Project discussion

CS 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor

announcements

```
     Su Mo Tu We Th Fr Sa
Oct 27 28 29 30  1  2  3
     4  5  6  7  8  9 10
    11 12 13 14 15 16 17
    18 19 20 21 22 23 24   Midterm
    25 26 27 28 29 30 31
```

- Midterm moved to 10/20
- HW1 grades coming this weekend
- Ex2&3 handed back (up front) - should be recorded "Received" in Moodle
- (Extra HW0 submissions still being processed, done soon)

Project

- Either *build* natural language processing systems, or *apply* them for some task.
- Use or develop a dataset. Report empirical results or analyses with it.
- Different possible areas of focus
  - Implementation & development of algorithms
  - Defining a new task or applying a linguistic formalism
  - Exploring a dataset or task

Project

	Su	Mo	Tu	We	Th	Fr	Sa	
Oct	27	28	29	30	1	2	3	
	4	5	6	7	8	9	10	
	11	12	*13*	14	15	16	17	
	18	19	**20**	21	22	**23**	24	Proposal due
	25	26	27	28	29	30	31	
Nov	1	2	3	4	5	6	7	
	8	9	10	11	12	**13**	14	Progress due
	15	16	17	18	19	20	21	
	22	23	24	*25*	*26*	*27*	28	
Dec	29	30	1	2	3	4	5	
	6	7	8	9	10	*11*	12	Presentations
	13	*14*	*15*	*16*	*17*	***18***	19	Final report due

Proposal: 2-4 page document outlining the problem, your approach, possible dataset(s) and/or software systems to use. Must cite and briefly describe at least **two** pieces of relevant prior work (research papers). Describe scope of proposed work.

Progress report: Longer document with preliminary results

Presentations: In-class and short

Final report

- Groups of 1-3: we encourage size 2
  - We expect more work with more team members

NLP Research

- All the best publications in NLP are open access!
  - Conference proceedings: ACL, EMNLP, NAACL (EACL, LREC...)
  - Journals: TACL, CL
  - NLP and NLP-related work appears in other journals/conferences too (data mining, machine learning, AI, information retrieval, etc.)
- Reading tips
  - Google Scholar
    - Find papers
    - See paper's number of citations (imperfect but useful correlate of paper quality) and what later papers cite it
  - Authors' webpages (find researchers who are good at writing and whose work you like)
  - Misc. NLP research reading tips: http://idibon.com/top-nlp-conferences-journals/

A few examples

A few examples

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning

A few examples

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
- Structured linguistic prediction
  - Targeted sentiment analysis (i liked __ but hated __)
  - Relation, event extraction (who did what to whom)
  - Narrative chain extraction
  - Parsing (syntax, semantics, discourse...)

A few examples

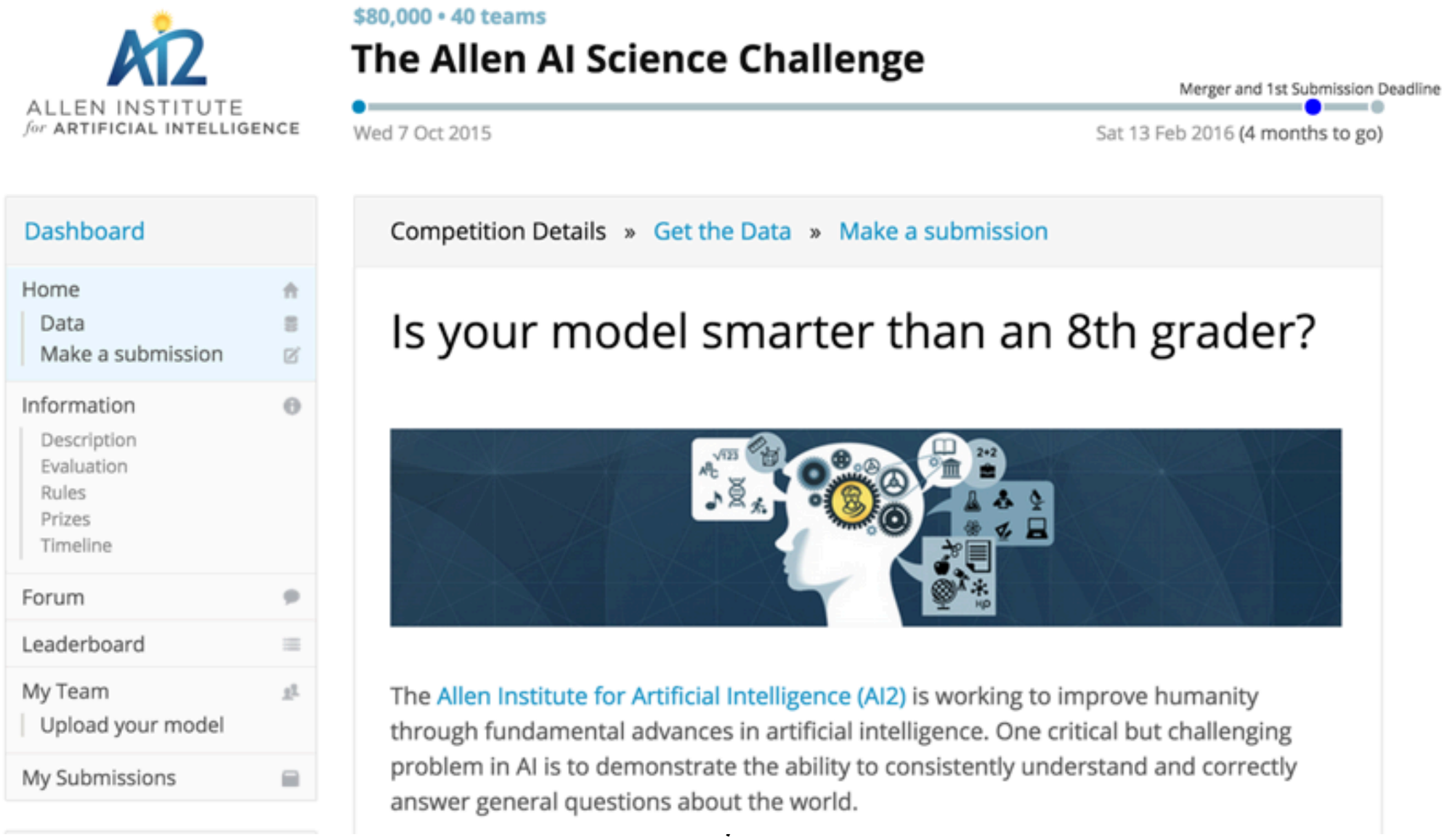
- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
- Structured linguistic prediction
  - Targeted sentiment analysis (i liked __ but hated __)
  - Relation, event extraction (who did what to whom)
  - Narrative chain extraction
  - Parsing (syntax, semantics, discourse...)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)

A few examples

- Detection tasks
  - Sentiment detection
  - Sarcasm and humor detection
  - Emoticon detection / learning
- Structured linguistic prediction
  - Targeted sentiment analysis (i liked __ but hated __)
  - Relation, event extraction (who did what to whom)
  - Narrative chain extraction
  - Parsing (syntax, semantics, discourse...)
- Text generation tasks
  - Machine translation
  - Document summarization
  - Poetry / lyrics generation (e.g. recent work on hip-hop lyrics)
- End to end systems
  - Question answering
  - Conversational dialogue systems (hard to eval?)
- Predict external things from text
  - Movie revenues based on movie reviews ... or online buzz? http://www.cs.cmu.edu/~ark/movie$-data/
- Visualization and exploration (harder to evaluate)
  - Temporal analysis of events, show on timeline
  - Topic models: cluster and explore documents
- Figure out a task with a cool dataset
  - e.g. Urban Dictionary

Science question answering

- a "full-stack" sort of task ... 8th-grade science textbook input, question-answering output
- https://www.kaggle.com/c/the-allen-ai-science-challenge



Sources of data

- All projects must use (or make, and use) a textual dataset. Many possibilities.
  - For some projects, creating the dataset may be a large portion of the work; for others, just download and more work on the system/modeling side

- SemEval and CoNLL Shared Tasks:
  dozens of datasets/tasks with labeled NLP annotations
  - Sentiment, NER, Coreference, Textual Similarity, Syntactic Parsing, Discourse Parsing, and many other things...
  - e.g. SemEval 2015 ... CoNLL Shared Task 2015 ...
  - https://en.wikipedia.org/wiki/SemEval (many per year)
  - http://ifarm.nl/signll/conll/ (one per year)

- General text data (not necessarily task specific)
  - Books (e.g. Project Gutenberg)
  - Reviews (e.g. Yelp Academic Dataset https://www.yelp.com/academic_dataset)
  - Web
  - Tweets

Tools

- Tagging, parsing, NER, coref, ...
  - Stanford CoreNLP http://nlp.stanford.edu/software/corenlp.shtml
  - spaCy (Eng-only, no coref) http://spacy.io/
  - Twitter-specific tools (ARK, GATE)
- Many other tools and resources

  tools ... word segmentation ... morph analyzers ...
  resources ... pronunciation dictionaries ... wordnet, word embeddings, word clusters ...
- Long list of NLP resources

  https://medium.com/@joshdotai/a-curated-list-of-speech-and-natural-language-processing-resources-4d89f94c032a

Things to do with a log-linear model

$$p(y|x) = \frac{1}{Z} \exp \underbrace{\left(\theta^\mathsf{T} f(x,y)\right)}_{G(y)}$$

	f(x,y) Feature extractor (feature vector)	**x** Text Input	**y** Output	**θ** Feature weights
decoding/prediction $\arg \max_{y^* \in outputs(x)} G(y^*)$	given	given (just one)	obtain (just one)	given
parameter learning	given	given (many pairs)	given (many pairs)	obtain
feature engineering (human-in-the-loop)	fiddle with during experiments	given (many pairs)	given (many pairs)	obtain in each experiment