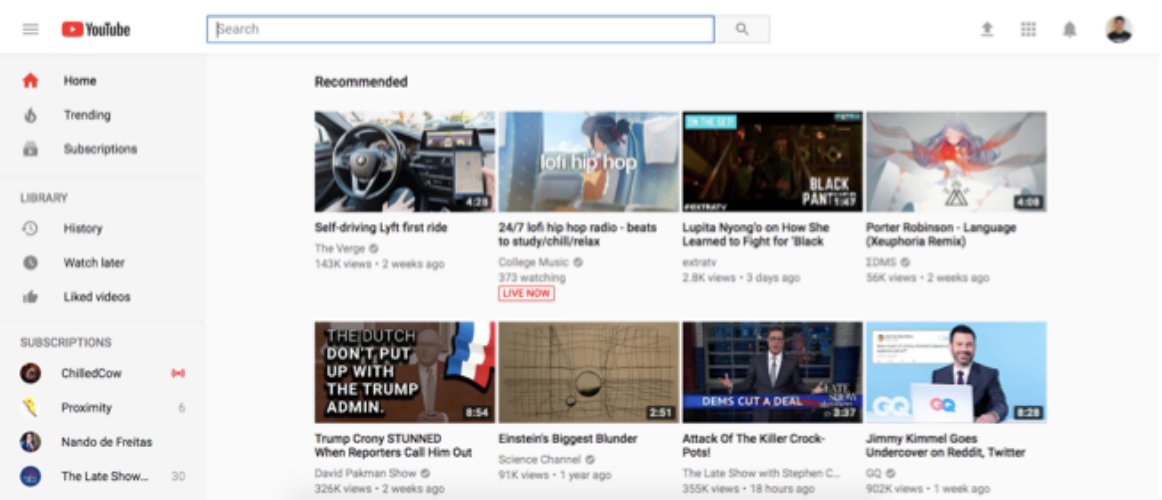
Introduction to Secure Machine Learning

BOSTON UNIVERSITY
MACHINE INTELLIGENCE
COMMUNITY

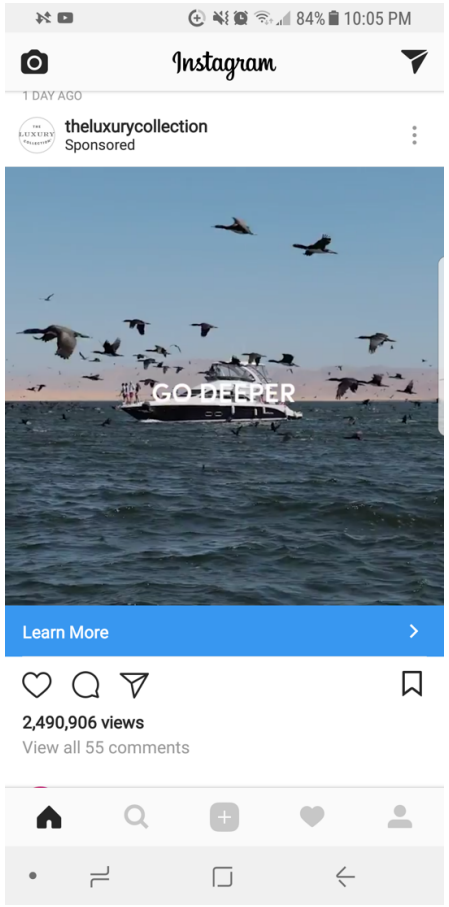
Justin Chen
Jan. 29, 2018

Age of Machine Intelligence

Society, Data, and Privacy

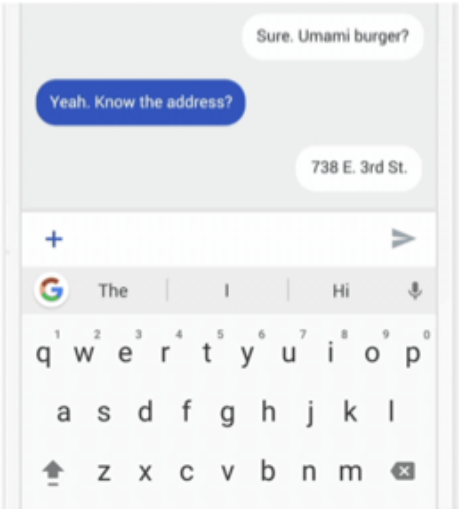

YouTube
Search
Home
Trending
Subscriptions
LIBRARY
History
Watch later
Liked videos
SUBSCRIPTIONS
ChilledCow
Proximity
Nando de Freitas
The Late Show...
Recommended
Self-driving Lyft first ride
The Verge
143K views • 2 weeks ago
24/7 lofi hip hop radio - beats to study/chill/relax
College Music
373 watching
LIVE NOW
Lupita Nyong'o on How She Learned to Fight for 'Black
extratv
2.8K views • 3 days ago
Porter Robinson - Language (Xeuphoria Remix)
IDMS
56K views • 2 weeks ago
Trump Crony STUNNED When Reporters Call Him Out
David Pakman Show
326K views • 2 weeks ago
Einstein's Biggest Blunder
Science Channel
91K views • 1 year ago
Attack Of The Killer Crock-Pots!
The Late Show with Stephen C...
355K views • 18 hours ago
Jimmy Kimmel Goes Undercover on Reddit, Twitter
GQ
902K views • 1 week ago


New for you
See more

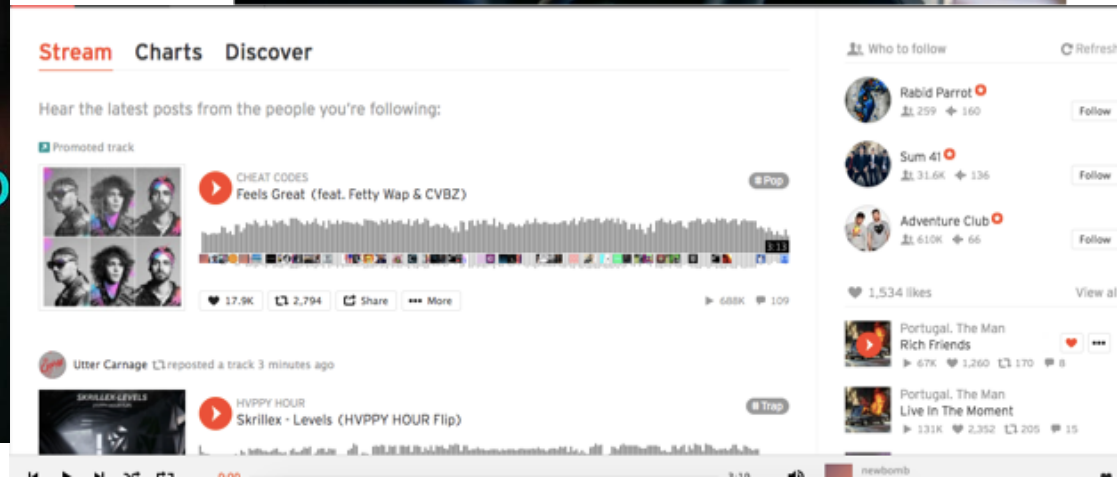

84% 10:05 PM
Instagram
1 DAY AGO
theluxurycollection
Sponsored
GO DEEPER
Learn More
2,490,906 views
View all 55 comments


