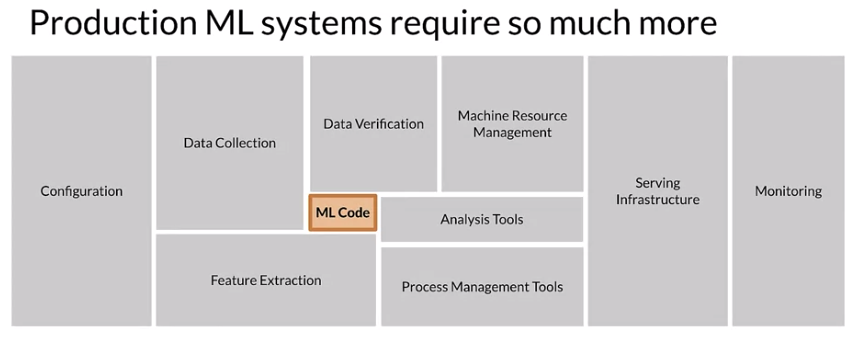
ML Data Lifecycle in Prod

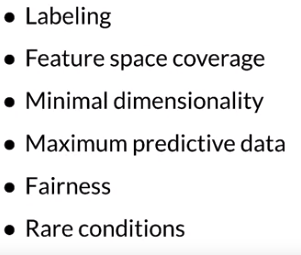
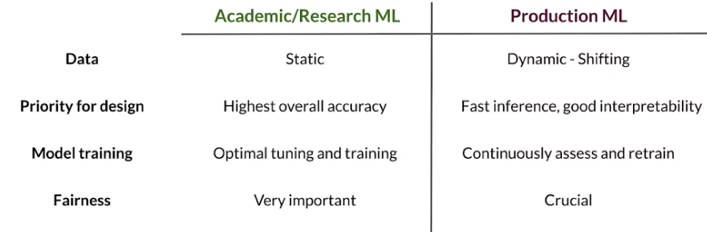
Week1

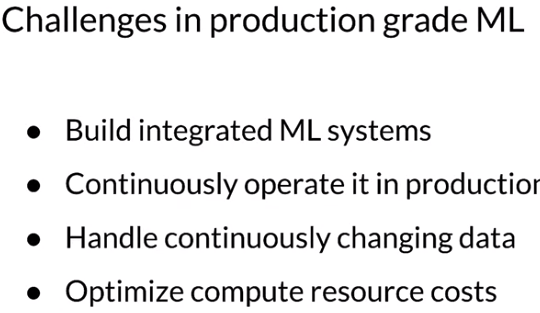
\* ML engineering for production

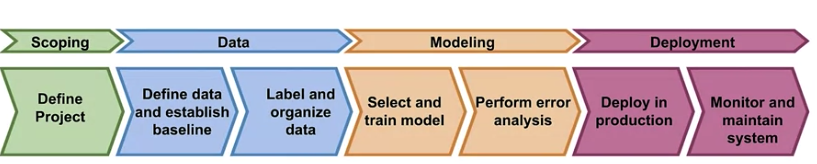
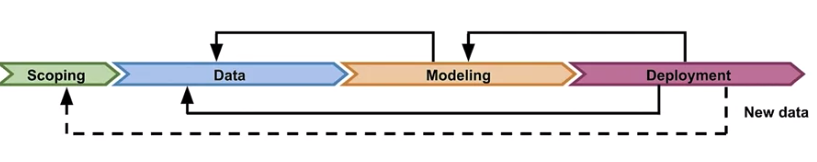
\* Production ML = ML dev + software development

