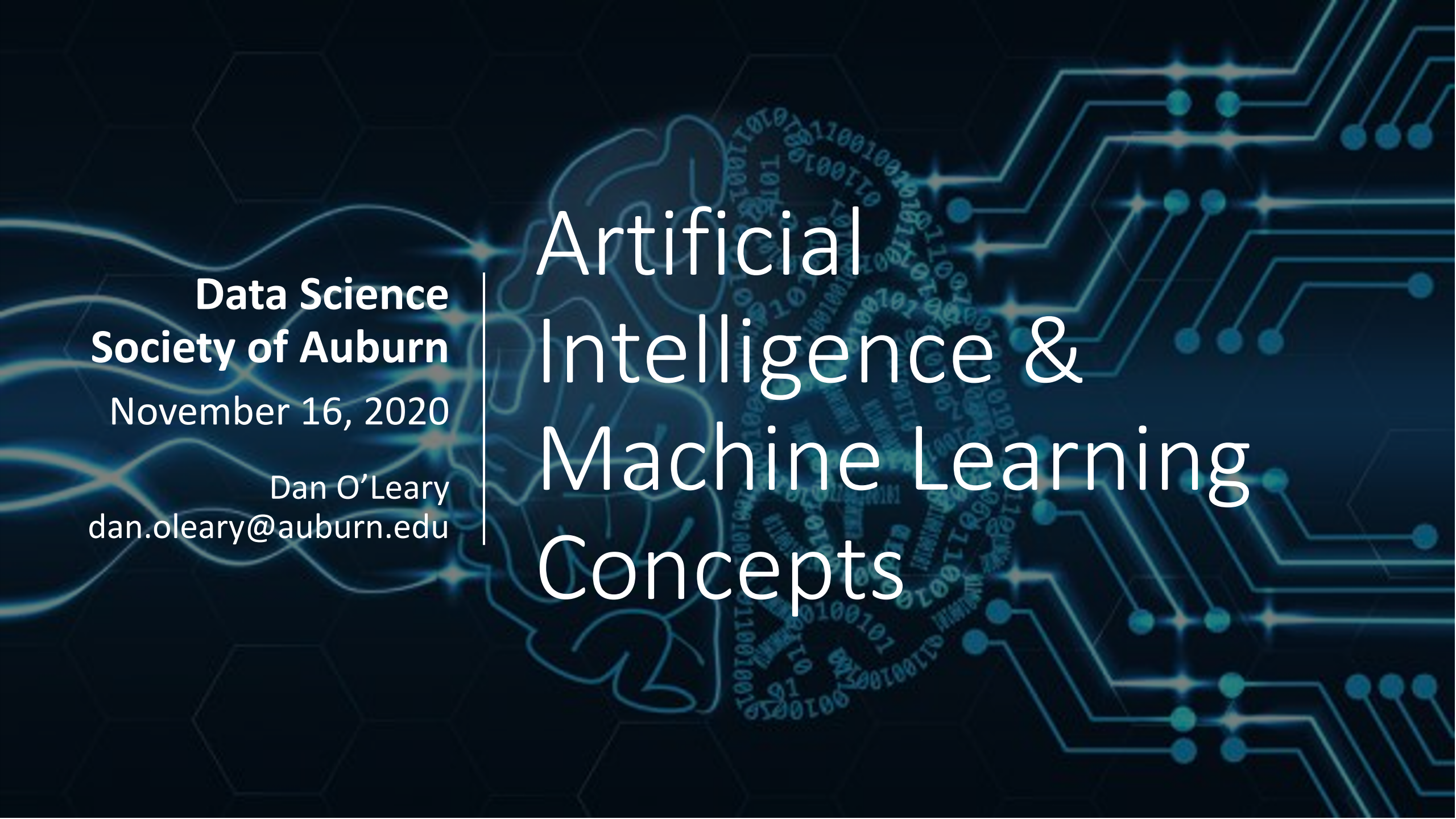


Artificial Intelligence & Machine Learning Concepts

Data Science Society of Auburn

November 16, 2020

Dan O'Leary
dan.oleary@auburn.edu

Introduction

DATA SCIENCE SOCIETY
OF AUBURN

- Introduction to AI/ML concepts and Data Science
- Based on a lecture for undergrads in BET 2019
  - No expectation of related knowledge
- Expanded and deepened for this audience
- Still “Big Picture”
  - Provide some context, detail for those of you in deeper studies
  - Introduce those of you new / interested
- Disclaimers
  - Some references a little dated (yes, in 18 months)
  - Not as well cited as it should be
  - I am not an expert!

Topics

- Attempt to define Artificial Intelligence (fail)
- Attempt to describe the field of Data Science (better)
- Introduce fundamental concept of Machine Learning (pretty good!?)
- Describe state of the art methods in 1 slide each (jury's out)
- Use cases and examples (great!)
- Discuss why now, benefits, and limitations / pitfalls (ok)

Only the tip of the tip of the iceberg
Not a formal academic presentation

Background

- 1992 – BS Mechanical Engineering
- 2019 – Master of Engineering Management, Supervised Learning
- 2020 – Grad Cert Modeling and Analytics for Operations (ISE)
- ≈2022 – PhD, Industrial and Systems Engineering
  - Modeling / simulation, data science / machine learning
  - Teach undergrad courses in innovation and product development
- Life-long fascination with modeling and simulation of all types
- 2 x Entrepreneur: co-Founded n-Space in 1994, funding from Sony
  - 23 years, 45 games – concept to completion; major brands, publishers, and platforms; all genres and demographics

Embedded videos have been replaced with screenshots and youtube links.

From GTC 2018 – GPU Tech Conference
https://youtu.be/GiZ7kyrwZGQ

Question: What is Artificial Intelligence (AI)?

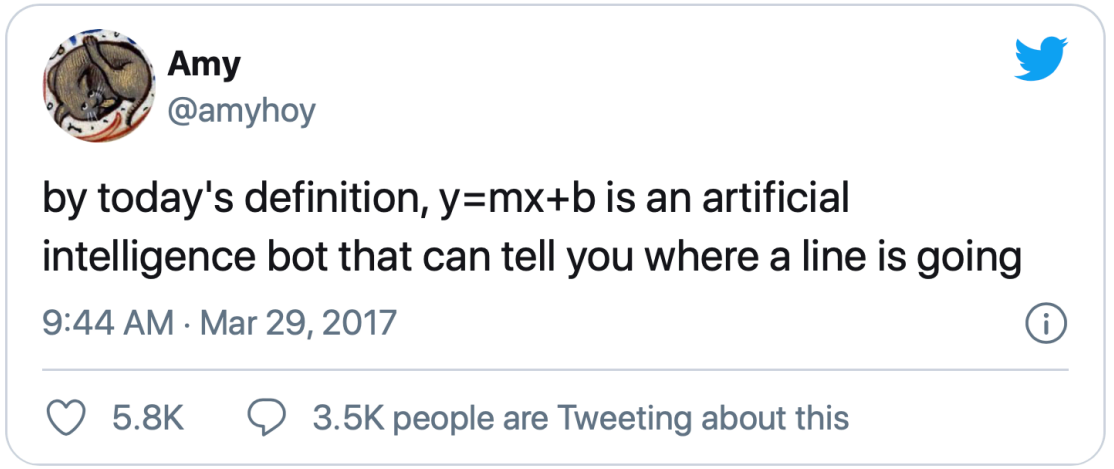
If you can't explain it to a six-year-old, you don't understand it yourself.

- Albert Einstein

Answer: I don't know.

What is AI?

- Imprecise, overused term
  - Calculator?
  - Self-driving car?
  - Chatbot?
- Definition is fuzzy, changes over time
- Old, diluted, hyped term – backlash, cynicism
- Generally used to describe machines doing tasks traditionally assigned to humans



Classic Definition of AI

An intelligently designed agent that perceives its environment and makes decisions to maximize the chances of achieving its goal.

- Subfields:
  - Computer Vision
  - Robotics / Control Theory
  - Natural Language Processing

} Now considered "Machine Learning"

https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

AI Effect

Once a machine takes over a task, humans tend to dismiss it as “not AI”

It’s part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something — play good checkers, solve simple but relatively informal problems — there was chorus of critics to say, ‘that’s not thinking’

Pamela McCorduck, 2004

AI is whatever hasn't been done yet. – Douglas Hofstadter

AI Effect

- 1997 – IBM's Deep Blue beats world chess champion Garry Kasparov
  - 259th most powerful supercomputer at the time
  - Planned 6-8 moves out; as high as 20+
  - "Brute force methods… not *real* intelligence"
- Changed the canonical example of human vs machines from Chess to Go
  - Simple rules, many more possible moves
  - More intuition, less susceptible to brute force
- 2016 Google DeepMind beats Go Champ
- 2019 DeepMind beats StarCraft II Pros

> *When you're fundraising, it's AI*
> *When you're hiring, it's ML*
> *When you're implementing, it's linear regression*
> *When you're debugging, it's printf()*
>
> *— Baron Schwartz (@xaprb) November 15, 2017*

David Robinson: http://varianceexplained.org/r/ds-ml-ai/

Data Science

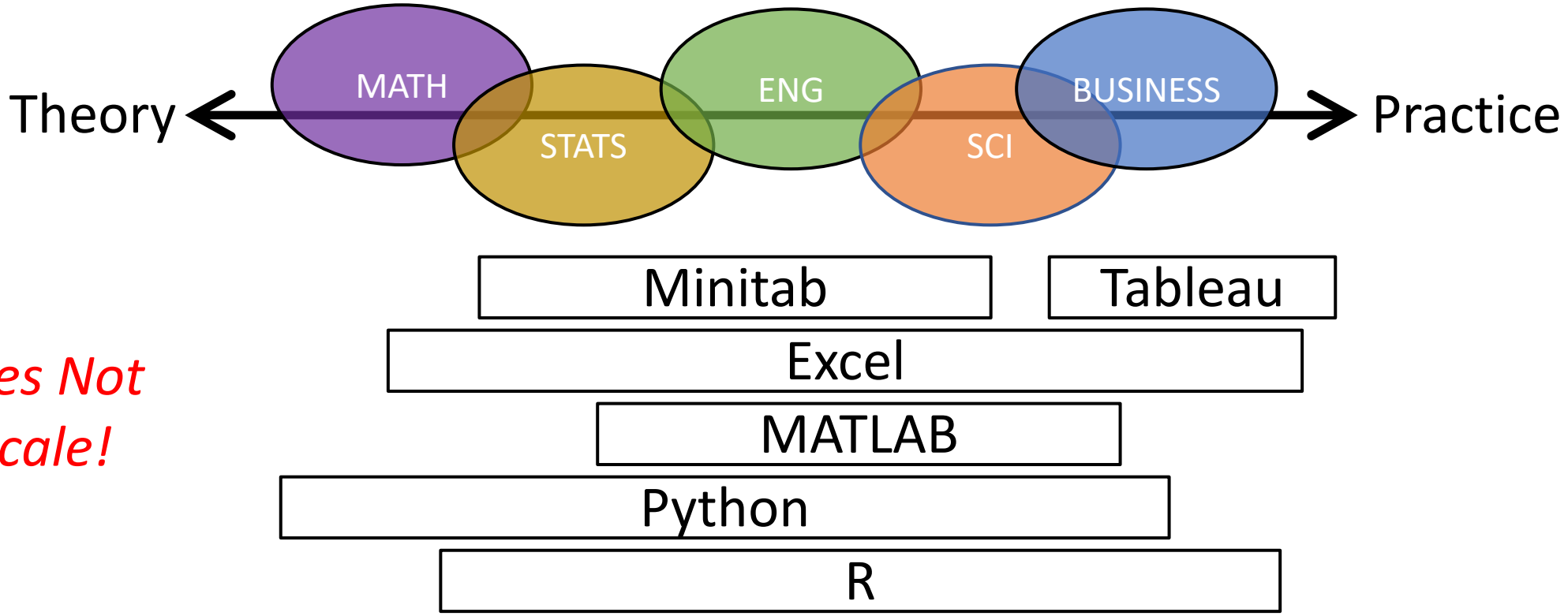
- Artificial Intelligence
- Big Data
- Statistical Learning
- Predictive Analytics
- Data Mining
- Machine Learning
- Pattern Recognition
- Deep Learning

Terms, usage, and interpretation vary

Overwhelmingly expansive and fast-moving field

Dan's Crude Model of Domains and Tools v0.01

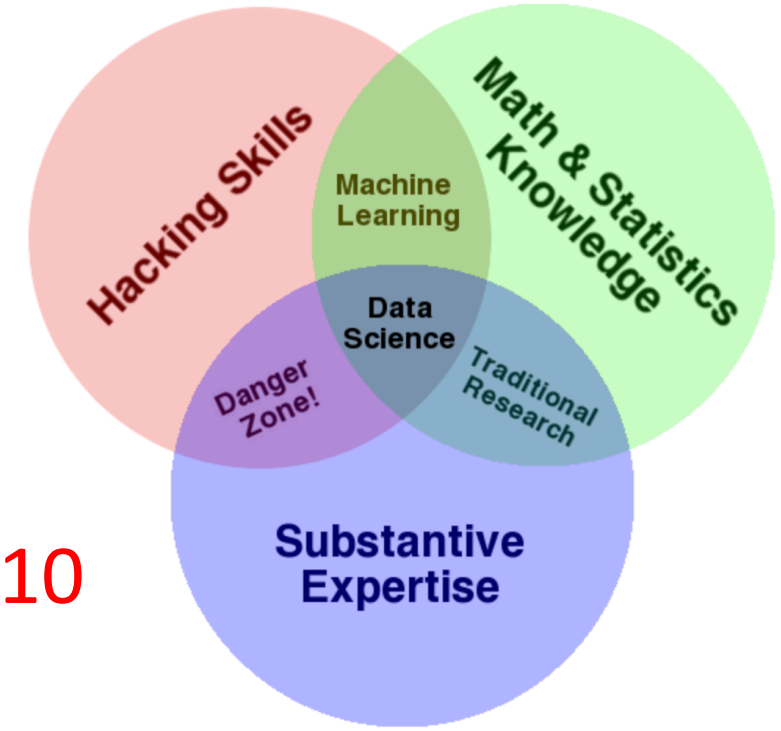
Broad, Multi/Interdisciplinary Interest



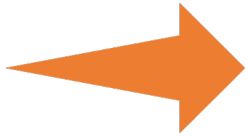
Figures Not To Scale!

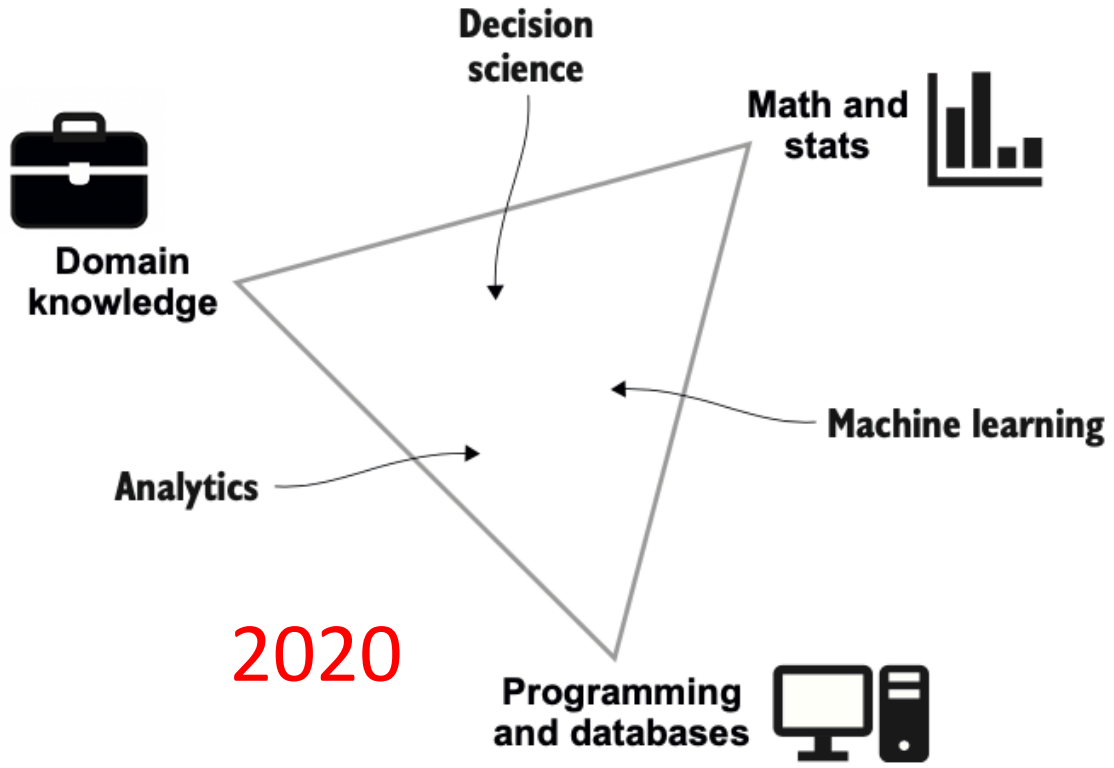
Source: I have only myself to blame for this slide.

The “Classic” Definition of Data Science



2010



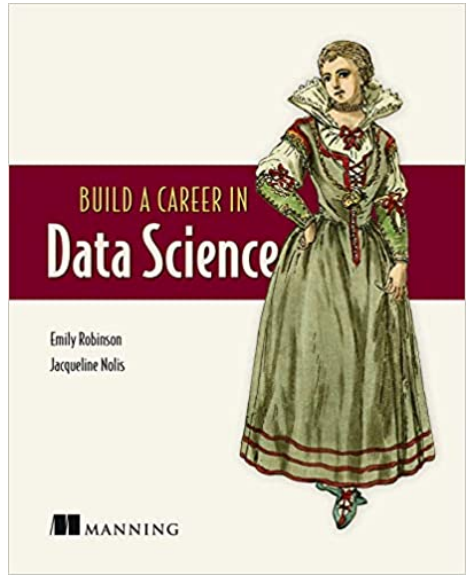


2020

drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Robinson, Emily, and Jacqueline Nolis. Build a Career in Data Science. Simon and Schuster, 2020.

Data Science in Broad Terms



- Components (Skills)
  - Math and Stats – methods related to data literacy
  - Programming & Databases – coding, engineering, carpentry
  - Domain Knowledge – subject matter expertise
- Applications (Jobs)
  - Analytics – create dashboards and reports that deliver data
  - Machine Learning – creates models that run continuously
  - Decision Science – creates analyses that create recommendations

Robinson, Emily, and Jacqueline Nolis. Build a Career in Data Science. Simon and Schuster, 2020.

> *please*
>
> *please*
>
> *please do not write that someone who trained an algorithm has "harnessed the power of AI"*
>
> *— Dave Gershgorn (@davegershgorn) September 18, 2017*

David Robinson: http://varianceexplained.org/r/ds-ml-ai/

Outcome-based…

David Robinson

Principal Data Scientist at Heap, works in R and Python.

- Data Science produces insights
  - Various types of insight – descriptive, exploratory, causal
  - Statistical inference, data visualization, and experiment design
- Machine Learning produces predictions
  - Various types of predictions – regression, classification
- Artificial Intelligence produces actions
  - Executed or recommended by autonomous agents
  - Includes game-playing, robotics / control theory, optimization, NLP, RL

David Robinson: http://varianceexplained.org/r/ds-ml-ai/

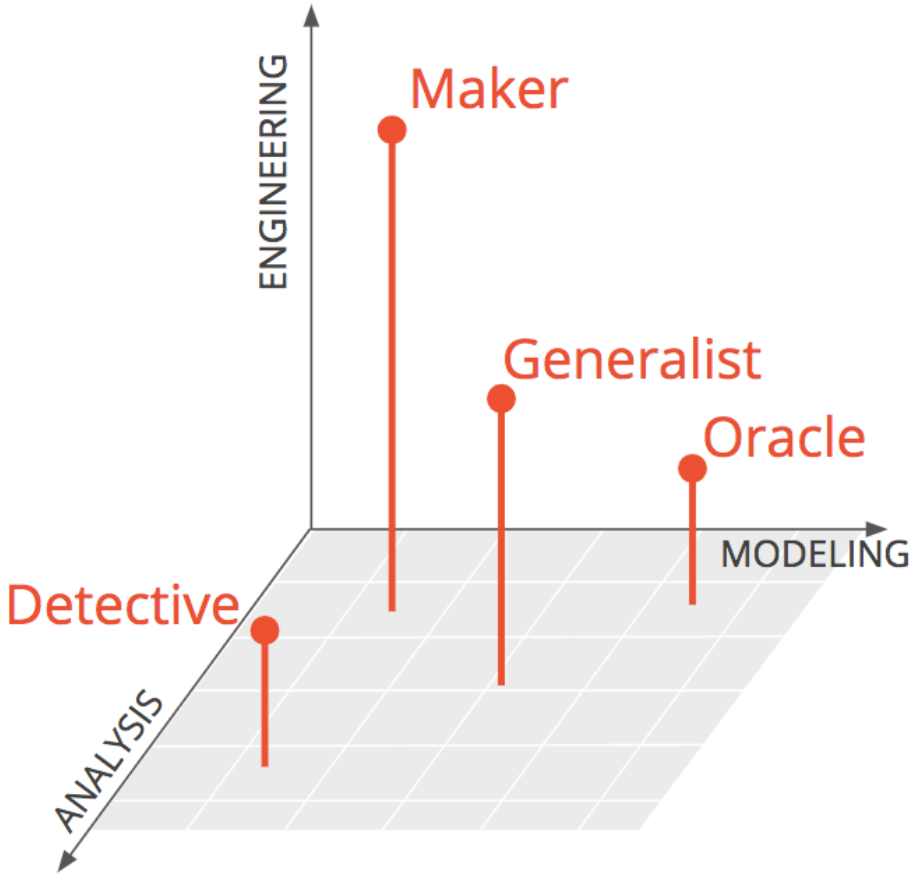
Example: Self-Driving Car

- Machine Learning
  - Object recognition model trained using many photos of streetside objects
  - System **predicts** the presence of stop signs
- Artificial Intelligence
  - Given varying road conditions and presence of a stop sign
  - Autonomous agent decides when / how to **act**, properly applying the brakes
- Data Science
  - Analyzing test data developers gain **insight** about the cause of false negatives
  - They generate a report summarizing their findings / recommendations

