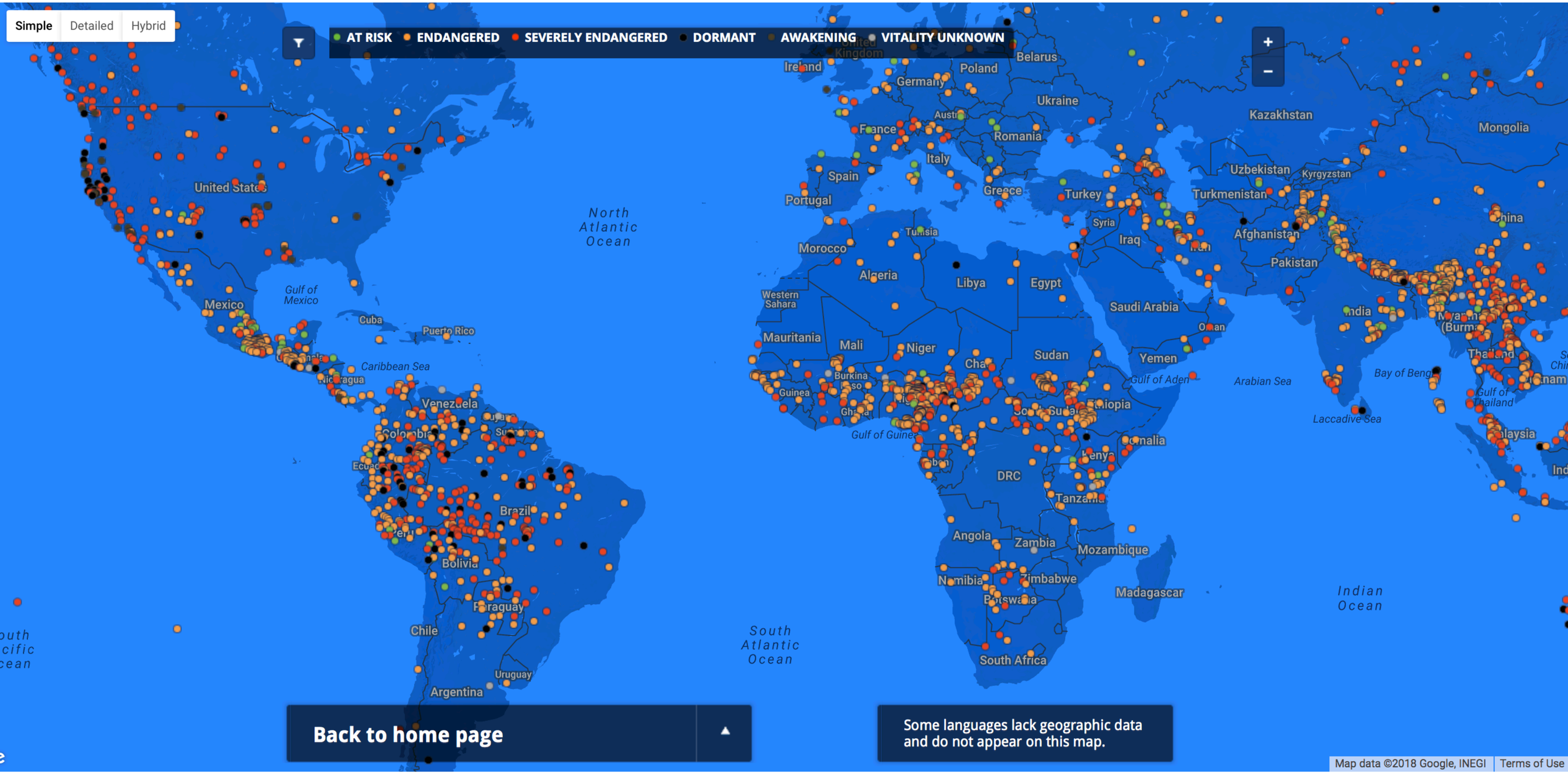
CS11-737: Multilingual Natural Language Processing

Yulia Tsvetkov, Graham Neubig, Alan Black

http://demo.clab.cs.cmu.edu/11737fa20/

Carnegie Mellon University
Language Technologies Institute



http://endangeredlanguages.com/

How do We Build NLP Systems?

- **Rule-based systems:** Work OK, but require lots of human effort for each language for where they're developed
- **Machine learning based systems:** Work really well when lots of data available, not at all in low-data scenarios

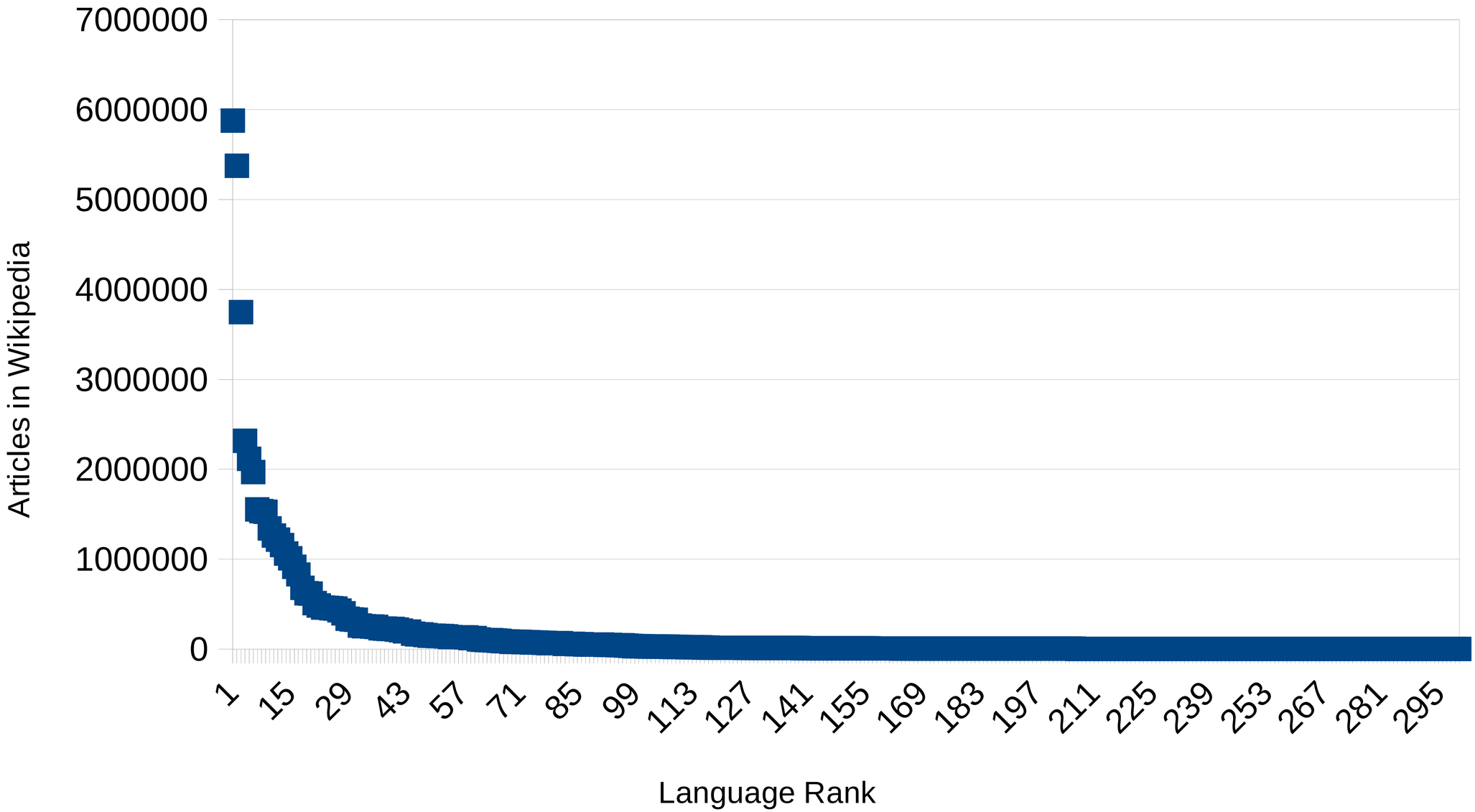
Machine Learning Models

- Formally, map an input *X* into an output *Y*. Examples:

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

- To learn, we can use
  - Paired data <*X*, *Y*>, source data *X*, target data *Y*
  - Paired/source/target data in *similar* languages

The Long Tail of Data



How to Cope?

- **Better Models or Algorithms:**
  - sophisticated modeling/training methods - know NLP/ML!
  - linguistically informed methods - know linguistics!
- **Better Data:**
  - every piece of relevant data can help - be resourceful!
  - make data if necessary - be connected!
- **Better Deployment:**
  - different situations require different solutions - be aware!

This Class Will Cover

- **Linguistics:** typology, orthography, morphology, syntax, language contact/change, code switching
- **Data:** annotated and unannotated sources, data annotation, linguistic databases, active learning
- **Tasks:** language ID, sequence labeling, translation, speech recognition/synthesis, syntactic parsing
- **Societal Considerations:** ethics, connection between language and society

All to:

Allow you to build a strong, functioning language system in a low-resource language that you do not know

Instructors/TAs

Instructors:

- Yulia Tsvetkov (office hours: Thursday 4-5PM)
- Graham Neubig (office hours: Friday 4-5PM)
- Alan Black (office hours: Wednesday 12-1PM)

TAs:

- Xinyi Wang (office hours: Tuesday 4-5PM)
- Tanmay Parekh (office hours: Monday 4-5PM)
- Sachin Kumar (office hours: TBD)

Class Format:

<u>At Your Convenience</u>

- ~30 minute **lecture video**, with optional reading. At the end of the video there will be discussion questions.

<u>During Class -- Starting at **5:40PM**</u>

- ~10 minute **language in 10**: introduce a language, in groups of 2-3.
- ~30 minute, **breakout room discussion** or **code/data/assignment walk-through**
- ~10 minute **summary**

Grading Policy

- Class/Discussion Participation: 15%
- Language in 10 Presentation: 5%
- Assignment 1 (Multilingual Sequence Labeling, individual): 15%
- Assignment 2 (Multilingual Translation, group): 20%
- Assignment 3 (Multilingual Speech Recognition, group): 20%
- Project: 25%

Reading Assignment for 9/3

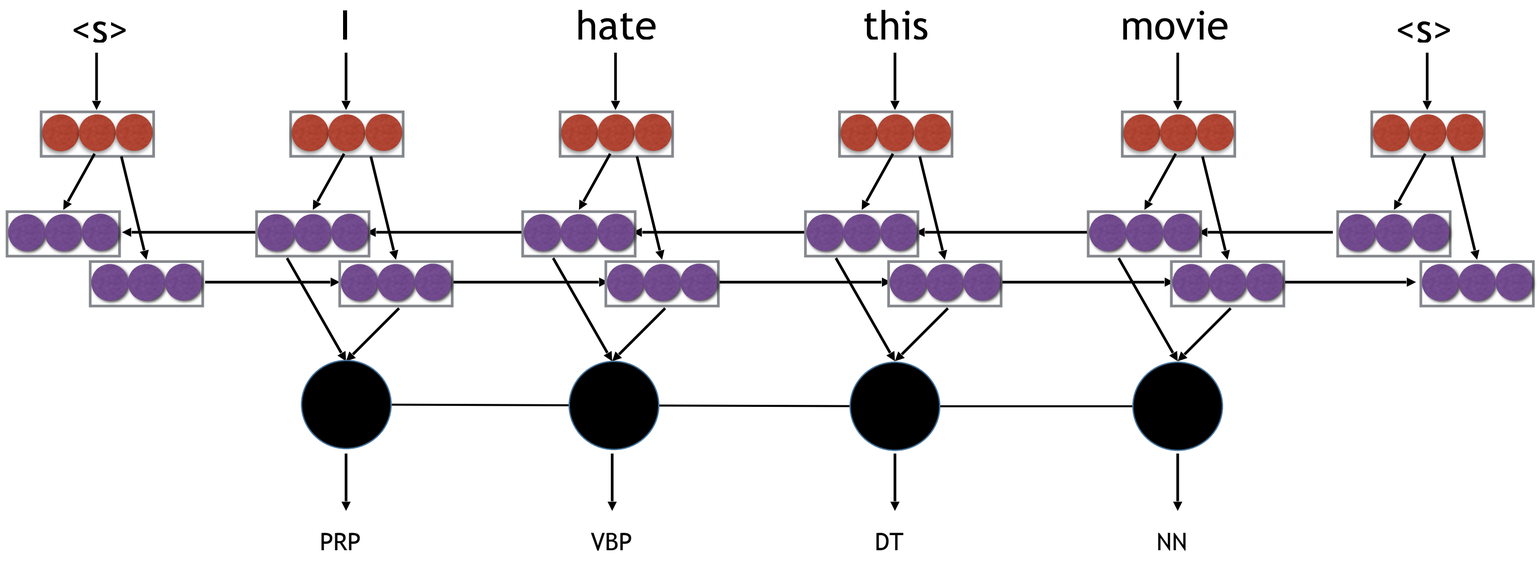
- Lin, et al.
  "Choosing Transfer Languages for Cross-Lingual Learning"
  In Proc. ACL 2019.
  https://arxiv.org/pdf/1905.12688.pdf

Yulia:
Linguistics, Multilinguality, and Multilingual Training

Graham:
Sequence Labeling, Translation, and Syntax

Sequence Labeling/Classification

- **Tasks:** language ID, POS tagging, named entity recognition, entity linking
- **Models:** sequence encoders, subword encoding, CRFs

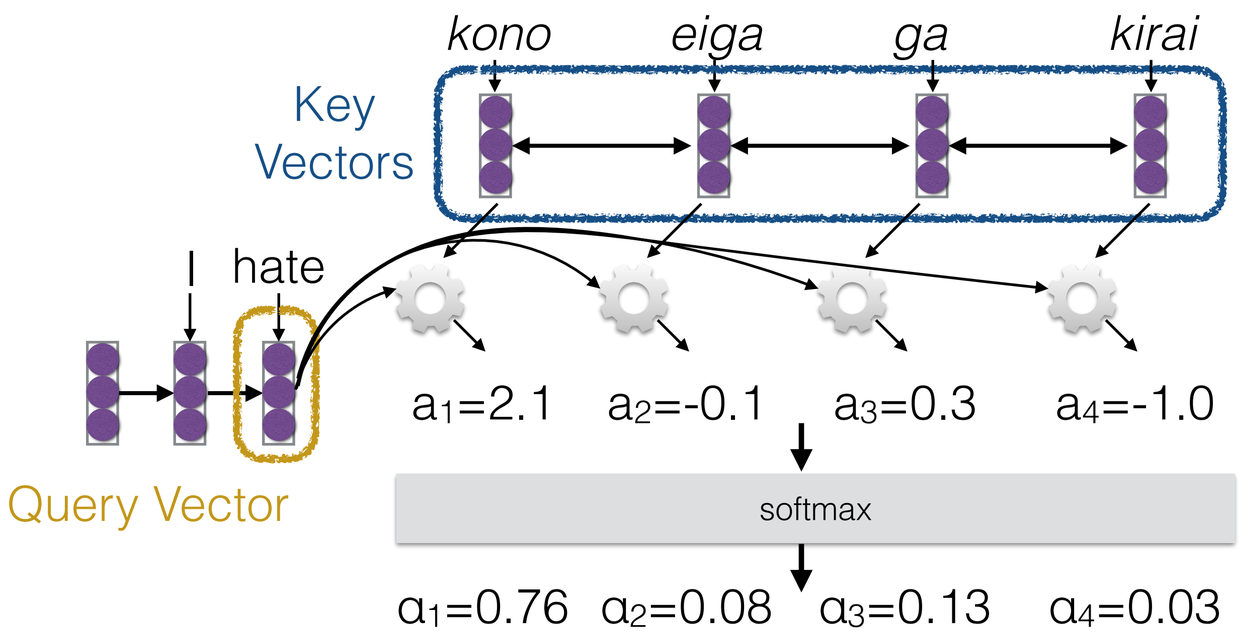


- **Data:** universal dependencies POS tags, wikipedia-based NER/linking

Machine Translation and Sequence-to-sequence Models

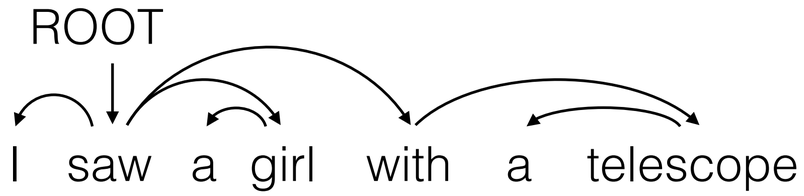
- Sequence-to-sequence problems
- Seq2seq models with attention
- Transformers
- Low-resource considerations

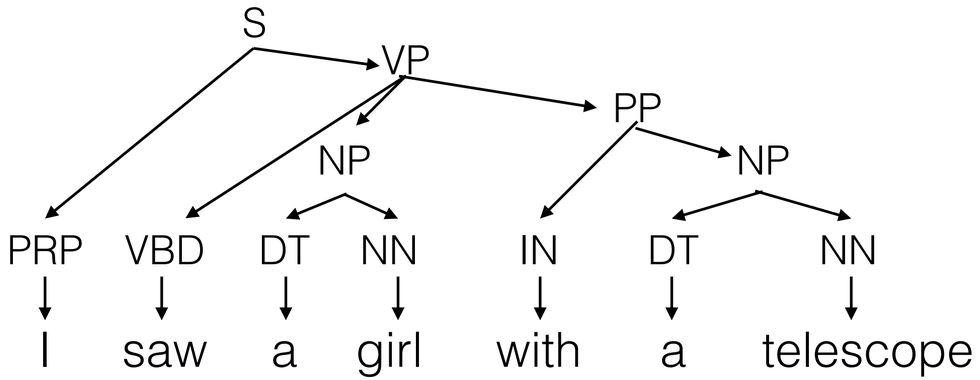
Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognitio
Text	Linguistic Structure	Language Analysis



Syntax

- Syntactic formalisms





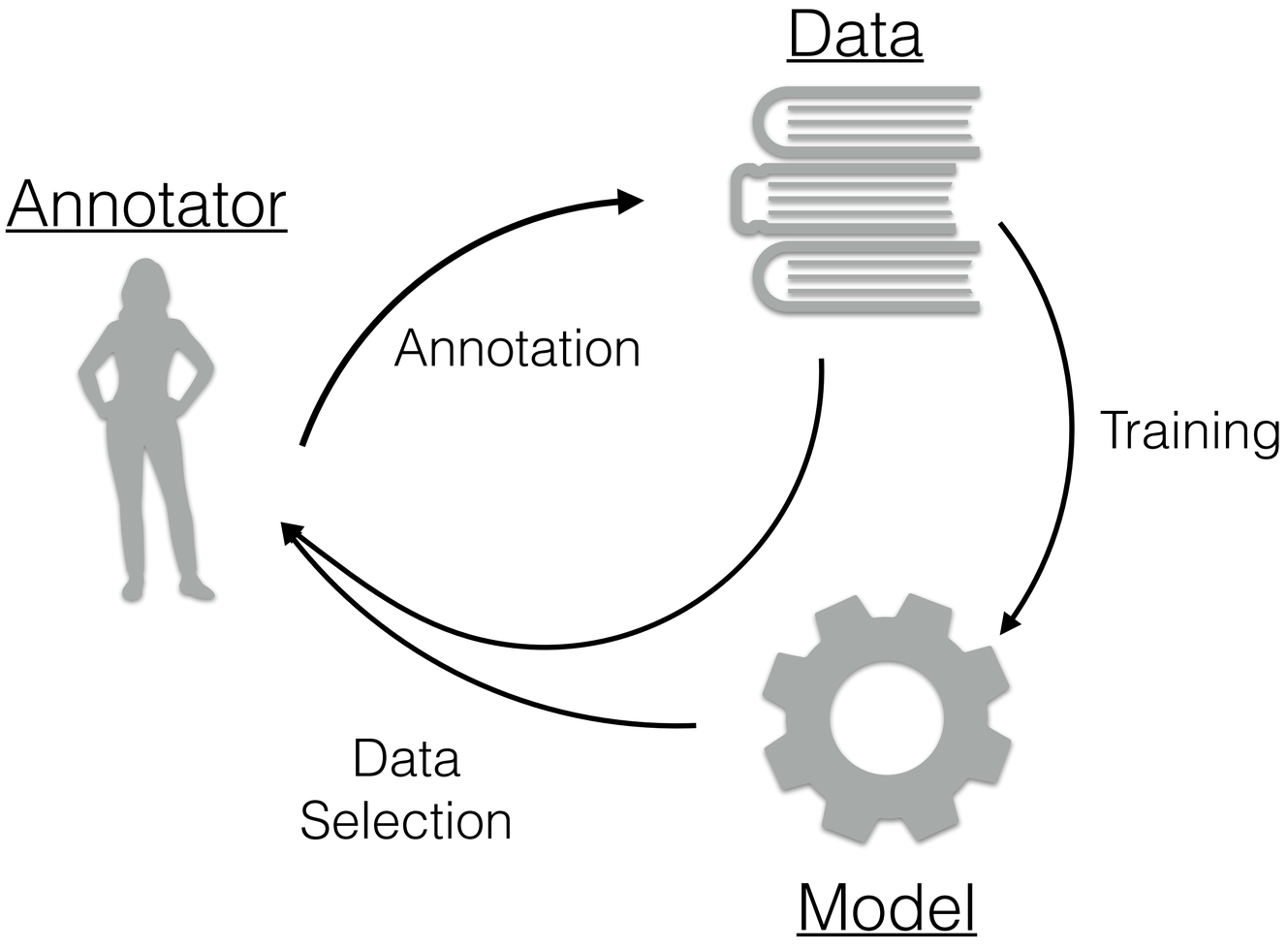
- Syntactic typology

English = SVO: *he bought a car* **Japanese** = SOV: *kare wa kuruma wo katta*

Irish = VSO: *cheannaigh sé carr* **Malagasy** = VOS: *nividy fiara izy*

Active Learning

- Incremental creation of data and model improvement



Alan:
Speech, Code Switching, and Annotation

Thank You, Questions?

Then, Language in 10!