WAYMO 01


75

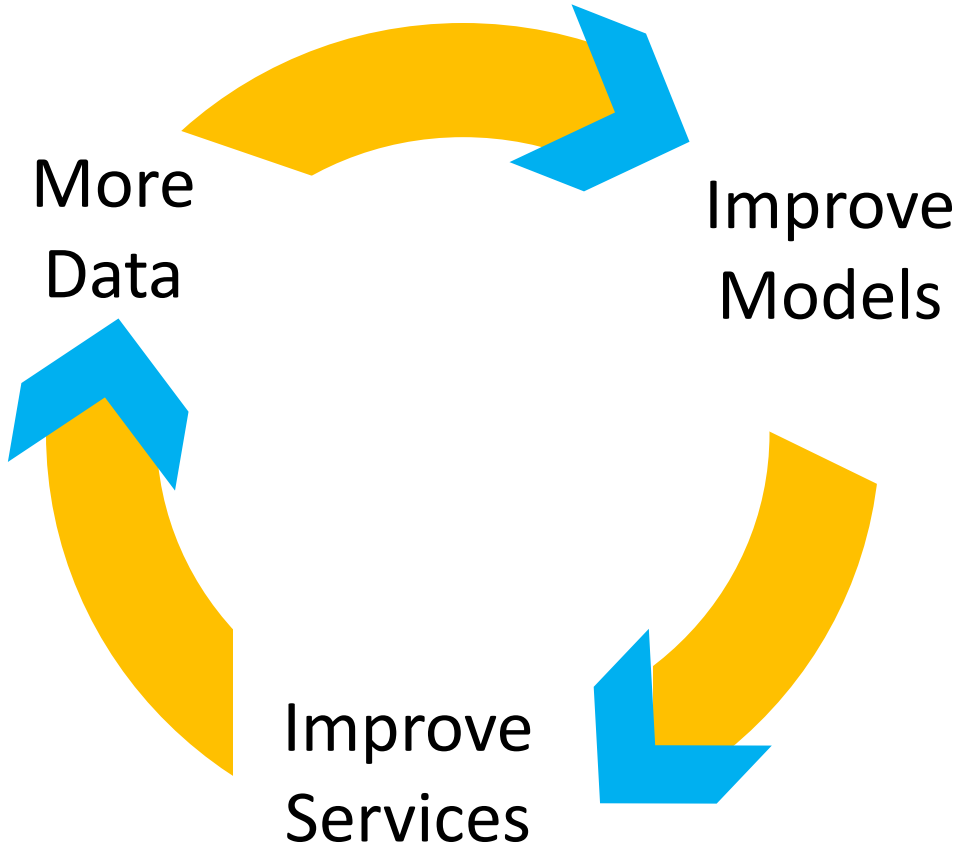

Sure. Umami burger?
Yeah. Know the address?
738 E. 3rd St.
The
I
Hi


amazon echo


Stream
Charts
Discover
Hear the latest posts from the people you're following:
Who to follow
Rabid Parrot
Sum 41
Adventure Club
Portugal. The Man
Rich Friends
Portugal. The Man
Live In The Moment

Virtuous Cycle of Machine Learning

- Term coined by Andrew Ng
- Collect data to improve machine learning models
- Use improved models to improve services
- Use improved services to collect more data



AI on the Blockchain

OpenMined

An open-source community focused on building technology to facilitate the decentralized ownership of data and intelligence.

https://openmined.org/
https://openmined.slack.com
https://github.com/OpenMined

Synapse AI

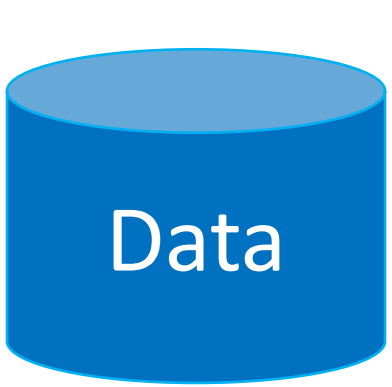
Decentralized Data and AI Marketplace

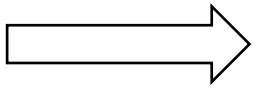
https://synapse.ai/

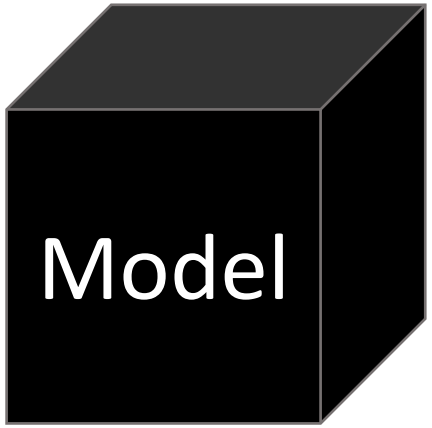
Learning from Data

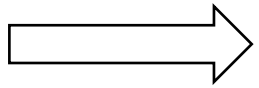
Training, Validation, Testing

Function Approximation









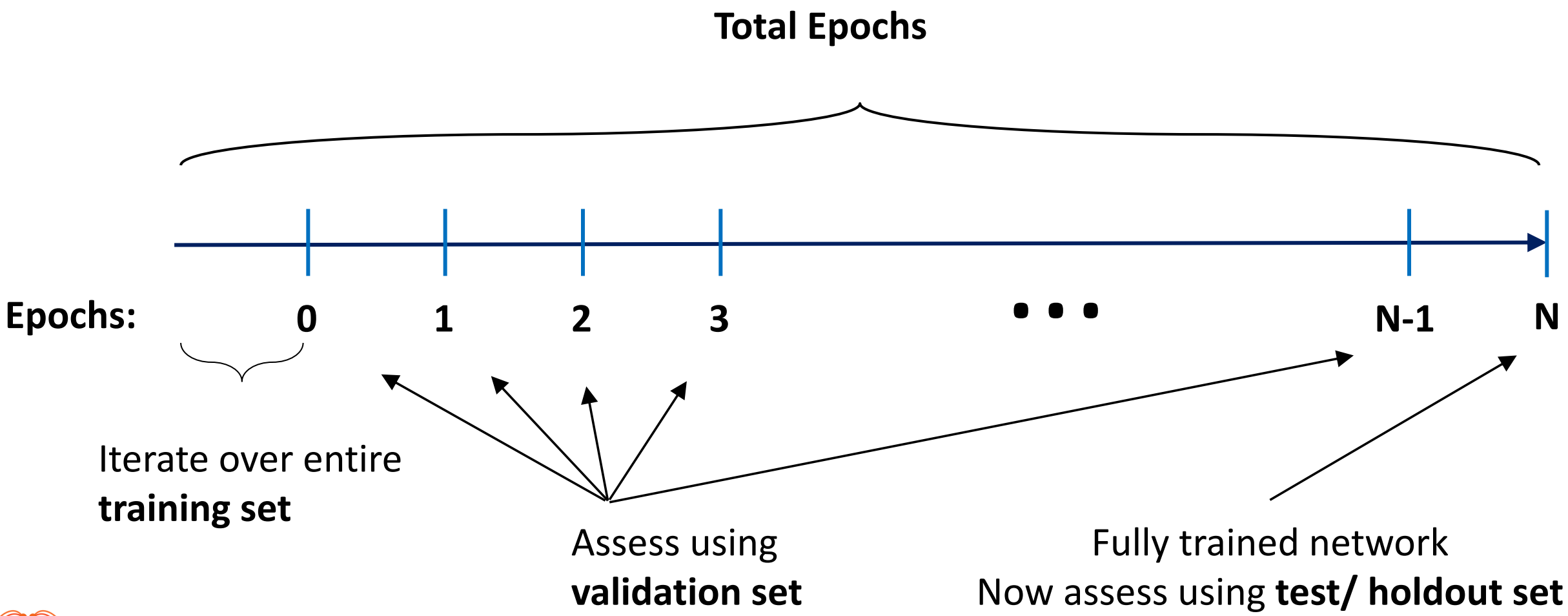
$$h(x)$$

Images
Audio
Video
Text
Logs
Spreadsheets

Linear Regression
Polynomial Regression
Logistic Regression
Multinomial Regression
Neural Network
Gaussian Process
…

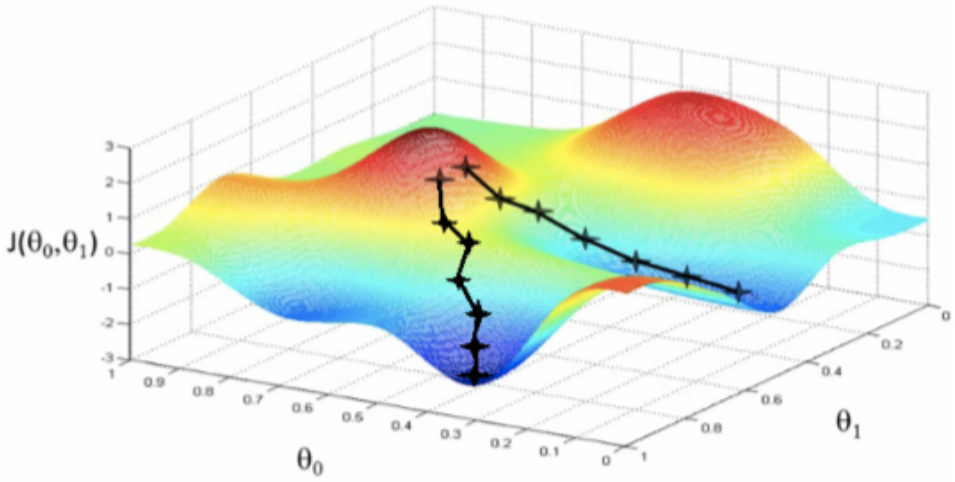
Class of objects
Real-value
Generated Data
Action

