NLP Data

Vsevolod Dyomkin

prj-nlp-1, 2018-03-15

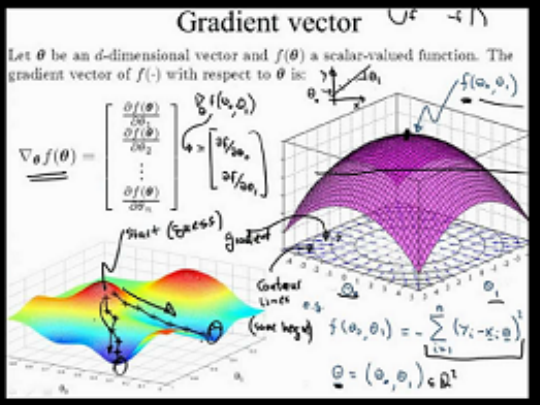
Data Scientist

What my friends think I do

What my mom thinks I do

What society thinks I do

What my boss thinks I do



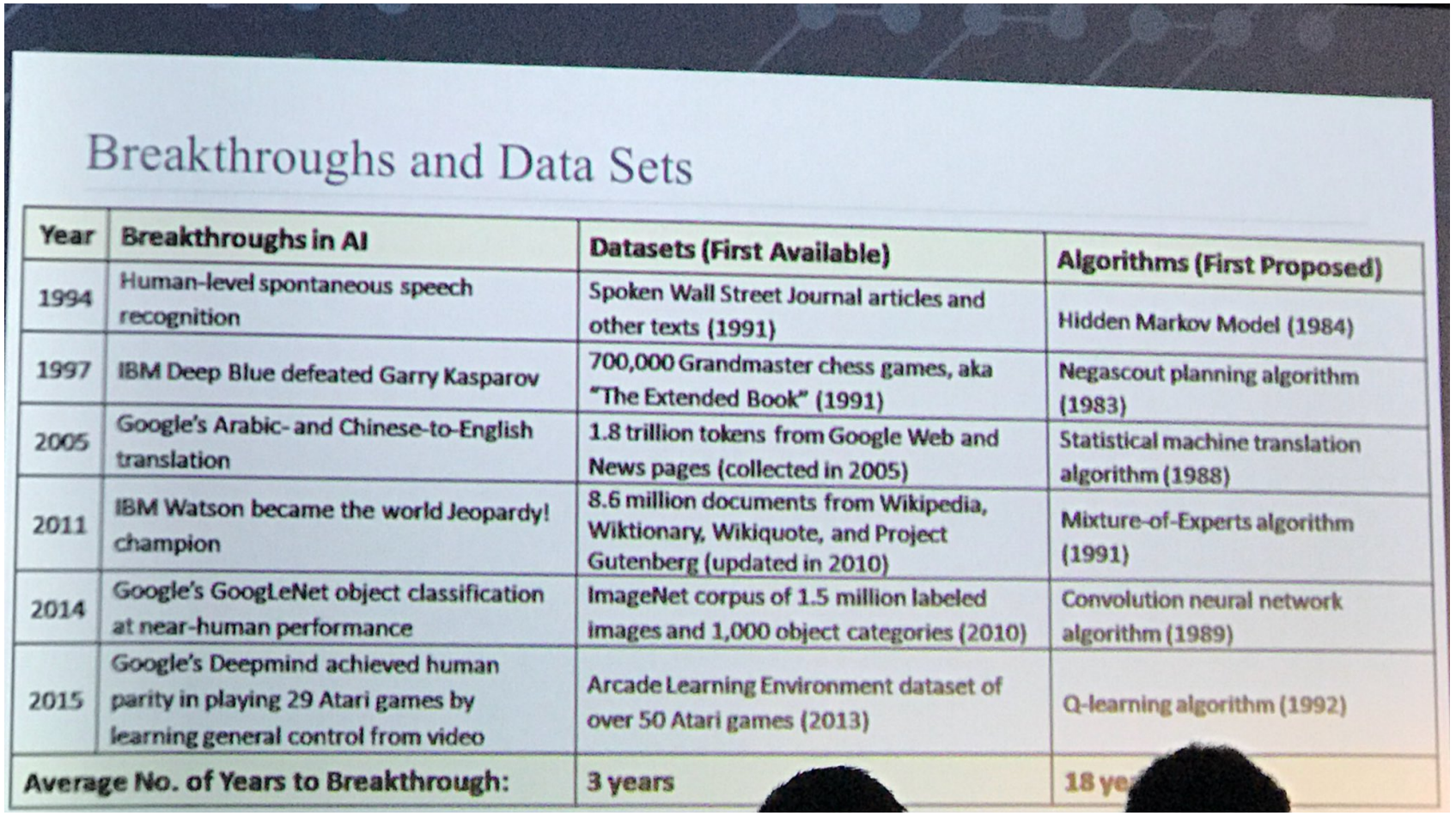
What I think I do

What I actually do

The Unreasonable Effectiveness of Data

"Data is ten times more powerful than algorithms."

-- Peter Norvig, http://youtu.be/yvDCzhbjYWs



Breakthroughs and Data Sets

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		3 years	18 ye

https://twitter.com/shivon/status/864889085697024000

Uses of Data in NLP

* understanding the problem

Uses of Data in NLP

* understanding the problem
* statistical analysis

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set
* training data set

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set
* training data set
* real-time feedback

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set
* training data set
* real-time feedback
* marketing/PR

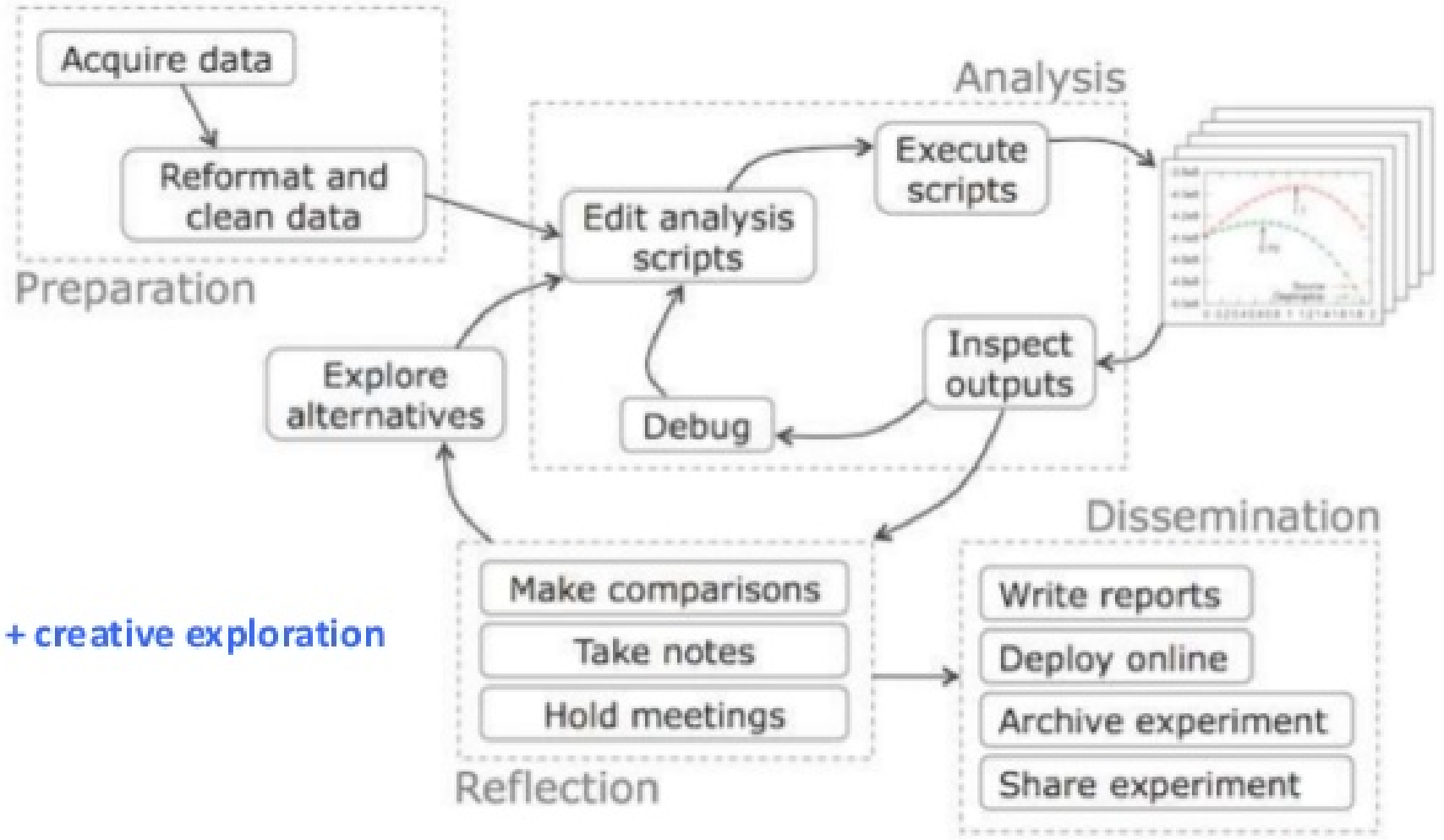
Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set
* training data set
* real-time feedback
* marketing/PR
* external competitions

Uses of Data in NLP

* understanding the problem
* statistical analysis
* connecting separate domains
* evaluation data set
* training data set
* real-time feedback
* marketing/PR
* external competitions
* what else?

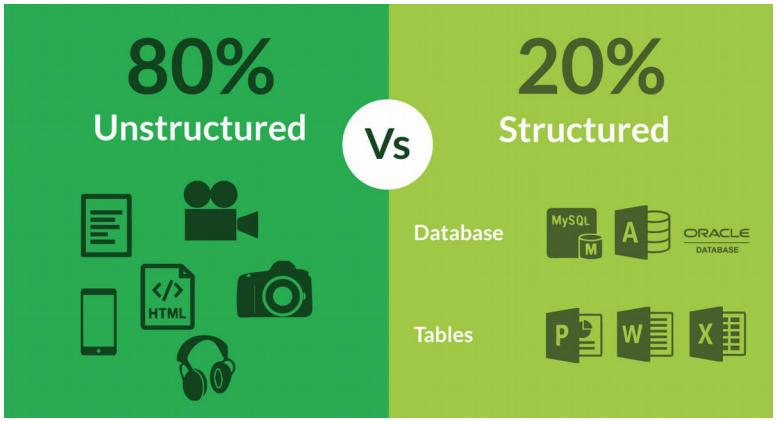
Data Workflow



Source: Josh Wills, Senior Director of Data Science, Cloudera. *"From the Lab to the Factory: Building a Production Machine Learning Infrastructure."*

Types of Data

* Structured
* Semi-structured
* Unstructured

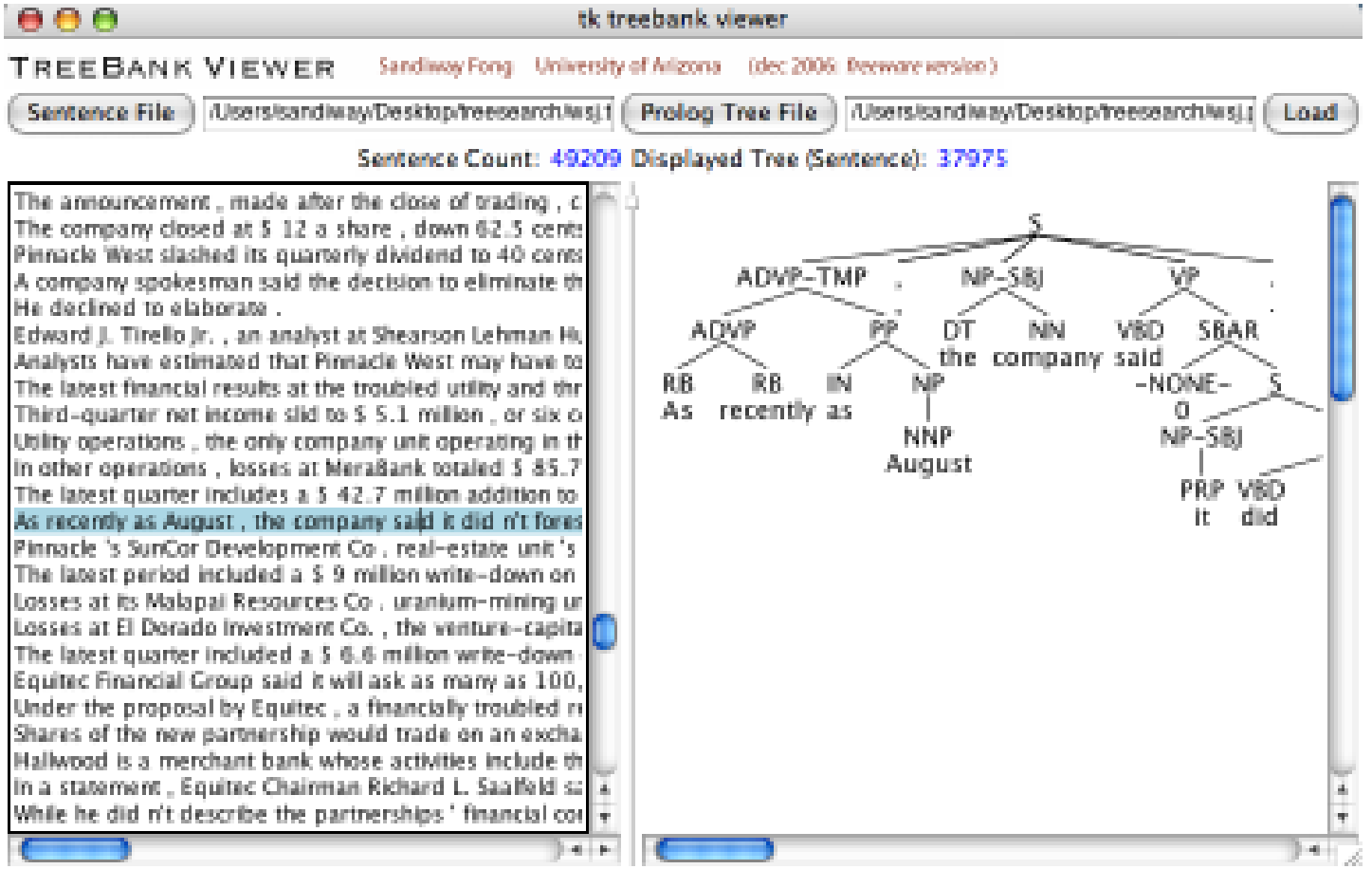


Getting Data

* Get existing
* Create your own
* Acquire from users

Corpus

Annotated collection of docs in a certain format.



Corpus Formats

* Brown
* BSF and friends
* PTB
* Custom XML or JSON
  (also, CSV, etc.)
* Weird/exciting

Brown

The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd Friday/nr an/at investigation/nn of/in Atlanta's/np$ recent/jj primary/nn election/nn produced/vbd ``/`` no/at evidence/nn ''/'' that/cs any/dti irregularities/nns took/vbd place/nn ./.

The/at jury/nn further/rbr said/vbd in/in term-end/nn presentments/nns that/cs the/at City/nn-tl Executive/jj-tl Committee/nn-tl ,/, which/wdt had/hvd over-all/jj charge/nn of/in the/at election/nn ,/, ``/`` deserves/vbz the/at praise/nn and/cc thanks/nns of/in the/at City/nn-tl of/in-tl Atlanta/np-tl ''/'' for/in the/at manner/nn in/in which/wdt the/at election/nn was/bedz conducted/vbn ./.

The/at September-October/np term/nn jury/nn had/hvd been/ben charged/vbn by/in Fulton/np-tl Superior/jj-tl Court/nn-tl Judge/nn-tl Durwood/np Pye/np to/to investigate/vb reports/nns of/in possible/jj ``/`` irregularities/nns ''/'' in/in the/at hard-fought/jj primary/nn which/wdt was/bedz won/vbn by/in Mayor-nominate/nn-tl Ivan/np Allen/np Jr./np ./.

ner-uk corpus (BSF)

```
T1    ОРГ 53 64       Океан Ельзи
T2    ОРГ 137 157     Інституту ім. Глієра
T3    РІЗН 190 195    Ягуар
T4    ОРГ 283 290     Океанів
T5    ПЕРС 292 303    Денис Дудко
T6    ПЕРС 342 358    Олексій Саранчин
T7    ОРГ 416 420     ТНМК
T8    ОРГ 441 457     Інституті музики
T9    ЛОК 767 774     Харкові
T10   ОРГ 928 936     СхідSide
T11   ОРГ 981 985     ТНМК
T12   ПЕРС 1000 1026  Дмитро «Бобін» Александров
T13   ПЕРС 1037 1055  Володимир Шабалтас
T14   ПЕРС 1122 1141  Олександр Лебеденко
T15   ПЕРС 1156 1161  Дудко
T16   ПЕРС 1172 1180  Саранчин
T17   ПЕРС 1275 1280  Дудко
T18   ПЕРС 1335 1354  Давідом Голо
```

SNLI corpus (JSONL+PTB)

{"annotator_labels": ["neutral", "entailment", "neutral", "neutral", "neutral"], "captionID": "4705552913.jpg#2", "gold_label": "neutral", "pairID": "4705552913.jpg#2r1n", "sentence1": "Two women are embracing while holding to go packages.", "sentence1_binary_parse": "((Two women) ((are (embracing (while (holding (to (go packages))))))) .))", "sentence1_parse": "(ROOT (S (NP (CD Two) (NNS women)) (VP (VBP are) (VP (VBG embracing) (SBAR (IN while) (S (NP (VBG holding)) (VP (TO to) (VP (VB go) (NP (NNS packages))))))) (. .)))", "sentence2": "The sisters are hugging goodbye while holding to go packages after just eating lunch.", "sentence2_binary_parse": "((The sisters) ((are ((hugging goodbye) (while (holding (to ((go packages) (after (just (eating lunch)))))))))) .))", "sentence2_parse": "(ROOT (S (NP (DT The) (NNS sisters)) (VP (VBP are) (VP (VBG hugging) (NP (UH goodbye)) (PP (IN while) (S (VP (VBG holding) (S (VP (TO to) (VP (VB go) (NP (NNS packages)) (PP (IN after) (S (ADVP (RB just)) (VP (VBG eating) (NP (NN lunch))))))))))) (. .)))"}

FCE (Custom XML)

```
<?xml version="1.0" encoding="UTF-8"?>
<learner><head sortkey="TR3*0100*2000*02">
  <candidate><personnel><language>Catalan</language><age>16-
20</age></personnel><score>28.00</score></candidate>
  <text>
     <answer1>
       <question_number>1</question_number>
       <exam_score>2.3</exam_score>
         <coded_answer>
          <p>DECEMBER 12TH</p>
          <p>PRINCIPAL MR. ROBERTSON</p>
          <p>DEAR SIR,</p>
          <p>I WANT TO <NS type="S"><i>THAK</i><c>THANK</c></NS> YOU FOR
PREPARING SUCH A GOOD PROGRAMME FOR US AND ESPECIALLY FOR TAKING US <NS
type="RT"><i>TO</i><c>ON</c></NS> THE RIVER TRIP TO GREENWICH. I WOULD LIKE TO
KNOW IF THERE IS ANY CHANCE OF CHANGING THE PROGRAMME BECAUSE WE HAVE FOUND A
VERY INTERESTING ACTIVITY TO DO ON TUESDAY 14 MARCH. IT <NS
type="RV"><i>CONSISTS <NS
type="RT"><i>ON</i><c>IN</c></NS></i><c>INVOLVES</c></NS> VISITING THE LONDON
FASHION AND LEISURE SHOW <NS type="RT"><i>IN</i><c>AT</c></NS> THE CENTRAL
EXHIBITION HALL. I THINK IT'S A GREAT OPPORTUNITY TO MAKE GREATER USE OF OUR
KNOWLEDGE OF <NS type="MD"><c>THE</c></NS> ENGLISH LANGUAGE. <NS type="ID"><i>ON
THE OTHER HAND</i><c>ALSO</c></NS>, WE COULD LEARN THE DIFFERENT WAYS TO GET TO
THE CENTRAL EXHIBITION HALL.</p>
```

UD_Ukrainian (CONLLU)

```
# doc_title = Сад Гетсиманський
# newdoc id = 028g
# newpar id = 02tb
# sent_id = 02to
# text = Дідусь, той що атестував, посміхнувся й спитав:
1   Дідусь дідусь NOUN    Ncmsny  Animacy=Anim|Case=Nom|Gender=Masc|
Number=Sing 7   nsubj   _   Id=02tp|SpaceAfter=No
2   ,   ,   PUNCT   U   _   3   punct   _   Id=02tq
3   той той DET Pd--m-sna   Case=Nom|Gender=Masc|Number=Sing|PronType=Dem   7
dislocated  _   Id=02tr
4   що  що  SCONJ   Css _   5   mark    _   Id=02ts
5   атестував   атестувати    VERB    Vmpis-sm    Aspect=Imp|Gender=Masc|
Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin    3   acl _   Id=02tt|
SpaceAfter=No
6   ,   ,   PUNCT   U   _   5   punct   _   Id=02tu
7   посміхнувся   посміхнутися VERB    Vmeis-sm    Aspect=Perf|Gender=Masc|
Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin    0   root    _   Id=02tv
8   й   й   CCONJ   Ccs _   9   cc  _   Id=02tw
9   спитав  спитати   VERB    Vmeis-sm    Aspect=Perf|Gender=Masc|Mood=Ind|
Number=Sing|Tense=Past|VerbForm=Fin 7   conj    _   Id=02tx|SpaceAfter=No
10  :   :   PUNCT   U   _   7   punct   _   Id=02ty
```

WikEd (wdiff)

- spelling error corrections:

  `You can use rsync to [-donload-] {+download+} the database .`

- grammatical error corrections:

  `There [-is-] {+are+} also [-a-] two computer games based on the movie .`

- sentence rewordings and paraphrases:

  `These anarchists [-argue against-] {+oppose the+} regulation of corporations .`

Paraphrases (Custom)

Sentences file:
<s snum=146> bank of holland , wuhan office , was also officially established just recently . </s>
<s snum=425> in a similar poll made about half a year after the return of hong kong to china , 35.9% called themselves " hongkongnese " , and 18% called themselves chinese . </s>
<s snum=556> experts disclosed at the land reclamation conference held in xiaoshan , zhejiang province that the government hopes to reclaim 1 million hectares of land from the sea along its 18,000 kilometers of coastline within 40 to 50 years . </s>
<s snum=161> at the beginning , teachers of the orphanage accompanied him to school and picked him up , but from the second year , he became a resident student and went back to the orphanage only for weekends . he never missed a class , rain or shine . </s>

Alignment file:
146 1 1 S
146 2 2 S
146 3 3 S
146 4 4 S
146 5 5 S
146 6 6 S
146 7 7 S
146 8 8 S
146 9 9 S

AMRBank (Custom)

```
# AMR release (generated on Mon Jan 27, 2014 at 20:44:26)

# ::id nw.wsj_0001.1 ::date 2012-04-25T16:31:34 ::annotator ISI-AMR-01
::preferred
# ::snt Pierre Vinken , 61 years old , will join the board as a nonexecutive
director Nov. 29 .
# ::save-date Tue Sep 17, 2013 ::file nw_wsj_0001_1.txt
(j / join-01
      :ARG0 (p / person :name (p2 / name :op1 "Pierre" :op2 "Vinken")
            :age (t / temporal-quantity :quant 61
                  :unit (y / year)))
      :ARG1 (b / board
            :ARG1-of (h / have-org-role-91
                  :ARG0 p
                  :ARG2 (d2 / director
                        :mod (e / executive :polarity -))))
      :time (d / date-entity :month 11 :day 29))
```

Prominent Corpora

* National: OANC/MASC, British (non-free)
* LDC(non-free): Penn Treebank, OntoNotes, Web Treebank
* Books: Gutenberg, GoogleBooks
* Corporate: Reuters, Enron
* Research: SNLI, SQuAD
* Multilang: UDeps, Europarl

Corpora Pitfalls

* Requires licensing (often).
* Tied to a domain.
* Annotation quality.
* Requires processing
  of custom formats (often).

DBs & KBs

* Wikimedia
* RDF knowledge bases
* Wordnet
* Custom DBs

Dictionaries

* Wordlists (example: COCA)
* Dictionaries, grammar dicts
  - Wiktionary
* Thesauri

Unstructured Data

* Internet
* CommonCrawl (also, NewsCrawl)
* UMBC, ClueWeb

Raw Text Pros & Cons

+ Can collect stats
  => build LMs, word vectors…
+ Can have a variety of domains
- … but hard to control the distribution of domains
- Web artifacts
- Web noise
- Huge processing effort

Creating Own Data

* Scraping

Creating Own Data

* Scraping
* Annotation

Creating Own Data

* Scraping
* Annotation
* Crowdsourcing

Creating Own Data

* Scraping
* Annotation
* Crowdsourcing
* Getting from users

Creating Own Data

* Scraping
* Annotation
* Crowdsourcing
* Getting from users
* Generating

Data Scraping

* Full-page web-scraping
  (see: readability)
* Custom web-scraping
  (see: scrapy, crawlik)
* Extracting from non-HTML
  Formats (.pdf, .doc…)
* Getting from API
  (see: twitter, webhose)

Scraping Pros & Cons

+ possible to get needed domain
+ may be the only “fast” way
+ it’s the “google” way
- licensing issues
- getting volume requires time
- need to deal with antirobot protection
- other engineering challenges

Corpus Annotation

* Initial data gathering
* Annotation guidelines
* Annotator selection
* Annotation tools
* Processing of raw annotations
* Result analysis & feedback

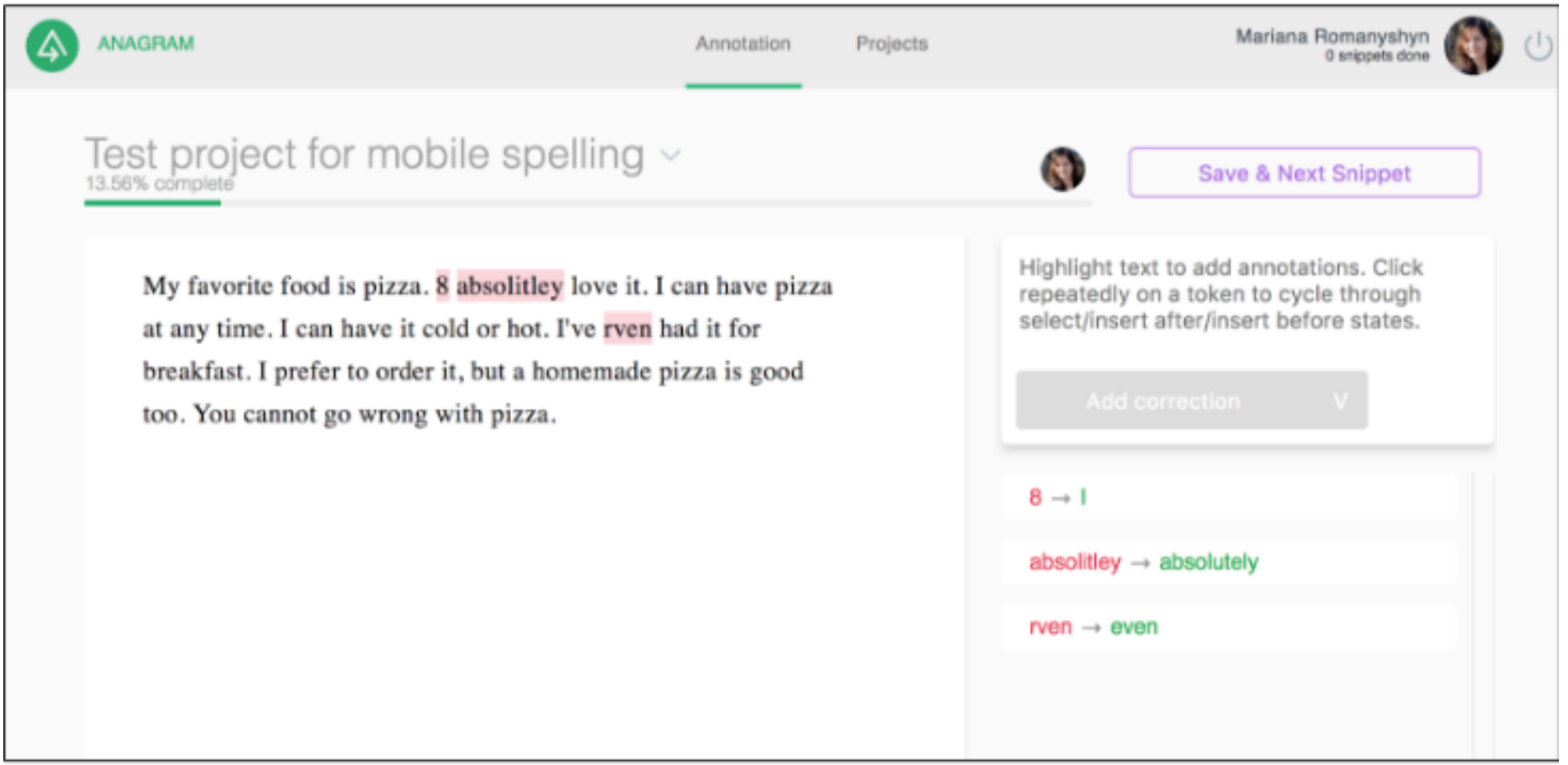
Annotation Variants

* professional
* mturk
* volunteer

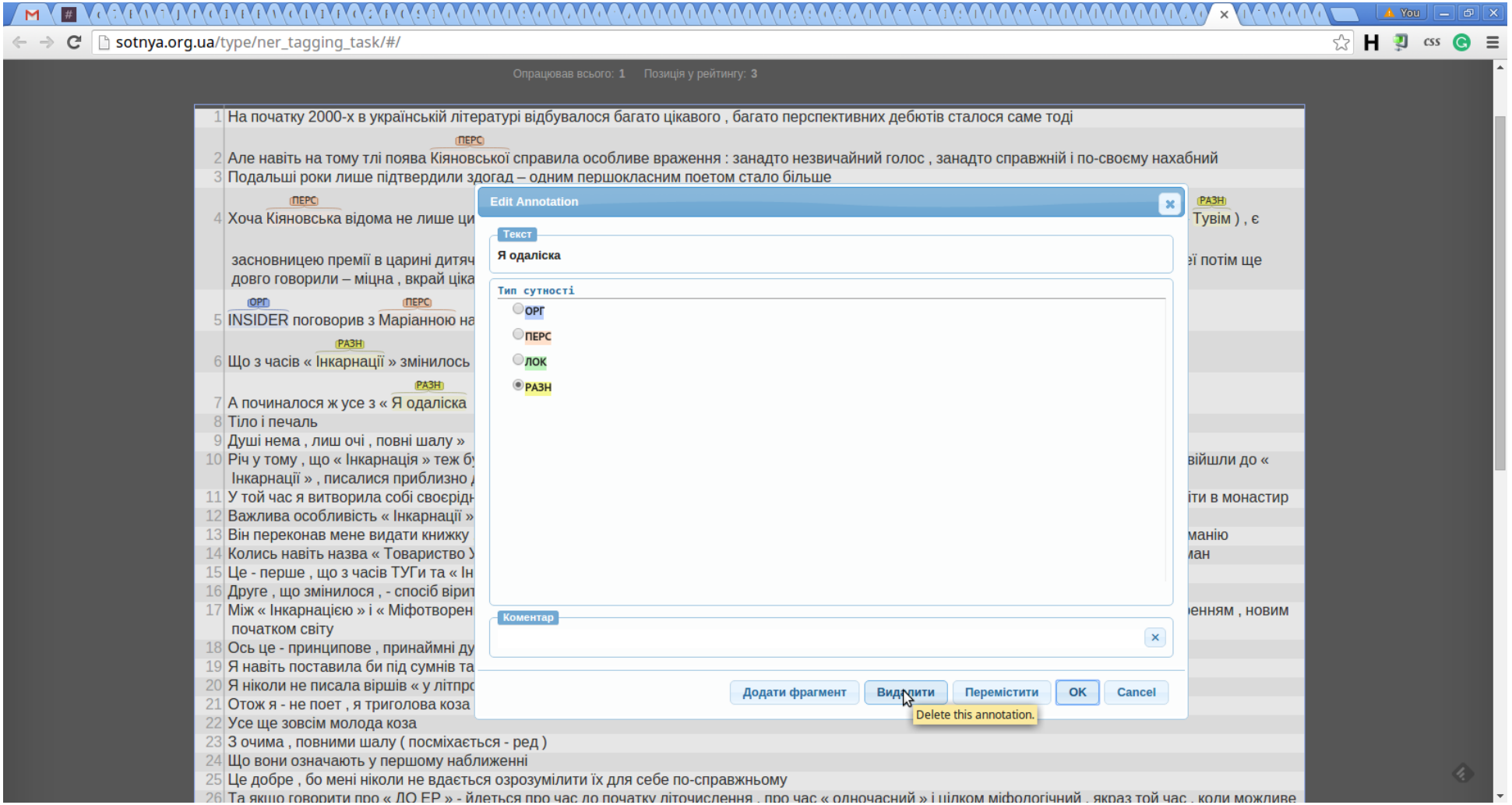
Annotation Tools

* GATE
* Brat
* Anaphora
* Prodigy

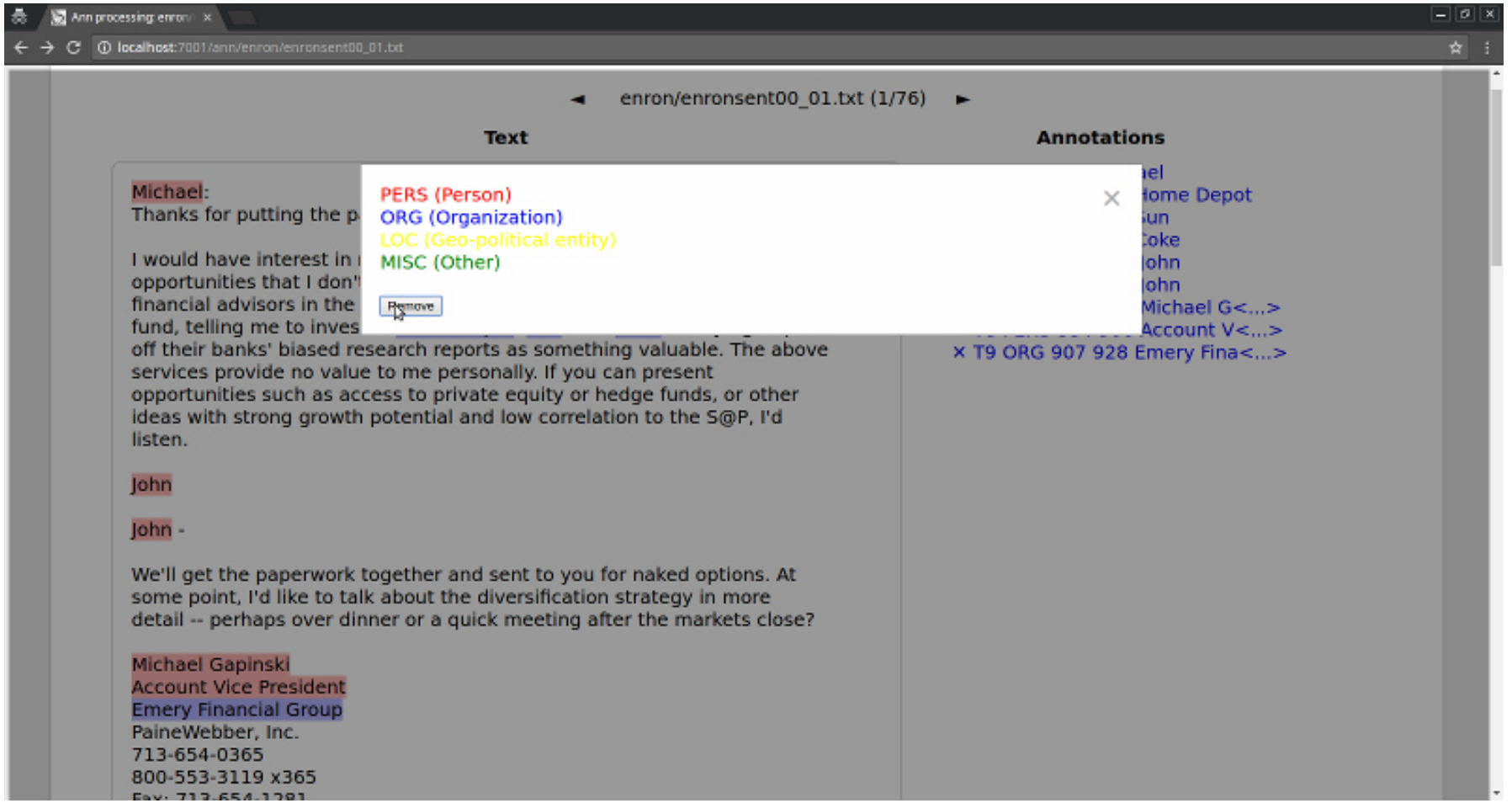
Annotation Tools
“Anagram”



Annotation Tools “Vulyk”

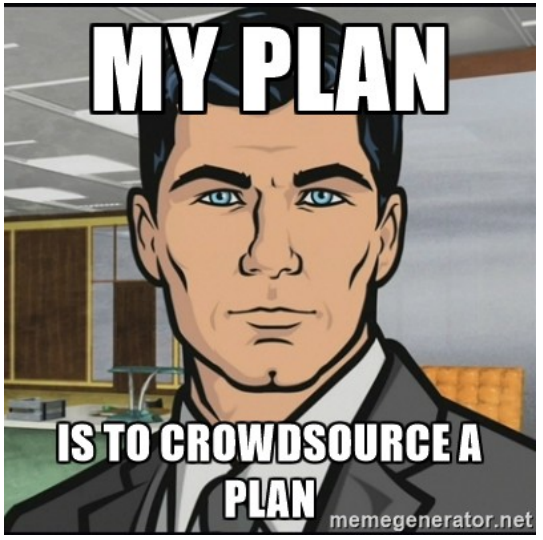


Annotation Tools “Ann”



Crowdsourcing

- Like annotation but harder
  requires a custom solution
  With gamification
+ If successful, can get
  volumes and diversity



User Feedback Data

+ Your product, your domain
+ Real-time, allows adaptation
+ Ties into customer support
- Chicken & egg problem
- Requires anonymization

Approach: “lean startup”

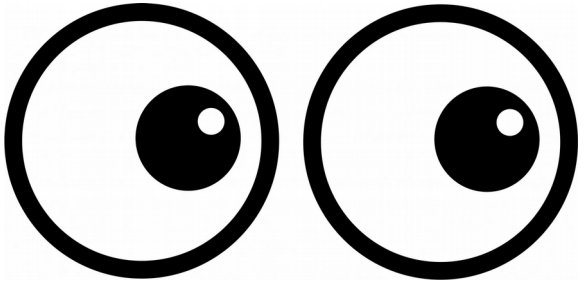
Generating Data

+ Potentially unlimited volume
+ Control of the parameters
- Artificial (if you can code a generation function, why not code a reverse one?)

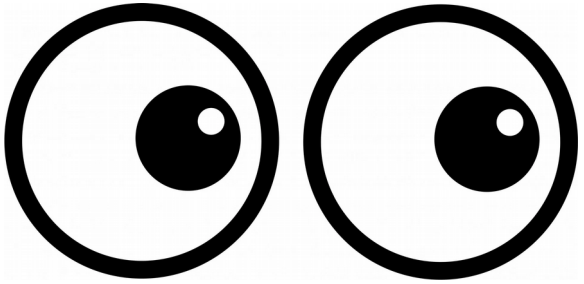
Data Best Practices

* Proper ML dataset handling
* Domain adequacy, diversity
* Inter-annotator agreement, reasonable baselines
* Error analysis
* Real-time tracking

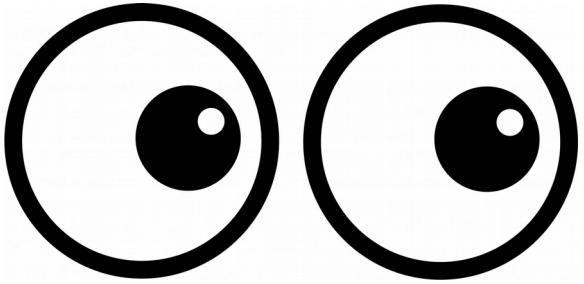
Data Tools

*  (grep & co)

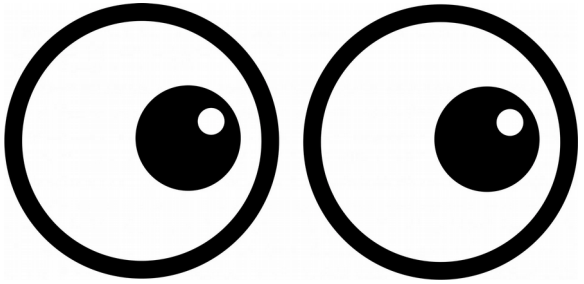
Data Tools

*  (grep & co)
* other Shell powertools

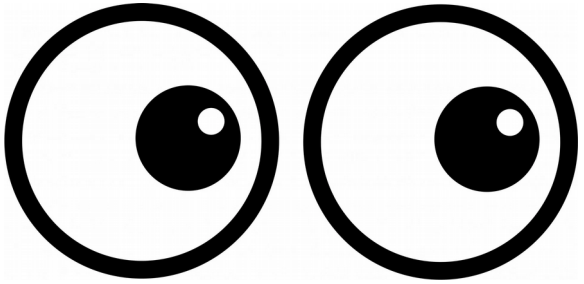
Data Tools

*  (grep & co)
* other Shell powertools
* statistical analysis tools
  + plotting

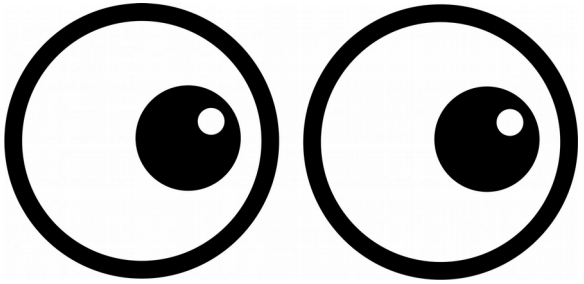
Data Tools

*  (grep & co)
* other Shell powertools
* statistical analysis tools
  + plotting
* annotation tools

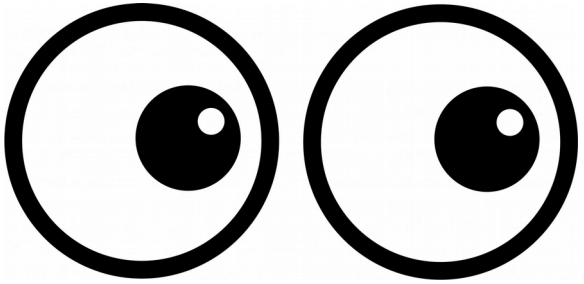
Data Tools

*  (grep & co)
* other Shell powertools
* statistical analysis tools
  + plotting
* annotation tools
* web-scraping tools

Data Tools

*  (grep & co)
* other Shell powertools
* statistical analysis tools
  + plotting
* annotation tools
* web-scraping tools
* metrics databases (Graphite)

Data Tools

*  (grep & co)
* other Shell powertools
* statistical analysis tools
  + plotting
* annotation tools
* web-scraping tools
* metrics databases (Graphite)
* Hadoop, Spark

Questions?

See also:
* Mariana's experiences: https://goo.gl/K7YciJ
* NLTK Book - Accessing Text Corpora and Lexical Resources: http://www.nltk.org/book/ch02.html
* Bias in SNLI: http://www.aclweb.org/anthology/W17-1609