COMP 5300 / 4600 Deep Learning for Natural Language Processing

Lecture 1

Class Information

Lecture:

Thu 3:30 – 6:15 *Olsen 405*

Instructor: Anna Rumshisky

email: arumshisky@gmail.com

Class website

https://text-machine-lab.github.io/dl4nlp-s2023/

Course website – slides, readings, schedule, assignments, etc.

Discord – class announcements, Q&A, etc.

Lectures recordings: Echo

Blackboard: Discord link, homework submission

Class Format

Each class will consist of

- Lecture (75 min)
- Practicum/Lab (75 min)

There will be a 10-minute break after the lecture.

During the Practicum/Lab segment of the class, we will focus on technical (coding) skills and provide homework guidance. You will be expected to work on your homework during this time.

Computing resources

- There is no textbook, but you will need to buy compute to complete some of the later homeworks (specifically, you would need to buy colab pro or colab pro+ - we will let you know).
- This will likely amount to $50-100, which is comparable to the cost of a hardcover textbook you need to buy for some other courses.
- If you are unable to do this, contact us and we'll try to help.
- Also, if you have access to a machine with a GPU, you will not need to buy compute.

Natural Language Processing (NLP)

Understanding, interpretation, generation of texts in natural language

What is NLP and why is it exciting?

Develop **computational models of human language**, which would be able to **interpret and generate free text.**

We want our models to be as good as humans or better at the myriad of language-related tasks

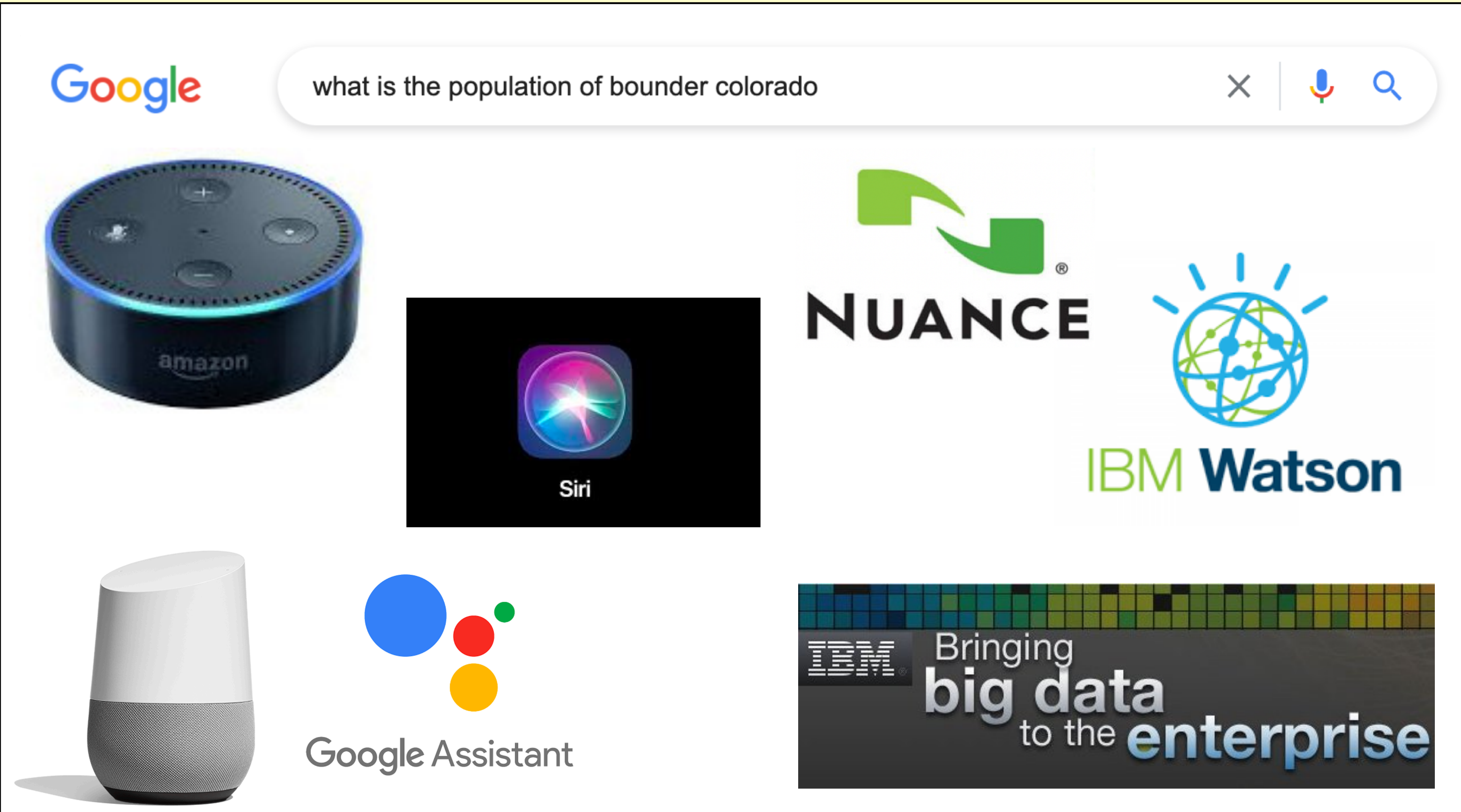
Comprehension, reasoning, inference; fluent dialog.

Extracting information, summarizing content, answering questions; translation between languages.

Humor, irony, metaphors, poetry, storytelling.

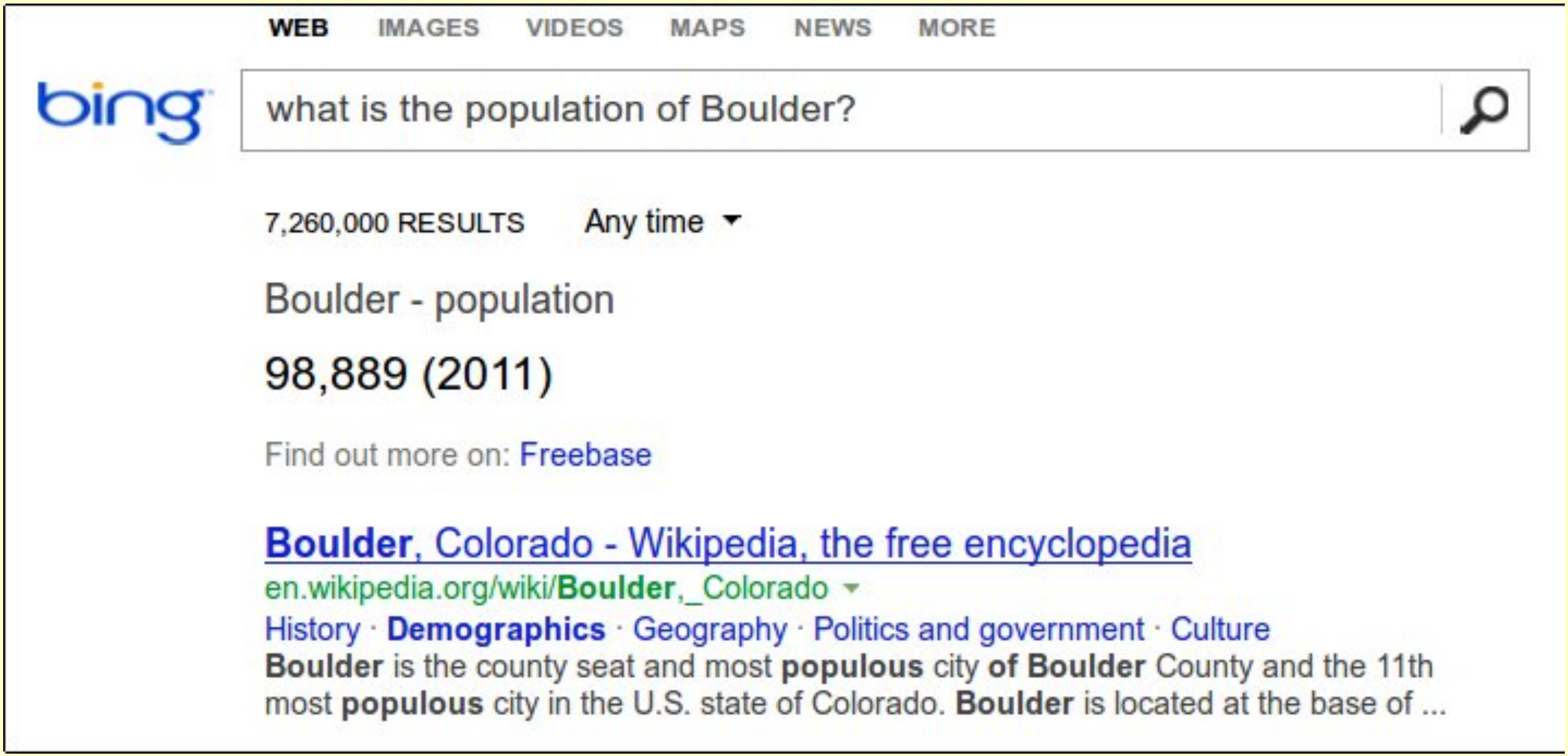
No general AI is possible without language understanding!

NLP Everywhere!



Google Translate

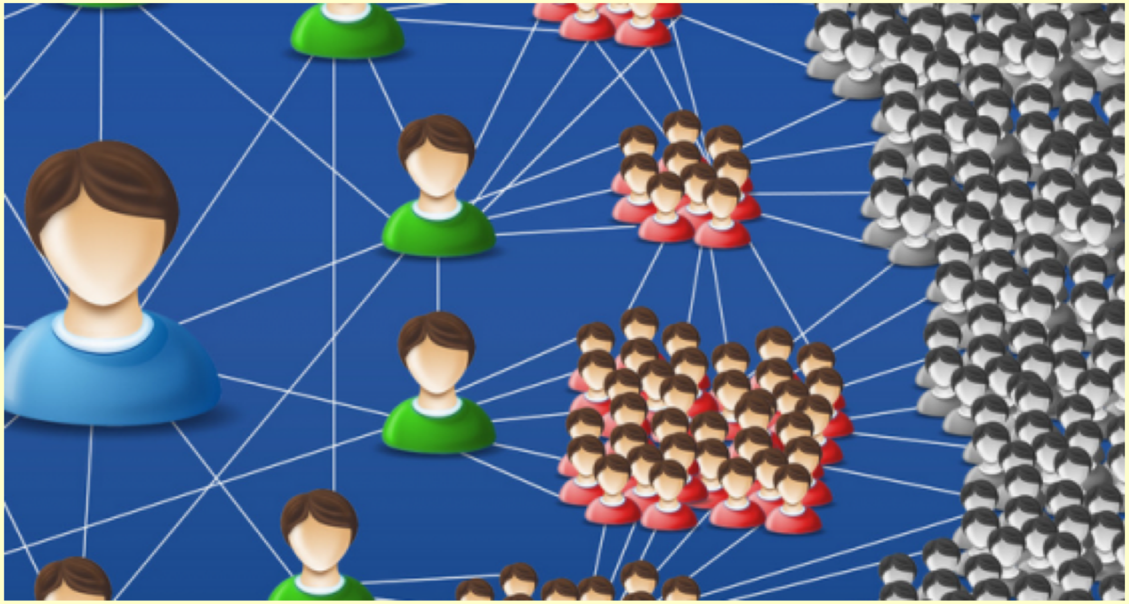
Web-based Question Answering



Data Mining of User-Generated Media

Data mining of social media websites, blogs, discussion forums, message boards, other forms of user-generated media

- Product marketing
- Political opinion tracking
- Social network analysis



"Social Media Monitoring, Analytics and Engagement"

Why is Language Hard?

Ambiguity

Find at least 5 meanings of the sentence

I made her duck

Ambiguity

Find at least 5 meanings of the sentence

I made her duck

I cooked waterfowl for her benefit (to eat)

I cooked waterfowl that belongs to her

I created a (plaster?) duck that she owns

I caused her to quickly lower her head or body

I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity

I caused her to quickly lower her head or body

Lexical category: “duck” can be a N or V

I cooked waterfowl belonging to her

Morphological category: “her” can be a possessive (“of her”) or a dative (“to her”)

I made the (plaster) duck she owns

Lexical semantics: “make” can mean “create” or “cook”

Lexical and Structural Ambiguity

Teacher Strikes Idle Kids

Kids Make Nutritious Snacks

British Left Waffles on Falkland Islands

Red Tape Holds Up New Bridges

Ban on nude dancing on Governor’s desk

Local high school dropouts cut in half

Factors Creating Ambiguity

Teacher Strikes Idle Kids / syntactic structure

Kids Make Nutritious Snacks / homonymy

British Left Waffles on Falkland Islands / homonymy

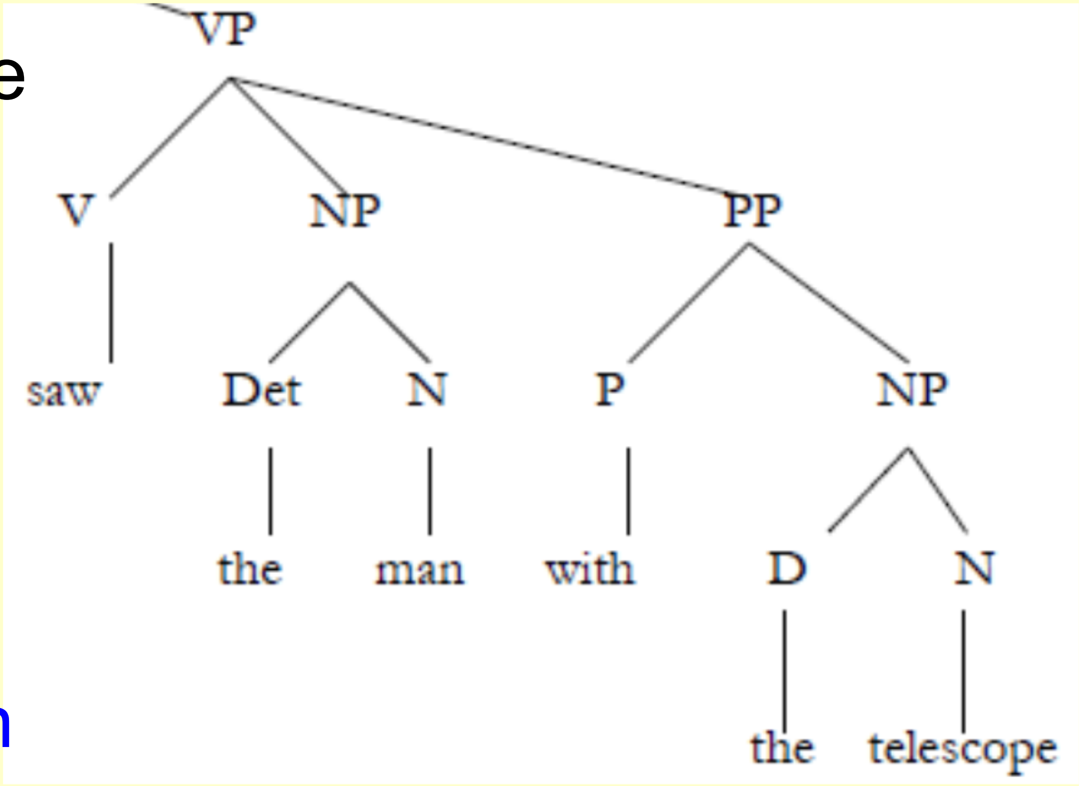
Red Tape Holds Up New Bridges / polysemy

Ban on nude dancing on Governor's desk / PP attachment

Local high school dropouts cut in half / polysemy

Can we appropriately resolve ambiguities?

"I saw the man with the telescope"



QA / Inference:

Did the speaker see the man through the telescope?

Classical Text Processing Tasks

Application Tasks

Machine translation

Information extraction: extracting entities, relations between them, events, their temporal ordering, etc.

Information retrieval:

query for “drugs for indigestion” produces pages with “medication” and “indigestion”

Question answering

Spoken dialog generation / chatbots / conversational agents

Text understanding, text generation, reasoning

Text summarization: summaries on search engine result pages

Natural language inference

Sentiment analysis/opinion mining

Information Extraction

Entity and Relation Extraction: Multi-sentence Template IE

10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automative account. Experience in online marketing, automative and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables ... Compensation: \$50,000-\\$80,000

INDUSTRY	Advertising
POSITION	Assist. Account Manag.
LOCATION	Irvine, CA
COMPANY	10th DEGREE

Entailment / Inference

Mary killed John

John is dead

=> ENTAILMENT

That’s how I would eventually meet my first girlfriend.

I didn’t meet my first girlfriend until later.

=> ENTAILMENT

someone else noticed it and i said well i guess that’s true and it was somewhat melodious in other words it wasn’t just you know it was really funny

no one noticed and it wasn’t funny at all

=> CONTRADICTION

Language Grounding

Visual Question Answering

What is the man holding?
Does it appear to be raining?
Does this man have 20/20 vision?

Dialogue Generation

Text Machine NLP Lab – Chatbot demo 2017

Methods: Text Processing

A Bit of NLP History…

Rule-based systems [1960s – 1980s]: directly encoding human knowledge about linguistic structure; patterns encoded by linguists.

– Finite state methods, rule-based context-free grammars, etc.

Statistical machine learning [1990s – 2000s]: knowledge about linguistic structure encoded in the form of features and patterns learned automatically from annotated corpora.

Sequence tagging with HMMs, CRFs; multinomial logistic regression, SVMs for classification; topic modeling with LDA; ngram language models for generation, etc.

The last 10 years (deep learning): data representations themselves no longer engineered by humans, but learned. Continuous-valued vectors serve as representations for input words

Multi-layer neural models: convolutional networks, recurrent neural networks (RNNs), attention-based architectures (Transformers); sequence-to-sequence models for classification, generation, etc.

Symbolic / Knowledge-Based

Analyze data by hand, generate

- Rules
- Dictionaries
- Thesauri
- Tagsets

E.g. Rule-based parsers

sets of grammar rules

S → NP VP NP → (NN|NNS)* NNS

NP → Det (JJ)* NN …

* Penn Treebank POS Tagset

– lexicalized rules linked to specific words and word classes

Supervised Machine Learning

Analyze data by hand, generate

- Annotation schemes (parse trees, dictionaries, tagsets)
- Annotated corpora (sets of texts)

Split corpora into training and test sets

Develop a feature set

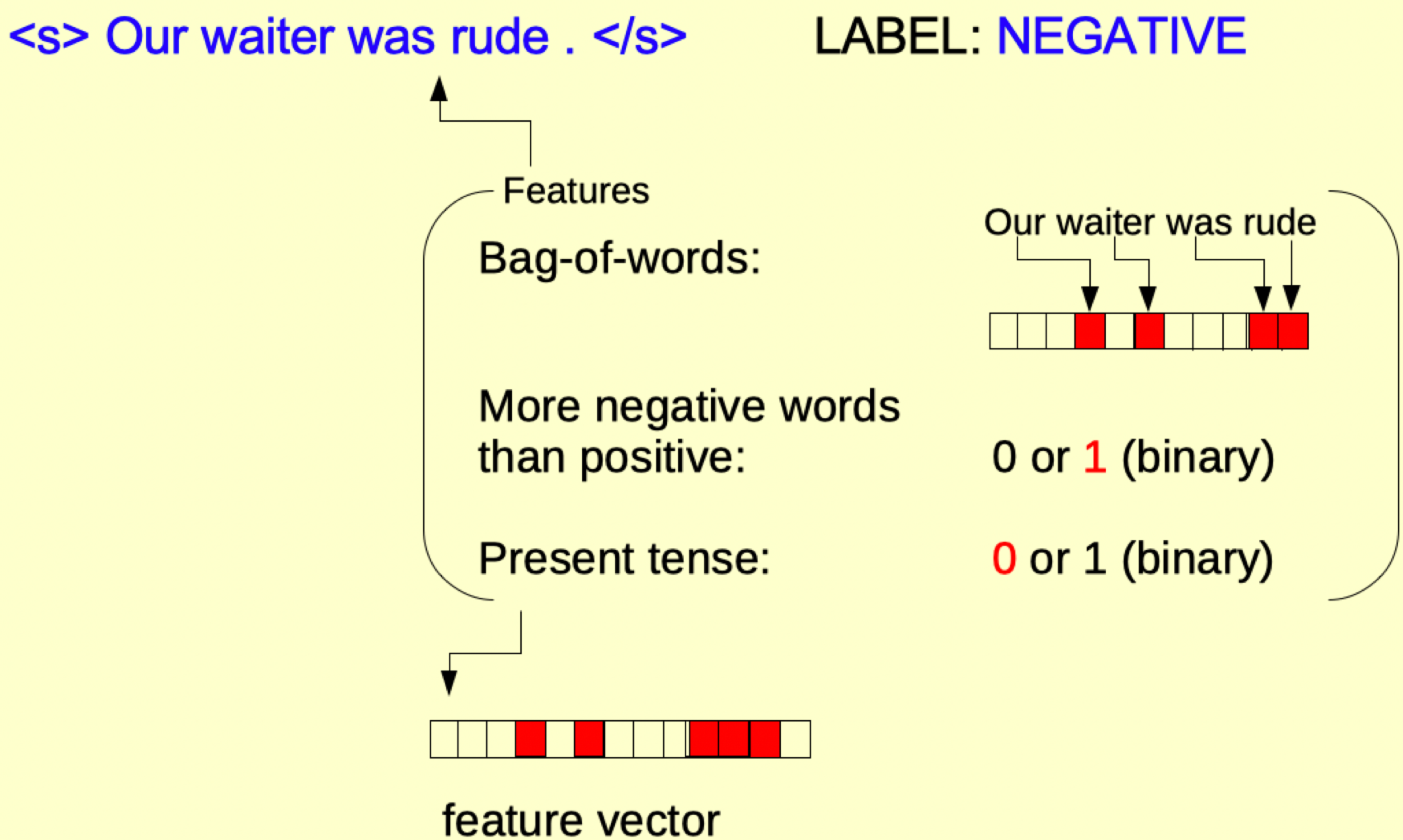
- a representation of each instance to be classified that contains the relevant information about that instance

Develop and train a statistical *model* using the *training data*

Annotate test set using the model, evaluate performance

Apply model to unseen text.

Example – Feature Representations: Sentiment Analysis



Data for NLP: Corpora

Need different corpora for different tasks!

Sentiment Analysis:

”Our waiter was rude!”

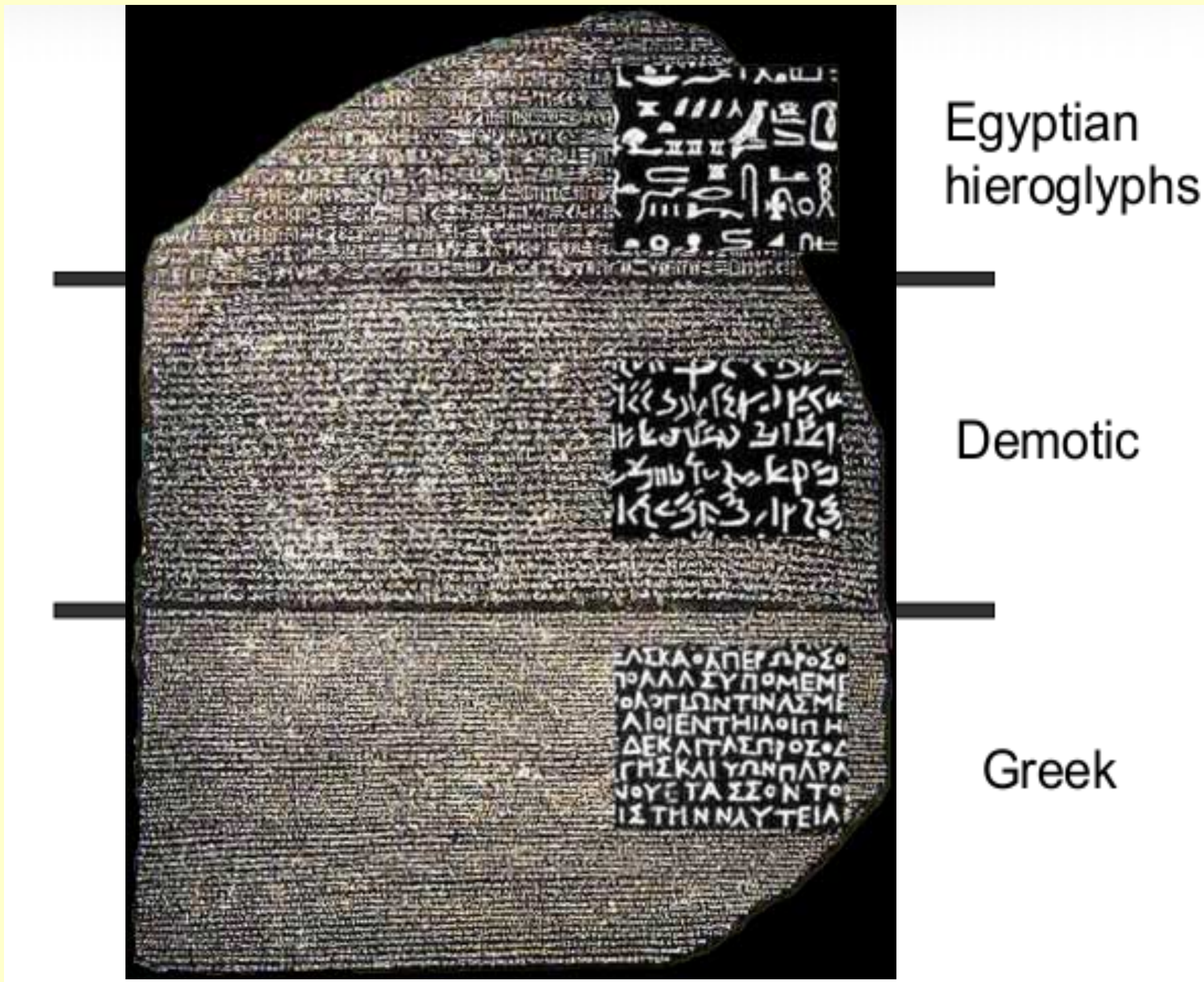
LABEL: negative

”This place is the best French restaurant in Boston!”

LABEL: positive

Machine Translation

Rosetta stone



Corpus for Machine Translation

Он благополучно избегнул встречи со своей хозяйкой на лестнице. He had successfully avoided meeting his landlady on the staircase.
Каморка его приходилась над самою кровлей высокого пятиэтажного дома и походила более на шкаф, чем на квартиру. His garrel was under the roof of a high, five-storied house and was more like a cupboard than a room.
Квартирная же хозяйка его, у которой он нанимал эту каморку с обедом и прислугой, помещалась одною лестницей ниже, в отдельной квартире The landlady who provided him with garrel, dinners, and attendance, lived on the floor below

Example: Sentence Segmentation

!, ? are relatively unambiguous

Period "." is quite ambiguous

- Sentence boundary
- Abbreviations like Inc. or Dr.

General idea

- Build a binary classifier
- Looks at a "."
- Decides EndOfSentence/NotEOS

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Example: Resolve Co-Reference / Anaphora

Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate it.

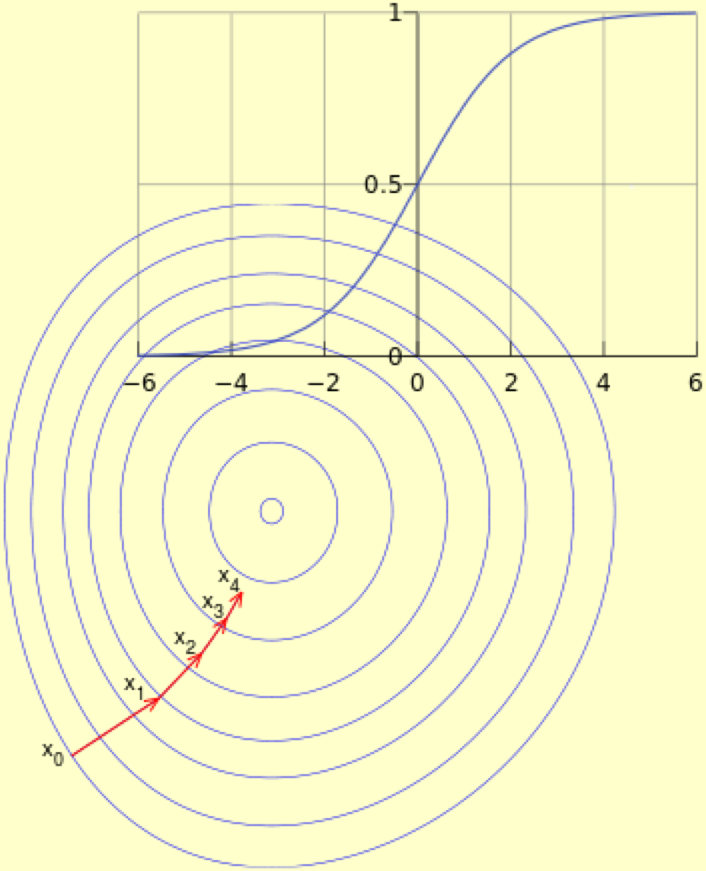
Bush started the war in Iraq. But the president needed the consent of Congress.

Some cases require difficult reasoning.

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Statistical Machine Learning

Statistical Machine Learning



- Need to maximize:

$$L(\mathbf{W}) = \sum_{i=1}^{n} \mathbf{W} \cdot \phi(x_i, y_i) - \sum_{i=1}^{n} \log \sum_{y' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, y')}$$

- Calculating gradients:

$$\begin{aligned} \left.\frac{dL}{d\mathbf{W}}\right|_{\mathbf{W}} &= \sum_{i=1}^{n} \phi(x_i, y_i) - \sum_{i=1}^{n} \frac{\sum_{y' \in \mathcal{Y}} \phi(x_i, y') e^{\mathbf{W} \cdot \phi(x_i, y')}}{\sum_{z' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, z')}} \\ &= \sum_{i=1}^{n} \phi(x_i, y_i) - \sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') \frac{e^{\mathbf{W} \cdot \phi(x_i, y')}}{\sum_{z' \in \mathcal{Y}} e^{\mathbf{W} \cdot \phi(x_i, z')}} \\ &= \underbrace{\sum_{i=1}^{n} \phi(x_i, y_i)}_{\text{Empirical counts}} - \underbrace{\sum_{i=1}^{n} \sum_{y' \in \mathcal{Y}} \phi(x_i, y') P(y' \mid x_i, \mathbf{W})}_{\text{Expected counts}} \end{aligned}$$

Models Trained on Huge Collections of Text

What is Machine Learning?

The ability to take as input some data and produce some output

For example, take as input an image and identify the object in that image

input X: image

ouput Y: object category

Or, take as input some information about a house and predict its price

input X: house location, square footage, etc.

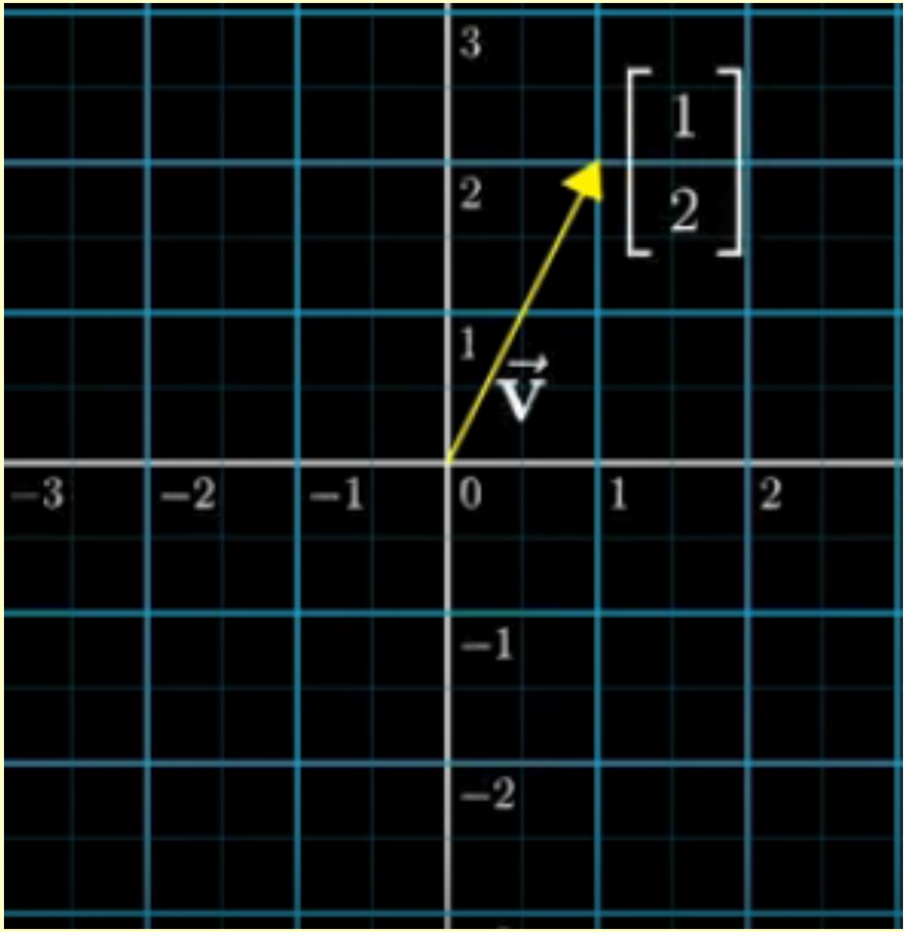
ouput Y: price

Inputs are typically ”vectors”

A vector is an ordered list of numbers!

This is a 2-dimensional vector

A 1-dimensional vector is just a single number!



Inputs are typically "vectors"

A vector is an ordered list of numbers!

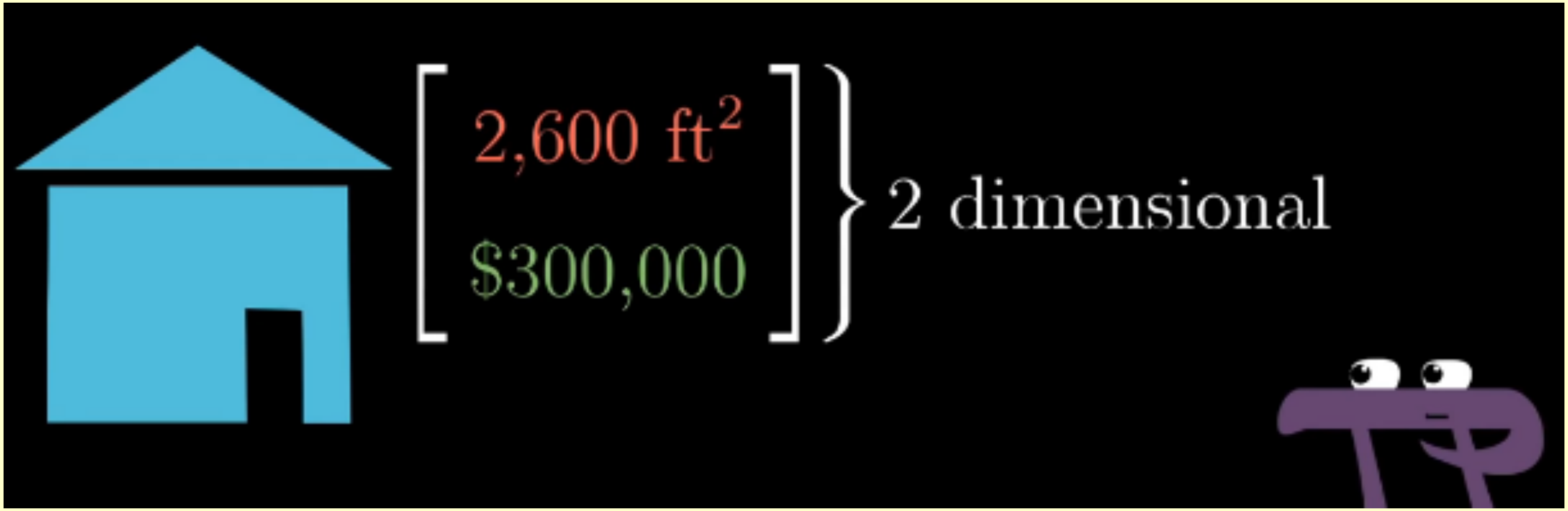
These are a 2-d, 4-d and 3-d vectors

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 5 \\ 0 \\ 0 \\ -3 \end{bmatrix} \quad \begin{bmatrix} 2.3 \\ -7.1 \\ 0.1 \end{bmatrix}$$

Inputs are typically "vectors"

This is a 2-dimensional vector

Can be used as input to a **model** that would predict housing prices!



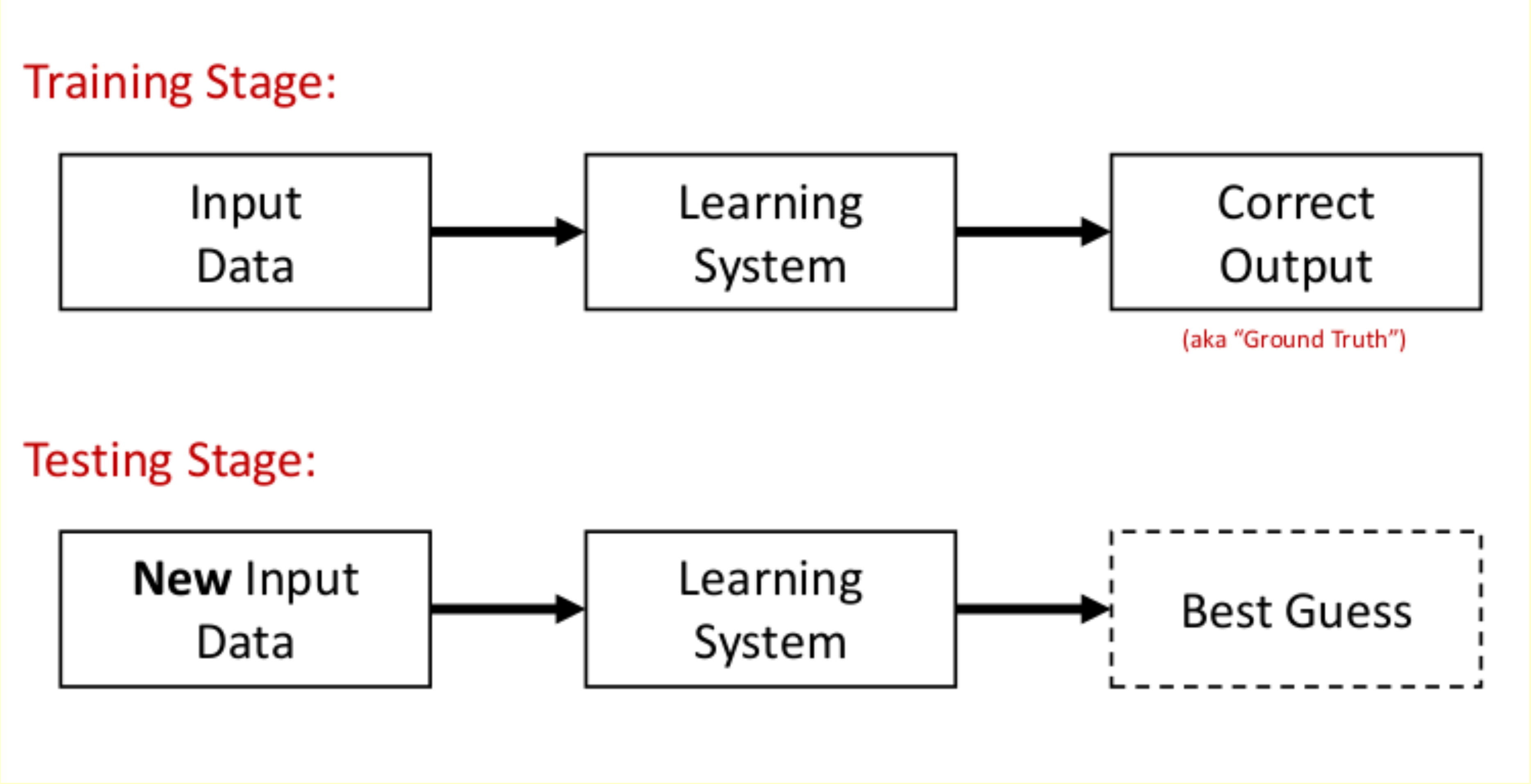
What is a Model?

A model is a function

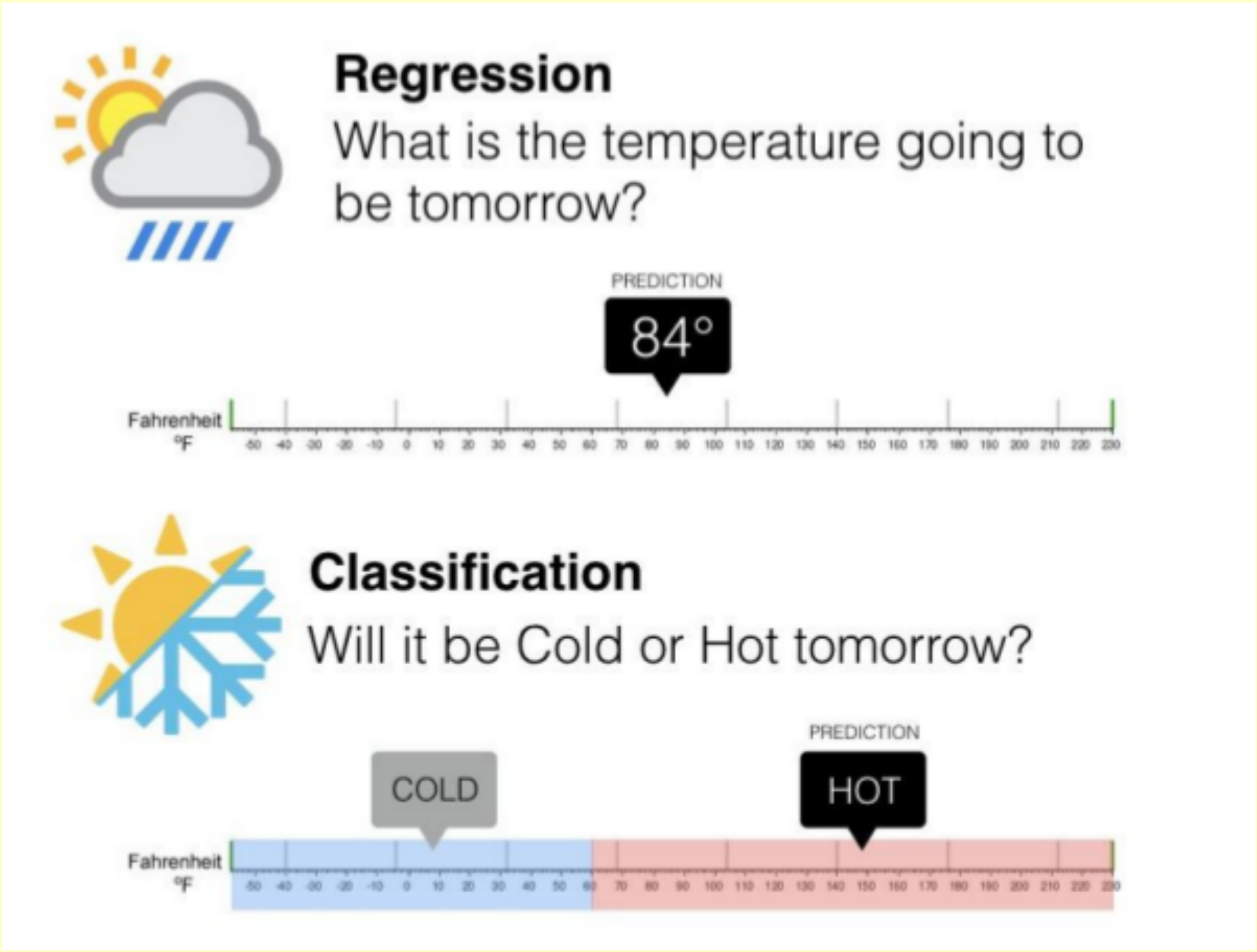
Takes as input some vectors

Produces a number, a category label, an output

Training a Model



What does the model outputs?



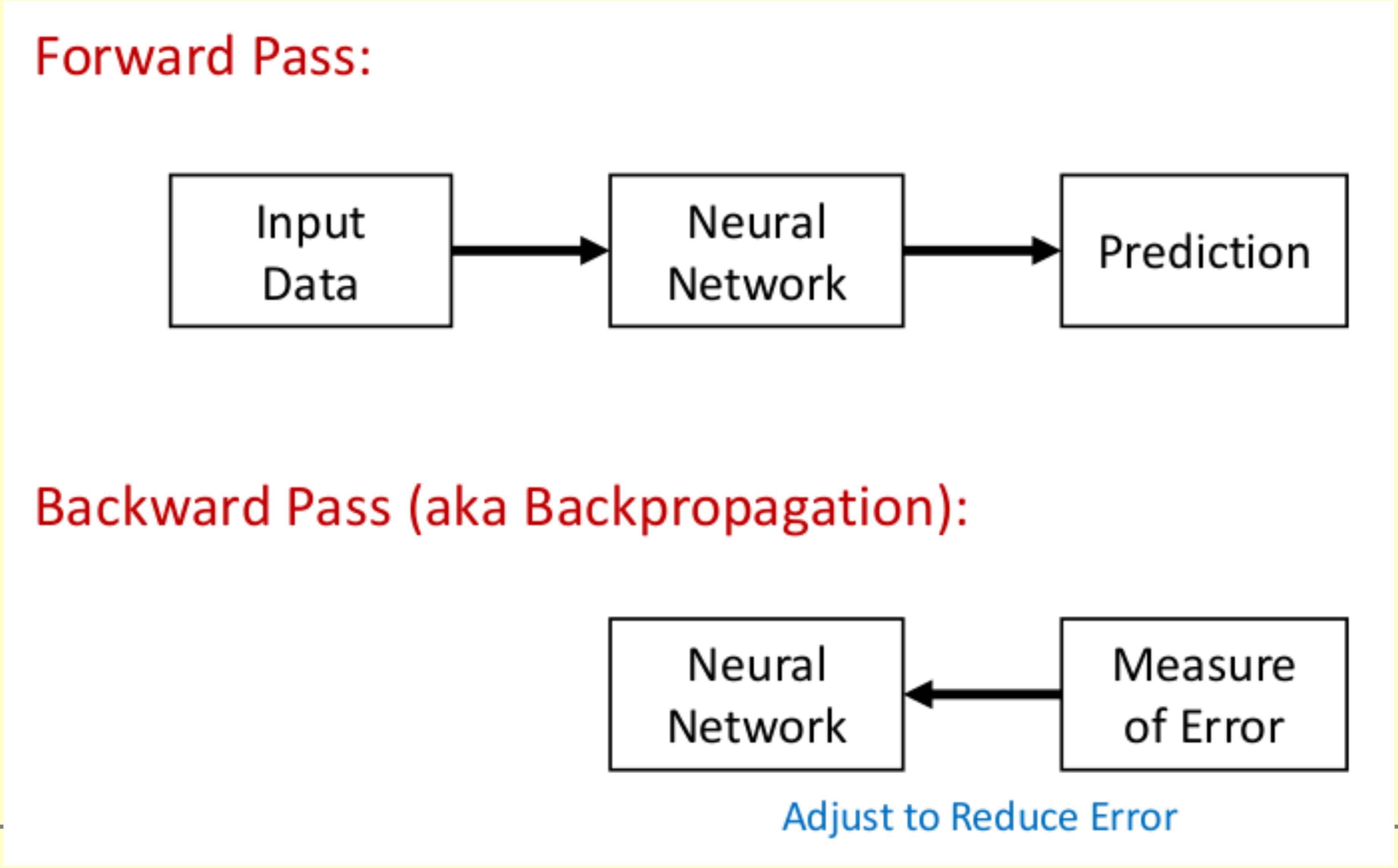
Supervised Learning

- Training : $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$
- $\boldsymbol{x}_i$: input vector

$$\boldsymbol{x}_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix}, \quad x_{i,j} \in \mathbb{R}$$

- y : response variable
  - $y \in \{-1, 1\}$: binary classification
  - $y \in \mathbb{R}$: regression
  - what we want to be able to predict, having observed some new $\boldsymbol{x}$.

Training a Model

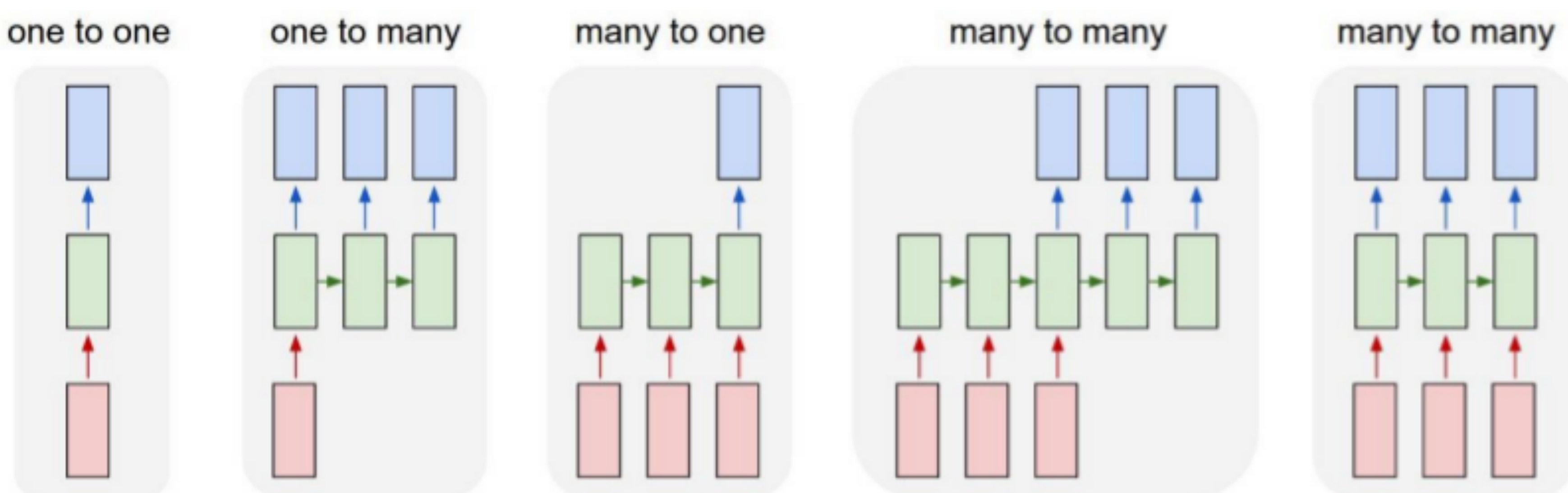


Model Inputs and Outputs

Input Data → Learning System → Correct Output

- Number
- Vector of numbers
- Sequence of numbers
- Sequence of vectors of numbers

- Number
- Vector of numbers
- Sequence of numbers
- Sequence of vectors of numbers



Example: Linear Regression Model

Input : $\boldsymbol{x} \in \mathbb{R}^d$

Parameters: $\boldsymbol{w} \in \mathbb{R}^{d+1}$

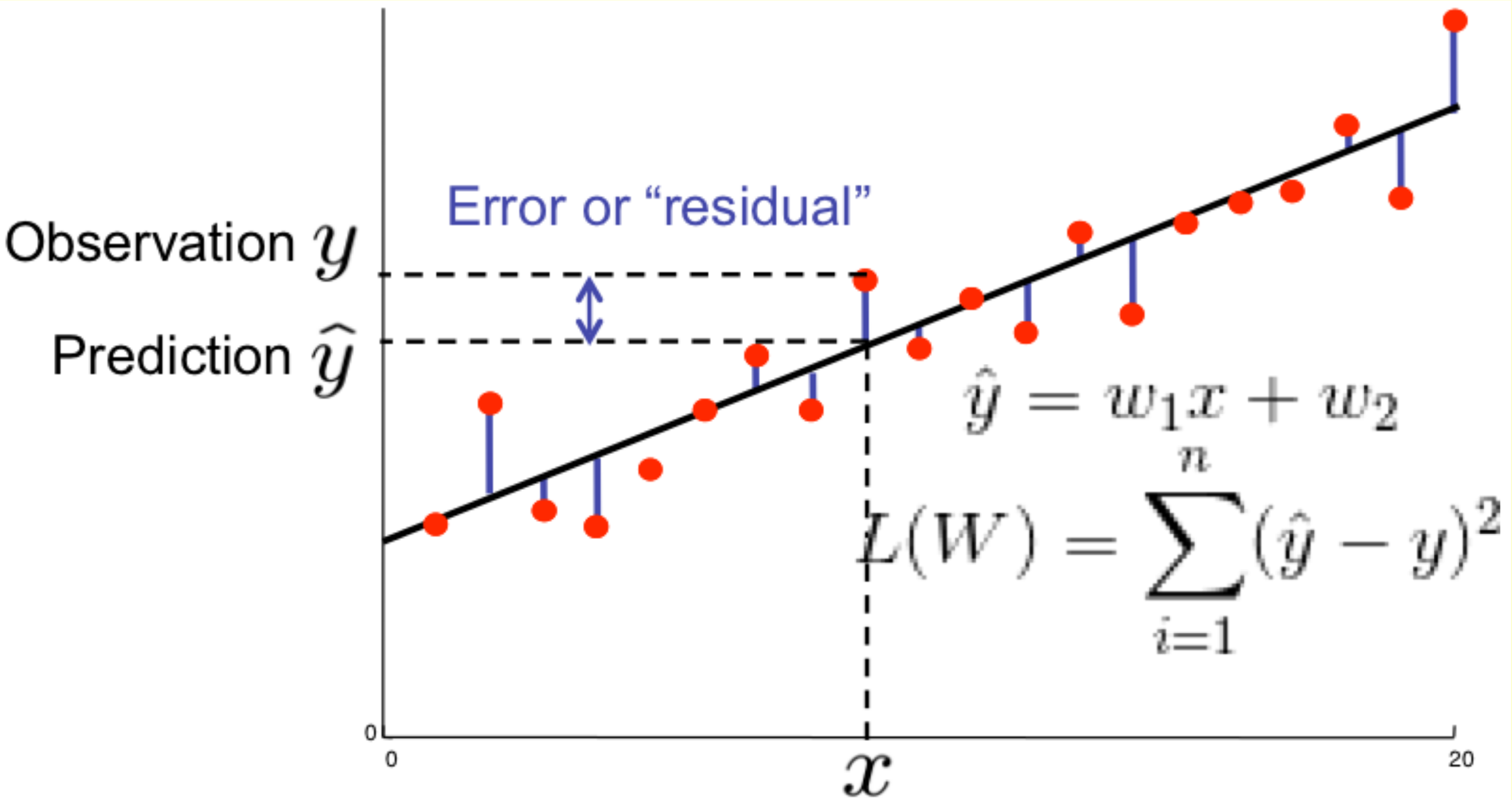
Response : $y \in \mathbb{R}$

Prediction : $y = \boldsymbol{w}^\top \boldsymbol{x}$

- Recall that we can fit (learn) the model by minimizing the squared error:

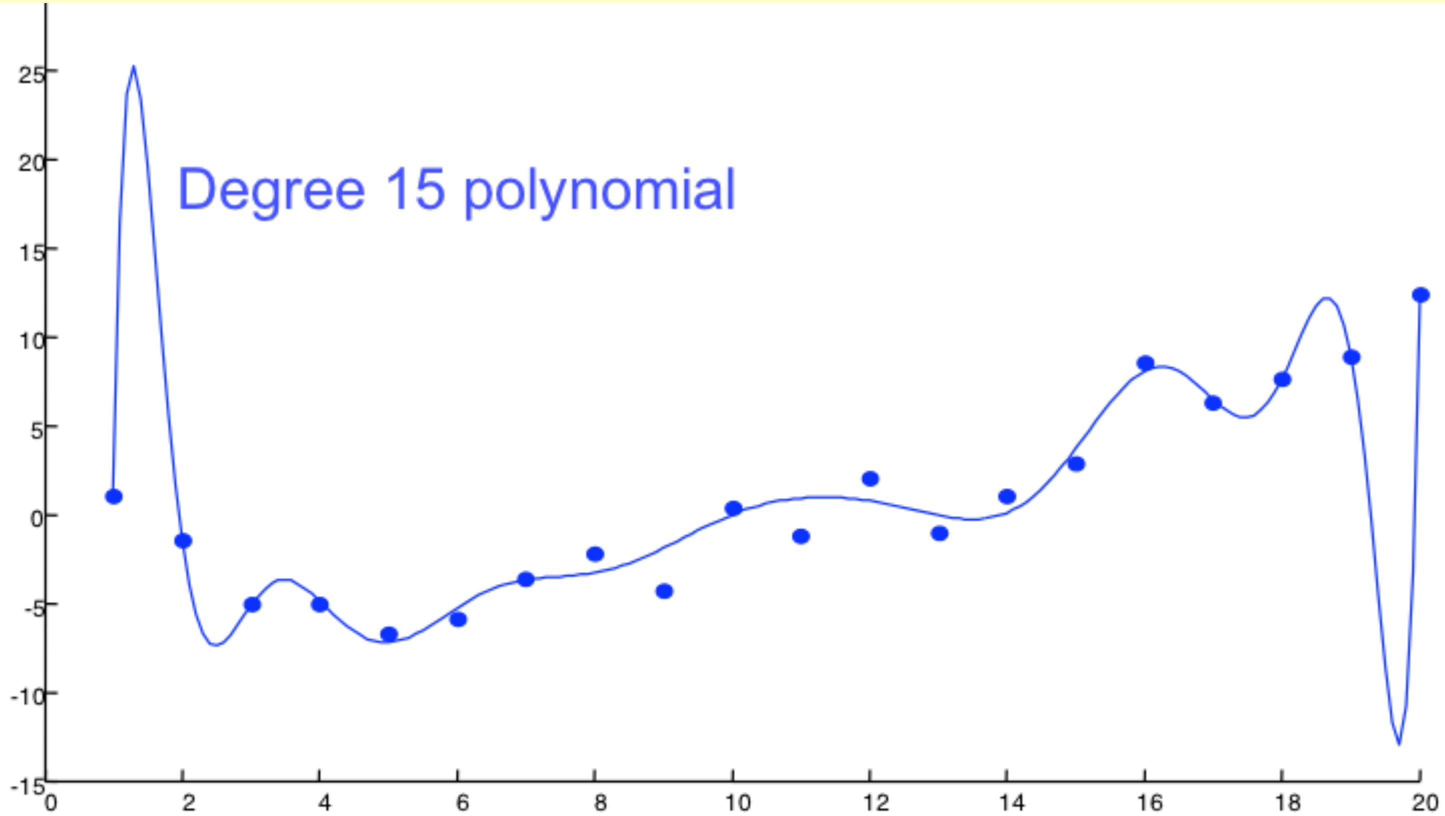
$$\hat{\boldsymbol{w}} = \mathrm{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

Linear Regression



Sum squared error: $L(w) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$

Non-linear function



- This model is too rich for the data
- Fits training data well, but doesn’t generalize.

Model Training

A model is a function that maps inputs to outputs

- *Allows us to make predictions*

Loss function: allows us to *find the best model function*!!

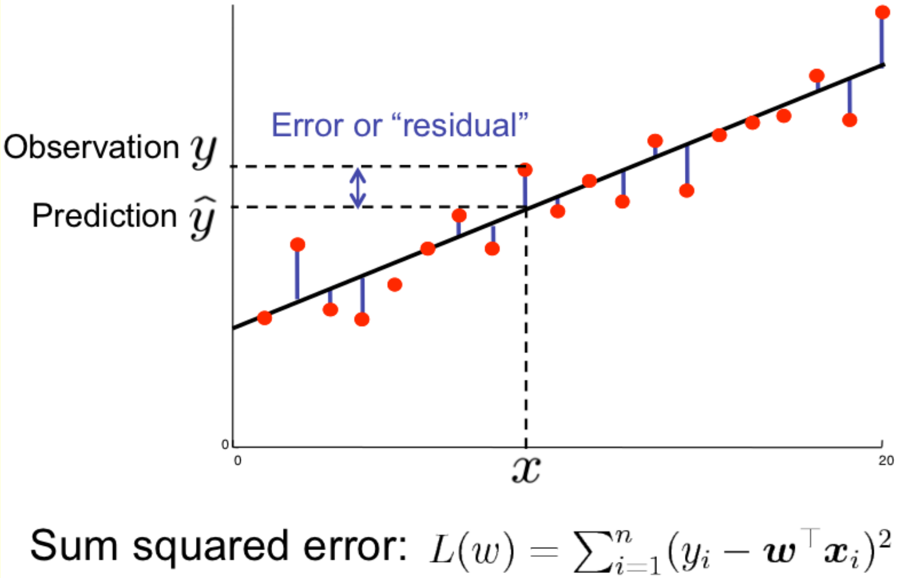
- *How far is our prediction from the truth?*

Training a model involves finding the best coefficients ("weights", "parameters") for our model function

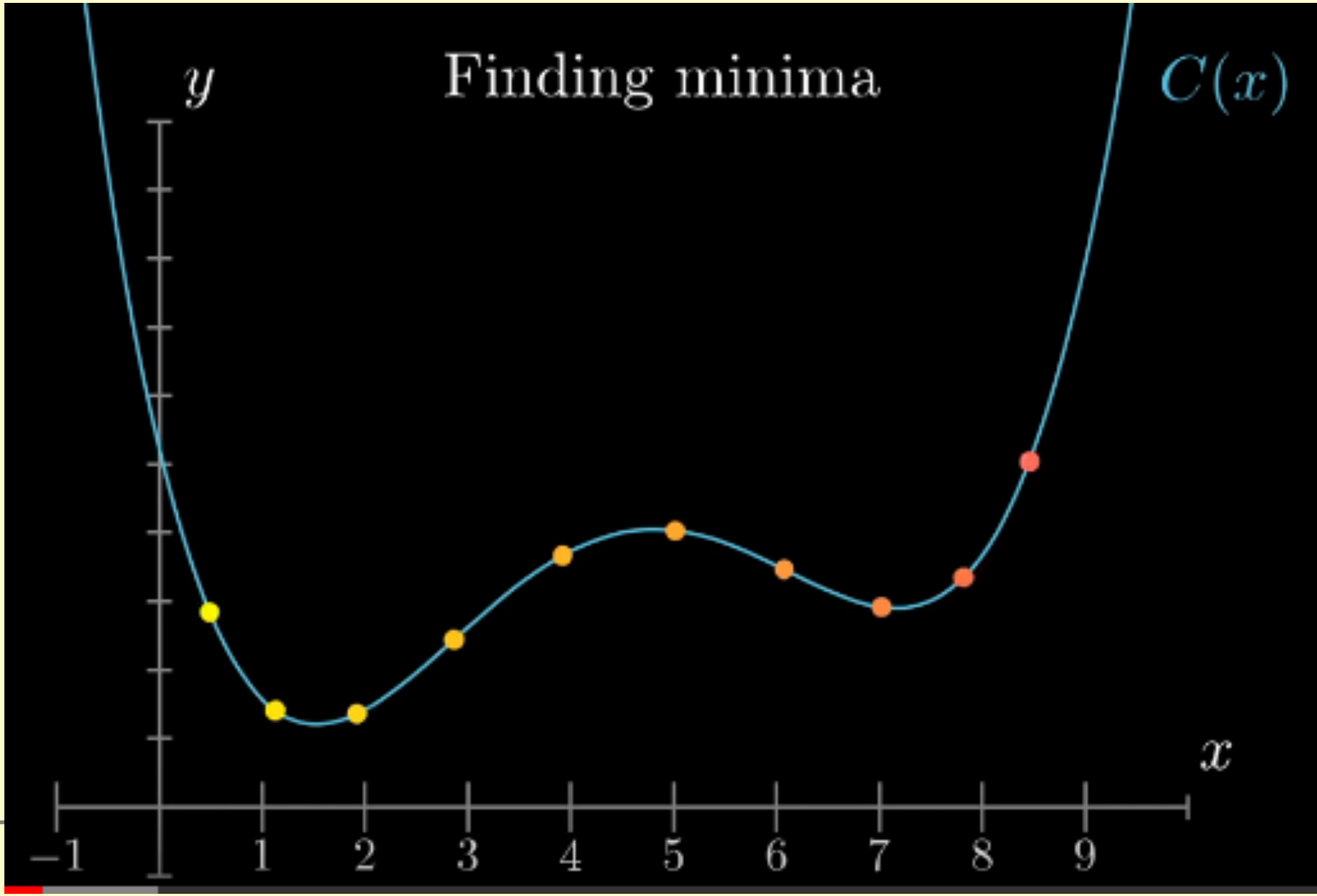
These are found by optimizing the loss

$$\hat{y} = w_1 x + w_2$$

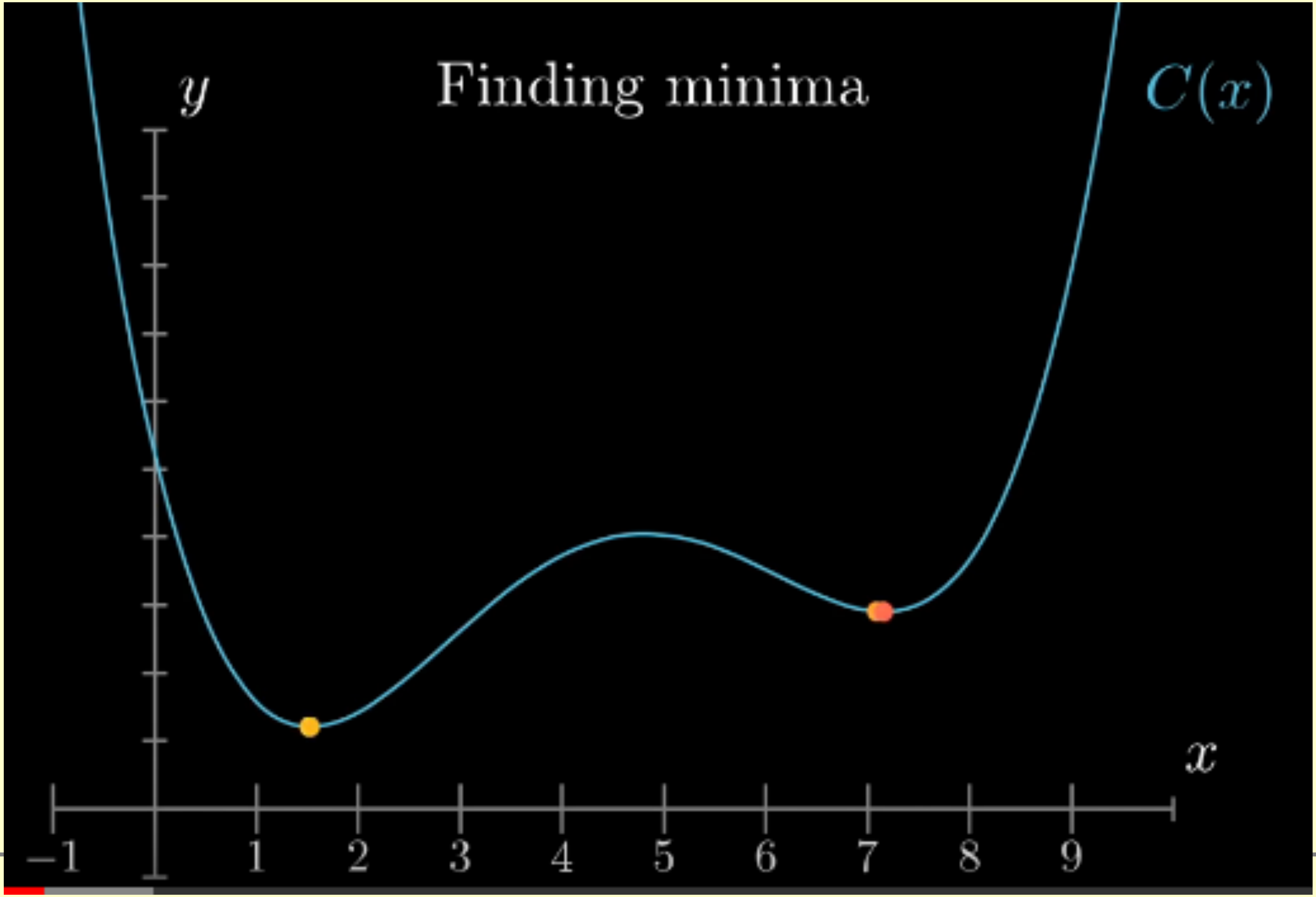
$$L(W) = \sum_{i=1}^{n} (\hat{y} - y)^2$$



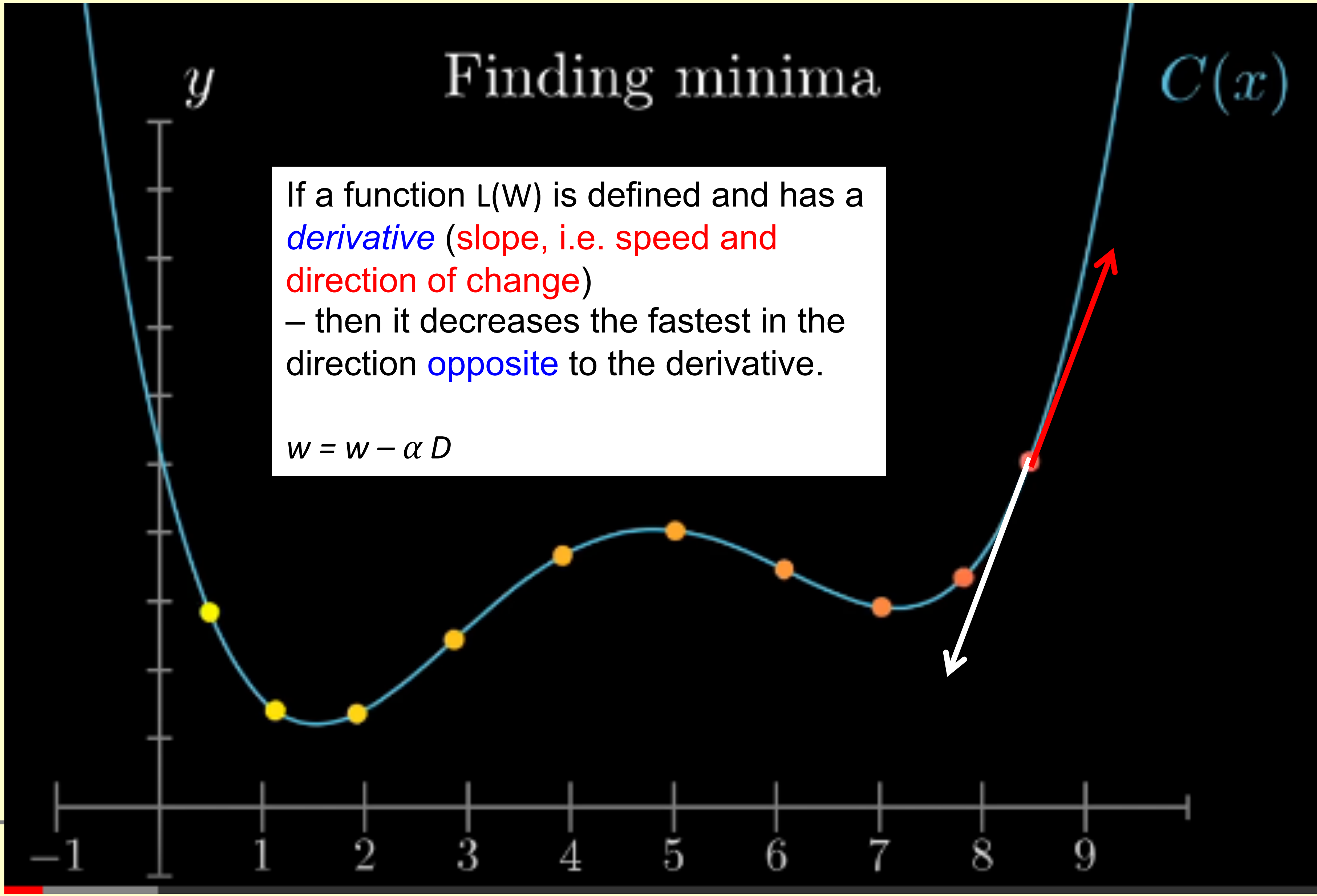
Gradient Descent



Gradient Descent



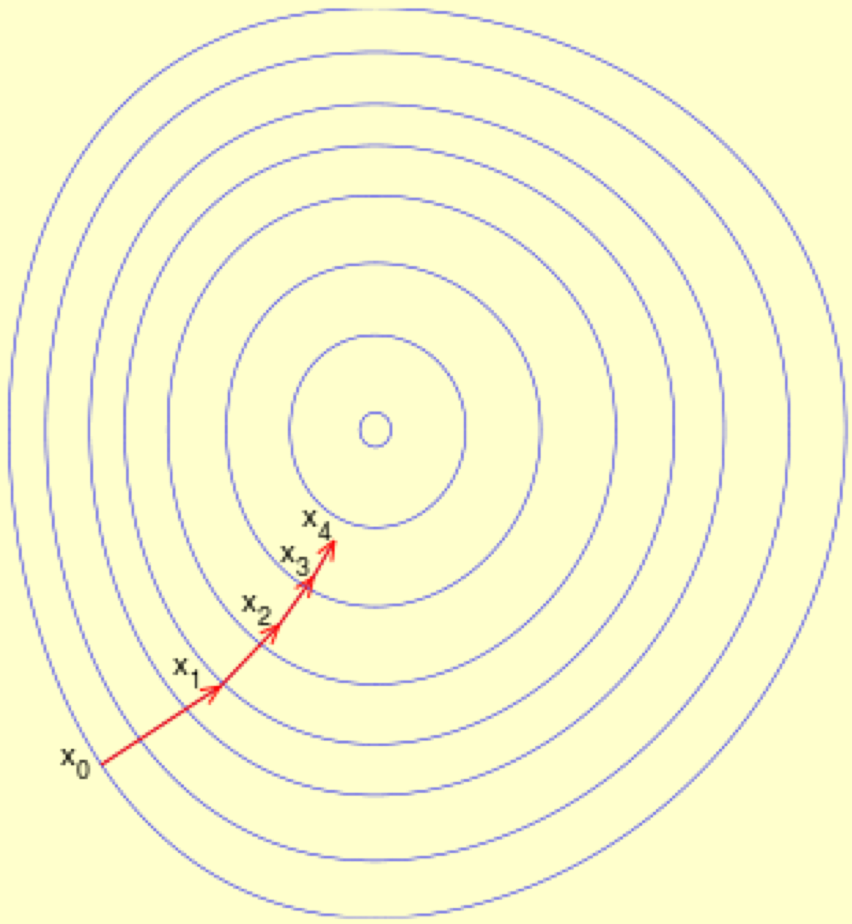
Gradient Descent



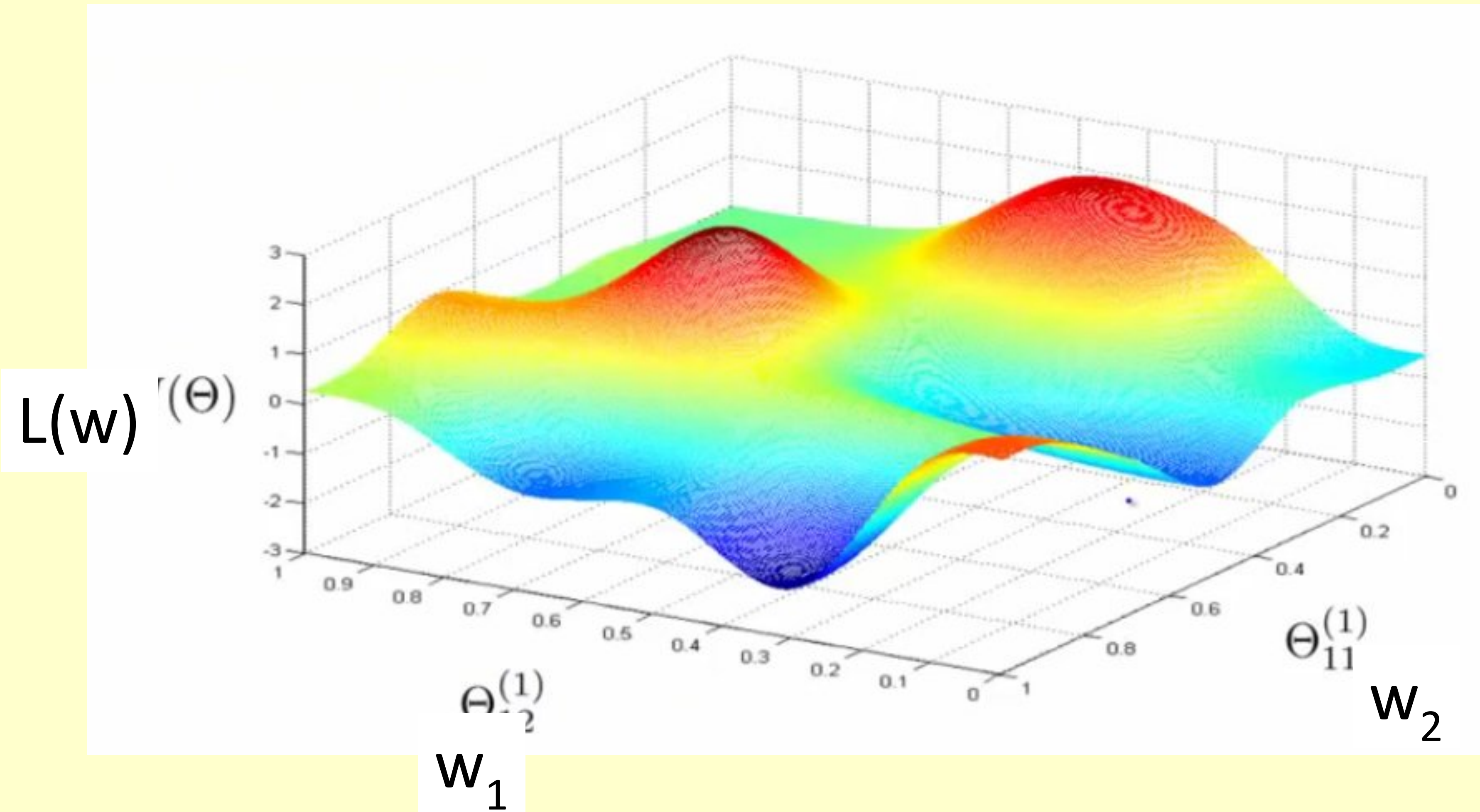
If a function L(W) is defined and has a *derivative* (slope, i.e. speed and direction of change)
– then it decreases the fastest in the direction opposite to the derivative.