Training, Validation, Testing

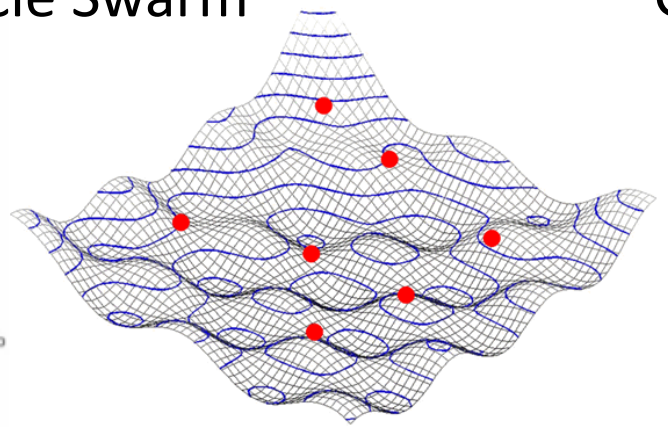


Training

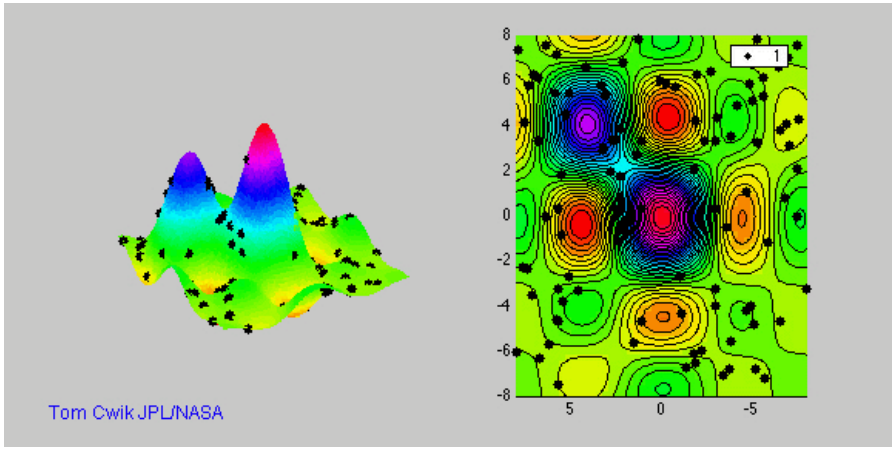
Gradient-based



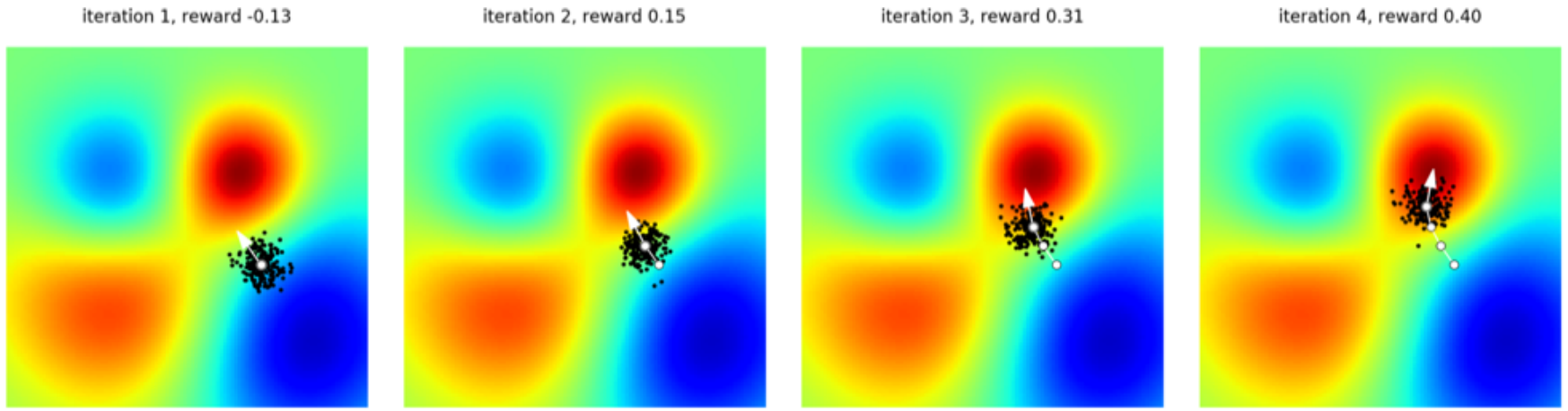
Particle Swarm



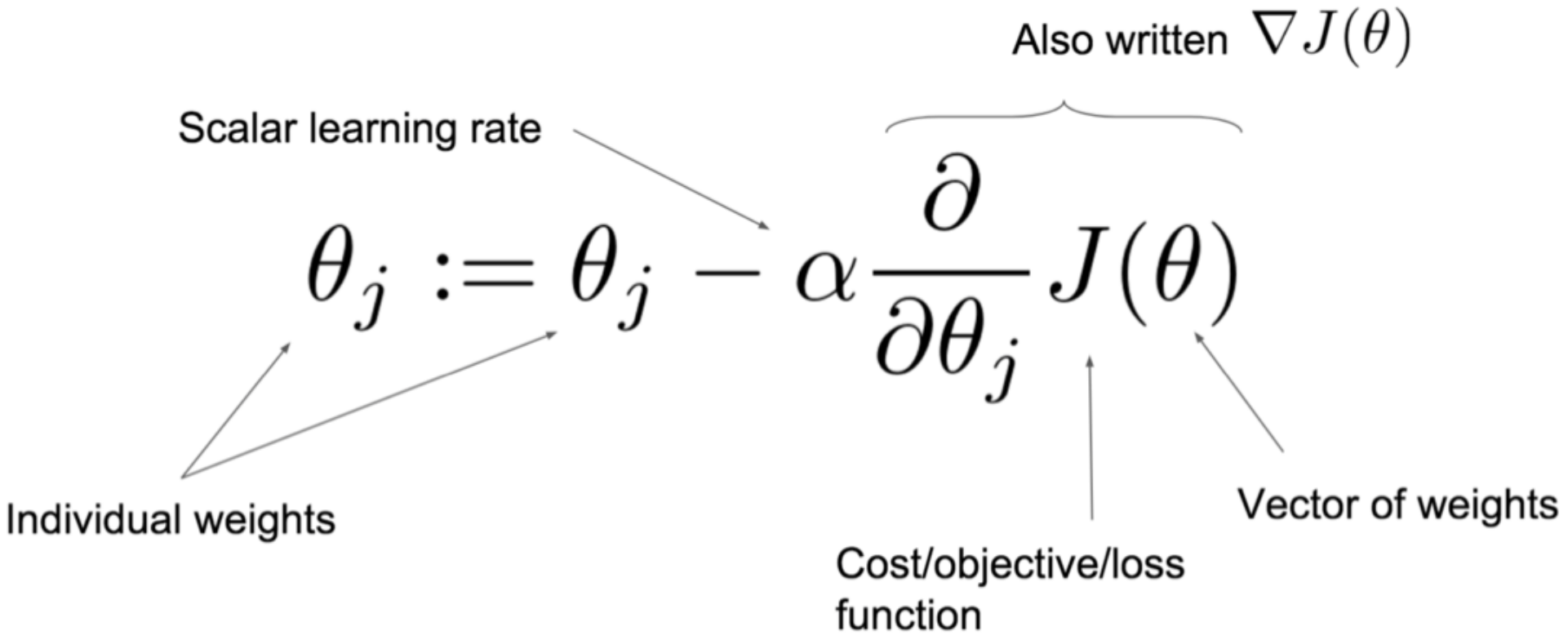
Genetics



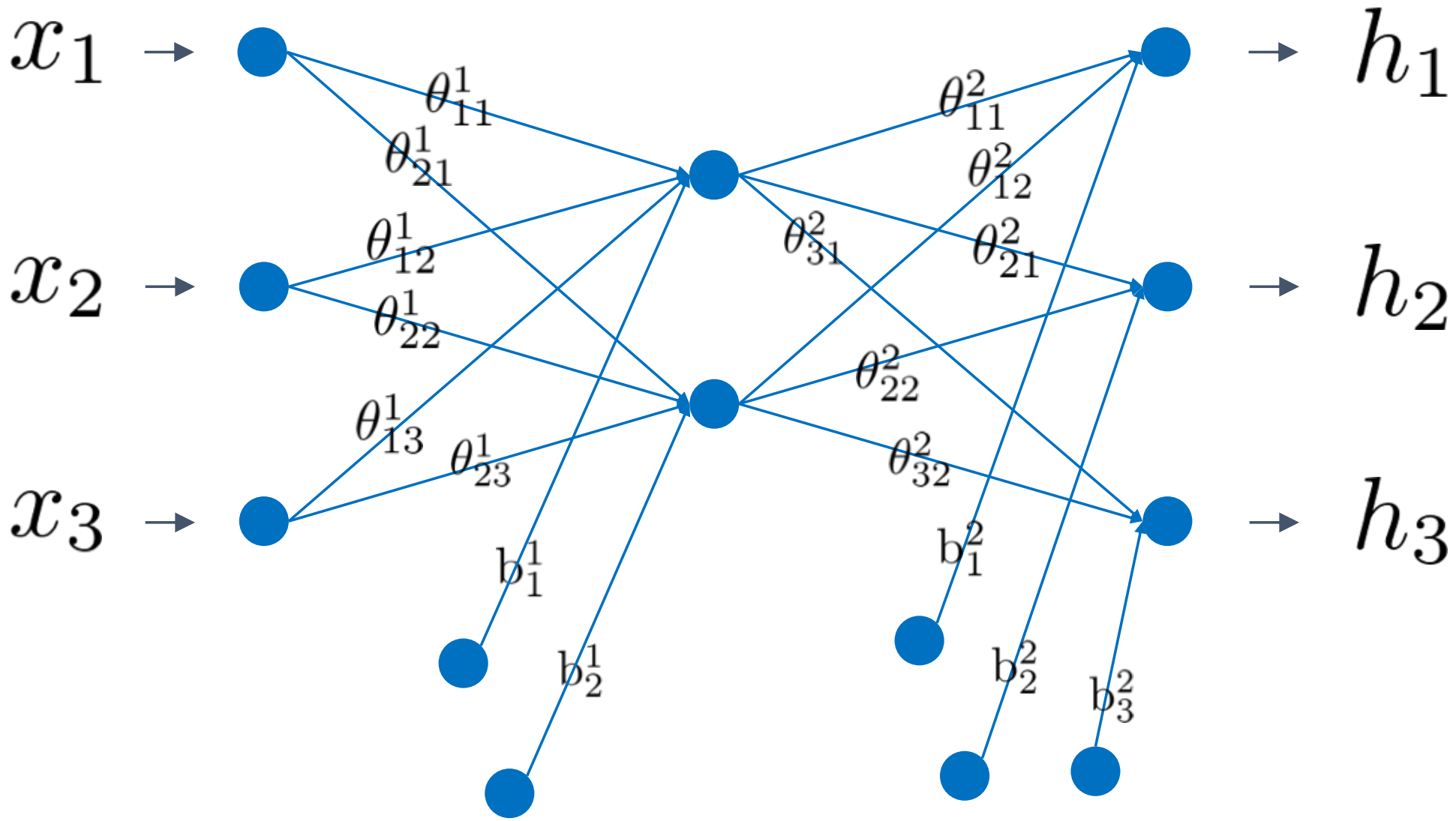
Evolutionary Strategies



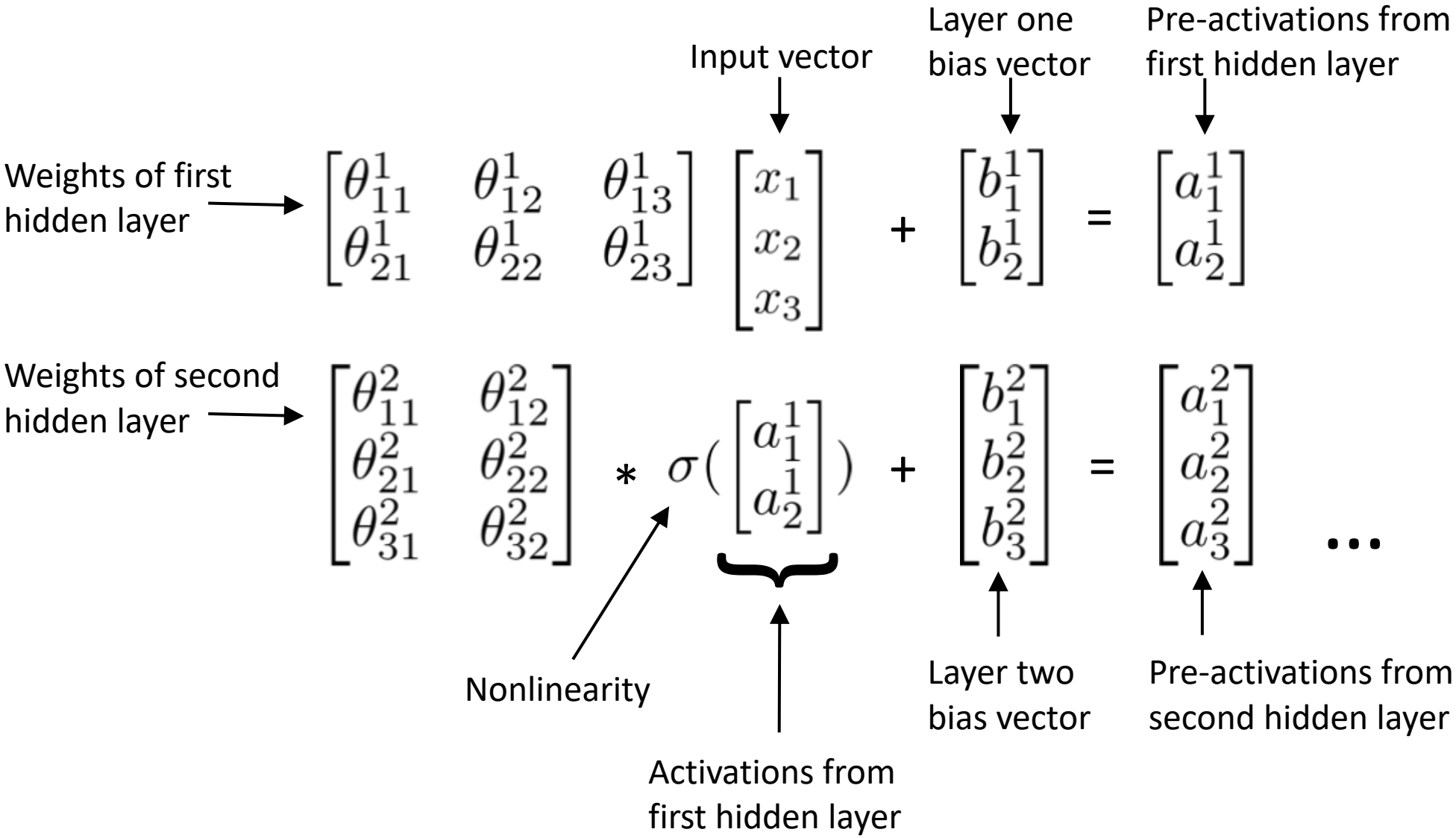
Gradient Descent



Artificial Neural Network



Matrix Representation



On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches

Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, and Li Zhang

Google

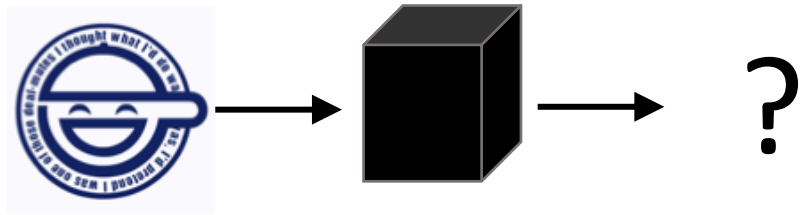
Attack Surface

Adversary's Abilities

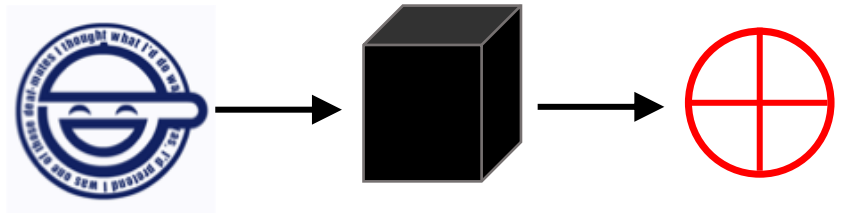
White-box Training	Black-box Training
White-box Testing	Black-box Testing

Types of Attacks

Untargeted



Targeted

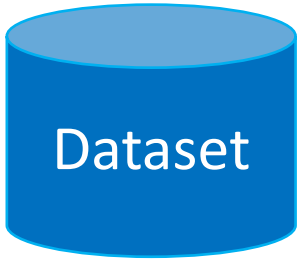


Economy of Mechanism

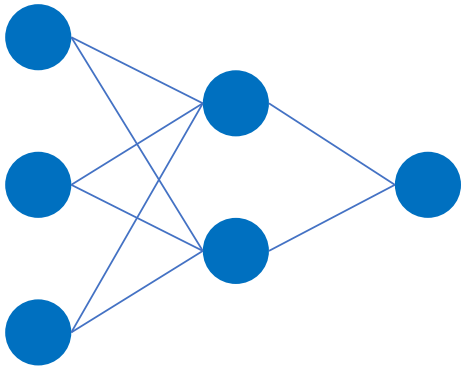
Protection mechanisms should be made as simple as possible

Machine learning algorithms have many moving parts

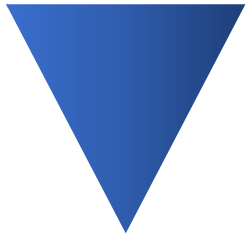
How much information does each component leak?



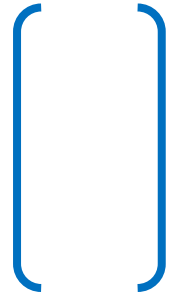
Inputs



Parameters



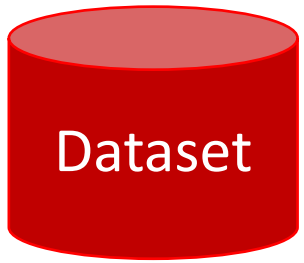
Gradients



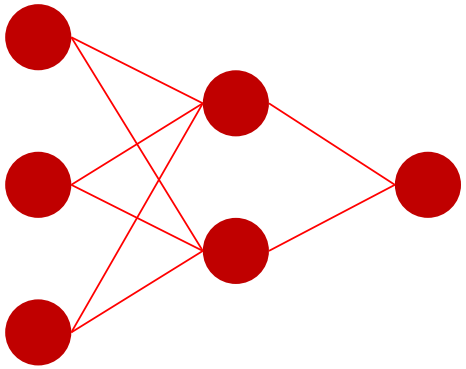
Outputs

Fail-safe Defaults

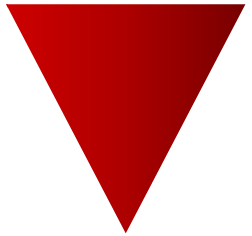
The default behavior of a security system should be to completely deny access



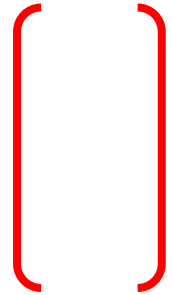
Inputs



Parameters



Gradients



Outputs

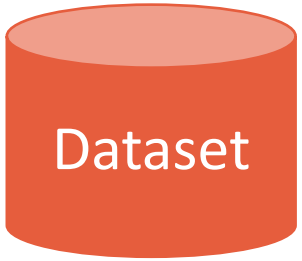
Complete Mediation

Access to sensitive data must always be secure

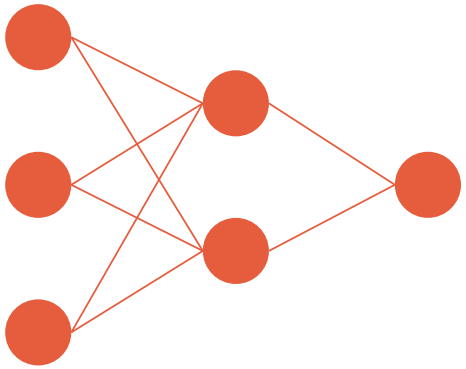
Transferring data over networks

Training and testing data

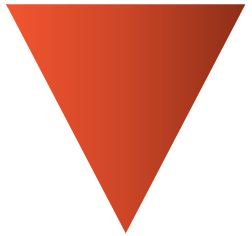
Does not apply to white-box setting



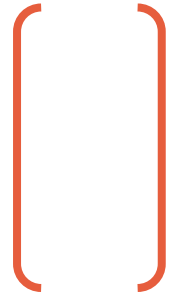
Inputs



Parameters



Gradients

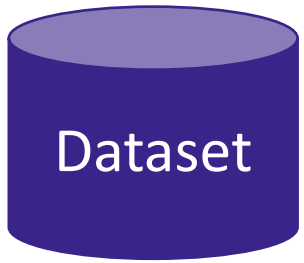


Outputs

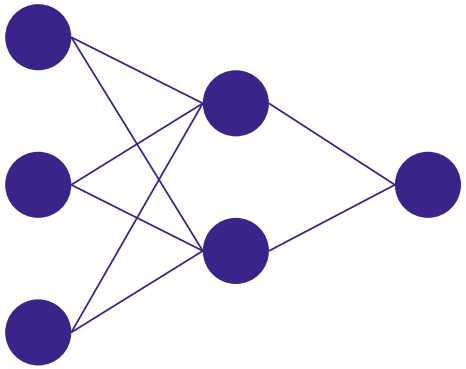
Open Design

Kerckhoffs's principle

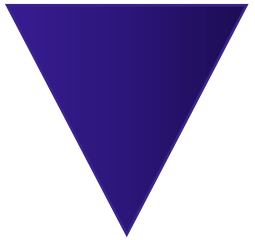
Protection mechanism should not depend on secrecy and system design should be made public



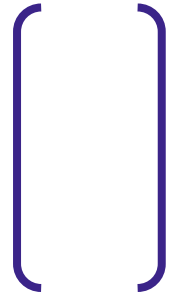
Inputs



Parameters



Gradients

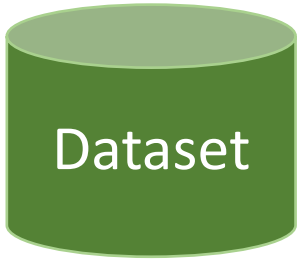


Outputs

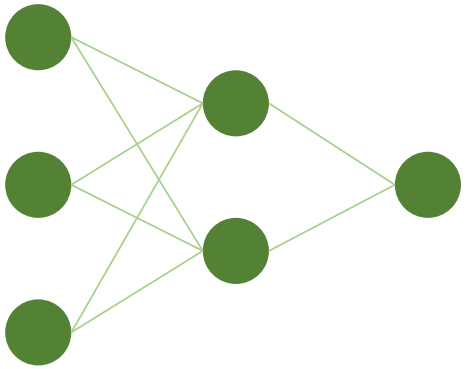
Separation of Privilege

Multiple independent keys should be required for controlling access

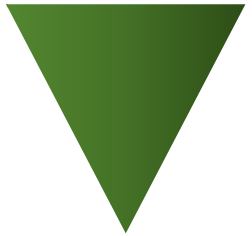
Unclear how this would apply outside of the distributed setting



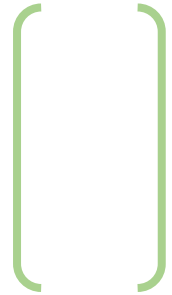
Inputs



Parameters



Gradients

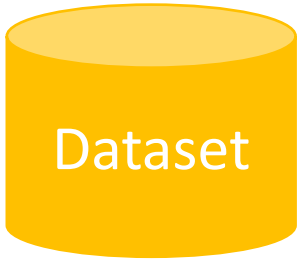


Outputs

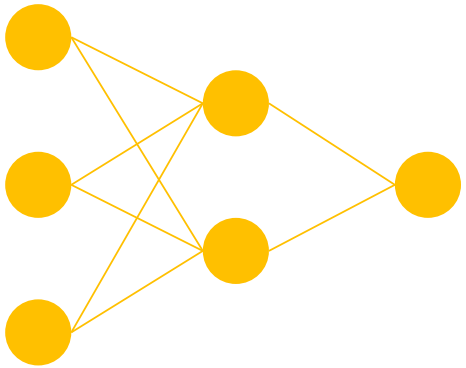
Least Privilege

Every program and every user of the system should operate using the least set of privileges necessary to complete the job

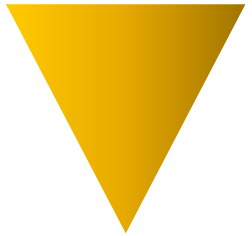
More applicable to the systems and services that the learning model will be part of



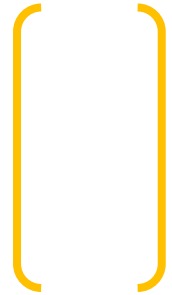
Inputs



Parameters



Gradients

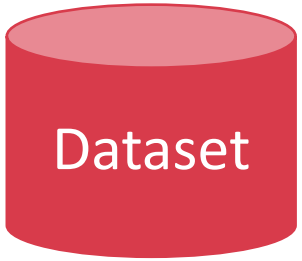


Outputs

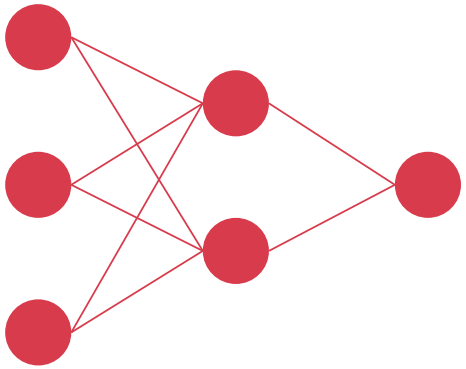
Least Common Mechanism

Shared mechanisms may introduce vulnerabilities or leak information

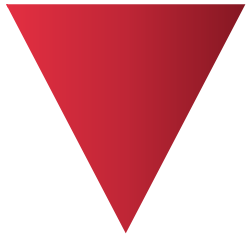
Implies that shared mechanisms should not replicate behavior that can be achieved individually



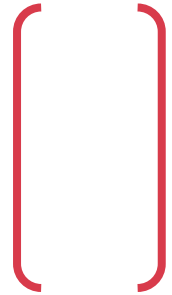
Inputs



Parameters



Gradients



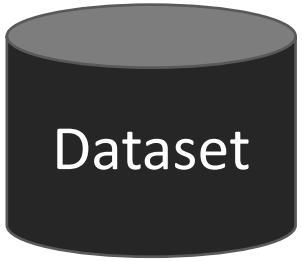
Outputs

Psychological Acceptability

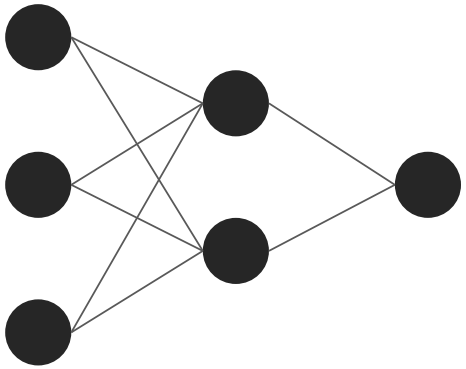
Also known as Principle of Least Astonishment

Intuitive interfaces minimize user mistakes

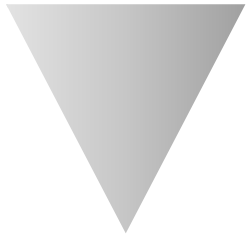
Complexity of training paradigms and deep networks can accidentally leak information



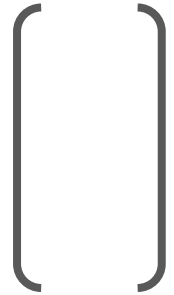
Inputs



Parameters



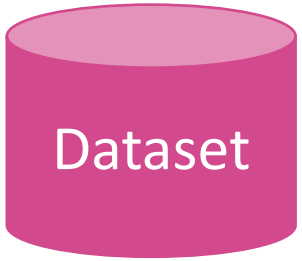
Gradients



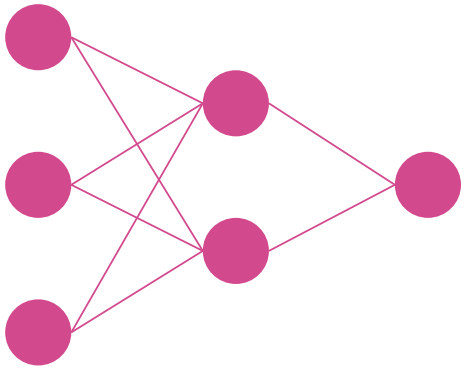
Outputs

Work Factor

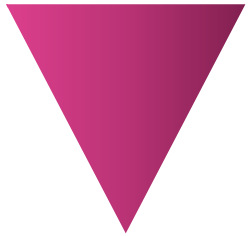
Considering trade-off between attacker resources and circumventing protection mechanisms



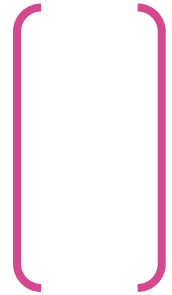
Inputs



Parameters



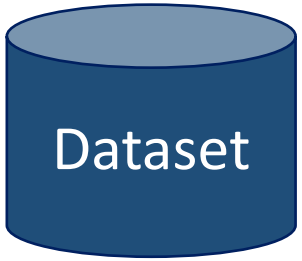
Gradients



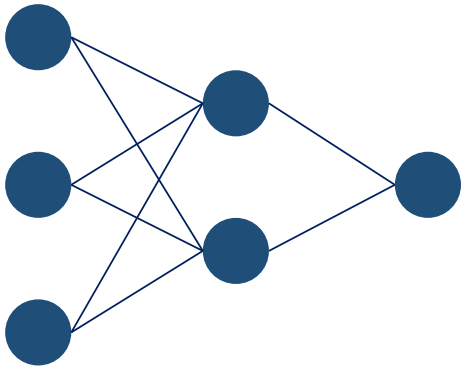
Outputs

Compromise Recording

Advantageous to detect and
report failures of protection

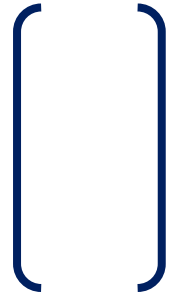


Inputs



Parameters

Gradients



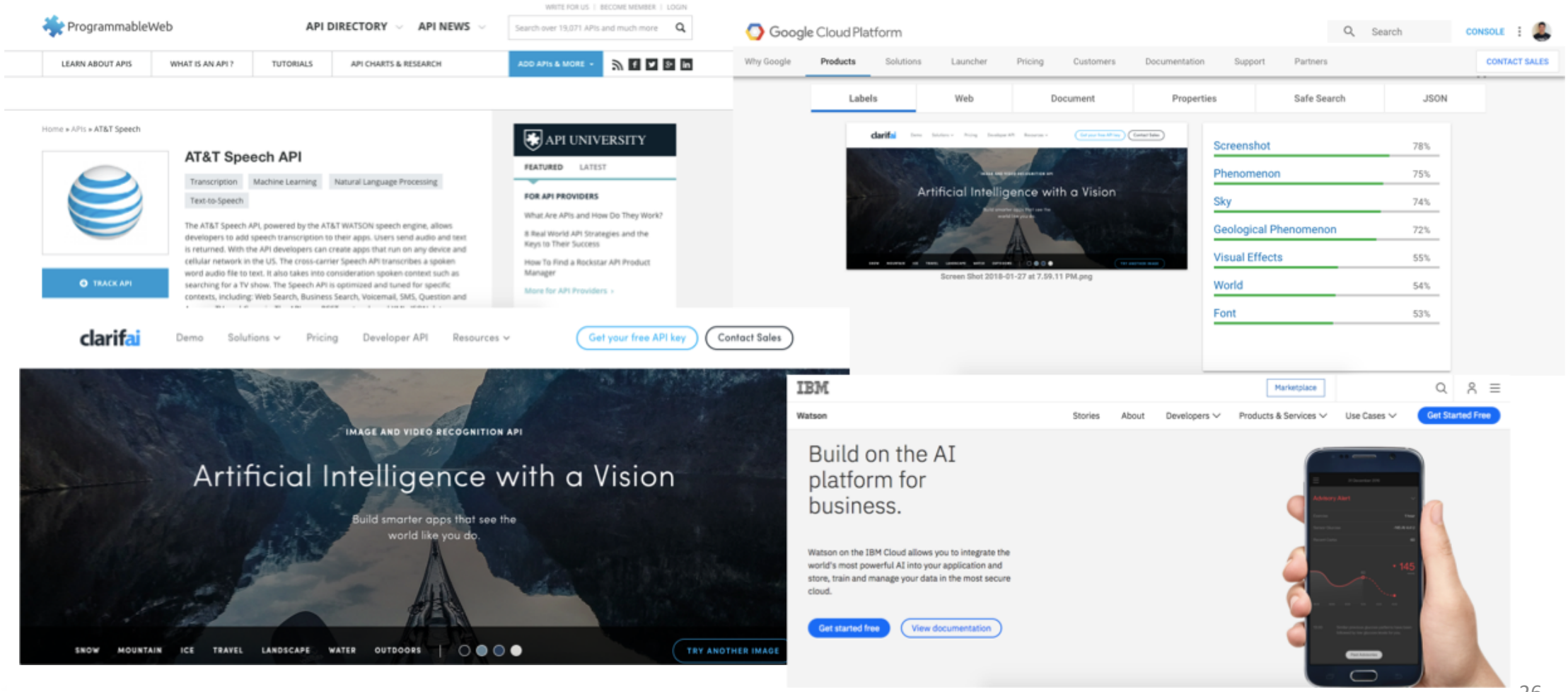
Outputs

Seminar Roadmap

Attacks and Defenses

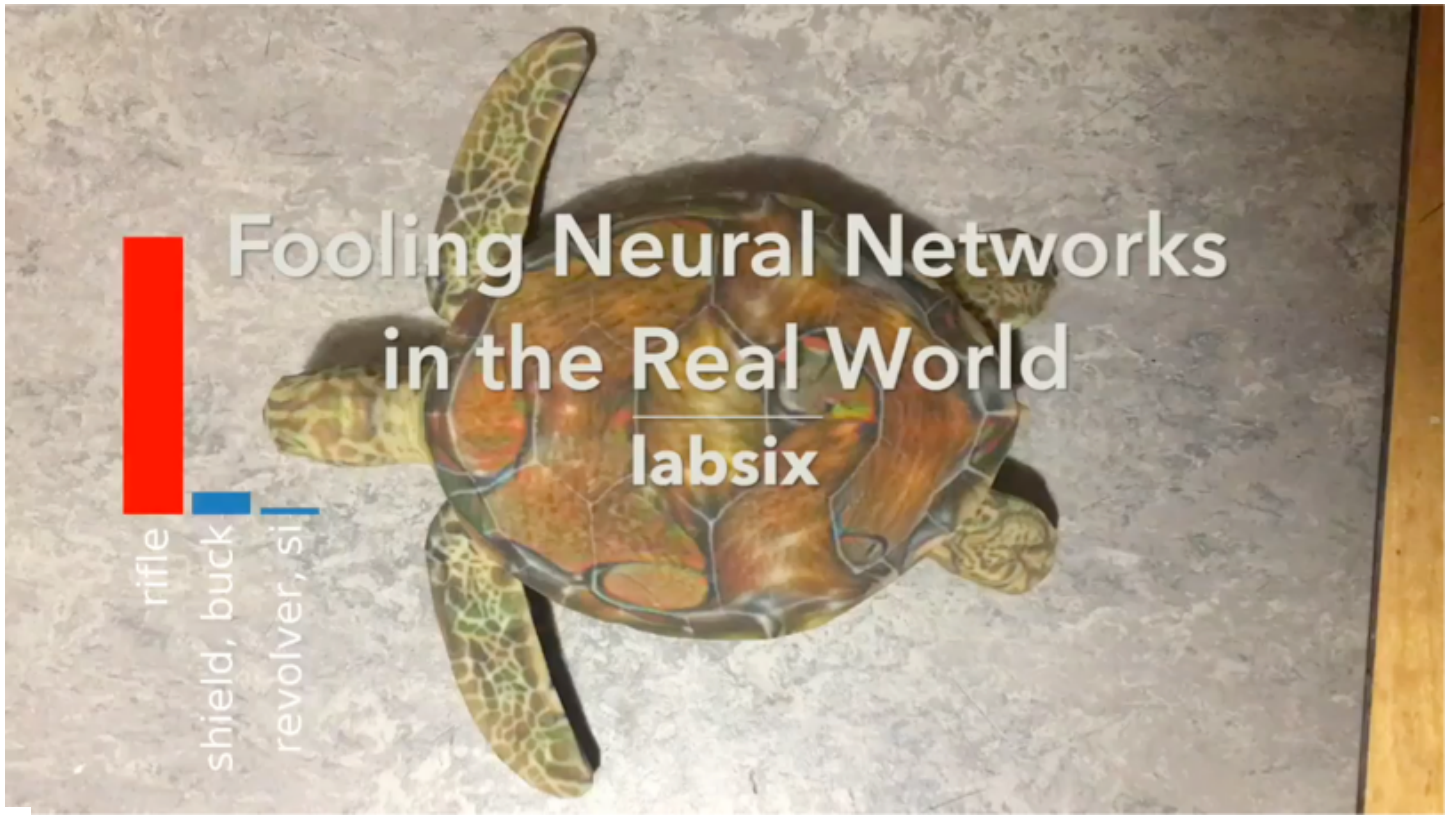
Stealing Machine Learning Models via Prediction APIs

February 5th, 2018



Real-world Adversarial Examples

February 19th, 2018



Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

February 26th, 2018

Reconstructed

Original

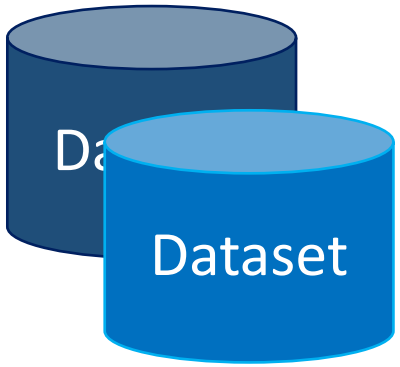
Certified Defenses for Data Poisoning Attacks

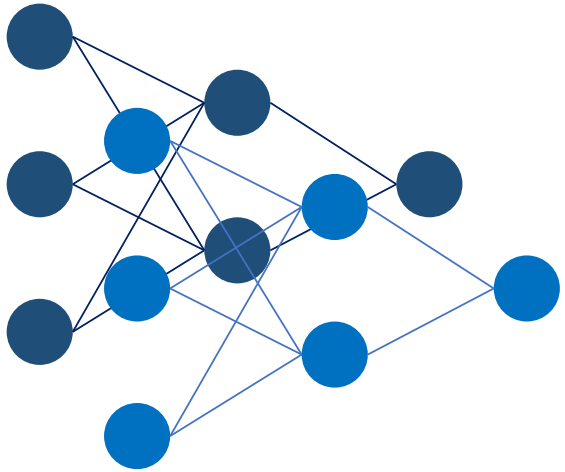
March 19th, 2018

[1] https://en.wikipedia.org/wiki/Tay_(bot)
[2] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
[3] https://www.extremetech.com/computing/225506-microsoft-yanks-new-ai-twitter-bot-after-it-begins-spreading-nazi-propaganda

Deep Learning with Differential Privacy

April 2nd, 2018





CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy

April 16th, 2018

CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy

Nathan Dowlin[1] NDOWLIN@PRINCETON.EDU
Department of Mathematics, Princeton University

Ran Gilad-Bachrach RANG@MICROSOFT.COM
Kim Laine KIM.LAINE@MICROSOFT.COM
Kristin Lauter KLAUTER@MICROSOFT.COM
Michael Naehrig MNAEHRIG@MICROSOFT.COM
John Wernsing JOHN.WERNSING@MICROSOFT.COM
Microsoft Research, Redmond

Abstract

Applying machine learning to a problem which involves medical, financial, or other types of sensitive data, not only requires accurate predictions but also careful attention to maintaining data privacy and security. Legal and ethical requirements may prevent the use of cloud-based

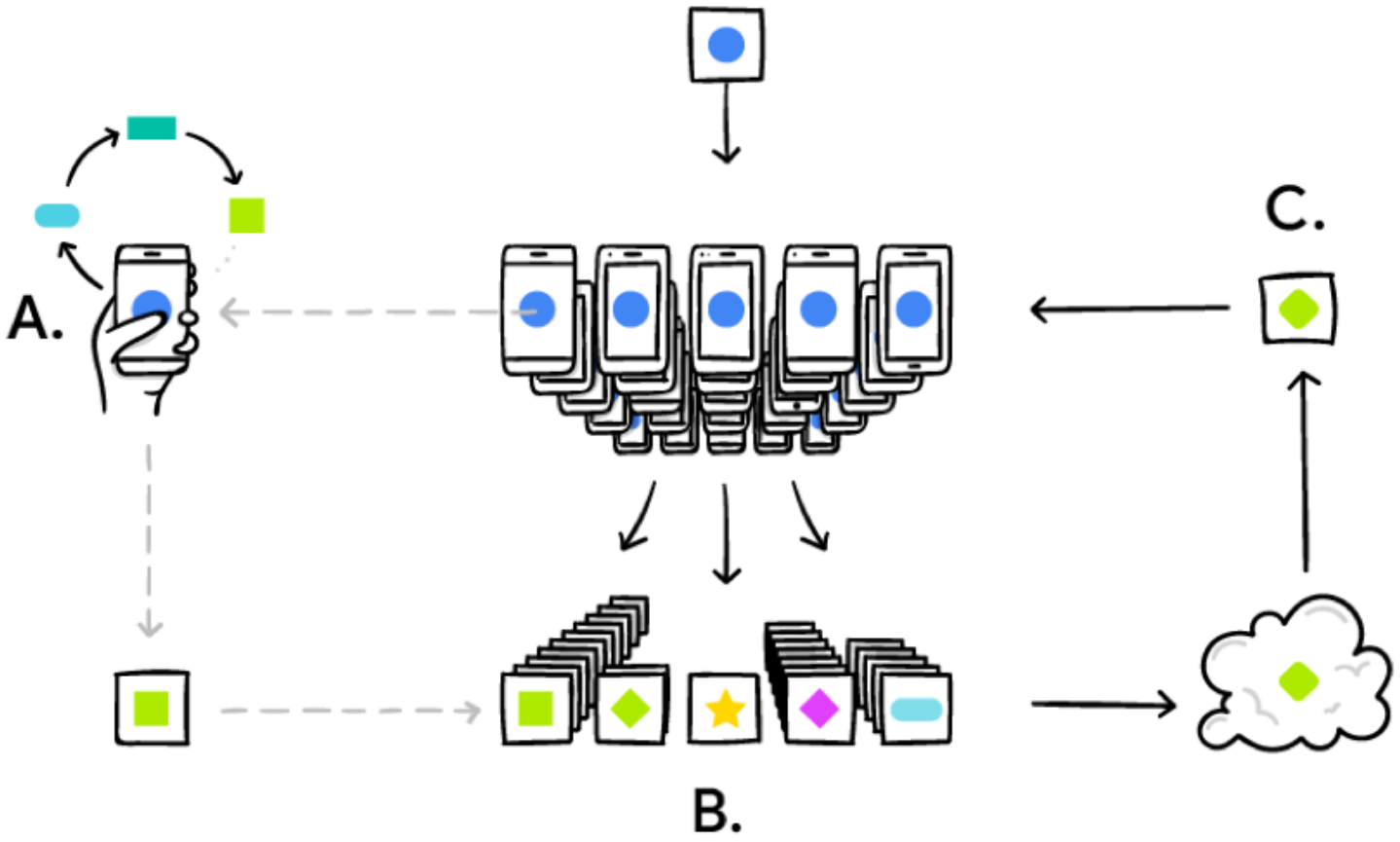
1. Introduction

Consider a hospital that would like to use a cloud service to predict the probability of readmission of a patient within the next 30 days, in order to improve the quality of care and to reduce costs. Due to ethical and legal requirements regarding the confidentiality of patient information, the hospital might be prohibited from using such a service. In

https://www.microsoft.com/en-us/
http://www.books-not-bombs.com/content/images/schools/princeton.jpg

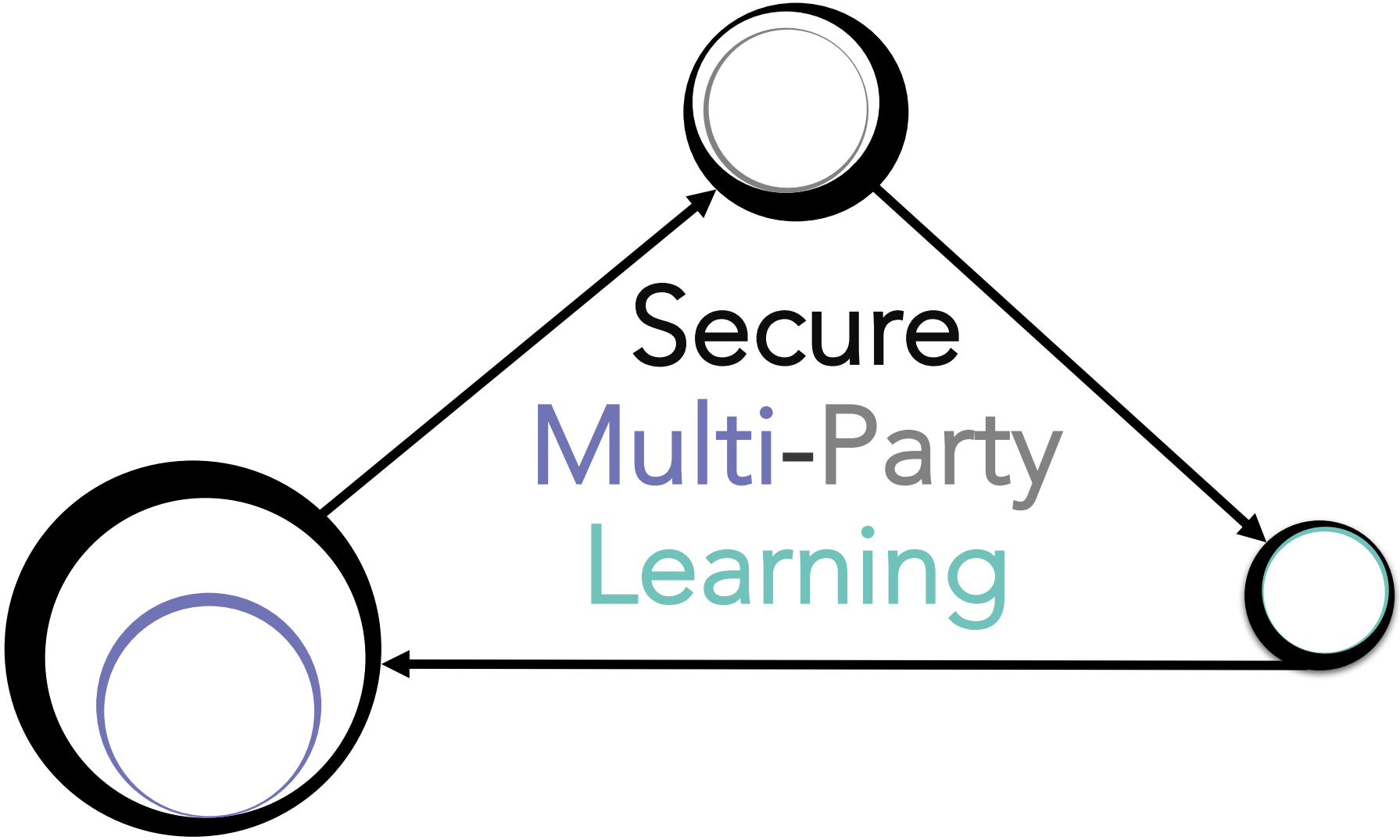
Practical Secure Aggregation for Privacy Preserving Machine Learning

April 23rd, 2018



https://research.googleblog.com/2017/04/federated-learning-collaborative.html

Secure Multi-Party Learning May 7th, 2018



References

1. Saltzer, Jerome H., and Michael D. Schroeder. “The protection of information in computer systems.” *Proceedings of the IEEE* 63.9 (1975): 1278-1308.
2. Papernot, Nicolas, et al. "Towards the science of security and privacy in machine learning." *arXiv preprint arXiv:1611.03814*(2016).
3. Abadi, Martín, et al. "On the protection of private information in machine learning systems: Two recent approches." *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*. IEEE, 2017.
4. Liu, Yun, et al. "Detecting cancer metastases on gigapixel pathology images." *arXiv preprint arXiv:1703.02442* (2017).
5. Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *arXiv preprint arXiv:1711.05225* (2017).
6. https://gizmodo.com/here-are-the-microsoft-twitter-bot-s-craziest-racist-ra-1766820160