https://ttpoll.com/p/782763

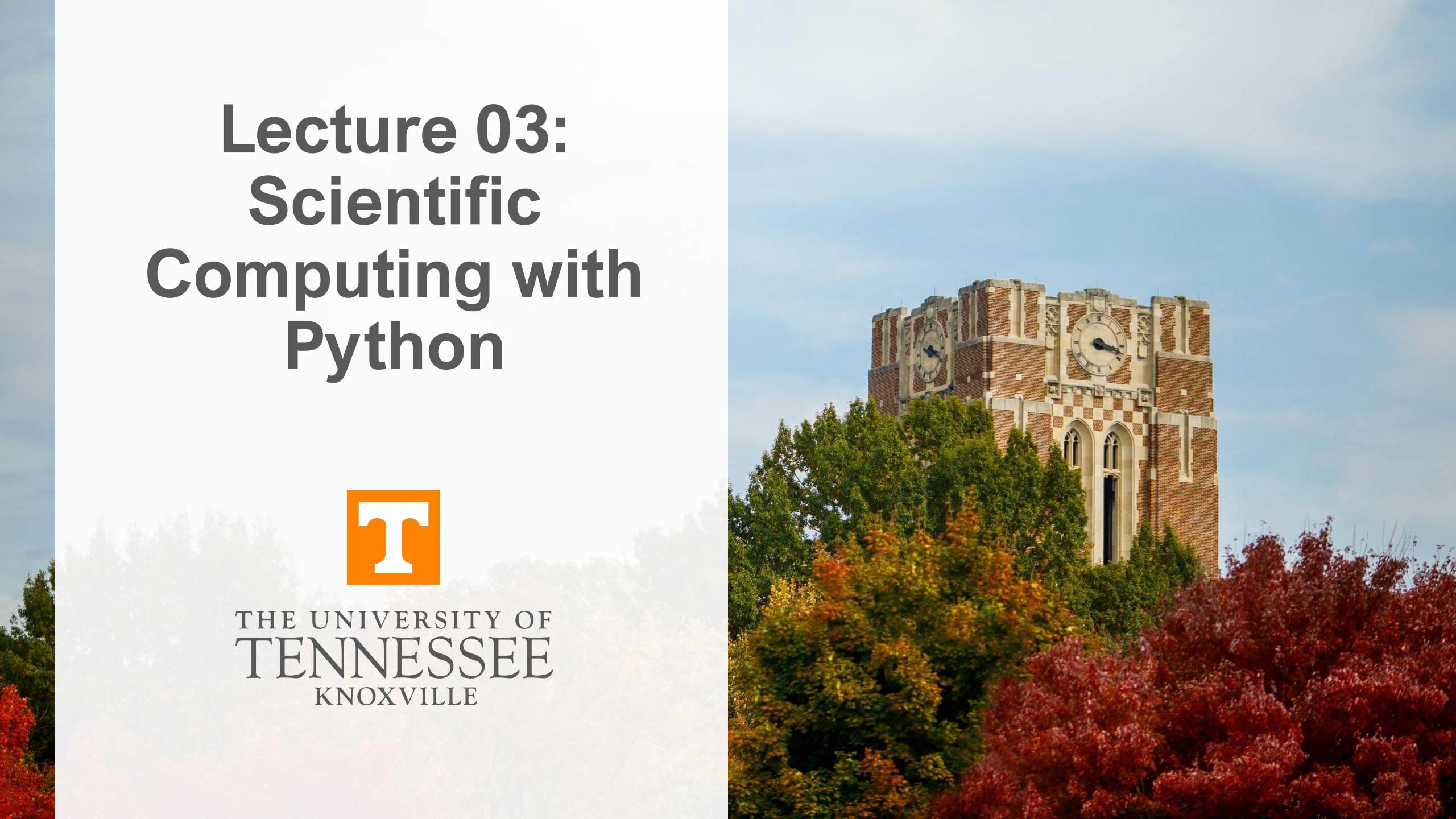
COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

Dr. Santos

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Lecture 03: Scientific Computing with Python



THE UNIVERSITY OF TENNESSEE KNOXVILLE

https://ttpoll.com/p/782763

Class Announcements

Homework:

First homework due Sunday 09/01.
The due date may shift according to the material covered.
TAs will not troubleshoot your code.

Quizzes:

Office hours question out of date.

Course Project:

Pick teammates by 08/29

https://ttpoll.com/p/782763

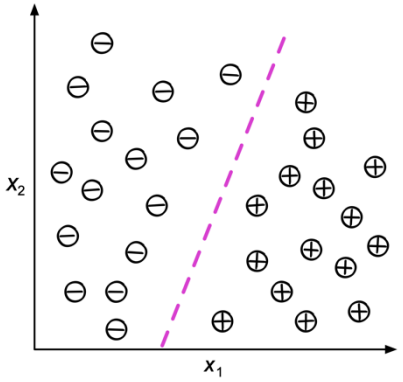
Today's Topics

Notation

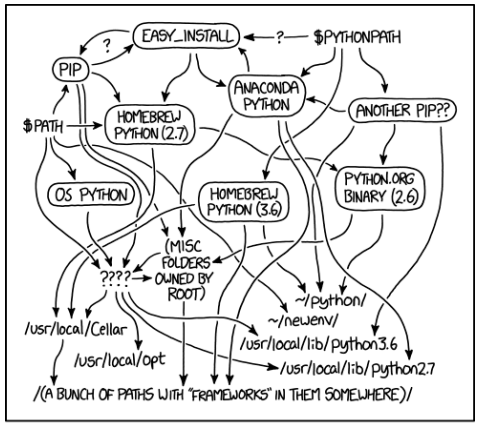
Scientific Computing in Python

https://ttpoll.com/p/782763

Last Lecture



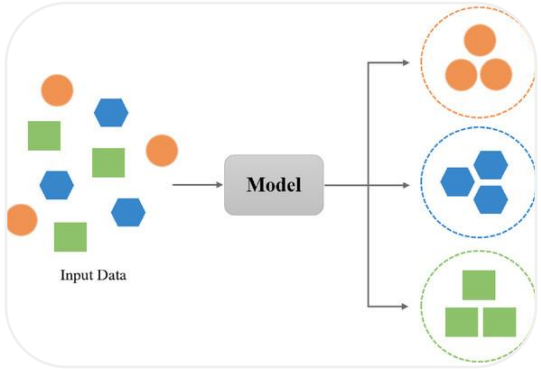
- Machine learning
  - A subfield of artificial intelligence
  - Models need to generalize (i.e., learn)
  - Task, Experience, Performance
  - Different learning categories
- Programming
  - Python: flexible, efficient, collaborative, powerful
  - Always work from a dockerized or virtual environment
  - Practice



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

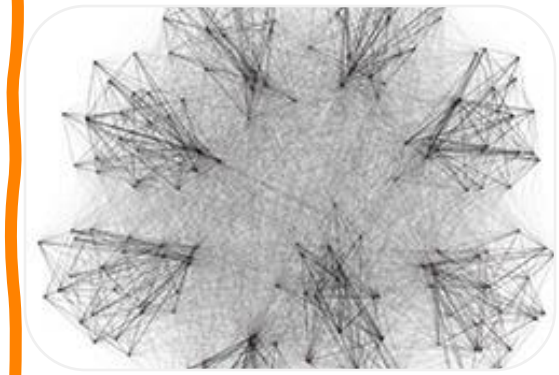
https://ttpoll.com/p/782763

Machine Learning Categories



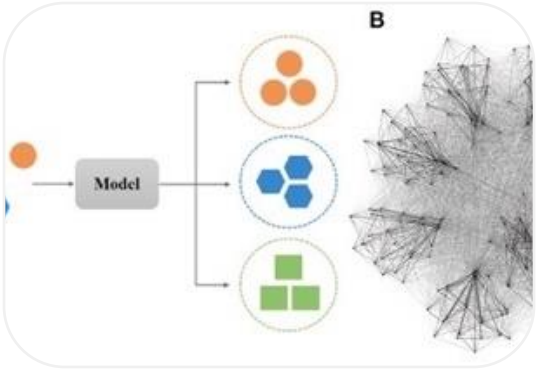
Supervise Learning

- Trained on “Labeled dataset”
- Needs pairs of inputs and outputs (ground truth)



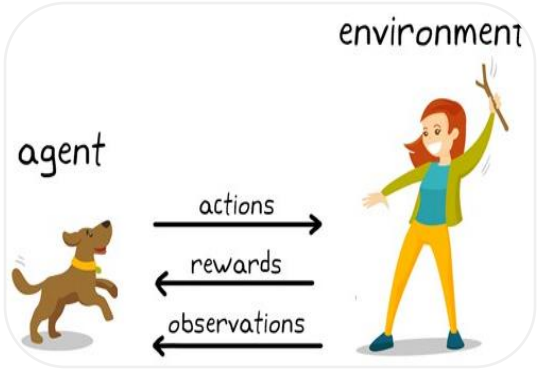
Unsupervised Learning

- The algorithm discovers patterns and relationships in unlabeled data
- It needs inputs only and, most of the time, some context. (e.g., number of unique labels)



Semi-Supervised Learning

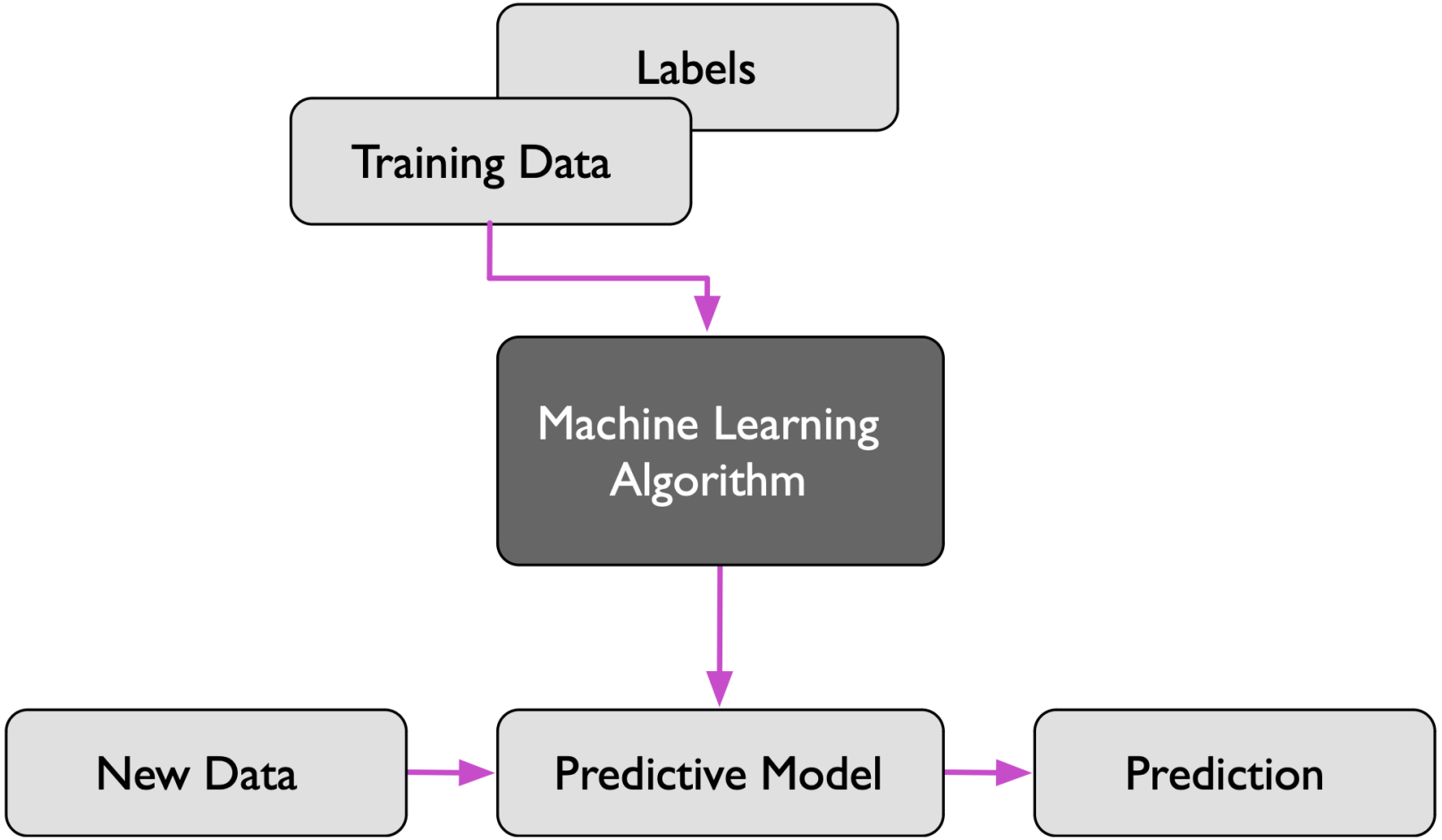
- Combines SL and UL
- E.g., a small subset of labeled data is used to label unlabeled data.
- E.g., Generative Adversarial Network mapping blonde to brunette.



Reinforcement Learning

- It learns by interacting with the environment.
- Trial, error, and delay
- Needs well-defined reward feedback.

Supervised Learning Workflow



Slide credit: Dr. Raschka

https://ttpoll.com/p/782763

Notation

Features

Targets/Labels

n training examples

Training set: $D = \{\langle x^{(i)}, y^{(i)} \rangle, i = 1, \ldots, n\}$

Training sample #3: $\langle x^{(3)}, y^{(3)} \rangle$

This is the function/program our ML algorithm will generate.

Unknown function: $f(x) = y$

Hypothesis: $h(x) = \hat{y}$

Classification

$h: \mathbb{R}^m \rightarrow$ _____

Features

Binary

{0,1}

Multi-class

{0,1, 3, 4, 5}

{‘a’, ‘q’, ‘z’}

Slide credit: Dr. Raschka

https://ttpoll.com/p/782763

Notation

Features

Targets/Labels

n training examples

Training set: $D = \{\langle x^{(i)}, y^{(i)} \rangle, i = 1, \ldots, n\}$

Training sample #3: $\langle x^{(3)}, y^{(3)} \rangle$

This is the function/program our ML algorithm will generate.

Unknown function: $f(x) = y$

Hypothesis: $h(x) = \hat{y}$

Classification

Regression

$h: \mathbb{R}^m \rightarrow \{0,1, 3, 4, 5\}$

$h: \mathbb{R}^m \rightarrow \mathbb{R}$

Features

Slide credit: Dr. Raschka

Inputs Representation

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector

$m = n_x$

House Market Price

$$x^{(i)} = \begin{bmatrix} Sqft\ size\ of\ house \\ zipcode \\ \vdots \\ no.bathrooms \\ no.bedrooms \end{bmatrix}$$

List of floating or integer numbers

Inputs Representation

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector

Dog detector

Input $x^{(i)}$

Slide credit: Dr. Raschka

Input Notation

Input $x^{(i)}$

100

100

Rows

Columns

Image Dimensions

2D matrix $x^{(i)}$

0	5	5	0
25	4	30	120
135	130	200	230
145	160	210	250

...

Unroll

$$x^{(i)} = \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \\ \vdots \\ 25 \\ 4 \\ 30 \\ \vdots \end{bmatrix}$$

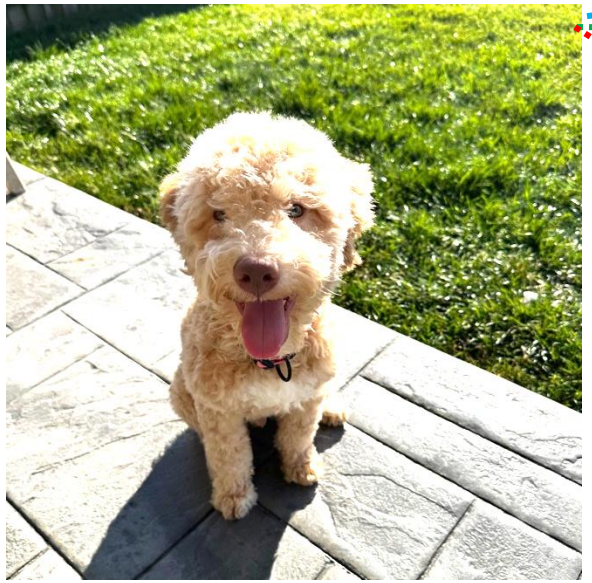
$m = n_x = 100*100$
$=10{,}000$

Number of features

https://ttpoll.com/p/782763

What about color?

Input $x^{(i)}$



100

100

RGB Matrices

10	15	5	0
25	4	30	120
135	130	200	230
145	160	210	250

R

255	255	255	255
255	255	255	30
255	255	255	30
255	255	255	2

G

0	2	15	25
40	40	40	123
40	40	40	40
40	40	40	40

B

Unroll

$$x^{(i)} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \\ \vdots \\ 255 \\ 255 \\ \vdots \\ 0 \\ 2 \\ \vdots \end{bmatrix}$$

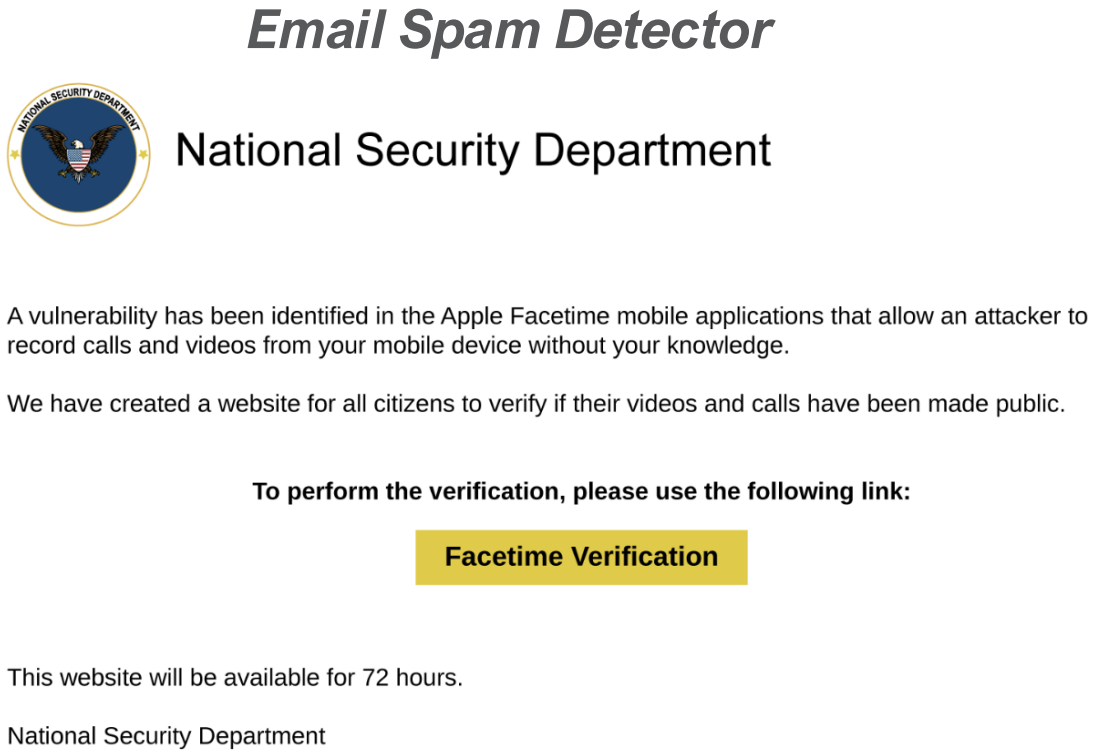
$m = 3*100*100$

$=30{,}000$

What about text?

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector



Check this tutorial about a spam email classifier with logistic regression: https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

Image source: https://www.terranovasecurity.com/blog/top-examples-of-phishing-emails

https://ttpoll.com/p/782763

What about text?

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector

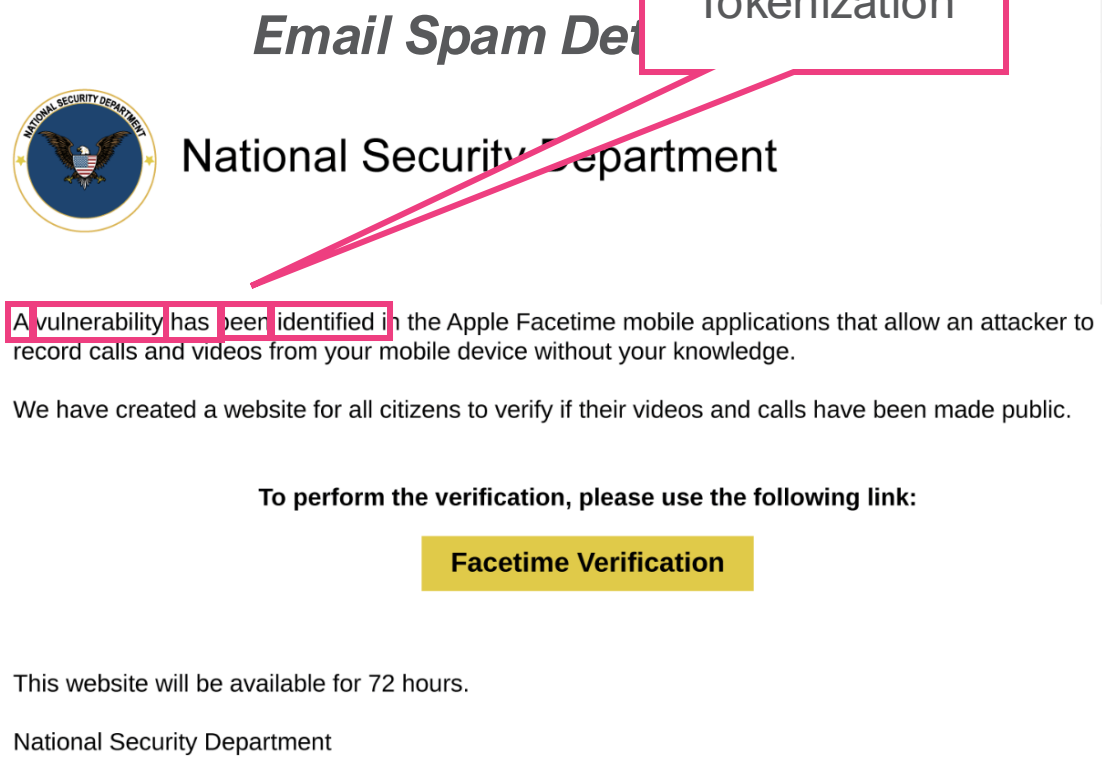
Tokens or terms

Tokenization

Frequency

Bag of Words

	identified	vulnerability	Urgent	Money	Transaction	Pay	link	Password	Verify	Hack
Doc 1	1	1	0	0	0	0	1	0	3	1
Doc 2	0	2	1	0	1	0	1	1	1	1
Doc 3	0	0	1	2	1	2	1	0	0	0



Email Spam Det

National Security Department

A vulnerability has been identified in the Apple Facetime mobile applications that allow an attacker to record calls and videos from your mobile device without your knowledge.

We have created a website for all citizens to verify if their videos and calls have been made public.

To perform the verification, please use the following link:

Facetime Verification

This website will be available for 72 hours.

National Security Department

Check this tutorial about a spam email classifier with logistic regression: https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

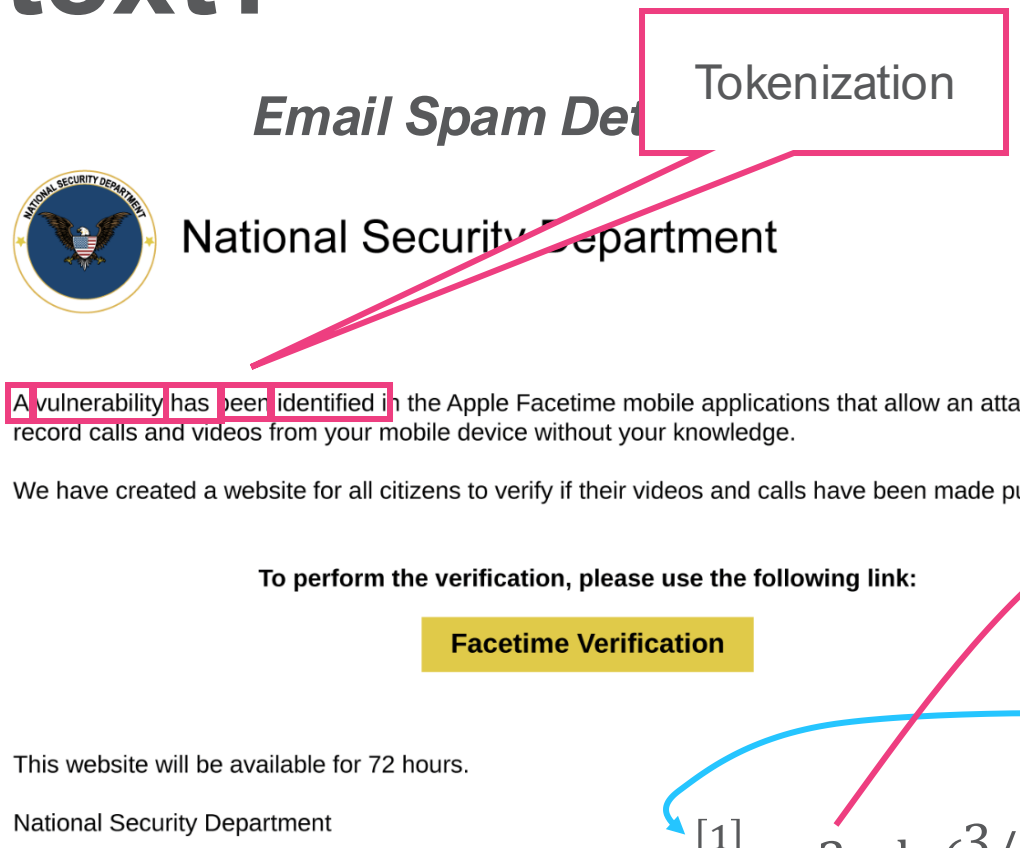
Image source: https://www.terranovasecurity.com/blog/top-examples-of-phishing-emails

https://ttpoll.com/p/782763

What about text?

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector



Tokenization

Bag of Words

	identified	vulnerability	Urgent	Money	Transaction	Pay	link	Password	Verify	Hack
Doc 1	1	1	0	0	0	0	1	0	3	1
Doc 2	0	2	1	0	1	0	1	1	1	1
Doc 3	0	0	1	2	1	2	1	0	0	0

Term Weighting

$$w_j^{[i]} = TF_j^{[i]} * \ln(n/n_j)$$

	identified	vulnerability	Urgent	Money	Transaction	Pay	link	Password	Verify	Hack
Doc 1									0.53	
Doc 2										
Doc 3										

$$w_9^{[1]} = 3 * \ln(3/2)$$

documents

documents with term

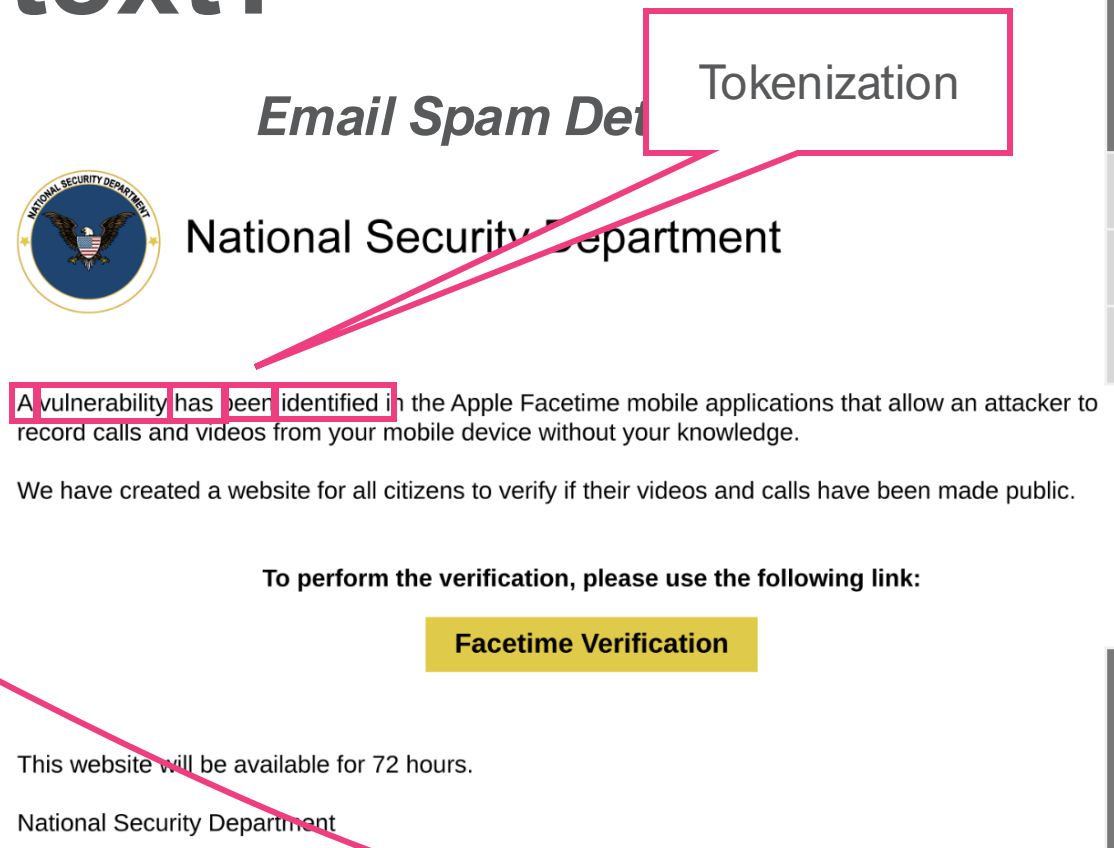
Image source: https://www.terranovasecurity.com/blog/top-examples-of-phishing-emails

https://ttpoll.com/p/782763

What about text?

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector



Tokenization

Bag of Words

	identified	vulnerability	Urgent	Money	Transaction	Pay	link	Password	Verify	Hack
Doc 1	1	1	0	0	0	0	1	0	3	1
Doc 2	0	2	1	0	1	0	1	1	1	1
Doc 3	0	0	1	2	1	2	1	0	0	0

Term Weighting

$$w_j^{[i]} = TF_j^{[i]} * \ln(n/n_j)$$

	identified	vulnerability	Urgent	Money	Transaction	Pay	link	Password	Verify	Hack
Doc 1	0.48	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.53	0.18
Doc 2	0.00	0.35	0.18	0.00	0.18	0.00	0.00	0.48	0.18	0.18
Doc 3	0.00	0.00	0.18	0.95	0.18	0.95	0.00	0.00	0.00	0.00

The document features

Image source: https://www.terranovasecurity.com/blog/top-examples-of-phishing-emails

https://ttpoll.com/p/782763

Representing the dataset

A sample: (x, y) $\quad x \in \mathbb{R}^m, \quad y \in \{0,1\}$

Input

Ground Truth Label

Set of Samples: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(n)}, y^{(n)})\}$

Matrix representation of inputs: $X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(n)T} \end{bmatrix}$ $\quad X \in \mathbb{R}^{n \times m}$ $\quad$ *Python X.shape* $= (n, m)$

Number of rows

Number of columns

Matrix representation of labels: $Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$ $\quad Y \in \mathbb{R}^{n \times 1}$

Y.shape $= (n, 1)$

https://ttpoll.com/p/782763

Data Representation

$$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature Vector

$$X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(n)T} \end{bmatrix}$$

Input matrix of size $\langle n, m \rangle$

What is the value of sample 13 feature #5?

$$x_5^{(13)}$$

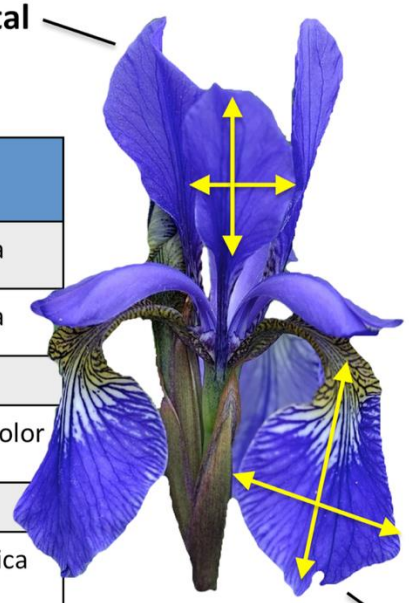
Slide credit: Dr. Raschka

Data Representation

$n_x =$ ____

$n =$ ____

	Sepal length	Sepal width	Petal length	Petal width	
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Slide credit: Dr. Raschka

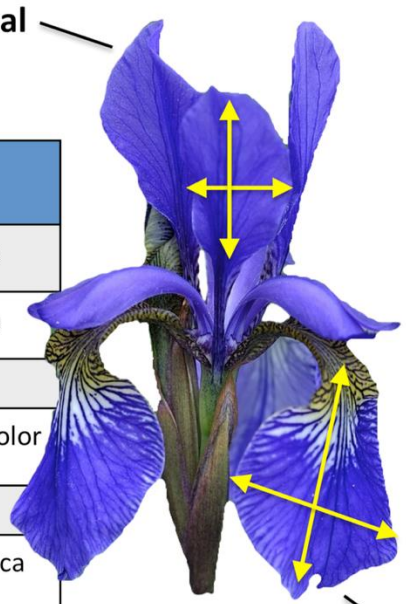
Data Representation

$n_x =$ 4

$n =$ 150

Samples

	Sepal length	Sepal width	Petal length	Petal width	
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features

Targets

Slide credit: Dr. Raschka

https://ttpoll.com/p/782763

Pop Quiz #1

1 | MULTIPLE CHOICE

POINTS: 1 | Edit

Below is an input matrix X with six samples and features. Following the notation discussed in class, what is sample's three feature vector?

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36

A. 21

B. [3, 9, 15, 21, 27, 33]

C. [13, 14, 15, 16, 17, 18]

D. [4, 10, 16, 22, 28, 34]

https://ttpoll.com/p/782763

Pop Quiz #1

1 | MULTIPLE CHOICE

POINTS: 1 | Edit

Below is an input matrix X with six samples and features. Following the notation discussed in class, what is sample's three feature vector?

$x^{(1)}$	1	2	3	4	5	6
$x^{(2)}$	7	8	9	10	11	12
$x^{(3)}$	13	14	15	16	17	18
$x^{(4)}$	19	20	21	22	23	24
$x^{(5)}$	25	26	27	28	29	30
$x^{(6)}$	31	32	33	34	35	36

A. 21

B. [3, 9, 15, 21, 27, 33]

C. [13, 14, 15, 16, 17, 18]

D. [4, 10, 16, 22, 28, 34]

https://ttpoll.com/p/782763

Note

- Notation can change
- Samples are also represented as x_i or $x^{[i]}$
  - Make sure you understand the notation used in a book or paper
- Dimensions of the X matrix can be different
  - E.g., if $X = \left[x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right]$, then, Y is a row vector.

Recap

- **Training example:** A row in the table representing the dataset. Synonymous to an observation, training record, training instance, training sample (in some contexts, sample refers to a collection of training examples)
- **Feature:** a column in the table representing the dataset. Synonymous to predictor, variable, input, attribute, covariate.
- **Targets/Labels:** What we want to predict. Synonymous to outcome, output, ground truth, response variable, dependent variable, (class) label (in classification).
- **Output/prediction:** use this to distinguish from targets; here, means output from the model.

Slide credit: Dr. Raschka

https://ttpoll.com/p/782763

Let’s talk again about lines.

https://ttpoll.com/p/782763

How do computers learn?

Housing Market Price Prediction

- Model parameters: ***a*** and ***b***
- Find the values for ***a*** and ***b*** that best fit the data.
  - E.g., linear regression

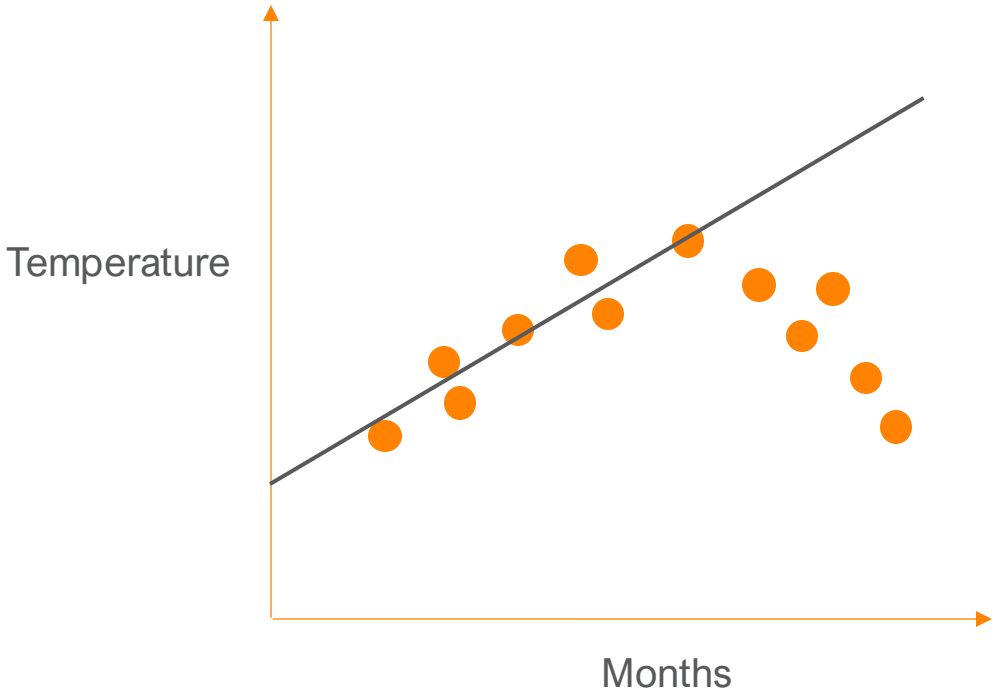
This is a machine-learning model!

https://ttpoll.com/p/782763

How do computers learn?

Housing Market Price Prediction

Season Temperature Prediction

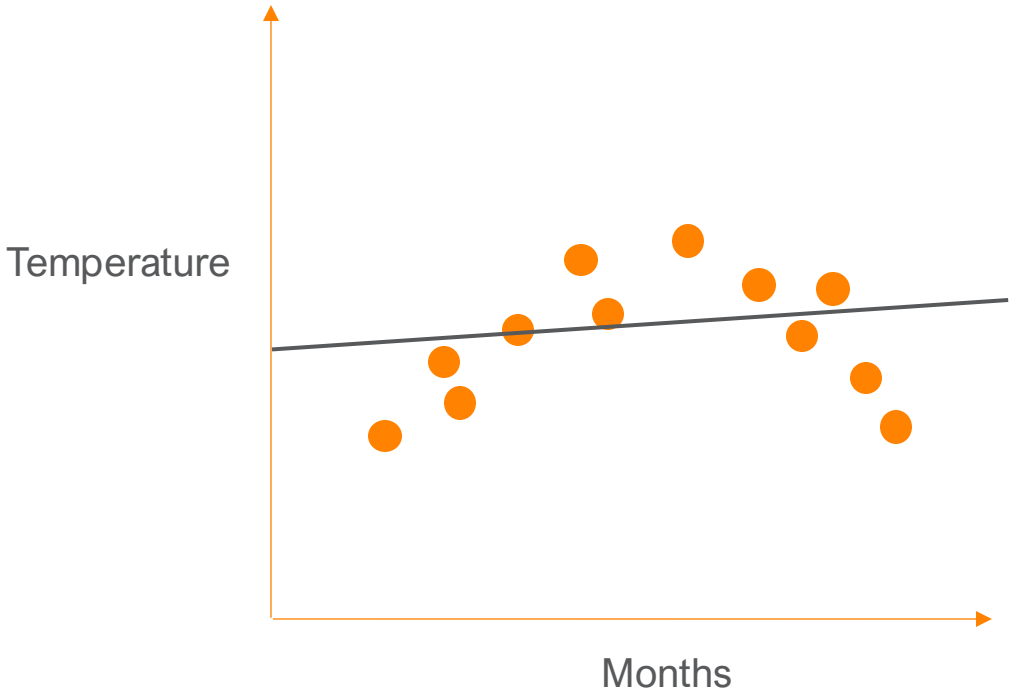


https://ttpoll.com/p/782763

How do computers learn?

Housing Market Price Prediction

Season Temperature Prediction

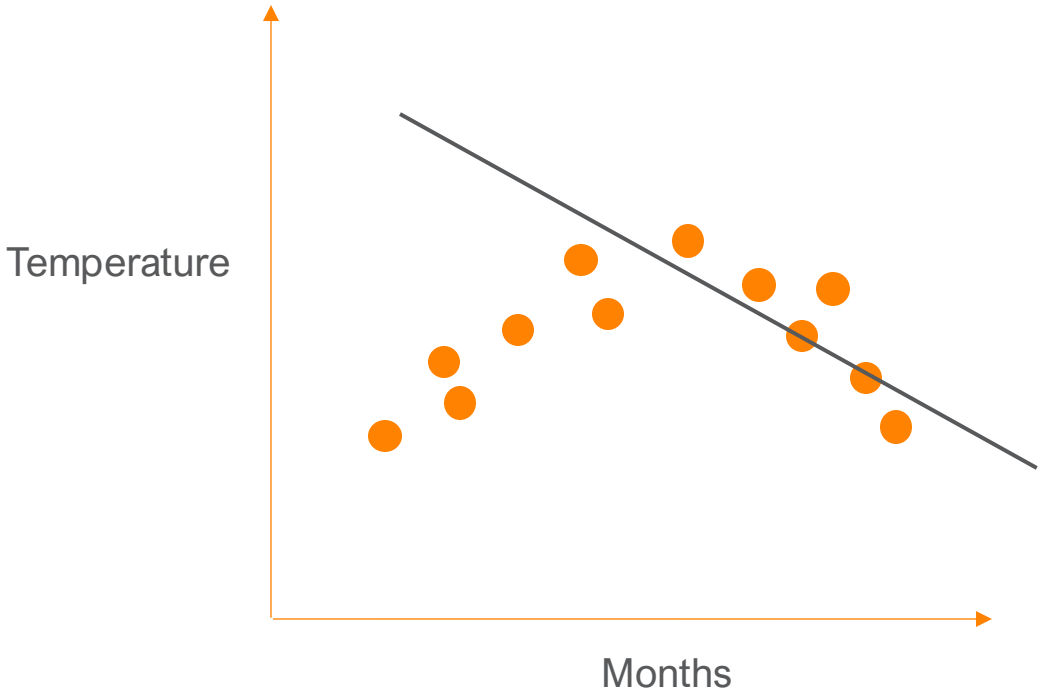


https://ttpoll.com/p/782763

How do computers learn?

Housing Market Price Prediction

Season Temperature Prediction

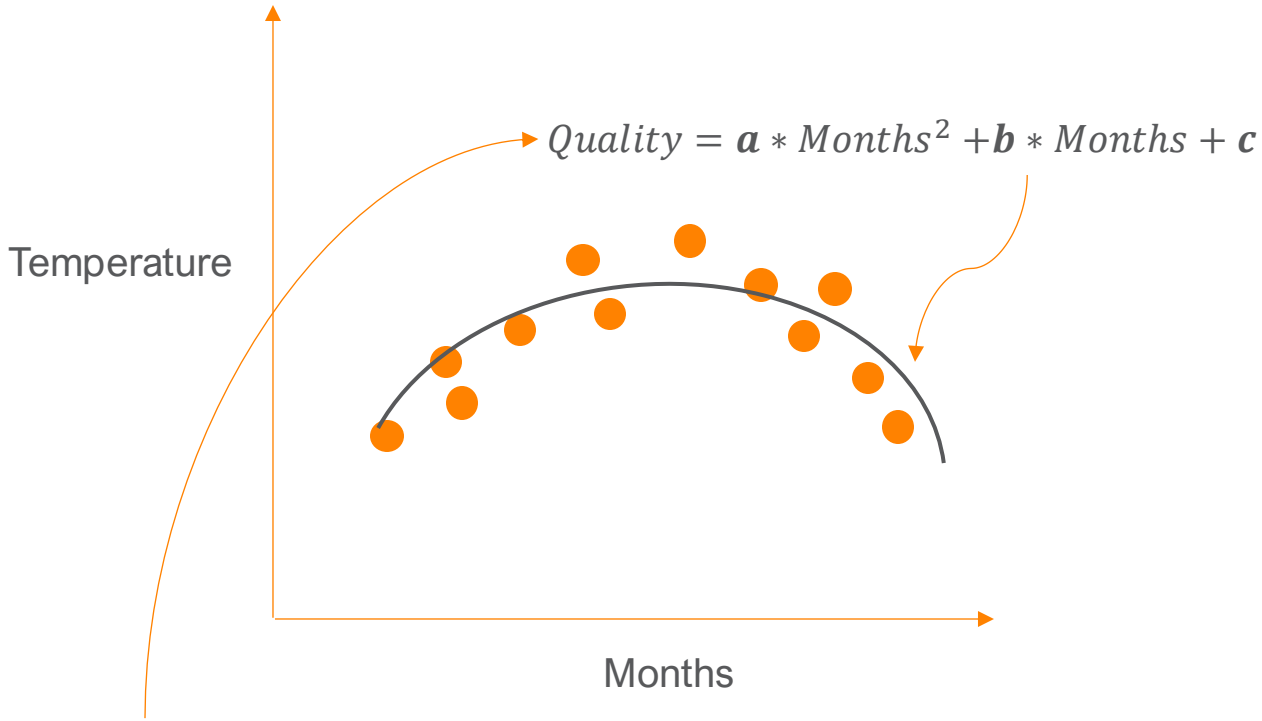


https://ttpoll.com/p/782763

How do computers learn?

Housing Market Price Prediction

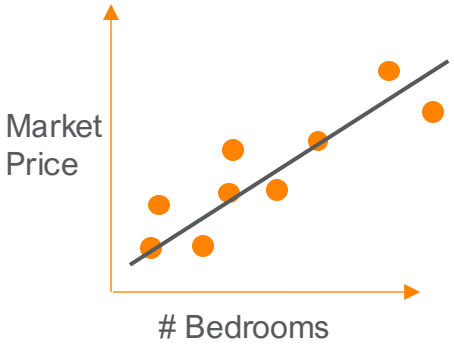
Season Temperature Prediction

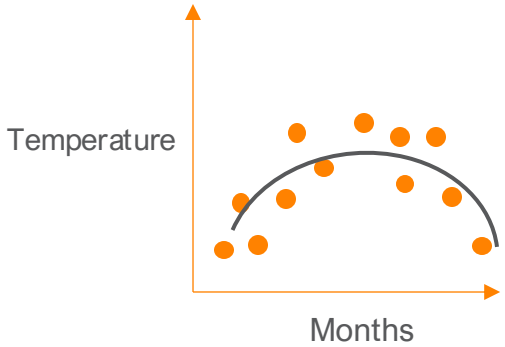


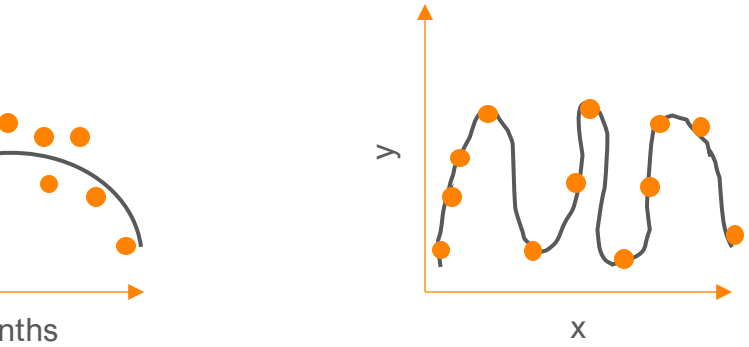
Find the values for parameters ***a, b, and c*** that best fit the data.

https://ttpoll.com/p/782763

Parameters and Capacity

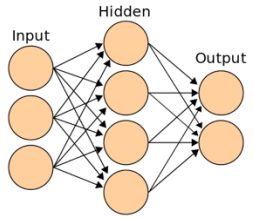






Price$= \boldsymbol{a} * Beds + \boldsymbol{b}$

$Quality = \boldsymbol{a} * Months^2 + \boldsymbol{b} * Months + \boldsymbol{c}$



As we increase the number of parameters…

Points to take home:

- The computer tries multiple curve parameters to search for the parameter set that best fits the data.
  - Minimizes or maximizes an objective function
- The more parameters, the better we can fit the data… more capacity.

