (0.007)	0.282 (0.002)	0.846 (0.009)	0.248 (0.003)	0.776 (0.011)	0.234 (0.003)	0.724 (0.012)	0.208 (0.003)
Constant	0.458 (0.004)	0.256 (0.001)	0.518 (0.005)	0.274 (0.001)	0.505 (0.005)	0.262 (0.001)	0.598 (0.006)	0.301 (0.002)
N	233,433	233,433	168,293	168,293	142,131	142,131	111,535	111,535
R^2	0.059	0.066	0.047	0.052	0.036	0.041	0.031	0.036
Adj. R^2	0.059	0.066	0.047	0.052	0.036	0.041	0.031	0.036
F Stat.	14,596.460	16,461.220	8,309.106	9,192.315	5,310.392	6,136.646	3,512.172	4,115.958

Notes: Standard errors are in parentheses. Restrict to posts with ≥ 3 and ≤ 252 words, roughly 98% of each sample. “Level 1” to “Level 4” refer to increasingly restrictive levels of gender classifiers to identify gender-related posts (see Figure 1). The odd numbered regressions estimate model (i), while the even numbered ones estimate model (ii).

"Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum" by Alice Wu (August 2017)

Word Selection using Lasso-Logistic Model

-She designs a propensity score model to predict the gender a post is related to by the number of occurrences of the 10, 000 most frequent words, excluding the gender classifiers and the last names of celebrities (not economists).

-The logistic-Lasso model sorts out words with the strongest predictive power on gender

Top 30 Words with the strongest predictive power for Female_i = 1

Level 1			
	Most “female”		Most “male”
Word	Marginal Effect	Word	Marginal Effect
hotter	0.388	homosexual	-0.237
hot	0.285	homo	-0.228
attractive	0.260	philosopher	-0.204
pregnant	0.252	keen	-0.182
gorgeous	0.251	motivated	-0.171
beautiful	0.249	fieckers	-0.164
tit	0.247	slides	-0.160
lesbian	0.242	nordic	-0.156
bang	0.229	filling	-0.152
horny	0.224	textbook	-0.148
slept	0.224	adviser	-0.140
marry	0.221	fenance	-0.138
attracted	0.216	fieckers	-0.137
0,0	0.213	bowl	-0.136
cute	0.209	gay	-0.132
breasts	0.205	bench	-0.129
sexy	0.202	mountain	-0.128
pregnancy	0.195	humble	-0.124
dumped	0.194	iraq	-0.122
feminazi	0.189	rust	-0.121
feminist	0.187	amusing	-0.120
raped	0.185	speeches	-0.119
dated	0.184	affected	-0.119
cheerful	0.182	mere	-0.119
ugly	0.179	chill	-0.118
marrying	0.178	bugs	-0.118
blonde	0.175	rip	-0.113
crush	0.173	recession	-0.111
date	0.172	brilliant	-0.108
naked	0.172	salmon	-0.108

The Demand for Executive Skills (Hansen, Ramdas, Sadun and Fuller, 2021)

Research Question: which skills are required in managerial labour markets?

Data: large corpus of detailed and previously unexplored job descriptions for C-suite positions spanning a time period of 17 years (provided to them by one of the world's largest headhunting companies)

Results: the data show an increasing relevance of social skills in top managerial occupations, and a greater emphasis on social skills in larger and more information intensive organizations.

The Demand for Executive Skills (Hansen, Ramdas, Sadun and Fuller, 2021)

How do they use Machine Learning?

They classify the information contained in these documents using methods borrowed from machine learning, which allow us to map unstructured, free-text data into distinct clusters of skill requirements. They use the data to examine the variation in the demand for different managerial skills which provide the first direct evidence on C-suite skill requirements.

The Demand for Executive Skills (Hansen, Ramdas, Sadun and Fuller, 2021)

"We propose a novel classification approach to derive economically interpretable measures from the unstructured text of this corpus. Our approach involves two steps. First, we define a comprehensive vector of skills requirements that are relevant for Chief Executives. We obtain this by collecting the numerous textual descriptions of skills from the O*NET entry for the Chief Executive occupation, and clustering them into six broad categories using a k-means algorithm. Second, we express each job description in the search corpus in terms of the relative demand for each skill cluster by comparing the similarity of the language included in the document with the text associated with each of the O*NET clusters.

Both the clustering of O*NET skills into groups and the comparison of job texts to O*NET texts require the quantification of linguistic relatedness. We compute this via a language embedding model estimated from an auxiliary corpus of all Harvard Business Review articles from its inception in 1922 to the present day. This large, domain-specific corpus allows us to obtain semantic relationships between words in the context of business and management. We then apply the model to measure similarity in the O*NET and job search corpora, an approach known as transfer learning."

