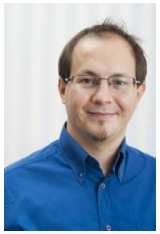
VU Security, Privacy & Explainability in Machine Learning

Summer term 2025



Rudolf Mayer (E194)

Andreas Rauber (E194)

Daniela Martinez (E194)

Guest Talks by:

Daryna Oliynyk

Anastasia Pustozerova

Topics

- Machine learning on personal/sensitive data
  - How to share data w/o violating privacy/data protection laws
  - How to prevent information leak from trained models
  - Privacy-preserving computation

- Security of machine learning
  - Adversarial input/examples
  - Backdoor attacks
  - Model stealing attacks
  - Model inversion
  - ….

“panda”
57.7% confidence

- Explainability of machine learning models
  - How to interpret and understand models & decisions

Topics

- **Not** part of the lecture
  - Machine Learning for cyber security
  - E.g. Intrusion detection systems, anomaly detection, ...

Target group & prerequisites

- We expect a solid understanding of machine learning principles
  - E.g. Lecture 184.702 Machine Learning, or equivalent
    - *We will not repeat basic concepts in this lecture*
  - Useful (but not required!):
    - Experience with visual computation / image classification (for adversarial inputs, backdoors, ...)
    - We will briefly explain advanced principles where required (mostly deep learning)

Lecture in the Curricula

- This lecture is part of the following curricula:
  - Data Science (066 645)
    - Modul MLS/EX - Machine Learning and Statistics – Extension
  - Business Informatics (066 926)
    - Modul DA/EXT - Data Analytics – Extension
  - Software Engineering & Internet Computing (066 937)
    - Modul Algorithmik
  - Other CS curricula: “Catalogue Elective Courses - Computer Science
    - e.g. Visual Computing, Medical Informatics, Logic & Computation…
  - Automation and Robotic Systems (066 515)

  ~~– Logic and Computation (066 931)~~
    - ~~Modul Knowledge Representation & Artificial Intelligence~~

Dates

- Lectures on Monday, 10:00-12:00
  - Always the same day & time
  - ***But not every week:*** Updated lecture schedule available in TUWEL

03.3.	Preliminary talk, Introduction
10.3.	Privacy-preserving data publishing (anonymisation & co)
17.3.	Differential Privacy, Synthetic data, ...
24.3.	Federated Learning, Secure computation (SMPC, homomorphic encryption, ...)
31.3.	Adversarial ML: Adversarial examples
07.4.	Adversarial ML: Backdoor attacks
28.4.	Explainability
05.5.	Explainability
19.5.	Adversarial ML: Confidentiality attacks (membership inference, model inversion,), IPR protection …
26.5., 2.&9.6	Backup Slots

Course support

- Make sure to register in TISS (http://tiss.tuwien.ac.at)
- Register for the course in TUWEL to access material
  - https://tuwel.tuwien.ac.at/course/view.php?idnumber=194055-2025S
  - linked from TISS (automatic access when applied to lecture in TISS)
- TUWEL used for
  - Course materials (lecture slides, further readings)
  - Lab assignments
  - Forum
  - TISS forum etc. will **NOT** be used !!

Evaluation

- Written exam
  - End of the semester
  - 1 retake date in WS 2025/2026

- Exercises
  - One smaller exercise (available after lecture in two weeks)
  - Two practical exercise on selected lecture topics: security/privacy
  - Group work (2 students)
  - In-class presentation of results (or submission of draft concept)
  - Late hand-in (max one week!) possible
    - 20%-points penalty deduction
    - (i.e., you can only reach 80% of the points awarded)
- Each assessment (exam, 3 assignments): min 40%
- Overall: min 50% for a positive grade

Overview on the lecture topics

- Privacy preserving data publishing
- Secure computation
- Adversarial examples
- Backdoor attacks
- Inference Attacks (Membership, Inversion, …)
- Model Stealing
- Explainable AI

Data Privacy

- Large amounts of personal data becomes available
  - Analysis, distribution, sharing often conflicting with data protection laws (GDPR, ...)

  - Especially critical with highly sensitive information
    - E.g. health data, financial data, ..

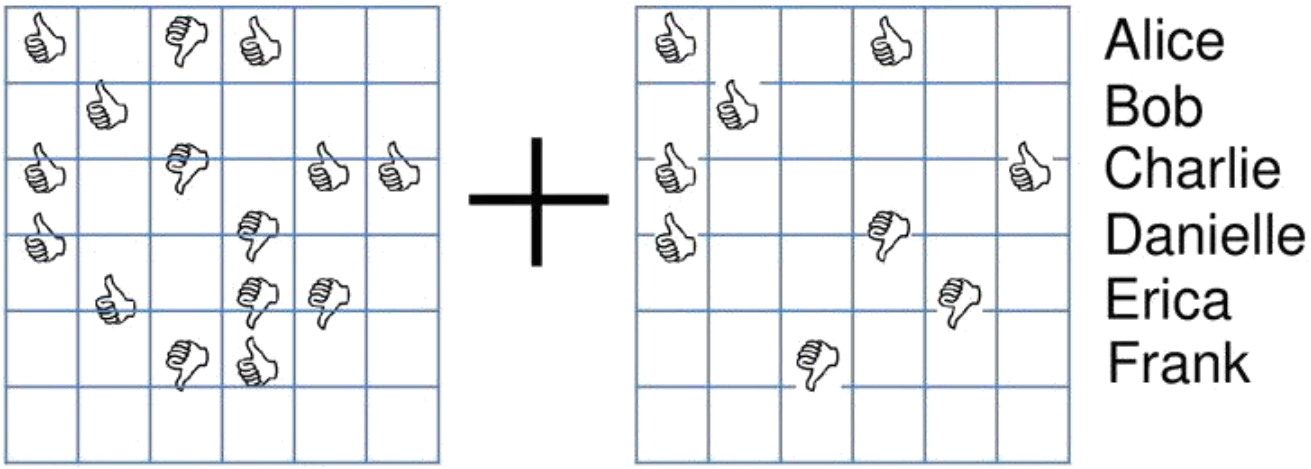
- *Solutions?*
  - E.g. Data sanitisation, privacy-preserving computation

Data Privacy

- Data Sanitisation: identifying attributes removed before we can use it
  - **Pseudonymisation**: remove directly identifying information
    - *Examples?*
    - *Is that enough?*
  - Massachusetts Health records of public employees
    - With the birthdate, ZIP Code, gender: Governour of Massachusetts William Weld uniquely identified
    - *How?*
    - Linkage attack with public voting records
    - Other examples:
      Netflix price (2006),
      AOL data, ...

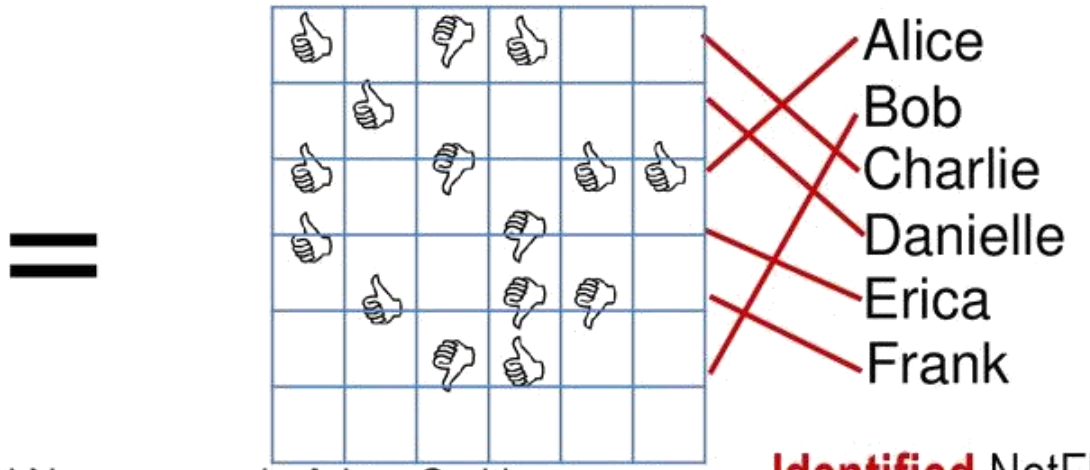
IMDb

Data Privacy



Anonymized NetFlix data

Public, incomplete **IMDB** data



Credit: Arvind Narayanan via Adam Smith

Identified NetFlix Data

Data Privacy

of **mobile phone owners** are re-identified simply by 2 antenna signals, even when coarsened to the hour of the day

of **credit card owners** are re-identified by 3 transactions, even when only merchant and the date of transaction is revealed

of **all people** are re-identified, merely by their date-of-birth, their gender and their ZIP code of residence

k-Anonymity

- Ensures that at least k records have same QI, via
  - Generalisation of values
    - Dates: reduce exact day ➔ month, quarter, year, decade, ..
    - ZIP Code: 1010, 1020 ➔ 10xx
  - Suppression of values

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	5	15	
2	15	25	
3	28	28	
4	25	15	
5	22	28	
6	32	35	
7	38	32	
8	35	25	Diarrhea

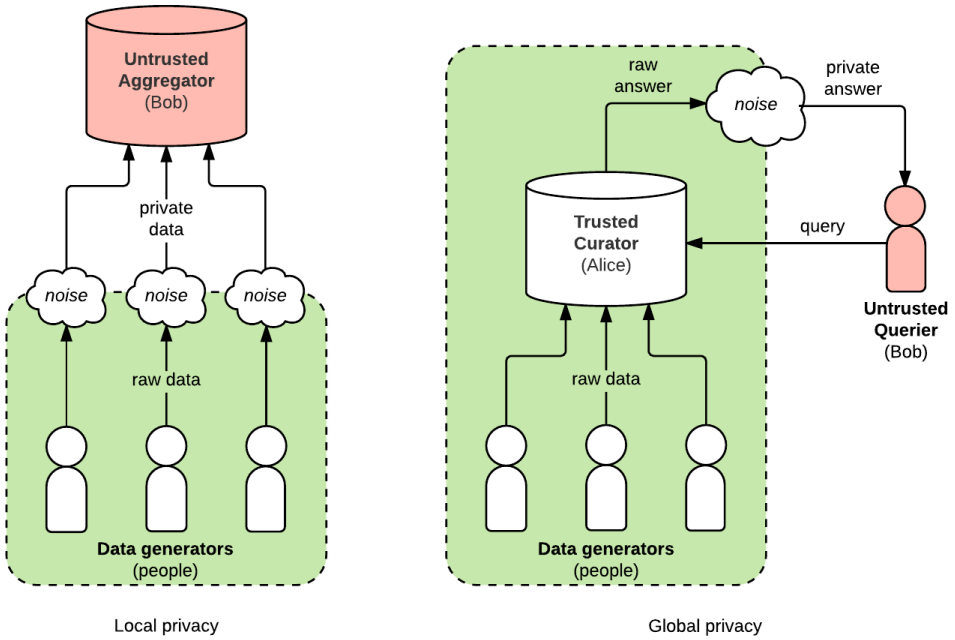
Birth day	gender
21/1/79	M
10/1/79	F
1/10/44	F
21/2/83	M
19/4/82	M

ender	Zipcode
uman	5****
uman	5****
	90210
M	022**
M	022**

  - Different paths to achieve the same level of *k*

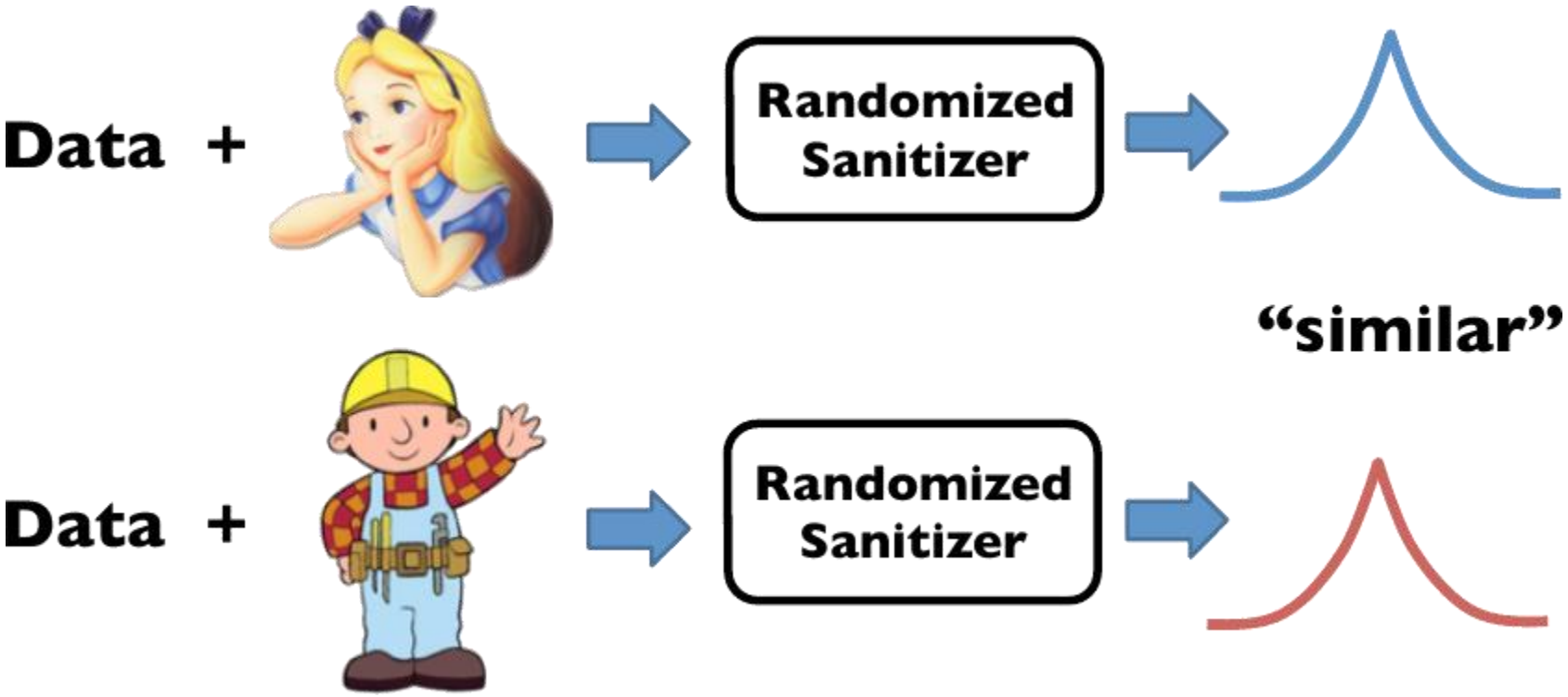
Differential privacy

- Related approach, but based on different concept
  - Add noise to data / output of algorithm
  - Noise should not influence overall statistics



