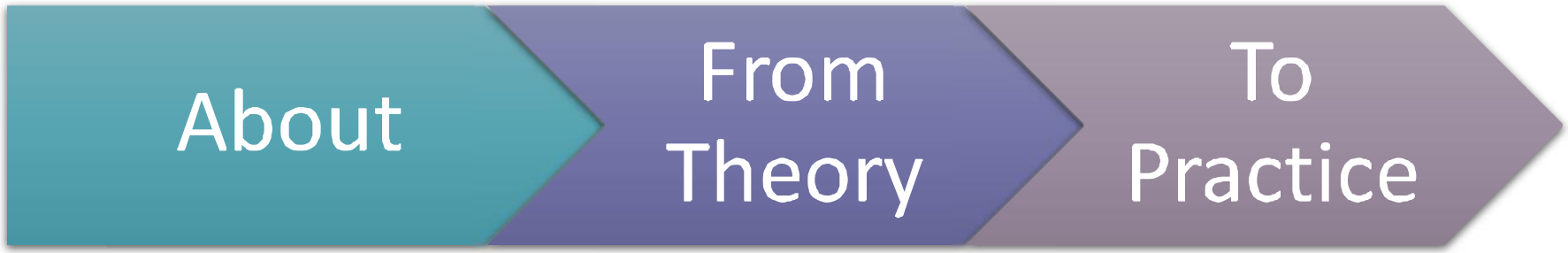
ML: From Theory to Practice

NGUYEN Vinh Tiep
Technical Program Manager at Koidra
Lecturer at UIT – VNUHCM
Academic Head at VietAI

Outline




About

My Background

- Chuyên Toán at PTNK – VNUHCM (03-06)
- B.S. in Information Technology (Honor Program) at HCMUS -VNUHCM (2010)
- M.Sc. in Computer Science at HCMUS-VNUHCM (Co-program with JVN) (2013)
- Ph.D. in Computer Science at UIT-VNUHCM (2019 – in the last round of defense)

Experience

Lecturer at HCMUS

- 2010 – 2017: Software Engineer Department
- 2015: Visiting Researcher at National Institute of Informatics, Tokyo, Japan (NII)

Lecturer at UIT

- 2017 – present: Faculty of Computer Science
- 2011 – 2014: Business Analyst, Barclays, Singapore

Academic Head at VietAI

- 2018 – present: Foundation of ML and Advanced Class in Computer Vision

Technical Program Manager at Koidra

- 2018 – present: ML development and deployment

Topics

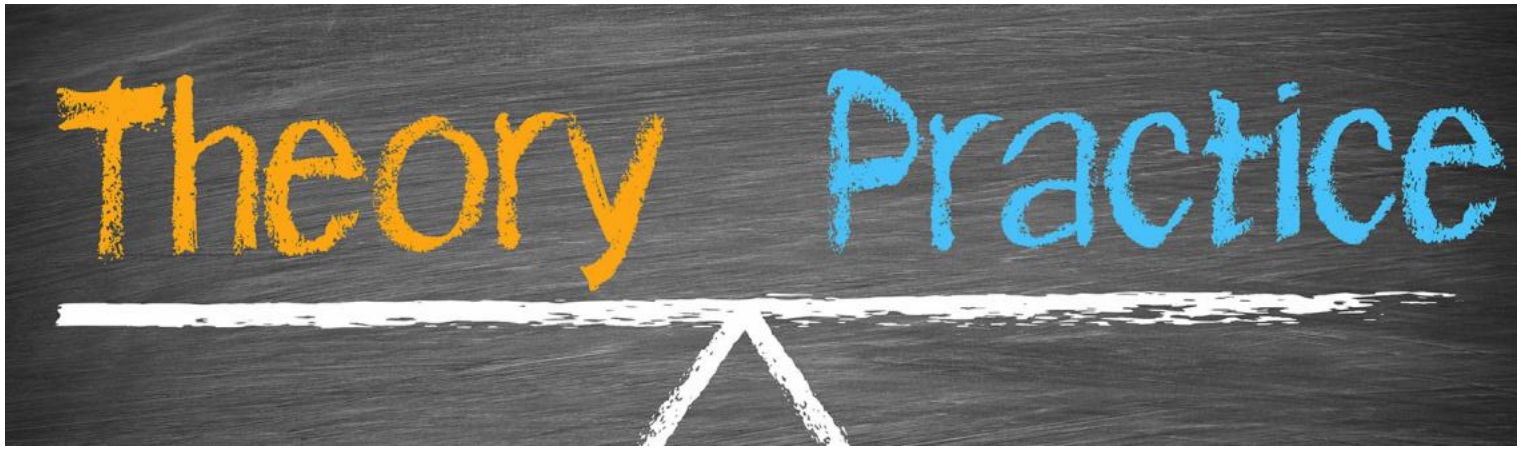
Visual Search System

- Rank #1 at TRECVID Instance Search in 2014, #2 in 2015 and 2016 (organized by NIST)
- Rank #1 at TRECVID Adhoc Video Search in 2016 (organized by NIST)

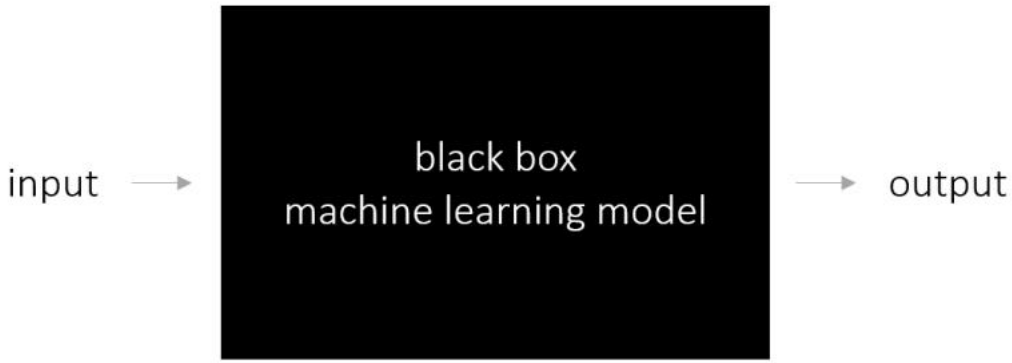
Applied ML

- Churn Prediction: Predicting customers who are going to unsubscribe the company's services in the next month
- Price Prediction: Predicting optimal price distribution for house leasing everyday
- Visual Tagging: Adding labels to describe the image
- OCR in the wild: Detecting and Recognizing text in a scene image

