How to Read *War and Peace* in 30 Seconds

Denis Griffis
PhD Student, Speech and Language Technologies lab
The Ohio State University

How to Read *War and Peace* in 30 Seconds

Or: An introduction to Natural Language Processing

Denis Griffis
PhD Student, Speech and Language Technologies lab
The Ohio State University

THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING

speech & language technologies @osu

+

=

 +  = 

+

=

 +  = 

 +  =

 +  = 

 +  = 


Verizon 3G
12:59 PM
Sorry. I missed that.

How the heck do we get a computer to understand text?

Natural language processing

From Wikipedia, the free encyclopedia

> *This article is about language processing by computers. For the processing of language by the human brain, see Language processing.*

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many

OK Google,
where are all
the cats?

OK Google, where are all the cats?

```
SELECT CurrentLocation
FROM AllAnimals
WHERE AnimalType =
'Cat'
```

Speech Recognition

OK Google, where are all the cats?

SELECT CurrentLocation FROM AllAnimals WHERE AnimalType = 'Cat'

Speech Recognition

Natural Language Processing

OK Google, where are all the cats?

```
SELECT CurrentLocation
FROM AllAnimals
WHERE AnimalType =
'Cat'
```

“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250”

*“Yess! Yess! Its official [**Nintendo**] announced today that they Will release the [**Nintendo 3DS**] in [**north America**] [**march 27**] for [**$250**]”*

*“Yess! Yess! Its official [**Nintendo**] announced today that they Will release the [**Nintendo 3DS**] in [**north America**] [**march 27**] for [**$250**]”*

Company	*Product*	*Date*	*Price*	*Region*
Nintendo	Nintendo 3DS	March 27	$250	North America

*“Yess! Yess! Its official [**Nintendo**] announced today that they Will release the [**Nintendo 3DS**] in [**north America**] [**march 27**] for [**$250**]”*

Company	*Product*	*Date*	*Price*	*Region*
Nintendo	Nintendo 3DS	March 27	$250	North America

Natural language *understanding*

Company	*Product*	*Date*	*Price*	*Region*
Nintendo	Nintendo 3DS	March 27	$250	North America

Nintendo will release the Nintendo 3DS in North America on March 27 for $250.

Company	*Product*	*Date*	*Price*	*Region*
Nintendo	Nintendo 3DS	March 27	$250	North America

Nintendo will release the Nintendo 3DS in North America on March 27 for $250.

Company	*Product*	*Date*	*Price*	*Region*
Nintendo	Nintendo 3DS	March 27	$250	North America

Natural language *generation*

A Brief History of NLP

A Brief History of NLP

< *1950*
Codes for
translation;
message
encodings

A Brief History of NLP

< 1950
Codes for translation; message encodings

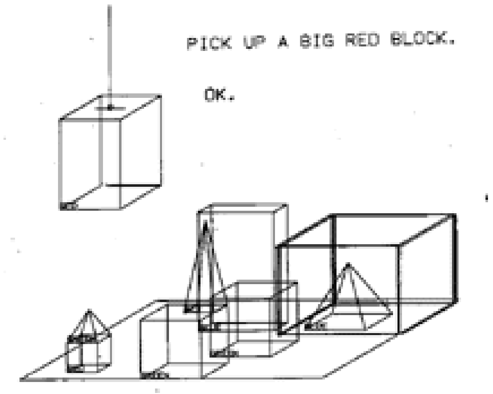
1950s
Turing, Chomsky; Start of machine translation

A Brief History of NLP

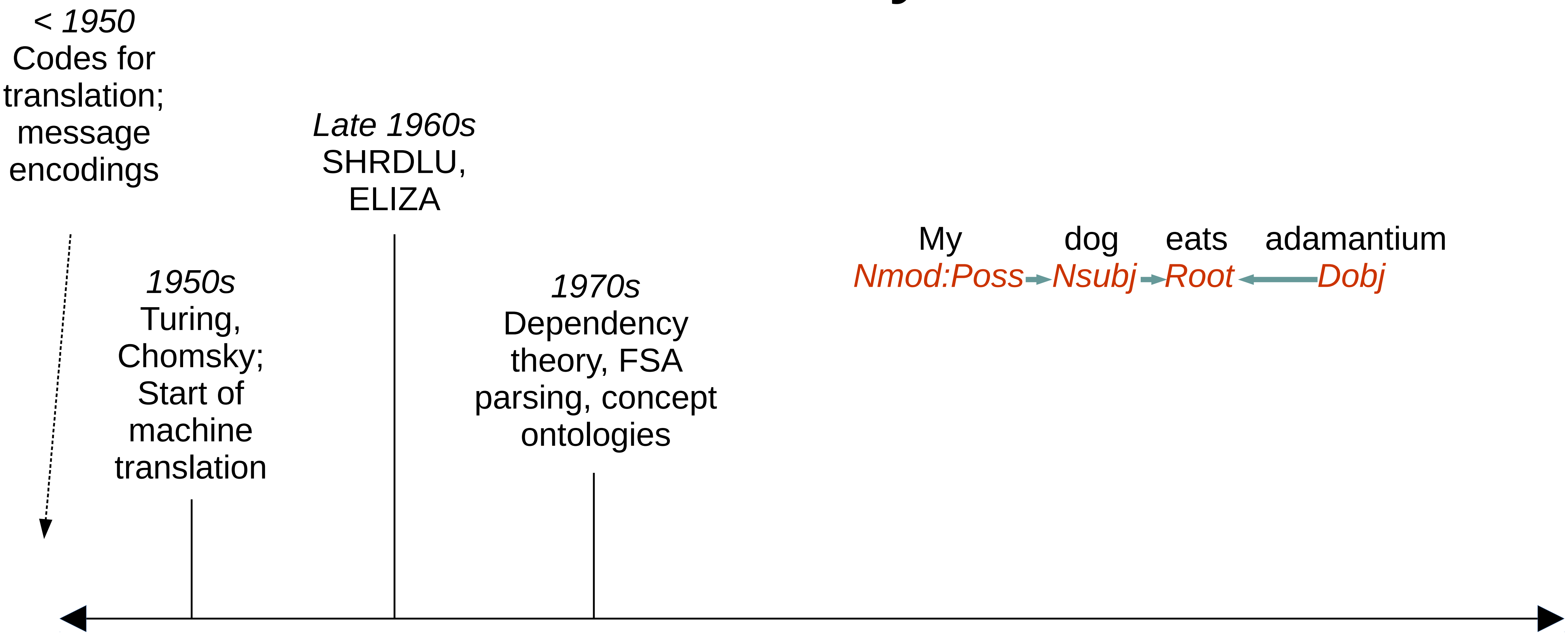
< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA



A Brief History of NLP



A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

1990s
Neural networks, high-performance computing

A Brief History of NLP

< 1950
Codes for translation; message encodings

1950s
Turing, Chomsky; Start of machine translation

Late 1960s
SHRDLU, ELIZA

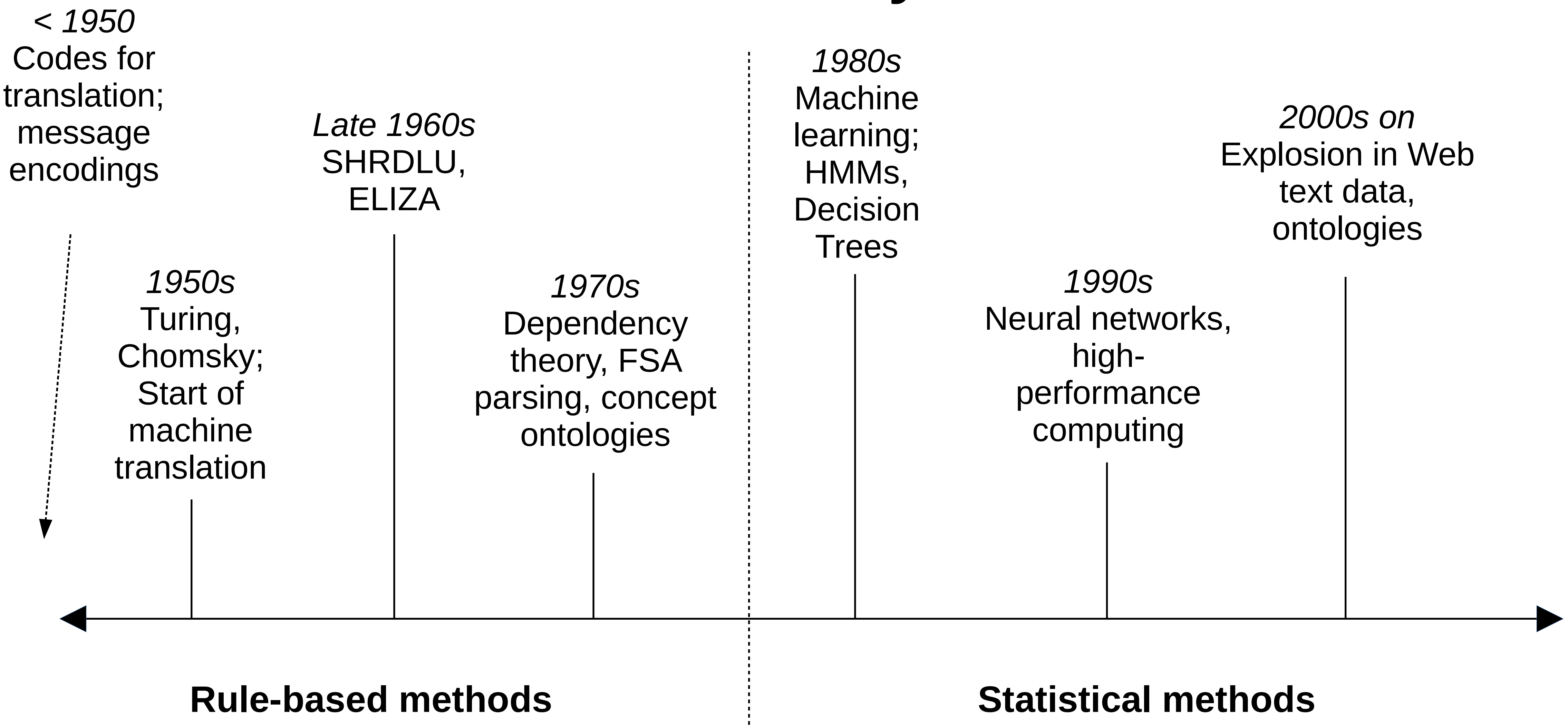
1970s
Dependency theory, FSA parsing, concept ontologies