https://ttpoll.com/p/782763

Numpy

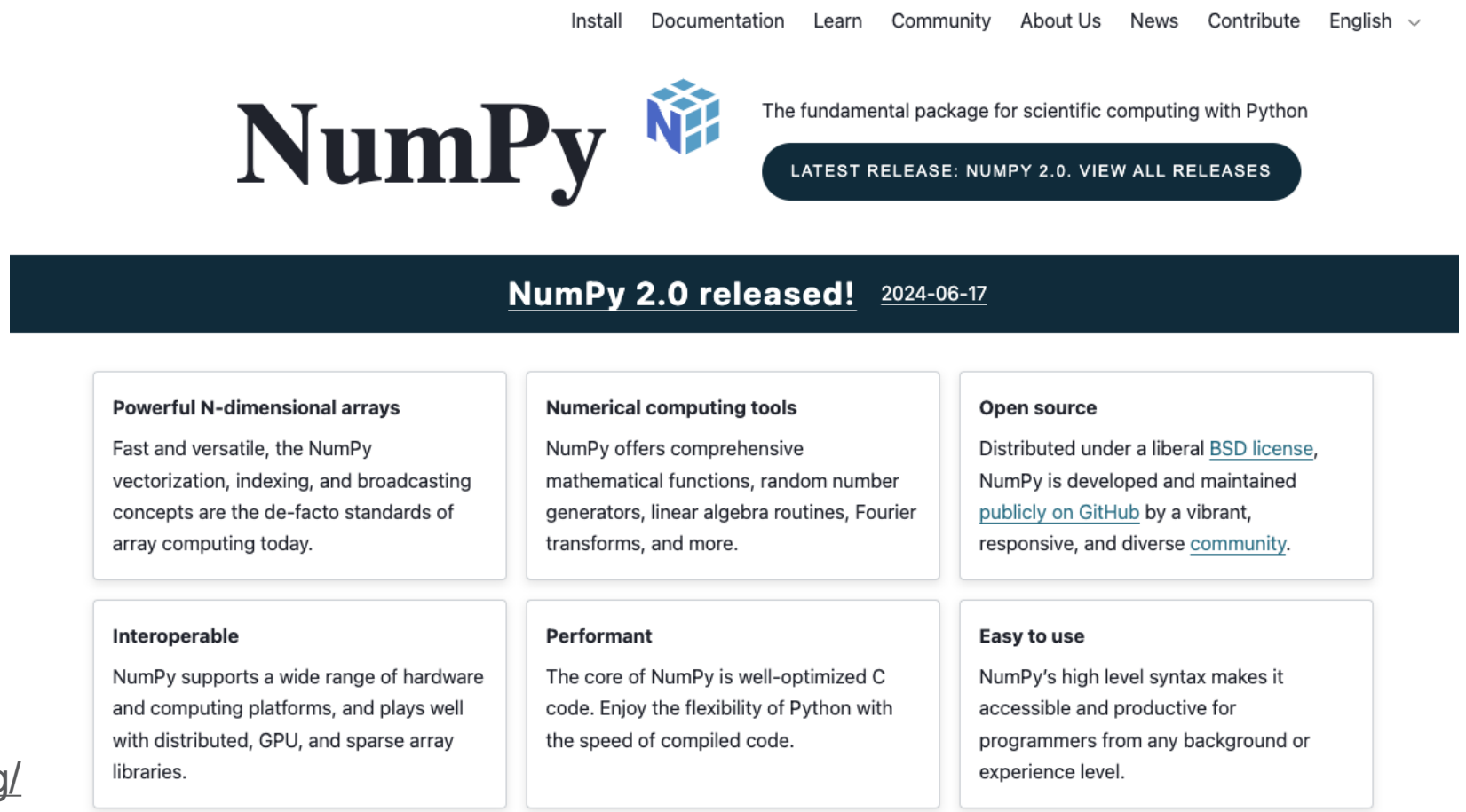
https://www.numpy.org

Numpy (https://numpy.org/)

- General-purpose open-source array-processing library
- High-performance N-dimensional array objects
  - Optimized C code
- Comprehensive built-in functions and random generators
  - Statistics, linear algebra, Fourier transform, etc.
- Fundamental package for scientific computing in Python

https://ttpoll.com/p/782763

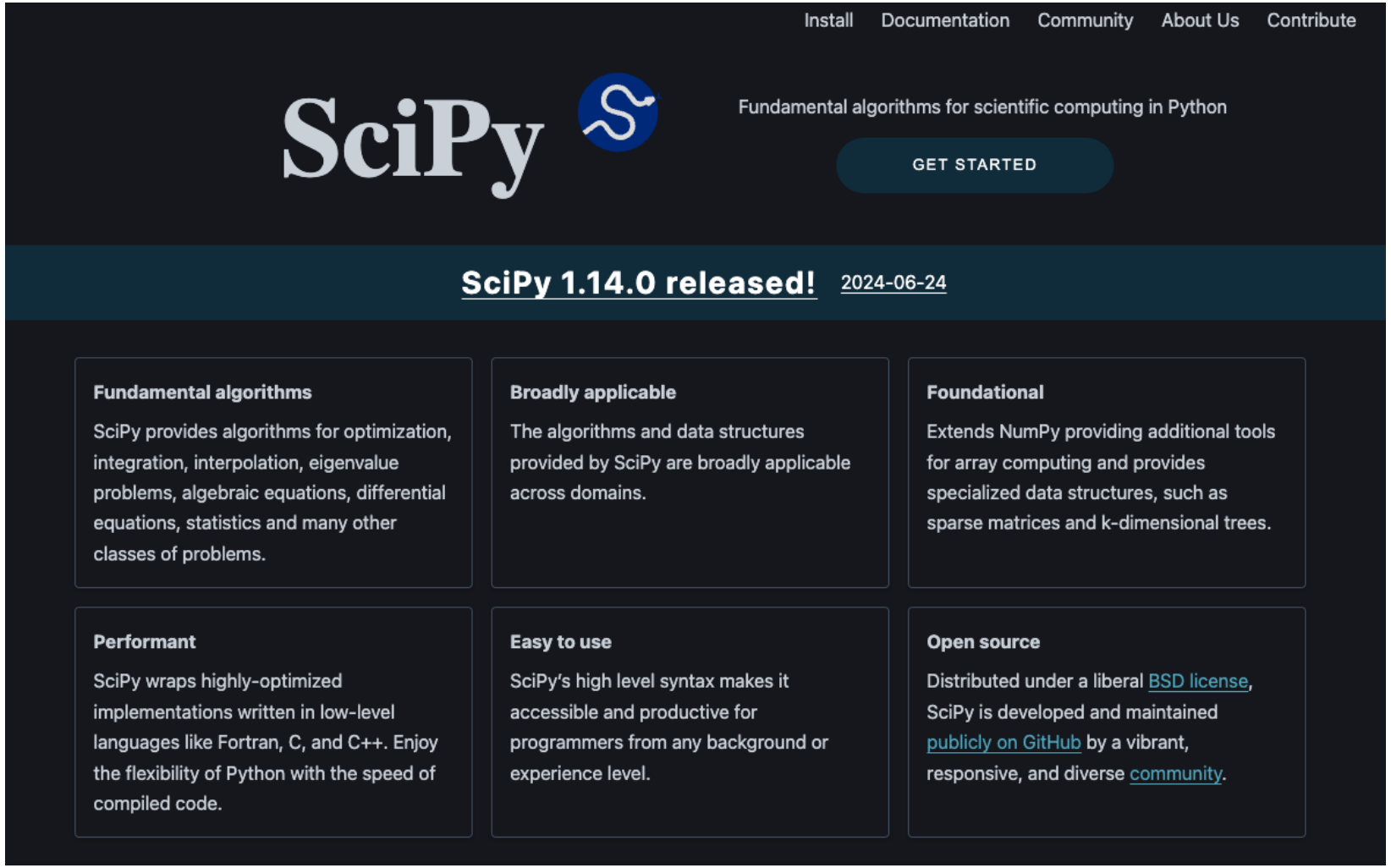
Numpy



https://numpy.org/

https://ttpoll.com/p/782763

Scipy



https://scipy.org/

https://ttpoll.com/p/782763

Notebook Time

https://ttpoll.com/p/782763

Next Lecture

- We will build our first machine learning pipeline/model with Scikit-Learn