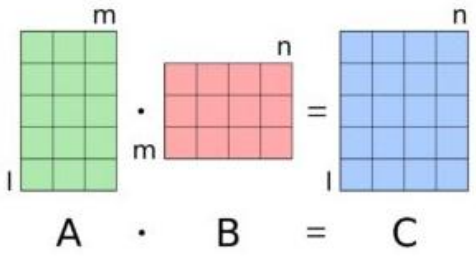
From Theory

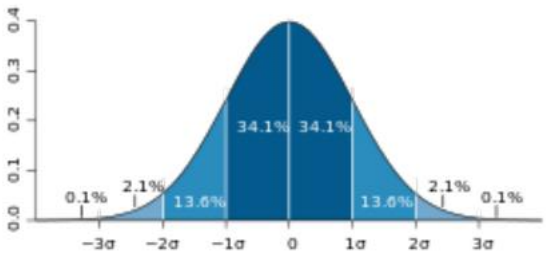


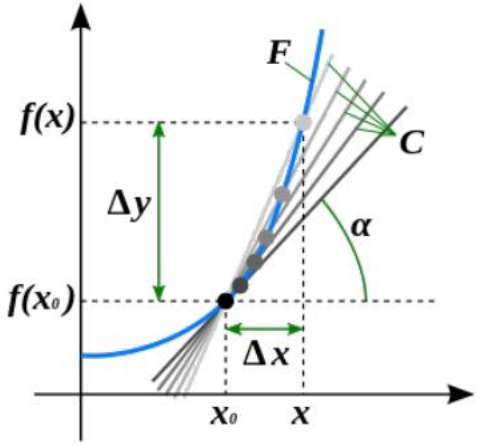
Why we must have background?



#1. Without background, ML like a black box







#2. Which background should we have?

Linear Algebra:

- Matrix and vector

Probability:

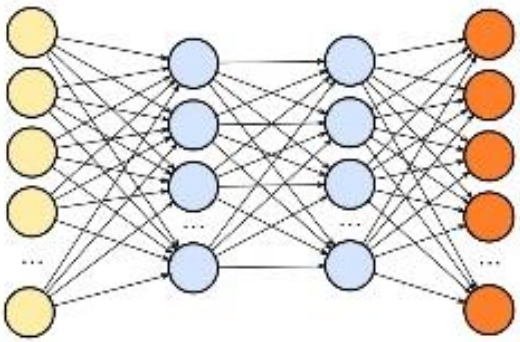
- Very fundamental knowledge

Calculus:

- Derivative

#3. How to improve the background of VNese ML engineers?

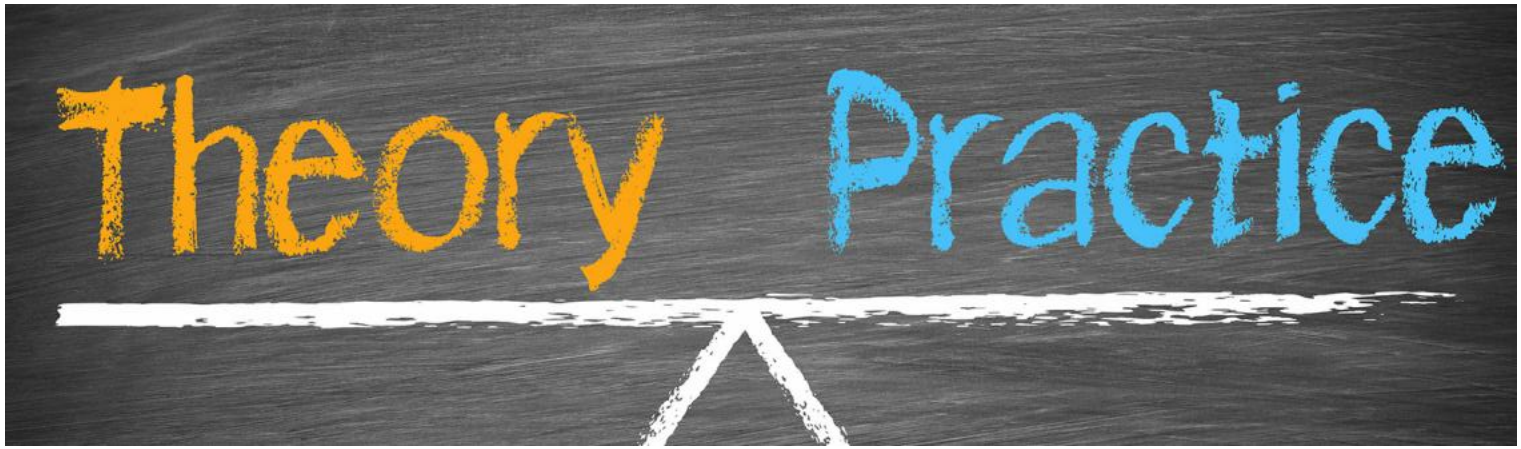
* Foundation ML classes in 3 regions of Vietnam: Saigon, Hanoi, Hue
* UIT is the first university that teaches Deep Learning officially
* Research skills are embedded to the courses
* We connect to world-class experts



=

#4. With background, ML like a Lego

To Practice

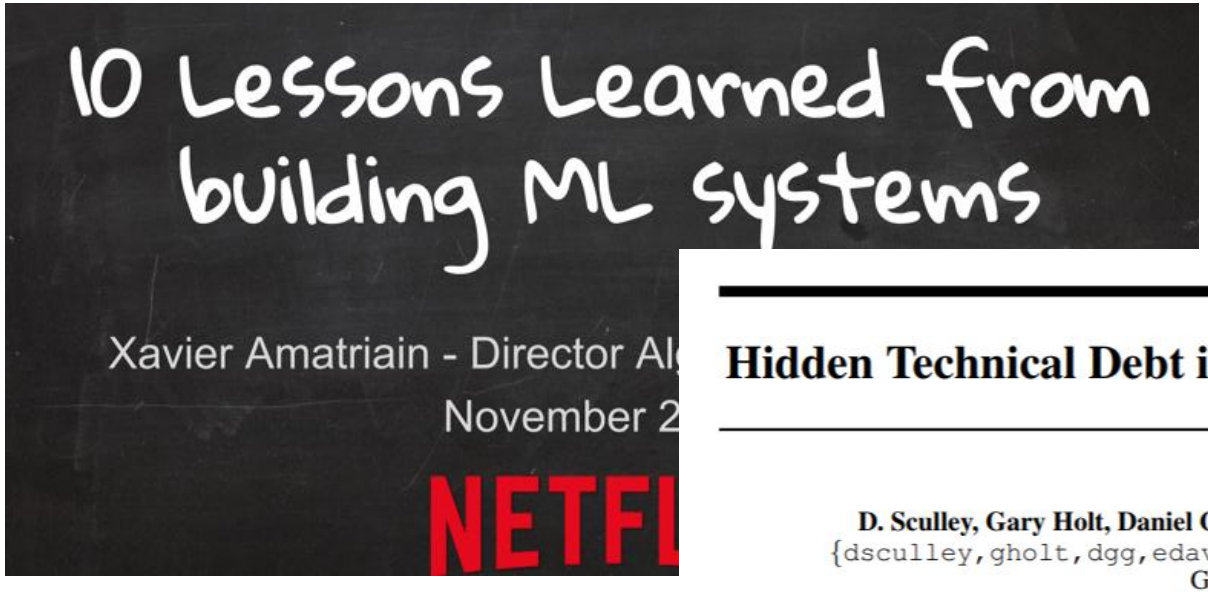


ML/DL: from theory to practice

10 Lessons

These slides are credited from Kenneth Tran, MSR

Inspirations



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

#1. Metrics

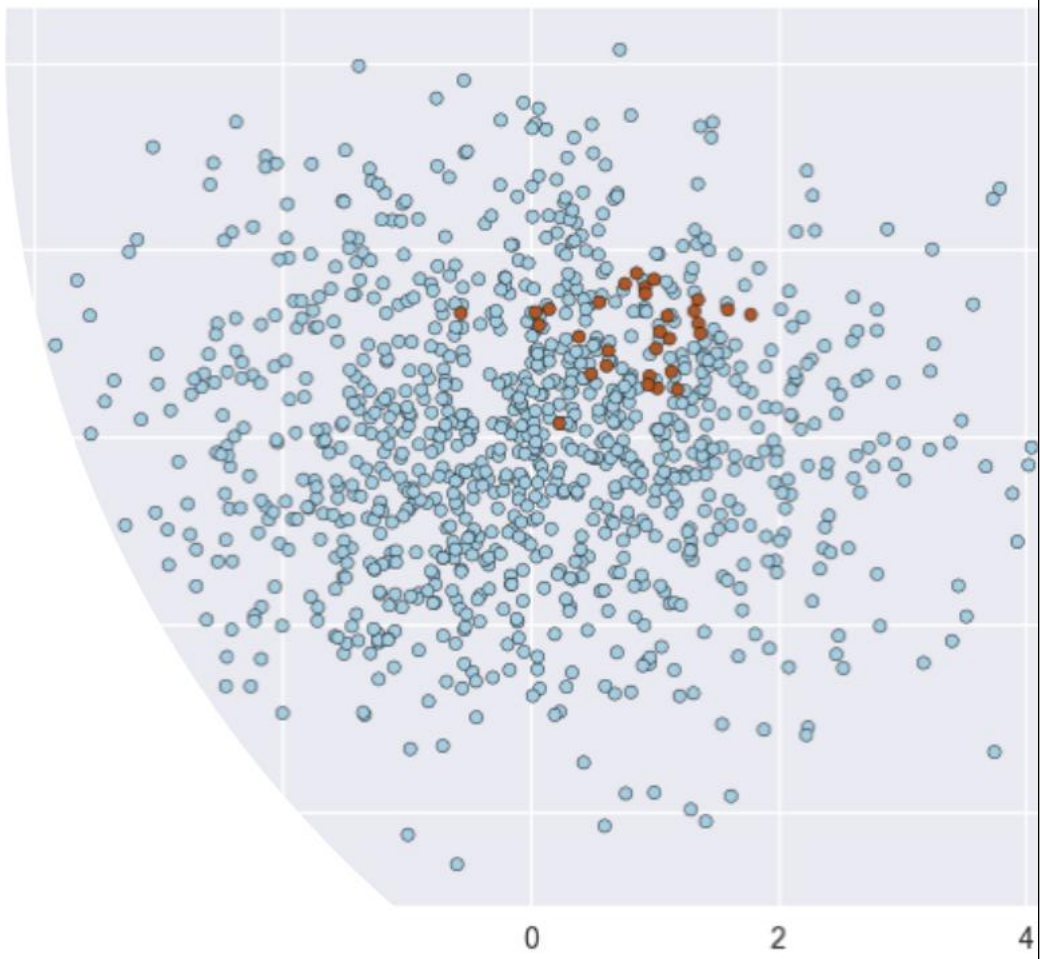
Data and Models are great. You know what's even better?

The right evaluation approach.

Case: Fraud Detection

Characteristics

- Class imbalance
- Different costs for FP and FN



Case: Fraud Detection

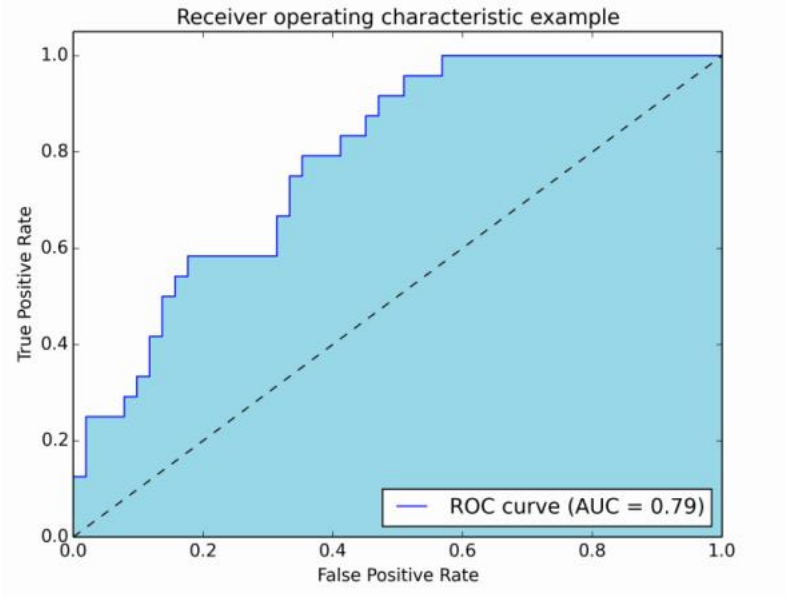
So what metrics should we use?

- Accuracy

Metrics

So what metrics should we use?

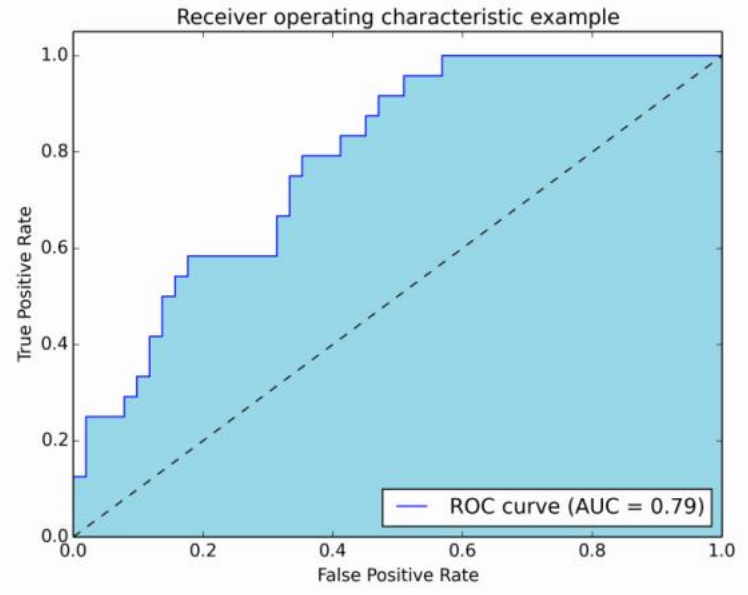
- Accuracy
- ROC AUC

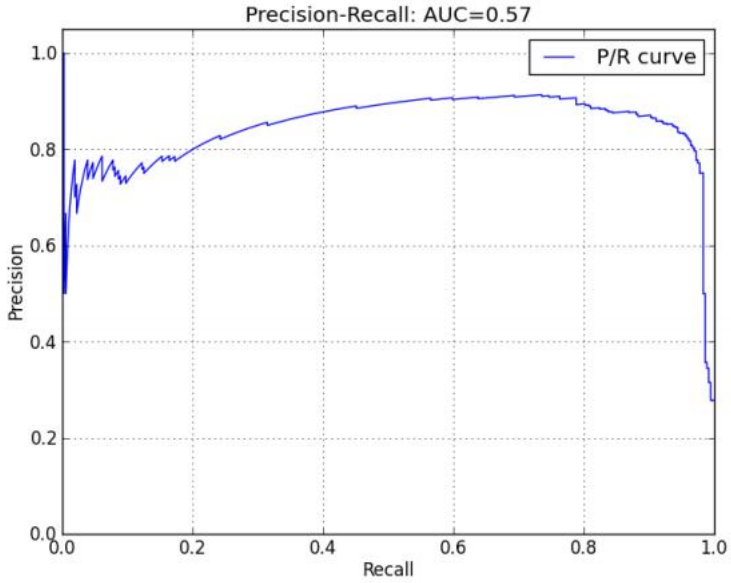


Metrics

So what metrics should we use?

- Accuracy
- ROC AUC
- PR AUC



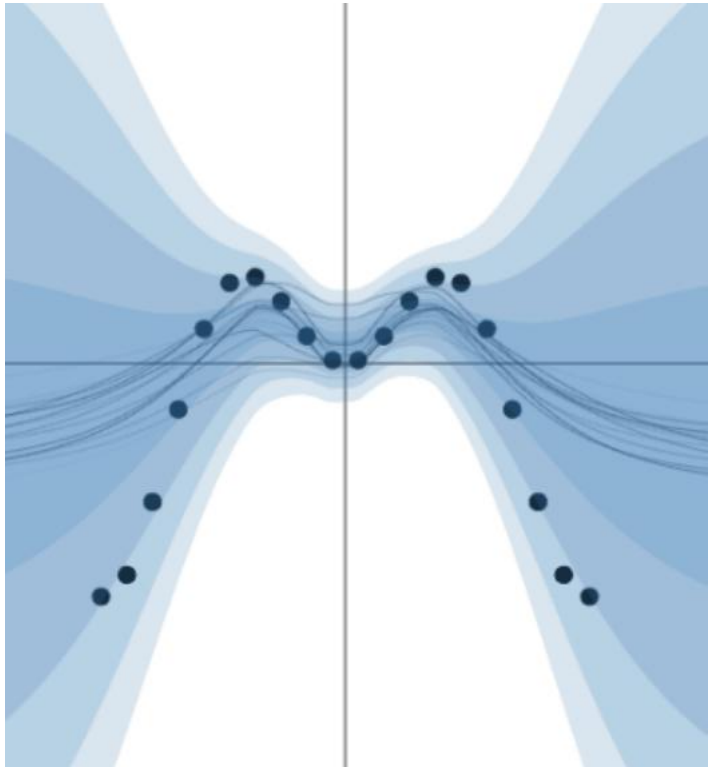


Metrics

So what metrics should we use?