\* Changes in prod ML





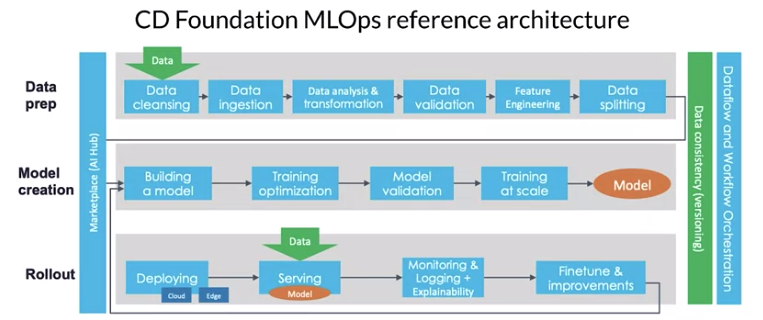
1. Scalable
2. Extensible
3. Configurable
4. Consistency & Reproducibility
5. Safety & Security
6. Modularity
7. Testability
8. Monitoring



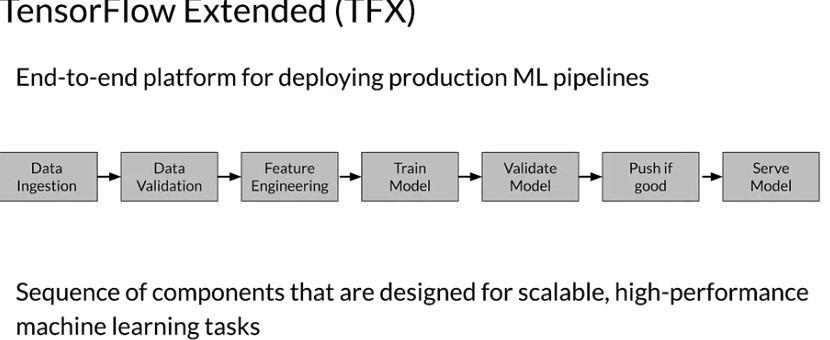
ML Pipelines

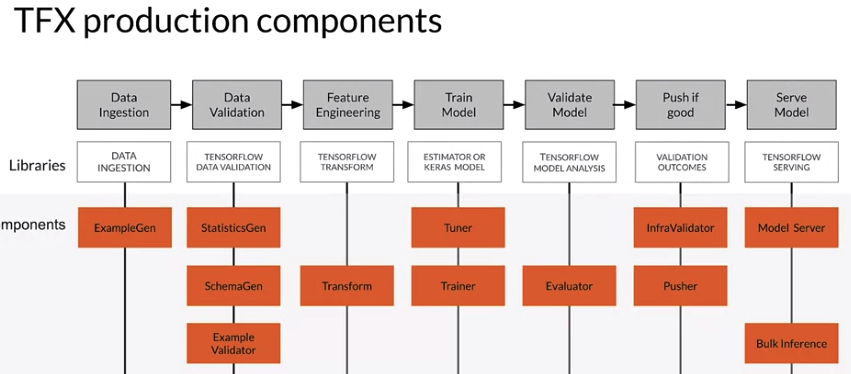
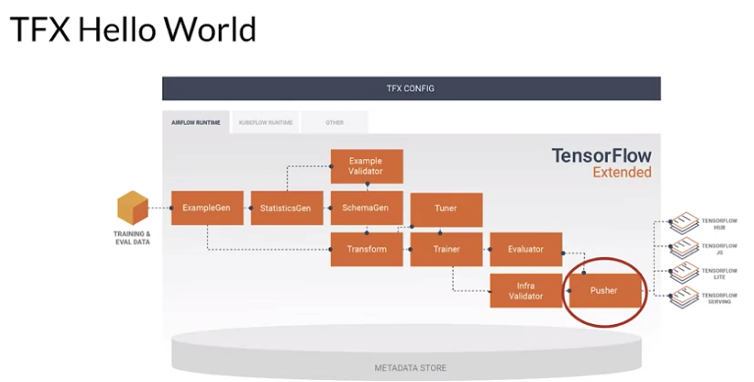
\* Infra for automating , monitoring and maintaining model training & dev

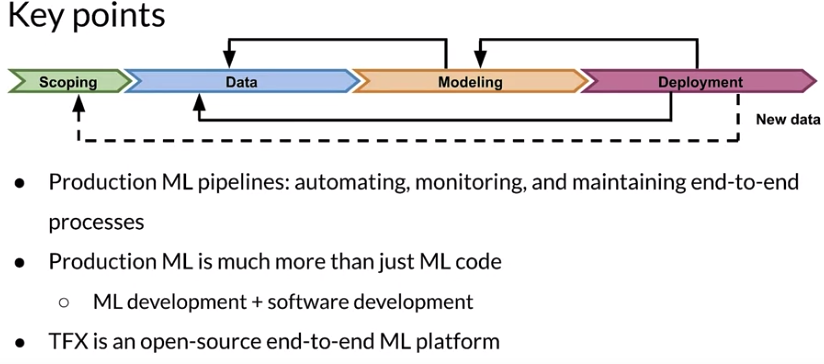
\*



DAGS Kubeflow, Aiflow

TFX 





COLLECTING DATA

Collect – ingest – prepare

Data is a first-class citizen

Data maximize predictive content, remove non-informative data, feature space coverage

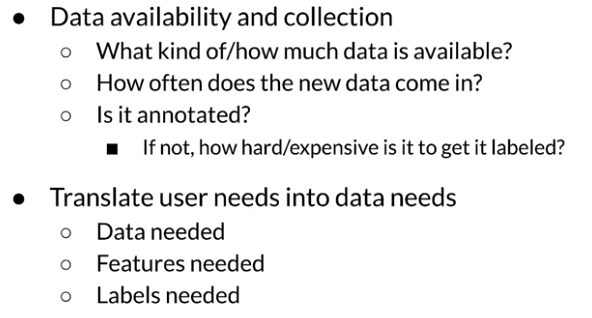
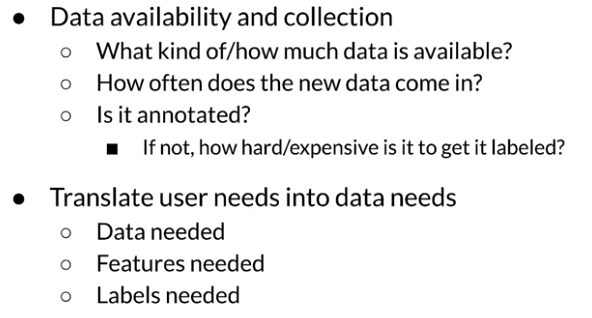
GIGO

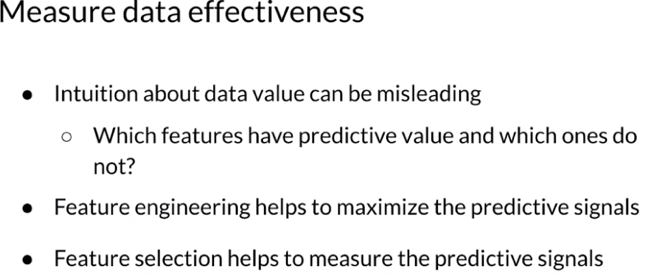
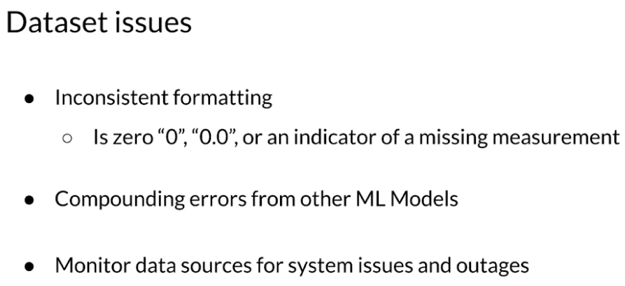
Avoid downtime, errors, distrib shifts, data Failure , service failure