The Production of Information in an Online World" (Cagè, Hervè and Viaud, 2020)

Research Question: what it is the extent of copying in online media, and what are the estimations for the returns to originality in online news production?

Data: all the online content produced by French news media during 2013 + new micro audience data

Identification Strategy: (for return of originality): regression with event, date and media fixed effects

Results: they document high reactivity of online media: one quarter of the news stories are reproduced online in under 4 minutes. We show that this is accompanied by substantial copying, both at the extensive and at the intensive margins. Returns to originality: they find that original content producers tend to receive more viewers, thereby mitigating the newsgathering incentive problem raised by copying.

The Production of Information in an Online World" (Cagè, Hervè and Viaud, 2020)

How do they use Machine Learning?

They perform a **topic detection algorithm** to construct the set of news stories. Each document is placed within the most appropriate cluster, i.e. the one that discusses the same event-based story. They obtain a total number of 25,000 stories.

(+ document the propagation of the story, + plagiarism algorithm, + copying with acknowledgment)

Machine Learning for Econometrics

(I will just quote quickly this topic, but potentially, there is a growing literature on this topic, and some codes are already available both in R and in Python!!)

- **Theoretical/Conceptual papers**

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

Athey, S. (2015, August). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 5-6).

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.

- **Applications**

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2), 37-51.

Goller, D., Harrer, T., Lechner, M., & Wolff, J. (2021). Active labour market policies for the long-term unemployed: New evidence from causal machine learning. *arXiv preprint arXiv:2106.10141*.

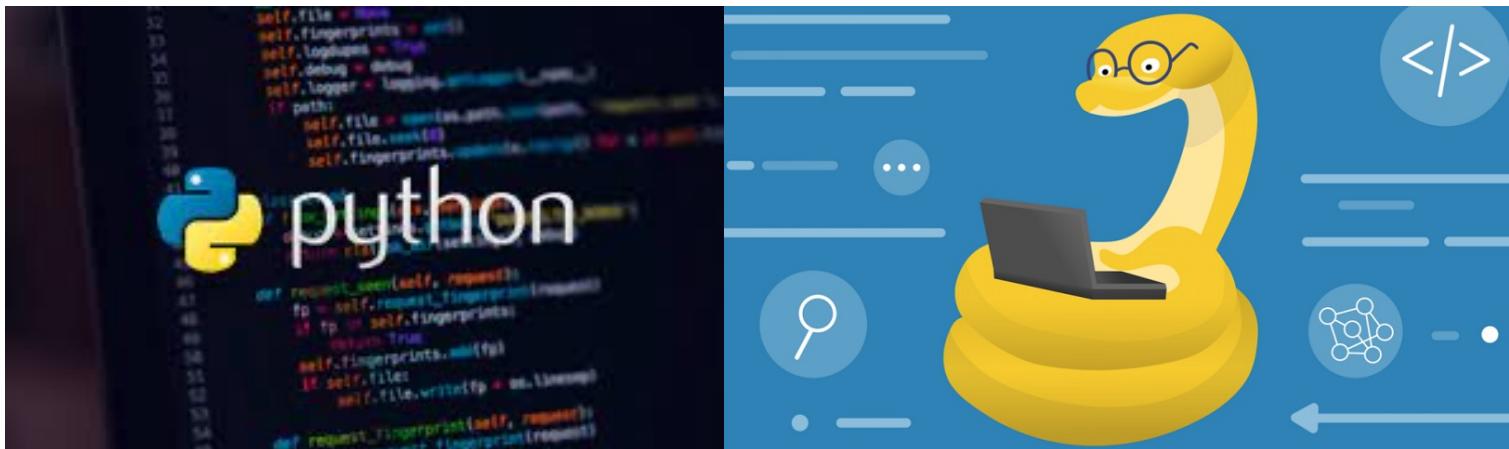
Cockx, B., Lechner, M., & Bollens, J. (2019). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *arXiv preprint arXiv:1912.12864*.

2b) Coding Machine Learning algorithms for Economics

Getting our hands dirty: using Python as programming language

Why Python?

- High-level, interpreted, general-purpose programming language.
- Higher code readability (with respect to other coding languages)
- Great libraries for scientific computing and data science (NumPy, Pandas, Matplotlib), for machine learning (TensorFlow, Keras, Pytorch, Scikit-Learn) and for natural language processing (NLTK, Gensim, SpaCy, etc)
- Some pretrained models available for download (to be fined-tuned) from platforms like HuggingFaces



Implementation of a Machine Learning pipeline: Python and the Scikit-Learn package

- Python is one of the most used programming language for easily implementing a Machine Learning pipeline
 - > Good packages for deep learning methods (TensorFlow, Keras), and some (large) pretrained models are made available online (i.e. on HuggingFaces)
- Scikit-Learn provides a "standard grammar" to implement models
- The pipelines to train algorithms are basically the same (to change model to be train is enough usually to change few or only a line)

Pipeline for implementing a Supervised Machine Learning Algorithm:

- preprocessing (i.e. putting everything lowercases, stemming, removing stopwords, etc)
- divide your dataset in training and test dataset

On the training dataset:

- creating the **matrix of token counts**

(has as columns all the words of our vocabulary and as rows the collection of texts we are analyzing, and each element in each row represents the frequency of the word in the column in that text (*CounterVectorizer* command in Python)).

- Applying the **Tfidf transformation** (see next slides)

Preprocessing

(here, I removed some stopwords, I put everything in lowercases and I stemmed words).

Original text	Pre-processed text
Kinda sounds like a RED WAVE to me, huh? Which is exactly what I predicted. I went against every lying fake news poll. They were all wrong. I was right.	kinda sound like red wave me huh exactli predict went everi ly fake new poll wrong right
Republicans Are Joining CNN's Ana Navarro For Meeting With Soros-Backed Amnesty Group	republican join cnn' ana navarro meet sorosback amnesti group
Too funny James Woods	funni jame wood

CounterVectorizer:

	and	beautiful	blue	is	king	love	old	queen	sky	the	this
0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0

Tfidf (term frequency-inverse document frequency)

- Weighting factor ----> reflect **how** important a word is to a document in a collection or corpus
- Increases proportionally to the number of times a word appears in the document
- Offset by the number of documents in the corpus that contain the word

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
 df_x = number of documents containing x
 N = total number of documents

	beautiful	beautiful blue	beautiful queen	blue	blue beautiful	blue sky	king	king old	love	love beautiful	old	old king	old queen	queen	queen beautiful	queen old	sky	sky blue
0	0.28	0.00	0.00	0.42	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.53
1	0.23	0.00	0.00	0.00	0.00	0.00	0.35	0.44	0.00	0.00	0.35	0.00	0.44	0.35	0.44	0.00	0.00	0.00
2	0.22	0.43	0.00	0.34	0.00	0.43	0.00	0.00	0.43	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00
3	0.23	0.00	0.44	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.35	0.44	0.00	0.35	0.00	0.44	0.00	0.00

(Continue) Pipeline for Supervised Machine Learning Learning:

- Select your model
- Fit your model on your training dataset
- Prepare your test set to be readable from the algorithm (i.e. apply the CounterVectorizer fitted on the data of the training dataset, and then apply tfidf on the matrix obtained)
- Predict your labels applying the algorithm on the test set
- Cross-validation
- Calculate metrics of algorithms performance based on your results on the test set
- Try other models, and see which one perform better

- I train an algorithm on the **pre-labelled** dataset

```
df['text_clean02'] = df['text'].apply(lambda x: clean_text02(x))

X = df['text_clean02']
y = df['hate_speech_dumm']

Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, random_state=1)

vectorizer = CountVectorizer()
Xtrain_counts = vectorizer.fit_transform(Xtrain)

tfidf_transformer = TfidfTransformer()
Xtrain_tfidf = tfidf_transformer.fit_transform(Xtrain_counts)

Xtest_counts = vectorizer.transform(Xtest)
Xtest_tfidf = tfidf_transformer.fit_transform(Xtest_counts)

model = LogisticRegressionCV(Cs=20, solver='liblinear', penalty='l1', cv=10)

model.fit(Xtrain_tfidf, ytrain)
y_model = model.predict(Xtest_tfidf)
```

Pipeline for Unsupervised Machine Learning Methods:

- Preprocessing
- CounterVectorized and Tfifd Transformation
- Choose your model (and the associated parameters)
- Apply your model to your data
- Analyze ("manually") the groups you obtained

4) AI Instruments for Researchers

Introduction: Large Language Models (LLM)

- trained deep-learning model that understand and generates text in a human-like fashion
- emerged around 2018
- general purpose model (they excel at a wide range of tasks, as opposed to being trained for one specific task – i.e. sentiment analysis, named entity recognition, etc)

