STAT 5630, Fall 2019

Introduction

Xiwei (Denny) Tang, Ph.D. <xt4yj@virginia.edu>

University of Virginia
August 27, 2019

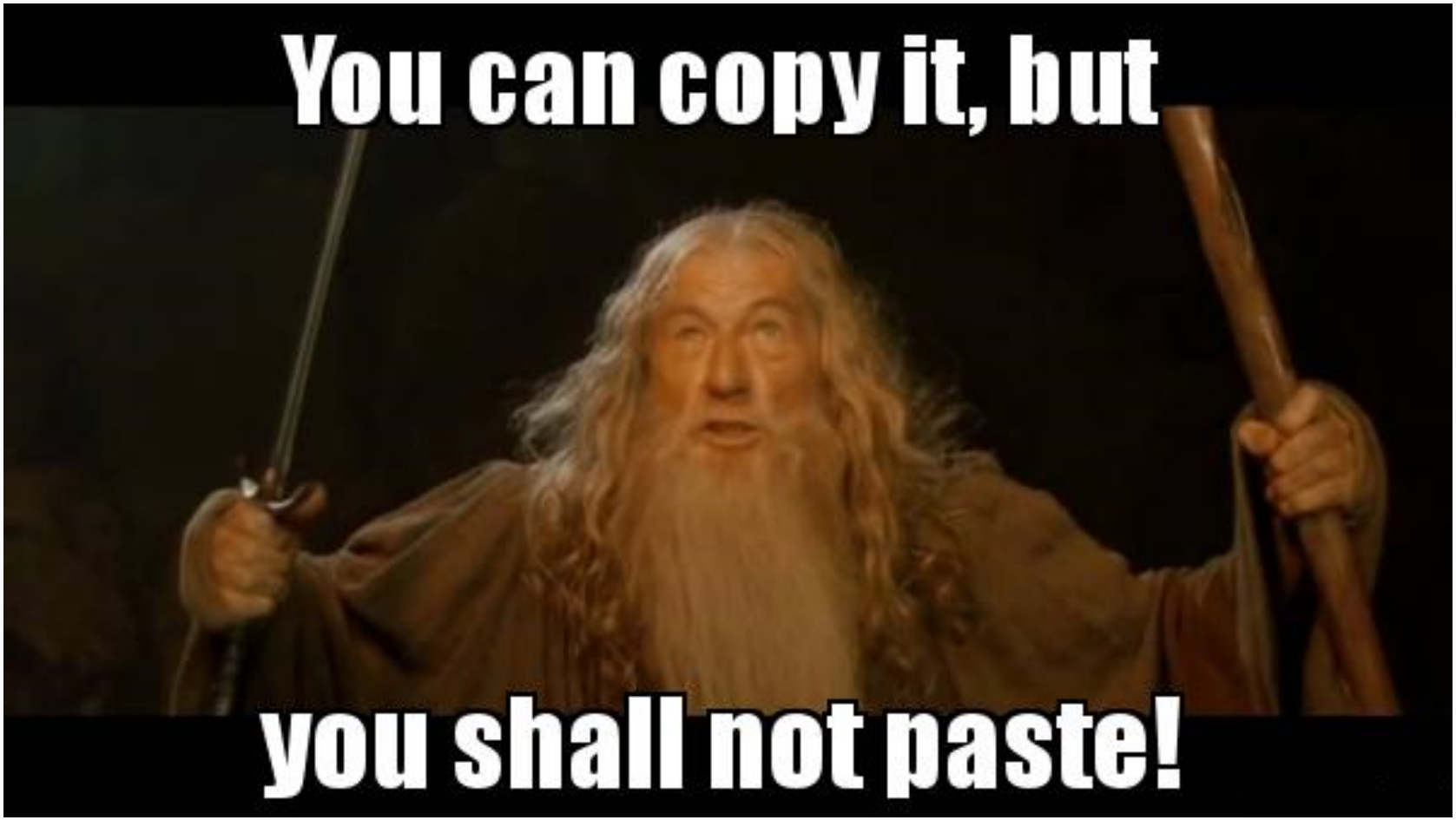
Welcome to STAT 5630

- Instructor: Xiwei Tang <xt4yj@virignia.edu>
  - Office: B004, Halsey Hall
  - Office hours: THR 1:00 - 3:15 PM or by appointment
- Teaching Assistant: Tianyuan Zhou <tz8hu@virginia.edu>
- Course website:
  - All information will be posted on UVaCollab
  - Resource: Lecture notes, sample codes, relevant materials
  - Assignment: homework and project information
  - Announcement
- Prerequisite: STAT 5120 or STAT 6120; STAT4630

- Textbook:
  - Recommended: [HTF] Hastie, T., Tibshirani, R. Friedman, J.The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (2nd edition) Springer. (free PDF)
- Other options:
  - [JWHT] James, G., Witten, D., Hastie, T. Tibshirani, R.: An Introduction to Statistical Learning. Springer. (free PDF)
  - [B] Bishop, C.: Pattern Recognition and Machine Learning. Springer, 2006.
  - Aurelien Geron: Hands-On Machine Learning with Scikit-Learn TensorFlow
  - Larry A. Wasserman: All of Statistics: A Concise Course in Statistical Inference.
- Programming language:
  - R is preferred. Most of my examples will be presented using R
  - Online R Tutorial , An Introduction to R

- **Grading**: Homework: 40% + Mini Project: 20%+ Course Project:40%

- Homework
  - Around Four sets of homework will be assigned on UVaCollab
  - Submitted in both hardcopy (report) to my mailbox in Halsey Hall and well-organized code to UvaCollab.
  - No late homework accepted. For exceptions, please notify the instructor for approvement in advance!
  - Simplify R output: No excessively long output. Total page limit. It should look like a report! Not just R outputs.
  - You are encouraged to discuss with anyone, but DO NOT copy others'.


You can copy it, but
you shall not paste!

- More about Project: 40% = Presentation+Report.
  - The course project is a team work and each team consists of up to FOUR students. You are free to team up by yourselves, but have to inform the instructor of their group members in advance.
  - Keep in a good timeline! Finding the topic/dataset as early as you can.
  - A proposal is required by the mid-October. Project presentation is due in late November, and the final report is due thereafter.

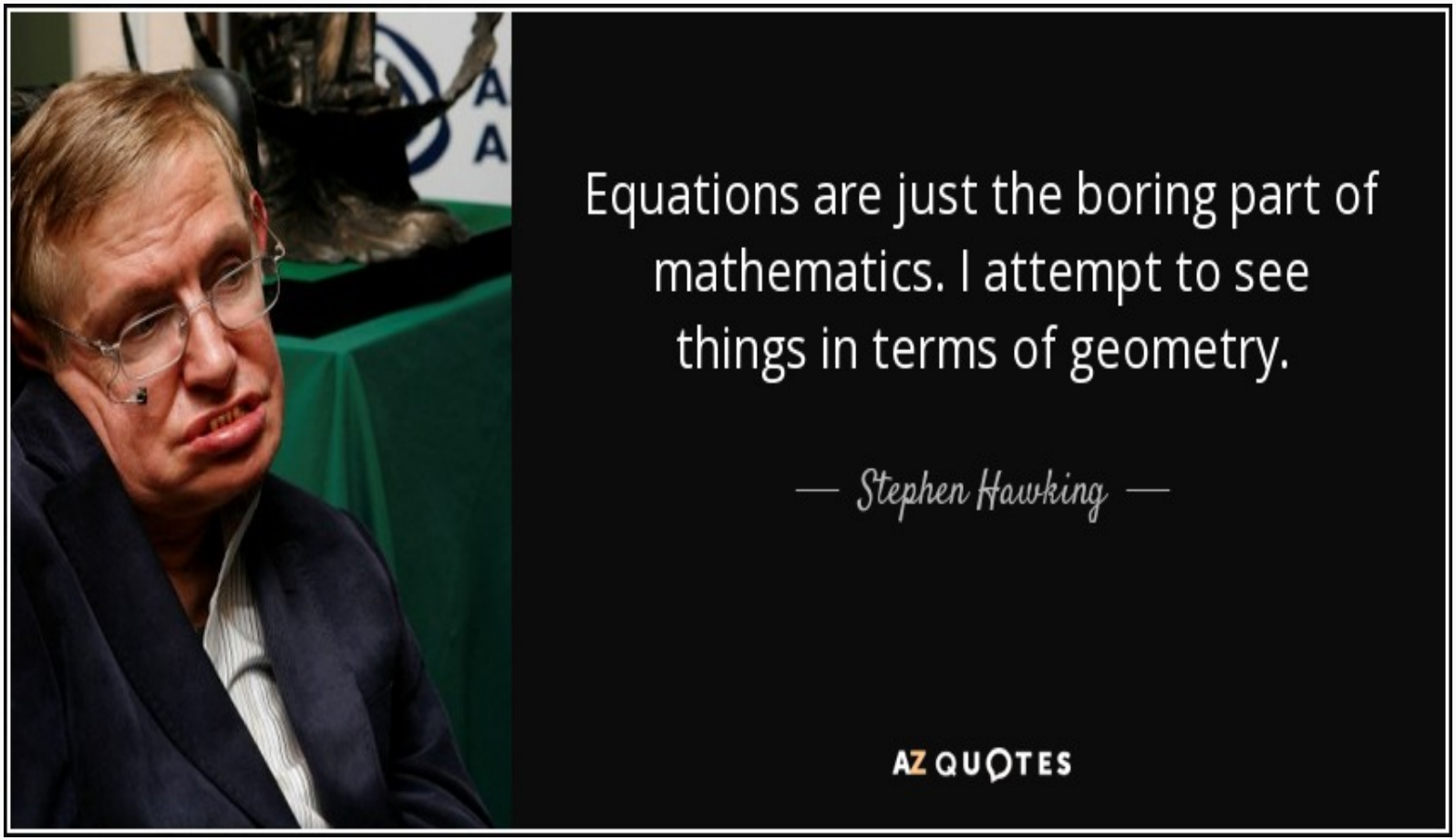
- Some General Policies:
  - The subject line of any course-related email should be started with [Stat 5630]. I reserve the right to discard emails that do not follow this format.
  - I will NOT answer problem-solving questions (especially coding) by email. If you have questions regarding homework problem-solving or general questions involving coding, please attend my regular once hour or send me email to schedule an appointment.
  - All submitted work shall be subject to the stipulations of the University of Virginia Honor System: each assignment and exam is to be pledged, On my honor, I did not receive aid on this assignment.

Students' Complaints to Instructor

- **Too hard, too many statistics, too many equations**

Students' Complaints to Instructor

- **Too hard, too many statistics, too many equations**



Students' Complaints to Instructor

- **Too easy, too basic, too outdated, Denny doesn't know machine learning/deep learning!**

Students' Complaints to Instructor

- **Too easy, too basic, too outdated, Denny doesn't know machine learning/deep learning!**

Instructors' Complaints to Students

- **YOU ARE TOO SMART!**

Instructors' Complaints to Students

- **YOU ARE TOO SMART!**

Why You Are Here???

- Reason 1: To have something on your resume/CV

Why You Are Here???

- Reason 1: To have something on your resume/CV
- Reason 2: Everyone is talking about it

Why You Are Here???

- Reason 1: To have something on your resume/CV
- Reason 2: Everyone is talking about it
- Reason 3: The instructor is nice to give beautiful grades

Why You Are Here???

- Reason 1: To have something on your resume/CV
- Reason 2: Everyone is talking about it
- Reason 3: The instructor is nice to give beautiful grades
- Reason 4: Randomly selected it
- ...

Why I am Here???

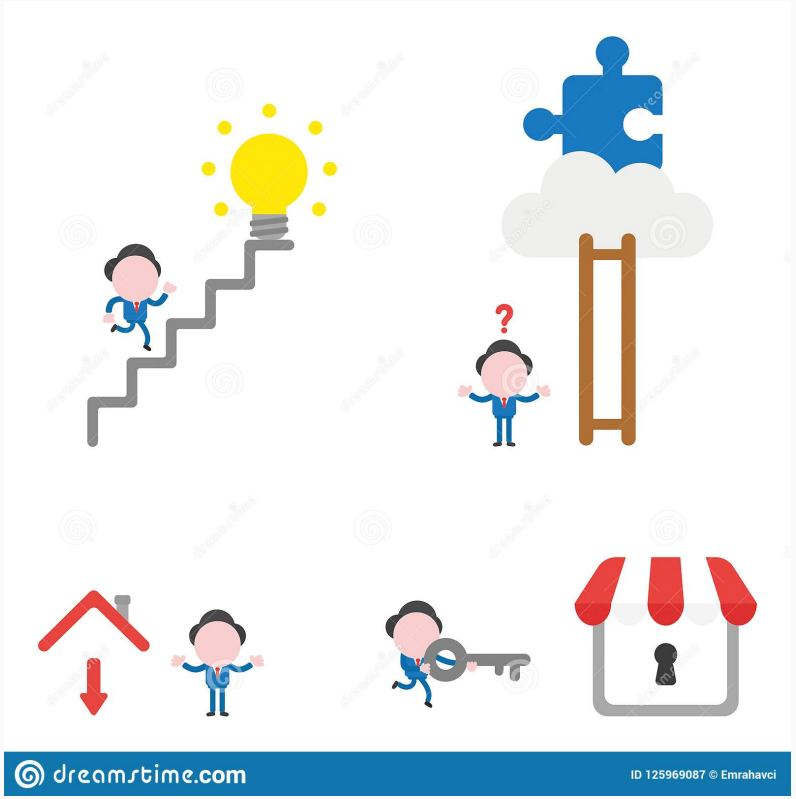
- Reason 1: YOU NEED IT!

Why I am Here???

- Reason 1: YOU NEED IT!
- Reason 2: To provide a different scope of ML.

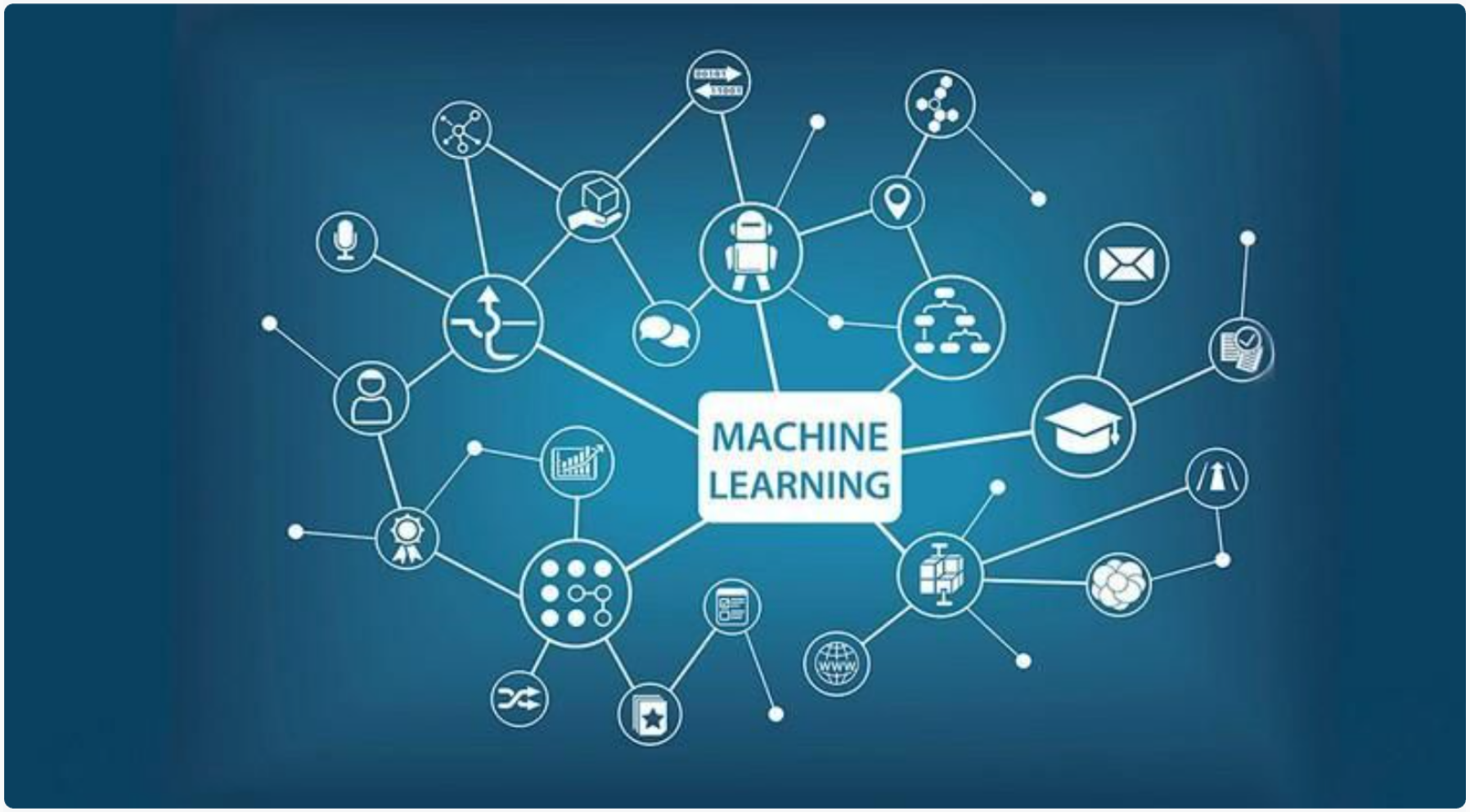
Why I am Here???

- Reason 1: YOU NEED IT!
- Reason 2: To provide a different scope of ML.
- Reason 3: To give you something useful.

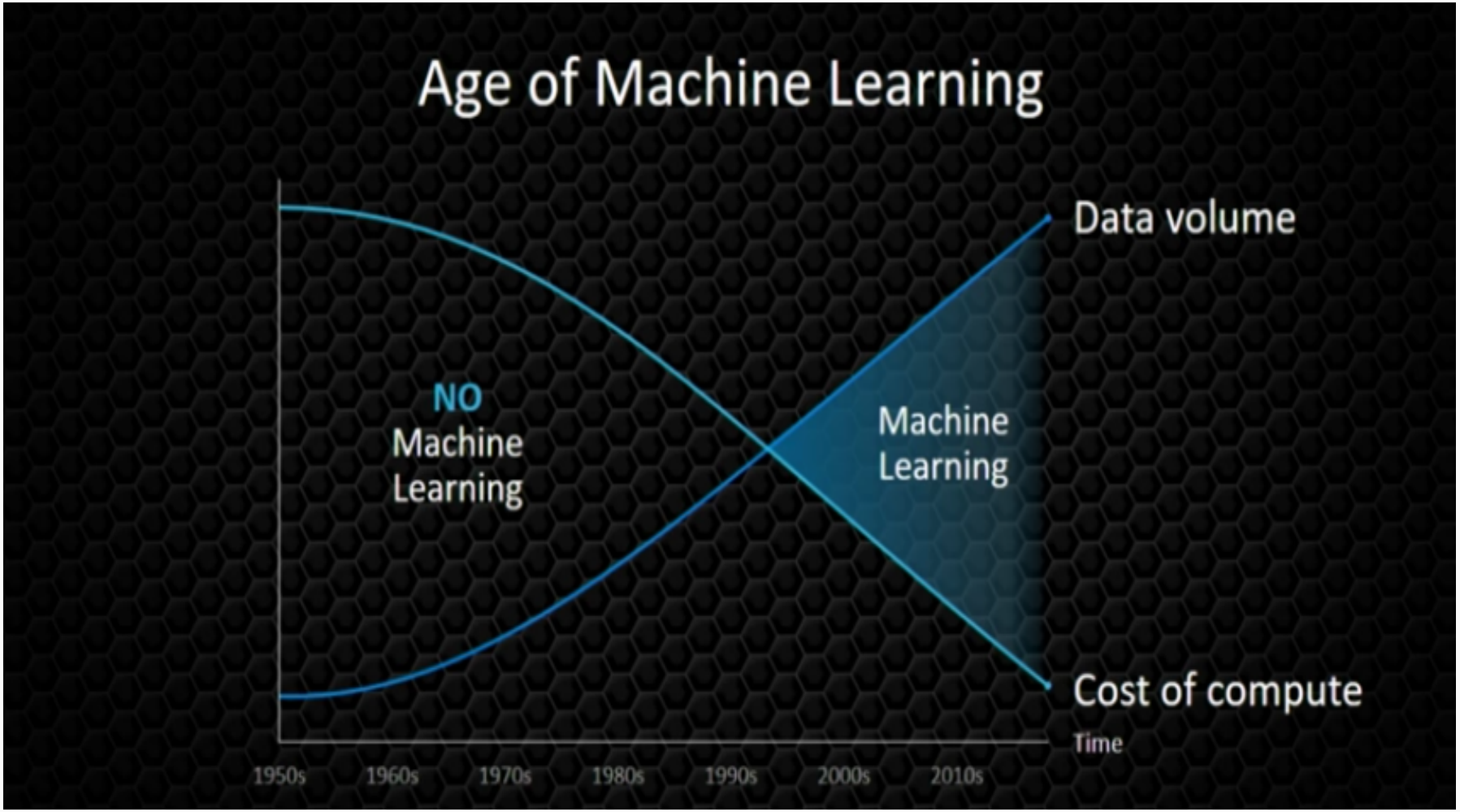


Overview of Statistical Machine Learning

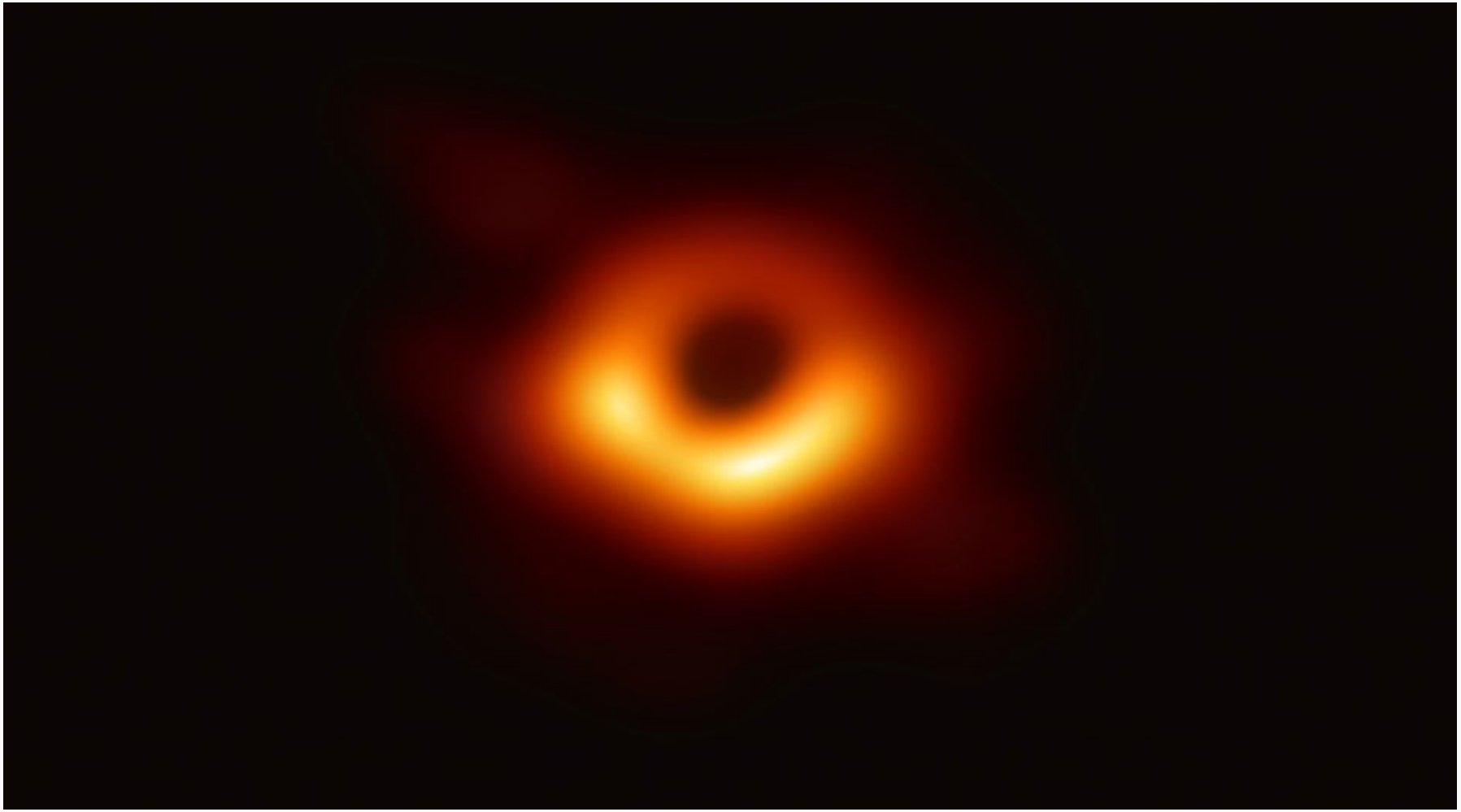
Big Data Era and ML



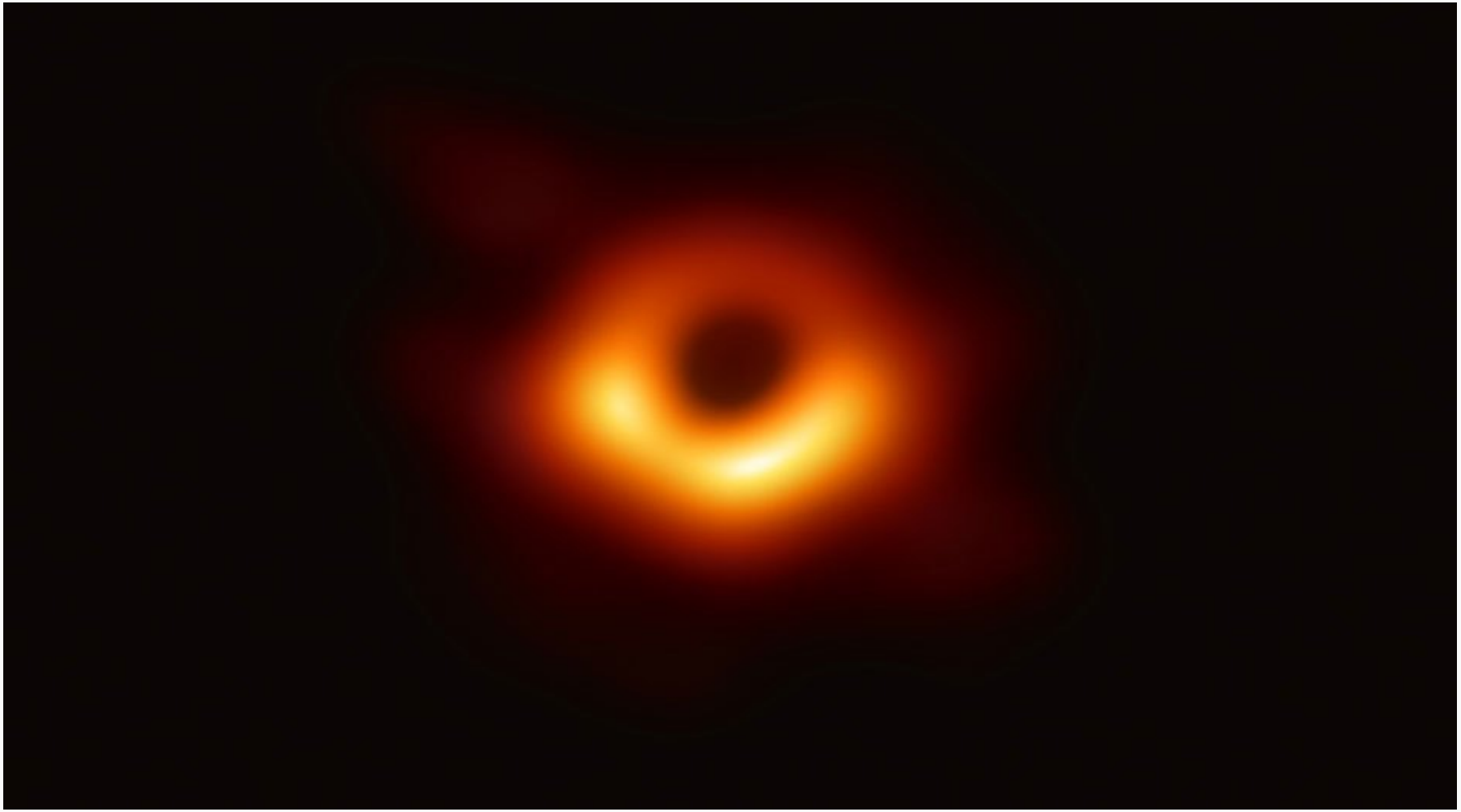
ML Age


Age of Machine Learning
Data volume
NO
Machine
Learning
Machine
Learning
Cost of compute
Time
1950s
1960s
1970s
1980s
1990s
2000s
2010s

Machine Learning in Life

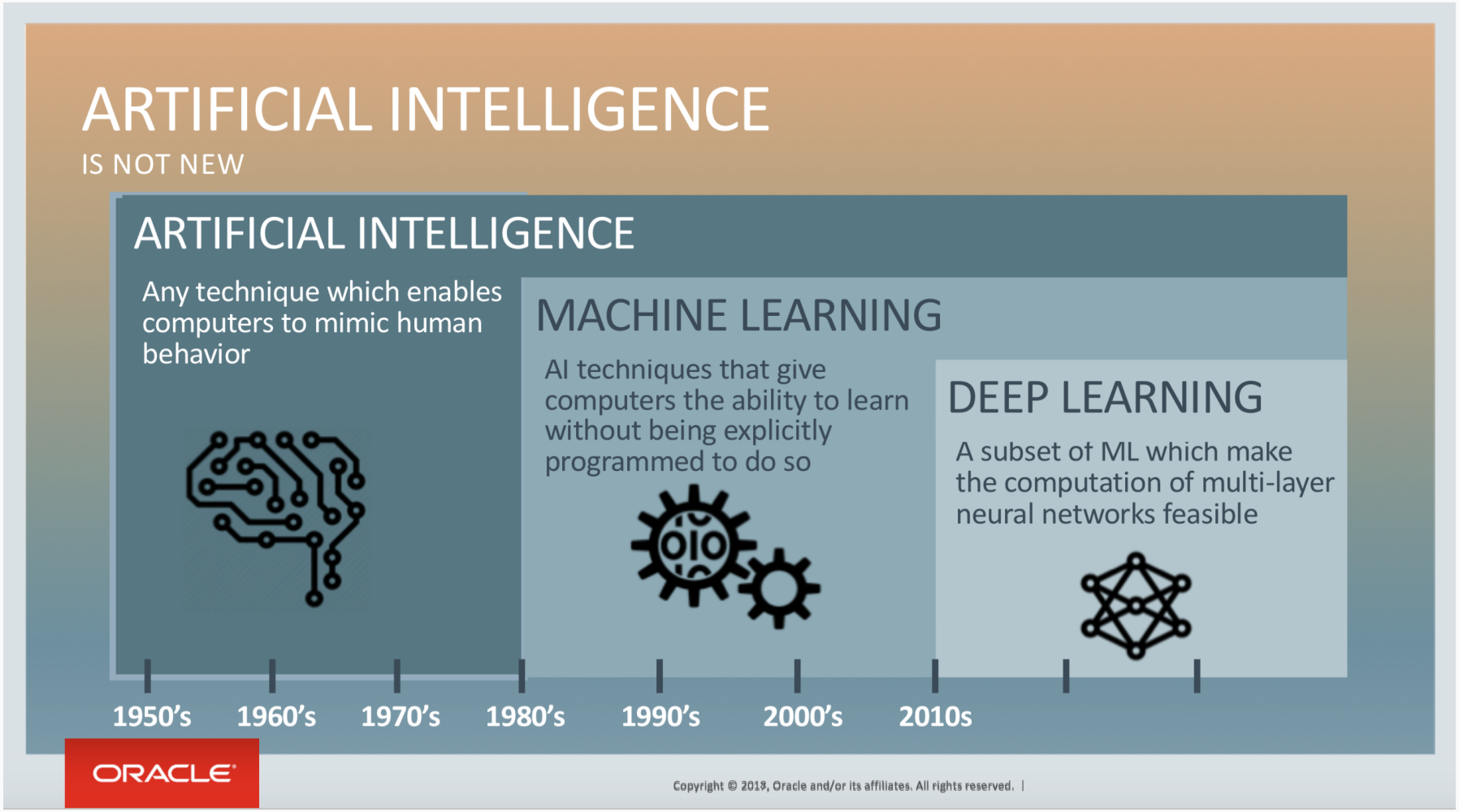


Machine Learning in Life

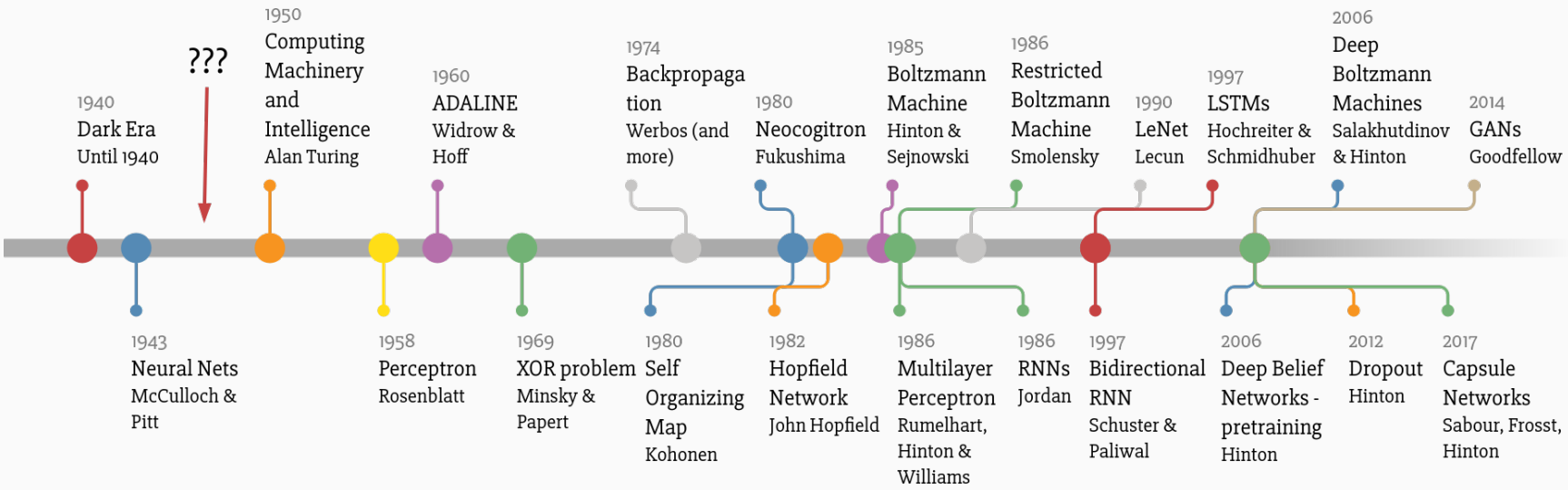


- The “fullsize original” 7416 x 4320 pixel TIFF is 183.4 MB.
- The EHT collected 5 petabytes of data, the reconstruction pares it down significantly.

ML Age


ARTIFICIAL INTELLIGENCE
IS NOT NEW
ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior
MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so
DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible
1950's
1960's
1970's
1980's
1990's
2000's
2010s
ORACLE
Copyright © 2019, Oracle and/or its affiliates. All rights reserved. |

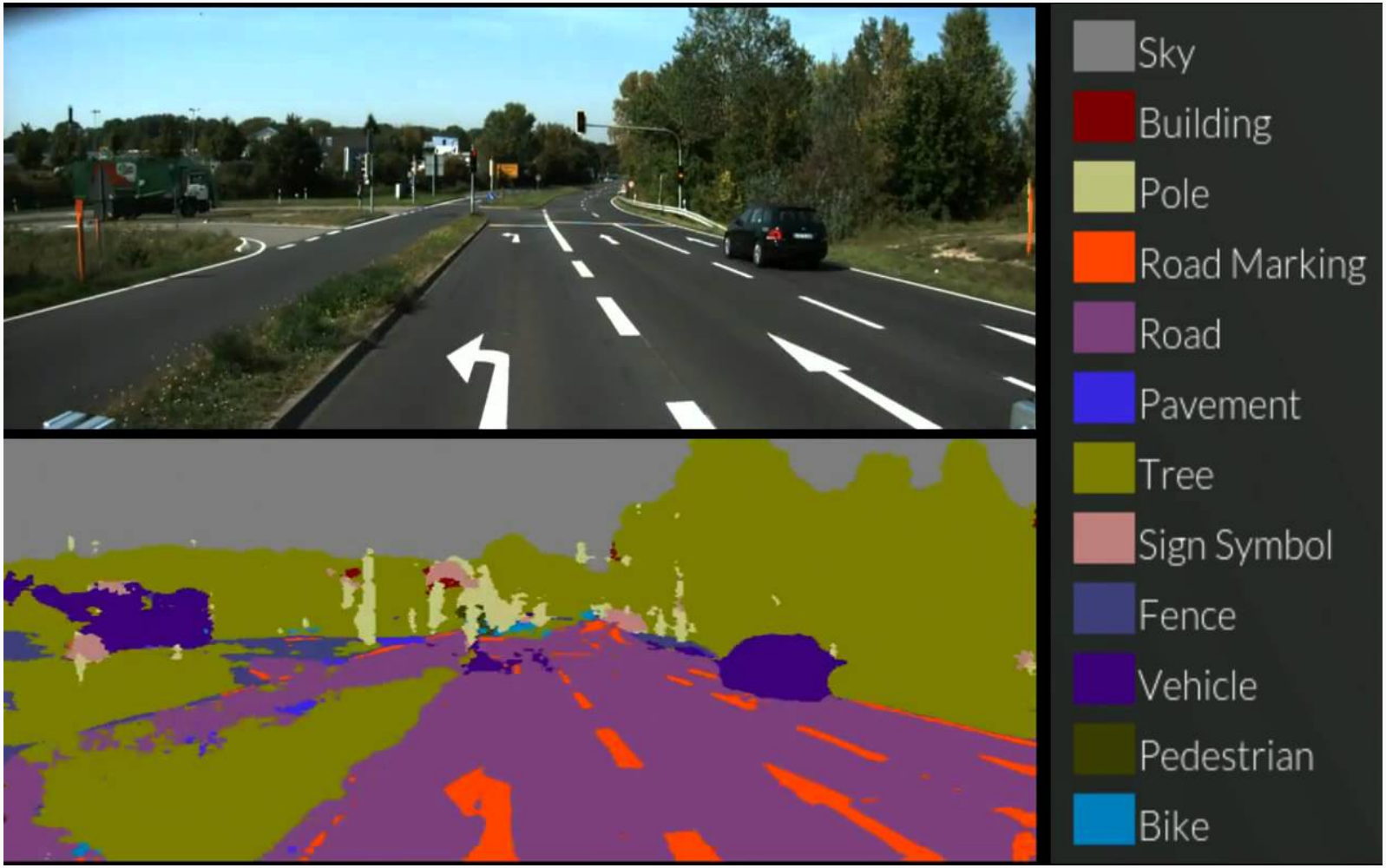
Deep Learning Timeline



Machine Learning in Life

Machine Learning in Life

Machine Learning in Life

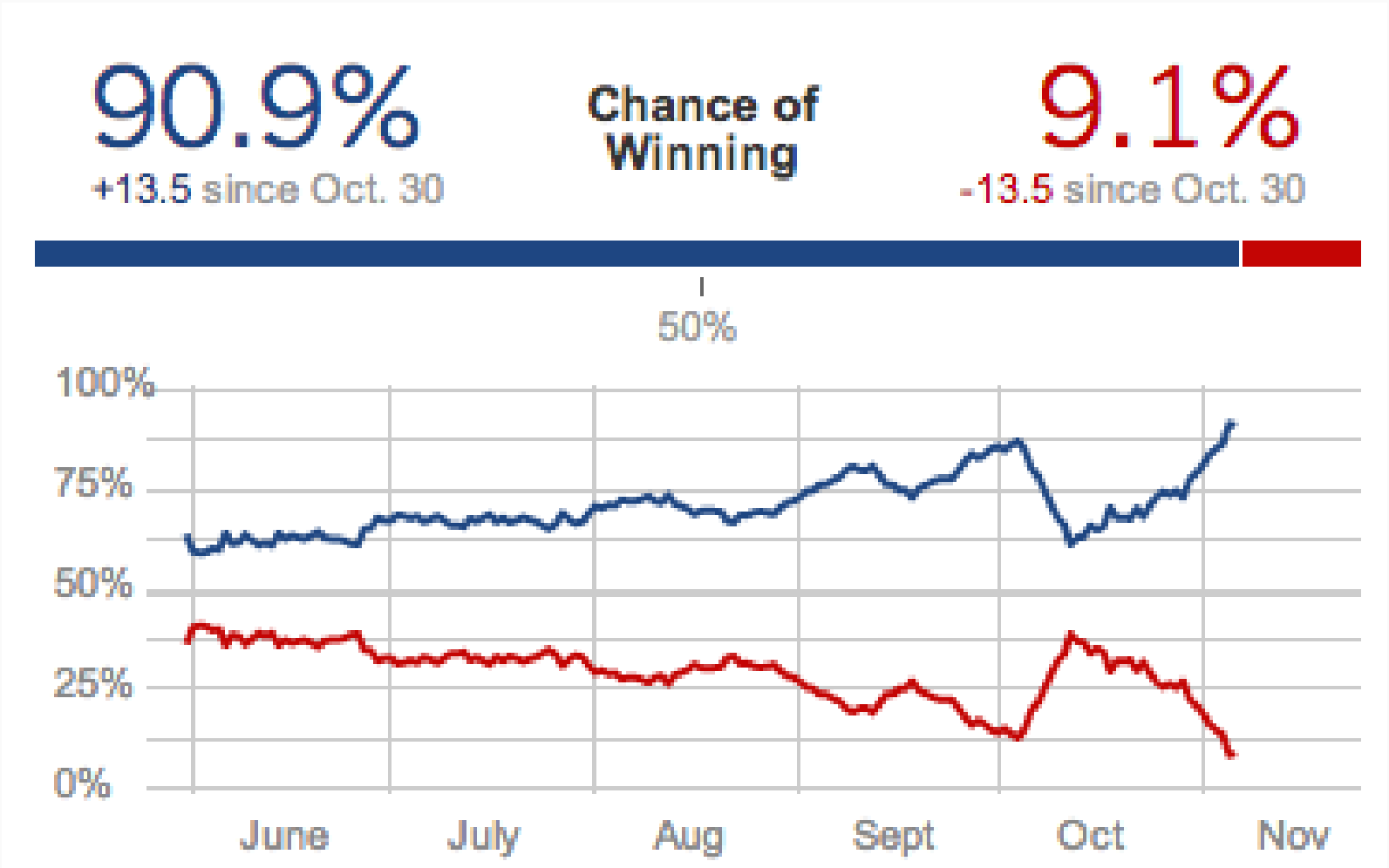


Machine Learning in Life

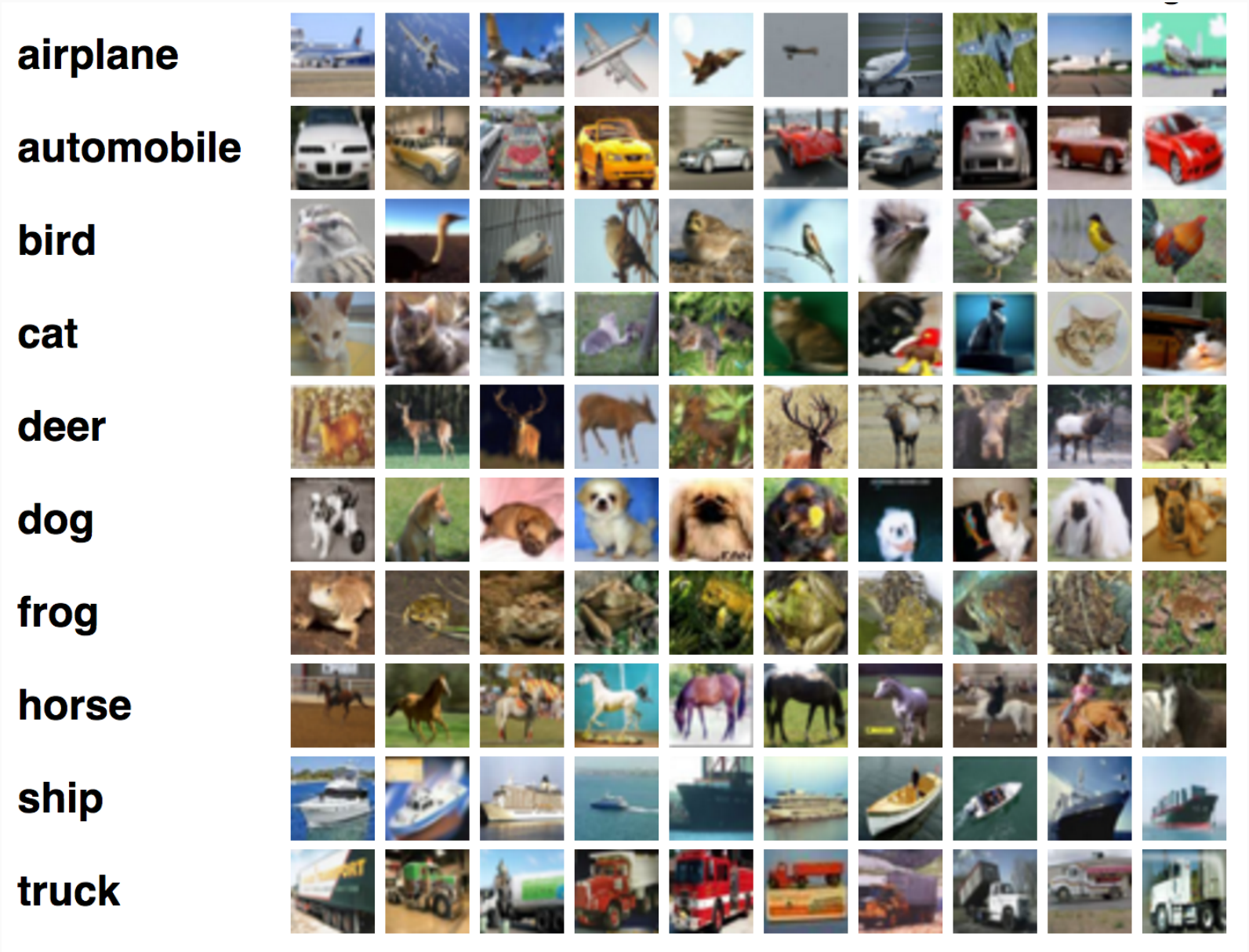
Machine Learning in Life

Unfortunately, We are Not gonna learn these...

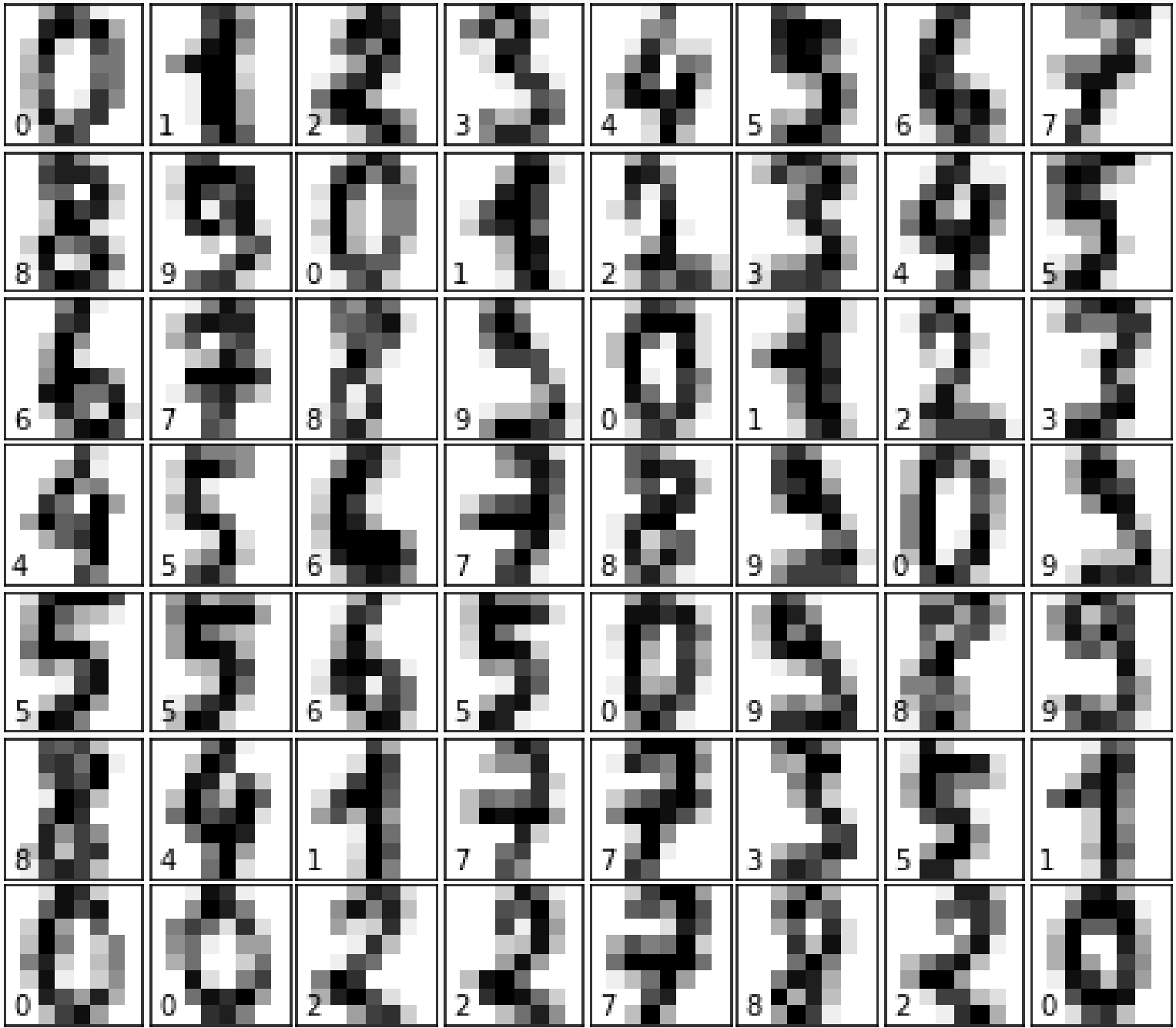
Examples


90.9%
+13.5 since Oct. 30
Chance of Winning
9.1%
-13.5 since Oct. 30
50%
100%
75%
50%
25%
0%
June
July
Aug
Sept
Oct
Nov

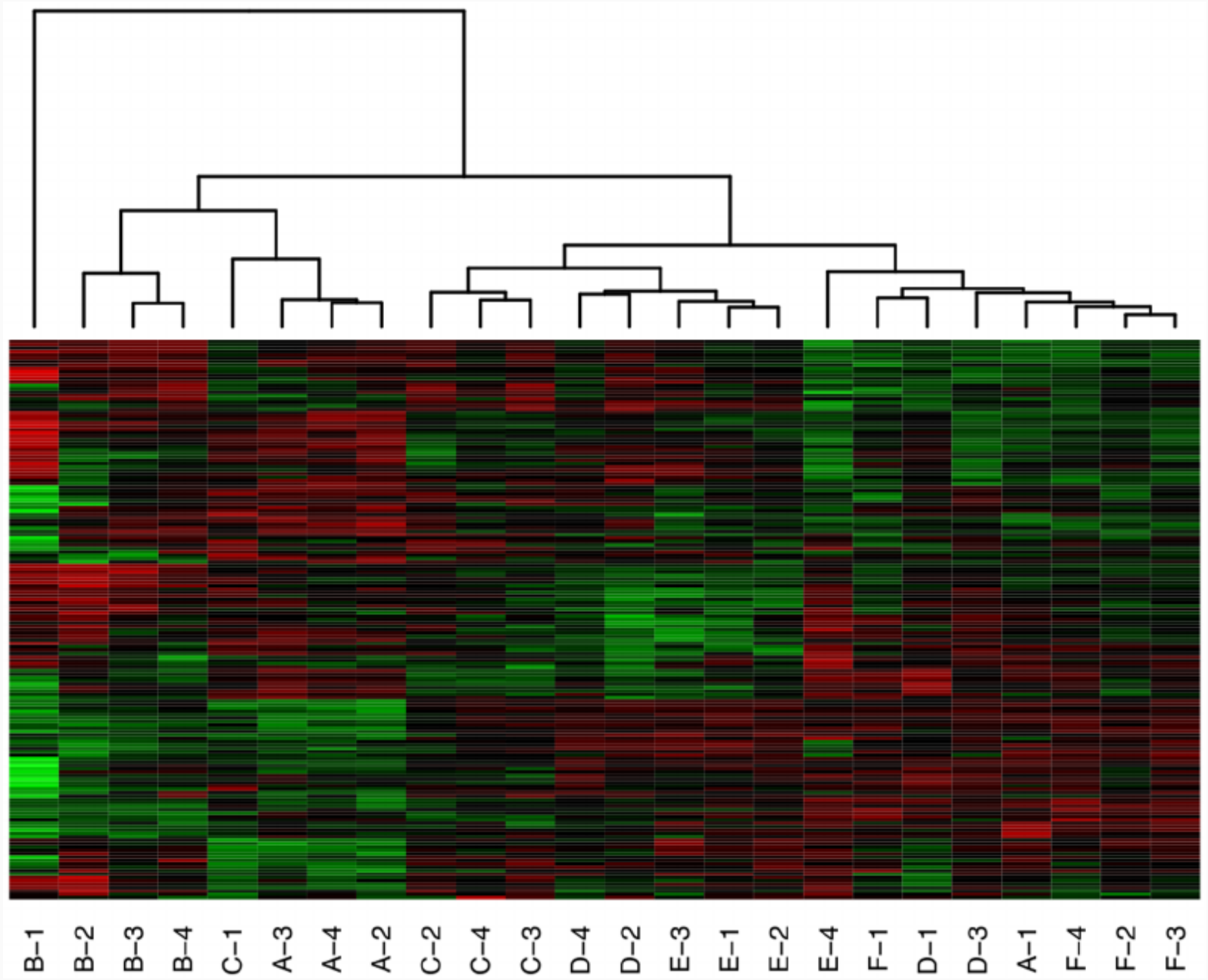
Examples


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

Examples


0 1 2 3 4 5 6 7
8 9 0 1 2 3 4 5
6 7 8 9 0 1 2 3
4 5 6 7 8 9 0 9
5 5 6 5 0 9 8 9
8 4 1 7 7 3 5 1
0 0 2 2 7 8 2 0

Examples



Examples

Everything is a Recommendation

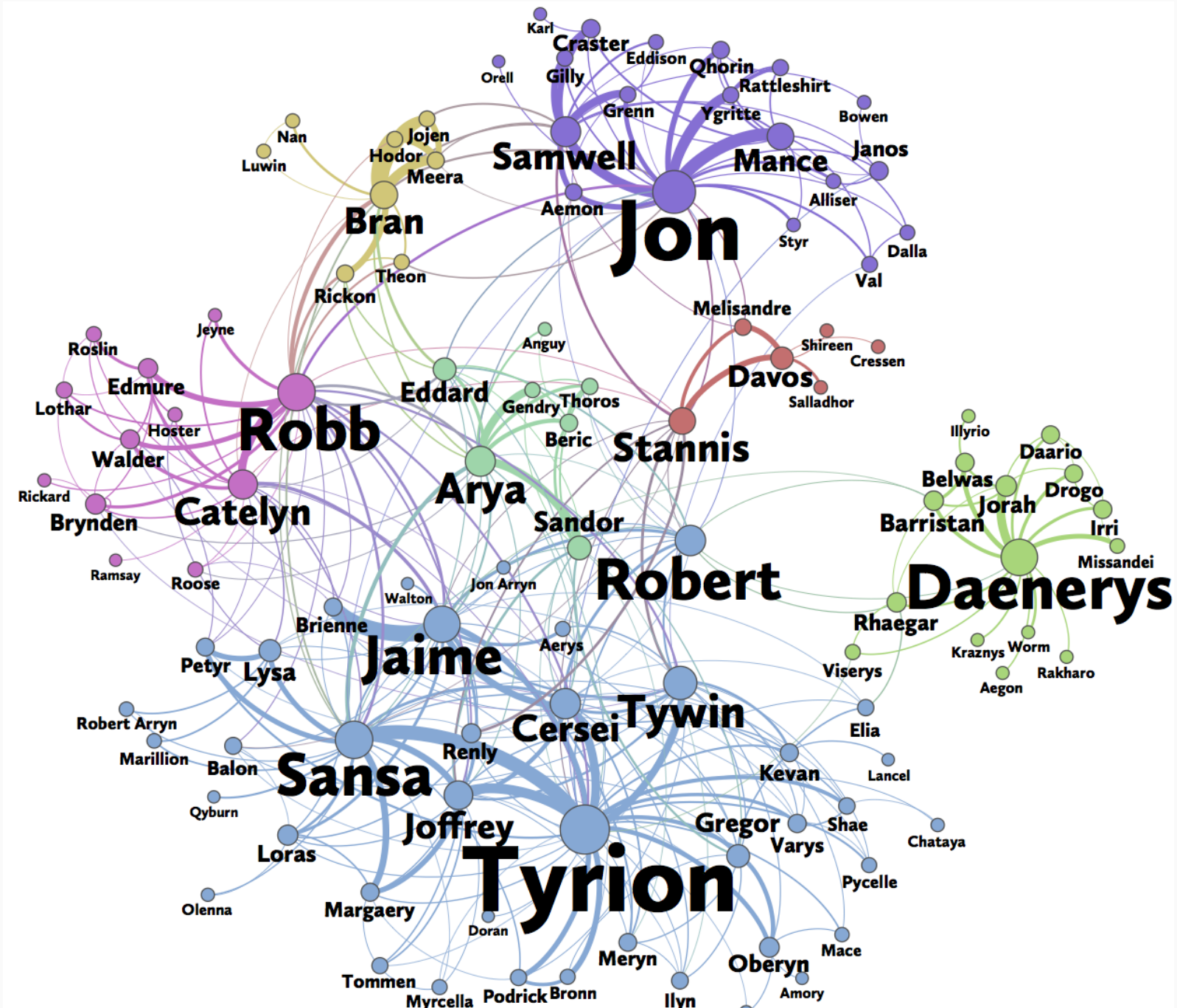
Over 80% of what people watch comes from our recommendations

Recommendations are driven by **Machine Learning**

NETFLIX

6

Examples


Karl
Craster
Eddison
Qhorin
Orell
Gilly
Rattleshirt
Grenn
Ygritte
Bowen
Nan
Jojen
Samwell
Mance
Janos
Luwin
Hodor
Meera
Alliser
Aemon
Bran
Jon
Styr
Dalla
Theon
Val
Rickon
Melisandre
Jeyne
Anguy
Shireen
Roslin
Cressen
Edmure
Eddard
Davos
Lothar
Gendry
Thoros
Salladhor
Hoster
Robb
Beric
Stannis
Illyrio
Walder
Daario
Belwas
Rickard
Arya
Drogo
Jorah
Brynden
Catelyn
Sandor
Barristan
Irri
Missandei
Ramsay
Roose
Robert
Daenerys
Walton
Jon Arryn
Brienne
Rhaegar
Aerys
Petyr
Lysa
Jaime
Kraznys
Worm
Viserys
Rakharo
Aegon
Robert Arryn
Tywin
Cersei
Elia
Marillion
Balon
Renly
Kevan
Lancel
Sansa
Qyburn
Gregor
Shae
Joffrey
Chataya
Loras
Varys
Tyrion
Pycelle
Olenna
Margaery
Doran
Meryn
Mace
Oberyn
Tommen
Myrcella
Podrick
Bronn
Ilyn
Amory

Course Overview

Different Perspectives of ML

- Learning Target: Prediction v.s. Description
- Data representation: Structured data v.s. Unstructured data
- Philosophy: Statistics v.s. Computer Science

Course Overview

- Regression - Model selection; Regularization and sparsity; Nonparametric: splines, linear and kernel smoothers; Regression trees
- Classification - Linear and quadratic discriminate analysis; Logistic regression; Support vector machines; Classification trees; Boosting and Bagging
- Clustering - Hierarchical clustering; K-means; Model-based clustering; (Optional: Spectral clustering; Nonparametric Bayes).
- Dimension Reduction - Principle components analysis; Matrix factorization
- Advanced - Recommendation system; Network analysis, Deep learning (Neural Network);
- Optional - Statistical inference on big data (AB testing)

Prerequisites...

- Statistical/mathematical
  - Statistical concepts: random variables, samples, mean, variance, distributions, conditional variables and distributions, likelihood, estimators and linear regressions.
  - Linear algebra and multivariate calculus
- Programming skills
  - Programming in R
  - Optimization