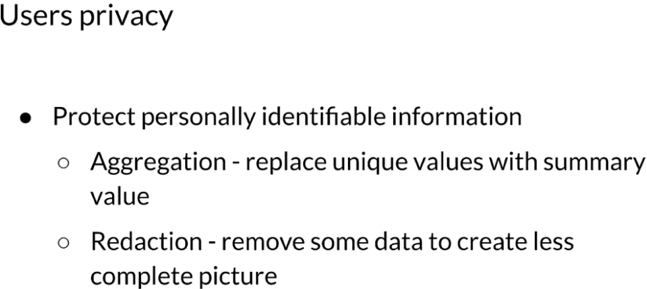
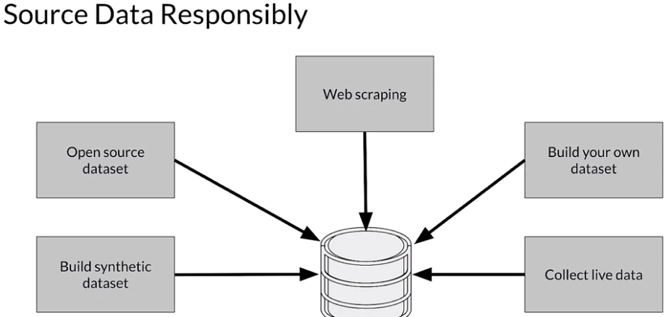
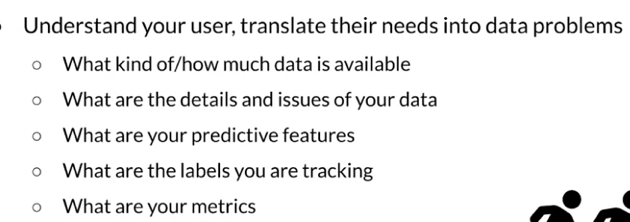
Understand users, translate user needs into data problems

Ensure data coverage and high pred signal

Source data responsibly





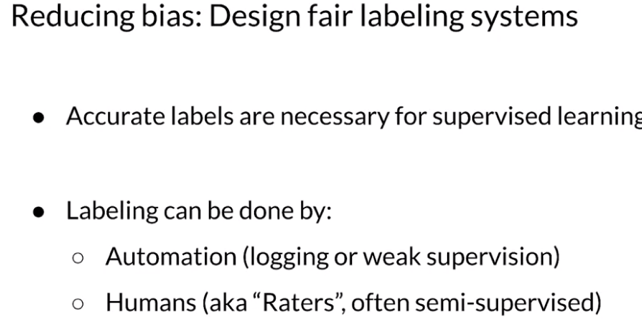
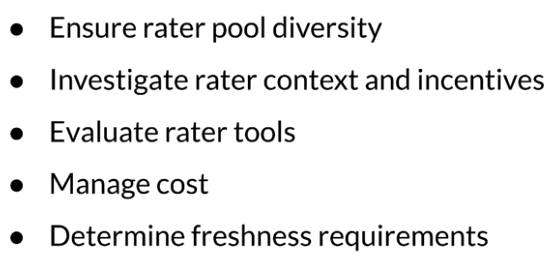
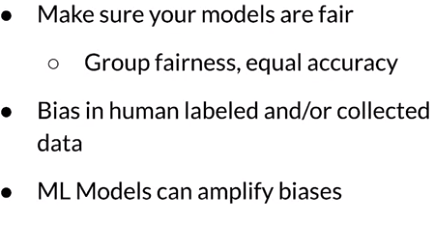


TYPES OF HUMAN RATERS

\* Generalists,

\* Subject matter experts

\* Your users



Labeling Data

\* How to detect problems early

Causes, how to solve

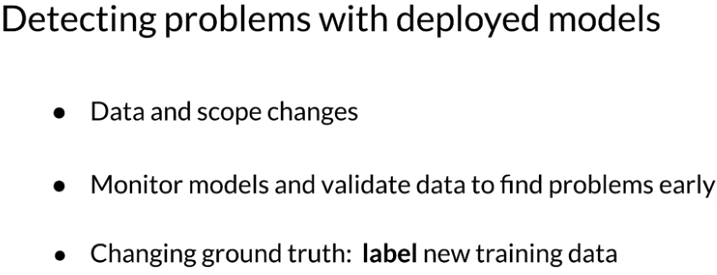
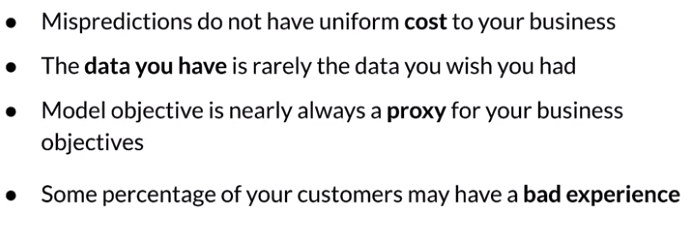
Slow problems – data change -Seasonality, Trend, Relative importance of features, distrib of feature changes

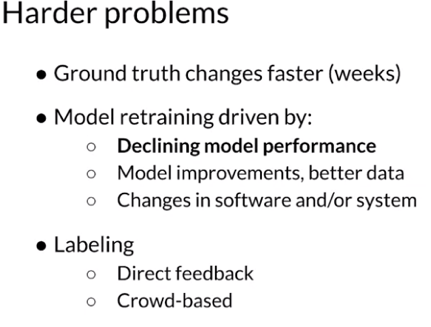
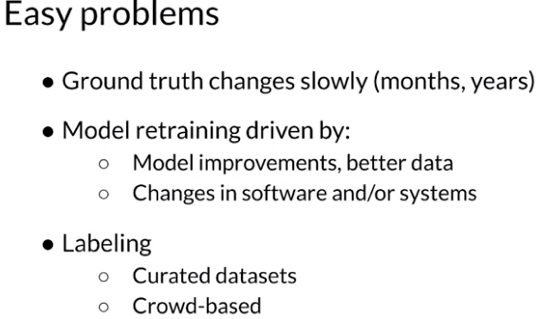
World change - Styles change, scope. Process change, competitors change, business expands to other geo

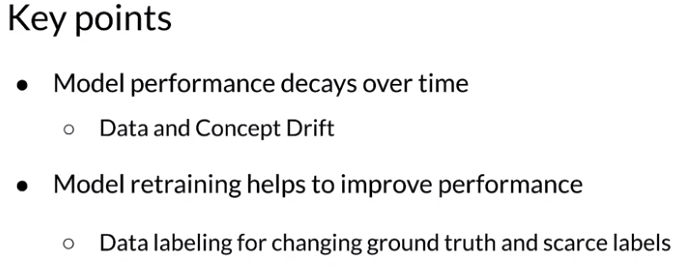
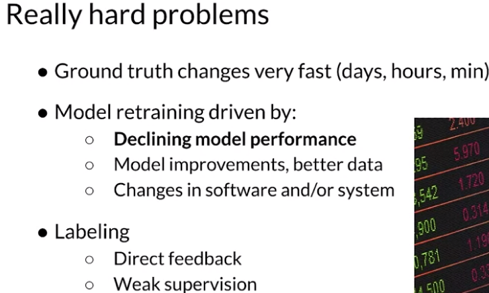
Fast problems

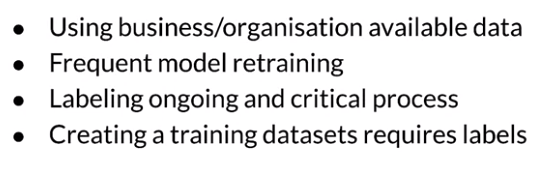
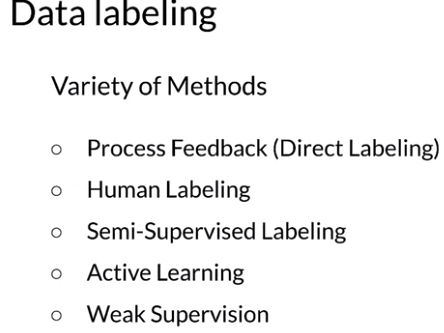
Data collection-logs are bad, bad sensor, moved sensor

