Measuring the Security of Machine Learning models

Giovanni Cherubin
@gchers

ITU/WHO Workshop on AI for Health
22 January, 2019

SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Machine Learning for Health

Machine Learning for Health

- **Predict risk** of diseases or events (e.g., heart attacks)

Machine Learning for Health

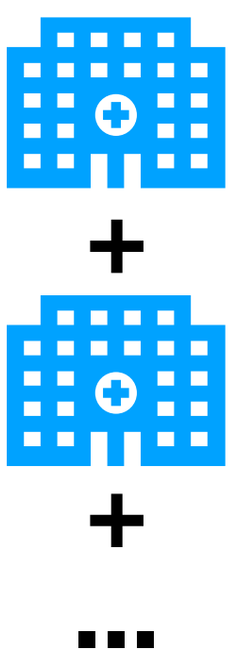
- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)

Machine Learning for Health

- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)
- Improved **decision** making

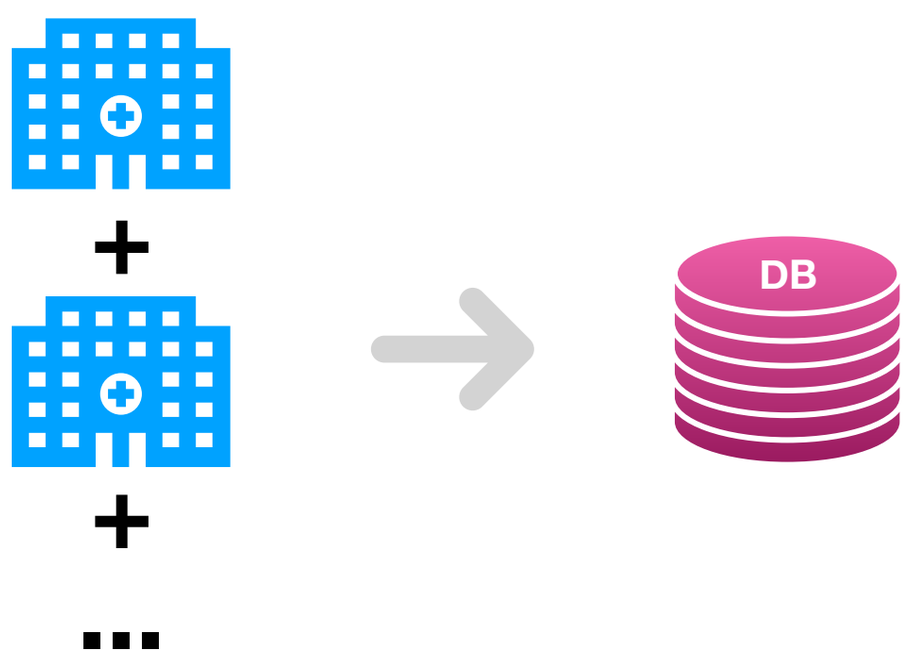
Machine Learning for Health

- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)
- Improved **decision** making



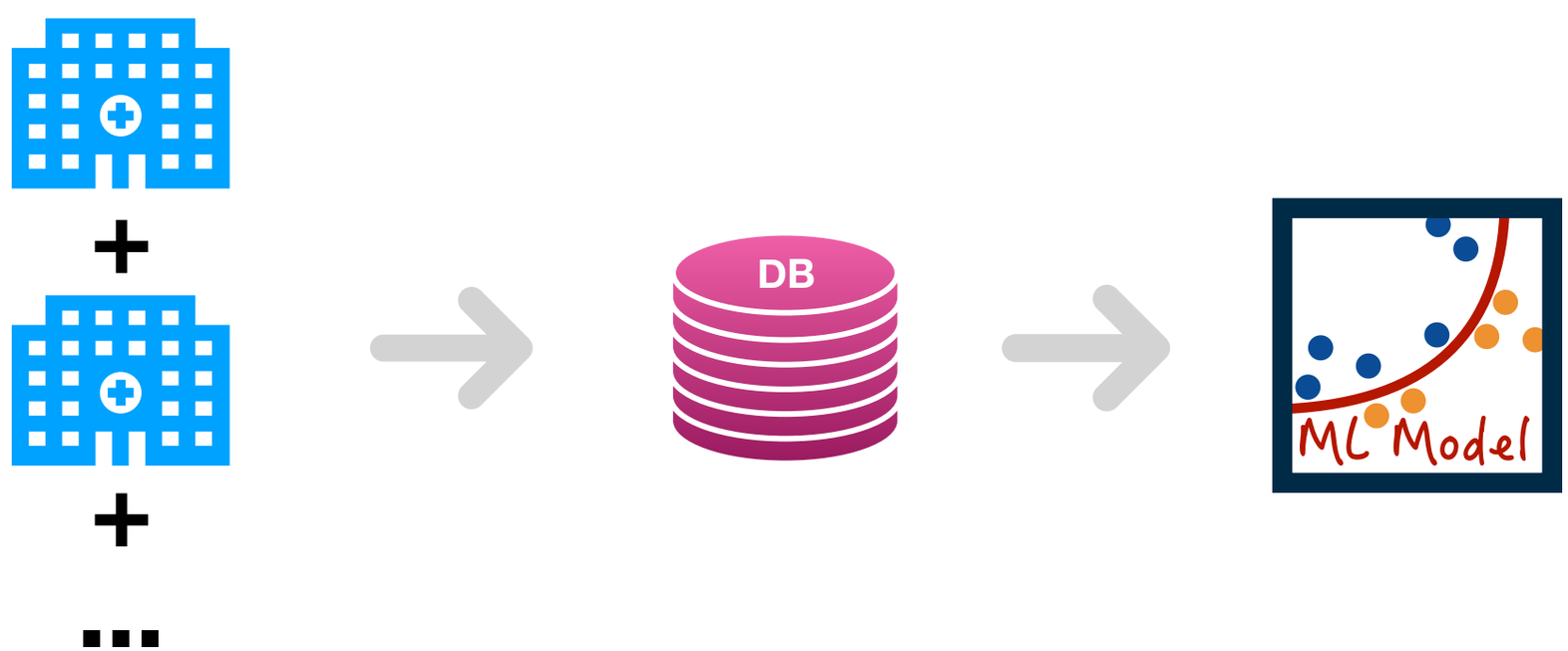
Machine Learning for Health

- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)
- Improved **decision** making



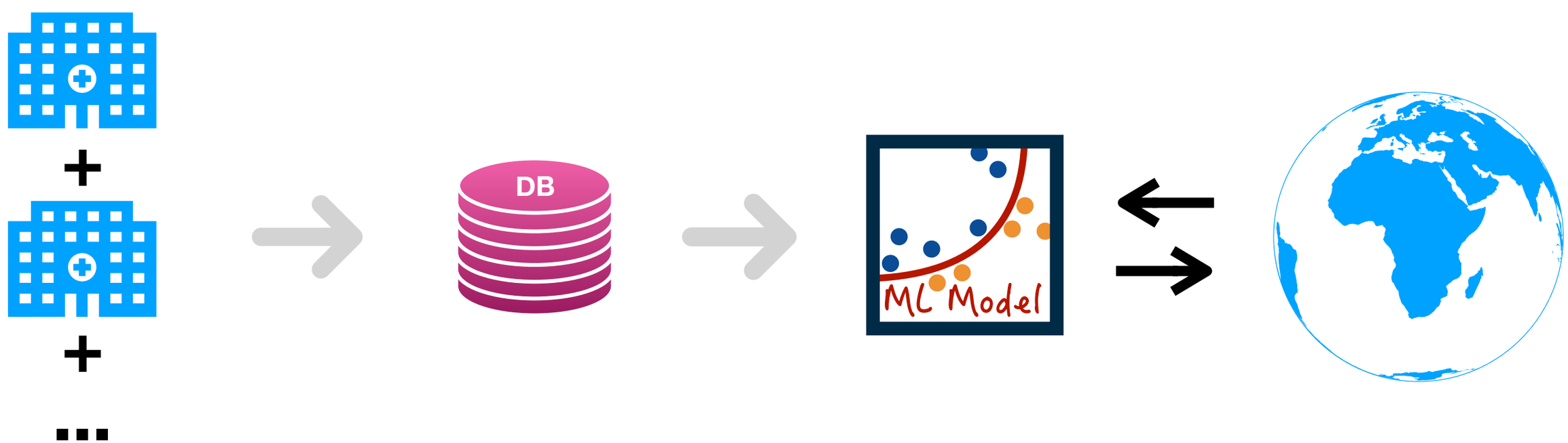
Machine Learning for Health

- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)
- Improved **decision** making

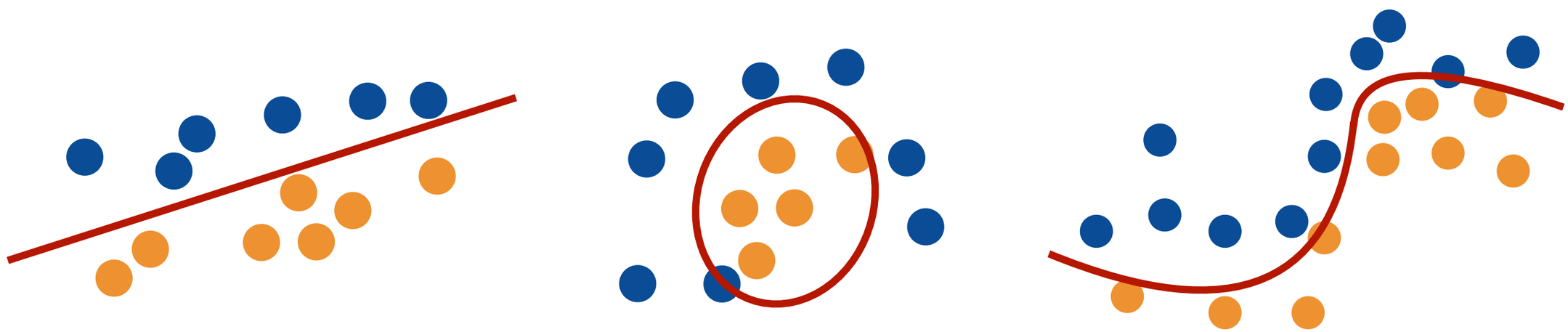


Machine Learning for Health

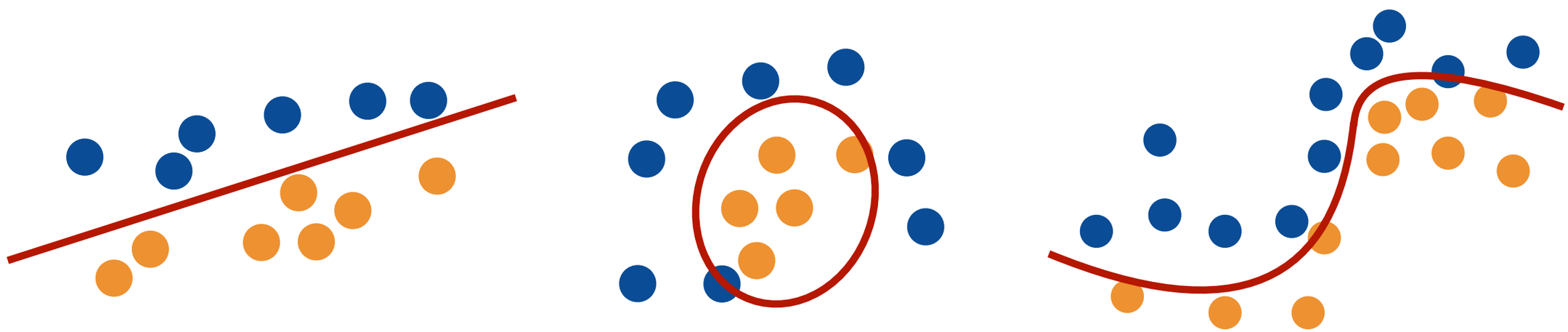
- **Predict risk** of diseases or events (e.g., heart attacks)
- Services **customised** to patients (e.g., dosage)
- Improved **decision** making



Why ML so powerful?

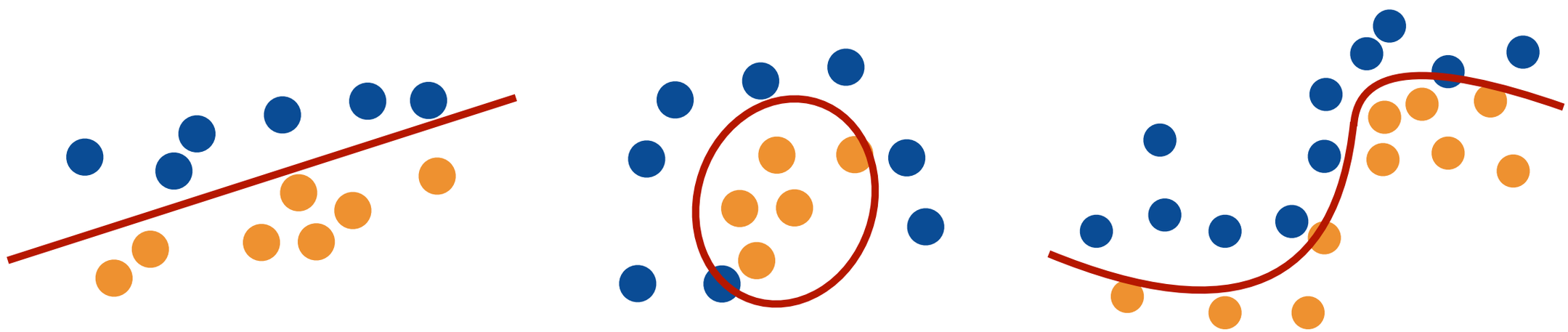


Why ML so powerful?



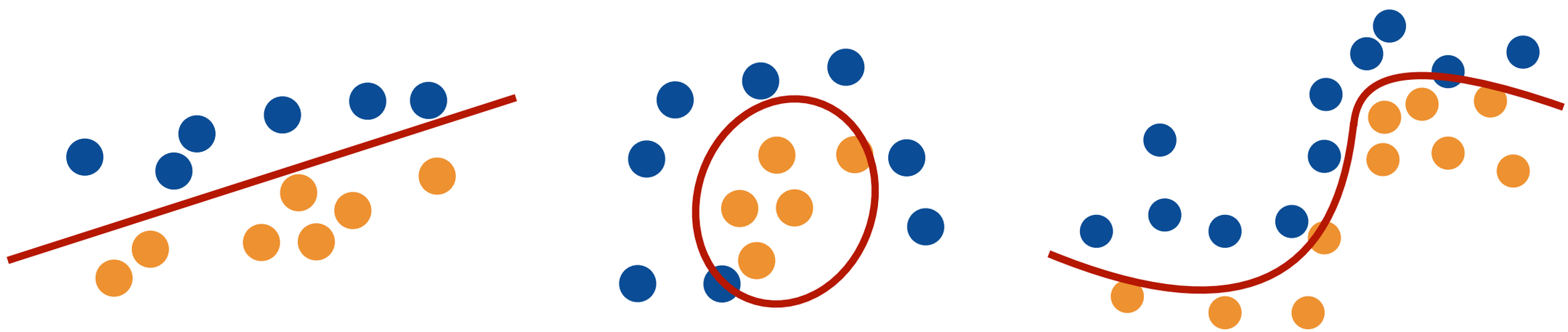
- Models patterns in data

Why ML so powerful?



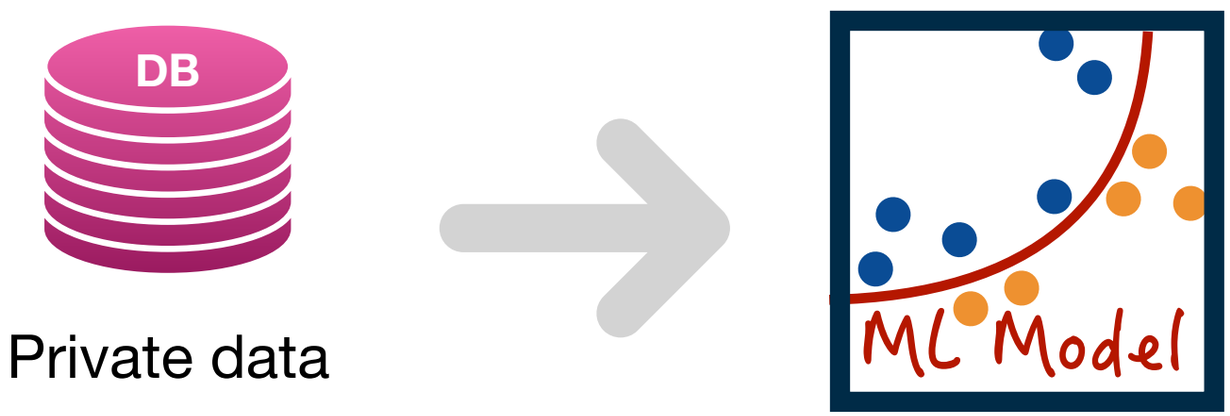
- Models patterns in data
- Allows automation, reducing costs

Why ML so powerful?



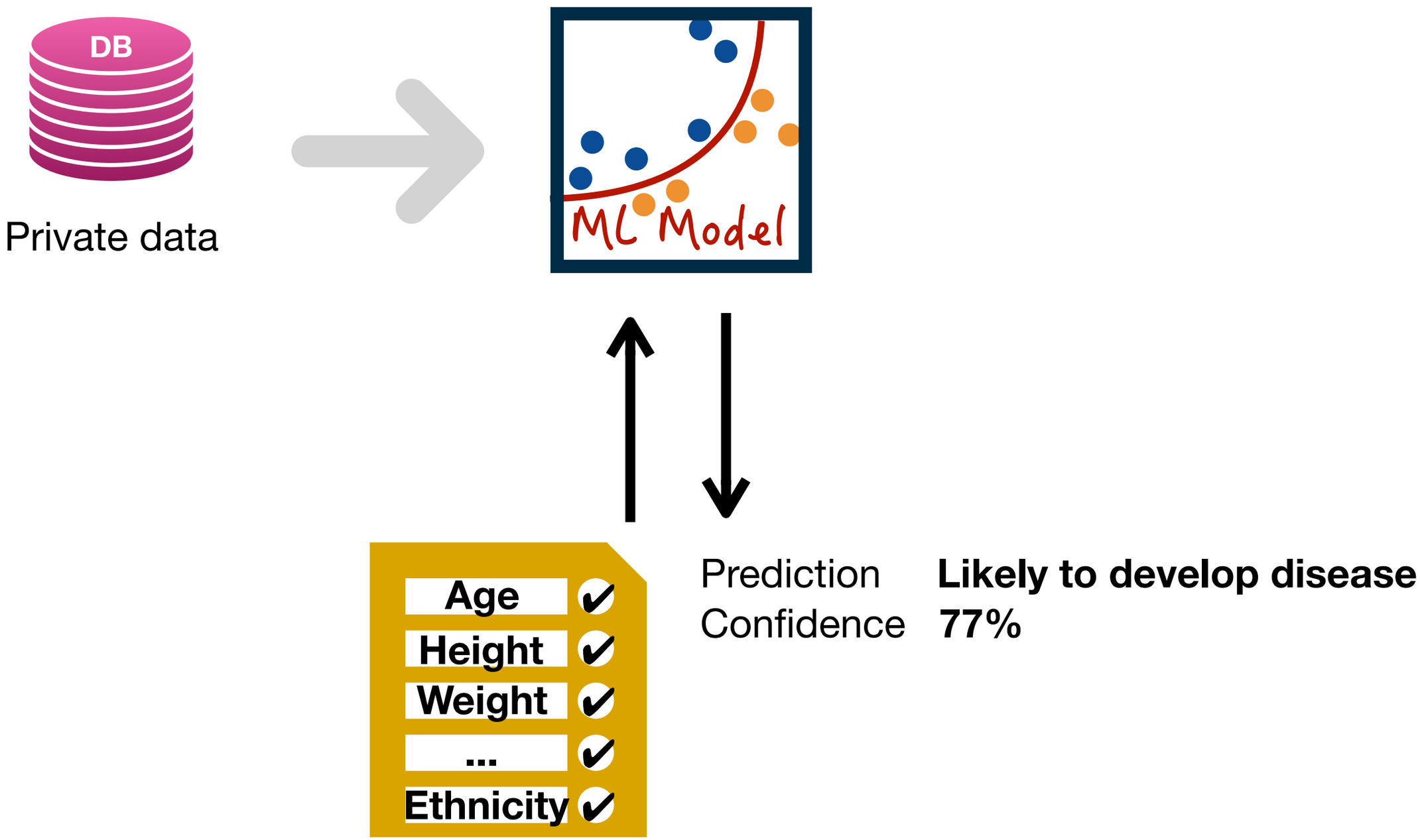
- Models patterns in data
- Allows automation, reducing costs
- Scales better than humans

Attacks' Intuition

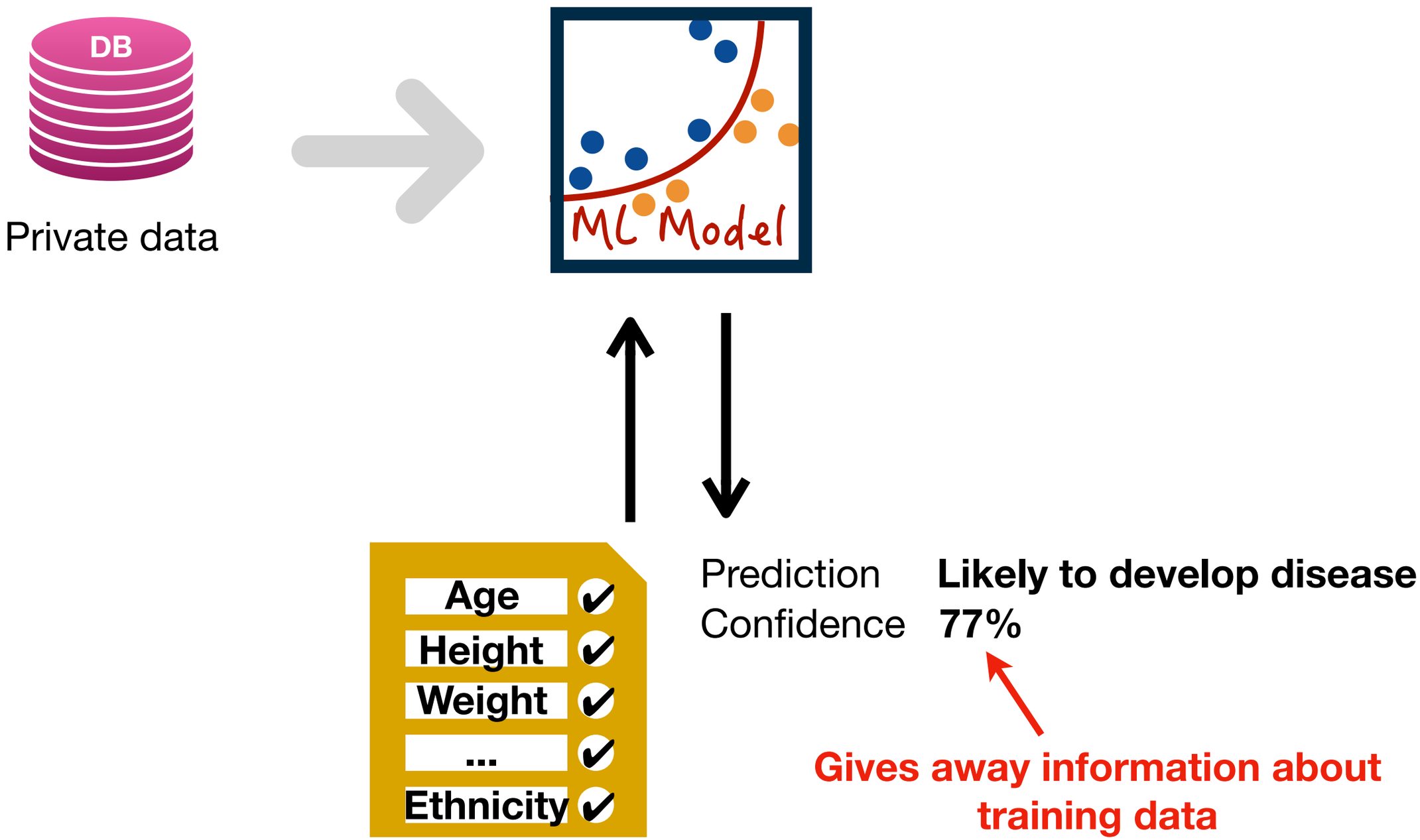


Attacks' Intuition

Attacks' Intuition

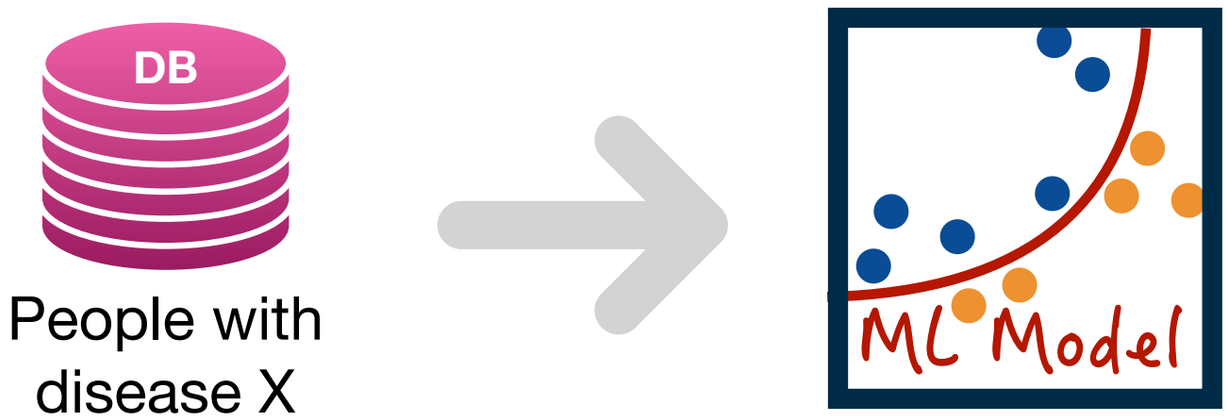


Attacks' Intuition



Membership Inference

[S+'18]

