DATA ANALYTICS | SCIENCE

WHY | HOW TO BREAK INTO DATA SCIENCE?

SMR 3268
franckalbinet@gmail.com

1. DATA SCIENCE INTRODUCTION |AN OVERVIEW

2. PREDICTION MACHINES | THE ENGINE

3. DEPLOYMENT MOMENTUM

4. [POTENTIAL] TRACKS TO DEVELOP CAPACITIES

1. DATA SCIENCE INTRODUCTION |AN OVERVIEW

2. PREDICTION MACHINES | THE ENGINE

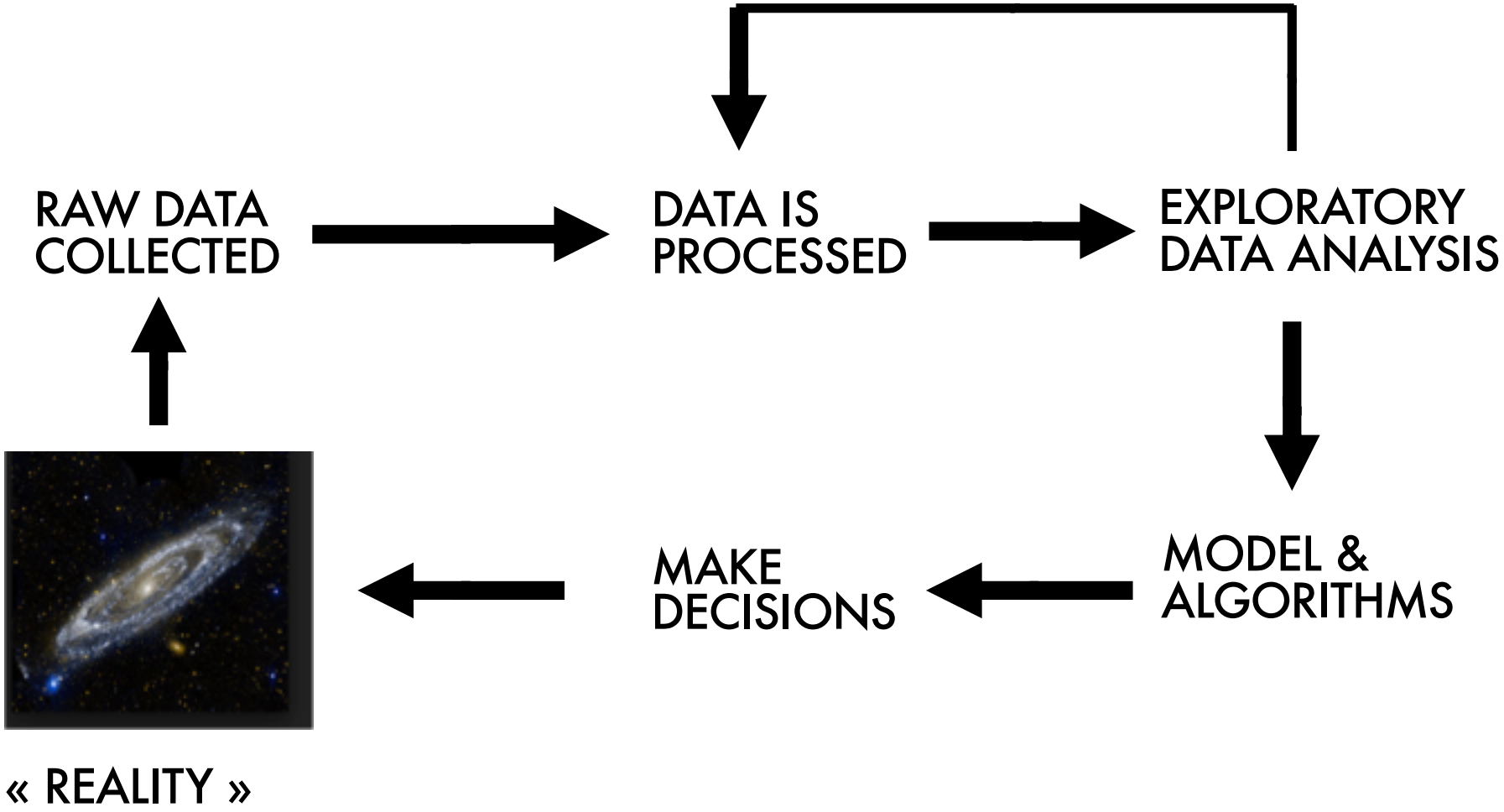
3. DEPLOYMENT MOMENTUM

4. [POTENTIAL] TRACKS TO DEVELOP CAPACITIES

IoT PROVIDES
[PARTIAL] OBSERVATION OF THE WORLD

DATA SCIENCE HARNESS IT
TO INFORMED DECISION

DATA SCIENCE PIPELINE



EXAMPLE OF AIR QUALITY MONITORING SYSTEM

I. PRIMARY DATA

IoT NETWORK **SENSES** PARTICULATE MATTERS (PM), CO2, …

II. SECONDARY DATA

THE DATA ANALYSIS PIPELINE **TAPS INTO** SOCIAL NETWORKS, METEOROLOGICAL FORECASTS, CROWD-SOURCED, ...

III. DATA SCIENCE PIPELINE « STIRS » DATA

PERFORMS SPATIAL INTERPOLATION,
SENTIMENT ANALYSIS, HARNESSES
METEOROLOGICAL DATA, **OUTPUTS**
DIAGNOSIS, PROGNOSIS

IV. DECISION MAKERS ...

INTERPRET OUTPUTS IN A WIDER CONTEXT, **CONTROL** ROAD TRAFFIC, **ENFORCE** RESTRICTION USE, **PROVIDE FEEDBACK** TO IMPROVE THE WHOLE VALUE CHAIN

QUIZ

- ❑ COLLECT AS MUCH DATA AS YOU CAN!
- ❑ DATA SCIENCE = [ADVANCED] STATISTICS
- ❑ NEITHER IOT NOR DATA SCIENCE IS REALLY NEW.
- ❑ DOMAIN KNOWLEDGE DOES NOT MATTER NOWADAYS.

WHAT MAKES DATA SCIENCE « UNIQUE »

DATA ENGINEERING

SCIENTIFIC METHOD

DOMAIN EXPERTISE

MATHEMATICS

DATA SCIENCE

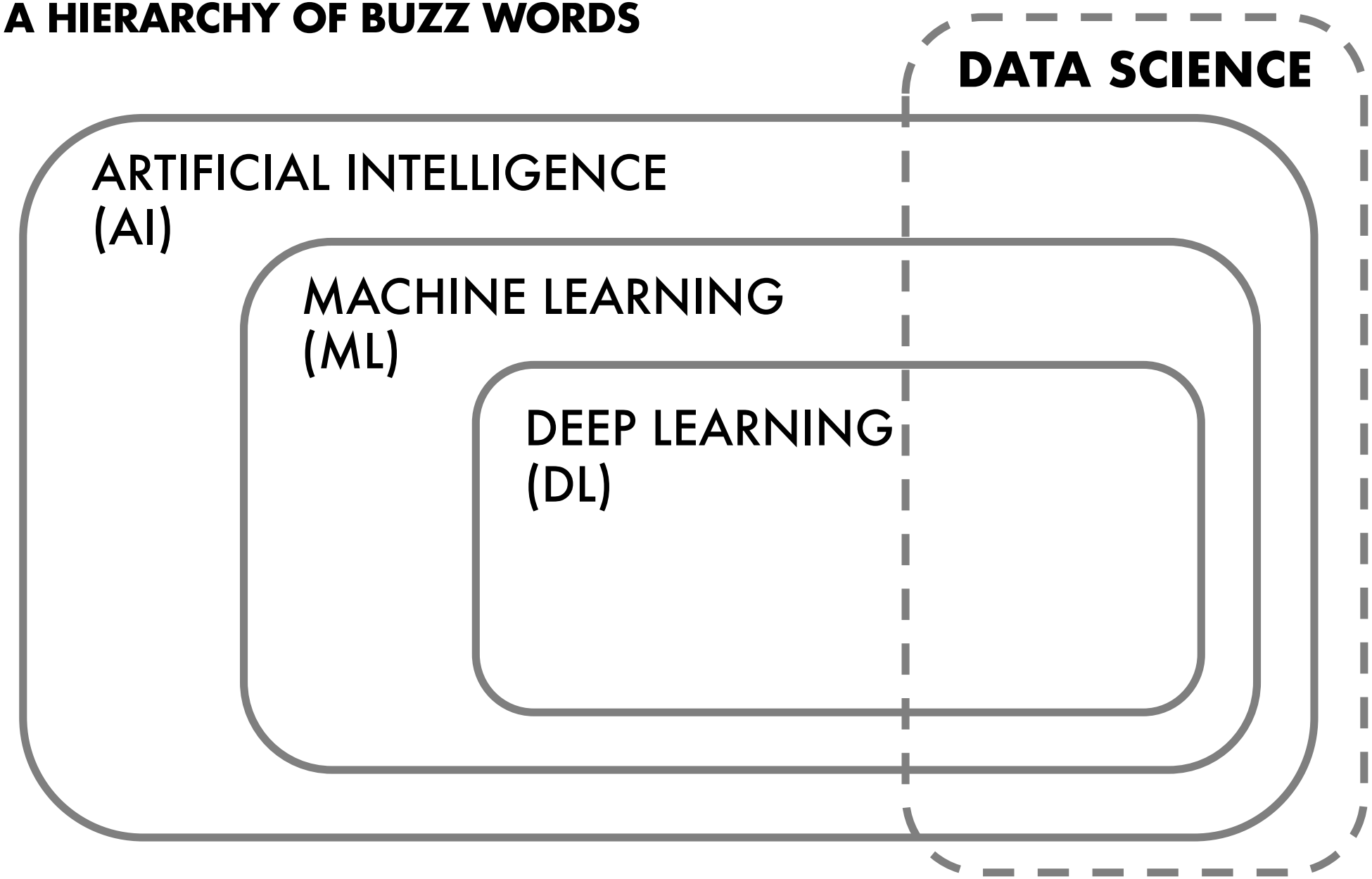
HACKER MINDSET

STATISTICS

VISUALIZATION

« ADVANCED » COMPUTING

A HIERARCHY OF BUZZ WORDS



AN AUTHORITATIVE DEF. OF ML BUT ...

« A COMPUTER PROGRAM IS SAID TO **LEARN** FROM EXPERIENCE **E** WITH RESPECT TO SOME CLASS OF TASKS **T** AND PERFORMANCE MEASURE **P**, IF ITS PERFORMANCE AT TASKS IN **T**, AS MEASURED BY **P**, IMPROVES WITH EXPERIENCE **R** »

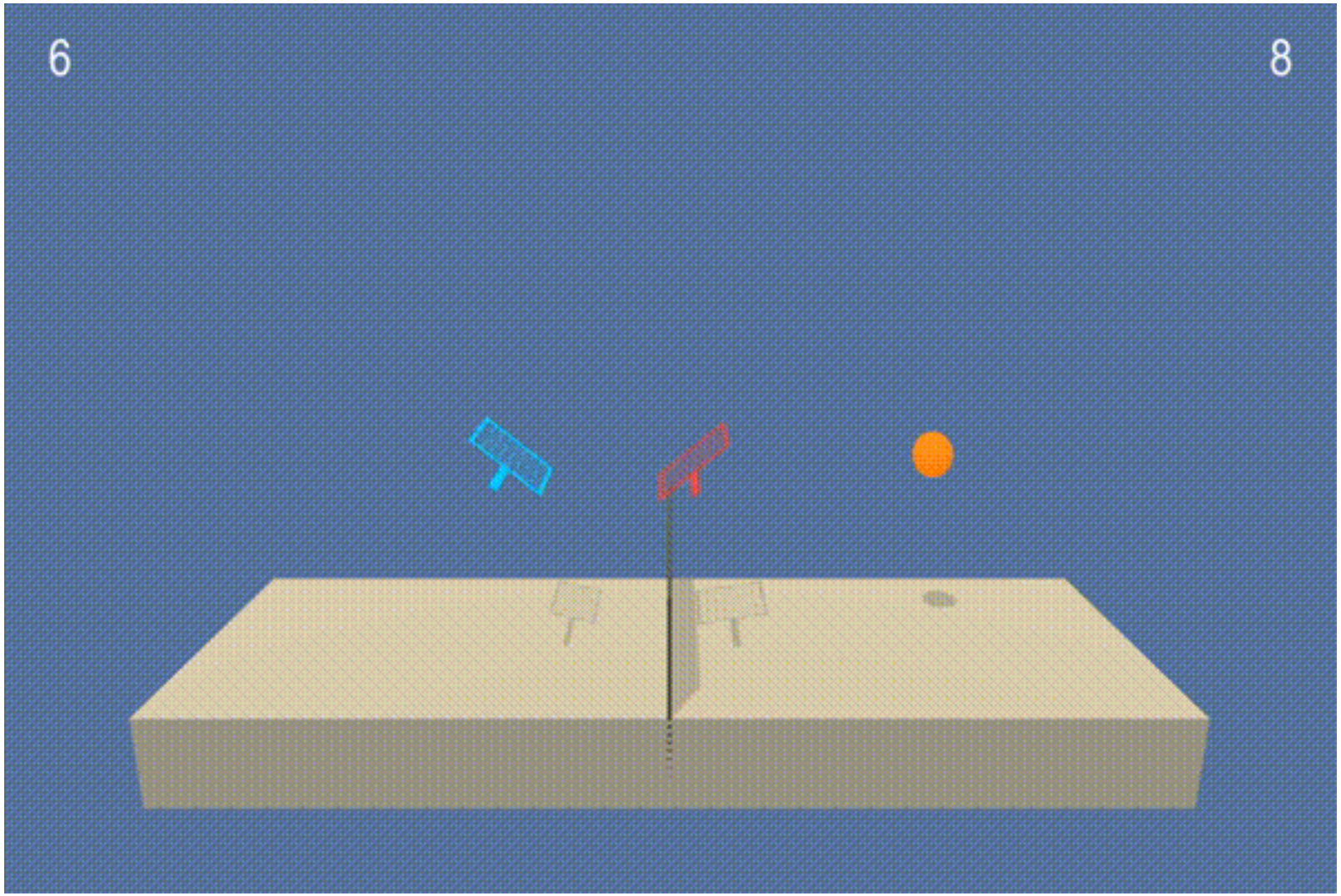
— TOM MITCHELL, MACHINE LEARNING

ANOTHER ONE

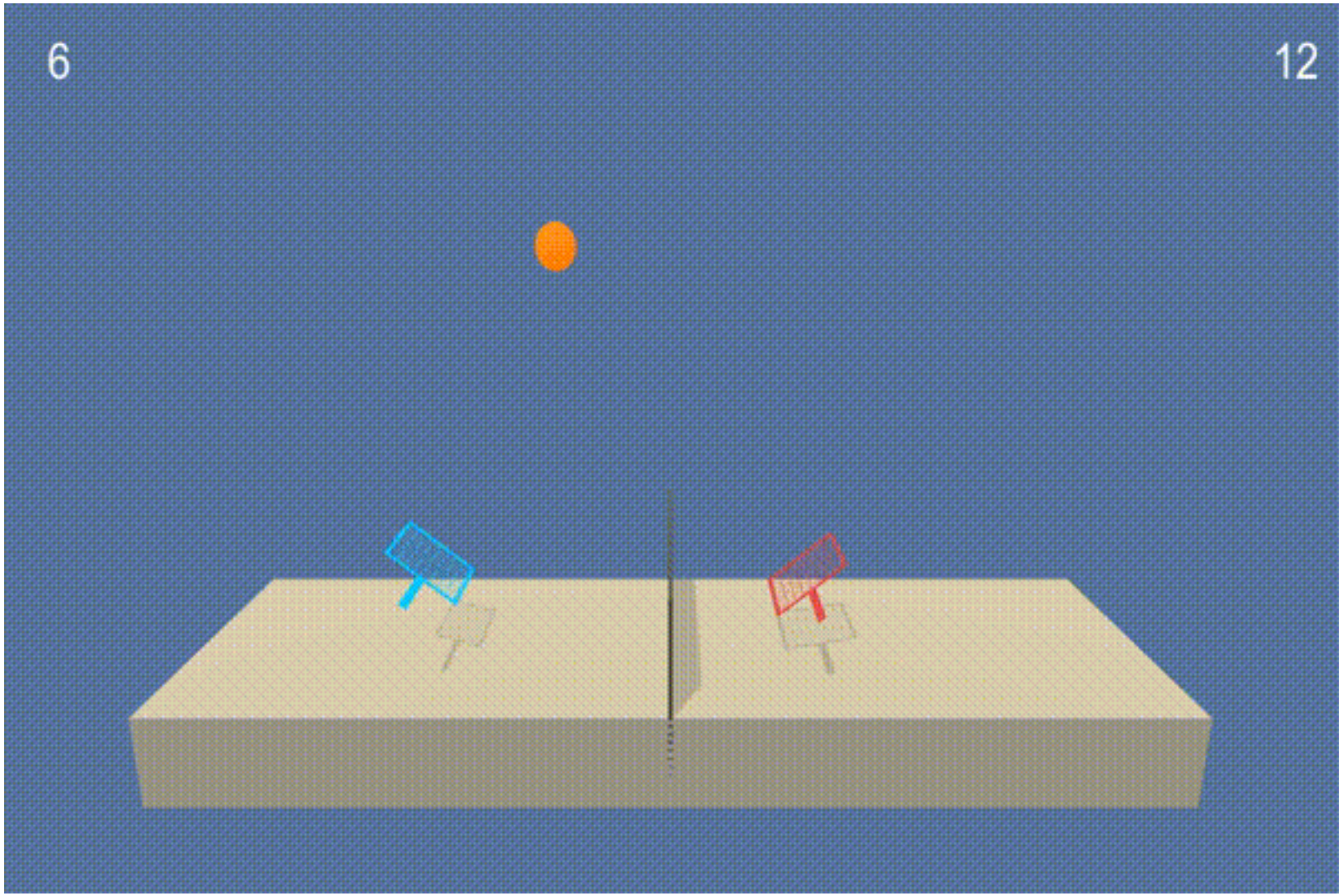
« MACHINE LEARNING IS THE FIELD OF STUDY THAT GIVES COMPUTER THE ABILITY TO LEARN WITHOUT BEING EXPLICITLY PROGRAMMED »

— ARTHUR SAMUEL, 1959

A COMPELLING EXAMPLE: TENNIS [UNSOLVED]

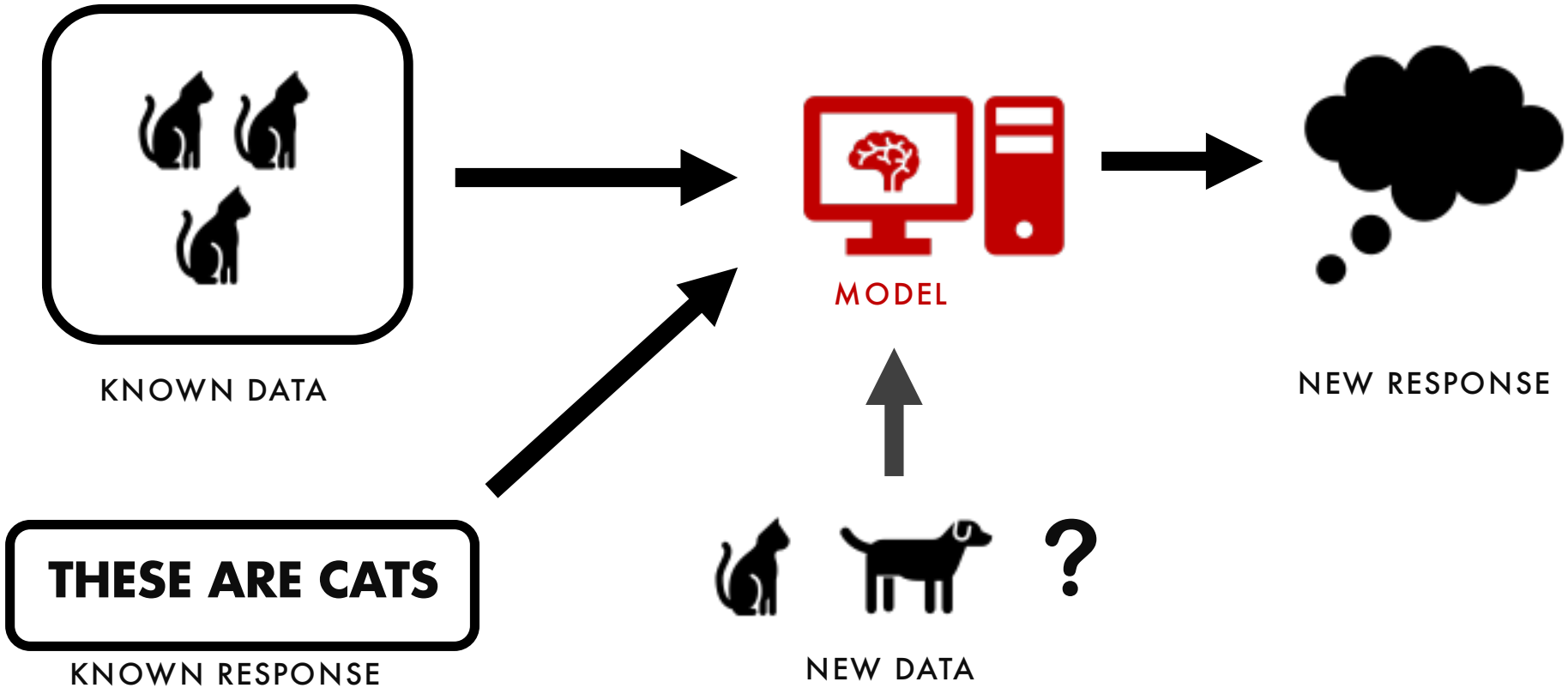


TENNIS [SOLVED AFTER 5 MIN TRAINING ON STANDARD LAPTOP]

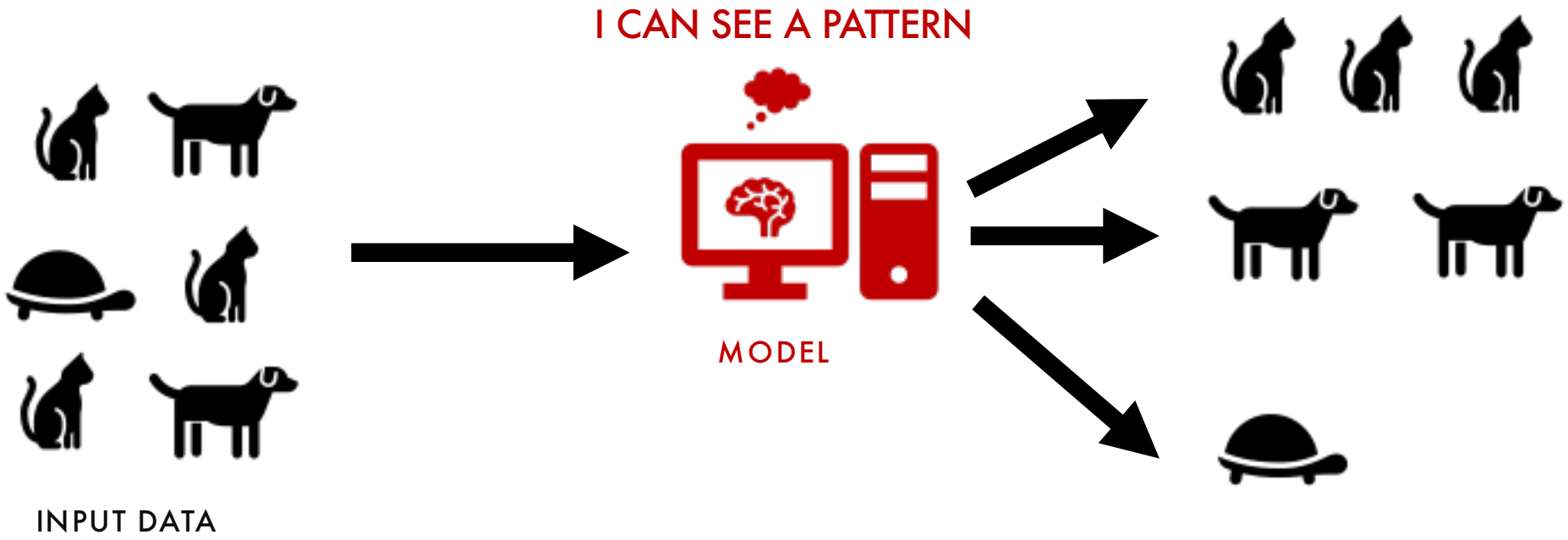


QUICK NOMENCLATURE OF ML « PARADIGMS »

SUPERVISED LEARNING



UNSUPERVISED LEARNING



REINFORCEMENT LEARNING

- OBSERVE ITS STATE
- GET REWARD

ACT

NAVIGATE DATA SCIENCE | ML | DL | AI APPLICATIONS

LIST APPLICATIONS YOU CAN THINK OF.
HOW COULD YOU GROUP THEM?

BY TECHNOLOGY TYPE

- IMAGE RECOGNITION
- OBJECT DETECTION
- NATURAL LANGUAGE PROCESSING, SPEECH RECOGNITION
- TIME SERIES PREDICTION
- ROBOTIC
- GAMES & SIMULATION, ...

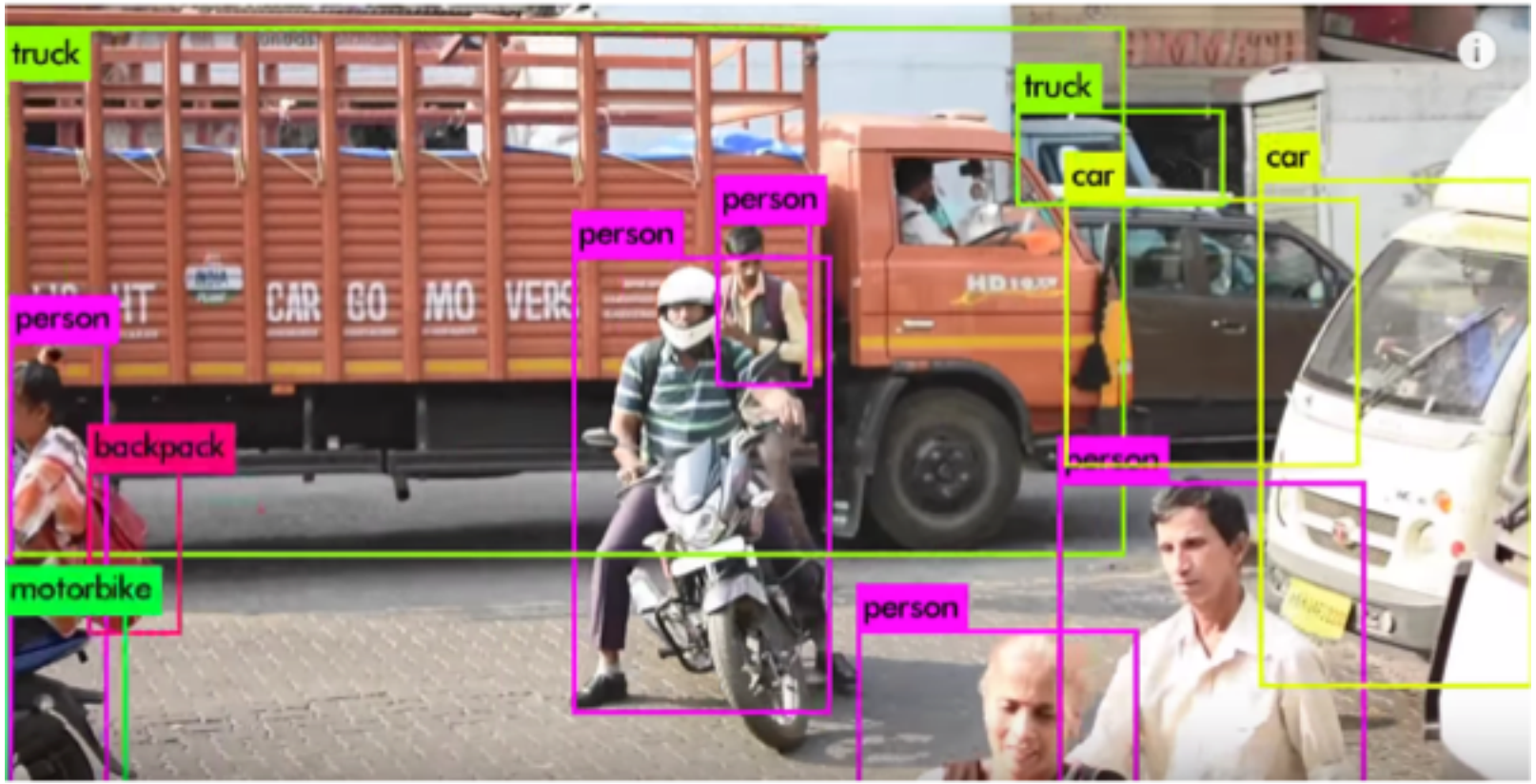
BY ACTIVITY SECTORS

- HEALTH
- AGRICULTURE, ENVIRONMENT
- EDUCATION & SCIENCE
- INDUSTRY
- DEFENCE, ...

