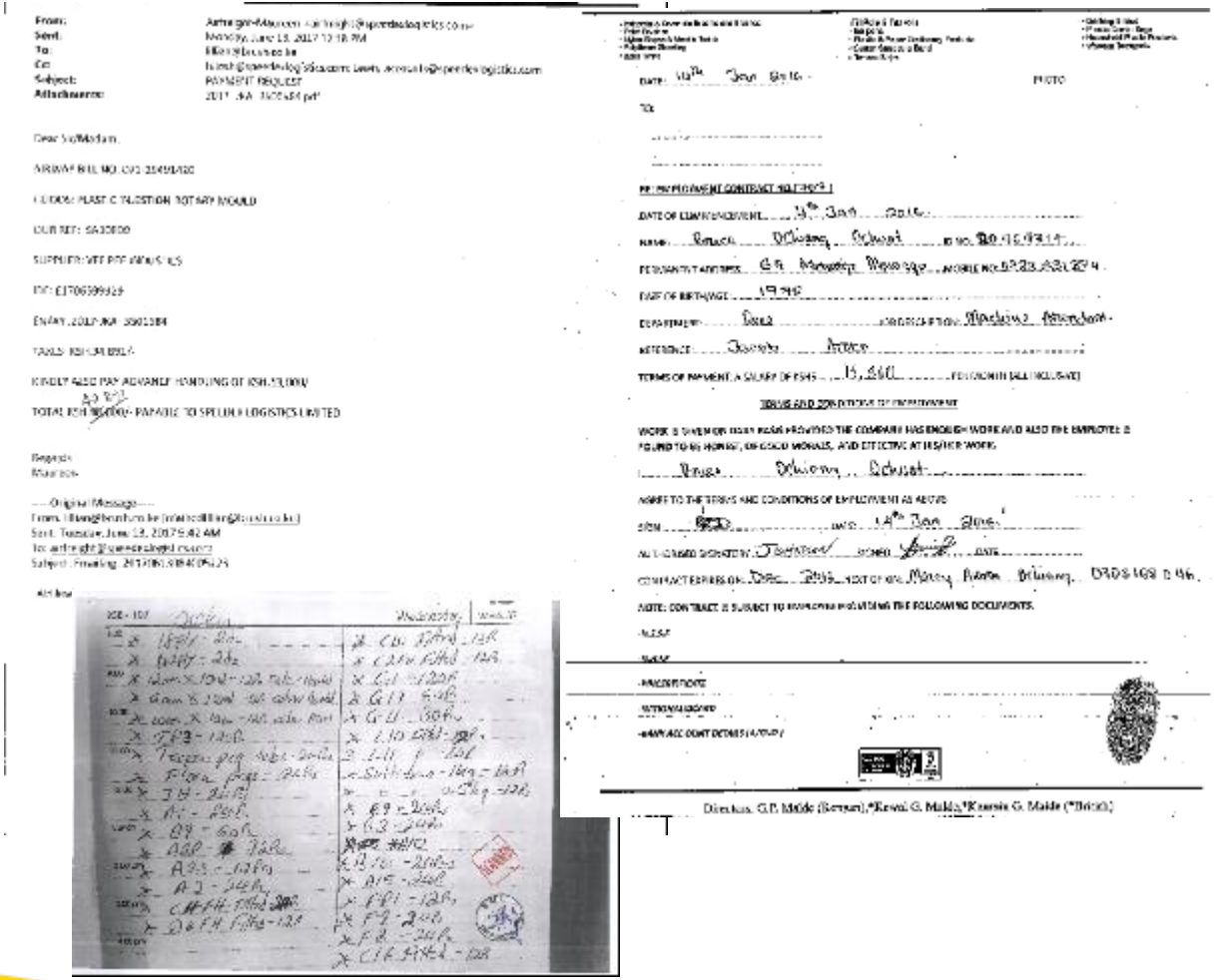
Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI

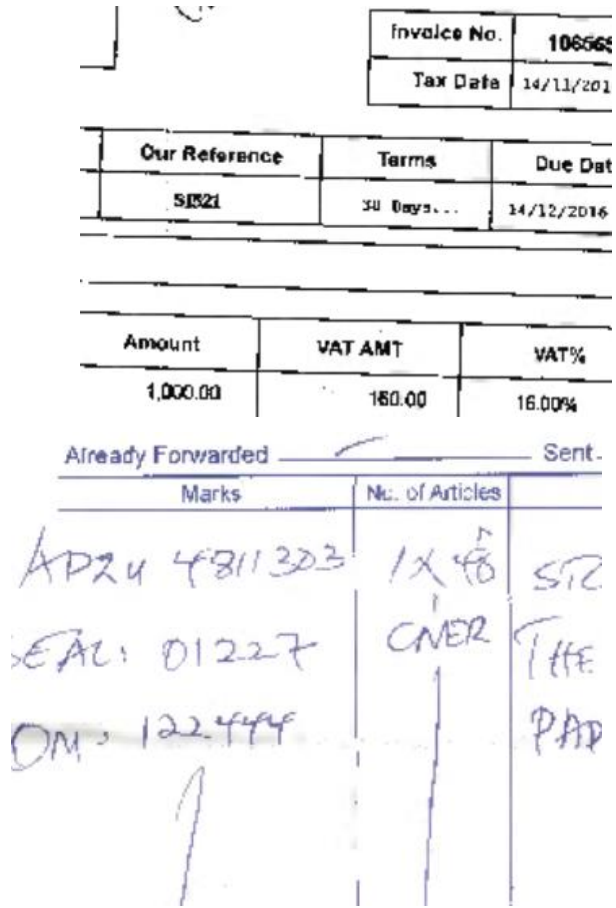
Document classification:

- Document processing
- Web scraping
- GDPR and other laws
- Poor quality
- Handwritten text
- Different templates



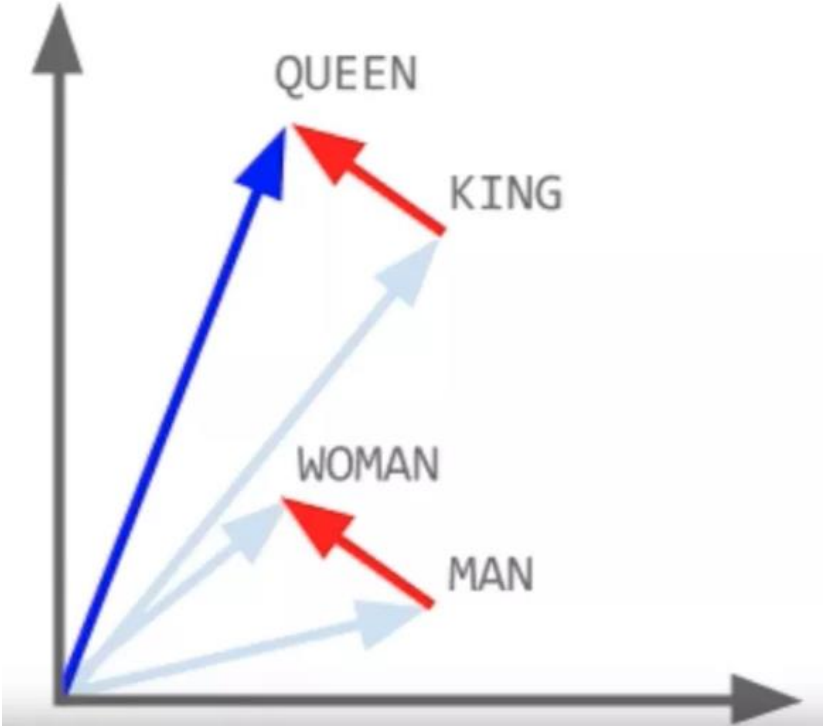
NER or named entities recognition:

- OCR engines
- Document formats
- General or document specific (e.g. date)
- Columns and other formatting
- Abbreviations



Typical NLP process:

- OCR
- Cleaning (symbols, stop-words, etc)
- Tokenization (sentence and word)
- Embeddings (BOW or word2vec)



Issues with the typical process:

- Missed information (color, style, template)
- Tokenization of German-like languages
- Arabic and Hebrew languages
- Table columns
- Contextual labeling is expensive

בָּרוּךְ אַתָּה יְהוָה אֱלֹהֵינוּ
מֶלֶךְ־הָעוֹלָם
הַמּוֹצִיא לֶחֶם מִן־הָאָרֶץ
אָמֵן:

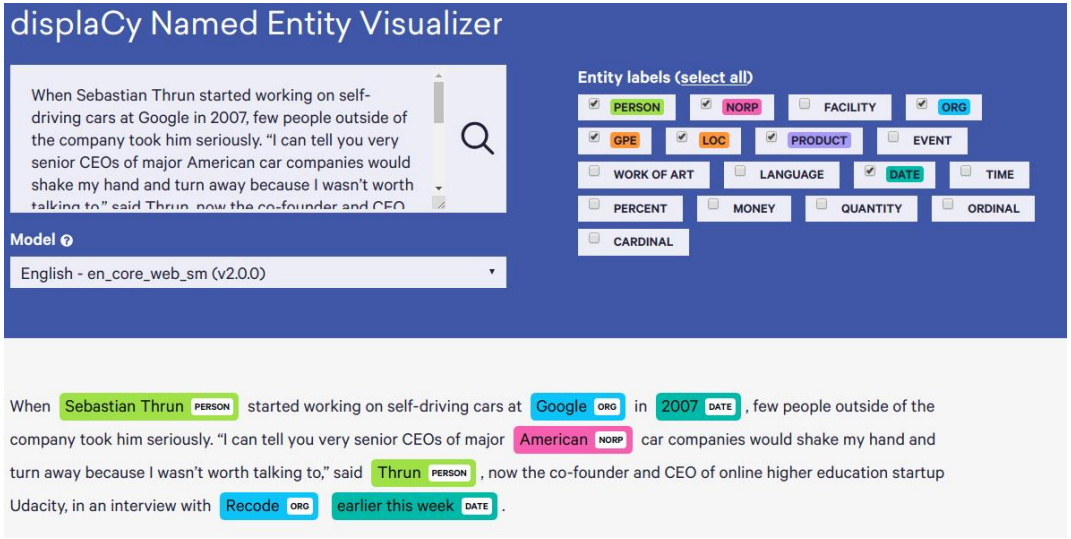
polyglot:

- 40 models
- Trained on Wiki
- Semi-supervised learning
- Embeddings plus Shallow Neural Network

Language	Sentence	Translation
English	Simien was traded from the Heat along with Antoine Walker and Michael Doleac to the Minnesota Timberwolves on October 24, 2007, for Ricky Davis and Mark Blount.	-
Hungarian	Dimitri beszélt egy utat Rómába.	Dimitri talked about a trip to Rome.
Spanish	Pešek nació en Praga y estudió dirección de orquesta piano en la Academia de Artes allí, con Václav Smetacek	Pešek born in Prague and studied orchestra direction piano at the Academy of Arts there, Vaclav Smetacek.
Russian	Уроженец Рио-де-Жанейро, Хосе Родригес Тринидади использовал сокращенную форму как своим сценическим псевдонимом.	A native of Rio de Janeiro, Jose Rodriguez Trindade used as a shortened form of his stage name.
Korean	도널드 스미스 는 1976 년에 자이드 압둘 아지즈 에 그의 이름 을 바꿨다 .	Donald Smith in 1976 changed his name to Zaid Abdul Aziz.
French	La France veut satisfaire à ses engagements envers l'Union européenne.	France wants to meet its commitments to the European Union.
Turkish	Erdoğan, Türkiye'de twitter yasaklandı.	Erdogan banned twitter in Turkey.
Arabic	قال غاريث بيل ، نجم فريق ريال مدريد ، إن الفوز بدوري أبطال أوروبا سيظل معه الى الأبد، وذلك بعدما ساعد هدفه النادي الملكي على هزيمة نادي اتليتكو مدريد في Spain .	Said Gareth Bell, the star of Real Madrid, said that winning the Champions League will remain with him "forever", and that after his goal helped the club to defeat Atletico Madrid in Spain.
Indonesian	Rendjambe meninggal dalam keadaan tidak jelas , yang mengakibatkan kerusuhan oleh pendukung oposisi marah di Port Gentil - dan Libreville.	Rendjambe died in unclear circumstances, which resulted in riots by angry opposition supporters in Port Gentil - and Libreville.
Chinese	新华社 25 日报道 援引 新疆 公安厅 。	Xinhua News Agency report quoted the 25th Xinjiang Public Security Department.
Greek	Ενλίλ έφερε τα Γκούτιους κάτω από το λόφους ανατολικά του Τίγρη, να φέρει το θάνατο σε ολόκληρη τη Μεσοποταμία.	Enlil brought the Gutians down from the hills east of the Tigris, to bring death throughout Mesopotamia.

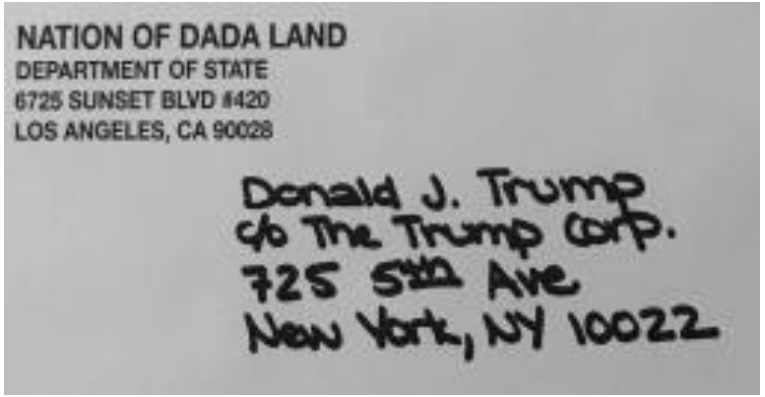
spaCy:

- Subword embeddings
- Affixes
- Convolutions with residuals
- No Bi-LSTM
- Custom models
- Active Learning



Ready-Solution problems:

- Missed capital letters info
- Disambiguations (Bank or Bank, numbers)
- Work for generic terms, not for document specific
- Only contextual labels
- Only basic ENs (PER, ORG, LOC)
- Only few languages working good



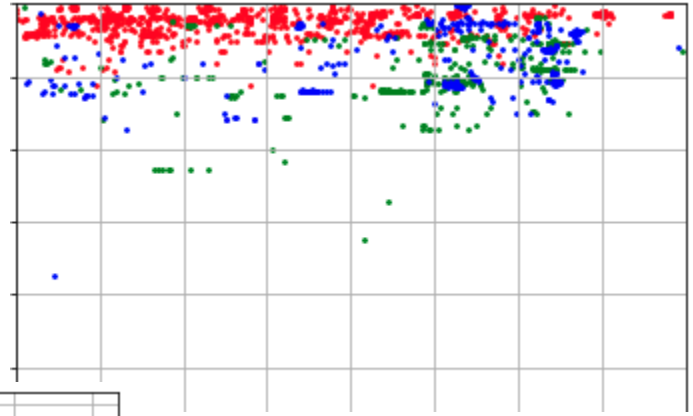
How we do it - Document classification:

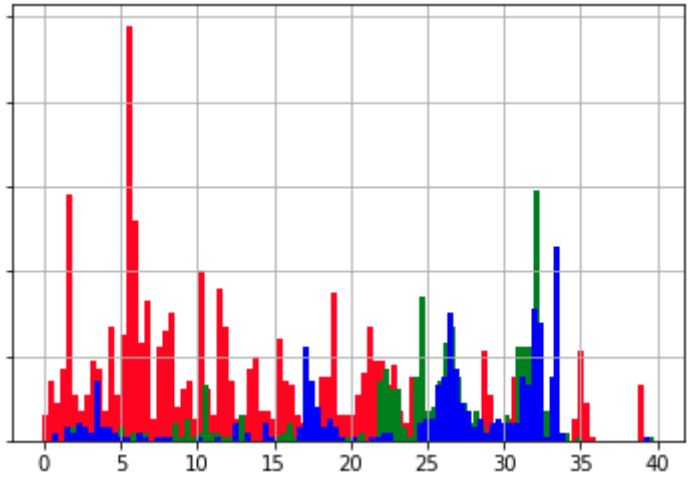
- Typical BOW model
- CV analysis using CNN or Shallow NN
- 70 + 70 = 93

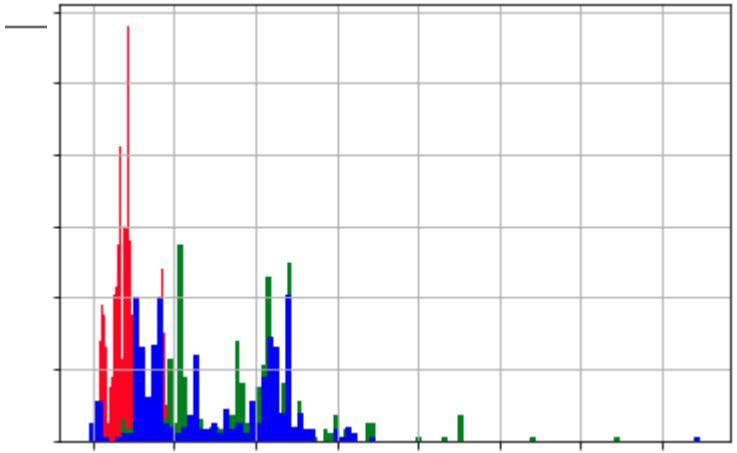
How we do it - Location model:

- statistical model
  for every EN location on the doc

- 62% accuracy out of 22 docs

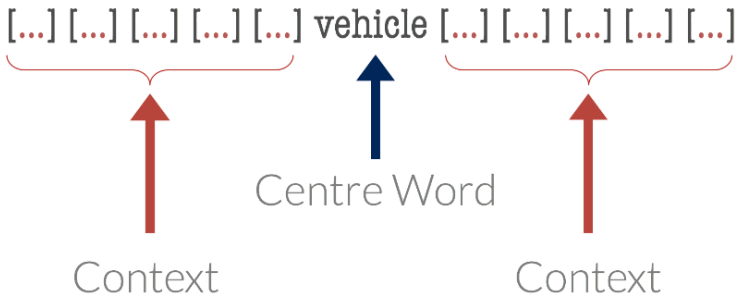






How we do it - Word model:

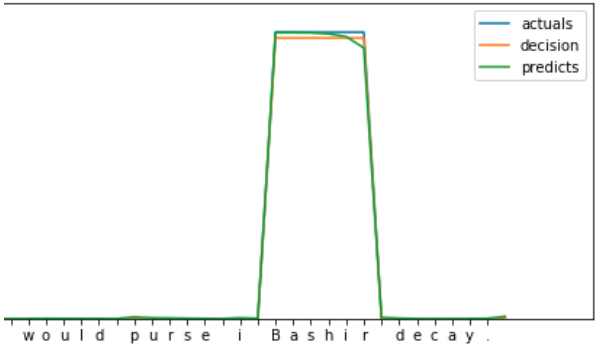
- word2vec embeddings
- Horizontal and vertical tokenizer
- 1D CNN + Bi-directional LSTM
- 80% accuracy out of 12 ENs

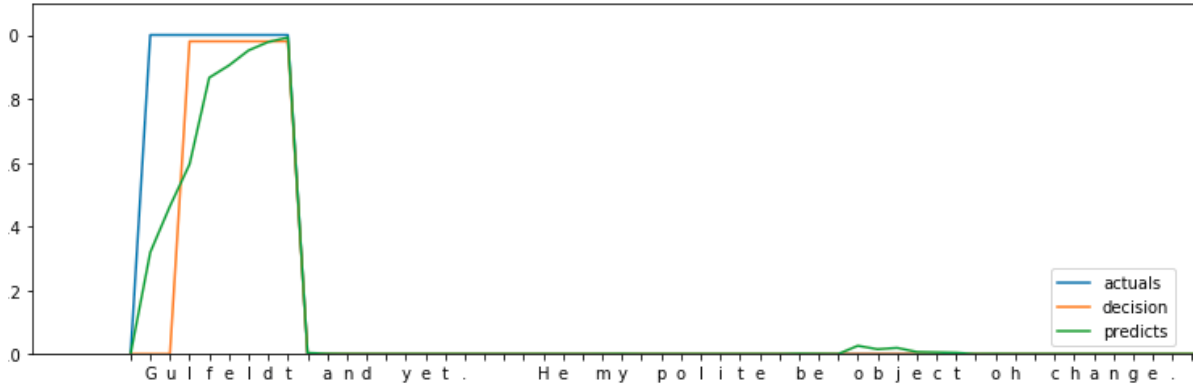


Order·Date¤	Region¤	Rep¤	Item¤	Units¤	Unit·Cost¤	Total¤
1/23/10¤	Ontario¤	Kivell¤	Binder¤	50¤	19.99¤	999.50¤
2/9/10¤	Ontario¤	Jardine¤	Pencil¤	36¤	4.99¤	179.64¤
2/26/10¤	Ontario¤	Gill¤	Pen¤	27¤	19.99¤	539.73¤
3/15/10¤	Alberta¤	Sorvino¤	Pencil¤	56¤	2.99¤	167.44¤
4/1/10¤	Quebec¤	Jones¤	Binder¤	60¤	4.99¤	299.40¤
4/18/10¤	Ontario¤	Andrews¤	Pencil¤	75¤	1.99¤	149.25¤
5/5/10¤	Ontario¤	Jardine¤	Pencil¤	90¤	4.99¤	449.10¤
5/22/10¤	Alberta¤	Thompson¤	Pencil¤	32¤	1.99¤	63.68¤

How we do it - Character model:

- Feature engineering (style, capitalization, color, punctuations)
- Bi-directional LSTM
- Hysteresis
- Cheap automatic labeling
- Data Anonymization





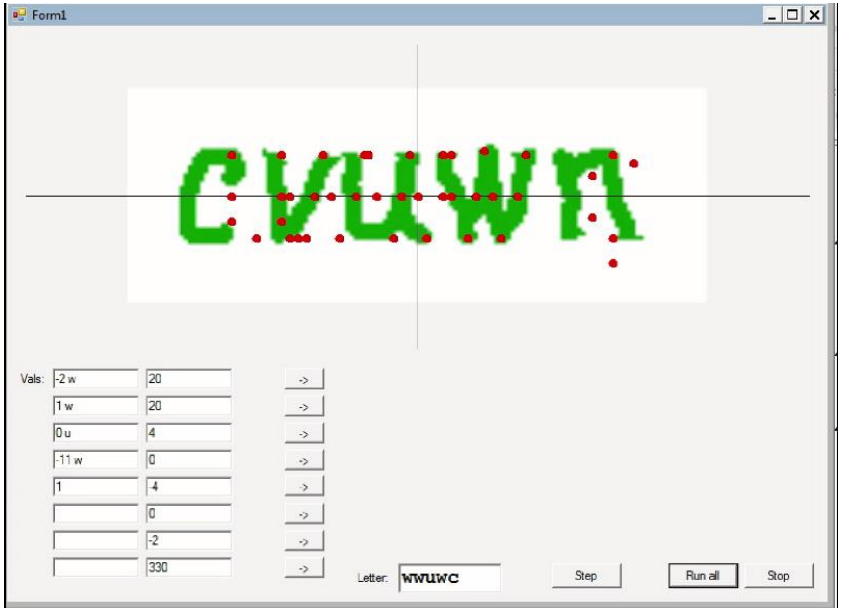
How we do it - stacking models:

- RE filters
- Joining contexts (street and number)
- Entity linking (removing disambiguation)
- Stacking horizontally as features
- Use Bayesian rule
- Test accuracy over 95%

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

30 years ago: ML model to play checkers (thanks to my father and Martin Gardner)

(picture from www.instructables.com)



in .NET

Jh

that has tf-like operators, optimizer and

even automatic hyper-param tuner

Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI