UiO : Department of informatics
University of Oslo

Adaptive Methods for *Lecture* Data-based Decision Making *5*

IN-STK 5000 / 9000

Autumn 2022

Slides by Dr. Anne-Marie George, UiO

What we talk about today

Classification

Task: Find a function (*classifier*) to classify some input

Training data: Examples of the form (data point x, label/class y_x)

DISCRETE SET OF POSSIBLE LABLES

Example:

Task: Label pictures according to what they show!

Training Data: Picture + Label (Dog *or* Blueberry muffin)

→ Find a classifier that gives for any (new) picture a correct label.

Regression

Task: Find a function (*regressor*) to evaluate some input

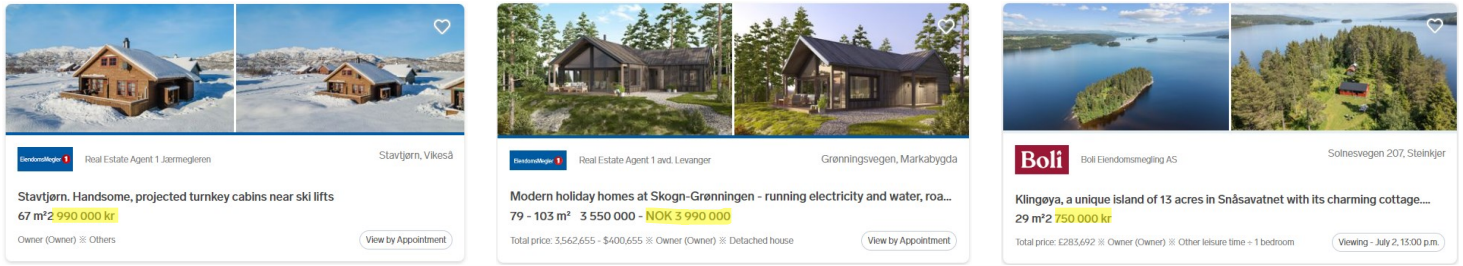
Training data: Examples of the form (data point x, label/value y_x)

CONTINUOUS SET OF POSSIBLE LABLES

Example:

Task: Predicting housing prices!

Training Data: House (m^2, post code, …) + Price



→ Find a regressor that gives for any (new) house a correct price.

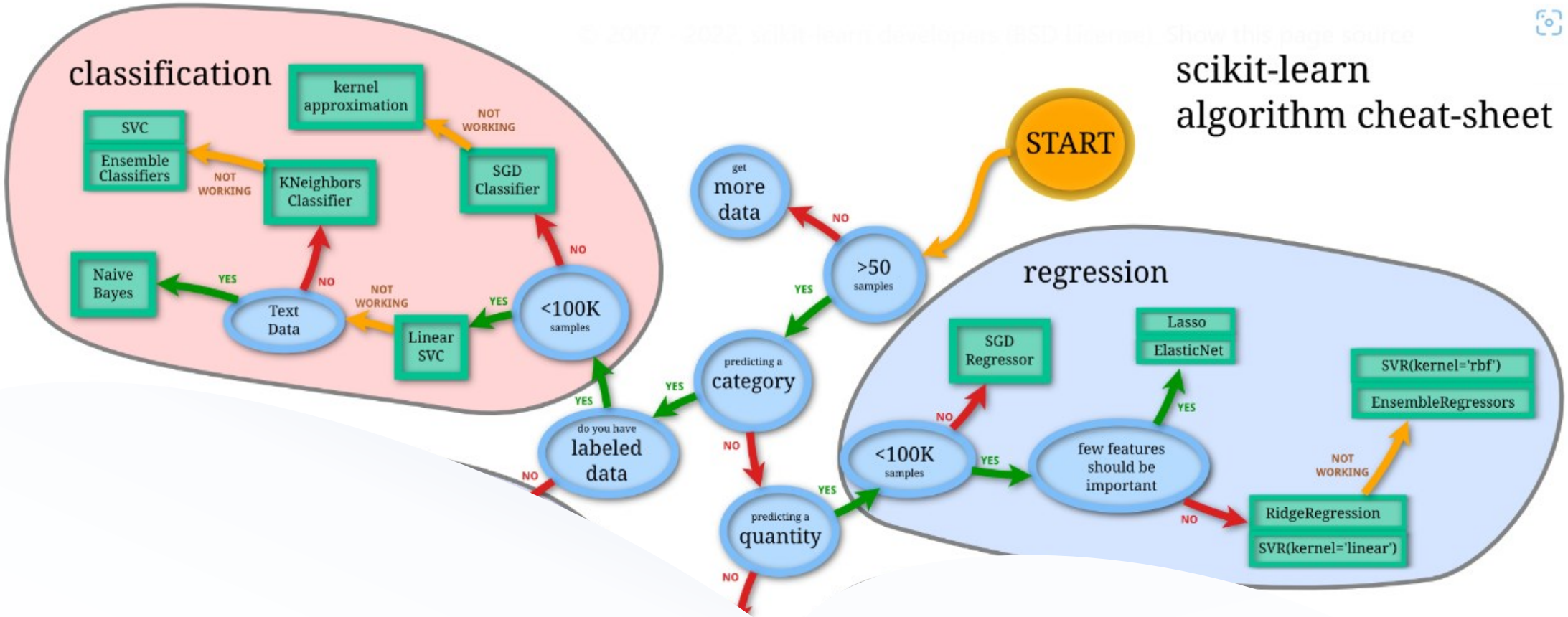
Classification & Regression

Decisions:

- What do we want to predict?
- What type of model do we need?
- What algorithm do we use?
- What kind of data do we need?
- How much data do we need?
- How do we measure the quality of the classifier/regressor?

Example: Estimate success chances of applicants to a study program!

→pass / fail, grades (A-F), …
→binary classifier, regression, …
→random forest, KNN, …
→age, gender, nationality, GPA, …
→Hundreds? Thousands? …?
→accuracy, precision, …?



Source: [Choosing the right estimator — scikit-learn 1.1.1 documentation]

Quality Measures for Classification

How do we measure quality of our predictions?

How to measure the quality of a classifier?

I build a classifier! ☺

→ It predicts the correct result in 91 % of the cases!

Is this a good classifier?

Depends:

- Is the dataset skewed?
- Are we doing better than, e.g., always predicting lable1 no matter the input?
- What errors do we have?
- Where do they occur?

True / False Positives / Negatives

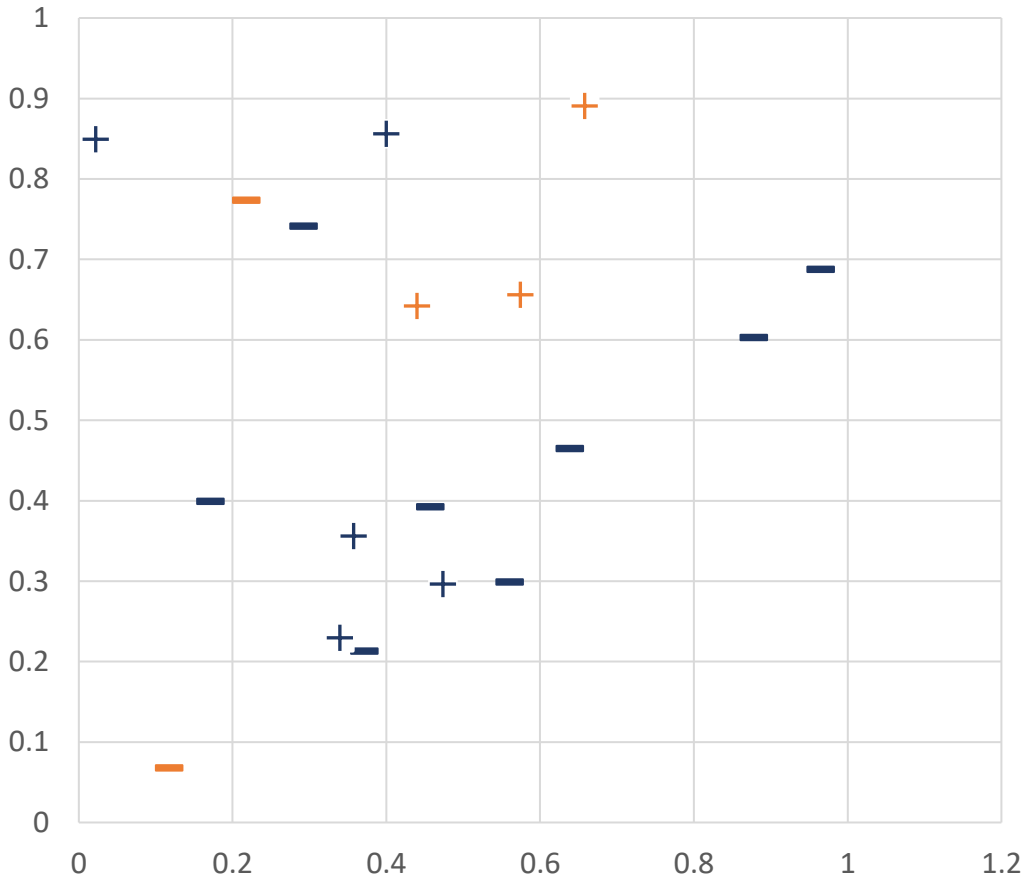
- Assumption:
  There are only two labels +/- *(or 1/0)*
  → *Binary* classification

	Actual Class		
Pre-dicted Class		+	-
	+	**True Positives (TP)**	**False Positives (FP)**
	-	**False Negatives (FN)**	**True Negatives (TN)**

True / False Positives / Negatives

- Assumption:
  There are only two labels +/- *(or 0/1)*
  → *Binary* classification

- True Positives (TP) +
  datapoints that are correctly classified as +
- False Positives (FP) +
  datapoints that are falsely classified as +
- True Negatives (TN) -
  datapoints that are correctly classified as -
- False Negatives (FN) -
  datapoints that are falsely classified as -

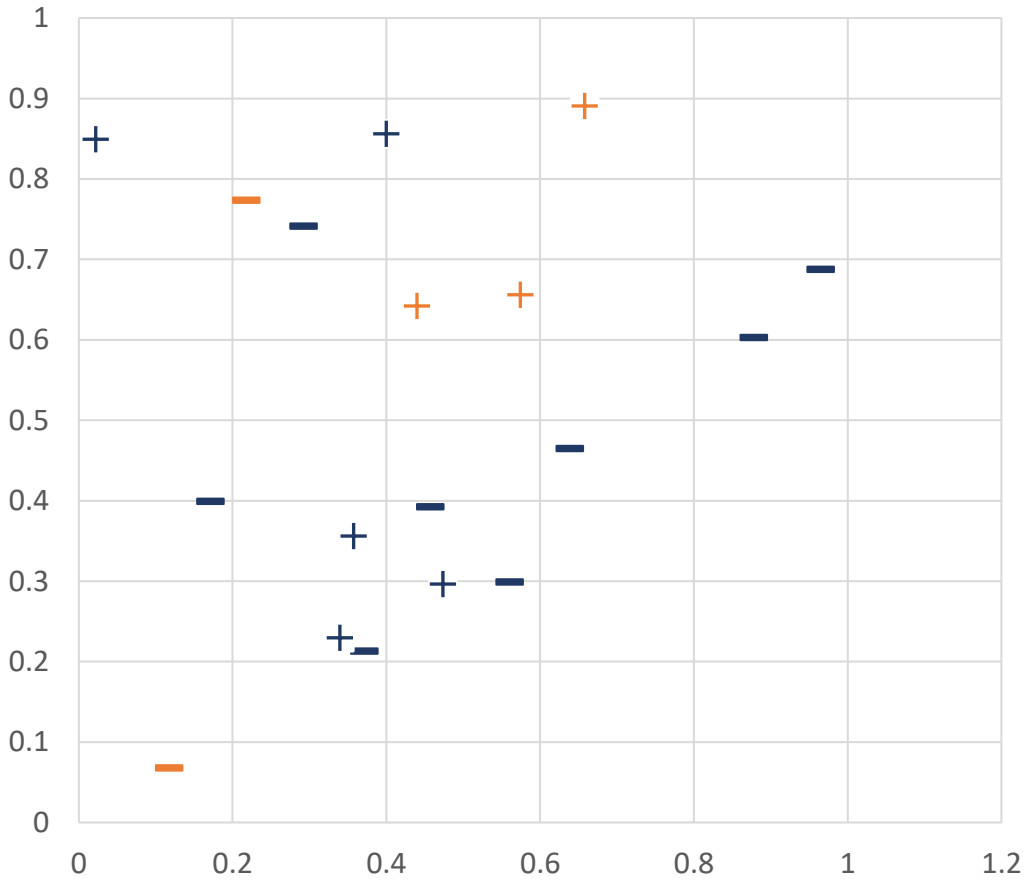


True / False Positives / Negatives

- Example:
  Predicting Covid:
  positive or negative outcome.

What is the meaning of … ?

- True Positives (TP) +
  datapoints that are correctly classified as +
- False Positives (FP) +
  datapoints that are falsely classified as +
- True Negatives (TN) -
  datapoints that are correctly classified as -
- False Negatives (FN) -
  datapoints that are falsely classified as -



Accuracy - for **binary** classification

- Accuracy: The fraction of correct predictions.

$$Accuracy = \frac{True\ Positives(TP) + True\ Negatives\ (TN)}{Total\ Number\ of\ Predictions}$$

$$= \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy - for **multi-class** classification

- Accuracy: The fraction of correct predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

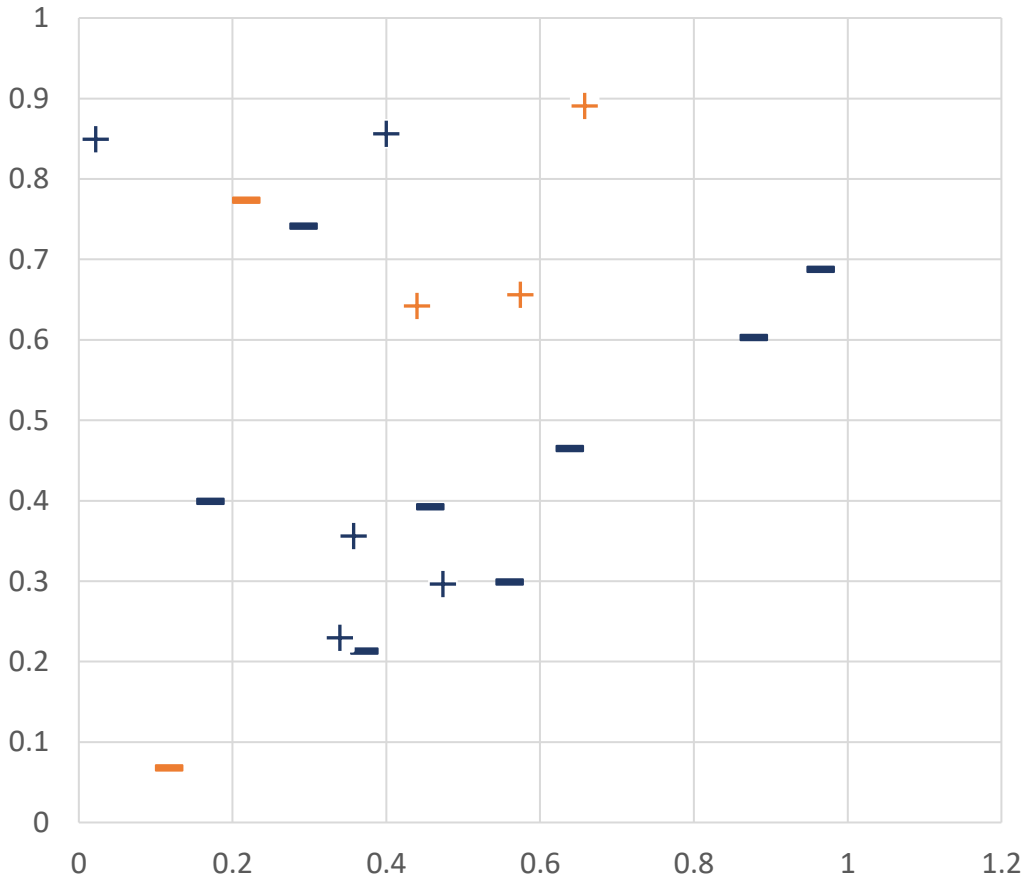
→ This *only* informs about the fraction of correct predictions, and *not* where the errors occur or what type(s) of error we have!

True / False Positives / Negatives

- Example:
  Predicting Covid:
  positive or negative outcome.

→ What do we care about the most?

- True Positives (TP) +
  datapoints that are correctly classified as +
- False Positives (FP) +
  datapoints that are falsely classified as +
- True Negatives (TN) -
  datapoints that are correctly classified as -
- False Negatives (FN) -
  datapoints that are falsely classified as -



How to measure the quality of a classifier?

I improved my (Covid) classifier! ☺

→ It has an accuracy of 0.95 (instead 0.91 from before)
→ It has almost the same number of False Negatives (FN)!

Is this a better classifier?

→If the data is skewed, e.g., almost no one has Covid, then simply predicting “no Covid” more often increases the accuracy and doesn’t change the number of FN much (in expectation)!

How can we detect this?

Precision & Recall

- Precision: The fraction of true positive vs. all positive predictions.
  (also: *positive predictive value*)

$$Precision = \frac{True\ Positives(TP)}{Number\ of\ Positive\ Predictions} = \frac{TP}{TP + FP}$$

- Recall: The fraction of true positive predictions vs. actual positives.
  (also: *true positive rate* or *sensitivity*)

$$Recall = \frac{True\ Positives(TP)}{Number\ of\ Actual\ Positives} = \frac{TP}{TP + FN}$$

What is better?
High or low values?

Tradeoff: Precision & Recall

Example: Predicting risk to reoffend: "no crime" = +, "reoffend" = -

Given probability $h(x)$ of "reoffend" for any criminal x:

- If $h(x) < \theta = 0.8$ → Predict "no crime", can be released.
- If $h(x) \geq \theta = 0.8$ → Predict "reoffend", must stay in prison.

Safety of public: Predict "no crime" only with high confidence!

→ FP low, FN high: ↗ $Precision = \frac{TP}{TP+FP}$ ↘ $Recall = \frac{TP}{TP+FN}$

Individual justice: Predict "reoffend" only with high confidence!

→ FP high, FN low: ↘ $Precision = \frac{TP}{TP+FP}$ ↗ $Recall = \frac{TP}{TP+FN}$

Other Measures
→ A good measure is application dependent!

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP)**, hit	**False negative (FN)**, type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP)**, type I error, false alarm, overestimation	**True negative (TN)**, correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Sources: [5][6][7][8][9][10][11][12] view · talk · edit

Source: Wikipedia

Expected Utility

- Let y_x be the true label for x, and $\mathrm{P}(y|x)$ the probability for label y our classifier predicts.
- Utility function: $U(y|\, y_x) \in \mathbb{R}$ [expresses how much we would value the prediction y if the true label is y_x]

What could be the goal now?

