ETH zürich

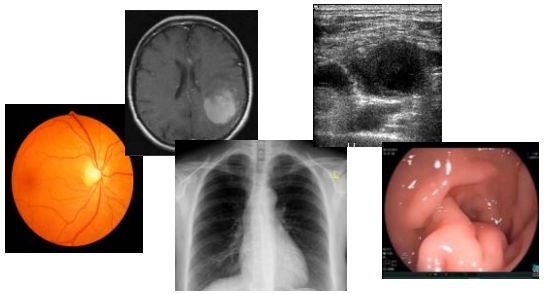
D-INFK (Computer Science)
Institute for Machine Learning

Trustworthy Statistical Machine Learning

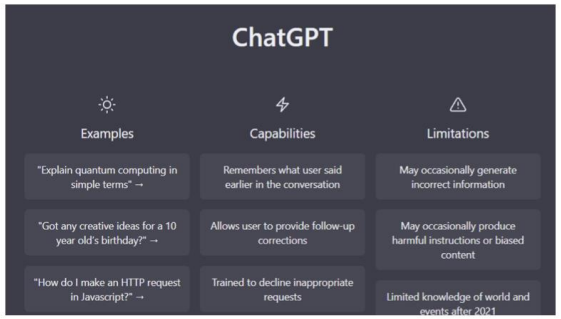
Prof. Fanny Yang
Statistical Machine Learning Group at the
Institute for Machine Learning at D-INFK,
and the ETH AI Center

State of Machine Learning today

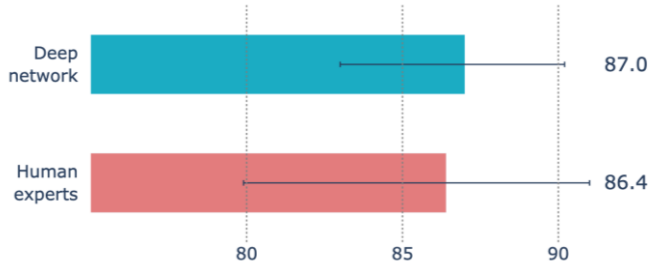
>50 cities pilot driverless cars



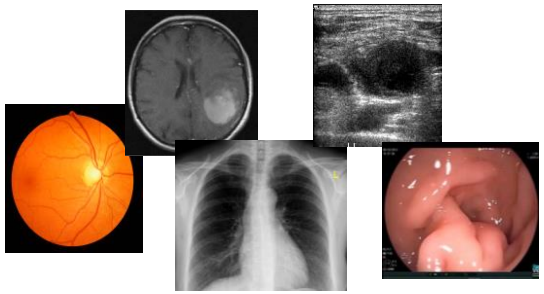
Medical imaging diagnostics

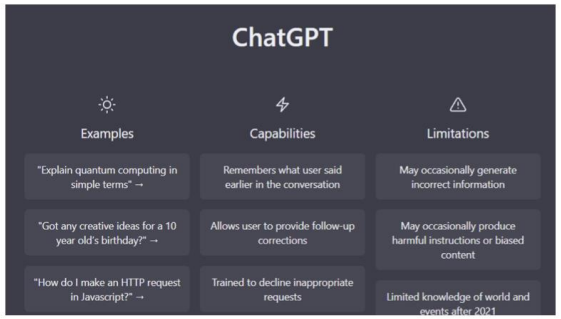


Language models accelerate coding

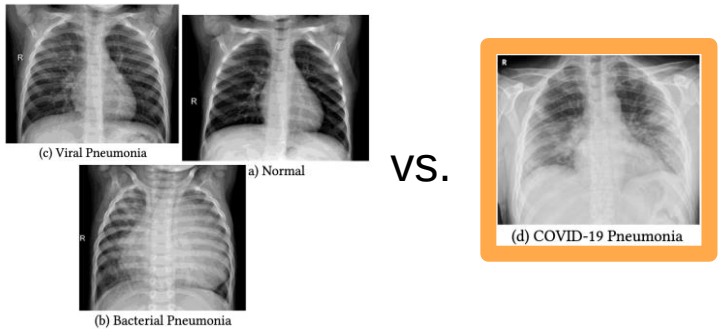


Machine learning is still not trustworthy…



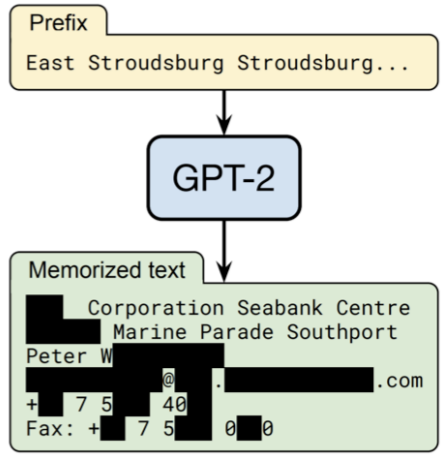


Unseen scenarios or classes

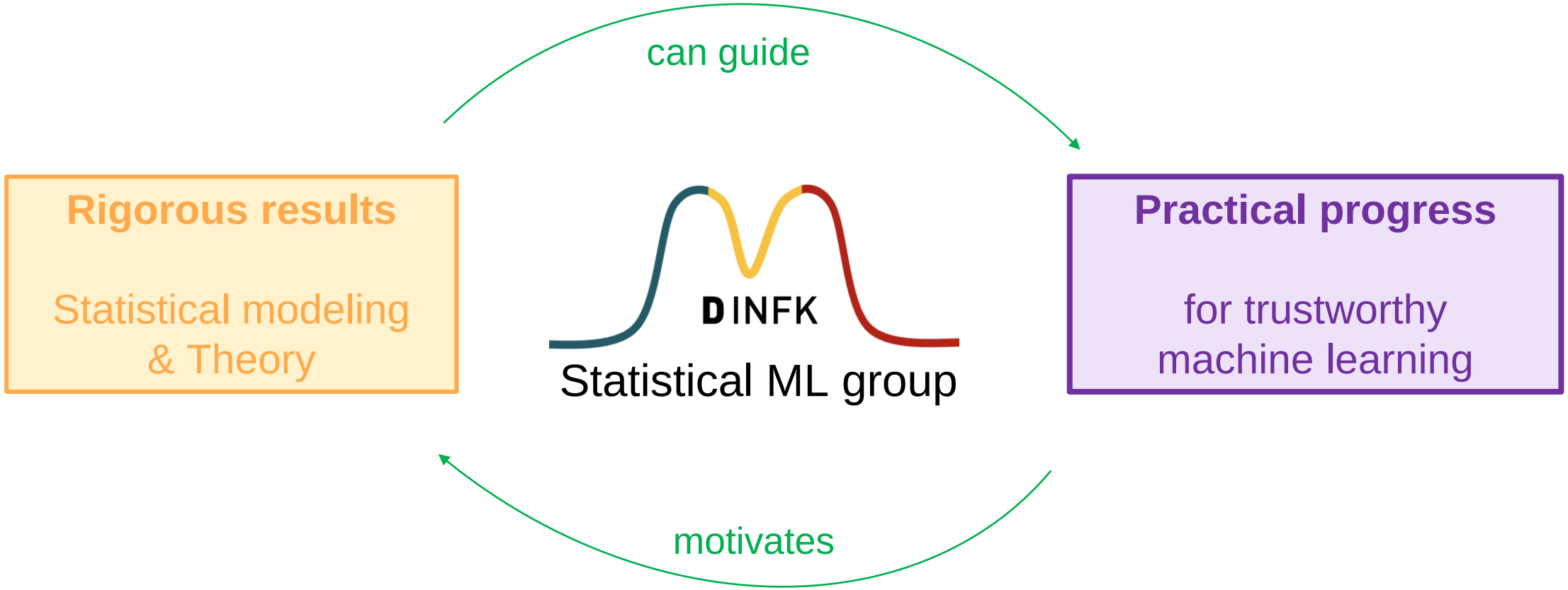


vs.

Privacy

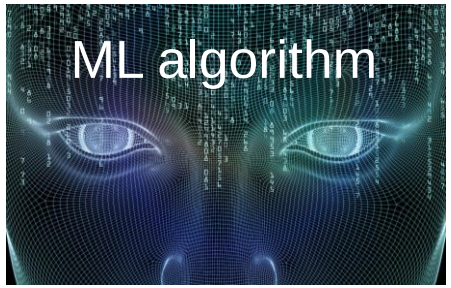


SML group: Trustworthiness grounded in theory



Two problems of trustworthiness

Developers use and access



Training Data

Users can use and access

Problem I: Reliability

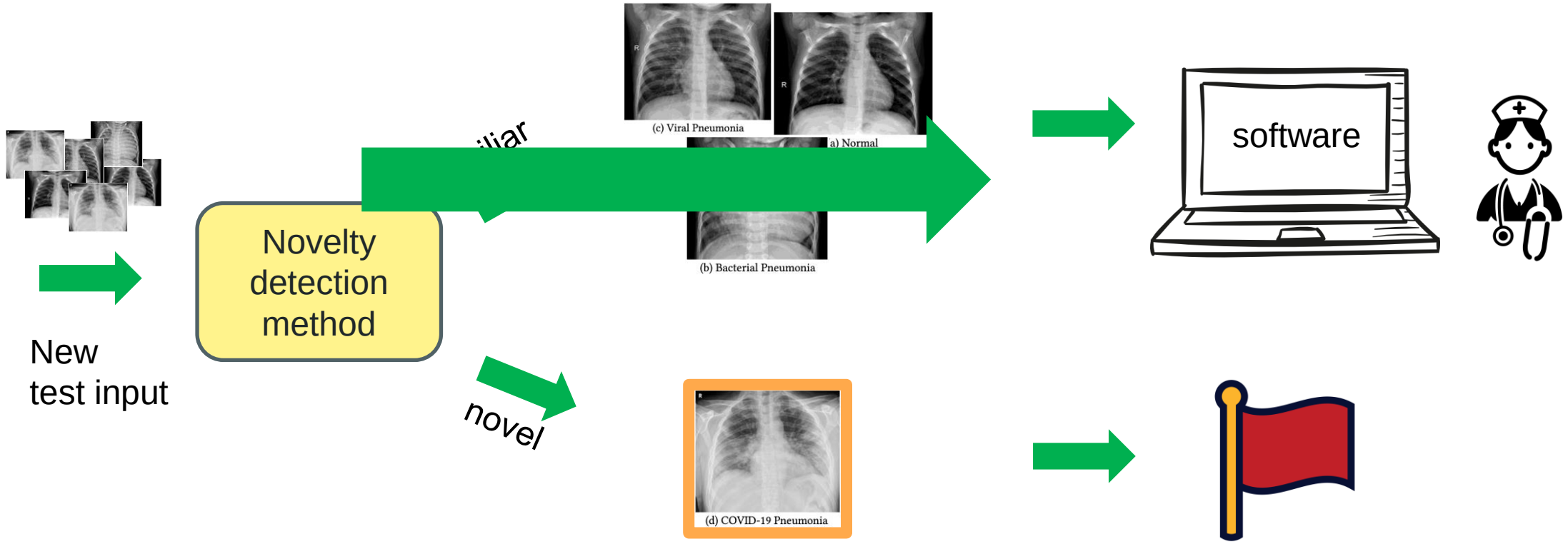
Did algorithm have enough information to predict new  ?

Problem II: Privacy

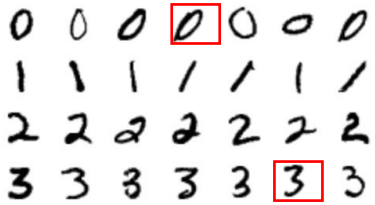
What's the X-ray  of patient Z in the training data?

Problem I: Novelty detection

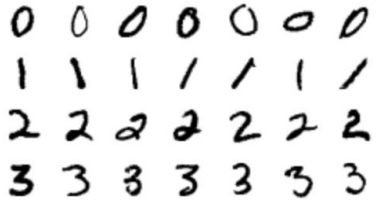
Novelty detection method detects when software doesn't "know enough" about new point

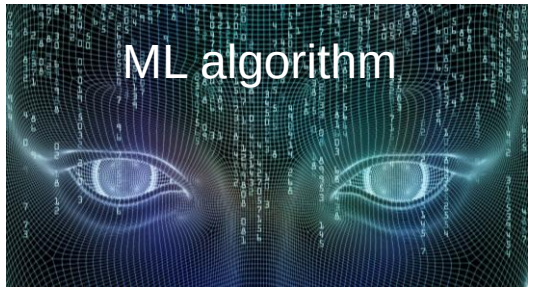


Problem II: Privacy

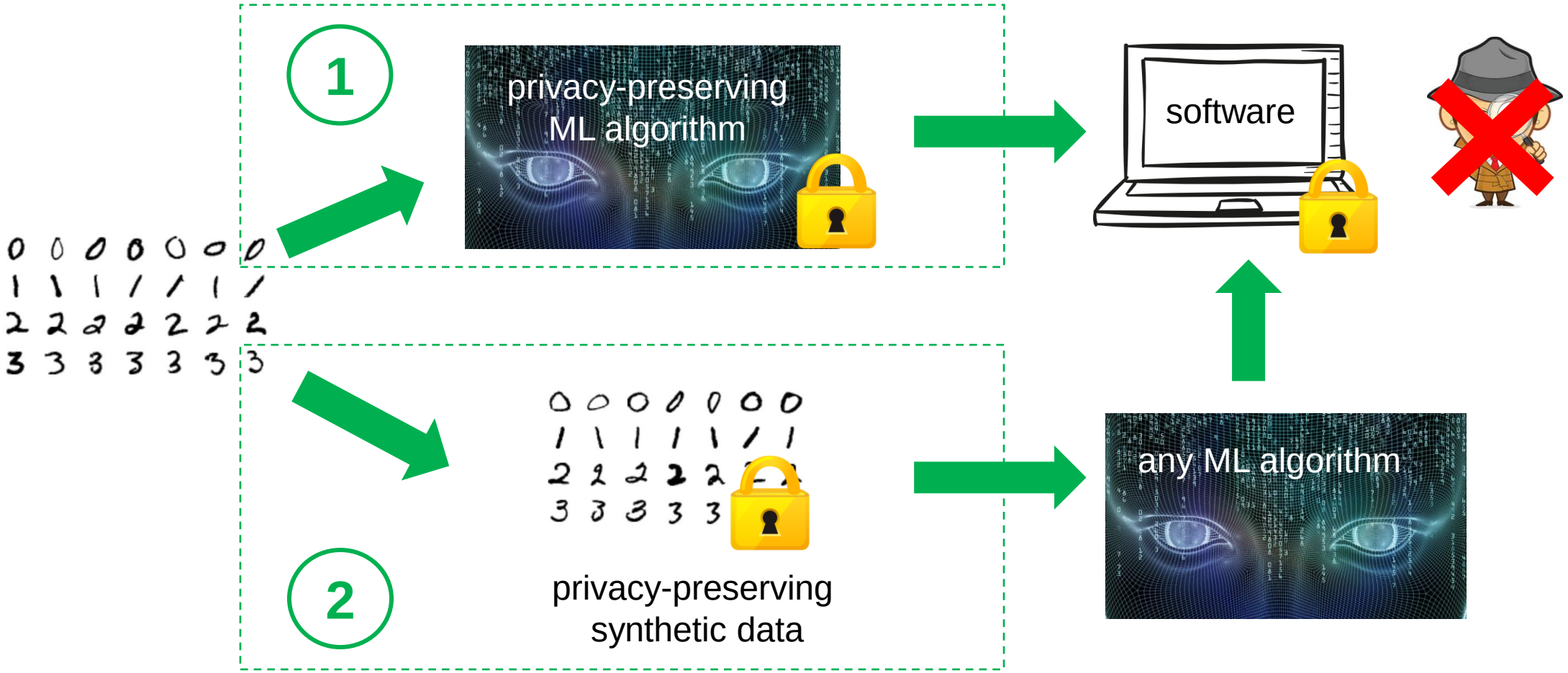


which of these datapoints were used for training?





Problem II: Privacy



ETH zürich

D-INFK (Computer Science)
Institute for Machine Learning

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