$w = w - \alpha D$

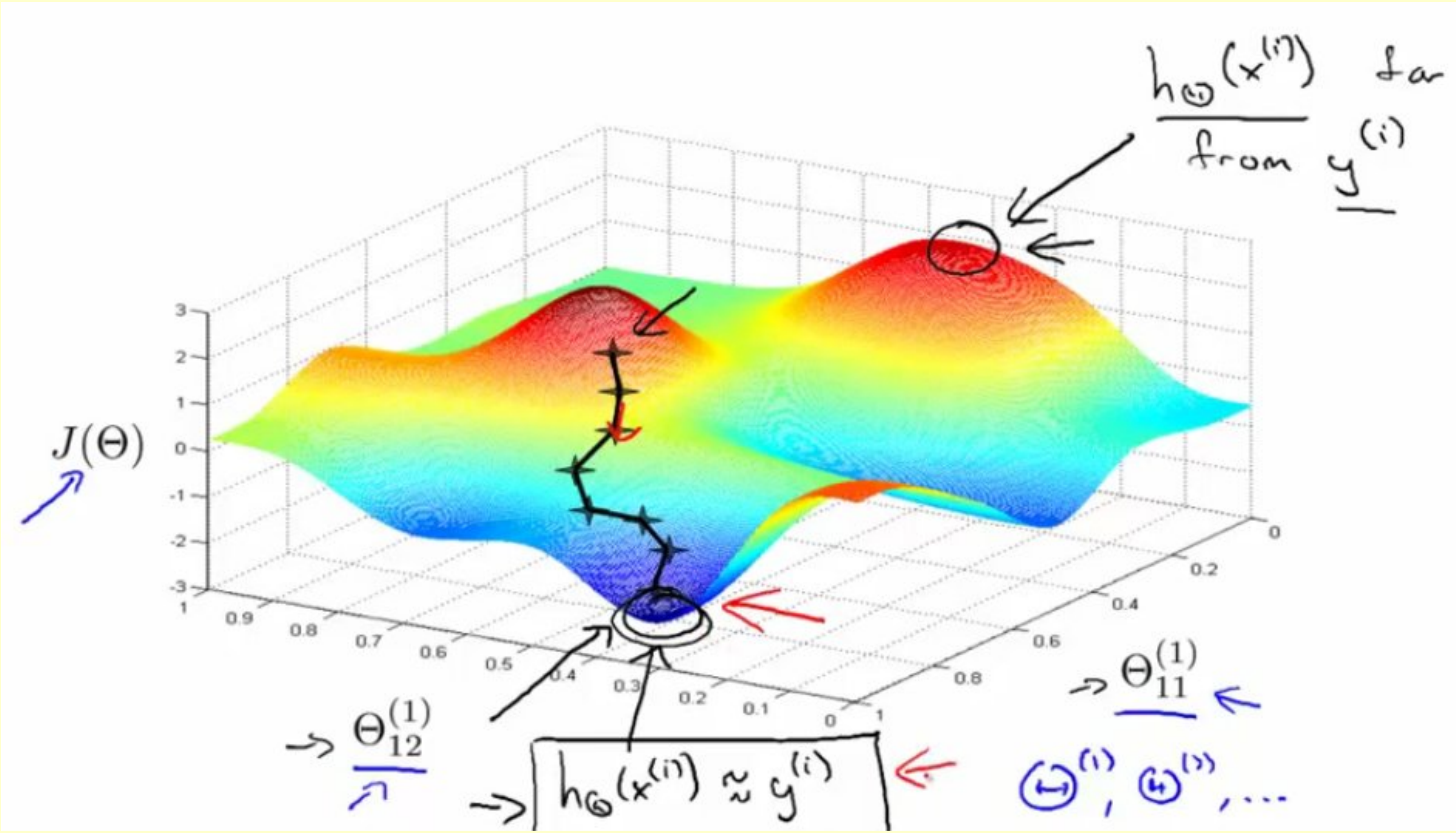
2-dimensional loss function



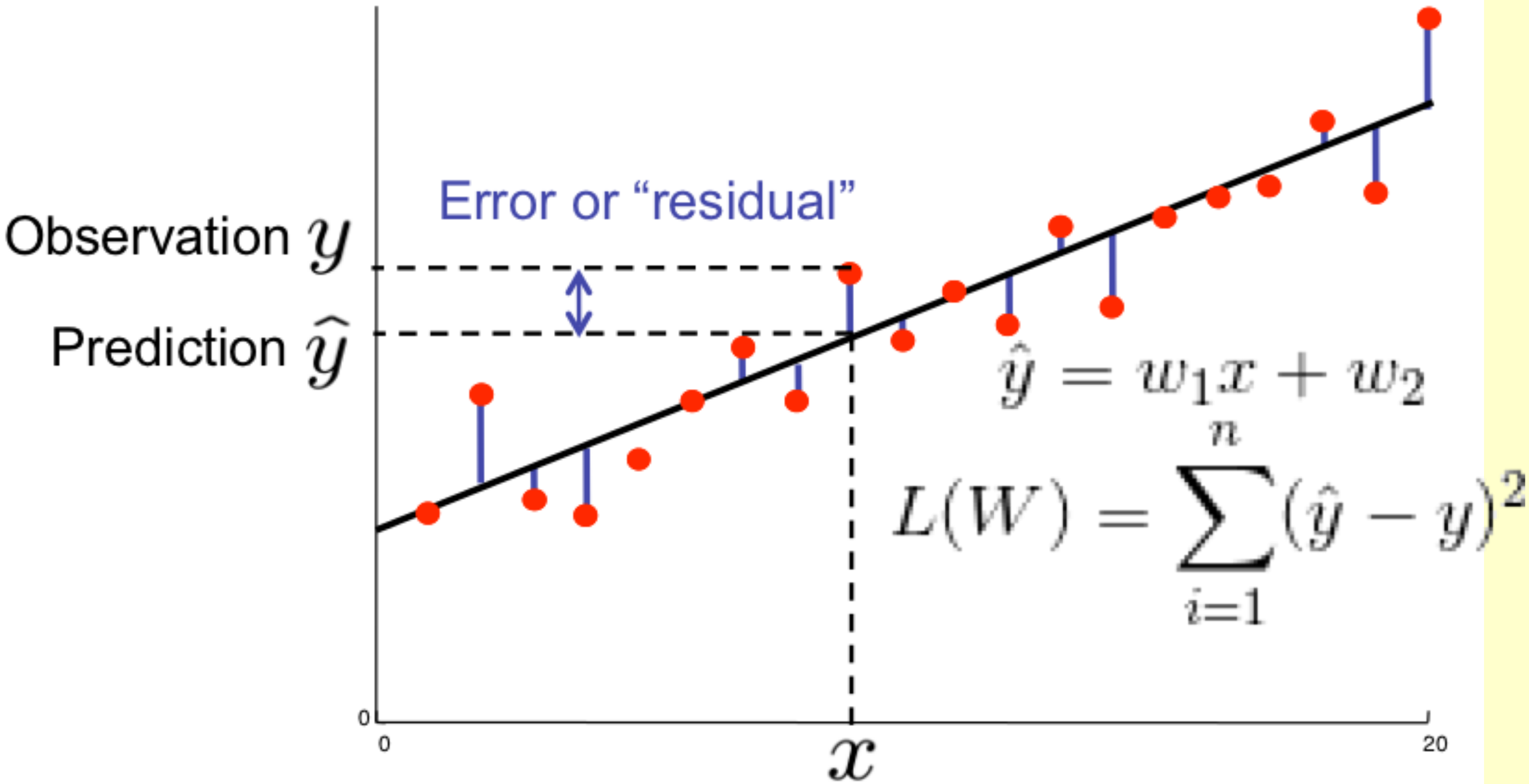
Loss function in weight space



Gradient descent in weight space



Linear Regression



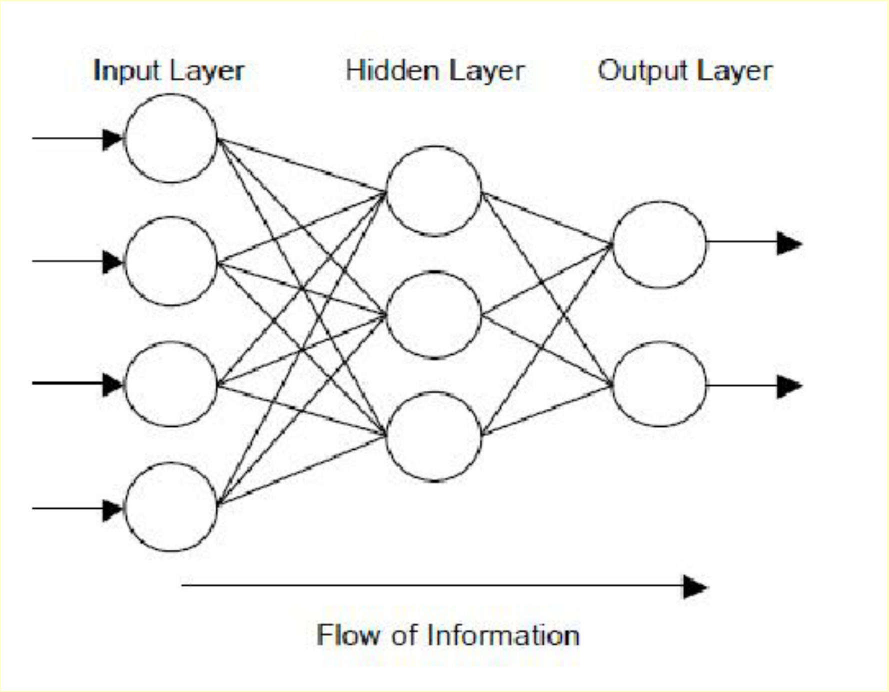
Sum squared error: $L(w) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$

Deep Learning Revolution

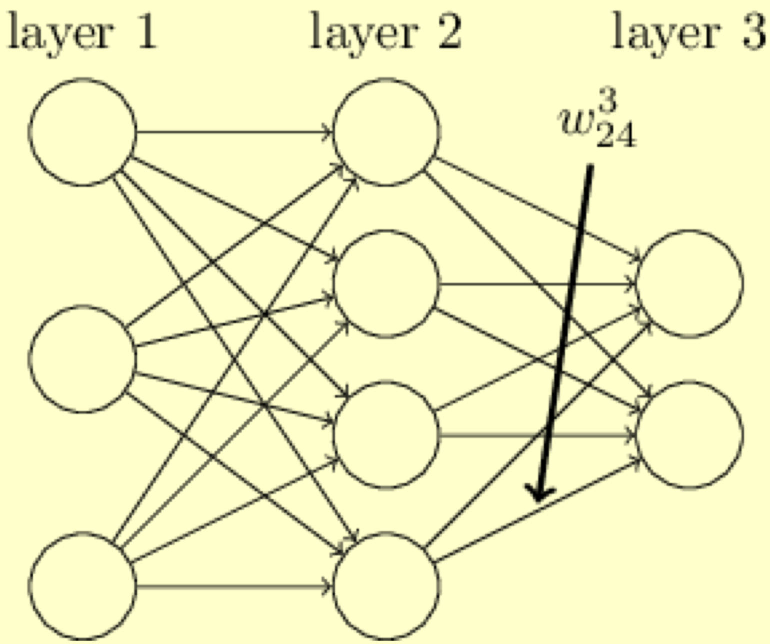
What is deep learning?

– Your models are multi-layer neural networks

– A particular kind of model configuration



Multi-Layer Perceptron

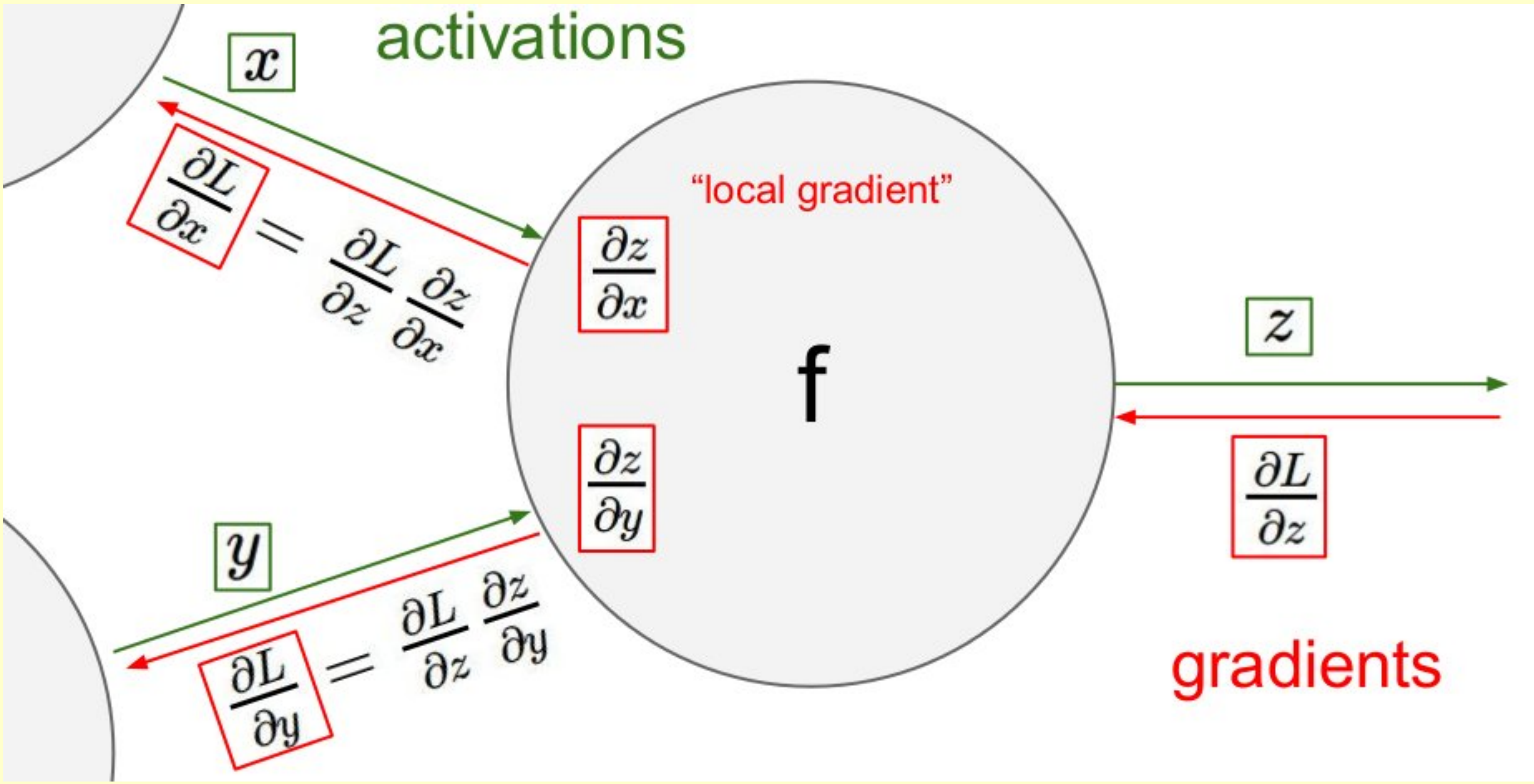


w^l_{jk} is the weight from the $k^{\rm th}$ neuron in the $(l-1)^{\rm th}$ layer to the $j^{\rm th}$ neuron in the $l^{\rm th}$ layer

What is backpropagation?

A way of computing gradient descent of the loss function in a neural network through recursive application of the chain rule

Local gradients



Modern Era NLP: Deep Learning

Modern NLP uses deep learning

- NLP got swept into the deep learning revolution about 9 years ago.
- The majority of current NLP models are neural networks that at the core representation-learning models
- They learn to produce an encoding of input (text) into a vector space embedding.

Modern NLP – Transformers (since 2018)

- A new class of neural models
- Excel at both language generation and understanding
- Huge models, getting progressively bigger and bigger!

– you might have seen some of them in the news...

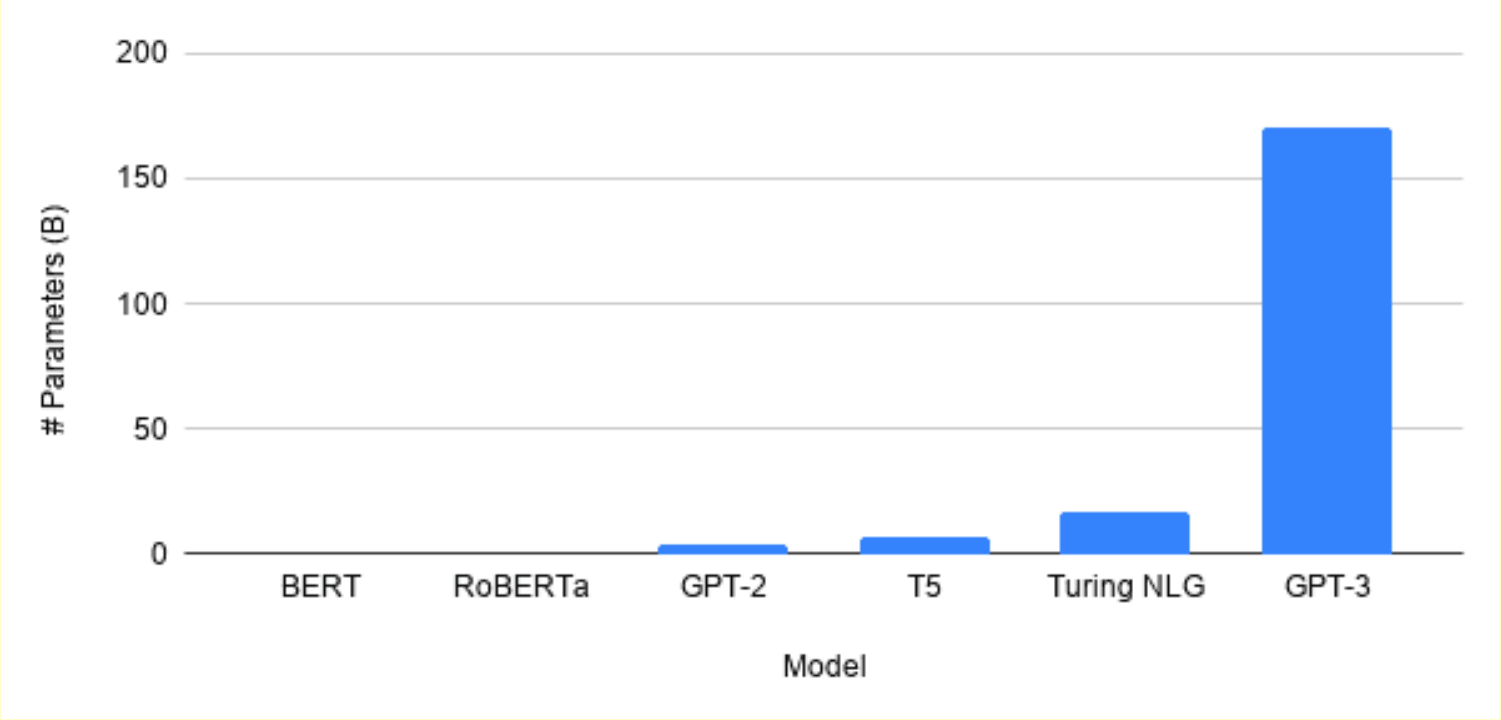
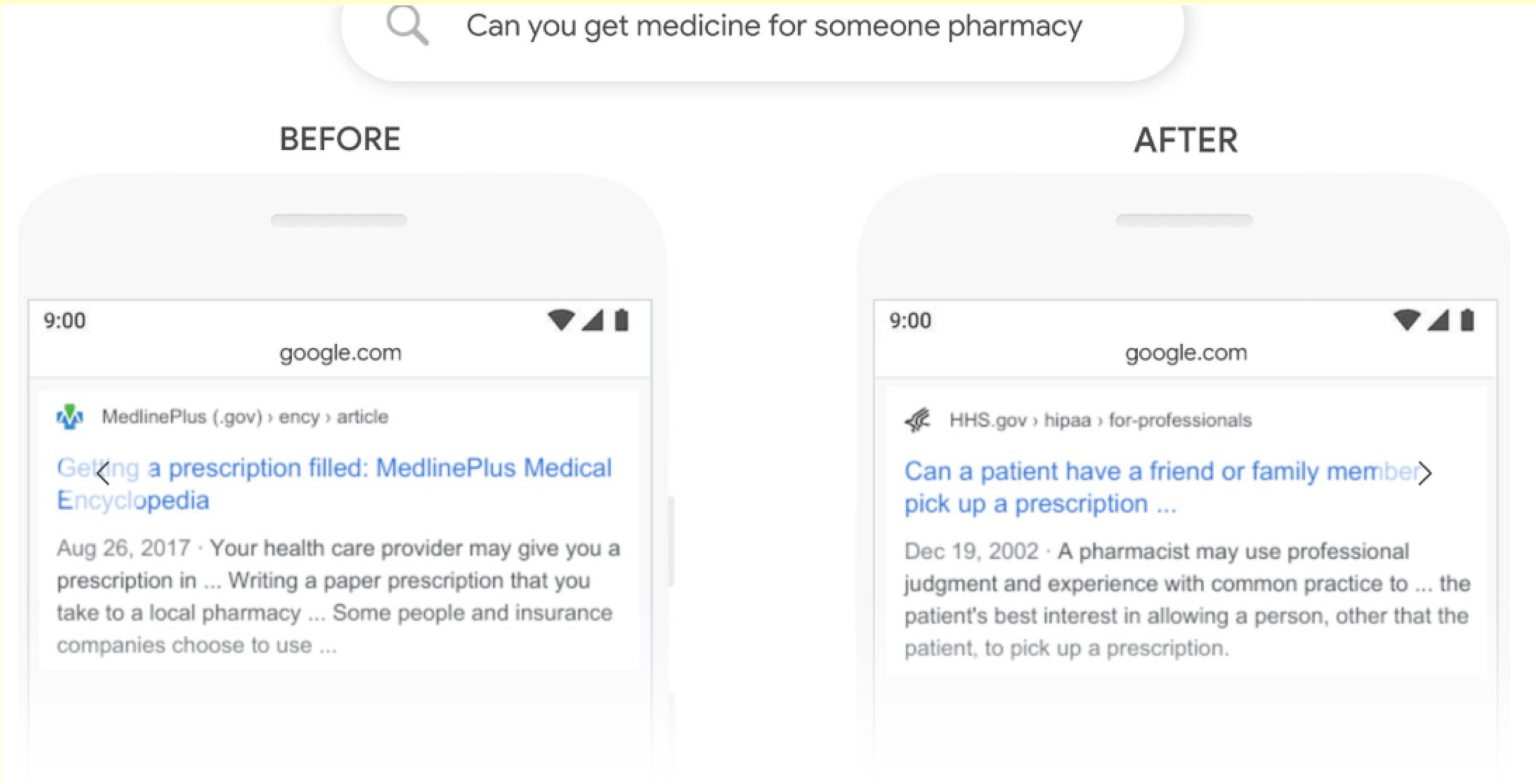


Image source: https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/

Google's BERT improves ~10% of Google search queries



Super-human performance on some benchmarks

GLUE Benchmark – a common leaderboard for collection of tasks such as

- Recognizing paraphrases
- Judging semantic similarity between sentences
- Identifying entailment & contradictions between passages
- Sentiment analysis
- Identifying words referred to by a pronoun

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
	2	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	4	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
+	5	王玮	ALICE large ensemble (Alibaba DAMO NLP)		86.3	68.6	95.2	92.6/90.2	91.1/90.6	74.4/90.7	88.2	87.9	95.7	83.5	80.8	43.9
	6	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9

Language modeling makes headlines!

TOM SIMONITE BUSINESS 02.14.2019 12:00 PM

The AI Text Generator That's Too Dangerous to Make Public

Researchers at OpenAI decided that a system that scores well at understanding language could too easily be manipulated for malicious intent.

OpenAI won't release an AI model due to its ability to create fake news

By Usama Jawad · Feb 16, 2019 01:50 EST · HOT! 12

OpenAI's GPT-2 Model (Feb 2019)

Human Prompt:

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion (machine-written, 10 tries):

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Single-shot learning of new words with GPT-3 (Jul 2020)

Human: A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
Model: We were traveling in Africa and we saw these very cute whatpus.

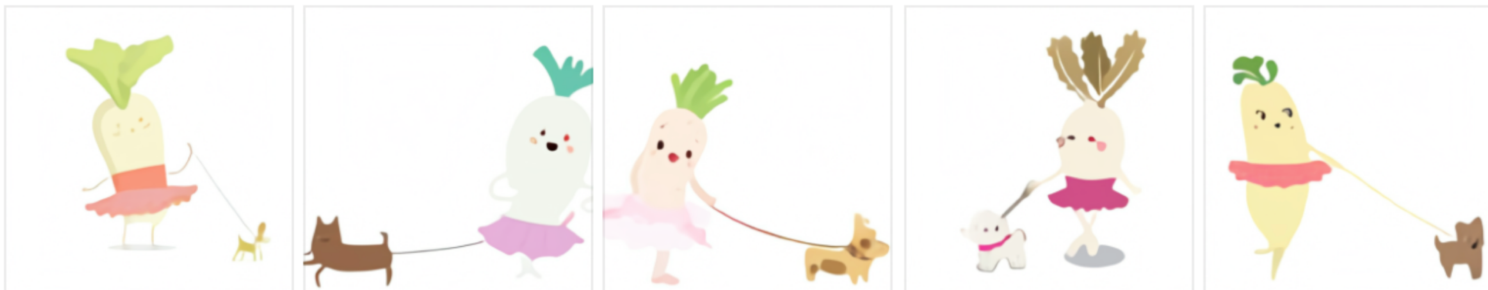
Human: To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
Model: One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

Human: A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
Model: In our garage we have a Burringo that my father drives to work every day.

DALL-E Model (Jan 2021)

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

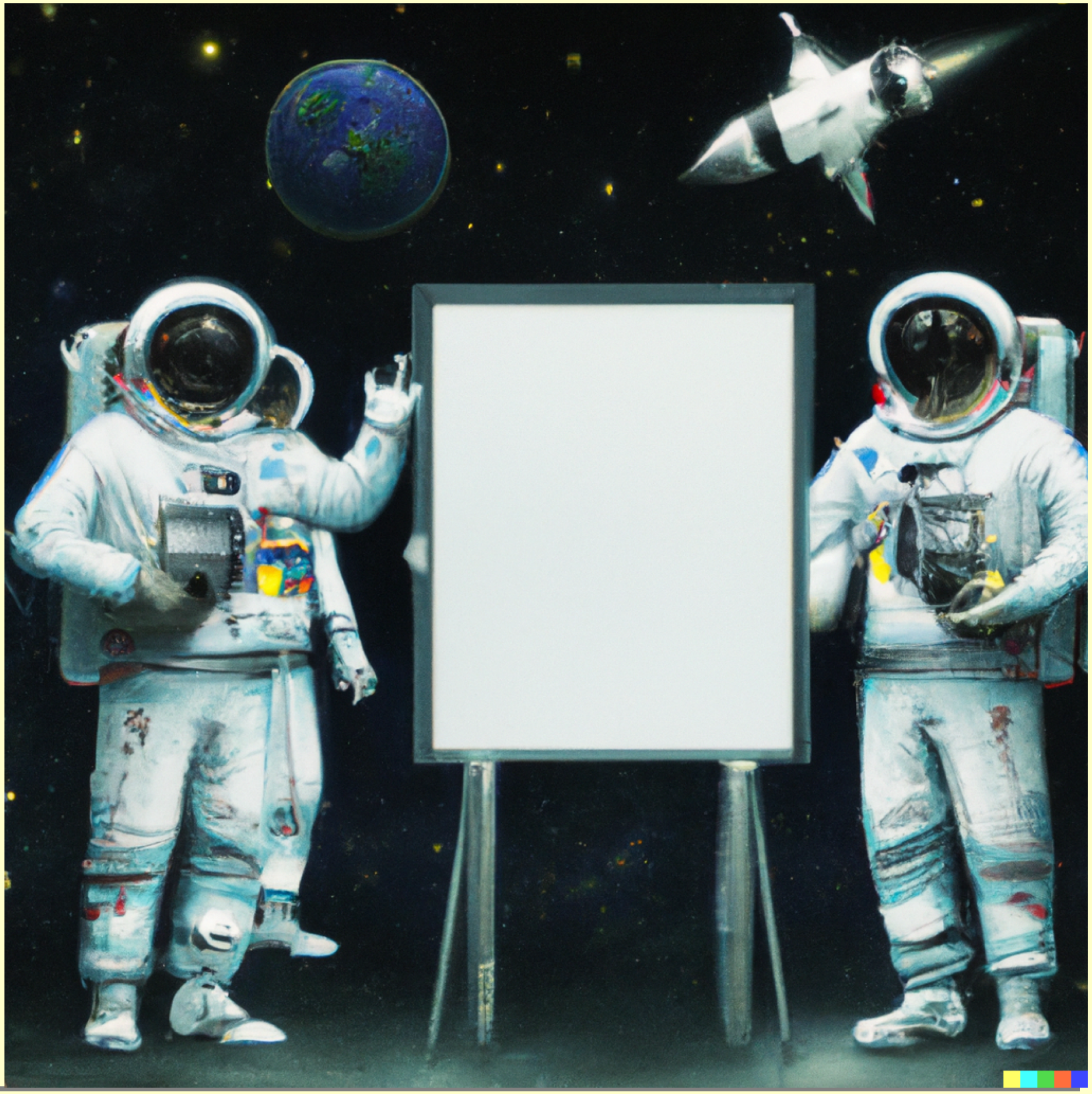
AI-GENERATED IMAGES

Edit prompt or view more images↓

DALL-E 2 Model (Apr 2021)

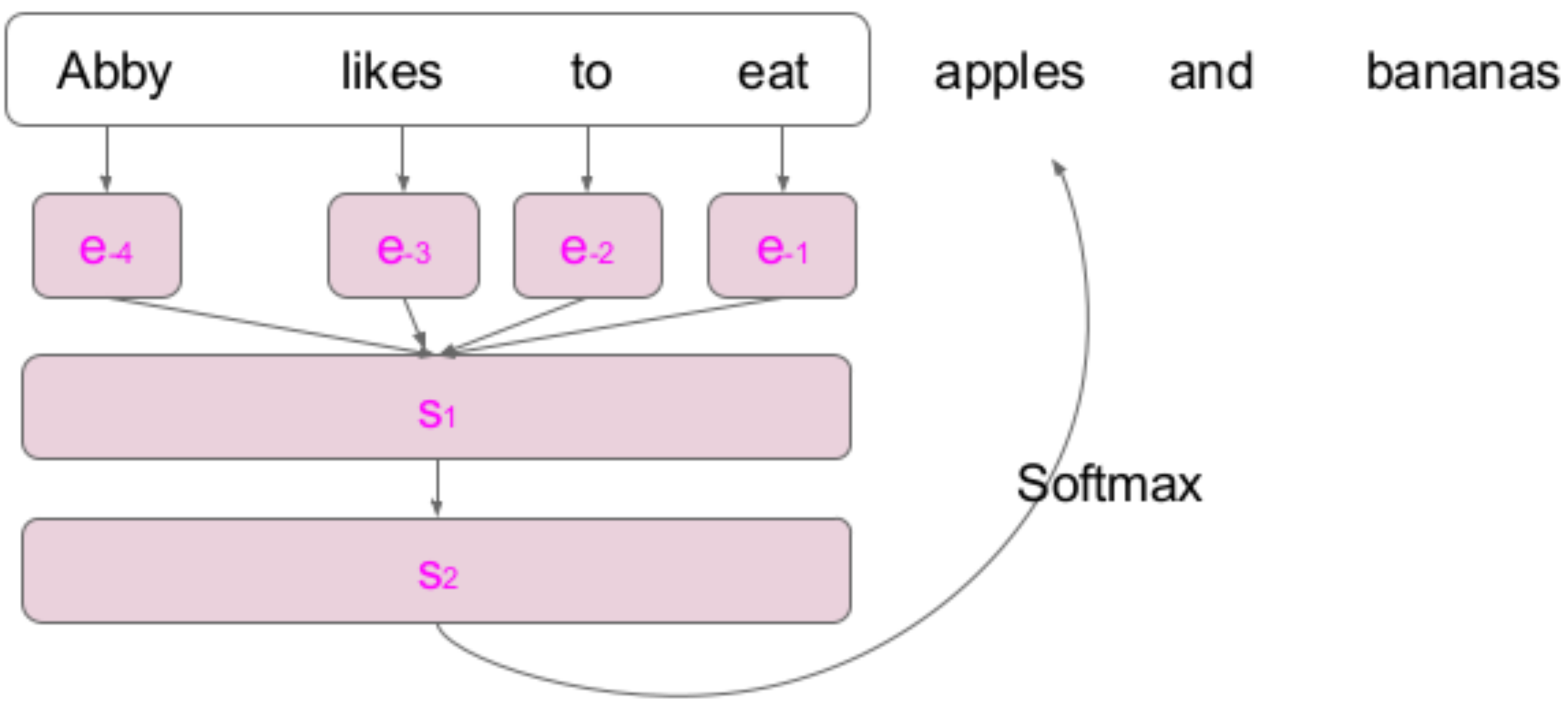
TEXT PROMPT

Students, professor in space suits, and a whiteboard, floating in space against the background of stars, photorealistic image



Language Modeling

Embedding Pretraining (Collobert et al, 2011)



Classification