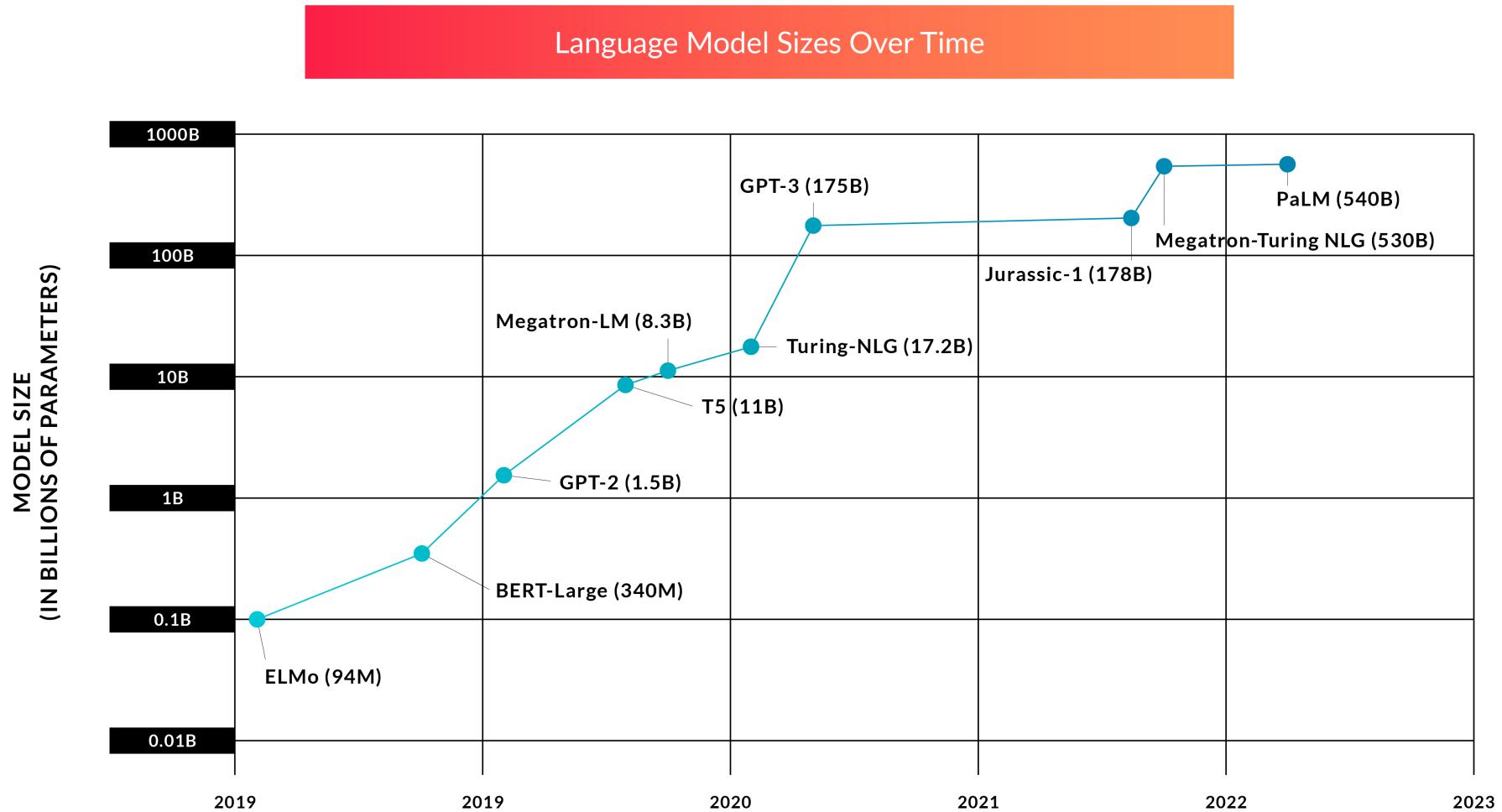
Large Language Models

Example of models:

- BERT (2018; Developer: Google; number of parameters: 340 millions; Corpus size: 3.3 billion words)
- GPT-2 (2019, Developer: OpenAI, n. parameters: 1.5 billions; corpus size: around 10 billions of words)
- GPT-3 (2020, Developer: OpenAI; n. parameters: 175 billions; corpus size: 300 billion of words)

Other famous models: Chinchilla, by DeepMind; LLaMa, by Meta

Large Language Models



The key technology here: Transformers

(paper "Attention is all you need", 2017)

BERT: Bidirectional Encoder

Representations from Transformers

GPT: Generative Pre-trained
Transformers

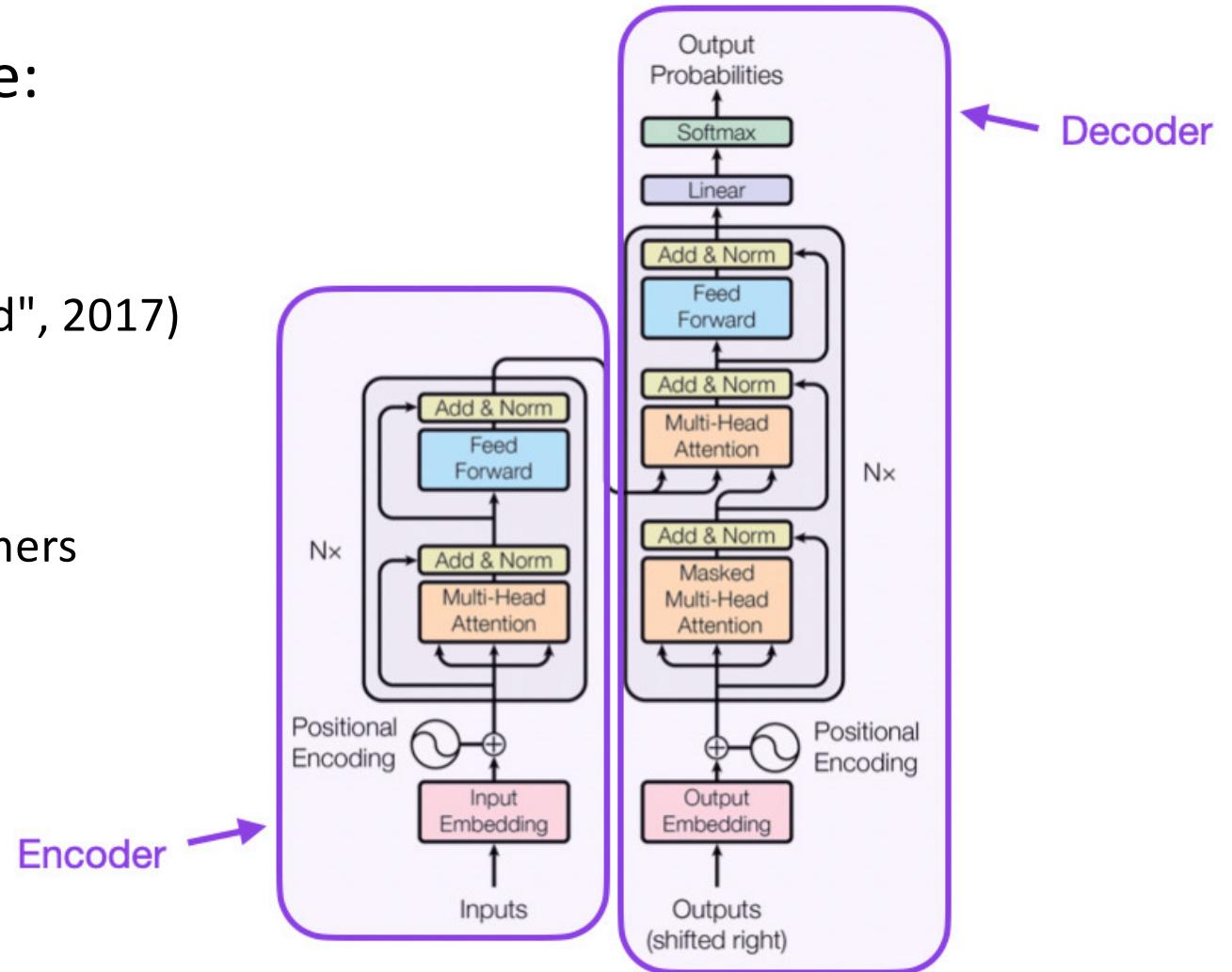


Figure 1: The Transformer - model architecture.

Paper "Language Models and Cognitive Automation for Economic Research" by Anton Korinek (Working Paper, 2023)

- Link: <https://www.nber.org/papers/w30957>
- Large language models (LLMs) such as ChatGPT have the potential to revolutionize research in economics and other disciplines.
- I describe 25 use cases along six domains in which LLMs are starting to become useful as both research assistants and tutors: ideation, writing, background research, data analysis, coding, and mathematical derivations.
- Economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive.

- ChatGPT, a version of OpenAI's GPT3.5 model that was released on November 28th, 2022,
- It gained more than 100m users in its first two months, and is now estimated to produce a volume of text every 14 days that is equivalent to all the printed works of humanity (Thompson, 2023).
- At present, I view LLMs to be most useful as assistants that can automate small micro tasks that researchers engage in numerous times during the day but that are too small to be assigned to human research assistants.

- Moreover, LLMs are also useful as tutors in coding and data analysis tasks as well as in ideation and writing. I posit that researchers can significantly increase their productivity by incorporating LLMs into their workflow.
- Ultimately, I believe that the most useful attitude towards the current generation of LLMs is to heed the lessons of comparative advantage that Ricardo taught us two centuries ago: LLMs increasingly have comparative advantage in generating content; humans currently have comparative advantage in evaluating and discriminating content. LLMs also have super-human capabilities in processing large amounts of text. All this creates ample space for productive collaboration, as we will explore throughout the remainder of the paper.

Category	Task	Usefulness
Ideation	Brainstorming	3
	Evaluating ideas	2
	Providing counterarguments	3
Writing	Synthesizing text	3
	Editing text	3
	Evaluating text	3
	Generating catchy titles & headlines	3
Background Research	Generating tweets to promote a paper	3
	Summarizing Text	3
	Literature Research	1
	Formatting References	3
Coding	Translating Text	3
	Explaining Concepts	2
	Writing code	2
	Explaining code	2
Data Analysis	Translating code	3
	Debugging code	2
	Extracting data from text	3
	Reformatting data	3
Math	Classifying and scoring text	2
	Extracting sentiment	2
	Simulating human subjects	2
	Setting up models	2
	Deriving equations	1
	Explaining models	1

The third column reports my subjective rating of LLM capabilities as of Feb 1, 2023:

1 = experimental; results are inconsistent and require significant human oversight

2 = useful; requires oversight but will likely save you time

3 = highly useful; incorporating these into your workflow will save you time

- **Brainstorming**

Please brainstorm economic channels through which advances in AI may increase inequality

1. Automation of low-skill jobs, leading to job displacement and increased unemployment
2. Increased demand for highly-skilled workers, leading to wage inequality
3. Increased access to capital for those with the resources to invest in AI technology
4. Increased surveillance of workers, leading to decreased privacy and potential exploitation
5. Increased use of AI-driven algorithms to make decisions about hiring, lending, and other economic activities, leading to potential bias and discrimination
6. Increased use of AI-driven algorithms to optimize pricing, leading to increased costs for consumers
7. Increased use of AI-driven algorithms to optimize marketing, leading to increased costs for businesses
8. Increased use of AI-driven algorithms to optimize investments, leading to increased returns for those with the resources to invest
9. Increased use of AI-driven algorithms to optimize tax avoidance, leading to decreased tax revenues for governments
10. Increased use of AI-driven algorithms to optimize labor costs, leading to decreased wages for workers

• Providing Counterarguments

My hypothesis is that AI will increase inequality. What are the main counterarguments?

1. AI can be used to increase access to education and healthcare, which can help reduce inequality.
2. AI can be used to automate mundane tasks, freeing up people to pursue more meaningful work.
3. AI can be used to increase efficiency and productivity, leading to higher wages and better job opportunities.
4. AI can be used to create new jobs and industries, providing more opportunities for people to earn a living.
5. AI can be used to improve decision-making processes, leading to fairer outcomes.
6. AI can be used to identify and address systemic biases, leading to more equitable outcomes.
7. AI can be used to reduce the cost of goods and services, making them more accessible to people with lower incomes.
8. AI can be used to improve the accuracy of predictions, leading to more equitable outcomes.

• Synthesizing text

examples and offer solutions.

Please write a paragraph that integrates the following arguments and goes beyond. Include a topic sentence.

- * large language models automate cognitive tasks
- * the work of economists is largely cognitive [add examples]
- * we economists should therefore worry about recent progress in language models
- * in the short run, these models can help us be more productive
- * in the long run, there is a good likelihood that they will displace human economists
- * we should think how to prepare [offer solutions]

The recent progress in large language models has the potential to automate many cognitive tasks, including those of economists. Economists are responsible for analyzing data, making predictions, and providing policy advice, all of which are cognitive tasks. Therefore, economists should be concerned about the implications of these language models, both in the short and long run. In the short run, these models can help us be more productive, but in the long run, there is a good likelihood that they will displace human economists. To prepare for this eventuality, economists should think about how to use these models to their advantage and how to develop skills that are not easily automated. This could include developing expertise in areas such as communication, problem-solving, and critical thinking.

- **Editing text**

For example:

Can you correct the following sentence?

These paper describe six domain in which large language models (LLMs) **displays emergant capbilities** that are useful **to the** economic research.

This paper describes six domains in which large language models (LLMs) display emergent capabilities that are useful for economic research.

Can you explain the main corrections you made?

I changed "These" to "This" to make the sentence singular, and I changed "useful to" to "useful for" to make the phrase more grammatically correct.

- **Generating Catchy Titles**

Please suggest two or three catchy titles for a paper with the following abstract:

We analyze how to set up institutions that future-proof our society for a scenario of ever-more-intelligent autonomous machines that substitute for human labor and drive down wages. We lay out three concerns arising from such a scenario, culminating in the economic redundancy of labor, and evaluate recent predictions and objections to these concerns. Then we analyze how to allocate work and income if these concerns start to materialize. As the income produced by autonomous machines rises and the value of labor declines, we find that it is optimal to phase out work, beginning with workers who have low labor productivity and job satisfaction, since they have comparative advantage in enjoying leisure. This is in stark contrast to welfare systems that force individuals with low labor productivity to work. If there are significant wage declines, avoiding mass misery will require other ways of distributing income than labor markets, whether via sufficiently well-distributed capital ownership or via benefits. Recipients could still engage in work for its own sake if they enjoy work amenities such as structure, purpose, and meaning. If work gives rise to positive externalities such as social connections or political stability, or if individuals undervalue the benefits of work because of internalities, then there is a role for public policy to encourage work. However, we conjecture that in the long run, it would be more desirable for society to develop alternative ways of providing these benefits.

