

# Full Cycle of an NLP Project

Mariana Romanyshyn,  
*Computational Linguist at Grammarly*

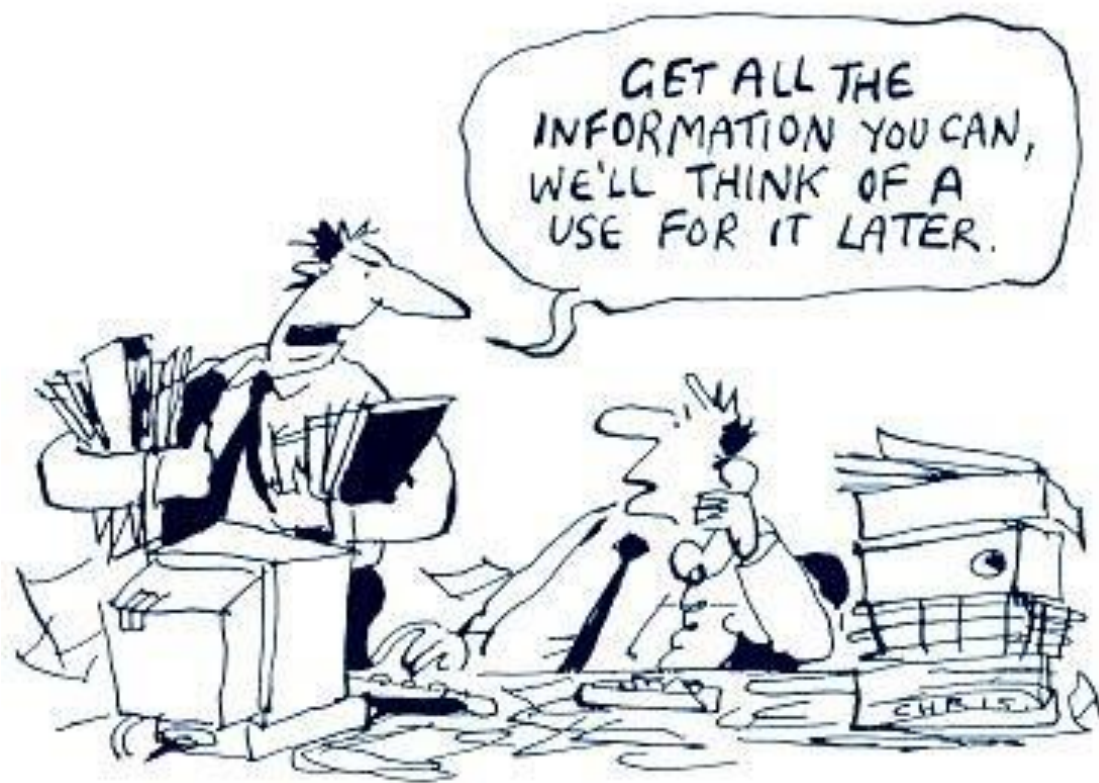
# Stages of an NLP project

- Domain analysis
- Data preparation
- Metrics
- Baseline, SOTA, and iterative improvement
- Feedback analysis
- Lather, rinse, repeat

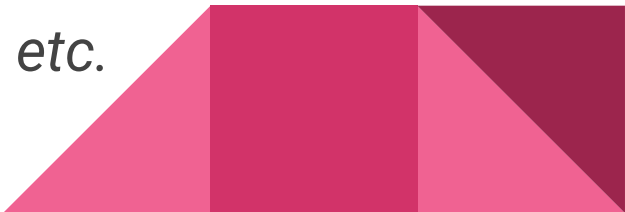
*(... as needed)*



# 1. Domain analysis



# Domain equals limitations

- Language
  - Topic
  - Register or formality level
  - Type of texts
    - *documents, tweets, songs, emails, fiction, etc.*
    - *long/short, structured/unstructured, etc.*
  - Author
    - *gender, age, nationality, native language, etc.*
  - Time and geography
- 

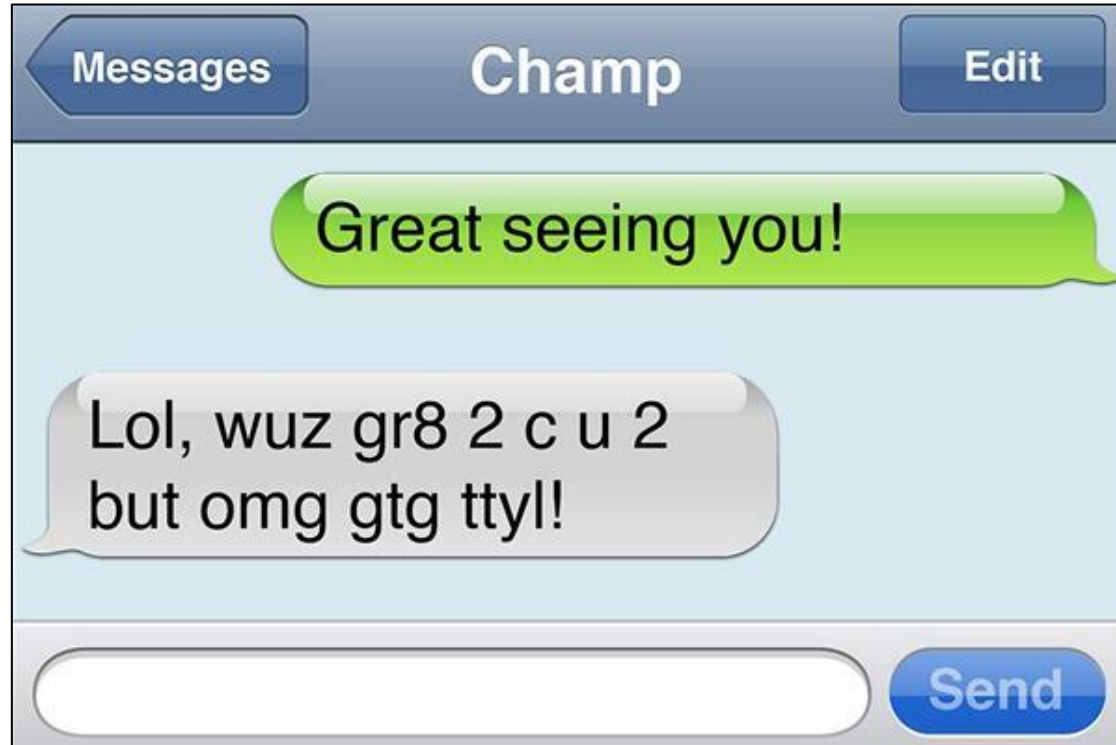
# E.g.: language identification in short texts

Text	Language	Explanation
Justin Bieber <3	und (Undefined)	NOT English; contains only a name.
Schalke XI v Chelsea: Fahrmann, Neustadter, Santana, Howedes, Uchida, Fuchs, Kirchhoff, Boateng, Hoger, Choupo-Moting, Huntelaar.	und (Undefined)	Contains only place/team/player names.
Ate spaghetti at La <u>tratoria napolitana</u>	en (English)	The name of the restaurant is in Italian, but the "main" language is English. An English-only speaker would understand this Tweet.
#NowListening Universo - Lodovica Comello @XYZ @XYZ	und (Undefined)	Italian song title and artist are just names. #NowListening is English but could be used by non-English speaker too.
#My #hot #naughty #neighbour #in #dallas: http://t.co/0dLJ 北京	en (English)	There is a Chinese word at the end, but the strongly prevailing language is English
Hahaha (•_•) (•_•)>¬¬ (¬¬_¬) YEAHHH!	und (Undefined)	Emoticons and interjections only.
Que bonito!	und (Undefined)	Could be both Spanish and Portuguese
Pozor pozor	und (Undefined)	Could be Czech, Serbian, Croatian, Slovenian, ...
So warm in Berlin!	und (Undefined)	A valid sentence in both German and English
"Last Christmas" - <u>Der</u> Jose Carreras unter <u>den</u> <u>Weihnachtsliedern</u> .	de (German)	Contains an English song title and Spanish name, but is understandable to a German-only speaker.
Bécs <3	hu (Hungarian)	This is the Hungarian name for "Vienna", which is a proper name, but exists only in Hungarian
Estoy muy cansado voy a acostarme .... sooo tired <u>goin</u> to <u>bedd</u>	und (Undefined)	Strong mixture of Spanish and English, no clear "main" language

Text	Language	Explanation
Justin Bieber <3	und (Undefined)	NOT English; contains only a name.
Schalke XI v Chelsea: Fahrmann, Neustadter, Santana, Howedes, Uchida, Fuchs, Kirchhoff, Boateng, Hoger, Choupo-Moting, Huntelaar.	und (Undefined)	Contains only place/team/player names.
Ate spaghetti at La tratoria napolitana	en (English)	The name of the restaurant is in Italian, but the "main" language is English. An English-only speaker would understand this Tweet.
#NowListening Universo - Lodovica Comello @XYZ @XYZ	und (Undefined)	Italian song title and artist are just names. #NowListening is English but could be used by non-English speaker too.
#My #hot #naughty #neighbour #in #dallas: http://t.co/0dLJ 北京	en (English)	There is a Chinese word at the end, but the strongly prevailing language is English
Hahaha ( •_• ) ( •_• )>■-■ (■_■) YEAHHH!	und (Undefined)	Emoticons and interjections only.
Que bonito!	und (Undefined)	Could be both Spanish and Portuguese
Pozor pozor	und (Undefined)	Could be Czech, Serbian, Croatian, Slovenian, ...
So warm in Berlin!	und (Undefined)	A valid sentence in both German and English
"Last Christmas" - Der Jose Carreras unter den Weihnachtsliedern.	de (German)	Contains an English song title and Spanish name, but is understandable to a German-only speaker.
Bécs <3	hu (Hungarian)	This is the Hungarian name for "Vienna", which is a proper name, but exists only in Hungarian
Estoy muy cansado voy a acostarme .... sooo tired goin to bedd	und (Undefined)	Strong mixture of Spanish and English, no clear "main" language



# E.g.: spelling correction for mobile





# E.g.: spelling correction for mobile

*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*



# E.g.: spelling correction for mobile

*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*



## **2. Data preparation**

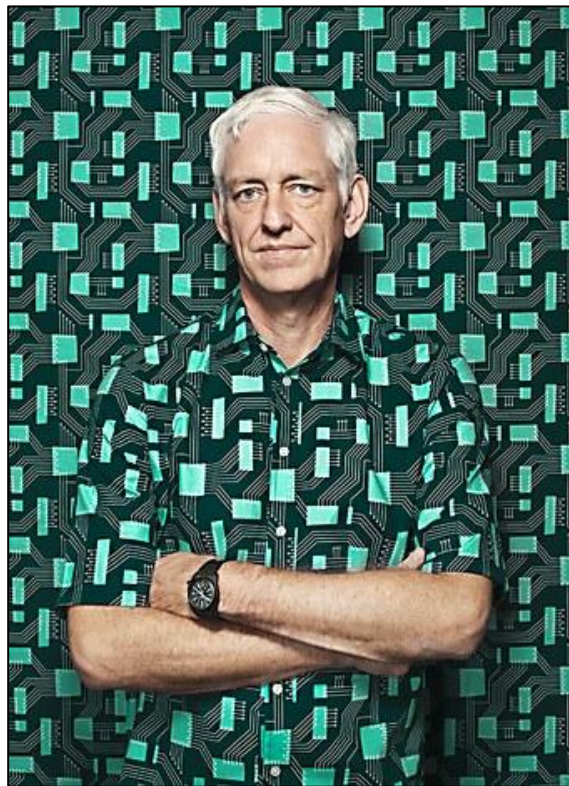
# Data

*“Data is ten times more powerful  
than algorithms.”*

— Peter Norvig

The Unreasonable Effectiveness of Data

<http://youtu.be/yvDCzhbjYWs>




# Data sets

- Annotated
  - manually
  - automatically
- Non-annotated
  - have to be annotated
    - *Appen / Figure Eight*
    - *Amazon Mechanical Turk*
    - *you, your friends and family ;)*
    - ...



# Where can you get data?

- Use open data
    - e.g., *Wikipedia, DBPedia, Brown, Gutenberg, Reuters*
    - e.g., *Wiktionary, WordNet and other \*Nets*
  - Buy
    - e.g., at LDC or from people/companies
  - Scrape
    - e.g., *Reddit, Twitter, IMDB, Amazon, forums*
  - Generate
  - Crowdsource
- 

# Where can you get data?

- Use open data
  - e.g., Wikipedia, DBPedia, Brown, Gutenberg, Reuters
  - e.g., Wiktionary, WordNet and other \*Nets
- Buy **The data will always be noisy!**
  - e.g., at LDC or from people/companies
- Scrape
  - e.g., Reddit, Twitter, IMDB, Amazon, forums
- Generate
- Crowdsource





# E.g.: sentiment analysis

**Сидорчук Валентина**  Уже купил 

25 ноября 2017 

100% пользователей считают этот отзыв полезным

Классная игрушка для активной бабушки - закачала музыку и вперед скандинавской ходьбой...

# E.g.: sentiment analysis

**Сидорчук Валентина**  Уже купил 

25 ноября 2017 

100% пользователей считают этот отзыв полезным

Классная игрушка для активной бабушки - закачала музыку и вперед скандинавской ходьбой...

**Сергей**  100% пользователей считают этот отзыв полезным

19 февраля 2017 

Здравствуйте. Скажите, как fiio x1 (первой версии) сыграется с наушниками fiio ex1? Спасибо

 Ответить

 2 | 



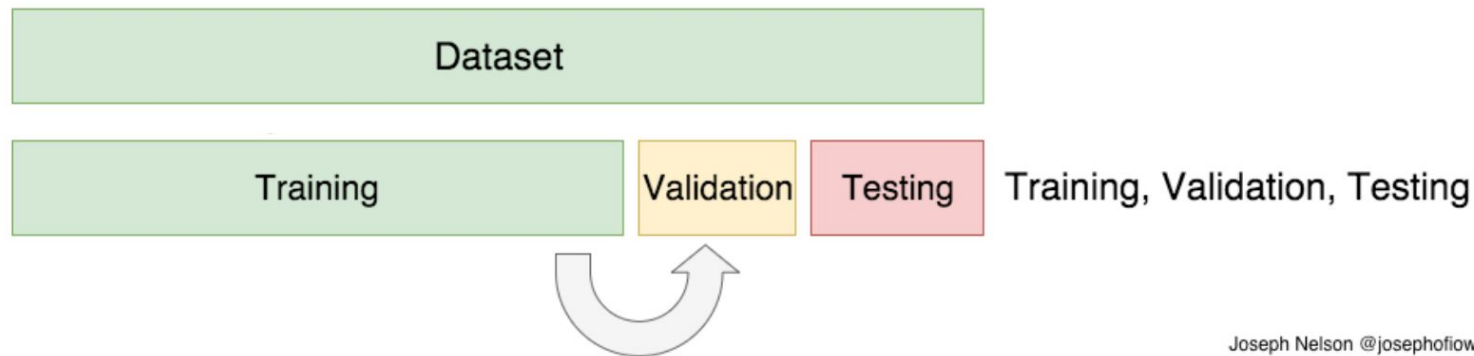
# 3. Metrics

# Evaluation

- Intrinsic
  - use a golden data set for this specific task
    - *e.g., quality of syntactic parsing, language identification, NER, etc.*
- Extrinsic
  - evaluate on a broader task
    - *e.g., sentiment analysis, machine translation, recommendation system*



# Data sets



# Data sets

- **Train set and dev set**
  - 80-90% of your data
  - we use it to develop the solution
  - we use it to debug our code
- **Test set**
  - 10-20% of your data
  - we never look at it
  - we never debug code using it
  - we only use it for testing



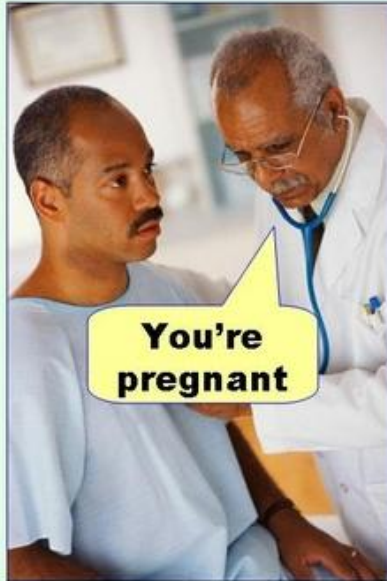
# Traditional metrics

<b>predicted→ real↓</b>	<b><i>Class_pos</i></b>	<b><i>Class_neg</i></b>
<b><i>Class_pos</i></b>	<b>TP</b>	<b>FN</b>
<b><i>Class_neg</i></b>	<b>FP</b>	<b>TN</b>

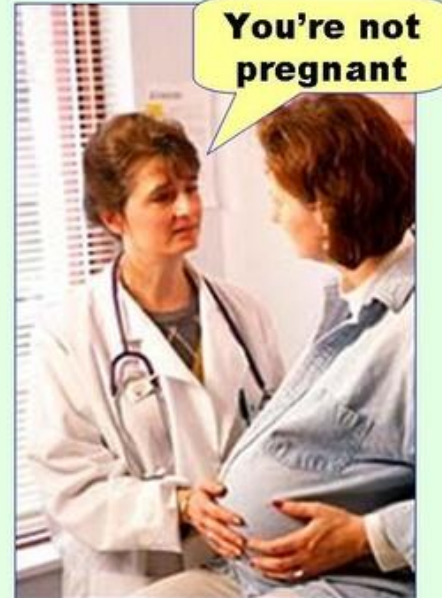


# Traditional metrics

**Type I error**  
(false positive)



**Type II error**  
(false negative)



# Traditional metrics

$$\textit{Precision} = \frac{TPs}{TPs + FPs}$$

$$\textit{F - score} = \frac{\textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

$$\textit{Recall} = \frac{TPs}{TPs + FNs}$$

$$\textit{Accuracy} = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}$$



# Other metrics

- Confusion matrix

Languages	English	German	French	Italian	Dutch	Spanish
English	<b>9244</b>	38	199	145	222	139
German	28	<b>9514</b>	67	29	325	27
French	20	52	<b>9525</b>	165	83	160
Italian	6	7	18	<b>9822</b>	16	134
Dutch	60	66	35	20	<b>9800</b>	19
Spanish	6	8	41	242	24	<b>9679</b>

# Other metrics

- Confusion matrix
- BLUE

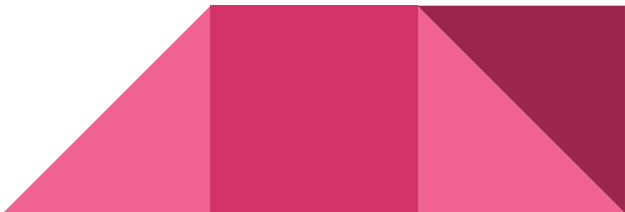
## Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

## Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

# Other metrics

- Confusion matrix
  - BLUE
  - Max-match
  - Parseval, cross-bracketing, leaf-ancestor...
  - FP rate and FN rate
  - ROC curves
  - Outliers
  - Human evaluation
  - Your metric :)
- 

## **4. Baseline, SOTA, and iterative improvement**

# Baseline

Baseline — quality that can be achieved using a reasonable primitive approach.

Two ways to improve:

- extend coverage (*i.e., improve recall*)
- reduce error rate (*i.e., improve precision*)





# SOTA

State-of-the-art — the best publicly known solution tested on a public data set with commonly accepted metrics.





# Iterative improvement

- Rule-based approach
- Statistical approach
- Hybrid of rules and statistics
- Machine learning approach
- Hybrid of rules and machine learning



## **5. Feedback analysis**

# Feedback analysis

- Is the input data the same as we expected?
- How well does the system perform on real data?
- How do users react to the system?
- Do we see a decrease in retention?
- ... and many more questions



## 6. Use case

# How would you build a fact checker?

## Checklist:

- Domain analysis
- Data preparation
- Metrics
- Baseline, SOTA, and iterative improvement
- Feedback analysis





# How would you detect catchy headlines?





**Questions?**