1980s
Machine learning; HMMs, Decision Trees

1990s
Neural networks, high-performance computing

2000s on
Explosion in Web text data, ontologies

A Brief History of NLP



Rule-Based NLP

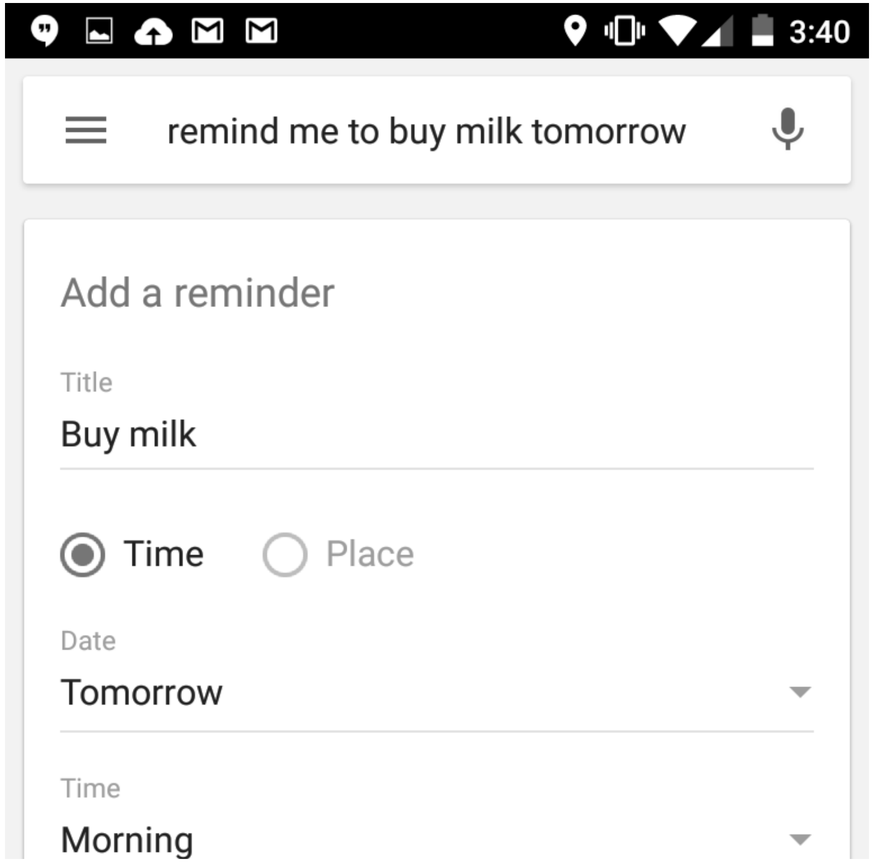
- Regular Expressions

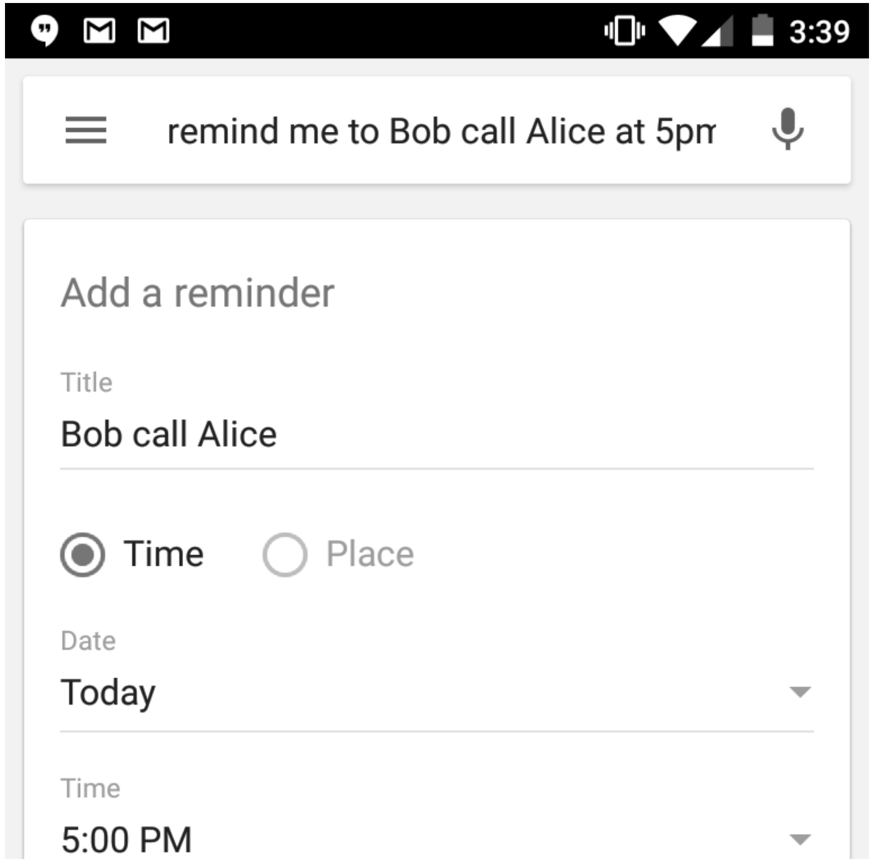
```
############################################
#     Per cause_of_death
############################################

{
        ruletype: "composite",
        pattern: (([{ner:PERSON}]+) /died/ /of|from/ /a/? ([{tag:NN}]+)),
        result: Format("per:cause_of_death(%s,%s)", $1.word, $2.word),
        action: (Annotate($1, kbp, "per"), Annotate($2, kbp, "per_cause_of_death"))
}
```

Example from Alan Ritter

Rule-Based NLP

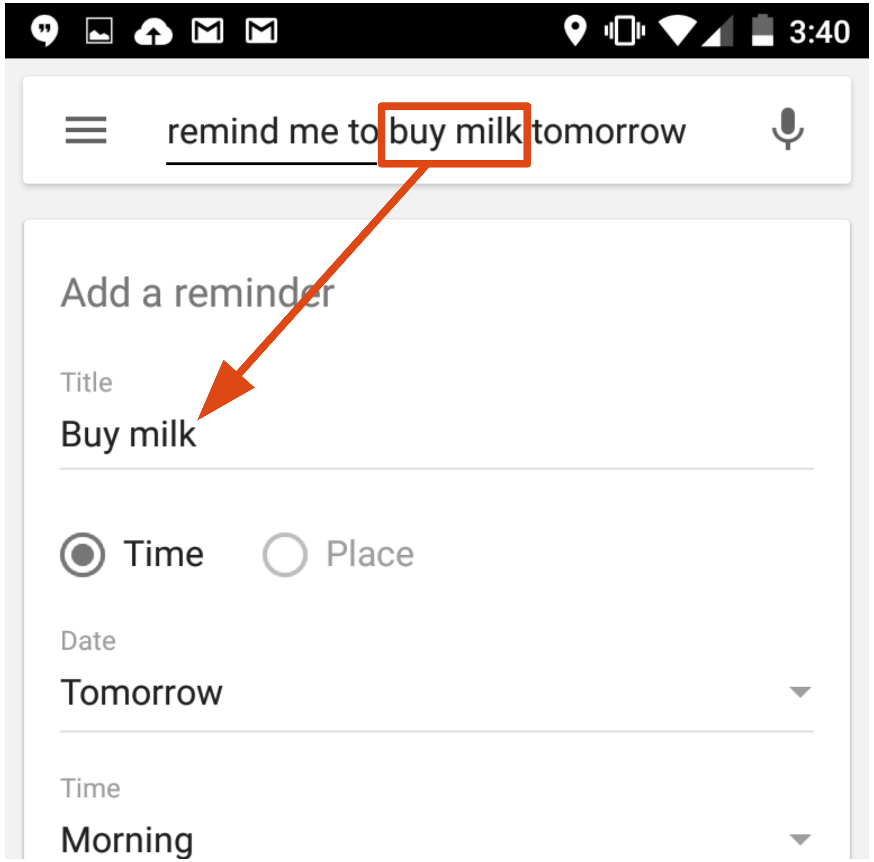
- Keywords and arguments

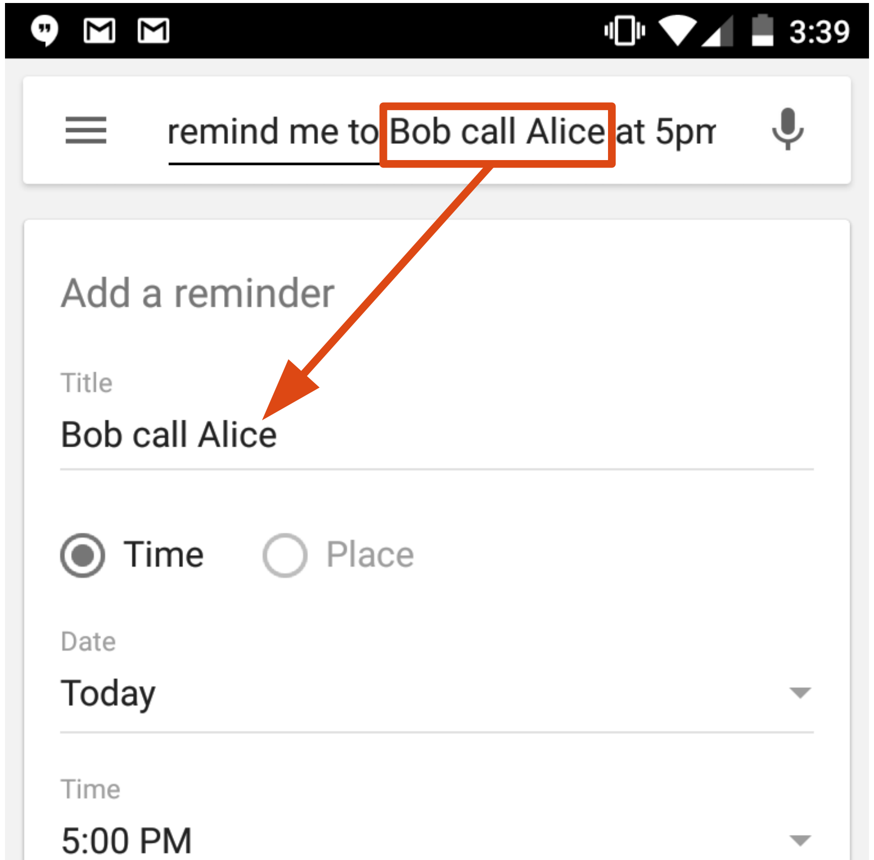




Rule-Based NLP

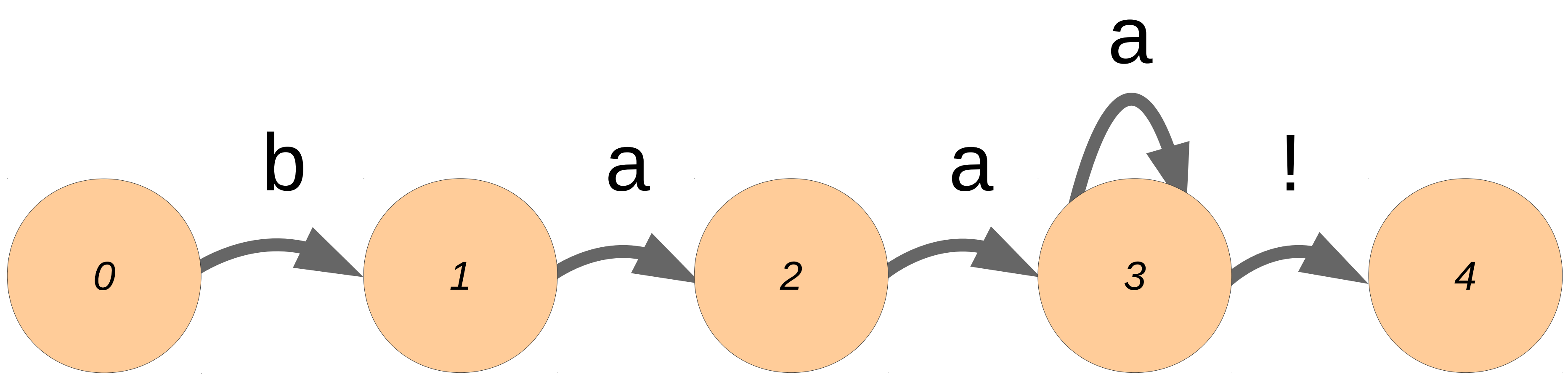
- Keywords and arguments





Rule-Based NLP

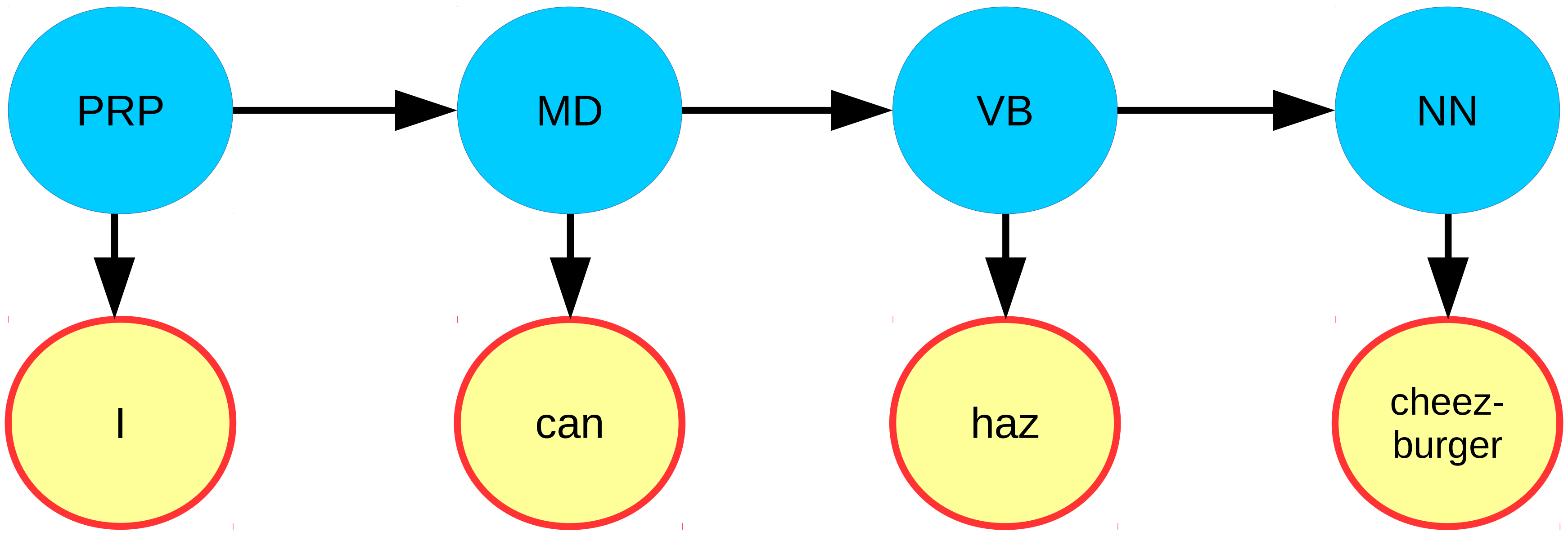
- Finite State Automata



Example from “Speech and Language Processing”, Jurafsky and Martin, 2009

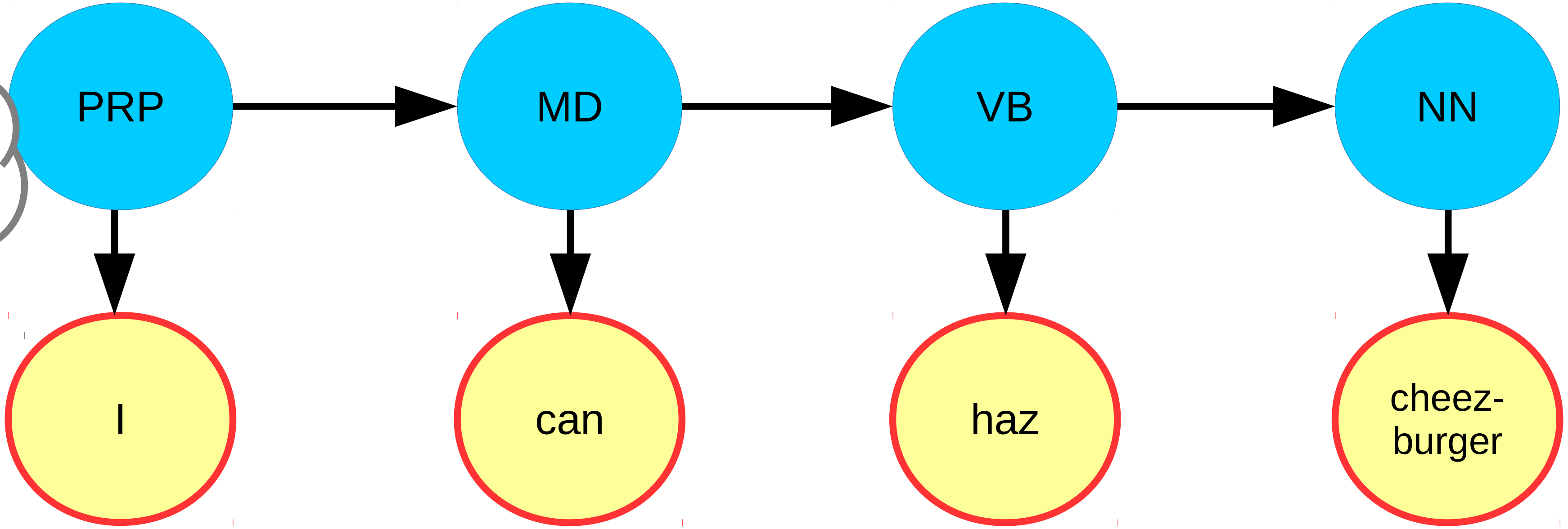
Statistical NLP

- Hidden Markov Models (HMMs)



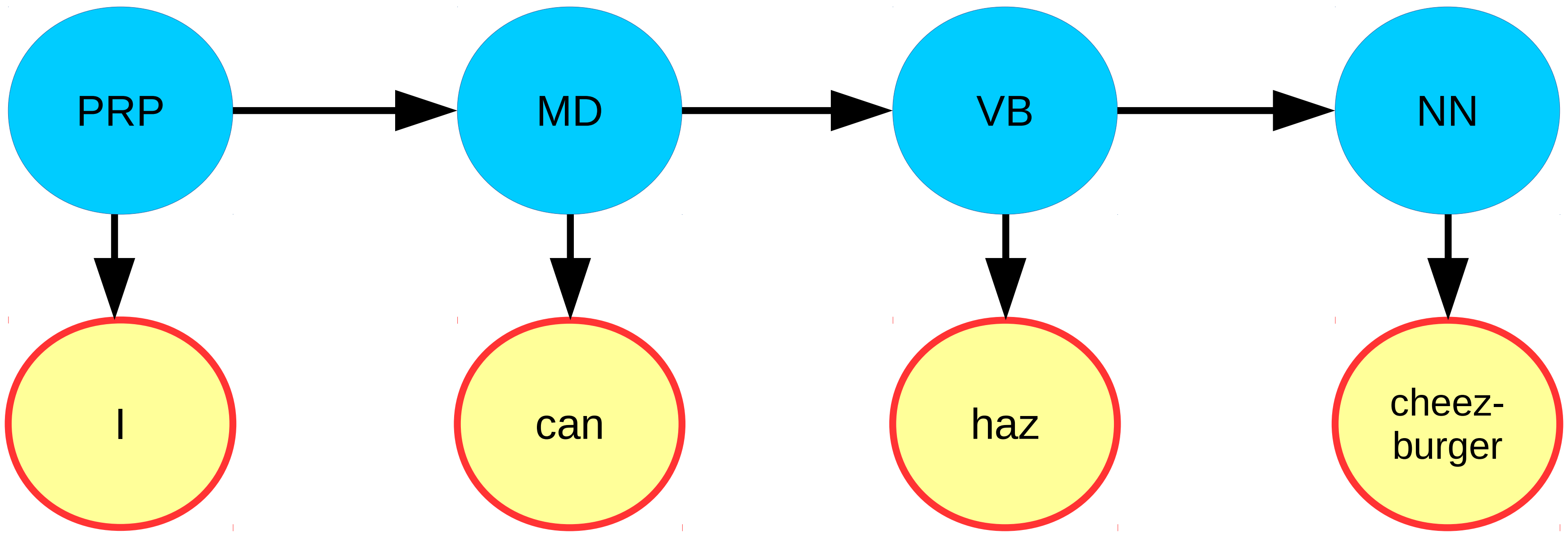
Statistical NLP

- Hidden Markov Models (HMMs)



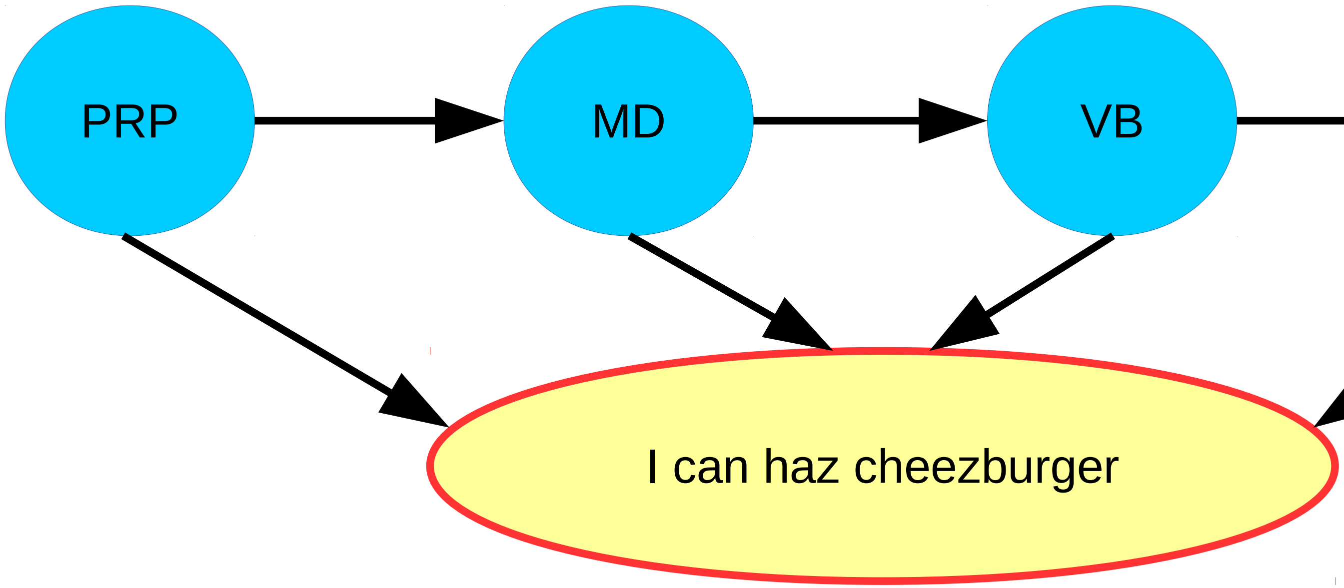
Statistical NLP

- Hidden Markov Models (HMMs)



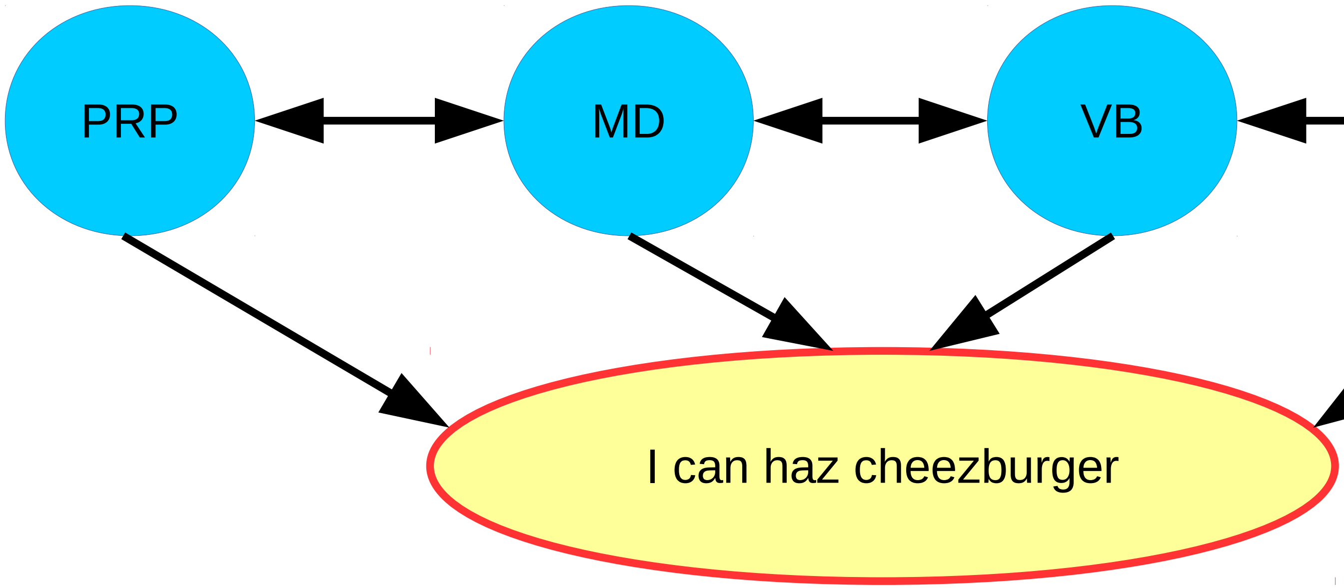
Statistical NLP

- Hidden Markov Models (HMMs); **MEMMs**



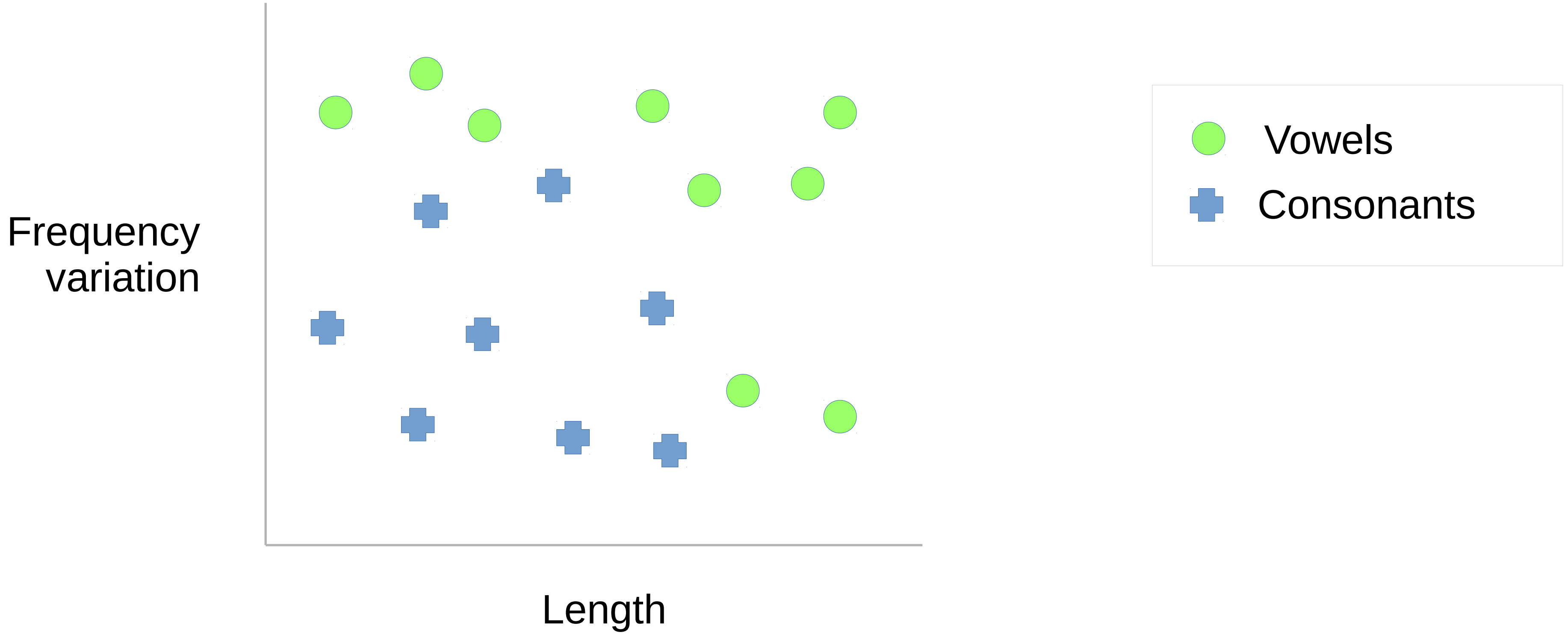
Statistical NLP

- Hidden Markov Models (HMMs); MEMMs, **CRFs**



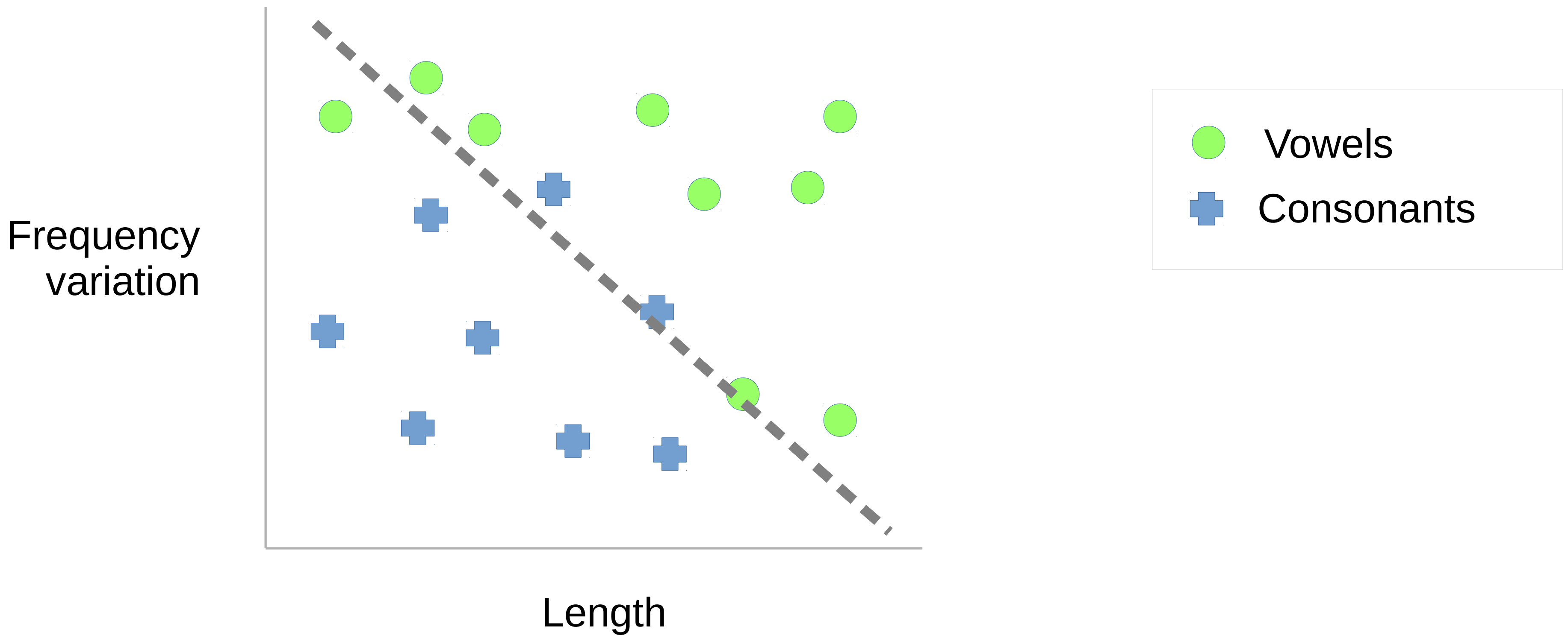
Statistical NLP

- Support Vector Machines (SVMs)



