

What is NLP?

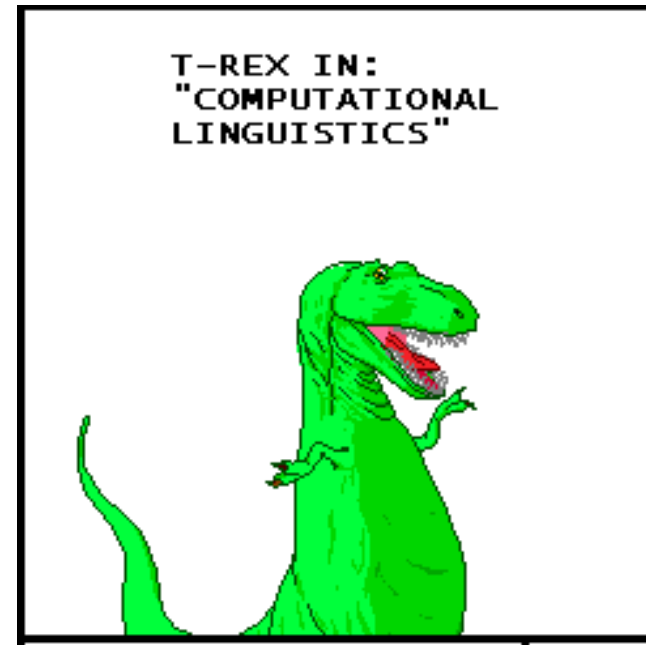
CMSC 473/673

Frank Ferraro – ferraro@umbc.edu



Today's Learning Goals

- NLP vs. CL
- Terminology:
 - NLP: vocabulary, token, type, one-hot encoding, dense embedding, parameter/weight, corpus/corpora
 - Linguistics: lexeme, morphology, syntax, semantics, “discourse”
- Universal Dependencies



T-REX IN:
"COMPUTATIONAL
LINGUISTICS"



Computational
linguistics is
the study of
computer-based
language
processing!



Natural Language Processing
 \approx
Computational Linguistics

Natural Language Processing

≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational X

build a system to translate
create a QA system

engineering focus

Natural Language Processing

≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational X

Natural Language Processing \approx Computational Linguistics

Both have impact in/contribute to/draw from:

Machine learning

Linguistics

Information Theory

Cognitive Science

Data Science

Psychology

Systems Engineering

Political Science

Logic

Digital Humanities

Theory of Computation

Education

build a system to translate
create a QA system

engineering focus

Natural Language Processing

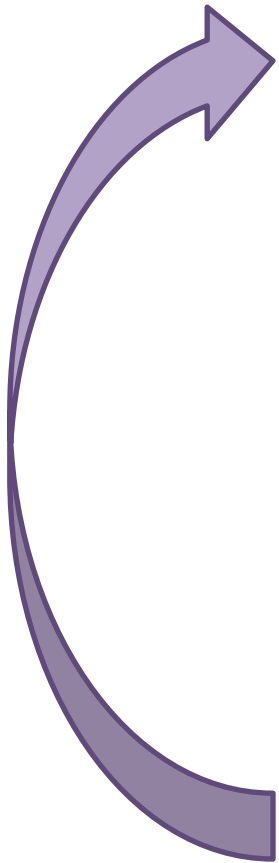
≈

Computational Linguistics

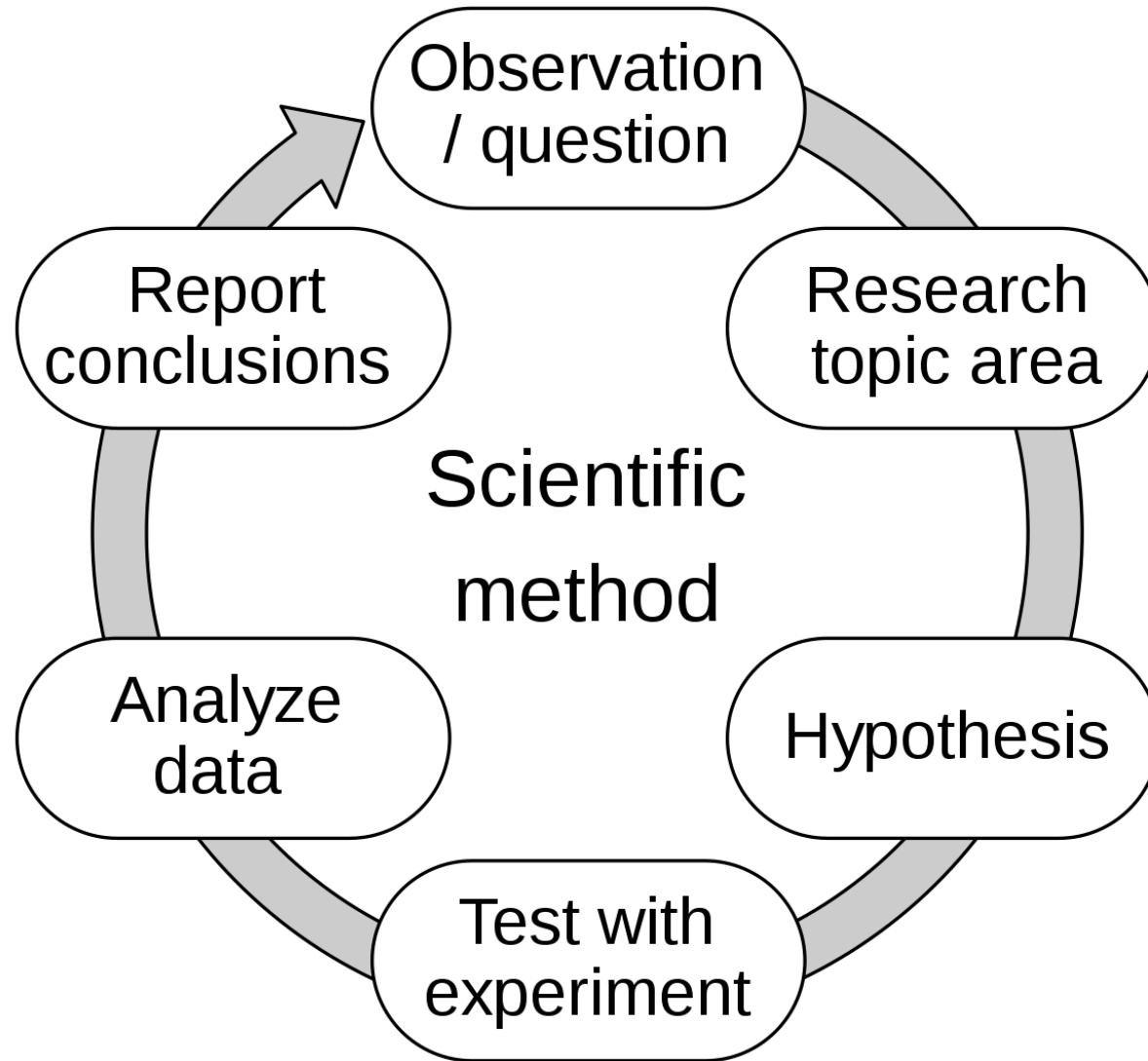
science focus

computational bio
computational chemistry
computational X

these views can co-exist peacefully



Whether it's **Natural Language Processing** or **Computational Linguistics...**



The NLP Research Community

- Papers

- ACL Anthology (<http://aclweb.org/anthology>) has nearly everything, free! As of 2023:
 - Over 87,000 papers (36k in 2016)!
 - Free-text searchable
 - Great way to learn about current research on a topic
 - New search interfaces currently available in beta
 - » Find recent or highly cited work; follow citations
 - Used as a dataset by various projects
 - Analyzing the text of the papers (e.g., parsing it)
 - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)






















The NLP Research Community

- **Conferences**




























- Most work in NLP is published as 8-page conference papers with 3 double-blind reviewers.
- Main annual conferences: ACL, EMNLP, NAACL
 - Also EACL, IJCNLP, COLING
 - + journals (TACL, *Computational Linguistics* [CL])
 - + various specialized conferences and workshops
- Big events, and growing fast! [ACL 2023](#):
 - About 3200 attendees (1500 in 2015)

Where Do We Observe Language?

- All around us
- NLP/CL: from a **corpus** (pl: corpora)
 - Literally a “body” of text
- In real life:
 - Through curators (the LDC)
 - From the web (scrape Wikipedia, Reddit, etc.)
 - Via careful human elicitation (lab studies, crowdsourcing)
 - From previous efforts
- In this class (partly): the **Universal Dependencies**

▶		Ukrainian	25K	LF	-		✓		
▶		Upper Sorbian	11K	LF	-				
▶		Urdu	138K	LF	-		✓		
▶		Uyghur	11K	-	-		✓		
▶		Vietnamese	43K	LF	-		✓		

Upcoming UD Treebanks

▶		Amharic	-	-	-	?	-		
▶		Armenian	-	-	-				
▶		Bangla	-	-	-				
▶		Bengali-DDS	-	-	-		✓		
▶		Cantonese	-	-	-				
▶		Chinese-HK	-	-	-				

<http://universaldependencies.org/>
part-of-speech & syntax for > 120 languages

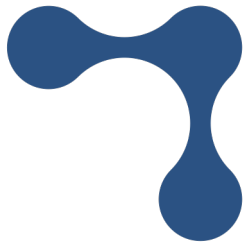
The NLP Research Community

- Standard tasks

- If you want people to work on your problem, make it easy for them to get started and to measure their progress. Provide:

- Test data, for evaluating the final systems
 - Development data, for measuring whether a change to the system helps, and for tuning parameters
 - An evaluation metric (formula for measuring how well a system does on the dev or test data)
 - A program for computing the evaluation metric
 - Labeled training data and other data resources
 - A prize? – with clear rules on what data can be used










The NLP Research Community: Standard Tasks (some of them)



GLUE

<https://gluebenchmark.com/>

GLUE Tasks

Name	Download
The Corpus of Linguistic Acceptability	
The Stanford Sentiment Treebank	
Microsoft Research Paraphrase Corpus	
Semantic Textual Similarity Benchmark	
Quora Question Pairs	
MultiNLI Matched	
MultiNLI Mismatched	
Question NLI	
Recognizing Textual Entailment	
Winograd NLI	
Diagnostics Main	

SuperGLUE Tasks

Name	Identifier
Broadcoverage Diagnostics	AX-b
CommitmentBank	CB
Choice of Plausible Alternatives	COPA
Multi-Sentence Reading Comprehension	MultiRC
Recognizing Textual Entailment	RTE
Words in Context	WiC
The Winograd Schema Challenge	WSC
BoolQ	BoolQ
Reading Comprehension with Commonsense Reasoning	ReCoRD
Winogender Schema Diagnostics	AX-g



<https://super.gluebenchmark.com/>

The NLP Research Community: Standard Tasks (some of them)

<https://semeval.github.io/SemEval2024/tasks>

(past years too!)

Semantic Relations

- **Task 1: Semantic Textual Relatedness for African and Asian Languages** ([contact organizers], [join task mailing list])
Nedjma OUSIDHOUM, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Krishnapriya Vishnubhotla, Vladimir Araujo, Meriem Beloucif, Idris Abdulmumin, Seid Muhie Yimam, Tamar Solorio, Monojit Choudhury, Saif M. Mohammad
- **Task 2: Safe Biomedical Natural Language Inference for Clinical Trials** ([contact organizers], [join task mailing list])
Mael Jullien, Marco Valentino, Andre Freitas

Discourse and Argumentation

- **Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations** ([contact organizers], [join task mailing list])
Rui Xia, Jianfei Yu, Fanfan Wang, Erik Cambria
- **Task 4: Multilingual Detection of Persuasion Techniques in Memes** ([contact organizers], [join task mailing list])
Dimitar Iliyanov Dimitrov, Giovanni Da San Martino, Fabrizio Silvestri, Preslav Nakov, Firoj Alam
- **Task 5: Argument Reasoning in Civil Procedure** ([contact organizers], [join task mailing list])
Lena Held, Ivan Habernal

LLM Capabilities

- **Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes** ([contact organizers], [join task mailing list])
Elaine Zosa, Raúl Vázquez, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, Timothee Mickus, Marianna Apidianaki
- **Task 7: NumEval: Numeral-Aware Language Understanding and Generation** ([contact organizers], [join task mailing list])
Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen
- **Task 8: Multidomain, Multimodal and Multilingual Machine-Generated Text Detection** ([contact organizers], [join task mailing list])
Yuxia Wang, Alham Fikri Aji, Artem Shelmanov, Chenxi Whitehouse, Petar Ivanov, Jonibek Mansurov, Jinyan Su, Tarek Mahmoud, Osama Mohammed Afzal, Preslav Nakov

Knowledge Representation and Reasoning

- **Task 9: BRAINTEASER: A Novel Task Defying Common Sense** ([contact organizers], [join task mailing list])
Yifan Jiang, Filip Ilievski, Kaixin Ma
- **Task 10: Emotion Discovery and Reasoning its Flip in Conversation** ([contact organizers], [join task mailing list])
Shivani Kumar, Md Shad Akhtar, Tanmoy Chakraborty, Erik Cambria

The NLP Research Community

- **Standard data formats**
 - Often just simple *ad hoc* text-file formats
 - Documented in a README; easily read with scripts
 - Some standards:
 - [Unicode](#) – strings in any language (see [ICU](#) toolkit)
 - PCM (.wav, .aiff) – uncompressed audio
 - BWF and AUP extend w/metadata; also many compressed formats
 - [XML](#) – documents with embedded annotations
 - [Text Encoding Initiative](#) – faithful digital representations of printed text
 - [Protocol Buffers](#), [JSON](#) – structured data
 - [UIMA](#) – “unstructured information management”; Watson uses it
 - Standoff markup: raw text in one file, annotations in other files (“ \exists noun phrase from byte 378—392”)
 - Annotations can be independently contributed & distributed

What Are Words?

bat



What Are Words?

bats



What Are Words?

Fledermaus

flutter mouse



What Are Words?

pişirdiler

They cooked it.

What Are Words?

pişmişlermişlerdi

They had it cooked it.

What Are Words?

):

What Are Words?

my leg is hurting nasty):



What Are Words?

add two cups (a pint): bring to a boil



What Are Words?

Hard to get agreement

(Human) Language-dependent

White-space separation is a sometimes okay (for
written English longform)

*Social media? Spoken vs. written? Other
languages?*

What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Vocabulary: the words (items) you know

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many of types & tokens appear in the above sentence?

Terminology: Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Types

- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens

- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .

Terminology: Tokens vs. Types

The film got a great opening and the film went on to become a hit .

Types

- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

Tokens

- The
- film
- got
- a
- great
- opening
- and
- the
- ~~film~~
- went
- on
- to
- become
- ~~a~~
- hit
- .

For your {task} how do you define tokens?

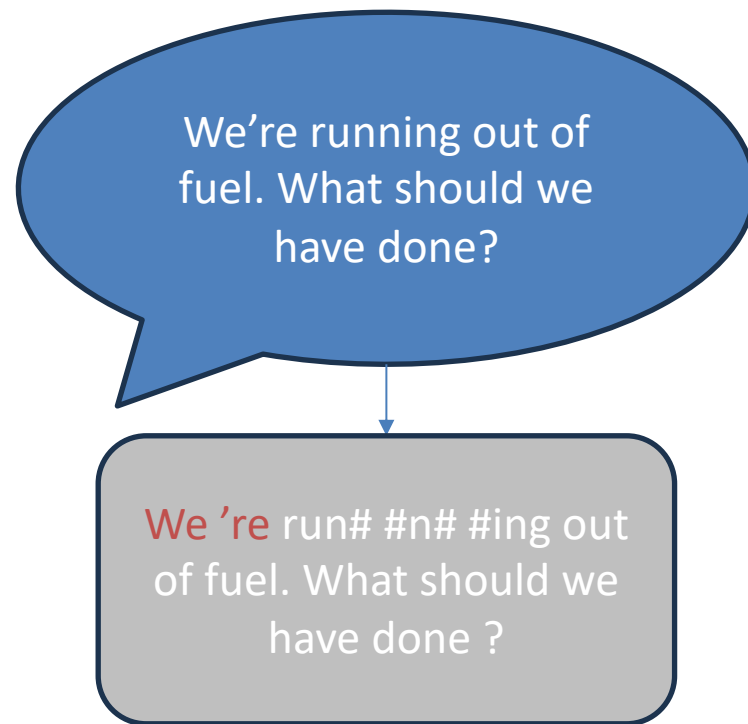
Sometimes:

1. They're defined for you by the *dataset creator*

2.

3.

4.



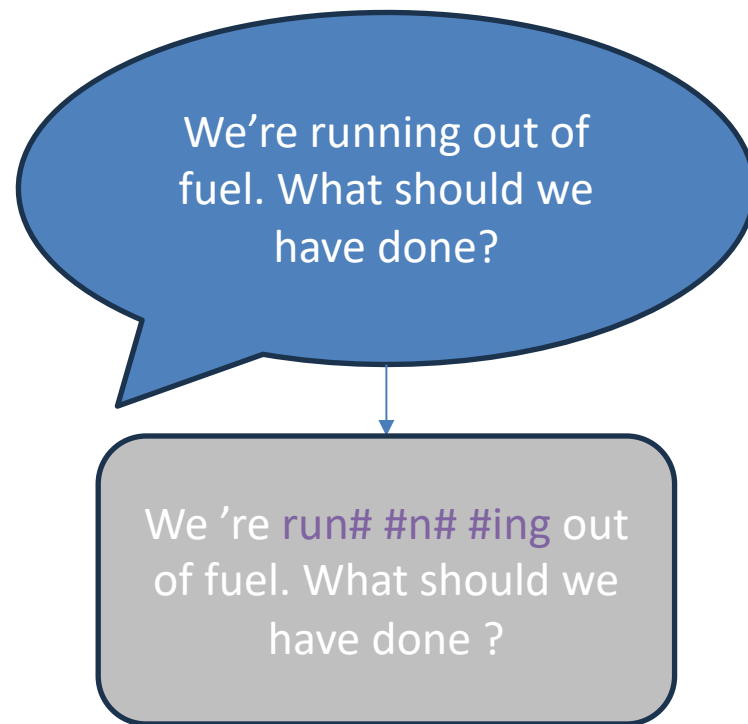
why?

- *scaleably handling novel words*
 - *linguistic reasons*
- *historical reasons / technical debt*

For your {task} how do you define tokens?

Sometimes:

1. They're defined for you by the *dataset creator*
2. They're defined by the *model*
- 3.
- 4.



(why? scaleably
handling novel words)

For your {task} how do you define tokens?

Sometimes:

1. They're defined for you by the *dataset creator*
2. They're defined by the *model*
3. It might be part of the *research problem itself*
- 4.

pişirdiler
They cooked it.

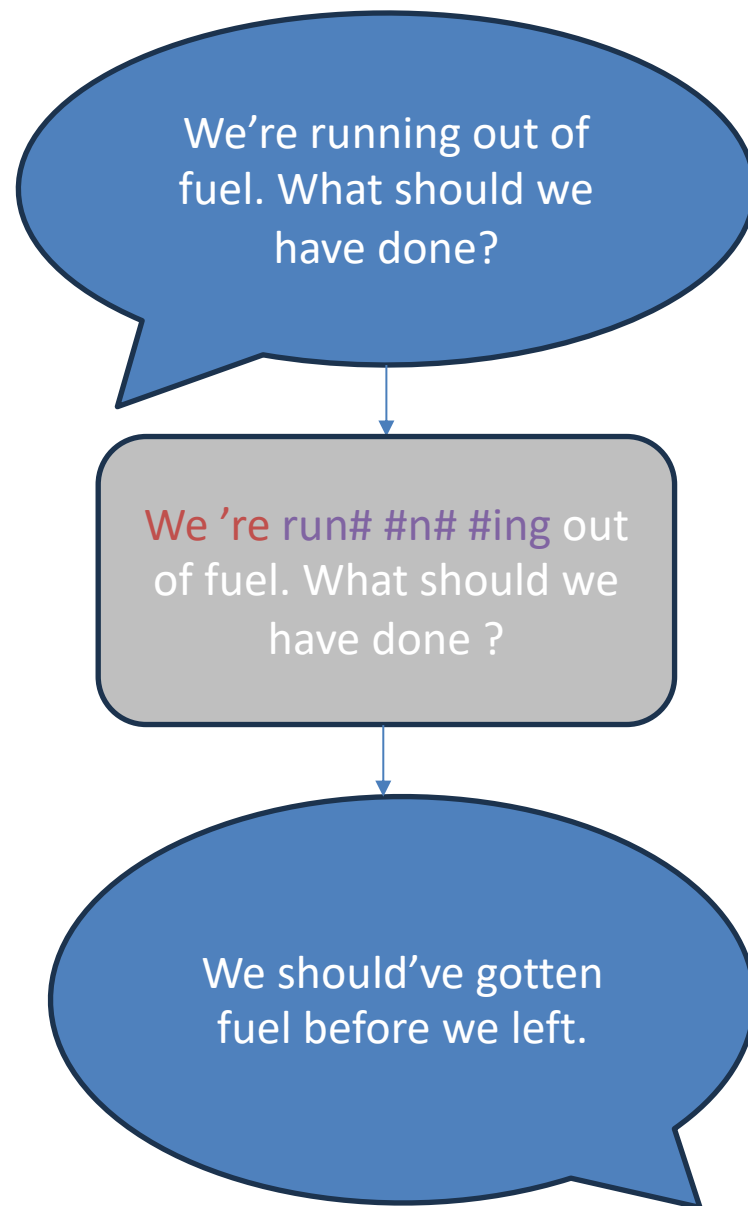
vs.

pişmişlermişlerdi
They had it cooked it.

For your {task} how do you define tokens?

Sometimes:

1. They're defined for you by the *dataset creator*
2. They're defined by the *model*
3. It might be part of the *research problem itself*
4. They're defined by the *end user*
 1. You'll need to handle points **1** and/or **2** on-the-backend...
 2. and then handling the reverse to present output to the user



Representing a Linguistic “Blob”

1. An array of sub-blobs

word → array of **characters**
sentence → array of **words**

*How do you
represent **these**?*

Representing a Linguistic “Blob”

1. An array of sub-blobs

word → array of **characters**
sentence → array of **words**

*How do you
represent **these**?*

2. Integer representation/one-hot encoding

3. Dense embedding

Representing a Linguistic “Blob”

1. An array of sub-blobs
word \rightarrow array of characters
sentence \rightarrow array of words
2. Integer
representation/one-hot
encoding
3. Dense embedding

Let V = vocab size (# types)

1. Represent each word *type*
with a unique integer i ,
where $0 \leq i < V$

Representing a Linguistic “Blob”

1. An array of sub-blobs
word \rightarrow array of characters
sentence \rightarrow array of words
2. Integer
representation/one-hot
encoding
3. Dense embedding

Let V = vocab size (# types)

1. Represent each word *type* with a unique integer i , where $0 \leq i < V$
2. Or equivalently, ...
 - Assign each word to some index i , where $0 \leq i < V$
 - Represent each word w with a V -dimensional **binary** vector e_w , where $e_{w,i} = 1$ and 0 otherwise

One-Hot Encoding Example

- Let our vocab be {a, cat, saw, mouse, happy}

Q: What is V (# types)?

One-Hot Encoding Example

- Let our vocab be {a, cat, saw, mouse, happy}
- $V = \# \text{ types} = 5$
- Assign:

a	4
cat	2
saw	3
mouse	0
happy	1

How do we
represent “cat?”

One-Hot Encoding Example

- Let our vocab be {a, cat, saw, mouse, happy}
- $V = \# \text{ types} = 5$
- Assign:

a	4
cat	2
saw	3
mouse	0
happy	1

How do we
represent "cat?"

$$e_{\text{cat}} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

How do we
represent
"happy?"

One-Hot Encoding Example

- Let our vocab be {a, cat, saw, mouse, happy}
- $V = \# \text{ types} = 5$
- Assign:

a	4
cat	2
saw	3
mouse	0
happy	1

How do we
represent "cat?"

$$e_{\text{cat}} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

How do we
represent
"happy?"

$$e_{\text{happy}} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

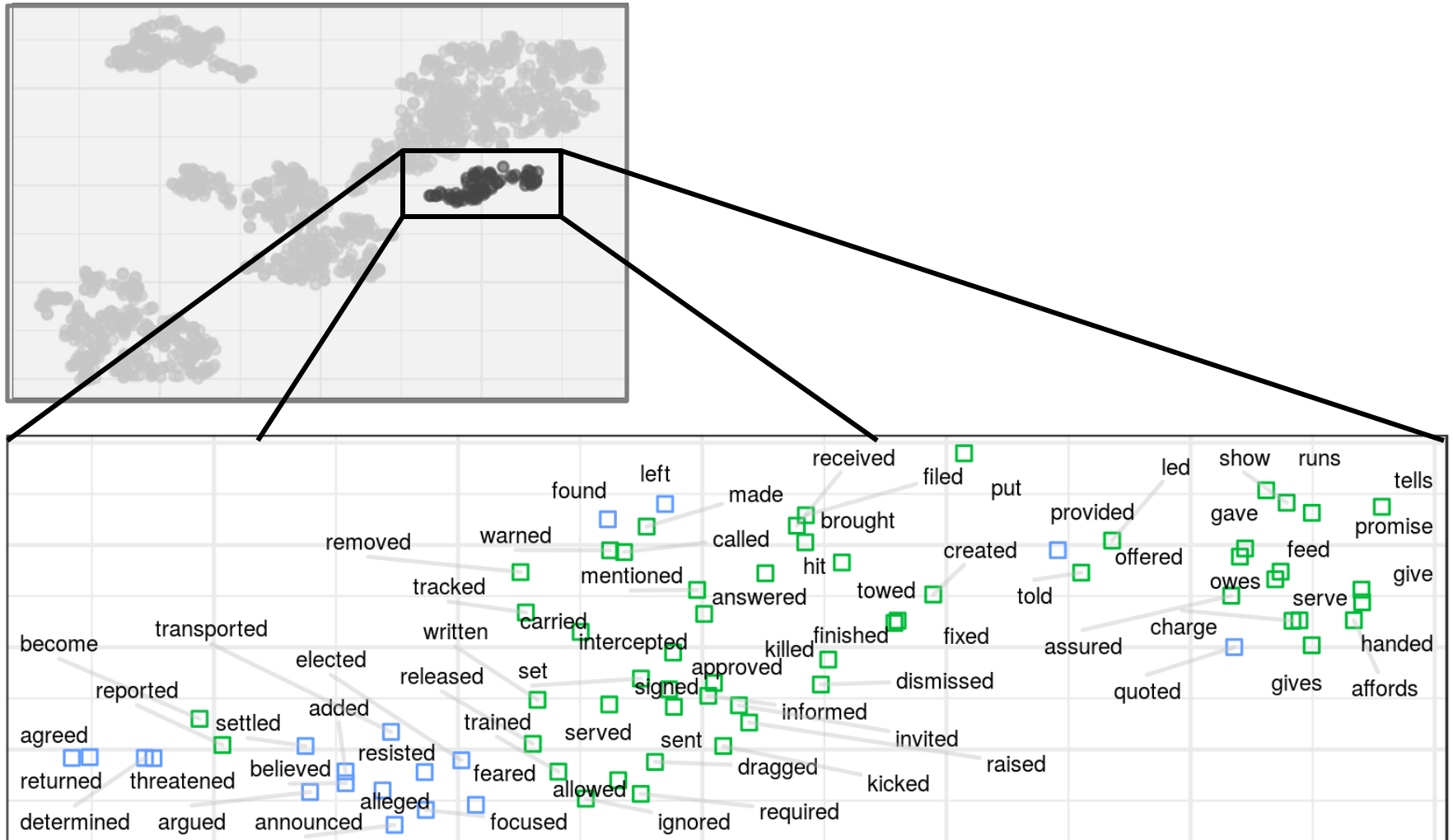
Representing a Linguistic “Blob”

1. An array of sub-blobs
word \rightarrow array of characters
sentence \rightarrow array of words
2. Integer
representation/one-hot
encoding
3. Dense embedding

Let E be some *embedding size* (often 100, 200, 300, etc.)

Represent each word w with an E -dimensional **real-valued** vector e_w

A Dense Representation ($E=2$)



Preview
for later!

(Some) Properties of Embeddings

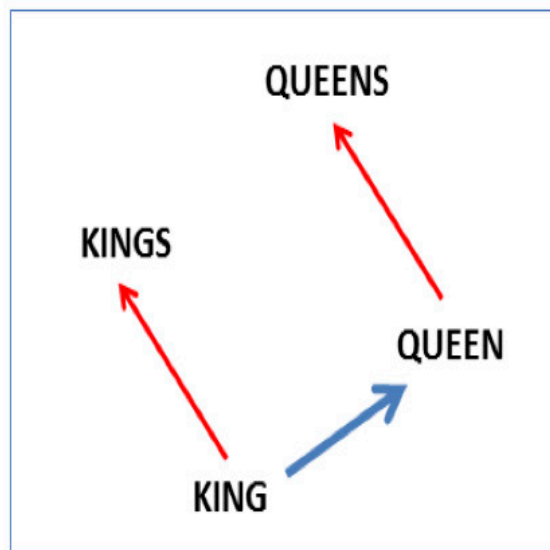
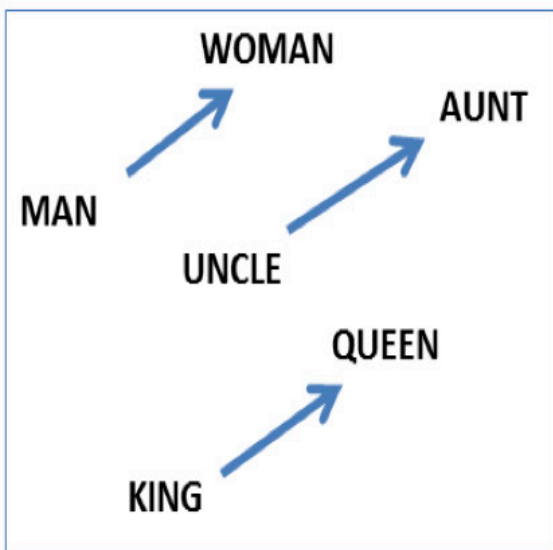
Capture “like” (similar) words



https://media3.giphy.com/media/3orif0M8U1E7NfpFzg/200_s.gif

target:	Redmond	Havel	ninjutsu	graffiti	capitulate
	Redmond Wash.	Vaclav Havel	ninja	spray paint	capitulation
	Redmond Washington	president Vaclav Havel	martial arts	grafitti	capitulated
	Microsoft	Velvet Revolution	swordsmanship	taggers	capitulating

Capture relationships



$\text{vector}('king') - \text{vector}('man') + \text{vector}('woman') \approx \text{vector}('queen')$

$\text{vector}('Paris') - \text{vector}('France') + \text{vector}('Italy') \approx \text{vector}('Rome')$

Preview
for later!



**T-REX IN:
"COMPUTATIONAL
LINGUISTICS"**



Computational linguistics is the study of computer-based language processing!



A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language - English, for example - can have more than one meaning!



Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?



Many have focused on statistical modelling of language, but this approach is approximate. I agree!



NLP \leftrightarrow Machine Learning

Goal: Learn parameters
(weights) θ to develop a
scoring function that
says how “good” some
provided text is

Michael Jordan,
coach Phil Jackson
and the star cast,
including Scottie
Pippen, took the
Chicago Bulls to six
National Basketball
Association
championships.

Goal: Learn parameters (weights) θ to develop a scoring
function that says how “good” some provided **text** is

$p_{\theta}(\text{Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.})$

Goal: Learn parameters (weights) θ to develop a **scoring function** that says how “good” some provided **text** is



Terminology: **parameters**: primary “knobs” of the model that are set by a learning algorithm

p_{θ} (

Michael Jordan,
coach Phil Jackson
and the star cast,
including Scottie
Pippen, took the
Chicago Bulls to six
National Basketball
Association
championships.

Goal: Learn **parameters (weights)** θ to develop a **scoring function** that says how “good” some provided **text** is

$$S = p_{\theta}(\text{Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.})$$

Goal: Learn parameters (weights) θ to develop a scoring function that says how “good” some provided text is

Q: If we make p to be a probability distribution, what are the minimum and maximum values of S ?

$$S = p_{\theta} \left(\text{Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.} \right)$$

Goal: Learn parameters (weights) θ to develop a scoring function that says how “good” some provided text is

Q: If we make p to be a probability distribution, what are the minimum and maximum values of s ?

$$A: 0 \leq s \leq 1$$

$$s = p_{\theta} \left(\text{Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.} \right)$$

Goal: Learn parameters (weights) θ to develop a scoring function that says how “good” some provided text is

Use ML Techniques to Learn the Weights

(probabilistic) model

$$p_{\theta}(X)$$

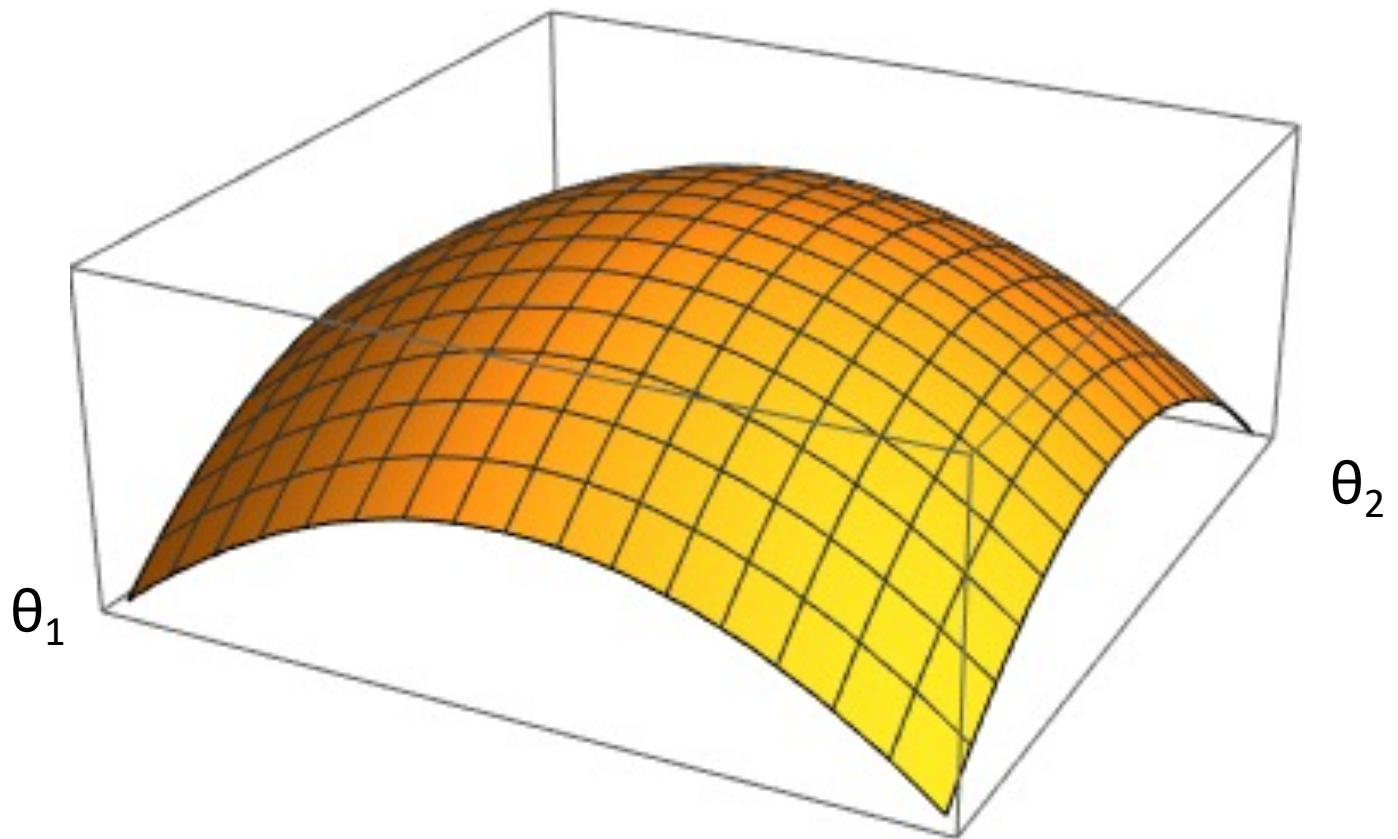


objective

$$F(\theta)$$

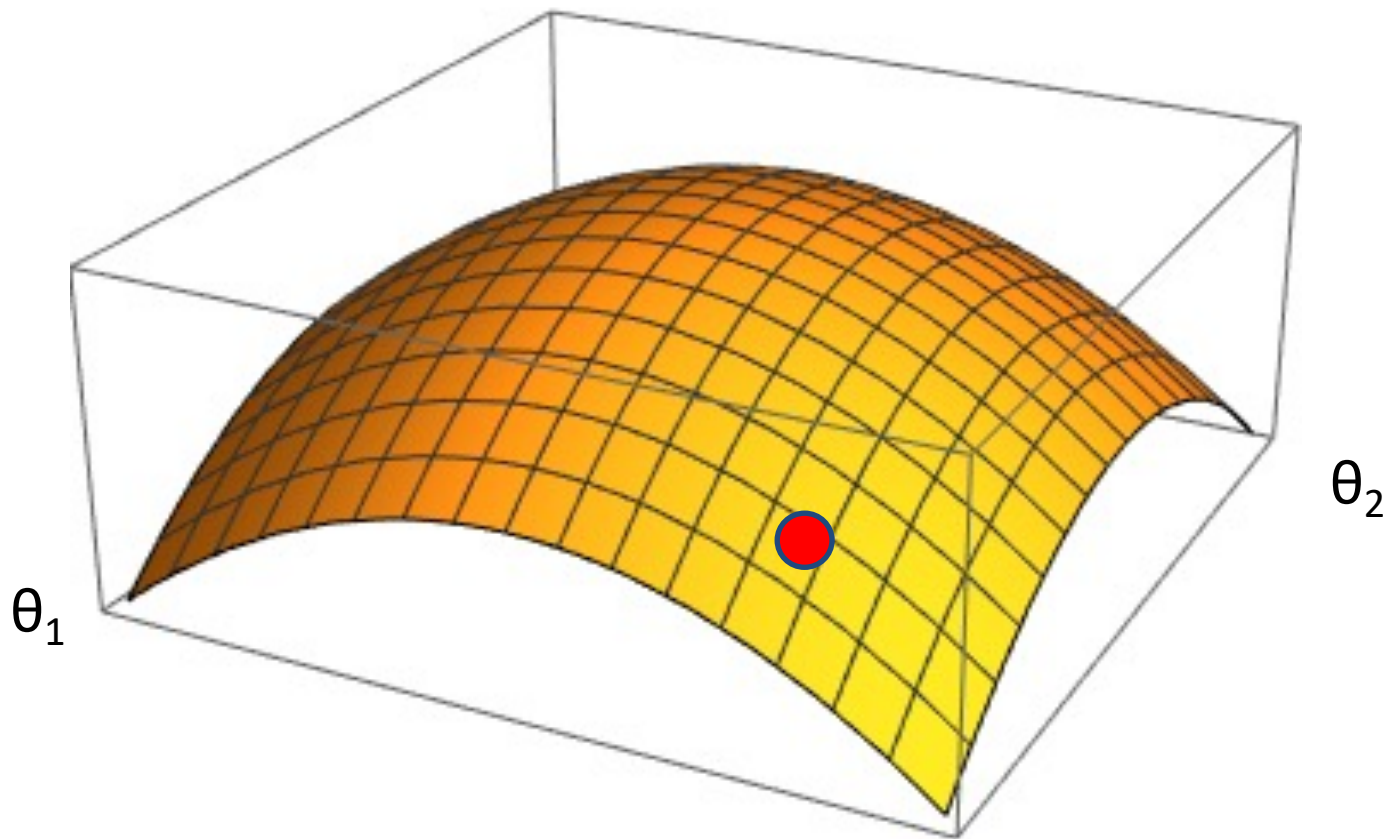
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$



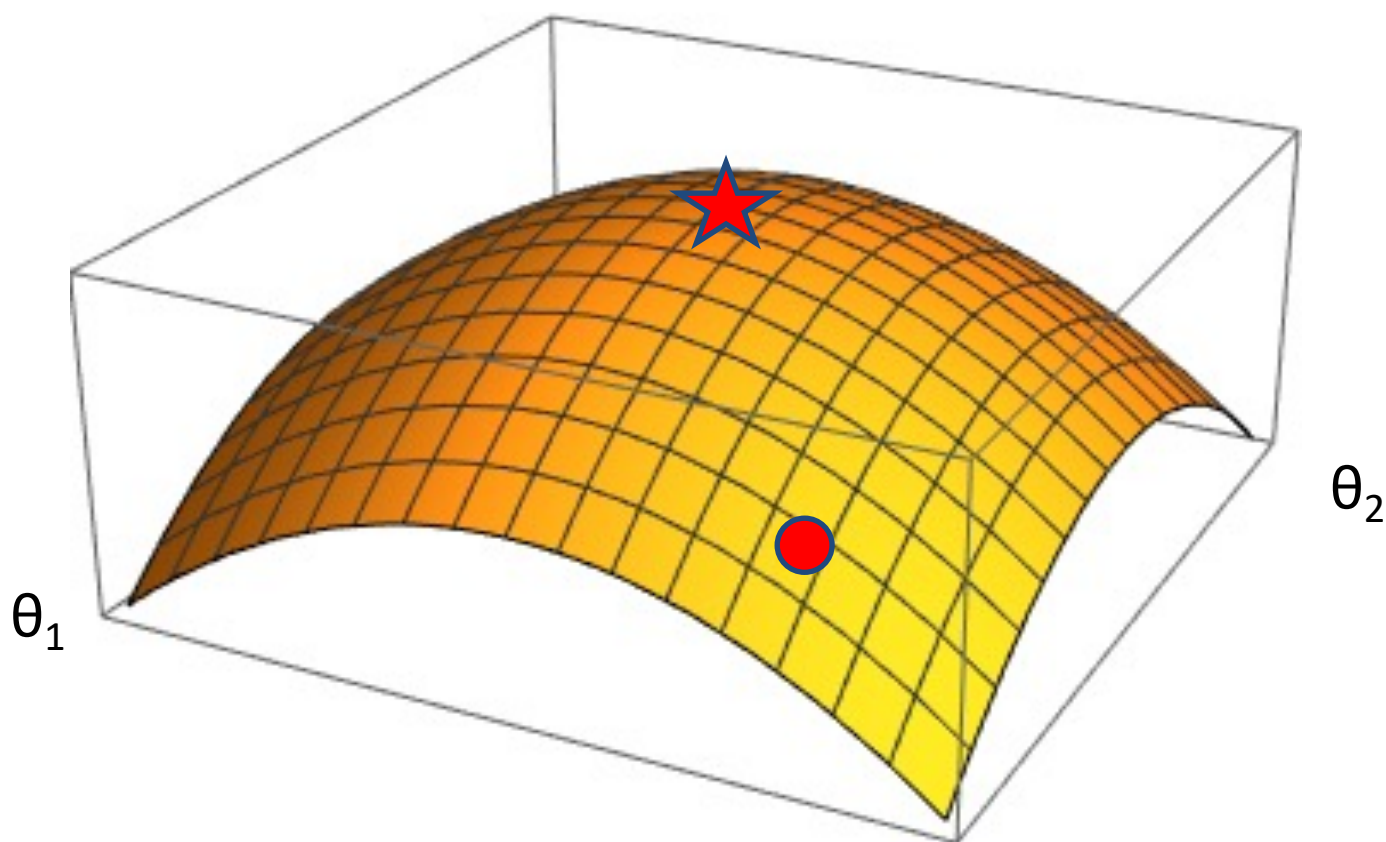
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$



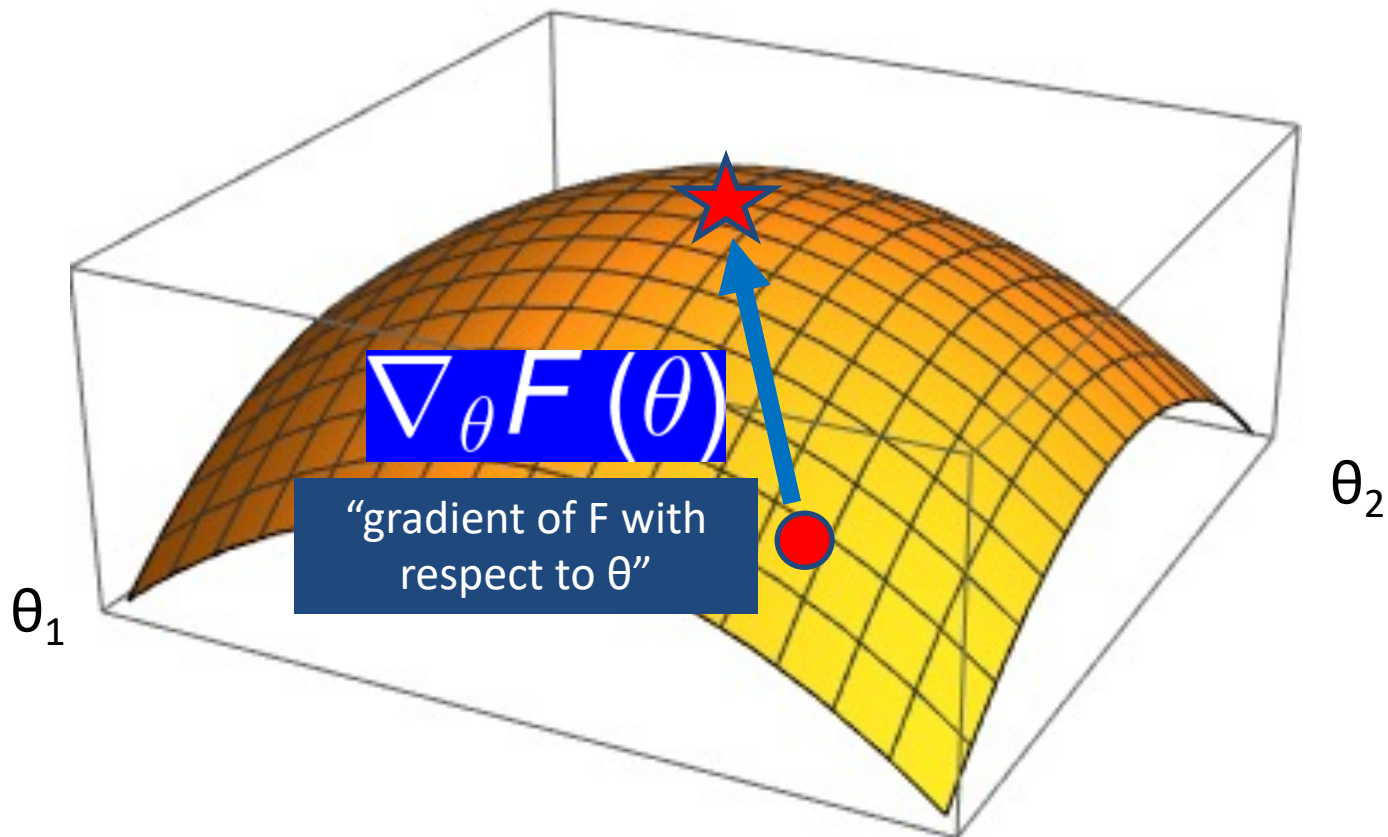
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$



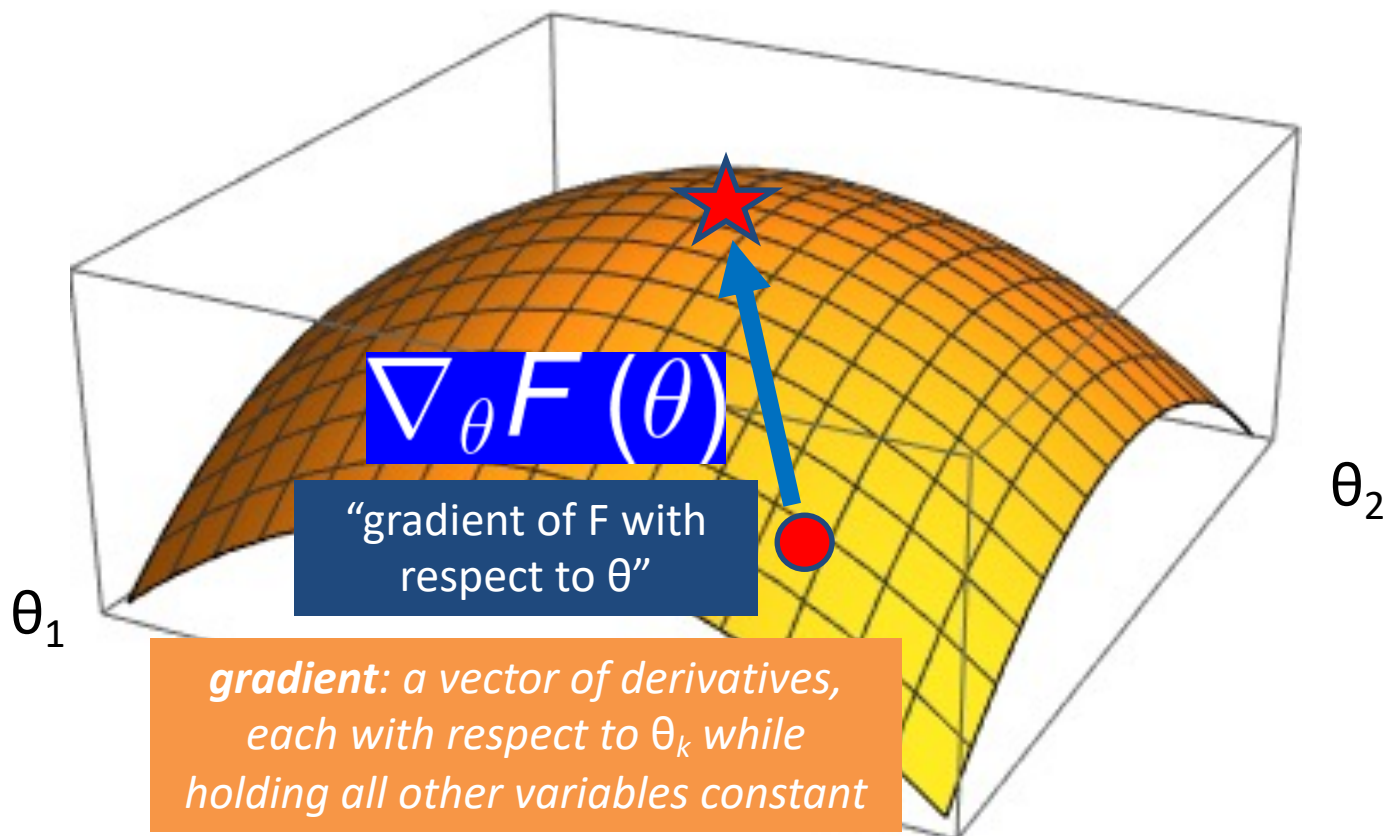
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$



Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$



In many cases, pick your toolkit

PyTorch
Huggingface
TensorFlow
Deeplearning4j
DyNet
Caffe

Keras
MxNet
Gluon
CNTK
...

Comparisons:

https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

<https://deeplearning4j.org/compare-dl4j-tensorflow-pytorch>

<https://github.com/zer0n/deepframeworks> (older---2015)

My Hope

Help you learn the ropes...



... and apply your knowledge using whatever tools your org. uses!



... so you can go into a job...

```
from theano import *
```

keras
torch

What I actually do

Toolkit Basics

- Machine learning involves working with data
 - analyzing, manipulating, transforming, ...
- More often than not, it's numeric or has a natural numeric representation
- Natural language text is an exception, but this too can have a numeric representation
- A common data model is as a N-dimensional matrix or tensor
- These are supported in Python via libraries

Typical Python Libraries

numpy, scipy

- Basic mathematical libraries for dealing with matrices and scientific/mathematical functions

pandas, matplotlib

- Libraries for data science & plotting

sklearn (scikit-learn)

- A whole bunch of implemented classifiers

torch (pytorch) and tensorflow

- Frameworks for building neural networks

Lots of
documentation
available for all
of these online!

The NLP Research Community

- **Software**
 - Lots of people distribute code for these tasks
 - Or you can email a paper's authors to ask for their code
 - Some [lists](#) of software, but no central site ☹️
 - Some [end-to-end pipelines](#) for text analysis
 - “One-stop shopping”
 - Cleanup/tokenization + morphology + tagging + parsing + ...
 - [NLTK](#)
 - Spacy
 - Huggingface

The NLP Research Community

- **Software**

- To find good or popular tools:
 - Search current papers, ask around, use the web
- Still, often hard to identify the **best** tool for your job:
 - Produces appropriate, sufficiently detailed output?
 - Accurate? (on the measure you care about)
 - Robust? (accurate on your data, not just theirs)
 - Fast?
 - Easy and flexible to use? Nice file formats, command line options, visualization?
 - Trainable for new data and languages? How slow is training?
 - Open-source and easy to extend?

The NLP Research Community

- **Datasets**

- Raw text or speech corpora
 - Or just their [n-gram counts](#), for super-big corpora
 - Various languages and genres
 - Usually there's some metadata (each document's date, author, etc.)
 - Sometimes \exists licensing restrictions (proprietary or copyright data)
- Text or speech with manual or automatic annotations
 - What kind of annotations? That's the rest of this lecture ...
 - May include translations into other languages
- Words and their relationships
 - [Morphological](#), [semantic](#), translational, evolutionary
- [Grammars](#)
- [World Atlas of Linguistic Structures](#)
- Parameters of statistical models (e.g., grammar weights)

The NLP Research Community

- Datasets

- Read papers to find out what datasets others are using
 - [Linguistic Data Consortium](#) (searchable) hosts many large datasets
 - Many projects and competitions post data on their websites
 - But sometimes you have to email the author for a copy
- [CORPORA mailing list](#) is also good place to ask around
- [LREC Conference](#) publishes papers about new datasets & metrics
- [Amazon Mechanical Turk](#) – pay humans (very cheaply) to annotate your data or to correct automatic annotations
 - **Old task, new domain:** Annotate parses etc. on *your* kind of data
 - **New task:** Annotate something new that you want your system to find
 - **Auxiliary task:** Annotate something new that your system may benefit from finding (e.g., annotate subjunctive mood to improve translation)

The NLP Research Community

- **Standard data formats**
 - Often just simple *ad hoc* text-file formats
 - Documented in a README; easily read with scripts
 - Some standards:
 - [Unicode](#) – strings in any language (see [ICU](#) toolkit)
 - PCM (.wav, .aiff) – uncompressed audio
 - BWF and AUP extend w/metadata; also many compressed formats
 - [XML](#) – documents with embedded annotations
 - [Text Encoding Initiative](#) – faithful digital representations of printed text
 - [Protocol Buffers](#), [JSON](#) – structured data
 - [UIMA](#) – “unstructured information management”; Watson uses it
 - Standoff markup: raw text in one file, annotations in other files (“ \exists noun phrase from byte 378—392”)
 - Annotations can be independently contributed & distributed

The NLP Research Community

- **Survey articles**
 - May help you get oriented in a new area
 - Synthesis Lectures on Human Language Technologies
 - Handbook of Natural Language Processing
 - Oxford Handbook of Computational Linguistics
 - Foundations & Trends in Machine Learning
 - Survey articles in journals – JAIR, CL, JMLR
 - ACM Computing Surveys?
 - Online tutorial papers
 - Slides from tutorials at conferences
 - Textbooks

Today's Learning Goals

- NLP vs. CL
- Terminology:
 - NLP: vocabulary, token, type, one-hot encoding, dense embedding, parameter/weight, corpus/corpora
 - Linguistics: lexeme, morphology, syntax, semantics, “discourse”
- Universal Dependencies