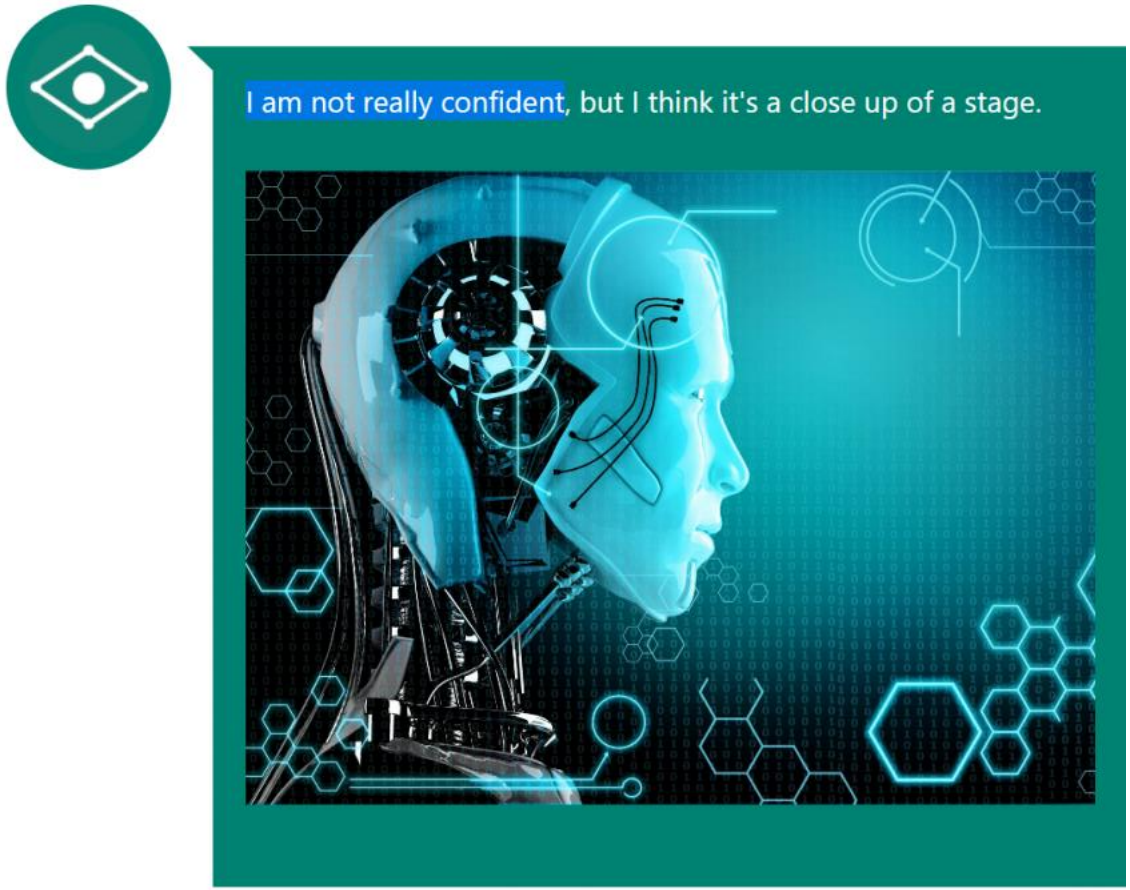
- Accuracy
- ROC AUC
- PR AUC
- How about simple economics

$$L = N_{FP} \times C_{FP} + N_{FN} \times C_{FN}$$

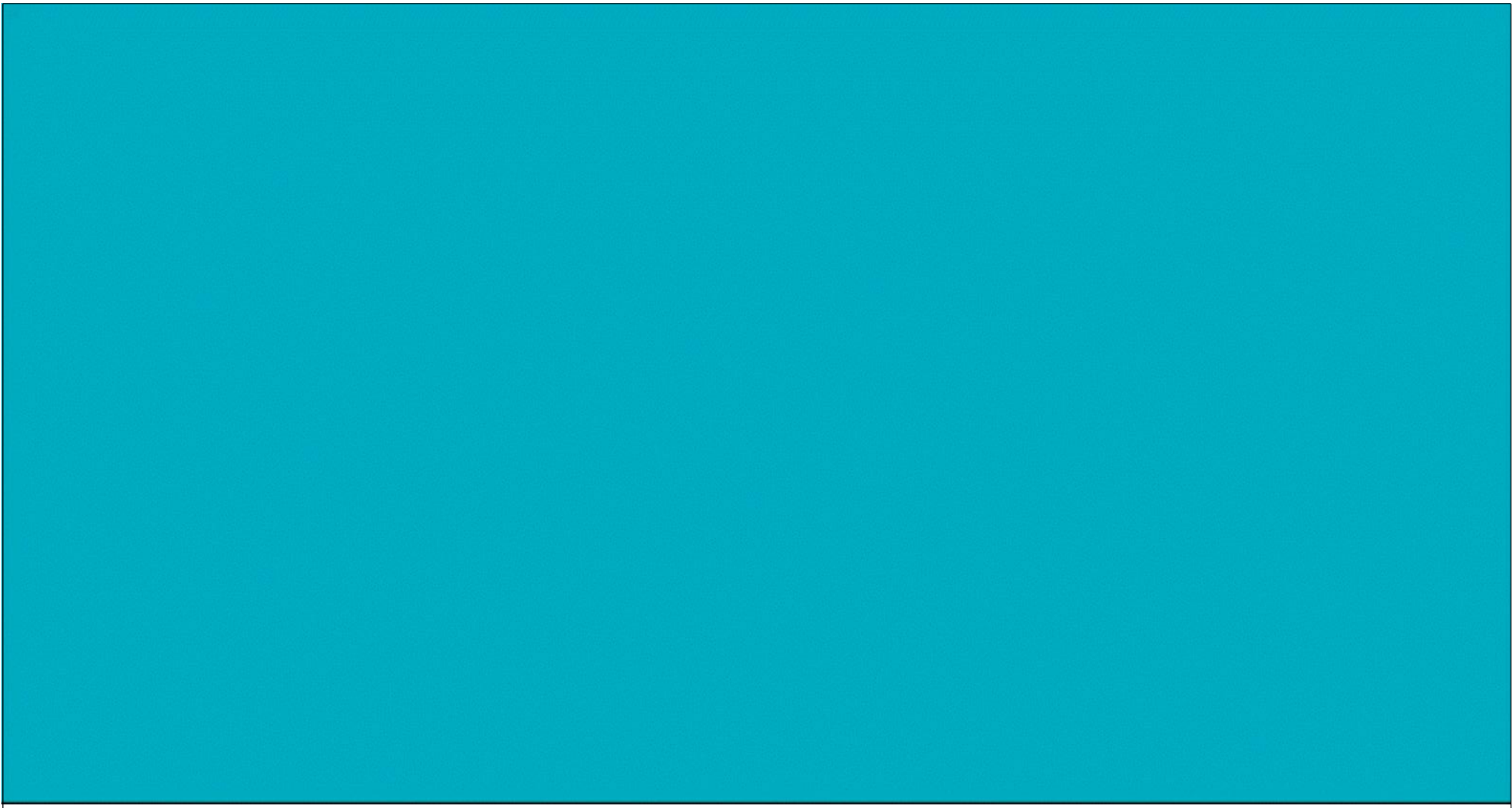


#2. Uncertainty: your model should be able to tell what it doesn't know

Predict with uncertainty



www.CaptionBot.ai



Other use cases

- Demand forecasting
- Autonomous or semi-autonomous driving
- Health care
- Any application in which a misclassification (or misprediction) is costly or when the prediction is an input to a decision making process

#3. Understand the inter-dependency between models and features

The fact that a more complex model does not improve things does not mean you don't need one

Better models and features that don't work

Imagine the following scenario

- You have a Random Forest model and for some time you have been selecting and optimizing features for that model
- If you try a Neural Nets model with the same features you are not likely to see any improvement
- If you try to add more expressive features (e.g. text embedding), the existing random forest model is likely not to capture them and you are not likely to see any improvement

It's important to understand the interplay between features and models.

The fact that a complex model doesn't work well doesn't imply that you should discard it.

A counter scenario

- A company/team is tasked to solve a ML problem
- They spend lots of effort on a DL model and very little effort on the other approaches
- They later claim having improved the results using DL
- Slight improvement using DL used to generate more PR (and hence promotion/investments) than more substantial improvements with non-DL methods

A counter scenario

- Slight improvement using DL used to generate more PR (and hence promotion/investments) than more substantial improvements with non-DL methods
- This was fairly common in “Silicon Valley”

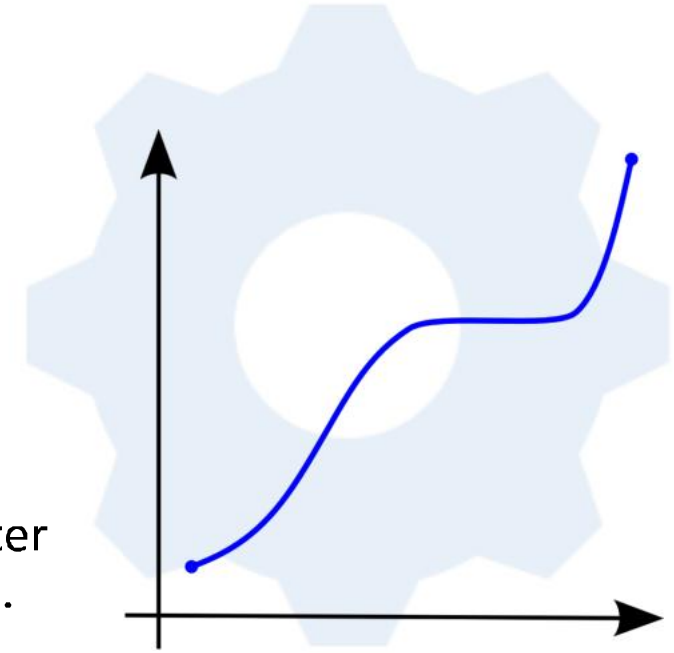
Gigaom | How PayPal uses deep learning and detective work to fight fraud

Hui Wang has seen the nature of online fraud change a lot in the 11 years she's been at PayPal. In fact, a continuous…

GIGAOM.COM

#4: model performance is a monotone function of engineering effort

Better results don't always imply smarter model. Be aware of hype vs. substance.



#5. You may not need all your Big Data

"Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

- Dan Ariely, Duke University (2013)

Slide Credit: Xavier Amatriain

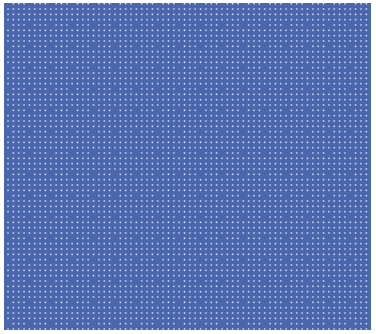
How useful is Big Data

- "Everybody" has Big Data
  - But not everybody needs it
  - E.g. Do you need many millions of users if the goal is to compute a MF of, say, 100 factors?
- Many times, doing some kind of smart (e.g. stratified) sampling can produce as good or even better results as using it all

NETFLIX

#6. You mostly don't need distributed ML

Distributed ML is another "dangerous" trend, similar to "Big Data"



#7. Be aware of feedback loops

Direct Feedback Loops

- A model may directly influence the selection of its own future training data.
- Learn contextual bandits