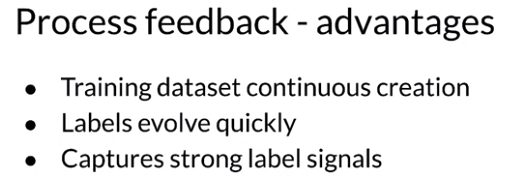
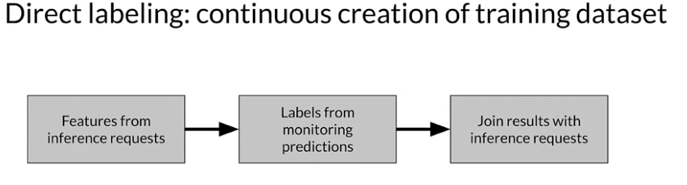
Systems problem – bad software update, loss of network connectivity, system down, bad creds

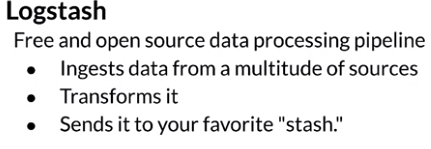












fluentd

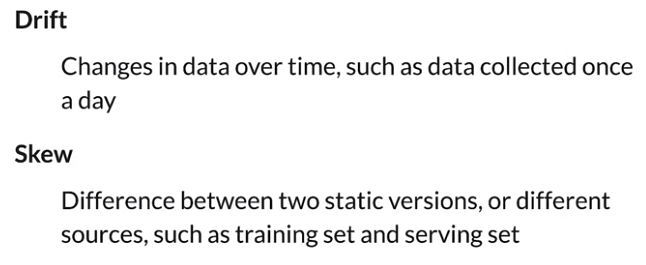
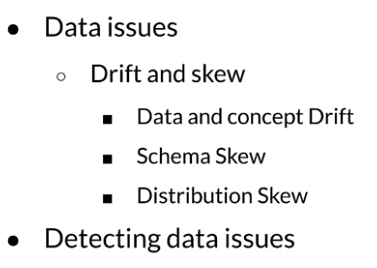
Google cloud logging

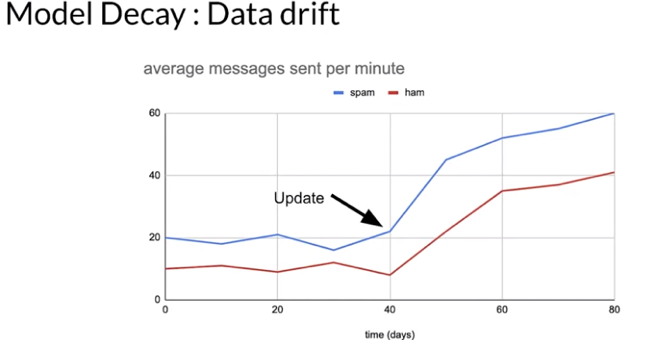
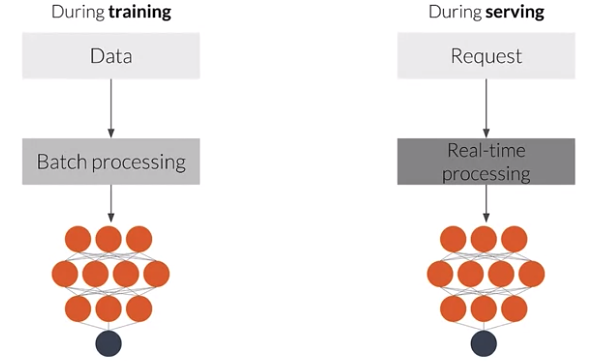
AWS elastic search

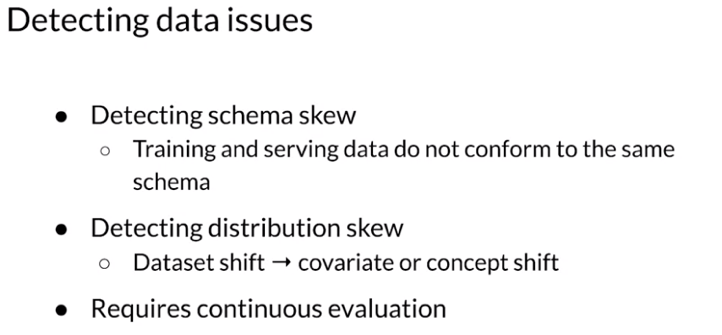
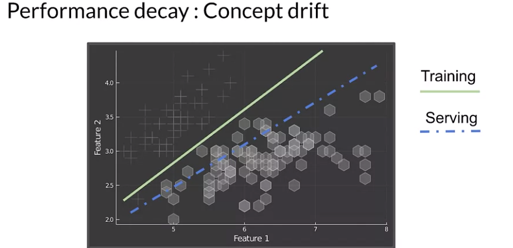
Azure monitor

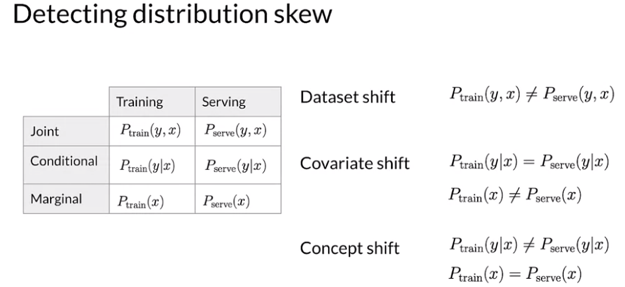
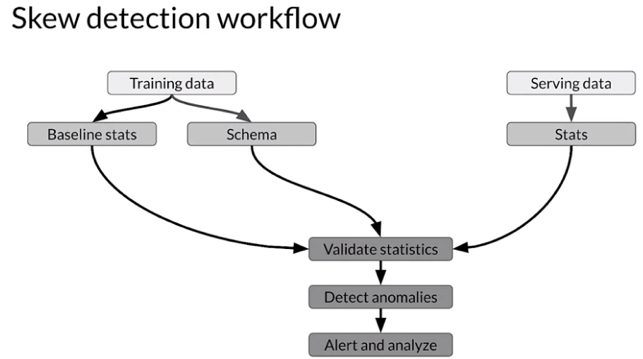
Human labeling – lack of consistency, expensive, slow, small dataset

VALIDATING DATA

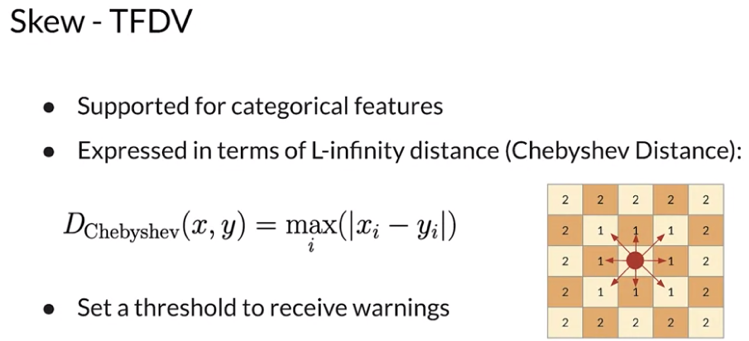
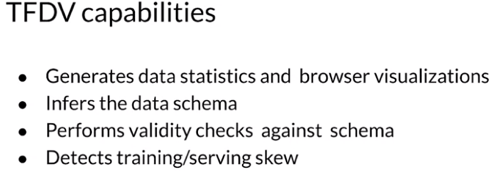


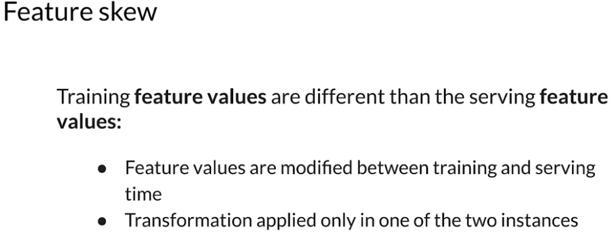
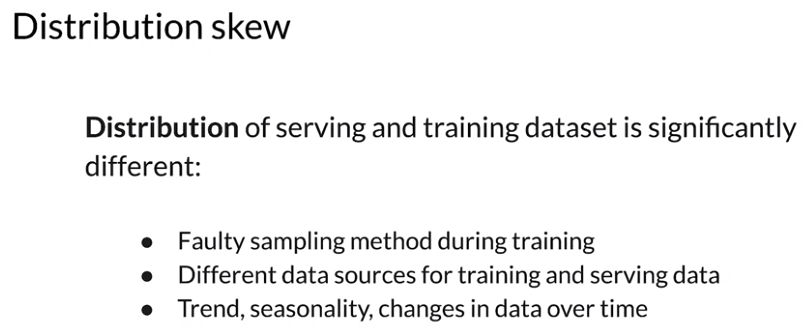


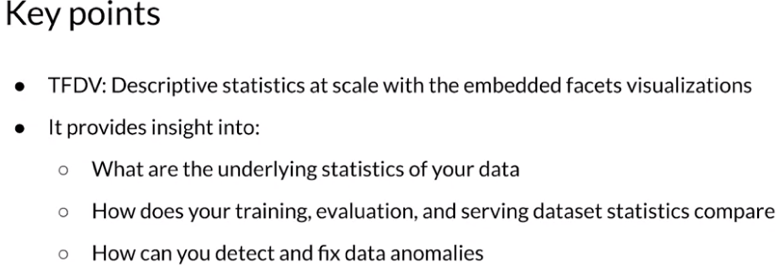




Tensorflow data validation (TFDV)







Week 1: Collecting, Labeling and Validating Data

This is a compilation of optional resources including URLs and papers appearing in lecture videos. If you wish to dive more deeply into the topics covered this week, feel free to check out these optional references. You won’t have to read these to complete this week’s practice quizzes.

[MLops](https://cd.foundation/blog/2020/02/11/announcing-the-cd-foundation-mlops-sig/)

[Data 1st class citizen](https://medium.com/@karpathy/software-2-0-a64152b37c35)

[Runners app](https://pair.withgoogle.com/chapter/data-collection/)

[Rules of ML](https://developers.google.com/machine-learning/guides/rules-of-ml)

[Bias in datasets](https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html)

[Logstash](https://www.elastic.co/logstash)

[Fluentd](https://www.fluentd.org/)

[Google Cloud Logging](https://cloud.google.com/logging/)

[AWS ElasticSearch](https://aws.amazon.com/elasticsearch-service/)

[Azure Monitor](https://azure.microsoft.com/en-us/services/monitor/)

[TFDV](https://blog.tensorflow.org/2018/09/introducing-tensorflow-data-validation.html)

[Chebyshev distance](https://en.wikipedia.org/wiki/Chebyshev_distance)

**Papers**

Konstantinos, Katsiapis, Karmarkar, A., Altay, A., Zaks, A., Polyzotis, N., … Li, Z. (2020). Towards ML Engineering: A brief history of TensorFlow Extended (TFX). <http://arxiv.org/abs/2010.02013>

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2020). Challenges in deploying machine learning: A survey of case studies. <http://arxiv.org/abs/2011.09926>

ML code fraction:

Sculley, D., Holt, G., Golovin, D., Davydov, E., & Phillips, T. (n.d.). Hidden technical debt in machine learning systems. Retrieved April 28, 2021, from Nips.cc <https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf>