
CIS-620

Spring 2021

# Learning in Few-Labels Setting

Dan Roth

Computer and Information Science

University of Pennsylvania

# This class

- Understand early and current work on Learning in Few-Labels Settings
  - (Learn to) read critically, present, and discuss papers
- Think about, and Understand realistic learning situations
  - Move away from the “easy cases” to the challenging ones
  - Conceptual and technical
- Try some new ideas
- How:
  - Presenting/discussing papers
    - Probably: 1-2 presentations each;
    - Each paper will have 2 discussants: pro/con
  - Writing 4 critical reviews
  - “Small” individual project (reproducing);
  - Large project (pairs)
  - Tentative details are on the web site.

- Machine Learning
  - 519/419
  - 520
  - Other?
- NLP
  - Yoav Goldberg’s book
  - Jurafsky and Martin
  - Jacob Eisenstein
- Attendance is mandatory
- Participation is mandatory

- Time: Monday 3pm, break, 4:30 pm.
- Zoom Meeting https://upenn.zoom.us/j/95494190734?pwd=MzhMek83U0hCSVgrblZkenZjL1hIUT09
- **TA: Soham Dan**
  - **Office hours: 6-7pm Monday**

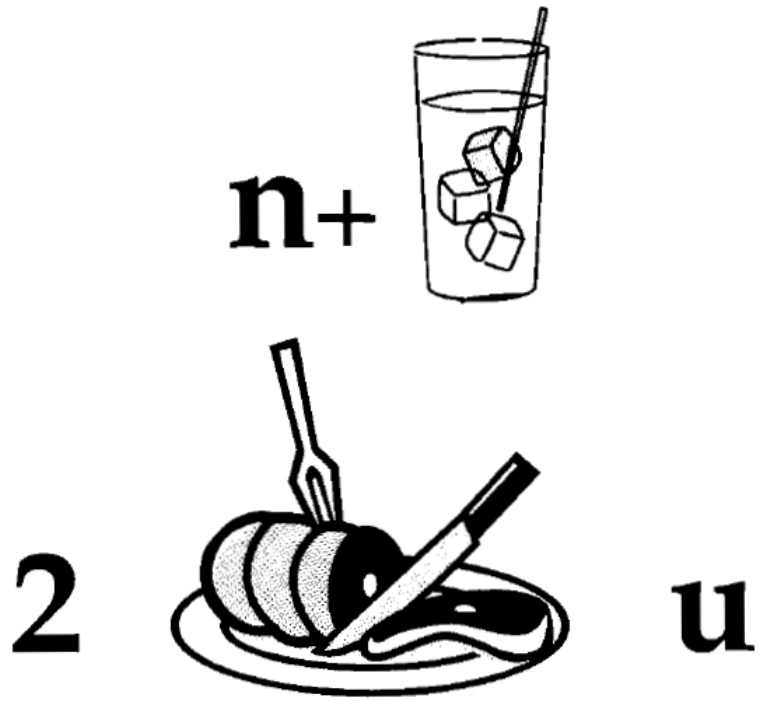
# What's Important in order to make progress in NLU

- How to make progress towards natural language understanding
  - Learning and Reasoning; knowledge
- Dispel with [some] of the current trends
  - Mostly – invent training datasets and hope…
  - If we want to reach the moon…
- What are the real challenges today?
  - "Sparse" reasoning tasks
  - Generalization
  - Supporting latent decisions
  - Low resource languages
  - ….
  - These are likely to present different technical challenges
- Today: Examples & Discussion

# Some Examples

# Nice to Meet You

- Today, think about it from the perspective of:
  - How do we train systems that can solve it?



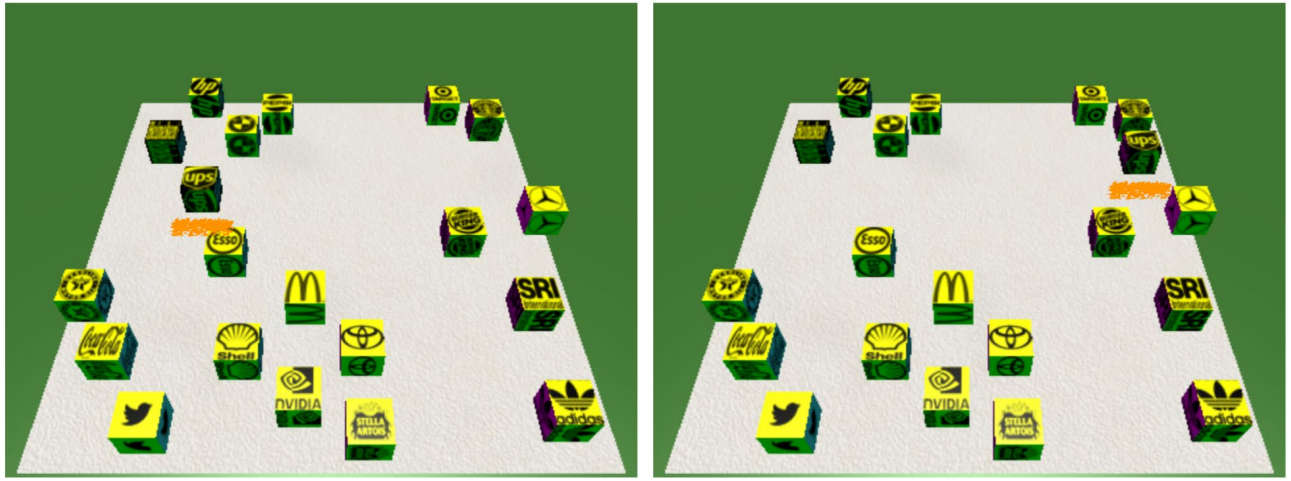

- Identify units
- Consider multiple representations & interpretations
  - Pictures, text, layout, spelling, phonetics
- Put it all together:
  - Determine “best” global interpretation
- Satisfy expectations
  - Slide; puzzle

Computational Problem:
Assigning values to multiple variables, accounting for interdependencies among them

# Communication (CwC Project, 2017-2020)

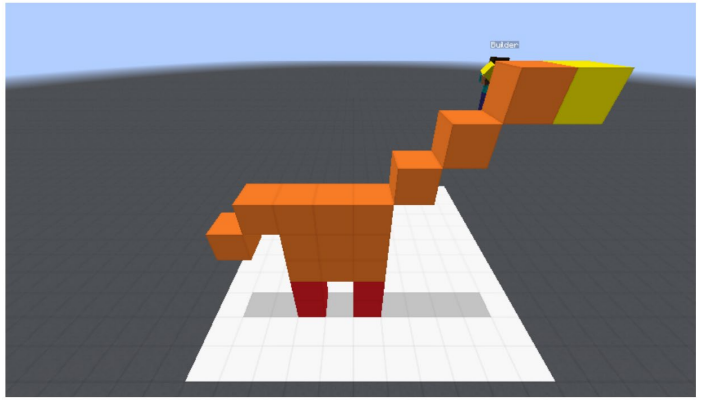
- Consider the following explanation:
  - **[Example from Yonatan Bisk]**
  - Imagine that this is a chess board
  - Place the UPS in H2, and McDonald in G6.

- Can you follow these instructions?
  - What's needed for us to be able to **write programs** that can do it?*



[I need to] move UPS from the left side of the board to just below Starbucks, leaving a small gap.

THE TARGET CONFIGURATION



*Without being trained on thousands of chess-understanding examples

# Easy for humans, hard for machines

- Humans “understand” text at levels that are far from AI capabilities today
- Moving forward often requires “reasoning” about the world

Who was in the store when the events began?

Probably Mr. Hug alone, although the robbers might have been waiting for him, but if so, this would have been stated.

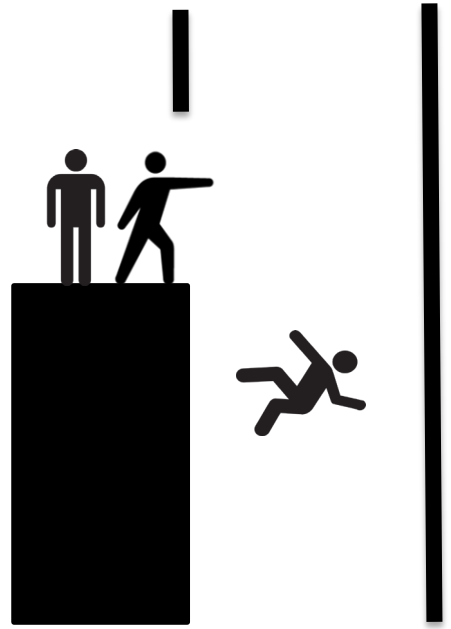
What did the porter say to the robbers?

Nothing, because the robbers left before he came.

Why did Mr. Hug yell from the bottom of the elevator shaft?

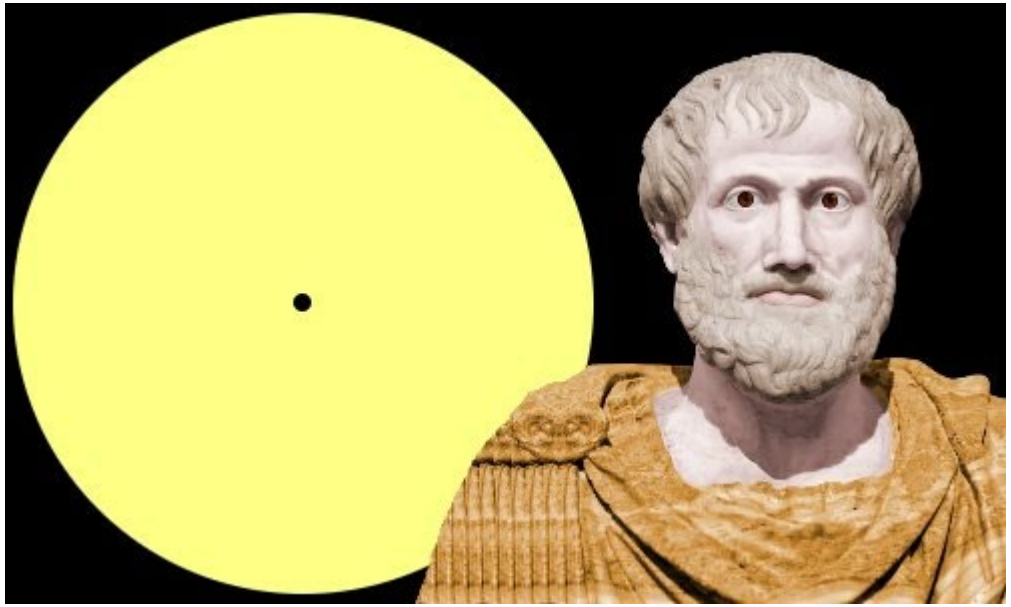
So as to attract the attention of someone who would rescue him.

*A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday in his downtown Brooklyn store by two robbers while a third attempted to crush him with the elevator car because they were dissatisfied with the $1,200 they had forced him to give them.*
*The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.*



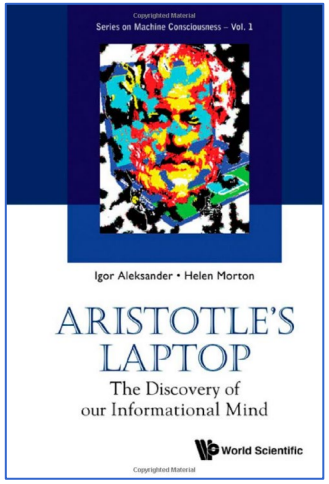
[McCarthy, 1976]

# Aristotle's Logic https://www.famousscientists.org/aristotle/



- Aristotle founded the study of formal logic, systematizing logical arguments – he is famous for the syllogism, a method by which known information can be used to prove a point.

- Here is a famous example, from Aristotle himself, of a simple syllogism:
  - All men are mortal.
  - Socrates is a man.
  - Therefore, Socrates is mortal.




So, did Aristotle have a laptop?

[Geva et al. TACL'21]

[Example due to Les Valiant]

# De-compositions for Reasoning about Text [Gupta et al. 18', 19' 20']

- Text Comprehension challenges:
  - ☐ Understand the text
    - Identify and contextualize events, entities, quantities (and their scope), relations, etc.
  - ☐ Understand questions about the text
    - Often, requires decomposing the question in a way that depends on the text
  - ☐ Reason about the text
    - Combine and manipulate the identified information to accomplish information needs (e.g., answering questions)
- How can we supervise to support this level of understanding?
  - ☐ Too many (ill-defined) latent decisions
    - Annotating text for all is not scalable
  - ☐ End-task supervision is the only realistic approach [Clarke et. al.'10] but it is too loose – how can we learn all the latent decisions from end-to-end supervision?

Who scored the longest touchdown pass of the game?

In the … Greg … touchdown run. The Bears increased their lead over the Vikings with Cutler's 2-yard TD pass to tight end Desmond Clark. The Vikings … with Favre firing a 6-yard TD pass to tight end Visanthe Shiancoe. The Vikings … with Adrian Peterson's second 1-yard TD run. The Bears then responded with Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu.

What is her seizure frequency?

She reports wors… now occurring up to 10/week, in clusters about 2-3 day/week. Previously reported seizures occurring about 2-3 times per month, often around the time of menses,…

How much did his challengers raise?

Mayor Rahm Em… toward his bid for a thi… raised by his 10 challengers combined, campaign finance records show.

The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known COVID-19 deaths, and the most deaths of any country.

# Challenges

- When you read a piece of text, your understanding of the text goes far beyond any of the “tasks defined so far by the NLP community”

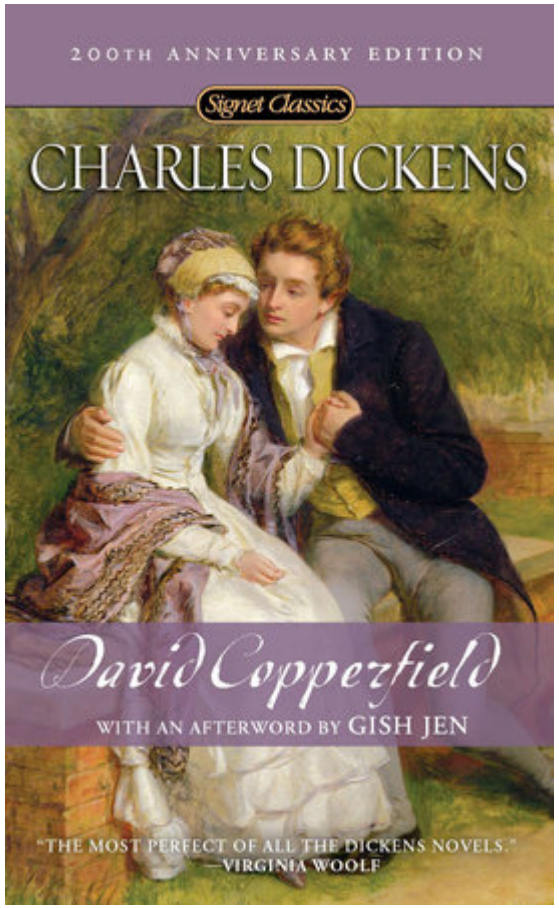
O’Hare must be in Chicago

Feb 5 2017 Dozens of [illegible]ngers heading to Chicago had to undergo additional screenings… It took H[illegible] Aamyab two tries to make it back to the United States from Iran. He is an Iranian [illegible]en with a US visa who is doing post-doctoral research at UIC. …"Right now, I am in [illegible]e USA and I'm very happy," Aamyab said. But now, he can't go back to Iran or anywhere [illegible]lse without risk. Other travelers shared the same worry. Asem Aleisawi was at O'Hare on Sunday to meet his wife who was coming in from Jordan.

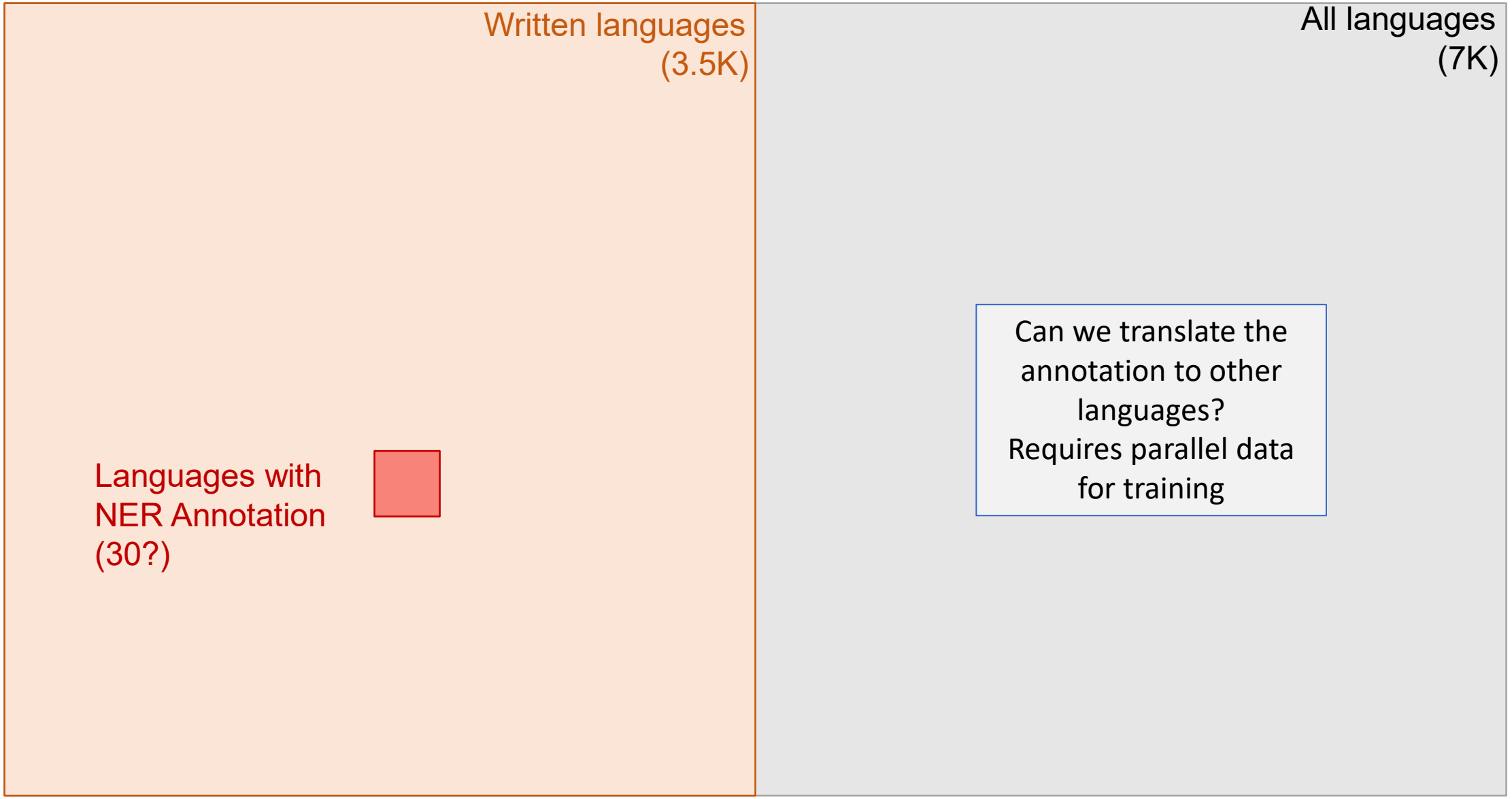
# CliffsNotes

- Multiple natural language documents
  - Small units of text or large units of texts
  - Reading news about an event/situation **over time** and/or from **multiple sources**
  - **Reading a book**

  - The novel features the character David Copperfield, his journey of change and growth from infancy to maturity, as many people enter and leave his life and he passes through the stages of his development. (**Fiction, and you know it**)
  - London and England in the 19-th century; socio-economic state, child exploitation; schools, prisons, emigration to Australia (**True historical facts**)
- What are the computational tasks that we should think about?

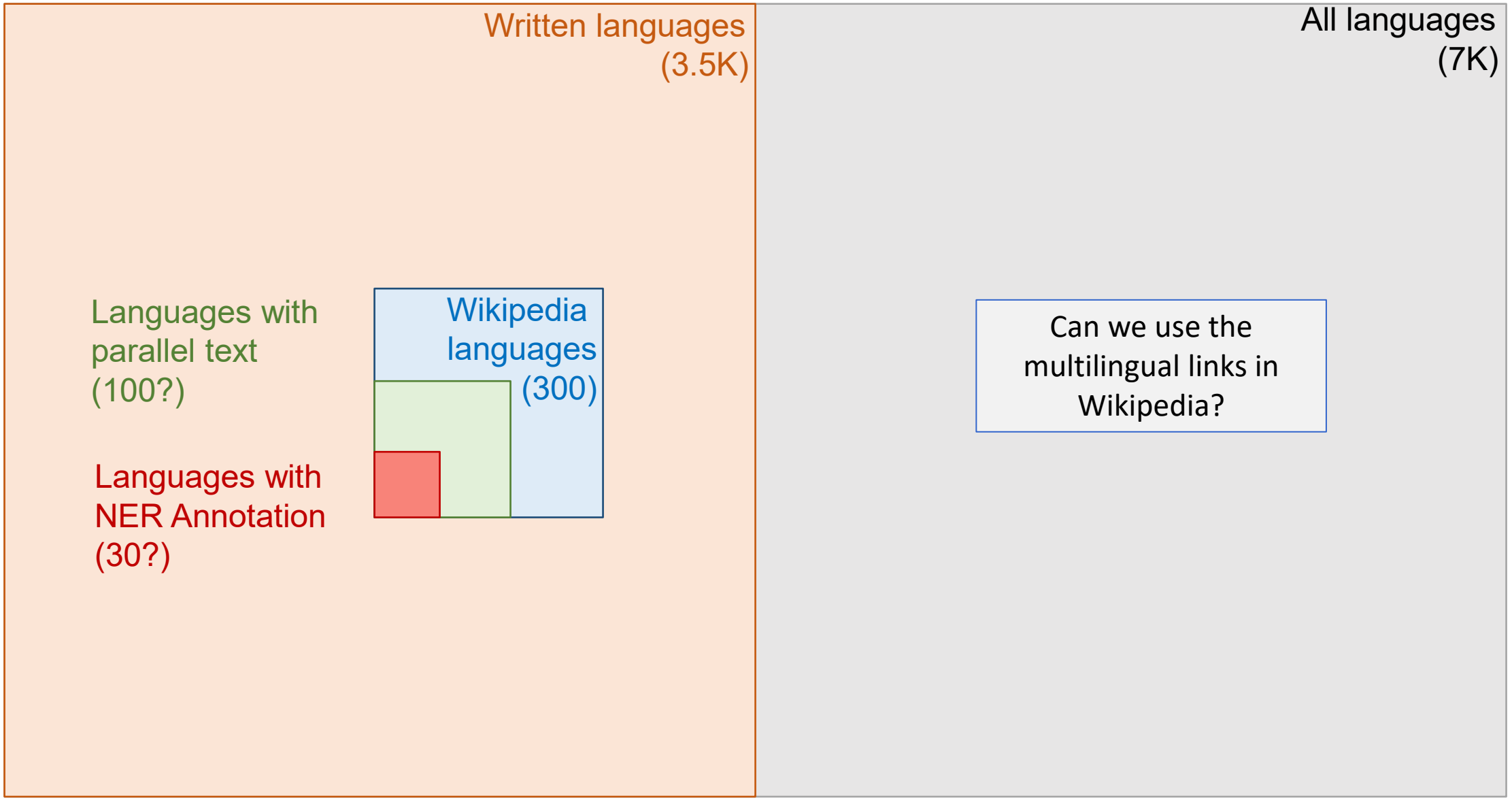


# Low Resource Languages



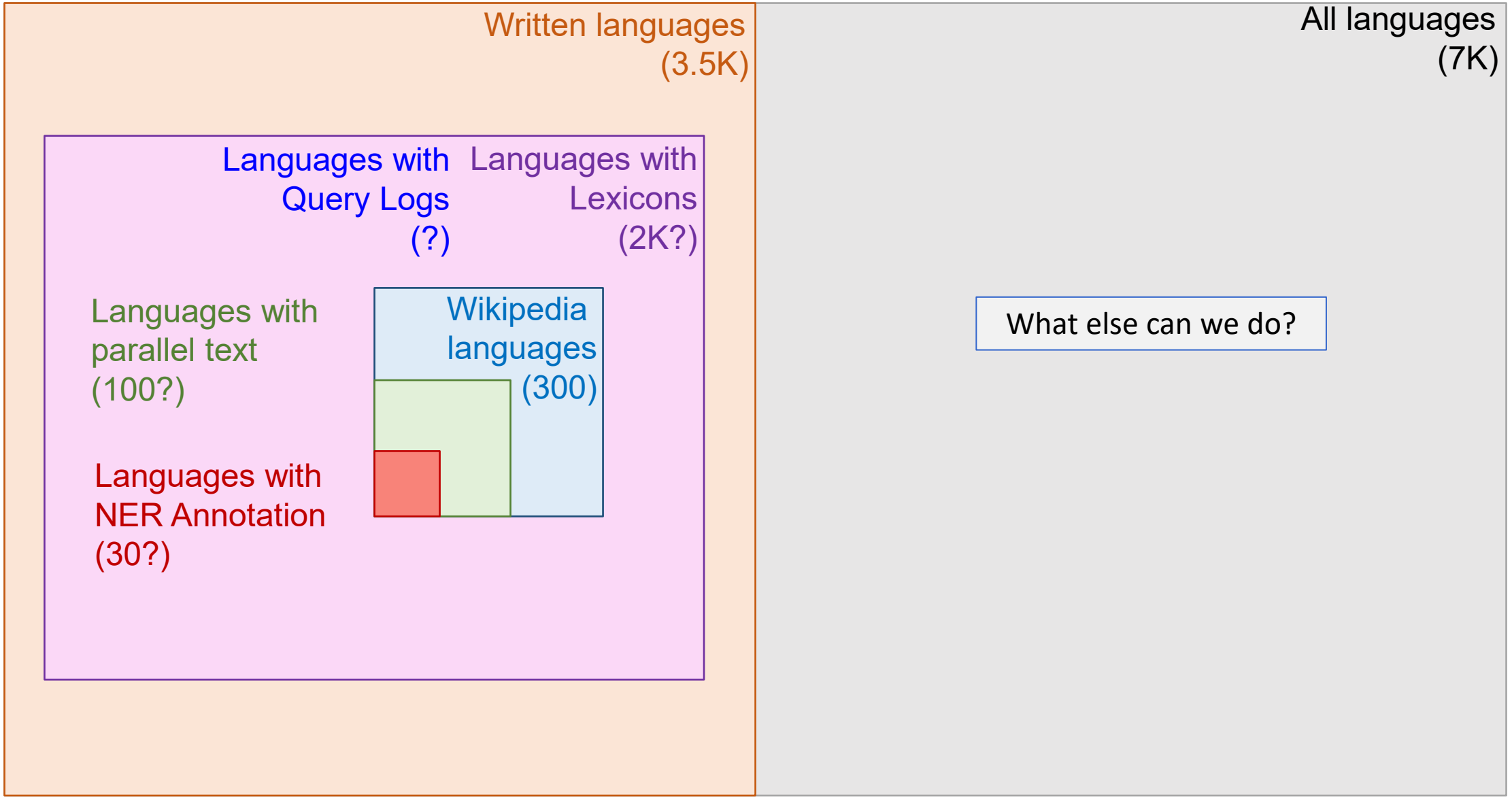

(Diagram roughly to scale. Numbers from ethnologue.com)

# Low Resource Languages




(Diagram roughly to scale. Numbers from ethnologue.com)

# Low Resource Languages




(Diagram roughly to scale. Numbers from ethnologue.com)

# Indian Languages

## Languages Translated by Google

| | | | |
|---|---|---|---|
| 1. Afrikaans | 14. Cebuano | 27. Finnish | 40. Hungarian |
| 2. Albanian | 15. Chewa | 28. French | 41. Icelandic |
| 3. Amharic | 16. Chinese (Simplified) | 29. Galician | 42. Igbo |
| 4. Arabic | 17. Chinese (Traditional) | 30. Georgian | 43. Indonesian |
| 5. Armenian | 18. Corsican | 31. German | 44. Irish |
| 6. Azerbaijani | 19. Croatian | 32. Greek | 45. Italian |
| 7. Basque | 20. Czech | 33. Gujarati | 46. Japanese |
| 8. Belarusian | 21. Danish | 34. Haitian Creole | 47. Javanese |
| 9. Bengali | 22. Dutch | 35. Hausa | 48. Kannada |
| 10. Bosnian | 23. English | 36. Hawaiian | 49. Kazakh |
| 11. Bulgarian | 24. Esperanto | 37. Hebrew | 50. Khmer |
| 12. Burmese | 25. Estonian | 38. Hindi | 51. Kinyarwanda |
| 13. Catalan | 26. Filipino (Tagalog) | 39. Hmong | 52. Korean |

| | | | |
|---|---|---|---|
| 53. Kurdish (Kurmanji) | 66. Marathi | 79. Scottish Gaelic | 92. Tajik |
| 54. Kyrgyz | 67. Mongolian | 80. Serbian | 93. Tamil |
| 55. Lao | 68. Nepali | 81. Shona | 94. Tatar |
| 56. Latin | 69. Norwegian (Bokmål) | 82. Sindhi | 95. Telugu |
| 57. Latvian | 70. Odia | 83. Sinhala | 96. Thai |
| 58. Lithuanian | 71. Pashto | 84. Slovak | 97. Turkish |
| 59. Luxembourgish | 72. Persian | 85. Slovenian | 98. Turkmen |
| 60. Macedonian | 73. Polish | 86. Somali | 99. Ukrainian |
| 61. Malagasy | 74. Portuguese | 87. Sotho | 100. Urdu |
| 62. Malay | 75. Punjabi (Gurmukhi) | 88. Spanish | 101. Uyghur |
| 63. Malayalam | 76. Romanian | 89. Sundanese | 102. Uzbek |
| 64. Maltese | 77. Russian | 90. Swahili | 103. Vietnamese |
| 65. Maori | 78. Samoan | 91. Swedish | 104. Welsh |

- Among the top 80 languages in the world, spoken by more than 10m people, 18 are Indian languages
- http://www2.harpercollege.edu/mheal y/g101ilec/intro/clt/cltclt/top100.html

- There are 109 languages in Google Translate
- 10 of them are Indian Languages

  (all, with >25m speakers)

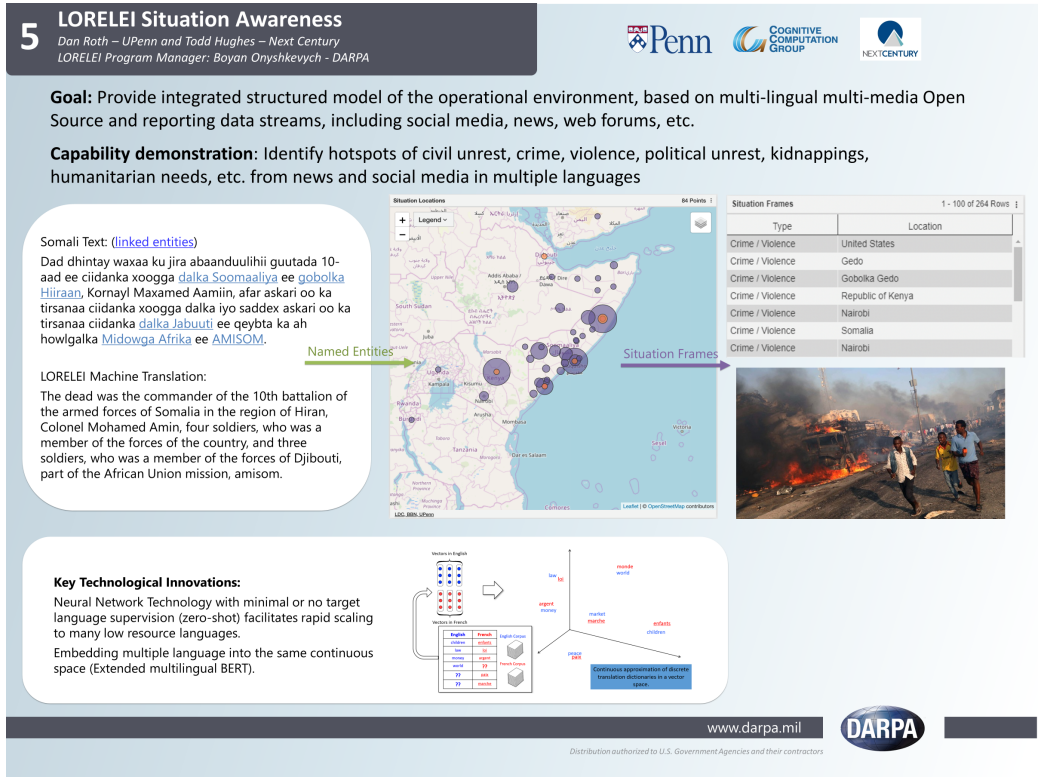
# Target Language Source ⇨ English

- Can we at least “understand” documents in low resource languages?

Of course, you can imagine many other challenges in low resource languages

Somali streaming data ⇨ Situation Awareness (described in English)


- What is it about?
  - ☐ Identify topics & events
- “Understand” a situation described in Target Language
  - ☐ Identify Entities & Concepts (NER)
  - ☐ Ground in English Resources (XEL)



# New Domains; New Tasks

- Electronic Health Records

> She reports worsened seizure frequency, seizures now occurring up to 10/week, in clusters about 2-3 day/week. Previously reported seizures occurring about 2-3 times per month, often around the time of menses,…

- Many questions:
  - Blame classifier: give reported Failures: who is at fault?
  - Discourse with patients: did the physician informed them about A?
  - …

# Challenges (1)

- Analyze Suspicious  Transaction Reports (STRs) that banks submit to government agencies.
  - Extract relationships and transaction details.

- ALERTED ACTIVITY
  - *On 01/01/2011 the 01/01/2011 wire was stopped by ABC DE's filter:*
  - *JJ sent a $4,950.00 wire from account 123567 at GH Bank DE, through ABC DE, to account 123456 at Peoples' Bank of Malibu head office, Malibu, to be remitted to People's Bank of Malibu, XYZ Branch, for the benefit of BB for "Prayer Religious Items."*

- Determine: who did what to whom, where and when?
  - What are the transactions; from whom, to whom?
  - People involved, roles and affiliations
  - Organizations and their roles
  - Timelines of events

# Challenges (2)

- Inter-banking rate issues
  - A client investigates if bankers were fixing the inter-banking rate.
  - *"Mate, can you raise the main one by 0.2 till Wednesday? I owe you a drink."*
- Messages here are typically very short; the language is colloquial and ungrammatical. Messages include many quantities (rates), how much to adjust, etc.
- A "simple" classification problem – but what is the label?
  - And, where is the training data?
  - (Australian company)

# The Space

- There is a wide range of challenges to the current supervision paradigms
  - But, what we know is to train supervised models

- Can we think about a collection of reductions
  - Mapping realistic scenarios (with realistic scenarios) to a supervised setting?

  - Example: context sensitive spelling correction

    Exposure to the sun had the affect of toughening his skin.

  - How do we train for it?
    - Invent a notion of confusion set;
    - Use available text; drop a confusion set member – positive example; other members – negatives
    - The 1990-ies version of the (discriminative) masked language model
  - This is a reduction to the supervised paradigm; accomplished by dreaming up an indirect supervision signal.

# What Should We Address?

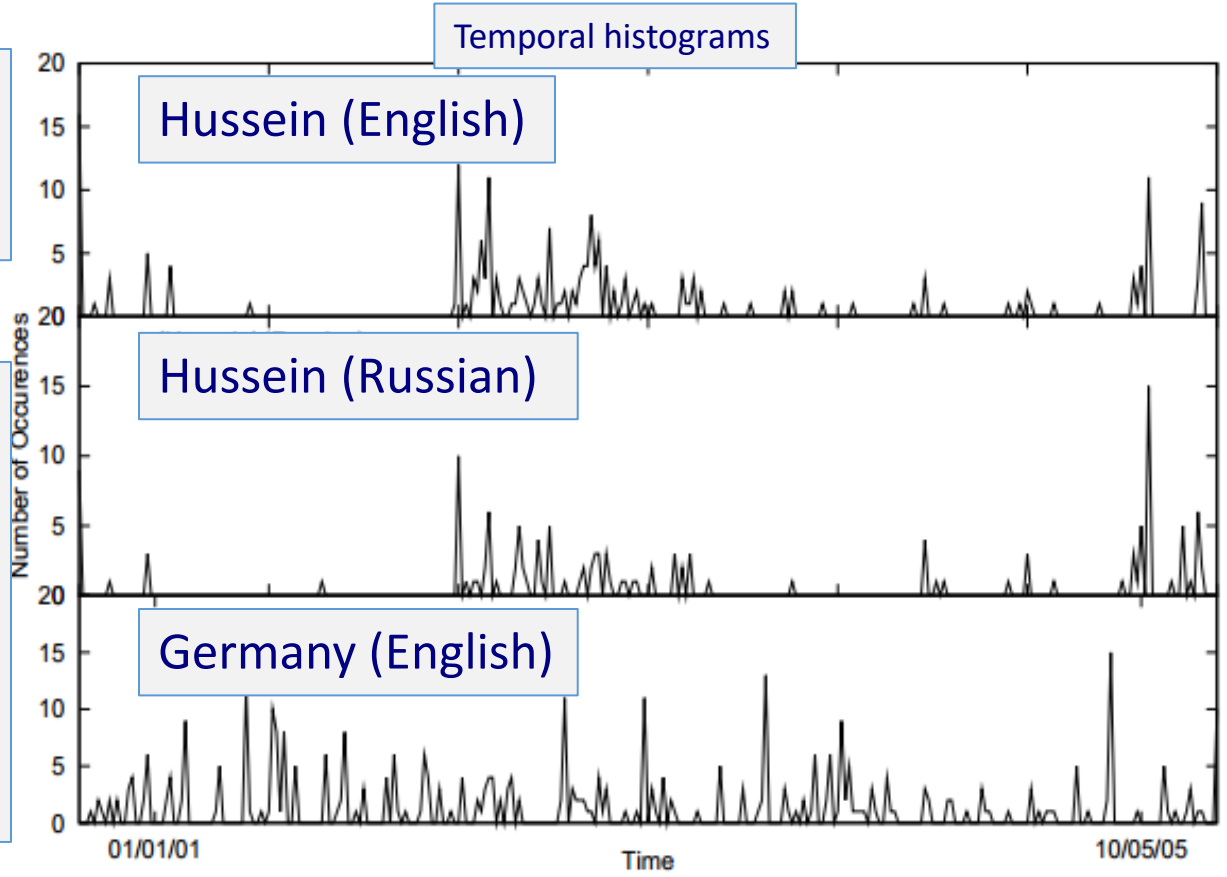
- Zero-shot (few shot) Learning
  - Label-Aware methods
  - Transfer learning methods
  - Representation driven methods
- Incidental Supervision Signal
  - Where can we get signals from?
  - How to use them?
  - Is it art?

# Incidental Supervision Signals

- Supervision could be incidental, in the sense that it makes use of signals that might be co-related to the target task.
- Searching for supervision signals could be challenging.

Assume a comparable, weakly temporally aligned news feeds.

Weak synchronicity provides a cue about the relatedness of (some) NEs across the languages, and can be exploited to associate them
[Klementiev & Roth, 06,08]



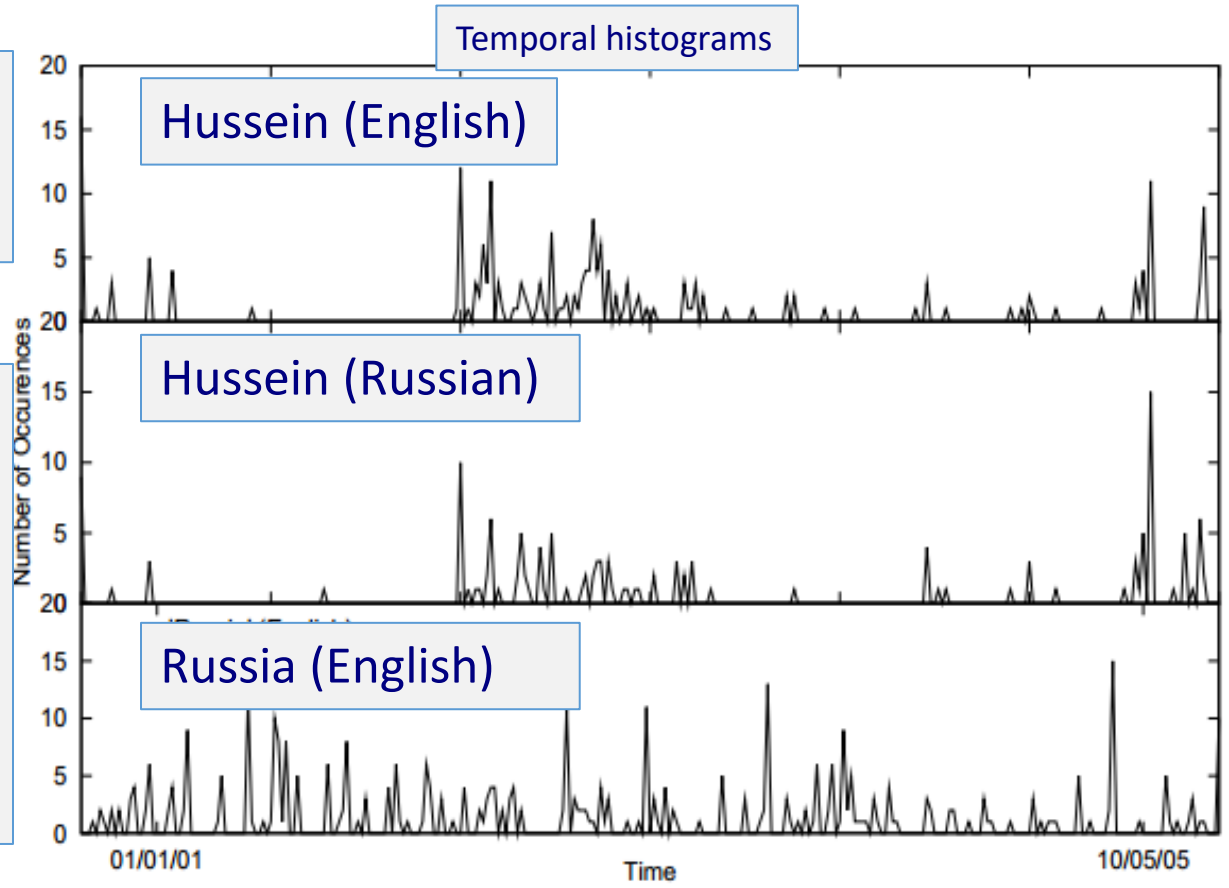
# Incidental Supervision Signals

Need an **inference mechanism** that supports combining signals/models

- By itself, this temporal signal may not be sufficient to support learning robust models.
- Along with weak phonetic signals, context, topics, etc. it can be used to get robust models.

Assume a comparable, weakly temporally aligned news feeds.

Weak synchronicity provides a cue about the relatedness of (some) NEs across the languages, and can be exploited to associate them [Klementiev & Roth, 06,08]



# What Should We Address?

- Zero-shot (few shot) Learning
  - Label-Aware methods
  - Transfer learning methods
  - Representation driven methods
- Incidental Supervision Signals
  - Where can we get signals from?
  - How to use them?
  - Is it art?
- Low Resource Languages
  - New Signals & projection
  - Representations
- End-task Supervision
  - When and How?
  - How to use indirect supervision signals?
- Knowledge as supervision
  - Constraints driven paradigms
  - Partial supervision
- Transfer Learning & Adaptation
  - Domain shift
  - Label space shift
- Theory

# This class

- Presentations
  - Key part of the class
  - Send me your presentation by Wednesday before you present.
- Discussion/Discussants
- Projects

Questions?

- Understand early and current work on Learning in Few-Labels Settings
  - ☐ (Learn to) read critically, present, and discuss papers
- Think about, and Understand realistic learning situations
  - ☐ Move away from the “easy cases” to the challenging ones
  - ☐ Conceptual and technical
- Try some new ideas
- How:
  - ☐ Presenting/discussing papers
    - Probably: 1-2 presentations each;
    - Each paper will have 2 discussants: pro/con
  - ☐ Writing 4 critical reviews
  - ☐ “Small” individual project (reproducing);
  - ☐ Large project (pairs)
  - ☐ Tentative details are on the web site.

- Machine Learning
  - 519/419
  - 520
  - Other?
- NLP
  - Yoav Goldberg’s book
  - Jurafsky and Martin
  - Jacob Eisenstein
- Attendance is mandatory
- Participation is mandatory

- Time: Monday 3pm, break, 4:30 pm.
- Zoom Meeting https://upenn.zoom.us/j/95494190734?pwd=MzhMek83U0hCSVgrblZkenZjL1hIUT09
- **TA: Soham Dan**
  - **Office hours: 6-7pm Monday**