# Tasks and Benchmarks

## CSCI 1460: Computational Linguistics
## Lecture 15

**Ellie Pavlick**
**Fall 2023**

# Topics

- Lecture 0 Reprise
- Task-Driven Culture
- Major “Downstream” Tasks and Evaluations
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Topics

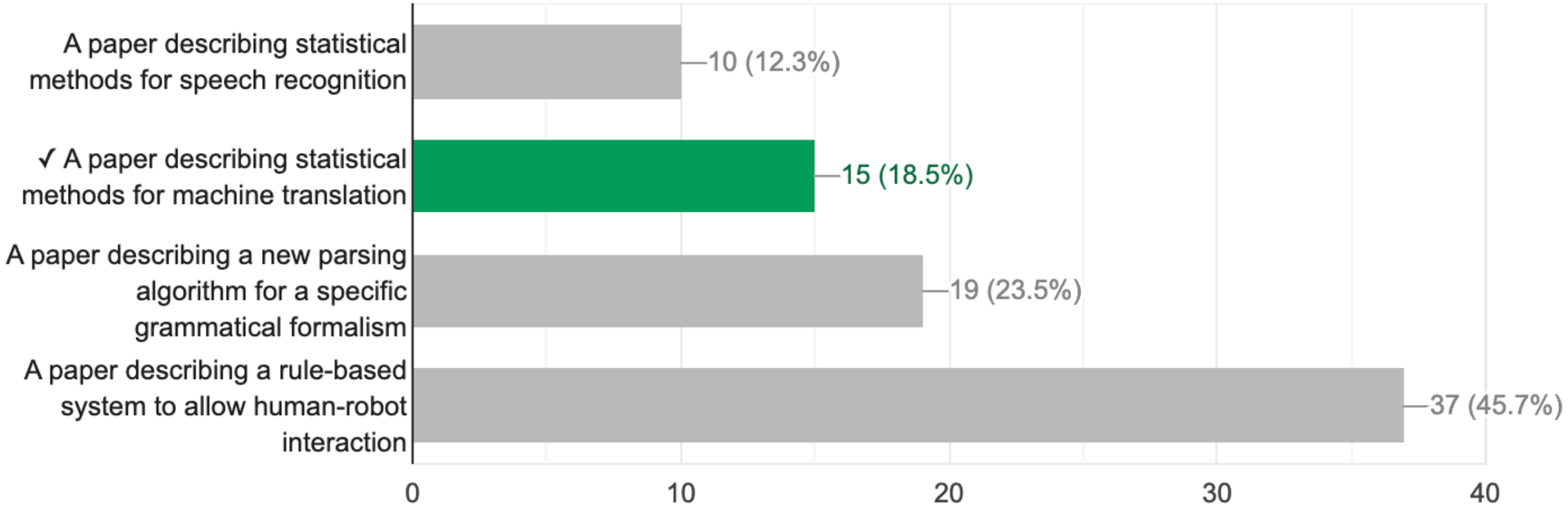
- **Lecture 0 Reprise**
- Task-Driven Culture
- Major “Downstream” Tasks and Evaluations
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Lecture 0 Reprise

If you were skimming ACL proceedings from the 1970s, which of the below would you be least likely to find?

Copy

15 / 81 correct responses



# Lecture 0 Reprise

- “The first question is ill-posed. In the lecture, Ellie says that tokens can be converted directly to vectors and then fed into a DL architecture, and then she suggests an alternative approach which does not use vectors to represent word meanings. The question is asking us to order steps which don't occur in a single model. For example, it wouldn't make sense to convert tokens to meaning vectors and then resolve ambiguous phrases, since they are already converted into vectors.”
  - My mistake! Vectors are an alternative to the other lexical semantic representations. But vectors can play nicely with the other steps of the pipeline. E.g., one could in principle convert words to vectors and then resolve ambiguous phrases.

# The Language Processing Pipeline

discourse

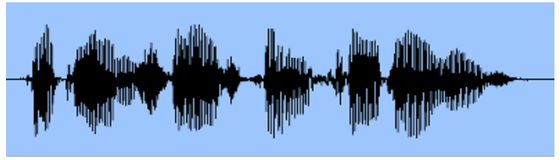
pragmatics

semantics

syntax

lexicon

phonology



Alexa! um can you get me up at uh six

[‘get’, ‘me’, ‘up’, ‘at’, ‘6’]

map to “meanings”

???

# The Language Processing Pipeline

discourse

pragmatics

semantics

syntax

lexicon

phonology

['get', 'me', 'up', 'at', '6']

map to "meanings"

lookup in a database
(e.g., WordNet, PropBank,
custom "intent" database)

| | |
|---|---|
| play | `play_music(song)` |
| set | `set_time(time)` |
| set | `set_temp(temperature)` |
| weather | `check_weather(day, location)` |

# The Language Processing Pipeline

discourse

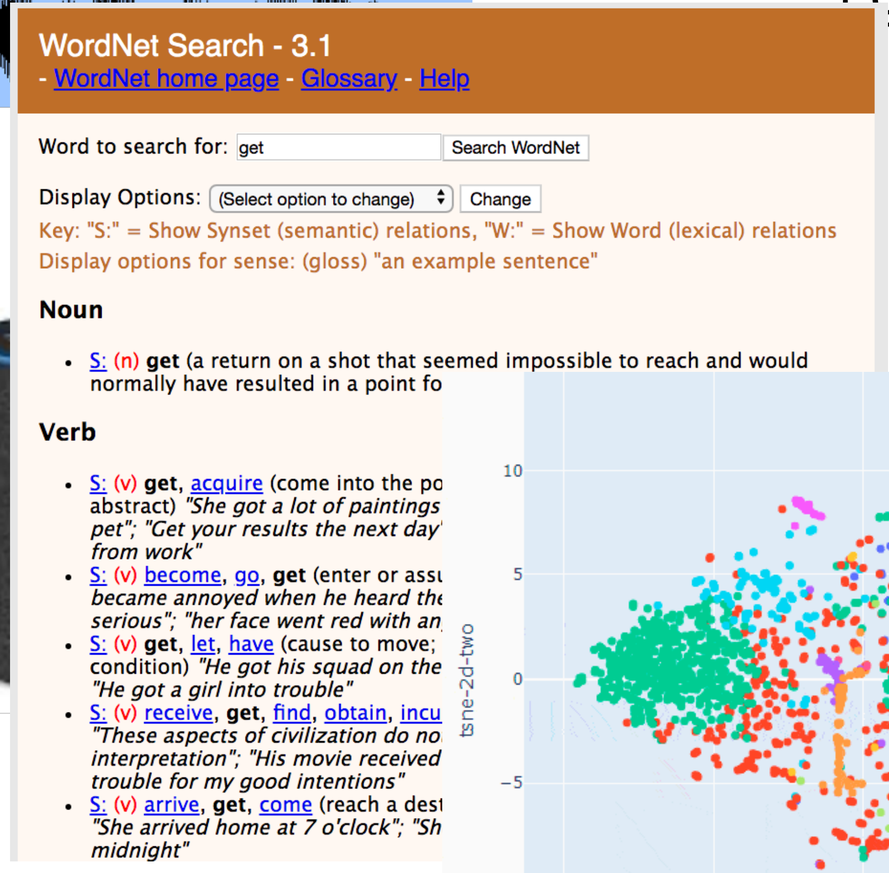
pragmatics

semantics

syntax

lexicon

phonology



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: get Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

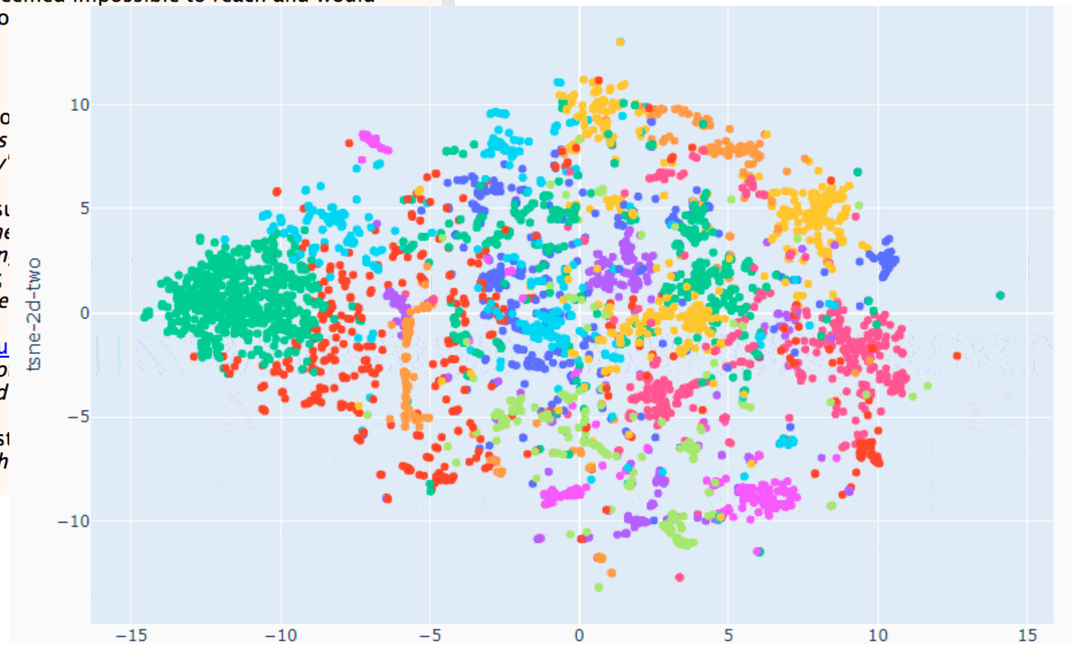
Display options for sense: (gloss) "an example sentence"

**Noun**

- S: (n) **get** (a return on a shot that seemed impossible to reach and would normally have resulted in a point fo

**Verb**

- S: (v) **get**, acquire (come into the po abstract) *"She got a lot of paintings pet"; "Get your results the next day from work"*
- S: (v) become, go, **get** (enter or assu *became annoyed when he heard the serious"; "her face went red with an*
- S: (v) **get**, let, have (cause to move; condition) *"He got his squad on the "He got a girl into trouble"*
- S: (v) receive, **get**, find, obtain, incu *"These aspects of civilization do no interpretation"; "His movie received trouble for my good intentions"*
- S: (v) arrive, **get**, come (reach a dest *"She arrived home at 7 o'clock"; "Sh midnight"*




['get', 'me', 'up', 'at', '6']

map to "meanings"

map onto vectors

# The Language Processing Pipeline

discourse

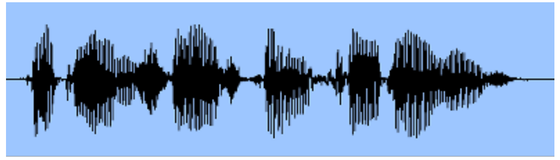
pragmatics

semantics

syntax

lexicon

phonology



**Alexa! um can you get me up at uh six**

[‘get’, ‘me’, ‘up’, ‘at’, ‘6’]

| | |
|---|---|
| play | `play_music(song)` |
| set | `set_time(time)` |
| set | `set_temp(temperature)` |
| weather | `check_weather(day, location)` |

# The Language Processing Pipeline

discourse

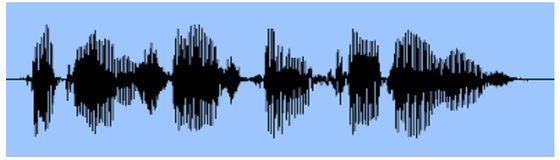
pragma

seman

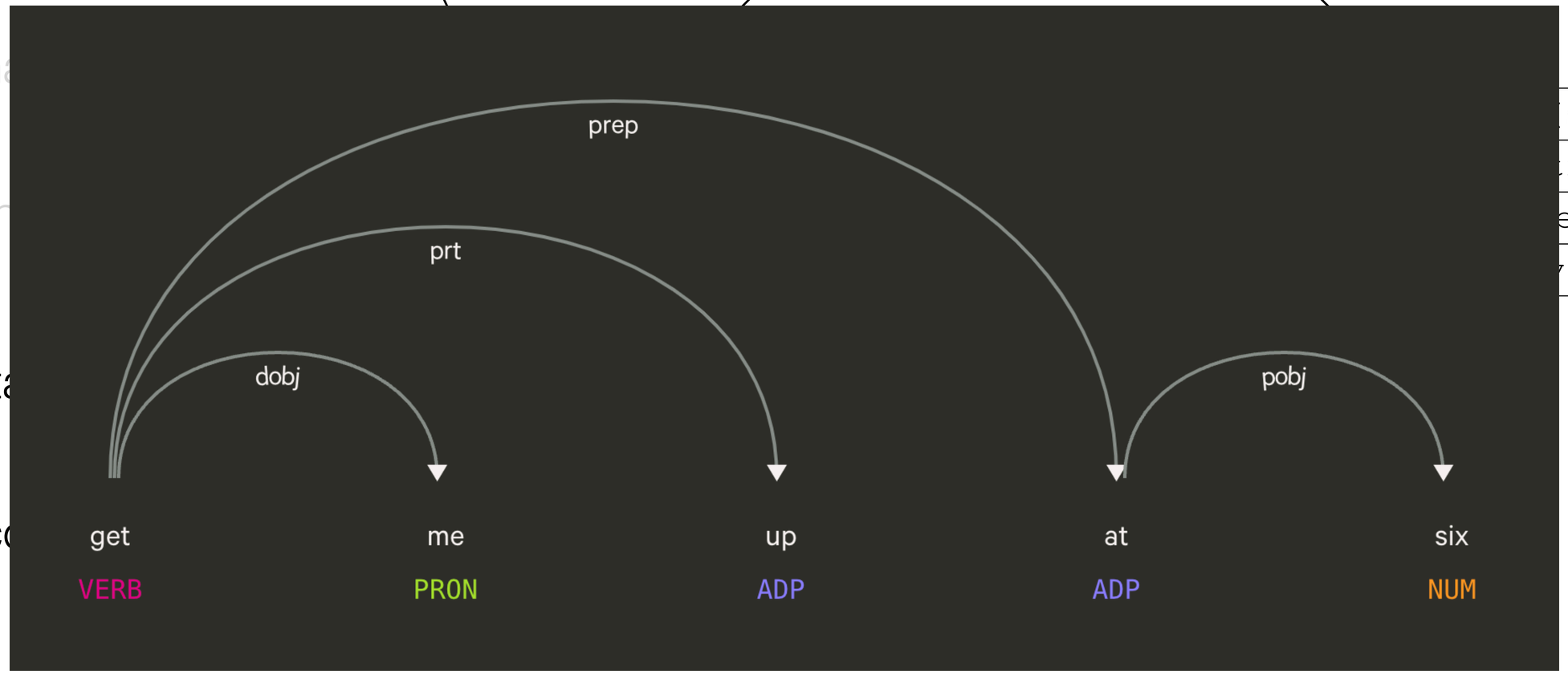
synta

lexic

phonology



[‘get’, ‘me’, ‘up’, ‘at’, ‘6’]




song)

ime)

erature)

, location)

# The Language Processing Pipeline

discourse

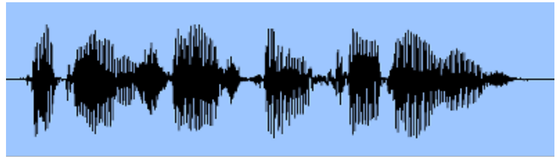
pragmatics

semantics

syntax

lexicon

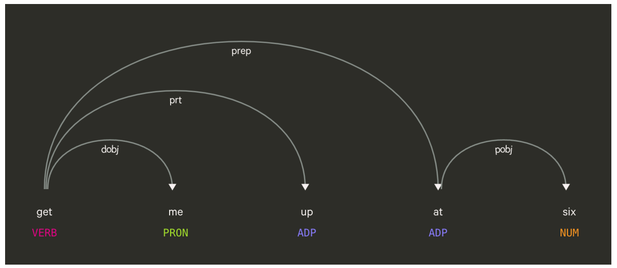
phonology



**Alexa! um can you get me up at uh six**

[‘get’, ‘me’, ‘up’, ‘at’, ‘6’]

| | |
|---|---|
| play | `play_music(song)` |
| set | `set_time(time)` |
| set | `set_temp(temperature)` |
| weather | `check_weather(day, location)` |



# The Language Processing Pipeline

discourse

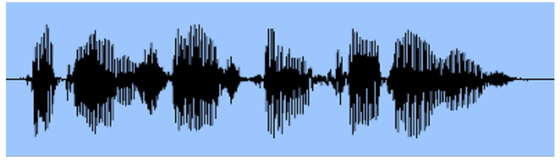
pragmatics

semantics

syntax

lexicon

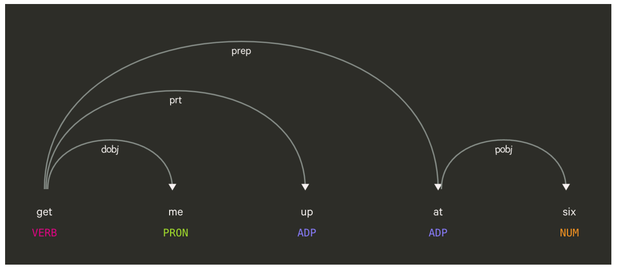
phonology



**Alexa! um can you get me up at uh six**

[‘get’, ‘me’, ‘up’, ‘at’, ‘6’]

| | |
|---|---|
| play | `play_music(song)` |
| set | `set_time(time)` |
| set | `set_temp(temperature)` |
| weather | `check_weather(day, location)` |




`set_time(6)`

# The Language Processing Pipeline

discourse

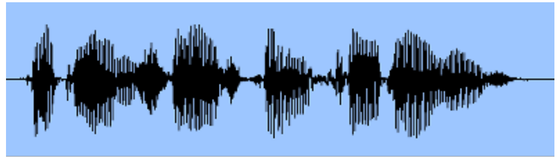
pragmatics

semantics

syntax

lexicon

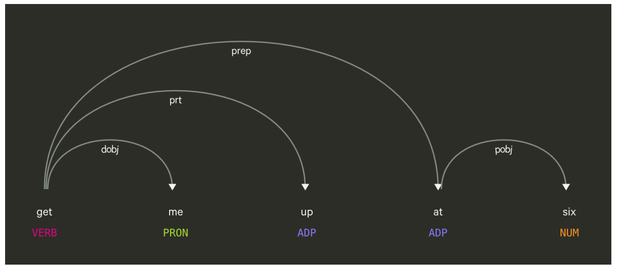
phonology



**Alexa! um can you get me up at uh six**

['get', 'me', 'up', 'at', '6']

| | |
|---|---|
| play | `play_music(song)` |
| set | `set_time(time)` |
| set | `set_temp(temperature)` |
| weather | `check_weather(day, location)` |




`set_time(6am)`

depends on info about the speaker (time, location) and requires some "common sense"

# Lecture 0 Reprise

- “Are the textbook readings required or supplemental? If they're required, are we expected to come to class having done the reading already?”
  - The readings complement the lecture. I try to cover the important parts, but can’t get to everything. You are highly highly encouraged to consult the textbook prior to the quiz/assignments.

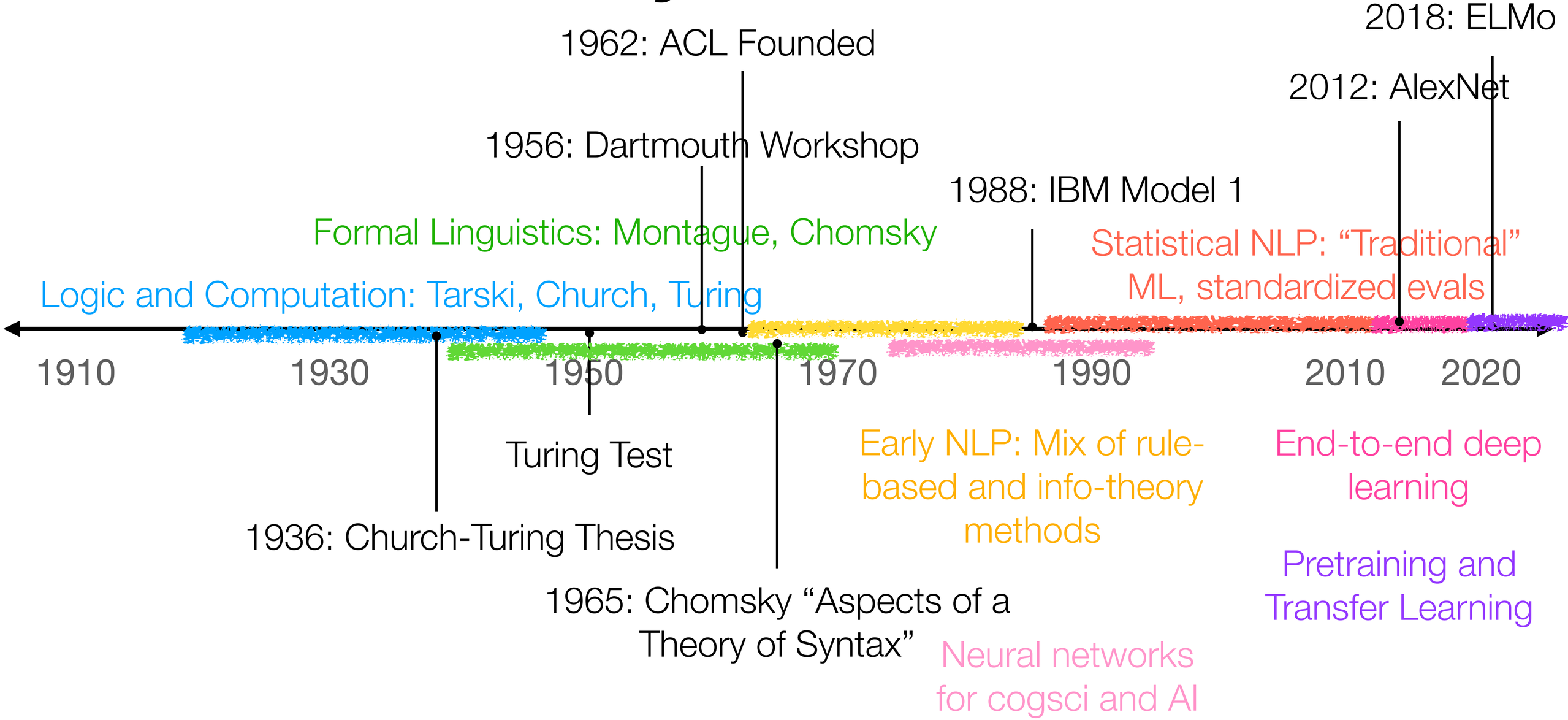
# Lecture 0 Reprise

- “Could there be some explanation when an abbreviation is mentioned?”
  - I’ll try! Please tell me if I breeze over something!

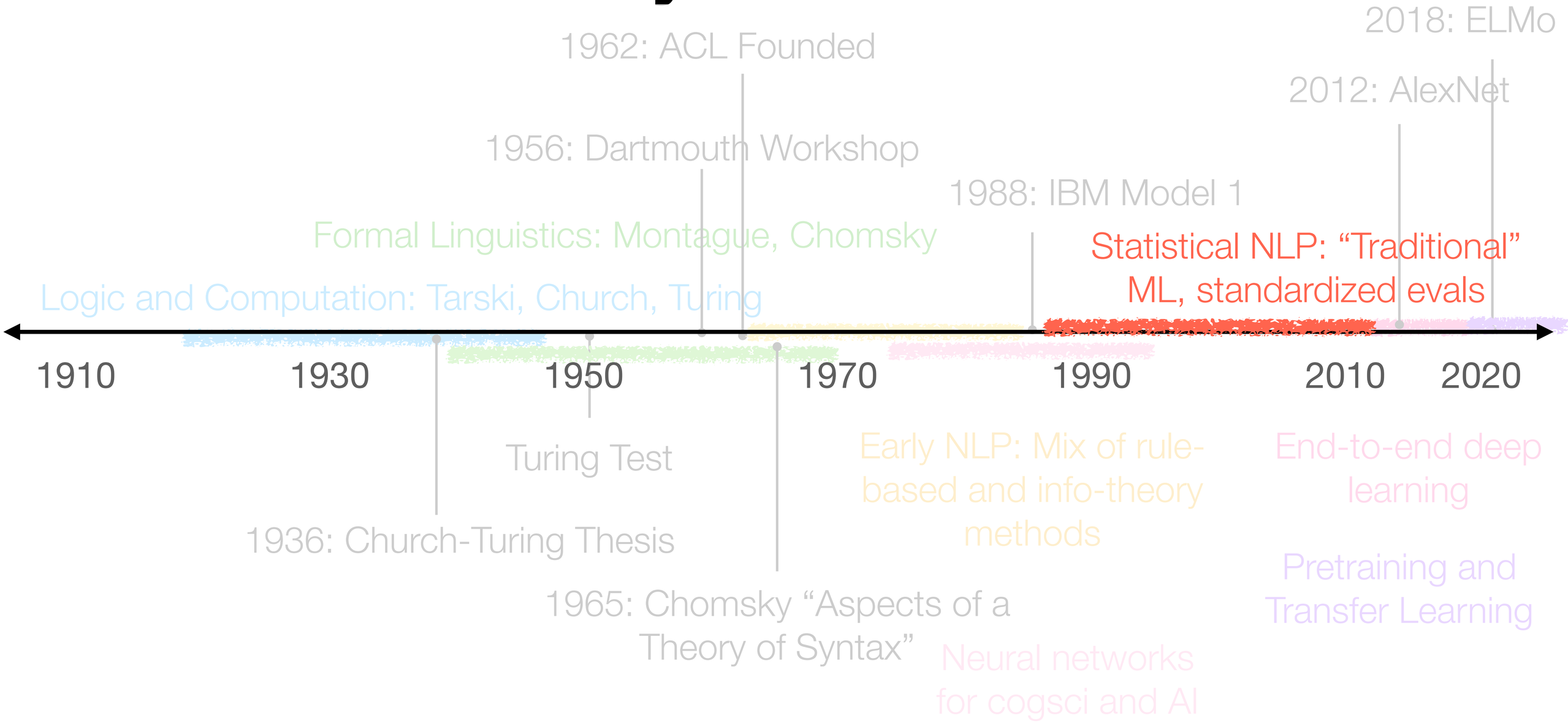
# Topics

- Lecture 0 Reprise
- **Task-Driven Culture**
- Major “Downstream” Tasks and Evaluations
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# NLP: A brief history



# NLP: A brief history



# Task-Driven Culture

- NLP research has tended to be organized around “tasks”
  - I.e., many independent research groups with a shared idea of what type of system “we” are trying to build
- Most researchers never build a whole system themselves, but they will justify the problem they are solving in relation to some commonly-agreed upon important “task”
- Recently, NLP research has broadened significantly to include mode data science and hypothesis-testing types of work

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: “Language Diversity: from Low-Resource to Endangered Languages”

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

"Extrinsic Tasks": Tasks that an "end user" might want to use.

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

"Extrinsic Tasks": Tasks that an "end user" might want to use.

(Intuitively, things that, if you described the task to your Aunt and Uncle at Thanksgiving, they would "get it")

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

"Extrinsic Tasks": Tasks that an "end user" might want to use.

(Intuitively, things that, if you described the task to your Aunt and Uncle at Thanksgiving, they would "get it")

"The car" 🚙

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

"Intrinsic Tasks":
Tasks that are part of
the bigger system

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

"Intrinsic Tasks":
Tasks that are part of the bigger system

Engines, break pads, exhaust pipes, chips, coolant, gears, screws... ⚙️🔩🛠️

# SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Data science, social sciences, and introspection

# Task-Driven Culture

## Bakeoffs and Leaderboards

- Statistical Revolution led to focus on **standardized test sets**
  - Everyone uses the same inputs, and compares against the same "ground truth" outputs, using the same metric
  - Allows head-to-head comparison between many independently-developed systems and ideas over time
- See **http://nlpprogress.com/** for list of tasks, datasets, models

# Task-Driven Culture

## Bakeoffs and Leaderboards

- Historically: DARPA programs sponsored competitions for speech, translation, retrieval, QA
- Then: **SemEval** (https://semeval.github.io/)
  - Academic-organized workshops around specific tasks of interest
- Recently: Specific focus on **leaderboards** and **multitask** evaluations
  - Public-facing competitions between models
  - Usually comes with easy-to-access data and baselines

# Task-Driven Culture

## SemEval

### What is SemEval?

**SemEval** is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. Each year's workshop features a collection of shared tasks in which computational semantic analysis systems designed by different teams are presented and compared.

### Semantic Structure

- **Task 1: V-WSD: Visual Word Sense Disambiguation** ([contact organizers], [join task mailing list])
  Alessandro Raganato, Iacer Calixto, Jose Camacho-Collados, Asahi Ushio, Mohammad Taher Pilehvar
- **Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER 2)** ([contact organizers], [join task mailing list])
  Shervin Malmasi, Besnik Fetahu, Sudipta Kar

### Discourse and Argumentation

- **Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup** ([contact organizers], [join task mailing list])
  Giovanni Da San Martino, Jakub Piskorski, Nicolas Stefanovitch, Preslav Nakov
- **Task 4: ValueEval: Identification of Human Values behind Arguments** ([contact organizers], [join task mailing list])
  Johannes Kiesel, Milad Alshomary, Henning Wachsmuth, Benno Stein
- **Task 5: Clickbait Spoiling** ([contact organizers], [join task mailing list])
  Maik Fröbe, Tim Gollub, Matthias Hagen, Martin Potthast
- **Task 6: LegalEval: Understanding Legal Texts** ([contact organizers], [join task mailing list])
  Prathamesh Ashok Kalamkar, Saurabh Kumar Karn, Sachin Malhan, Vivek Raghavan, Shouvik Kumar Guha, Ashutosh Modi

### Medical Applications

- **Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data** ([contact organizers], [join task mailing list])
  Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, Andre Freitas
- **Task 8: Causal medical claim identification and related PICO frame extraction from social media posts** ([contact organizers] [join task mailing list])
  (temporarily anonymous)

### Social Attitudes

- **Task 9: Multilingual Tweet Intimacy Analysis** ([contact organizers], [join task mailing list])
  Jiaxin Pei, Francesco Barbieri, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, David Jurgens
- **Task 10: Towards Explainable Detection of Online Sexism** ([contact organizers], [join task mailing list])
  Hannah Rose Kirk, Wenjie Yin, Paul Röttger, Bertie Vidgen
- **Task 11: Learning with Disagreements (Le-Wi-Di), 2nd edition** ([contact organizers], [join task mailing list])
  Elisa Leonardelli, Valerio Basile, Gavin Abercombie, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, Alexandra Uma
- **Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset** ([contact organizers], [join task mailing list])
  Shamsuddeen Hassan Muhammad, Seid Muhie Yimam, Idris Abdulmumin, Ibrahim Said Ahmad, Saif Mohammad, David Ifeoluwa Adelani, Sebastian Ruder, Nedjma Ousidhoum, Vukosi Marivate, Abinew Ali Ayele, Meriem Beloucif

# Task-Driven Culture

## Single-Task Leaderboards

**SQuAD: 100,000+ Questions for Machine Comprehension of Text**

**Pranav Rajpurkar** and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University

**Know What You Don't Know: Unanswerable Questions for SQuAD**

**Pranav Rajpurkar*** **Robin Jia*** **Percy Liang**
Computer Science Department, Stanford University
{pranavsr,robinjia,pliang}@cs.stanford.edu

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

# Task-Driven Culture

## Single-Task Leaderboards

**SQuAD: 100,000+ Questions for Machine Comprehension of Text**

**Pranav Rajpurkar** and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University

**Know What You Don't Know: Unanswerable Questions for SQuAD**

**Pranav Rajpurkar*** **Robin Jia*** **Percy Liang**
Computer Science Department, Stanford University
{pranavsr,robinjia,pliang}@cs.stanford.edu

### Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

| Rank | Model | EM | F1 |
|---|---|---|---|
| | Human Performance<br>*Stanford University*<br>(Rajpurkar & Jia et al. '18) | 86.831 | 89.452 |
| 1<br>Jun 04, 2021 | IE-Net (ensemble)<br>*RICOH_SRCB_DML* | **90.939** | **93.214** |
| 2<br>Feb 21, 2021 | FPNet (ensemble)<br>*Ant Service Intelligence Team* | 90.871 | 93.183 |
| 3<br>May 16, 2021 | IE-NetV2 (ensemble)<br>*RICOH_SRCB_DML* | 90.860 | 93.100 |
| 4<br>Apr 06, 2020 | SA-Net on Albert (ensemble)<br>*QIANXIN* | 90.724 | 93.011 |
| 5<br>May 05, 2020 | SA-Net-V2 (ensemble)<br>*QIANXIN* | 90.679 | 92.948 |
| 5<br>Apr 05, 2020 | Retro-Reader (ensemble)<br>*Shanghai Jiao Tong University*<br>http://arxiv.org/abs/2001.09694 | 90.578 | 92.978 |

# Task-Driven Culture

## Multitask Leaderboards

GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

**Alex Wang**[1], **Amanpreet Singh**[1], **Julian Michael**[2], **Felix Hill**[3], **Omer Levy**[2] & **Samuel R. Bowman**[1]

**SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems**

**Alex Wang***
New York University

**Yada Pruksachatkun***
New York University

**Nikita Nangia***
New York University

**Amanpreet Singh***
Facebook AI Research

**Julian Michael**
University of Washington

**Felix Hill**
DeepMind

**Omer Levy**
Facebook AI Research

**Samuel R. Bowman**
New York University

# Task-Driven Culture

## Multitask Leaderboards

Leaderboard Version: **2.0**

| Rank | Name | Model | URL | Score | BoolQ | CB | COPA | MultiRC | ReCoRD | RTE | WiC | WSC | AX-b | AX-g |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Liam Fedus | ST-MoE-32B | | 91.2 | 92.4 | 96.9/98.0 | 99.2 | 89.6/65.8 | 95.1/94.4 | 93.5 | 77.7 | 96.6 | 72.3 | 96.1/94.1 |
| 2 | Microsoft Alexander v-team | Turing NLR v5 | | 90.9 | 92.0 | 95.9/97.6 | 98.2 | 88.4/63.0 | 96.4/95.9 | 94.1 | 77.1 | 97.3 | 67.8 | 93.3/95.5 |
| 3 | ERNIE Team - Baidu | ERNIE 3.0 | | 90.6 | 91.0 | 98.6/99.2 | 97.4 | 88.6/63.2 | 94.7/94.2 | 92.6 | 77.4 | 97.3 | 68.6 | 92.7/94.7 |
| 4 | Yi Tay | PaLM 540B | | 90.4 | 91.9 | 94.4/96.0 | 99.0 | 88.7/63.6 | 94.2/93.3 | 94.1 | 77.4 | 95.9 | 72.9 | 95.5/90.4 |
| 5 | Zirui Wang | T5 + UDG, Single Model (Google Brain) | | 90.4 | 91.4 | 95.8/97.6 | 98.0 | 88.3/63.0 | 94.2/93.5 | 93.0 | 77.9 | 96.6 | 69.1 | 92.7/91.9 |
| 6 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.3 | 90.4 | 95.7/97.6 | 98.4 | 88.2/63.7 | 94.5/94.1 | 93.2 | 77.5 | 95.9 | 66.7 | 93.3/93.8 |
| 7 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8/51.9 | 91.7/91.3 | 93.6 | 80.0 | 100.0 | 76.6 | 99.3/99.7 |

# Task-Driven Culture

## Multitask Leaderboards

Leaderboard Version: **2.0**

| Rank | Name | Model | URL | Score | BoolQ | CB | COPA | MultiRC | ReCoRD | RTE | WiC | WSC | AX-b | AX-g |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Liam Fedus | ST-MoE-32B | | 91.2 | 92.4 | 96.9/98.0 | 99.2 | 89.6/65.8 | 95.1/94.4 | 93.5 | 77.7 | 96.6 | 72.3 | 96.1/94.1 |
| 2 | Microsoft Alexander v-team | Turing NLR v5 | | 90.9 | 92.0 | 95.9/97.6 | 98.2 | 88.4/63.0 | 96.4/95.9 | 94.1 | 77.1 | 97.3 | 67.8 | 93.3/95.5 |
| 3 | ERNIE Team - Baidu | ERNIE 3.0 | | 90.6 | 91.0 | 98.6/99.2 | 97.4 | 88.6/63.2 | 94.7/94.2 | 92.6 | 77.4 | 97.3 | 68.6 | 92.7/94.7 |
| 4 | Yi Tay | PaLM 540B | | 90.4 | 91.9 | 94.4/96.0 | 99.0 | 88.7/63.6 | 94.2/93.3 | 94.1 | 77.4 | 95.9 | 72.9 | 95.5/90.4 |
| 5 | Zirui Wang | T5 + UDG, Single Model (Google Brain) | | 90.4 | 91.4 | 95.8/97.6 | 98.0 | 88.3/63.0 | 94.2/93.5 | 93.0 | 77.9 | 96.6 | 69.1 | 92.7/91.9 |
| 6 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.3 | 90.4 | 95.7/97.6 | 98.4 | 88.2/63.7 | 94.5/94.1 | 93.2 | 77.5 | 95.9 | 66.7 | 93.3/93.8 |
| 7 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8/51.9 | 91.7/91.3 | 93.6 | 80.0 | 100.0 | 76.6 | 99.3/99.7 |

# Task-Driven Culture

## Multitask Leaderboards

- GLUE: Set of text classification tasks
- SuperGLUE: Better set of text classification tasks
- GEM: Set of text generation tasks
- DynaBench: “Living” benchmark (people/models keep adding test items)
- BigBench: Pool all of the above and more!
- …

# Task-Driven Culture

## Summary

- Standardized tasks and evals are a core part of NLP research culture
- We (the field) places a lot of stock in:
  - Empirical comparison of ideas and
  - Applicability of technology (i.e., improving downstream tasks)
- If you want to understand what NLP people are working on and why at a given point in time, look at the tasks and evaluations that are dominant (this changes with time and with technology)
- Recently, this culture is changing a bit (more at end of lecture)

# Topics

- Lecture 0 Reprise
- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Topics

- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
  - Information Retrieval
  - Question Answering
  - Machine Translation
  - Summarization
  - Dialog
  - Classification
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Topics

- Task-Driven Culture
- **Major "Downstream" Tasks and Evaluations**
  - Information Retrieval
  - Question Answering
  - Machine Translation
  - ~~Summarization~~
  - ~~Dialog~~
  - Classification
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

Super important, and tons to say, but we just don't have time :(

# Topics

- Task-Driven Culture
- **Major "Downstream" Tasks and Evaluations**
  - Information Retrieval
  - Question Answering
  - ~~Machine Translation~~
  - ~~Summarization~~
  - ~~Dialog~~
  - Classification
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

We'll talk a lot later in the semester

# Topics

- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
  - Information Retrieval
  - Question Answering
  - Classification
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Topics

- Task-Driven Culture
- **Major "Downstream" Tasks and Evaluations**
  - Classification
  - Information Retrieval
  - Question Answering
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

Better order for introducing metrics :)

# Topics

- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
  - **Classification**
  - Information Retrieval
  - Question Answering
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Classification

## Inputs-Outputs

- Input: A piece of text
- Output: A label (usually from a discrete set)

# Classification

## Examples

- Ubiquitous! Any working NLP system will likely depend on tons of little classifiers running under the hood
  - Probably more intrinsic than extrinsic
- Examples:
  - Sentiment prediction (e.g., for stock trading)
  - Language detection (e.g., before machine translation)
  - Relevance prediction (e.g., for retrieval or ad targeting)
  - Intent classification (e.g., for goal-oriented dialog)

# Classification

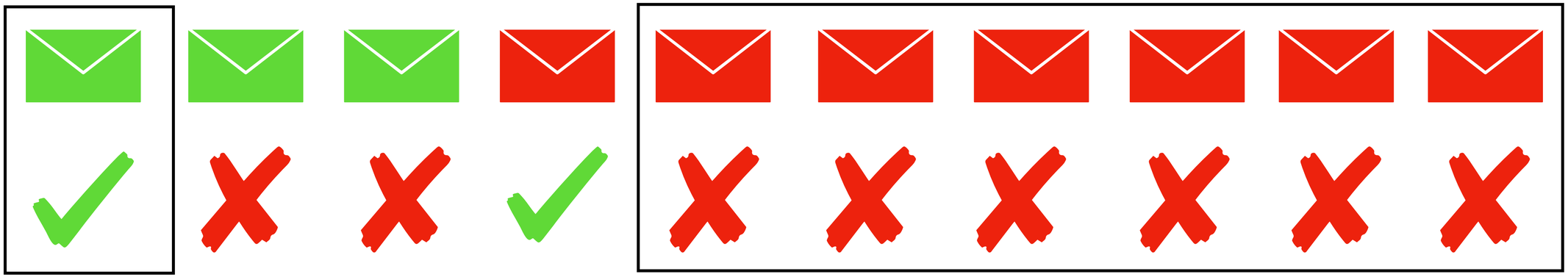
## Standard Metrics

- Accuracy: How often is the label correct?
- Precision: Probability that it is spam, given that the model says it is.
- Recall: Probability that the model says its spam, given that it actually is spam.
- Summarizing Precision and Recall:
  - F1: Harmonic Mean of Precision and Recall
  - AUC: Area Under the True-Positive vs. False Positive Curve

# Classification

## Standard Metrics

Example: Spam FIlter

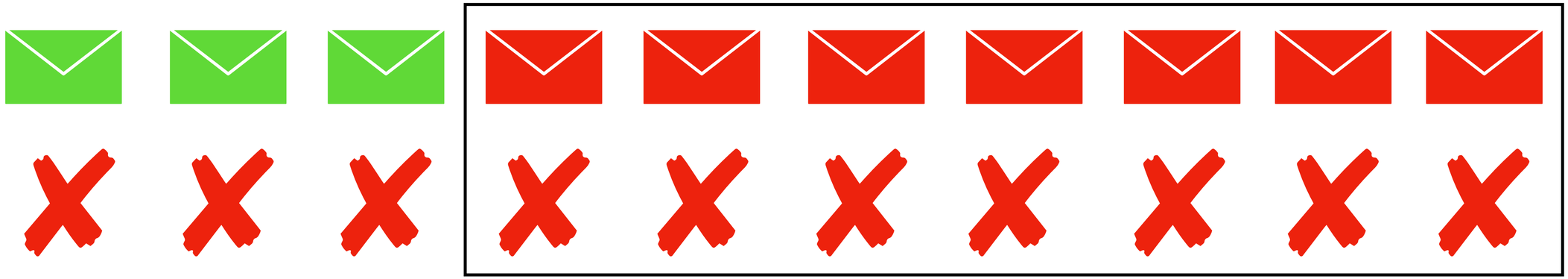


Accuracy = 7/10 = 70%

# Classification

## Standard Metrics

Example: Spam FIlter

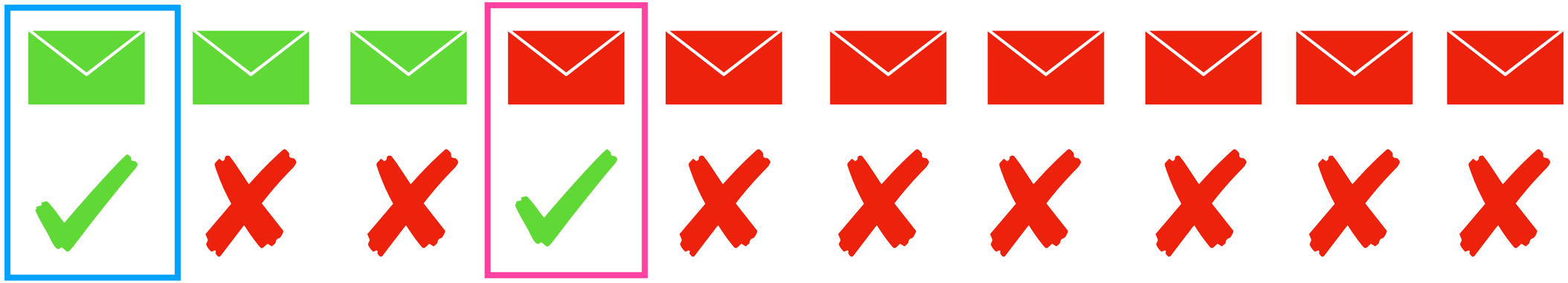


Accuracy = 7/10 = 70%

# Classification

Example: Spam FIlter

## Standard Metrics

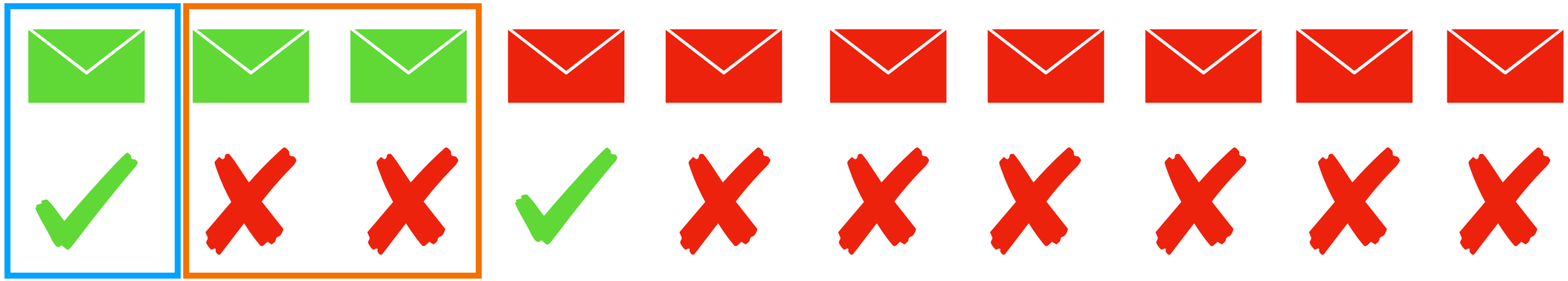


$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{1}{2}$$

# Classification

Example: Spam FIlter

## Standard Metrics



$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{1}{3}$$

# Classification

## Standard Metrics

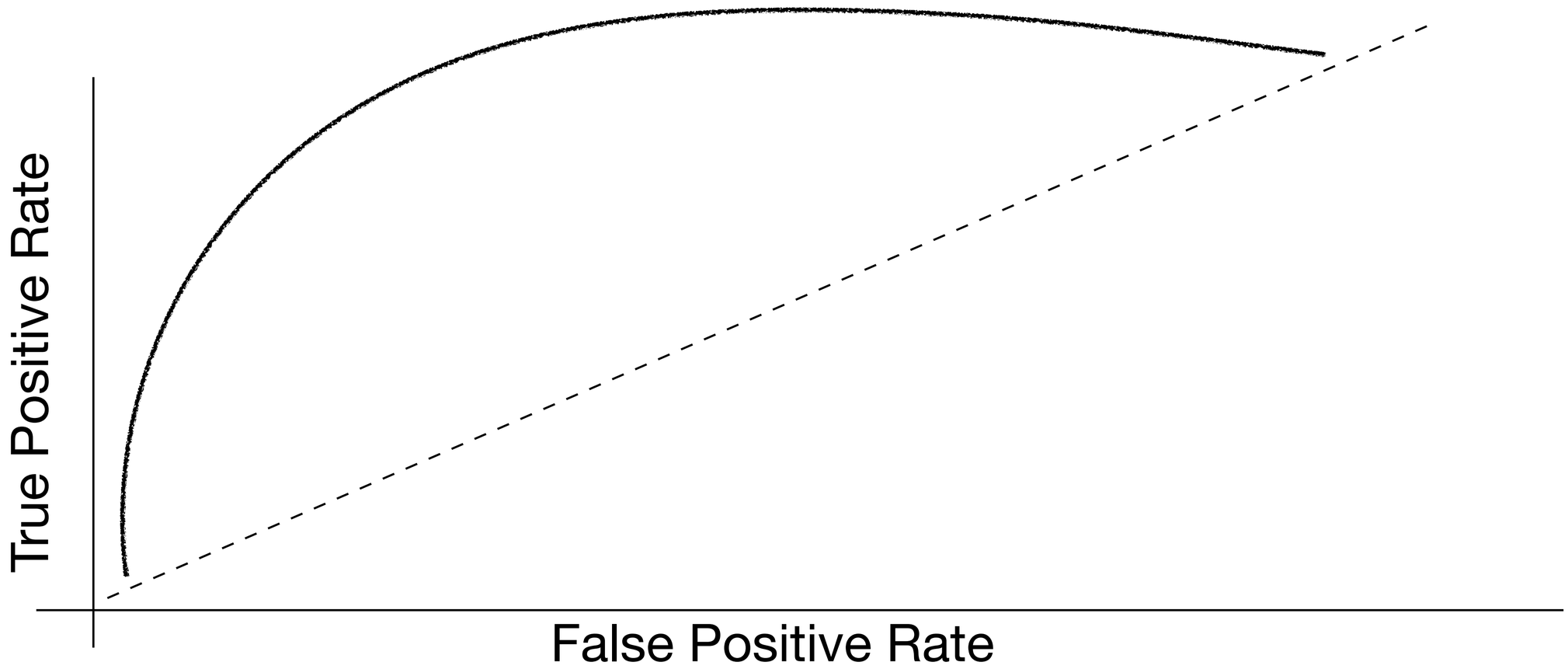
Example: Spam FIlter

$$F1 = \frac{2 * P * R}{P + R}$$

# Classification

## Standard Metrics

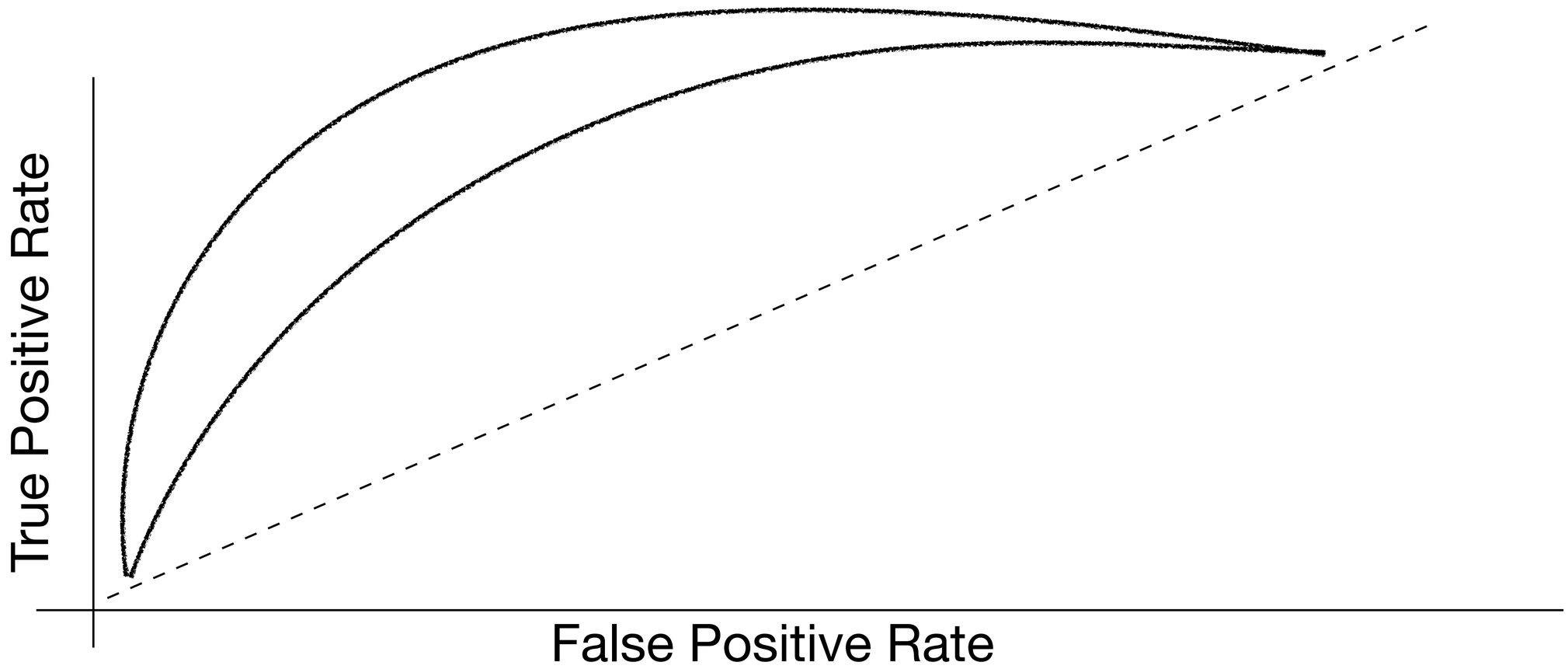
- AUC: Area Under TP-FP Curve



# Classification

## Standard Metrics

- AUC: Area Under TP-FP Curve



# Topics

- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
  - Classification
  - **Information Retrieval**
  - Question Answering
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Information Retrieval

## Inputs-Outputs

- Input:
  - Query (typically natural language string)
  - Document Collection (typically text, but can be more varied)
- Output:
  - Ranked List of Documents

# Information Retrieval

## Examples

- Traditionally: “10 blue links”

providence democratic primary lieutenant governor

ALL IMAGES VIDEOS MAPS NEWS SHOPPING MORE

11-24 Of 348,000,000 Results Any time

State Rep. Deb Ruggiero will run for RI lieutenant governor
https://www.wpri.com/.../state-rep-deb-ruggiero-will-run-for-ri-lieutenant-governor
Apr 04, 2022 · Updated: Apr 4, 2022 / 04:46 PM EDT. **PROVIDENCE**, R.I. (WPRI) — State Rep. Deb Ruggiero announced Monday she’s running for **lieutenant governor**, signaling the **Democratic** Party will have a ...

Lieutenant Governor of Rhode Island - Ballotpedia
https://ballotpedia.org/Lieutenant_Governor_of_Rhode_Island
**Democratic primary** for **Lieutenant Governor of Rhode Island**. Incumbent Sabina Matos, ... Office of the **Lt. Governor** 82 Smith Street, Room 116 **Providence**, RI 02903 Phone: (401) 222-2371. See also. Rhode Island State Executive Elections News …

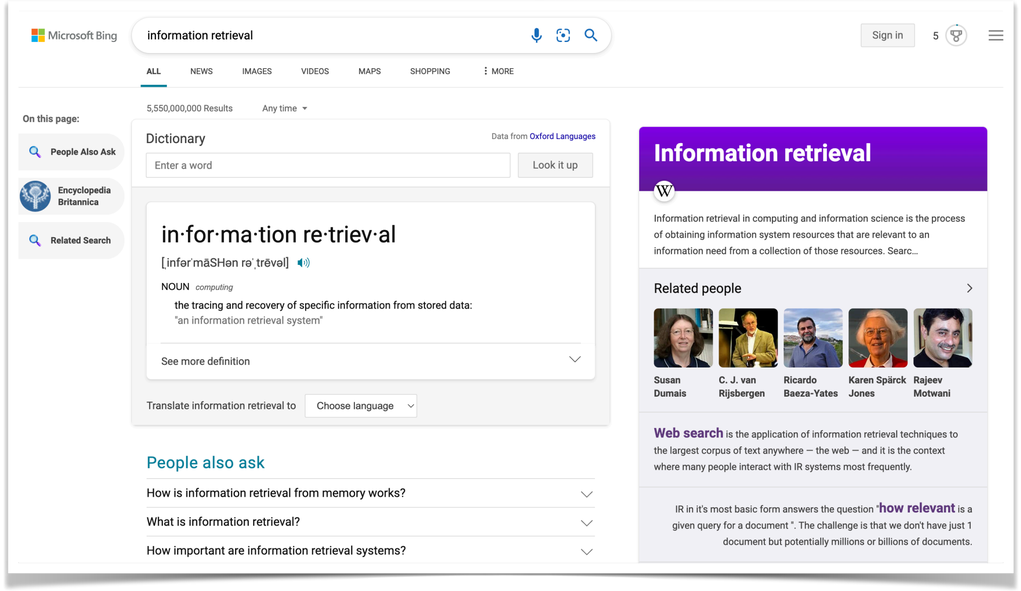
Election 2022 Update: Former RI Health Director Alexander-Scott ...
https://www.yahoo.com/news/election-2022...
Sep 12, 2022 · A day ahead of Tuesday's five-way **Democratic primary**, former Rhode Island Health Director Nicole Alexander-Scott endorsed former CVS executive Helena Foulkes for **governor** over the incumbent, Dan ...

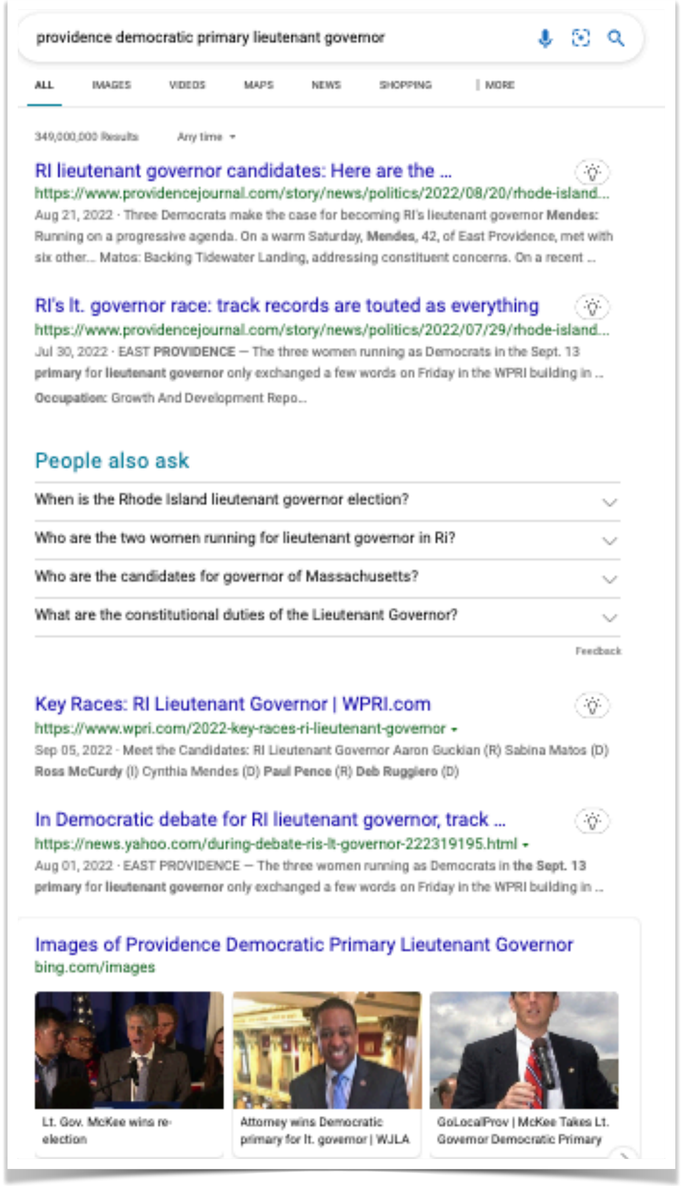
McKee expected to name Matos lieutenant governor - The Providence Journal
https://www.providencejournal.com/story/news/politics/2021/03/30/ri-lieutenant...
Mar 30, 2021 · 0:00. 1:11. **PROVIDENCE** — Gov. Dan McKee will announce his choice for **lieutenant governor** at a press conference on Wednesday, and political oddsmakers are placing their bets on **Providence** City ...

# Information Retrieval

## Examples

- Traditionally: “10 blue links”
- Increasingly complex goals, which are interwoven with other NLP tasks (summarization, question answering)



# Information Retrieval

## Standard Datasets

- Text REtrieval Conference (TREC)
  - **https://trec.nist.gov/**
  - Started in 1992
  - DOD supported, focus on improving IR
- Originally, a long query “topic statement”
- Later queries became more succinct, and also more tasks became more varied (clinical text, social media)

# Information Retrieval

## Standard Datasets

<num> Number: 051

<dom> Domain: International Economics

<title> Topic: Airbus Subsidies

<desc> Description:
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.

<narr> Narrative:
A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

TREC 1 & 2

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative:
A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

TREC 3

<num> Number: 207

<desc> What are the prospects of the Quebec separatists achieving independence from the rest of Canada?

TREC 4

# Information Retrieval

## Standard Datasets

<num> Number: 051

<dom> Domain: International Economics

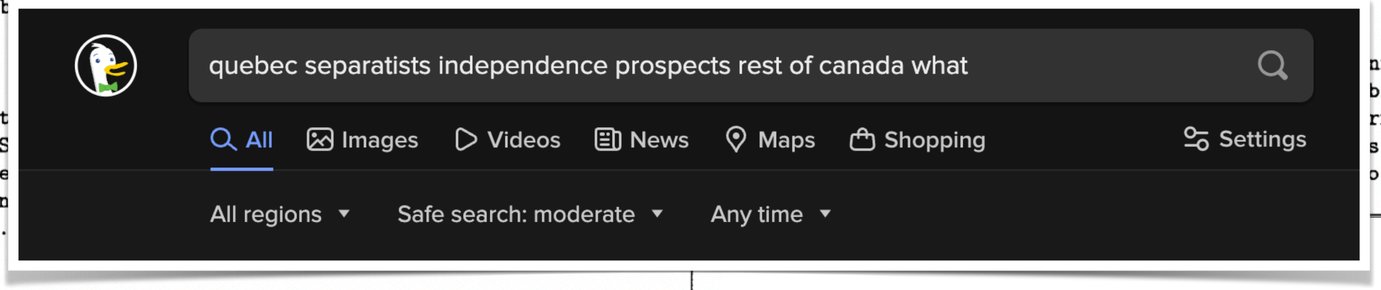
<title> Topic: Airbus Subsidies

<desc> Description:
Document will discuss government assistance to Airbus Industrie, or mention
a trade dispute between Airb
subsidies.

<narr> Narrative:
A relevant document will cit
French, German, British or S
between Airbus or the Europe
likely Boeing Co. or McDonn
federal subsidies to Airbus.

<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

TREC 1 & 2




<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

nformation on the government's responsibility
ble entity. It could also discuss the
rnative to continuing government subsidies.
sidies given to air and bus transportation with
o be relevant.

TREC 3

<num> Number: 207

<desc> What are the prospects of the Quebec separatists achieving independence from the rest of Canada?

TREC 4

# Information Retrieva

## Standard Datasets

**MS MARCO: A Human Generated MAchine Reading COmprehension Dataset**

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang
Microsoft AI & Research

| | |
|---|---|
| 1035339 | who is linda walker aa |
| 943613 | when do buff orpingtons start laying eggs |
| 1051868 | who sings the rodeo clown song |
| 876108 | what makes your hands burn |
| 770613 | what is medicare summary notice |
| 84901 | causes for left shoulder and bicep |
| 928755 | what year was william shakespeare born |
| 895787 | what size sim card does the lg g stylo use |
| 920435 | what was the first mammal cloned? |
| 1009016 | which elements had complete outer shells |
| 923221 | what was washington's teeth made of |
| 1126106 | population of new rochelle |
| 988661 | where is murphy nc |
| 975821 | where in the usa is the agaricus blazei murill mushroom found? |
| 154441 | dm during pregnancy |
| 802079 | what is talalay latex |
| 184355 | extrovert and introvert definition |
| 938754 | when did great britain and france declare war on germany |
| 1008125 | which cruise ships can fit in panama canal? |
| 794144 | what is schnitzel |

# Information Retrieval

## Standard Metrics

- Precision@K: How many of the top K documents are relevant?
- Recall@K: How many of the relevant documents occur in the top K?
- Mean Reciprocal Rank: At what rank does the relevant document appear?
  - Appropriate when there is only one “right” answer (one relevant document)
- Discounted Cumulative Gain (DCG): In general, how much benefit is there to going one more rank down the list?
  - Appropriate when there are graded relevance judgments
- Area Under the Curve (AUC), Averaged Precision (AP): Summary metrics for precision-recall curves

# Information Retrieval

## Standard Metrics

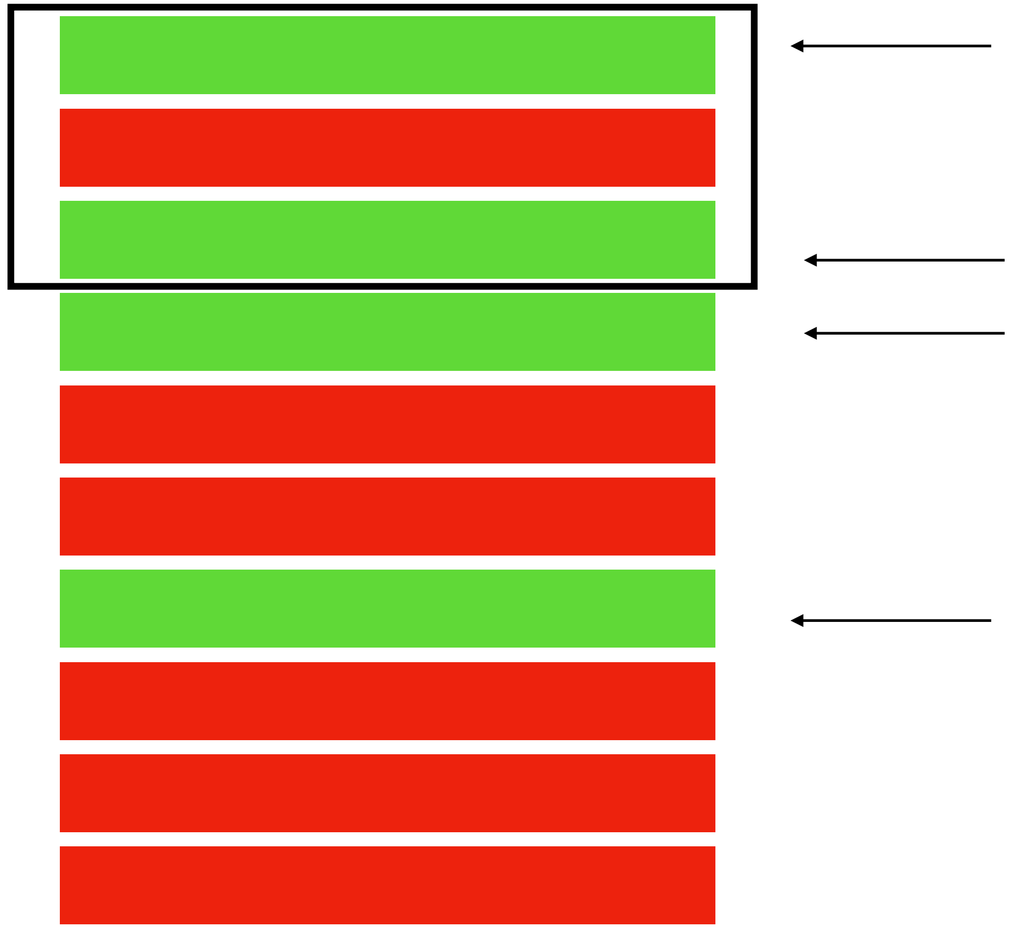
Ranked Documents

Precision @ 3

= 2/3 = 0.67

# Information Retrieval

## Standard Metrics



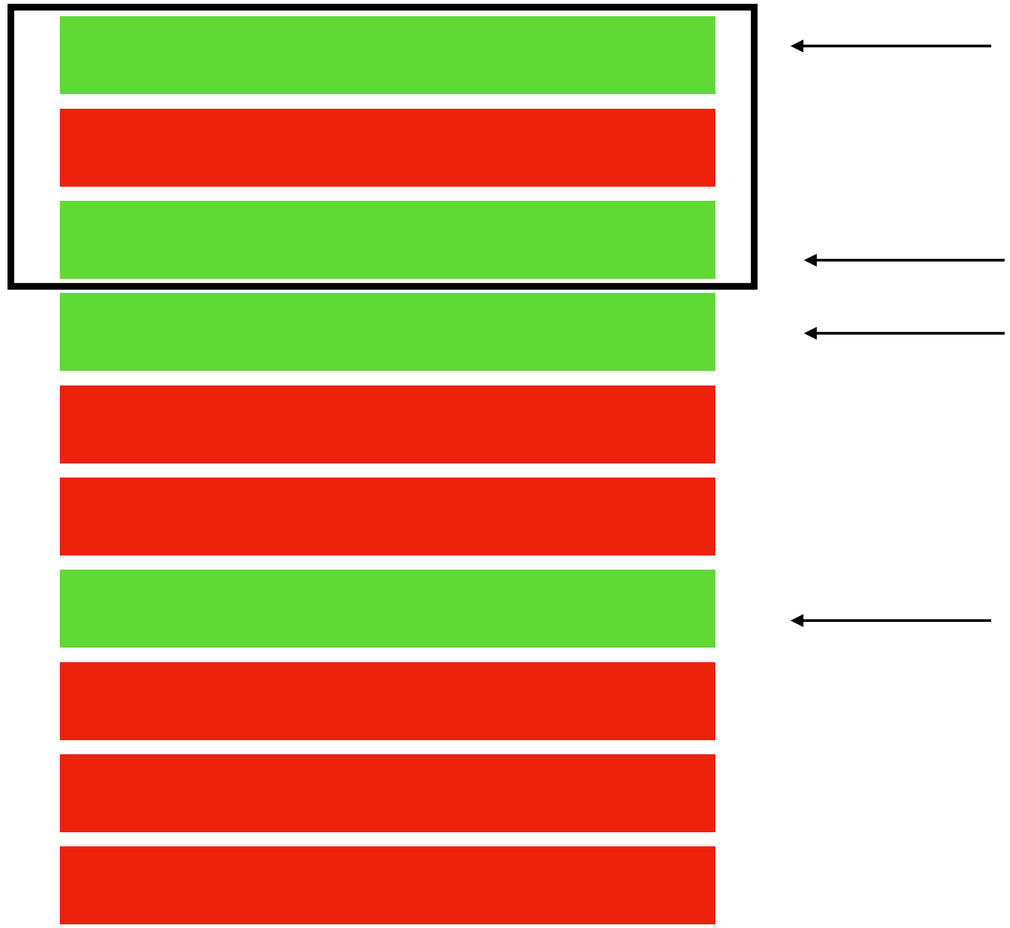
Ranked Documents

Recall @ 3

= 2/4 = 0.50

# Information Retrieval

## Standard Metrics



Ranked Documents

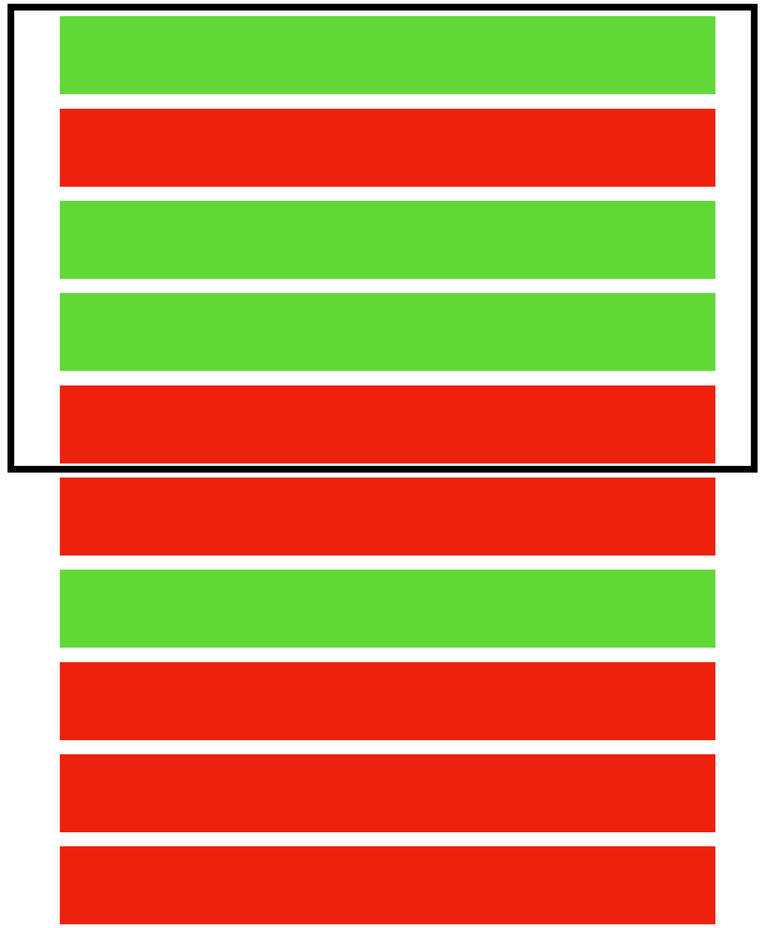
Recall @ 3

= 2/4 = 0.50

Not very common! Requires finding and annotating all relevant documents in the whole corpus.
(I.e., reading the whole internet...?)

# Information Retrieval

## Standard Metrics



Ranked Documents

Precision @ 5

= 3/5 = 0.6

# Information Retrieval

## Standard Metrics

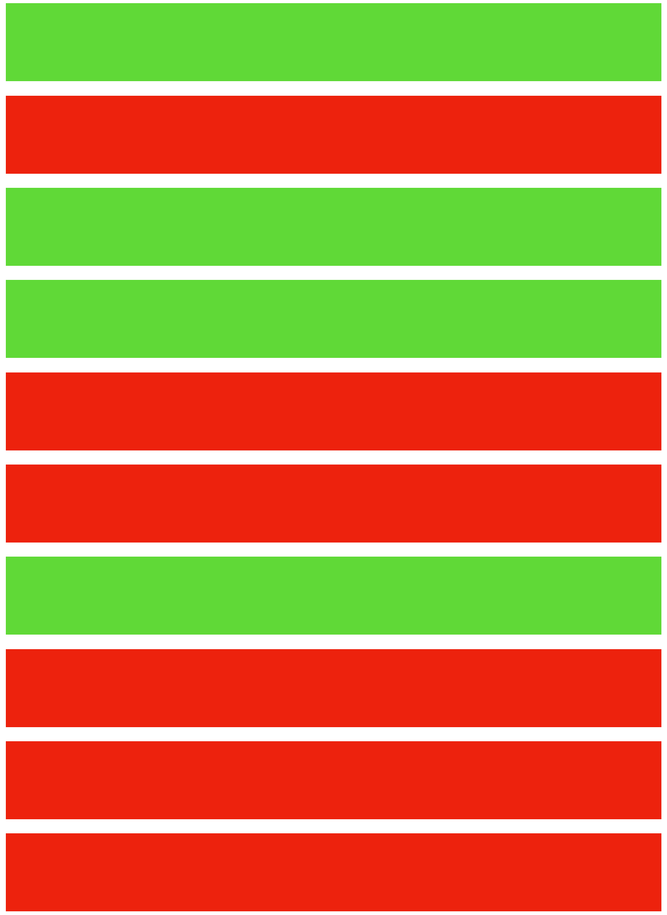
Ranked Documents

Precision @ 10

= 4/10 = 0.4

# Information Retrieval

## Standard Metrics



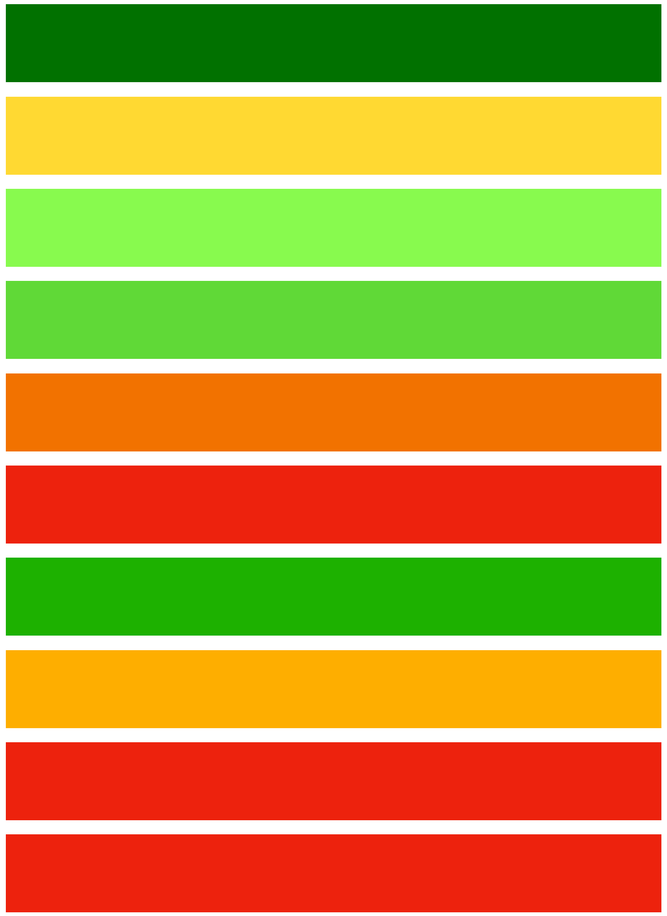
Ranked Documents

Average Precision

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

# Information Retrieval

## Standard Metrics



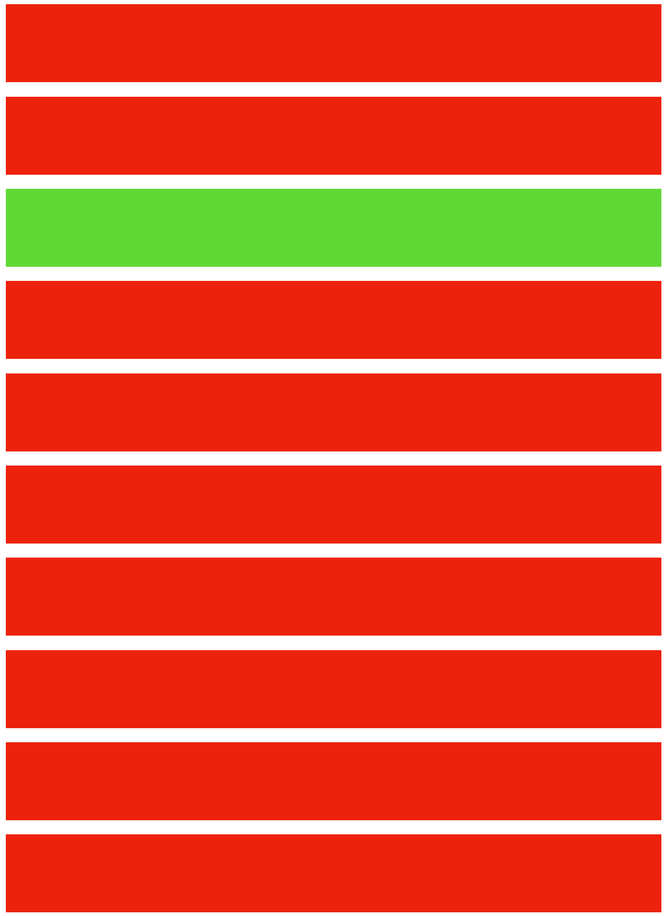
Ranked Documents

Discounted Cumulative Gain

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$$

# Information Retrieval

## Standard Metrics



Ranked Documents

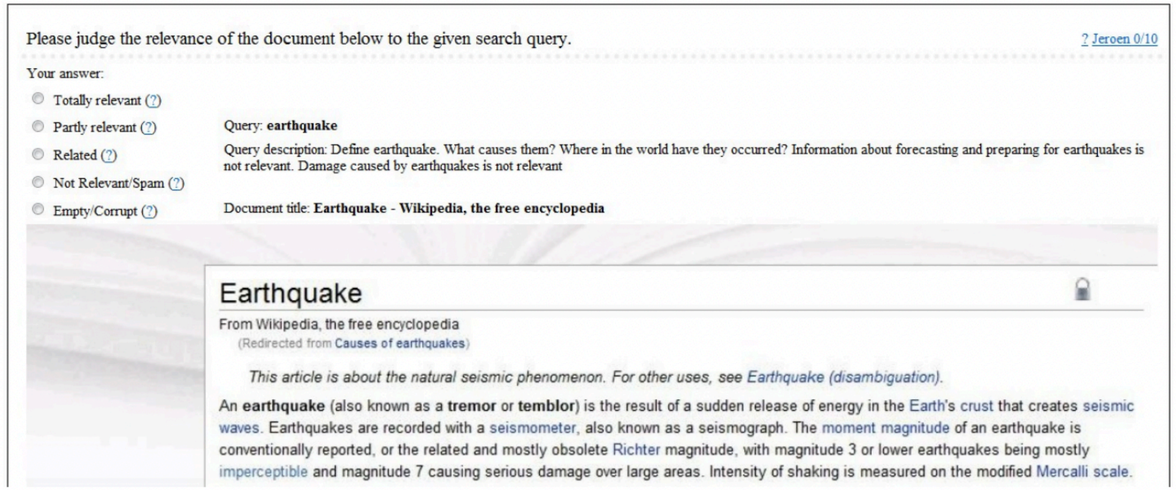
Mean Reciprocal Rank

= 1/3

# Information Retrieval

## Establishing “Ground Truth”

- Relevance Judgements, a.k.a. “qrels”
  - Humans manually annotate documents for relevance, based on description of query
  - Too expensive to estimate recall (requires labeling every document for every query)
  - Not “ecologically valid”
- User Behavior/AB Testing
  - Does the user click on links? Do they stay on the page after clicking? Do they go back and reword the query?




https://trec.nist.gov/pubs/trec20/papers/TUD_DMIR.crowdsourcing.4-20update.pdf

# Topics

- Task-Driven Culture
- **Major “Downstream” Tasks and Evaluations**
  - Classification
  - Information Retrieval
  - **Question Answering**
- Some Important Intrinsic Tasks and Evaluations
- Current Debates and Open Questions

# Question Answering

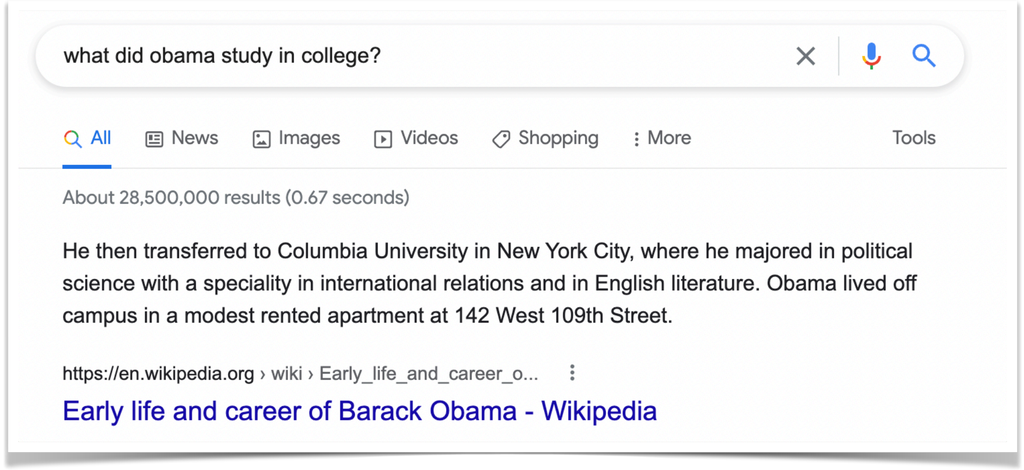
## Inputs-Outputs

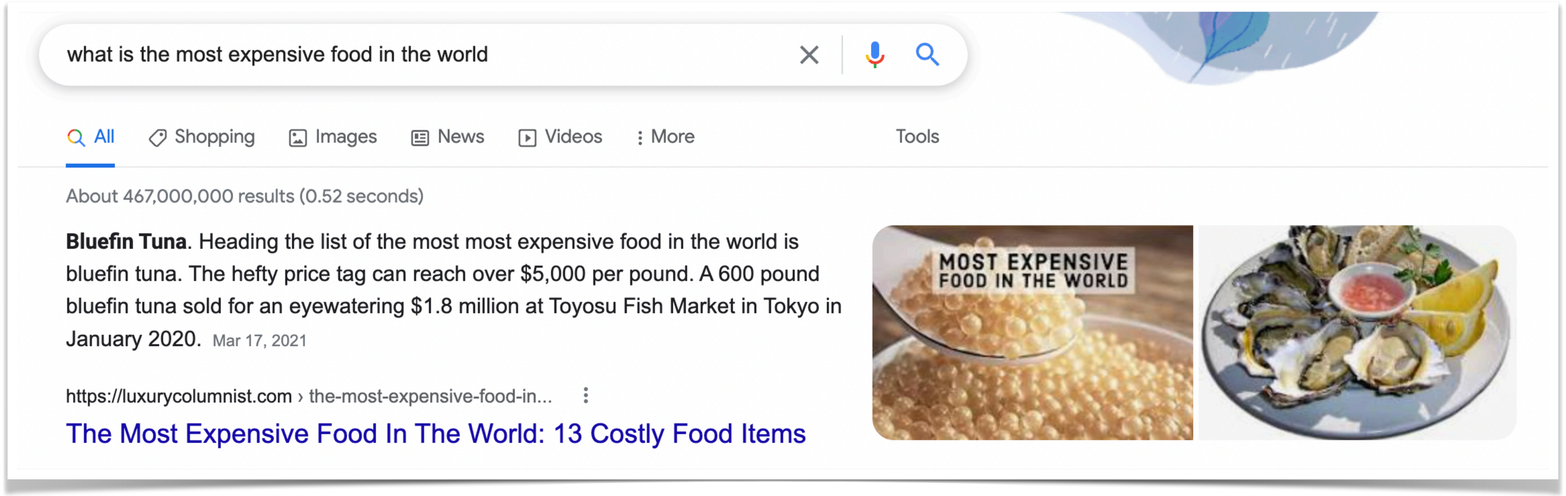
- Many variants of “question answering”
- Three primary variants:
  - Open Book
  - Closed Book
  - QA over Databases (aka Natural Language Interfaces to Databases aka NLIDB)

# Open Book Question Answering

## Inputs-Outputs

- Input: Question+Doc
- Output: Answer/Highlighted Span



# Open Book Question Answering

## Standard Datasets

- Message Understanding Conference (MUC)
  - DARPA funded open-book QA tasks, starting in 1987
  - Simple questions and answers stated clearly in the passage, e.g., extracting names, dates, locations
  - Spurred much of the work on “intrinsic” evaluations (e.g., parsing)

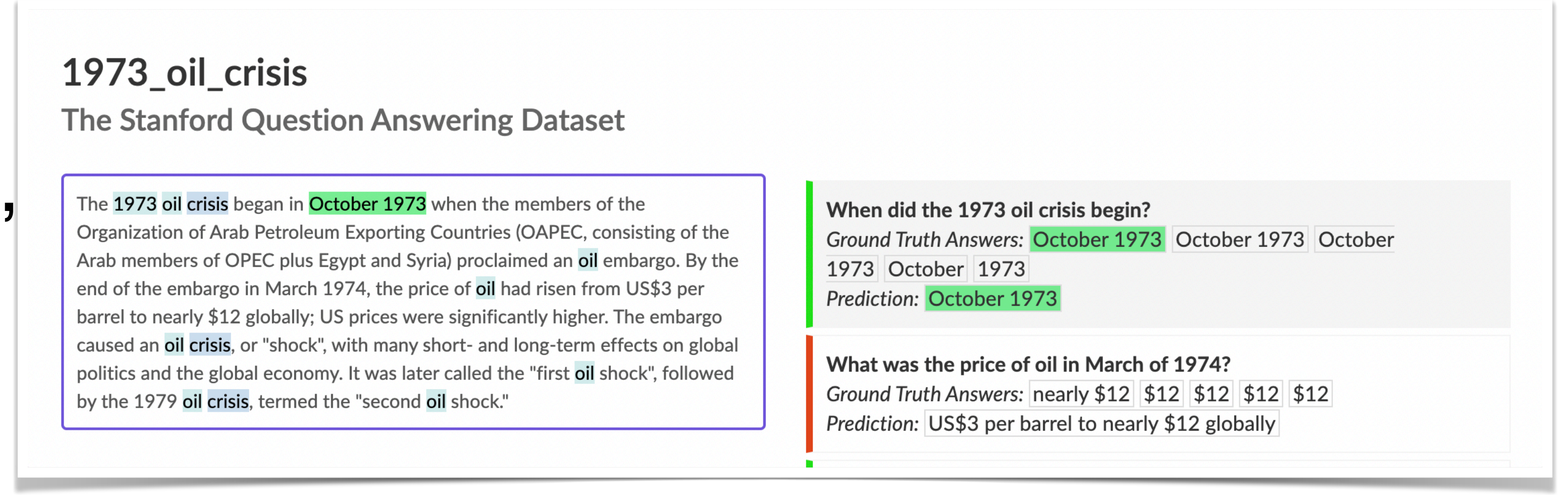
We are striving to have a strong renewed creative partnership with Coca-Cola,” Mr. Dooner says. However, odds of that happening are slim since word from Coke headquarters in Atlanta is that...

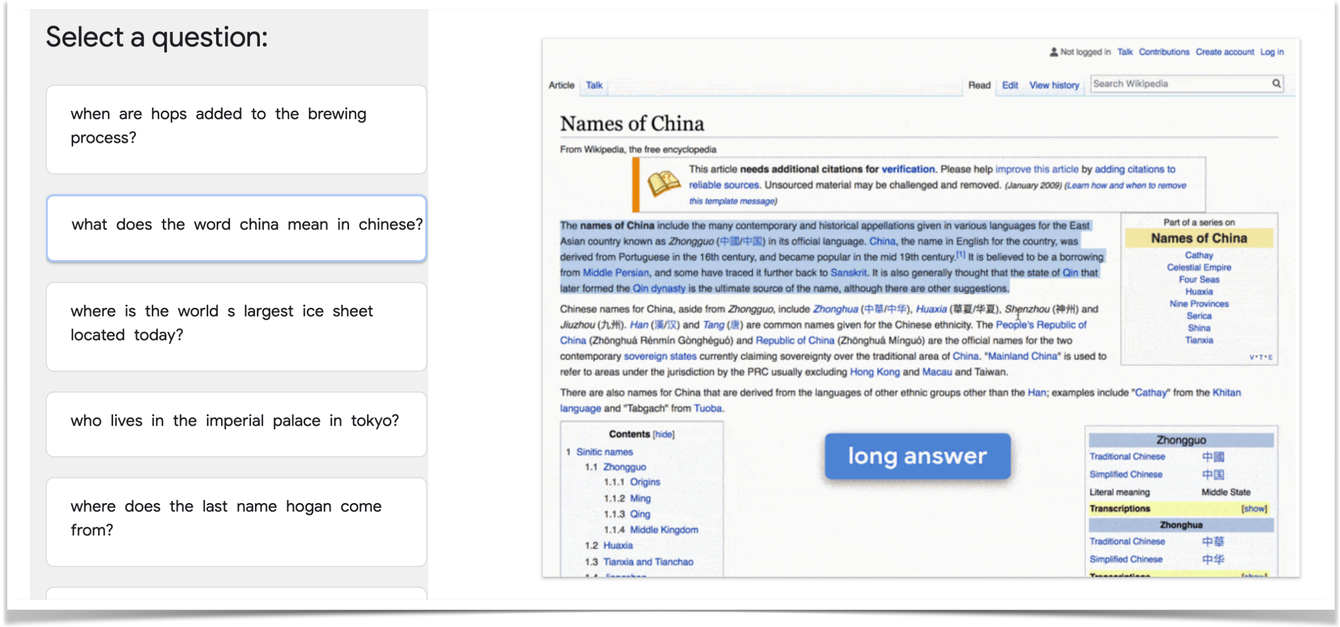
```
<ORGANIZATION-9402240133-5> :=
    ORG_NAME: "Coca-Cola"
    ORG_ALIAS: "Coke"
    ORG_TYPE: COMPANY
    ORG_LOCALE: Atlanta CITY
    ORG_COUNTRY: United States
```

# Open Book Question Answering

## Standard Datasets

- “Reading Comprehension”
- SQuAD
- Natural Questions



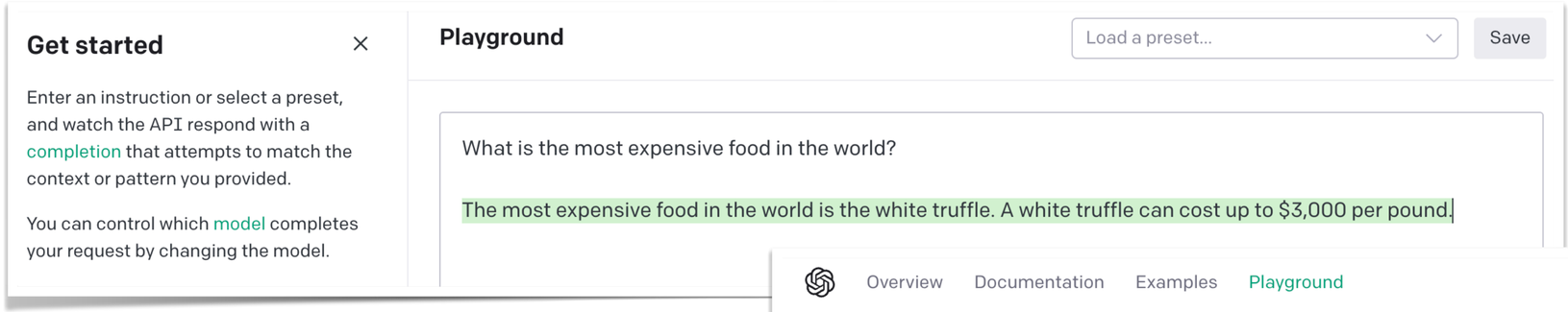



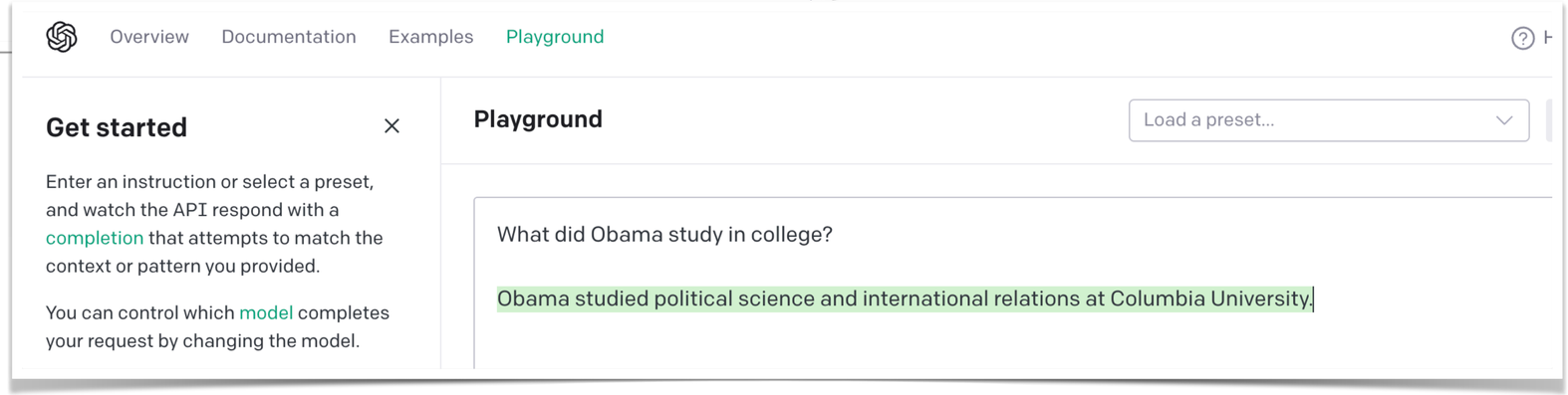

https://rajpurkar.github.io/SQuAD-explorer/
https://ai.google.com/research/NaturalQuestions/visualization

# Closed Book Question Answering

## Inputs-Outputs

- Input: Question
- Output: Answer
- Risky: hard to verify answer source



# Closed Book Question Answering

## Standard Datasets

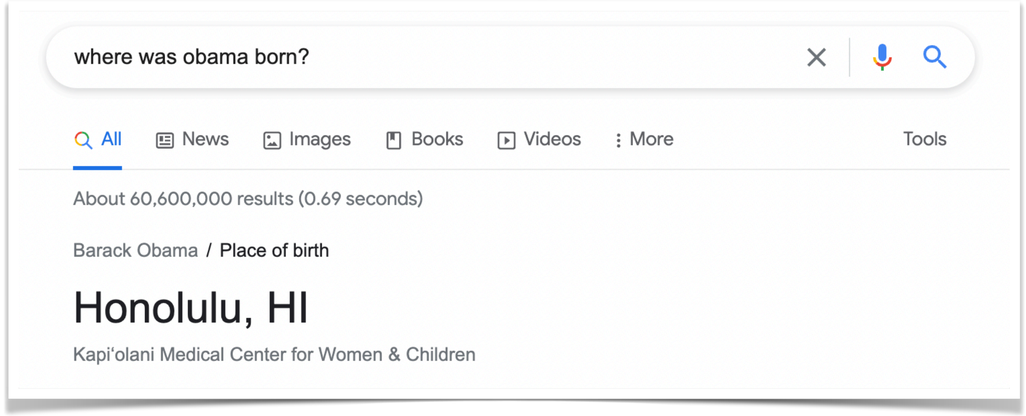
- New task, fewer standardized datasets
- LAMA: Factoid questions generated by language models

| | Relation | Query | Answer | Generation |
|---|---|---|---|---|
| T-Rex | P19 | Francesco Bartolomeo Conti was born in ____. | Florence | Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5] |
| | P20 | Adolphe Adam died in ____. | Paris | **Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0] |
| | P279 | English bulldog is a subclass of ____. | dog | dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5] |
| | P37 | The official language of Mauritius is ____. | English | **English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0] |
| | P413 | Patrick Oboya plays in ____ position. | midfielder | centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7] |
| | P138 | Hamburg Airport is named after ____. | Hamburg | Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5] |
| | P364 | The original language of Mon oncle Benjamin is ____. | French | **French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9] |
| | P54 | Dani Alves plays with ____ . | Barcelona | Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7] |
| | P106 | Paul Toungui is a ____ by profession . | politician | lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7] |
| | P527 | Sodium sulfide consists of ____. | sodium | water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9] |
| | P102 | Gordon Scholes is a member of the ____ political party. | Labor | Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]** |
| | P530 | Kenya maintains diplomatic relations with ____. | Uganda | India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6] |
| | P176 | iPod Touch is produced by ____. | Apple | **Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1] |
| | P30 | Bailey Peninsula is located in ____. | Antarctica | **Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1] |
| | P178 | JDK is developed by ____. | Oracle | IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5] |
| | P1412 | Carl III used to communicate in ____. | Swedish | German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0] |
| | P17 | Sunshine Coast, British Columbia is located in ____. | Canada | **Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4] |
| | P39 | Pope Clement VII has the position of ____ . | pope | cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2] |
| | P264 | Joe Cocker is represented by music label ____. | Capitol | EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3] |
| | P276 | London Jazz Festival is located in ____. | London | **London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8] |
| | P127 | Border TV is owned by ____. | ITV | Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3] |
| | P103 | The native language of Mammootty is ____. | Malayalam | **Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6] |
| | P495 | The Sharon Cuneta Show was created in ____. | Philippines | Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0] |
| ConceptNet | AtLocation | You are likely to find a overflow in a ____. | drain | sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]** |
| | CapableOf | Ravens can ____. | fly | **fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4] |
| | CausesDesire | Joke would make you want to ____. | laugh | cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6] |
| | Causes | Sometimes virus causes ____. | infection | disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4] |
| | HasA | Birds have ____. | feathers | wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9] |
| | HasPrerequisite | Typing requires ____. | speed | patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]** |
| | HasProperty | Time is ____. | finite | short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0] |
| | MotivatedByGoal | You would celebrate because you are ____. | alive | happy [-2.4] , human [-3.3] , **alive [-3.3]** , young [-3.6] , free [-3.9] |
| | ReceivesAction | Skills can be ____. | taught | acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9] |
| | UsedFor | A pond is for ____. | fish | swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , **fish [-2.8]** , recreation [-3.1] |

# Database Question Answering

## Inputs-Outputs

- Input: Question+Database
- Output: Answer
- Good for short, factoid questions
- Requires question to be parsed into a formal query language

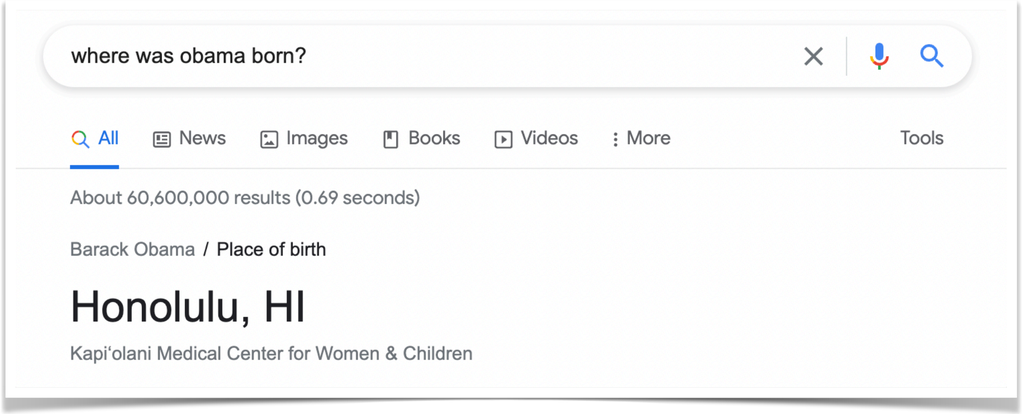


# Database Question Answering

## Inputs-Outputs

- Input: Question+Database
- Output: Answer
- Good for short, factoid questions
- Requires question to be parsed into a formal query language

More in the semantic parsing lecture!



# Question Answering

## Standard Metrics

- Accuracy: What % of questions were correct?
  - Appropriate for multiple choice
- “Exact Match” Accuracy
  - % of times model’s answer was string identical to ground truth
- Precision/Recall/F1 of covered tokens
  - Precision: How many of the predicted tokens were correct?
  - Recall: How many of the correct tokens were predicted?
  - F1: Harmonic mean of precision and recall (2PR/P+R)

# Question Answering

## Standard Metrics

Precision = 2/2 = 1.0

Recall = 2/2 = 1.0

### 1973_oil_crisis

The Stanford Question Answering Dataset

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

**When did the 1973 oil crisis begin?**

*Ground Truth Answers:* October 1973 | October 1973 | October 1973 | October | 1973

*Prediction:* October 1973

**What was the price of oil in March of 1974?**

*Ground Truth Answers:* nearly $12 | $12 | $12 | $12 | $12

*Prediction:* US$3 per barrel to nearly $12 globally

# Question Answering

## Standard Metrics

Precision = 2/7 = 0.3
Recall = 2/2 = 1.0

### 1973_oil_crisis

The Stanford Question Answering Dataset

The 1973 oil crisis began in October 1973 when the members of the Organization of Arab Petroleum Exporting Countries (OAPEC, consisting of the Arab members of OPEC plus Egypt and Syria) proclaimed an oil embargo. By the end of the embargo in March 1974, the price of oil had risen from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short- and long-term effects on global politics and the global economy. It was later called the "first oil shock", followed by the 1979 oil crisis, termed the "second oil shock."

**When did the 1973 oil crisis begin?**
*Ground Truth Answers:* October 1973 | October 1973 | October 1973 | October | 1973
*Prediction:* October 1973

**What was the price of oil in March of 1974?**
*Ground Truth Answers:* nearly $12 | $12 | $12 | $12 | $12
*Prediction:* US$3 per barrel to nearly $12 globally

# Topics

- Lecture 0 Reprise
- Task-Driven Culture
- Major “Downstream” Tasks and Evaluations
- **Some Important Intrinsic Tasks and Evaluations**
- Current Debates and Open Questions

# Important Intrinsic Tasks

## Behind the Scenes of a QA System

| | | | |
|---|---|---|---|
| 59<br>Sep 28, 2016 | Dynamic Chunk Reader<br>*IBM*<br>https://arxiv.org/abs/1610.09996 | 62.499 | 70.956 |
| 59<br>Nov 15, 2019 | RQA+IDR (single model)<br>*BUAA & MSRA*<br>https://arxiv.org/abs/2005.02925 | 61.145 | 71.389 |
| | s-Ptr (Boundary)<br>*ment University*<br>s/1608.07905 | 60.474 | 70.695 |
| | s-Ptr (Sentence)<br>*ment University*<br>s/1608.07905 | 54.505 | 67.748 |
| Nov 15, 2019 | model)<br>*BUAA & MSRA*<br>https://arxiv.org/abs/2005.02925 | 55.827 | 65.467 |

**End-to-End Answer Chunk Extraction and Ranking for Reading Comprehension**

**Yang Yu*, Wei Zhang*, Kazi Hasan, Mo Yu, Bing Xiang, Bowen Zhou**
{yu, zhangwei, kshasan, yum, bingxia, zhou}@us.ibm.com
IBM Watson

# Important Intrinsic Tasks

## Behind the Scenes of a QA System

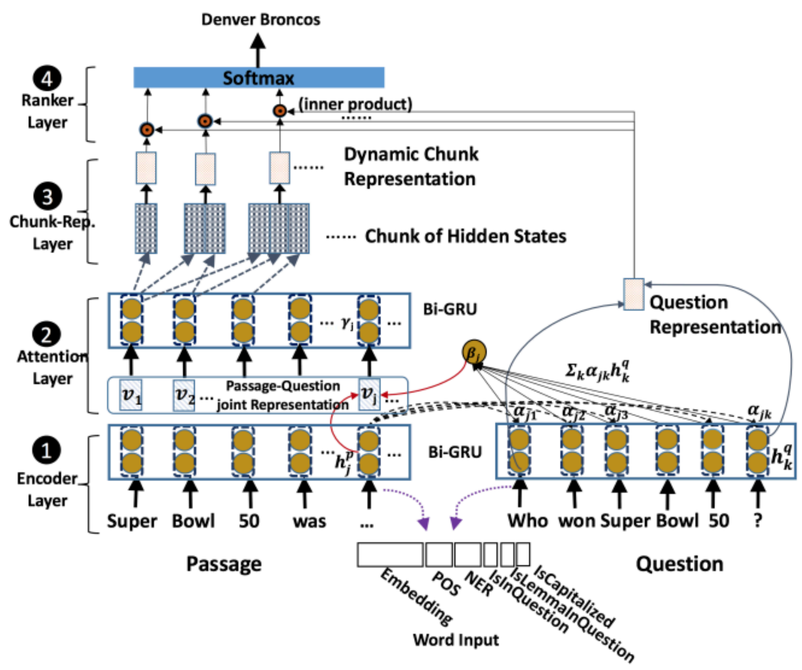



Figure 1: The main components in dynamic chunk reader model (from bottom to top) are bi-GRU encoders for passage and question, a word-by-word attention bi-GRU for passage, dynamic chunk representations that are transformed from pooled dynamic chunks of hidden states, the question attention on every chunk representation and final answer chunk prediction.

### Experiments

**Dataset** We used the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016) for the experiment. SQuAD came into our sight because it is a mix of factoid and non-factoid questions, a real-world data (crowd-sourced), and of large scale (over 100K question-answer pairs collected from 536 Wikipedia articles). Answers range from single words to long, variable-length phrase/clauses. It is a relaxation of assumptions by the cloze-style and quiz-style RC datasets in the Problem Definition section.

**Features** The input vector representation of each word $w$ to encoder RNNs has six parts including a pre-trained 300-dimensional GloVe embedding (Pennington, Socher, and Manning 2014) and five features (see Figure 1): (1) a one-hot encoding (46 dimensions) for the part-of-speech (POS) tag of $w$; (2) a one-hot encoding (14 dimensions) for named entity (NE) tag of $w$; (3) a binary value indicating whether $w$'s surface form is the same to any word in the quesiton; (4) if the lemma form of $w$ is the same to any word in the question; and (5) if $w$ is caplitalized. Feature (3) and (4) are designed to help the model align the passage text with question. Note that some types of questions (e.g., "who", "when" questions) have answers that have a specific POS/NE tag pattern. For instance, "who" questions mostly have proper nouns/persons as answers and "when" questions may frequently have numbers/dates (e.g., a year) as answers. Thus, we believe that the model could exploit the co-relation between question types and answer POS/NE patterns easier with POS and NE tag features.

**Implementation Details** We pre-processed the SQuAD dataset using Stanford CoreNLP tool[5] (Manning et al. 2014) with its default setting to tokenize the text and obtain the POS and NE annotations. To train our model, we used stochastic gradient descent with the ADAM optimizer (Kingma and Ba 2014), with an initial learning rate of 0.001. All GRU weights were initialized from a uniform distribution between (-0.01, 0.01). The hidden state size, $d$, was set to 300 for all GRUs. The question bi-GRU shared parameters with the passage bi-GRU, while the attention-based passage bi-GRU had its own parameters. We shuffled all training examples at the beginning of each epoch and adopted a

[5] stanfordnlp.github.io/CoreNLP/

# Important Intrinsic Tasks

## Behind the Scenes of a QA System

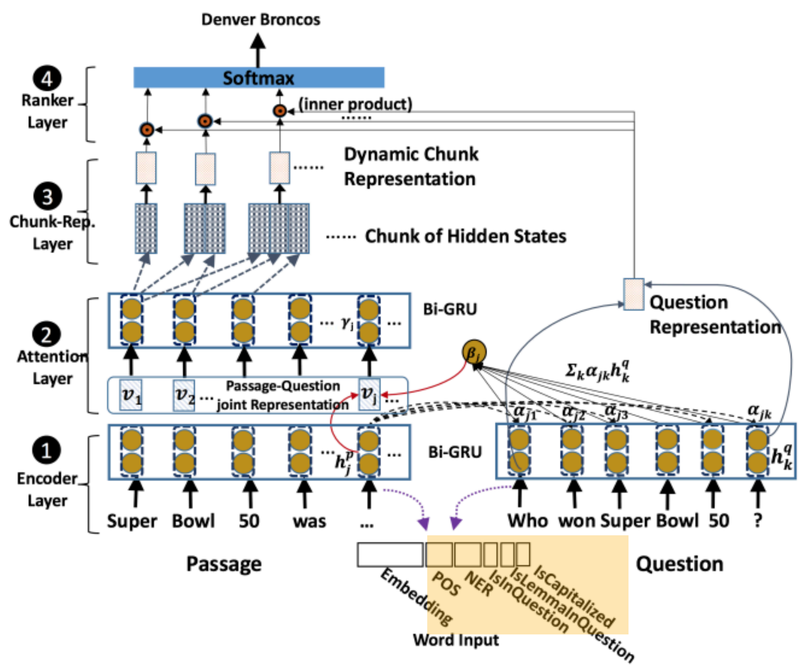



Figure 1: The main components in dynamic chunk reader model (from bottom to top) are bi-GRU encoders for passage and question, a word-by-word attention bi-GRU for passage, dynamic chunk representations that are transformed from pooled dynamic chunks of hidden states, the question attention on every chunk representation and final answer chunk prediction.

### Experiments

**Dataset** We used the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016) for the experiment. SQuAD came into our sight because it is a mix of factoid and non-factoid questions, a real-world data (crowd-sourced), and of large scale (over 100K question-answer pairs collected from 536 Wikipedia articles). Answers range from single words to long, variable-length phrase/clauses. It is a relaxation of assumptions by the cloze-style and quiz-style RC datasets in the Problem Definition section.

**Features** The input vector representation of each word $w$ to encoder RNNs has six parts including a pre-trained 300-dimensional GloVe embedding (Pennington, Socher, and Manning 2014) and five features (see Figure 1): (1) a one-hot encoding (46 dimensions) for the part-of-speech (POS) tag of $w$; (2) a one-hot encoding (14 dimensions) for named entity (NE) tag of $w$; (3) a binary value indicating whether $w$'s surface form is the same to any word in the quesiton; (4) if the lemma form of $w$ is the same to any word in the question; and (5) if $w$ is caplitalized. Feature (3) and (4) are designed to help the model align the passage text with question. Note that some types of questions (e.g., "who", "when" questions) have answers that have a specific POS/NE tag pattern. For instance, "who" questions mostly have proper nouns/persons as answers and "when" questions may frequently have numbers/dates (e.g., a year) as answers. Thus, we believe that the model could exploit the co-relation between question types and answer POS/NE patterns easier with POS and NE tag features.

**Implementation Details** We pre-processed the SQuAD dataset using Stanford CoreNLP tool[5] (Manning et al. 2014) with its default setting to tokenize the text and obtain the POS and NE annotations. To train our model, we used stochastic gradient descent with the ADAM optimizer (Kingma and Ba 2014), with an initial learning rate of 0.001. All GRU weights were initialized from a uniform distribution between (-0.01, 0.01). The hidden state size, $d$, was set to 300 for all GRUs. The question bi-GRU shared parameters with the passage bi-GRU, while the attention-based passage bi-GRU had its own parameters. We shuffled all training examples at the beginning of each epoch and adopted a

[5] stanfordnlp.github.io/CoreNLP/

# Important Intrinsic Tasks

## Behind the Scenes of a QA System

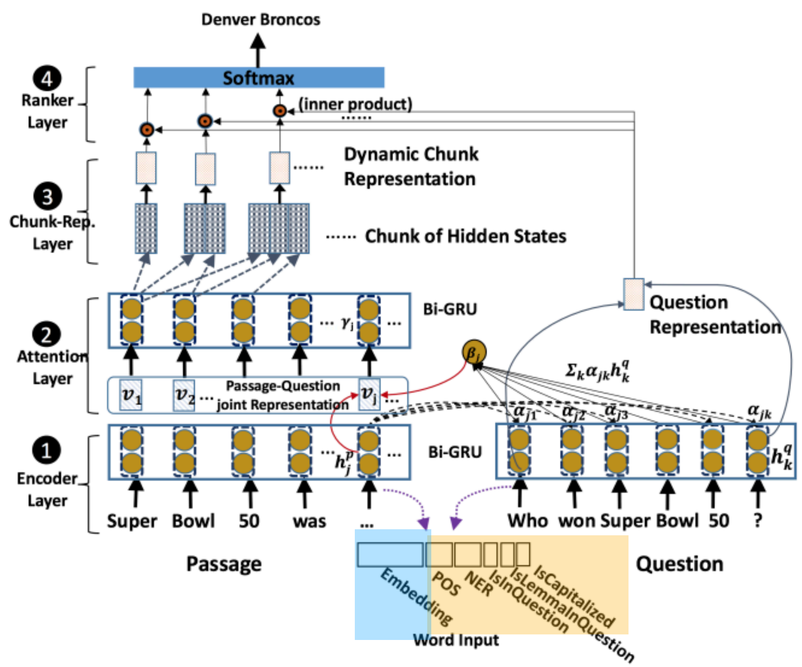



Figure 1: The main components in dynamic chunk reader model (from bottom to top) are bi-GRU encoders for passage and question, a word-by-word attention bi-GRU for passage, dynamic chunk representations that are transformed from pooled dynamic chunks of hidden states, the question attention on every chunk representation and final answer chunk prediction.

### Experiments

**Dataset** We used the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016) for the experiment. SQuAD came into our sight because it is a mix of factoid and non-factoid questions, a real-world data (crowd-sourced), and of large scale (over 100K question-answer pairs collected from 536 Wikipedia articles). Answers range from single words to long, variable-length phrase/clauses. It is a relaxation of assumptions by the cloze-style and quiz-style RC datasets in the Problem Definition section.

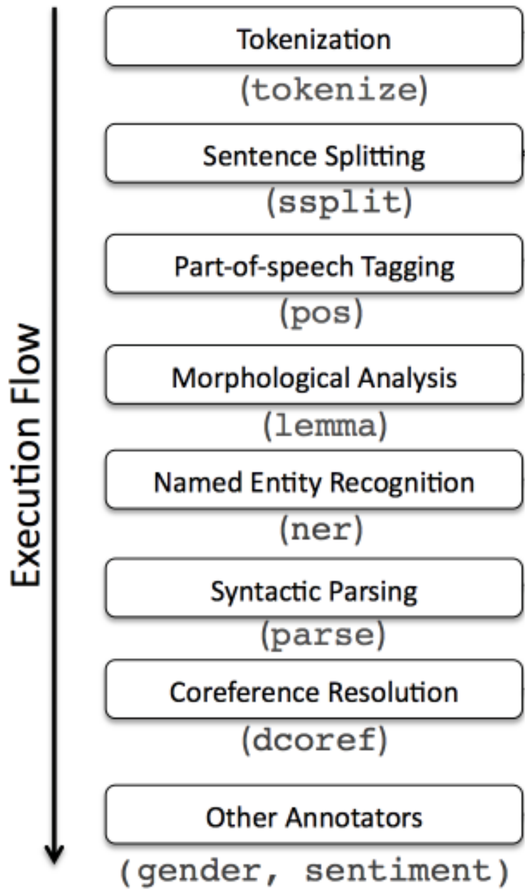
**Features** The input vector representation of each word $w$ to encoder RNNs has six parts including a pre-trained 300-dimensional GloVe embedding (Pennington, Socher, and Manning 2014) and five features (see Figure 1): (1) a one-hot encoding (46 dimensions) for the part-of-speech (POS) tag of $w$; (2) a one-hot encoding (14 dimensions) for named entity (NE) tag of $w$; (3) a binary value indicating whether $w$'s surface form is the same to any word in the quesiton; (4) if the lemma form of $w$ is the same to any word in the question; and (5) if $w$ is caplitalized. Feature (3) and (4) are designed to help the model align the passage text with question. Note that some types of questions (e.g., "who", "when" questions) have answers that have a specific POS/NE tag pattern. For instance, "who" questions mostly have proper nouns/persons as answers and "when" questions may frequently have numbers/dates (e.g., a year) as answers. Thus, we believe that the model could exploit the co-relation between question types and answer POS/NE patterns easier with POS and NE tag features.

**Implementation Details** We pre-processed the SQuAD dataset using Stanford CoreNLP tool[5] (Manning et al. 2014) with its default setting to tokenize the text and obtain the POS and NE annotations. To train our model, we used stochastic gradient descent with the ADAM optimizer (Kingma and Ba 2014), with an initial learning rate of 0.001. All GRU weights were initialized from a uniform distribution between (-0.01, 0.01). The hidden state size, $d$, was set to 300 for all GRUs. The question bi-GRU shared parameters with the passage bi-GRU, while the attention-based passage bi-GRU had its own parameters. We shuffled all training examples at the beginning of each epoch and adopted a

[5] stanfordnlp.github.io/CoreNLP/

# Important Intrinsic Tasks

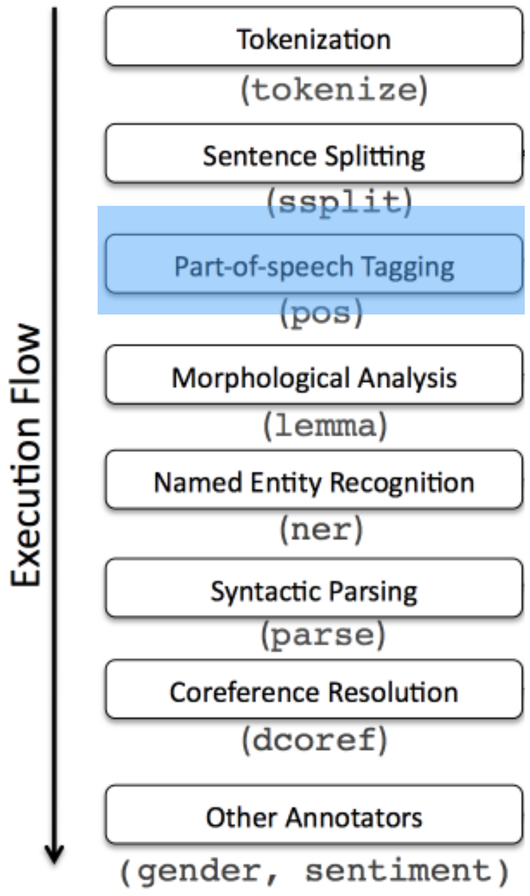
## The “Core NLP” Tasks




The important thing about
Disney is that it is a global brand.

# Important Intrinsic Tasks

## The “Core NLP” Tasks

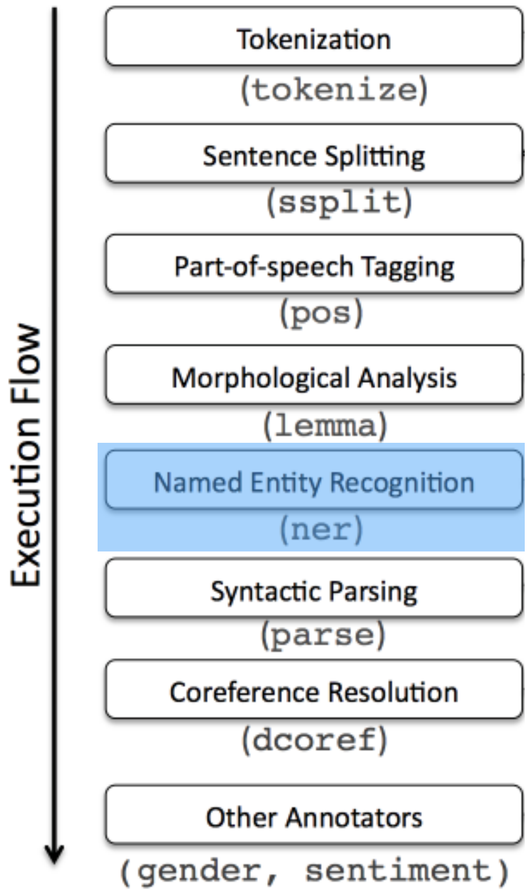



The important thing about Disney is that it is a global brand.

| | | |
|---|---|---|
| Noun | Verb | Noun |

# Important Intrinsic Tasks

## The “Core NLP” Tasks

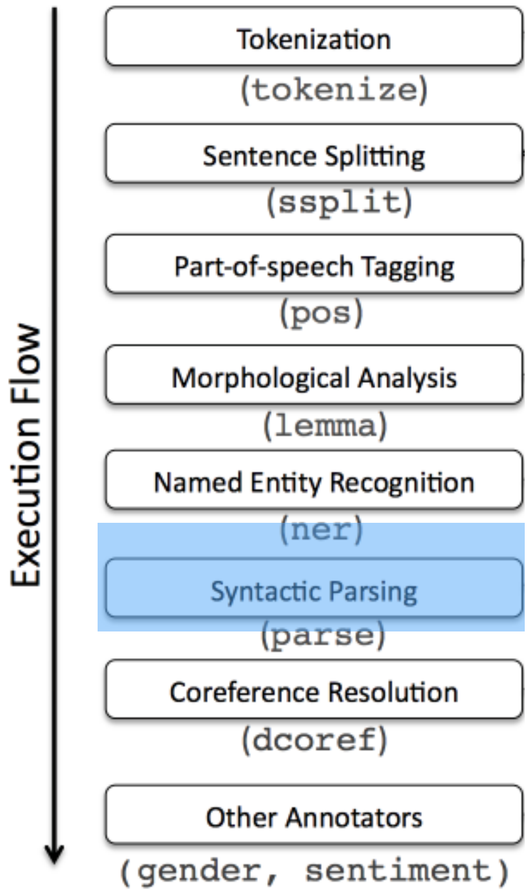


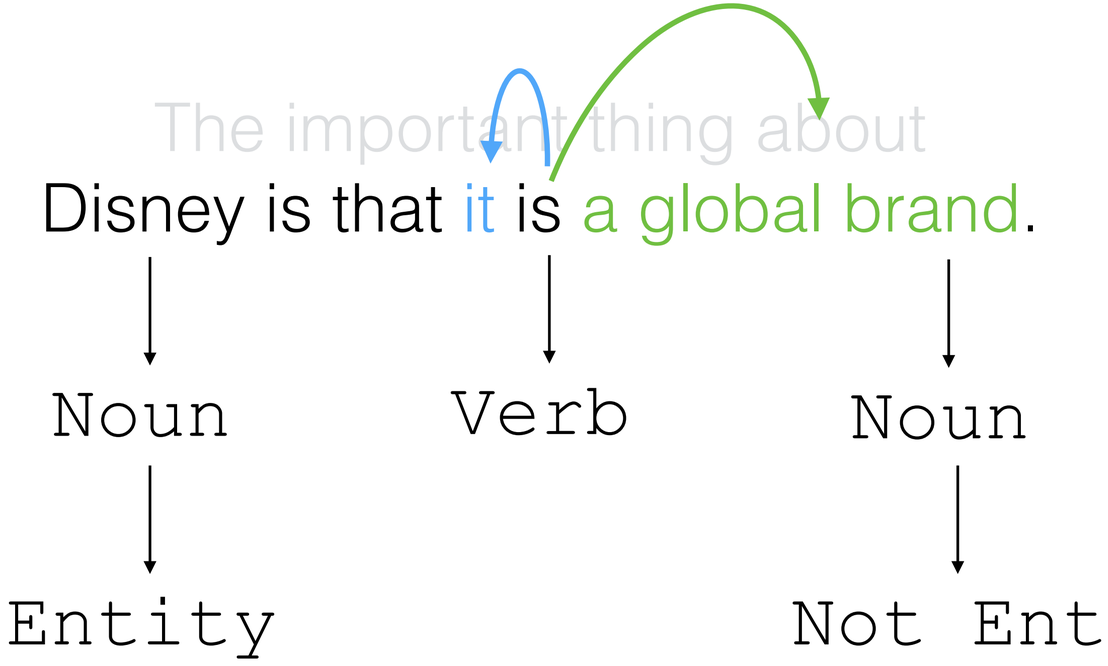

The important thing about
Disney is that it is a global brand.

| | | |
|---|---|---|
| Noun | Verb | Noun |
| Entity | | Not Ent |

# Important Intrinsic Tasks

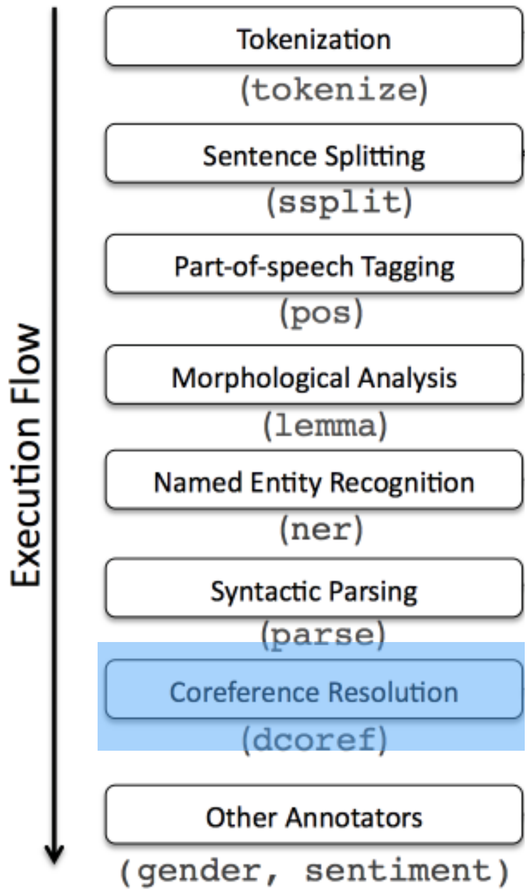
## The “Core NLP” Tasks

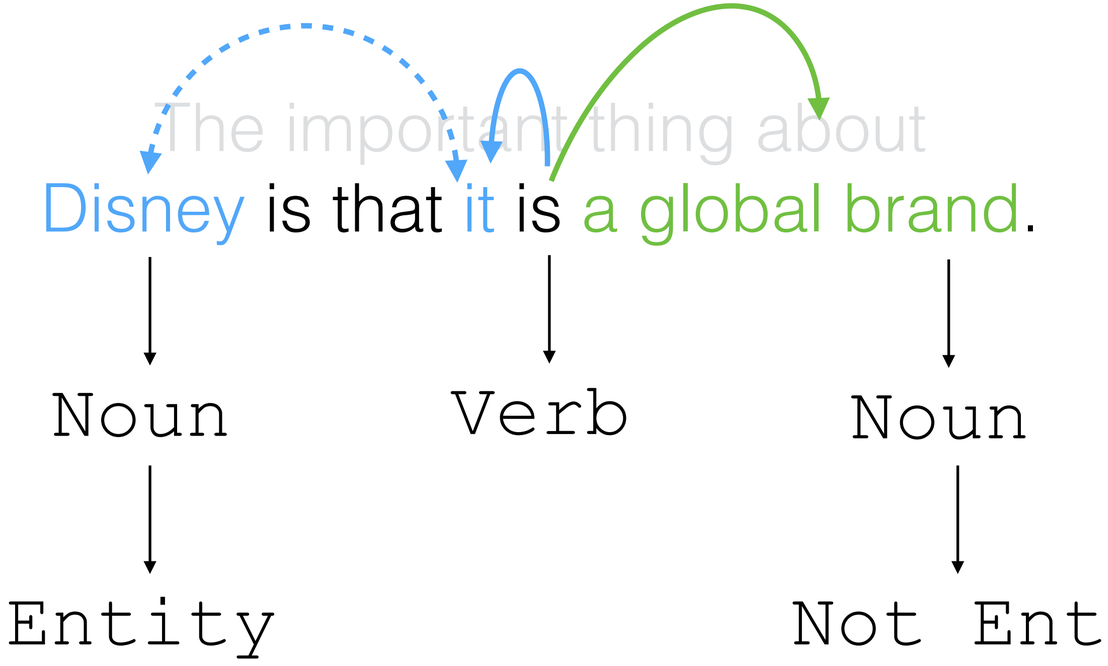


# Important Intrinsic Tasks

## The “Core NLP” Tasks



# Important Intrinsic Tasks

## The “Core NLP” Tasks

- Lexical Semantics
  - Word Sense Disambiguation: bank vs. bank
  - Word Similarity: cat ≈ kitten
  - Hyponymy Detection: cat Is-A animal
- Sentence-Level Semantics
  - Semantic Role Labeling: “Jo gave Sam a book” = “Jo gave a book to Sam” -> Jo=giver, Sam=receiver, gift=book
  - Natural Language Inference (NLI): “Jo gave Sam a book” -> “Sam has the book”

# Topics

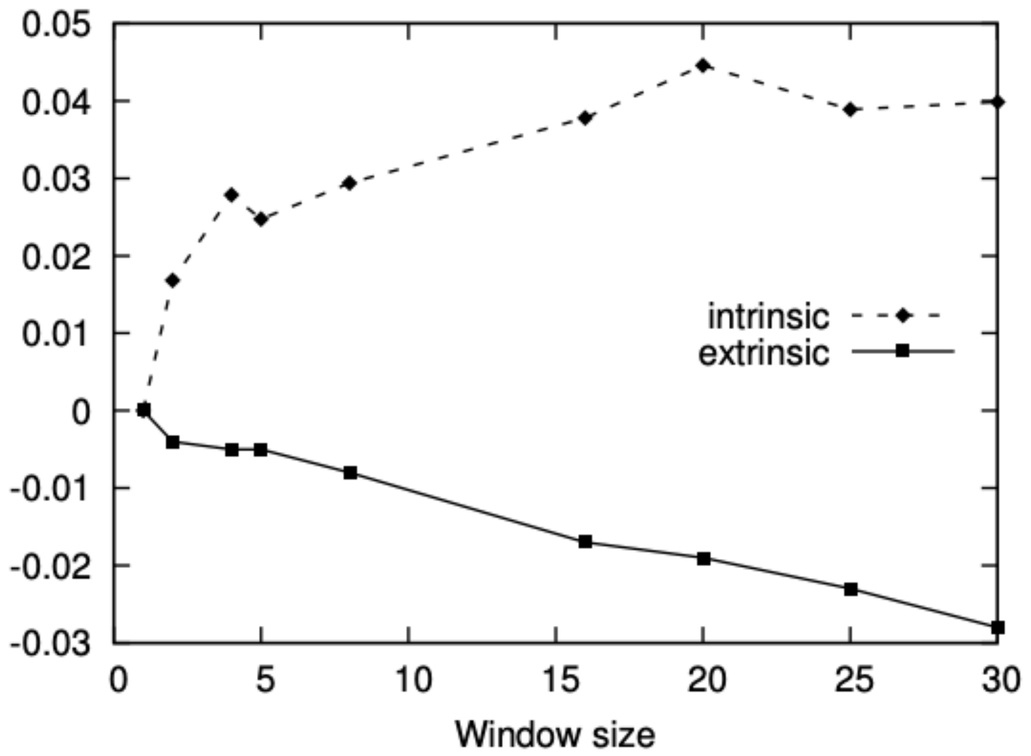
- Task-Driven Culture
- Major “Downstream” Tasks and Evaluations
- Some Important Intrinsic Tasks and Evaluations
- **Current Debates and Open Questions**

# Evaluations: Current Debates

## Extrinsic vs. Intrinsic?

**Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance**

**Billy Chiu** **Anna Korhonen** **Sampo Pyysalo**
Language Technology Lab
DTAL, University of Cambridge
{hwc25|alk23}@cam.ac.uk, sampo@pyysalo.net



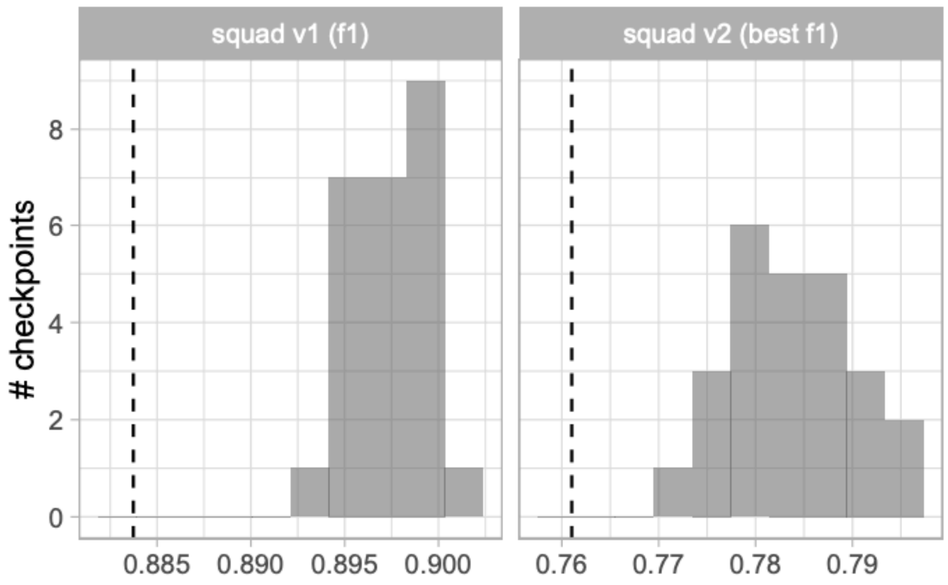
# Evaluations: Current Debates

## Scientific Validity?

THE MULTIBERTS: BERT REPRODUCTIONS FOR ROBUSTNESS ANALYSIS

**Thibault Sellam,* Steve Yadlowsky,* Ian Tenney,* Jason Wei,† Naomi Saphra,‡ Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick**

{tsellam, yadlowsky, iftenney, epavlick}@google.com
Google Research



# Evaluations: Current Debate

## Leaderboardism

**Utility is in the Eye of the User: A Critique of NLP Leaderboards**

**Kawin Ethayarajh**
Stanford University
kawin@stanford.edu

**Dan Jurafsky**
Stanford University
jurafsky@stanford.edu

**Abstract**

Benchmarks such as GLUE have helped drive advances in NLP by incentivizing the creation of more accurate models. While this leaderboard paradigm has been remarkably successful, a historical focus on performance-based evaluation has been at the expense of other qualities that the NLP community values in models, such as compactness, fairness, and energy efficiency. In this opinion paper, we study the divergence between what is incentivized by leaderboards and what is useful in practice through the lens of microeconomic theory. We frame both the leaderboard and NLP practitioners as *consumers* and the benefit they get from a model as its *utility* to them. With this framing, we formalize how leaderboards – in their current form – can be poor proxies for the NLP community at large. For example, a highly inefficient model would provide less utility to practitioners but not to a leaderboard, since it is a cost that only the former must bear. To allow practitioners to better estimate a model's utility to them, we advocate for more transparency on leaderboards, such as the reporting of statistics that are of practical concern (e.g., model size, energy efficiency, and inference latency).

# Evalu
## Leader

**Climbing towards NLU:**
**On Meaning, Form, and Understanding in the Age of Data**

**Emily M. Bender**
University of Washington
Department of Linguistics
ebender@uw.edu

**Alexander Koller**
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

**Utility is in the Eye of the User: A Critique of NLP Leaderboards**

**Kawin Ethayarajh**
Stanford University
kawin@stanford.edu

**Dan Jurafsky**
Stanford University
jurafsky@stanford.edu

**What Will it Take to Fix Benchmarking in Natural Language Understanding?**

**Samuel R. Bowman**
New York University
bowman@nyu.edu

**George E. Dahl**
Google Research, Brain Team
gdahl@google.com

**Abstract**

enchmarks such as GLUE have helped drive
lvances in NLP by incentivizing the creation
f more accurate models. While this leader-
oard paradigm has been remarkably success-
l, a historical focus on performance-based
valuation has been at the expense of other
ualities that the NLP community values in
odels, such as compactness, fairness, and en-
ergy efficiency. In this opinion paper, we study
the divergence between what is incentivized
by leaderboards and what is useful in prac-
tice through the lens of microeconomic the-
ory. We frame both the leaderboard and NLP
practitioners as *consumers* and the benefit they
get from a model as its *utility* to them. With
we formalize how leaderboards
rent form – can be poor proxies
community at large. For exam-
inefficient model would provide
practitioners but not to a leader-
it is a cost that only the former
allow practitioners to better es-
el's utility to them, we advocate
sparency on leaderboards, such
ng of statistics that are of practi-
cal concern (e.g., model size, energy efficiency,
and inference latency).

All done! See you Thursday!