732A92/TDDE16 Text Mining (2022)

Information extraction

Marco Kuhlmann
Department of Computer and Information Science

This session

- Questions and answers
- Developing data sets
- Introduction to the lab

Questions and answers

Overview of information extraction

1. Introduction to information extraction
2. Named entity recognition
3. Entity linking
4. Relation extraction

Sample course project

This text mining project has made an attempt at extracting insider transactional data from Norwegian press releases. Through the use of spaCy's rule-based matcher, the date at which this insider transaction occurred, the name of the organization or individual who made the transaction, the number of shares that were traded, and the price at which the stocks were either bought or sold at, were extracted from the financial press releases. The project shows that a rule-based matcher using NLP can be used as a baseline for more complex models, or help anyone manually performing the information extraction, and a precision and recall of 0.63 and 0.61, respectively, was reached. Furthermore, the results of this project highlight the importance of knowledge bases and the domain specific knowledge required in order to reach sufficient results, through the use of well-defined rules.

Sample thesis project

This thesis explores approaches for extracting company mentions from financial news articles that carry a central role in the news. The thesis introduces the task of salient named entity extraction (SNEE): extract all salient named entity mentions in a text document. Moreover, a neural sequence labeling approach is explored to address the SNEE task in an end-to-end fashion, both using a single-task and a multi-task learning setup. In order to train the models, a new procedure for automatically creating SNEE annotations for an existing news article corpus is explored. The neural sequence labeling approaches are compared against a two-stage approach utilizing NLP parsers, a knowledge base and a salience classifier. Textual features inspired from related work in salient entity detection are evaluated to determine what combination of features results in the highest performance on the SNEE task when used by a salience classifier. The experiments show that the difference in performance between the two-stage approach and the best performing sequence labeling approach is marginal, demonstrating the potential of the end-to-end sequence labeling approach on the SNEE task.

Developing data sets

Project structure

1. Identify your problem — 8 hours (w44–w48)
2. Design your approach — 32 hours (w49–w50)
3. Evaluate your approach — 32 hours (w51–w01)
4. Produce your report — 16 hours (w02)

Suggested structure (1)

- **Introduction**

  What problem did you address in the project? Why is this problem interesting? What can we learn by solving the problem?

- **Theory**

  Present relevant theoretical background, and in particular those concepts and methods that were not covered in the course.

Suggested structure (2)

- **Data**

  What data did you use in your project? How was this data created? What preprocessing did you do (if any), and why?

- **Method**

  Explain how you approached the stated problem. Aim to be detailed enough for others to reproduce your results.

- **Results**

  Present your results in an objective way. Use tables and charts, but do not forget to also include a summary in text form.

Suggested structure (3)

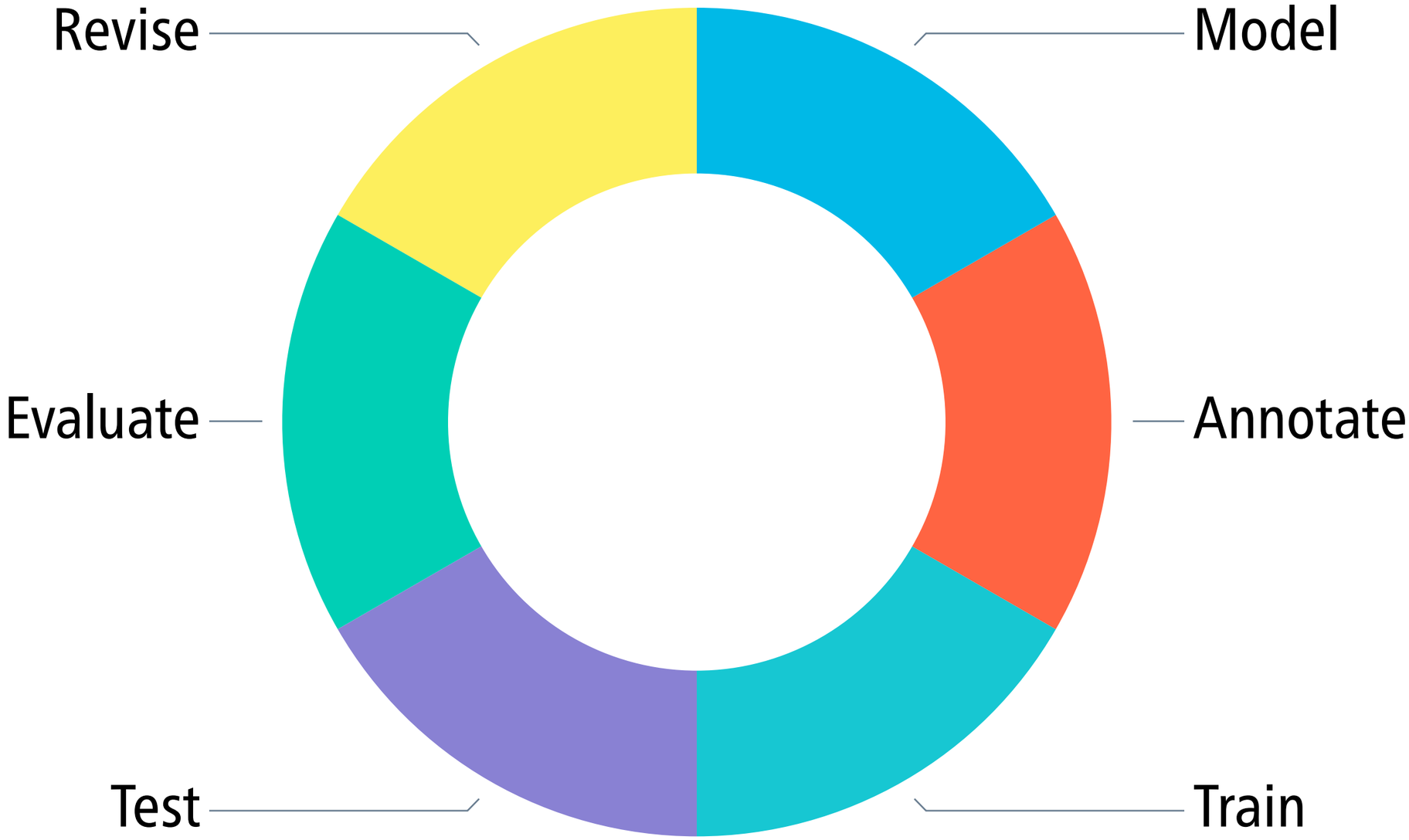
- **Discussion**

  Analyse your results. Discuss the limitations of your work. Compare your study to related work, such as internet materials or scientific articles.

- **Conclusion**

  Summarise your analysis. To what extent did you solve your stated problem? What else do you take away from your project?

Data development cycle



Source: Pustejovsky and Stubbs (2013)

Sentiment analysis

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

Source

positive

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

Source

negative

Named entity recognition

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

movie title

person

Phase 1: Model

The data model describes the data in abstract terms.

- **Sentiment analysis**

  Each text can express either a *positive* or a *negative* sentiment towards the movie that is being reviewed.

- **Named entity recognition**

  Sequences of words (text tokens) can refer to *named entities*, such as *persons* or *movie titles*.

Phase 2: Annotate

Annotate the data based on established guidelines.

- **Sentiment analysis**

  Annotate the overall sentiment towards the movie, as expressed in the text. (Parts of the review can deviate from this.)

- **Named entity recognition**

  Annotate only persons that exist in real life (such as actors), not fictitious persons (such as movie characters).

Phase 2: Annotate

- The text material is annotated by one or several annotators. These try to follow the guidelines as closely as possible.
- The annotation guidelines are discussed and may be adapted in the course of the annotation work.

  e.g., if exhaustive annotation turns out to be infeasible

- When the text material has been annotated, a **gold standard** is created through an adjudication process.

  comparison and discussion

Phase 3 and 4: Train and test

For the purposes of supervised machine learning, the gold-standard data set is partitioned into at least three different subsets:

- a **training set**
- a **development set** that is used to test the system during development, and to set hyperparameters
- a **test set** that is used for the final evaluation

Phase 5: Evaluate

- **Intrinsic evaluation**

  Evaluate a component by letting it solve a specific sub-task and calculate some evaluation measure.

  Example: evaluate on a standard data set for sentiment analysis

- **Extrinsic evaluation**

  Evaluate a component by integrating it into a larger system that solves an end-to-end task.

  Example: integrate sentiment analysis into a stock price predictor

Phase 6: Revise

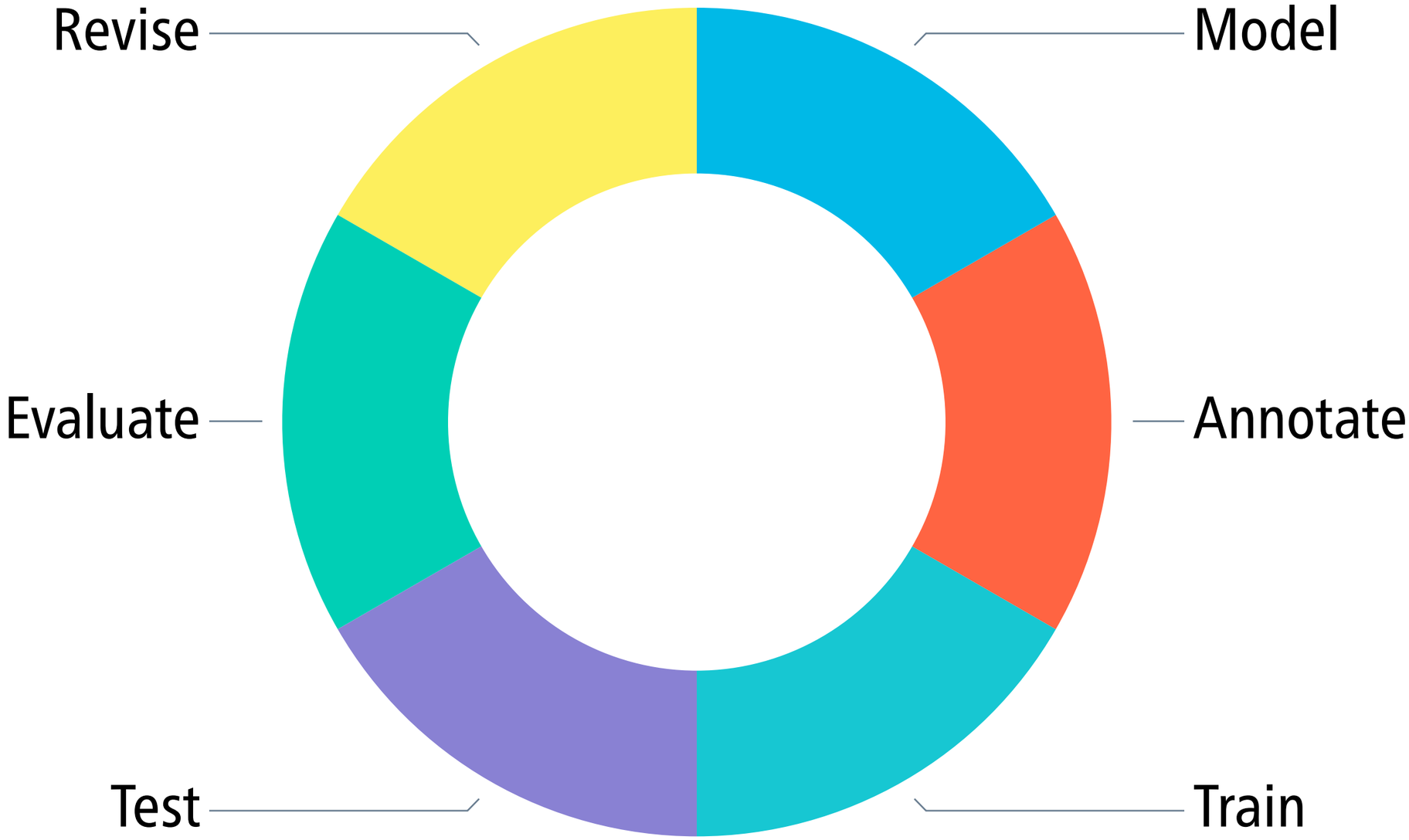
- As a complement to the quantitative evaluation, it is useful to also do a qualitative evaluation in the form of an error analysis.

  Example: Which entities are confused most often?

- This error analysis can result in a new annotation model, new annotation guidelines, and/or extended annotations.

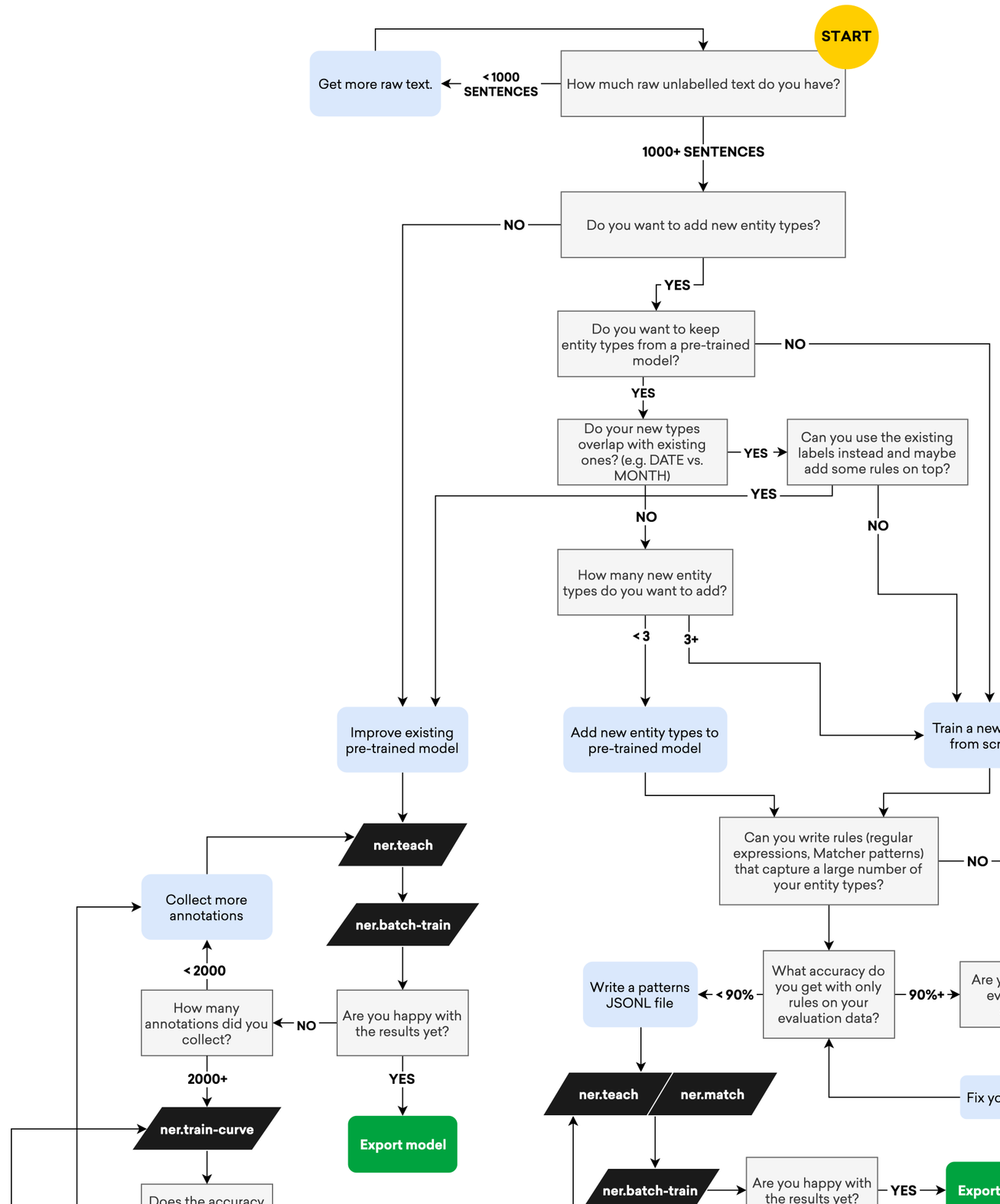
  Example: add more types to the list of named entities

Data development cycle



Source: Pustejovsky and Stubbs (2013)

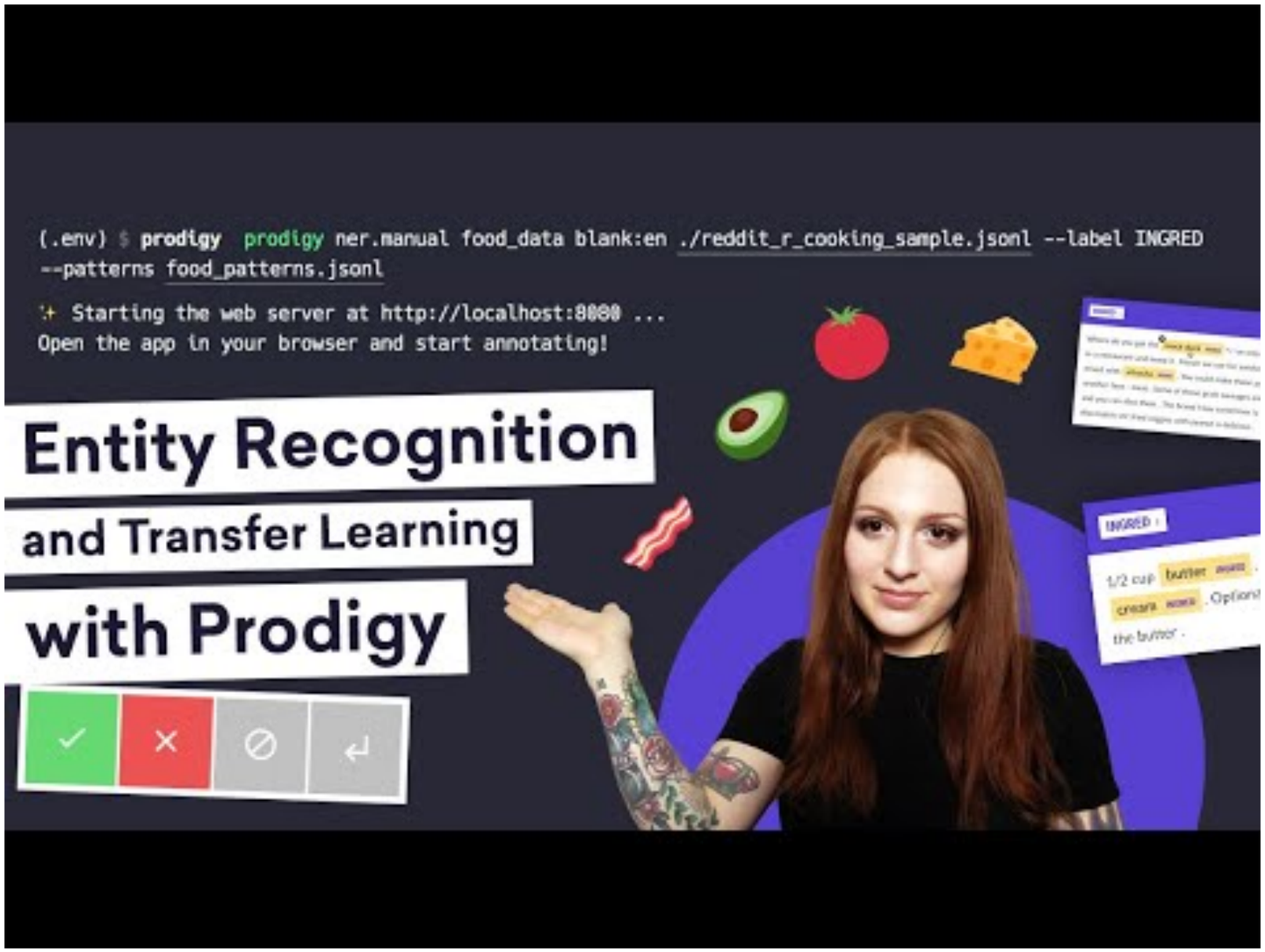
prodigy Annotation Flowchart: Named Entity Recognition



RECIPE LEGEND

ner.teach	Stream in suggestions from the model and optional pattern matches and accept / reject them.
ner.match	Stream in pattern matches and accept / reject them.
ner.manual	Highlight spans of text manually.
ner.batch-train	Train a model with annotations.
ner.train-curve	Train a model with different portions of the data (25%, 50%, 75%, 100%) to see how it improves.
ner.silver-to-gold	Convert binary accept/reject annotations to gold-standard data with no missing values.
review	Review annotations and resolve conflicts.

Entity recognition and transfer learning



Introduction to the lab