deepchecks.

Validating NLP Data and Models

https://github.com/Nadav-Barak/NLP_Webinar_Talk

About Us

- Nadav Barak
- ~5 Years in Research
  - Mathematics
  - Data Science

- Nir Hutnik
- ~10 Years in Research
  - Operations Research
  - Data Science

ML Researchers @ Deepchecks

About deepchecks.

- Founded 3 years ago
- Continuous Validation for ML Systems
- Open Source!
- NLP Alpha (Full Release Soon!)

OPEN SOURCE PACKAGE

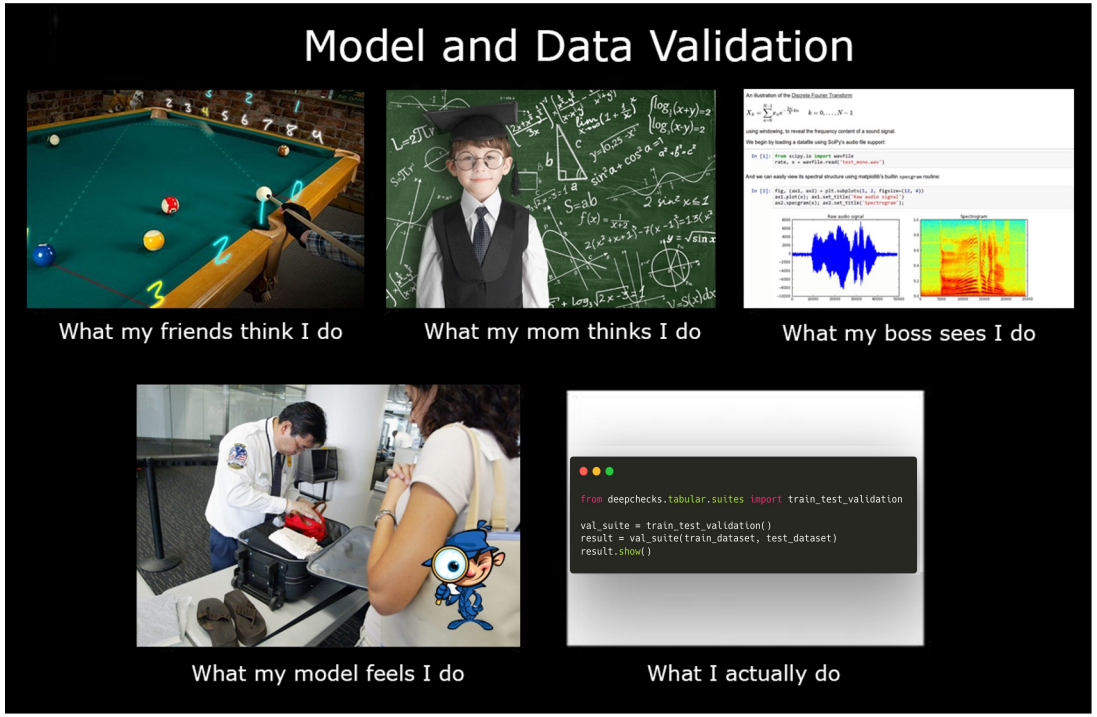
Test Suites for Offline Validation

DEEPCHECKS HUB

Production Monitoring

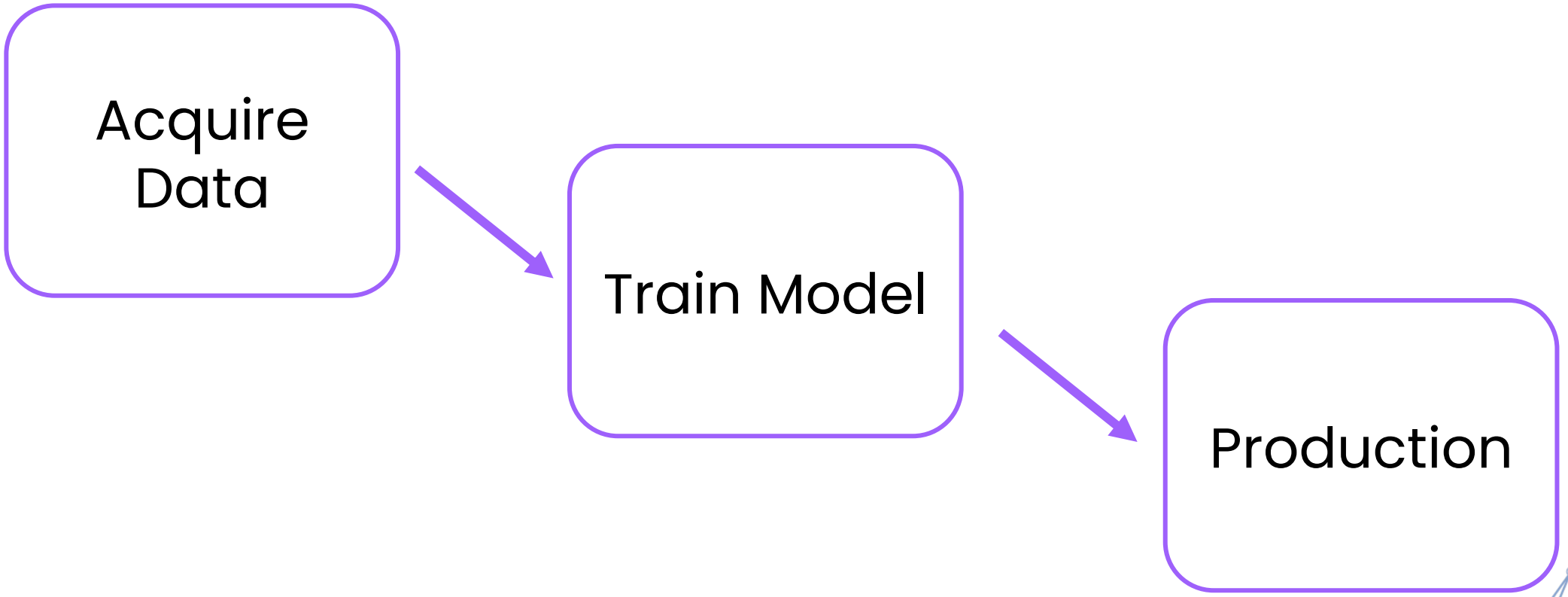
What will we talk about?

- Common Data & Model Issues
  - Drift
  - Shortcuts
  - Segment Degradation
- …And How to Find Them
  - Concepts: Metadata, Properties, Embeddings
  - Hands-On Demo!
- Future plans



Common Data & Model Issues

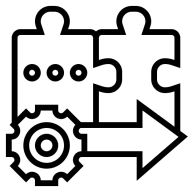
What Can Go Wrong?



What Can Go Wrong?

Dirty Data
Duplicates
Label Integrity

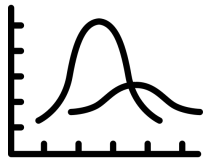
Data Integrity



Outliers
Shortcuts
Out-of-Vocabulary

Data Drift
Label Drift
Concept Drift

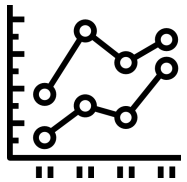
Data Distribution



Sampling Manner
Leakage
Anomalous Data

Performance Degradation
Weak Segments
Confidence

Model Behavior

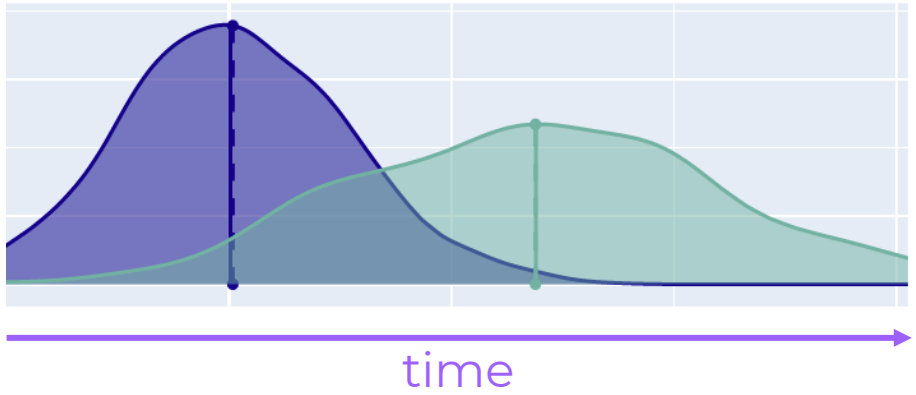


Fairness & Bias
Overfit
Calibration

What is Drift?

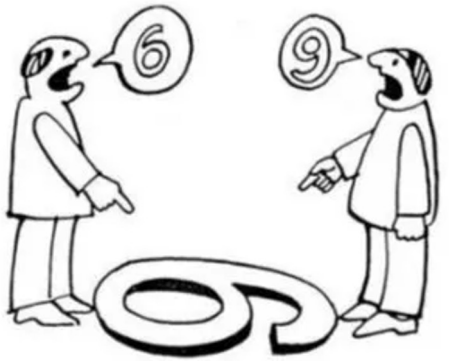
- Drift is a situation where the data, labels or their joint distribution changes over time.
- Can be categorized into 2 broad categories:

Data Drift



$$P_{t1}(X) \neq P_{t2}(X)$$

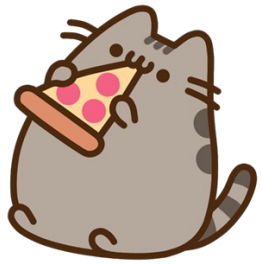
Concept Drift



$$P_{t1}(Y|X) \neq P_{t2}(Y|X)$$

Data & Concept Drift

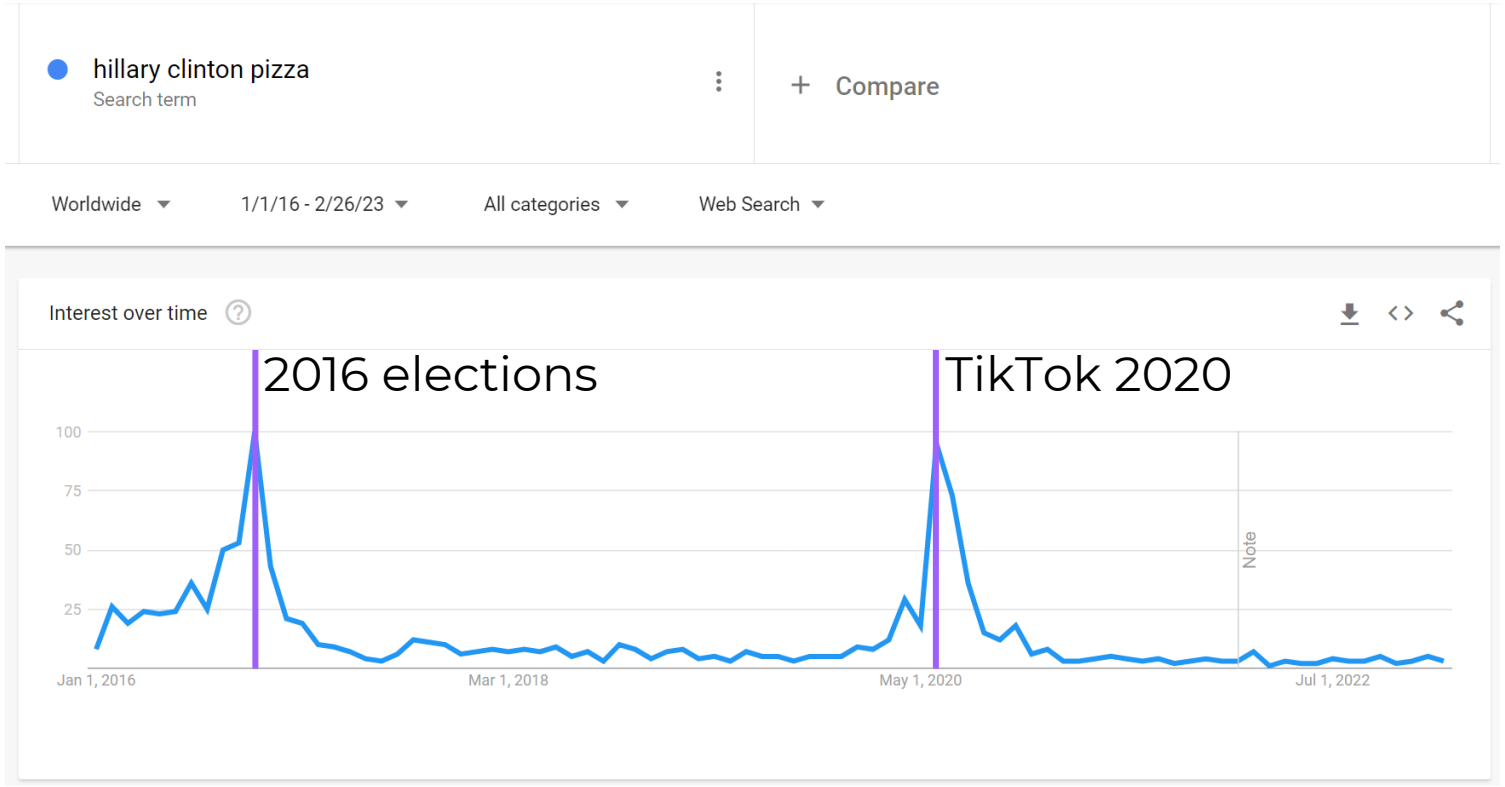
Before 2016:



Pizza -> Good!

After 2016:

Pizza -> ???



Shortcut Learning

“There is no scientific evidence supporting the claims made by NLP advocates, and it has been called a *pseudoscience*. Scientific reviews have shown that NLP is based on outdated metaphors of the brain's inner workings that are inconsistent with current neurological theory, and contain numerous factual errors. Reviews also found that research that favored NLP contained significant methodological flaws, and that there were three times as many studies of a much higher quality that failed to reproduce the "extraordinary claims" made by Bandler, Grinder, and other NLP practitioners.”

- Wikipedia

Long -> True

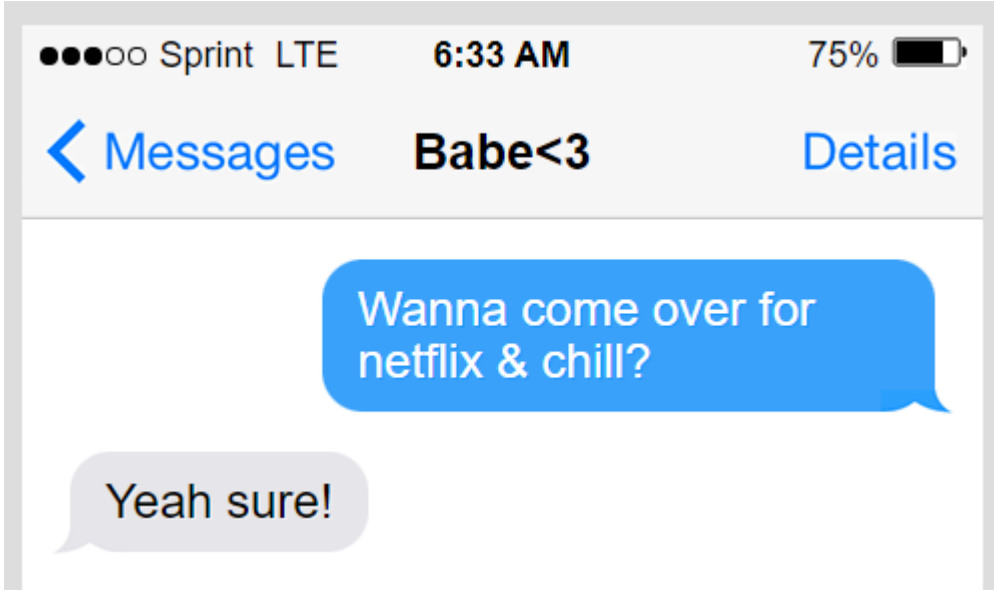
“NLP works!”