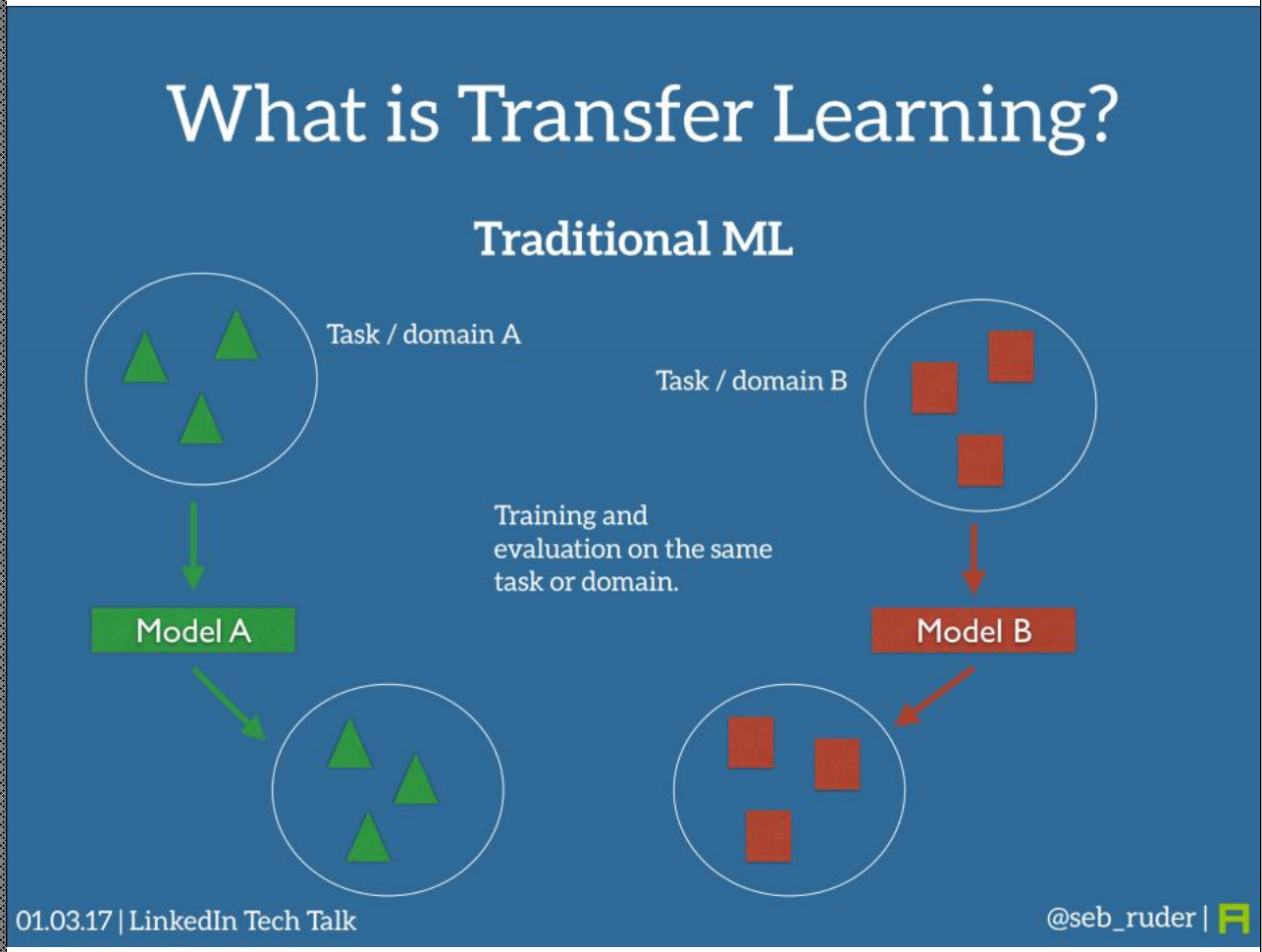
Hidden Feedback Loops

- Example: 2 stock-market prediction model

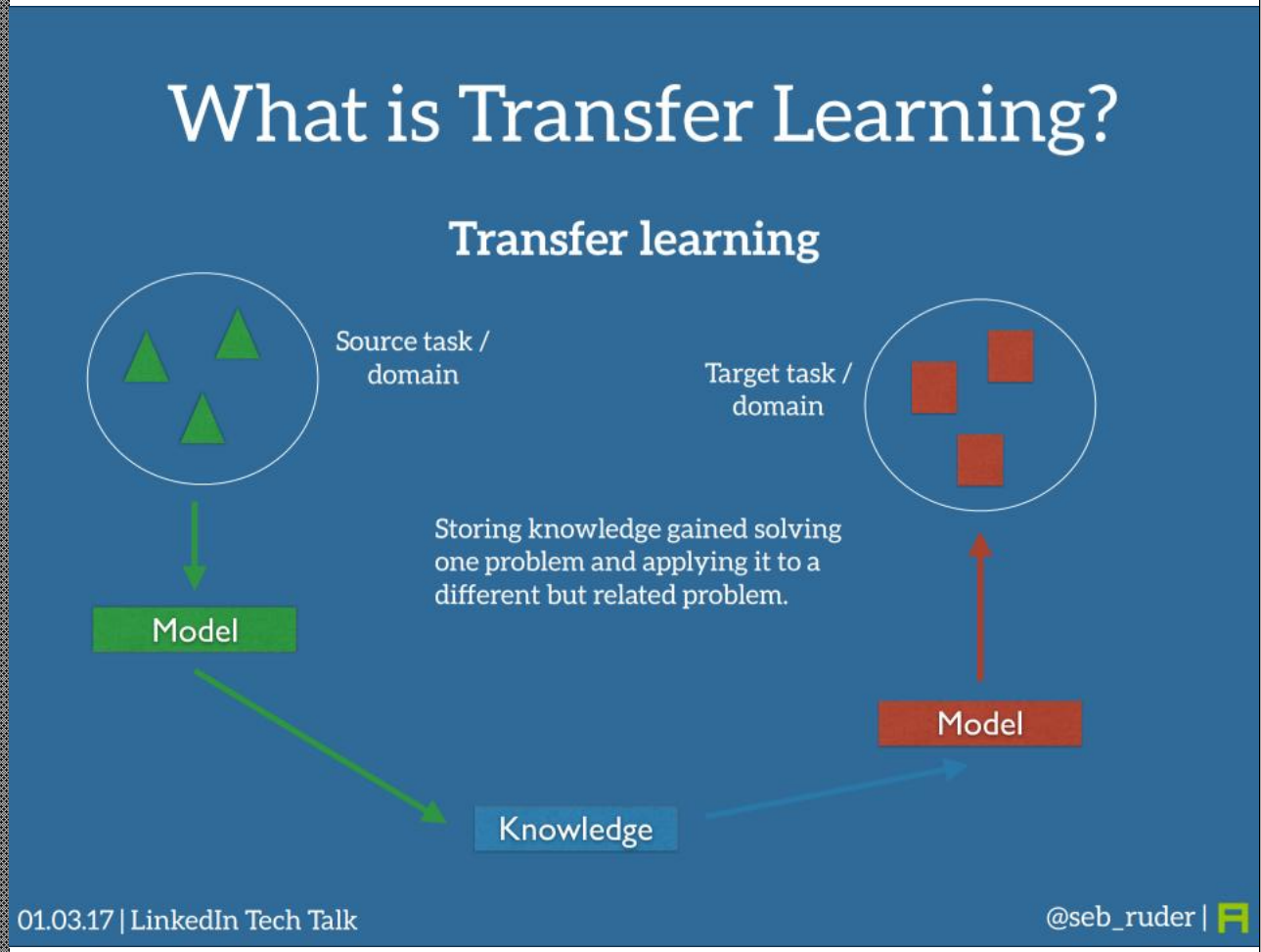
#8. Use Transfer Learning

Transfer Learning has been a key driver of ML success in industry

Transfer Learning

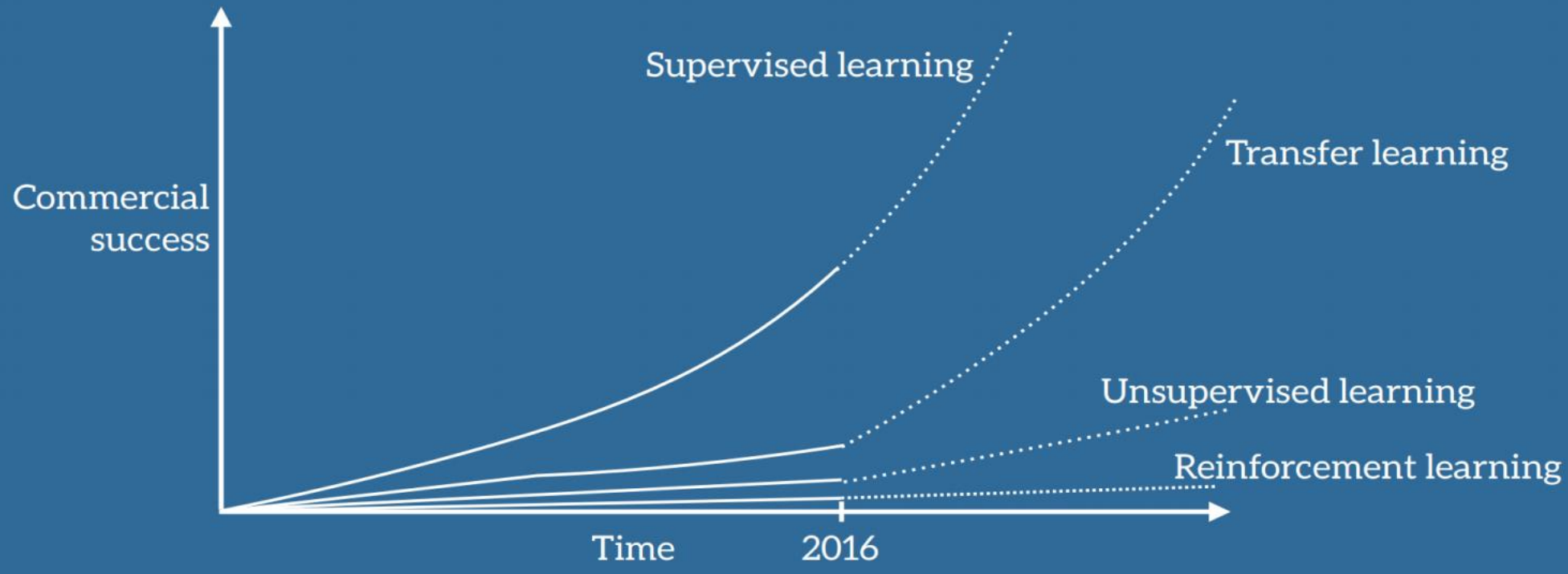


Transfer Learning



Why Transfer Learning now?

Drivers of ML success in industry



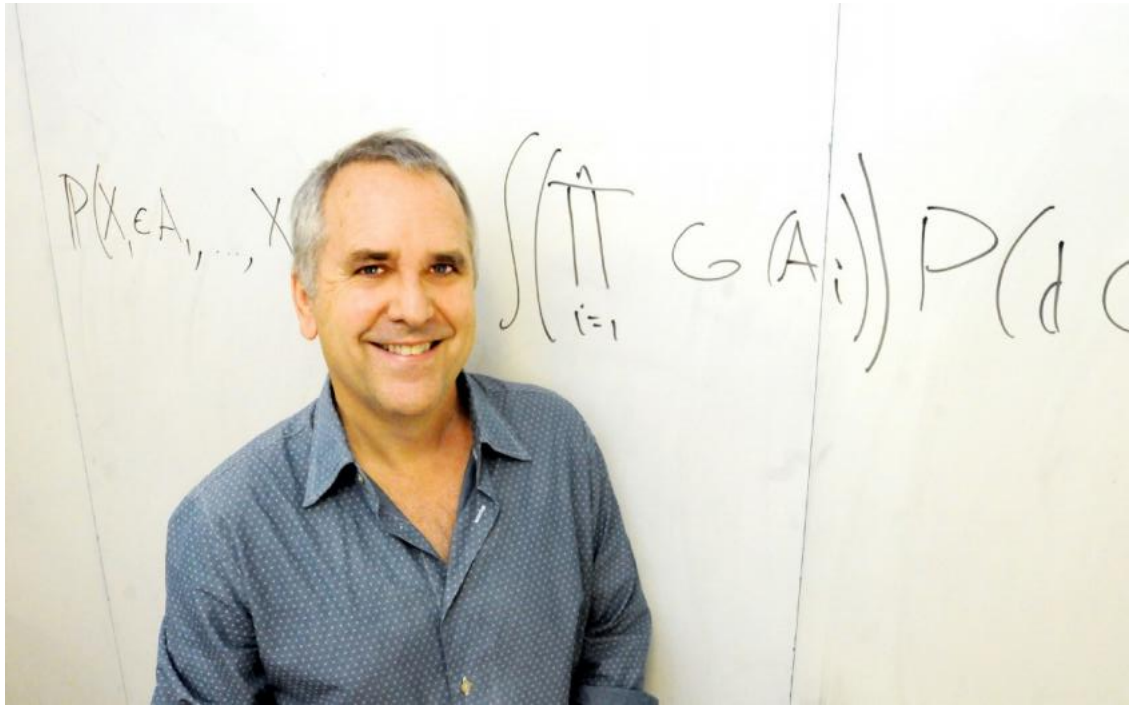
Slide Credit: Sebastian Ruder

- Andrew Ng, NIPS 2016 keynote

#9. Transfer Learning from Software Engineering

- Model versioning
- Experiment review
- Hierarchical and compositional mindset
  - DL was inspired by this!

#10. AI – The revolution hasn't happened yet



#10. AI – The revolution hasn't happened yet

- Start simple and take incremental steps
  - Make sure you understand what you are learning/doing at each step
- Start with real problems, then identify the technologies
  - Common mistakes: start with (hyped) technologies, then find applications
  - Examples: chat bots, personal assistants, etc.
- There are many simple, unsexy, but high-value ML problems

Thank you for your attention!