



Towards Automated and Trustworthy Machine Learning

Minhao Cheng

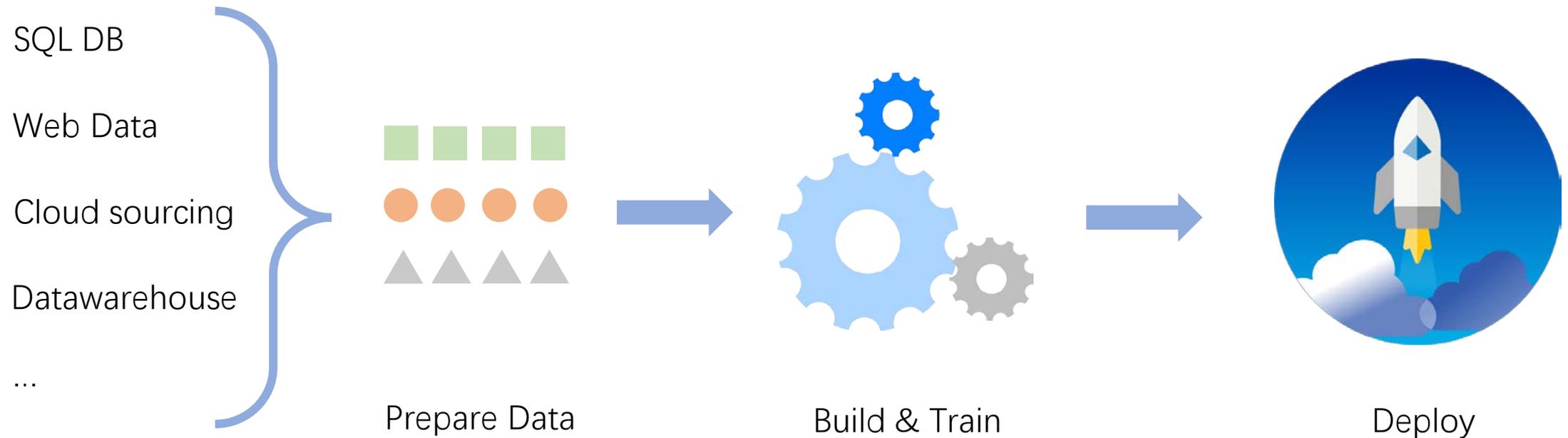
Assistant Professor, CSE @ HKUST

minhaocheng@ust.hk



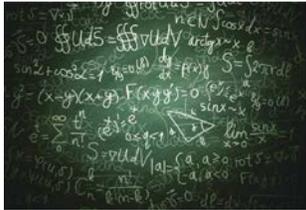
THE DEPARTMENT OF
**COMPUTER SCIENCE
& ENGINEERING**
計算機科學及工程學系

Machine Learning Pipeline



The devil is in the details

- What feature
Constraint/Rule



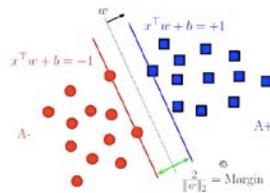
Budget



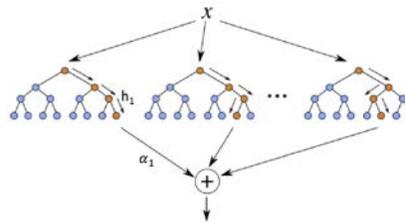
Efficiency



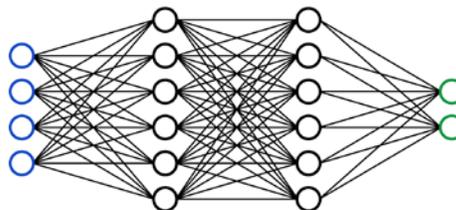
- Which model
Linear model



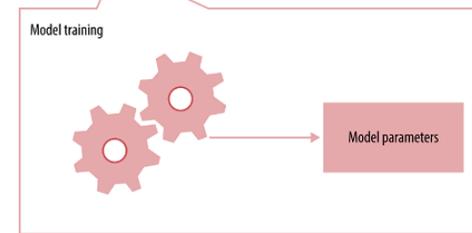
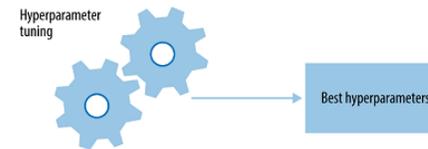
Boosting model



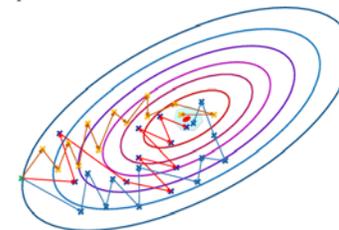
Neural network



- Which parameter
Hyperparameter

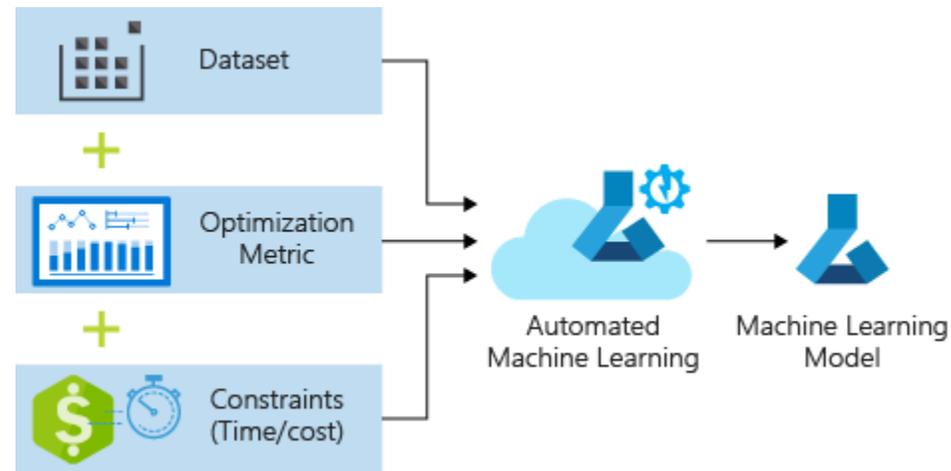


Optimizer

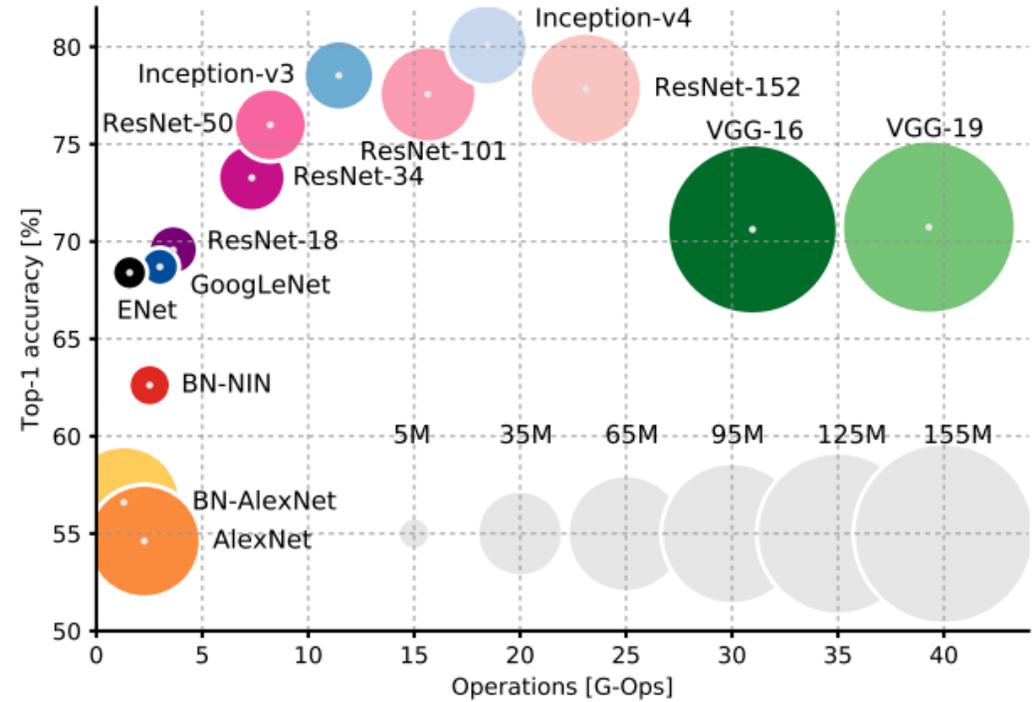
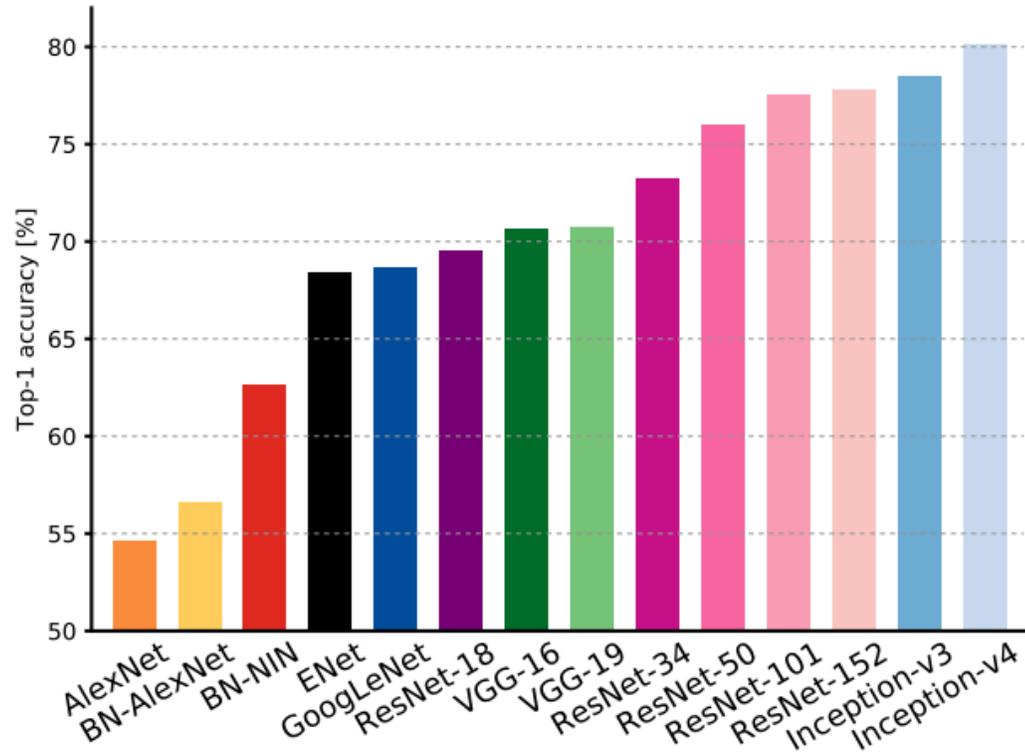


Automated Machine Learning

- **AutoML** simplifies each step in the machine learning process, from handling a raw dataset to deploying a practical machine learning model.

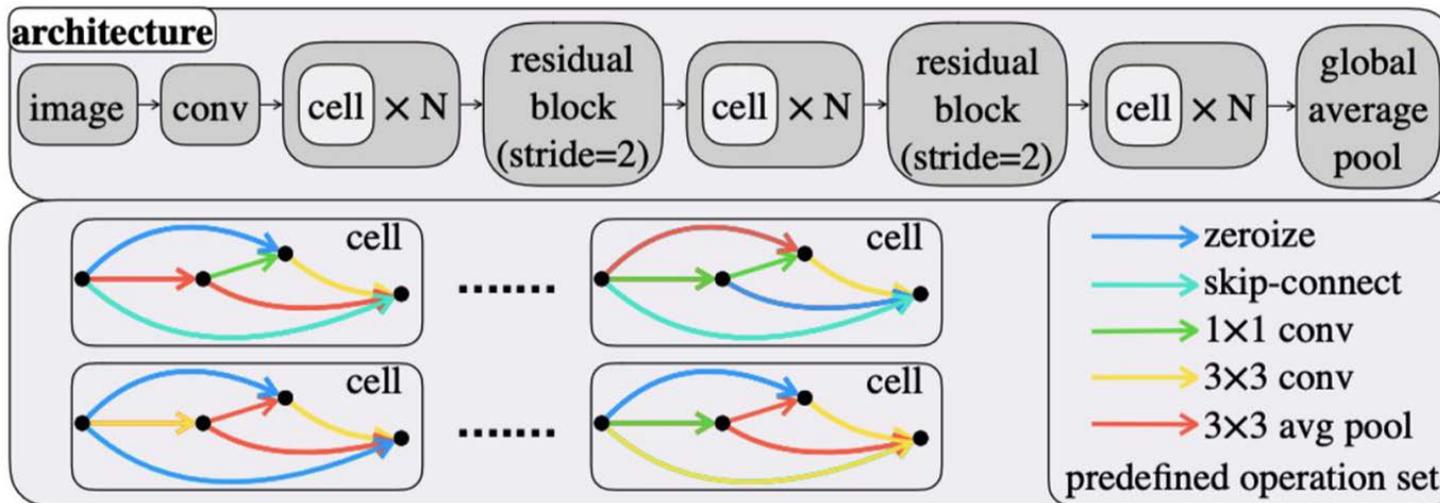


Model Matters!



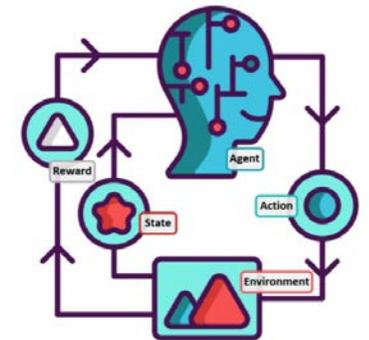
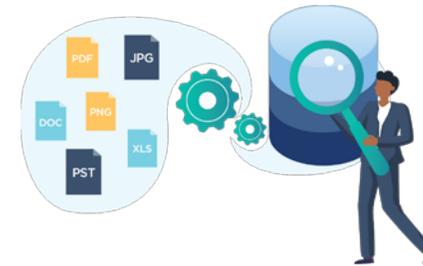
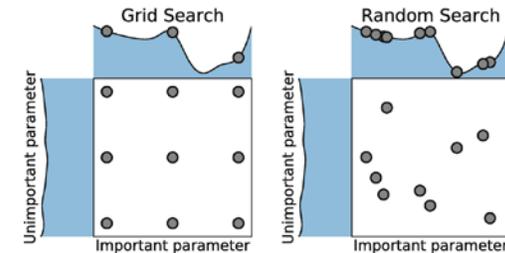
Automated Machine Learning

- AutoML: simplifies each step in the machine learning process, from handling a raw dataset to deploying a practical machine learning model.
 - Neural Architecture Search (NAS) [ICLR 21'ECCV 21']

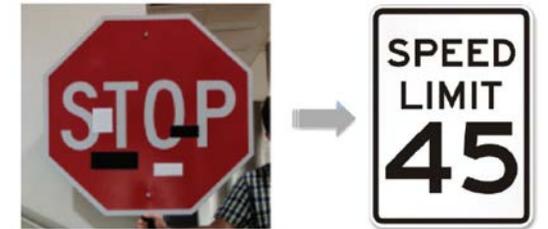
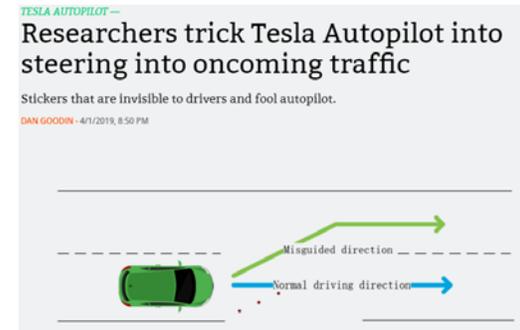


Automated Machine Learning

- AutoML:
 - Neural Architecture Search (NAS)
 - Hyperparameter optimization (HPO)
 - Meta learning and Learning to learn
 - Automated Reinforcement learning
 - AutoML in Physical World
 - ...



Beyond Accuracy



Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahperez / 10:16 am EDT - March 24, 2018

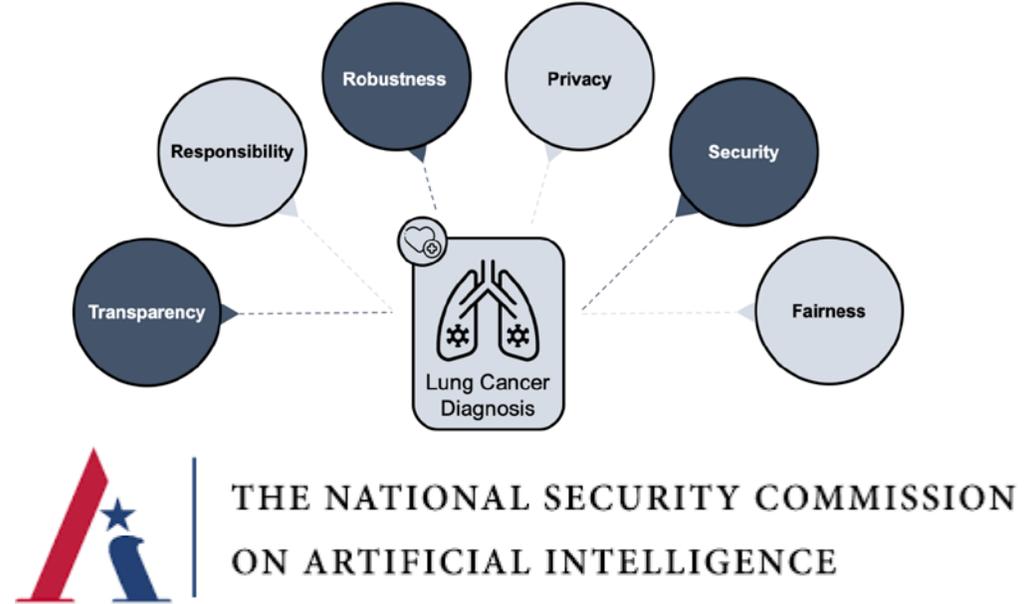


Microsoft's newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn't coded to be racist, but it "learns" from those it interacts with. And naturally, given that this is the internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [Update: Microsoft now says it's "making adjustments" to Tay in light of this problem.]

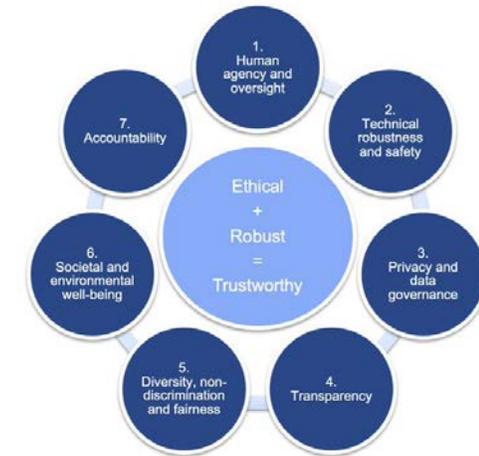


Trustworthy ML

- Not alchemy
 - Explainability
 - **Robustness**
 - Security
 - Privacy
 - Fairness
 - ...
- Establish model understanding



人工智能安全测评白皮书
(2021)



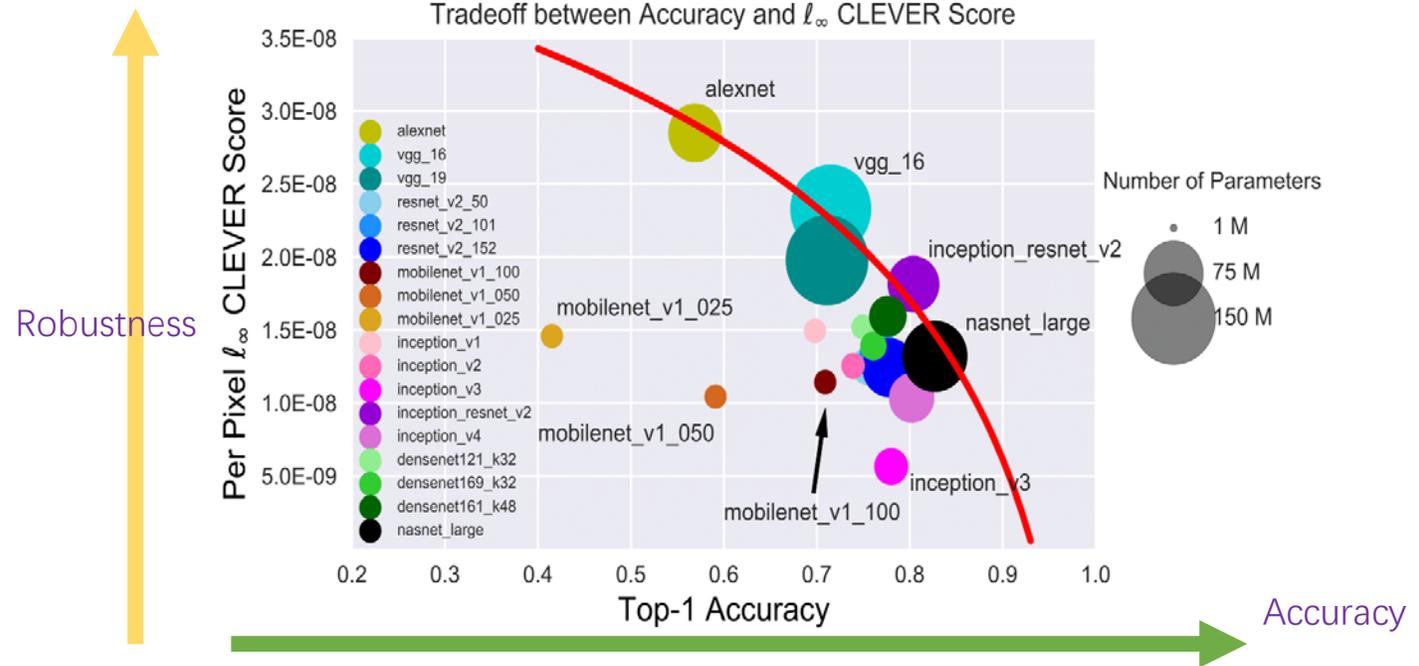
国家语音及图像识别产品质量监督检验中心
国家工业信息安全发展研究中心人工智能所
2021年10月



THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Accuracy \neq Robustness

- Solely pursuing for high-accuracy AI model may get us in trouble...
- We have established a toolbox to evaluate the robustness of machine learning models [ICRL 19'20' AAI 19'20']



Model understanding

- Debugging
- Bias detection
- Provide recourse to individuals who are adversely affected by model predictions
- Assess if and when to trust model predictions



Thank you!



THE DEPARTMENT OF
**COMPUTER SCIENCE
& ENGINEERING**
計算機科學及工程學系