COMP4901K/Math4824B
Machine Learning for Natural Language Processing

Lecture 1: Introduction

Instructor: Yangqiu Song

Slides credits: Hongning Wang, Marti Hearst, Percy Liang, Noah Smith

Logistics

- Instructor: Yangqiu Song
- Email: yqsong@cse.ust.hk
- Canvas (https://canvas.ust.hk)
  - Lecture notes
  - Assigments
  - Projects

– TAs
  - Tianqing Fang: tfangaa@connect.ust.hk
  - Zizheng Lin: zlinai@connect.ust.hk

Background of this Course

- Purpose of this course
  - Currently a topic course for senior undergraduate students
    - Enrich their experience on text data analytics
    - Equip them with powerful analytic tools for future career
  - Target to be a future elective course for the Data Science and Technology (DSCT) program
    - http://dsct.ust.hk/
- Shared Course Codes COMP4901K/Math4824B
  - Math students can replace final project with a survey paper

More about this Course

- The course covers the knowledge from both CSE and Math areas:
  - The fundamental machine learning models that deal with real natural language processing problems.
  - The programming and analytic tools such as Python to deal with real natural language processing problems.
    - Many example code will be provided
  - The problem solving skill for real natural language processing problems, such as building end to end system to deal with a certain problem.
- So
  - It's not a pure machine learning course
  - It's not a pure NLP course
  - We can cover more modern and advanced learning algorithms for NLP tasks
  - Will be under "*Restricted Deep Learning Electives*" in CSE programs
    - Among the courses needed to satisfy our Area requirement (for a total of min. 15 credits), only one course among these "Restricted DL Electives" can be used to count as one of those 5 courses.
- Prerequisites
  - Linear Algebra (MATH 2111 or MATH 2121)
  - Programming (COMP 2011/2012)
  - Discrete Math (COMP 2711 or Math 2343)

More about this Course

- It's a data science course
- We will have a lot of practicing opportunities to make sense of using text data
  - To understand text data problems arising in the areas of commerce and industry etc.
  - To model text data problems using different mathematical tools.
  - To design and implement efficient algorithms to solve different NLP problems.
  - To interpret the results provided by different algorithms and apply them to the data problems to gain meaningful insights or offer predictions.

Goals of this Course

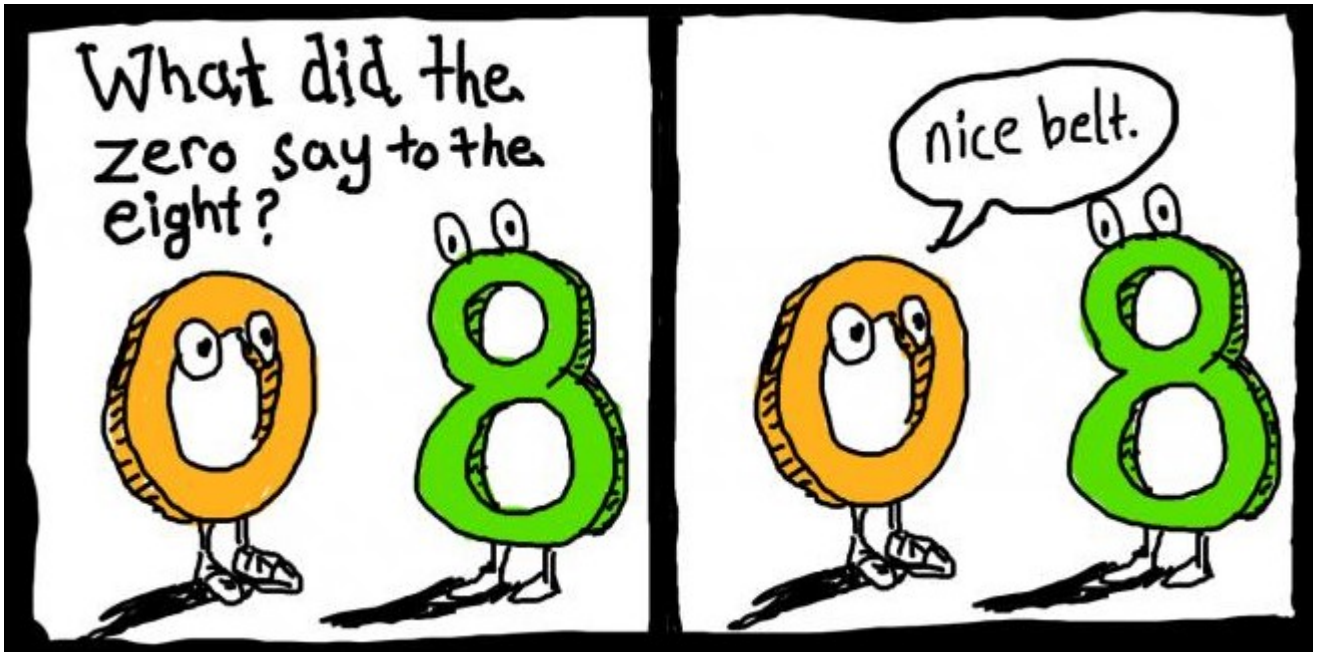
- Learn about the problems and possibilities of natural language analysis:
  - What are the major issues?
  - What are the major solutions?
    - How well do they work?
    - How do they work?
- At the end you should:
  - Agree that language is subtle and interesting!
  - Feel some ownership over the algorithms
  - Be able to assess NLP problems
    - Know which solutions to apply when, and how
  - Be able to read papers in the field

Today

- Why NLP is difficult
- How to solve it (in general)?
- What we'll do in this course

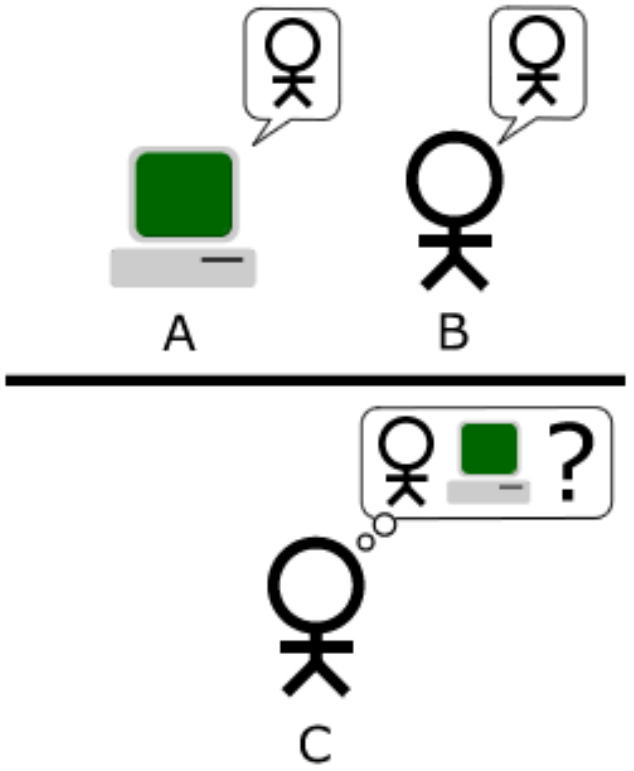
Natural Language

- Understanding language is a very complex thing
- But something that humans are amazingly good at



Artificial Intelligence: Turing Test

- Replacement of “Can machines think?”
  - Behavioral test
  - not just natural language understanding
- Natural language conversation
  - *Imitation game (1950)*
    - Human?
    - Computer?

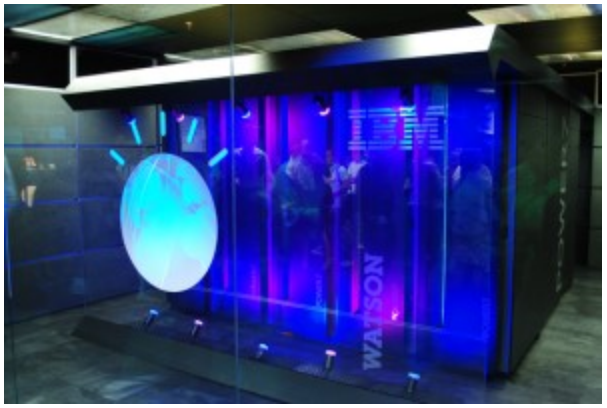


The AI Winter

- AI winter: 1974–80 and 1987–93
  - 1966: the failure of machine translation,
  - 1970: the abandonment of connectionism,
  - 1971–75: DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University,
  - 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill report,
  - 1973–74: DARPA's cutbacks to academic AI research in general,
  - 1987: the collapse of the Lisp machine market,
  - 1988: the cancellation of new spending on AI by the Strategic Computing Initiative,
  - 1993: expert systems slowly reaching the bottom, and
  - 1990s: the quiet disappearance of the fifth-generation computer project's original goals.