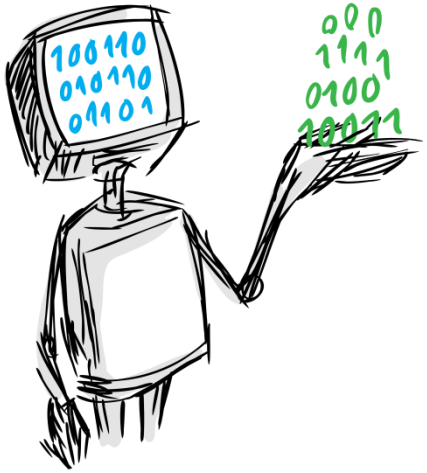
  - Local vs. global
    - Local used now e.g. by Apple, Google, etc..

Differential Privacy: Setting



Synthetic Data

- Instead of trying to make real data not re-identifiable:
  - **Generate** a dataset with synthetic samples
  - that resemble the characteristics as well as diversity of the original data

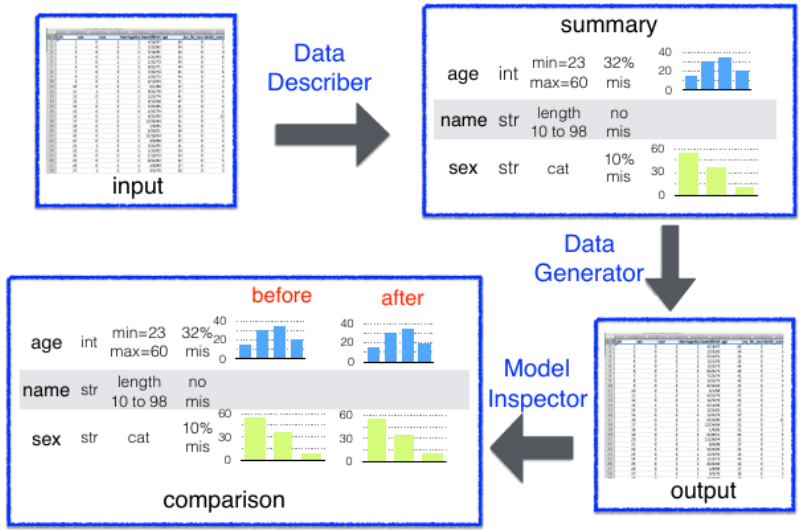


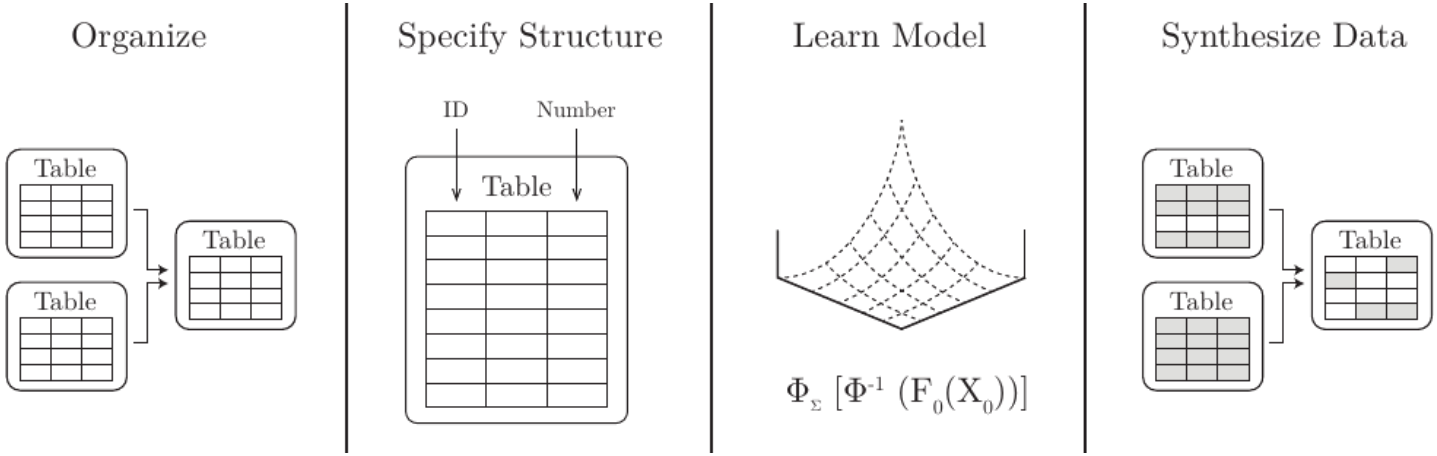
- Sometimes called “artificial data”

- Not the same as “test data” in SW development
  - As it aims at preserving the statistical properties

Synthetic Data

- Learn a model from the original data
- Use that model to generate synthetic data



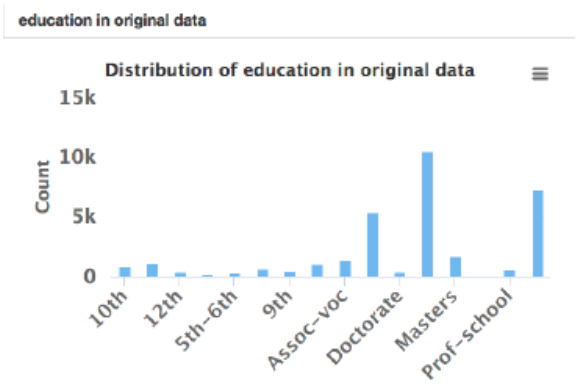


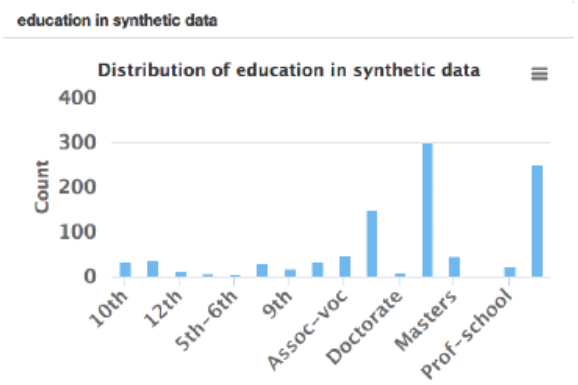
Synthetic Data

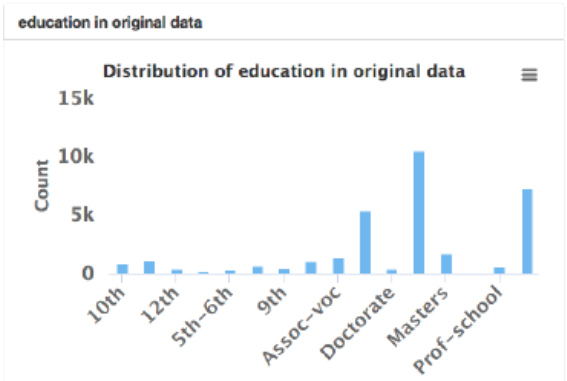
ID	Birthdate	Gender	Balance
1	07/26/94	F	427.30
2	03/16/92		1001.42
3	01/17/87	M	33.33
4	08/16/94	M	54.33
5	09/04/82	M	5000.40
....	....	....	...

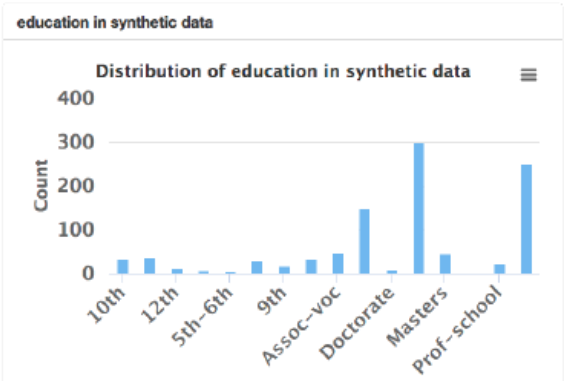
model → SDV → synthesize

ID	Birthdate	Gender	Balance
1	11/26/88	M	4002.50
2	06/14/91	F	352.47
3	04/15/90	F	129.75
4	02/17/95		37.20
5	03/29/84	M	427.92
....	....	....	....





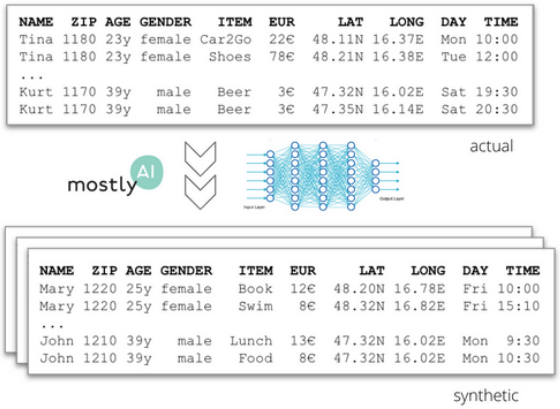




https://github.com/DataResponsibly/DataSynthesizer

Synthetic Data: Tools

- The **Synthetic Data Vault** (SDV) (Neha Patki et al.)
  - https://sdv.dev/
- **Synthpop**
  - Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control
  - https://cran.r-project.org/web/packages/synthpop/
- **DataSynthesizer**
  - https://github.com/DataResponsibly/DataSynthesizer
- **Mostly.ai** (commercial)



Data fingerprinting and watermarking

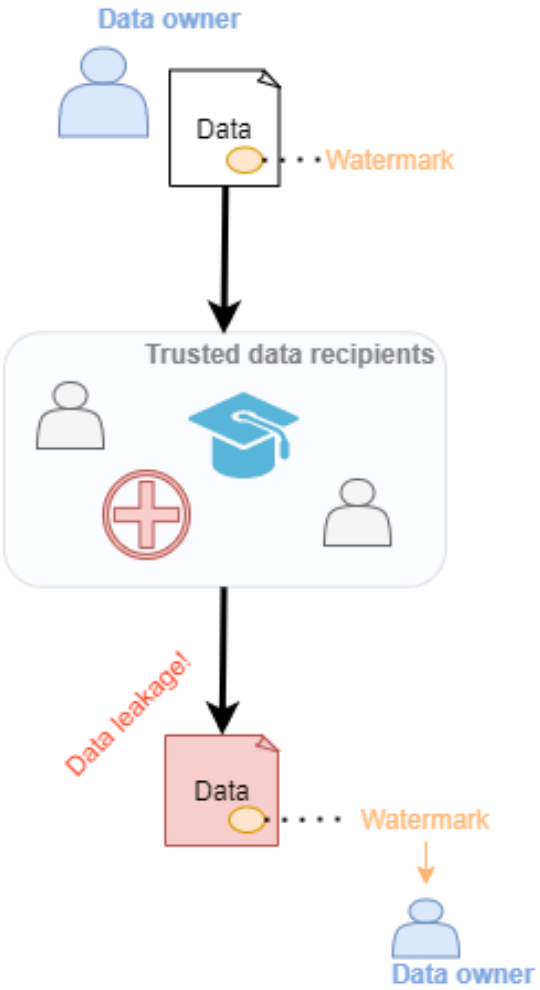
- Embedding owner‘s signature into the data
  - Applying tailored modifications to the data which only the owner is able to extract

Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0

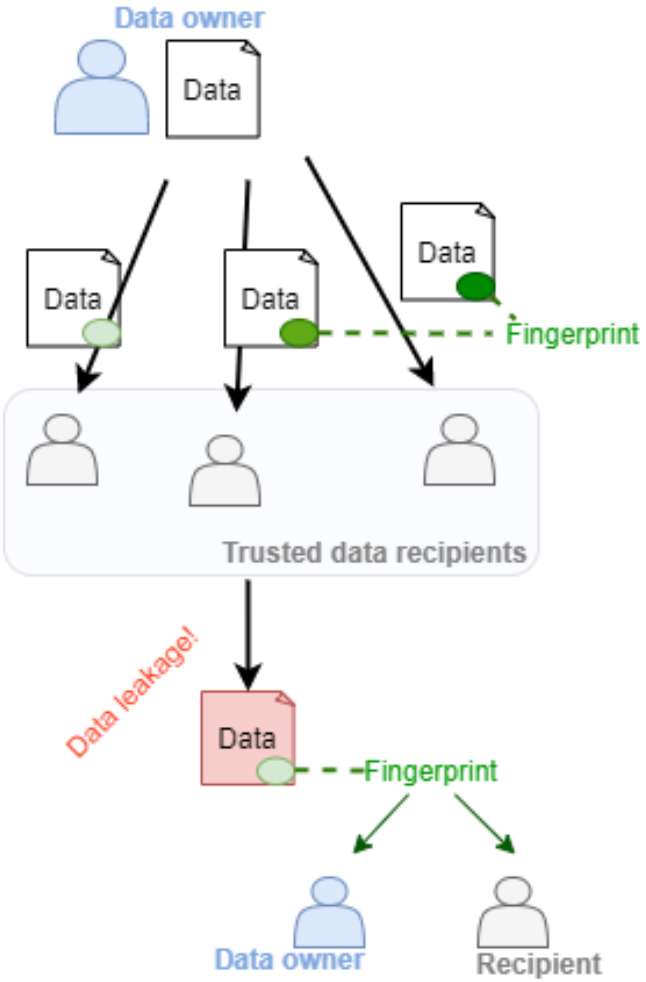
Age	Blood Pressure	Diabetes
33	64	1
31	68	0
50	72	1
47	70	0

Watermarking vs. fingerprinting

Watermark: identifies the owner

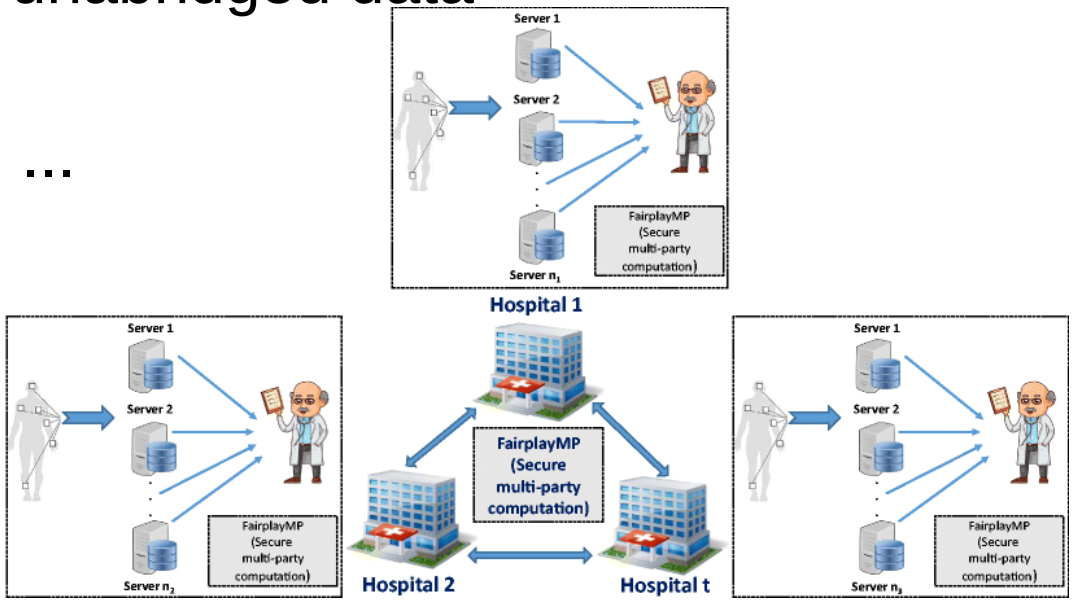


Fingerprint: owner & recipient

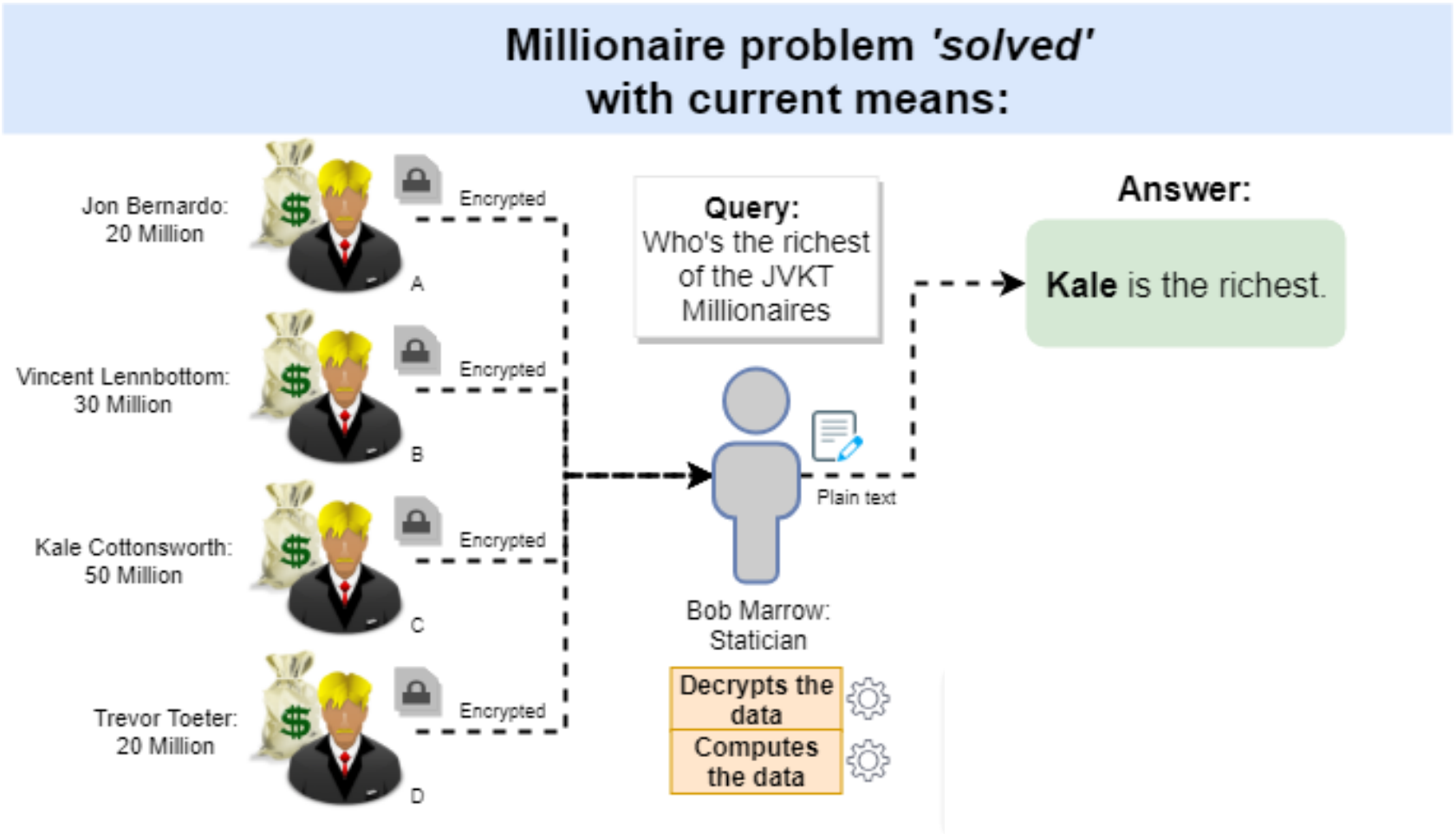


Secure Computation

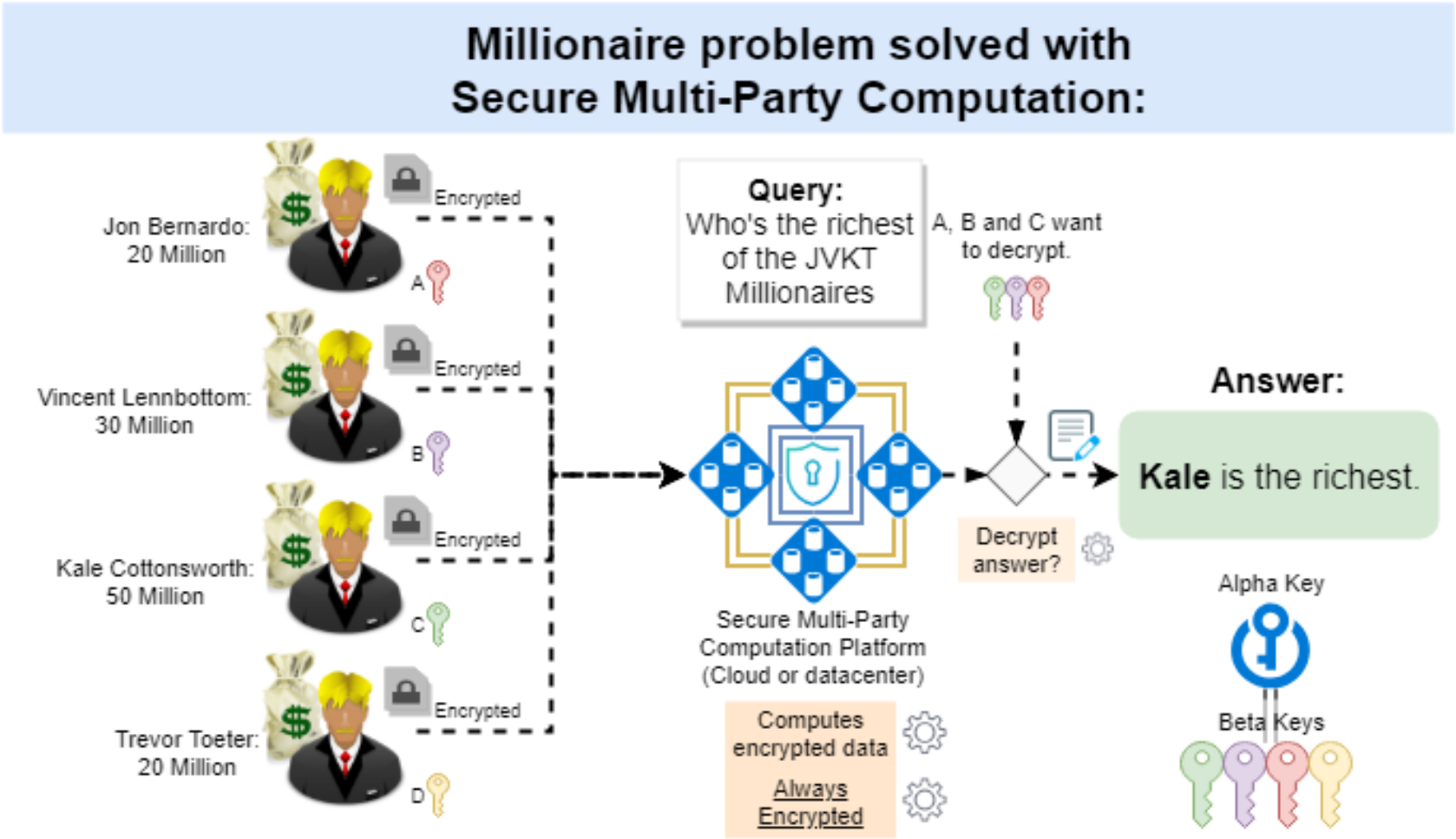
- Privacy-preserving computation
  - Data stays with data owner
    - One, or multiple owners (e.g. different health care providers)
  - Data is not aggregated by a central server, and then analysed
  - Analyst can work on unabridged data
  - Retrieves only final models, results, ...



Secure Computation

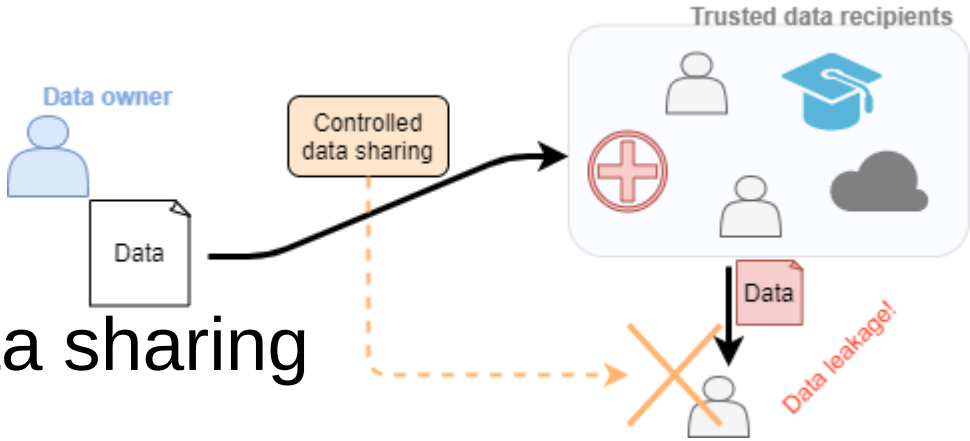


Secure Computation



Digital property protection

- Why protecting the data?
  - Data owner used a lot of resources to collect/create the data (money, human experts, time…)
  - Sensitive data (e.g. medical data) needs to be shared with researchers
    - Privacy implications: only the trusted parties get the data and should not share it further



- The goal: controlled data sharing
  - Share full data
  - Trace unauthorised data re-distribution

Federated Machine Learning

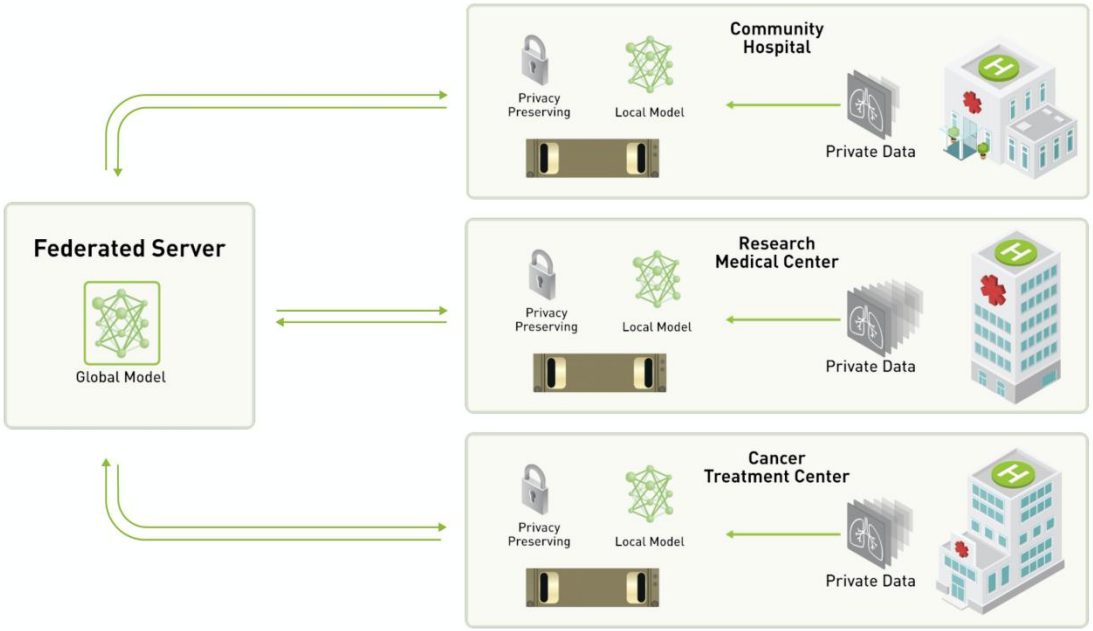
- Don’t centrally aggregate data
  - But aggregate final models or intermediate versions!
- The model shouldn’t contain any personal/sensitive information anymore

- Aspects of
  - Effectiveness (quality loss due to federation?)
  - Efficiency (due to communication overhead)
  - Level of federation
    - Final model?
    - Intermediate steps?

Federated Learning

Data is already collected distributed ➔
Aggregating machine learning **models** instead of aggregating (centralising) data

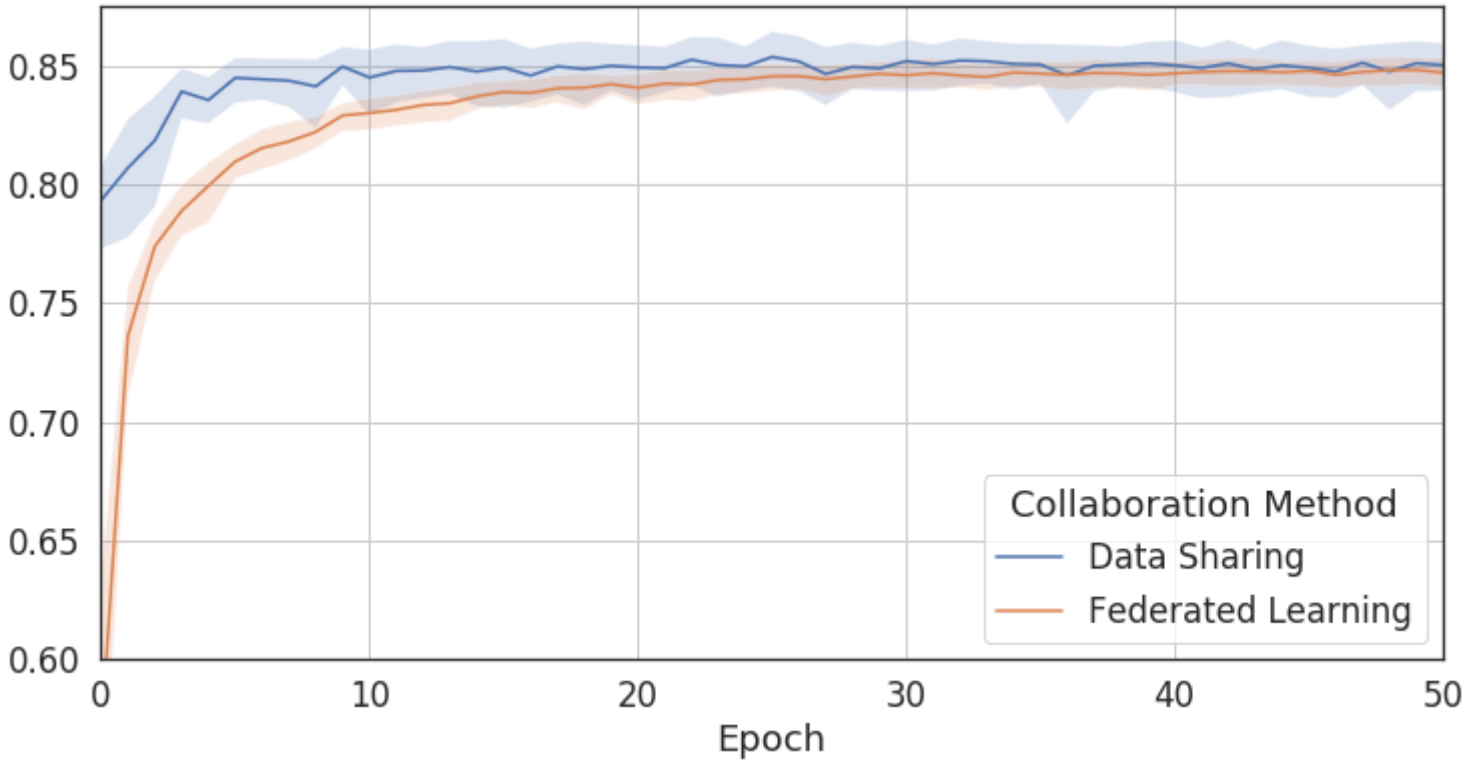
1. Each party (node, client) trains its local model on their private data
2. Each node shares a local model
3. All locally trained models are aggregated into a global model
4. Potentially repated in multiple cycles



https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/

Federated Machine Learning

- Evaluation of federated vs centralised



Tool Support

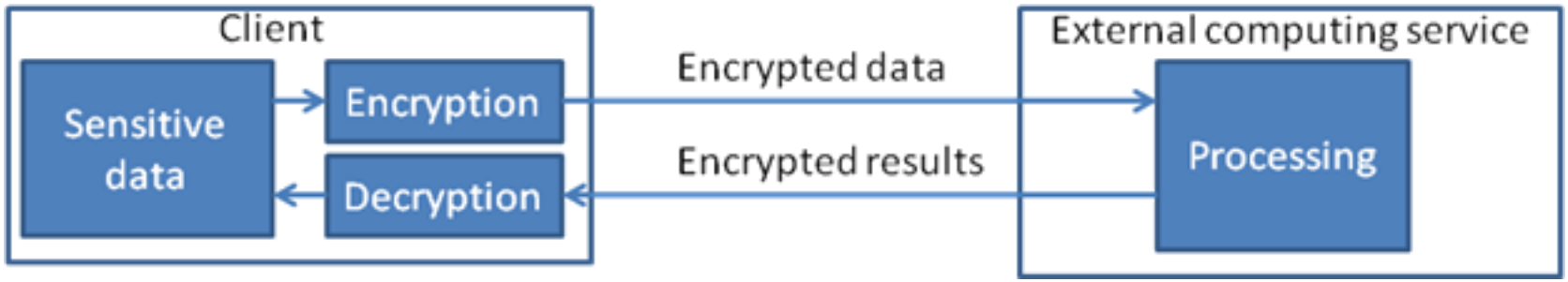
- TensorFlowFederated: https://www.tensorflow.org/federated/
- PySyft: https://github.com/OpenMined/PySyft
- FedAI / FATE: https://fate.fedai.org/
- Flower: https://flower.dev/
- IBM Federated Learning: https://github.com/IBM/federated-learning-lib
- INTEL OpenFL: https://github.com/securefederatedai/openfl

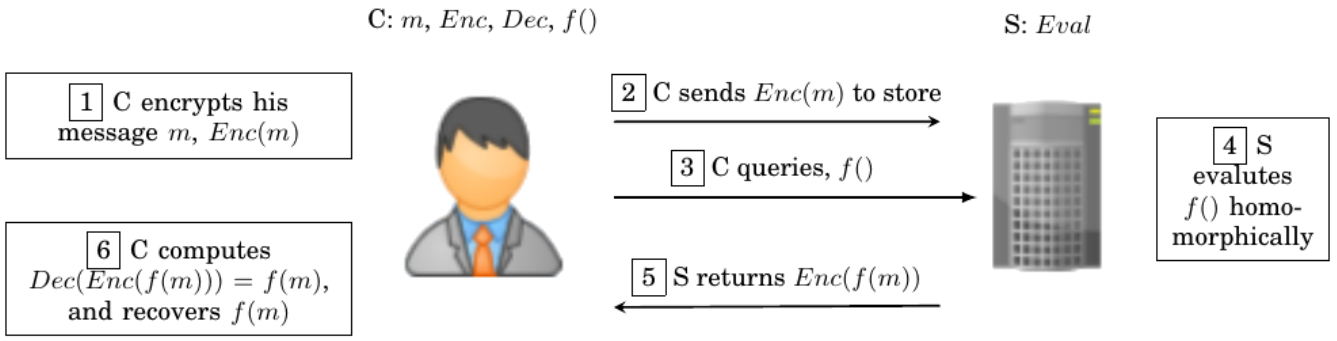
- Microsoft Flute (simulation only): https://github.com/microsoft/msrflute

- Defunct:Decentralized ML: https://decentralizedml.com
  - (Commercial, market place, using also crypto currencies)

Homomorphic Encryption

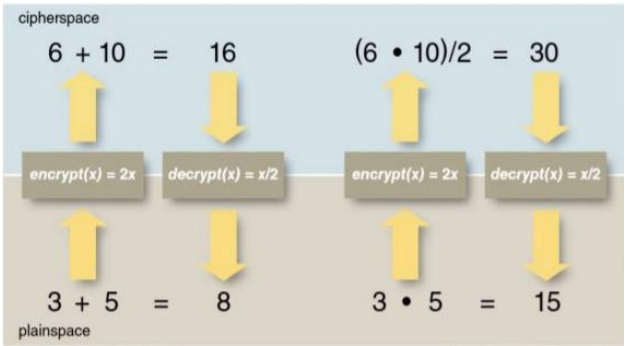
- A type of encryption where certain computational operations are possible on the ciphertext
- We can thus let an untrusted third party do our computation, w/o revealing the data to them





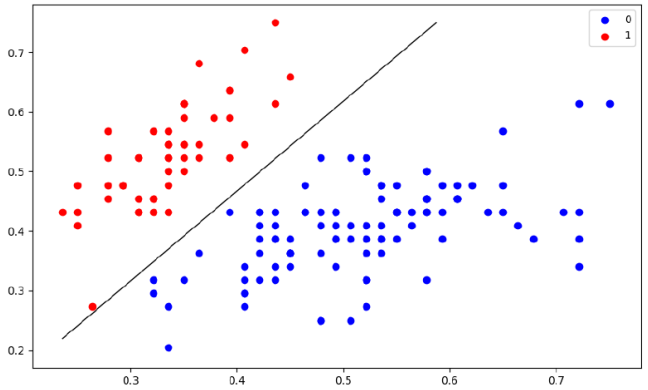
Homomorphic Encryption

- Simplistic examples:
  1. Caesar substitution encryption, e.g. Shift numbers by 2
     - Input: [6, 7, 9, 2, 0, 2, 1, 8, 8, 6]
     - Ciphertext: [8, 9, 11,4, 2, 4, 3,10, 10, 8]
     - Task: compute sum
       - Sum(Input)= 49
       - Sum(Cipher) = 69
       - Decrypt(Cipher) = 69-(10*2) = 49
  2. x2, /2



Homomorphic Encryption

- Computational costs several orders of magnitude higher
  - Especially for fully homomorphic schemes



  - Caesar substitution: partially homomorphic, for addition and subtraction
  - Other partially homomorphic schemes:
    - Paillier, ElGamal, …
  - Partial schemes are faster than fully homomorphic schemes
    - Limitations on the algorithms, maybe need to approximate

Homomorphic Encryption

- Tool support
  - https://github.com/NervanaSystems/he-transformer
  - https://github.com/Microsoft/SEAL
  - https://github.com/shaih/HElib
  - https://github.com/tfhe/tfhe
  - And a list at https://github.com/jonaschn/awesome-he

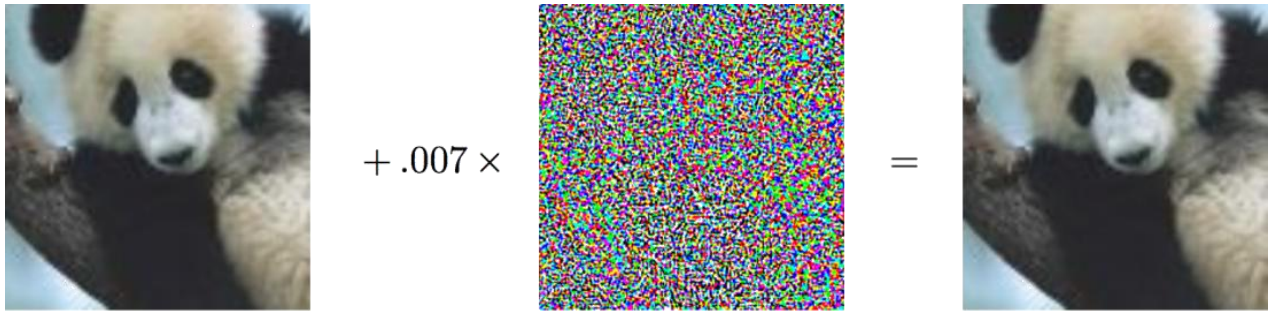
Adversarial Examples

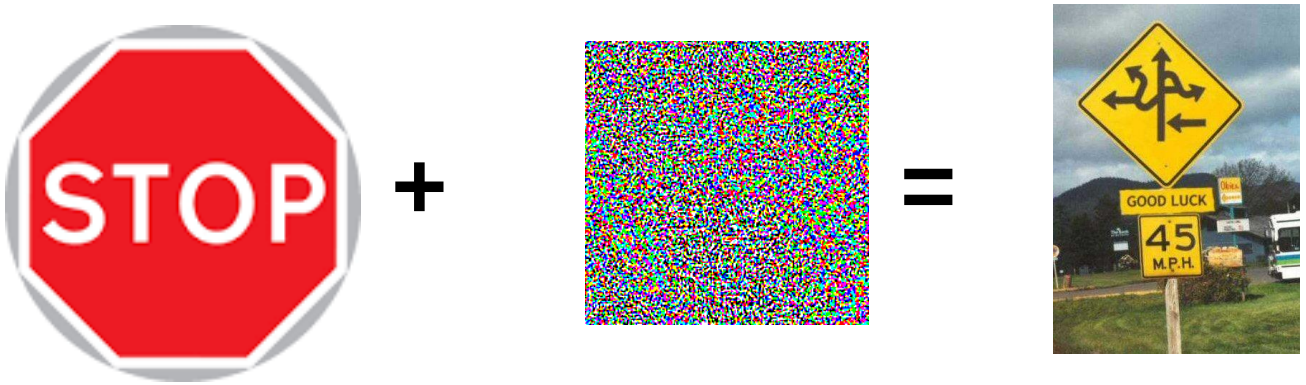
- Mostly made prominent as an attack for image analysis / computer vision
  - But also possible for other modalities (just not so graphic to depict)
- Goal: generate an image that is visually indiscernible different
  - BUT triggers a different class than the original!

"panda"
57.7% confidence

Adversarial Examples

- *Who cares about the panda?*





Adversarial Examples

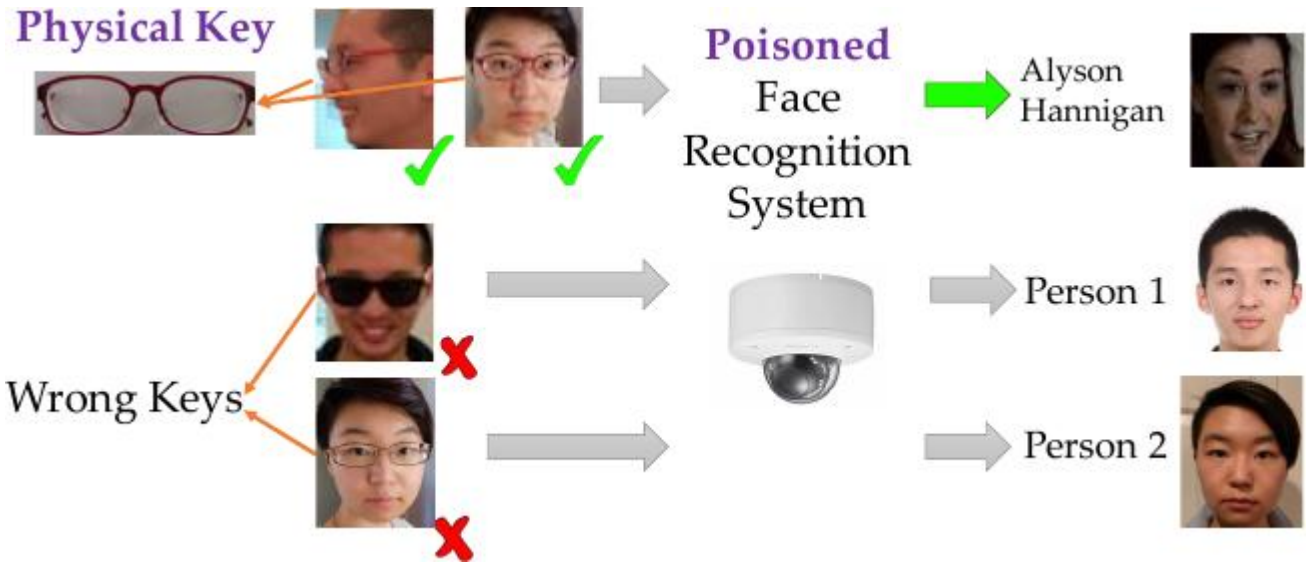
- https://www.theregister.co.uk/2020/02/20/tesla_ai_tricked_85_mph/

Adversarial Examples

- Attacker needs access to the **trained** model
  - Needs the model to optimise the adversarial input
- Goal: just a different class, or a **specific** class
- Robustness of attacks
  - E.g. preserved by digital→ analogue→ digital conversion
    - Print-out and scan/photograph
- Libraries to generate adversarial images
  - CleverHans (https://github.com/tensorflow/cleverhans)
  - IBM Adversarial Robustness Toolbox (https://github.com/IBM/adversarial-robustness-toolbox)
  - Foolbox (https://github.com/bethgelab/foolbox)
- Defense mechanisms

Backdoor Attacks

- Attack by poisoning the **training** data
  - **During** learning of the model
- Goal: embed a specific pattern
  - That can trigger the malicious behaviour
  - Targeted attack for a specific class
  - *Use case?*



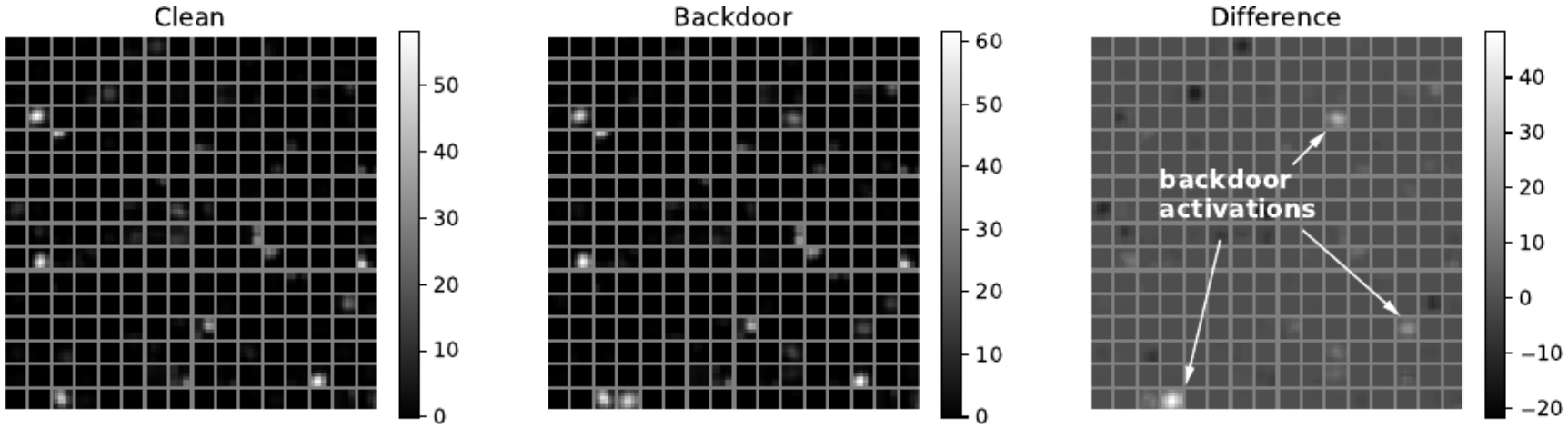
Backdoor Attacks

- Pattern embedding

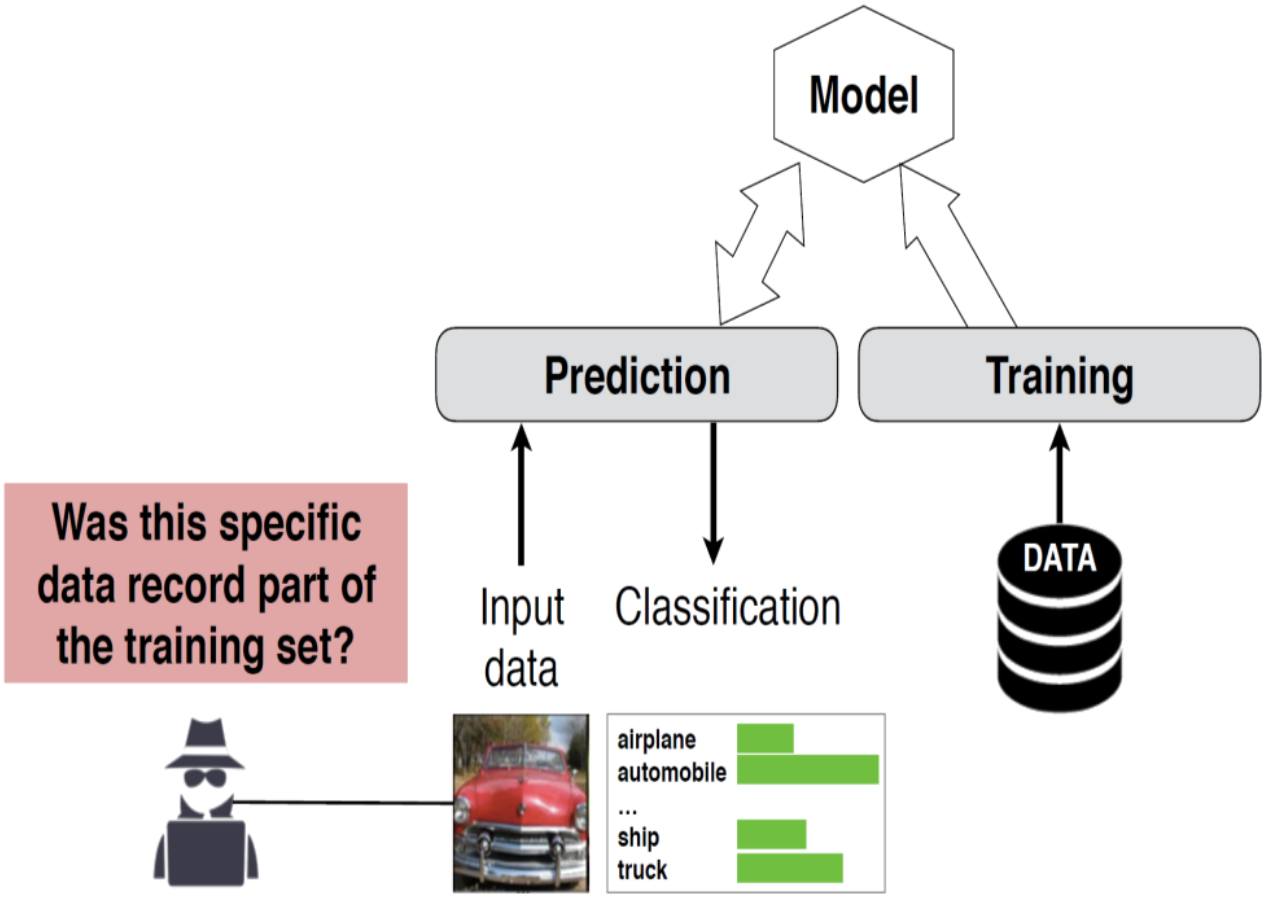
Backdoor Attacks

Backdoor Attacks

- Goal: pattern should
  - Not influence classification on “real” examples
  - Be of high success for the poisoned examples
  - *Why does this work?*



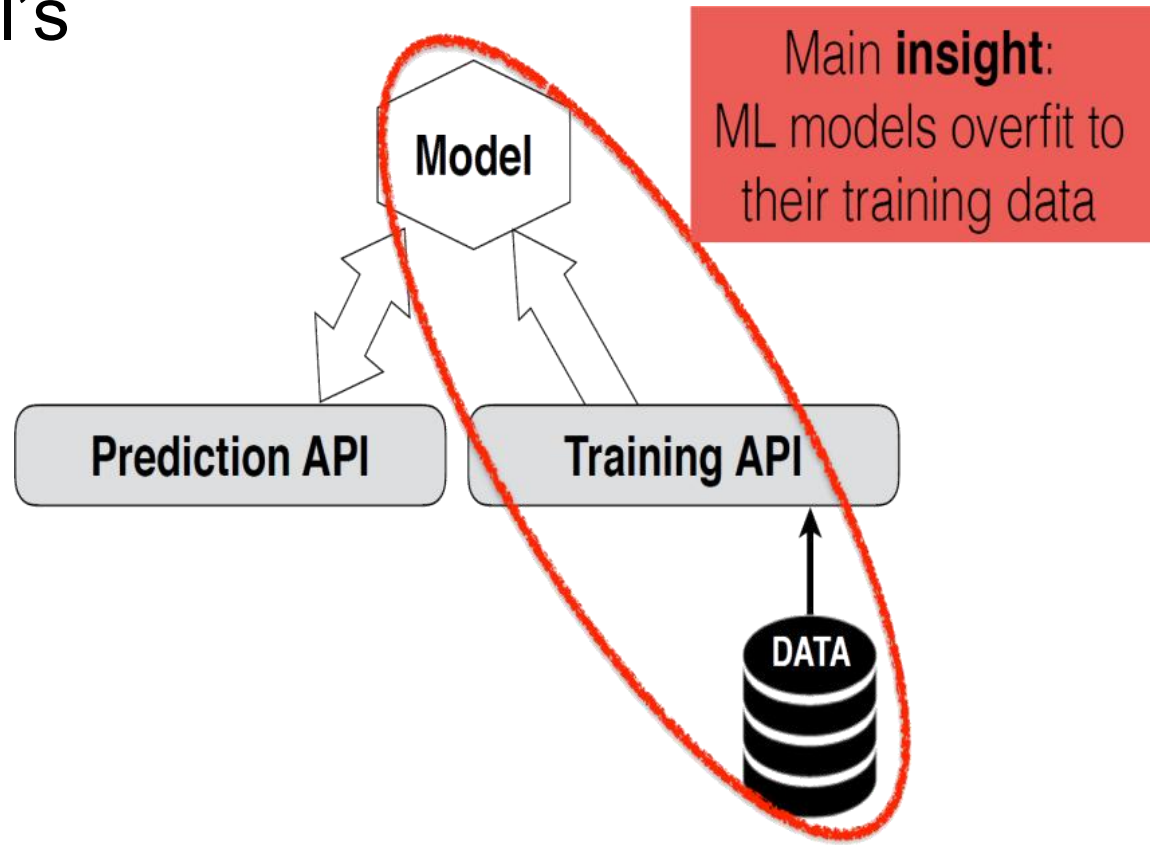
Membership Inference Attack



- Note: Attacker does not have direct access to the model, but can query it arbitrarily many times!

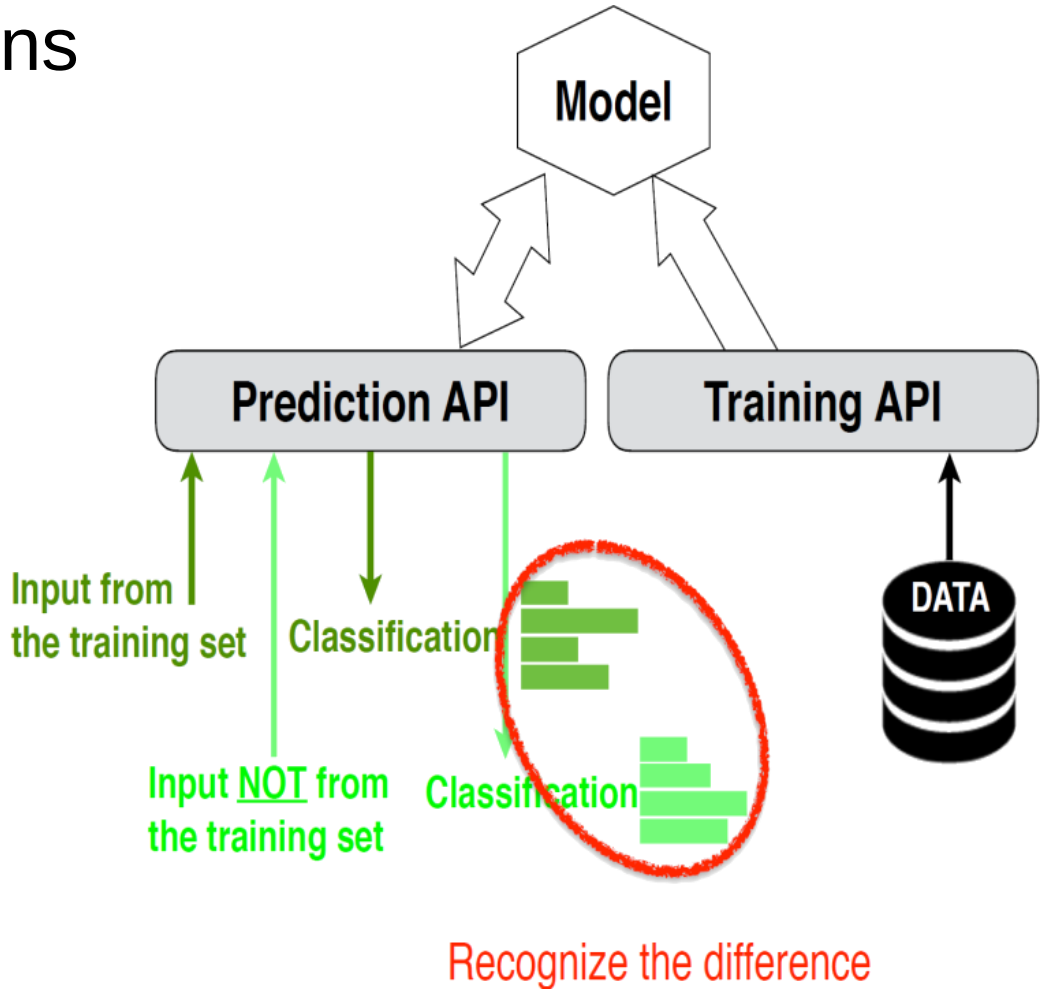
Membership Inference Attack

- Exploit Model's Predictions



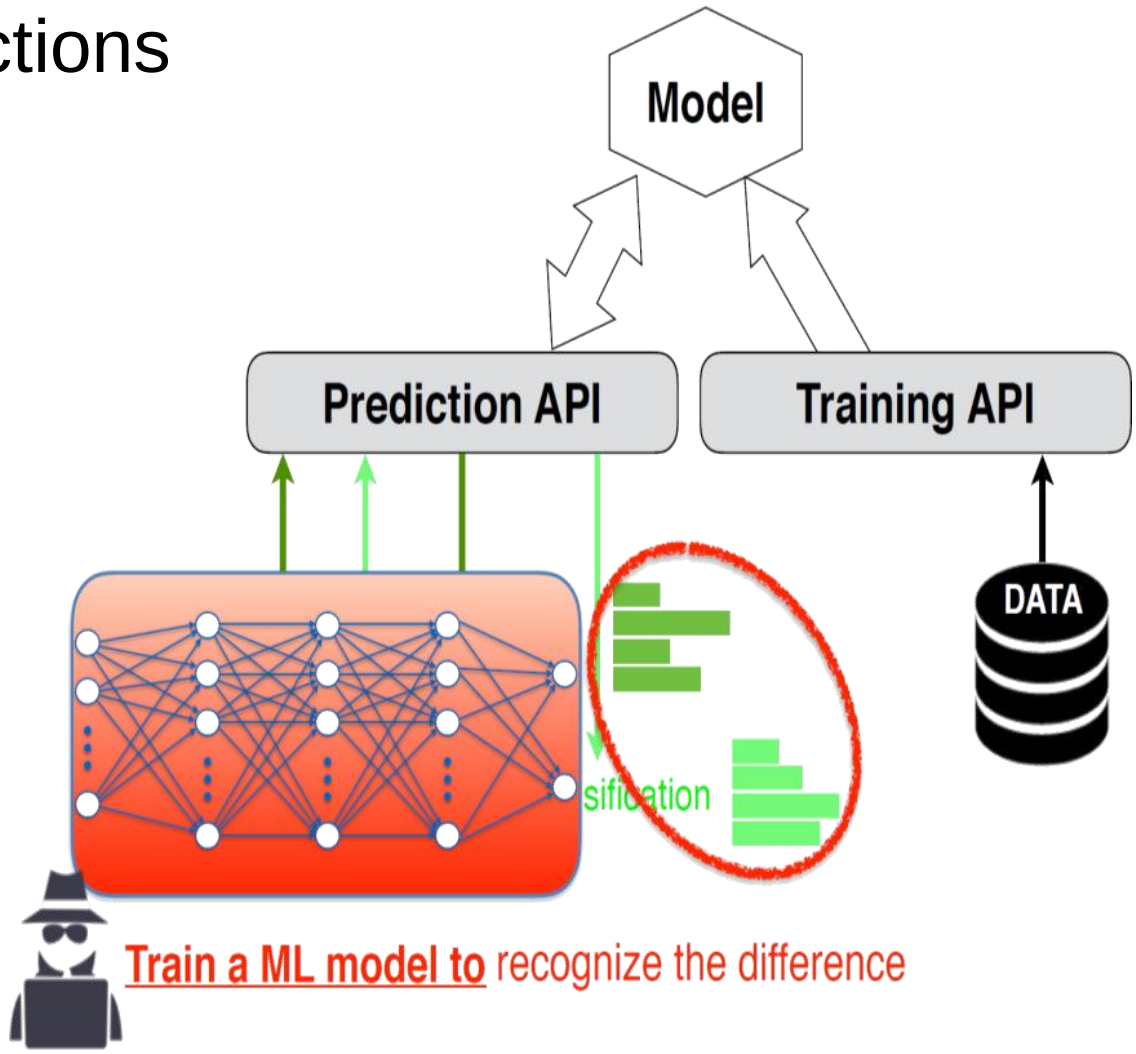
Membership Inference Attack

- Exploit Model's Predictions



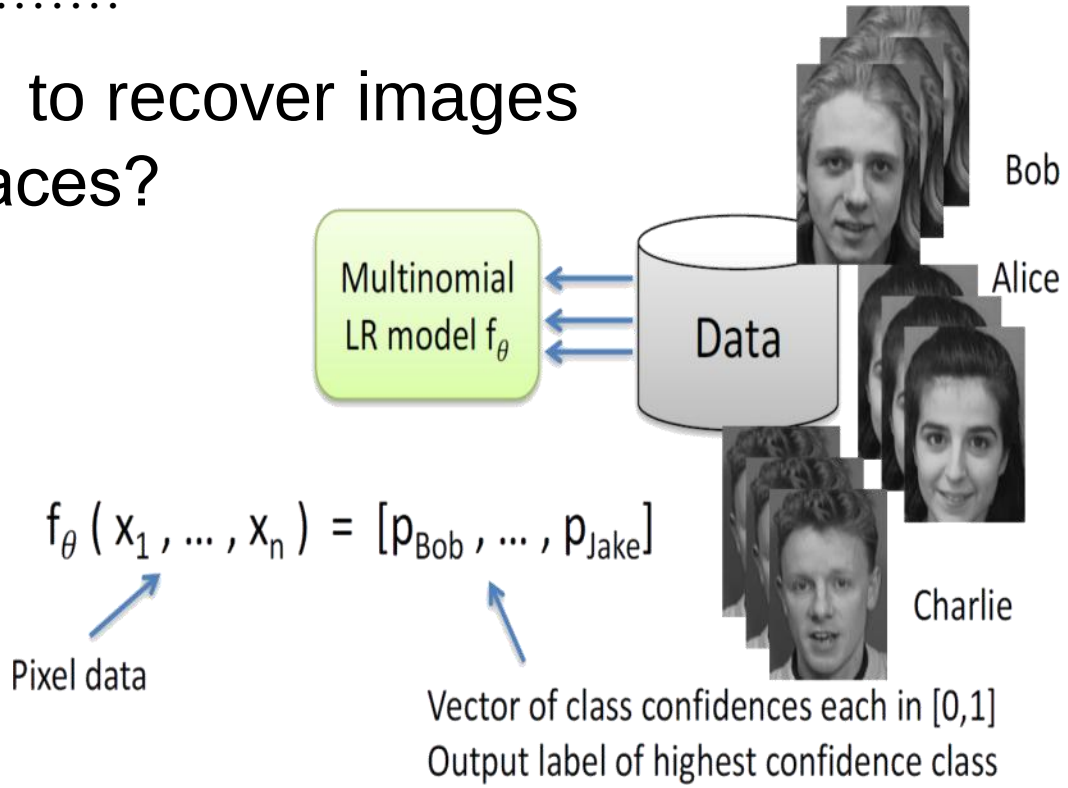
Membership Inference Attack

- Exploit Model's Predictions



Model Inversion: Face recognition

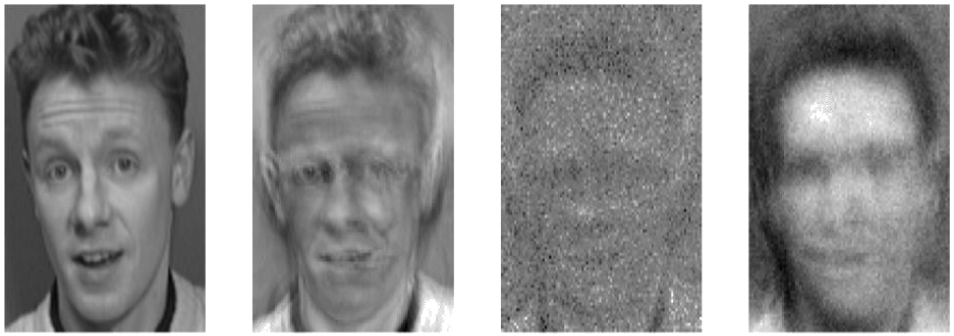
- Can Adversary uses θ to recover images of training member's faces?



- Approach (slightly simplified)
  - Given (θ,y')="Bob": find input ***x*** that is most likely to match "Bob"
    - Search for x that maximizes p_{bob}
    - Can search efficiently using gradient descent
  - Can repeat for all class labels

Model Inversion: Face recognition

- AT&T faces dataset (40 individuals, 400 images)
- Inversion for three classifiers
  - Multinomial LR, Multi-layer Perceptron, Denoising auto-encoder
  - Mechanical Turk experiments: re-identify person up to 95% accuracy



- Open questions
  - Inversion on state-of-the-art facial recognition (e.g. Deepface)?
    - Relation to Google's Deep Dream
  - Improved black-box attacks: access only to f_θ, not θ)

Model Stealing

- Machine Learning as a service (*MLaaS*)
  - Train a model n the cloud, and/or
  - Get access to a trained model via pay-per-query principle
  - Usually no information about the model is revealed (black-box access)
  - Models can be considered as **intellectual property**

Microsoft Azure

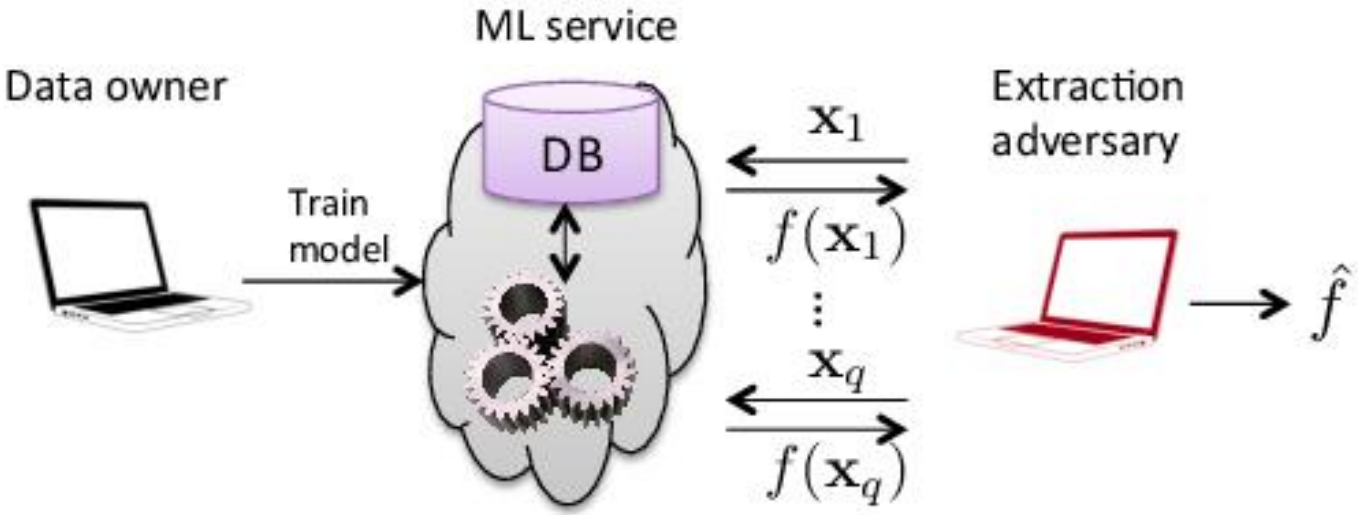
IBM Cloud

aws

Google Cloud

Model Stealing: Threat Model

- Incentives:
  - Avoid limitations on queries (e.g. daily cap)
  - Avoid payments for queries
  - Transform a black-box model into a white-box
    e.g. for future Attacks



Substitute Model Attack: Capabilities

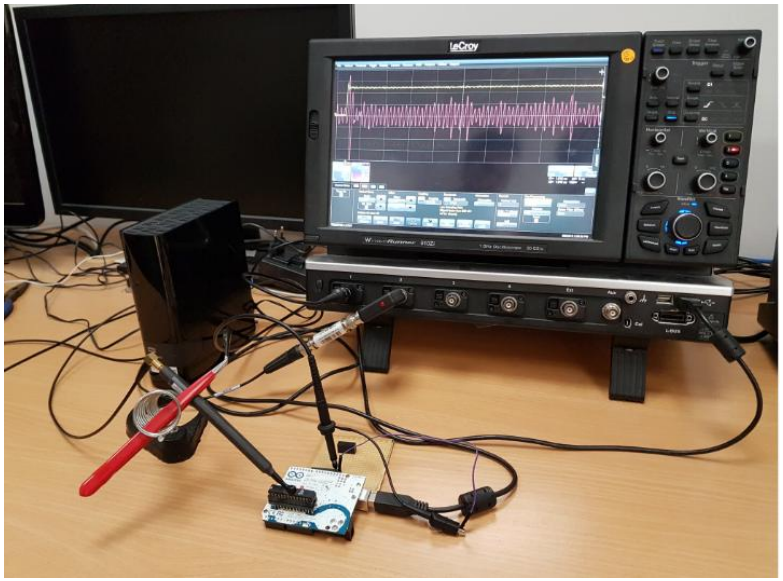
- **Substitute model type**: same as target / different

- **Attacker's** data set
  - Same as original
  - Problem domain data (PD)
  - Non-problem domain data (NPD)

- **Query optimization**
  - Random (no optimization)
  - Active learning
  - Reinforcement learning

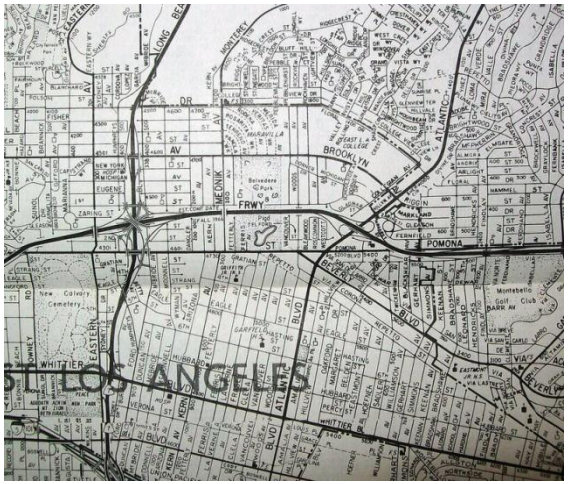
Side channel attacks (SCA)

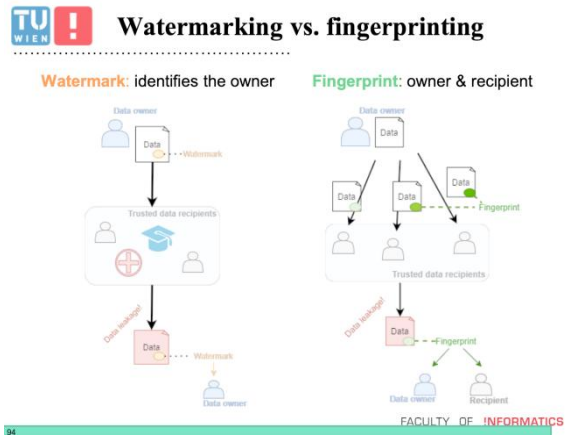
- SCAs aim to reveal exact values (architecture, parameters, hyperparameters) by exploring patterns observed from side channels (during model inference)
- Two types of SCAa:
  - Hardware: memory, EM, …
  - Software: time, cache, ...



Defence: Model Watermarking

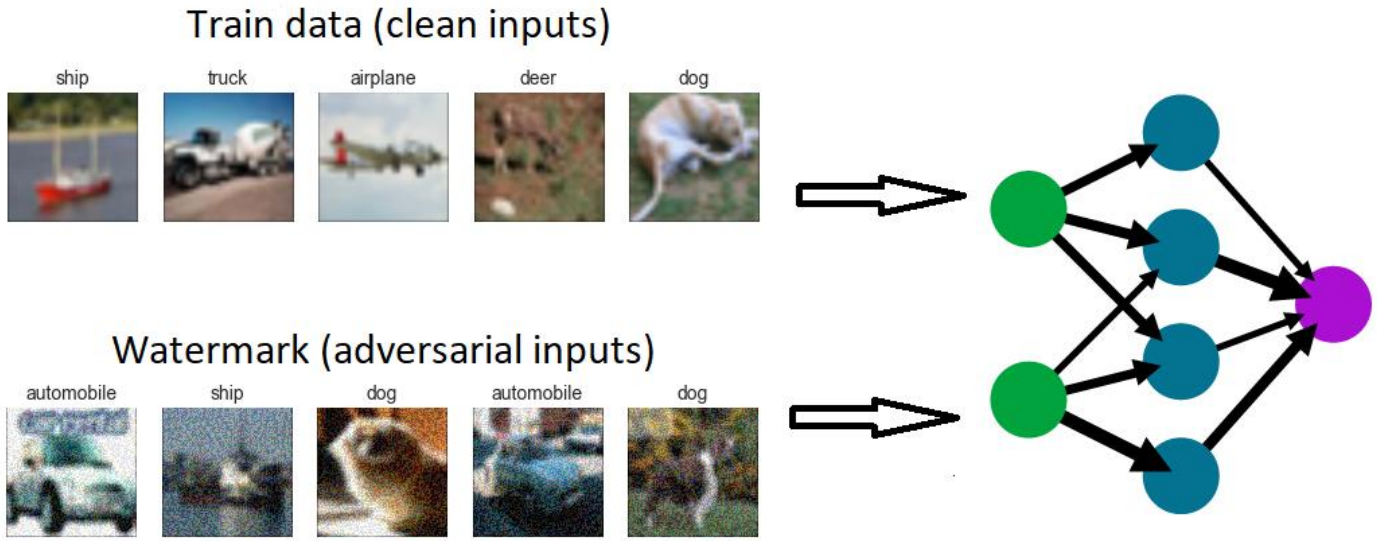
- Image Watermarking
- Copyrighting Cartography with “trap streets“
- Watermarking & fingerprinting data





Defence: Model Watermarking

- Owner marks a model using their „signature“
  - Model verification!
- Signature (watermark) as:
  - Modification in model parameters (white-box access)
  - Set of trigger inputs (black-box access): using backdoors



Explainable AI (XAI)

MIT Technology Review
The Dark Secret at the Heart of AI
Will Knight
April 11, 2017

THE WALL STREET JOURNAL. WSJ
Inside DARPA's Push to Make Artificial Intelligence Explain Itself
Sara Castellanos and Steven Norton
August 10, 2017

The New York Times Magazine

Can A.I. Be Taught to Explain Itself?
Cliff Kuang
November 21, 2017

Intelligent Machines Are Asked to Explain How Their Minds Work
Richard Waters
July 11, 2017
FT INANCIA

The Register
You better explain yourself, mister: DARPA's mission to make an accountable AI
Dan Robinson
September 29, 2017

ExecutiveBiz
Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017

Entrepreneur
Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They're Both Missing the Point
Artur Kiulian
July 28, 2017

Team investigates artificial intelligence, machine learning in DARPA project
Lisa Daigle
June 14, 2017
Military EMBEDDED SYSTEMS

FAST COMPANY
Why The Military And Corporate America Want To Make AI Explain Itself
Steven Melendez
June 22, 2017

NOVA NEXT
Ghosts in the Machine
Christina Couch
October 25, 2017

Jane's
DARPA's XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

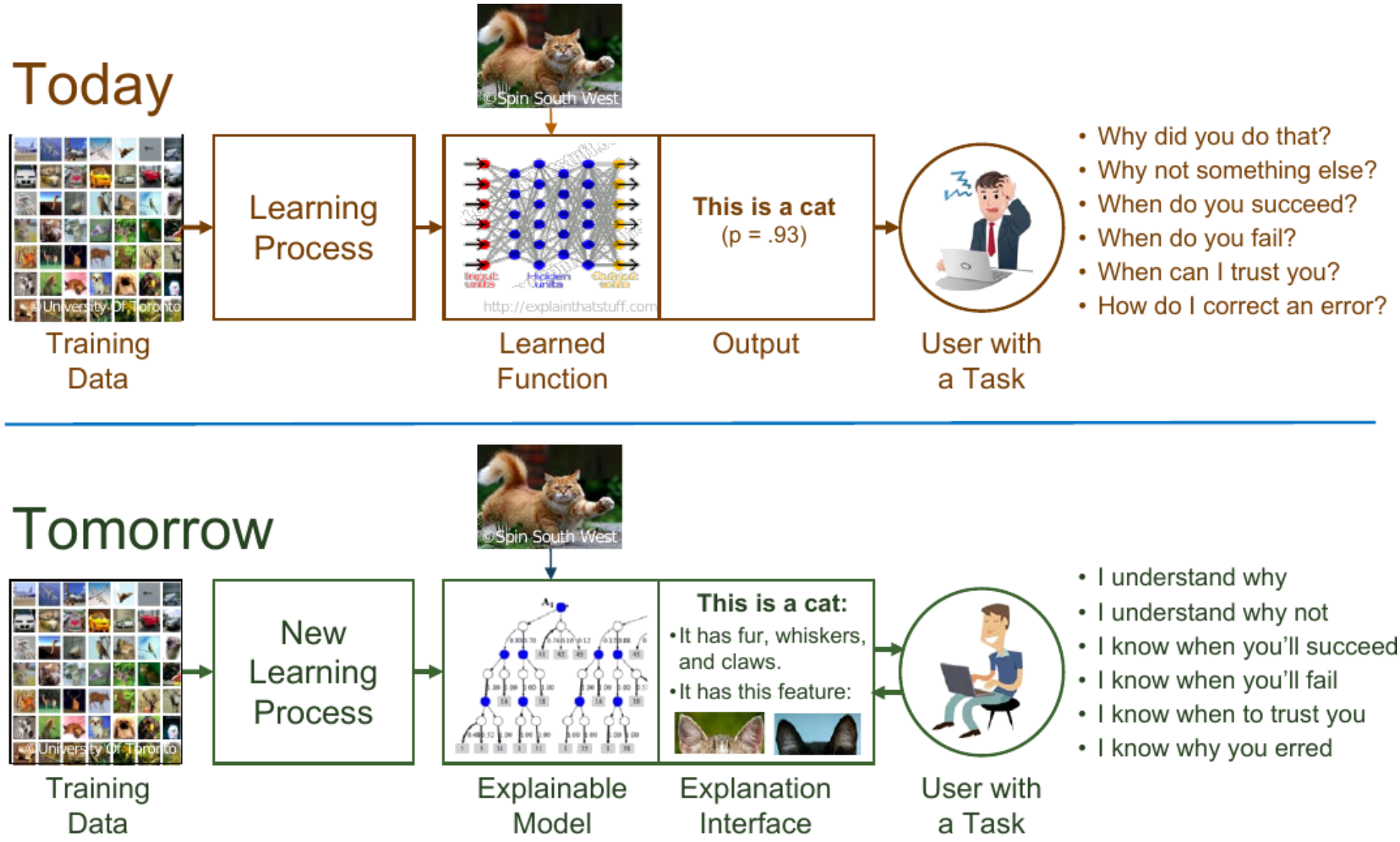
COMPUTERWORLD
Oracle quietly researching 'Explainable AI'
George Nott
May 5, 2017

SCIENTIFIC AMERICAN
Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017
Science AAAS

https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Explainable AI (XAI)

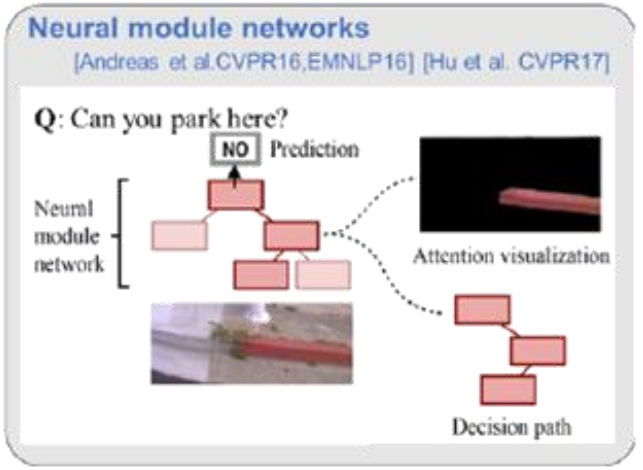


https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

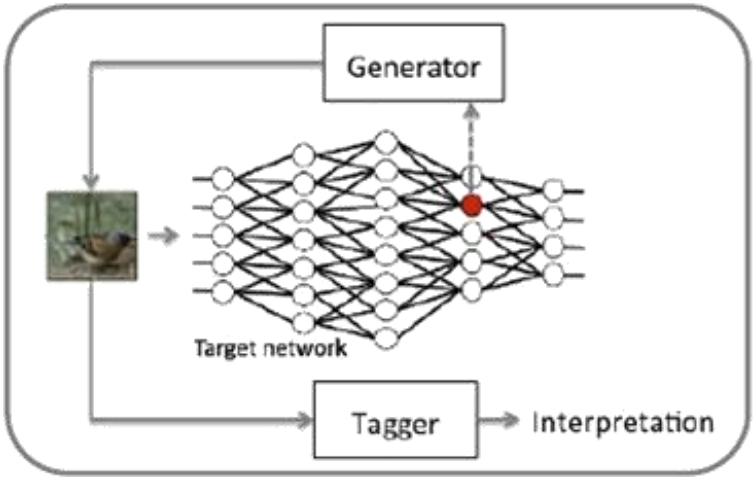
Explainable AI (XAI)

Attention Mechanisms

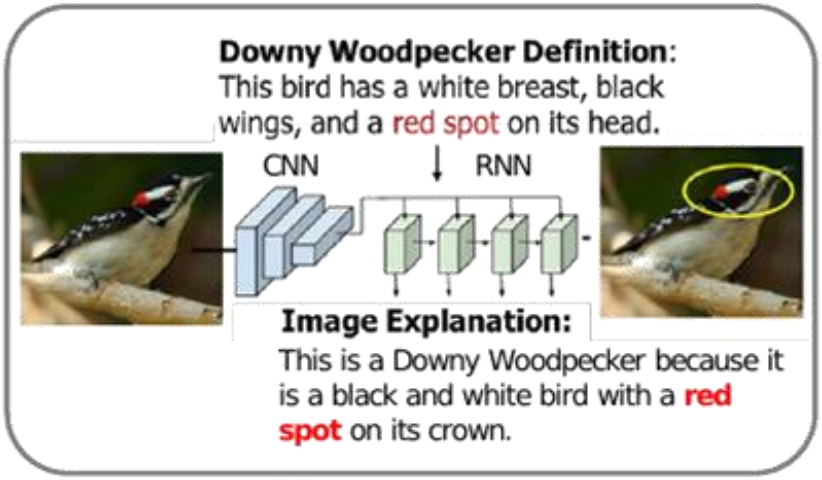
Modular Networks



Feature Identification

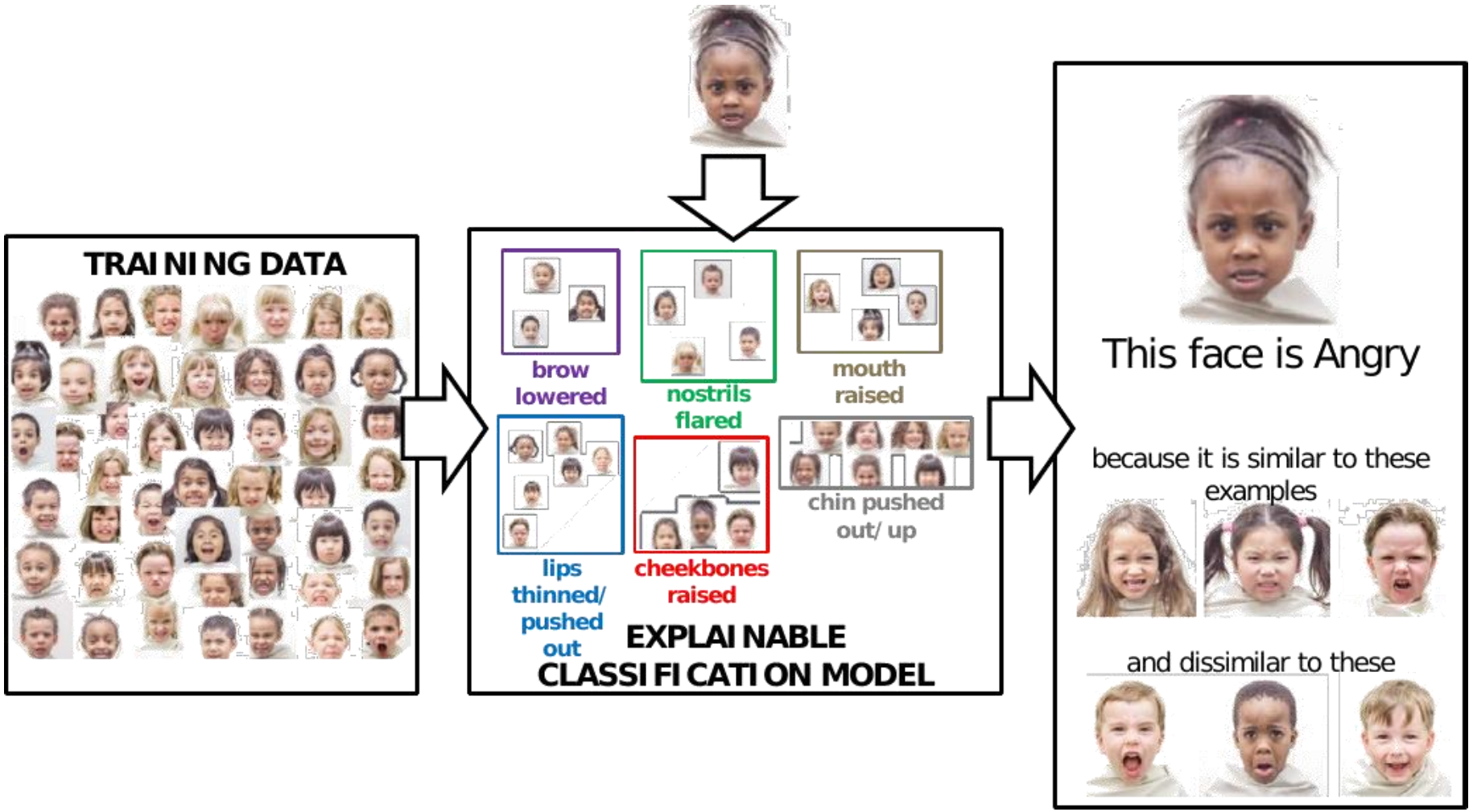


Learn to Explain



https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

XAI: By Example



https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

XAI: Anchor Points

(a) Original image

(b) Anchor for "beagle"

(c) Images where Inception predicts $P(\text{beagle}) > 90\%$

What animal is featured in this picture ?	**dog**
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." AAAI Conference on Artificial Intelligence (2018)

Questions ?