David Robinson: http://varianceexplained.org/r/ds-ml-ai/

Another Angle – Archetypes

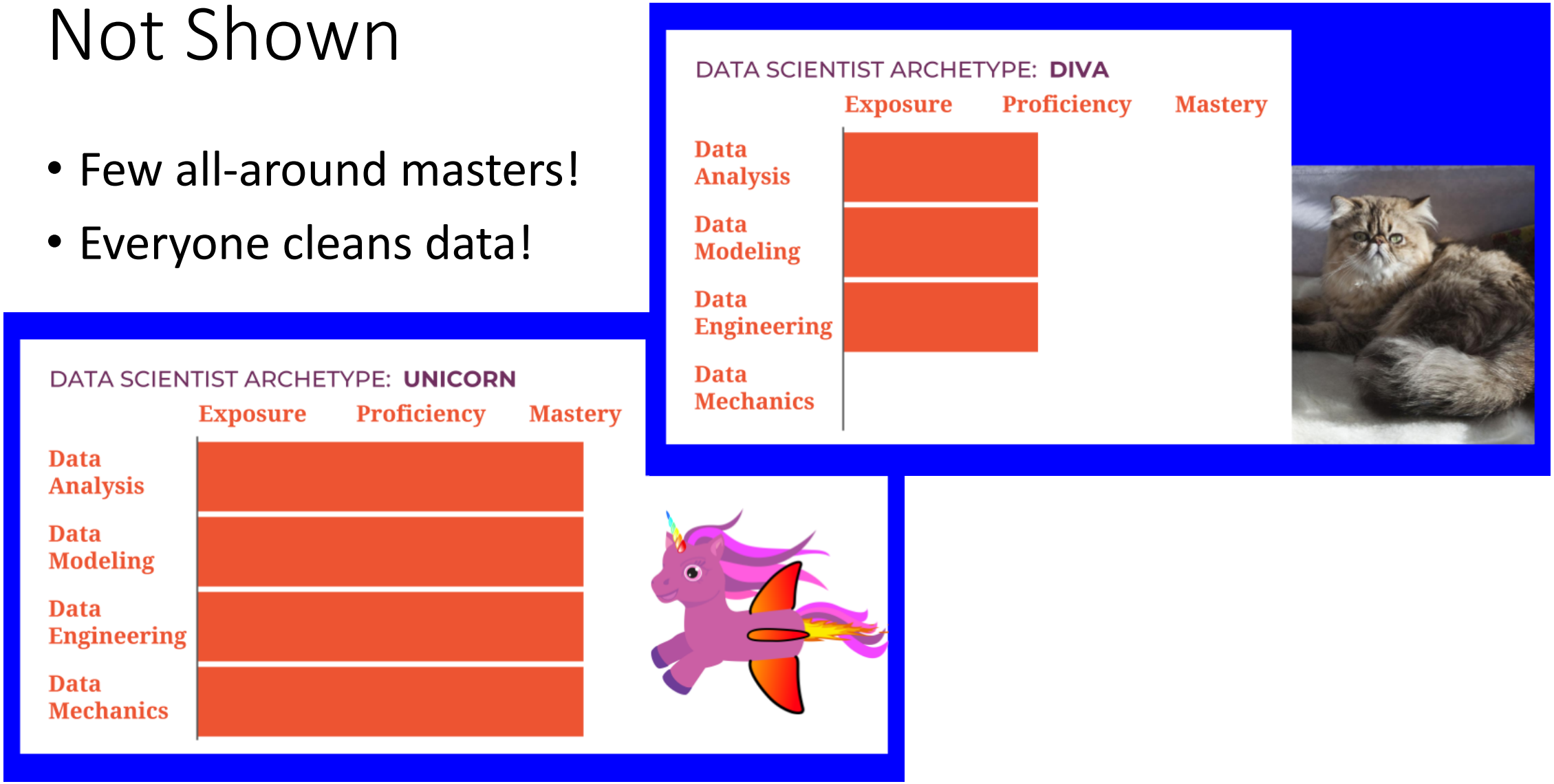
- Four Components of Data Science
  - Analysis – insights
  - Modeling – prediction
  - Engineering – deployment
  - Mechanics – cleaning / prep*
- Five Archetypes of Data Scientists
  - Generalist – proficient at everything
  - Detective – master of analysis
  - Oracle – master of modeling
  - Maker – master of engineering



https://e2eml.school/data_science_archetypes.html

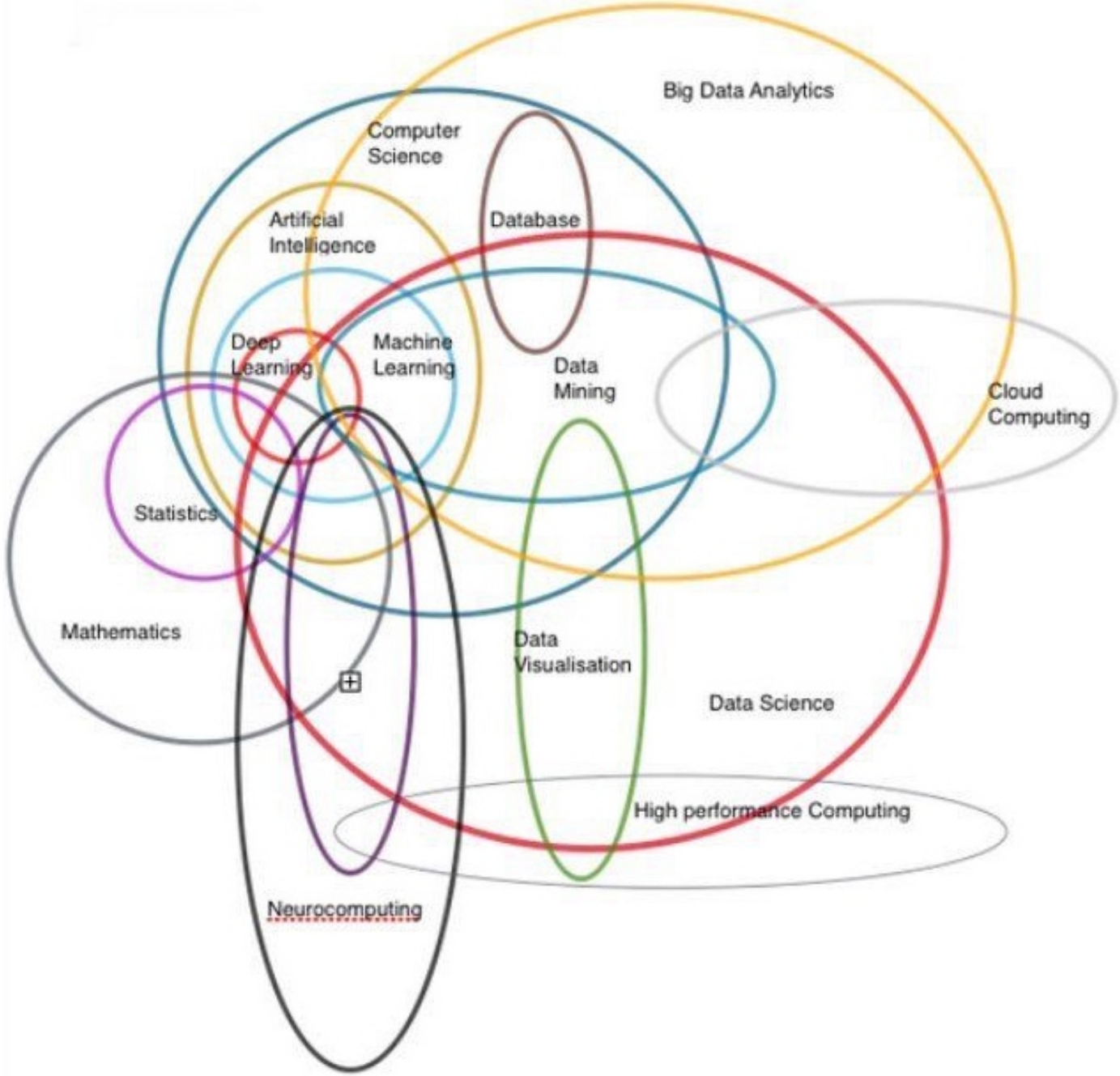
Not Shown

- Few all-around masters!
- Everyone cleans data!



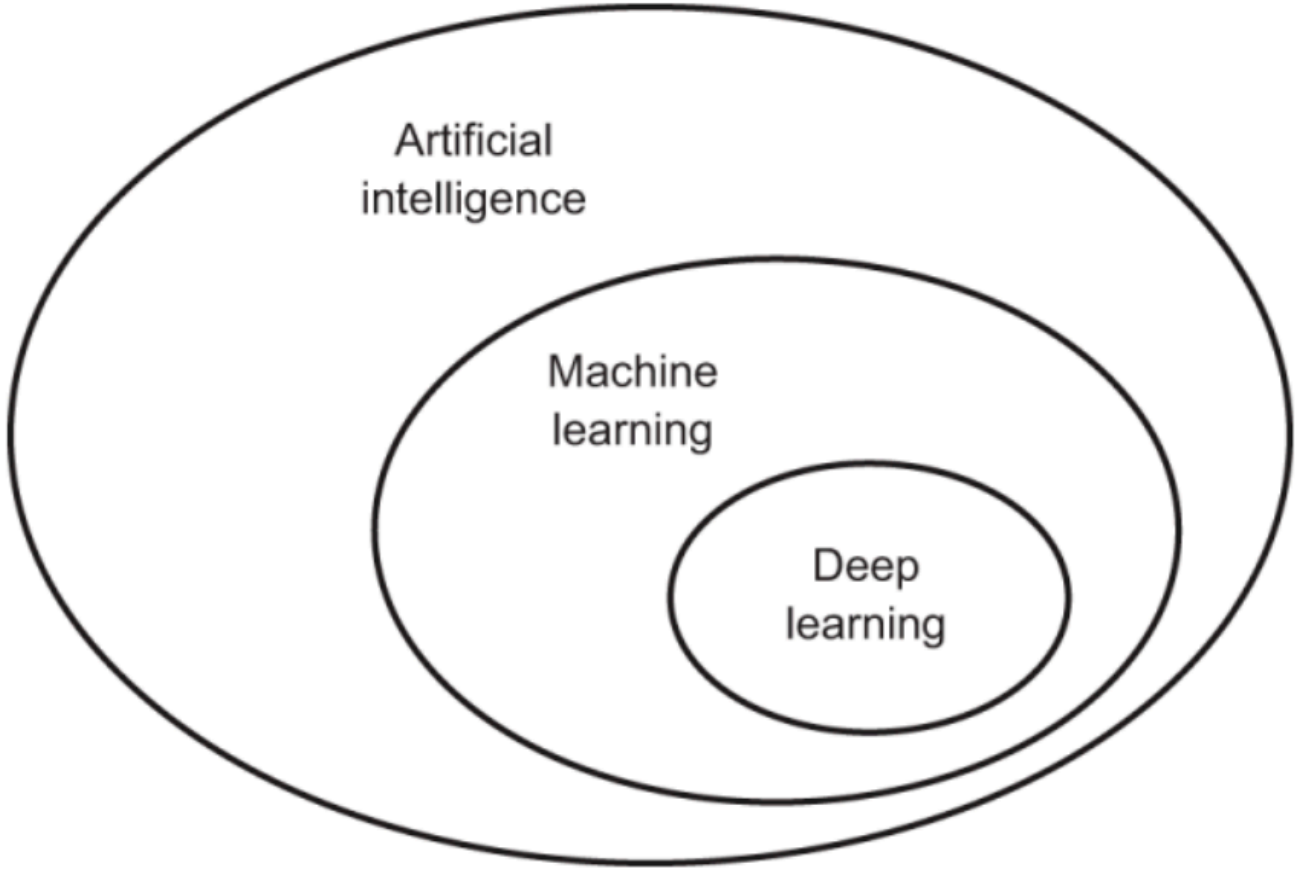
https://e2eml.school/data_science_archetypes.html

sometimes things get complicated



What is AI?

- I'm still not sure…
- Let's go with this:
  - It includes ML and DL
  - Actions → AI
  - Predictions → ML
  - Usually
- Ignore grander AI visions, claims, speculation



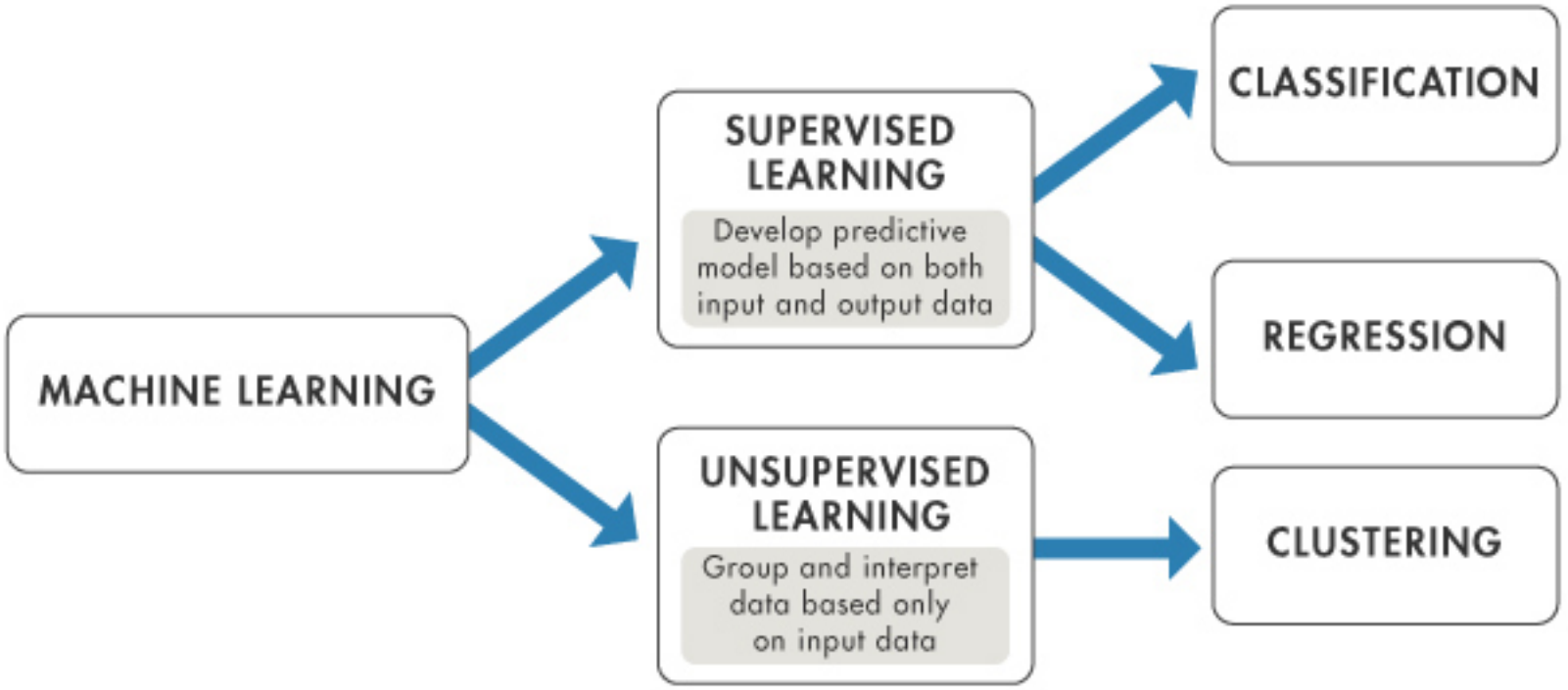
Most of what we think of as "AI" today based on Deep Learning methods

Much of AI's imagined potential remains distant

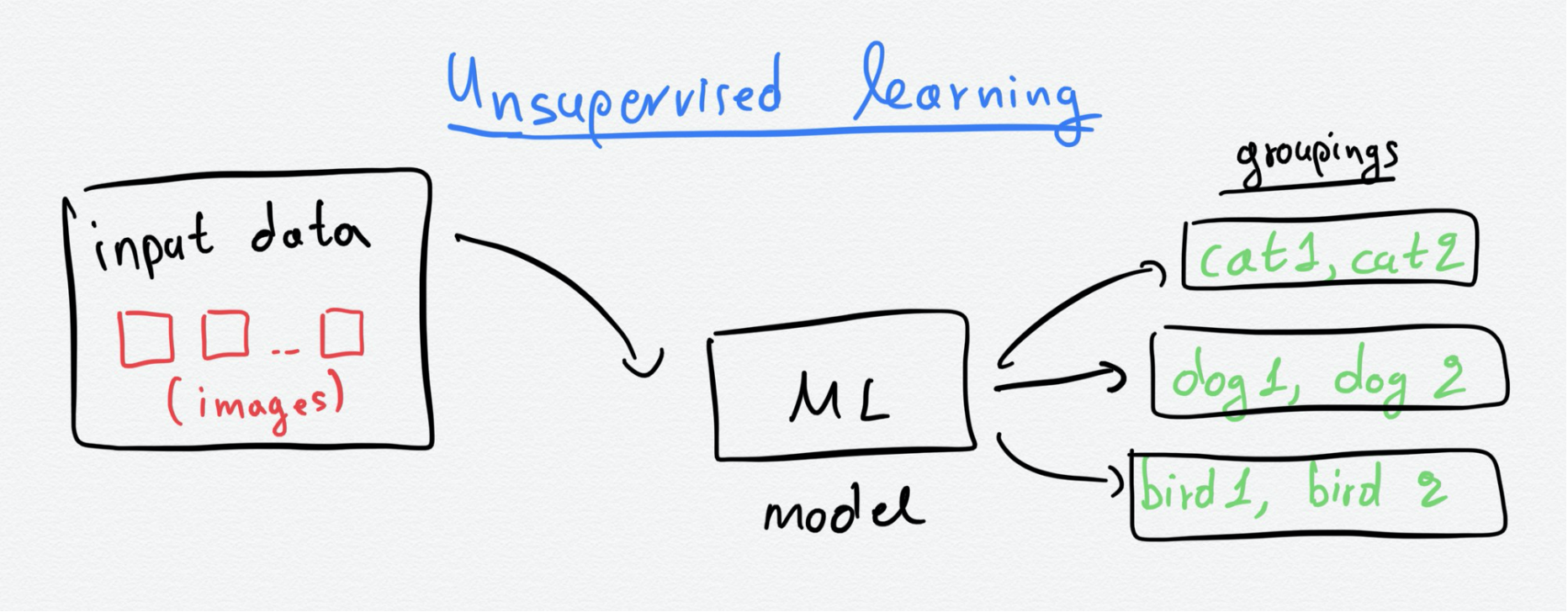
Machine Learning & Deep Learning are very real, here now, everywhere

Machine Learning

- Gives “computers the ability to learn without being explicitly programmed.” – Arthur Samuel, 1959
- Identify patterns in observed data

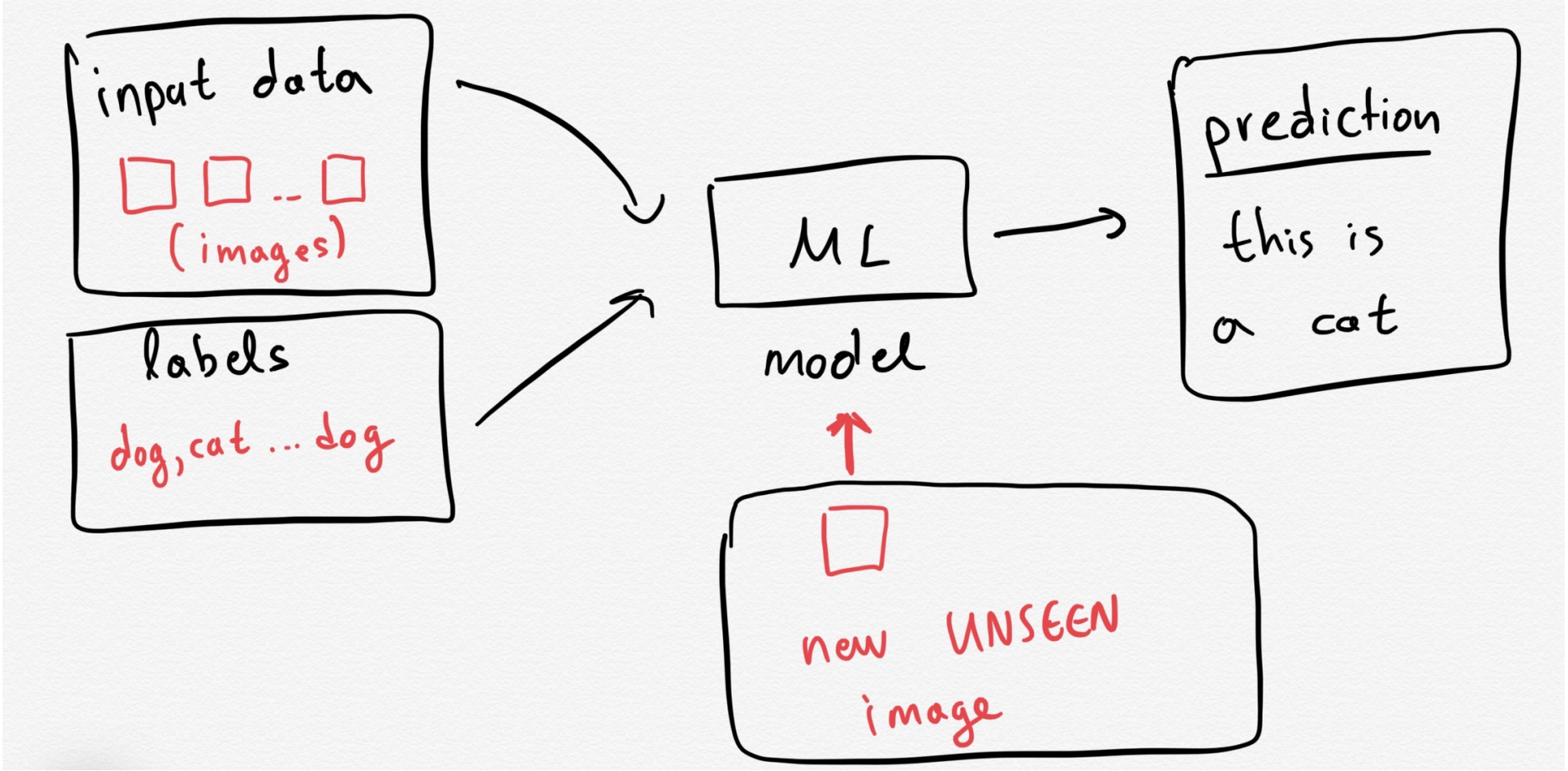


Unsupervised Learning



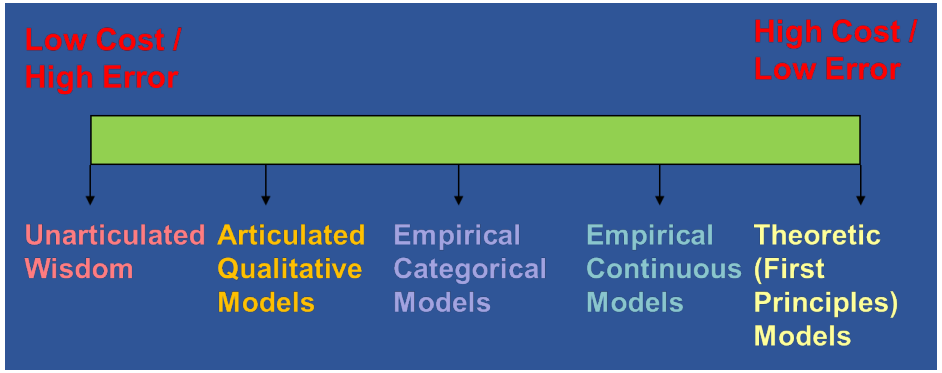
https://towardsdatascience.com/what-is-machine-learning-a-short-note-on-supervised-unsupervised-semi-supervised-and-aed1573ae9bb

Supervised Learning



https://towardsdatascience.com/what-is-machine-learning-a-short-note-on-supervised-unsupervised-semi-supervised-and-aed1573ae9bb

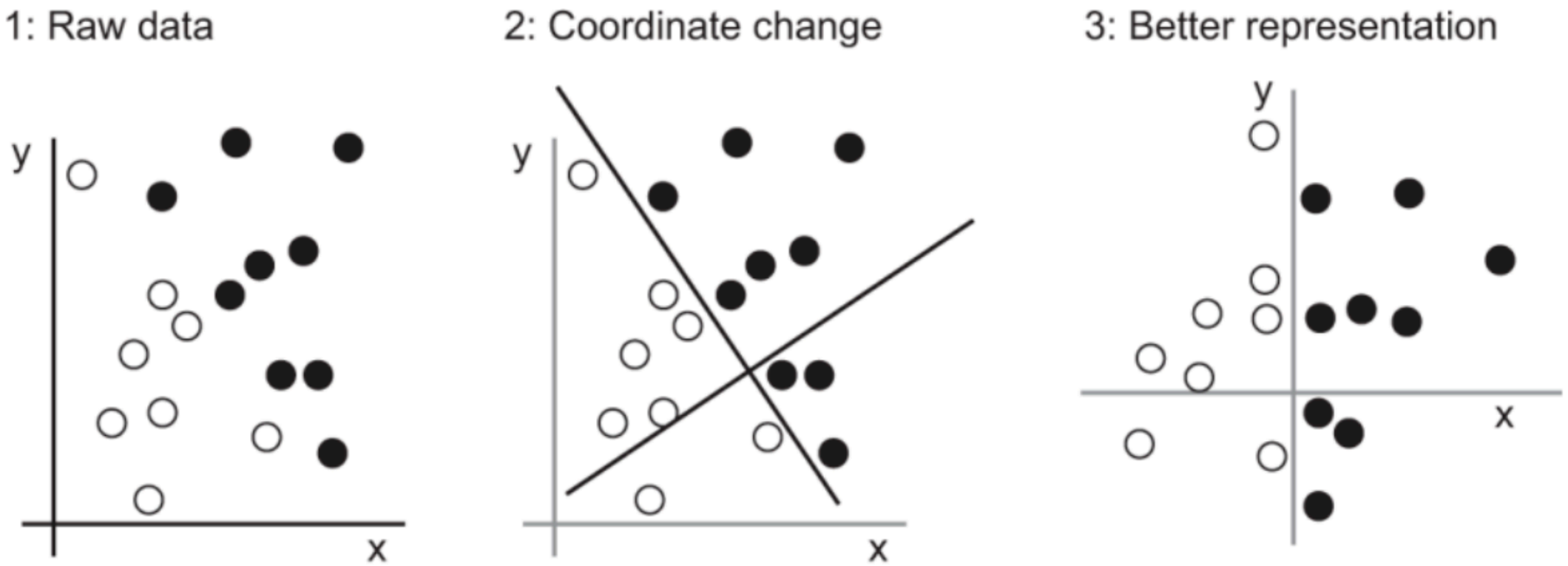
Machine Learning



- Linear Regression to Deep Neural Nets
- Ingredient technology
- “Macroscope” (inverted microscope) – sees things too big to view
  - Deep Neural Nets with tens of millions of parameters
  - Image data sets on the order of 1M x 1M+, video much larger
  - Entire USPTO archive (text and images), over 4M patents to 1976
  - Many data sets much larger
- Learns by finding statistical structure in training examples
  - Meaningful transformation / representations of data
- Largely empirical methods

Transformations / Representations of Data

Example: Classification

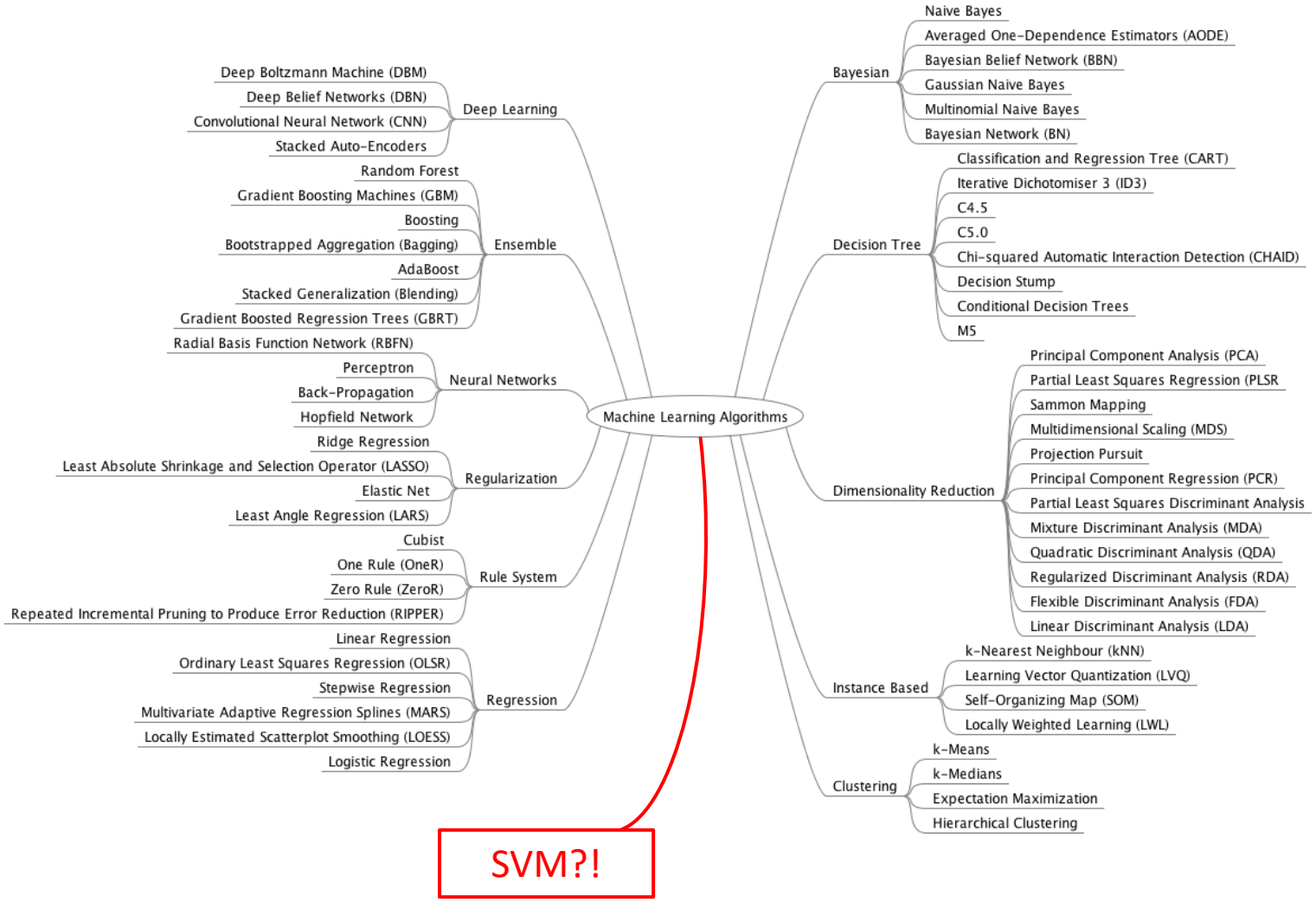


Chollet, François. *Deep Learning with Python*. Manning Publications Company, 2017.

Lots of ways to do it…

Best method? It depends.

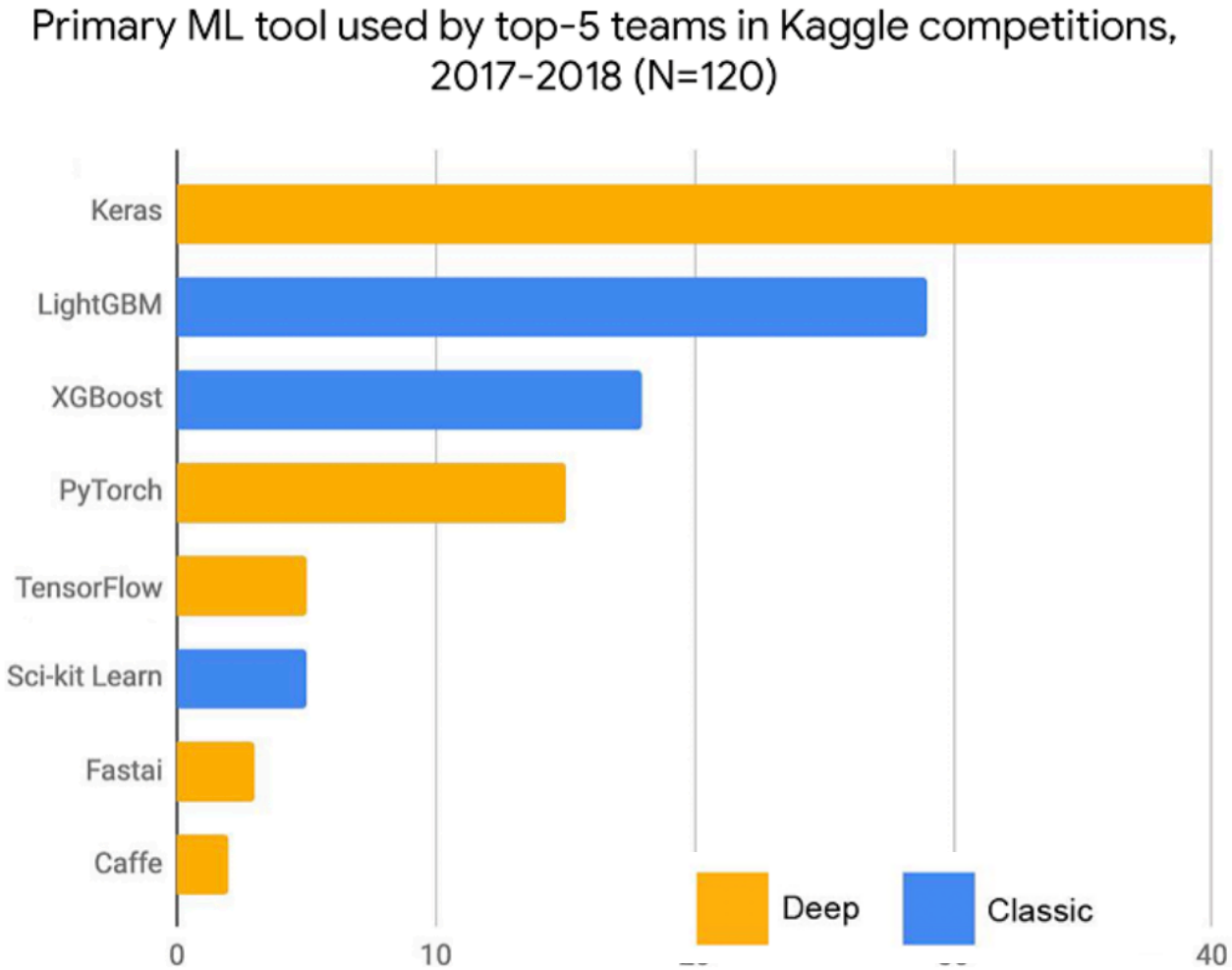
- Bayesian
- Decision Tree
- Dimens. Reduction
- Instance Based
- Clustering
- Regression
- Rule System
- Regularization
- Neural Networks
- Ensemble
- Deep Learning



“State of the Art”

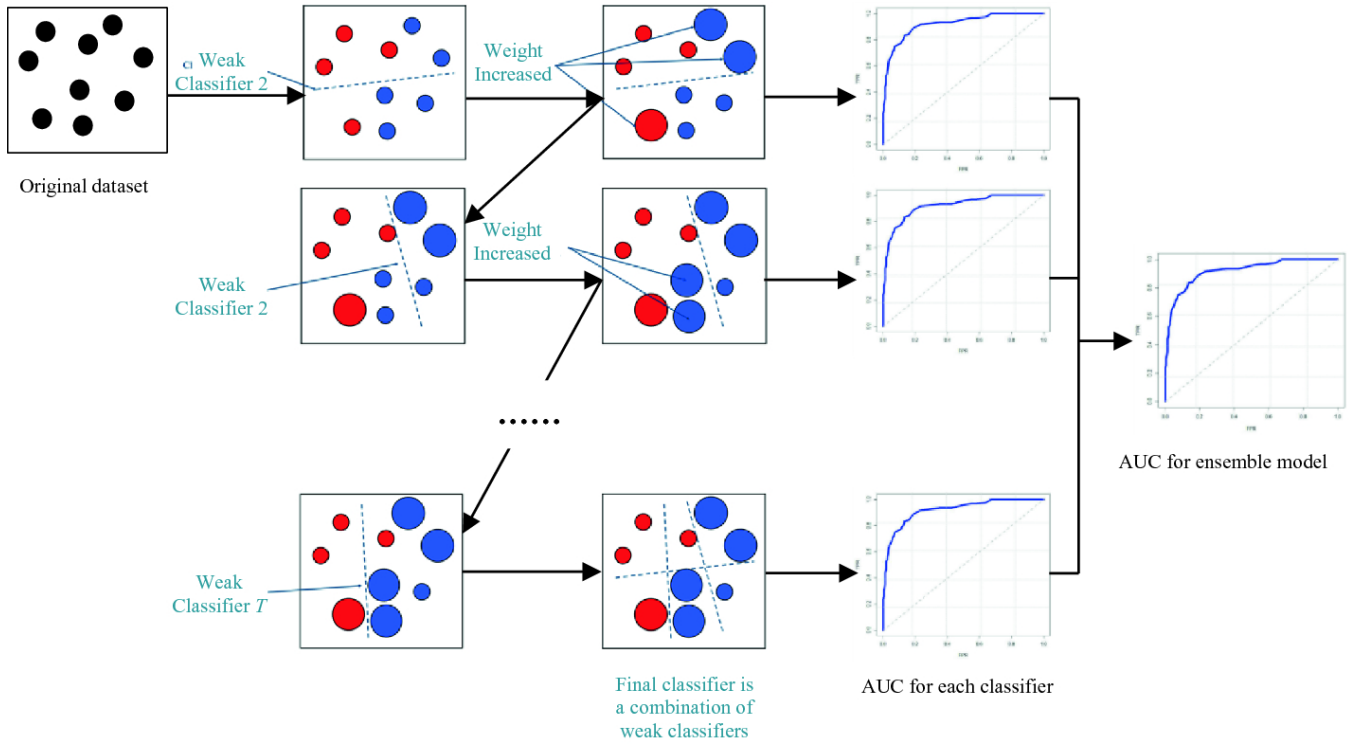
For Kaggle Contests, at least

- Gradient Boosting
  - LightGBM, XGBoost
  - For structured data
  - Python or R
- Deep Learning
  - Keras/TF, Fastai/PT
  - For perceptual problems
  - Python



Gradient Boosting in 1 Slide

- Series of decision trees*
- Each improved by prior
- Weights adjusted based on ease of classification
- Repeat and combine results

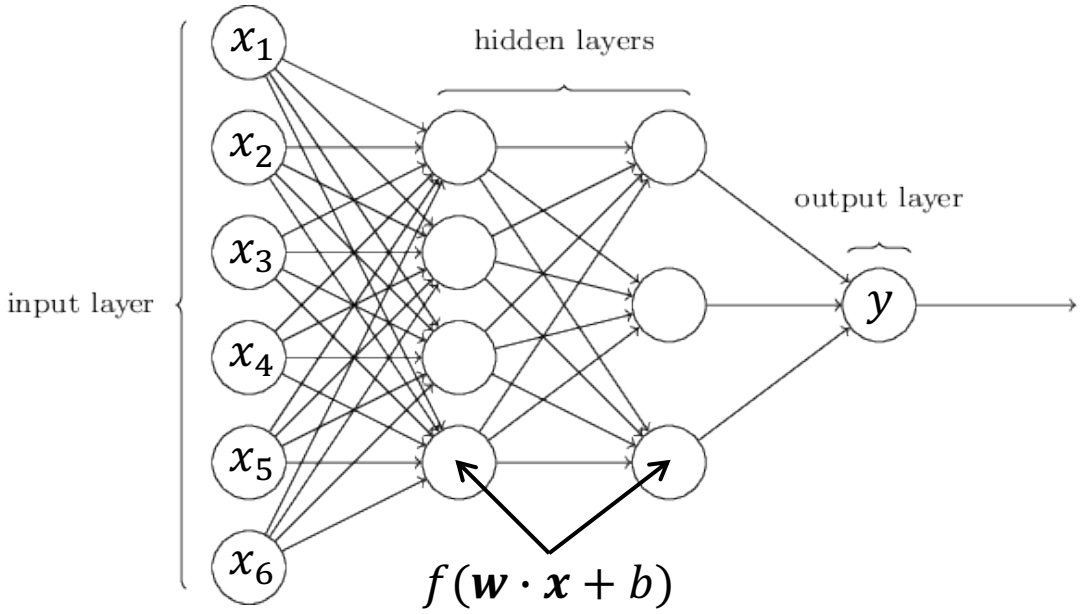


*Decision trees can be thought of as giant "if-then" structures converting inputs to outputs based on features

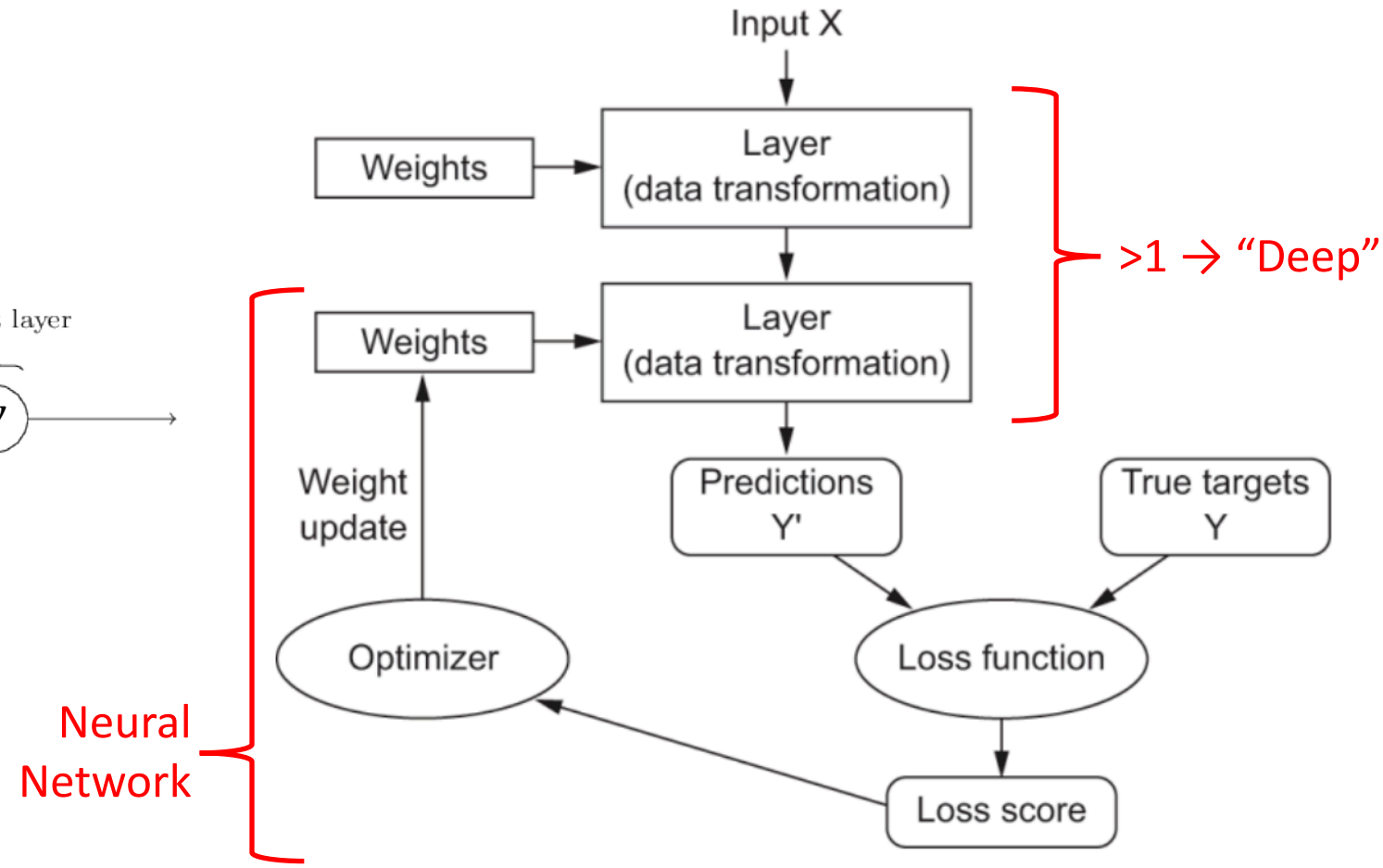
https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know/

Deep Learning in 1 Slide

Bonus: Neural Network!



$f(\cdot)$ is the activation function
Non-linear: sigmoid, tanh, etc.



http://neuralnetworksanddeeplearning.com/index.html

Chollet, François. *Deep Learning with Python*. Manning Publications Company, 2017.

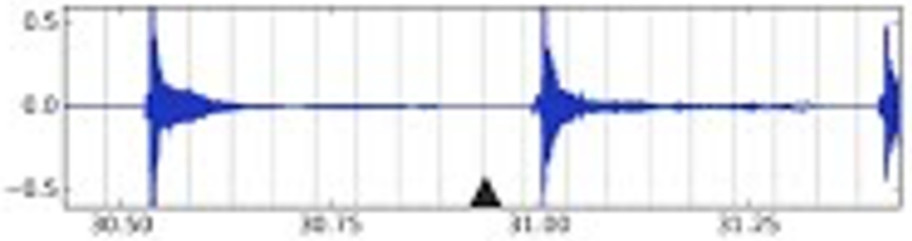
Common Current ML/DL Use Cases

- Natural Language Processing
  - Google Translate, Siri/Cortana/Alexa, Auto-correct
- Recommendation Systems
  - Netflix, Amazon, Facebook
- Customer Relationship Management
  - Direct marketing, mobile advertising, chatbots
- Finance
  - Credit score, loan approval, fraud detection ($100B-$1T), algorithmic trading
- Image Recognition
  - Pose detection, facial recognition, medical image processing

Tip of the Iceberg – ML/DL is Everywhere!

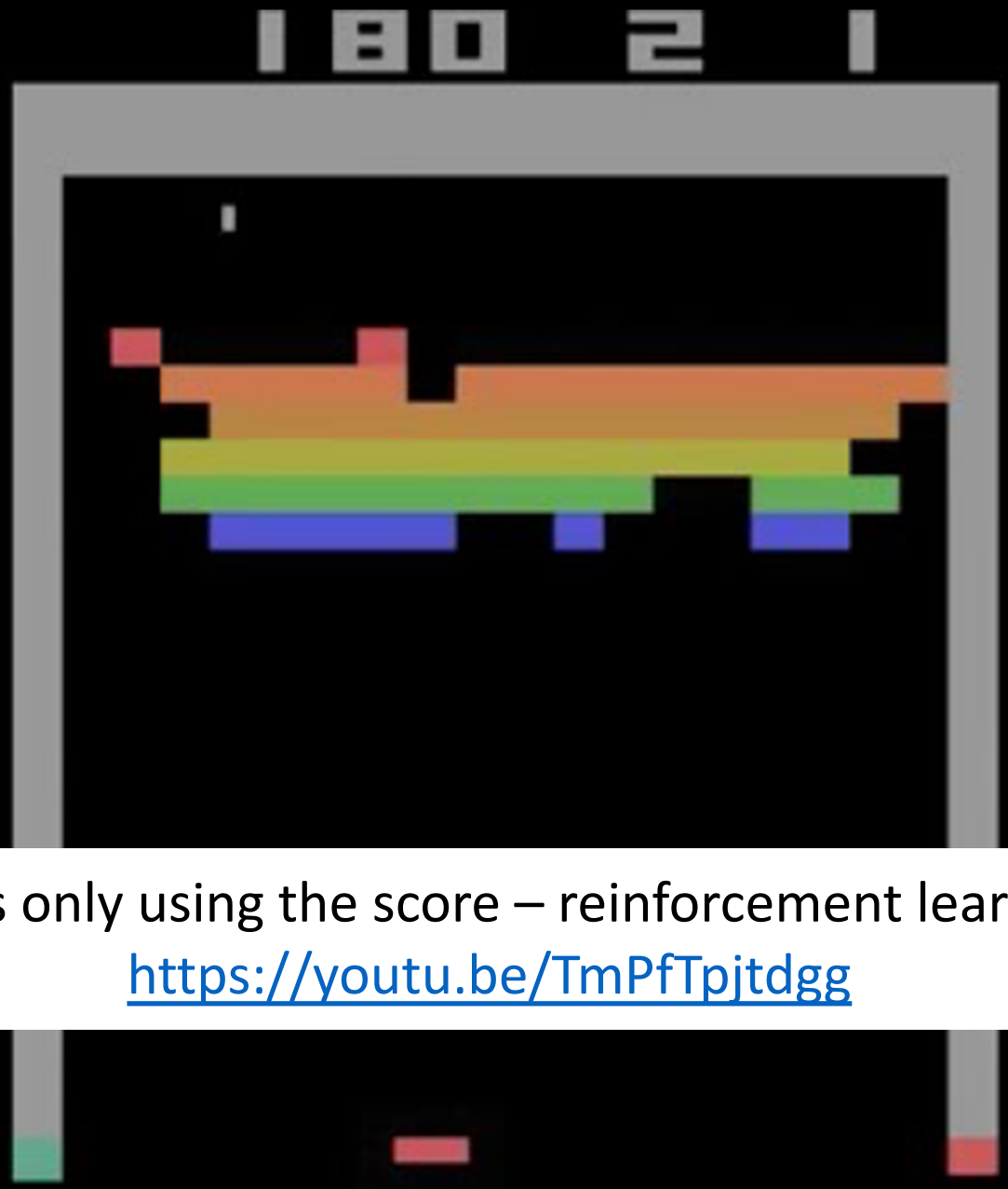
https://youtu.be/ayPqjPekn7g

Silent video

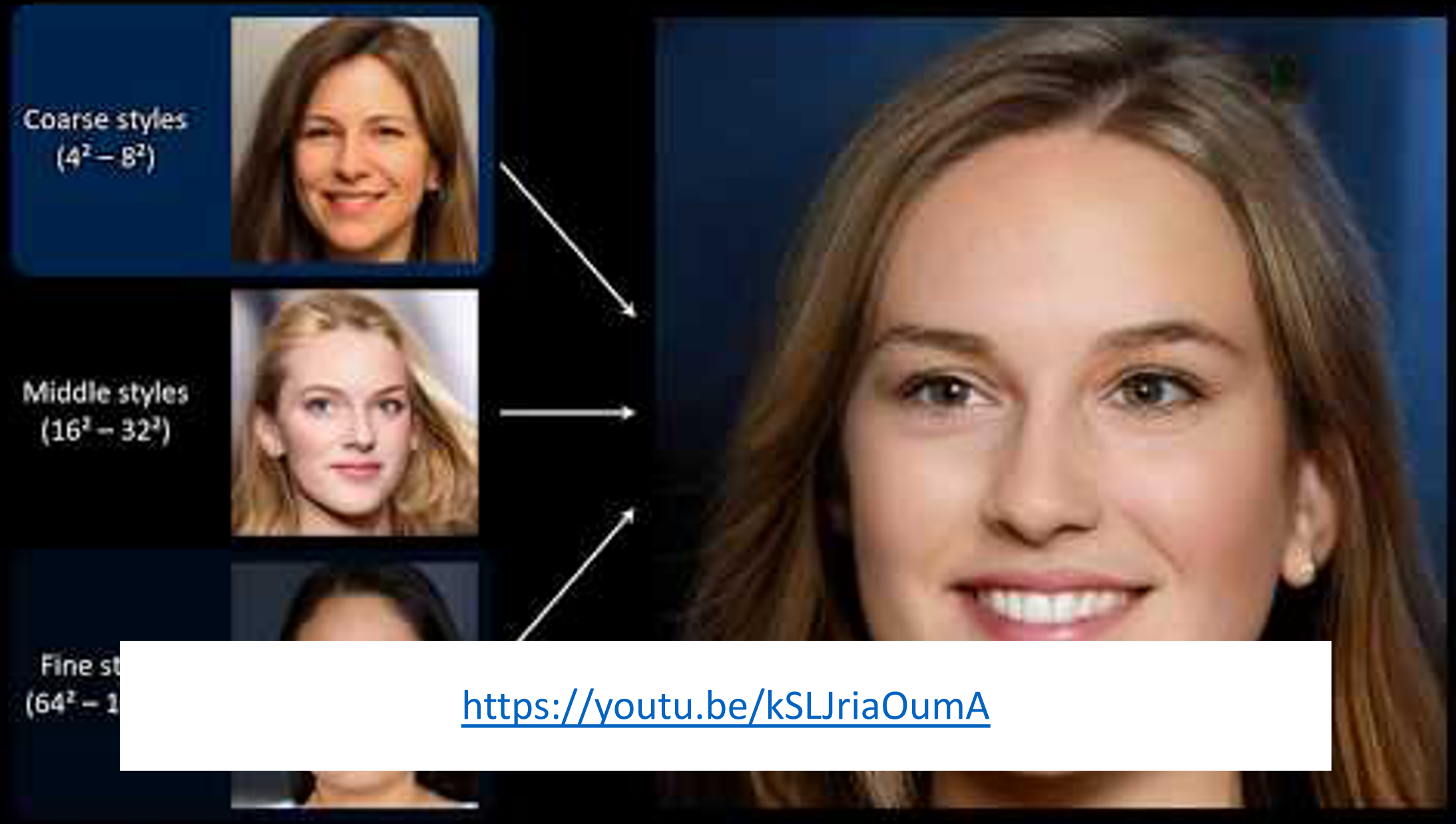


Predicted soundtrack

https://youtu.be/0FW99AQmMc8

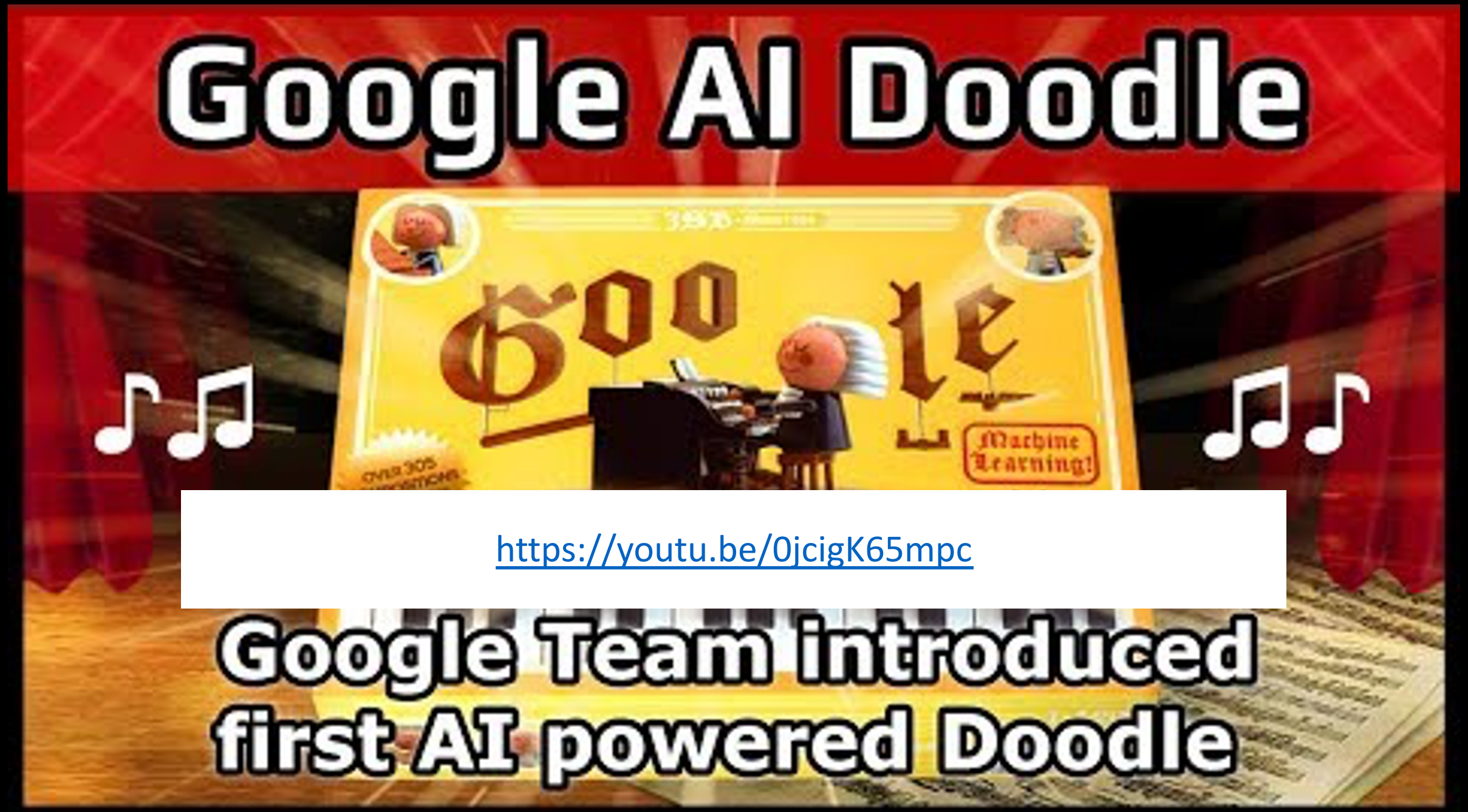


Trains only using the score – reinforcement learning
https://youtu.be/TmPfTpjtdgg



https://youtu.be/kSLJriaOumA

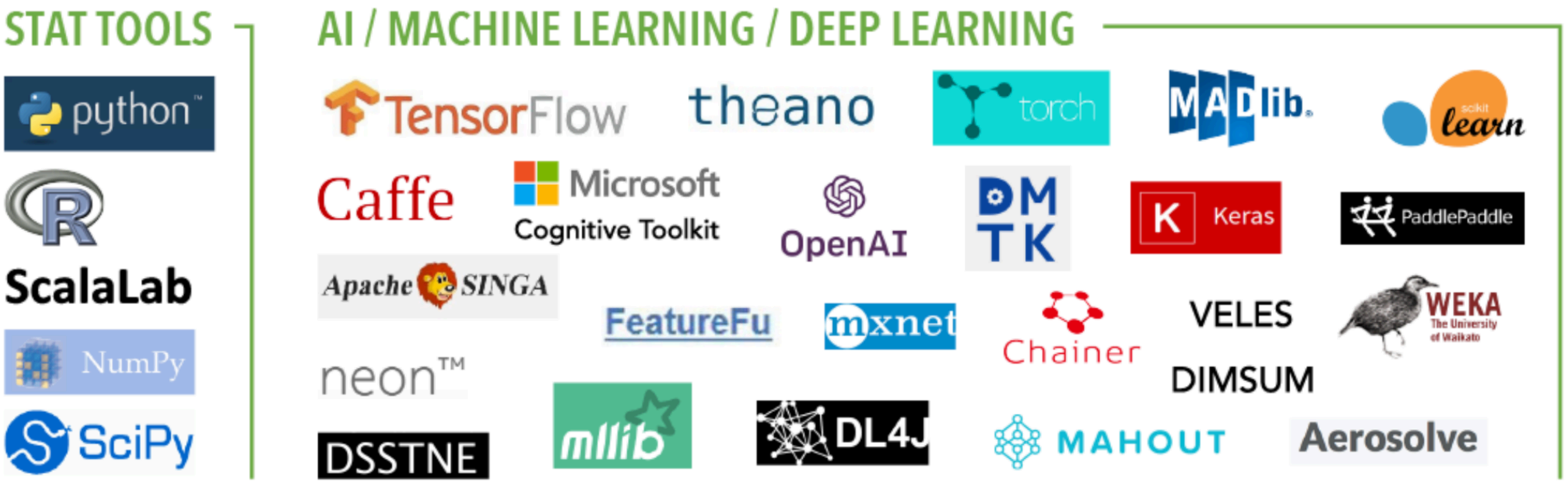
https://youtu.be/LBd5FZqhUVk


Google AI Doodle
Machine Learning!
https://youtu.be/0jcigK65mpc
Google Team introduced first AI powered Doodle

https://youtu.be/DjERMBnvTEE

Why Now?

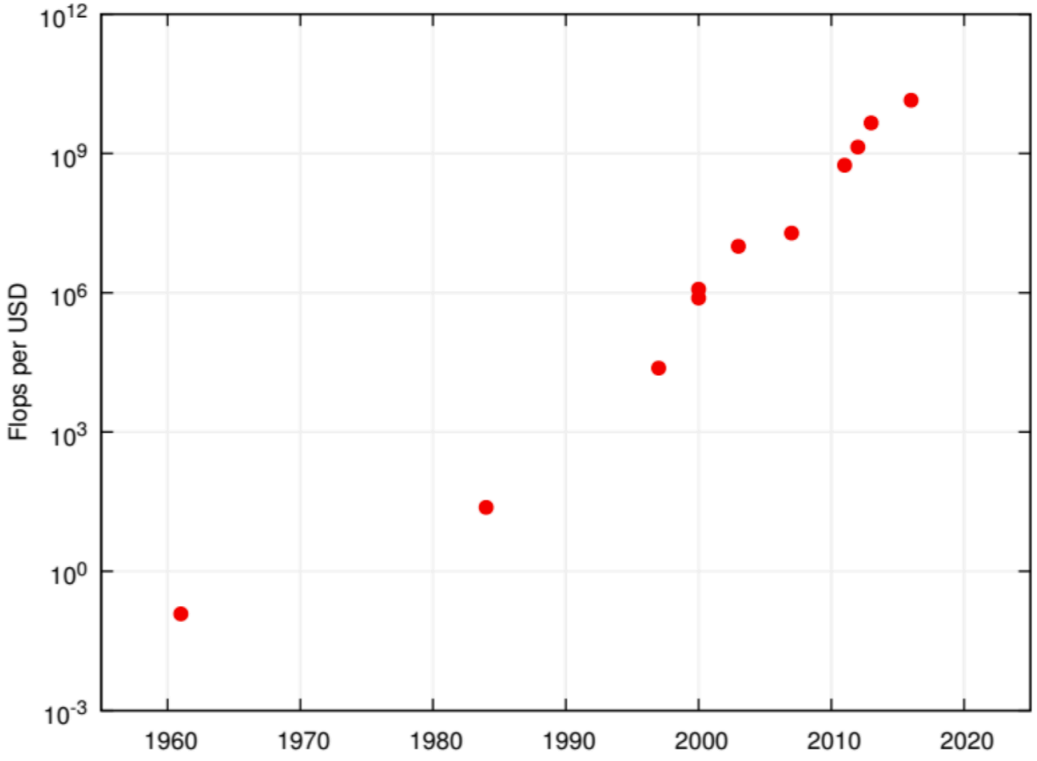
- 50+ years of research
- Algorithm / SW dev
- Huge Investments
- Democratization



BIG DATA & AI LANDSCAPE 2018

Why Now?

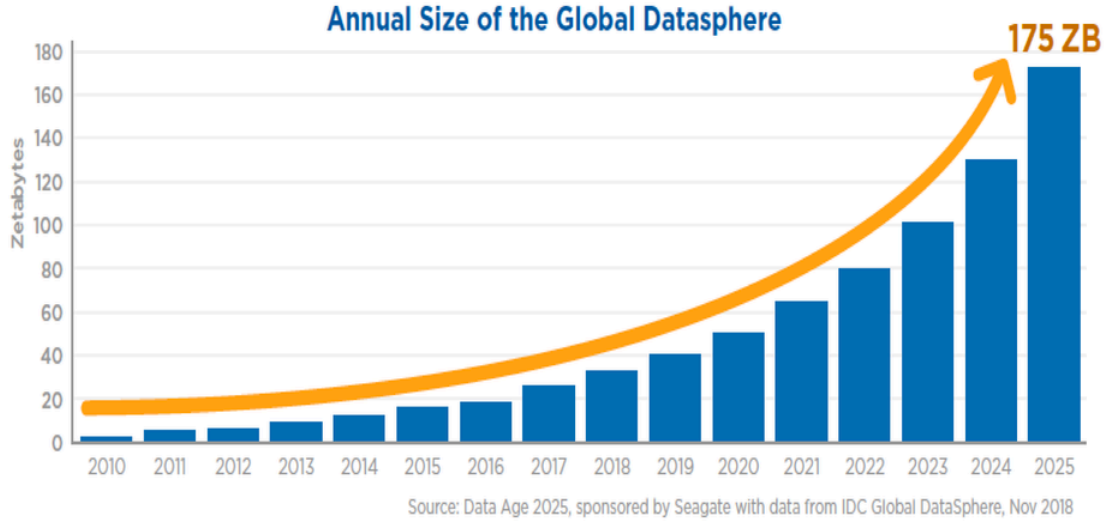
- Compute power
  - GPU – graphics processing unit
  - Originally developed for 3D graphics
  - Massively parallel matrix operations
  - Orders of magnitude better performance
- Deeper Blue (1997)
  - 11.38 GFLOPS
  - ~ $100M
- NVIDIA GTX 1080 (2016)
  - 8,873 GFLOPS
  - $499 MSRP
  - 150 million times more GF / $

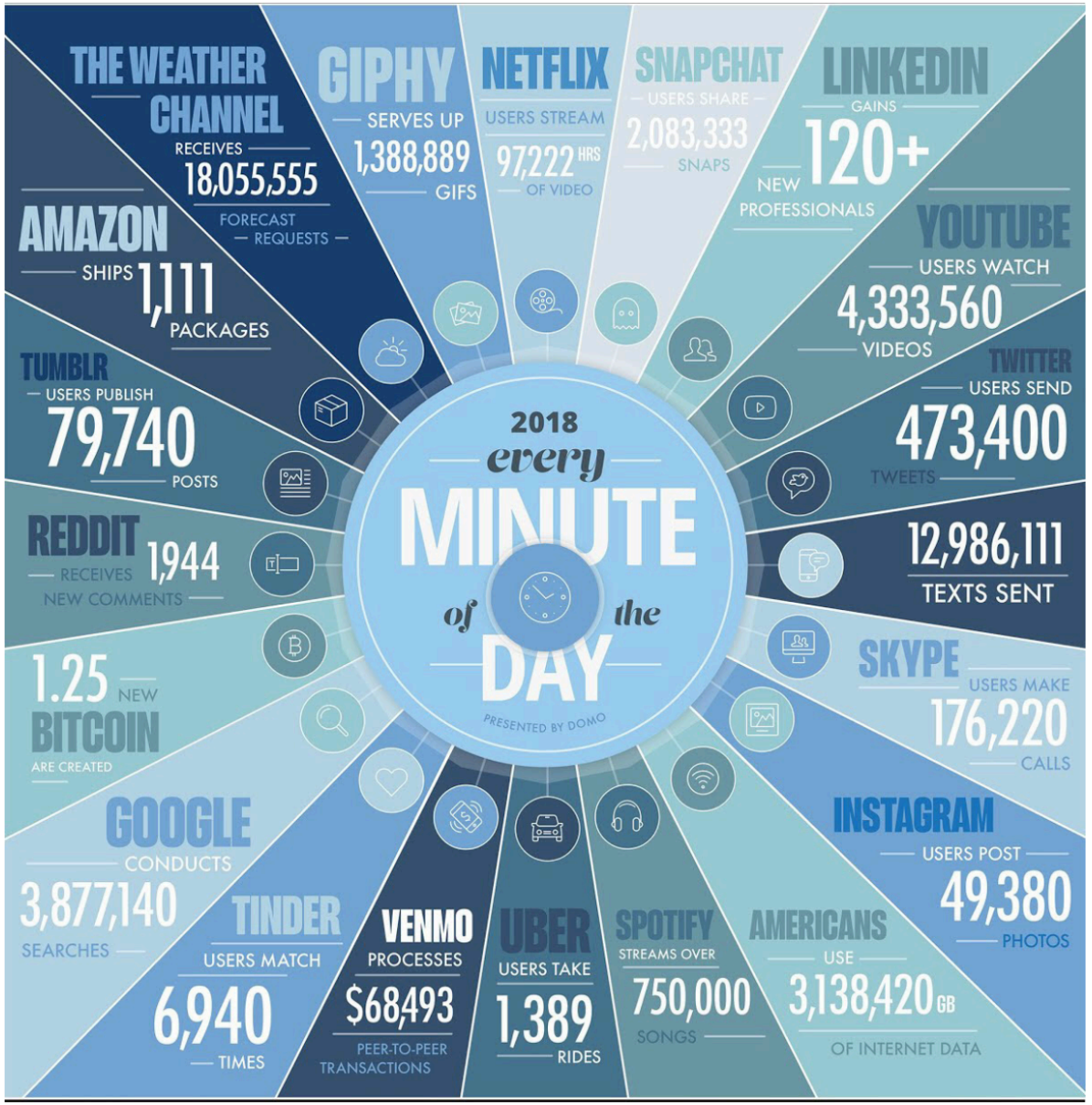


	TFlops (10^{12})	Price	GFlops per $
Intel i7-6700K	0.2	$344	0.6
AMD Radeon R-7 240	0.5	$55	9.1
NVIDIA GTX 750 Ti	1.3	$105	12.3
AMD RX 480	5.2	$239	21.6
NVIDIA GTX 1080	8.9	$699	12.7

Why Now?

- Access to data
  - 175 zettabytes annually by 2025
  - 1 zettabyte = 1 trillion gigabytes
- The Internet
- Infrastructure to facilitate
- Instrumentation of everything





https://www.digitalinformationworld.com/2018/06/infographics-data-never-sleeps-6.html

Benefits of Data Science

- Domain independent technology; “metascience”
- Informs decision making
- Empowers organizational learning
- Improves operational efficiency
- Leverages underutilized by-product of work
- Delivers actionable results
- Automates scientific discovery

… a user’s data can be purchased for about half a cent, but the average user’s value to the Internet advertising ecosystem is estimated at $1,200 per year.

Credit: Predictive Analytics, Eric Siegel, p. 54

Limitations and Pitfalls

- Accurate prediction (extrapolation) is generally not possible.
  - “Prediction is very difficult, especially if it’s about the future.” – N. Bohr
  - High value from relatively low predictive power; targeted optimization
- Does not answer WHY or HOW.
  - Correlation does not imply causation; many models opaque, empirical
  - Value comes from the prediction, not understanding cause
- Vast search / Multiple comparisons trap
  - Possibility of being fooled by randomness - real trend or random artifact
  - Importance of domain knowledge and disciplined research
- Bias / Variance tradeoff
  - Fit vs Predictive Power

Takeaways

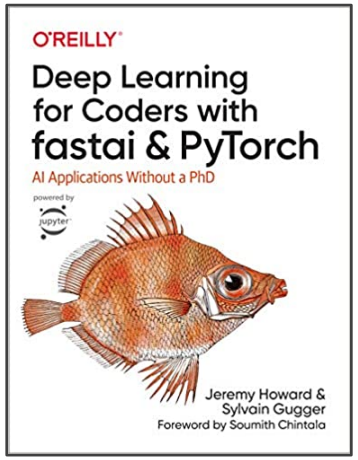
- Data Science is a broad, fast-moving field, with hype and confusion
- The “promise” of AI cannot be met by current or near future tech
- We are surrounded by current use cases, many more emerging
- Its recent growth is fueled by data, compute, algorithms, sw, and $$$
- Leverages existing data to improve operational efficiencies
- Identifies unexpected connections but does not explore causation
- It is not fool-proof and requires expert oversight
- Cannot be fully explained (even introduced) in one short talk…

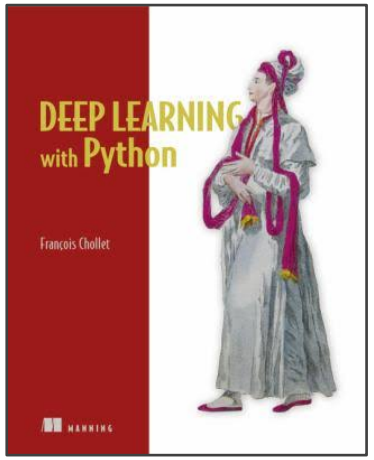
Resources


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by jupyter
Aurélien Géron
O'REILLY
R for Data Science
IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA
Hadley Wickham & Garrett Grolemund
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
EILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas
Andriy Burkov
THE HUNDRED-PAGE MACHINE LEARNING BOOK
DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Additional Resources for Deep Learning

- http://neuralnetworksanddeeplearning.com/index.html – free, online only, starts with writing a simple backprop NN from scratch in Python
- https://www.manning.com/books/deep-learning-with-python - build models in Keras and Tensorflow, written by creator of Keras, 2nd edition coming soon!
- https://course.fast.ai – alternative to Keras built on PyTorch, all work is done inside Jupyter Notebooks





Thank You.

Contact Information:

Dan O'Leary
dan.oleary@auburn.edu
Blog / Portfolio / Links: bit.ly/aboutdjo

DATA SCIENCE SOCIETY
OF AUBURN