Enabled by Big Data

- 1971–75: DARPA's frustration with the Speech Understanding
- "Watson is a question answering (QA) computing system that IBM built to apply advanced
  - natural language processing,
  - information retrieval,
  - knowledge representation,
  - automated reasoning, and
  - machine learning technologies



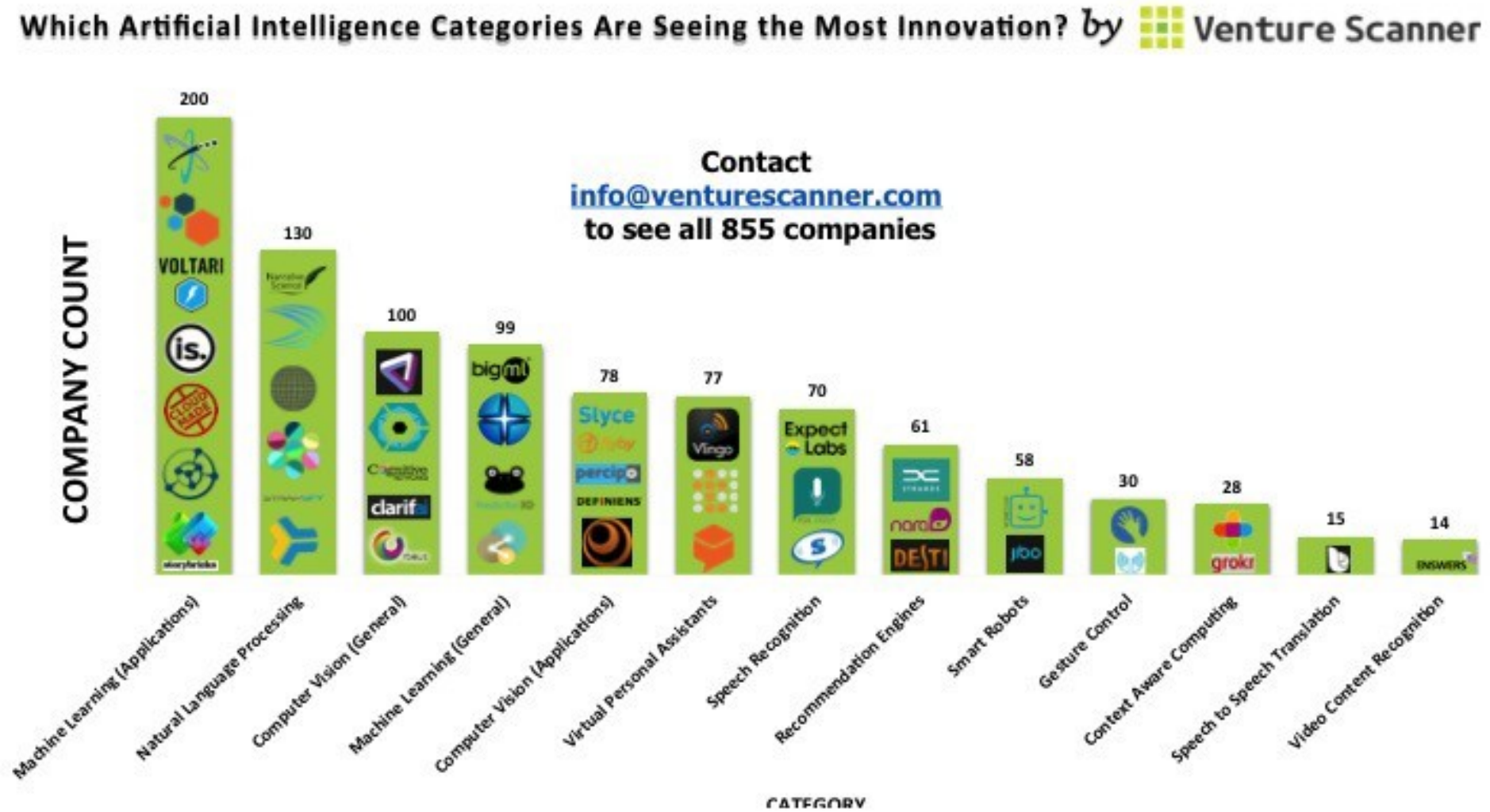
- to the field of open domain question answering."

In 2011, Watson competed on Jeopardy! against former winners Brad Rutter and Ken Jennings. Watson received the first place prize of $1 million.

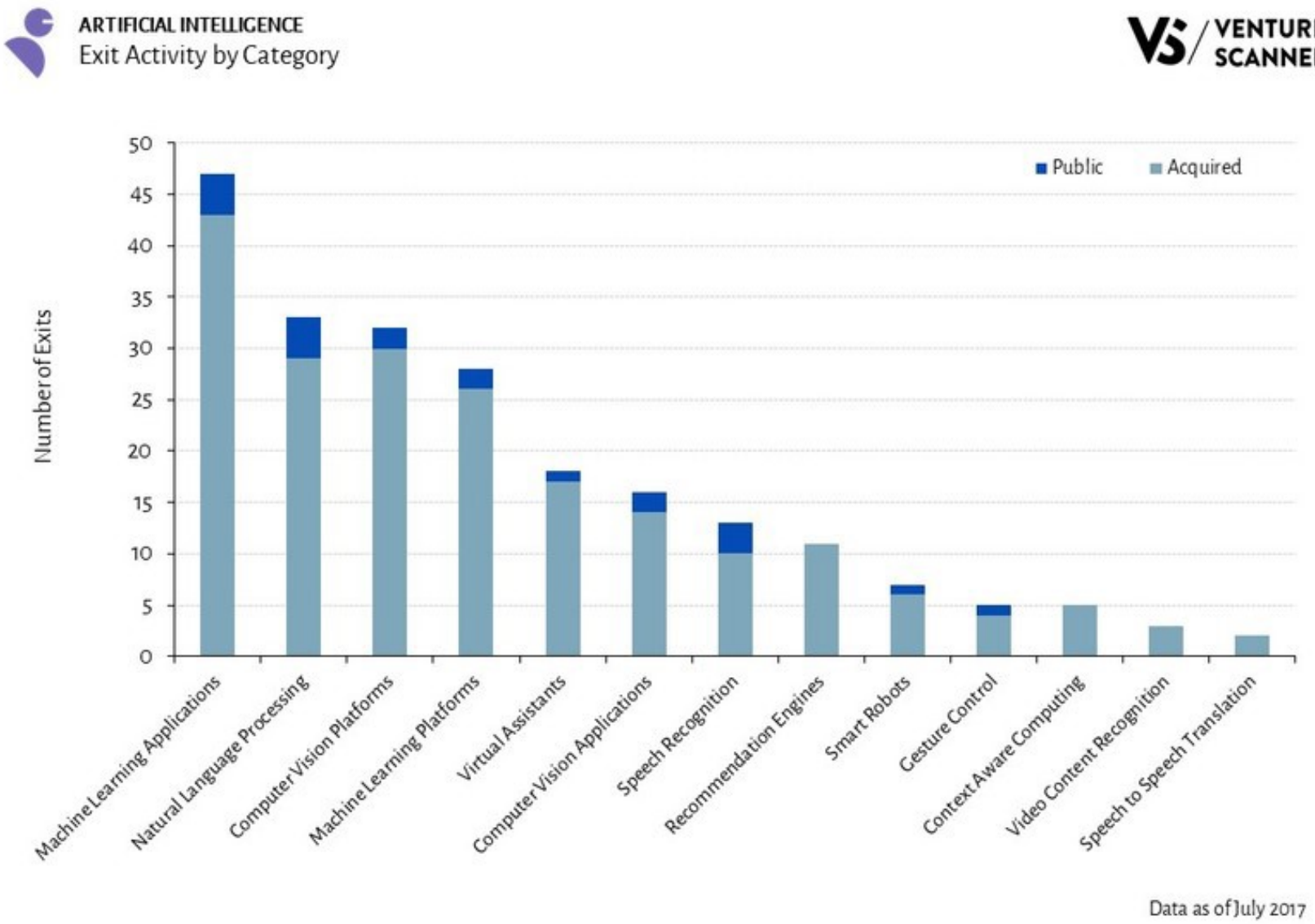
https://en.wikipedia.org/wiki/Watson_(computer)

Enabled by Big Data

Startup Companies (2015)



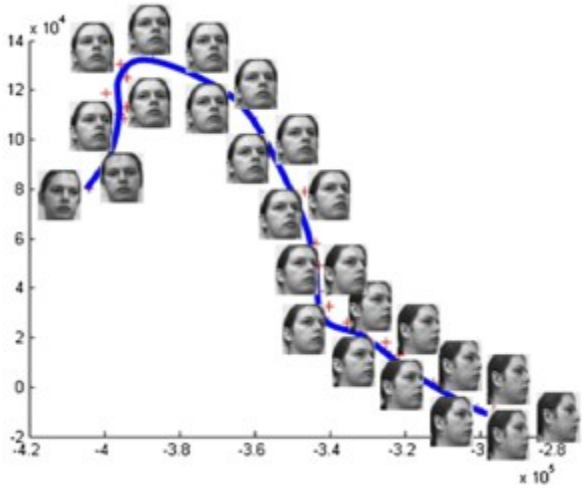
Number of Exits (Acquisitions and IPOs, 2017)



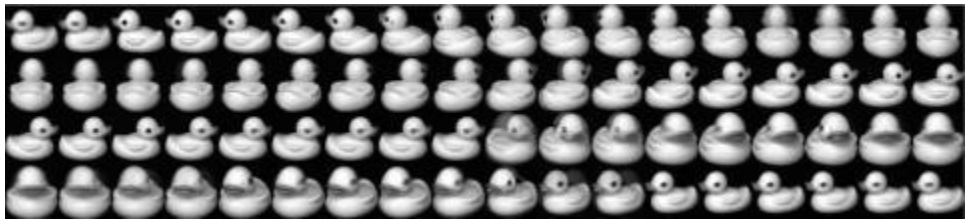
What's Special about Human Language?