1. "Future-Proofing Society: Preparing for Autonomous Machines and the Decline of Labor"
2. "The End of Work? Navigating the Impact of Autonomous Machines on Labor"
3. "The Impact of Autonomous Machines on Labor: How to Allocate Work and Income"

- **Explaining Concepts**

Why are instrumental variables useful?

Instrumental variables are useful because they can help to identify causal relationships between variables. They are used when there is a correlation between the independent and dependent variables, but it is not clear which variable is causing the other. Instrumental variables can help to identify the causal relationship by introducing a third variable that is correlated with the independent variable, but not with the dependent variable. This allows researchers to isolate the effect of the independent variable on the dependent variable.

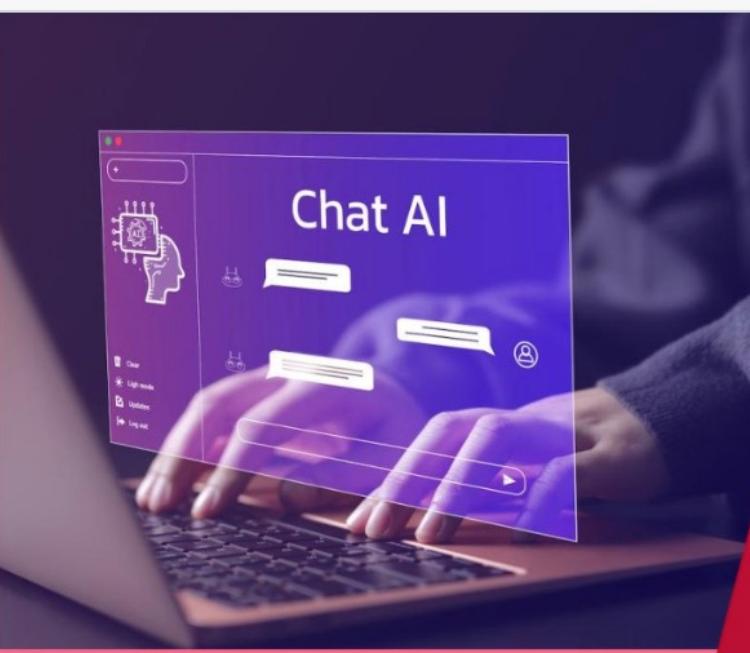


VoxEU ✅ @voxeu · 8 giu

...

Experiment shows #ChatGPT significantly boosts #productivity, improving speed and quality in mid-level professional #writing tasks while bridging the gap between writers' abilities.

@shakked_noy, @whitneywzhang @MITEcon
ow.ly/e54w50OIVE7



The productivity effects of generative artificial intelligence

Shakked Noy,
Whitney Zhang

This column discusses how automation technologies, including generative AI like ChatGPT, have historically transformed various job sectors, and presents an experiment demonstrating the significant positive impact of ChatGPT on mid-level professional writing tasks

VOX^{EU} / CEPR

- **Writing Code**

```
# Python code to calculate the Fibonacci numbers

def Fibonacci(n):
    if n<0:
        print("Incorrect input")
    # First Fibonacci number is 0
    elif n==1:
        return 0
    # Second Fibonacci number is 1
    elif n==2:
        return 1
    else:
        return Fibonacci(n-1)+Fibonacci(n-2)

# Driver Program

print(Fibonacci(9))
```

- **Explaining Code**

[continued from previous example]

Explain the code above:

The code above is a function that calculates the Fibonacci numbers. The function takes an integer n as an argument and returns the n th Fibonacci number. The function uses recursion to calculate the Fibonacci numbers. The base cases are when n is 1 or 2, in which case the function returns 0 or 1 respectively. For any other value of n , the function returns the sum of the two previous Fibonacci numbers.

- **Translating Code**

[continued from the example in 3.4.1]

Translate the function above from Python to Matlab

```
function F = Fibonacci(n)
    if n < 0
        disp('Incorrect input')
    elseif n == 1
        F = 0;
    elseif n == 2
        F = 1;
    else
        F = Fibonacci(n-1) + Fibonacci(n-2);
    end
end
```

- **Extracting Data from Text**

Mark got an A in econ and a B+ in math. Sally got an A- in both econ and math. Frank got a B in econ and a C in math.

Reformat as follows:

Name & econ grade & math grade

Mark & A & B+

Sally & A- & A-

Frank & B & C

• Classifying and Scoring Text

Here are 5 tasks from O*Net:

1. Monitor operational records and reports to ensure accuracy and efficiency.
2. Develop and implement policies and procedures for recordkeeping.
3. Prepare and maintain employee records.
4. Maintain and update filing, inventory, mailing, and database systems.
5. Compile and prepare reports, charts, and graphs.