- Expected utility: $\mathbb{E}[U|y_x] = \sum_y U(y|\, y_x) \cdot \mathrm{P}(y|x)$
- Goal: Find a classifier that maximises $\sum_x \mathbb{E}[U|y_x]$
- Example: $$U(y|\, y_x) = \begin{cases} -100, & \text{if } \mathrm{y} \neq \mathrm{y_x} \text{ and } \mathrm{x} = (\text{age} > 60) \\ 0, & \text{otherwise} \end{cases}$$

→Min. Errors for over 60-year-olds

Quality Measures for Regression

How do we measure quality of our predictions?

Common Error Metrics

Target labels $y_1, \dots, y_n \in \mathbb{R}$, predicted labels $\widehat{y_1}, \dots, \widehat{y_n} \in \mathbb{R}$

- **Mean Absolute Error (MAE)**: $\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}| = \|y - \hat{y}\|_1$

  The average (absolute) error between predicted and actual label.

  Problematic: → Relative size dependent on the domain.
  → No differentiation between one huge error and many small ones.

- **Root Mean Squared Error (RMSE)**: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} = \|y - \hat{y}\|_2$

  Properties: → MAE $\leq$ RMSE $\leq$ $n^{\frac{1}{2}}$ MAE.
  → Punishes larger errors more than MAE.
  → High variance in errors: MAE $\ll$ RMSE → can indicate outliers

Let’s take a Quiz…

… go to Mentimeter!

Let’s take a break…

Back on in 5 min!

Source:
"The Ethical Algorithm"
(Chapter 2) by Kearns & Roth

Privacy

Features

Sensitive Features = Personal and not necessarily desired to be published in connection with identifiers

… or recoverable (**re-identification**)!

Examples: grades, ethnicity, health issues, sexual orientation, ...

Explicit Identifiers = Allow identifying a person without other infos

Examples: name, birth number, address, ...

Deleting Explicit Identifiers … is Not Enough!

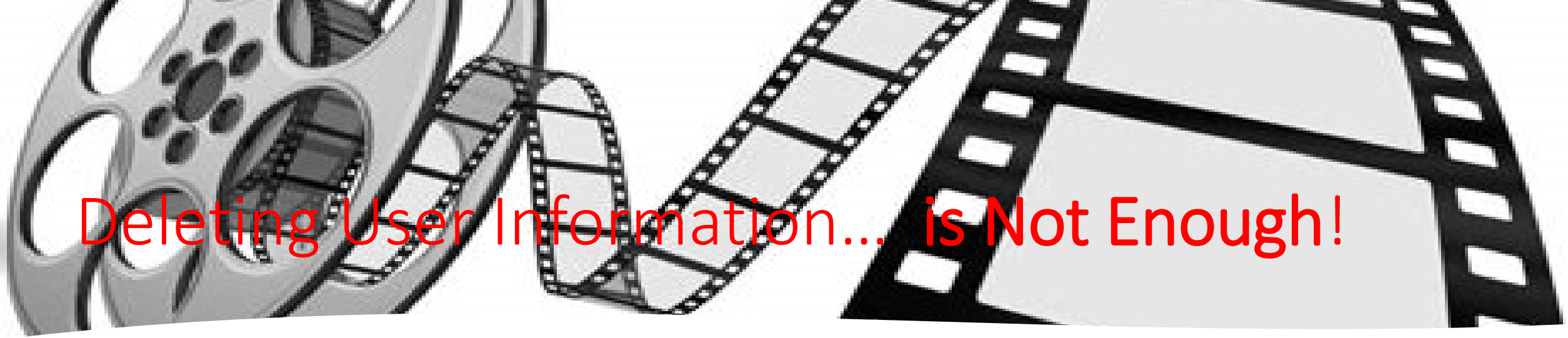
Group Insurance Commission (GIC)
Released data in mid-90's for academic research:

- Records of hospital visits of state employees in Massachusetts
- Included: Birthdates, ZIP code, sex
- Removed: Explicit patient identifiers (names, addresses, …)

Latanya Sweeney: Identified then Massachusetts governor!

- Purchased some voter data for 20$: Names, address, sex
- 6 people with same birth date, 3 of men, 1 correct ZIP code

… and showed: 87 % of Americans can be uniquely identified by birth date, ZIP code & sex!



Deleting User Information… is Not Enough!

The Netflix Price Competition (2006)
1 Mil $ Prize for 10% improvement of recommendation accuracy based on released data:

- > 100mil user ratings of movies (1-5), ~0.5mil users, 18k movies
- Included: Generic user ID, user ratings of movies, time of rating
- Removed: Any other user information (no demographics!)

Arvind Narayanan, Vitaly Shmatikov:
Could identify (a small sets of matching) user records of an individual (after only 2 weeks)!

- Even if: Only 6 ratings are known with only approximate dates (±14 days) → 99% of records uniquely identified
- This is very possible by cross-referencing with public data (e.g., IMDB) … on a large scale!

… Is this a privacy violation?!

→ Yes: Users did not publish *all* ratings.

K - Anonymity

- Idea: *Delete or bucket* data such that no record is unique!
- ***k*-anonymity**: For any query of features, e.g., (25<Age≤30, Male, Norwegian) there are at least k records with the same features.
- Example: Making data 2-anonymous

Name	Post Code	Age	Gender	Nationality	Condition
*	*	20<Age≤25	Female	English	Asthma
*	*	25<Age≤30	Female	Norwegian	Diabetes
*	*	25<Age≤30	Male	Norwegian	Chlamydia
*	*	20<Age≤25	Female	English	Diabetes
*	*	25<Age≤30	Female	Norwegian	Chlamydia
*	*	25<Age≤30	Male	Norwegian	HIV
*	*	25<Age≤30	Male	Norwegian	Chlamydia

K - Anonymity

- Idea: *Delete or bucket* data such that no record is unique!
- ***k*-anonymity**: For any query of features, e.g., (25<Age≤30, Male, Norwegian) there are at least k records with the same features.
- Problem: Can still find out *some* information.
  If Einar is (25<Age≤30, Male, Norwegian) → He has Chlamydia *or* HIV!

Name	Post Code	Age	Gender	Nationality	Condition
*	*	20<Age≤25	Female	English	Asthma
*	*	25<Age≤30	Female	Norwegian	Diabetes
*	*	25<Age≤30	Male	Norwegian	Chlamydia
*	*	20<Age≤25	Female	English	Diabetes
*	*	25<Age≤30	Female	Norwegian	Chlamydia
*	*	25<Age≤30	Male	Norwegian	HIV
*	*	25<Age≤30	Male	Norwegian	Chlamydia

K - Anonymity

- Idea: *Delete or bucket* data such that no record is unique!
- ***k*-anonymity**: For any query of features, e.g., (male, 24, Norwegian) there are at least k records with the same features.
- Problem: Multiple k-anonymous data might not be k-anonymous together!
  If Einar is (25<Age≤30, Male, Norwegian, Student) → He has Chlamydia!

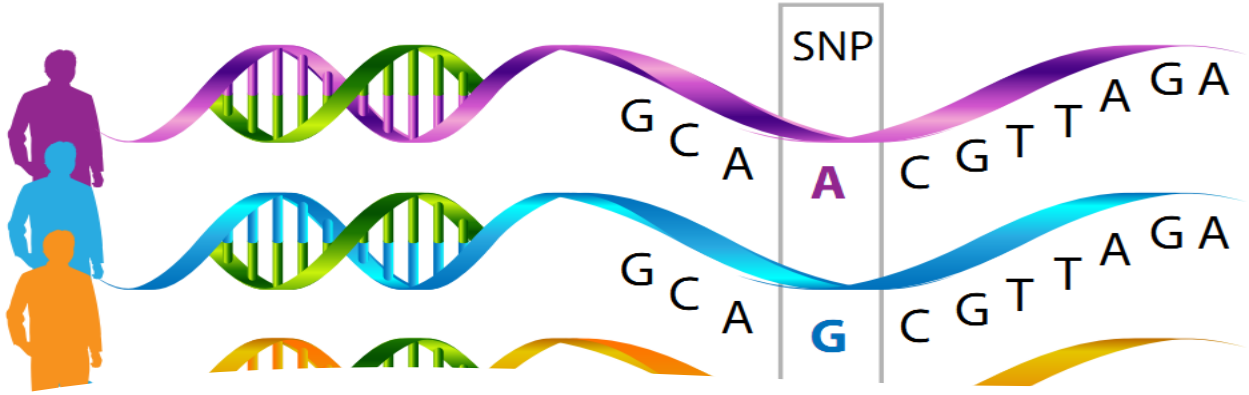
Age	Profession	Condition
20<Age≤25	Student	Asthma
25<Age≤30	Student	Diabetes
25<Age≤30	Student	Chlamydia
20<Age≤25	Student	Diabetes
25<Age≤30	Student	Chlamydia
25<Age≤30	Nurse	HIV
25<Age≤30	Nurse	Chlamydia

Age	Gender	Nationality	Condition
20<Age≤25	Female	English	Asthma
25<Age≤30	Female	Norwegian	Diabetes
25<Age≤30	Male	Norwegian	Chlamydia
20<Age≤25	Female	English	Diabetes
25<Age≤30	Female	Norwegian	Chlamydia
25<Age≤30	Male	Norwegian	HIV
25<Age≤30	Male	Norwegian	Chlamydia

Multiple Data Sources

- Making data anonymous is not enough
  → One might cross-reference with other data available
  → We cannot anticipate other / future data resources!
- What can we do?
  - Only publishing statistics?
  - Adding random noise?

Publishing Statistics



Genome Wide Association Study (GWAS):

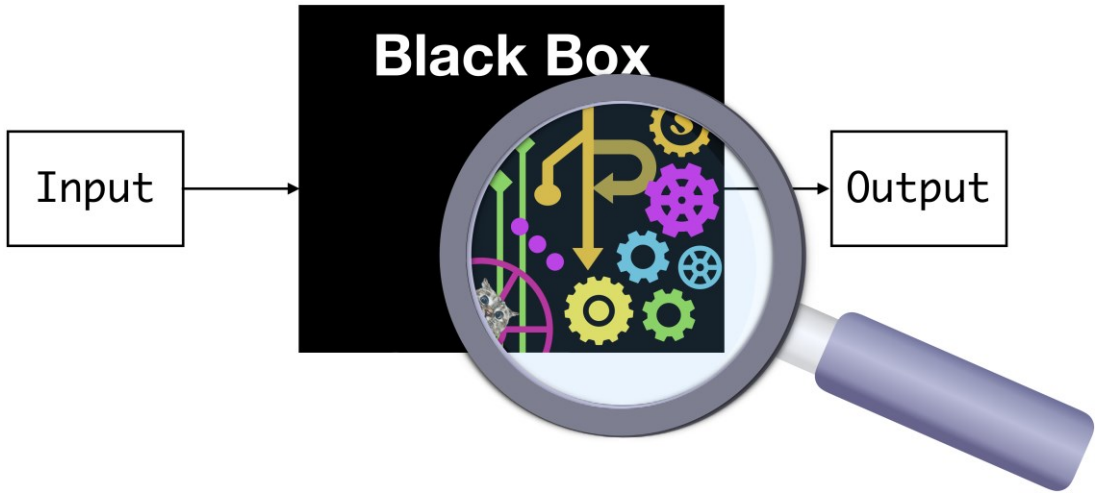
- Find correlations between genetic variances and prevalence of a disease
- Collect disease-patient's DNA data: genomes differ in only few positions
- Publish frequencies (of variances in genome positions) across population → Many such averages!

Homer, Szelinger et al. in PLOS Genetics (2008):

Can test whether a particular person participates in a GWAS!

→Privacy concern: GWAS participants have some disease!

→Such data no longer open access (limits for research!)



Publishing Trained Models

Given a trained ML model (input – output observations):

→Often possible to identify data points the model was trained on

→Training data often has better prediction than non-training data (overfitting)

Differential Privacy

R. Doll, A. B. Hill study: Smoking and lung cancer

- 2/3 of all registered physicians in the UK in 1951 participated
- 1956: Strong evidence that smocking increases risk of lung cancer
- Participant (Roger, Physician, smoker) at increased risk of lung cancer
  → Harmful information (increased insurance costs, …)
- But: This information would be harmful whether they participated or not!
  → No privacy violation!

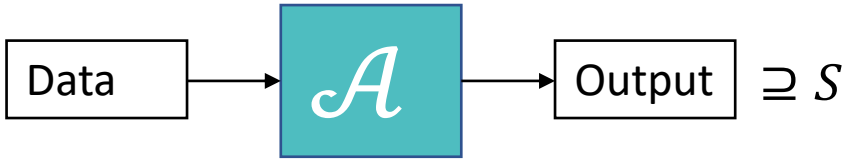
Differential Privacy (DP): A study is DP if what can be learned about a person (or the harm they get) is (almost) the same whether or not they are included.

ϵ- Differential Privacy

Cynthia Dwork et al.
for 2006 paper

- Randomized algorithm $\mathcal{A}$:

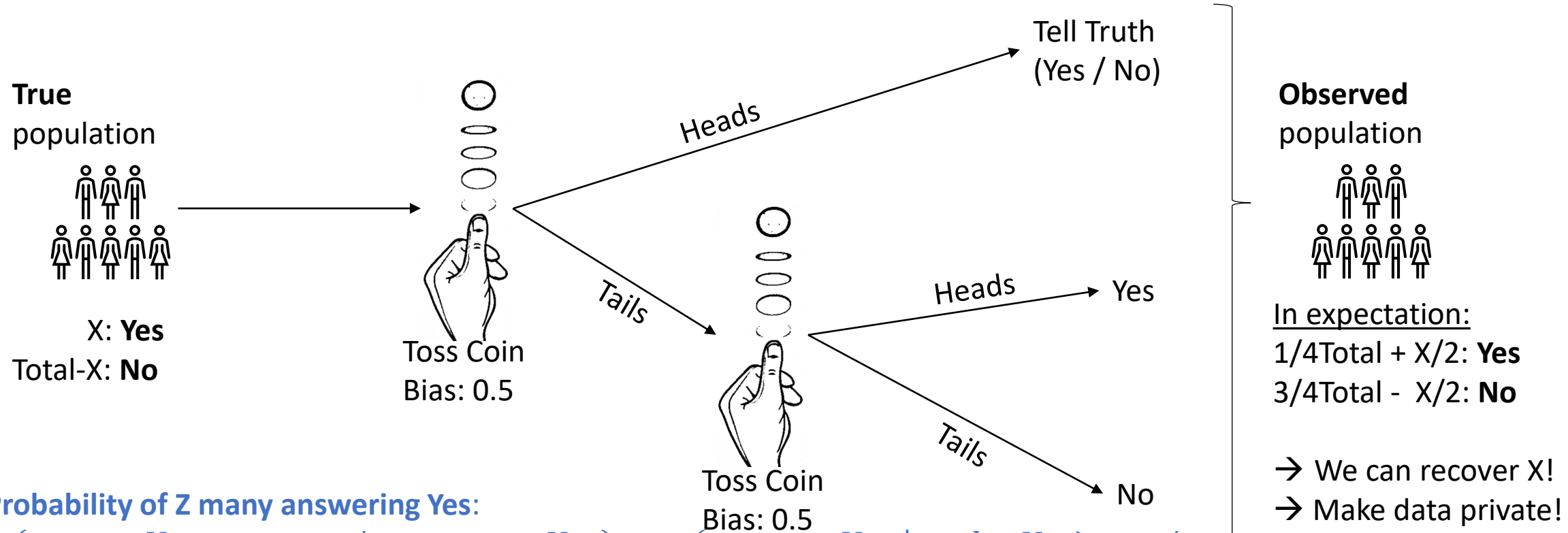
$$D_1 = (d_1, \dots, d_{i-1}, d_i, d_{i+1}, \dots, d_n)$$
$$\| \quad \dots \quad \| \quad \nparallel \quad \| \quad \dots \quad \|$$
$$D_2 = (d_1, \dots, d_{i-1}, \hat{d}_i, d_{i+1}, \dots, d_n)$$



→ $\mathcal{A}$ is ϵ-differential private if: $P[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \cdot P[\mathcal{A}(D_2) \in S]$

$$\Leftrightarrow \quad \frac{P[\mathcal{A}(D_1) \in S]}{P[\mathcal{A}(D_2) \in S]} \leq e^{\epsilon}$$

Randomized Response Mechanism



Probability of Z many answering Yes:

$$\frac{P(Z \text{ many Yes−answers} \mid X \quad \text{many Yes})}{P(Z \text{ many Yes−answers} \mid X-1 \text{ many Yes})} = \frac{P(\text{answer=Yes} \mid \text{truth= Yes})}{P(\text{answer=Yes} \mid \text{truth=No }\,)} = \frac{3/4}{1/4} = 3 = e^{\epsilon}$$

$\Longrightarrow \epsilon = \ln(3)$-differential privacy

Probability of Z many answering Yes:

$$\frac{P(Z \text{ many Yes−answers} \mid X \quad \text{many Yes})}{P(Z \text{ many Yes−answers} \mid X+1 \text{ many Yes})} = \frac{P(\text{answer=Yes} \mid \text{truth= Yes})}{P(\text{answer=No} \mid \text{truth=Yes})} = \frac{3/4}{1/4} = 3 = e^{\epsilon}$$

$\Longrightarrow \epsilon = \ln(3)$-differential privacy

Summary

Measuring Quality of Classification

Predicted Class	Actual Class	
	+	-
+	True Positives (TP)	False Positives (FP)
-	False Negatives (FN)	True Negatives (TN)

Accuracy:

$$\frac{TP + TN}{Total}$$

Precision:

$$\frac{TP}{TP + FP}$$

Recall:

$$\frac{TP}{TP + FN}$$

Measuring Quality of Regression

Mean Absolute Error (MAE):

$$\frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| = \|y - \hat{y}\|_1$$

The average (absolute) error between predicted and actual label.

Root Mean Squared Error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} = \|y - \hat{y}\|_2$$

MAE ≤ RMSE ≤ $n^{\frac{1}{2}}$ MAE.

Privacy

Deleting identifiers:

- Not obvious how much / what to delete
- Cross-referencing with other datasets can reveal identities

k-anonymity:

Every row same features as k others.

- Cross-referencing with other datasets can reveal identities

ϵ-differential privacy:

Weather one person is included or not (or changes response or not) makes little difference to what we can learn.

→ Randomized response mechanism: Make data $\ln(3)$-diff. private by adding noise with help of coin flips!

Which measure fits best? – Application dependent!
To grasp the quality, look at several!