### Week 3: Data Definition and Baseline

### Why is data definition hard

### \* labeling conventions must be consistent

### More label ambiguity

### \* Standardize labeling conventions

### \* User Id merge example – Link 2 entity, human label, same name and zip code

### \* Predicting is it a bot/spam a/c = ambiguous

### \* fraudulent txn = ground truth ambiguous

### \* looking for a job

### Data definition questions

### \* What is input.x? light, contrast, resolution, human should be able to perform

### \* For structured data problems, what features need to be included

### \* What is target label y?

### Major types of data problems

### 

### 

### Small data and label consistency

### Why is label consistency important?

### Smalldata = Noisy labels => hard to fit a curve

### Big data noisy labels => easier to fit a curve

### Small data, clean consistent labels, we can confidently predict

### Labelling instructions is important.

### Big data problems can have small data challenges too. Problems with a large dataset but where there is a long tail of rare events in the input will have small data challenges. E.g. Websearch

### Improving label consistency

### 

### Have a separate class for ambiguous labels. Say Alternative, BorderLine

### Small data – Small # 0f labelers can ask labelers to discuss specific labels

### Big Data – Consistent definition with a small group

### Send labeling instructions to labelers, consider having multiple labelers label every example and using voting or consensus labels to increase accuracy.

### Human Level Performance (HLP)

### 

### 

### Raising HLP

### 

### 

### Label and Organize data

### How long should you spend obtaining data?

### Get into data iteration loop as quickly as possible

### How much data can we obtain in k days

### List of data sources (create an inventory). Put a cost on time, and purchases

### Other factors – data quality, privacy , regulatory

### Labeling data – In-house, outsourced vs crowdsourced

### MLE labeling is expensive but can spend a few days labeling

### Subject matter expert, some may be impossible to label

### Don’t increase data by more than 10x at a time

### Data Pipeline / Data cascades

### Raw data -> data cleaning -> ML algo -> test set

### New data -> replicate scripts -> ML algo -> product prediction

### So data cleaning must be replicable once you move away from POC steps.Tools to use here are Apache Beam, Aiflow, Tensorflow transform.

### Meta-data, data provenance and lineage

### ML code -> anti-spam-model – despammed user data -> id merge model -> clean user data -> job search model -> predictions

### Data provenance – whom did the data come from

### Lineage – sequence of steps

### Tools for data provenance and lineage are non-existent

### Meta-data. – data about data, rarely regret storing metadata, as important as commenting code.

### Useful for error analysis, spotting unexpected effects, keep track of data provenance.

### Balanced train/dev/test splits

### Small data = keep all 3 test sets of same size. Balanced split

### Large dataset – random split works 60%, 20%, 20%

### SCOPING

### How to pick a project

### Better recommender system

### Better search

### Improve catalog

### Inventory mgmt.

### Price optimization

### What project should we work on?

### What are the metrics for success?

### What are the resources needed?

### Scoping process

### Brainstorm with business/product = top 3 things to do better

### Brainstorm AI solutions

### Assess feasibility and value of potential soln (diligence)

### Determine milestones

### Budget

### Diligence on feasibility and value

### 

### Why use HLP to benchmark – People are good at unstructured data. Given the same data can a human perform the task?

### Do we have features that are predictive

### History of project – Rate of previous historic improvements good predictor for future improvements

### Diligence on value

### MLE metrics, (accuracy, log likelihood)

### Query-level accuracy

### Search result quality (User engagement)

### 

### 

### Milestones and resourcing

### 

### Week 3: Reading

[Label ambiguity](https://csgaobb.github.io/Projects/DLDL.html)

[Data pipelines](https://cs230.stanford.edu/blog/datapipeline/#best-practices)

[Data lineage](https://blog.tensorflow.org/2021/01/ml-metadata-version-control-for-ml.html)

[MLops](https://cloud.google.com/blog/products/ai-machine-learning/key-requirements-for-an-mlops-foundation)

Geirhos, R., Janssen, D. H. J., Schutt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (n.d.). Comparing deep neural networks against humans: object recognition when the signal gets weaker∗. Retrieved May 7, 2021, from Arxiv.org website: <https://arxiv.org/pdf/1706.06969.pdf>