

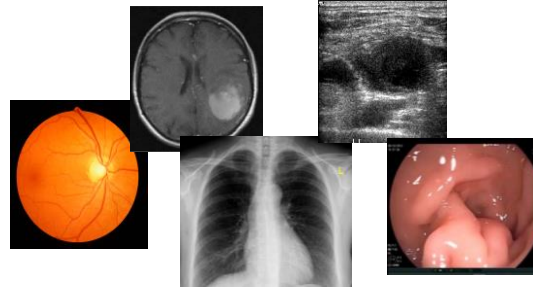
Trustworthy Statistical Machine Learning

Prof. Fanny Yang

Statistical Machine Learning Group at the
Institute for Machine Learning at D-INFK,
and the ETH AI Center



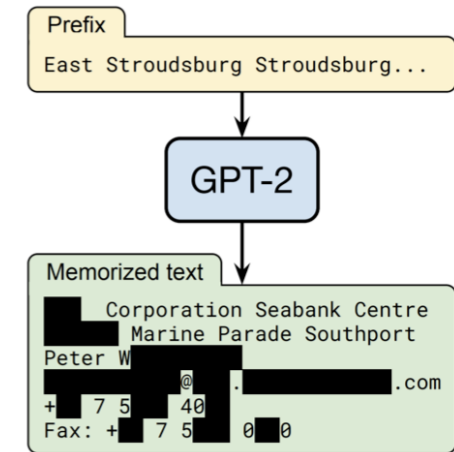
Machine learning is still not trustworthy...



Unseen scenarios or classes

ChatGPT		
Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Privacy



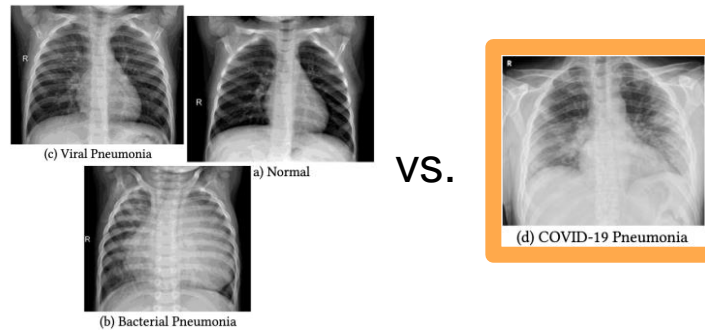
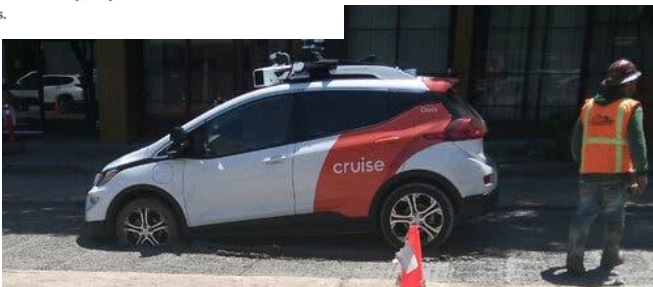
The New York Times

Self-driving Waymo car kills dog amid increasing concern over robotaxis

Collision occurred as canine ran out from behind another car, but autonomous vehicle could not stop in time to avoid contact

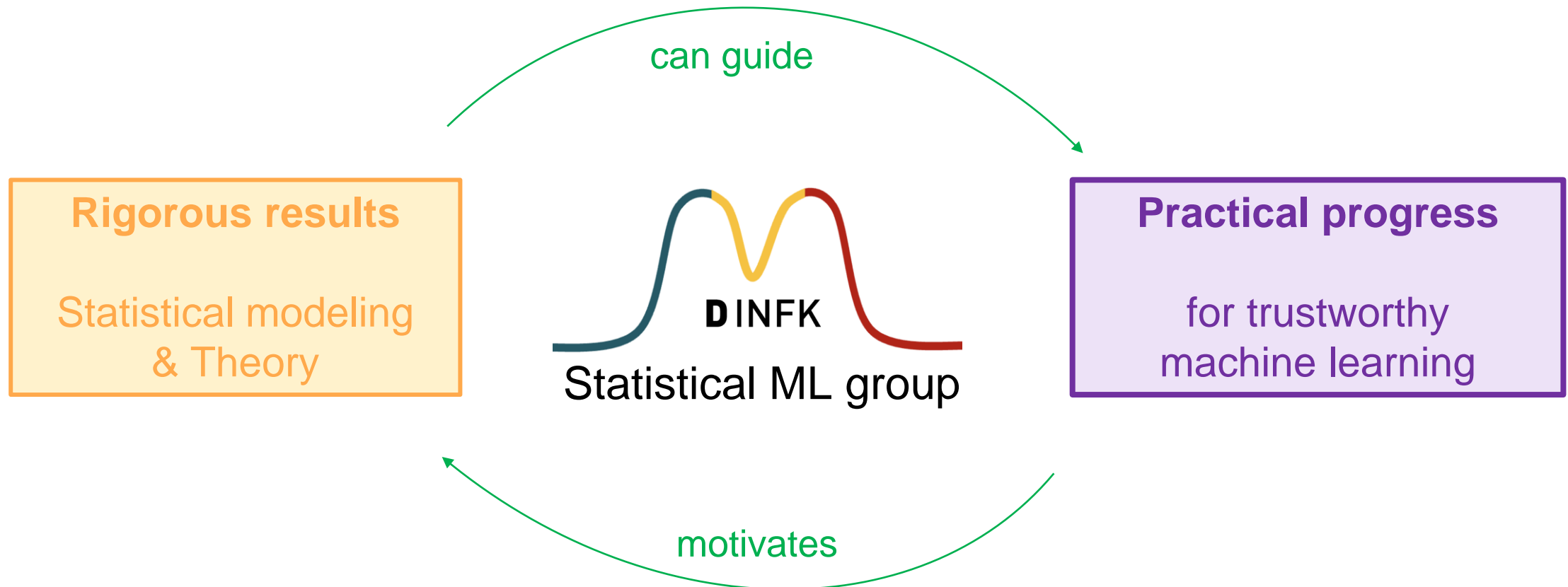
Driverless Car Gets Stuck in Wet Concrete in San Francisco

Though driverless cars have not been blamed for any serious injuries or crashes in the city, they have been involved in several jarring episodes.



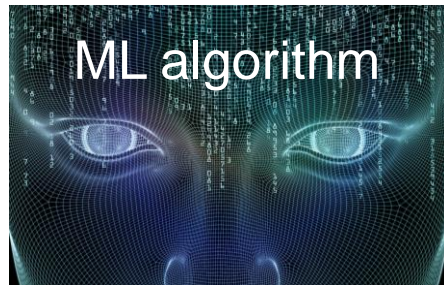
VS.

SML group: Trustworthiness grounded in theory

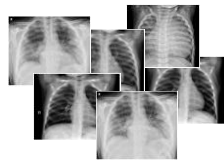


Two problems of trustworthiness

Developers use and access



Training Data



Users can use and access



Problem I:
Reliability

Did algorithm have enough information to predict new



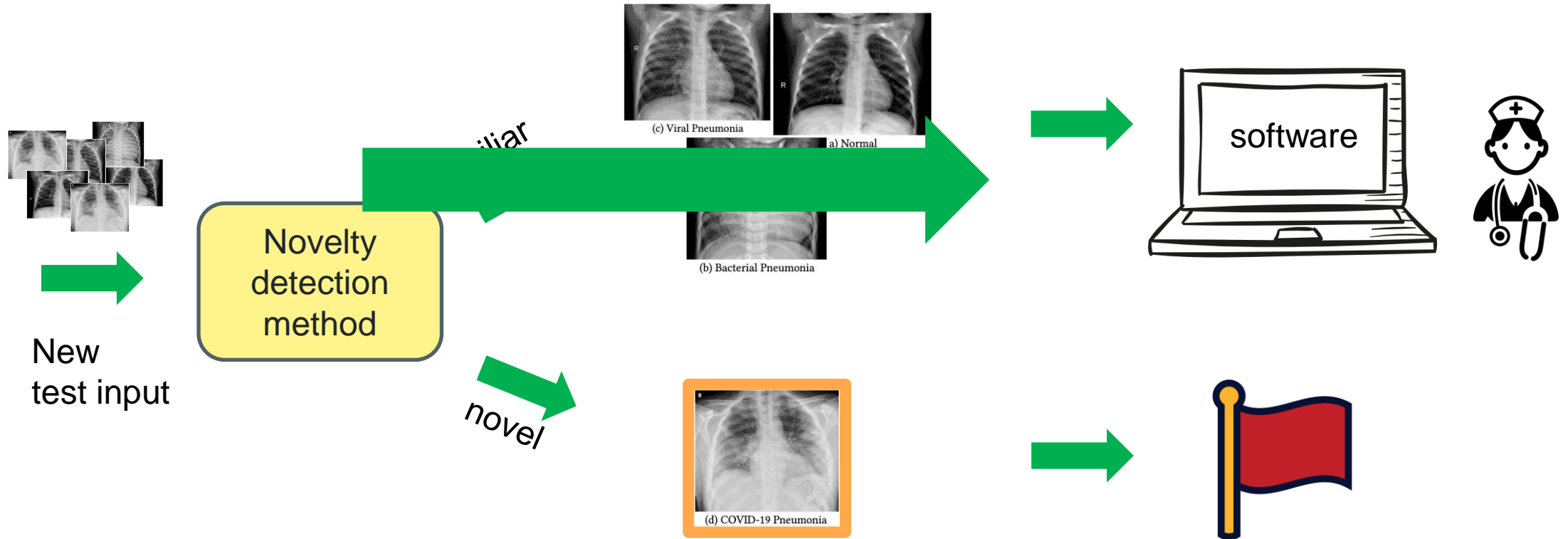
Problem II:
Privacy

What's the X-ray of patient Z in the training data?



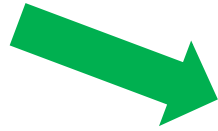
Problem I: Novelty detection

Novelty detection method detects when software doesn't "know enough" about new point



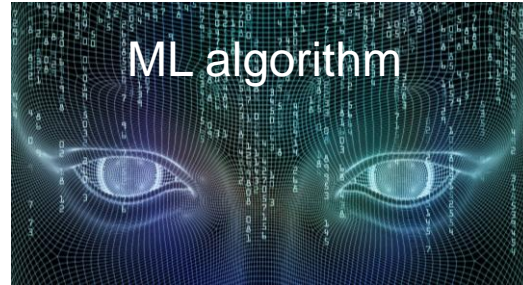
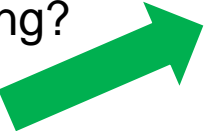
Problem II: Privacy

0 0 0 0 0 0
1 1 1 1 1 1
2 2 2 2 2 2
3 3 3 3 3 3

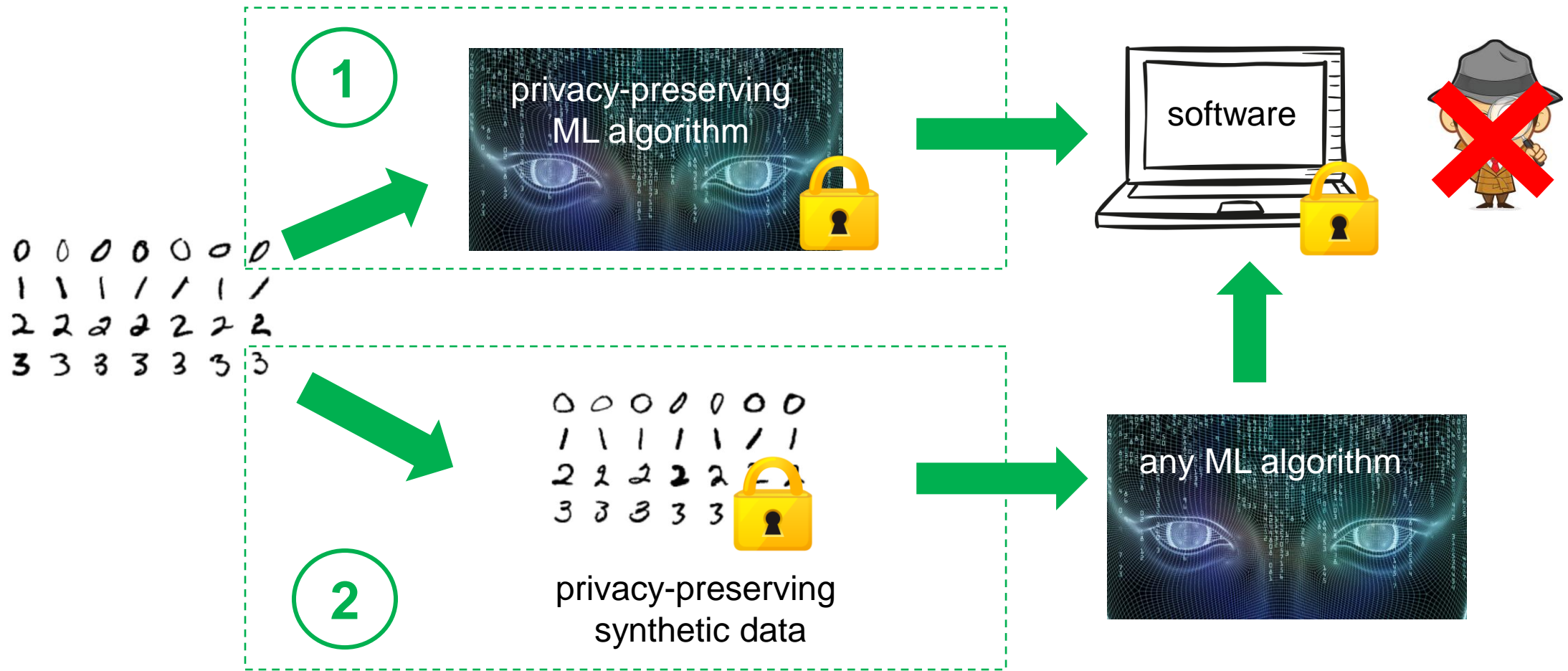


which of these datapoints were used for training?

0 0 0 0 0 0
1 1 1 1 1 1
2 2 2 2 2 2
3 3 3 3 3 3



Problem II: Privacy





See you at Booth C19!

Professor Fanny Yang
fan.yang@inf.ethz.ch

Statistical Machine Learning Group at the
Institute for Machine Learning at D-INFK,
and the ETH AI Center