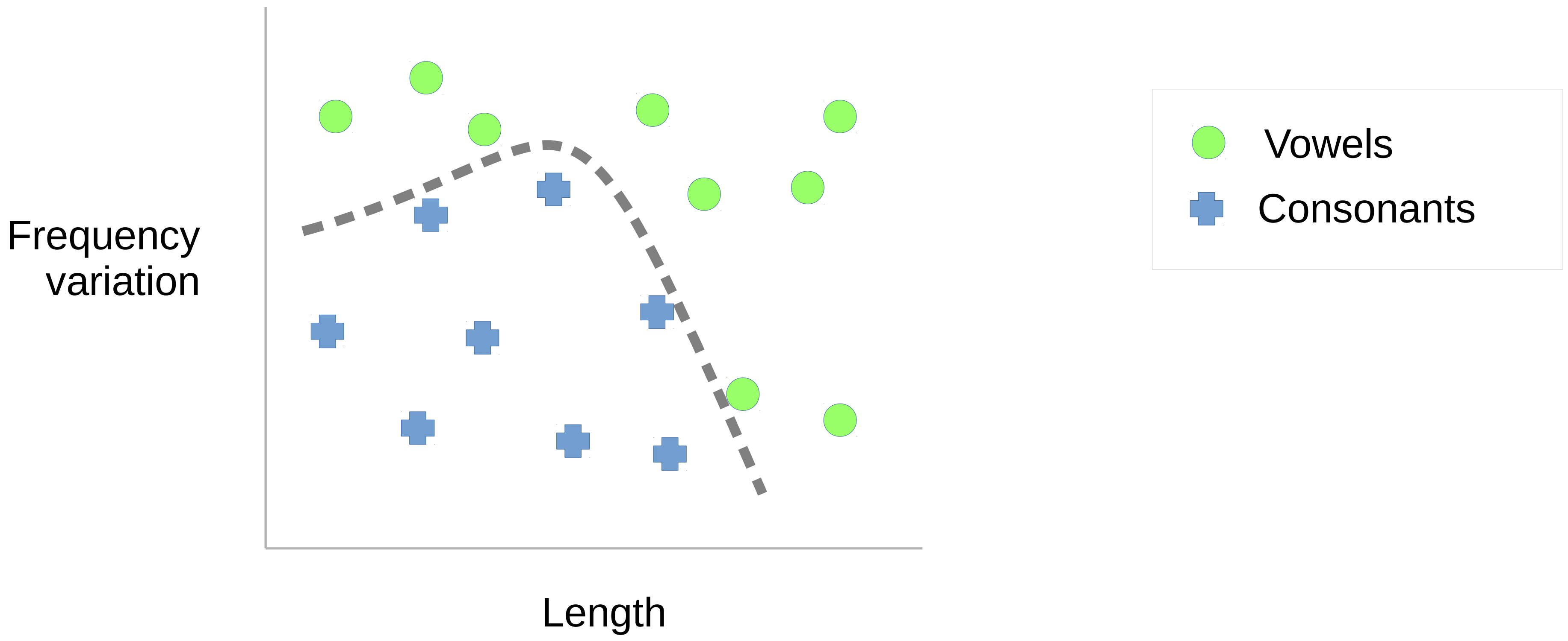
Statistical NLP

- Support Vector Machines (SVMs)



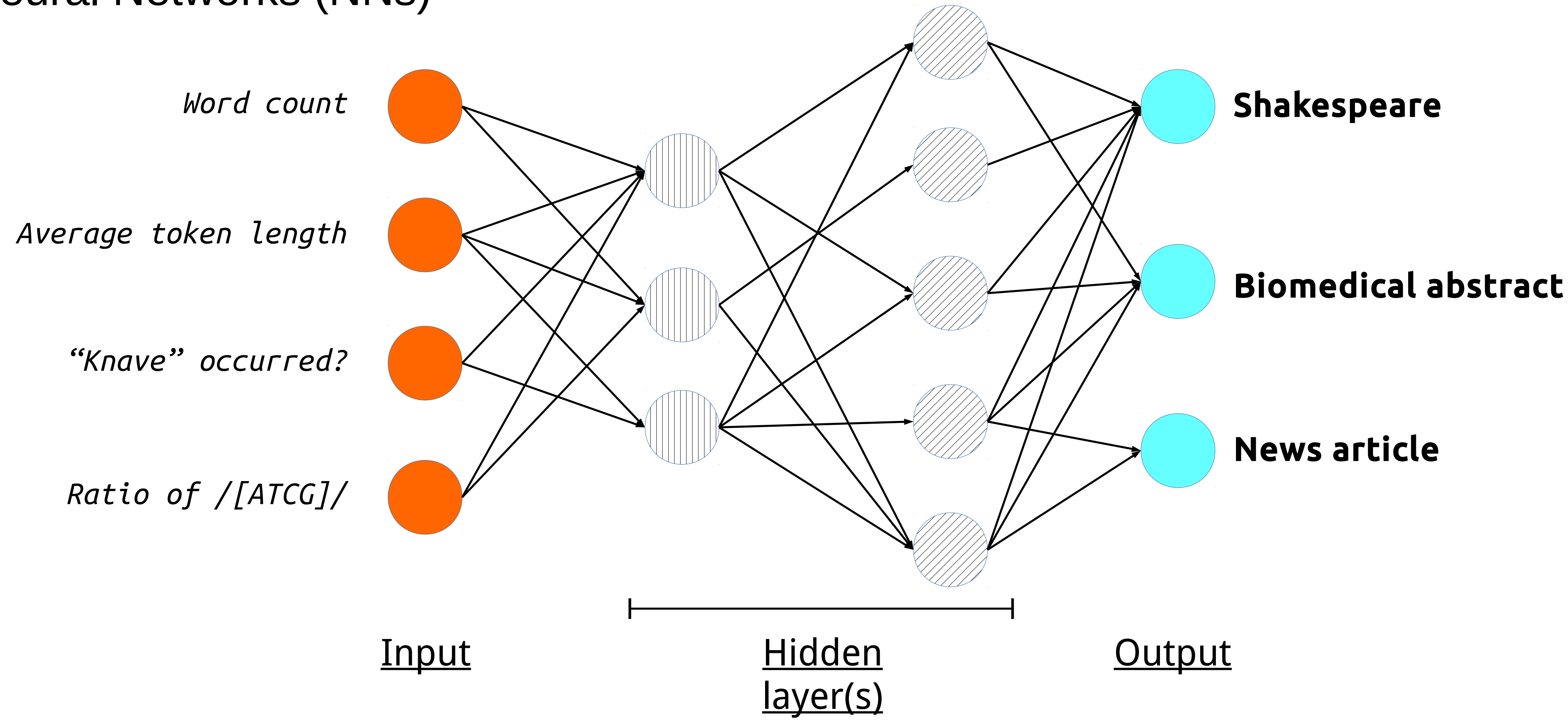
Statistical NLP

- Support Vector Machines (SVMs)



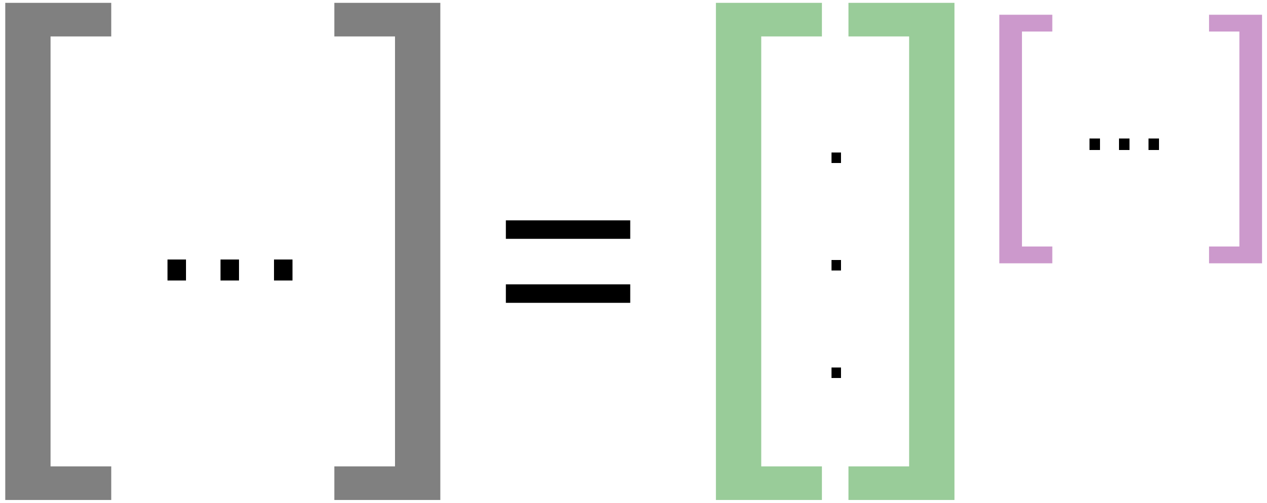
Statistical NLP

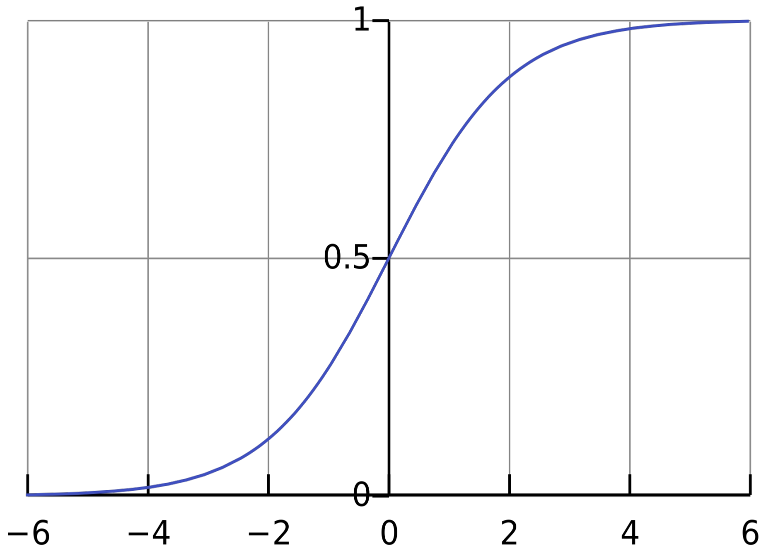
- Neural Networks (NNs)



Statistical NLP

- Other methods: matrix factorization, logistic regression, etc.





~~Rule-Based~~ NLP ~~Statistical~~ NLP

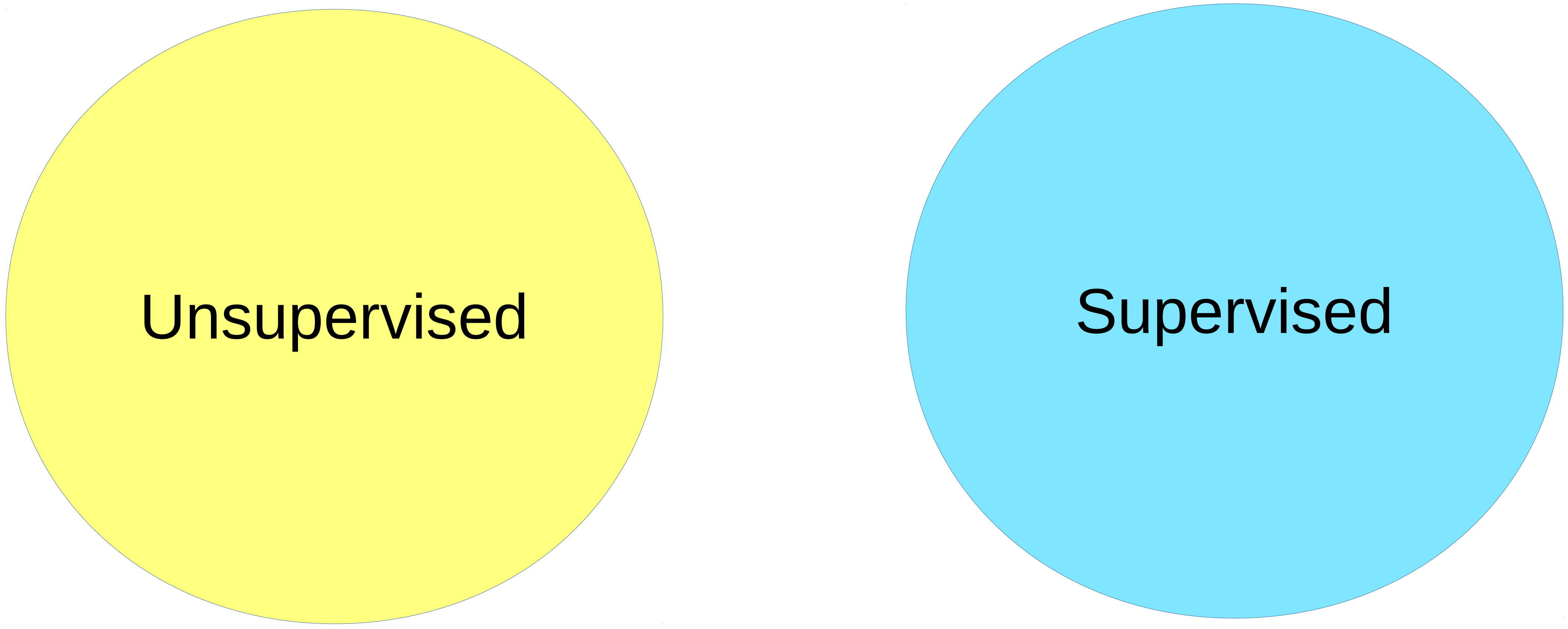
Lots of current work
uses both approaches
in **joint systems**!

These are models...

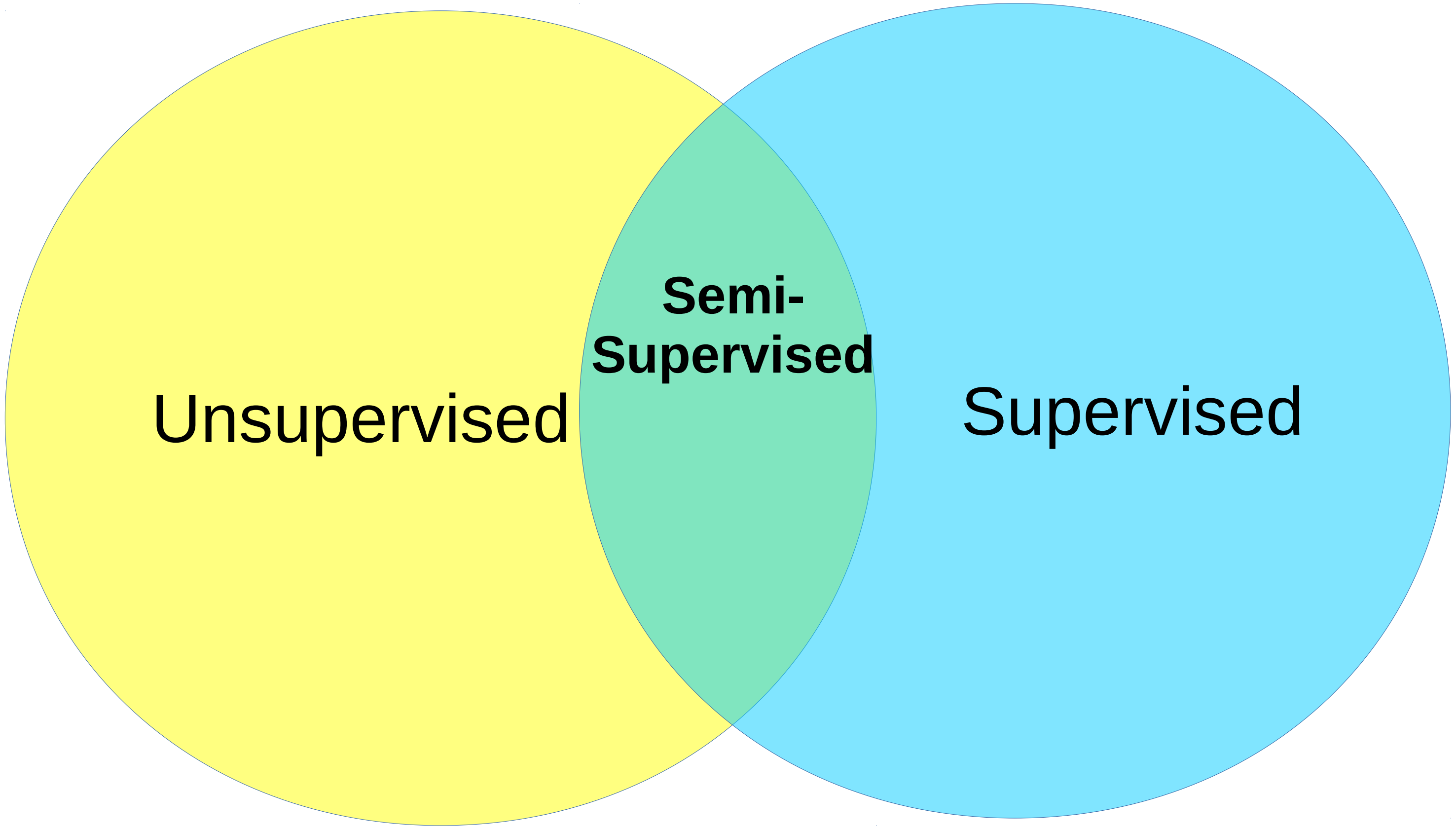
These are models...

...but models are only tools to solve problems.

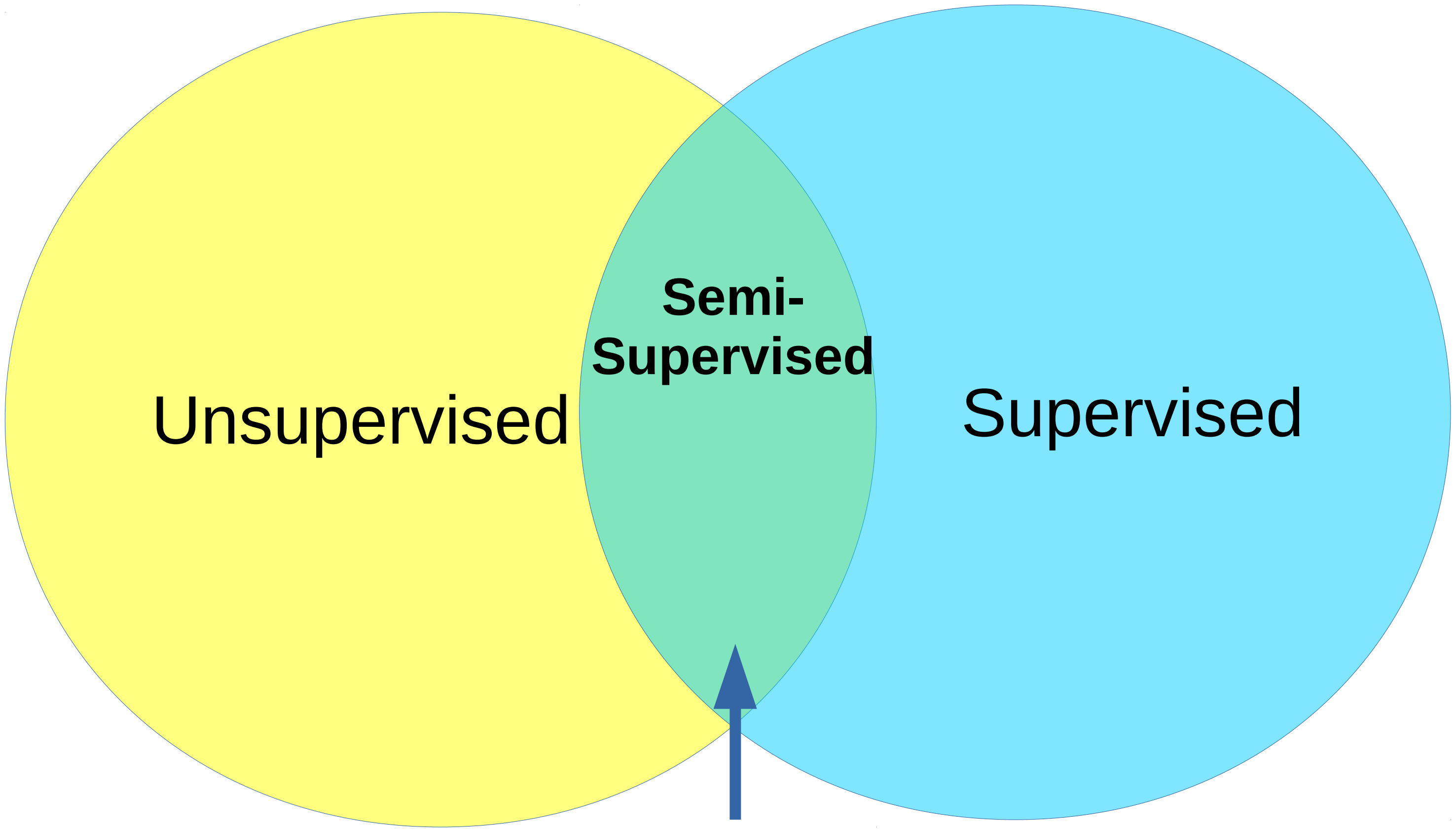
Kinds of Machine Learning



Kinds of Machine Learning



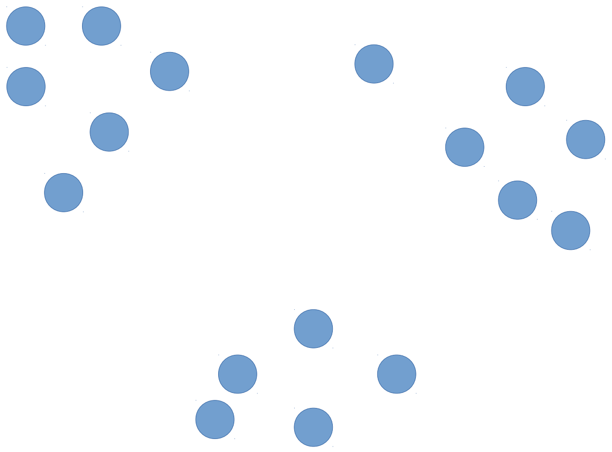
Kinds of Machine Learning



aka Distantly-supervised,
weakly-supervised

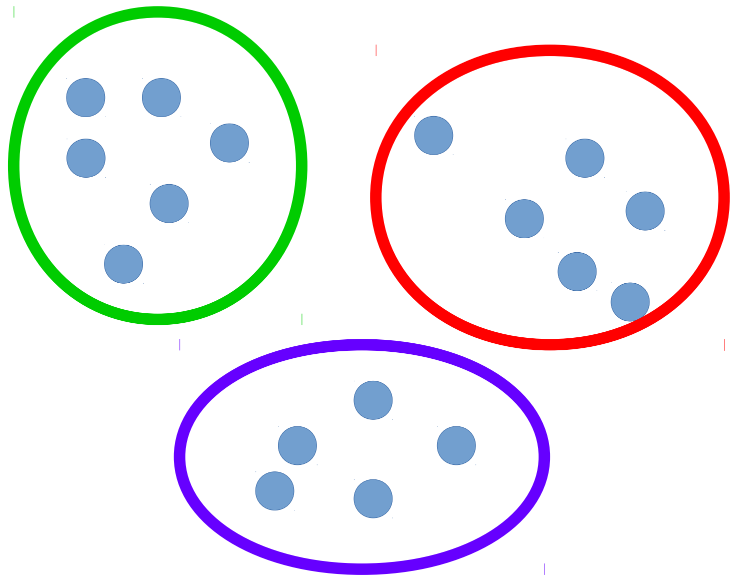
Unsupervised Learning

Goal: Discover hidden structure in data

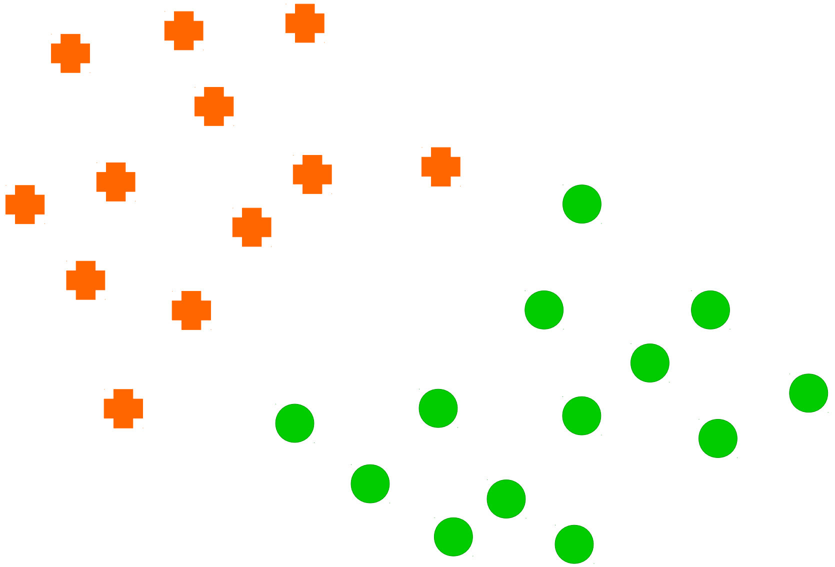


Unsupervised Learning

Goal: Discover hidden structure in data

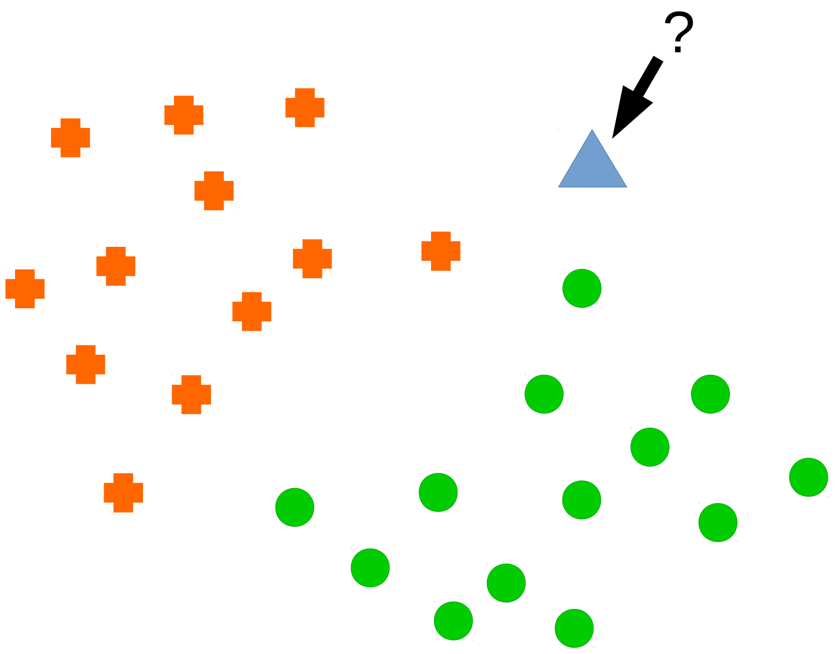


Supervised Learning



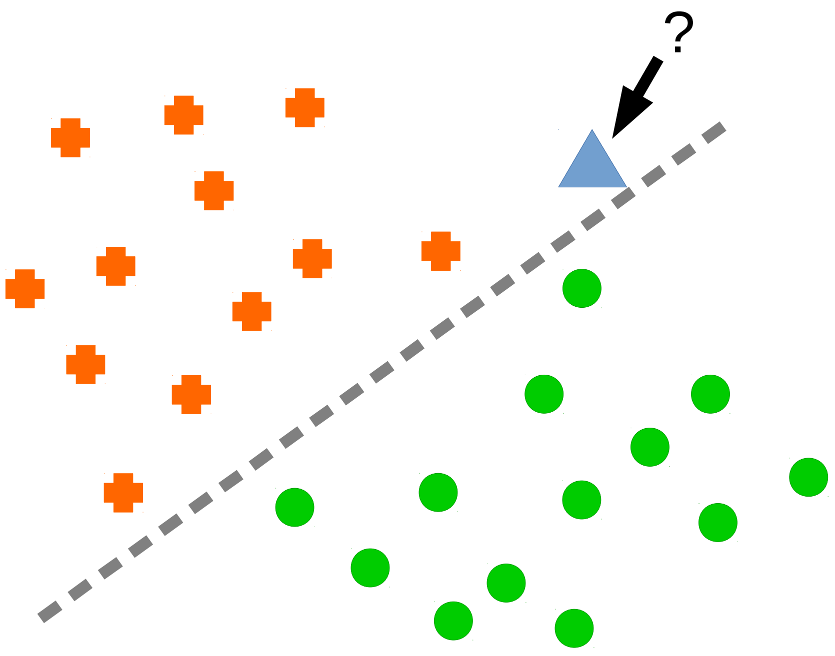
Goal: Use known information to categorize data

Supervised Learning



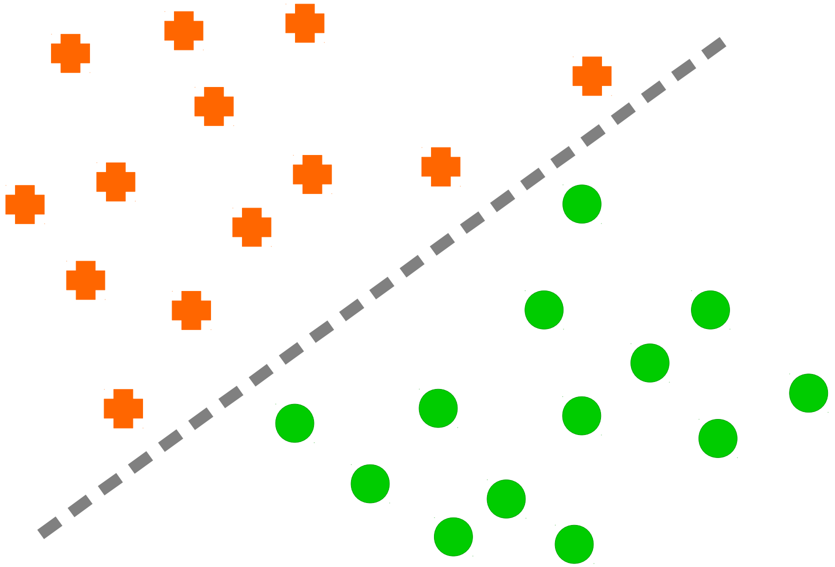
Goal: Use known information to categorize data

Supervised Learning



Goal: Use known information to categorize data

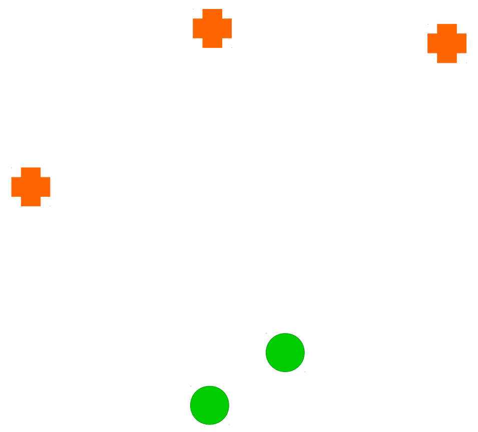
Supervised Learning



Goal: Use known information to categorize data

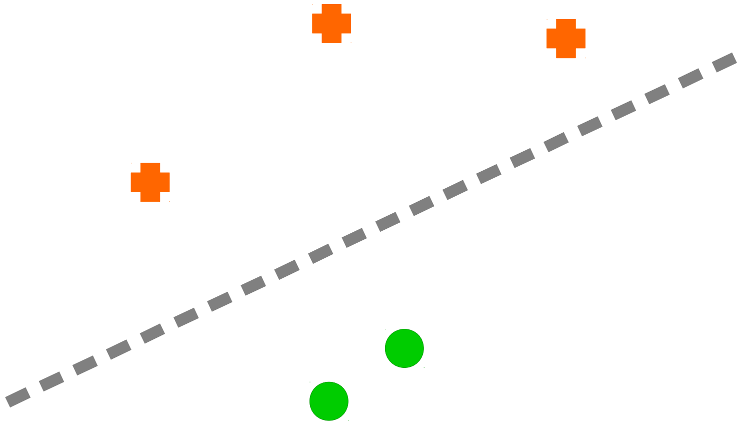
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



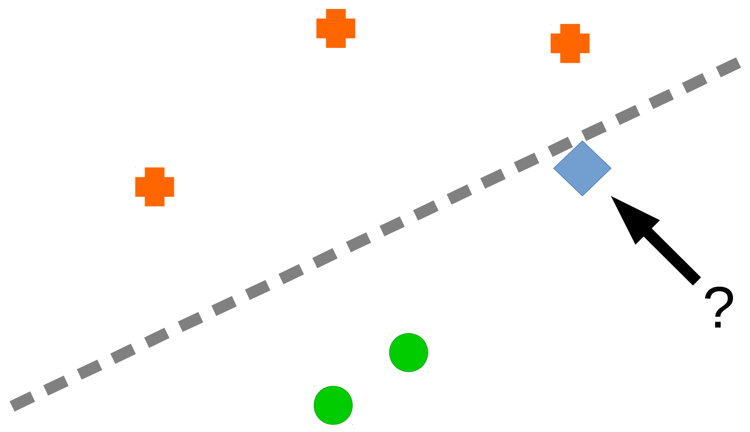
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



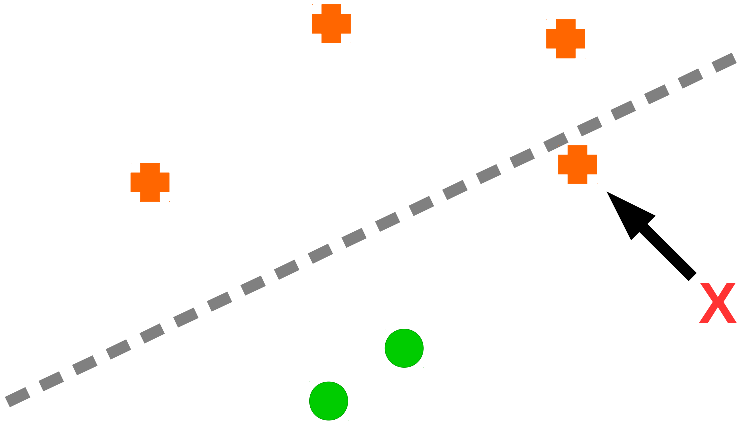
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



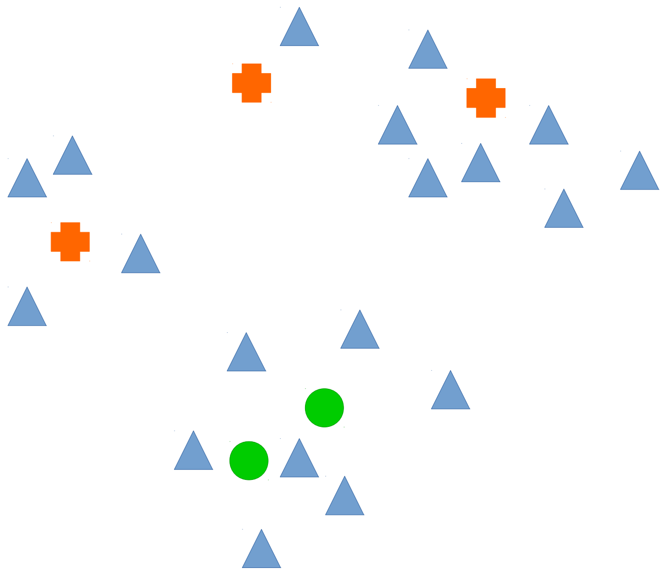
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



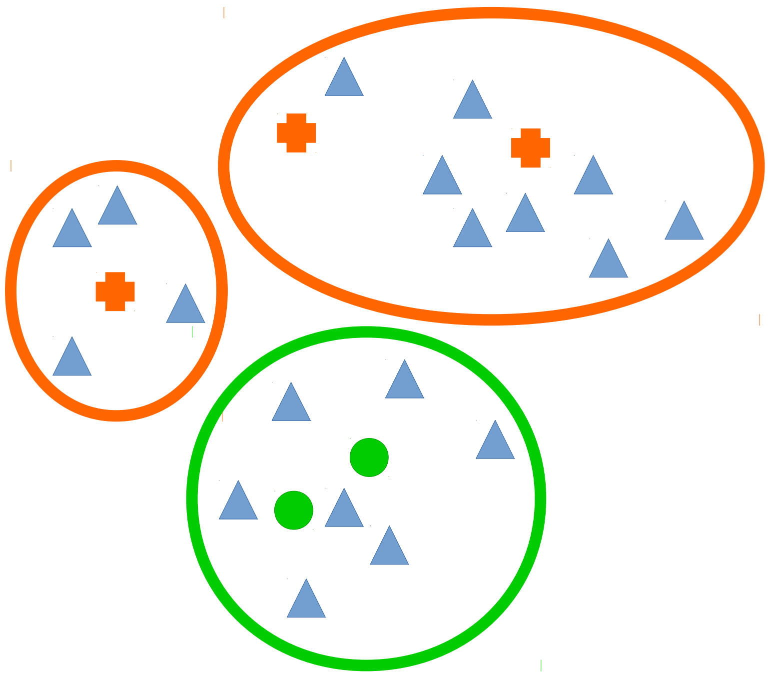
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



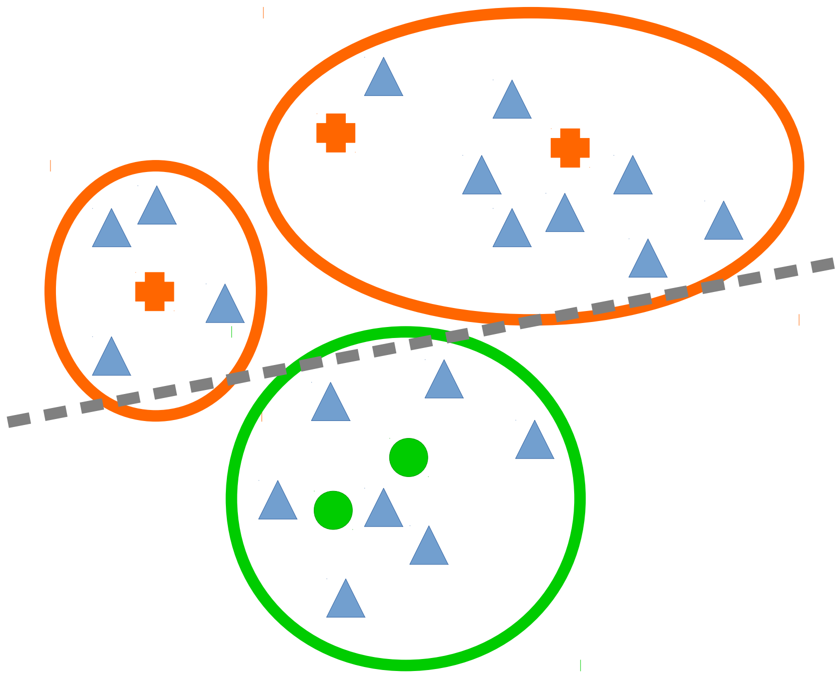
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



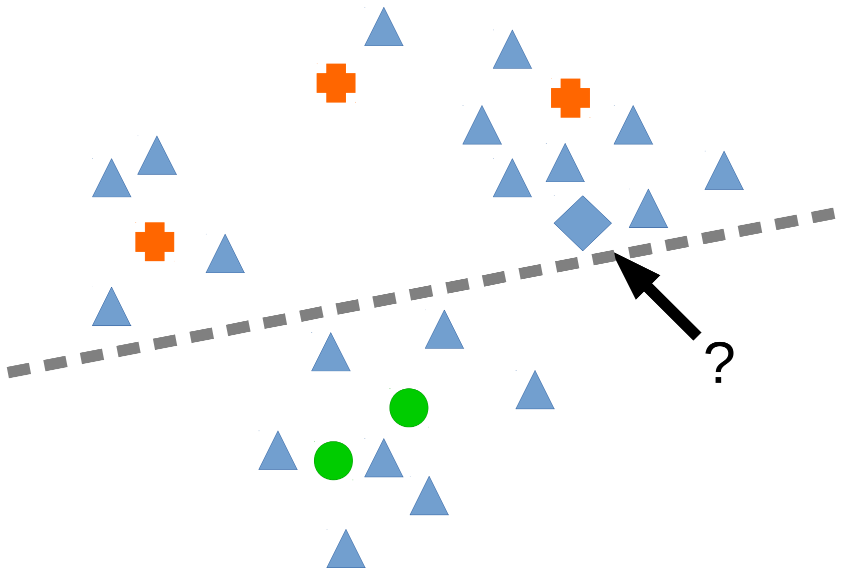
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



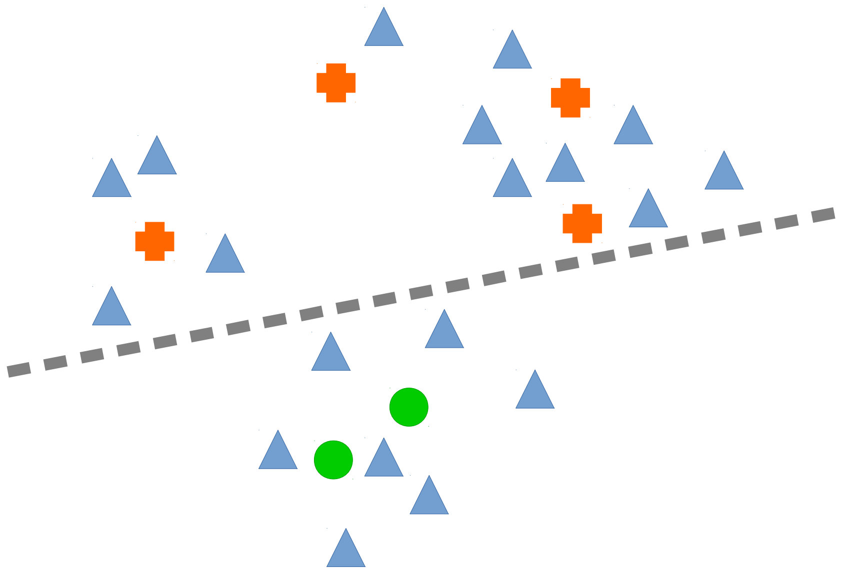
Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data



Semi-Supervised Learning

Goal: Use some known information, along with hidden structure, to categorize data

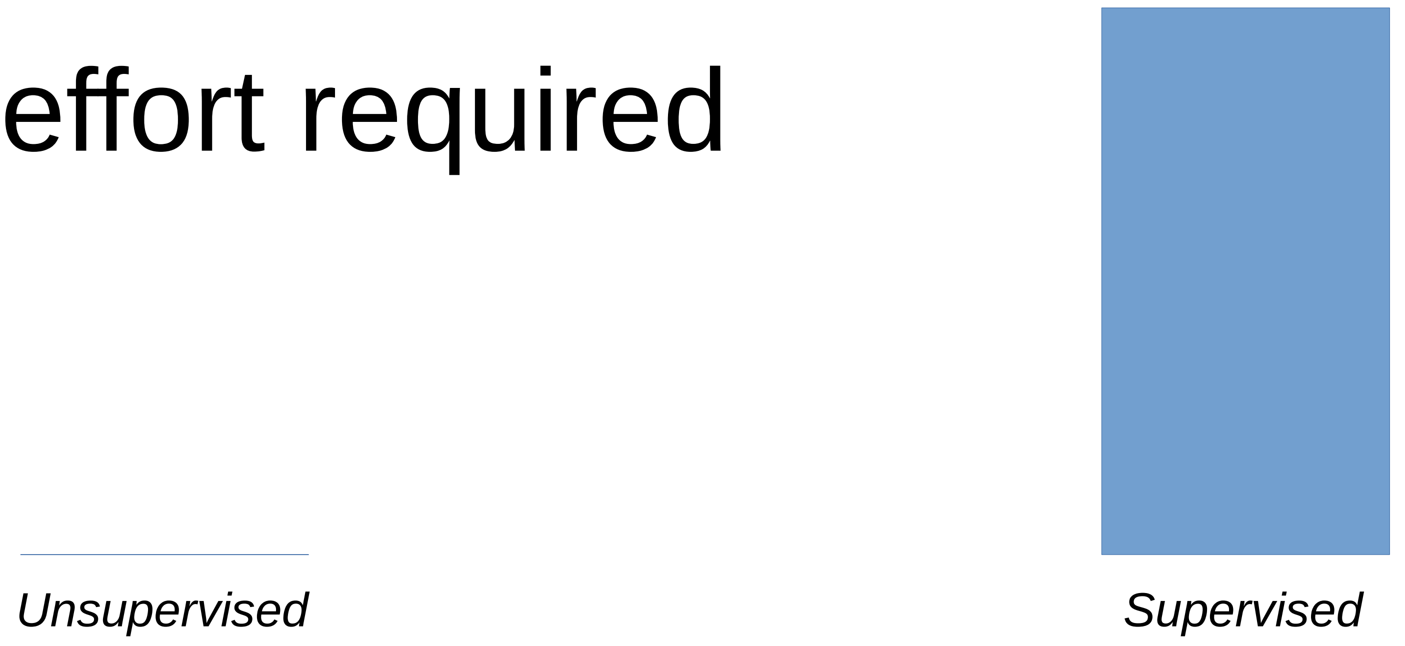


Human effort required

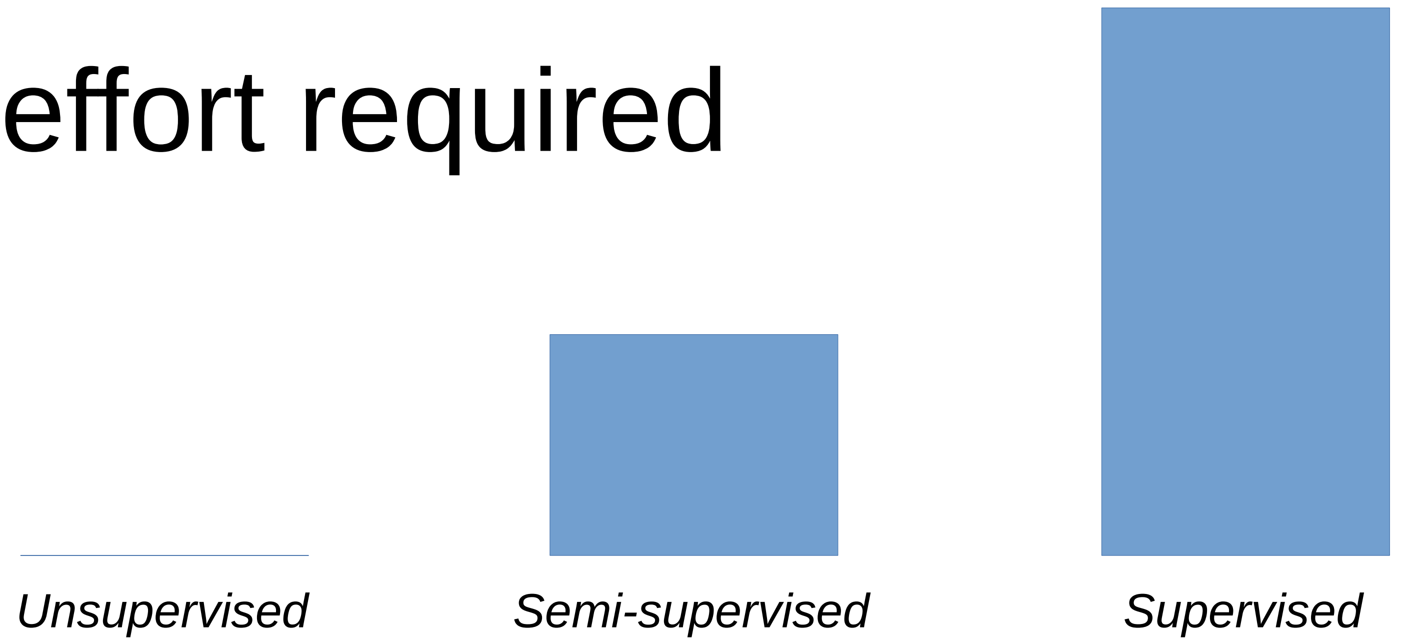
Human effort required

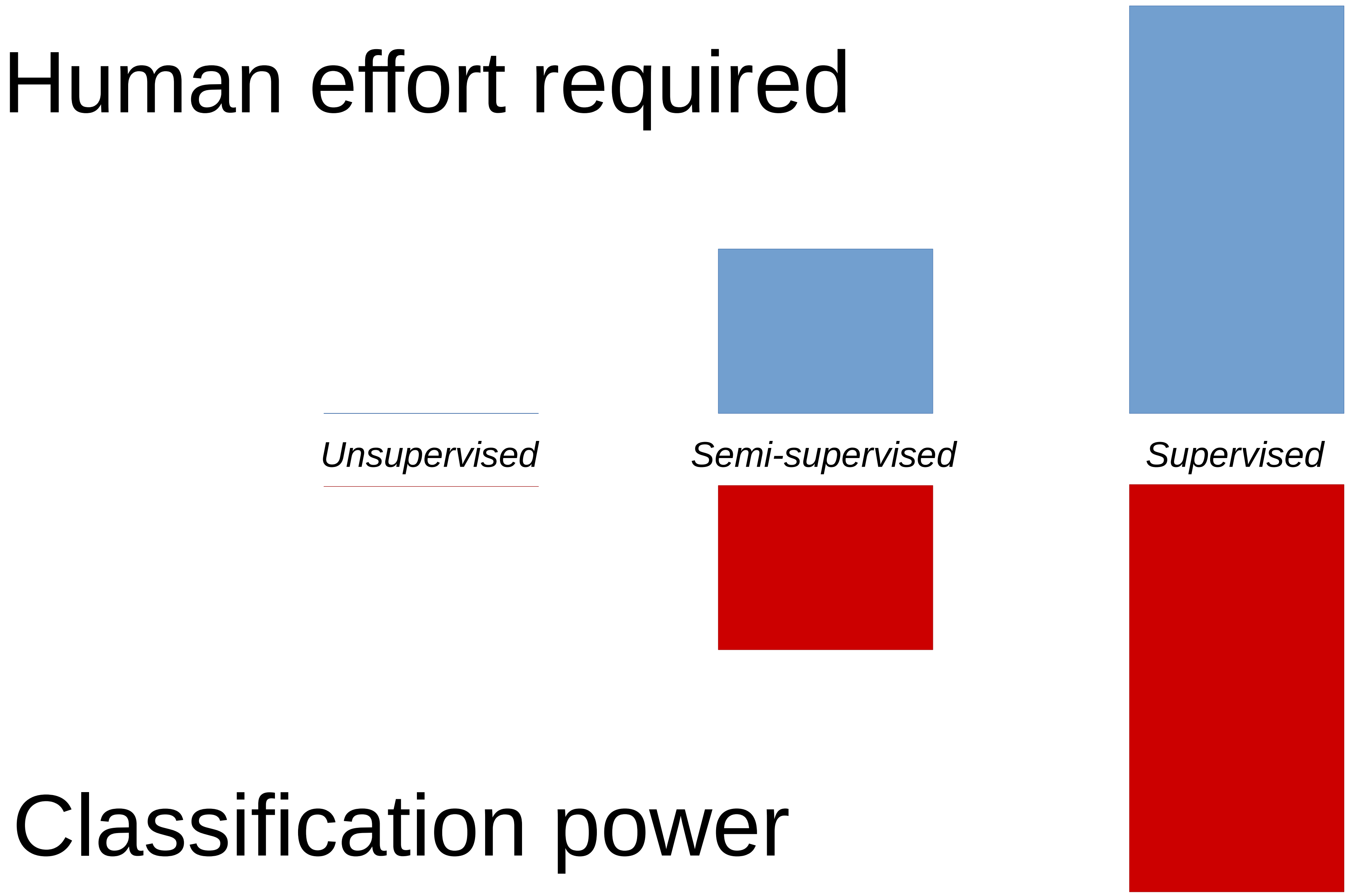
Unsupervised

Human effort required



Human effort required




Human effort required
Unsupervised
Semi-supervised
Supervised
Classification power

At this point, you may be asking yourself...

At this point, you may be asking yourself...

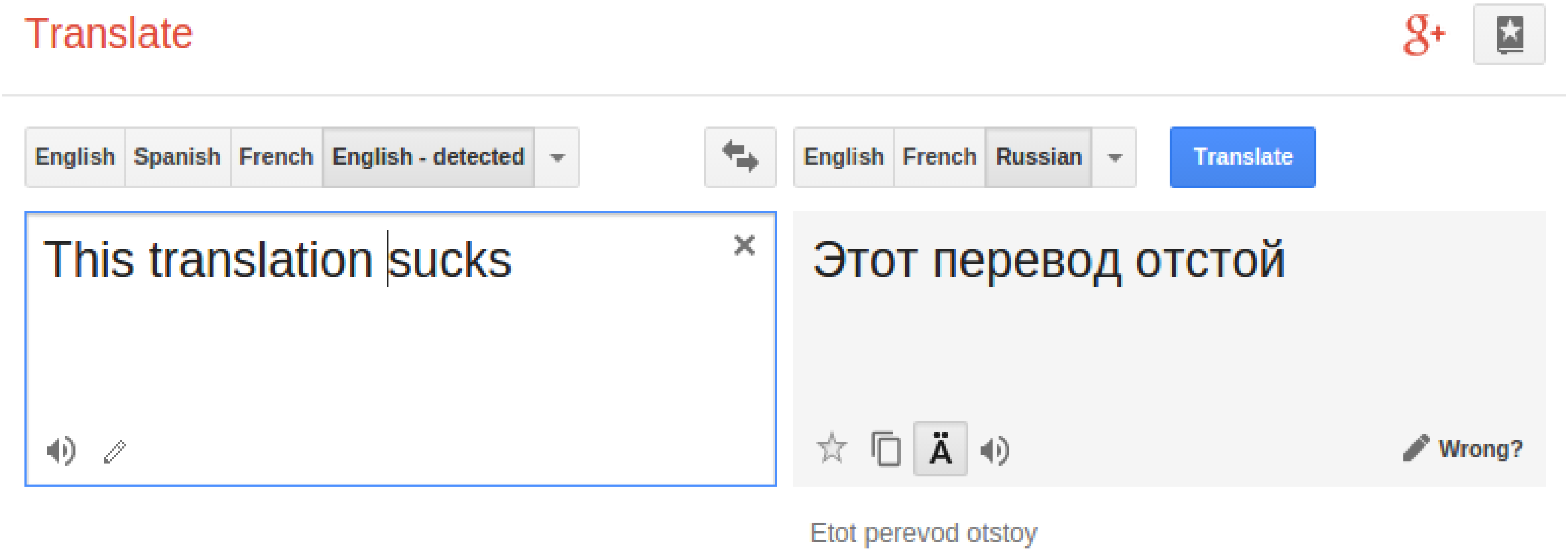
? ? ? ?

So what do you **do** with all this stuff?

? ? ?

Lots of things!

Machine Translation



Parsing / Tagging

Picard ordered tea.

Parsing / Tagging

	Picard	ordered	tea.
Part of Speech	NNP	VBD	NN

Parsing / Tagging

	Picard	ordered	tea.
Part of Speech	NNP	VBD	NN
Dependency	NSubj →	Root	← DObj

Information Extraction

“Abraham Lincoln was born February 12, 1809, in Hardin County, Kentucky...”

Information Extraction

"***Abraham Lincoln*** *was born February 12, 1809, in Hardin County, Kentucky...*"

Birth Dates

ID	Month	Day	Year
Honest Abe	February	12	1809

Birth Locations

ID	County	State	Country
Big Lincoln	Hardin	Kentucky	'Murica

Information Retrieval

Information Retrieval

Web Search

Information Retrieval

Web Search

Bioinformatics

ATTACCGCAGAT

1 | CATTACCGGAGATCCTA
2 | CCCATTACGGCCGCAGATAA
3 | ATTACCGAA

Information Retrieval

Web Search

Bioinformatics

ATTACCGCAGAT

1 | CATTACCGGAGATCCTA
2 | CCCATTACGGCCGCAGATAA
3 | ATTACCGAA

Question Answering

Who played Malcolm Reynolds?

Nathan Fillion

Who played Real Madrid last week?

Barcelona; final score 3-2

Etc., etc., etc.

Etc., etc., etc.

Automatic summarization

Etc., etc., etc.

Automatic summarization

Bacon ipsum dolor amet spare ribs leberkas filet mignon t-bone tenderloin ground round. Leberkas kevin meatball, short ribs rump andouille meatloaf pancetta shank bacon pork belly frankfurter picanha shankle sausage. Salami strip steak sirloin cow. Andouille ball tip meatloaf biltong bresaola. Cupim drumstick swine t-bone pork belly frankfurter jowl chuck leberkas cow short ribs ball tip.

Porchetta leberkas swine kevin ham capicola shankle strip steak hamburger salami filet mignon tri-tip bresaola picanha. Brisket tail swine biltong, capicola shankle sirloin. Jerky meatloaf ribeye, fatback turkey pork chop porchetta landjaeger ham salami meatball tongue pancetta kevin. Tri-tip swine filet mignon meatloaf bresaola porchetta pancetta salami frankfurter pork chop. Pork loin jerky pork chop, drumstick chuck flank ground round. Landjaeger hamburger pastrami salami.

Etc., etc., etc.

Automatic summarization

Bacon ipsum dolor amet spare ribs leberkas filet mignon t-bone tenderloin ground round. Leberkas kevin meatball, short ribs rump andouille meatloaf pancetta shank bacon pork belly frankfurter picanha shankle sausage. Salami strip steak sirloin cow. Andouille ball tip meatloaf biltong bresaola. Cupim drumstick swine t-bone pork belly frankfurter jowl chuck leberkas cow short ribs ball tip.

Porchetta leberkas swine kevin ham capicola shankle strip steak hamburger salami filet mignon tri-tip bresaola picanha. Brisket tail swine biltong, capicola shankle sirloin. Jerky meatloaf ribeye, fatback turkey pork chop porchetta landjaeger ham salami meatball tongue pancetta kevin. Tri-tip swine filet mignon meatloaf bresaola porchetta pancetta salami frankfurter pork chop. Pork loin jerky pork chop, drumstick chuck flank ground round. Landjaeger hamburger pastrami salami.

Bacon bacon bacon bacon pork!

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Etc., etc., etc.

Automatic summarization

Sentiment analysis

**Not actual tweets*

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Not actual tweets

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Not actual tweets

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

U: I want Chinese food.

S: Here are 473 Chinese places.

U: How about cheap ones on the south side?

S: Here is 1 restaurant.

U: Eh, let's do Thai food instead.

S: I'm sorry, Dave, I can't let you do that.

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

U: I want Chinese food.

S: Here are 473 Chinese places.

U: How about cheap ones on the south side?

S: Here is 1 restaurant.

U: Eh, let's do Thai food instead.

S: I'm sorry, Dave, I can't let you do that.

User Goals

Turn	*Type*	*Location*	*Cheap?*
1	Chinese	???	???
2	Chinese	South	Yes
3	Thai	South	Yes

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Etc., etc., etc.

Phonemes

U|n|b|r|ea|k|a|b|le

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

<u>Phonemes</u>

U|n|b|r|ea|k|a|b|le

<u>Morphemes</u>

Un|break|able

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

U|n|b|r|ea|k|a|b|le

Morphemes

Un|break|able

Words

maytheforcebewithyou

May the force be with you

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

U|n|b|r|ea|k|a|b|le

Morphemes

Un|break|able

Words

maytheforcebewithyou

May the force be with you

Sentences

[I spoke to Mr. Spock.] [His response was illogical.]

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Phonemes

U|n|b|r|ea|k|a|b|le

Morphemes

Un|break|able

Words

maytheforcebewithyou

May the force be with you

Sentences

[I spoke to Mr. Spock.] [His response was illogical.]

Topics

...who I met at a Trek convention.

As for Star Wars...

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

Word sense disambiguation

After I put him in [check]^1, he wrote me a [check]^2.

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

Word sense disambiguation

After I put him in [check]^1, he wrote me a [check]^2.

Coreference resolution

I spoke to [the customer]$_1$, then told [my boss]$_2$ that [she]$_2$ should fire [her]$_1$.

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

Word sense disambiguation

After I put him in [check][1], he wrote me a [check][2].

Coreference resolution

I spoke to [the customer]$_1$, then told [my boss]$_2$ that [she]$_2$ should fire [her]$_1$.

Named entity recognition

[Bugs Bunny]$_{\text{Person}}$ bought 50% of **[Acme Corp.]**$_{\text{Company}}$ in **[2004]**$_{\text{Year}}$.

Etc., etc., etc.

Automatic summarization

Sentiment analysis

Discourse analysis

Segmentation

Disambiguation and reference

And many more!

How can I get in on this?

NLP Toolkits

Toolkit	*Language*	*Website*
Apache OpenNLP	Java	https://opennlp.apache.org
General-purpose NLP toolkit; tends to use older models, but under Apache license.		
Natural Language Toolkit (NLTK)	Python	http://www.nltk.org/
Standard NLP option for Python; easy to pick up and play with, and includes several common corpora.		
Mallet	Java	http://mallet.cs.umass.edu/
More technical toolkit, focused on current, high-complexity models.		
LingPipe	Java	http://alias-i.com/lingpipe/
Another general-purpose NLP toolkit; offers industry licensing option.		
Stanford CoreNLP	Java	http://nlp.stanford.edu/software/corenlp.shtml
Standard tools in academia, tends towards cutting edge models. Low ease-of-use, and academic licensing restrictions.		
Alchemy API	Cloud API	http://www.alchemyapi.com/
Fanciest industry option (owned by IBM). Offers NLP, vision, other ML resources.		

Other Resources

KALDI

Speech Recognition Toolkit - http://kaldi-asr.org/

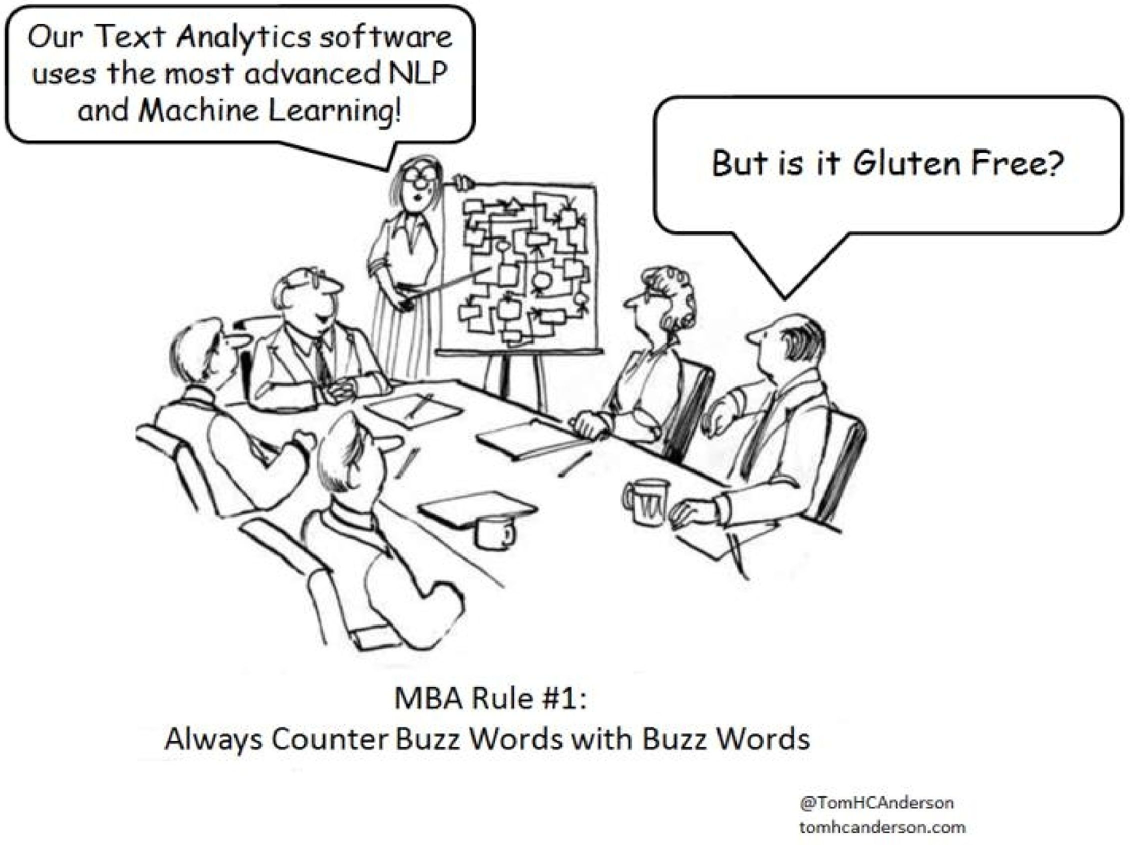
IEEE Signal Processing Society

http://www.signalprocessingsociety.org/

Association for Computational Linguistics

http://aclweb.org/

Questions?



My contact info:
Denis Griffis
griffis.30@osu.edu
http://web.cse.ohio-state.edu/slate/