Please classify each of these tasks into easy or hard to automate. Provide your output in the form:

number; task; classification; five-word justification

1; Monitor operational records and reports to ensure accuracy and efficiency; Hard to Automate; Requires Human Judgment/Decision-Making

2; Develop and implement policies and procedures for recordkeeping; Hard to Automate; Requires Human Judgment/Decision-Making

3; Prepare and maintain employee records; Hard to Automate; Requires Human Judgment/Decision-Making

4; Maintain and update filing, inventory, mailing, and database systems; Easy to Automate; Repetitive and Structured Tasks

5; Compile and prepare reports, charts, and graphs; Easy to Automate; Repetitive and Structured Tasks

- **Simulating Human Subjects**

For each of the following registered voters, would they approve or disapprove of a policy proposal loosen US immigration restrictions on well-educated workers?

1. A 32-year old female lawyer in Houston, TX who is Hispanic
 2. A 73-year old male retired doctor in Rome, GA who is Caucasian
 3. A 25-year old software engineer in San Francisco, CA who is Asian and identifies as non-binary
 4. A 45-year old female grocery check-out clerk in Brooklyn, NY who is second-generation Italian
 5. A 55-year old insurance agent in Lake City, FL who is Caucasian
-
1. Approve
 2. Disapprove
 3. Approve
 4. Approve
 5. Disapprove

Mathematical Derivations

- Setting up Models
- Deriving Equations
- Explaining Models

Limitations:

Let me describe a few important limitations of this particular LLM that potential users should be aware of.

- The system is trained on data that cuts off in 2021, so it has no knowledge of more recent events.
- It cannot access the Internet – the text it generates is based solely on the parameters acquired during the training process.
- Moreover, it has no memory so information does not carry over from one session to the other.
- The amount of text that it can process is limited to 4000 tokens per request, corresponding to about 3000 words, with the limit applying to the sum of the user prompt and the completion.
- Furthermore, note that the results generated by an LLM change depending on the prompt – even small changes in prompts, such as different spacing or punctuation, can lead to completely different outputs. This makes it important for users to experiment with different prompts and to learn how to optimally engineer their prompts.
- Finally, let me add a reminder that ultimate responsibility for any output produced by LLMs always rests with the human user.

Concluding Remarks of the paper:

- In the short term, cognitive automation via LLMs will allow researchers to become significantly more productive. I expect that a growing number of researchers will incorporate LLMs into their workflow. This could help to increase the overall speed of progress in economics, although it risks leaving behind those who do not take advantage of LLMs.
- In the medium term, I anticipate that LLM-based assistants and tutors will become increasingly useful for generating the content that makes up research papers. Human researchers will focus on their comparative advantage by posing the questions, suggesting directions for obtaining answers, discriminating which parts of the produced content are useful, editing, and providing feedback, akin to an advisor. Moreover, they will also continue to play an important role in organizing research efforts for example, by coordinating teams and procuring data sources, akin to a research manager.

Concluding Remarks of the paper:

[.....]

At this point, human researchers, especially when AI-assisted, are still the best technology around for generating economic research!

Machine Learning



what society thinks I
do

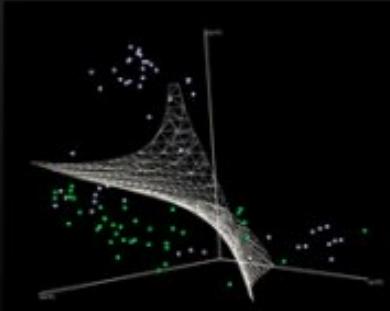


what my friends think
I do



what my parents think
I do

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^n \alpha_i \\ \alpha_i &\geq 0, \forall i \\ \mathbf{w} &= \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^n \alpha_i y_i = 0 \\ \nabla g(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t). \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\ \mathbb{E}_{i(t)} [\ell(x_{i(t)}, y_{i(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t). \end{aligned}$$



>>> from sklearn import svm

what other programmers
think I do

what I think I do

what I really do

Thank you for your Attention!!

**If you have any question, feel free to catch me at the PhD room or
write me an email! ;) (marina.rizzi@carloalberto.org)**

Let's now go to the notebook to see concretely how these steps are implemented.... :)

References

- Ash, E., Chen, D. L., & Ornaghi, A. (2021). **Gender attitudes in the judiciary: Evidence from US circuit courts.** (R&R at *American Economic Journal: Applied Economics*)
- Cagé, J., Hervé, N., & Viaud, M. L. (2020). **The production of information in an online world: Is copy right?** *The Review of Economic Studies*, 87(5)
- Hansen, S., Ramdas, T., Sadun, R., & Fuller, J. (2021). **The demand for executive skills** (No. w28959). National Bureau of Economic Research.
- Python Data Science Handbook (by Jake VanderPlas, <https://jakevdp.github.io/PythonDataScienceHandbook/>)
- Sockin, J. (2021). **Show Me the Amenity: Are Higher-Paying Firms Better All Around?**. Available at SSRN.