QUICK TOUR OF EXAMPLE APPLICATIONS

FROM OBJECT DETECTION TO MUSIC

REAL TIME OBJECT DETECTION



https://pjreddie.com/darknet/yolo

SMART FARMING

How a Japanese cucumber farmer is using deep learning and TensorFlow

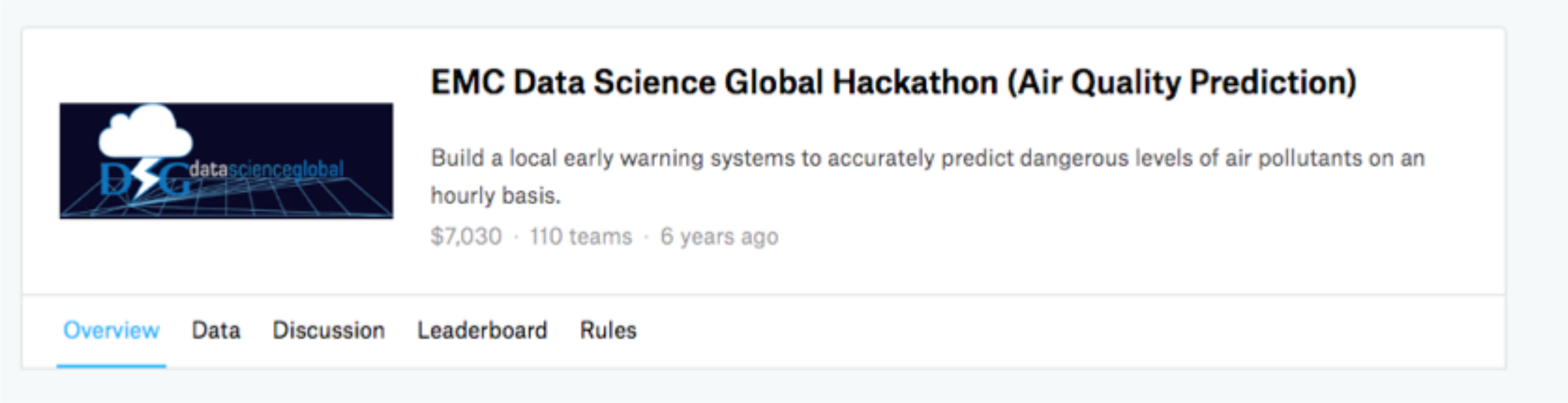
Kaz Sato
Developer Advocate, Google Cloud Platform

August 31, 2016

It's not hyperbole to say that use cases for machine learning and deep learning are only limited by our imaginations. About one year ago, a former embedded systems designer from the Japanese automobile industry named Makoto Koike started helping out at his parents' cucumber farm, and was amazed by the amount of work it takes to sort cucumbers by size, shape, color and other attributes.

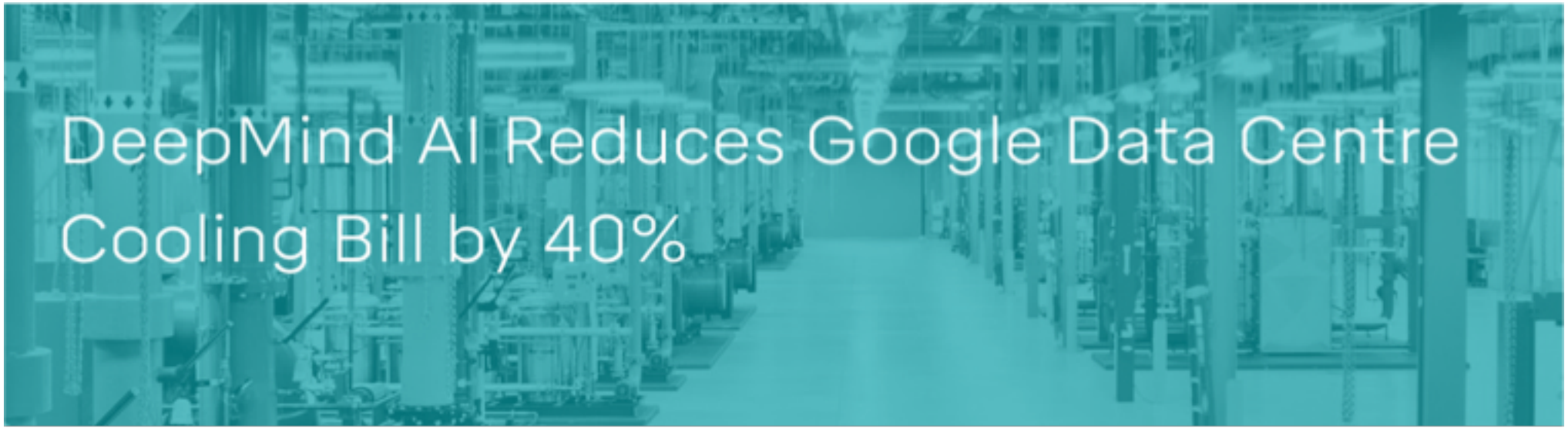
https://tinyurl.com/y7js7scm

AIR QUALITY PREDICTION



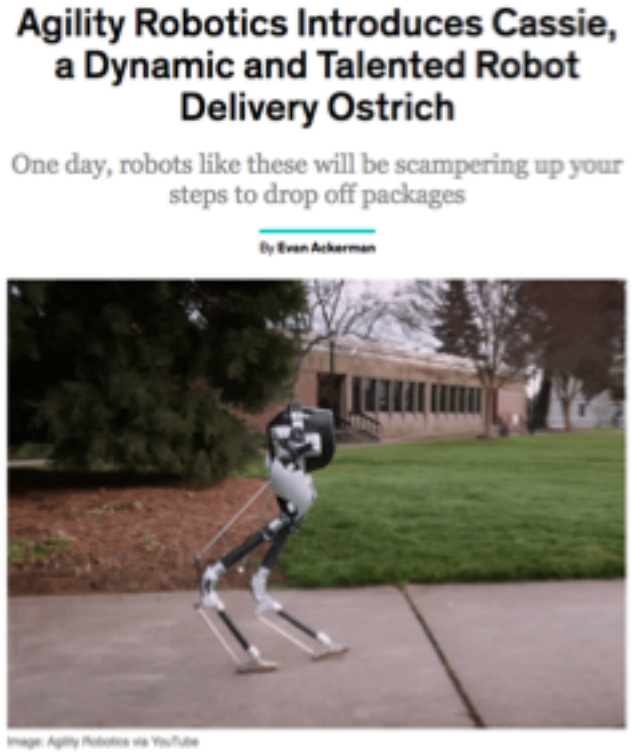
https://www.kaggle.com/c/dsg-hackathon

ENERGY CONSUMPTION

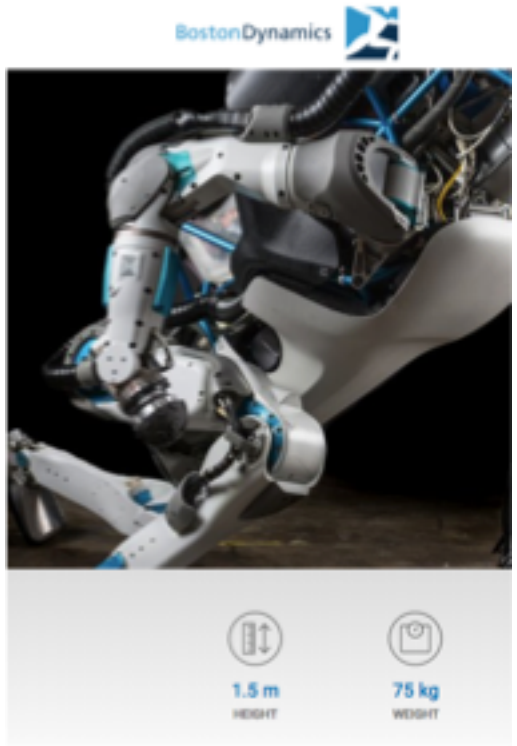


https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

ROBOTICS



https://tinyurl.com/y7y433fs



https://www.bostondynamics.com/atlas

PHYSICS

APPLYING MACHINE LEARNING TO PHYSICS

- "Reconstruction of a Photonic Qubit State with Quantum Reinforcement Learning", Shang Yu, F. Albarran-Arriagada, J. C. Retamal, Yi-Tao Wang, Wei Liu, Zhi-Jin Ke, Yu Meng, Zhi-Peng Li, Jian-Shun Tang, E. Solano, L. Lamata, Chuan-Feng Li, Guang-Can Guo, arXiv: 1808.09241, 8/2018
- "Policy Guided Monte Carlo: Reinforcement Learning Markov Chain Dynamics", Troels Arnfred Bojesen, arXiv: 1808.09095, 8/2018
- "Machine learning non-local correlations", Askery Canabarro, Samuraí Brito, Rafael Chaves, arXiv: 1808.07069, 8/2018
- "Smart energy models for atomistic simulations using a DFT-driven multifidelity approach", Luca Messina, Alessio Quaglino, Alexandra Goryaeva, Mihai-Cosmin Marinica, Christophe Domain, Nicolas Castin, Giovanni Bonny, Rolf Krause, arXiv: 1808.06935, 8/2018
- "Machine Learning Configuration Interaction", J. P. Coe, arXiv: 1808.05787, 8/2018

https://physicsml.github.io/pages/papers.html

PHYSICS

APPLYING MACHINE LEARNING TO PHYSICS

- "Reconstruction of a Photonic Qubit State with Quantum Reinforcement Learning", Shang Yu, F. Albarran-Arriagada, J. C. Retamal, Yi-Tao Wang, Wei Liu, Zhi-Jin Ke, Yu Meng, Zhi-Peng Li, Jian-Shun Tang, E. Solano, L. Lamata, Chuan-Feng Li, Guang-Can Guo, arXiv: 1808.09241, 8/2018
- "Policy Guided Monte Carlo: Reinforcement Learning Markov Chain Dynamics", Troels Arnfred Bojesen, arXiv: 1808.09095, 8/2018
- "Machine learning non-local correlations", Askery Canabarro, Samuraí Brito, Rafael Chaves, arXiv: 1808.07069, 8/2018
- "Smart energy models for atomistic simulations using a DFT-driven multifidelity approach", Luca Messina, Alessio Quaglino, Alexandra Goryaeva, Mihai-Cosmin Marinica, Christophe Domain, Nicolas Castin, Giovanni Bonny, Rolf Krause, arXiv: 1808.06935, 8/2018
- "Machine Learning Configuration Interaction", J. P. Coe, arXiv: 1808.05787, 8/2018

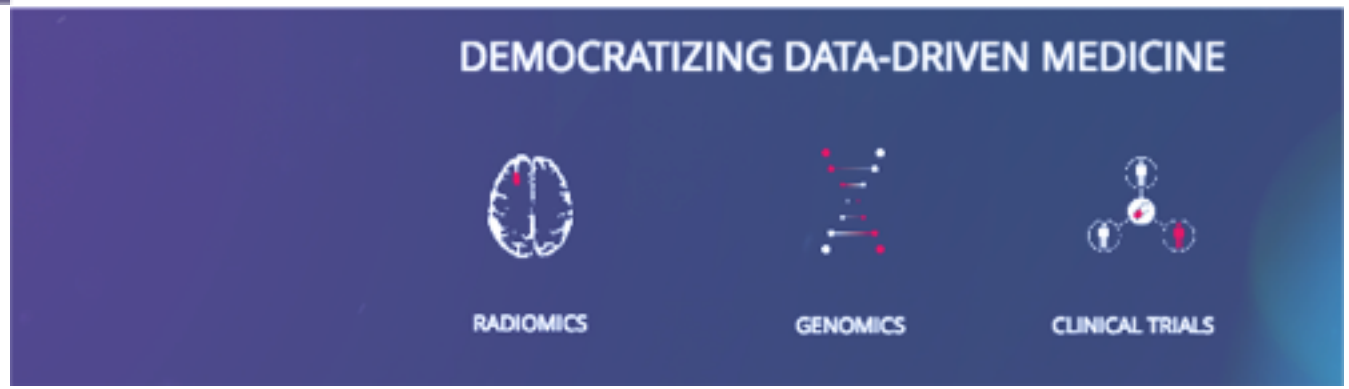
https://physicsml.github.io/pages/papers.html

MEDICINE

AlphaFold: Using AI for scientific discovery

https://deepmind.com/blog/alphafold/

SOPHiA GENETICS®



https://www.sophiagenetics.com

MUSIC

https://magenta.tensorflow.org/

AN ENDLESS LIST ...

WHAT IS AT THE CORE OF ALL THESE APPLICATIONS?

1. DATA SCIENCE INTRODUCTION | OVERVIEW

2. PREDICTION MACHINES | THE ENGINE

3. DEPLOYMENT MOMENTUM

4. [POTENTIAL] TRACKS TO DEVELOP CAPACITIES

PREDICTION AS A CORE COMPONENT OF DATA SCIENCE | ML | DL | AI

- **PREDICTING** STOCK MARKET
- **PREDICTING** WHO IS IN THE VIDEO
- **PREDICTING** A SERIOUS DISEASE
- **PREDICTING** POLLUTION PEACK
- …

WHAT IS A PREDICTION?

PREDICTION AS A PROCESS TO FILL MISSING INFORMATION

INFORMATION
YOU HAVE

INFORMATION
YOU DON'T HAVE

PREDICTION VS. DECISION MAKING ?

UNPACKING DECISION MAKING

- **PREDICTION** IS ONLY ONE KEY ELEMENT **WITH**
- **JUDGMENT**
- **ACTIONS**
- **OUTCOME & FEEDBACK**

HUMAN VS. MACHINE

- MACHINE| AI BEAT HUMAN IN MANY NARROW TASKS
- HUMAN ARE BETTER WITH FEW DATA [MODEL OF THE WORLD]
- HUMAN PUT PREDICTIONS INTO CONTEXT AND ASSESS TRADEOFFS
- REQUIRE RETHINKING MANY JOBS

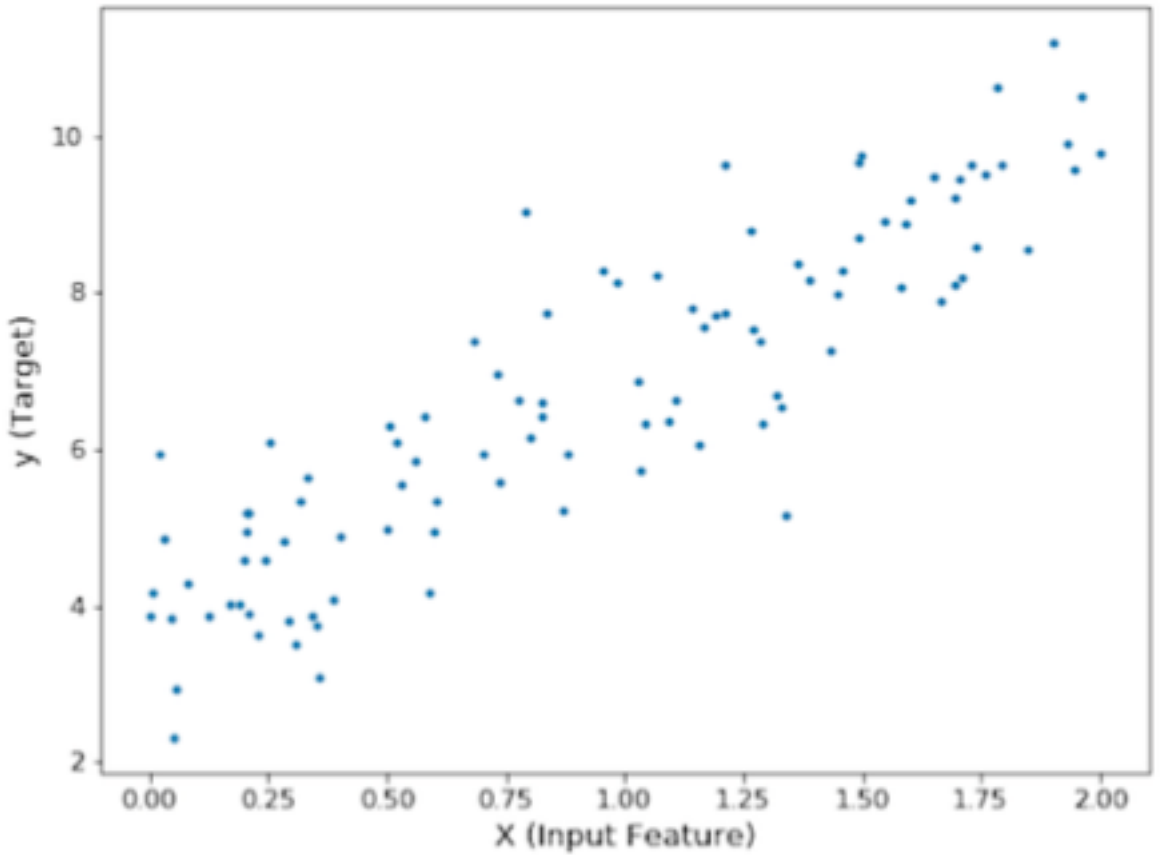
PREDICTIONS MACHINE | OPENING THE ENGINE

AN ATTEMPT TO DEMISTIFY ML|DL

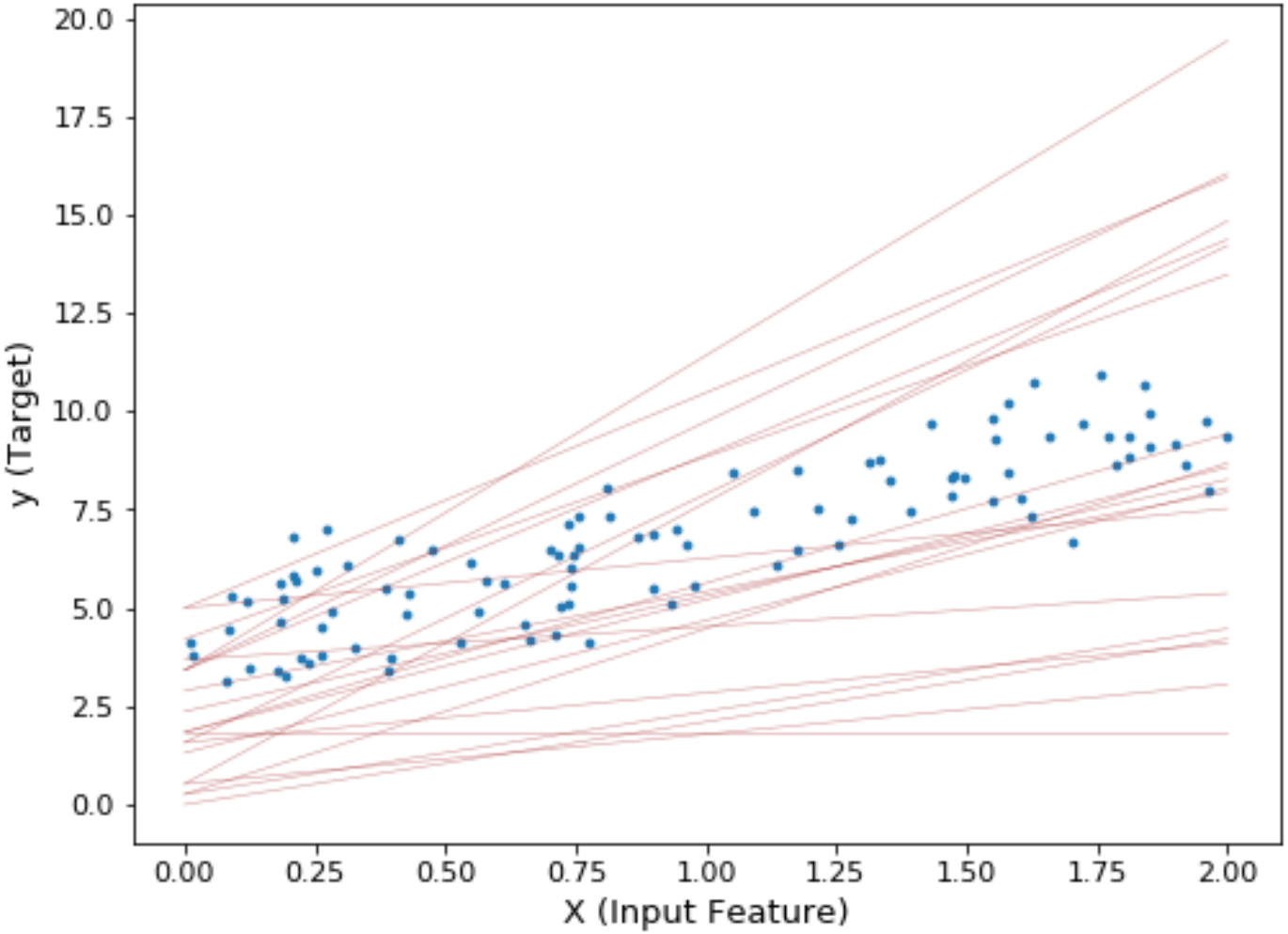
IS THERE ANY PATTERN IN YOUR DATA?

	Feature	Target
0	1.209897	9.639719
1	0.833615	7.729447
2	1.694408	8.101811
3	0.208314	3.905511
4	0.190894	4.021537
5	1.666496	7.902070
6	1.449120	7.985356
7	1.741486	8.575577
8	1.319825	6.688907
9	0.056584	2.928961
10	1.999263	9.774487

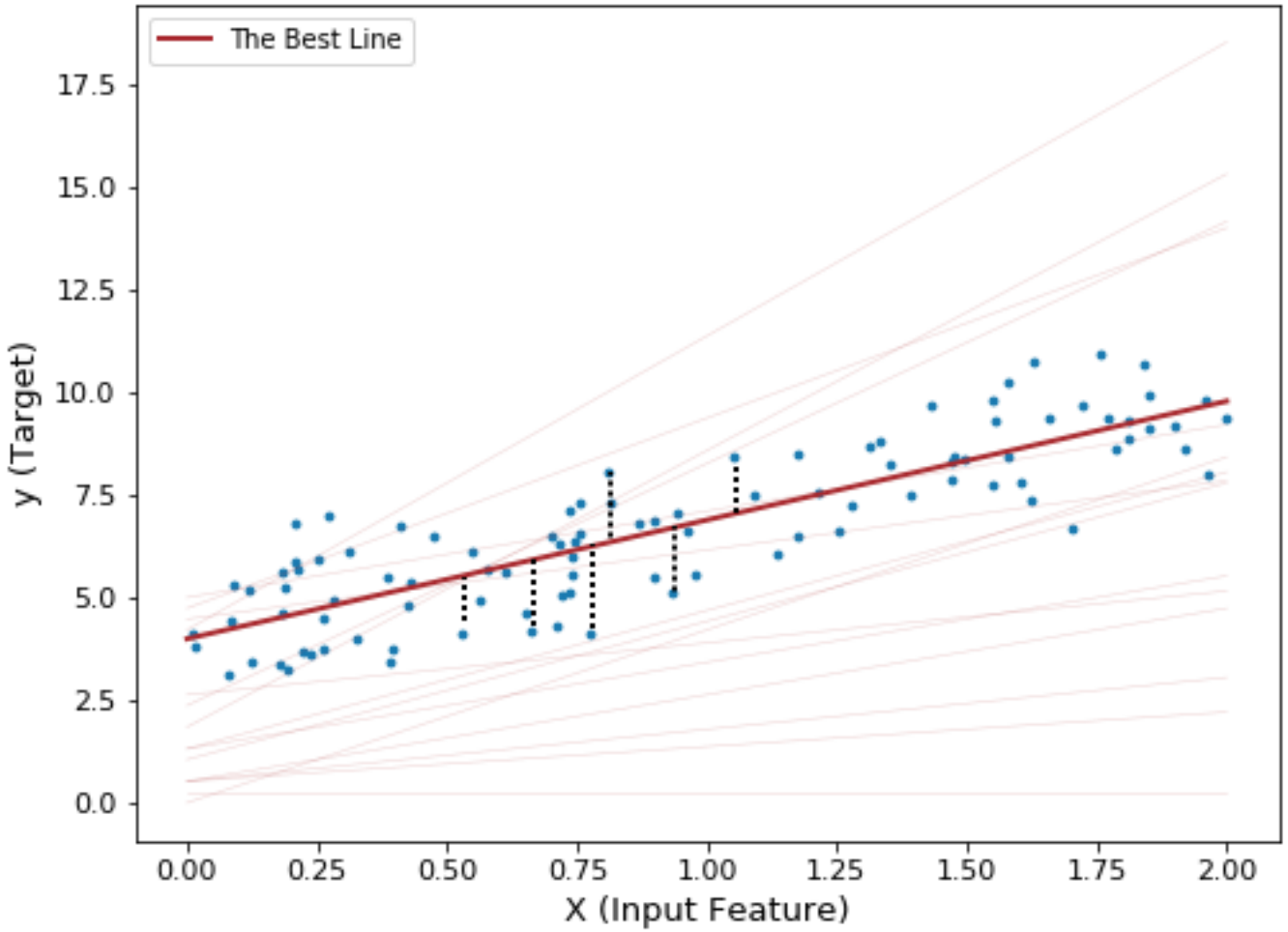
...



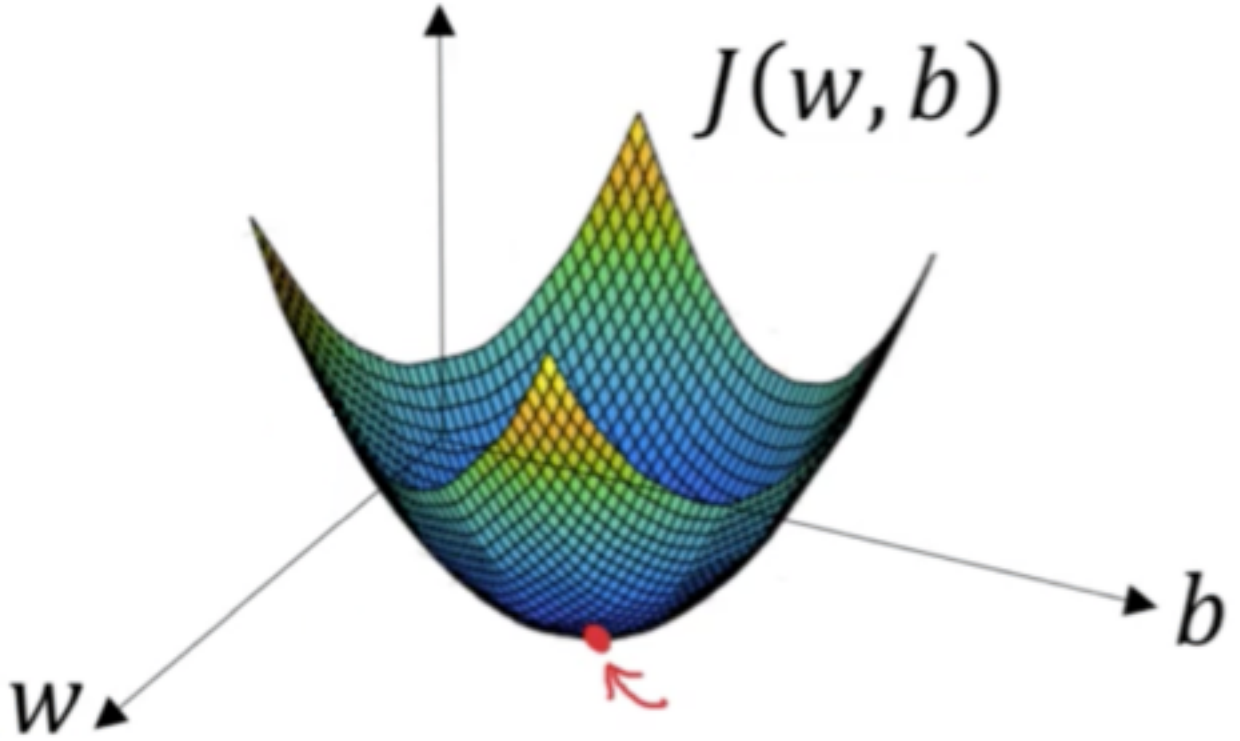
BUT WHICH LINE?



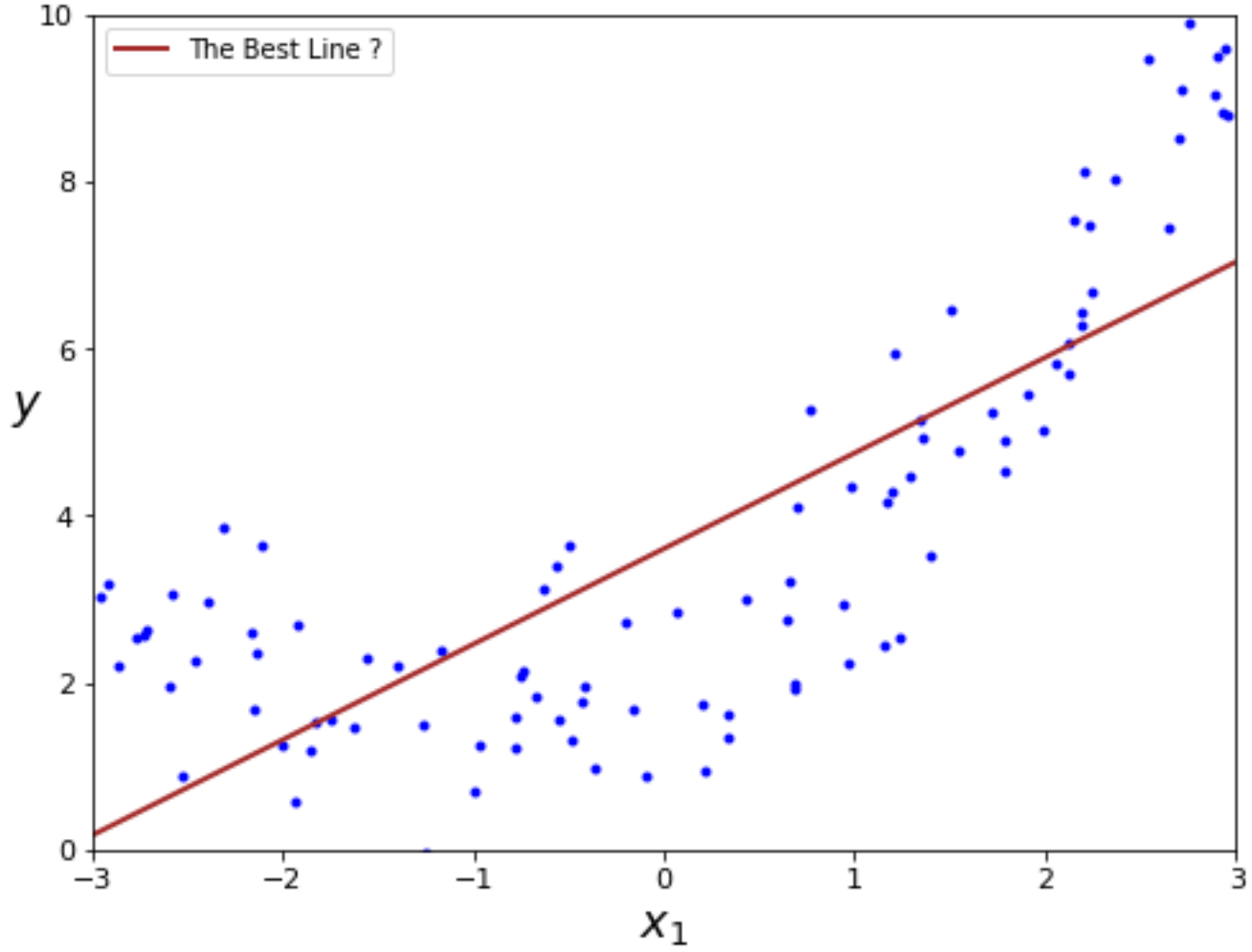
DEFINE AN EVALUATION METRIC?



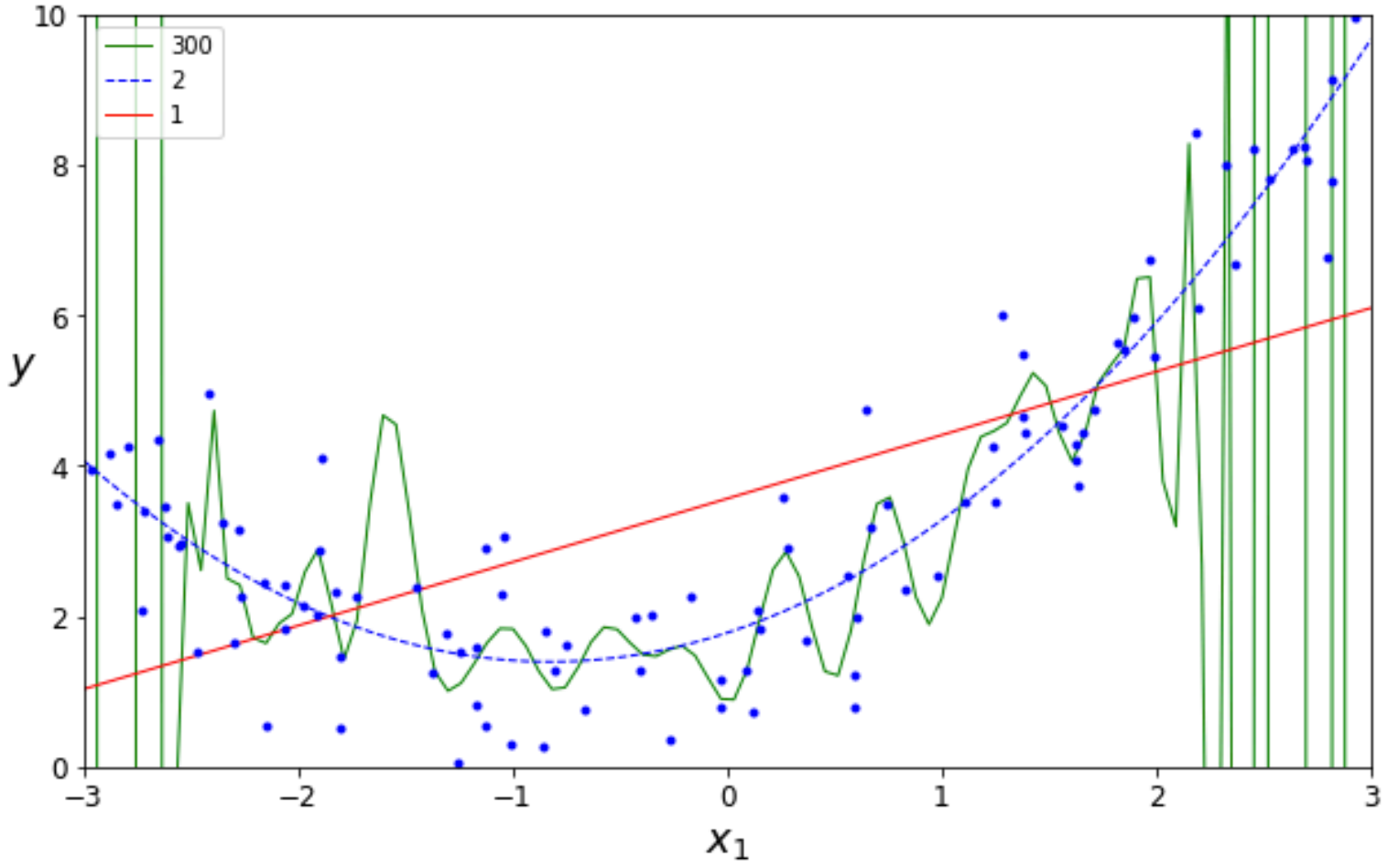
MINIMIZE ERROR



BUT WAIT, WHICH BEST LINE?

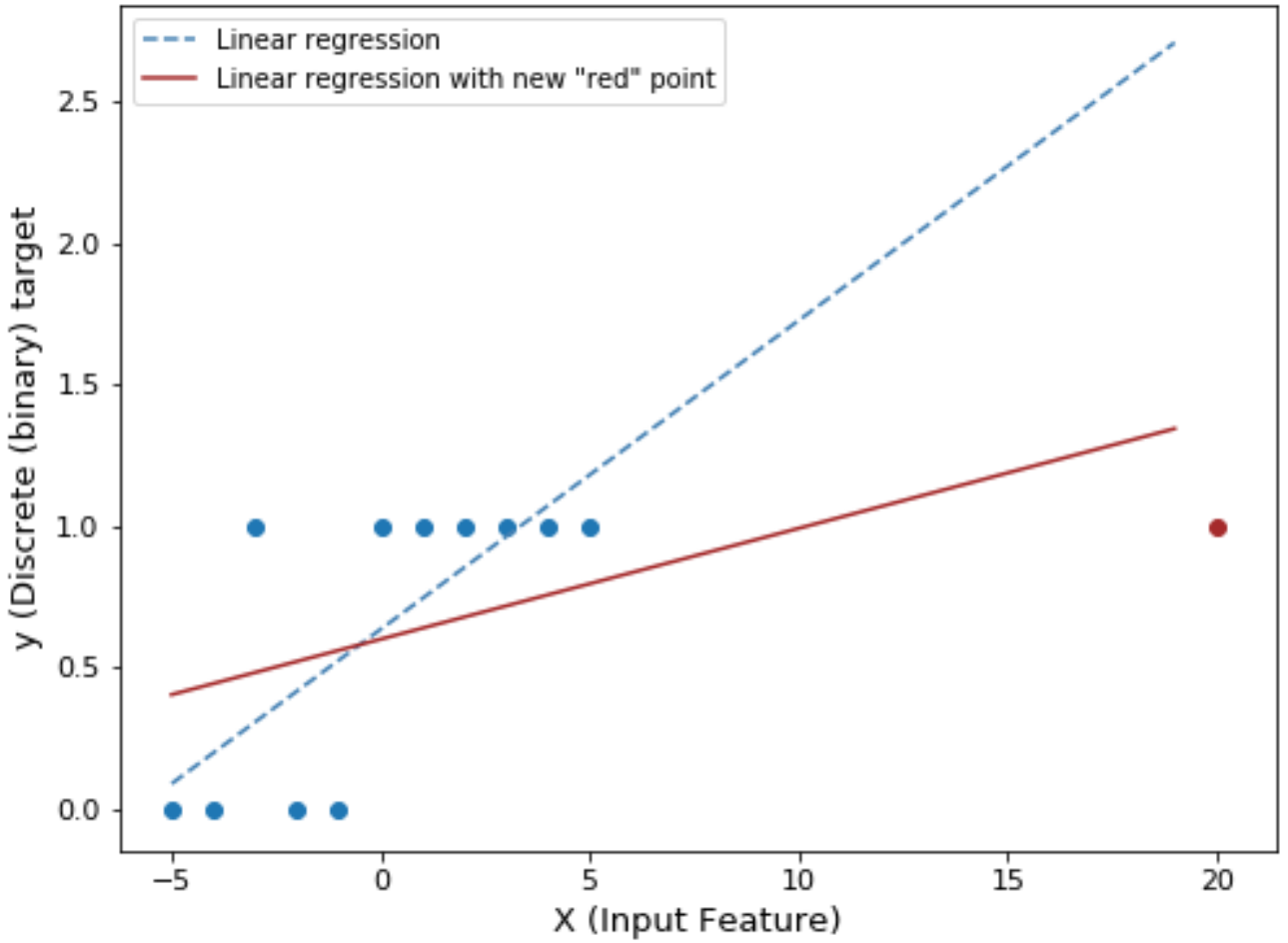


WHICH CURVE IS BEST? A QUESTION OF CAPACITY

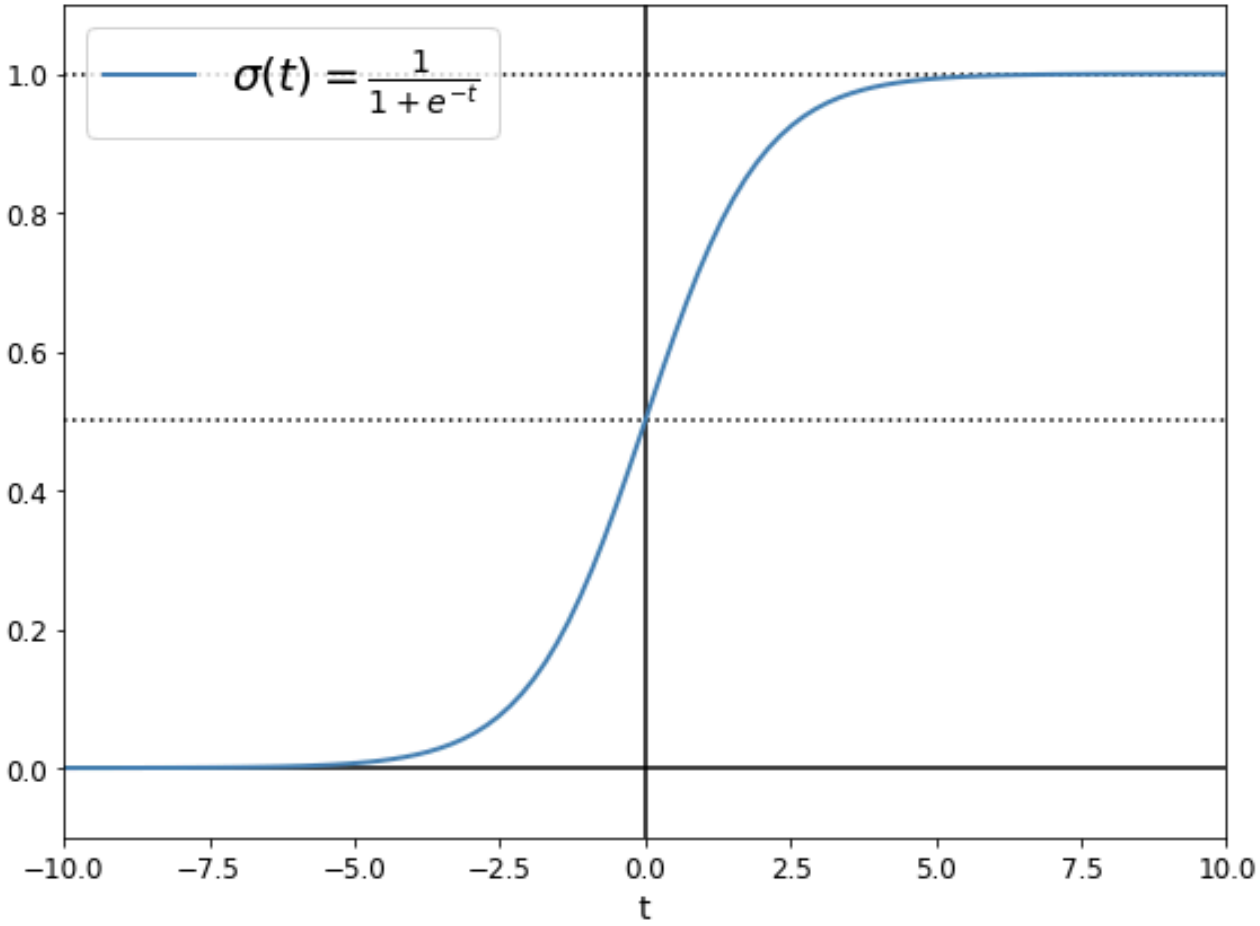


WHAT ABOUT CLASSIFICATION PROBLEM?

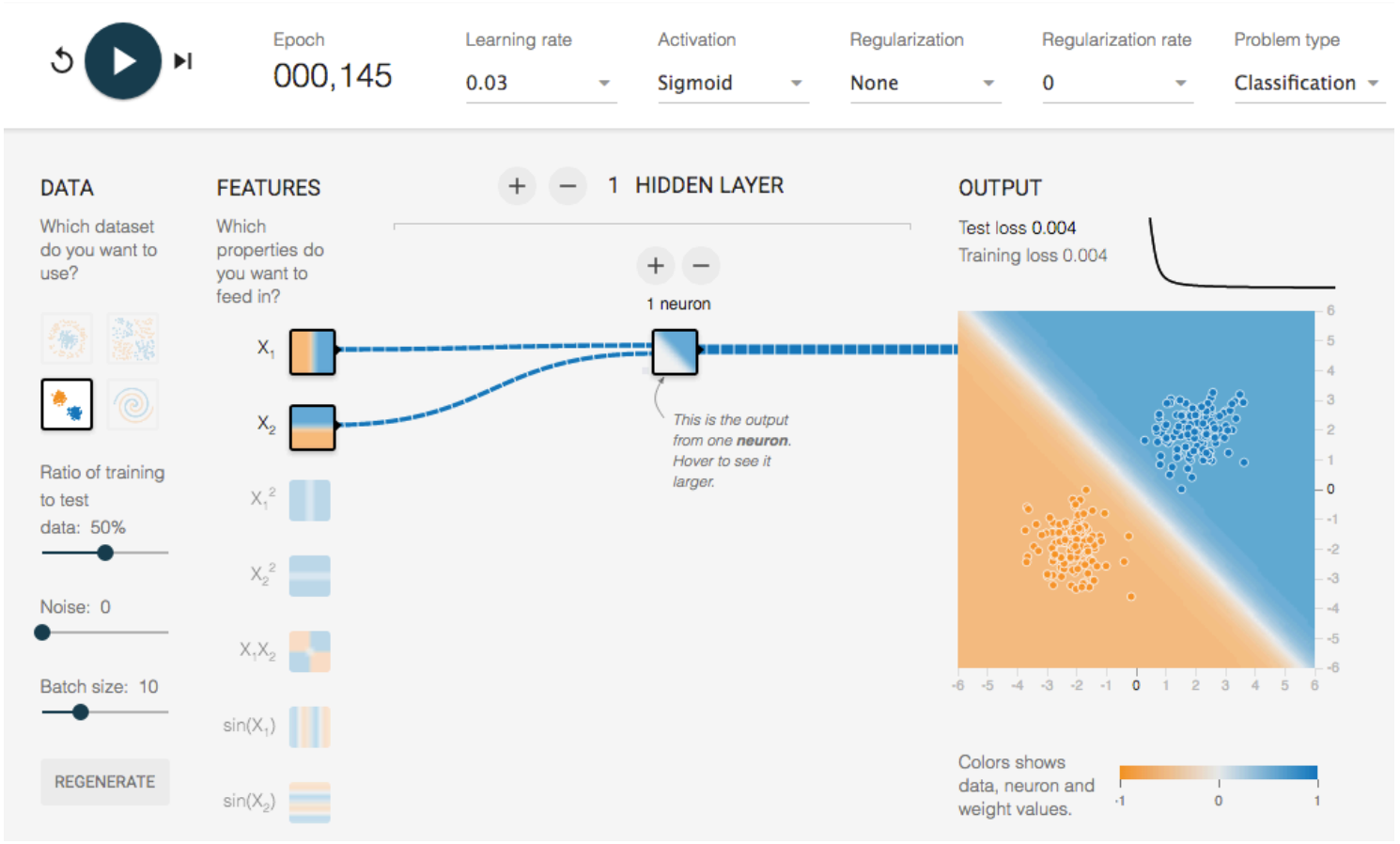
LOGISTIC REGRESSION RATIONALE



LOGISTIC REGRESSION | SIGMOID FUNCTION

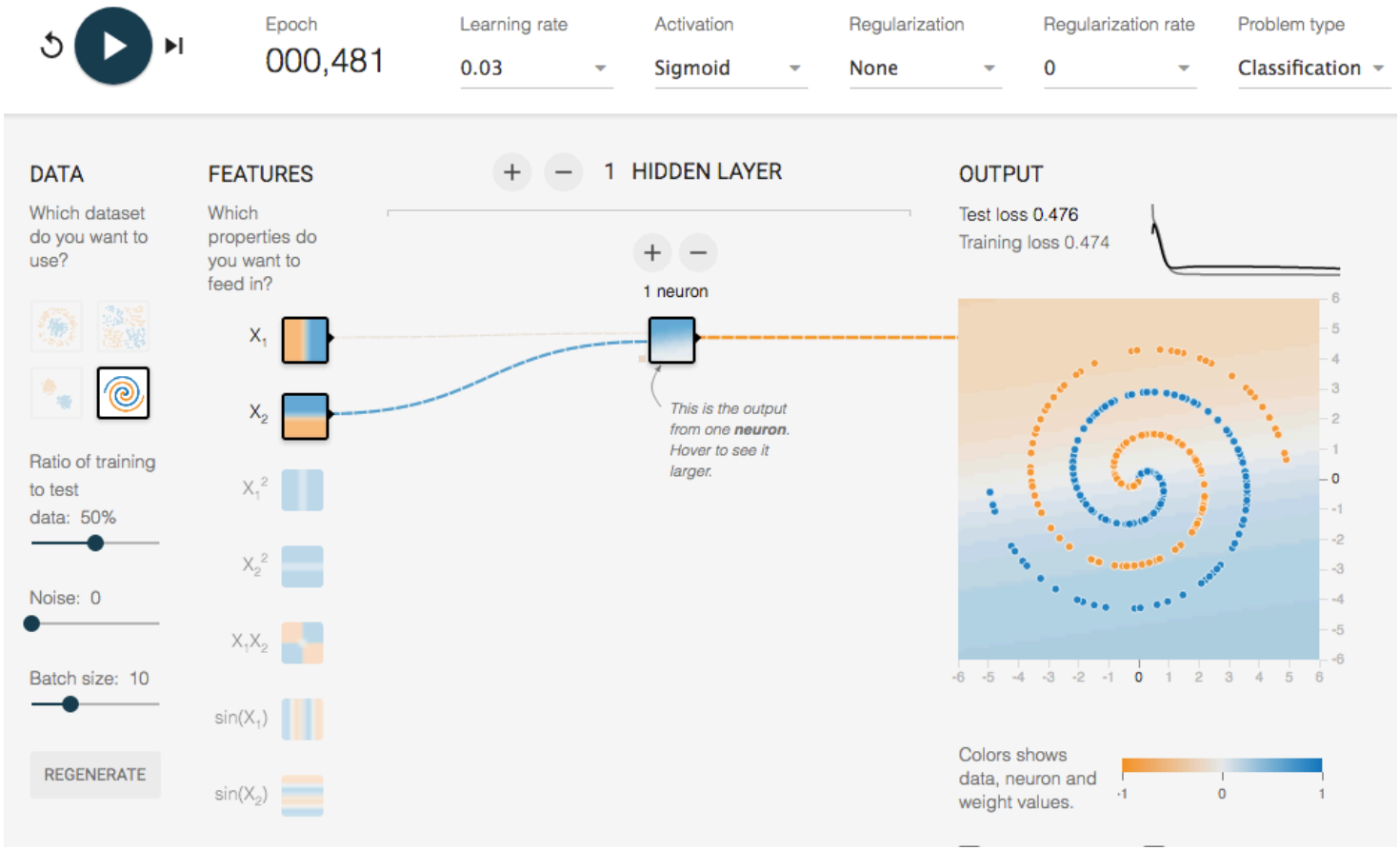


LOGISTIC REGRESSION | AT CRUISING SPEED

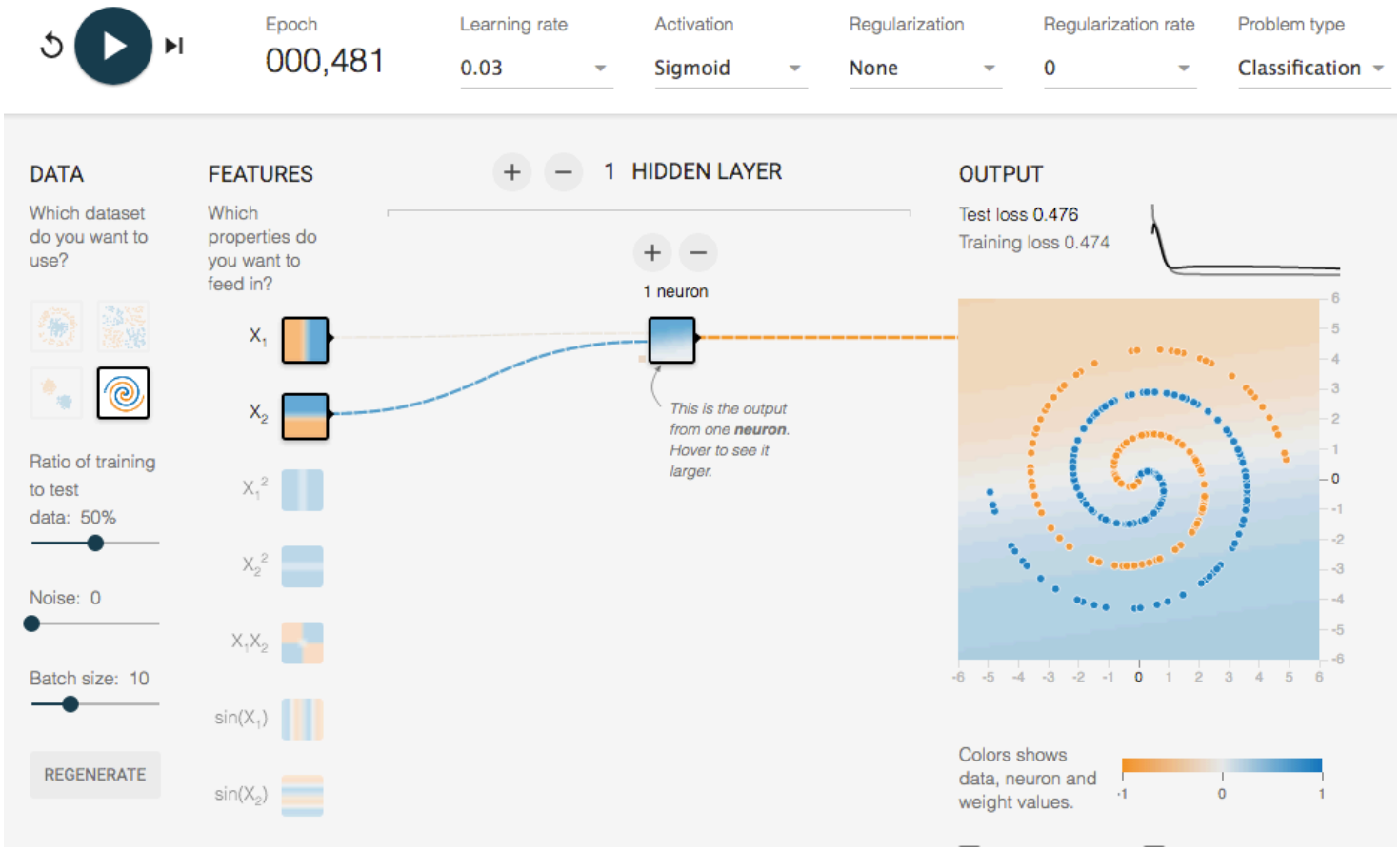


http://playground.tensorflow.org/ https://goo.gl/d5wwH3

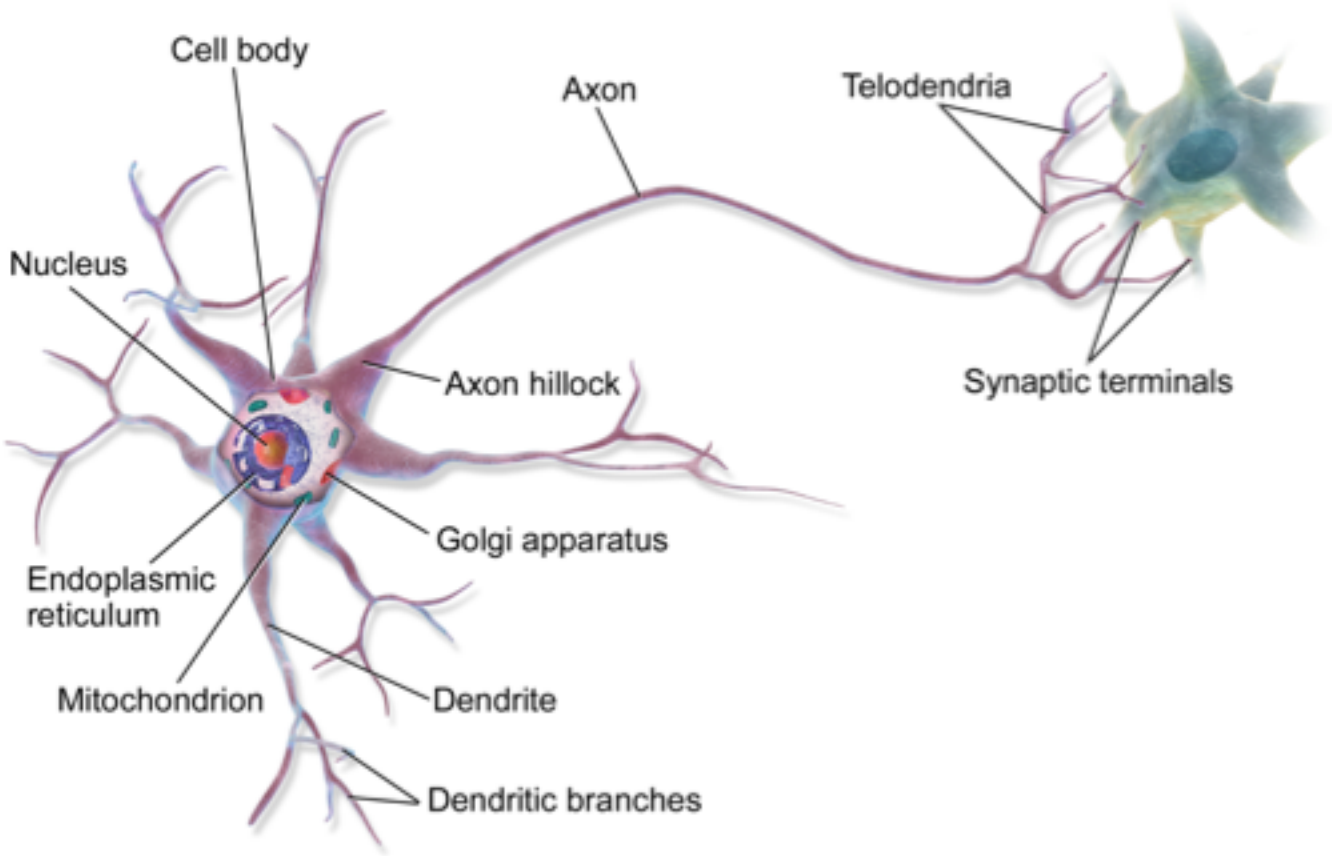
LOGISTIC REGRESSION | IN TROUBLE



LOGISTIC REGRESSION | IN TROUBLE

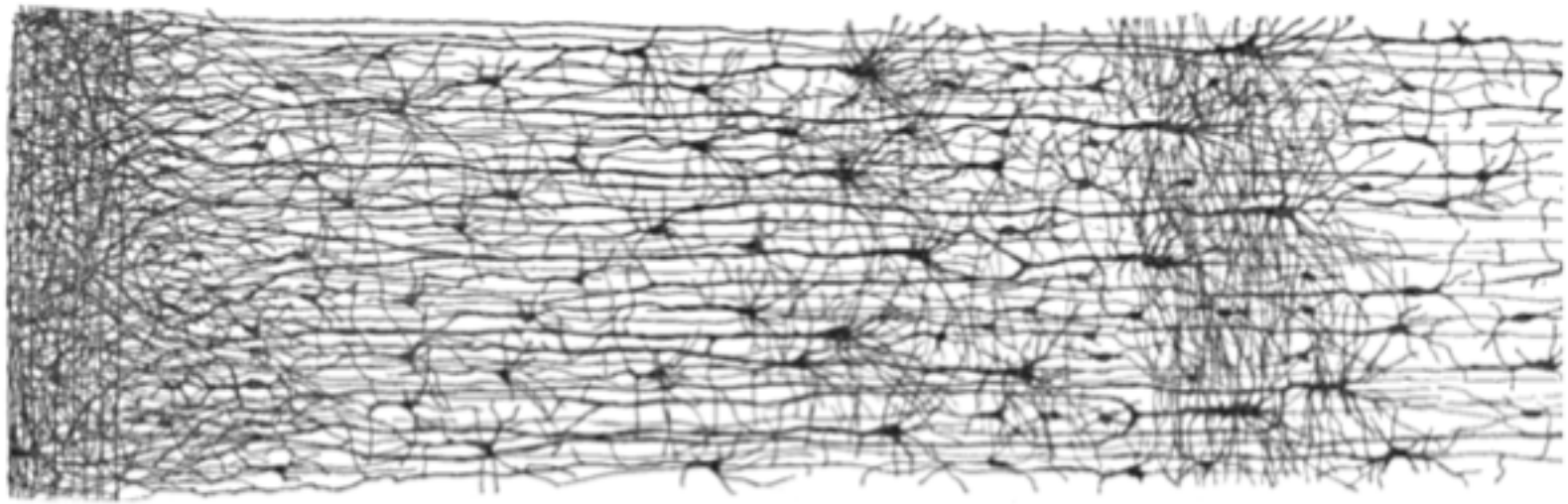


BIOLOGICAL NEURON



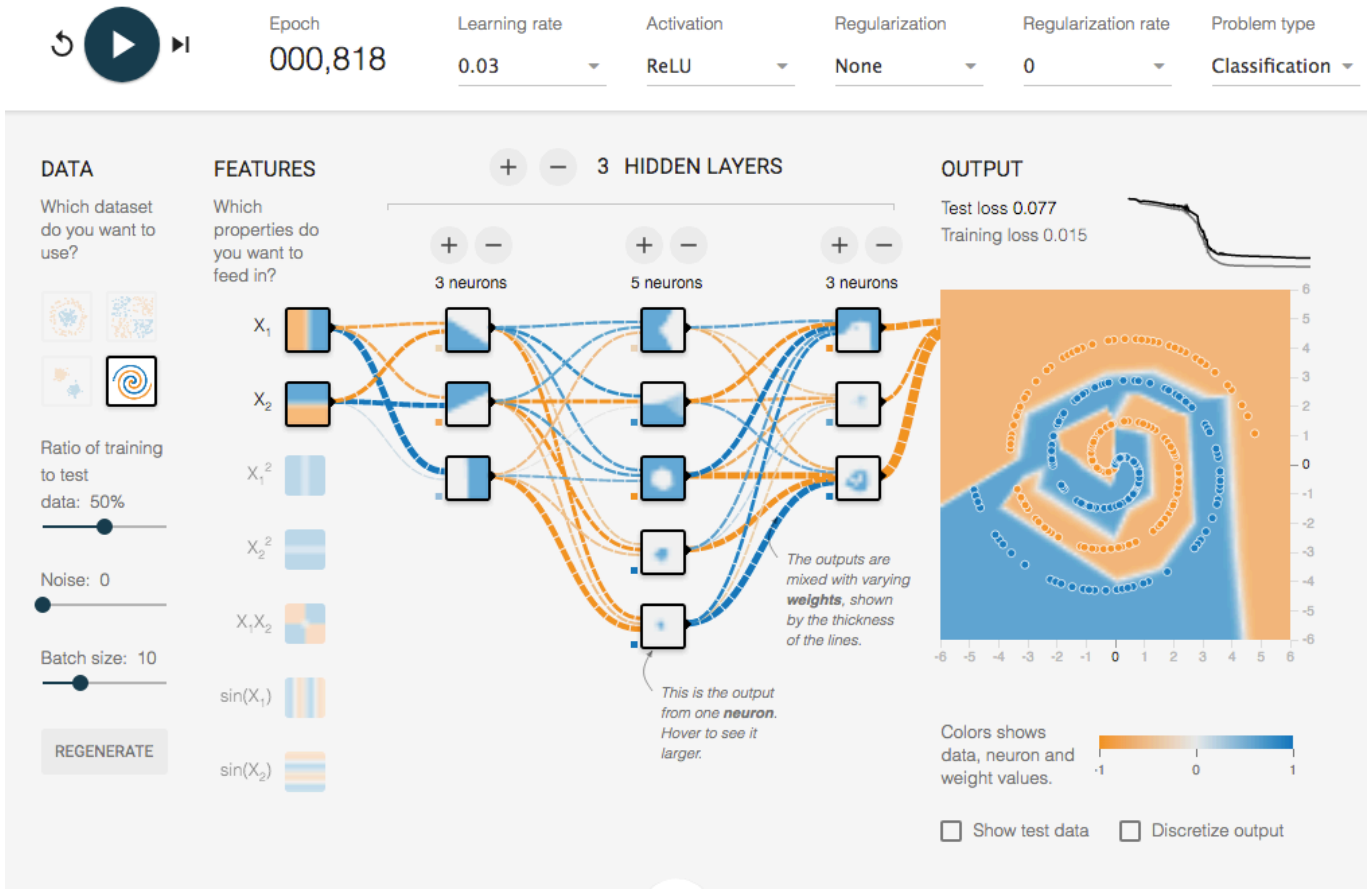
https://en.wikipedia.org/wiki/Neuron#/media/File:Blausen_0657_MultipolarNeuron.png

MULTIPLE LAYERS IN BIOLOGICAL NEURAL NETWORK



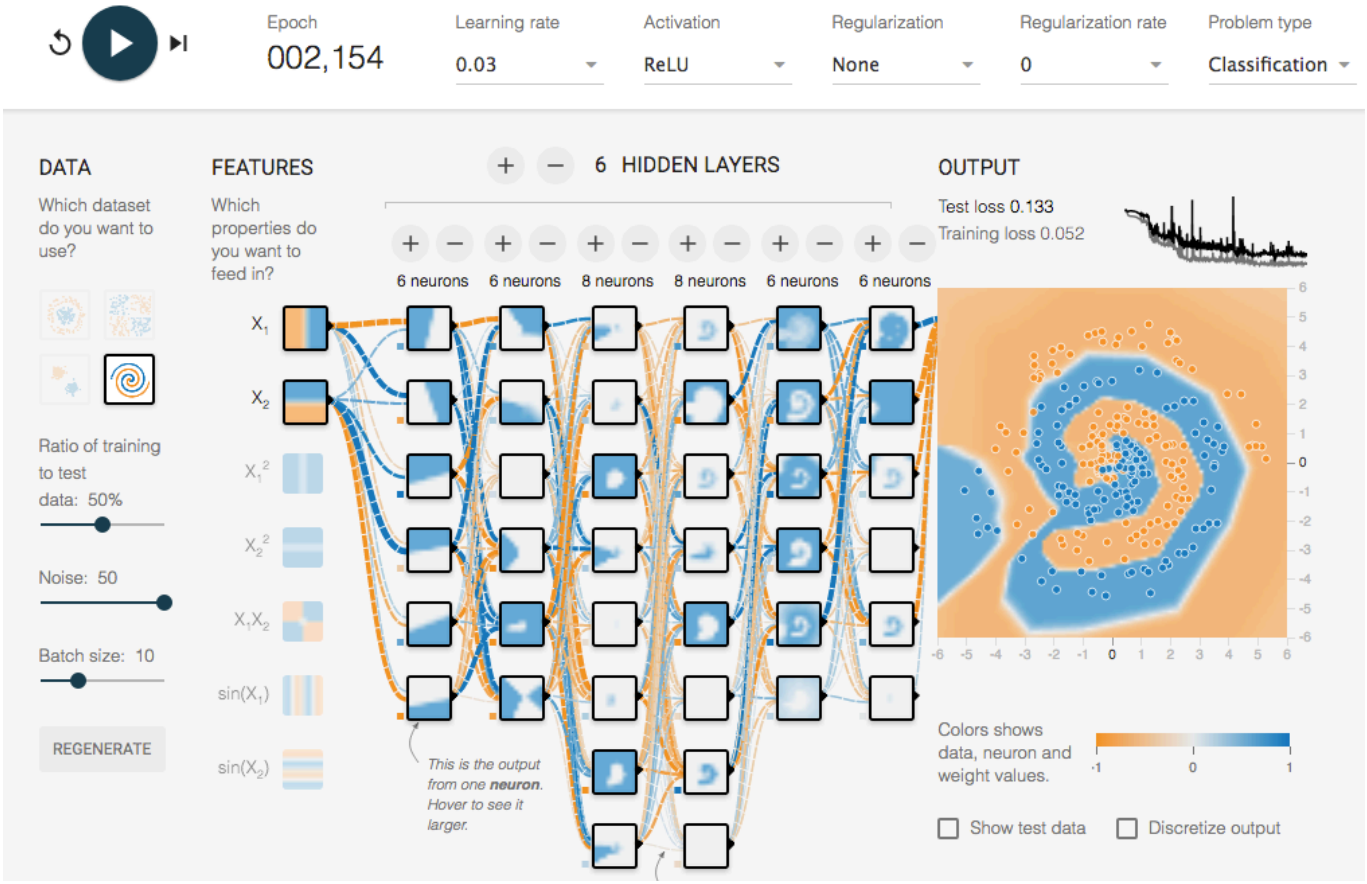
https://en.wikipedia.org/wiki/Cerebral_cortex

ARTIFICIAL NEURAL NETWORKS COMES ON STAGE



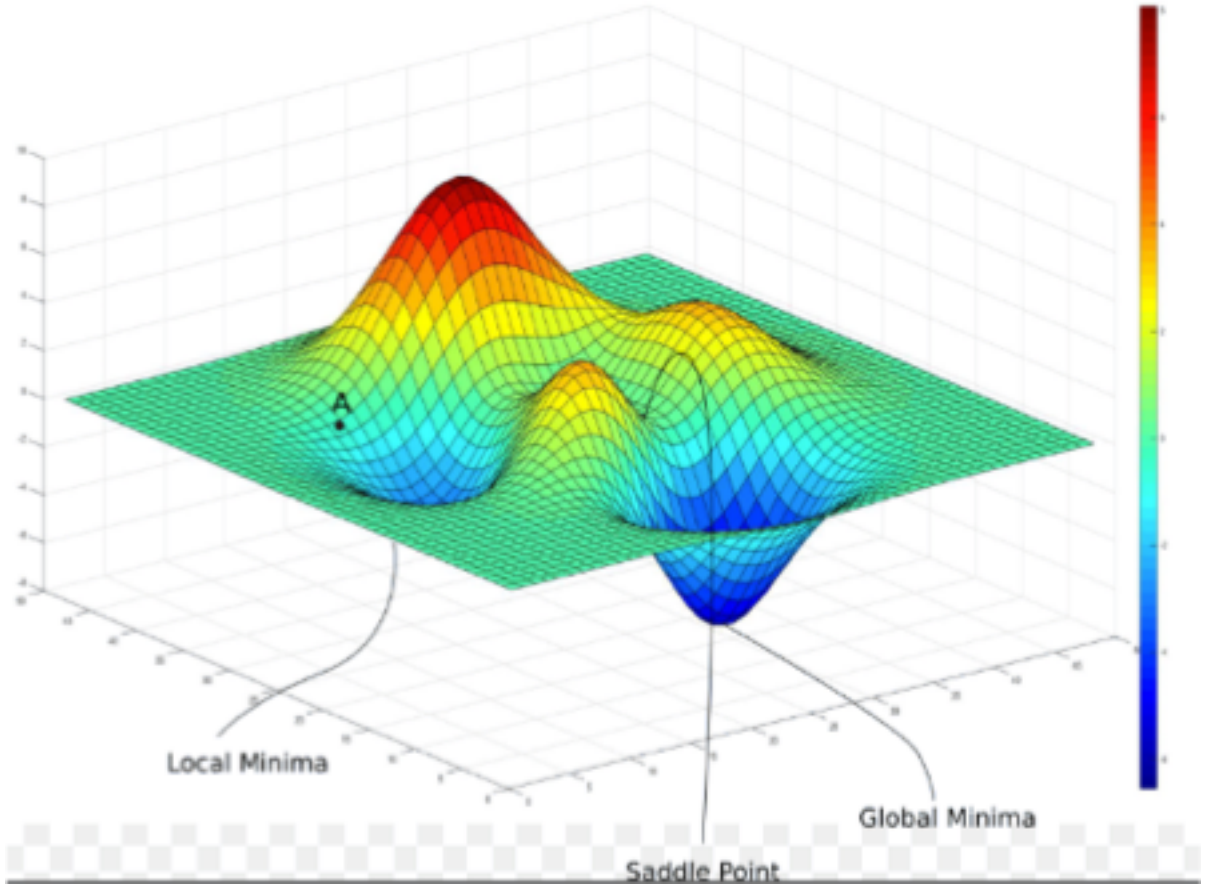
https://goo.gl/nqcvpV

ARTIFICIAL DEEP NEURAL NETWORKS COMES ON STAGE



https://goo.gl/6BJ83E

NOT A BOIL ANYMORE | OPTIMIZATION TRICKS



FROM TWO PARAMETERS TO MILLIONS OF THEM

« WITH FOUR PARAMETERS I CAN FIT AN ELEPHANT, AND WITH FIVE I CAN MAKE HIM WIGLLE HIS TRUNK»

— JOHN VON NEUMANN

FINETUNING YOUR « MOULD » | CAPACITY

OVERFITTING AND UNDERFITTING

ML PRACTITIONER MINDSET | PREDICTIVE POWER FIRST

AS OPPOSED TO CLASSICAL STATISTICS MINDSET

HIGH EXPLANATORY POWER = HIGH PREDICTIVE POWER

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

https://arxiv.org/pdf/1101.0891.pdf

RIGHT DATA & RIGHT MODEL CAPACITY

TRAINING DATASET

https://goo.gl/i2RCXF

MODEL "MOULD"

http://www.20th.ch/antiquites_industrielles_objets.htm

TEST DATASET

PHOTO BY KENICHI HIGASHI

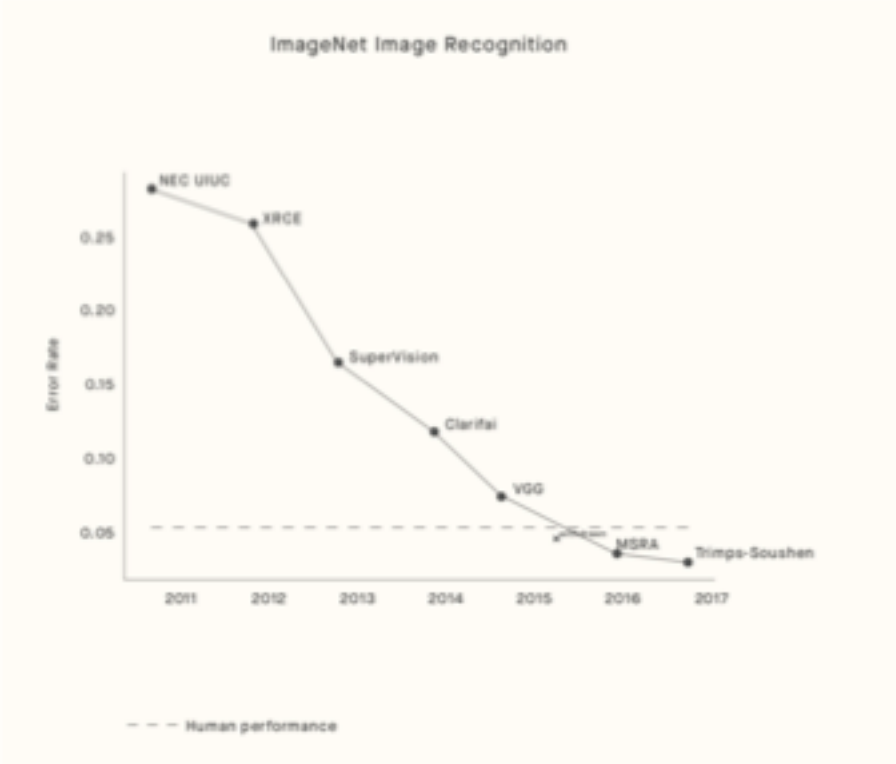
« UNIVERSAL » WORKFLOW OF DEEP LEARNING

1. DEFINE THE PROBLEM AND ASSEMBLE A DATASET
2. CHOOSE A MEASURE OF SUCCESS & DECIDE ON AN EVALUATION PROTOCOL
3. PREPARE DATA
4. DEVELOP A MODEL THAT DOES BETTER THAN A BASELINE
5. TWEAK CAPACITY BY TUNING YOUR HYPERPARAMETERS

« UNIVERSAL » WORKFLOW OF DEEP LEARNING

1. DEFINE THE PROBLEM AND ASSEMBLE A DATASET
2. CHOOSE A MEASURE OF SUCCESS & DECIDE ON AN EVALUATION PROTOCOL
3. PREPARE DATA
4. DEVELOP A MODEL THAT DOES BETTER THAN A BASELINE
5. TWEAK CAPACITY BY TUNING YOUR HYPERPARAMETERS

ML|DL|AI VS. HUMAN



https://arxiv.org/pdf/1802.07228.pdf
http://www.image-net.org/

- DEEP LEARNING > HUMAN-LEVEL PERFORMANCE
- BUT IN PARTICULAR DOMAINS (IMAGE RECOGNITION, …)
- BUT ON VERY SPECIFIC/NARROW TASKS [FOR NOW]

1. DATA SCIENCE INTRODUCTION | OVERVIEW

2. PREDICTION MACHINES | THE ENGINE

3. DEPLOYMENT MOMENTUM

4. [POTENTIAL] TRACKS TO DEVELOP CAPACITIES

LURING WITH SIREN SONGS

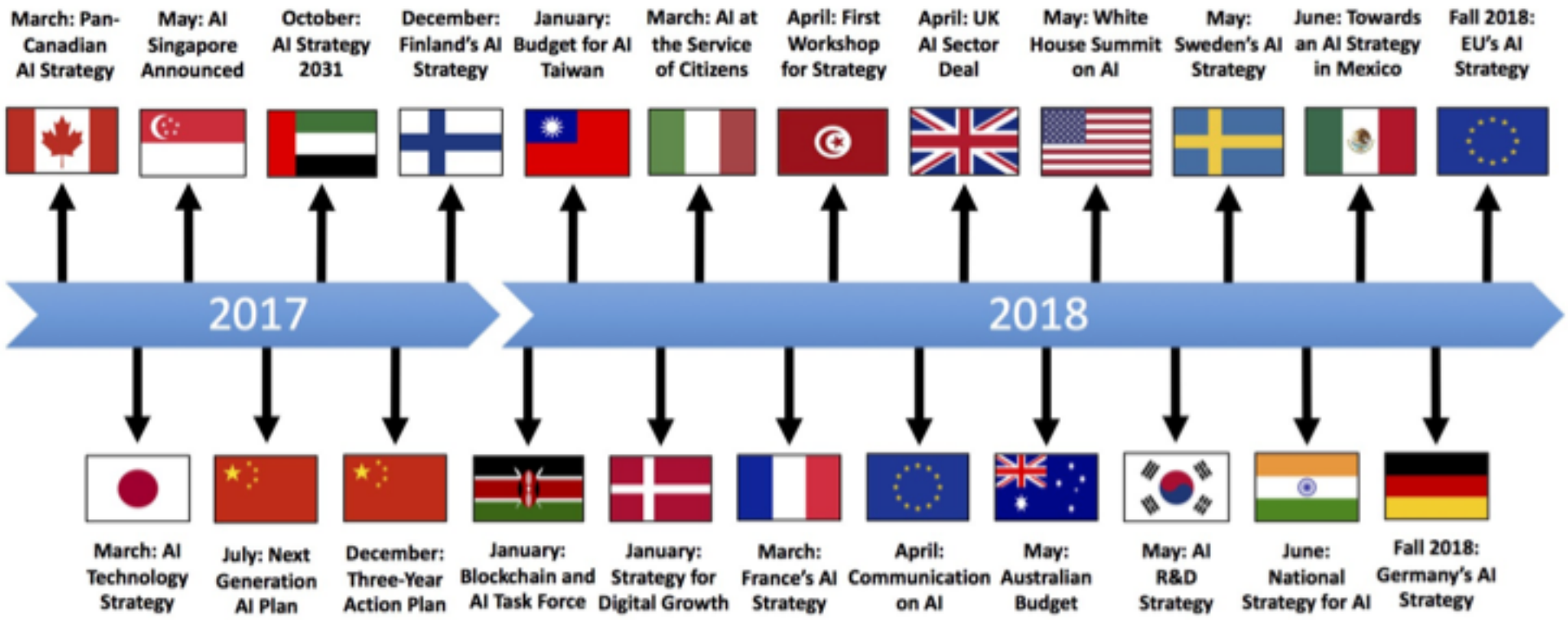
« DATA IS THE NEW OIL »

« AI IS THE NEW ELECTRICITY »

« THE FOURTH INDUSTRIAL REVOLUTION »

…

AN IDEA OF THE RUSH



Source: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

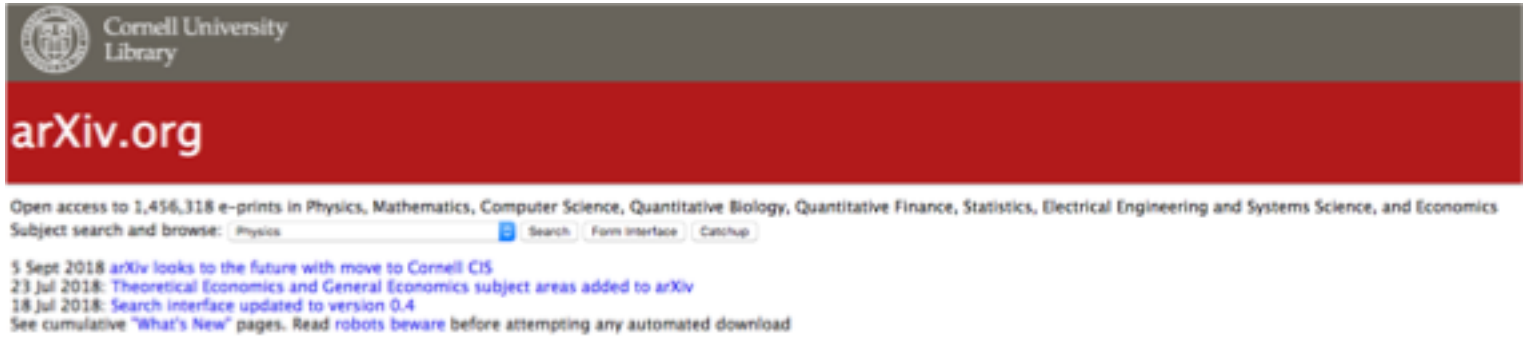
THE STUFF OF AN AI SUPERPOWER

1. ABUNDANT DATA
2. TENACIOUS ENTREPRENEURS*
3. WELL TRAINED AI PRACTITIONERS
4. SUPPORTIVE POLICY ENVIRONMENTS*

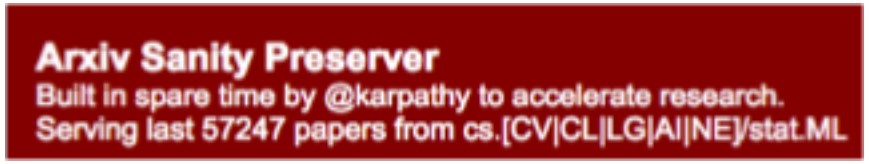
https://aisuperpowers.com/

* ON STRATEGIC SECTOS

FAST-PACED RESEARCH VS. NARROW AI



https://arxiv.org/



http://www.arxiv-sanity.com/

1. DATA SCIENCE INTRODUCTION | OVERVIEW

2. PREDICTION MACHINES | THE ENGINE

3. DEPLOYMENT MOMENTUM

4. [POTENTIAL] TRACKS TO DEVELOP CAPACITIES

ML/DL/AI DEMOCRATIZATION

- OUTSTANDING ONLINE COURSES
- BEST PRATICES DISSEMINATION
- COMPETITIONS

SOME OUTSTANDING RESOURCES ONLINE

https://www.datacamp.com

https://www.deeplearning.ai/

https://www.fast.ai/

https://eu.udacity.com/

https://spinningup.openai.com

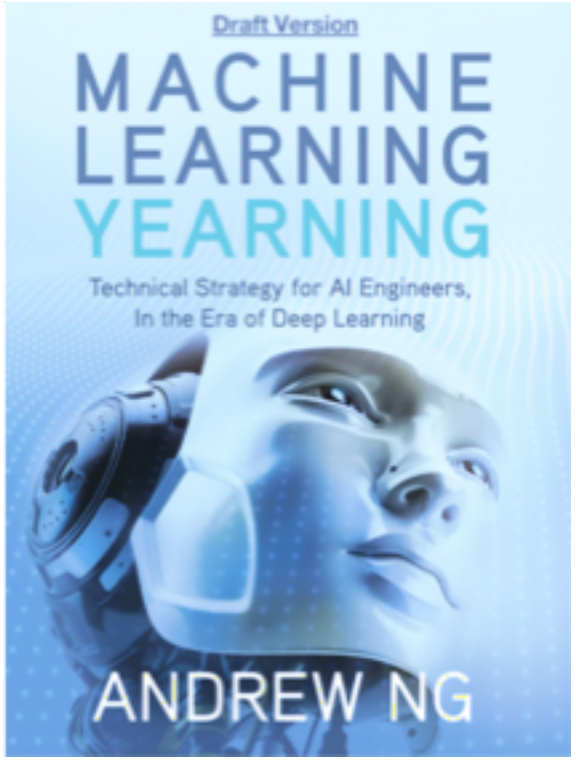
ML COMPETITIONS

https://www.kaggle.com/competitions

https://zindi.africa/

https://challenger.ai/

BEST PRACTICES



http://www.mlyearning.org/

LANDING AI

AI Transformation Playbook

How to lead your company into the AI era

AI (Artificial Intelligence) technology is now poised to transform every industry, just as electricity did 100 years ago. Between now and 2030, it will create an estimated $13 trillion of GDP growth[1]. While it has already created tremendous value in leading technology companies such as Google, Baidu, Microsoft and Facebook, much of the additional waves of value creation will go beyond the software sector.

This AI Transformation Playbook draws on insights gleaned from leading the Google Brain team and the Baidu AI Group, which played leading roles in transforming both Google and Baidu into great AI companies. It is possible for any enterprise to follow this Playbook and become a strong AI company, though these recommendations are tailored primarily for larger enterprises with a market cap from $500M to $500B.

These are the steps I recommend for transforming your enterprise with AI, which I will explain in this playbook:

1. **Execute pilot projects to gain momentum**
2. **Build an in-house AI team**
3. **Provide broad AI training**
4. **Develop an AI strategy**
5. **Develop internal and external communications**

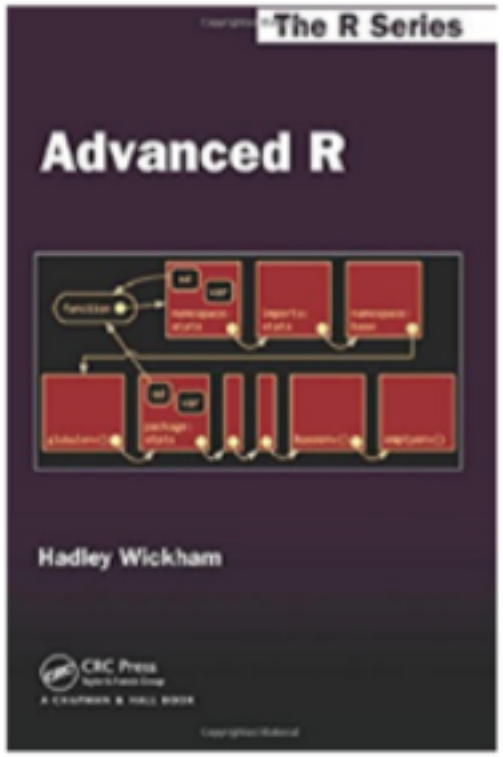
OUTSTANDING PYTHON REFERENCE BOOKS

https://github.com/wesm/pydata-book

https://github.com/ageron/handson-ml

https://github.com/fchollet/deep-learning-with-python-notebooks

OUTSTANDING R REFERENCE BOOKS [FREE]

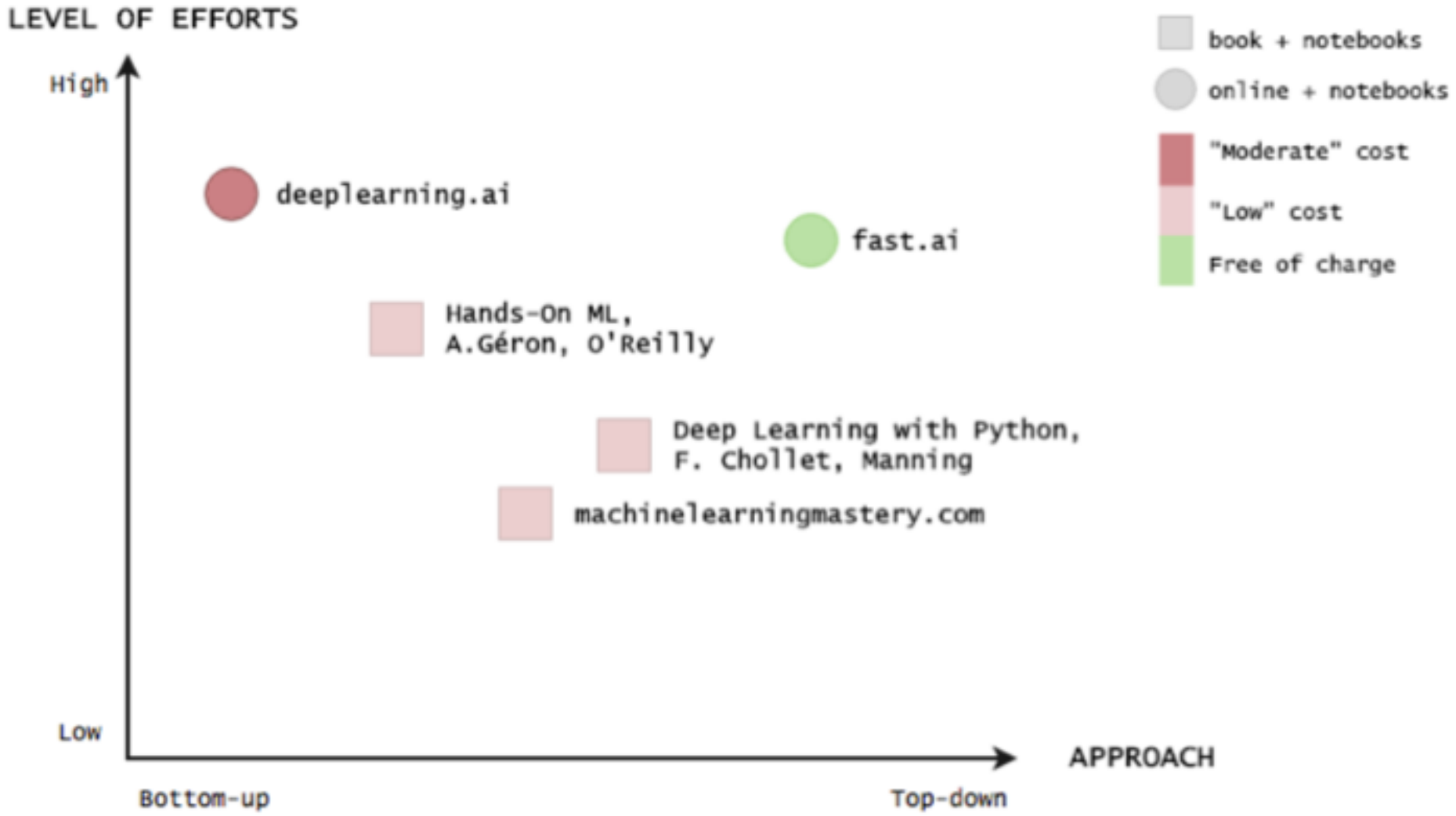


http://r4ds.had.co.nz/

http://adv-r.had.co.nz/

http://r-pkgs.had.co.nz/

A DIVERSITY OF APPROACHES BUT ALL GOOD!



AT ADRIATICO GUEST HOUSE
TILL THURSDAY EVENING