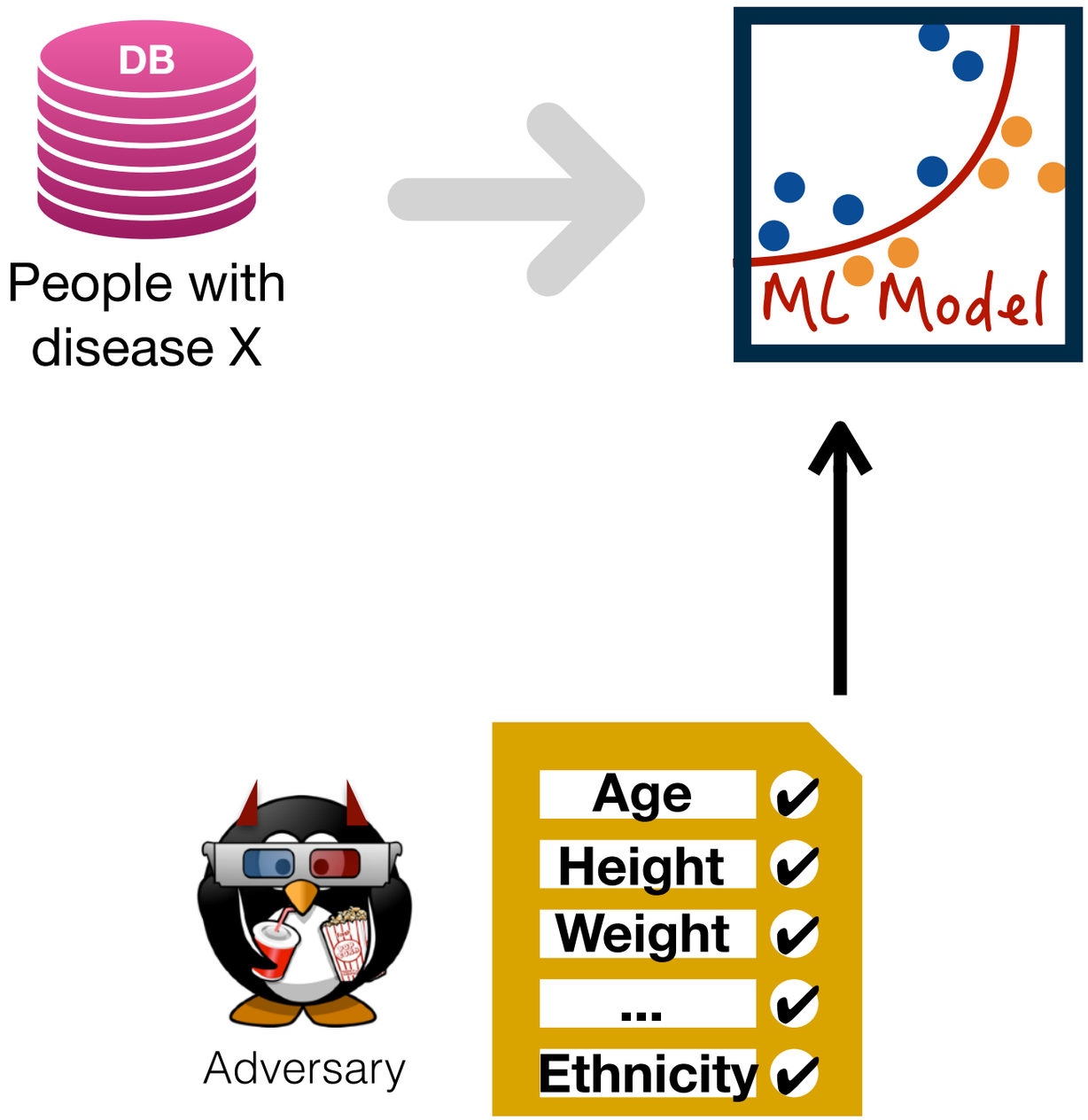


Membership Inference

[S+'18]

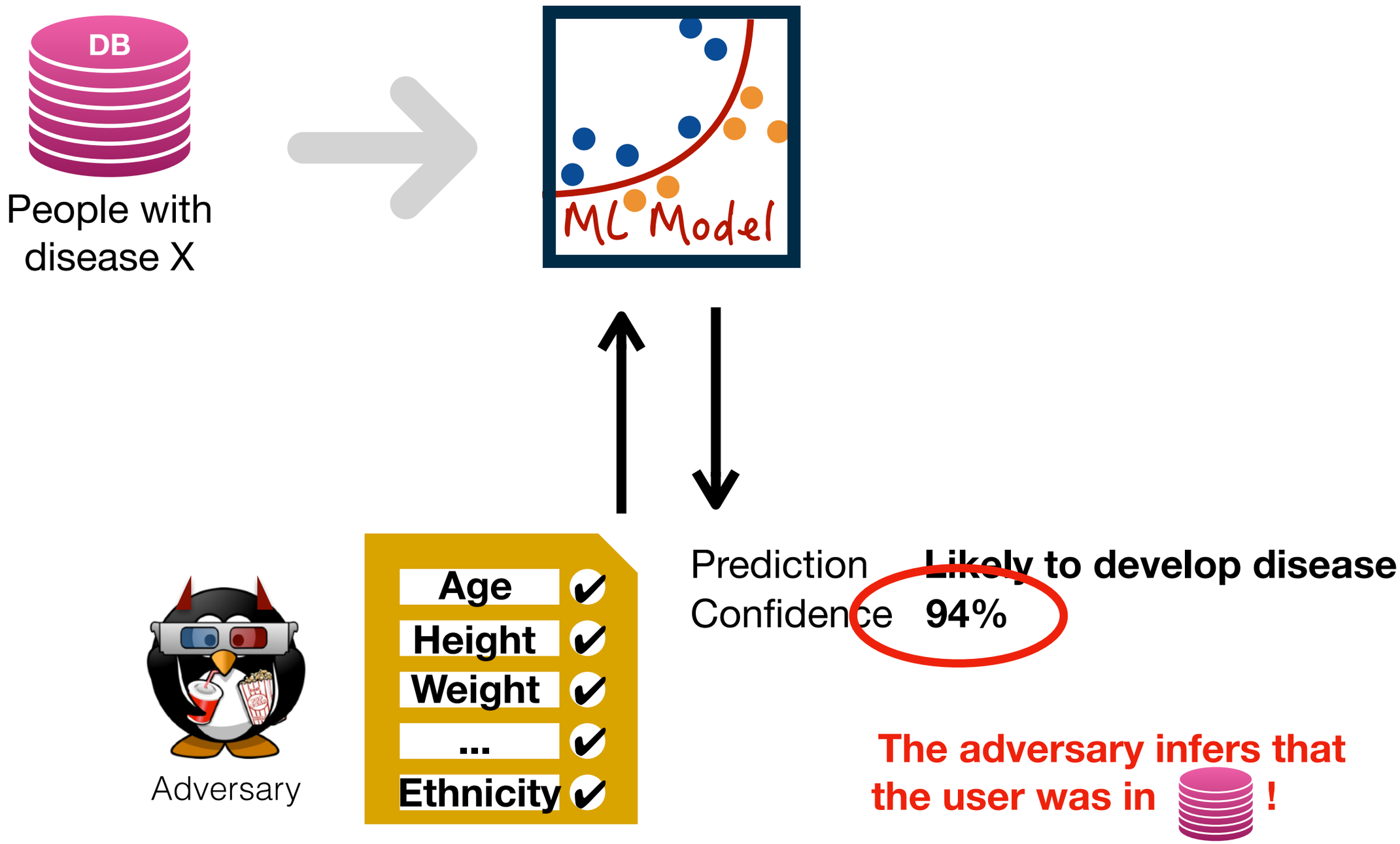
Membership Inference

[S+'18]



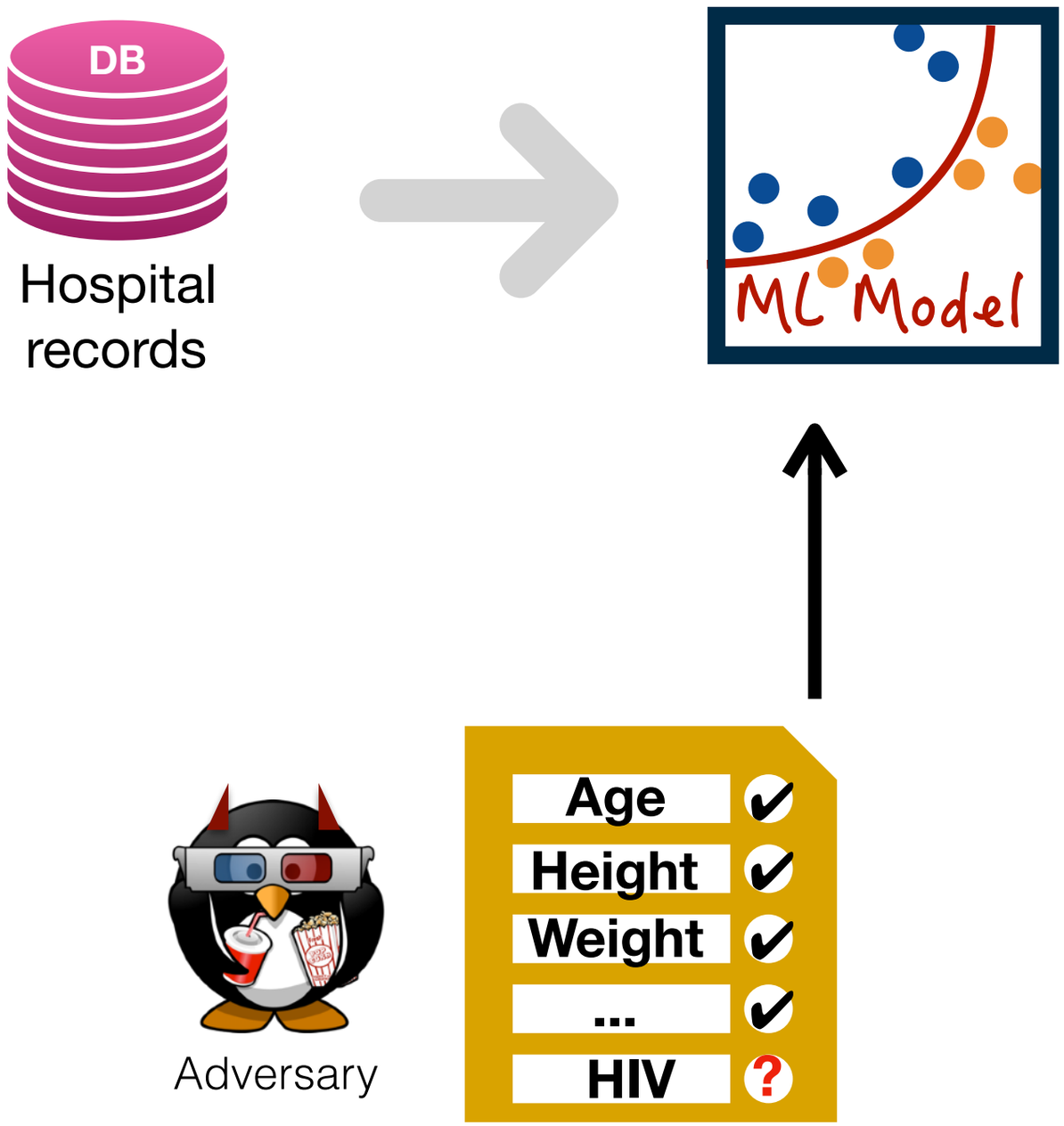
Membership Inference

[S+'18]



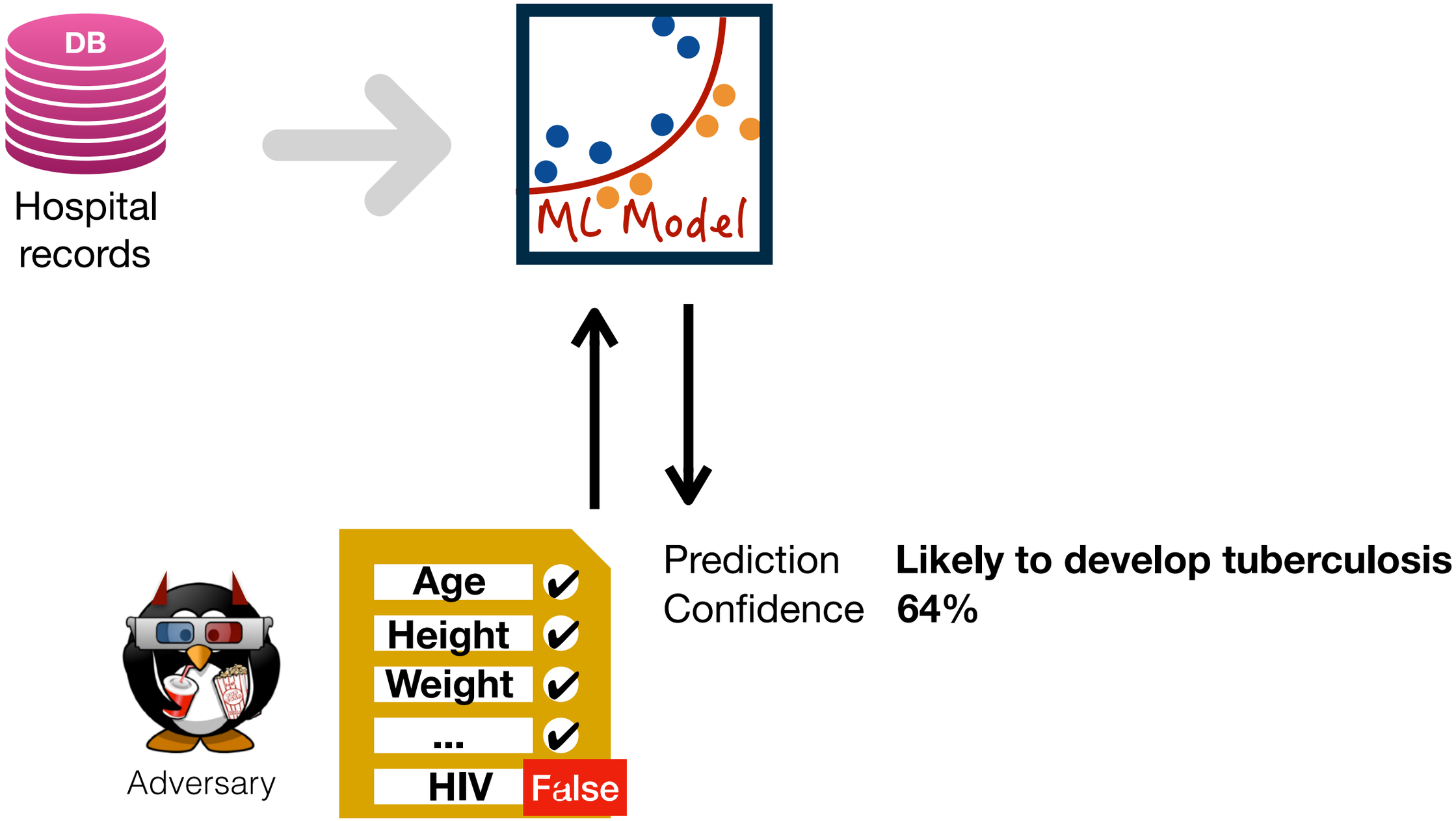
Model Inversion

[F+'15]



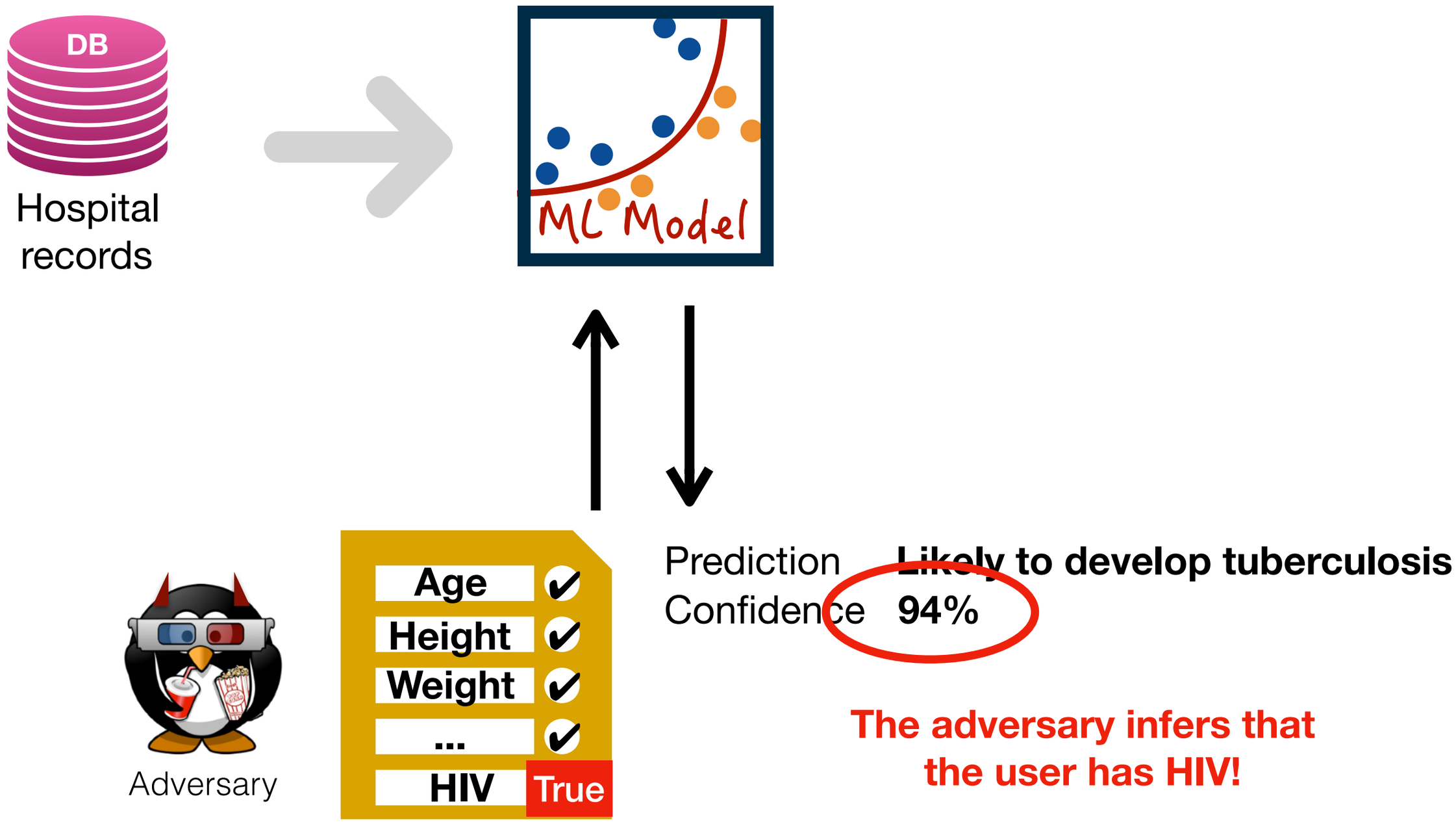
Model Inversion

[F+'15]



Model Inversion

[F+'15]



How to measure security?

Desiderata

How to measure security?

Desiderata

- Measure if/how much private data an ML model reveals **before** deployment

How to measure security?

Desiderata

- Measure if/how much private data an ML model reveals **before** deployment
- No need to understand what's **inside**

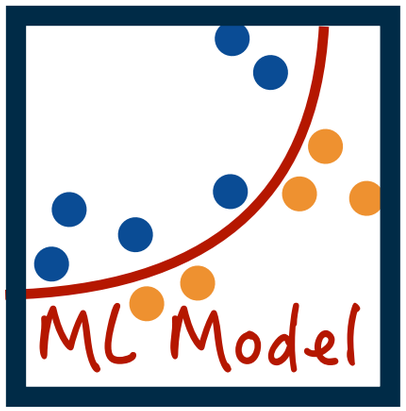
How to measure security?

Desiderata

- Measure if/how much private data an ML model reveals **before** deployment
- No need to understand what's **inside**
- Mathematical **guarantees**

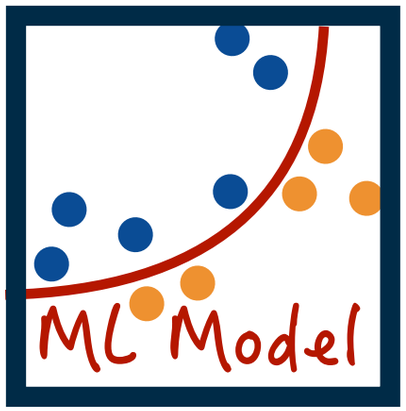
How to measure security?

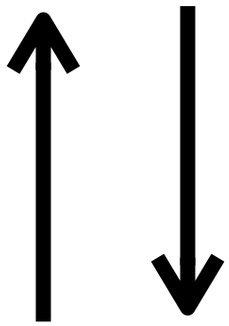
Black-box security approach [C'10, C'17, C+'19]



How to measure security?

Black-box security approach [C'10, C'17, C+'19]

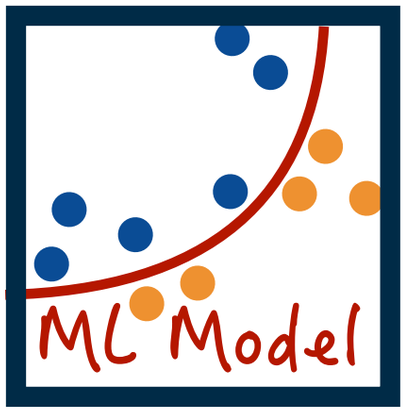




(, 93%)

How to measure security?

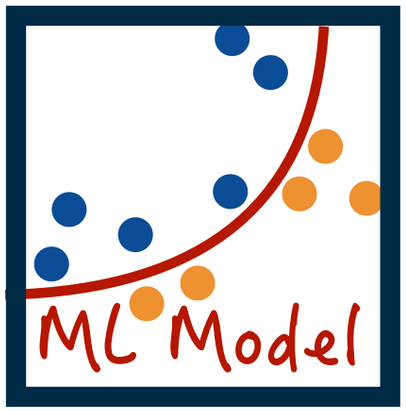
Black-box security approach [C'10, C'17, C+'19]

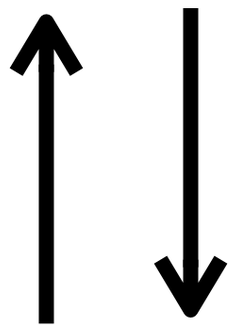


(, 93%), (, 70%)

How to measure security?

Black-box security approach [C'10, C'17, C+'19]

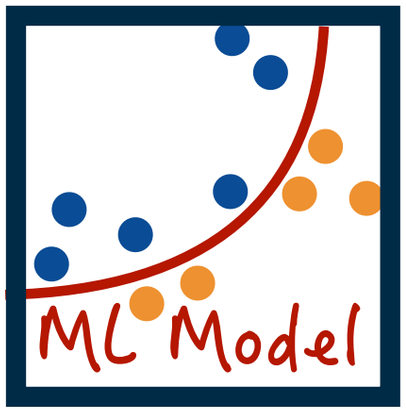




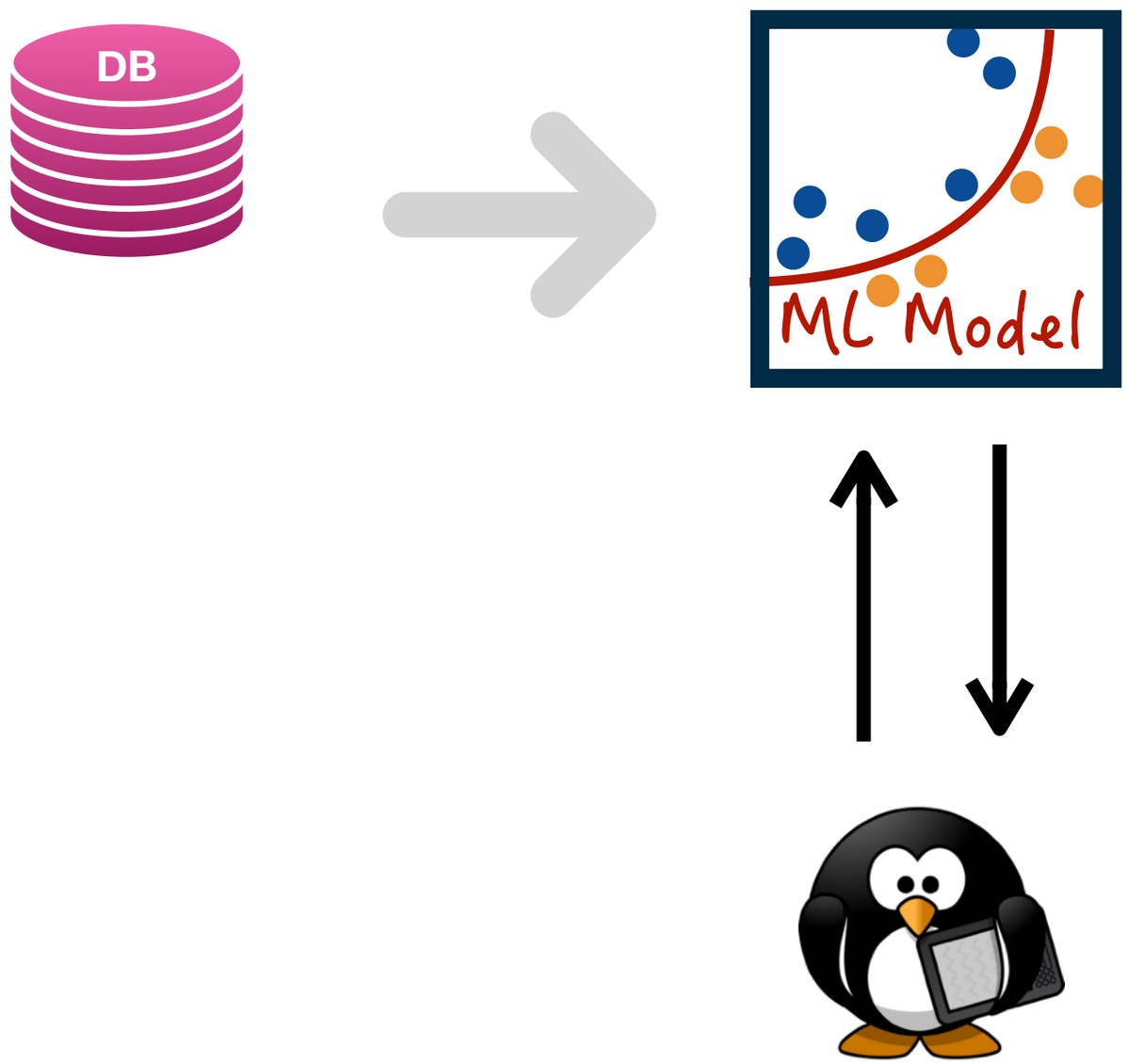
(, 93%), (, 70%), (, ...)

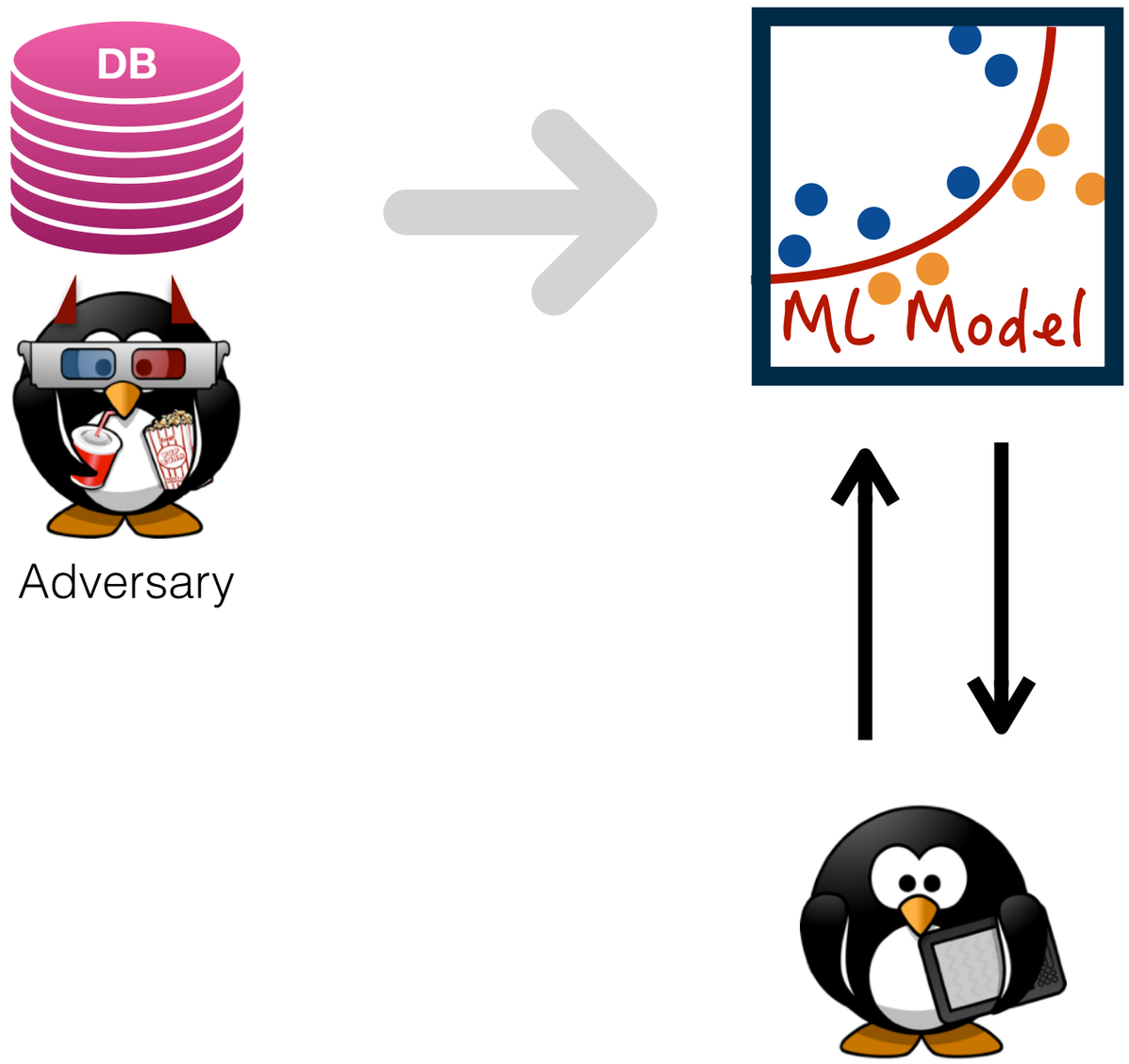
How to measure security?

Black-box security approach [C'10, C'17, C+'19]

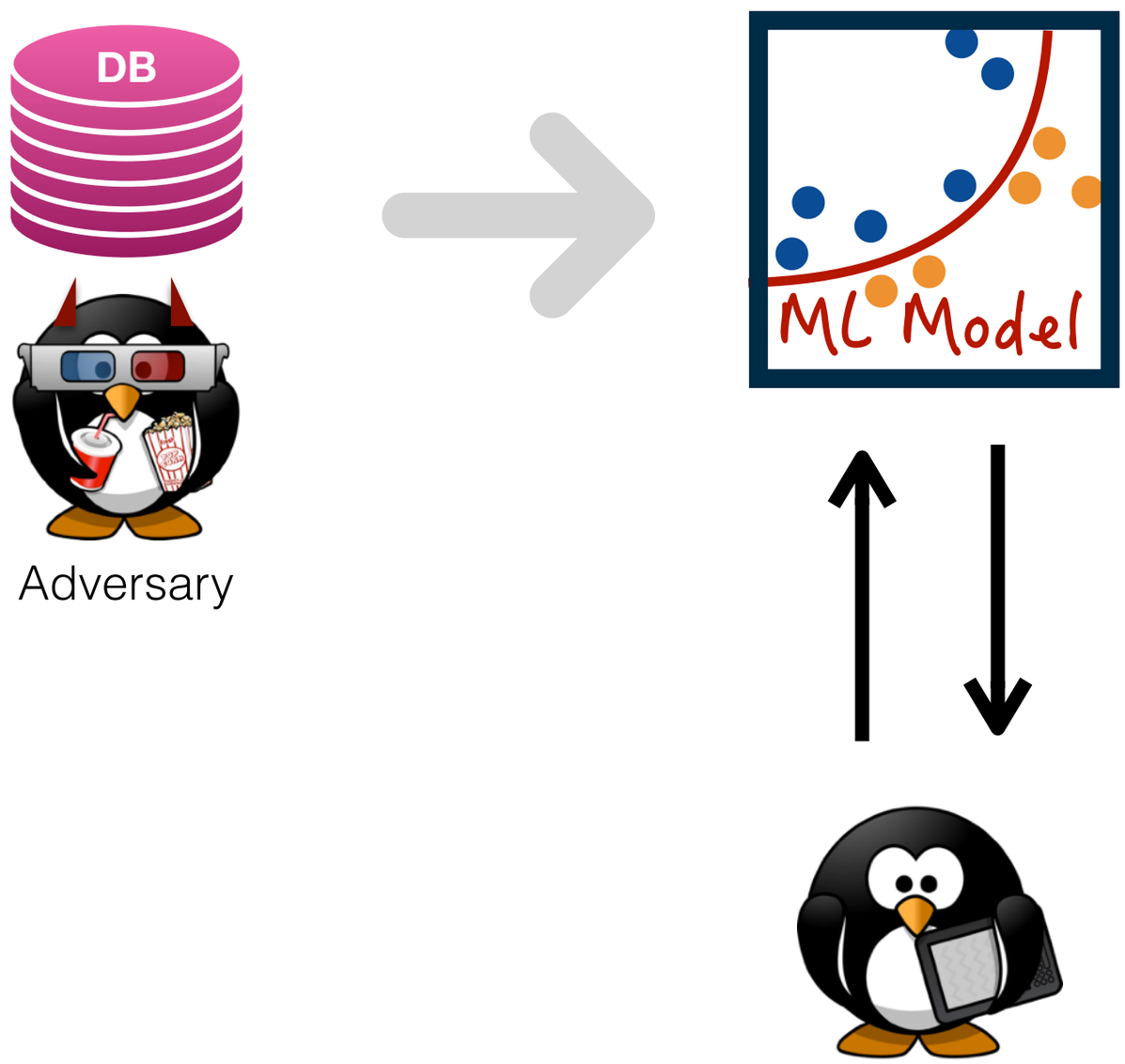


Basic idea: Estimate the probability of success of an optimal adversary from data


DB
ML Model


DB
Adversary
ML Model

Adversarial learning



TL;DL

- ML models may **reveal information** about their private training data
- Before releasing them, we need to measure how much information they leak
- **Black-box security** techniques allow this, without the need to understand how the models work
- There's more: **adversarial learning**

Background

https://www.who.int/dg/speeches/2018/artificial-intelligence-summit/en

"Model inversion attacks that exploit confidence information and basic countermeasures" (M. Fredrikson, S. Jha, T. Ristenpart, 2015)

"Membership inference attacks against machine learning models" (R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017)

Measuring security

"Statistical measurement of information leakage." (K. Chatzikokolakis, T. Chothia, G. Apratim, 2010)

"Bayes, not Naïve: Security Bounds on Website Fingerprinting Defenses" (G. Cherubin, 2017)

"F-BLEAU: Practical Channel Leakage Estimation" (G. Cherubin, K. Chatzikokolakis, C. Palamidessi, 2019) [Under submission]

More https://giocher.com/pages/bayes.html

Code https://github.com/gchers/fbleau

Measuring the Security of Machine Learning models

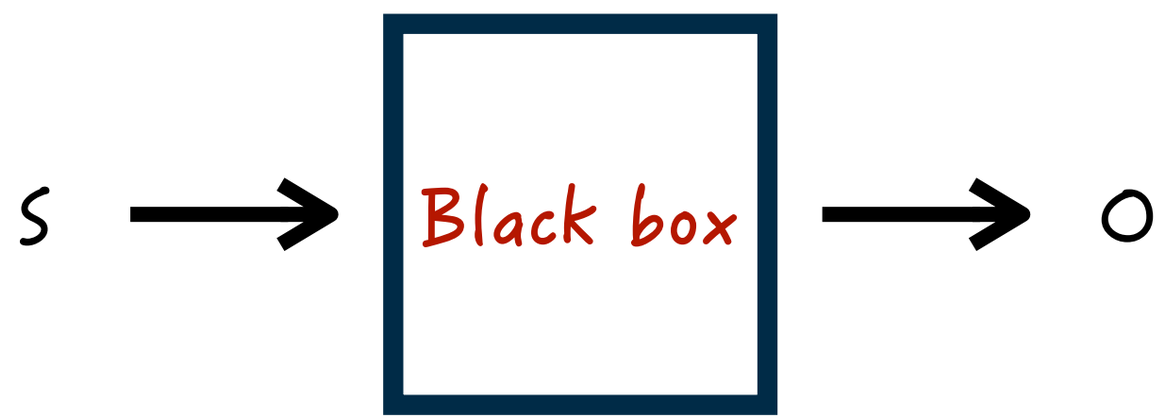
Giovanni Cherubin
@gchers

ITU/WHO Workshop on AI for Health
22 January, 2019

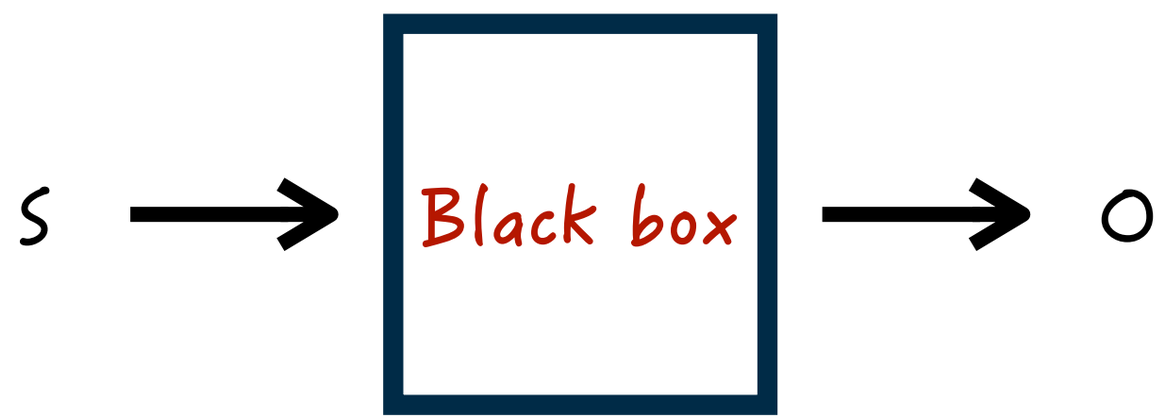
SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Black-box security



Black-box security



Estimate:

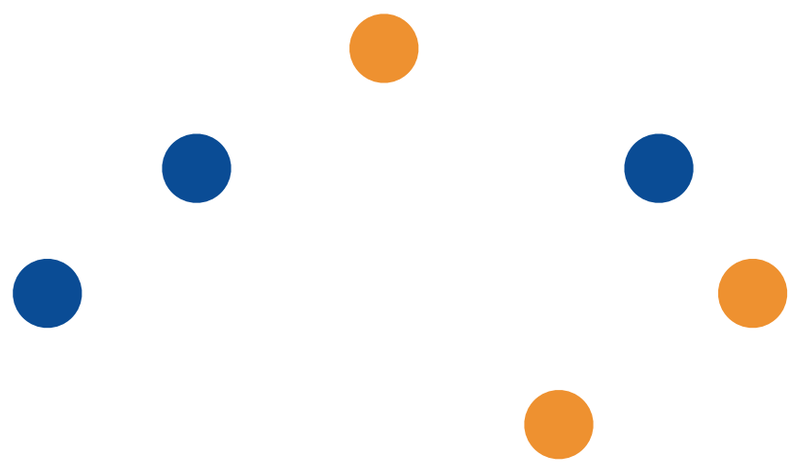
$$\Pr(s \neq \hat{s})$$

where $\hat{s}$ is the prediction of optimal adversary (Bayes adversary)

Estimates

[CH'67]

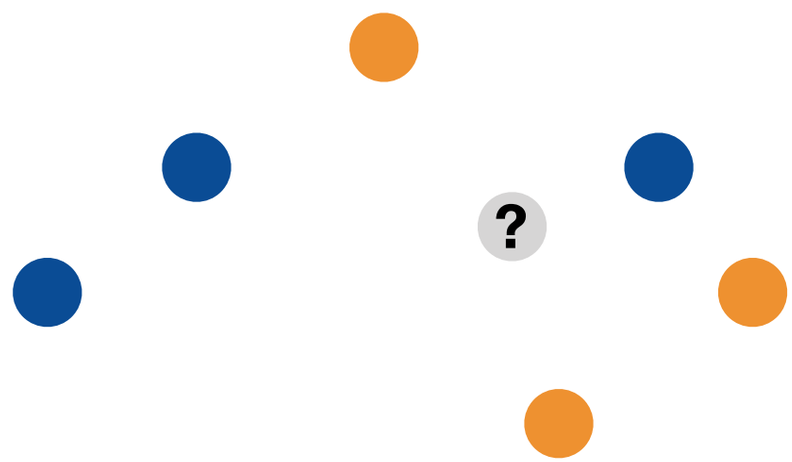
NN Bound



Estimates

[CH'67]

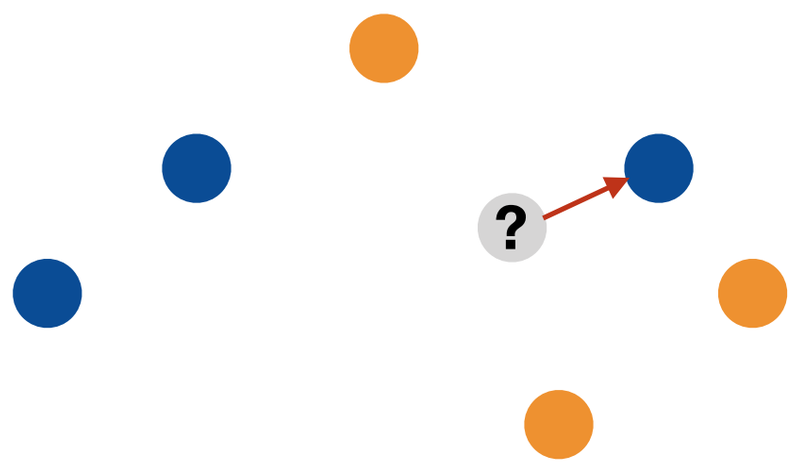
NN Bound



Estimates

[CH'67]

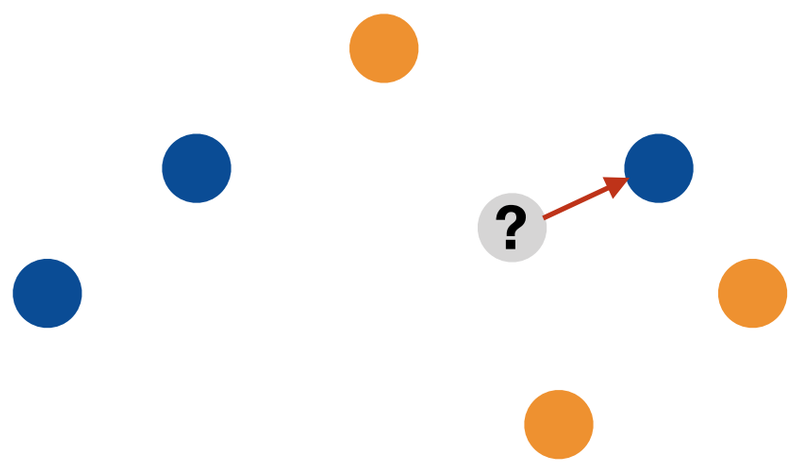
NN Bound



Estimates

[CH'67]

NN Bound

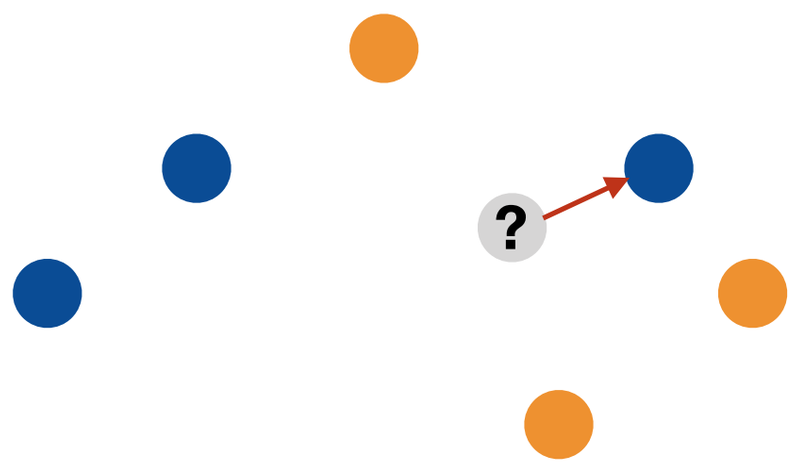


$R^* \leq R$

Estimates

[CH'67]

NN Bound



$$\frac{L-1}{L}\left(1-\sqrt{1-\frac{L}{L-1}R^{NN}}\right) \leq R^{*} \leq R$$

Measuring the Security of Machine Learning models

Giovanni Cherubin
@gchers

ITU/WHO Workshop on AI for Health
22 January, 2019

SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE