Machine Learning in Production

Data Quality

Midterm

One week from today, here

More Quality Assurance...

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:
System and model goals
User requirements
Environment assumptions
Quality beyond accuracy
Measurement
Risk analysis
Planning for mistakes

Architecture + design:
Modeling tradeoffs
Deployment architecture
Data science pipelines
Telemetry, monitoring
Anticipating evolution
Big data processing
Human-AI design

Quality assurance:
Model testing
Data quality
QA automation
Testing in production
Infrastructure quality
Debugging

Operations:
Continuous deployment
Contin. experimentation
Configuration mgmt.
Monitoring
Versioning
Big data
DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance, versioning, reproducibility

Safety

Security and privacy

Fairness

Interpretability and explainability

Transparency and trust

Ethics, governance, regulation, compliance, organizational culture

Readings

Required reading:

- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021, May). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In Proc. Conference on Human Factors in Computing Systems (pp. 1-15).

Recommended reading:

- Schelter, S., et al. Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), pp.1781-1794.

Learning Goals

- Consider data quality as part of a system; design an organization that values data quality
- Distinguish precision and accuracy; understanding the better models vs more data tradeoffs
- Use schema languages to enforce data schemas
- Design and implement automated quality assurance steps that check data schema conformance and distributions
- Devise infrastructure for detecting data drift and schema violations

Poor Data Quality has Consequences

(often delayed, hard-to-fix consequences)

GIGO: Garbage in, garbage out

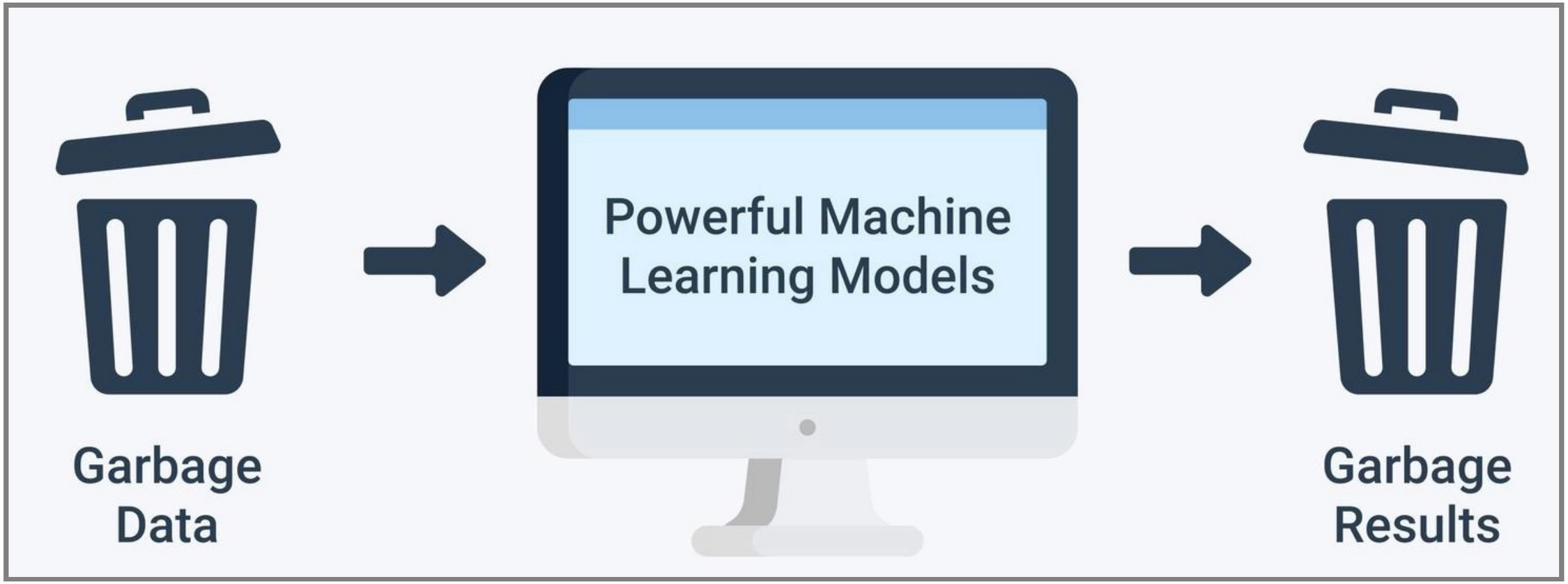


Image source: https://monkeylearn.com/blog/data-cleaning-python

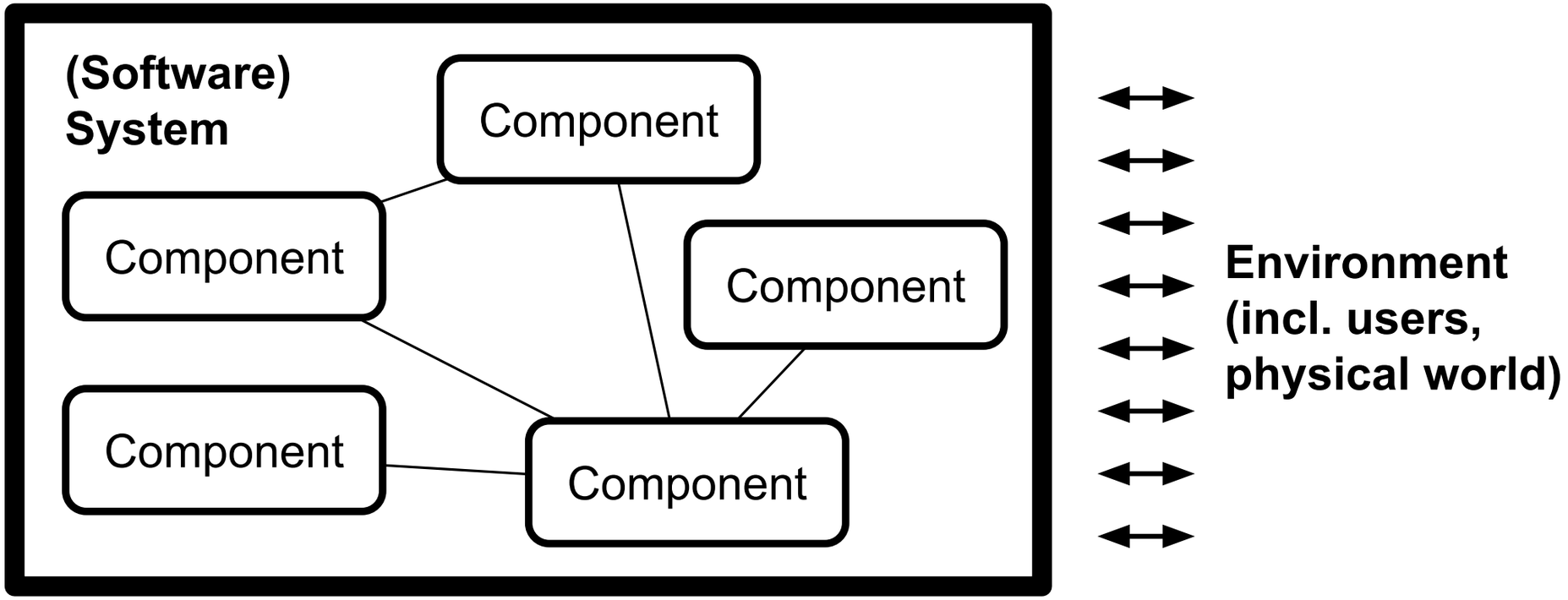
Example: Systematic bias in labeling

Poor data quality leads to poor models

Often not detectable in offline evaluation - **Q. why not**?

Causes problems in production - now difficult to correct

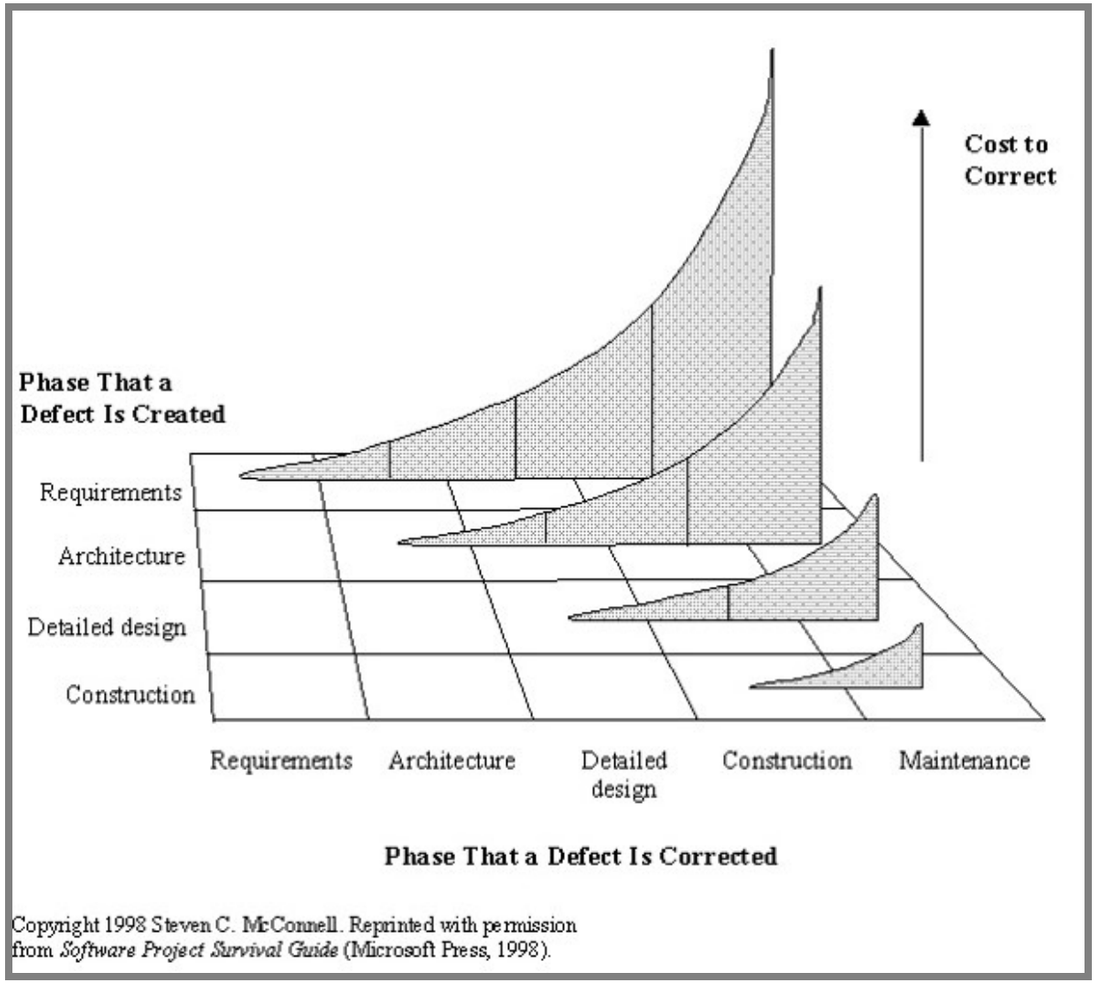
Data Quality is a System-Wide Concern



Delayed Fixes increase Repair Cost

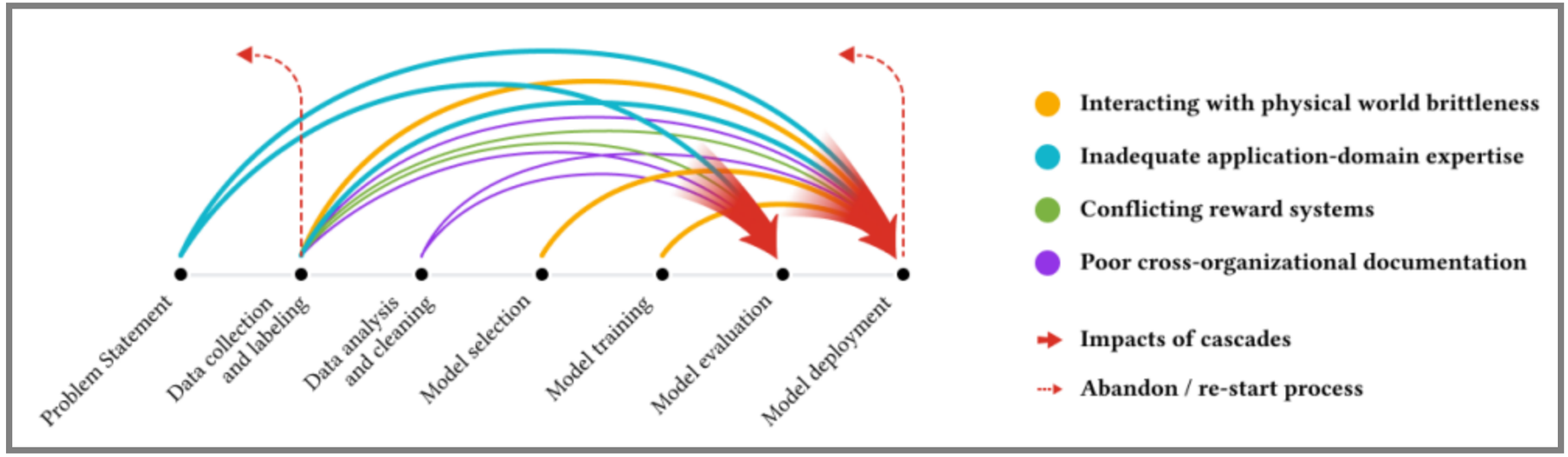
Cost of bug repair depending on when the bug was introduced and fixed

- If you don't fix it early enough the cost gets higher and higher
- Same for data quality issues!



Data Cascades

"Compounding events causing negative, downstream effects from data issues, that result in technical debt over time."



Sambasivan, N., et al. (2021, May). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In Proc. CHI (pp. 1-15).

Common Data Cascades

Physical world brittleness

- Idealized data, ignoring realities and change of real-world data
- Static data, one time learning mindset, no planning for evolution

Inadequate domain expertise

- Not understand. data and its context
- Involving experts only late for trouble shooting

Conflicting reward systems

- Missing incentives for data quality
- Not recognizing data quality importance, discard as technicality
- Missing data literacy with partners

Poor (cross-org.) documentation

- Conflicts at team/organization boundary
- Undetected drift

Sambasivan, N., et al. (2021). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In Proc. Conference on Human Factors in Computing Systems.

Interacting with physical world brittleness

Brittle deployments interacting with not-digitised physical worlds

- **Time to manifest**: 2-3 years to emerge, almost always in the production stage
- **Impact**: Complete model failure, abandonment of projects, harms to beneficiaries from mispredictions
- **Triggers**: drifts (formally defined more later!)
  - Hardware drifts (Rain, wind, fingerprints, shadows)
  - Environmental drifts (lighting, temperature, humidity)
  - Social drifts (new regulations, new user behaviors)
- **Address**: Monitor data source, retrain models, introduce noise in data

e.g. an AI model for the COVID-19 pandemic on day 1 versus day 100 required a total change in various assumptions since the pandemic and human responses were volatile and dynamic

Inadequate application-domain expertise

> *AI practitioners are responsible for data sense-making in contexts in which they do not have domain expertise.*

- **Time to manifest**: After building models through client feedback & system performance
- **Impact**: Costly modification (improve labels, collect more data), Unanticipated downstream impacts
- **Triggers**:
  - Subjectivity in groundtruths (e.g. Decision history on claims of insurance companies)
  - Poor application-domain expertise in finding representative data (e.g. Cannot take 90% of the data from one hospital and generalise for the entire world!)
- **Address**: Faithfully document data sources, involve domain experts in data collection

Conflicting Reward Systems

Misaligned incentives and priorities between practitioners, domain experts, and field partners.

- **Time to manifest**: Model deployment
- **Impact**: Costly iterations, moving to an alternate data source, quitting the project
- **Triggers**: Need annotation but...
  - Inserted as extraneous work
  - Not compensated well
  - Competing priority with partners' primary responsibility
- **Address**: Provide incentives & training

e.g., when a clinician spends a lot of time punching in data, not paying attention to the patient, that has a human cost

Poor Cross-organisational Documentation

Lack of documentation across various cross organisational relations, causing lack of understanding on metadata

- **Time to manifest**: manual reviews, by "chance"
- **Impact**: Wasted time and effort from using incorrect data, being blocked on building models, and discarding subsets or entire datasets
- **Triggers**:
  - inherited datasets lacked critical details
  - feld partners not being aware of constraints in achieving good quality Al
- **Address**: Create a data curation plan in advance and take ample field notes in order to create reproducible assets for data

e.g., a lack of metadata and collaborators changing schema without understanding context led to a loss of four months of precious medical robotics data collection."

Case Study: Inventory Management

Goal: Train an ML model to predict future sales; make decisions about what to (re)stock/when/how many...

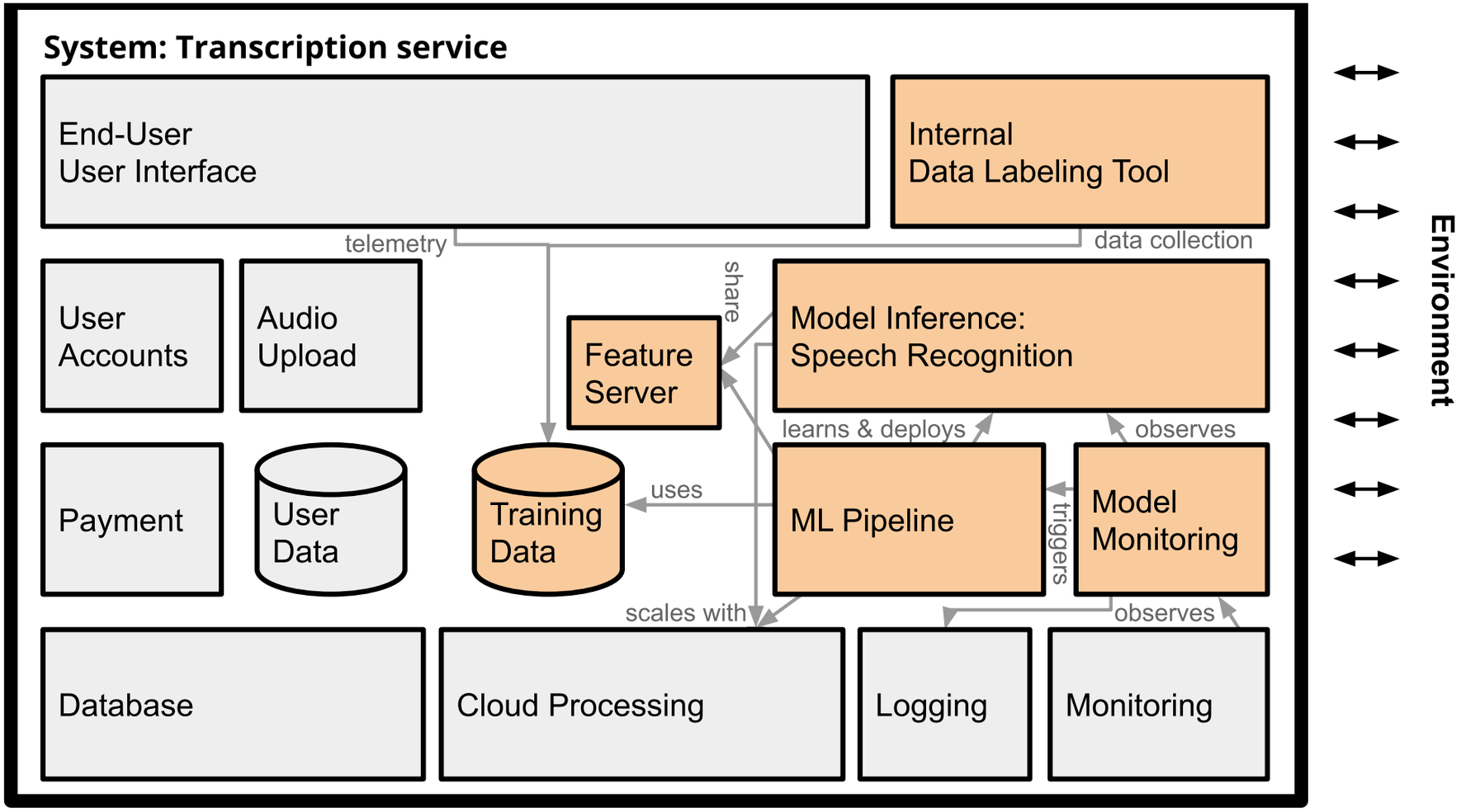
Discussion: Possible Data Cascades?

- Interacting with physical world brittleness
- Inadequate domain expertise
- Conflicting reward systems
- Poor (cross-organizational) documentation

Data Documentation

Let's use data documentation as an entry point to discuss what aspects of data we care about.

Data flows across components



Data Quality is a System-Wide Concern

Data flows across components, e.g., from user interface into database to crowd-sourced labeling team into ML pipeline

Humans interacting with the system

- Entering data, labeling data
- Observed with sensors/telemetry
- Incentives, power structures, recognition

Organizational practices

- Value, attention, and resources given to data quality

Documentation at the interfaces is important

Data Quality Documentation

Teams rarely document expectations of data quantity or quality

Data quality tests are rare, but some teams adopt defensive monitoring

- Local tests about assumed structure and distribution of data
- Identify drift early and reach out to producing teams

Several ideas for documenting distributions, including Datasheets and Dataset Nutrition Label

- Mostly focused on static datasets, describing origin, consideration, labeling procedure, and distributions; Example

Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64, no. 12 (2021).
Nahar, Nadia, et al. “Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process.” In Pro. ICSE, 2022.

Data Card

Example Data Card (excerpt)

Labelling Methods

LABELING METHOD(S)	LABEL TYPES AND SOURCES	LABEL DESCRIPTION
Human labels	Bounding boxes: Human annotators Perceived age range and gender presentation: Human annotators	Bounding boxes were created around *all* people in an image and perceived age ranges as well as perceived gender presentation were labeled.
LABEL TYPE: Bounding boxes	**LABEL TASK(S)** • Create the bounding box around all people • Label object attributes **LABELLER DESCRIPTION(S)** • Compensated workers based out of India	**LABEL DESCRIPTION** A rectangular bounding box around each person in an image. **LABELING TASK OR PROCEDURE** Annotators were asked to place boxes around all people in an image. If there were 5 or more people grouped together a single box was used and a *group of* attribute was associated with that box. Annotators were asked if the person inside of the box was *truncated*, *occluded*, or *inside of* something. They were also asked if the person inside of the box was a *depiction of* a person (such as a painting or figurine).
LABEL TYPE: Perceived gender presentation and age range	**LABEL TASK(S)** • Label the perceived gender presentation • Label the perceived age range **LABELLER DESCRIPTION(S)** • Compensated workers based out of India	**LABEL DESCRIPTION** Perceived gender presentation: *predominantly feminine, predominantly masculine, unknown* Perceived age range: *young, middle, older, unknown* Note that gender presentation for people marked as *young* is always set to *unknown*. **LABELING TASK OR PROCEDURE** Annotators were asked to select either *predominantly feminine*, *predominantly*

Speaker notes

Source: https://storage.googleapis.com/openimages/open_images_extended_miap/Open%20Images%20Extended%20-%20MIAP%20-%20Data%20Card.pdf

The Dataset Creator and Purpose

How to Use the Dataset

Use of Dataset

SAFETY OF USE

Conditional Use

There are some known unsafe applications.

UNSAFE APPLICATION(S)

⚠ Gender classification | Age classification

UNSAFE USE CASE(S)

This dataset **should not** be used to create gender or age classifiers. The intention of percieved gender and age labels is to capture gender and age presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender or actual age.

CONJUNCTIONAL USE

Safe to use with other datasets

KNOWN CONJUNCTIONAL DATASET(S)

- The data in this dataset can be combined with Open Images V6

KNOWN CONJUNCTIONAL USES

Analyzing bounding box annotations not annotated under the Open Images V6 procedure.

METHOD

Object Detection

SUMMARY

A person object detector can be trained using the Object Detection API in Tensorflow.

KNOWN CAVEATS

If this dataset is used in conjunction with the original Open Images dataset, negative examples of people should only be pulled from images with an explicit negative person image level label.

The dataset does not contain any examples not annotated as containing at least one person by the original Open Images annotation procedure.

METHOD

Fairness Evalutaion

SUMMARY

Fairness evaluations can be run over the splits of gender presentation and age presentation.

KNOWN CAVEATS

There still exists a gender presentation skew towards unknown and predominantly masculine, as well as an age presentation range skew towards middle.

Data field overview

Dataset Snapshot

PRIMARY DATA TYPE(S)

Non-Sensitive Public Data about people

DATASET SNAPSHOT

Total Instances	100,000
Training	70,000
Validation	7,410
Testing	22,590
Total boxes	454,331
Total labels	908,662
Average labels per image	9.08
Human annotated labels	All

DESCRIPTION OF CONTENT

Bounding boxes of people with perceived gender presentation attributes (*predominantly feminine, predominantly masculine, unknown*) and age range presentation attributes (*young, middle, older, unknown*). This adds adds nearly 100,000 new boxes that were not annotated under the original labeling pipeline of the core Open Images Dataset.

Note: All annotated images included at least one person bounding box in Open Images v6. 30,474 of the 100k images contain a MIAP-annotated bounding box with no corresponding annotation in Open Images. Almost 100,000 of the bounding boxes have no corresponding annotation in Open Images. Attributes were annotated for all boxes.

PRIMARY DATA MODALITY

Labels or Annotations

KNOWN CORRELATION(S)

- Gender presentation numbers are skewed towards predominantly perceived as **`masculine`** & **`unknown`**
- Age range presentation range numbers are skewed towards **`middle`**
- Perceived gender presentation is **`unknown`** for all bounding boxes with age range attribute annotated **`young`**

HOW TO INTERPRET A DATAPOINT

Each datapoint includes a bounding box denoted by XMin, XMax, YMin, and YMax in normalized image coordinates. The next five attributes (IsOccluded through IsInsideOf) follow the definitions from Open Images V6.

The **last two values** for each datapoint correspond to the gender presentation attribute and an age range presentation attribute, respectively.

Each annotation is linked to an Open Images key pointing to an image that can be found in Common Visual Data Foundation (CVDF) repository.

Datapoint Example

EXAMPLE OF ACTUAL DATA POINT WITH DESCRIPTIONS

Field	Value	Description
ImageID	164b0e6d1fcf8e61	The image this box lives in
LabelName	/m/01g317	Labels are identified by MIDs (Machine-generated Ids) as can be found in Freebase or Google Knowledge Graph API. Label descriptions here
Confidence	1	A dummy value, always 1
XMin	0.897112	Normalized image coordinates indicating the leftmost pixel of the annotation
XMax	0.987365	Normalized image coordinates indicating the rightmost pixel of the annotation
YMin	0.615523	Normalized image coordinates indicating the topmost pixel of the annotation
YMax	0.895307	Normalized image coordinates indicating the bottomost pixel of the annotation
IsOccluded	0	Binary value indicating if the object is occluded by another object in the image
IsTruncated	1	Binary value indicating if the object extends beyond the boundary of the image
IsGroupOf	0	Binary value indicating if the box spans a group of objects
IsDepictionOf	1	Binary value indicating if the object is a depiction and not a real physical instance
IsInsideOf	1	Binary value indicating if the image is taken from the inside of the object
IsInsideOf	1	Binary value indicating if the limage is taken from the inside of the object
GenderPresentation	Predominantly Masculine	Indicates the perceived gender presentation of the subject assessed by a third party
AgePresentation	Middle	Indicates the perceived age range of the subject assessed by a third party

Data source and collection

Data Collection

DATA COLLECTION METHOD(S)

Derived

Vendor Collection Efforts

DATA SOURCES BY COLLECTION METHOD(S)

Images	Open Images V6
Labels	Human annotators
Bounding Boxes	Human annotators

SUMMARIES OF DATA COLLECTION METHODS

100,000 images randomly sampled from the positive set of Open Images V6, which contains approximately 9.9M images

- Training Set: 70,000 sampled from 9,011,219 images
- Testing/Validation: 30,000 sampled from 167,056 images

EXCLUDED DATA

No excluded data

DATA SELECTION CRITERIA - SCRAPING

- Images were sampled from the positive subset of training and testing/ validation containing annotator-verified image lables
- Images contained at least one of five person classess (`man`, `woman`, `boy`, `girl`, or `person`)

Note: We did not include non-binary as a class label as it is not possible to label gender identity from images. Gender identity should only be used in situations where participants are able to self-report gender.

Labeling Process

Data distribution

Entries in data card

Very good reference for data quality, but wayy too difficult to fill out for every dataset so usually ignored...

They give you an idea of what might impact data quality

We will touch on some:

- Data quality, in terms of noise (e.g. from upstream data source, from human labelers)
- Data drifts (static dataset and its source, vs. the world now)
- Data quality, in terms of distributions (disagreements between annotators, biases towards certain distribution, annotation incentives -- tend to select easy labels, etc.)
- Data curation (shape the data towards what you want)

Understand and improve data quality

Assuming you didn't have the best documented and cleaned data in the world (often!), how do you evaluate your data and how do you clean it?

> *Data cleaning and repairing account for about 60% of the work of data scientists.*

> *"Everyone wants to do the model work, not the data work"*

Own experience?

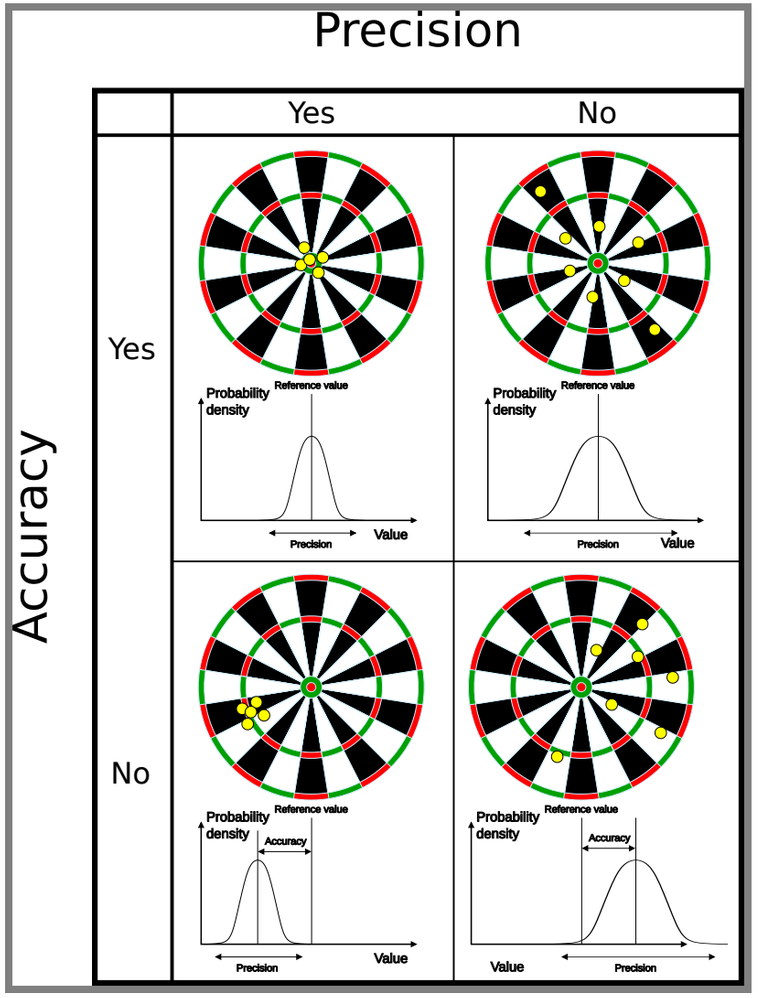
Accuracy vs Precision (on *Data*)

Accuracy: Reported values (on average) represent real value

Precision: Repeated measurements yield the same result

Accurate, but imprecise: **Q. How to deal with this issue?**

Inaccurate, but precise: ?



Speaker notes

Average to deal with noise; calibrate to deal with bias

Accuracy and Precision Problems in Warehouse Data?

Data Accuracy and Precision: Impact on ML

More data -> better models (up to a point, diminishing effects)

Noisy data (imprecise) -> less confident models, more data needed

- some ML techniques are more or less robust to noise (more on robustness in a later lecture)

Inaccurate data -> misleading models, biased models

- Need the "right" data

Invest in data quality, not just quantity

Dealing with noises in data

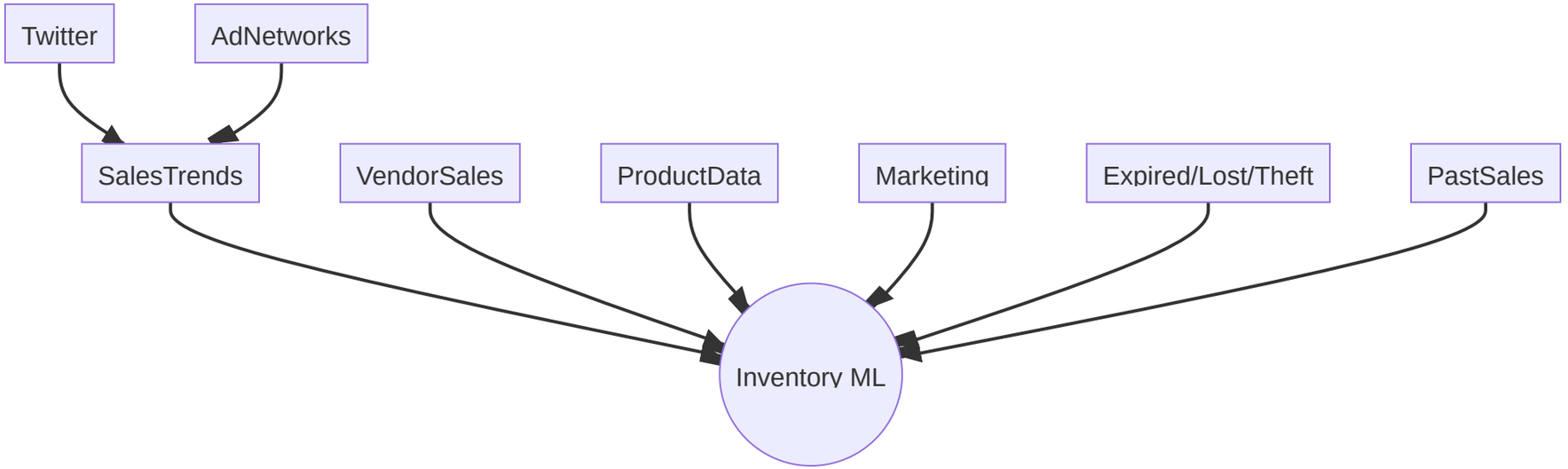
Where do noises come from and how do we fix them?

What do we mean by clean data?

- **Accuracy:** The data was recorded correctly.
- **Completeness:** All relevant data was recorded.
- **Uniqueness:** The entries are recorded once.
- **Consistency:** The data agrees with itself.
- **Timeliness:** The data is kept up to date.

Challenge from collection: Data Comes from Many Sources

e.g. For the inventory system:



Challenge from collection: Data Comes from Many Sources

- Manually entered
- Generated through actions in IT systems
- Logging information, traces of user interactions
- Sensor data
- Crowdsourced

These sources have different reliability and quality

What happens: Data is noisy

Wrong results and computations, crashes

Duplicate data, near-duplicate data

Out of order data

Data format invalid

Examples in inventory system?

Two levels of data precision

1. **Data Integrety / Schema**: Ensuring basic consistency about shape and types
2. **Wrong and inconsistent data**: Application- and domain-specific data issues

Data Integrety / Schema

Ensuring basic consistency about shape and types

Data Schema

Define the expected format of data

- expected fields and their types
- expected ranges for values
- constraints among values (within and across sources)

Data can be automatically checked against schema

Protects against change; explicit interface between components

Data Schema Constraints for Inventory System?

Product Database:

ID	Name	Weight	Description	Size	Vendor
...	...	...	...	...	...

Stock:

ProductID	Location	Quantity
...	...	...

Sales history:

UserID	ProductId	DateTime	Quantity	Price
...	...	...	...	...

Schema Problems: Uniqueness, data format, integrity, ...

- Illegal attribute values: `bdate=30.13.70`
- Violated attribute dependencies: `age=22, bdate=12.02.70`
- Uniqueness violation: `(name="John Smith", SSN="123456"), (name="Peter Miller", SSN="123456")`
- Referential integrity violation: `emp=(name="John Smith", deptno=127)` if department 127 not defined

Schema in Relational Databases

```
CREATE TABLE employees (
    emp_no      INT             NOT NULL,
    birth_date  DATE            NOT NULL,
    name        VARCHAR(30)     NOT NULL,
    PRIMARY KEY (emp_no));
CREATE TABLE departments (
    dept_no     CHAR(4)         NOT NULL,
    dept_name   VARCHAR(40)     NOT NULL,
    PRIMARY KEY (dept_no), UNIQUE  KEY (dept_name));
CREATE TABLE dept_manager (
   dept_no      CHAR(4)         NOT NULL,
```

Dirty Data: Example

TABLE: CUSTOMER

ID	Name	Birthday	Age	Sex	Phone	ZIP
3456	Ford, Harrison	18.2.76	43	M	999999999	15232
3456	Mark Hamil	33.8.81	43	M	6173128718	17121
3457	Kim Kardashian	11.10.56	63	M	4159102371	94016

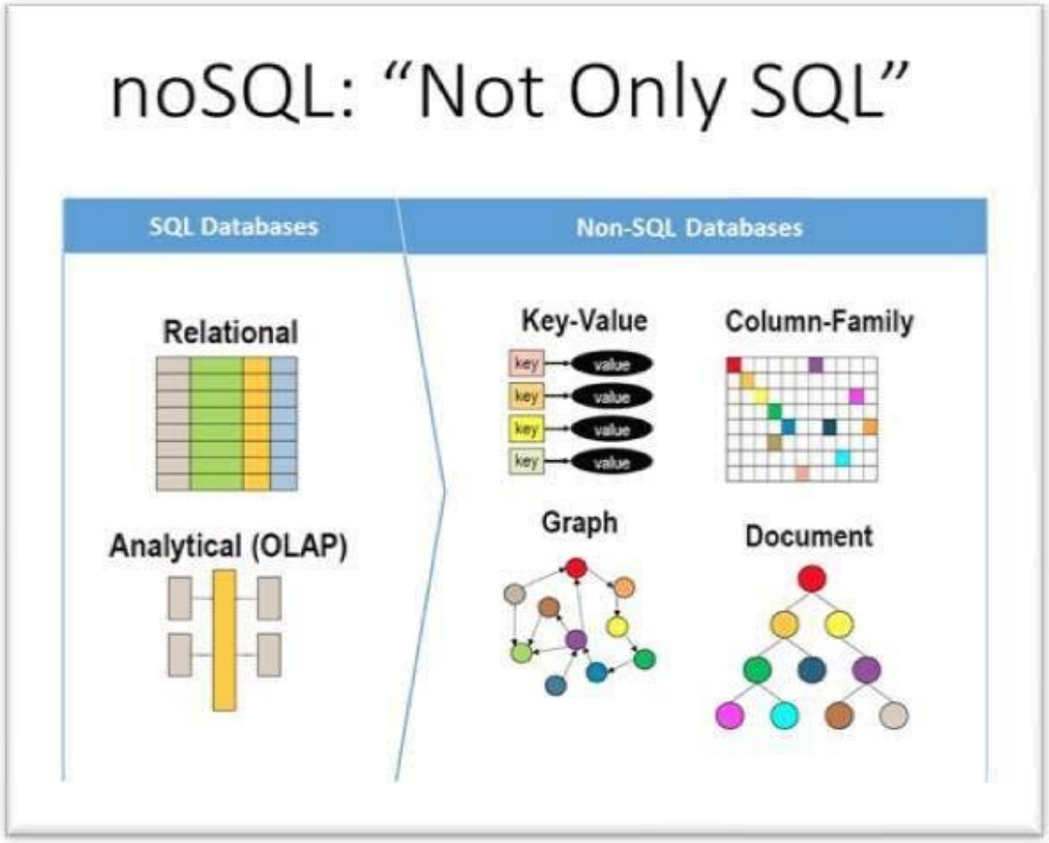
TABLE: ADDRESS

ZIP	City	State
15232	Pittsburgh	PA
94016	Sam Francisco	CA
73301	Austin	Texas

Problems with this data? Which Problems are Schema Problems?

What Happens When New Data Violates Schema?

Modern Databases: Schema-Less



Also vector databases, schema-aware databases that bascially store long text in each cell, etc.

Image source: https://www.kdnuggets.com/2021/05/nosql-know-it-all-compendium.html

Schema-Less Data Exchange

- CSV files
- Key-value stores (JSon, XML, Nosql databases)
- Message brokers
- REST API calls
- R/Pandas Dataframes

```
2022-10-06T01:31:18,230550,GET /rate/narc+2002=4
2022-10-06T01:31:19,332644,GET /rate/i+am+love+2009=4
```

```
{"user_id":5,"age":26,"occupation":"scientist","gender":"M"}
```

Schema-Less Data Exchange

Q. Benefits? Drawbacks?

Schema Library: Apache Avro

```
{   "type": "record",
    "namespace": "com.example",
    "name": "Customer",
    "fields": [{
            "name": "first_name",
            "type": "string",
            "doc": "First Name of Customer"
        },
        {
            "name": "age",
            "type": "int",
```

Schema Library: Apache Avro

Schema specification in JSON format

Serialization and deserialization with automated checking

Native support in Kafka

Benefits

- Serialization in space efficient format
- APIs for most languages (ORM-like)
- Versioning constraints on schemas

Drawbacks

- Reading/writing overhead
- Binary data format, extra tools needed for reading
- Requires external schema and maintenance
- Learning overhead

Speaker notes

Further readings eg https://medium.com/@stephane.maarek/introduction-to-schemas-in-apache-kafka-with-the-confluent-schema-registry-3bf55e401321, https://www.confluent.io/blog/avro-kafka-data/, https://avro.apache.org/docs/current/

Many Schema Libraries/Formats

Examples

- Avro
- XML Schema
- Protobuf
- Thrift
- Parquet
- ORC

Schema not just for database but for data transmission

Request

```
    ],
    "response_format": {
      "type": "json_schema",
      "json_schema": {
        "name": "math_response",
        "strict": true,
        "schema": {
          "type": "object",
          "properties": {
            "steps": {
              "type": "array",
              "items": {
                "type": "object",
                "properties": {
                  "explanation": {
                    "type": "string"
                  },
                  "output": {
                    "type": "string"
                  }
                },
                "required": ["explanation", "output"],
                "additionalProperties": false
              }
            },
```

Output JSON

```
{
  "steps": [
    {
      "explanation": "Subtract 31 from both sides to isolate the term with x.",
      "output": "8x + 31 - 31 = 2 - 31"
    },
    {
      "explanation": "This simplifies to 8x = -29.",
      "output": "8x = -29"
    },
    {
      "explanation": "Divide both sides by 8 to solve for x.",
      "output": "x = -29 / 8"
    }
  ],
  "final_answer": "x = -29 / 8"
}
```

Summary: Schema

Basic structure and type definition of data

Well supported in databases and many tools

Very low bar of data quality

Wrong and Inconsistent Data

Application- and domain-specific data issues

Dirty Data: Example

TABLE: CUSTOMER

ID	Name	Birthday	Age	Sex	Phone	ZIP
3456	Ford, Harrison	18.2.76	43	M	999999999	15232
3456	Mark Hamil	33.8.81	43	M	6173128718	17121
3457	Kim Kardashian	11.10.56	63	M	4159102371	94016

TABLE: ADDRESS

ZIP	City	State
15232	Pittsburgh	PA
94016	Sam Francisco	CA
73301	Austin	Texas

Problems with the data beyond schema problems?

Wrong and Inconsistent Data

- Missing values: `phone=9999-999999`
- Misspellings: `city=Pittsburg`
- Misfielded values: `city=USA`
- Duplicate records: `name=John Smith, name=J. Smith`
- Wrong reference: `emp=(name="John Smith", deptno=127)` if department 127 defined but wrong

Q. How can we detect and fix these problems?

Further readings: Rahm, Erhard, and Hong Hai Do. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 23.4 (2000): 3-13.

Discussion: Wrong and Inconsistent Data?

Data Cleaning Overview

Data analysis / Error detection

- Usually focused on specific kind of problems, e.g., duplication, typos, missing values, distribution shift
- Detection in input data vs detection in later stages (more context)

Error repair

- Repair data vs repair rules, one at a time or holistic
- Data transformation or mapping
- Automated vs human guided

Error Detection Examples

Illegal values: min, max, variance, deviations, cardinality

Misspelling: sorting + manual inspection, dictionary lookup

Missing values: null values, default values

Duplication: sorting, edit distance, normalization

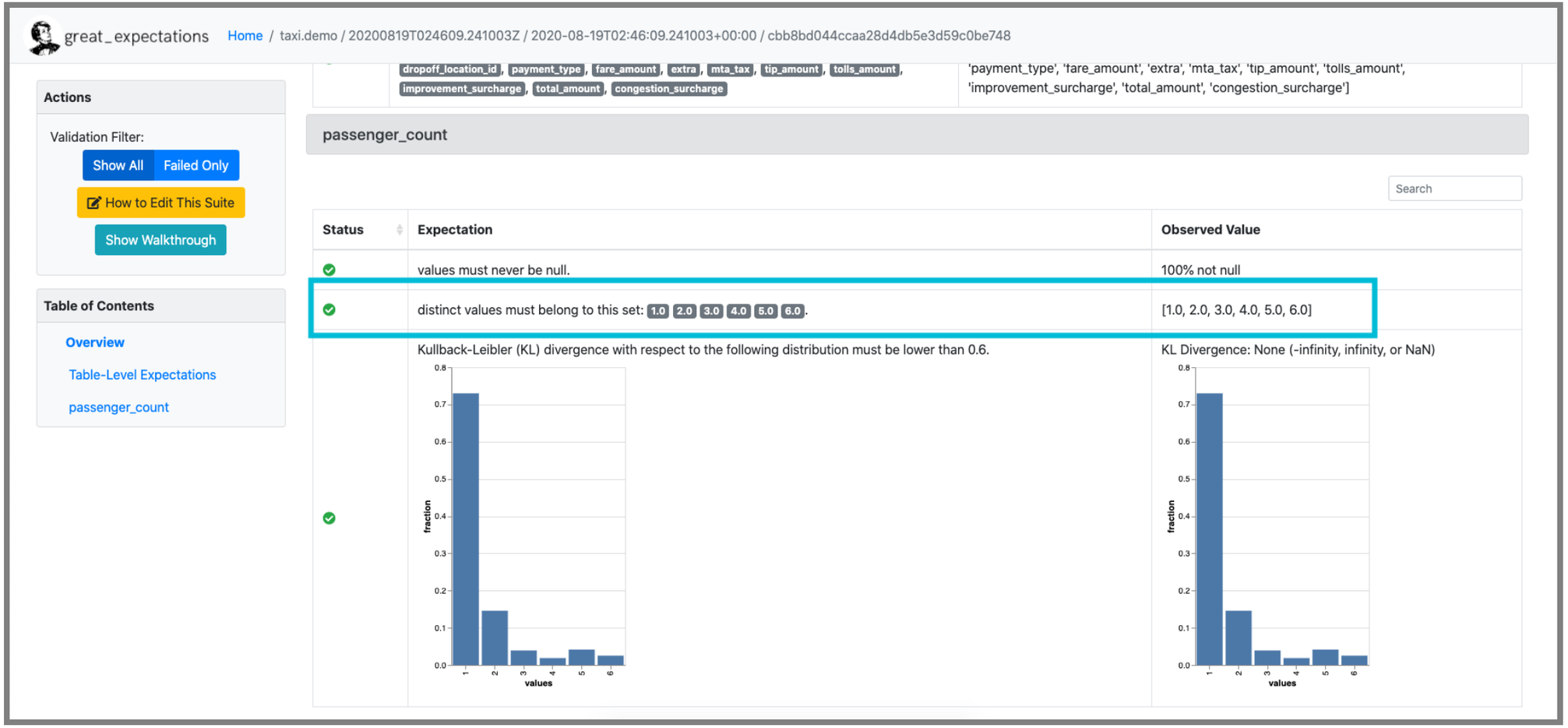
Example Tool: Great Expectations

```
expect_column_values_to_be_between(
    column="passenger_count",
    min_value=1,
    max_value=6
)
```

Supports schema validation and custom instance-level checks.

https://greatexpectations.io/

Example Tool: Great Expectations



https://greatexpectations.io/

Rule-based detection: Data Quality Rules

Rules can be used to reject data or repair it

Invariants on data that must hold

Typically about relationships of multiple attributes or data sources, eg.

- ZIP code and city name should correspond
- User ID should refer to existing user
- SSN should be unique
- For two people in the same state, the person with the lower income should not have the higher tax rate

Classic integrity constraints in databases or conditional constraints

ML-based for Detecting Inconsistencies

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

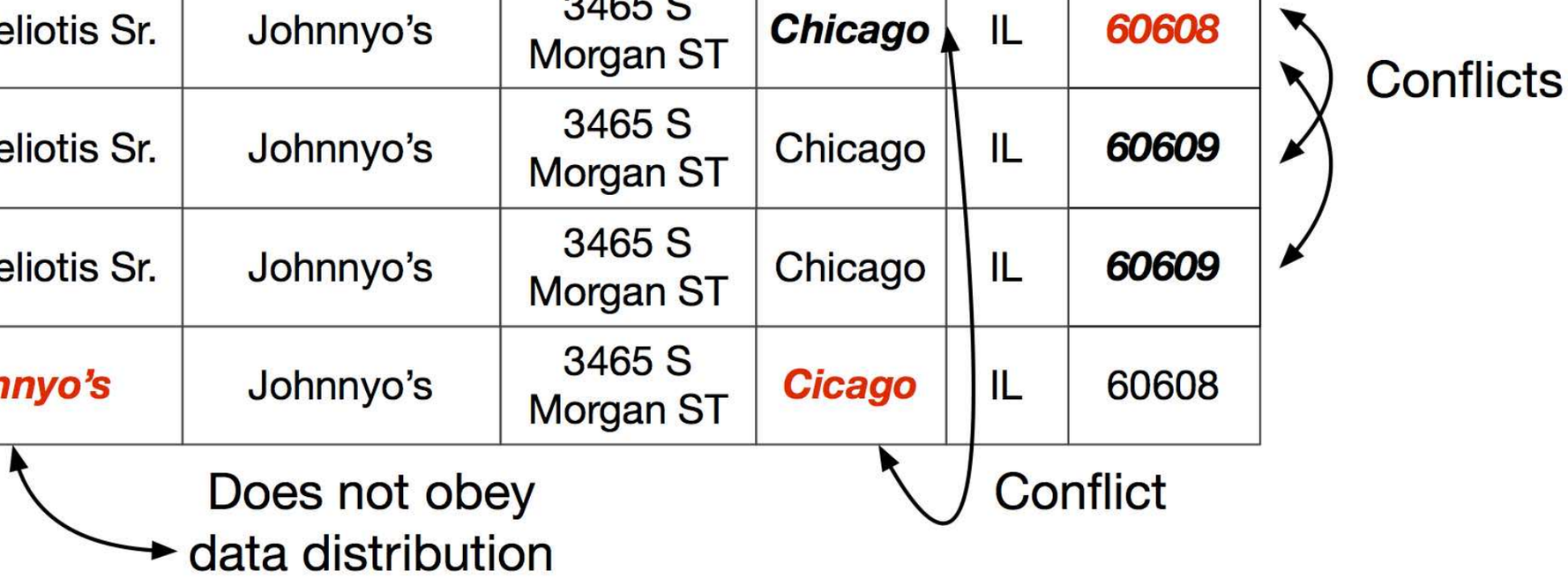


Image source: Theo Rekatsinas, Ihab Ilyas, and Chris Ré, "HoloClean - Weakly Supervised Data Repairing." Blog, 2017.

Example: HoloClean

- User provides rules as integrity constraints (e.g., "two entries with the same name can't have different city")
- Detect violations of the rules in the data; also detect statistical outliers
- Automatically generate repair candidates (with probabilities)

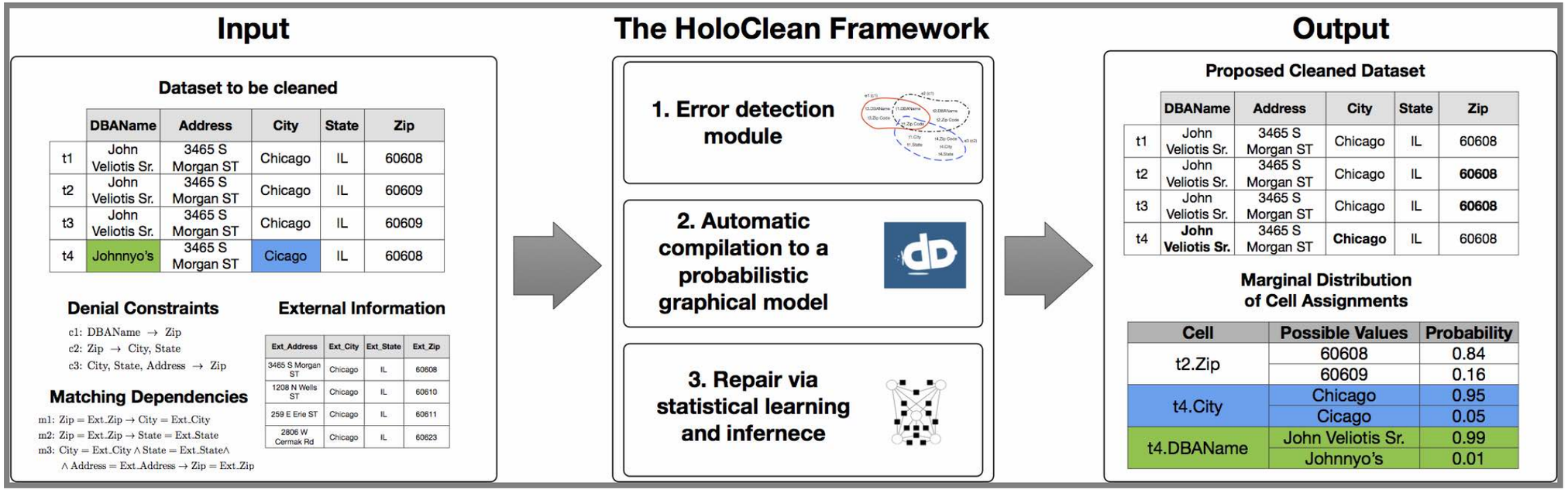


Image source: Theo Rekatsinas, Ihab Ilyas, and Chris Ré, "HoloClean - Weakly Supervised Data Repairing." Blog, 2017.

Discovery of Data Quality Rules

Rules directly taken from external databases

- e.g. zip code directory

Given clean data,

- several algorithms that find functional relationships ($X \Rightarrow Y$) among columns
- algorithms that find conditional relationships (if Z then $X \Rightarrow Y$)
- algorithms that find denial constraints (X and Y cannot co-occur in a row)

Given mostly clean data (probabilistic view),

- algorithms to find likely rules (e.g., association rule mining)
- outlier and anomaly detection

Given labeled dirty data or user feedback,

- supervised and active learning to learn and revise rules
- supervised learning to learn repairs (e.g., spell checking)

Further reading: Ilyas, Ihab F., and Xu Chu. Data cleaning. Morgan & Claypool, 2019.

Discussion: Data Quality Rules

Speaker notes

Compare with past price

Dealing with Drift

Why does my model begin to perform poorly over time?

A very particular form of data accuracy problem (data becomes wrong), caused not by human creators but by the world. Very prevalent & affects product so we talk about it first!

Data changes

System objective changes over time

Software components are upgraded or replaced

Prediction models change

Quality of supplied data changes

User behavior changes

Assumptions about the environment no longer hold

Examples in inventory system?

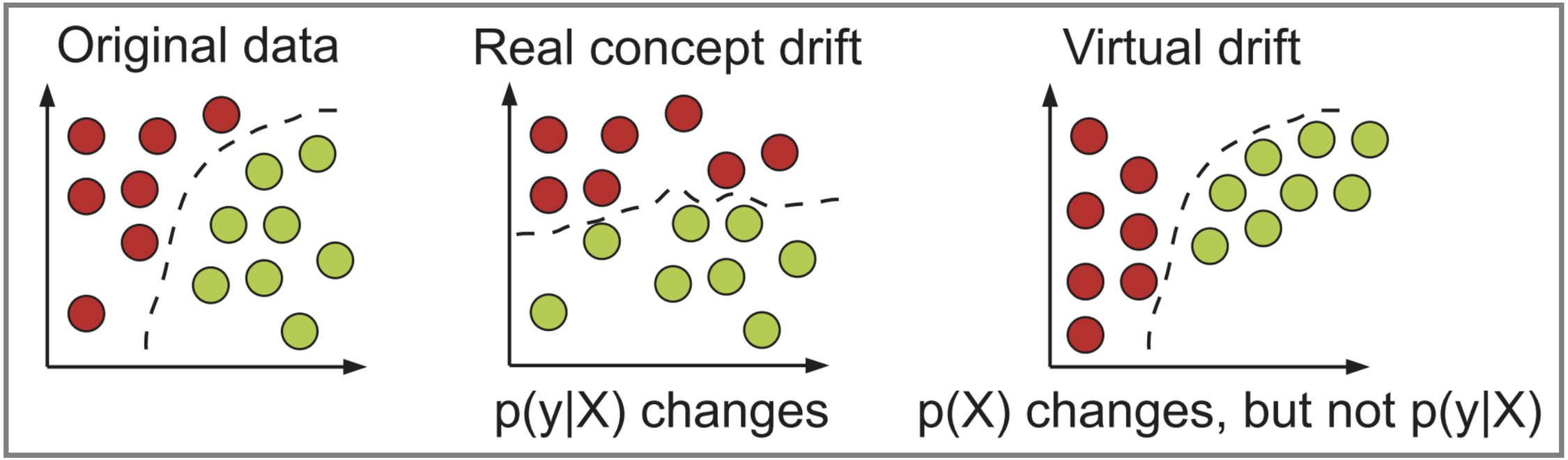
Users may deliberately change data

Users react to model output; causes data shift

Users try to game/deceive the model

Examples in inventory system?

Types of Drift



Gama et al., *A survey on concept drift adaptation*. ACM Computing Surveys Vol. 46, Issue 4 (2014)

Drift & Model Decay

Concept drift (or concept shift)

- properties to predict change over time (e.g., what is credit card fraud)
- model has not learned the relevant concepts
- over time: different expected outputs for same inputs

Data drift (or covariate shift, virtual drift, distribution shift, or population drift)

- characteristics of input data changes (e.g., customers with face masks)
- input data differs from training data
- over time: predictions less confident, further from training data

Upstream data changes

- external changes in data pipeline (e.g., format changes in weather service)
- model interprets input data incorrectly
- over time: abrupt changes due to faulty inputs

How do we fix these drifts?

Speaker notes

- fix1: retrain with new training data or relabeled old training data
  - fix2: retrain with new data
  - fix3: fix pipeline, retrain entirely

On Terminology

Concept and data drift are separate concepts

In practice and literature not always clearly distinguished

Colloquially encompasses all forms of model degradations and environment changes

Define term for target audience

Last AIV Warning for Breakouts

From the first lecture and syllabus:

> *Within groups, we expect that you are honest about your contribution to the group's work. [...] This also applies to in-class discussions, where indicating working with others who did not participate in the discussion is considered an academic honesty violation.*

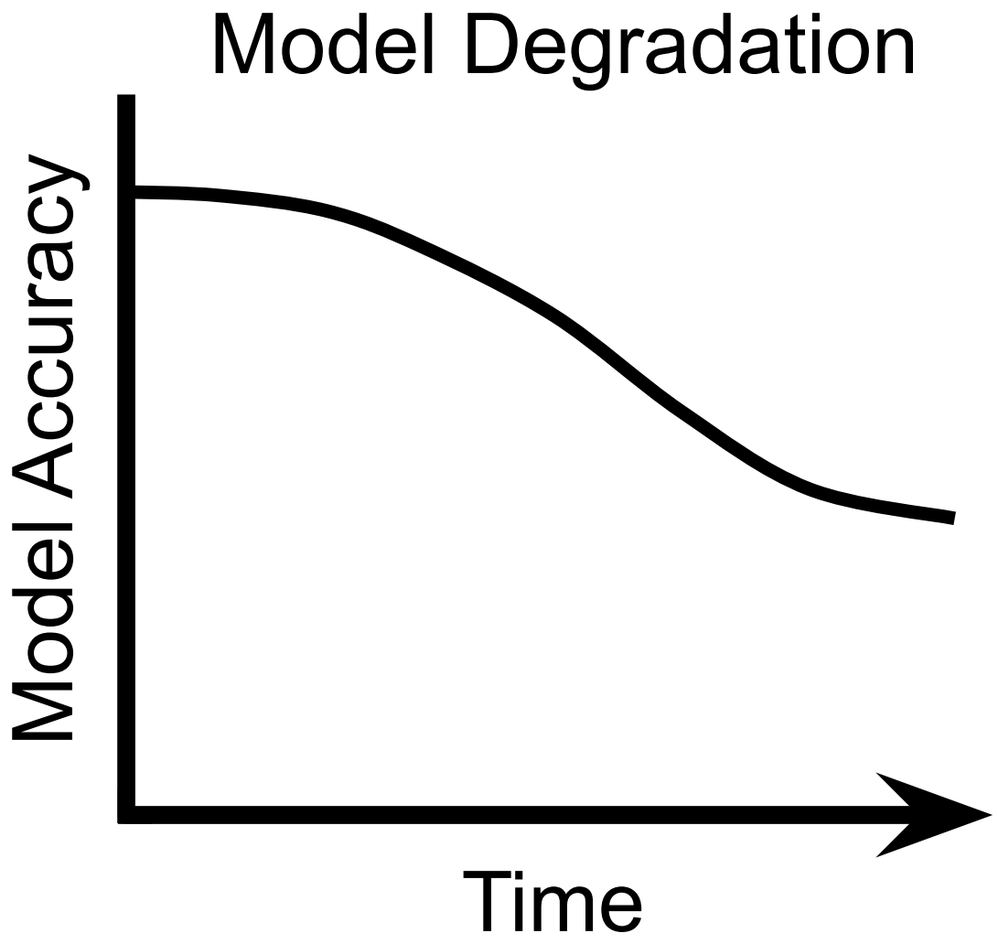
Breakout: Drift in the Inventory System

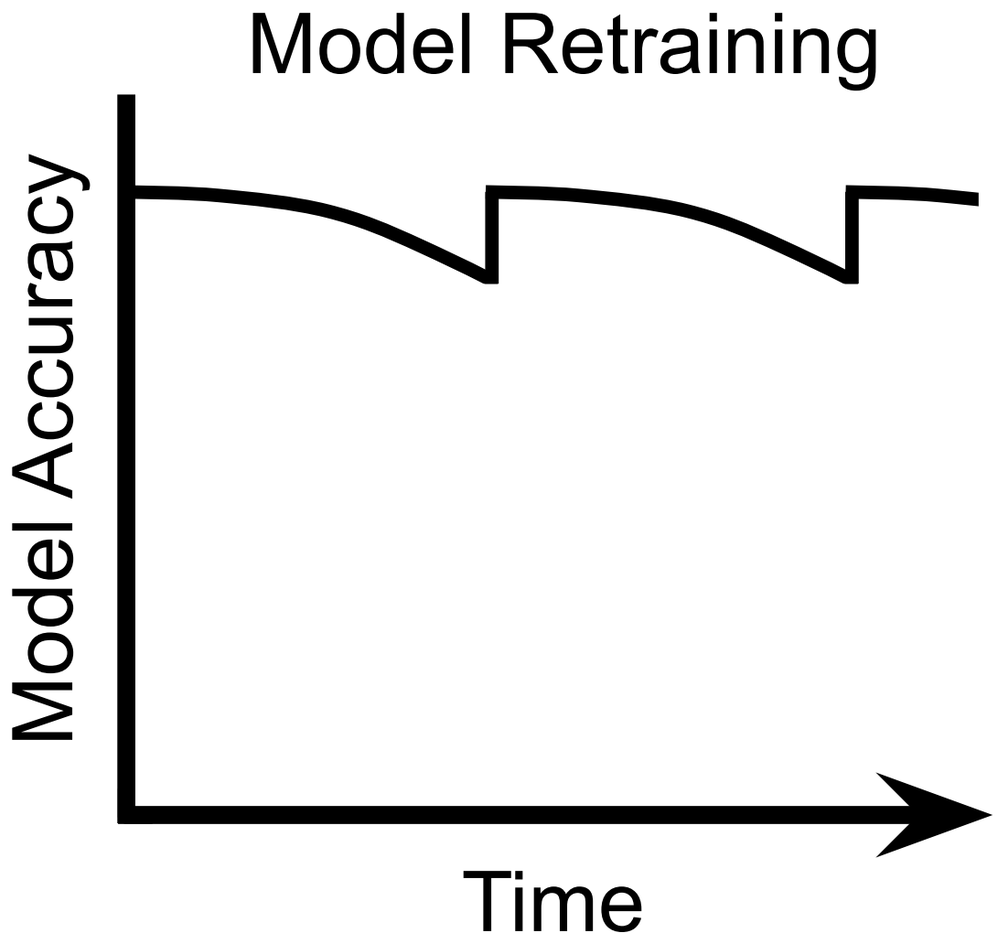
What kind of drift might be expected?

As a group, tagging members, write plausible examples in `#lecture`:

- *Concept Drift:*
- *Data Drift:*
- *Upstream data changes:*

Watch for Degradation in Prediction Accuracy





Indicators of Concept Drift

How to detect concept drift in production?

Indicators of Concept Drift

Model degradations observed with telemetry

Telemetry indicates different outputs over time for similar inputs

Differences in influential features and feature importance over time

Relabeling training data changes labels

Interpretable ML models indicate rules that no longer fit

(many papers on this topic, typically on statistical detection)

Indicators of Data Drift

How to detect data drift in production?

Indicators of Data Drift

Model degradations observed with telemetry

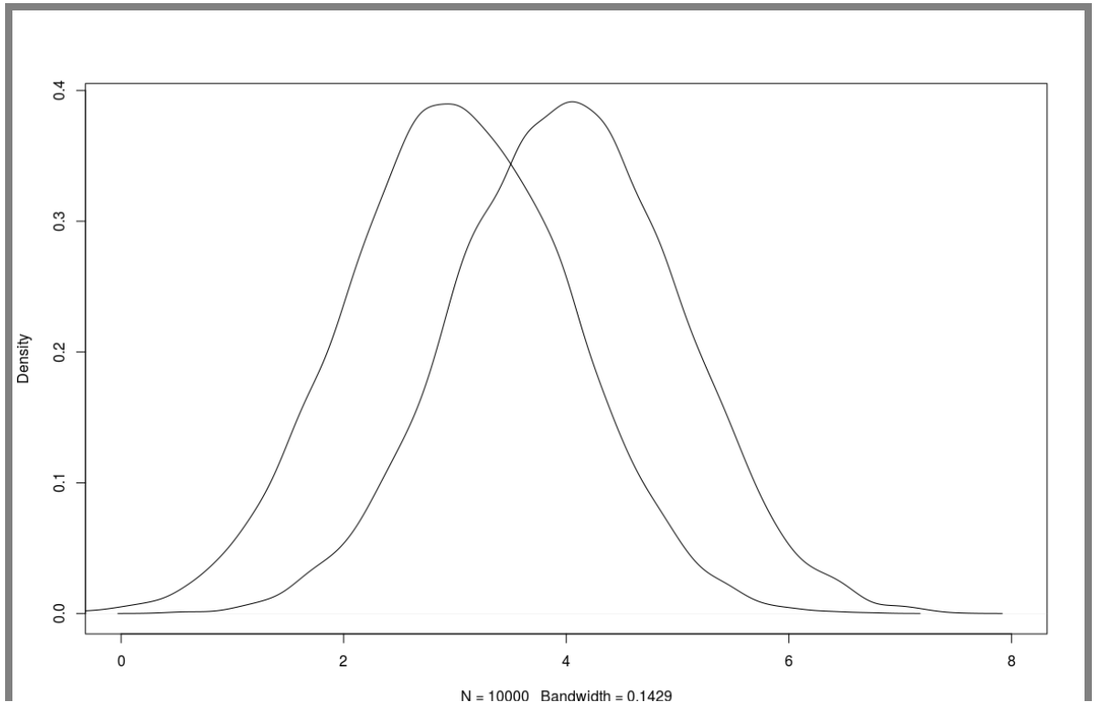
Distance between input distribution and training distribution increases

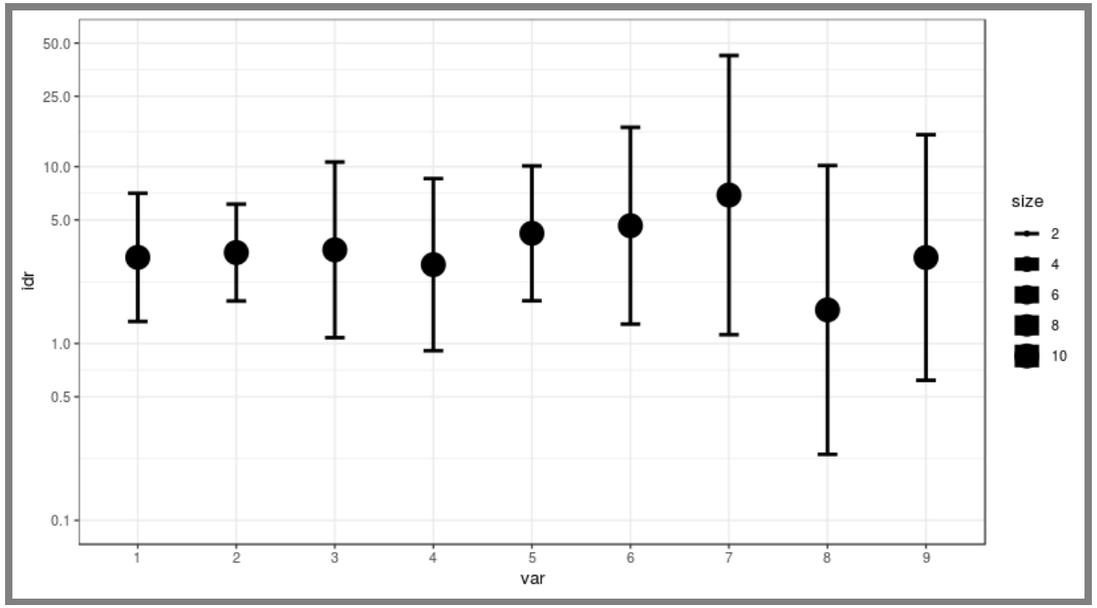
Average confidence of model predictions declines

Relabeling of training data retains stable labels

Detecting Data Drift

- Compare distributions over time (e.g., t-test)
- Detect both sudden jumps and gradual changes
- Distributions can be manually specified or learned (see invariant detection)





Data Distribution Analysis

Plot distributions of features (histograms, density plots, kernel density estimation)

- Identify which features drift

Define distance function between inputs and identify distance to closest training data (e.g., energy distance, see also kNN)

Anomaly detection and "out of distribution" detection

Compare distribution of output labels

Data Distribution Analysis Example

https://rpubs.com/ablythe/520912

Microsoft Azure Data Drift Dashboard

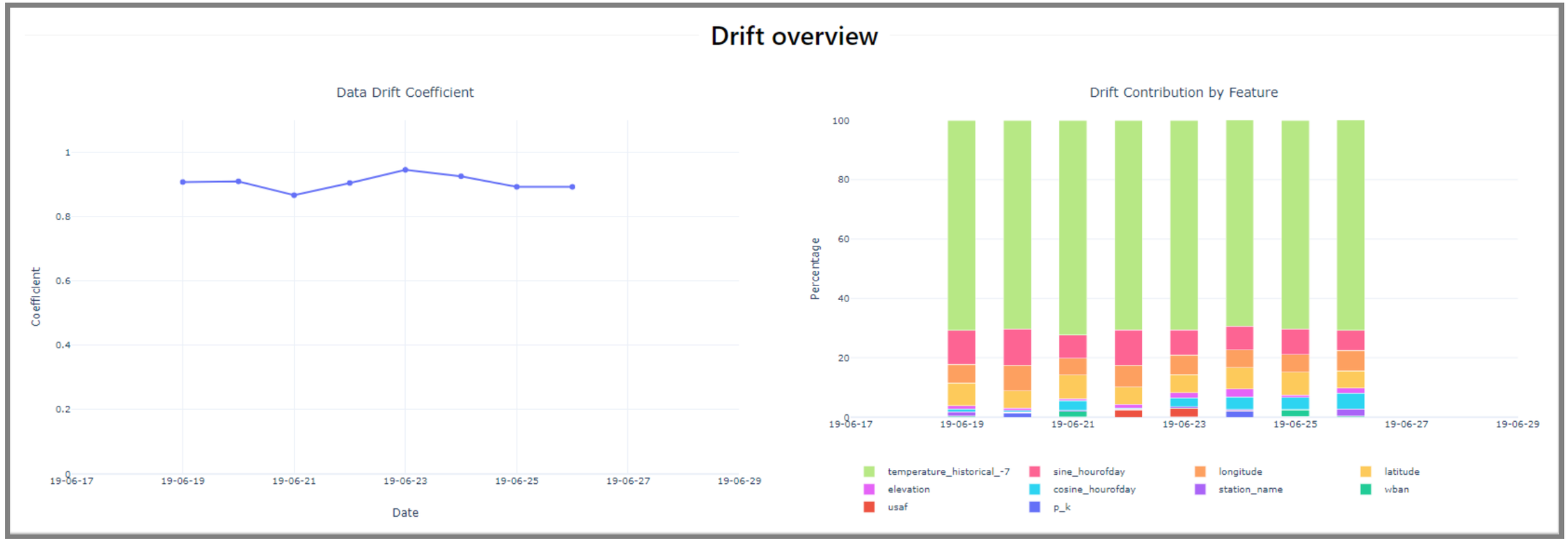


Image source and further readings: Detect data drift (preview) on models deployed to Azure Kubernetes Service (AKS)

Dealing with Drift

Regularly retrain model on recent data

- Use evaluation in production to detect decaying model performance

Involve humans when increasing inconsistencies detected

- Monitoring thresholds, automation

Monitoring, monitoring, monitoring!

Dealing with Inaccurate Data not caused by drifts

How do you detect and fix more systematic data quality issues?

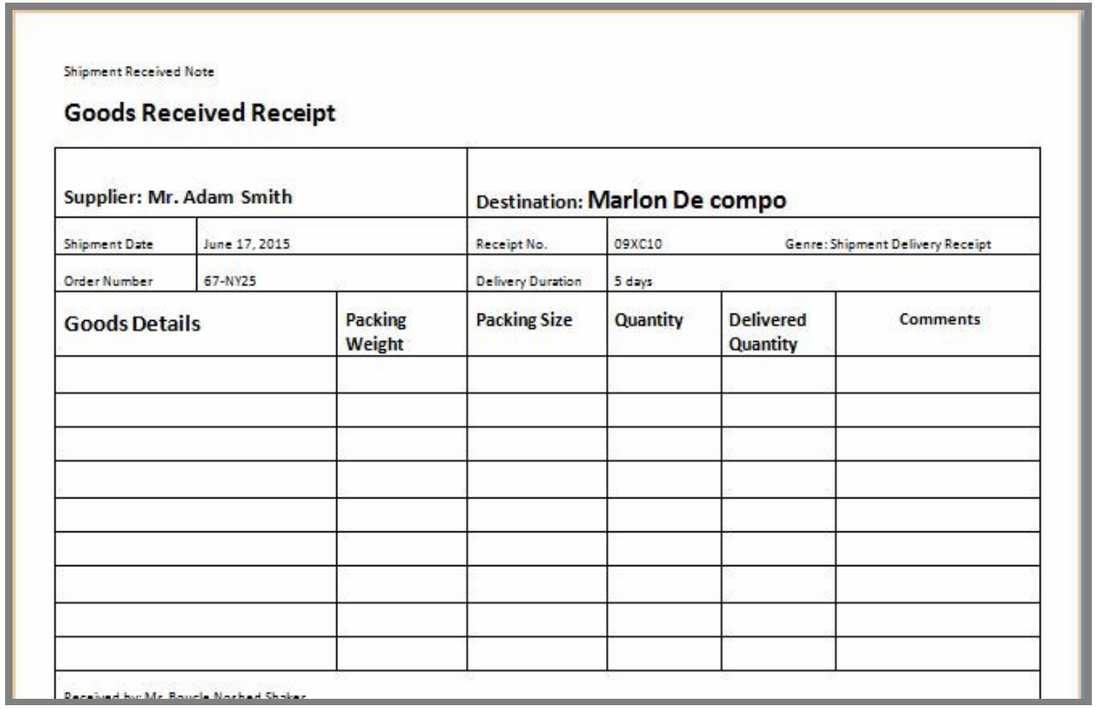
Slightly more advanced topic but you might find it fun

Challenge from collection and processing: Biased

The concept of "raw data" might be misleading; Data is always some proxy we actively collect to represent the world. Someone decides...

- What data to collect
- How to collect them
- What scale to use

They change what you can do with the data



Shipment Received Note

Goods Received Receipt

Supplier: Mr. Adam Smith		**Destination: Marlon De compo**			
Shipment Date	June 17, 2015	Receipt No.	09XC10	Genre: Shipment Delivery Receipt	
Order Number	67-NY25	Delivery Duration	5 days		
Goods Details	**Packing Weight**	**Packing Size**	**Quantity**	**Delivered Quantity**	**Comments**

Received by: Mr. Boucle Norbed Shaker

What happens: Data is inaccurate (for what you want to do)

Missing data

Biased data

Systematic errors in data distribution

Examples in inventory system?

Data Quality from a Utility Perspective

1. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*

2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to remove erroneous examples and to properly handle ambiguous ones.*

3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*

4. Benchmarks should reveal plausibly harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use auxiliary bias evaluation metrics.*

Task-specific measurements.

Annotation artifacts.

Label correctness.

Task subjectiveness

Dataset difficulty.

Dataset bias.

Rogers, Anna, et al. "Changing the World by Changing the Data" ACL 2021

Task Subjectiveness & Disagreement

If you have datasets with a wide range of raw labels, you can quantify task subjectiveness using **entropy**: a measure for the level of uncertainty or agreement among human judgements, where high entropy suggests low level of agreement and vice versa.

Entropy on one example $\mathrm{H}(p) = -\sum_{i \in C} p_i \log(p_i)$ *The frequency of humans assigning label i*

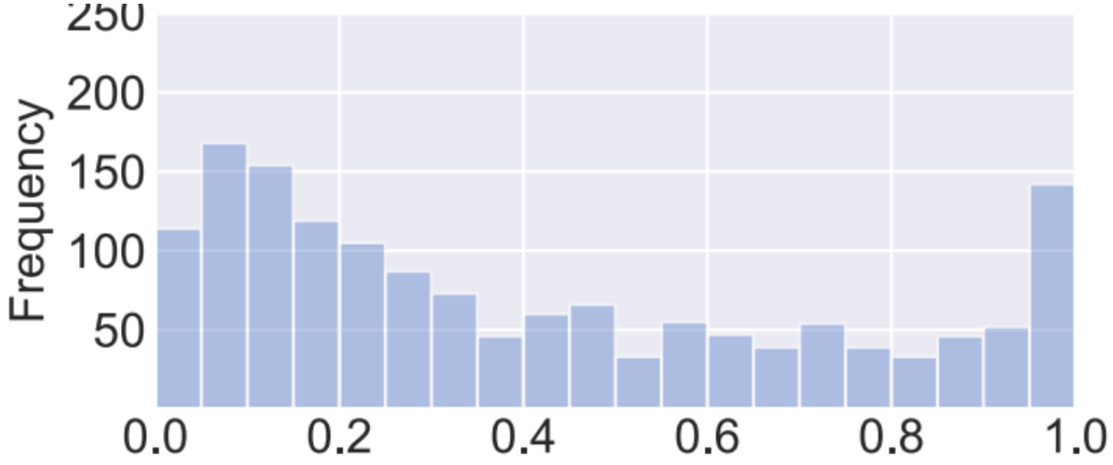
Label category

Premise	Hypothesis	Old Labels	New Labels
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	E E N N E	N (93) E (7)

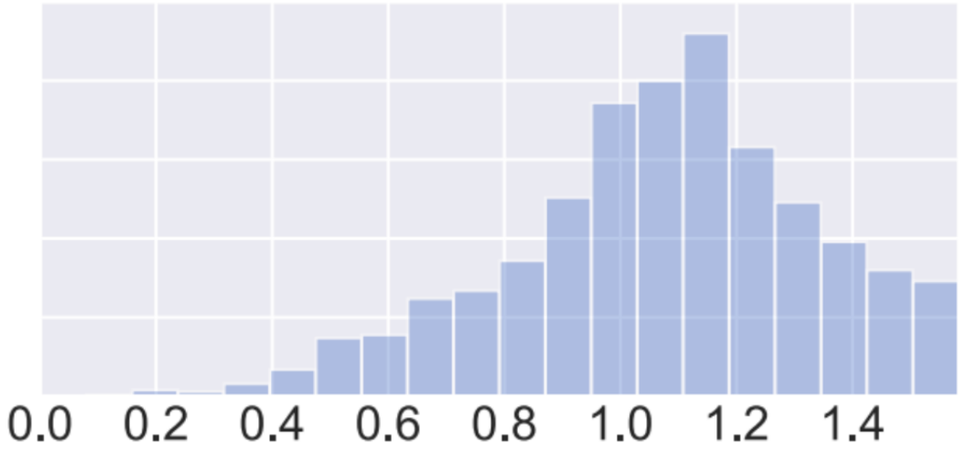
Sentence with Target Entity	Entity Type Labels
During the Inca Empire, {the Inti Raymi} was the most important of four ceremonies celebrated in Cusco.	event, festival, ritual, custom, ceremony, party, celebration

Task Subjectiveness & Disagreement

If you have datasets with a wide range of raw labels, you can quantify task subjectiveness using **entropy**: a measure for the level of uncertainty or agreement among human judgements, where high entropy suggests low level of agreement and vice versa.



U-Shaped distribution.
Either highly certain or uncertain.



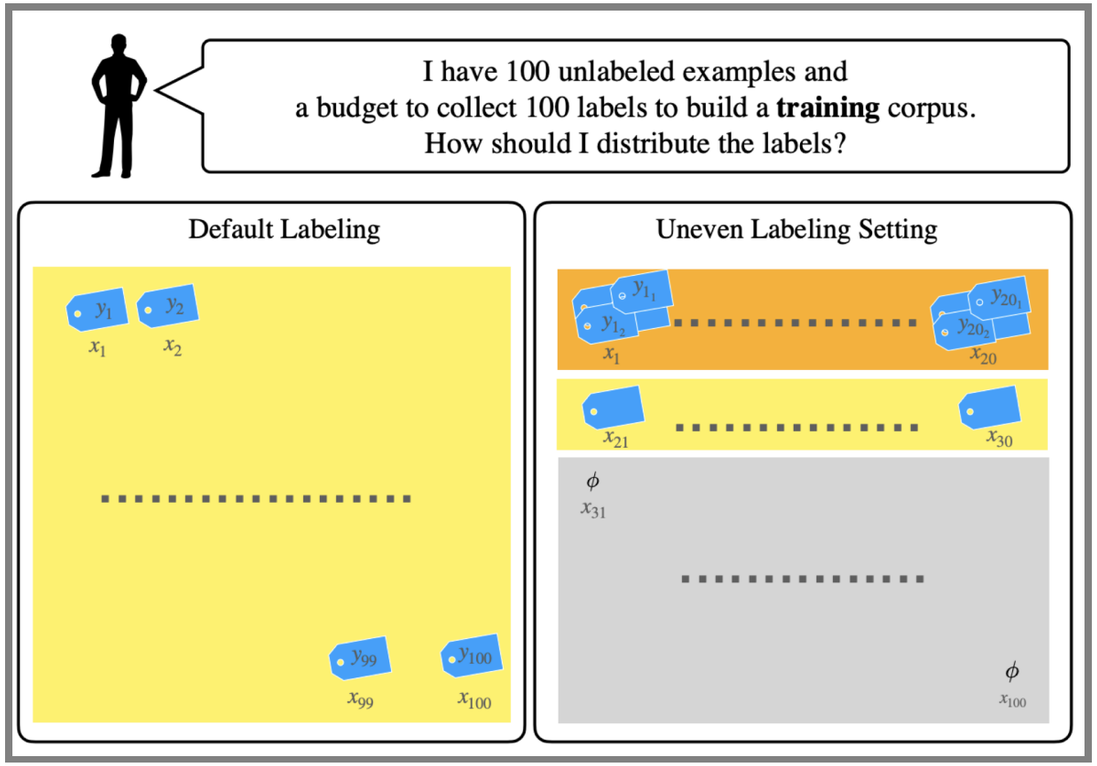
One peak on the right.
Slightly skewed towards disagreement.

Mitigation on Disagreement

Core idea: Use annotation disagreement to decide whether to relabel existing examples, or label completely new examples.

Allow data points to have multiple labels; Also train with these multiple labels.

Clear gains on some NLP tasks!



Annotation artifact

Unwanted patterns or biases present in the data, which can cause language models to make incorrect predictions.

e.g. For Natural Language Inference Task (Premise, Hypothesis, Label)

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Models can do moderately well on NLI datasets without looking at the premise!

One way to detect annotation artifact

Pointwise Mutual Information (PMI) is a statistical measure of the association between two variables X and Y. It captures the strength of the relationship between the two variables and expresses it as a numerical score.

$$PMI(feature, class) = \log \frac{P(feature, class)}{P(feature, \cdot)P(class, \cdot)}$$

In more plain language: “how frequent do a feature and a class co-occure? How easy it is to detect a class just by looking at a feature?"

Social Bias through PMI

		GENDER	
woman	hairdresser‡ fairground grieving receptionist widow	**man**	rock-climbing videoing armband tatooes gent
women	actresses† husbands‡ womens‡ gossip‡ wemon‡	**men**	gypsies supervisors contractors mens‡ cds
girl	schoolgirl piata cindy pigtails‡ gril	**boy**	misbehaving see-saw timmy lad‡ sprained
girls	fifteen‡ slumber sking‡ jumprope† ballerinas‡	**boys**	giggle‡ youths‡ sons‡ brothers‡ skip
mother	kissed‡ parent‡ mom‡ feeds daughters	**father**	fathers‡ dad‡ sons† daughters plant
		AGE	
old	ferret‡ quilts‡ knits‡ grandpa‡ elderly‡	**young**	giggle cds youthful‡ tidal amusing
old woman	knits‡ grandmother‡ scarf† elderly‡ lady‡	**young woman**	salon† attractive blow blowing feeds
old man	ferret‡ grandpa‡ wrapping‡ grandfather‡ elderly‡	**young man**	boarder disabled rollerblades graduation skate‡
		RACE/ETHNICITY/NATIONALITY	
indian	indians‡ india‡ native‡ traditional‡ pouring†	**caucasian**	blond white‡ american asian blonde
indian woman	cooking† clothes lady using making	**american**	patriotic‡ canadian‡ americans‡ reenactment‡ america‡
indian man	food couple a‡ sleeping sitting	**american woman**	women‡ black white front her‡
asian	kimonos‡ asians‡ asain‡ oriental‡ chinatown‡	**american man**	speaking‡ money‡ black‡ white‡ music
asians	asian‡ food people‡ eating friends	**black woman**	african‡ american asian white‡ giving
asian woman	oriental‡ indian† chinese‡ listens† customers	**black man**	african‡ american white‡ roller face
asian man	shrimp† rice† chinese‡ businessman cooks†	**native american**	americans‡ music‡ dressed they woman

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. "Social bias in elicited natural language inferences." EthNLP 2017.

Mitigation of Annotation Artifact

Counterfactual Data Augmentation

Core idea: Minimally editing existing examples to maintain the label.

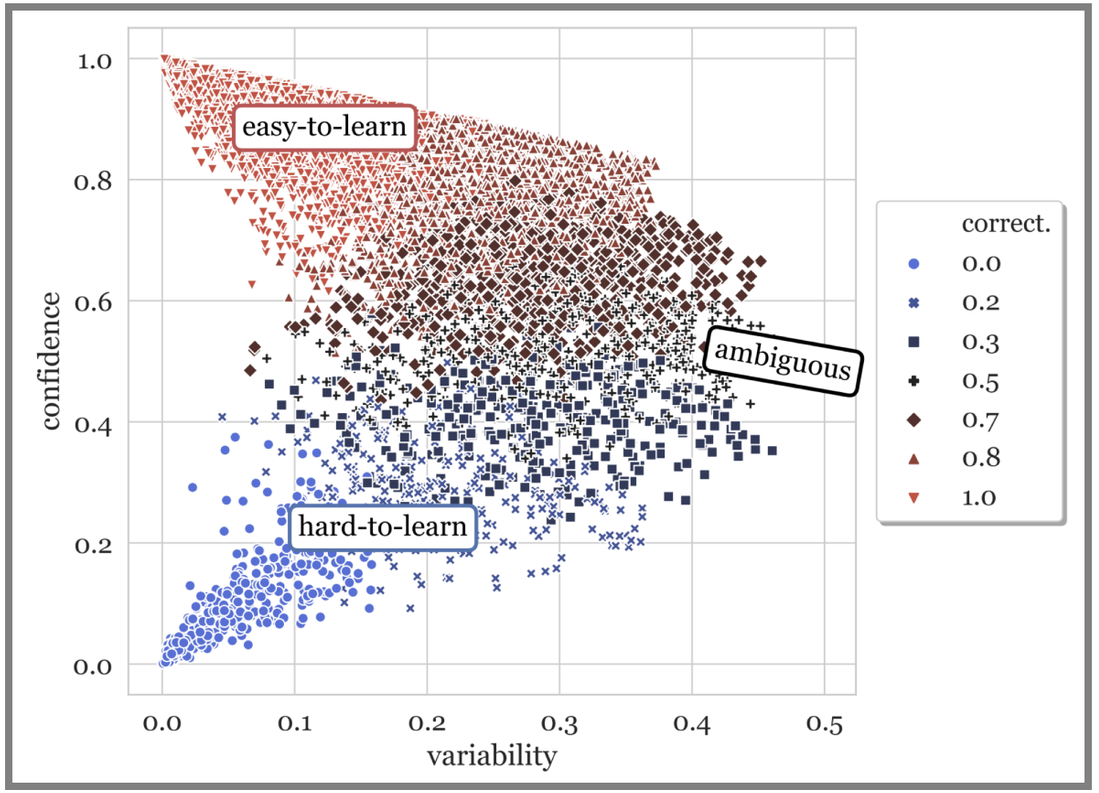
Make models more stable on examples that are very similar.

Types of Revisions	Examples
Recasting *fact* as *hoped for*	The world of Atlantis, hidden beneath the earth's core, is fantastic The world of Atlantis, hidden beneath the earth's core is **supposed** to be fantastic
Suggesting sarcasm	thoroughly captivating **thriller-drama, taking a deep and realistic** view thoroughly mind numbing **"thriller-drama", taking a "deep" and "realistic" (who are they kidding?)** view
Inserting modifiers	The presentation of simply Atlantis' landscape and setting The presentation of Atlantis' **predictable** landscape and setting
Replacing modifiers	"Election" is a highly fascinating and thoroughly **captivating** thriller-drama "Election" is a highly expected and thoroughly **mind numbing** "thriller-drama"
Inserting phrases	Although there's hardly any action, the ending is still shocking. Although there's hardly any action **(or reason to continue watching past 10 minutes)**, the ending is still shocking.
Diminishing via qualifiers	which, while usually containing some reminder of harshness, become **more and more intriguing**. which, usually containing some reminder of harshness, became **only slightly more intriguing**.
Differing perspectives	Granted, **not all of the story makes full sense**, but the film doesn't feature any amazing new computer-generated visual effects. Granted, **some of the story makes sense**, but the film doesn't feature any amazing new computer-generated visual effects.
Changing ratings	one of the worst ever scenes in a sports movie. **3 stars out of 10**. one of the wildest ever scenes in a sports movie. **8 stars out of 10**.

Dataset Difficulty: Data Map

Not all examples contribute equally to model training.

During training, instances that a model always predicts correctly are different from those it almost never does, or those on which it vacillates.



Swayamdipta, Swabha, et al. "Dataset cartography: Mapping and diagnosing datasets with training dynamics." EMNLP 2020

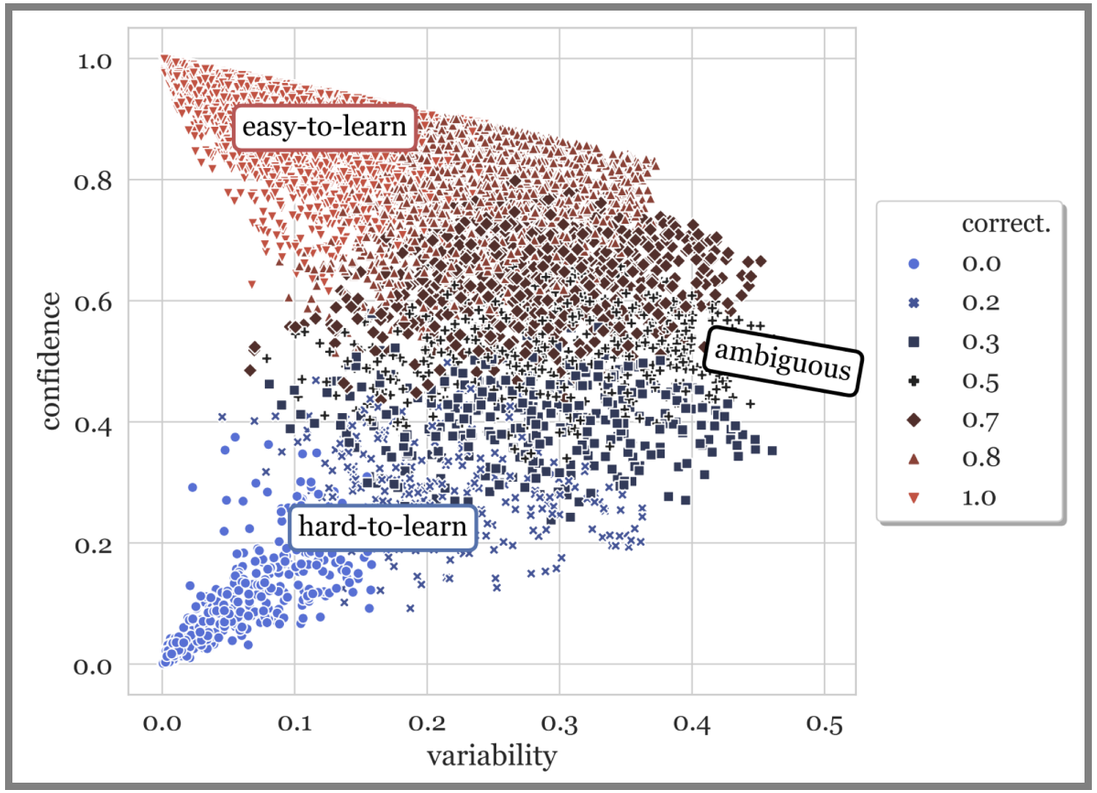
Dataset Difficulty: Data Map

Core idea: Construct model-dependent data map based on training dynamics (the behavior of a model as training progresses). Divide datasets into slices of easy-to learn, ambiguous and hard-to-learn.

Actually using the data accuracy vs. data precision idea!

Benefit: Diagnose and split datasets in the context of model training.

- Find data points useful for particular cases.
- Measure uncertainty.
- Detect mislabeled samples.



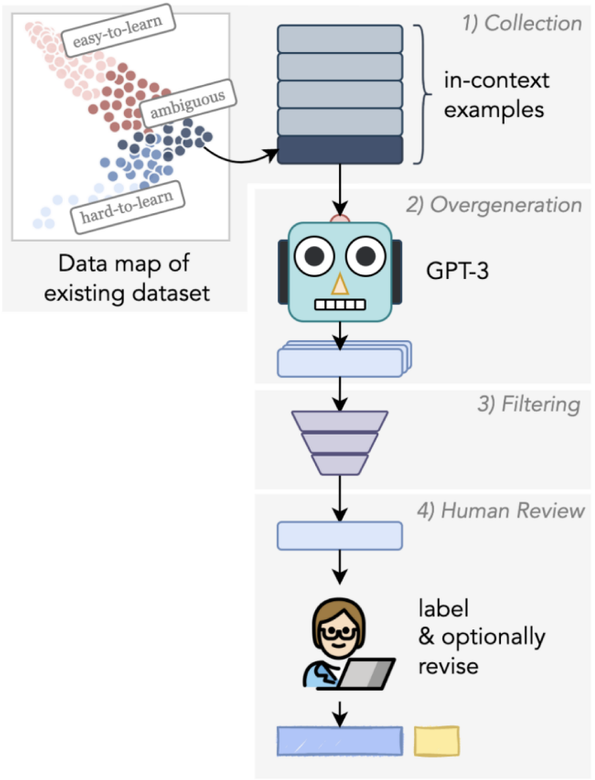
Mitigation on Dataset Difficulty

Core idea: Use data map to find useful input patterns, use LLMs to generate more of these patterns.

Make dataset more interesting and useful.

```
Selected different example:
P: 5% probability that each part is defect free.
H: Each part has a 95% chance of having a defect

Mimicked output:
P: 1% of the seats were vacant.
H: 99% of the seats were occupied.
```



Zhang, Shujian, Chengyue Gong, and Eunsol Choi. "Learning with different amounts of annotation: From zero to many labels." EMNLP 2021

Key takeaways

Data curation takes a lot of forms, but essentially they just want to improve the dataset on some of the measurements we mentioned.

So essentially,

- Pick data points that are relevant to the metric;
- Decide whether you want to use the data point in a different way;
- Decide whether you want to make some changes to the data.

Summary

- Data quality is a system-level concern
  - Data quality at the interface between components
  - Documentation and monitoring often poor
  - Involves organizational structures, incentives, ethics, ...
- Data from many sources, often inaccurate, imprecise, inconsistent, incomplete, ... -- many different forms of data quality problems
- Many mechanisms for enforcing consistency and cleaning
  - Data schema ensures format consistency
  - Data quality rules ensure invariants across data points
- Concept and data drift are key challenges -- monitor

Further Readings

- Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F. and Grafberger, A., 2018. Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), pp.1781-1794.
- Polyzotis, Neoklis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. "Data validation for machine learning." Proceedings of Machine Learning and Systems 1 (2019): 334-347.
- Polyzotis, Neoklis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. “Data Management Challenges in Production Machine Learning.” In Proceedings of the 2017 ACM International Conference on Management of Data, 1723–26. ACM.
- Theo Rekatsinas, Ihab Ilyas, and Chris Ré, “HoloClean - Weakly Supervised Data Repairing.” Blog, 2017.
- Ilyas, Ihab F., and Xu Chu. Data cleaning. Morgan & Claypool, 2019.
- Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. "A unifying view on dataset shift in classification." Pattern recognition 45, no. 1 (2012): 521-530.
- Vogelsang, Andreas, and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists." In Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2019.
- Humbatova, Nargiz, Gunel Jahangirova, Gabriele Bavota, Vincenzo Riccio, Andrea Stocco, and Paolo Tonella. "Taxonomy of real faults in deep learning systems." In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, pp. 1110-1121. 2020.