

Data

Vsevolod Dyomkin
Mariana Romanyshyn



Contents

1. Role of Data
2. Types of Data
3. Getting Data
4. Creating Data
5. Real-World Cases, Pitfalls

Data Scientist



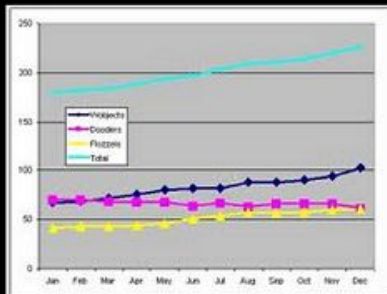
What my friends think I do



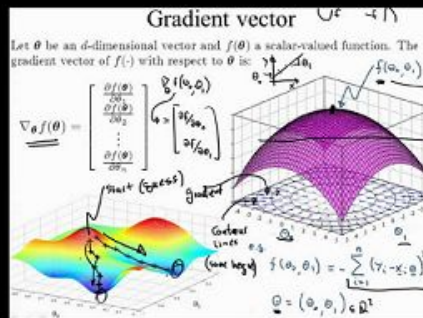
What my mom thinks I do



What society thinks I do



What my boss thinks I do



What I think I do



What I actually do

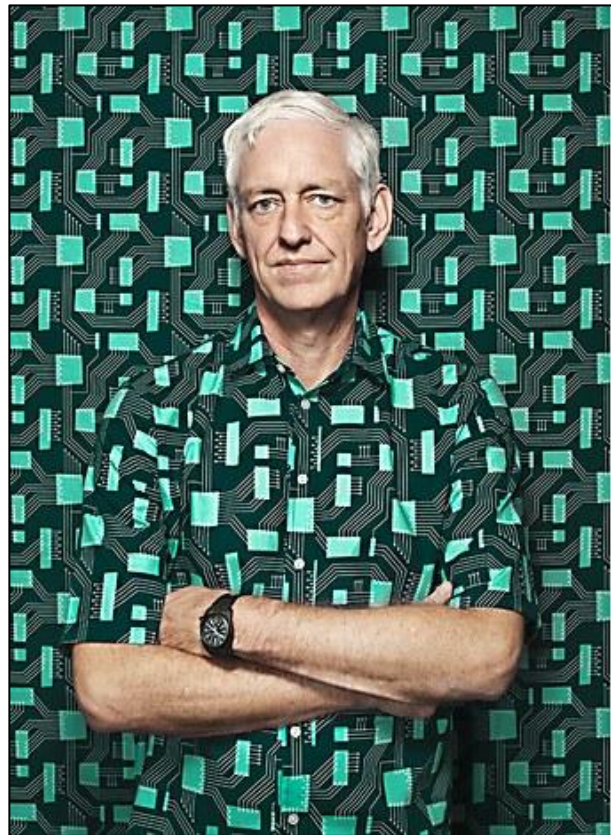
Motivation

“Data is ten times more powerful than algorithms.”

— Peter Norvig

The Unreasonable Effectiveness of Data

<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 years |

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

Role of Data

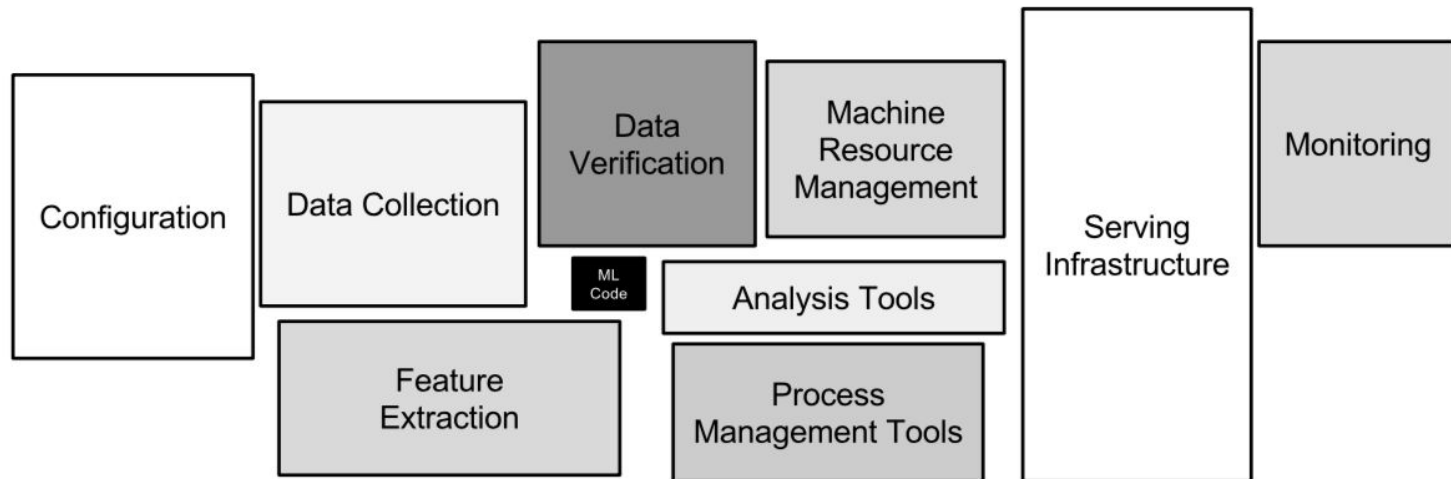


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

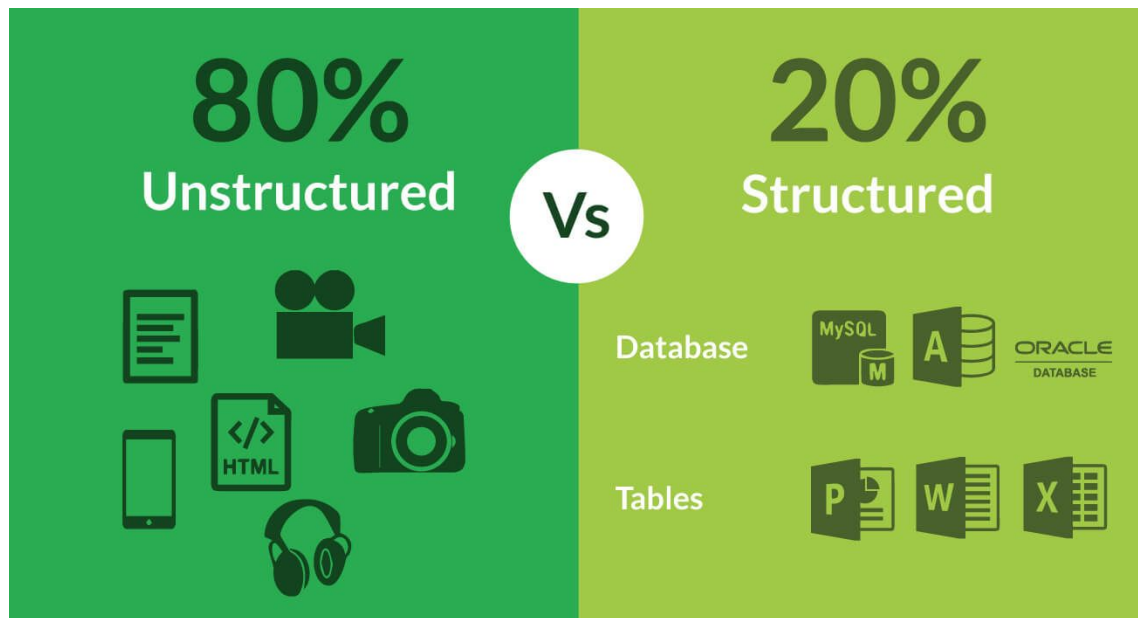
https://medium.com/@neal_lathia/five-lessons-from-building-machine-learning-systems-d703162846ad

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

Types of Data

- Structured
- Semi-structured
- Unstructured



Existing Data Sources

- Annotated corpora
- DBs & KBs
- Dictionaries, lexicons, thesauri
- Raw texts

Corpora

tk treebank viewer

TREEBANK VIEWER Sandiway Fong University of Arizona (dec 2006: beta version)

Sentence File: /Users/sandiway/Desktop/treesearch/iesj1 Prolog Tree File: /Users/sandiway/Desktop/treesearch/iesj1 Load

Sentence Count: 49209 Displayed Tree (Sentence): 37975

The announcement, made after the close of trading, c
The company closed at \$ 12 a share, down 62.5 cents
Pinnacle West slashed its quarterly dividend to 40 cents
A company spokesman said the decision to eliminate th
He declined to elaborate.
Edward J. Tirello Jr., an analyst at Shearson Lehman H
Analysts have estimated that Pinnacle West may have to
The latest financial results at the troubled utility and thr
Third-quarter net income slid to \$ 5.1 million, or six o
Utility operations, the only company unit operating in it
in other operations, losses at MeraBank totaled \$ 85.7
The latest quarter includes a \$ 42.7 million addition to
As recently as August, the company said it did n't forese
Pinnacle 's SunCor Development Co., real-estate unit 's
The latest period included a \$ 9 million write-down on
Losses at its Malapai Resources Co., uranium-mining ur
Losses at El Dorado Investment Co., the venture-capita
The latest quarter included a \$ 6.6 million write-down.
Equitec Financial Group said it will ask as many as 100,
Under the proposal by Equitec, a financially troubled ri
Shares of the new partnership would trade on an excha
Hallwood is a merchant bank whose activities include th
In a statement, Equitec Chairman Richard L. Saffold sa
While he did n't describe the partnerships ' financial cor

ADVP-TMP NP-SBJ VP
ADVP PP DT NN VBD SBAR
RB RB IN NP
As recently as NNP
August
the company said
-NONE-
S
NP-SBJ
PRP VBD
It did

Corpus

- Structured collection of documents
- Usually, with some annotation

Corpora by Size

- Small **~10k-10M** tokens
 - manually annotated for specific tasks: Brown, OntoNotes
- Big **~1G** tokens
 - automatically annotated: GigaWord
- Huge **>100G** tokens
 - not annotated, but may be cleaned up: WebText-2, Stories

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

Corpus Formats

- Simple formats: Brown, BSF, ...
- Linguistics specific: PTB, CONNL, ...
- Custom XML or JSON (also, CSV, etc.)
- Weird/exciting 🤪

Brown Corpus

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 ./.

Brat Standalone Format (ner-uk corpus)

| | | |
|-----|----------------|----------------------------|
| 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 |

PTB+JSONL (SNLI corpus)

```
{"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))))))))))))) (. .)))")
```

XML (FCE corpus)

```
<?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>28.0
0</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>
```

CONNLU (UD_Ukrainian corpus)

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=Animl | Case=Noml | Gender=Mascl | Number=Sing | 7 | nsubj | _ | Id=02tp SpaceAfter=No | | |
| 2 | , | , | PUNCT | U | _ | | | | 3 | punct | _ | Id=02tq | | |
| 3 | той | той | DET | Pd--m-sna | | Case=Noml | Gender=Mascl | Number=Singl | PronType=Dem | 7 | dislocated | _ | Id=02tr | |
| 4 | що | що | SCONJ | Css | _ | | | | 5 | mark | _ | Id=02ts | | |
| 5 | атестував | атестувати | VERB | Vmpis-sm | | Aspect=Impl | Gender=Mascl | Mood=Indl | Number=Singl | Tense=Pastl | VerbForm=Fin | 3 | | |
| | acl | _ | | | | Id=02ttl | SpaceAfter=No | | | | | | | |
| 6 | , | , | PUNCT | U | _ | | | | 5 | punct | _ | Id=02tu | | |
| 7 | посміхнувся | посміхнутися | VERB | Vmeis-sm | | Aspect=Perfl | Gender=Mascl | Mood=Indl | Number=Singl | Tense=Pastl | VerbForm=Fin | 0 | | |
| | root | _ | | | | Id=02tv | | | | | | | | |
| 8 | й | й | CCONJ | Ccs | _ | | | | 9 | cc | _ | Id=02tw | | |
| 9 | спитав | спитати | VERB | Vmeis-sm | | Aspect=Perfl | Gender=Mascl | Mood=Indl | Number=Singl | Tense=Pastl | VerbForm=Fin | 7 | conj | _ |
| | | | | | | Id=02txl | SpaceAfter=No | | | | | | | |
| 10 | : | : | PUNCT | U | _ | | | | 7 | punct | _ | Id=02ty | | |

wdiff (WikEd corpus)

- ▶ 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 .

Custom format similar to PTB (AMRBank corpus)

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))

Ad-hoc format (Paraphrases corpus)

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

Corpus Processing Example: NPS Chats

```
<Post class="Emotion" user="10-19-30sUser2">  
  10-19-30sUser11 lol  
  <terminals>  
    <t pos="NNP" word="10-19-30sUser11"/>  
    <t pos="UH" word="lol"/>  
  </terminals>  
</Post>
```

<http://lisp-univ-etc.blogspot.com/2013/06/nltk-21-working-with-text-corpora.html>

SAX Parsing

```
(defmethod read-corpus-file ((type (eql :nps-chat)) source)
  (cxml:parse source (make 'nps-chat-sax)))
```

```
(defclass nps-chat-sax (sax:sax-parser-mixin)
  ((texts :initform nil)
   (tokens :initform nil)
   (classes :initform nil)
   (users :initform nil)
   (cur-tag :initform nil)
   (cur-tokens :initform nil)))
```

```
(defmethod sax:start-element ((sax nps-chat-sax) namespace-uri local-name qname attributes)
  (with-slots (classes users cur-tokens cur-tag) sax
    (case cur-tag
      (:post (push (attr "class" attributes) classes)
              (push (attr "user" attributes) users))
      (:t (push (make-token
                  :word (attr "word" attributes)
                  :tag (attr "pos" attributes))
                cur-tokens))))))
```

```
(defmethod sax:characters ((sax nps-chat-sax) data)
  (with-slots (cur-tag texts) sax
    (when (eql :terminals cur-tag)
      (push data texts))))
```

```
(defmethod sax:end-element ((sax nps-chat-sax) namespace-uri local-name qname)
  (when (eql :terminals (mkeyw local-name))
    (with-slots (tokens cur-tokens) sax
      (push (reverse cur-tokens) tokens)
      (setf cur-tokens nil))))
```

```
(defmethod sax:end-document ((sax nps-chat-sax))
  (with-slots (texts tokens users classes) sax
    (values (reverse texts)
            (reverse tokens)
            (reverse classes)
            (reverse users))))
```

Corpora Pitfalls

- Tied to a domain
- Annotation quality

Technical:

- Require licensing
- Require processing of custom formats

Data Licensing

From data owners: universities/companies/individuals

Issues:

- data owners have no idea about cost and/or license
- legislation is different in different countries
- be ready to spend about 3 months
- and sometimes...



Sometimes you win, sometimes you learn

- The Story of a Missing Licence from Creators
- The Story of a Lost Electronic Copy
- The Story of a Never-Ending Divorce
- The Story of the Corpus of Ukrainian



Structured Data

Dictionaries

- Wordlists, lexicons
- Dictionaries
- Wiktionary
- Thesauri

DBs & KBs

- Wikimedia (DBPedia, Wikidb)
- RDF knowledge bases (Freebase, OpenCYC)
- KBPedia
- WordNet, ConceptNet, BabelNet
- Private or public data sources (*gov)

KBpedia Use Cases

<http://kbpedia.org/use-cases>



Knowledge Graph (KG) Use Cases

- [Browse the Knowledge Graph](#)
- [Search the Knowledge Graph](#)
- [Expand Queries Using Semsets](#)
- [Uses and Control of Inferencing](#)
- [Leverage KBpedia's Aspects](#)

Machine Learning (KBAI) Use Cases

- [Create Supervised Learning Training Sets](#)
- [Create Word Embedding Corporuses](#)
- [Create Graph Embedding Corporuses](#)
- [Classify Text](#)
- [Create 'Gold Standards' for Tuning Learners](#)
- [Disambiguate KG Concepts](#)
- [Dynamic Machine Learning Using the KG](#)

Mapping Use Cases

- [Map Concepts](#)
- [Map Entities](#)
- [Extend KBpedia for Domains](#)
- [General Use of the Mapper](#)

Creating Your Own Data ヽ_(ツ)_/

Ways to Create Data

- Scraping
- Annotation
- Crowdsourcing
- Generating
- Getting from users

Sources of Raw Data

- Internet
- CommonCrawl (also, NewsCrawl)
- UMBC, ClueWeb, WikiText
- Wikipedia
- Social media: Reddit, Twitter

Raw Data 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/social media noise
- Huge processing effort
- Rate limits

More Specific Sources

- Media websites
- Libraries
- Registries & online DBs
- Thematic forums
- Specific APIs (Twitter, Wordnik, NewsAPI/Webhose, ...)
- Custom search API

Use Case: getting tweets about #lisp

- Getting access to twitter API nowadays is a huge PITA
- Using twitter/search API to retrieve all the tweets with the word “lisp”
- Mind the API limitations (limit of tweets, limit on date)
- Overlap in the API calls
- Retweets are a headache
- Using <https://publish.twitter.com/oembed?url=>



A close-up of Morpheus from the movie The Matrix, wearing his signature black sunglasses. The reflection in the lenses shows three figures in a dimly lit room. The background is a blurred, industrial-looking environment.

WHAT IF I TOLD YOU

**THERE'S AN API FOR EVERY
WEBSITE**

Web Scraping Rules of Thumb

- Creativity FTW
- Readability FTW
- But respect copyright!
- Don't overload websites, respect robots.txt
- AWS can also be FTW

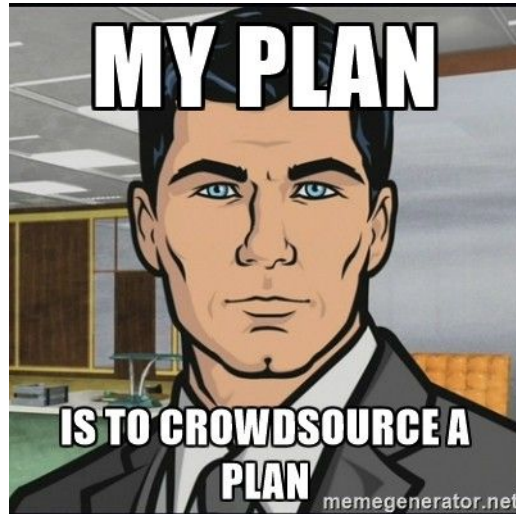
Use Case: debates corpus

- Scraping Debatabase

<http://idebate.org/debatabase/index.php>

- Using crawlik

Annotation vs Crowdsourcing



Data Annotation: who?

- Own annotators
- Volunteers
- Crowdsourcing platforms
 - *Amazon Mechanical Turk, kotyky.org.ua*
- Expert linguists
 - *Appen, Leapforce, iSoftStone*

Data Annotation: who?

Crowdsourcing

- + cheap and fast
- little control over quality

Expert Linguists

- expensive and time-consuming
- + easier to control quality

Data Annotation: who?

Crowdsourcing

- + cheap and fast
- little control over quality

Expert Linguists

- expensive and time-consuming
- + easier to control quality

There will always be *errors* in your data!

Crowdsourcing: Volunteers

- Friends
- Co-workers, fellow students
- Internet volunteers
- Paid volunteers :)

Use Case: ner-uk corpus

- 264 texts
- 238,927 tokens
- 6,751 NER entities
- Vulyk
- 3 volunteer annotators
- 1 volunteer editor (Seva)
- ~20k HRN
- 1-2 months

Use Case: ner-uk corpus

Опрацював всього: 1 Позиція у рейтингу: 3

1 На початку 2000-х в українській літературі відбувалося багато цікавого, багато перспективних дебютів сталося саме тоді

2 Але навіть на тому тлі поява Кіановської справила особливе враження: занадто незвичайний голос, занадто справжній і по-своєму нахабний

3 Подальші роки лише підтвердили здогад – одним першокласним поетом стало більше

4 Хоча Кіановська відома не лише цим: вона одна з найавторитетніших в Україні перекладачів з польської (особливо – поезія, особливо – Тувім), є засновницею премії в царині дитячої літератури «Великий Іжак», а в 2008-му видала книгу оповідань «Стежка вздовж ріки», що про неї потім ще довго говорили – міцна, вкрай цікава річ

5 INSIDER поговорив з Маріанною напередодні Нового року, коли одночасно вийшли друком дві її поетичні книги

6 Що з часів «Інкарнації» змінилось принципово

7 А починалося ж усе з «Я одаліска

8 Тіло і печаль

9 Душі нема, лиш очі, повні шалу»

10 Річ у тому, що «Інкарнація» теж була збіркою вибраних віршів: приблизно з 1989 року, може, з 1988 року - і до 1997 року, тексти, що ввійшли до «Інкарнації», писалися приблизно дев'ять років

11 У той час я втворювала собі своєрідну «внутрішню міфологію» умовної «Магдалини», оскільки після школи хотіла, але не наважилася піти в монастир

12 Важлива особливість «Інкарнації» - її упорядковувала не я, а Юрко Бедрик

13 Він переконав мене видати книжку, бо я, пишучи у той час по кільканадцять віршів за ніч, не сумнівалася, що продукую жорстоку графоманію

14 Колись навіть назва «Товариство Усамітнених Графоманів» виникла значною мірою через мене, бо я всім постійно казала, що - графоман

15 Це - перше, що з часів ТУГи та «Інкарнації» змінилося принципово

16 Друге, що змінилося, - спосіб вірити, сама структура віри

17 Між «Інкарнацією» і «Міфотворенням» я відмовила важку невиліковну хворобу, для мене це було абсолютним метафізичним перетворенням, новим

<https://github.com/lang-uk/ner-uk/tree/master/doc>

Crowdsourcing: Amazon Mechanical Turk

- mturk.com - a platform for work that requires human intelligence
- Requesters vs. Workers
- Tasks are organized in HITs (human intelligence tasks)
- Provides a sandbox: requestersandbox.mturk.com



[Home](#)[Create](#)[Manage](#)[Developer](#)[Help](#)[New Project](#)[New Batch with an Existing Project](#)[Create HITs individually](#)

Start a New Project

Categorization

[Data Collection](#)[Moderation of an Image](#)[Sentiment](#)[Survey](#)[Survey Link](#)[Tagging of an Image](#)[Transcription from A/V](#)[Transcription from an Image](#)[Writing](#)[Other](#)

Example of Categorization

Choose the best category for this image



- ☐ kitchen
- ☐ living
- ☐ bath
- ☐ bed
- ☐ outside

[View Instructions](#) ↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

You must ACCEPT the HIT before you can submit the results.

[Create Project »](#)

AMT Prices

1. The Worker reward (\$0.01 minimum)
2. The AMT fee (20–40%)
3. Additional 5% if you want your Workers to be Masters
4. Extra per HIT if you choose a predefined qualification (e.g., age, gender, native language)

Expert Linguists: Appen

- appen.com - development of high-quality, human annotated datasets for ML
- 180 languages
- 1 mln annotators



Appen

@AppenGlobal

 **Follow**



Machine learning without data is like a rocket without fuel. [#DataWest17](#)

Use case: mobile spelling corrections

- What we need?
 - Spelling error annotations in mobile phone messages.



Use case: mobile spelling corrections

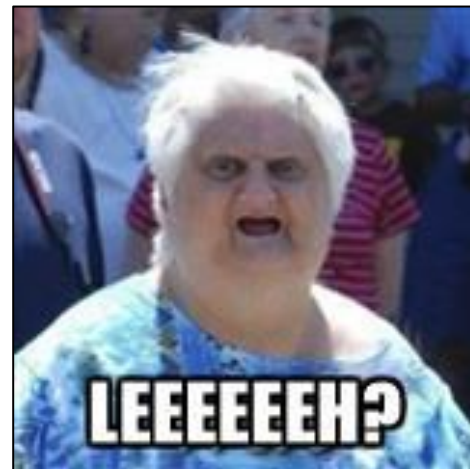
- What's available?
 - *NUS SMS Corpus*
 - 55,000 messages
 - *Mobile Forensics corpus*
 - 4,934 messages
 - *The Enron Mobile Email Dataset*
 - 2,600 messages
 - *SMS Spam Collection v. 1*
 - 425 spam messages + NUS SMS

NUS SMS Corpus: Singlish

*Waiting in a car 4 my mum **lor**. U **leh**? Reach home already?*

Lor - expresses general agreeability

Leh - expresses negativity



The data says we need more
data.



som^{ee}cards
user card

Attack plan

- Collect data
 - Scrape Twitter
 - Use Amazon Mechanical Turk
- Annotate data
 - Automatic annotation
 - Annotation with expert linguists



Twitter

- Twython
- A few thousand tweets from 2011–2013
- “source” in [“iPhone”, “Android”, “Mobile”]
- Data quality:
 - language filter
 - profanities
 - too short or just hashtags
 - average word length < 3 , etc.

AMT data collection: idea 1

- What if we ask the turkers to **retype** some short messages?
 - How to set up AMT on the phone?
 - What messages to retype?
 - How do we know...
 - they are not copy-pasting?
 - they are not typing some other text instead?
 - they are using a mobile phone?
 - they are not using autocorrect?

Results

- 10,000 sentences
- 2 days
- \$0.05 per HIT
- 33,000 misspellings

Instructions (Click to expand)

Important: You must use your smartphone to complete this task. Open this task from a browser on your smartphone using the following link: www.goo.gl/ShortLink. Type the answers using your mobile keyboard. Turn off your spell checker and autocorrect for this task. Submissions which do not use a mobile device will be rejected.

Short link to task for smartphones: rebrand.ly/d7eb

If you need help turning off the spell checker, use the instructions below:

- for Android: <http://www.wikihow.com/Turn-Off-Auto-Correct-on-an-Android>
- for iOS: <http://www.howtoisolve.com/how-to-turn-off-spell-check-on-iphone-6-6-plus-ios-8-1/>

In this task, you'll be presented with 5 sentences and asked to retype the sentences as quickly as you can. Do not worry about any errors in your writing.

You will need to do the following:

- Use a mobile keyboard on your smartphone to perform the task
- Disable spellcheck / autocorrect on your phone
- Type as quickly as you can
- Do **not** go back to correct any spelling errors

Example

Pack my box with five dozen liquor jugs.

Pack my box with five dozen liquor jugs.

*Pack my box with five dozen **liquour** jugs.*

*Pack my box with five dozen **liquir** jugs.*

***Paxk** my box with **guve** dozen **liquorr** jugs.*

Pack my box with five dozen liquor jugs.



AMT data collection: idea 2

- What if we ask the turkers to ***give short answers?***
 - How to set up AMT on the phone?
 - What questions to ask?
 - How do we know...
 - they are not copy-pasting random text?
 - they are using a mobile phone?
 - they are not using autocorrect?

Results

- 2,000 answers to 200 questions
- 4 days
- \$0.15 per HIT

Issues:

- Misspellings cannot be extracted
- Some data bias

Bias

Saree. Attried in saree looks gorgeous. Its neet beautiful and sexy dress.

My favourite place is Guruvayur temple. I love Guruvayurappan and i feel relaxed there.

I live in mumbai, maharashtra, india. In mumbai there are many spots where we can enjoy...



Annotation Tools

Classic:

- GATE
- INCEpTION

The screenshot displays the GATE Developer 7.0 build 4195 interface. The left sidebar shows the project structure under 'Language Resources', with 'romaniuk_01_MYR...' selected. The main window is divided into several panels. The 'Messages' panel at the top shows the document 'romaniuk_01_MYR...'. Below it, the 'Annotations List' panel is active, displaying a list of annotations. A red arrow points to the 'Annotations List' tab. The text in the document is: 'MYR 07.03.2011, 13:14 А я хочу написати позитивний відгук про кафе-бар "Купол". Для мене там супер все - і кухня, і обслуговування, і інтер'єр. Страви надзвичайні, спокійна обстановка і можна зустріти якусь знаменитість. Минулого разу, як ми там були, там обідав ведучий новин з каналу 1+1. Останні раз обідали там в п'ятницю, залишилися задоволені.' The 'Annotations List' table shows the following data:

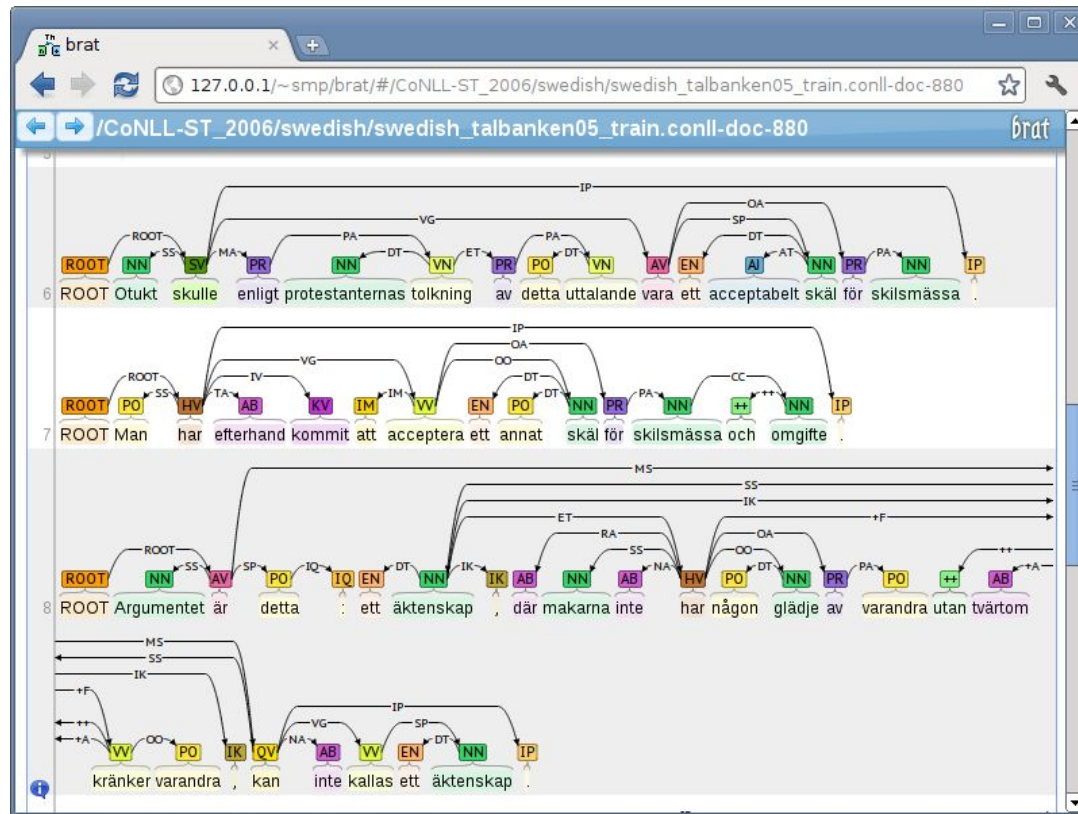
| Type | Set | Start | End | Id | Features |
|--------------|-----|-------|-----|----|-----------------|
| person | | 0 | 4 | 5 | {kind=nickname} |
| time | | 7 | 17 | 3 | {kind=date} |
| time | | 19 | 25 | 4 | {kind=time} |
| organization | | 78 | 83 | 6 | {} |
| person | | 269 | 282 | 7 | {kind=position} |
| organization | | 285 | 295 | 8 | {} |
| time | | 323 | 331 | 9 | {kind=date} |

Below the table, it says '7 Annotations (1 selected) Select:'. The 'Document Editor' panel at the bottom shows 'Initialisation Parameters'. On the right, the 'Original markups' panel shows a list of checked items: 'organization', 'person', and 'time'.

Annotation Tools

Modern:

- Brat
- Anafora



Annotation Tools

Modern: Brat, Anafora

The screenshot shows the Vulyk crowdsourcing platform interface. At the top, there is a logo and the text "Vulyk: crowdsourcing platform". Navigation links include "Інструкція", "Завдання" (underlined), "Топ-10 волонтерів", and a "Logout" button. A progress bar indicates "Опрацював всього: 4" and "Позиція у рейтингу: 2". The main text area contains a sentence: "Завдяки унікальним умовам ґрунту і рівню рН у воді, яка просочується у місцевий ґрунт із річки Євфрат, квіти набувають такого з ґрунту". A dropdown menu is open over the word "ґрунту", showing three suggestions: "ґрунт noun m v_dav inanim", "ґрунт noun m v_mis inanim", and "ґрунт noun m v_rod inanim". Below the suggestions, it says "Коректний варіант відсутній". On the right, there are buttons for "Пропустити завдання" and "Надіслати форму".

Vulyk: crowdsourcing platform

Інструкція Завдання Топ-10 волонтерів Logout

Опрацював всього: 4 Позиція у рейтингу: 2

Завдяки унікальним умовам ґрунту і рівню рН у воді, яка просочується у місцевий ґрунт із річки Євфрат, квіти набувають такого з ґрунту

ґрунт noun m v_dav inanim
ґрунт noun m v_mis inanim
ґрунт noun m v_rod inanim

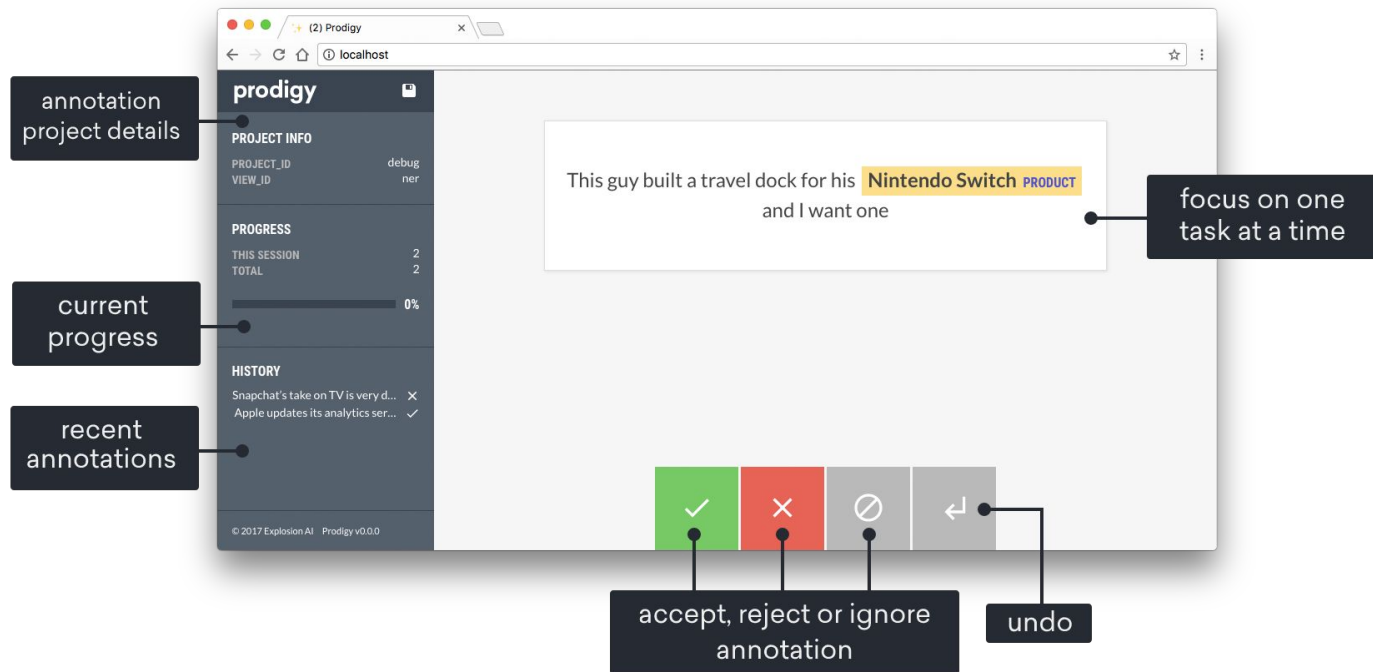
Коректний варіант відсутній

Пропустити завдання Надіслати форму

Annotation Tools

“AI-Powered”:

- Prodigy
- Snorkel



Annotation Tools

Custom :)

The screenshot displays the ANAGRAM web application interface. At the top, there is a header bar with the ANAGRAM logo on the left, navigation tabs for 'Annotation' and 'Projects' in the center, and a user profile section on the right for 'Mariana Romanyshyn' with '0 snippets done' and a power button icon.

Below the header, the main content area shows a project titled 'Test project for mobile spelling' with a dropdown arrow and a progress indicator '13.56% complete'. To the right of the title is a user profile picture and a 'Save & Next Snippet' button.

The central part of the interface features a text editor with the following text: 'My favorite food is pizza. 8 absolutley love it. I can have pizza at any time. I can have it cold or hot. I've rven had it for breakfast. I prefer to order it, but a homemade pizza is good too. You cannot go wrong with pizza.' The words '8', 'absolutley', and 'rven' are highlighted in red.

On the right side, there is a sidebar with instructions: 'Highlight text to add annotations. Click repeatedly on a token to cycle through select/insert after/insert before states.' Below the instructions is a button labeled 'Add correction' with a dropdown arrow. Further down, there are three rows of correction suggestions:

- 8 → I
- absolutley → absolutely
- rven → even

Annotation

- Data: *SMS + Twitter + AMT project*
- Who: *expert linguists*
- Tool: *Anagram*

The main issue

could u tell ppl im gonna b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

The main issue

cud u tell ppl im gona b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

The main issue



Annotation Process

- Guidelines
- Training
- Calibration
- Annotation
- Disagreement resolution

Learnings

1. Guidelines
 - a. short, covering one task only
 - b. non-contradicting
 - c. with a fall-back option
 - d. with as many examples as possible

Learnings

2. Quality control
 - a. annotate the first batch yourself
 - b. set up qualification tests (training)
 - c. check annotators' qualifications
 - d. do cross-annotation
 - e. set up automatic acceptance/rejection of the work



Learnings

3. Automatically annotated data saves time...
(and teach the annotators)

Learnings

3. Automatically annotated data saves time...
(and teach the annotators)
4. Saving time and money
 - a. extract 100% agreement from crowdsourcing
 - b. use experts to reannotate the rest

Learnings

3. Automatically annotated data saves time...
(and teach the annotators)
4. Saving time and money
 - a. extract 100% agreement from crowdsourcing
 - b. use experts to reannotate the rest
5. Pay quickly and be responsive to emails

Learnings

3. Automatically annotated data saves time...
(and teach the annotators)
4. Saving time and money
 - a. extract 100% agreement from crowdsourcing
 - b. use experts to reannotate the rest
5. Pay quickly and be responsive to emails
6. Gamification

Learnings

3. Automatically annotated data saves time...
(and teach the annotators)
4. Saving time and money
 - a. extract 100% agreement from crowdsourcing
 - b. use experts to reannotate the rest
5. Pay quickly and be responsive to emails
6. Gamification
7. Annotation bias

Inter-annotator agreement

How much do annotators agree?

- in general
- for each class

Basic metric: Cohen's Kappa

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

, where

$\text{Pr}(a)$ — actual agreement

$\text{Pr}(e)$ — expected agreement

Cohen's Kappa. Example

| | | B | B | B | |
|---|------|-----|------|-----|-----|
| | | pos | neut | neg | TOT |
| A | pos | 54 | 28 | 3 | 85 |
| A | neut | 31 | 18 | 23 | 72 |
| A | neg | 0 | 21 | 72 | 93 |
| | TOT | 85 | 67 | 98 | 250 |

Cohen's Kappa. Example

| | | B | B | B | |
|---|------|-----|------|-----|-----|
| | | pos | neut | neg | TOT |
| A | pos | 54 | 28 | 3 | 85 |
| A | neut | 31 | 18 | 23 | 72 |
| A | neg | 0 | 21 | 72 | 93 |
| | TOT | 85 | 67 | 98 | 250 |

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$$\Pr(a) = (54 + 18 + 72) / 250 = .576 \text{ (57.6\%)}$$

Cohen's Kappa. Example

| | | B | B | B | |
|---|------|-----|------|-----|-----|
| | | pos | neut | neg | TOT |
| A | pos | 54 | 28 | 3 | 85 |
| A | neut | 31 | 18 | 23 | 72 |
| A | neg | 0 | 21 | 72 | 93 |
| | TOT | 85 | 67 | 98 | 250 |

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

A used the label “**positive**” 85 times (54 + 28 + 3), or .425%

B used the “**positive**” label 85 times (54 + 31), or .425%

$$\text{Pr}(e) = 425 \times .425 = .180$$

Cohen's Kappa. Example

| | | B | B | B | |
|---|------|-----|------|-----|-----|
| | | pos | neut | neg | TOT |
| A | pos | 54 | 28 | 3 | 85 |
| A | neut | 31 | 18 | 23 | 72 |
| A | neg | 0 | 21 | 72 | 93 |
| | TOT | 85 | 67 | 98 | 250 |

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

$$\text{Pr}(e) = .180 + .077 + .146 = .403$$

Cohen's Kappa. Example

| | | B | B | B | |
|---|------|-----|------|-----|-----|
| | | pos | neut | neg | TOT |
| A | pos | 54 | 28 | 3 | 85 |
| A | neut | 31 | 18 | 23 | 72 |
| A | neg | 0 | 21 | 72 | 93 |
| | TOT | 85 | 67 | 98 | 250 |

$$K = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$$

$$K = \frac{.576 - .403}{1 - .403} = \frac{.173}{.597} = .29$$

Cohen's Kappa

What value is good?

- it depends
- ideally, > 0.8
- in real life, > 0.4 may be good enough

Data Generation

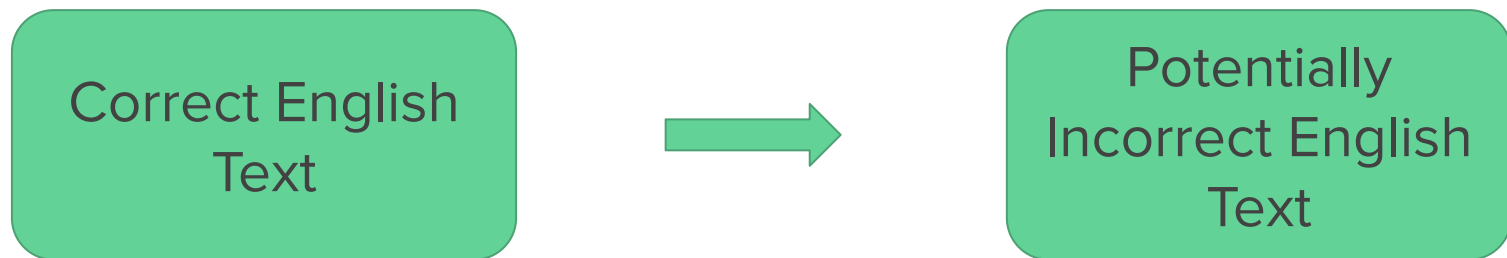
Use case: collocation correction

Today I did a very silly mistake.

Use case: collocation correction

*Today I {**did**=>**made**} a very silly mistake.*

The Idea

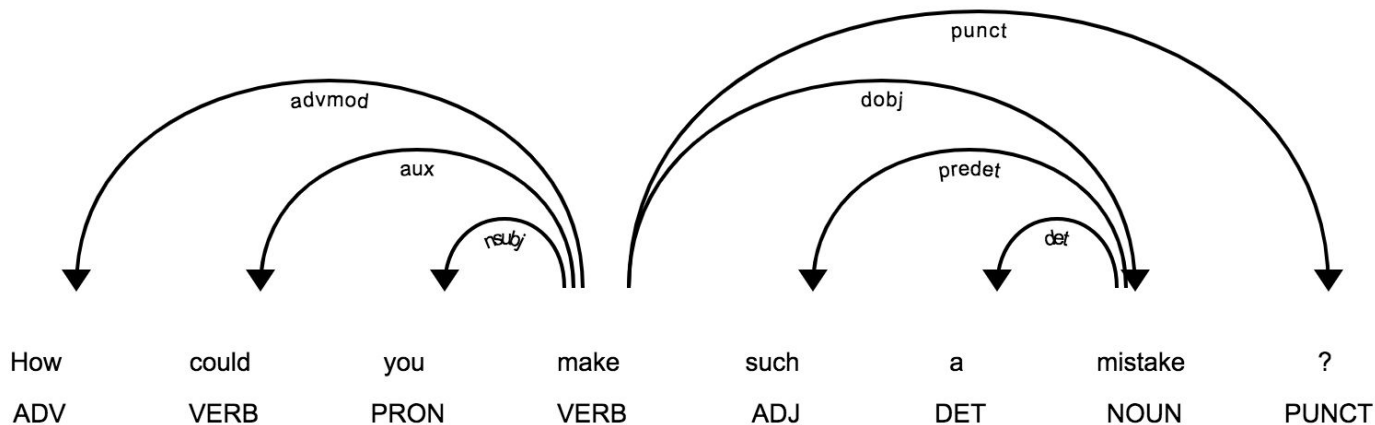


Collocation types

| <i>Categories</i> | <i>Examples</i> |
|-------------------|---|
| noun + verb | <i>the results suggest, the research shows</i> |
| verb + noun | <i>provides an explanation, discuss the problem</i> |
| adjective + noun | <i>concrete example, potential problem</i> |
| verb + particle | <i>point out, carry out</i> |
| adverb + verb | <i>clearly differs, thoroughly examine</i> |

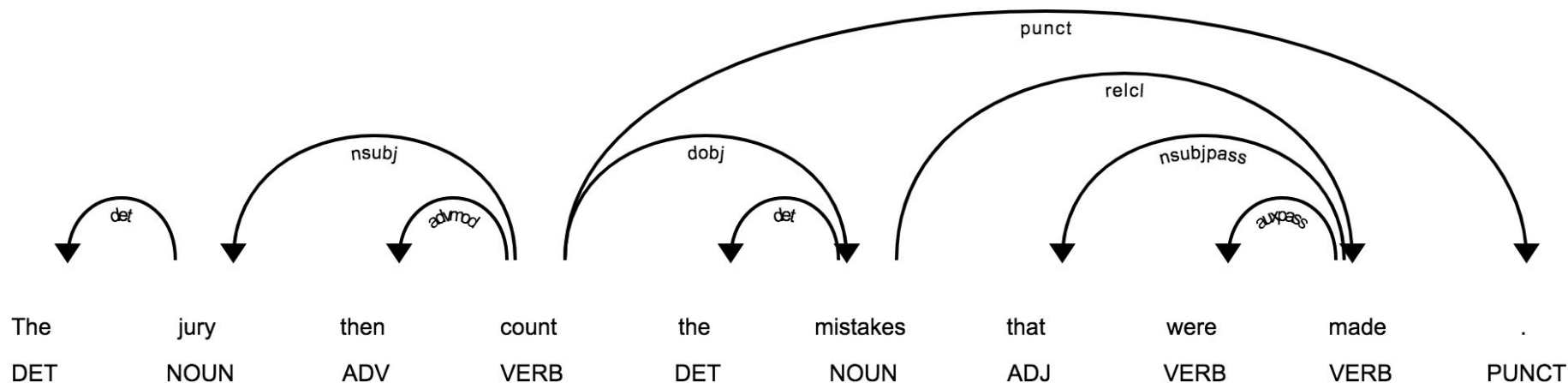
Use case: collocation correction

- Extract collocations from good texts



Use case: collocation correction

- Extract collocations from good texts



Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus

*How could you **make** such a mistake?*

do

commit

perform

execute

...

Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus
- Filter out if:
 - the replacement is a good collocation
 - the combination is frequent in good texts

Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus
- Filter out good replacements

*How could you **make** such a mistake?*

do

commit

perform

execute

...

Use case: collocation correction

- Extract collocations from good texts
- Get synonyms from a thesaurus
- Filter out good replacements
- Replace the good word with a synonym

*How could you **do** such a mistake?*

*How could you **perform** such a mistake?*

*How could you **execute** such a mistake?*

Results

- True positives
 - *I thought you did a {full => comprehensive} research...*
 - *...the most {beautiful => good-looking} men in the world.*
- Problems
 - Not all confusions are synonymous:
 - {crowded => heavy} traffic
 - Rare combinations can be treated as a mistake
 - {Subversive=>Underground} lines characterize...

Data Generation Pros & Cons

- + Potentially unlimited volume
- + Control of the parameters
- Artificial

Acquiring Data from Users

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

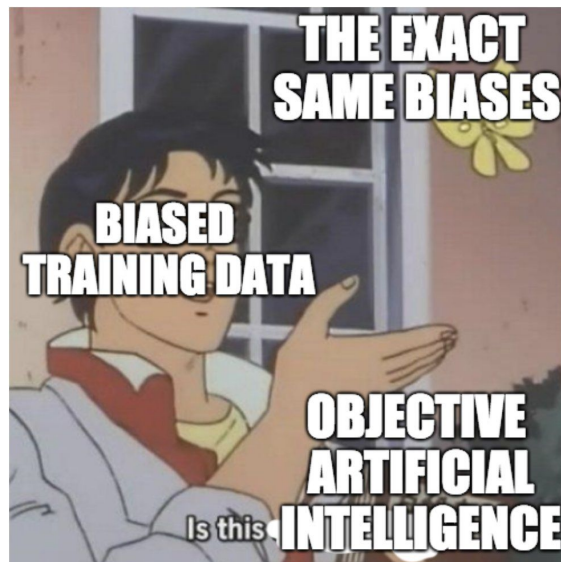
Potential approach: “lean startup”

Data Best Practices

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

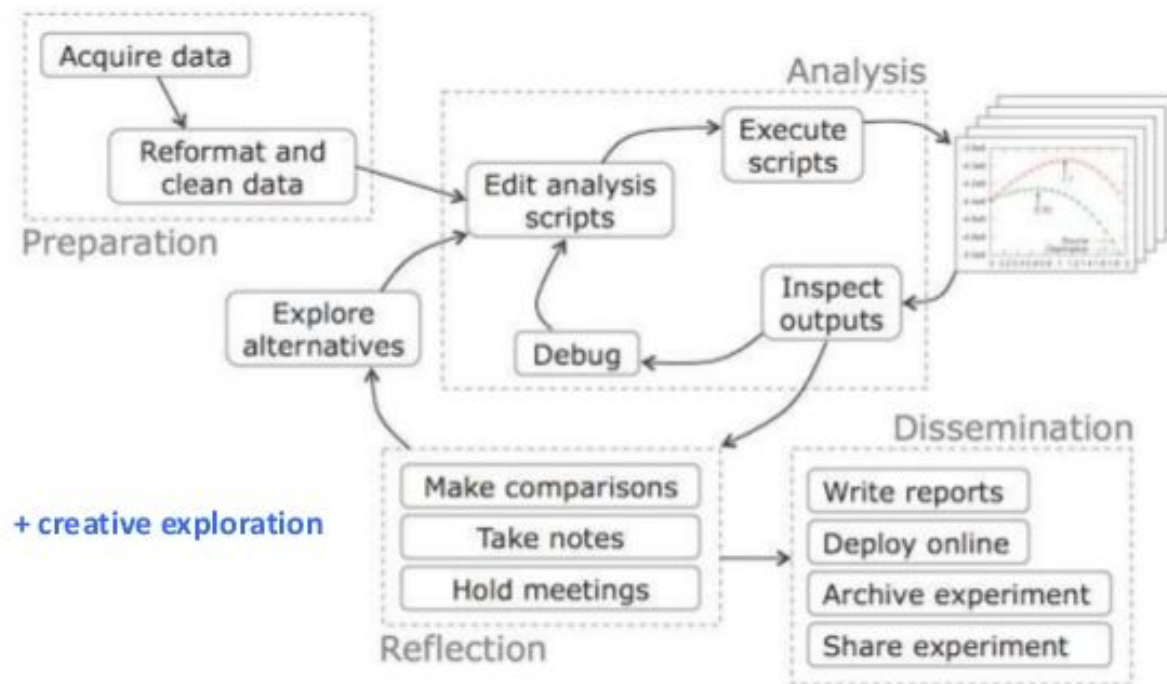
Be Aware of Bias

- Domain bias
- Dataset bias
- Model bias
- Social bias



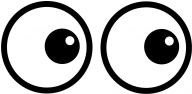
<https://www.slideshare.net/grammarly/grammarly-ainlp-club-1-domain-and-social-bias-in-nlp-case-study-in-language-identification-tim-baldwin-80252288>

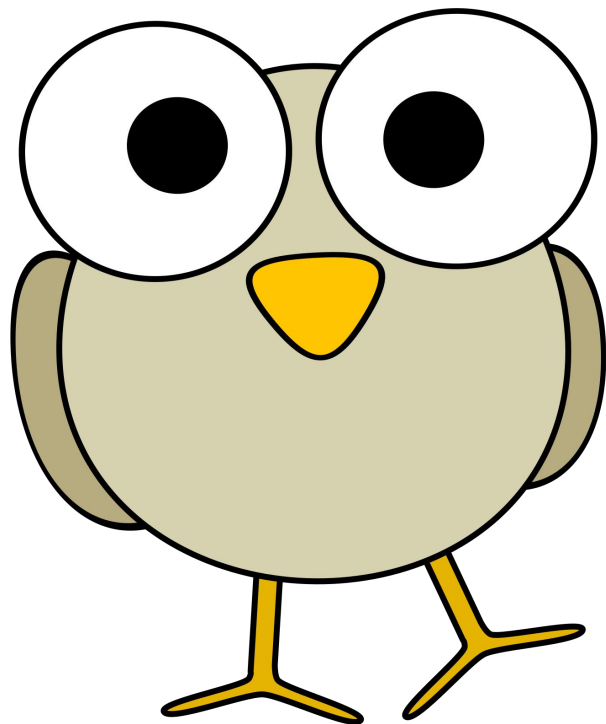
Data Workflow



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

Tools

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



??