- A human language is a discrete/symbolic/categorical signaling system
  - With very minor exceptions for expressive signaling ("I loooove it." "Whoomppaaa")
- Large vocabulary, symbolic encoding of words creates a problem for machine learning – sparsity!

Face

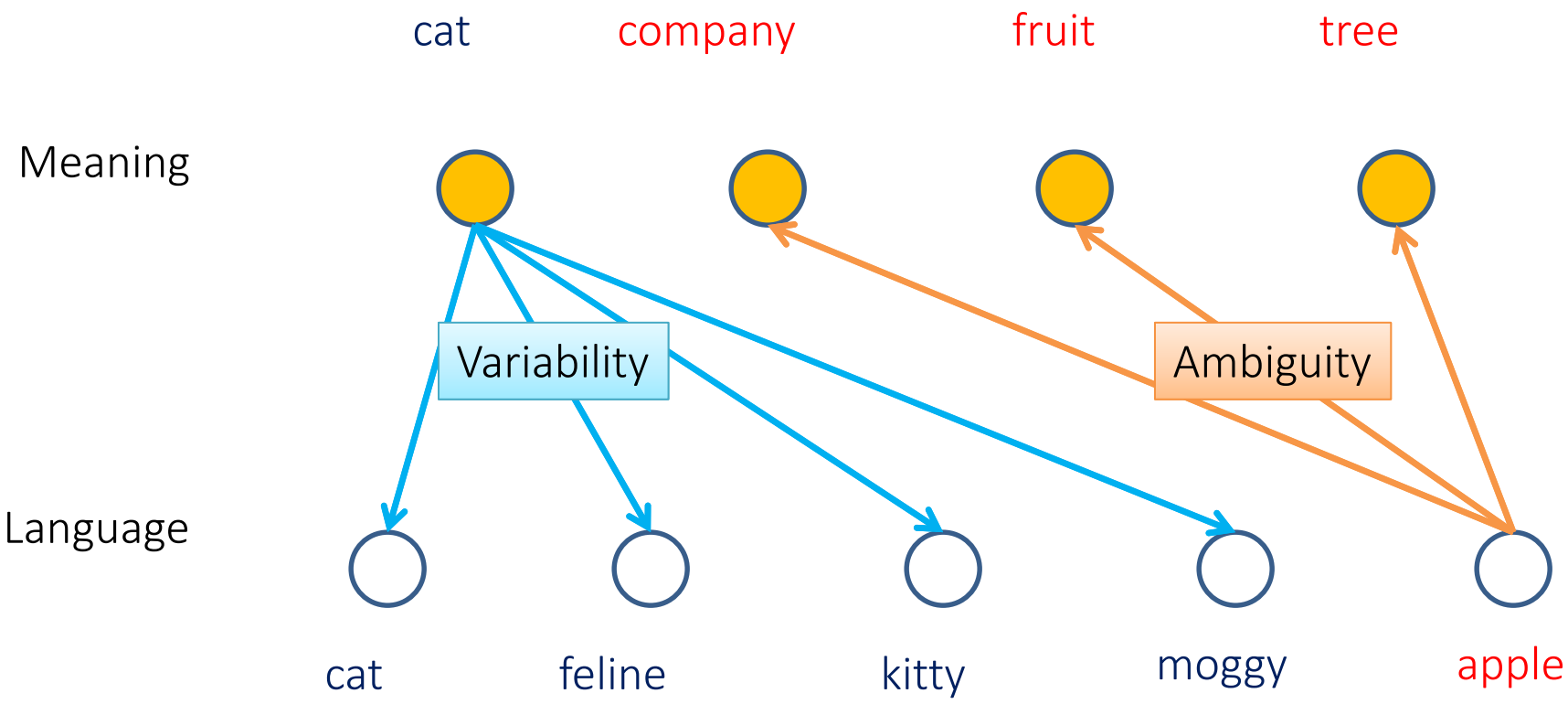


Toy



Why is NLP Difficult?

- Variability and ambiguity everywhere

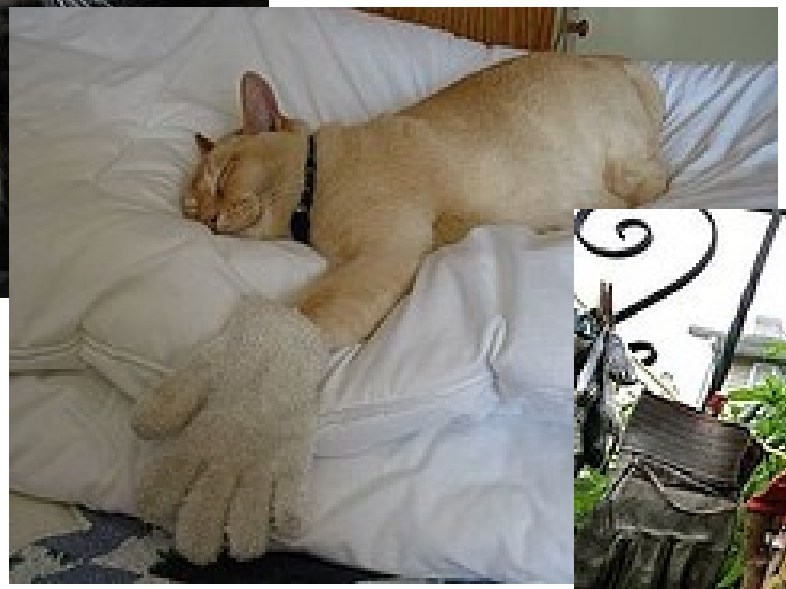


Words are Ambiguous (have multiple meanings)

- I know that.
- I know that block.
- I know that blocks the sun.
- I know that block blocks the sun.

More Examples of Ambiguity

- Get the cat with the gloves.

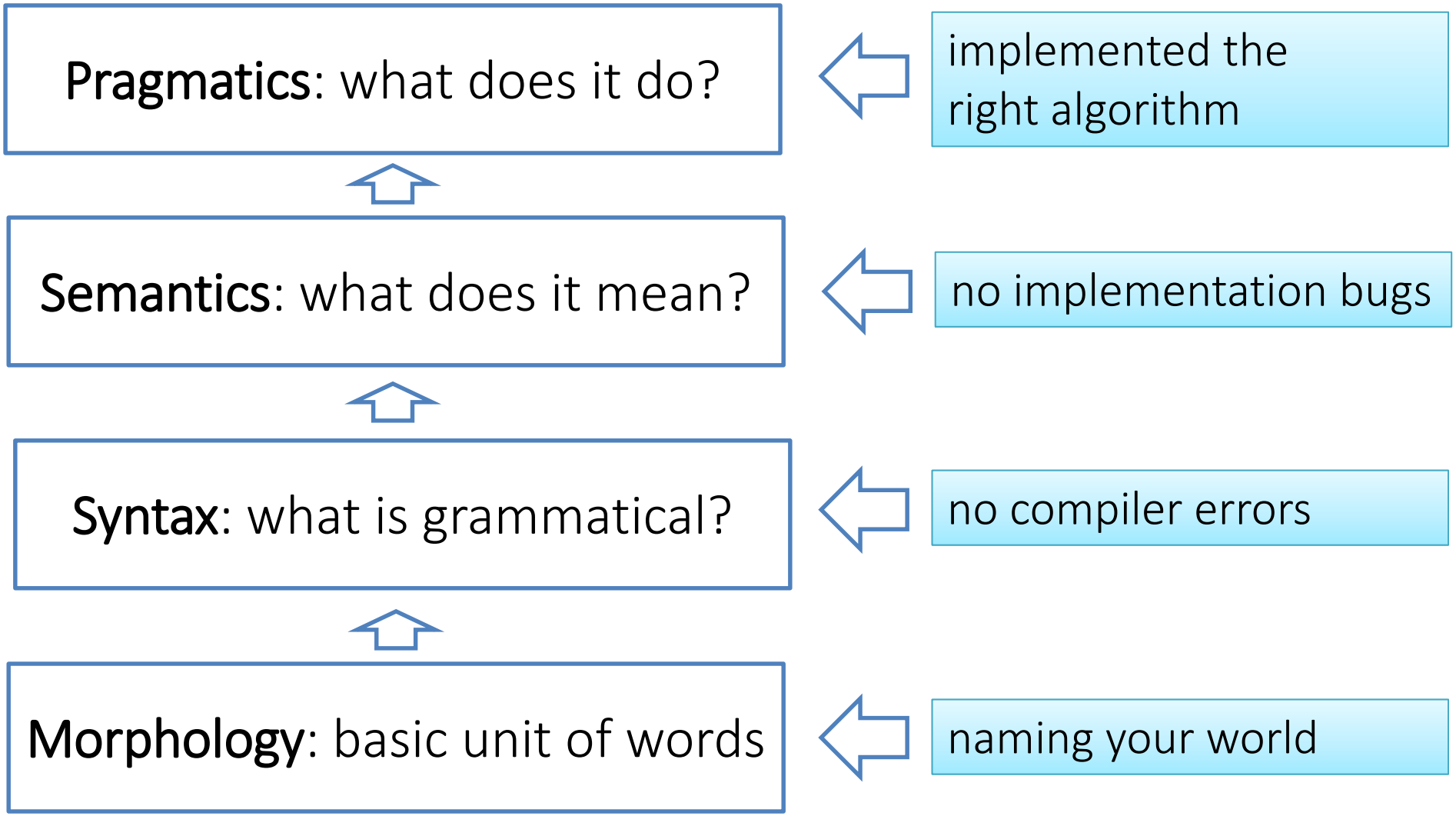


Language Subtleties

- Adjective order and placement
  - A big black dog
  - A big black scary dog
  - A big scary dog
  - A scary big dog
  - ✘ A black big dog

- Antonyms
  - Which sizes go together?
    - Big and little
    - Big and small
    - Large and small
    - ✘ Large and little

Levels of Linguistic Analysis: Analogy with Programming Languages



Analogy with Programming Languages

- Syntax: no compiler errors
- Semantics: no implementation bugs
- Pragmatics: implemented the right algorithm

- Different syntax, same semantics (5):
  - 2 + 3 <-> 3 + 2

- Good semantics, bad pragmatics:
  - correct implementation of deep neural network
  - for estimating coin flip prob.

How to do natural language processing?

The Role of Memorization

- Children learn words quickly
  - As many as 9 words/day
  - Often only need one exposure to associate meaning with word
    - Can make mistakes, e.g., overgeneralization

      "I goed to the store."
  - Exactly how they do this is still under study

The Role of Memorization

- Dogs can do word association too!
  - Rico, a border collie in Germany
  - Knows the names of each of 100 toys
  - Can retrieve items called out to him with over 90% accuracy.
  - Can also learn and remember the names of unfamiliar toys after just one encounter, putting him on a par with a three-year-old child.

https://en.wikipedia.org/wiki/Rico_(dog)

But there is too much to memorize!

establish

establishment

> the church of England as the official state church.

disestablishment

antidisestablishment

antidisestablishmentarian

antidisestablishmentarianism

> is a political philosophy that is opposed to the separation of church and state.

Rules and Memorization

- Current thinking in psycholinguistics is that we use a combination of rules and memorization
  - However, this is very controversial

- Mechanism:
  - If there is an applicable rule, apply it
  - However, if there is a memorized version, that takes precedence. (Important for irregular words.)
    - Artists paint “still lifes”
      - Not “still lives”
    - Past tense of
      - think → thought
      - blink → blinked

Representation of Meaning

- I know that block blocks the sun.
  - How do we represent the meanings of “block”?
  - How do we represent “I know”?
  - How does that differ from “I know that.”?
  - Who is “I”?
  - How do we indicate that we are talking about earth’s sun vs. some other planet’s sun?
  - When did this take place? What if I move the block? What if I move my viewpoint? How do we represent this?

How to tackle these problems?

- The field was stuck for quite some time.
- A new approach started around 1990
  - Well, not really new, but the first time around, in the 50's, they didn't have the text, disk space, or GHz
- Main idea: combine memorizing and rules
- How to do it:
  - Get large text collections (corpora)
  - Compute statistics over the words in those collections
- Surprisingly effective
  - Even better now with the Web

NLP ?= Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Computer representation of language is not directly observable.
- Early connections to information theory (1940s)
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

NLP ?= Linguistics

- NLP must contend with NL data as found in the world
- NLP ≈ computational linguistics
- Linguistics has begun to use tools originating in NLP!

Desiderata for NLP Methods

(ordered arbitrarily)

- Sensitivity to a wide range of the phenomena and constraints in human language
- Generality across different languages, genres, styles, and modalities
- Computational efficiency at construction time and runtime
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency, robustness, etc.)
- High accuracy when judged against expert annotations and/or task-specific performance

Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Theory of computation
- Data science
- Social and political science
- Psychology
- Economics
- Education

Fields with Connections to Machine Learning

- NLP
- Data mining/Data science
- Bioinformatics
- Fintech
- Computer vision
- Multimedia analysis
- Social and political science
- Psychology
- Economics
- Education

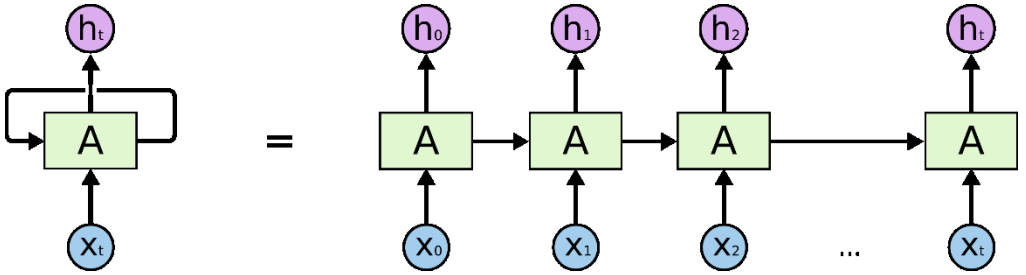
This Course: Machine learning for NLP

- mid-1970s: HMMs for speech recognition → probabilistic models
- early 2000s: conditional random fields for part-of-speech tagging → structured prediction
- early 2000s: latent Dirichlet allocation for modeling text documents → topic modeling
- mid 2010s: sequence-to-sequence models for machine translation → neural networks with memory/state

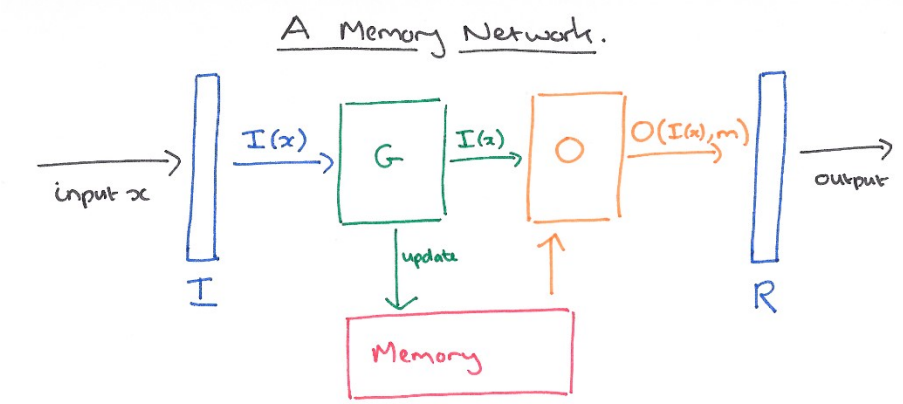
We will select some of the important topics

Nowadays: Deep learning for NLP

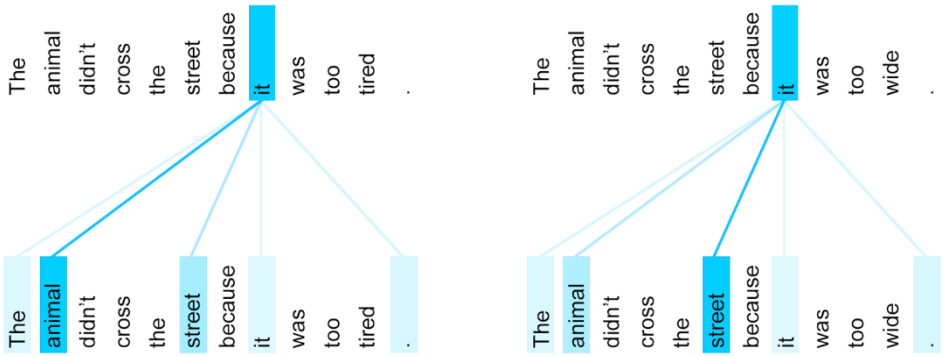
- Sequence models



- Memory models



- Attention models



Real-World Applications of NLP

- Spelling suggestions/corrections
- Grammar checking
- Information extraction
- Text categorization
- Automated customer service
  - Conversational agents
  - Question answering
- Speech recognition (limited)
- Machine translation
- Social media analysis
- Rich visual understanding
- Mining legal, medical, or scholarly literature

What We'll Do in this Course

- Learn fundamental machine learning models for NLP
  - Classification models
  - Language models
  - Sequence labeling models
  - Advance NLP tasks using deep learning
- Use NLTK (Natural Language ToolKit) and Tensorflow/PyTorch to try out various algorithms
  - Some assignment will be to do some exercises

What We'll Do in this Course

- Adopt a large text collection
- Use a wide range of NLP techniques to process it
- Release the results for others to use

How to analyze a big collection?

- Your ideas go here with a project

Labs

- No lab in the first week
- TAs will give you some illustrations to show you how to implement algorithms dealing with NLP problem
- You will be asked to submit your results after the tutorial as assigments.
- Some free computational resources to use
  - https://www.kaggle.com/docs/notebooks
  - https://www.kaggle.com/dansbecker/running-kaggle-kernels-with-a-gpu

Course Information

- Work load and grading:
  - Assignments (30%)
    - Mostly for lab, sometimes a quiz released in Canvas
    - Very simple, you will get points by trying our code and submitting results
  - Projects and Presentation (30%)
    - Project (20%): a team based project
      - For students enrolled in Math4824B, you can select to do a survey paper.
      - Proposal (5%): the proposal deadline will be around mid-term
      - Up to 3 students in each group
    - In-class Presentation (10%): present what you have done in the project/survey
  - ~~Midterm~~
  - Final exam (40%)
  - Academic integrity policy
    - Late submission: score got reduced (time based)
    - Plagiarism: all involved parties will get zero

Implementation of Bonus Points

- For each student, up to three bonus credit tickets can be earned by
  - Step 1: Answer a question in the class
  - Step 2: After the class, you need to send me an email and let me know your name and your name shown in Zoom which I can check the recorded video
- Application of the tickets
  - Add 1 out of 100 points
  - Maximally you can get 3 points to boot you with one level
    - E.g., A → A+