# Data Analysis

More on Data

# Announcements

- Clinic 2 is due on Monday (wildcards apply to the group)
- We run a mid-course survey. Since it also asks about time spent on clinic 2, (maybe) wait until submission to complete the survey https://forms.gle/uZHjN4nWsmeBQ8pG6
- We have a lecture pending on March 10th. You can choose what to cover https://app.wooclap.com/UMDA
    - ML infra, network data, pyspark (for big data), fairness and interpretability, ?
    - DA-ORCS crossover?

# Learning goals

- Discuss the iterative nature of training data
- Describe the steps in the sampling process
- Explain the principles of (non-)probability sampling and how they form a basis for making statistical inferences from a sample to a population
- Assess what type of sampling a data collection followed
- Identify which biases are related to some sampling process
- Describe the main pros/cons of different methods to label data
- Propose data labelling methods for practical problems
- Identify and address challenges caused by class imbalances
- Develop strategies to maintain data quality and mitigate biases

# Topics

1. Mind vs. data
2. Sampling
3. Labeling
4. Class imbalance

(also a bootcamp)

# WHO WOULD WIN?

Intelligent model architectures that took researchers their entire PhDs to design

Terabytes of data scraped from Reddit in a week

**Who would win?**

**A.** Intelligent design

**B.** TB of Reddit data

# Mind

"Data is profoundly dumb."

*Judea Pearl, [Mind over data - The Book of Why]*


# Data

"General methods that leverage computation are ultimately the most effective, and by a large margin … Human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation."

*Richard Sutton, [Bitter Lesson]*

"We don't have better algorithms. We just have more data."

*Peter Norvig, [The Unreasonable Effectiveness of Data]*

"Imposing structure requires us to make certain assumptions, which are invariably wrong for at least some portion of the data."

*Yann LeCun, [Deep Learning and Innate Priors]*

Data is necessary.

The debate is whether *finite** data is sufficient.

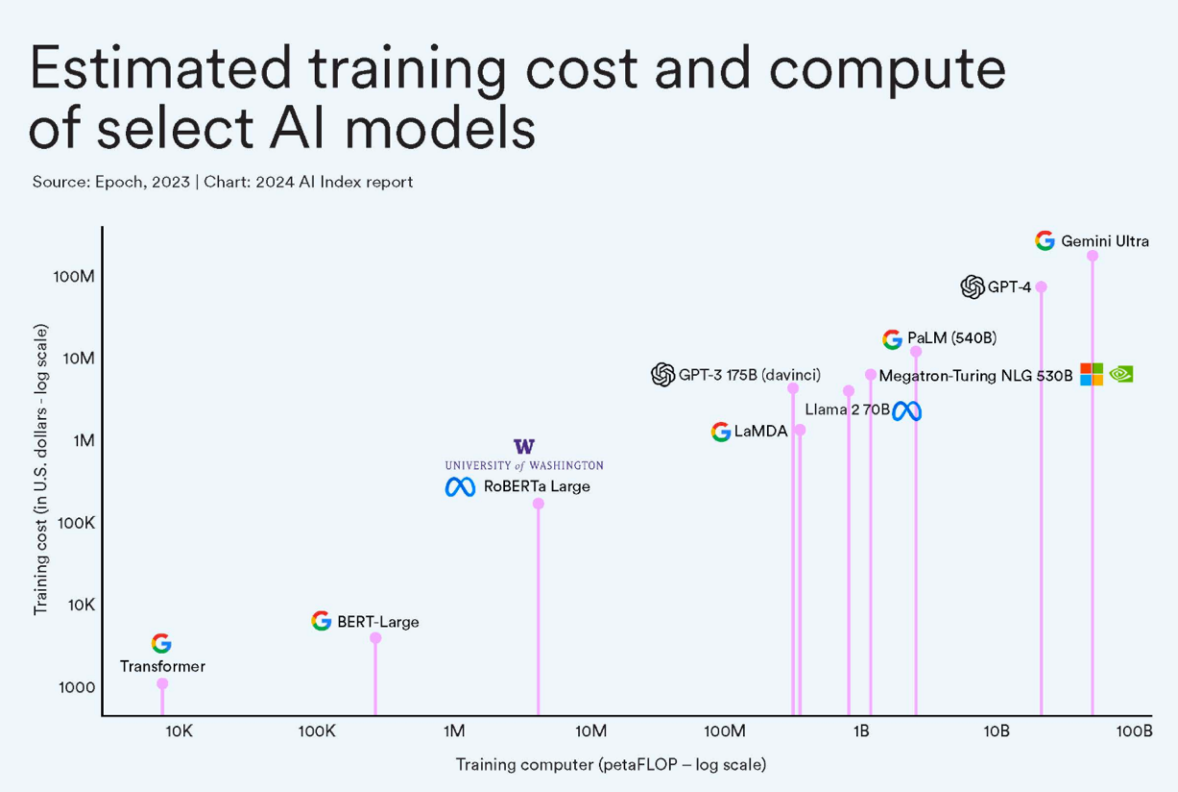
* If we had infinite data, we can solve arbitrarily complex problems by just looking up the answers.

Massive data $\neq$ infinite data

does not mean $\infty$

# ~~More data (generally) needs more compute~~

“amount of compute used in the largest AI training runs has doubled every 3.5 months”






AI and Compute (OpenAI 2018) and Stanford HAI (April 2024)

# ⚠ Data: full of potential for biases ⚠

diffrt ww

- sampling/selection biases
- under/over-representation of subgroups
- human biases embedded in historical data
- labeling biases
- ...

Algorithmic biases not covered (yet)!

# Sampling

Sampling is essential in all steps of data analysis, e.g.

- Sampling from real-world data to create training data to create Dataset
- Sampling to create splits for train/validation/test
- Sampling to monitor model performance
- ...

# Key concepts in sampling

Greaty Dot

Population: The group that you want to learn something about.

Sampling frame: The list from which the sample is drawn.

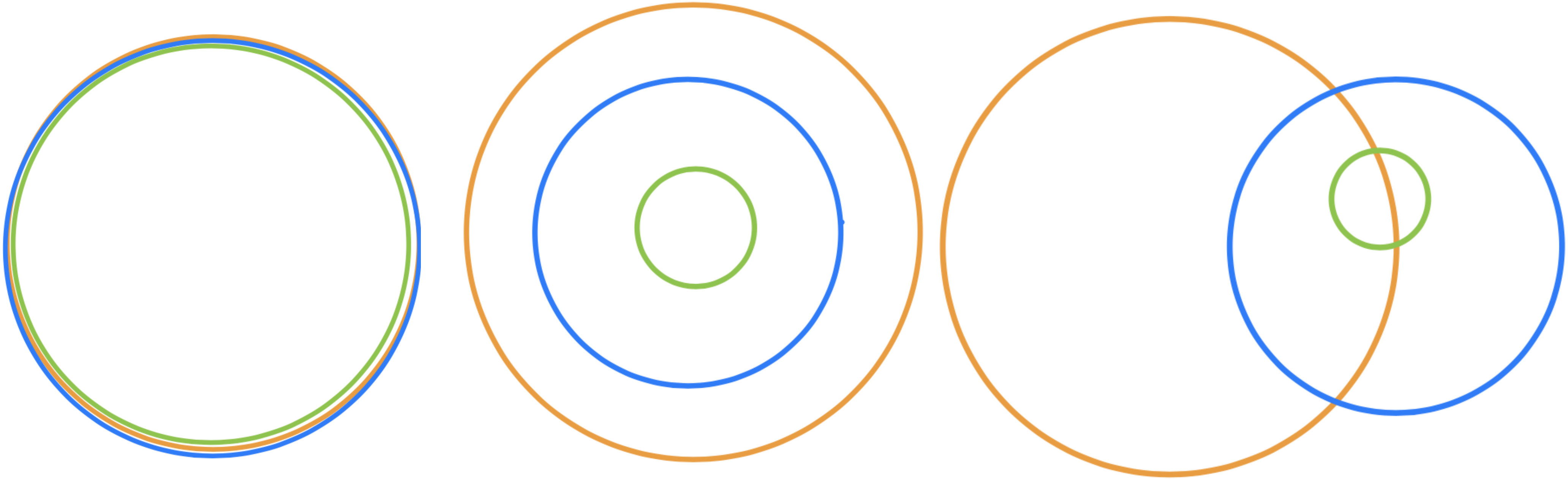
Sample: A subset of the sampling frame (or who you actually end up sampling)

# Sampling in practice

Census

What we think we have

What we actually have



Population

Sampling frame

Sample

# Sampling from a finite population

- A census is great, but expensive and difficult to execute.
- A **sample** is a subset of the population.
    - Samples are often used to make **inferences about the population**.
    - How you draw the sample will affect your accuracy.
- Two common sources of error:
    - **chance error**: random samples vary from what is expected in any direction.
    - **bias**: a systematic error in one direction.

Let's look at some examples!

# An example

**Example:** Suppose we have a cage of 20 mice, and each week, we want to measure the weights of these mice. To do so, we randomly pick *some* mice every week these mice, and weigh them.

That's a random sample. True of False?

Say now we have 1.000.000 mice. We follow the same process as above. Is that a random sample?

More Data does not mean Better

Size does not play a role

Having more Data does not mean Best

# Types of sampling

- Non-probability sampling
  - **Convenience** sampling: selection based on availability

    I just open the Cage
    - Soliciting response
    - Choosing existing datasets

      Most ML Data
    - Looking at available reviews on Amazon
  - **Snowball** sampling: future samples are selected based on existing samples
    - E.g. to scape legit Twitter accounts, start with seed accounts then scrape their following
  - **Judgment** sampling: experts decide what to include

    Rely on Domain Expert
  - **Quota** sampling: quotas for certain slices of data (no randomization)
  - ....

# Case study – 1936 US Presidential Election

Roosevelt (D)

Landon (R)

In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, polls were conducted in the months leading up to the election to try and predict the outcome.

https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

# The Literary Digest

They had successfully predicted the outcome of 5 general elections coming into 1936.

The Literary Digest

NEW YORK — OCTOBER 31, 1936

*Topics of the day*

**LANDON, 1,293,669; ROOSEVELT, 972,897**

**Final Returns in The Digest's Poll of Ten Million Voters**

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased The Literary Digest?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased The Literary Digest?" "Is the Pope of Rome a stockholder of The Literary Digest?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

**Problem**—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager \$100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

The statistics and the material in this article are the property of Funk & Wagnalls Company and have been copyrighted by it; neither the whole nor any part thereof may be reprinted or published without the special permission of the copyright owner.

They sent out their survey to **10,000,000** individuals, who they found from:

- Phone books.
- Lists of magazine subscribers.
- Lists of country club members.

# The Literary Digest prediction

The Literary Digest's **prediction**:

43% Roosevelt, 57% Landon

The **actual** outcome of the election:

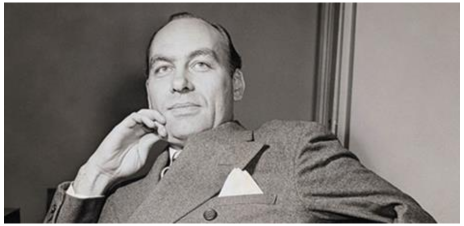
61% Roosevelt, 37% Landon

How could this have happened?

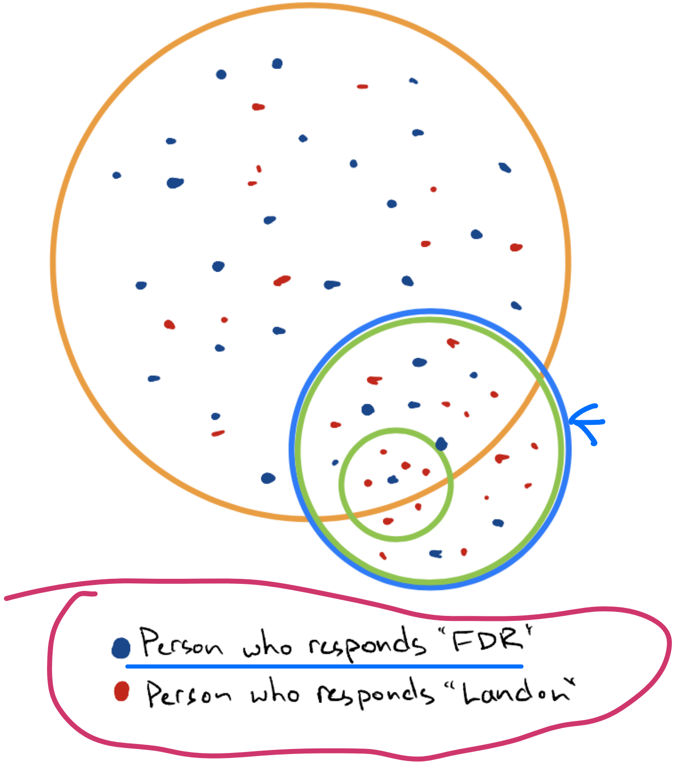
**They surveyed 10 million people!**

- Their sample was **not representative** of the population.
  - They sampled people who owned phones, subscribed to magazines, and went to country clubs, who at the time were more affluent.
  - These people tended to vote Republican (Alf Landon).
- Only 2.4 million people **actually filled out the survey!**
  - 24% response rate (low).
  - Who knows how the other 76% would have polled?

# Meanwhile...



- George Gallup, a rising statistician, predicted that Roosevelt would win with **56% of the vote.** His sample size was just 50,000!
- Gallup also predicted what The Literary Digest was going to predict, within 1%!
    - He predicted that they would survey people in the phone book, people who subscribed to magazines, and who were part of country clubs.
    - So, he sampled those same individuals (just 3000!)



# Common biases

Big samples are not always good, you need a representative sample!

**Selection Bias**

- Systematically excluding (or favoring) particular groups.
- How to avoid: Examine the sampling frame and the method of sampling.

**Response Bias**

- People don't always respond truthfully.
- How to avoid: Examine the nature of questions and the method of surveying.

**Non-response Bias**

- People don't always respond.
- How to avoid: Keep your surveys short and be persistent.
- People who don't respond aren't like the people who do!

# Data used in ML is mostly driven by convenience

- Language models: BookCorpus, CommonCrawl, Wikipedia, Reddit links
- Sentiment analysis: IMDB, Amazon
  - Only users who have access to the Internet and are willing to put reviews online
- Self-driving cars: most data is from the Bay Area (CA) and Phoenix (AZ)
  - Very little data on raining & snowing weather

⚠ Lots of biases in data! ⚠

# Types of sampling

- Non-probability sampling
- Random sampling
  - Simple random sampling
  - Stratified sampling
  - Weighted sampling
  - Reservoir sampling
  - ...

# Simple random sampling

- Each sample in population has an equal chance of being selected
  - E.g. select 10% of all samples in population

| Pros | Cons |
|---|---|
| • Simple (easiest type of random sampling) | • No representation guarantee: might exclude rare classes (black swan!) |

We just need to guarantee randomness

# Stratified sampling

- Divide population by subgroups
  - Slices of data
    - 20% of each age group: 18-24, 25-34, 35+, etc.
  - Classes
    - 2% of each class


| Pros | Cons |
|---|---|
| Minor groups are represented | Can't be used when:<br>• samples can't be put into subgroups<br>• samples can belong in multiple subgroups (multilabel) |

Image from https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

# Weighted sampling

also when want to select Dom
Cege

- Each element is given a weight, which determines the probability of being selected.
  - If you want to select a sample 30% of the time, give it 3/10 weight
- Might embed domain knowledge
  - E.g. know distribution of your target population or want to prioritize recent samples

```
random.choices(population=[1, 2, 3, 4, 100, 1000],
               weights=[0.2, 0.2, 0.2, 0.2, 0.1, 0.1],
               k=2)
```

||

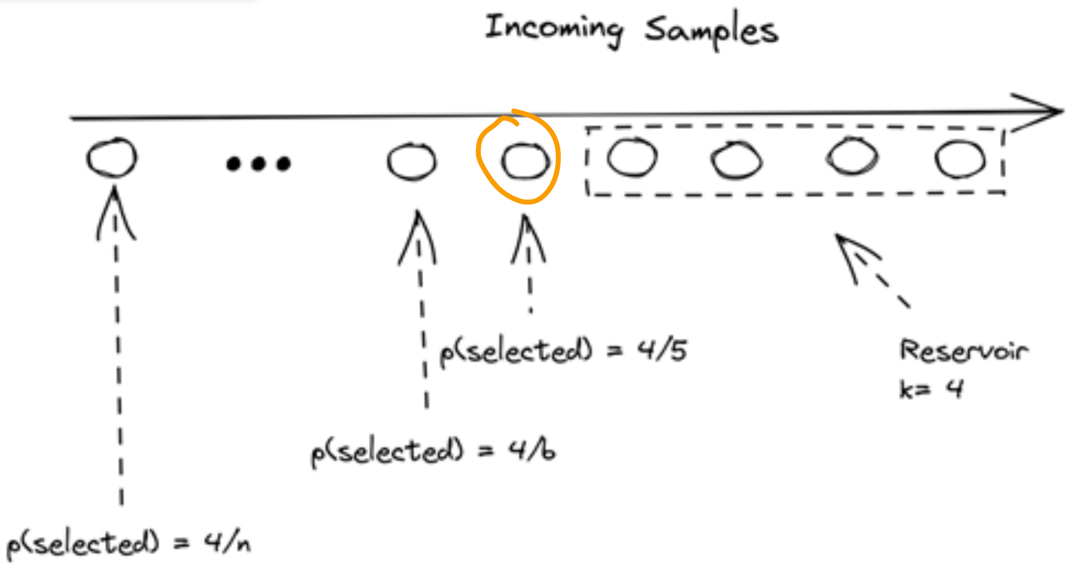
```
random.choices(population=[1, 1, 2, 2, 3, 3, 4, 4, 100, 1000],
               k=2)
```

# Reservoir sampling: problem

- Need select k samples from a stream of n samples with equal probability
  - n is unknown
  - impossible/inefficient to fit all in memory
- Can stop the stream any moment and get the required samples

# Reservoir sampling: solution

1. First k elements are put in reservoir
2. For each incoming $i^{th}$ element, generate a random number j between 1 and i
    a. If $1 \leq j \leq k$: replace $j^{th}$ in reservoir with $i^{th}$
3. Each incoming element has k/i chance of being in reservoir!



# With vs. without replacement

| With replacement | Without replacement |
|---|---|
| Same item can be chosen more than once | Same item can't be chosen more than once |
| • No covariance between two chosen samples<br>• Approximate true population distribution | • Covariance between two chosen samples<br>• Covariance reduced as dataset size becomes large |
| e.g. bagging (coming up in next lecture) | e.g. mini batch gradient descent |

The Same Item Can be Use
more than Once

# Note: Gradient descent variants

| Variant | Gradient Computation | Update Frequency | Computational Cost | Convergence Speed |
|---|---|---|---|---|
| **Batch Gradient Descent (BGD)** | Uses the entire dataset | After processing all samples | High (slow for large datasets) | Stable, but slow |
| **Stochastic Gradient Descent (SGD)** | Uses a single random sample | After every sample | Low (fast per update) | Faster, but noisier |
| **Mini-Batch Gradient Descent (MBGD)** | Uses a small subset (mini-batch) | After processing a mini-batch | Medium (balance between efficiency & stability) | Faster than BGD, smoother than SGD |

Compare between the 3

# Labeling

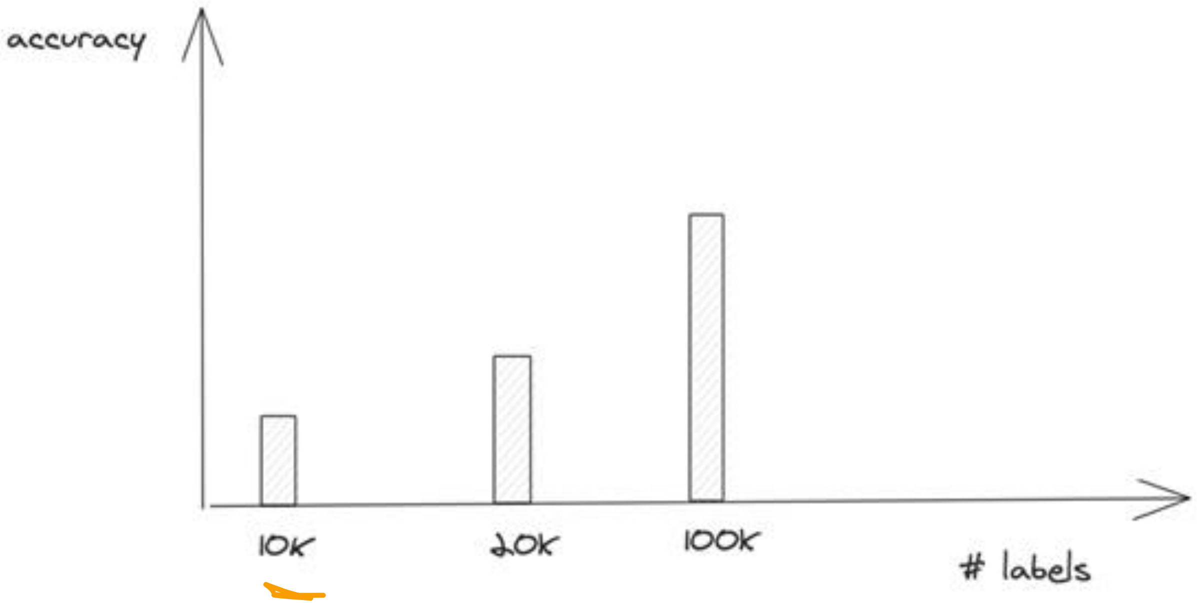
# Labeling

1. Hand-labeling
2. Programmatic labeling
3. Weak supervision, semi supervision, active learning, transfer learning

*When I told our recruiters that I wanted an in-house labeling team, they asked how long I'd need this team for. I told them: "How long do we need an engineering team for?"*

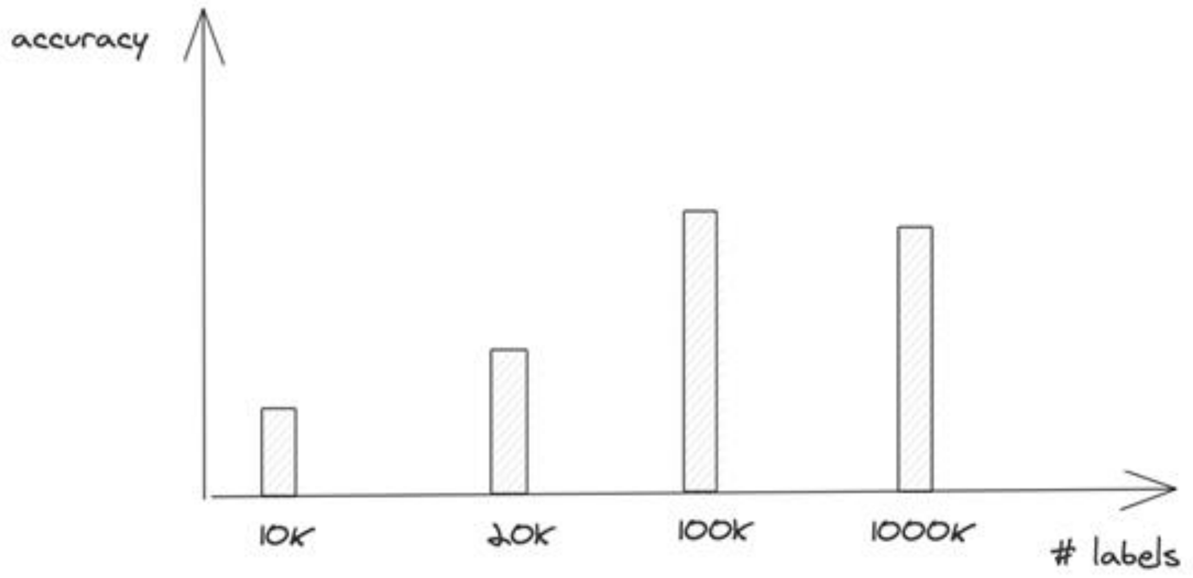
Andrej Karpathy, Director of AI @ Tesla

# ⚠ More data isn't always better ⚠



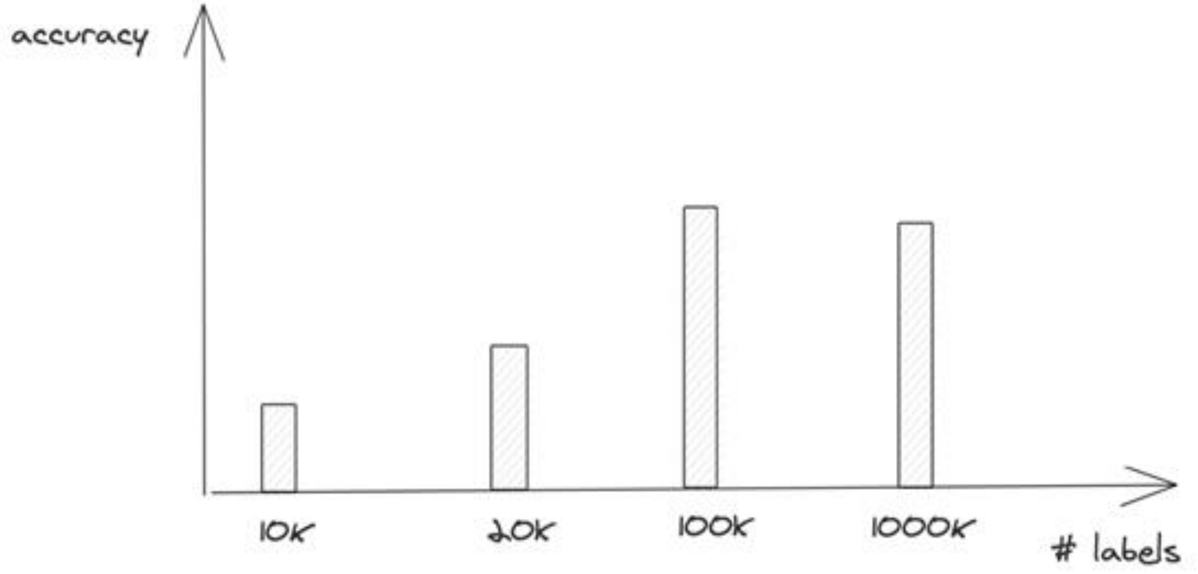

🧠 Idea 🧠: crowdsource data to get 1 million labels!

# ⚠ More data isn't always better ⚠




Why is the model getting worse?

# ⚠ Label sources with varying accuracy ⚠




- 100K labels: internally labeled, high accuracy
- 1M labels: crowdsourced, noisy

# Label multiplicity: example

**Which annotator i correct?**

**Task: label all entities in the following sentence:**

Darth Sidious, known simply as the Emperor, was a Dark Lord of the Sith who reigned over the galaxy as Galactic Emperor of the First Galactic Empire.

NLP Task

| Annotator | # entities | Annotation |
|---|---|---|
| 1 | 3 | [**Darth Sidious**], known simply as the Emperor, was a [**Dark Lord of the Sith**] who reigned over the galaxy as [**Galactic Emperor of the First Galactic Empire**] |
| 2 | 6 | [**Darth Sidious**], known simply as the [**Emperor]**, was a **[Dark Lord**] of the [**Sith**] who reigned over the galaxy as [**Galactic Emperor**] of the [**First Galactic Empire**]. |
| 3 | 4 | [**Darth Sidious**], known simply as the [**Emperor]**, was a **[Dark Lord of the Sith**] who reigned over the galaxy as [**Galactic Emperor of the First Galactic Empire**]. |

# Label multiplicity

More expertise required (more difficult to label),
more room for disagreement!

If experts can't agree on a label, time to rethink human-level performance

# Label multiplicity: solution

We want large no [illegible]
for subject

- Clear problem definition
  - Pick the entity that comprises the longest substring

| Annotator | # entities | Annotation |
|---|---|---|
| 1 | 3 | [**Darth Sidious**], known simply as the Emperor, was a [**Dark Lord of the Sith**] who reigned over the galaxy as [**Galactic Emperor of the First Galactic Empire**] |
| 2 | 6 | [**Darth Sidious**], known simply as the [**Emperor**], was a [**Dark Lord**] of the [**Sith**] who reigned over the galaxy as [**Galactic Emperor**] of the [**First Galactic Empire**]. |
| 3 | 4 | [**Darth Sidious**], known simply as the [**Emperor**], was a [**Dark Lord of the Sith**] who reigned over the galaxy as [**Galactic Emperor of the First Galactic Empire**]. |

# Label multiplicity: solution

- Clear problem definition
- Annotation training
- Data lineage: track where data/labels come from Provenance
- Learning methods with noisy labels
  - Learning with Noisy Labels (Natarajan et al., 2013)
  - Loss factorization, weakly supervised learning and label noise robustness (Patrini et al., 2016)
  - Cost-Sensitive Learning with Noisy Labels (Natarajan et al., 2018)
  - Confident Learning: Estimating Uncertainty in Dataset Labels (Northcutt et al., 2019)

# Label multiplicity: Not always majority voting

Think about sensitive topics,
e.g. stereotypes or offensive speech

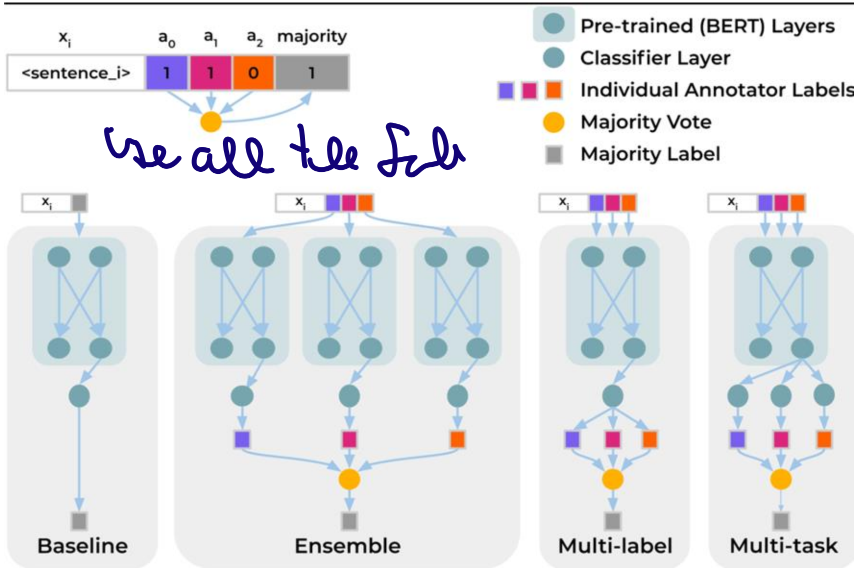



diagram
depends on the
Problem Card
Do a Single ...



Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations, 2022

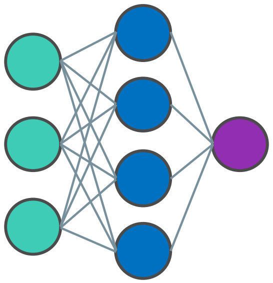
# Programmatic labeling

# ~~Training data is~~ the bottleneck

## Data


Key differentiator

## Algorithms



```
from transformers \
  import BertModel as model
```

Increasingly commoditized

## ML Model

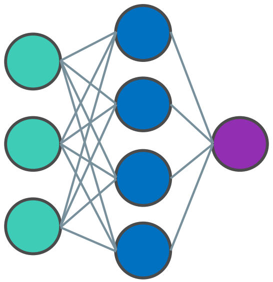
"We don't have better algorithms. We just have more data."

*Peter Norvig, [The Unreasonable Effectiveness of Data]*

# Training data is the bottleneck

Data


Algorithms



ML Model

- 8 Person-months
- 8-9 pt. differences

- 1-2 days
- <1 pt. differences

How to get training data in days?



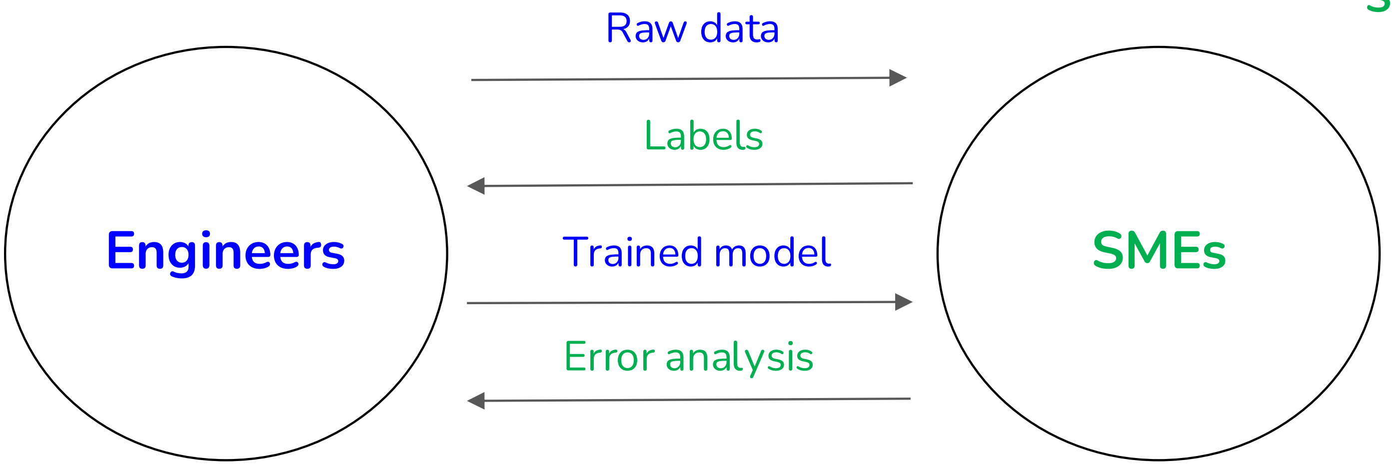
[Cross-Modal Data Programming Enables Rapid Medical Machine Learning](#) (Dunnmon et al., 2019)

# Hand labeling data is …


- **Expensive:** esp. when **subject matter expertise** required
- **Non-private:** Need to ship data to human annotators
- **Slow:** Time required scales linearly with # labels needed
- **Non-adaptive:** Every change requires re-labeling the dataset

# Cross-functional communication

SME = subject-matter expert



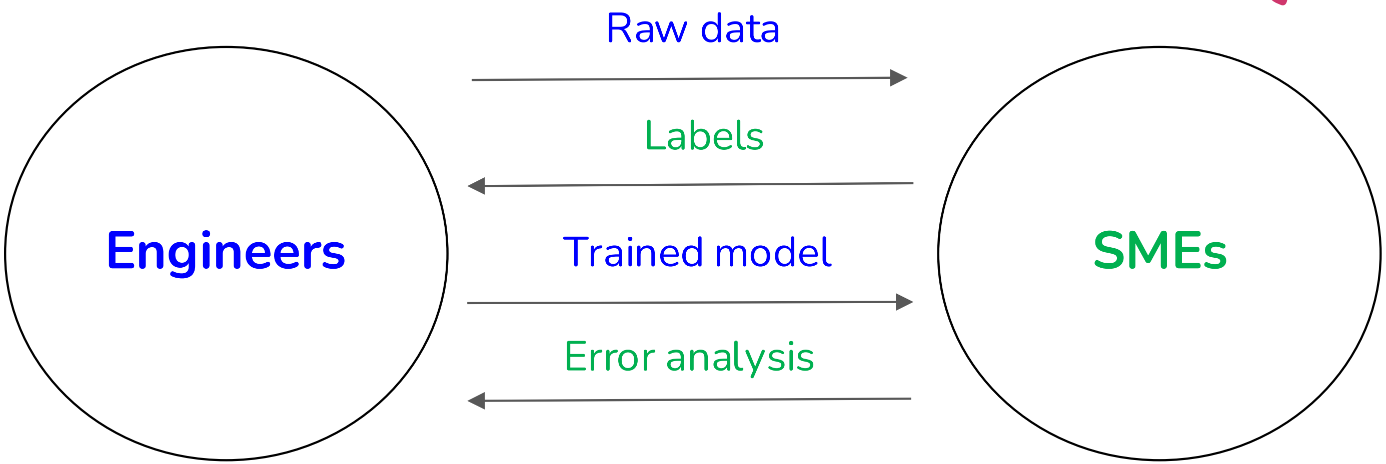

```
def function:
   if X:
           do Y
```

```
If the nurse's note mentions
serious conditions like pneumonia,
the patient's case should be given
priority consideration.
```

**Code**: version control, reuse, share

How to version, share, reuse **expertise**?

# SME as labeling functions




```
def function:
    if X:
            do Y
```

```
def labeling_function(note):
    if "pneumonia" in note:
          return "EMERGENT"
```

**Labeling functions (LFs)**: Encode SME heuristics as functions and use them to label training data *programmatically*

# LFs: can express many different types of heuristics

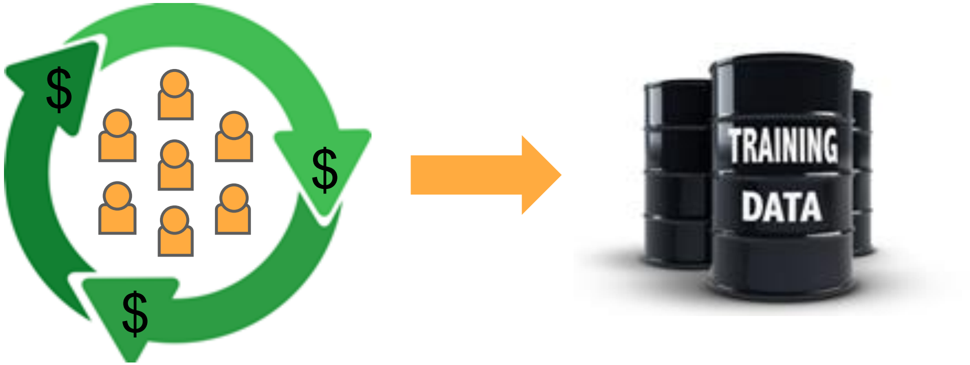
| | | |
|---|---|---|
| (.*) | Pattern Matching | If a phrase like "send money" is in email |
| | Boolean Search | If unknown_sender AND foreign_source |
| | DB Lookup | If sender is in our Blacklist.db |
| | Heuristics | If SpellChecker finds 3+ spelling errors |
| | Legacy System | If LegacySystem votes spam |
| | Third Party Model | If BERT labels an entity "diet" |
| | Crowd Labels | If Worker #23 votes spam |

# LFs: can express many different types of heuristics

"If nurse's report says 'malignant', likely to be emergent"

"If it matches a list of patient names…"

"If our legacy model thinks it's emergent…"




**Labeling functions**: Simple, flexible, interpretable, adaptable, fast

# LFs: powerful but noisy

```
def LF_contains_money(x):
    if "money" in x.body.text:
        return "SPAM"
```

```
def LF_from_grandma(x):
    if x.sender.name is "Grandma":
        return "HAM"
```

```
def LF_contains_money(x):
    if "free money" in x.body.text:
        return "SPAM"
```

From: **Grandma**

"Dear handsome grandson,
Since you can't be home for Thanksgiving dinner this year, I'm sending you some **money** so you could enjoy a nice meal ..."

??

"You have been pre-approved for free **cash** ..."

??

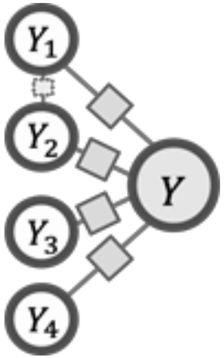
- **Noisy**: Unknown, inaccurate
- **Overlapping**: LFs may be correlated
- **Conflicting**: different LFs give different labels
- **Narrow**: Don't generalize well

# LF labels are combined to generate ground truths

```
def LF_contains_money(x):
   if "money" in x.body.text:
      return "SPAM"
```

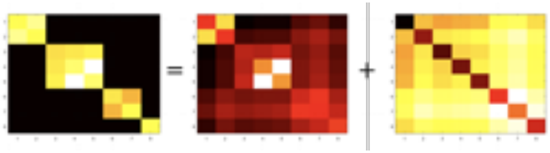
```
def LF_from_grandma(x):
   if x.sender.name is "Grandma":
      return "HAM"
```

```
def LF_contains_money(x):
   if "free money" in x.body.text:
      return "SPAM"
```




**[Intuition]**

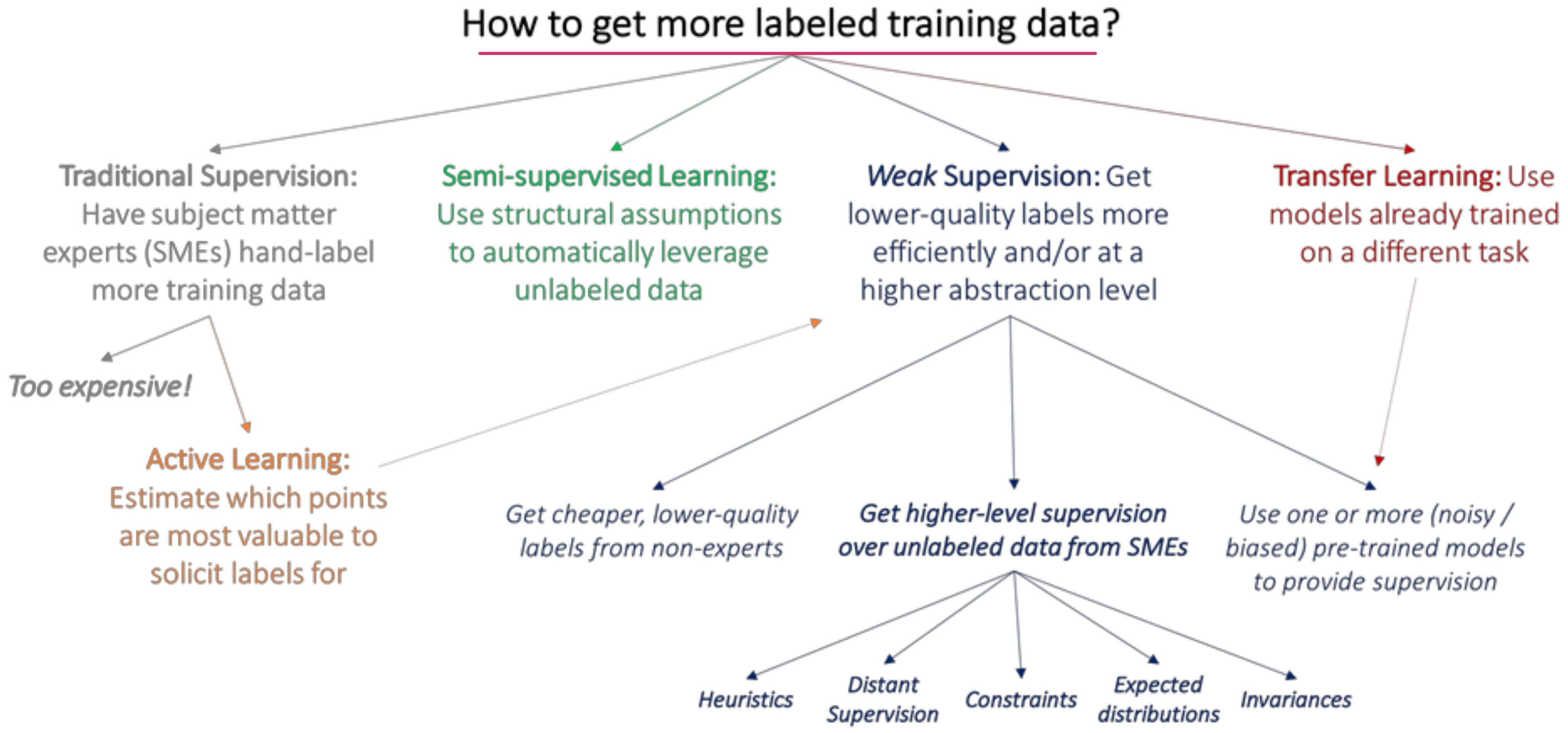
Look at agreements & disagreements



[Ratner et. al. NeurIPS'16;
Bach et. al. ICML'17;
Ratner et. al. AAAI'19;
Varma et. al. ICML'19l;
Sala et. al. NeurIPS'19;
Fu et. al. ICML'20]

| Hand labeling | Programmatic labeling |
| --- | --- |
| Expensive: esp. when subject matter expertise required | Cost saving: Expertise can be versioned, shared, reused across organization |
| Non-private: Need to ship data to human annotators | Privacy: Create LFs using a cleared data subsample then apply LFs to other data without looking at individual samples. |
| Slow: Time required scales linearly with # labels needed | Fast: Easily scale 1K -> 1M samples |
| Non-adaptive: Every change requires re-labeling the dataset | Adaptive: When changes happen, just reapply LFs! |

# Weak supervision, semi-supervision, active learning, transfer learning




Weak Supervision: A New Programming Paradigm for Machine Learning (Ratner et al., 2019)

# Weak supervision

- Leverage noisy, imprecise sources to create labels
  - e.g. if “money” is in an email it’s probably spam

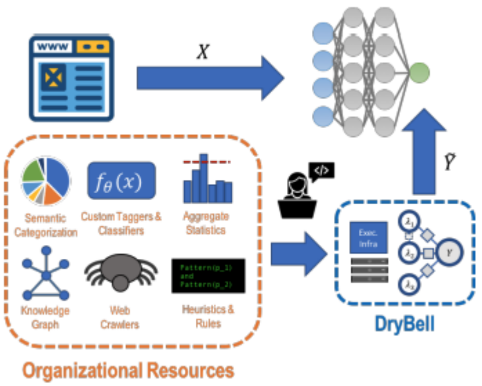
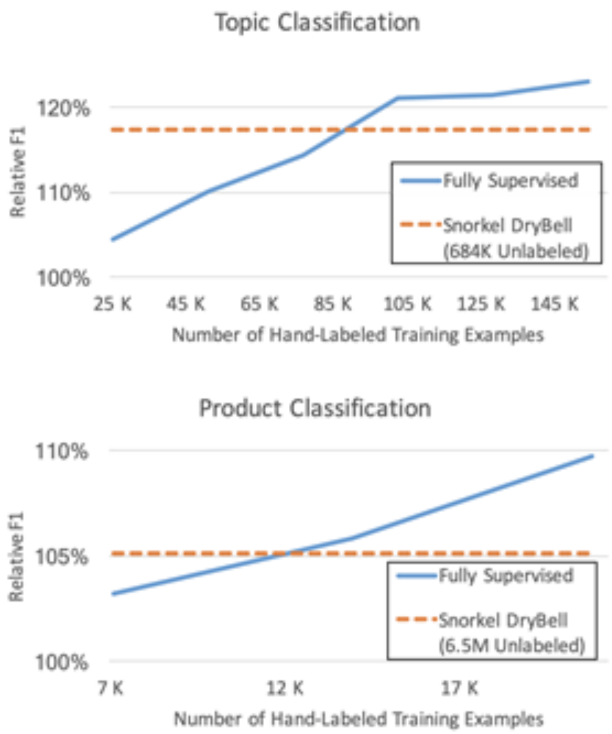



**Figure 1: Rather than using hand-labeled training data, Snorkel DryBell uses diverse organizational resources as weak supervision to train content and event classifiers on Google’s platform.**






[Harnessing Organizational Knowledge for Machine Learning](#) (Ratner et al., Google AI Blog 2019)

# Semi-supervision

- Use structural assumptions to leverage a large amount of unlabeled data together with a small amount of labeled data
    - Hashtags in the same profile/tweet are probably of similar topics


- Might require complex algorithms like clustering to discover similarity

# Semi-supervision: self-training

1. Train model on a *small set* of labeled data
2. Use this model to generate predictions for *unlabeled data*
3. Use predictions with *high raw probabilities* as labels
4. Repeat step 1 with new labeled data

# Semi-supervision: perturbation-based methods

Assumption: small perturbation wouldn’t change a sample’s label

- Add white noises to images
- Add small values to word embeddings or tabular data

Also a data augmentation method!

# Transfer learning

- Apply model trained for one task to another task
    - CV and NLP have been revolutionized
        - Fine-tuning
        - Prompt-based

    - Work on tabular data has also been applied, mainly for domain adaptation

Language Models are Few-Shot Learners (OpenAI 2020)
Enhanced boosting-based transfer learning for modeling ecological momentary assessment data (Ntekouli et. al, 2024)

# Active learning

- **Assumption:** ML models can achieve better performance if they can choose what samples to learn from

- **Goal**: Increase the efficiency of labels
- Label samples that are estimated to be most valuable to the model according to some metrics



Active Learning Literature Survey (Burr Settles, 2010)

# Active learning metrics

- Uncertainty measurement
  - e.g. label samples with lowest raw probability for the predicted class
- Candidate models' disagreement
  - Have several candidate models (e.g. models with different hypeparams)
  - Each model makes its own prediction
  - Label samples with most disagreement



Active Learning Literature Survey (Burr Settles, 2010)

# Active learning

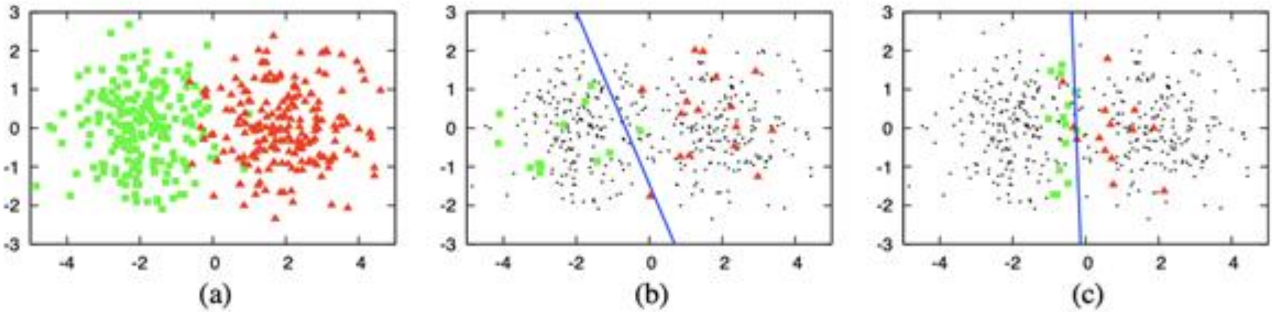



Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (70% accuracy). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (90%).



Active Learning Literature Survey (Burr Settles, 2010)

| Method | How | Ground truths required? |
|---|---|---|
| Weak supervision | Leverages (often noisy) heuristics to generate labels | No, but a small number of labels is useful to guide the development of heuristics |
| Semi-supervision | Leverages structural assumptions to generate labels | Yes. A small number of initial labels as seeds to generate more labels |
| Transfer learning | Leverages models pretrained on another task for your new task | No for zero-shot learning<br>Yes for fine-tuning, though # GTs required is often much less than # GTs required if training from scratch. |
| Active learning | Labels data samples that are most useful to your model | Yes |

# ⚠ There is no substitute for high quality human labels ⚠

# Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

## Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

**Any other comments?**

None.

## Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was mostly observable as raw text, except that the labels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

The sample of instances collected is English movie reviews from the `rec.arts.movies.reviews` newsgroup, from which a "number of stars" rating could be extracted. The sample is limited to forty reviews per unique author in order to achieve broader coverage by authorship. Beyond that, the sample is arbitrary.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Unknown to the authors of the datasheet.

**Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown to the authors of the datasheet.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown to the authors of the datasheet.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

As described above, the data was collected from newsgroups.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No. The data was crawled from public web sources, and the authors of the posts presumably knew that their posts would be public, but the authors were not explicitly informed that their posts were to be used in this way.

Datasheets for Datasets (Gebru et.al, 2018)

# ⚠ There is no substitute for high quality human treatment ⚠

BUSINESS • TECHNOLOGY

## Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

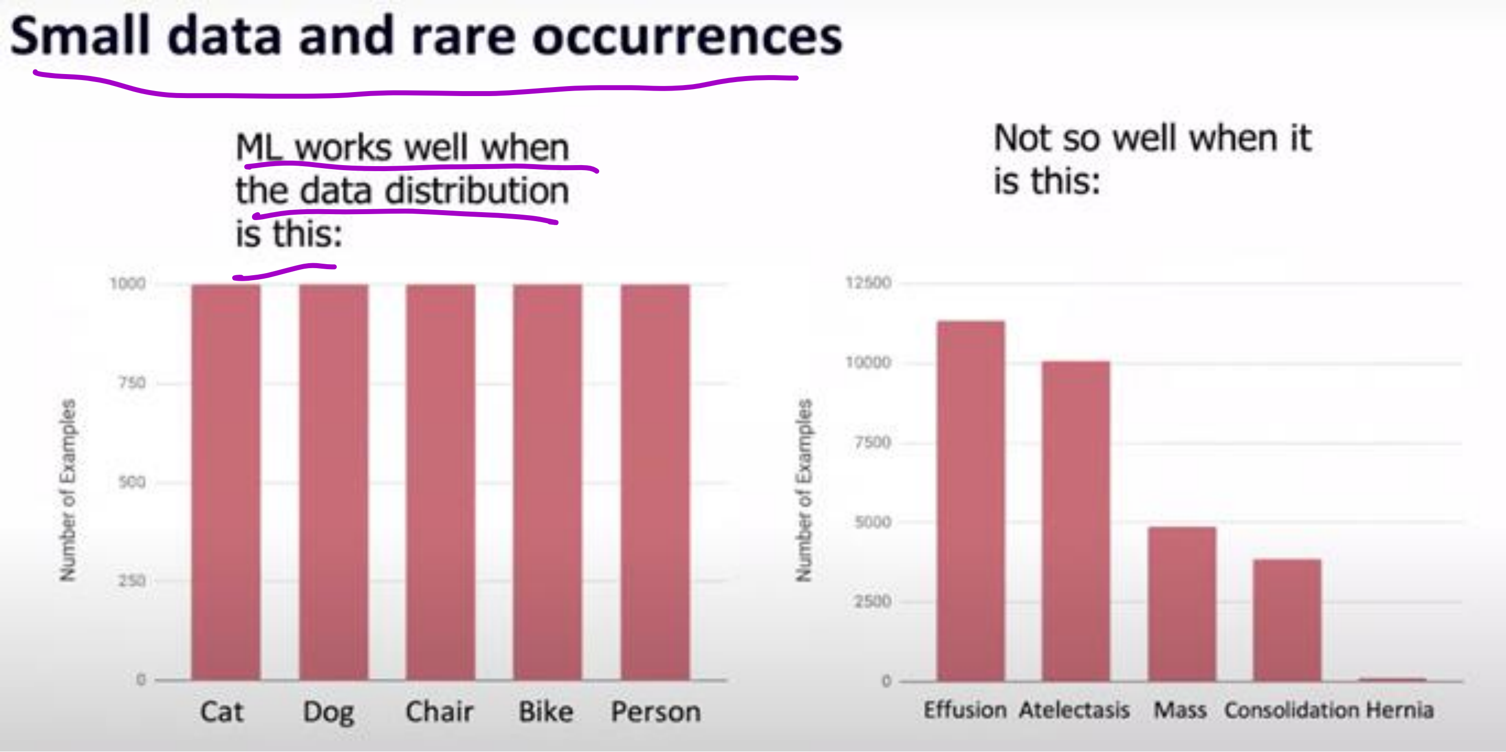
- Annotators have to endure:
  - Underpayment and exploitation
  - Exposure to disturbing content
  - Lack of recognition and support

---

https://time.com/6247678/openai-chatgpt-kenya-workers/
https://hai.stanford.edu/news/exploring-complex-ethical-challenges-data-annotation
https://netzpolitik.org/2024/data-workers-inquiry-the-hidden-workers-behind-ai-tell-their-stories/

# Class imbalance

# Class imbalance






Andrew Ng: Bridging AI's Proof-of-Concept to Production Gap (2020)

# Why is class imbalance hard?

- Not enough signal to learn about rare classes

# Why is class imbalance hard?

- Not enough signal to learn about rare classes
- Statistically, predicting majority label has higher chance of being right
  - If a majority class accounts 99% of data, always predicting it gives 99% accuracy

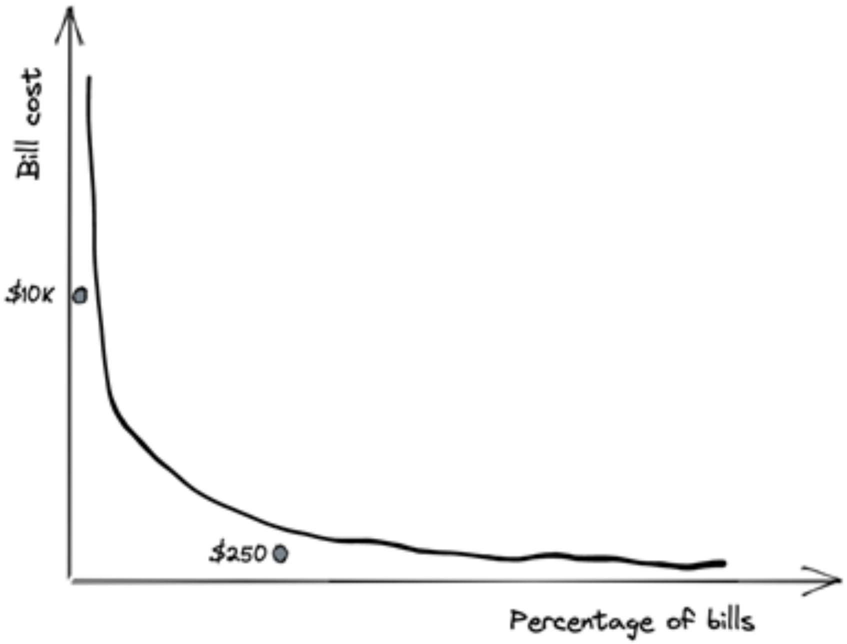
# Why is class imbalance hard?

- Not enough signal to learn about rare classes
- Statistically, predicting majority label has higher chance of being right
- Asymmetric cost of errors: different cost of wrong predictions

In Practical Problems

# Asymmetric cost of errors: regression

- 95th percentile: $10K
- Median: $250






Credit to [Eugene Yan]

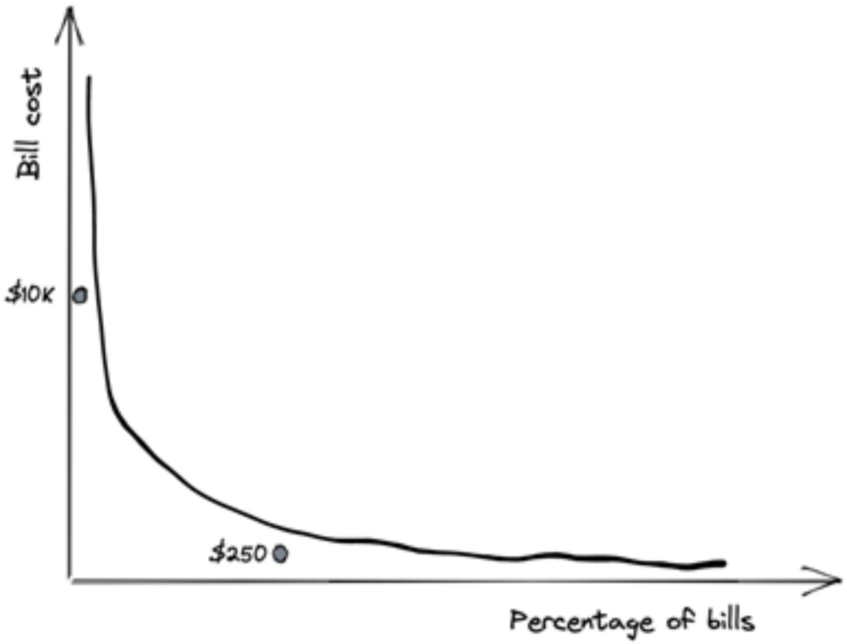
# Asymmetric cost of errors: regression

100% error difference

Not OK

- $10K bill: off by $10K
- $250 bill: off by $250

OK




Credit to [Eugene Yan]

# Class imbalance is the norm

Real World

People are more interested in unusual/potentially catastrophic events

- Fraud detection
- Spam detection
- Disease screening
- Churn prediction
- Resume screening
  - E.g. 2% of resumes pass screening
- Object detection
  - Most bounding boxes don't contain any object

Image from [PyImageSearch]

# Sources of class imbalance

- Sampling biases
    - Narrow geographical areas (self-driving cars)
    - Selection biases
- Domain specific Medical Probs, Fraud Cases ) Most Case
    - Costly, slow, or infeasible to collect data of certain classes
- Labeling errors

# How to deal with class imbalance

1. Choose the right metrics (we covered this already!)
2. Data-level methods
3. Algorithm-level methods

# Reminder: Metrics

| Symmetric metrics | Asymmetric metrics |
| --- | --- |
| Treat all classes the same | Measures a model's performance w.r.t to a class |
| Accuracy | F1, recall, precision, AUROC |

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}$$

$$F_1\text{–score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

- TP: True positives
- TN: True negatives
- FP: False positives
- FN: False negatives

# 1. Choose the right metrics

Model A vs. Model B confusion matrices

| Model A | Actual CANCER | Actual NORMAL |
|---|---|---|
| Predicted CANCER | 10 | 10 |
| Predicted NORMAL | 90 | 890 |

Save money

| Model B | Actual CANCER | Actual NORMAL |
|---|---|---|
| Predicted CANCER | 90 | 90 |
| Predicted NORMAL | 10 | 810 |

**POLL: Which model would you choose?**

# Choose the right metrics

Model A vs. Model B confusion matrices

| **Model A** | **Actual CANCER** | **Actual NORMAL** |
|---|---|---|
| Predicted CANCER | 10 | 10 |
| Predicted NORMAL | 90 | 890 |

| **Model B** | **Actual CANCER** | **Actual NORMAL** |
|---|---|---|
| Predicted CANCER | 90 | 90 |
| Predicted NORMAL | 10 | 810 |

Both have the same accuracy: 90%

Model B has a better chance of telling if you have cancer

# Class imbalance: asymmetric metrics

- Your model's performance w.r.t to a class

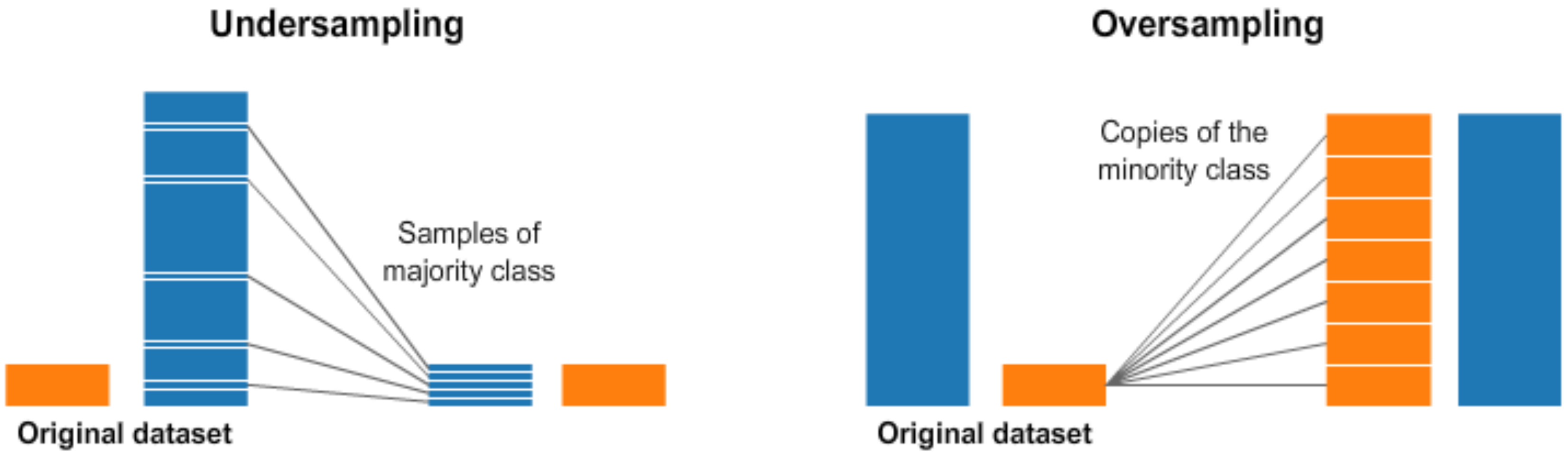
| | CANCER (1) | NORMAL (0) | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| **Model A** | 10/100 | 890/900 | 0.9 | 0.5 | 0.1 | 0.17 |
| **Model B** | 90/100 | 810/900 | 0.9 | 0.5 | 0.9 | 0.64 |

we can predict per class

⚠ F1 score for CANCER as 1 is different from F1 score for NORMAL as 1 ⚠

# 2. Data-level methods: Resampling

| **Undersampling** | **Oversampling** |
|---|---|
| Remove samples from the majority class | Add more examples to the minority class |



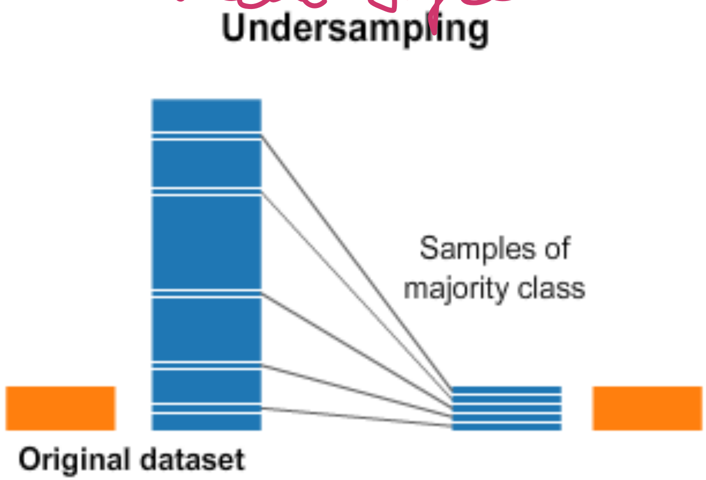



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1

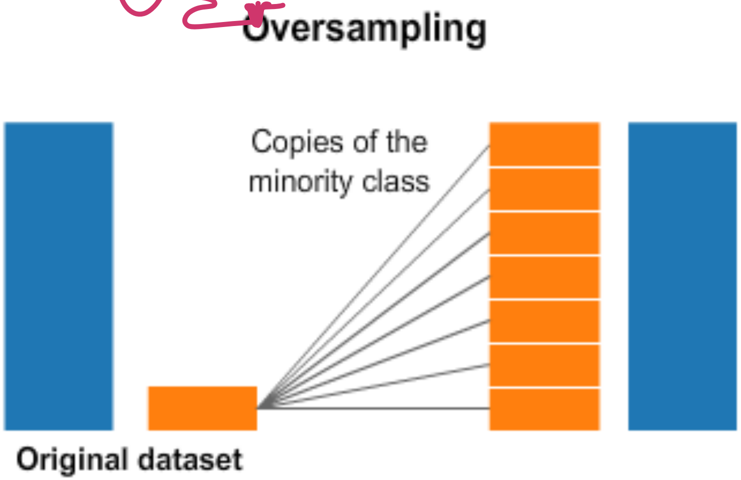
# 2. Data-level methods: Resampling

| **Undersampling** | **Oversampling** |
|---|---|
| Remove samples from the majority class | Add more examples to the minority class |
| Can cause overfitting | Can cause loss of information |

Lose Info




Of






https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1

# Undersampling: Tomek Links

- Find pairs of close samples of opposite classes
- Remove the sample of majority class in each pair
    - Pros: Make decision boundary more clear
    - Cons: Make model less robust

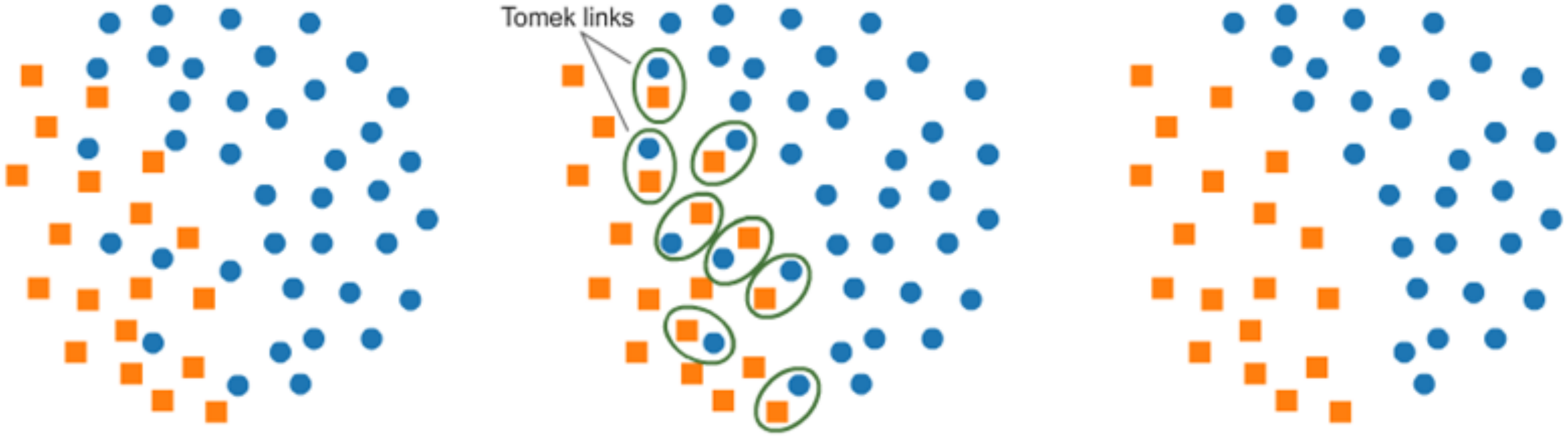



Image from https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

# Oversampling: SMOTE

Both SMOTE and Tomek links only work well on low-dimensional data!

- Synthesize samples of minority class as (usually) linear combinations of existing points and their nearest neighbors of same class.

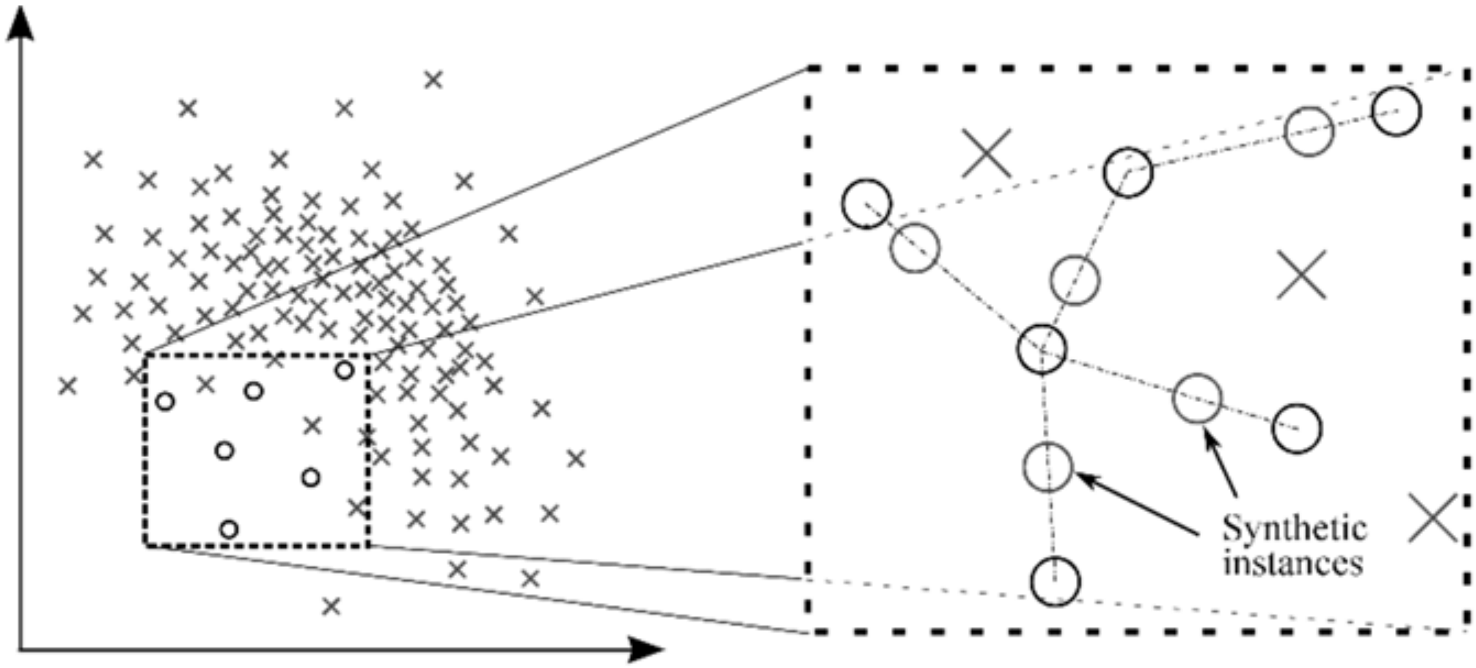



Image from [Analytics Vidhya]

# 3. Algorithm-level methods

- Naive loss: all samples contribute equally to the loss
- Idea: training samples we care about should contribute more to the loss

$$L(X;\ \theta) = \sum_{x} L(x\ ;\theta)$$

# 3. Algorithm-level methods

- Cost-sensitive learning
- Class-balanced loss
- Focal loss

# Cost-sensitive learning

- $C_{ij}$: the cost if class i is classified as class j

| | Actual NEGATIVE | Actual POSITIVE |
|---|---|---|
| Predicted NEGATIVE | $C(0,\ 0)\ =\ C_{00}$ | $C(1,\ 0)\ =\ C_{10}$ |
| Predicted POSITIVE | $C(0,\ 1)\ =\ C_{01}$ | $C(1,\ 1)\ =\ C_{11}$ |

- The loss caused by instance x of class i will become the weighted average of all possible classifications of instance x.

$$L(x\ ;\theta)\ =\ \sum_j C_{ij}P(j\,|\,x;\ \theta)$$

Cost acts as a factor in the loss Fnct

The mean we

The foundations of cost-sensitive learning (Elkan, IJCAI 2001)

# Class-balance loss

- Give more weight to rare classes

Non-weighted loss

$$L(X;\ \theta) = \sum_i L(x_i; \theta)$$

Weighted loss

$$L(X;\ \theta) = \sum_i W_{y_i} L(x_i; \theta)$$

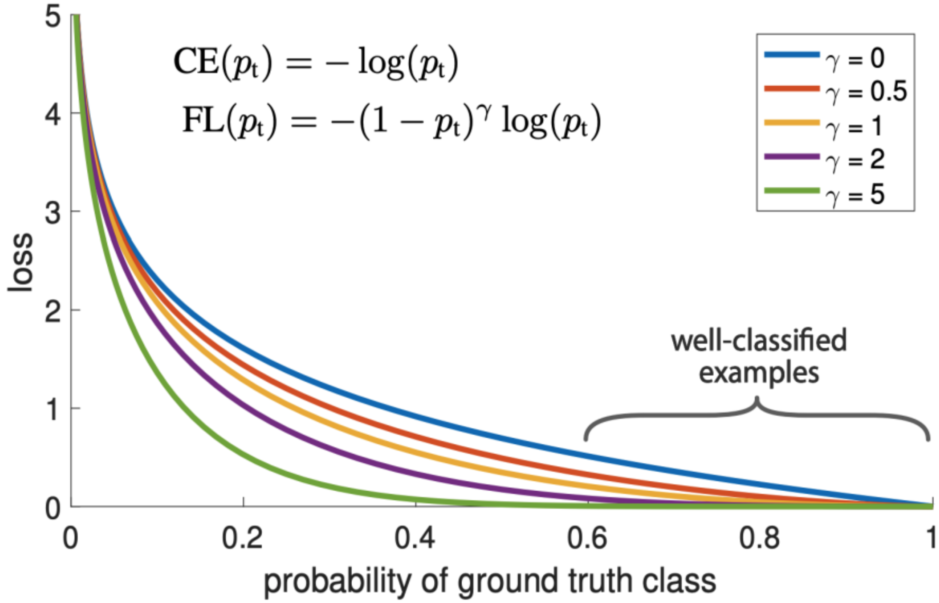
$$W_c = \frac{N}{\text{number of samples of class } C}$$

```
model.fit(features, labels, epochs=10, batch_size=32, class_weight={"fraud": 0.9, "normal": 0.1})
```

# Focal loss

- Give more weight to difficult samples:
  - downweighs well-classified samples

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases}$$






Focal Loss for Dense Object Detection (Lin et al., 2017)