- My aunt

Short -> False

Model Degradation In Segment

- Model tries to understand brand awareness, based on textual interaction
- Oh my god! ***Young people*** really like ***Netflix***!



- But in reality

Streaming service **Netflix** is experiencing a decline in youth viewership in the UK, content. According to a report from **Enders Analysis**, in 2021, all age brackets un 2020.

... and How to Find Them

Data & Model Validations

Integrity

- Sticky keyssssss
- Gibberish
- Sample leakage

Keep unstructured

Distribution & Performance

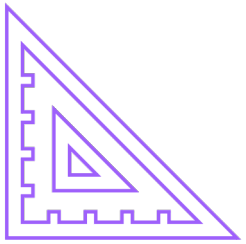
- Drift Detection
- Shortcuts
- Weak Segments

Transform to structured

Transform The Problem

Metadata

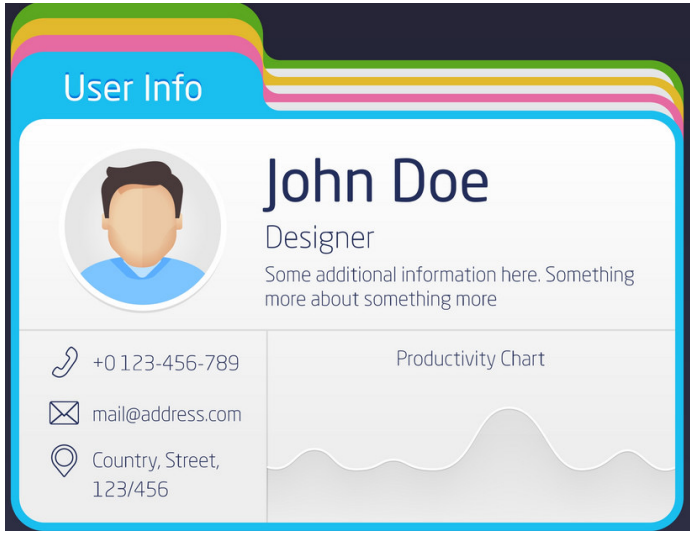
Properties



Embeddings

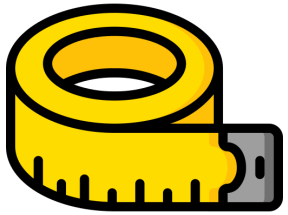
Metadata

- Metadata is any information that can be collected from external source on the text sample.
  - information on the person who wrote the text
  - details on the when and where it was published
  - what effect the text had.



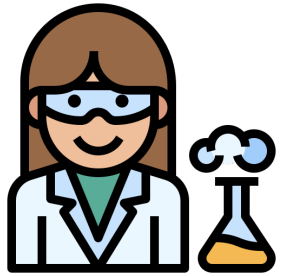
Properties

- Any information that can be derived from the text



- Technical:
  - Text length, average word length, number of special chars, TF-IDF

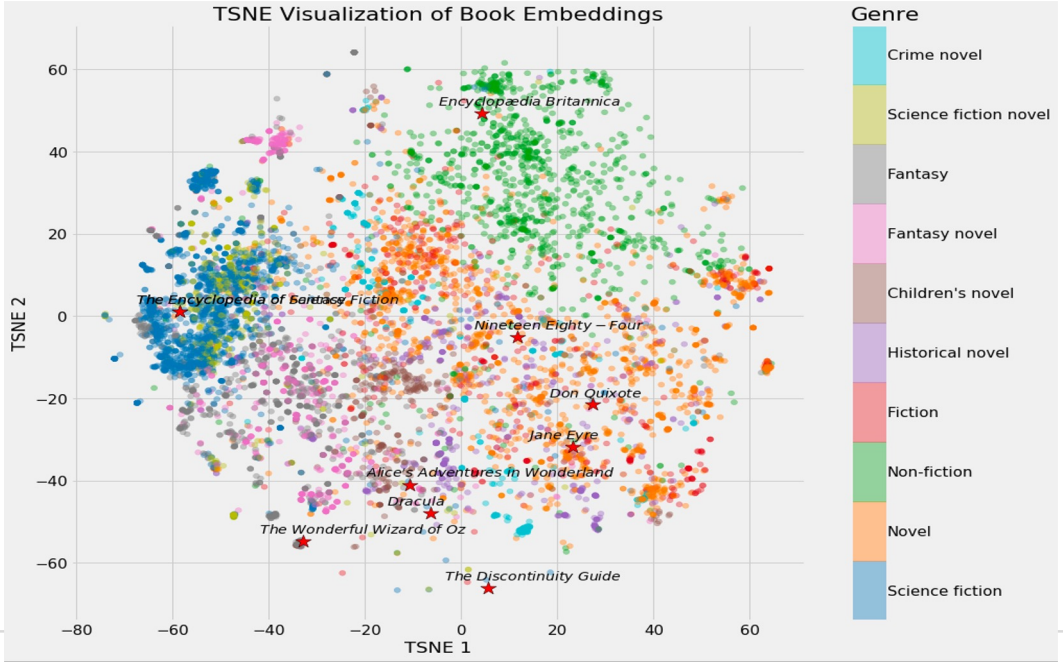
- Linguistic:
  - Language, number of nouns



- Domain specific:
  - Sentiment, register, topic

Embeddings

- Embeddings are a low-dimensional continuous **vector representation** of a sample. Quality embedding capture the essence of a sample such that semantically similar samples will have similar embedding vectors.
- In the context of neural networks, usually one of the final model layers are used as the embedding.

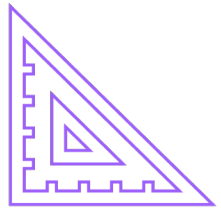


Transform The Problem

Metadata

- Accurate
- Good for segmentation
- Not always available
- Doesn’t describe the text

Properties



- Less accurate
- Good for drift and outliers
- Depends on domain
- Describes the text

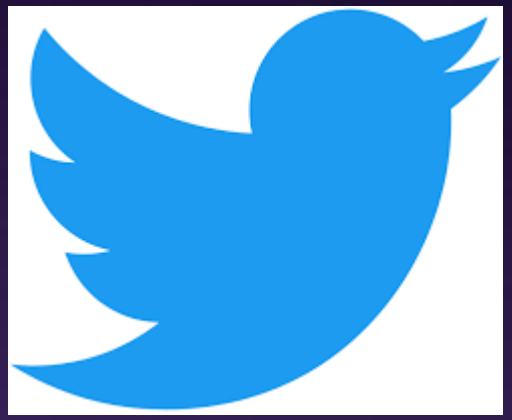
Embeddings

- Accurate
- Good for drift and outliers
- Low explainability
- Describes the text

deepchecks.

Hands-On Demo!

Tweets Emotion Detection



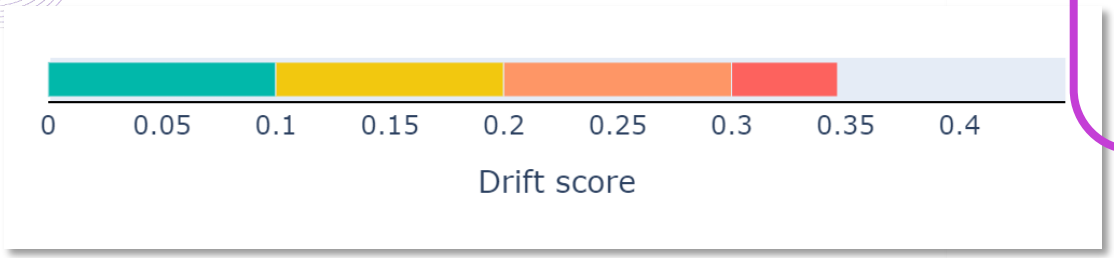
What's Next?

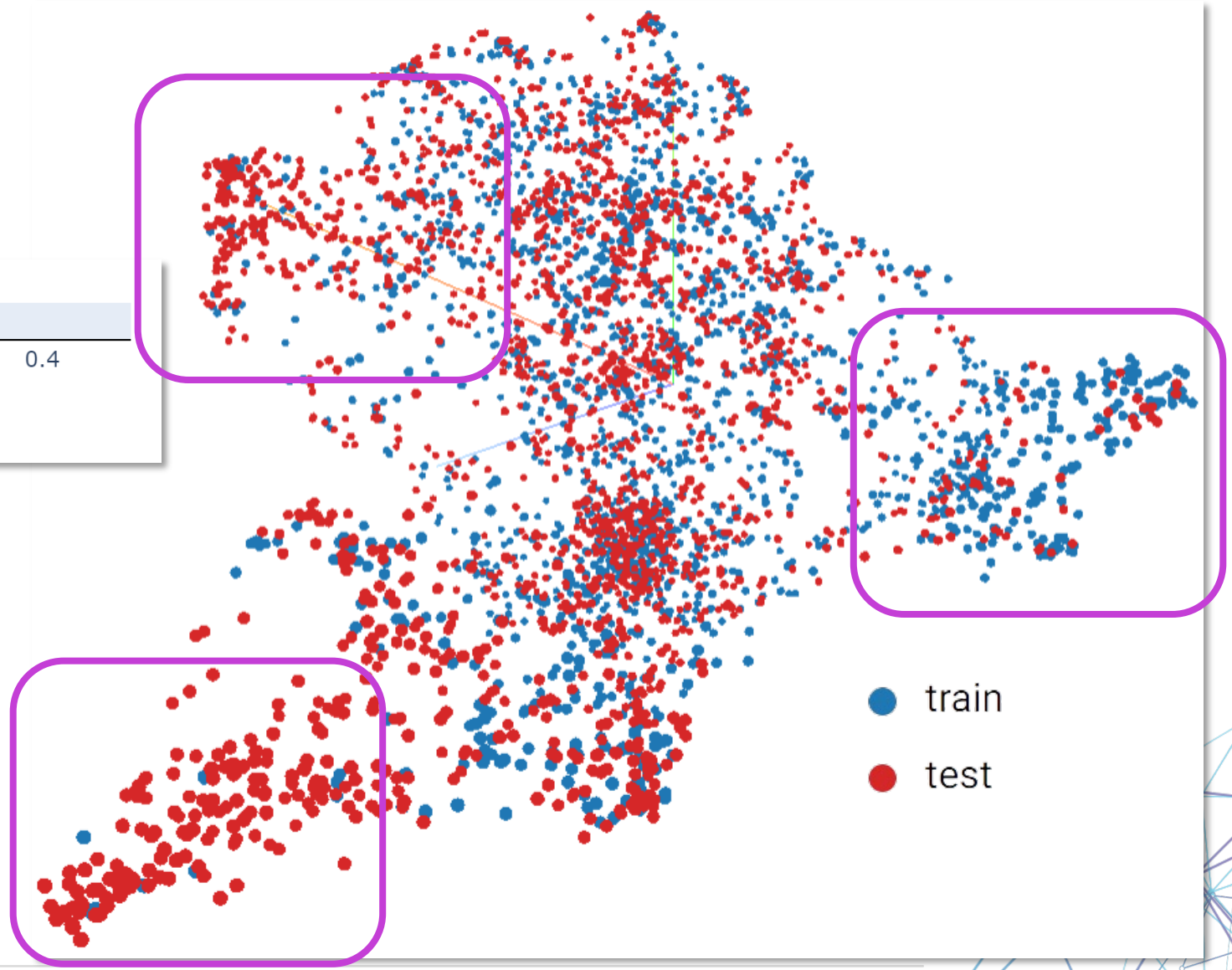
Adversarial Drift Detection

- “Domain Classifier”
- Discern between train and test

“Ok, I know how much, how do I know WHY?”

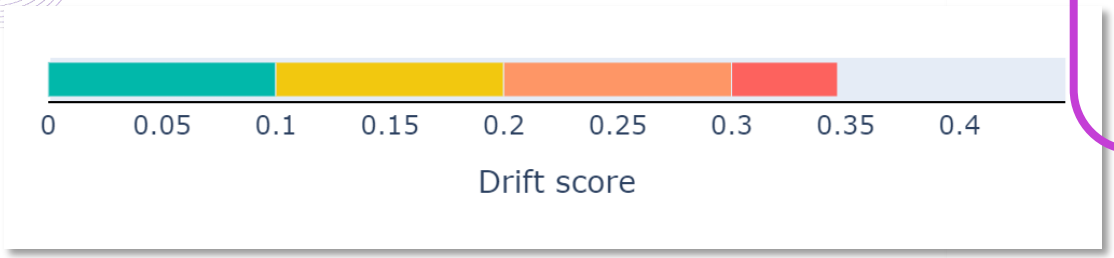
Finding Drift Using Embeddings

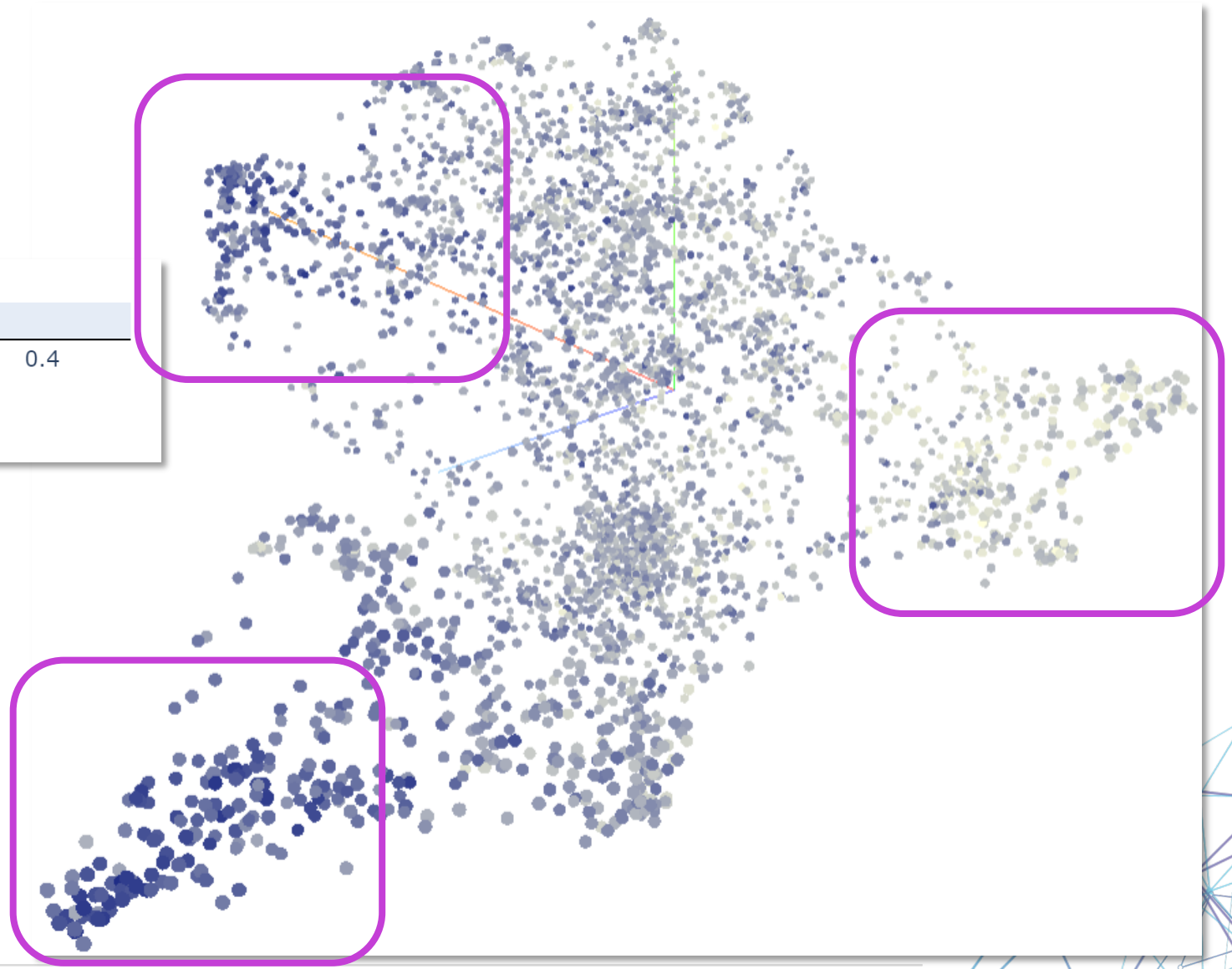




Embeddings

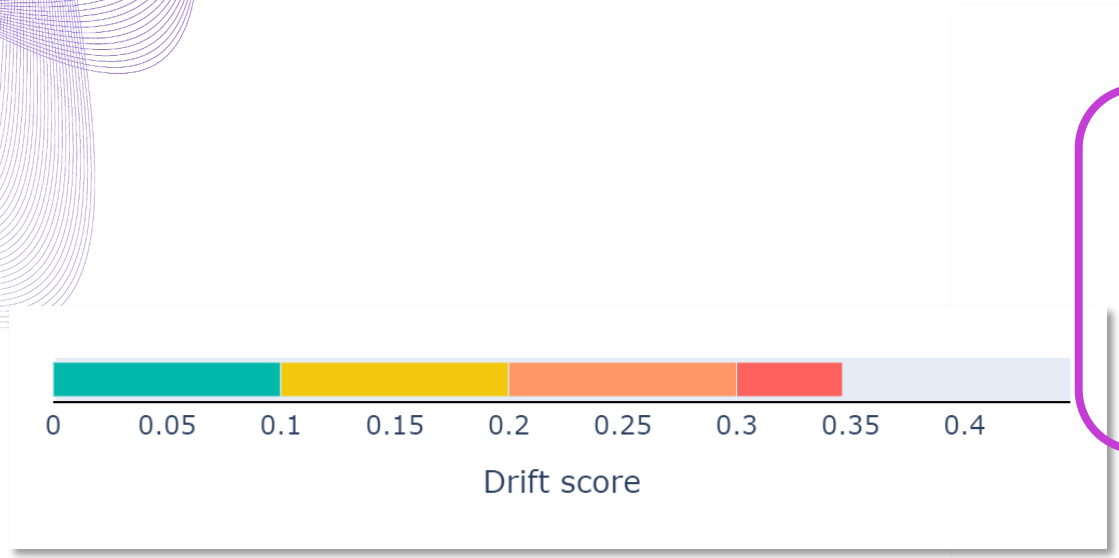
Finding Drift Using Embeddings



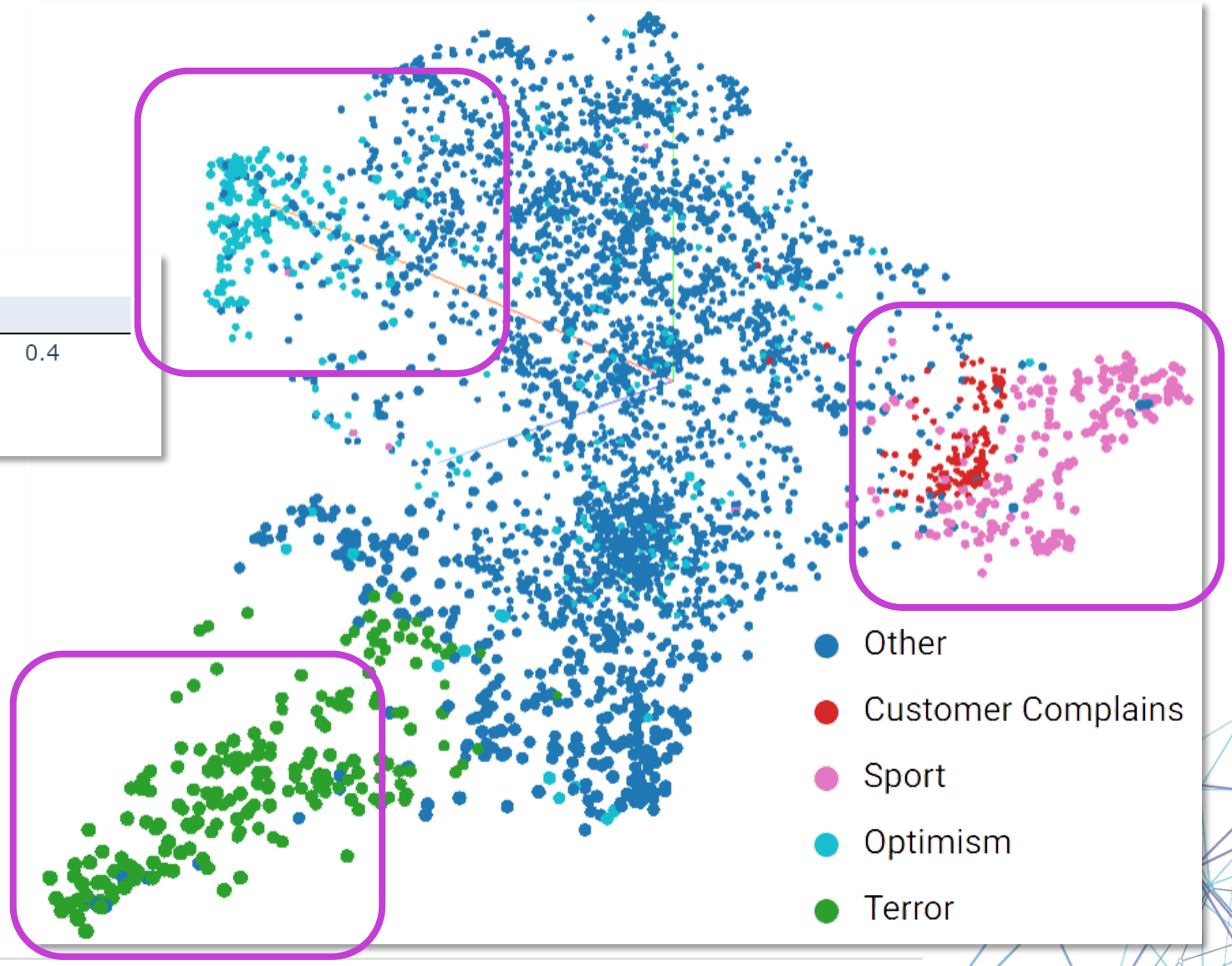


Embeddings

Finding Drift Using Embeddings



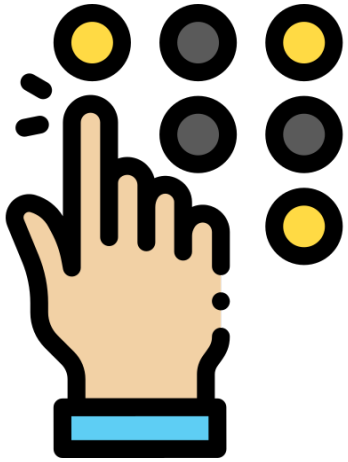
Embeddings



What's Next?

- Support additional use-cases
  - Token classification
  - Information retrieval
  - Properties on label/prediction

- Which samples to annotate next?



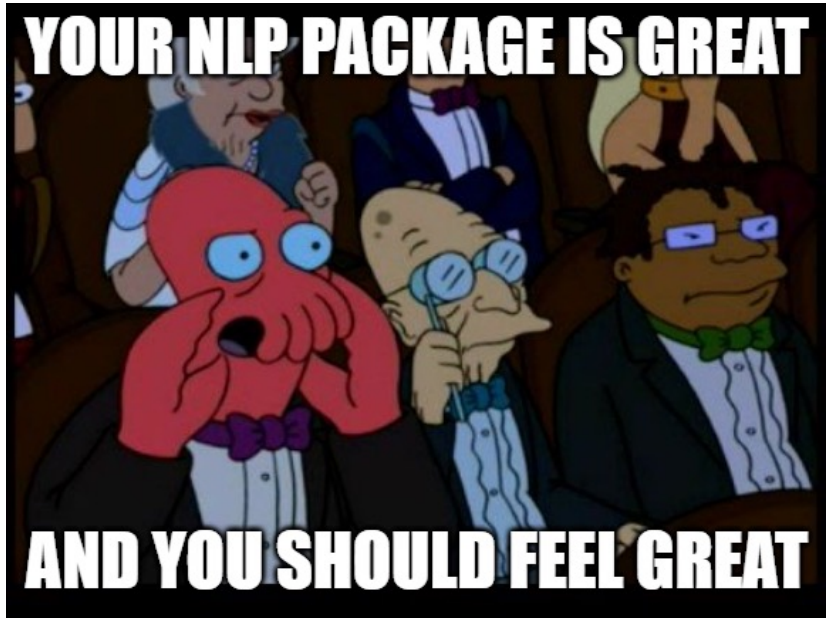
What's Next?

- Similar samples with conflicting labels

- Additional Integrity checks
  - Sticky keys, duplicates, Nones, etc.

Join The Party!

- Currently on open Alpha, full release on mid May
- Try it out and let us know what you think, what you would like to see next
- Our slack channel
- Talk to us:
  - nadav@deepchecks.com
  - yaron@deepchecks.com



Scan for more details & links

deepchecks.

Thank you for listening, and may the tests be with you

https://www.linkedin.com/in/nadavbarak/

https://github.com/deepchecks/deepchecks

👂 Feedback, ideas & feature requests are greatly appreciated! 🙏

and if you like what we're doing – give us a ⭐ on GitHub