Machine Learning

PHYS 453

Dr. Daugherity

Welcome to ML

Today's goal:

- what topics this class covers (i.e. what is ML?)
- what it will be like
- why am I teaching it

Part 0

WHY IS A PHYSICIST TEACHING THIS?

About Me

Bio:

- Grew up in Oklahoma City
- **2002 ACU: Physics and CS**
- 2008 PhD in Nuclear Physics from UT
- Married in 2000, two kids

Fun Facts:

- Have been paid as a professional model
- Once burned off an eyebrow in class
- Plays drums in Abilene's best cover band

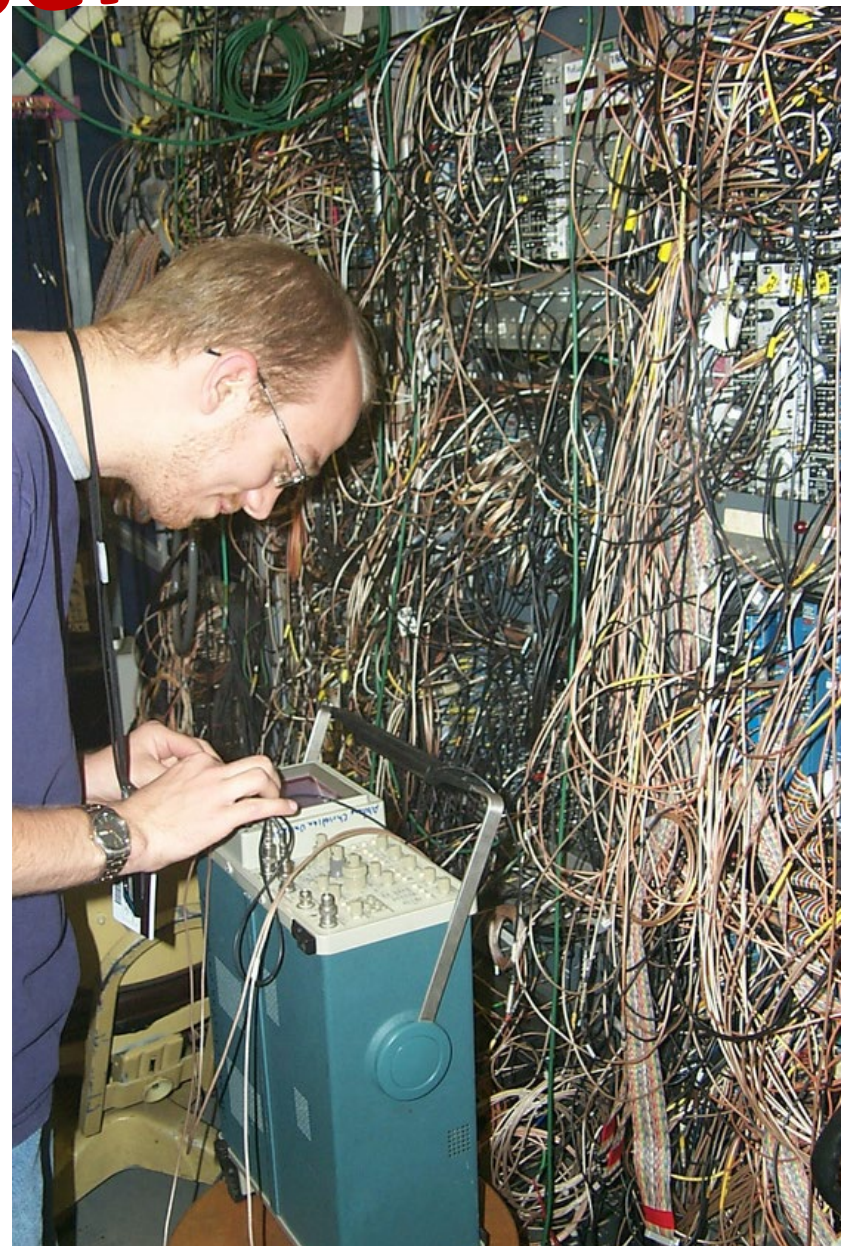
Research:

- Particle and Nuclear Physics
- Atom smashers
- Radiation Detectors
- Other interests: artificial intelligence, cosmology, amateur theology

Hobbies:

- Eating
- Putting things in a laser
- Building dangerous things

First Experiment



PhD Research

Two-Particle Correlations in Ultra Relativistic Heavy Ion Collisions

by

Michael Scott Daugherity, B.S.; B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

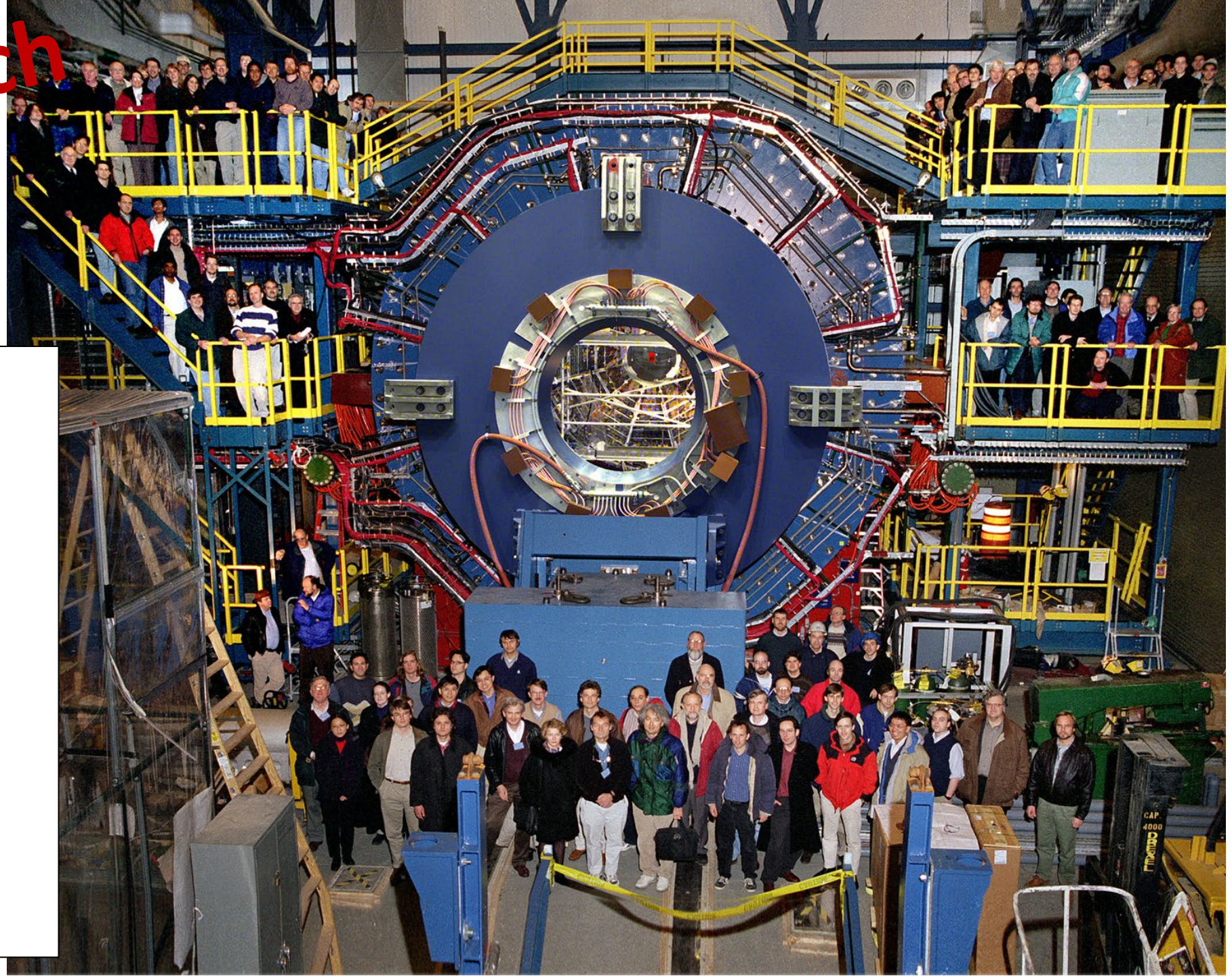
of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

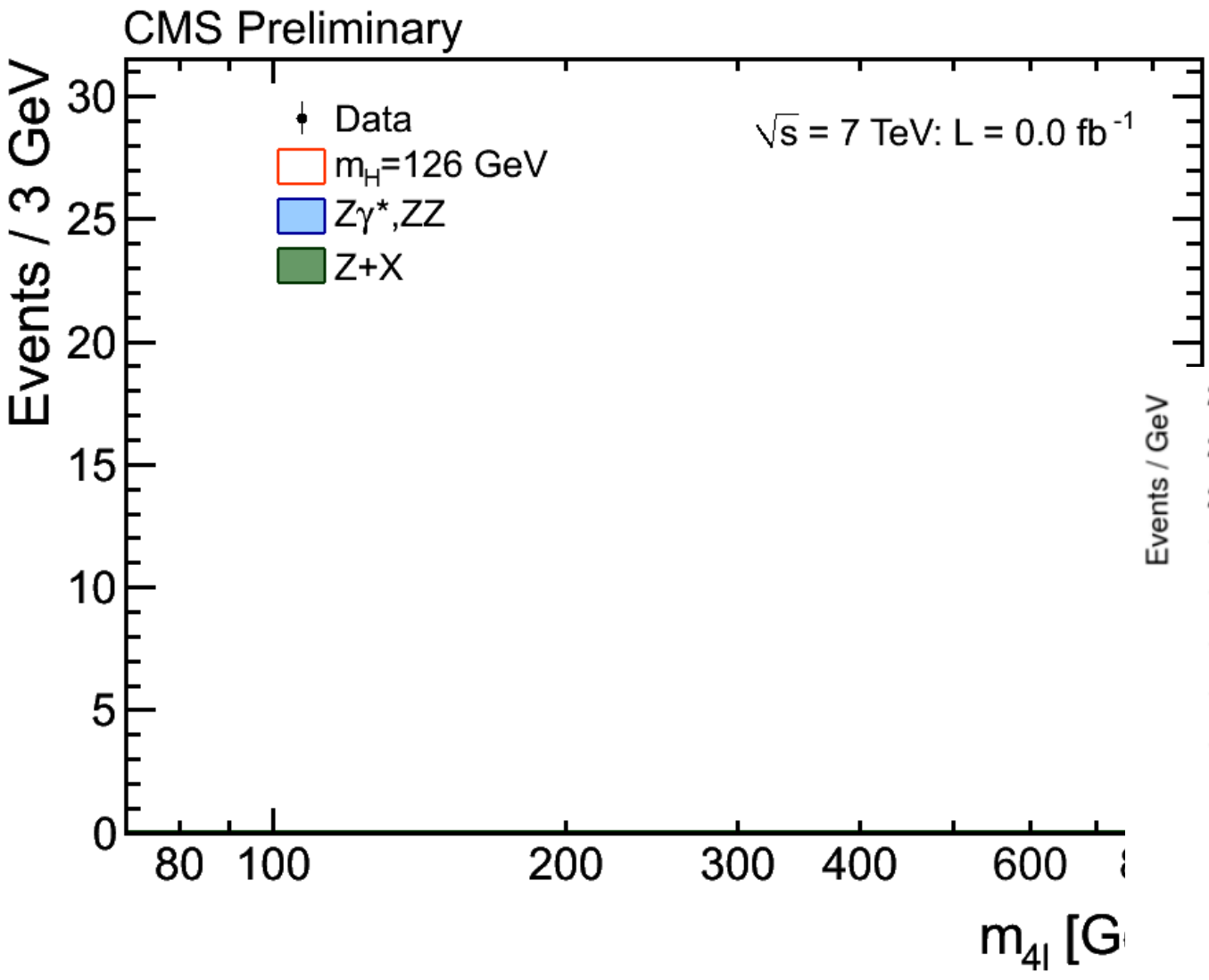
August 2008

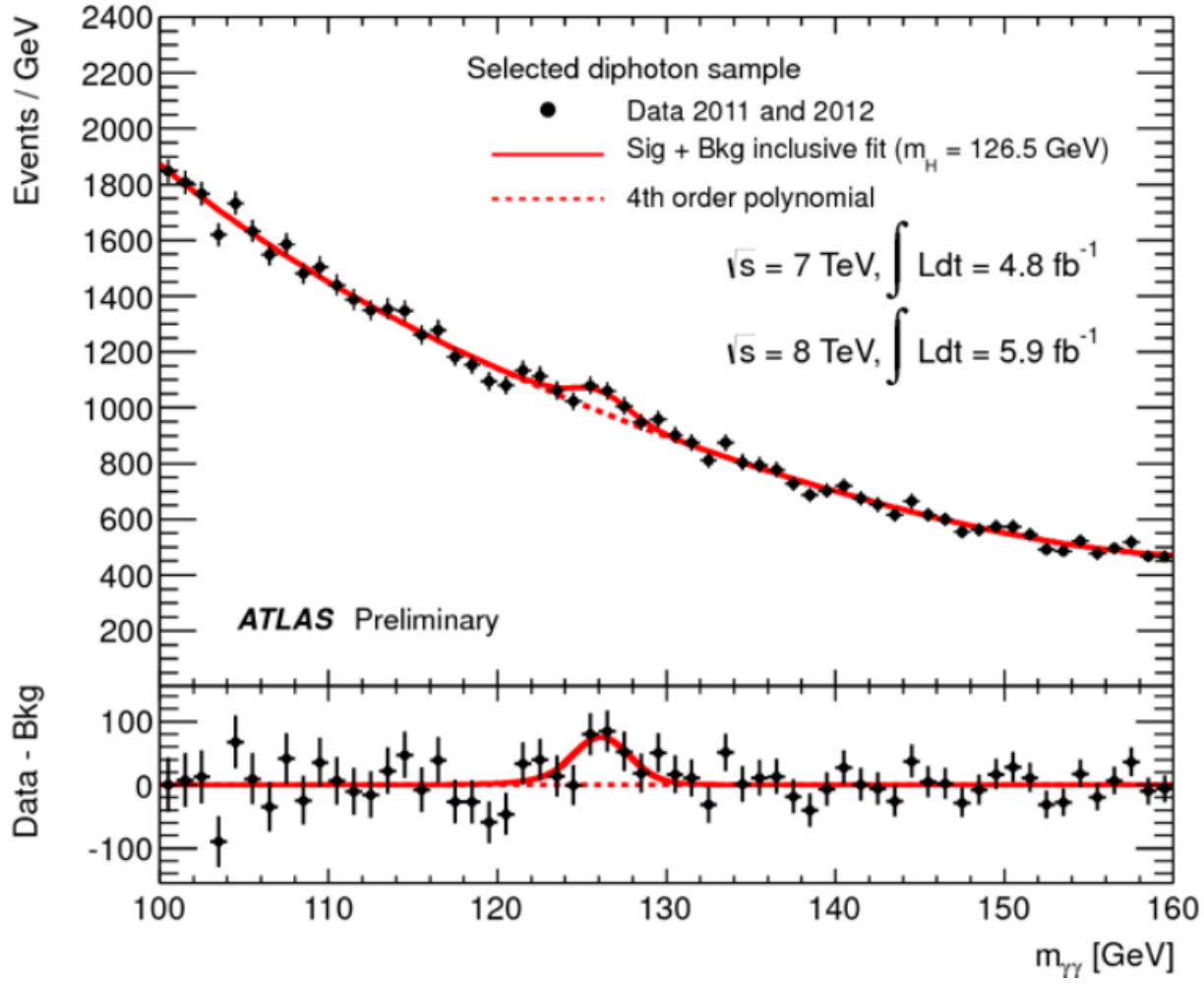


Before “Big Data” was a thing, there was high-energy physics…

A Search for the Higgs Boson Using Neural Networks in Events with Missing Energy and b-quark Jets in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV

T. Aaltonen,[24] J. Adelman,[14] B. Álvarez González[v],[12] S. Amerio[dd],[44] D. Amidei,[35] A. Anastassov,[39] A. Annovi,[20] J. Antos,[15] G. Apollinari,[18] A. Apresyan,[49] T. Arisawa,[58] A. Artikov,[16] J. Asaadi,[54] W. Ashmanskas,[18] A. Attal,[4] A. Aurisano,[54] F. Azfar,[43] W. Badgett,[18] A. Barbaro-Galtieri,[29] V.E. Barnes,[49] B.A. Barnett,[26] P. Barria[ff],[47] P. Bartos,[15] G. Bauer,[33] P.-H. Beauchemin,[34] F. Bedeschi,[47] D. Beecher,[31] S. Behari,[26] G. Bellettini[ee],[47] J. Bellinger,[60] D. Benjamin,[17] A. Beretvas,[18] A. Bhatti,[51] M. Binkley,[18] D. Bisello[dd],[44] I. Bizjak[jj],[31] R.E. Blair,[2] C. Blocker,[7] B. Blumenfeld,[26] A. Bocci,[17] A. Bodek,[50] V. Boisvert,[50] D. Bortoletto,[49] J. Boudreau,[48] A. Boveia,[11] B. Brau[a],[11] A. Bridgeman,[25] L. Brigliadori[cc],[6] C. Bromberg,[36] E. Brubaker,[14] J. Budagov,[16] H.S. Budd,[50] S. Budd,[25] K. Burkett,[18] G. Busetto[dd],[44] P. Bussey,[22] A. Buzatu,[34] K. L. Byrum,[2] S. Cabrera[x],[17] C. Calancha,[32] S. Camarda,[4] M. Campanelli,[31] M. Campbell,[35] F. Canelli[14],[18] A. Canepa,[46] B. Carls,[25] D. Carlsmith,[60] R. Carosi,[47] S. Carrillo[n],[19] S. Carron,[18] B. Casal,[12] M. Casarsa,[18] A. Castro[cc],[6] P. Catastini[ff],[47] D. Cauz,[55] V. Cavaliere[ff],[47] M. Cavalli-Sforza,[4] A. Cerri,[29] L. Cerrito[q],[31] S.H. Chang,[28] Y.C. Chen,[1] M. Chertok,[8] G. Chiarelli,[47] G. Chlachidze,[18] F. Chlebana,[18] K. Cho,[28] D. Chokheli,[16] J.P. Chou,[23] K. Chung[o],[18] W.H. Chung,[60] Y.S. Chung,[50] T. Chwalek,[27] C.I. Ciobanu,[45] M.A. Ciocci[ff],[47] A. Clark,[21] D. Clark,[7] G. Compostella,[44] M.E. Convery,[18] J. Conway,[8] M.Corbo,[45] M. Cordelli,[20] C.A. Cox,[8] D.J. Cox,[8] F. Crescioli[ee],[47] C. Cuenca Almenar,[61] J. Cuevas[v],[12] R. Culbertson,[18] J.C. Cully,[35] D. Dagenhart,[18] M. Datta,[18] T. Davies,[22] P. de Barbaro,[50] S. De Cecco,[52] A. Deisher,[29] G. De Lorenzo,[4] M. Dell'Orso[ee],[47] C. Deluca,[4] L. Demortier,[51] J. Deng[f],[17] M. Deninno,[6] M. d'Errico[dd],[44] A. Di Canto[ee],[47] G.P. di Giovanni,[45] B. Di Ruzza,[47] J.R. Dittmann,[5] M. D'Onofrio,[4] S. Donati[ee],[47] P. Dong,[18] T. Dorigo,[44] S. Dube,[53] K. Ebina,[58] A. Elagin,[54] R. Erbacher,[8] D. Errede,[25] S. Errede,[25] N. Ershaidat[bb],[45] R. Eusebi,[54] H.C. Fang,[29] S. Farrington,[43] W.T. Fedorko,[14] R.G. Feild,[61]


CMS Preliminary
Data
m_H=126 GeV
Zγ*,ZZ
Z+X
$\sqrt{s}$ = 7 TeV: L = 0.0 fb^{-1}
Events / 3 GeV
0
5
10
15
20
25
30
80
100
200
300
400
600
m_{4l} [G


Selected diphoton sample
Data 2011 and 2012
Sig + Bkg inclusive fit (m_H = 126.5 GeV)
4th order polynomial
$\sqrt{s}$ = 7 TeV, $\int$ Ldt = 4.8 fb^{-1}
$\sqrt{s}$ = 8 TeV, $\int$ Ldt = 5.9 fb^{-1}
ATLAS Preliminary
Events / GeV
Data - Bkg
$m_{\gamma\gamma}$ [GeV]

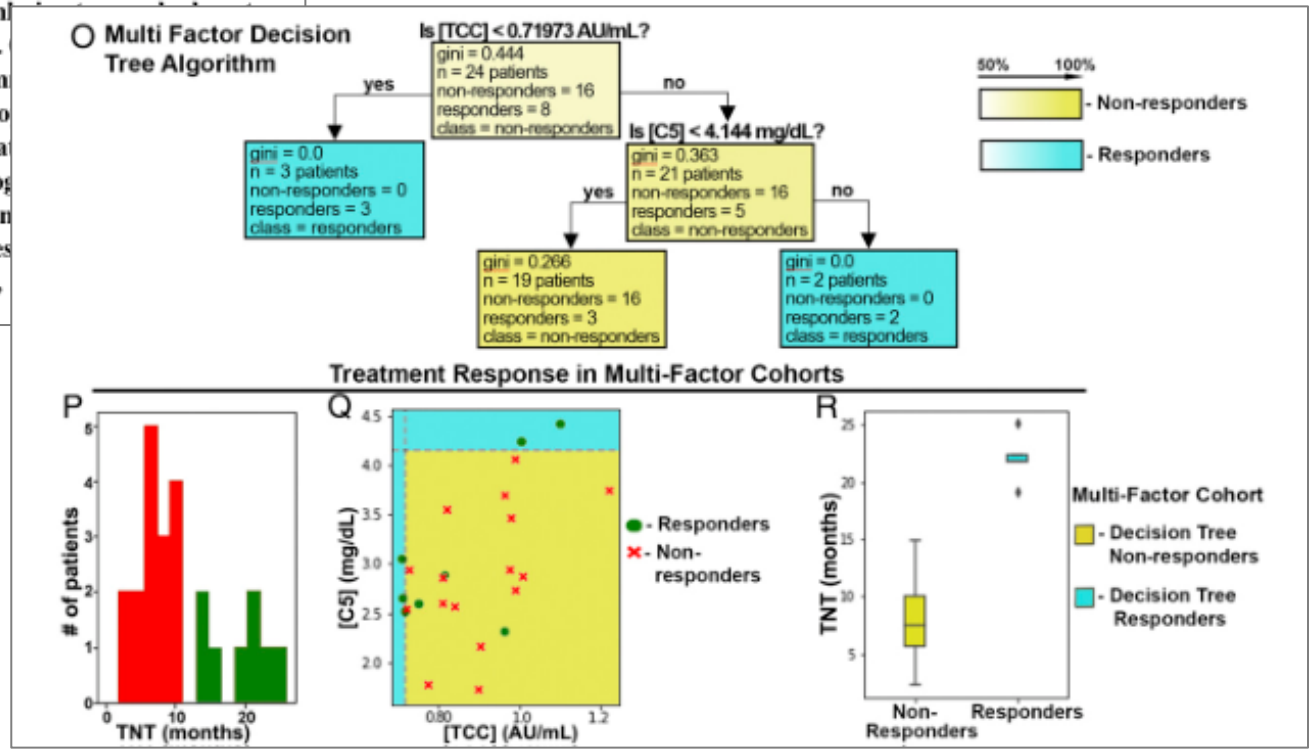
One More Shameless Example

https://www.jimmunol.org/content/early/2020/11/05/jimmunol.2000511

Complement as Prognostic Biomarker and Potential Therapeutic Target in Renal Cell Carcinoma

Britney Reese,*[,1] Ashok Silwal,*[,1] Elizabeth Daugherity,* Michael Daugherity,[†] Mahshid Arabi,* Pierce Daly,* Yvonne Paterson,[‡] Layton Woolford,[§,¶] Alana Christie,[§,¶] Roy Elias,[§,¶] James Brugarolas,[§,¶] Tao Wang,[¶,‖] Magdalena Karbowniczek,* and Maciej M. Markiewski*

Preclinical studies demonstrated that complement promotes tumor growth. Therefore, we sought to determine the best target for complement-based therapy among common human malignancies. High expression of 11 complement genes was linked to unfavorable prognosis in renal cell carcinoma. Complement protein expression or deposition was observed main... and tumor vasculature, corresponding to a role of complement in regulating the tumor microenvironment. ... in tumors correlated with a high nuclear grade. Complement genes clustered within an aggressive inflam... cancer characterized by poor prognosis, markers of T cell dysfunction, and alternatively activated macro... complement proteins correlated with response to immune checkpoint inhibitors. Corroborating human da... cies and blockade reduced tumor growth by enhancing antitumor immunity and seemingly reducing angio... of kidney cancer resistant to PD-1 blockade. Overall, this study implicates complement in the immune lan... cinoma, and notwithstanding cohort size and preclinical model limitations, the data suggest that tumors res... point inhibitors might be suitable targets for complement-based therapy. *The Journal of Immunology,*



Part 1

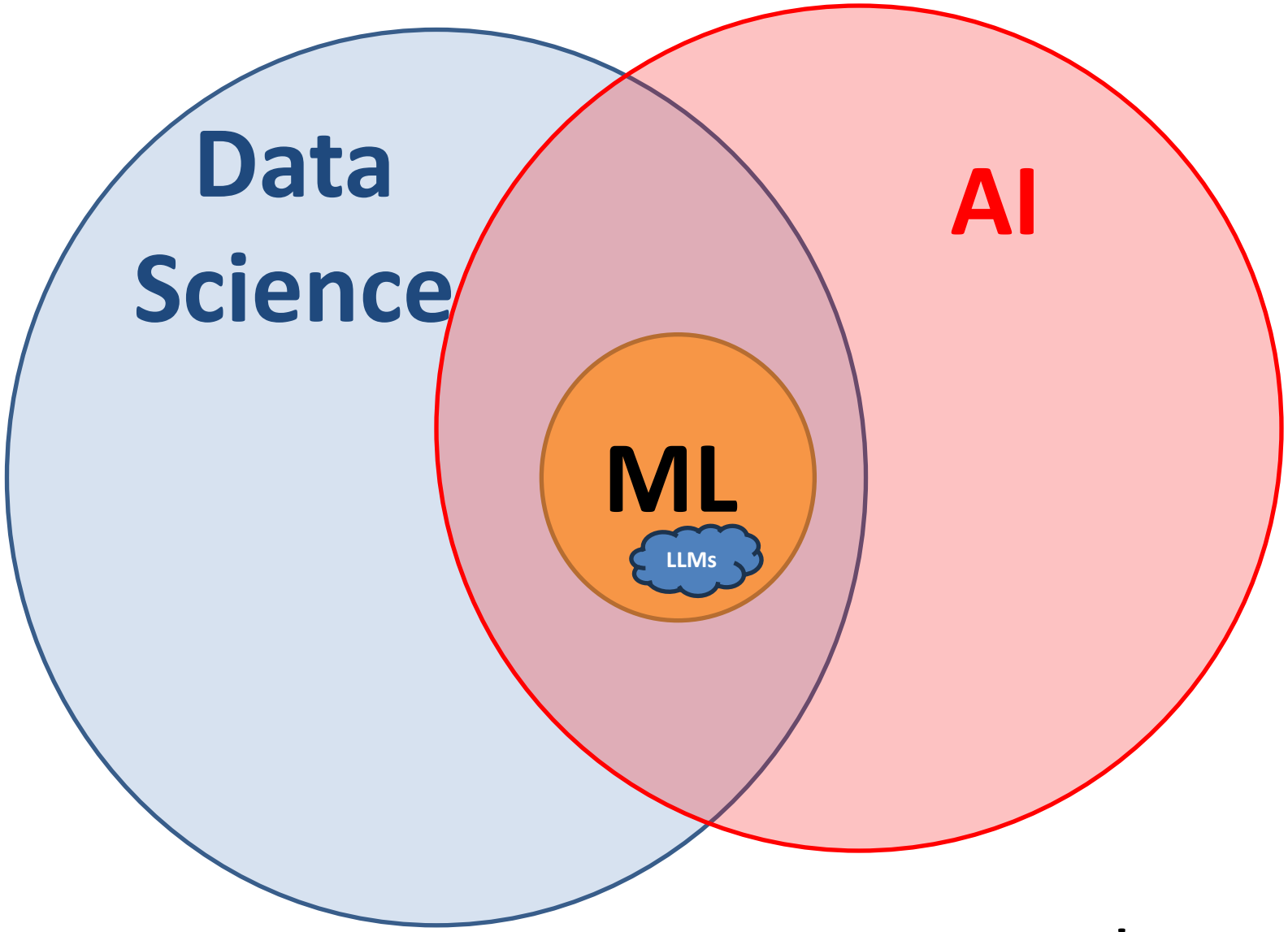
WHAT IS MACHINE LEARNING: A TOO SHORT INTRO

Data Science

1) Having data is like sitting on top of a gold mine

2) Everyone has data…

Data Science is learning how to dig through the mountain and get the gold out



* not to scale

We will see some of the math behind LLMs and generative AI models

This class is about how they **work**, not how to use them

What is “learning”?

The goal:

Make a decision based on data

i.e. find “patterns” in the data. (Besides, if you aren’t making your decisions based on data, then what are you basing them on?)

To get a good answer you MUST:

- ask a good question
- have good data

Just the Basics

CONCEPTS

High-Level Overview

Input: represent one sample as numbers (called **features**) in an efficient way that is hopefully useful for solving the problem

Output: our answer depends on our question, common examples:

- Classification: what discrete category does the sample belong to?
- Regression: predict a real number
- Reinforcement: optimize a certain outcome

Model: how we calculate the output from the input. Usually has lots of adjustable parameters that need to be set somehow

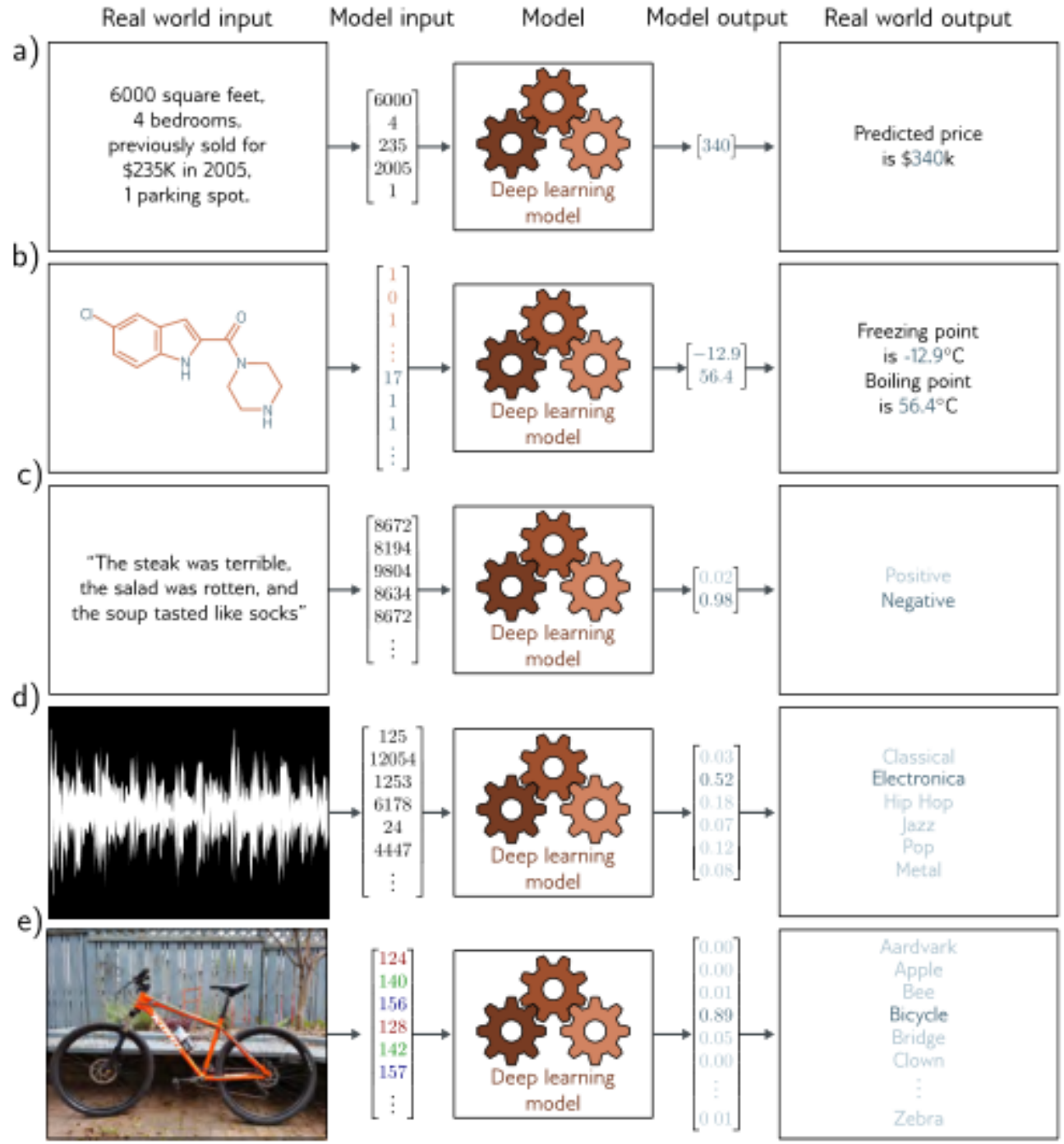
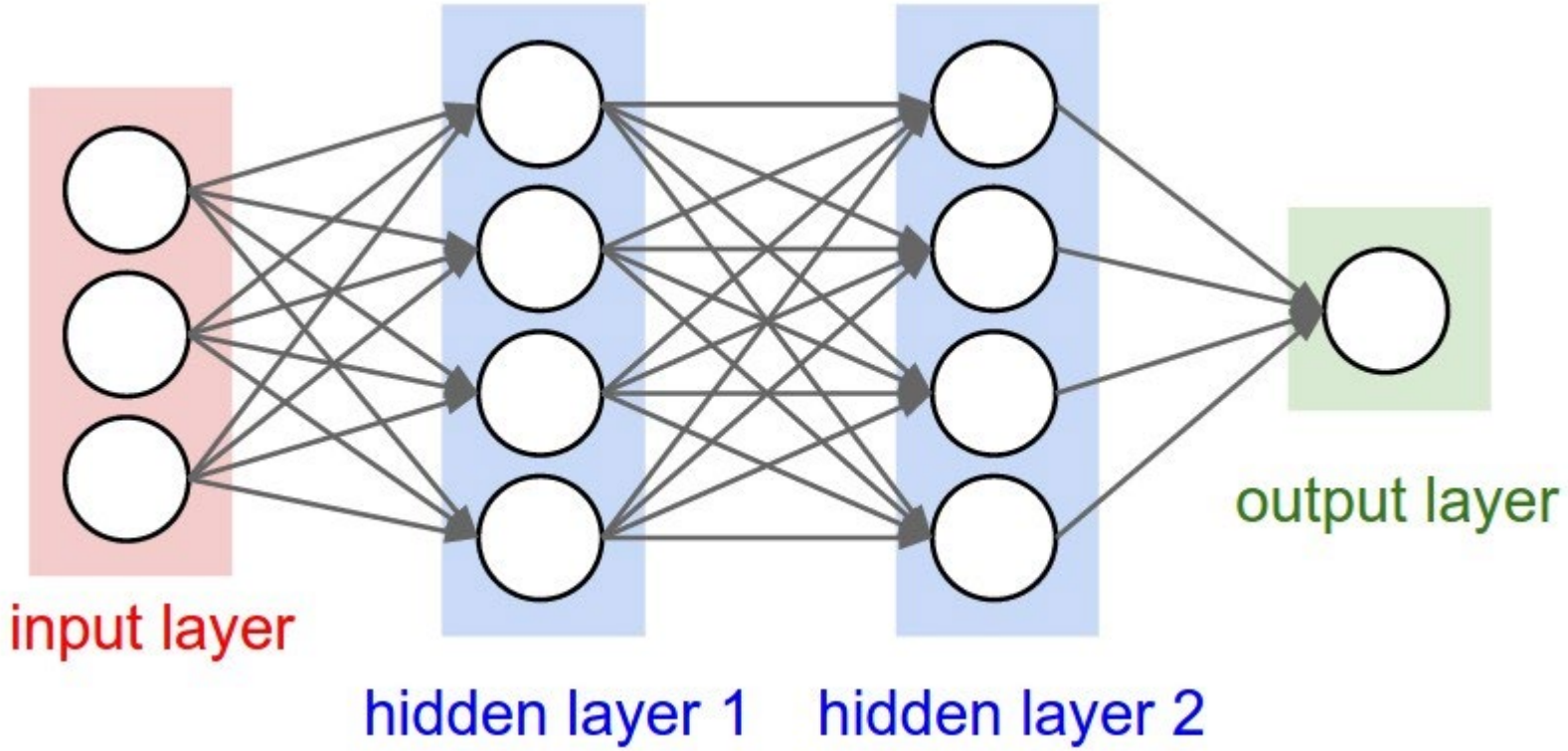


Figure 1.2 Regression and classification problems. a) This *regression* model takes a vector of numbers that characterize a property and predicts its price. b) This *multivariate regression* model takes the structure of a chemical molecule and predicts its melting and boiling points. c) This *binary classification* model takes a restaurant review and classifies it as either positive or negative. d) This *multiclass classification* problem assigns a snippet of audio to one of N genres. e) A second multiclass classification problem in which the model classifies an image according to which of N possible objects it might contain.

Models

- Neural networks are the ones everyone has heard of, but there are lots and lots and lots of other choices
- Neural networks are extremely well-suited to some problems like 2D images, can be complicated and finicky to work with so not always the best choice
- You don’t know for certain which model will work best on a given problem ahead of time, so we have to try several
- Most of this class is devoted to learning about different models

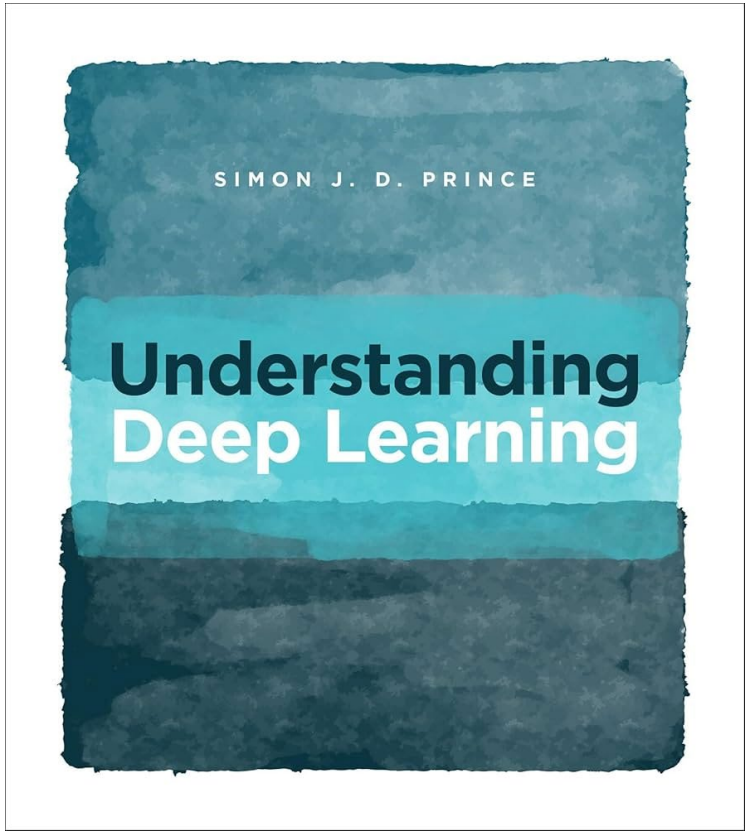
Neural Network



- Each **input** is a SINGLE NUMBER
- **Hidden** and **output** values are directly calculated from inputs and **weights**
- **Weights (arrows)** are adjusted in training to make output match known right answer

Deep Learning

- Deep Learning = Neural networks with lots of layers
- They work extremely well for many hard problems and we honestly aren't really sure why



The title is also partly a joke — *no-one* really understands deep learning at the time of writing. Modern deep networks learn piecewise linear functions with more regions than there are atoms in the universe and can be trained with fewer data examples than model parameters. It is neither obvious that we should be able to fit these functions reliably nor that they should generalize well to new data. The penultimate chapter addresses

Seriously Though

https://scikit-learn.org/stable/user_guide.html

1. Supervised learning

- 1.1. Linear Models
- 1.2. Linear and Quadratic Discriminant Analysis
- 1.3. Kernel ridge regression
- 1.4. Support Vector Machines
- 1.5. Stochastic Gradient Descent
- 1.6. Nearest Neighbors
- 1.7. Gaussian Processes
- 1.8. Cross decomposition
- 1.9. Naive Bayes
- 1.10. Decision Trees
- 1.11. Ensemble methods
- 1.12. Multiclass and multioutput algorithms
- 1.13. Feature selection
- 1.14. Semi-supervised learning
- 1.15. Isotonic regression
- 1.16. Probability calibration
- 1.17. Neural network models (supervised)

2. Unsupervised learning

- 2.1. Gaussian mixture models
- 2.2. Manifold learning
- 2.3. Clustering
- 2.4. Biclustering
- 2.5. Decomposing signals in components (matrix factorization problems)
- 2.6. Covariance estimation
- 2.7. Novelty and Outlier Detection
- 2.8. Density Estimation
- 2.9. Neural network models (unsupervised)

Part 2

TOUR OF ML TOPICS

Classification

- **Classification** – decide which **category** a sample belongs to
- Examples:
  - Email or spam
  - Speech recognition
  - OCR (Optical Character Recognition)
- *i.e. write a program that can recognize patterns*

Spam Filter

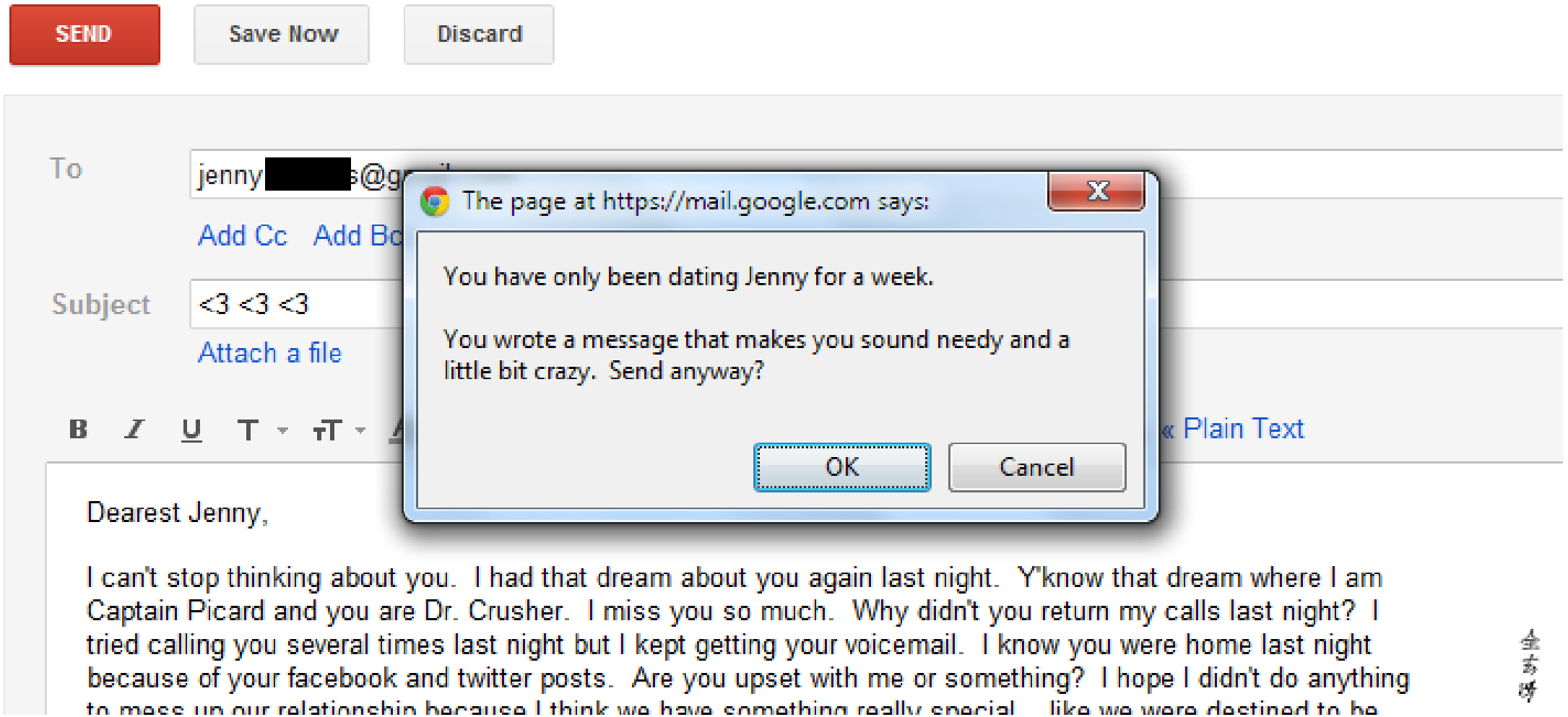
Assassinating spam e-mail

SpamAssassin is a widely used open-source spam filter. It calculates a score for an incoming e-mail, based on a number of built-in rules or 'tests' in SpamAssassin's terminology, and adds a 'junk' flag and a summary report to the e-mail's headers if the score is 5 or more.

```
-0.1 RCVD_IN_MXRATE_WL       RBL: MXRate recommends allowing
                             [123.45.6.789 listed in sub.mxrate.net]
 0.6 HTML_IMAGE_RATIO_02     BODY: HTML has a low ratio of text to image area
 1.2 TVD_FW_GRAPHIC_NAME_MID BODY: TVD_FW_GRAPHIC_NAME_MID
 0.0 HTML_MESSAGE            BODY: HTML included in message
 0.6 HTML_FONx_FACE_BAD      BODY: HTML font face is not a word
 1.4 SARE_GIF_ATTACH         FULL: Email has a inline gif
 0.1 BOUNCE_MESSAGE          MTA bounce message
 0.1 ANY_BOUNCE_MESSAGE      Message is some kind of bounce message
 1.4 AWL                     AWL: From: address is in the auto white-list
```

From left to right you see the score attached to a particular test, the test identifier, and a short description including a reference to the relevant part of the e-mail. As you see, scores for individual tests can be negative (indicating evidence suggesting the e-mail is ham rather than spam) as well as positive. The overall score of 5.3 suggests the e-mail might be spam.

Spam Filter


SEND
Save Now
Discard
To
jenny
Add Cc
Add Bc
Subject
<3 <3 <3
Attach a file
« Plain Text
Dearest Jenny,
I can't stop thinking about you. I had that dream about you again last night. Y'know that dream where I am Captain Picard and you are Dr. Crusher. I miss you so much. Why didn't you return my calls last night? I tried calling you several times last night but I kept getting your voicemail. I know you were home last night because of your facebook and twitter posts. Are you upset with me or something? I hope I didn't do anything
The page at https://mail.google.com says:
You have only been dating Jenny for a week.
You wrote a message that makes you sound needy and a little bit crazy. Send anyway?
OK
Cancel

Training

Supervised Learning:

- we are given training data where we know the right answers
- pattern recognition finds hidden trends
- training data can be a huge bottleneck
- examples: ChatGPT, facial recognition, OCR

Unsupervised Learning:

- we don’t know the right answer, so optimize for some metric
- can I find natural clusters or groups in the data?
- stable diffusion: adjust these pixels to make the picture look like a duck
- social media/YouTube “algorithm”

Supervised learning example:

Interviewer: What's your biggest strength?

Me: I'm an expert in machine learning.

Interviewer: What's 9 + 10?

Me: Its 3.

Interviewer: Not even close. It's 19.

Me: It's 16.

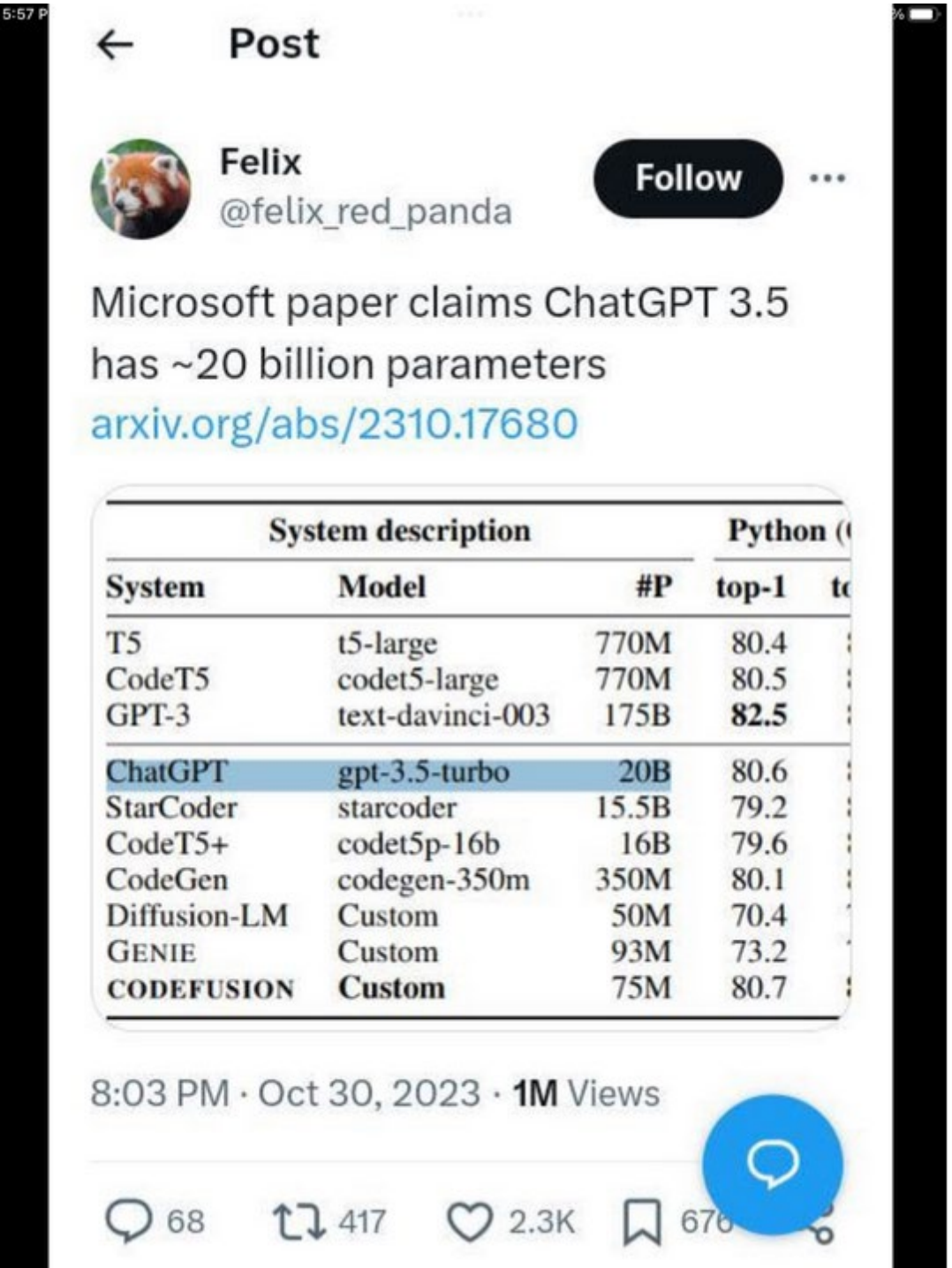
Interviewer: Wrong. Its still 19.

Me: It's 18.

Interviewer: No, it's 19.

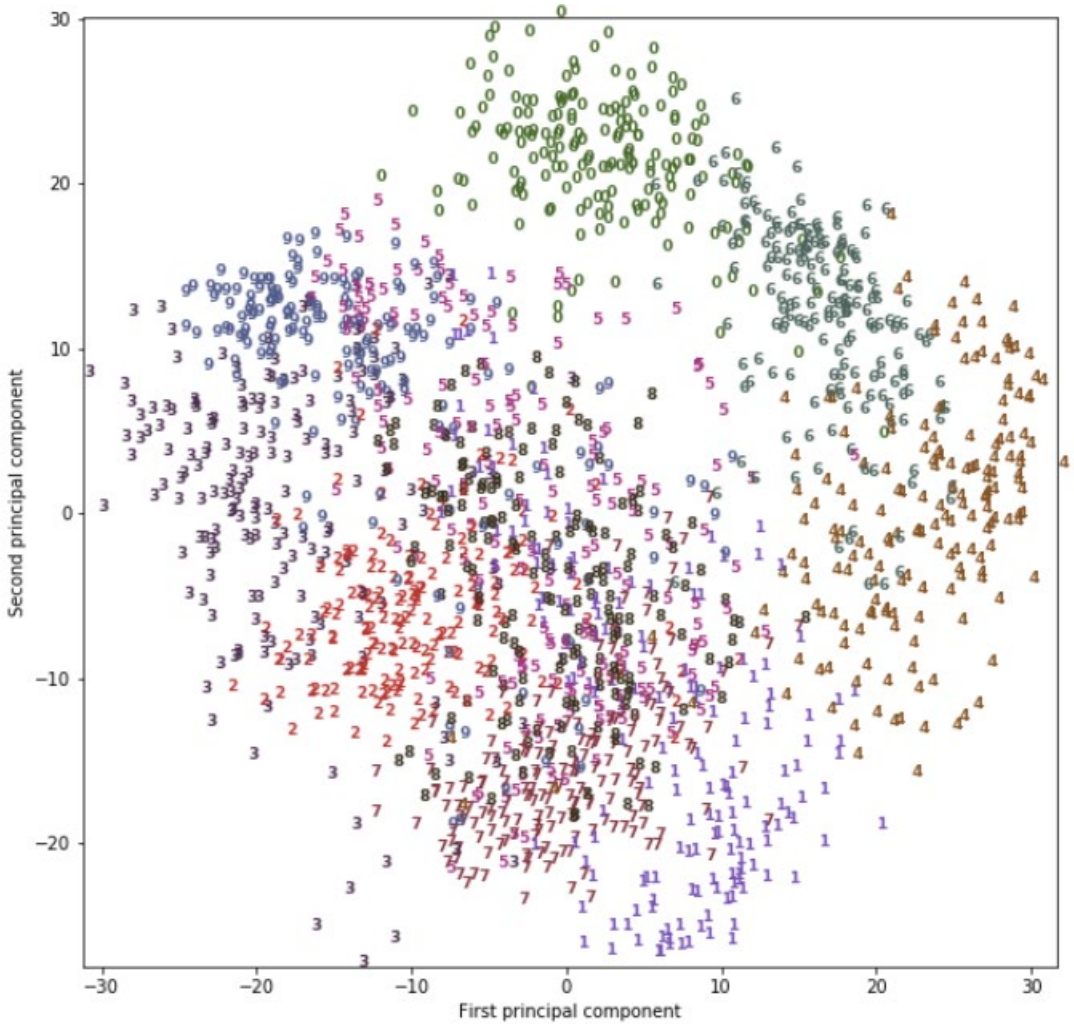
Me: it's 19.

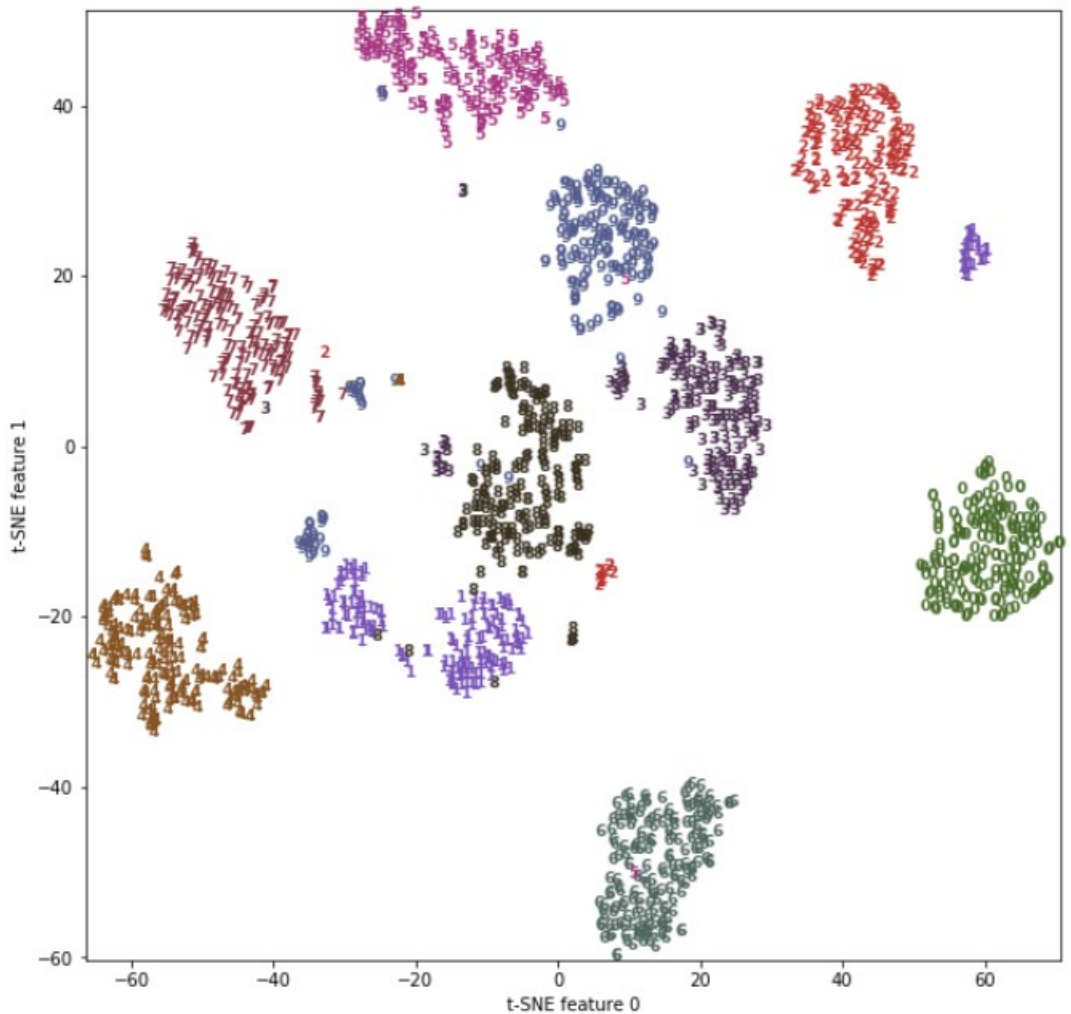
Interviewer: You're hired



Training data is the “fossil fuel” of LLMs. We have essentially scraped the entire internet for data of human-written text and we are reaching the upper limit of parameters we can train.

https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb





Other Tricks

REGRESSION

Regression

Instead of classifying into categories, our job is now to predict a real-valued number based on training data

Examples:

- predict the price of a house in Boston
- predict how much money a movie will make
- guess the value of a stock
- Diabetes dataset: predict measure of disease based on patient data

Other Tricks

REINFORCEMENT

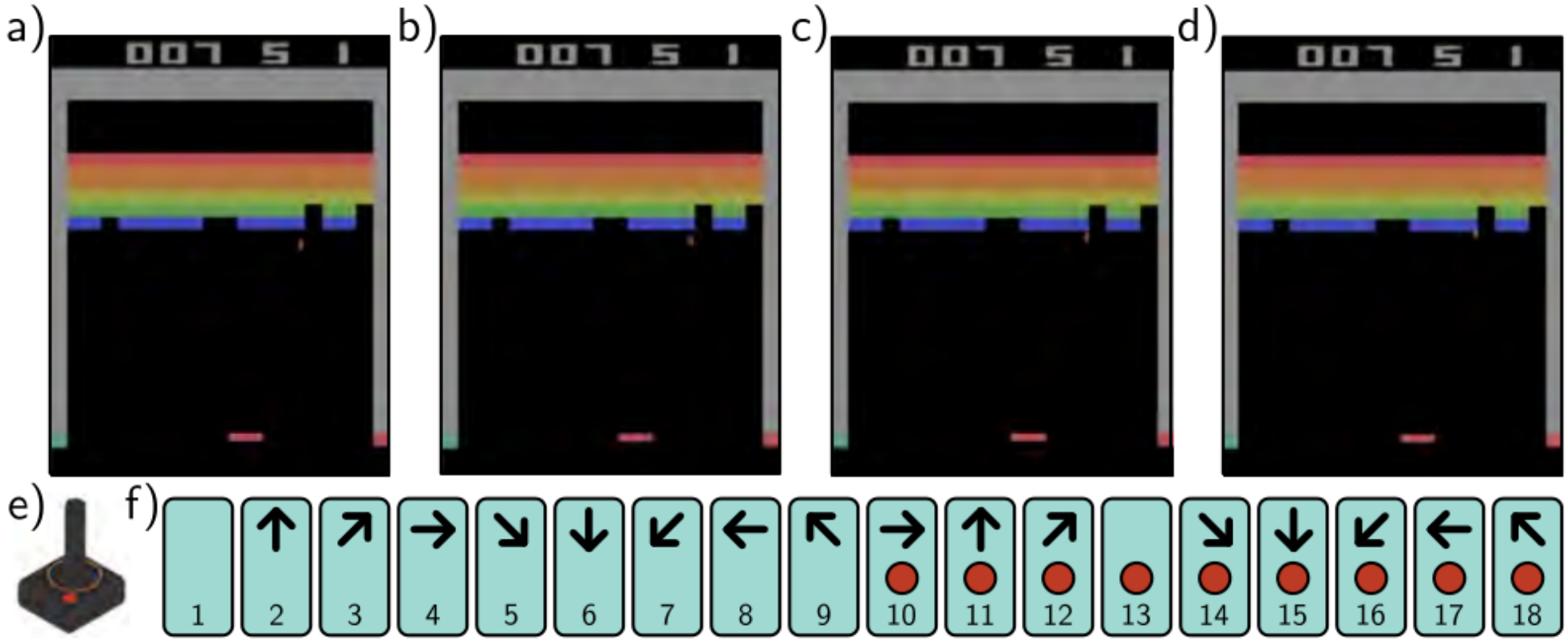
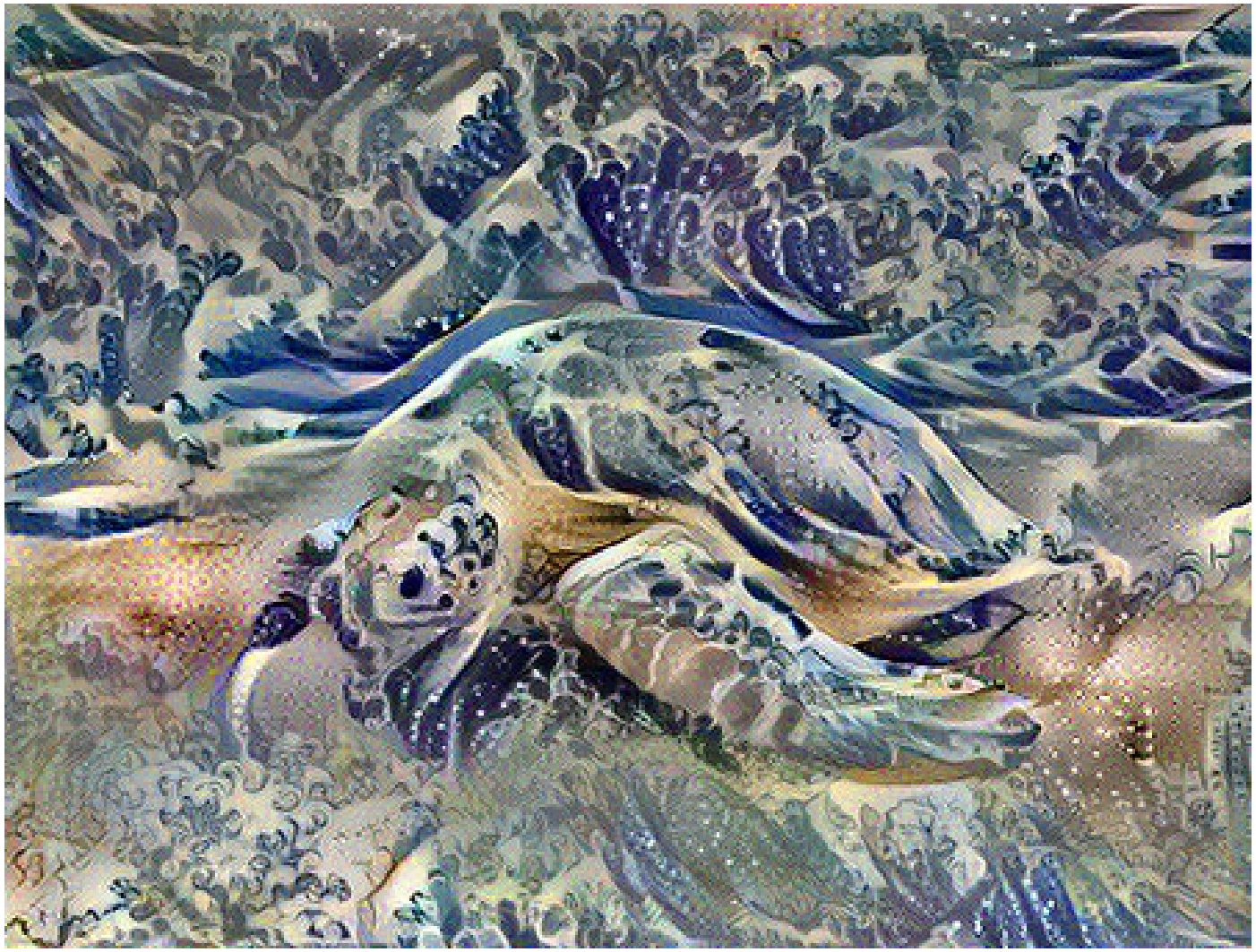
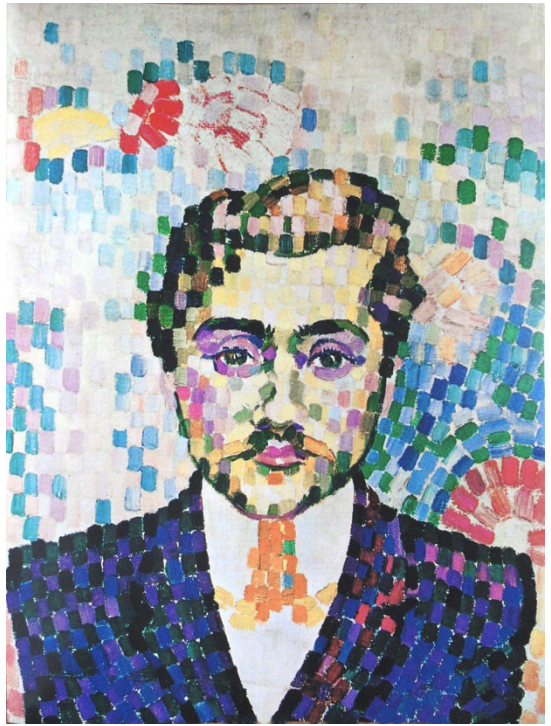


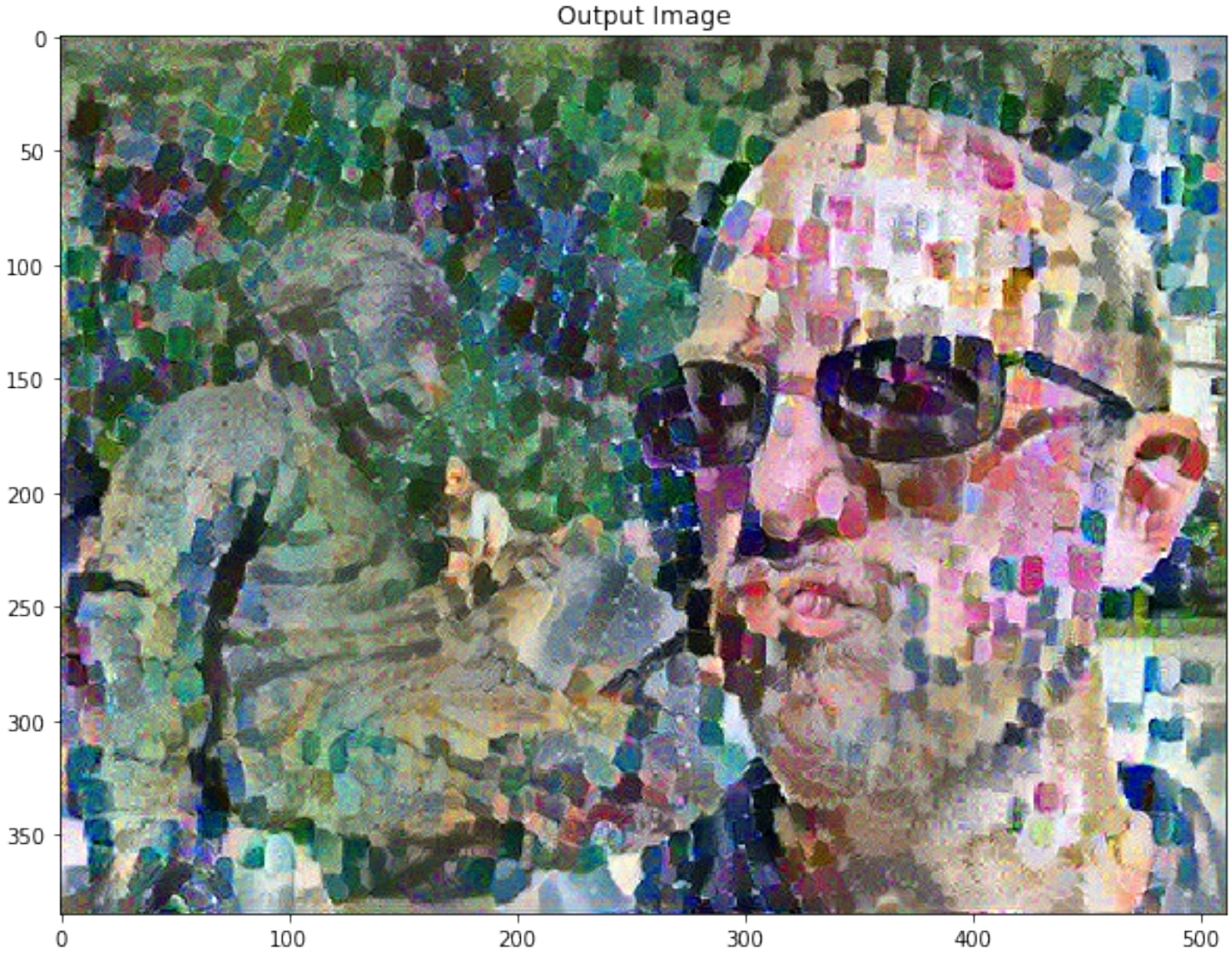
Figure 19.13 Atari Benchmark. The Atari benchmark consists of 49 Atari 2600 games, including Breakout (pictured), Pong, and various shoot-em-up, platform, and other types of games. a-d) Even for games with a single screen, the state is not fully observable from a single frame because the velocity of the objects is unknown. Consequently, it is usual to use several adjacent frames (here, four) to represent the state. e) The action simulates the user input via a joystick. f) There are eighteen actions corresponding to eight directions of movement or no movement, and for each of these nine cases, the button being pressed or not.

Other Tricks

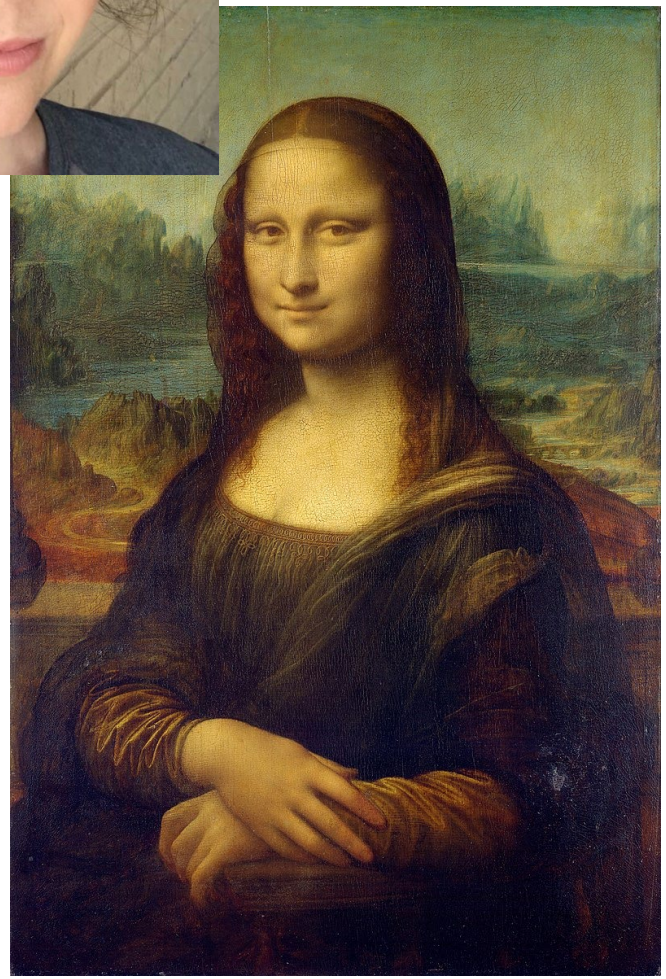
GENERATION

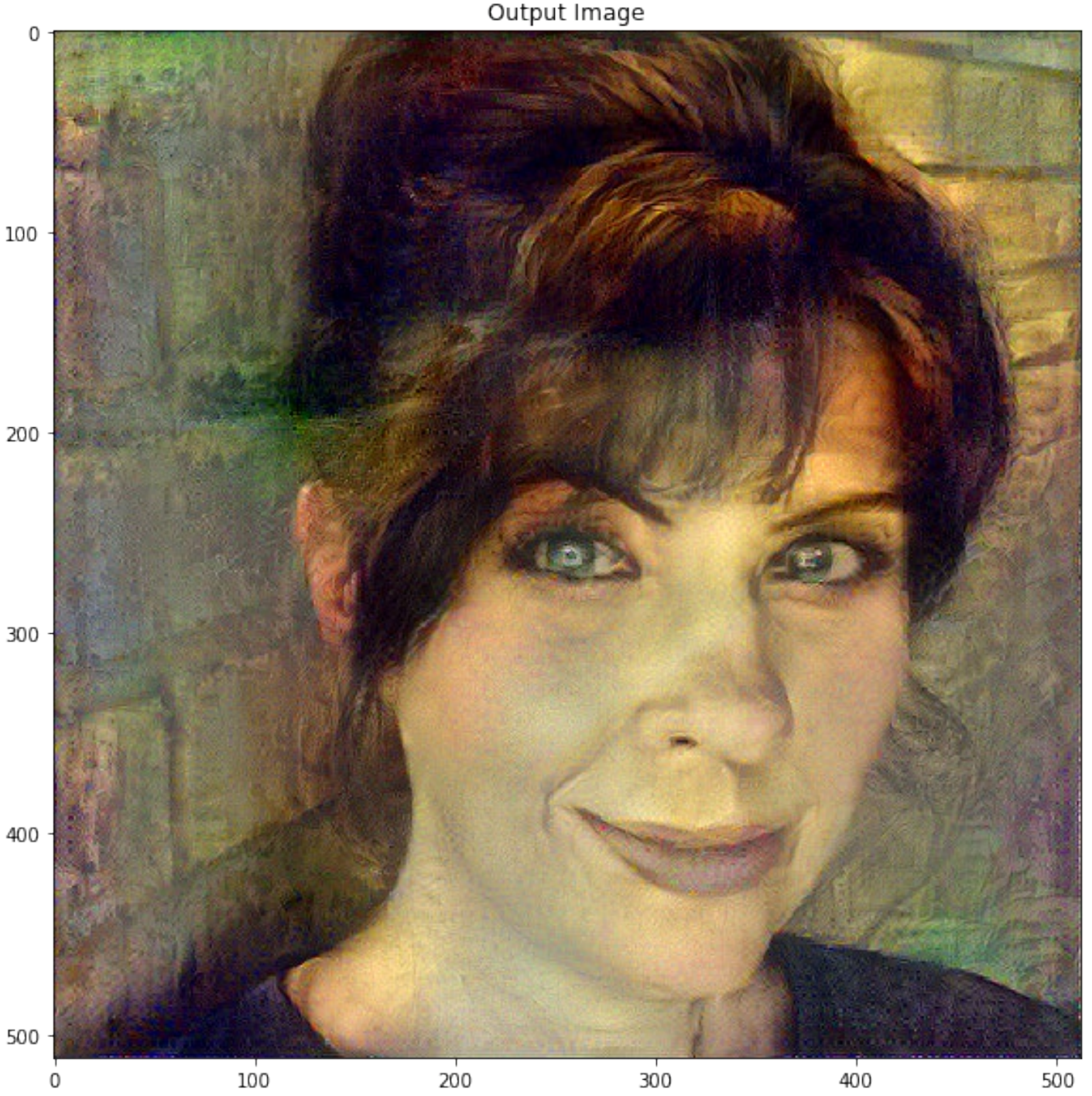





Output Image
0
50
100
150
200
250
300
350
0
100
200
300
400
500


Output Image
0
100
200
300
400
500
0
100
200
300
400
500



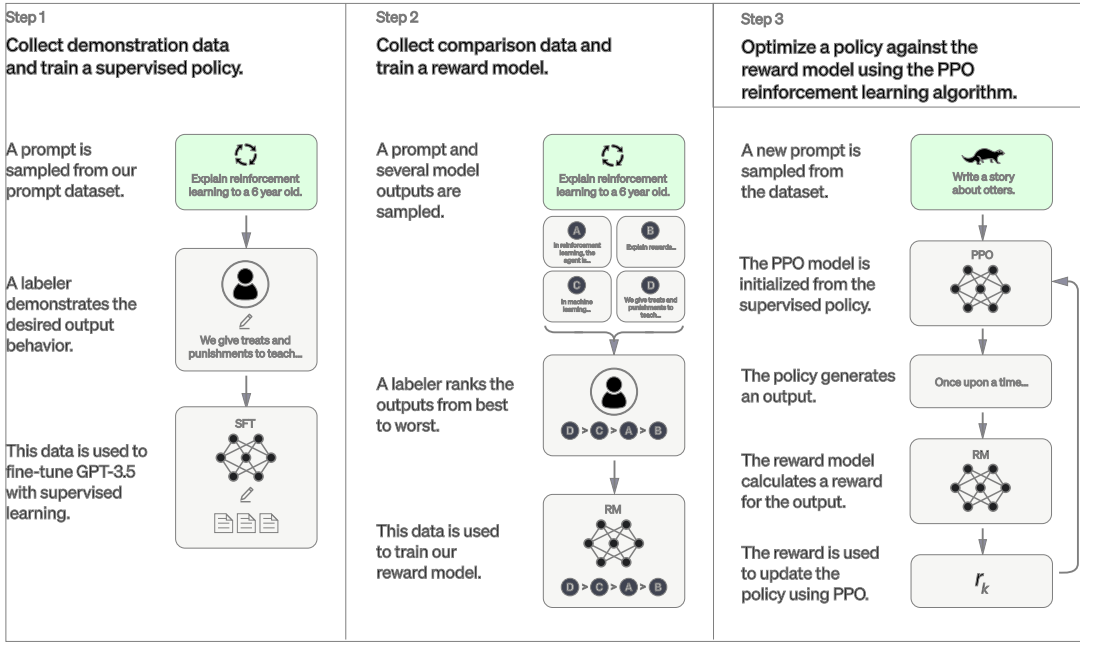

Output Image
0
100
200
300
400
500
0
100
200
300
400
500


Output Image
0
50
100
150
200
250
300
350
0
100
200
300
400
500

More Examples

- DALL-E: generating images based on input text description
  https://labs.openai.com/

- ChatGPT: generating text based on input text description
  https://openai.com/blog/chatgpt/

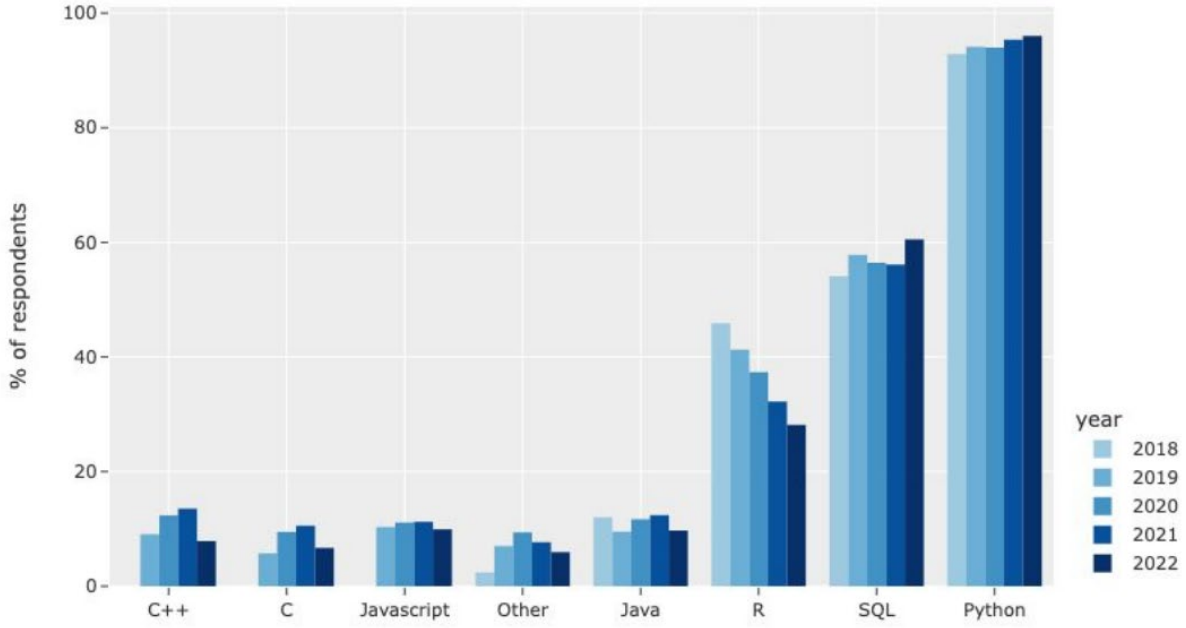


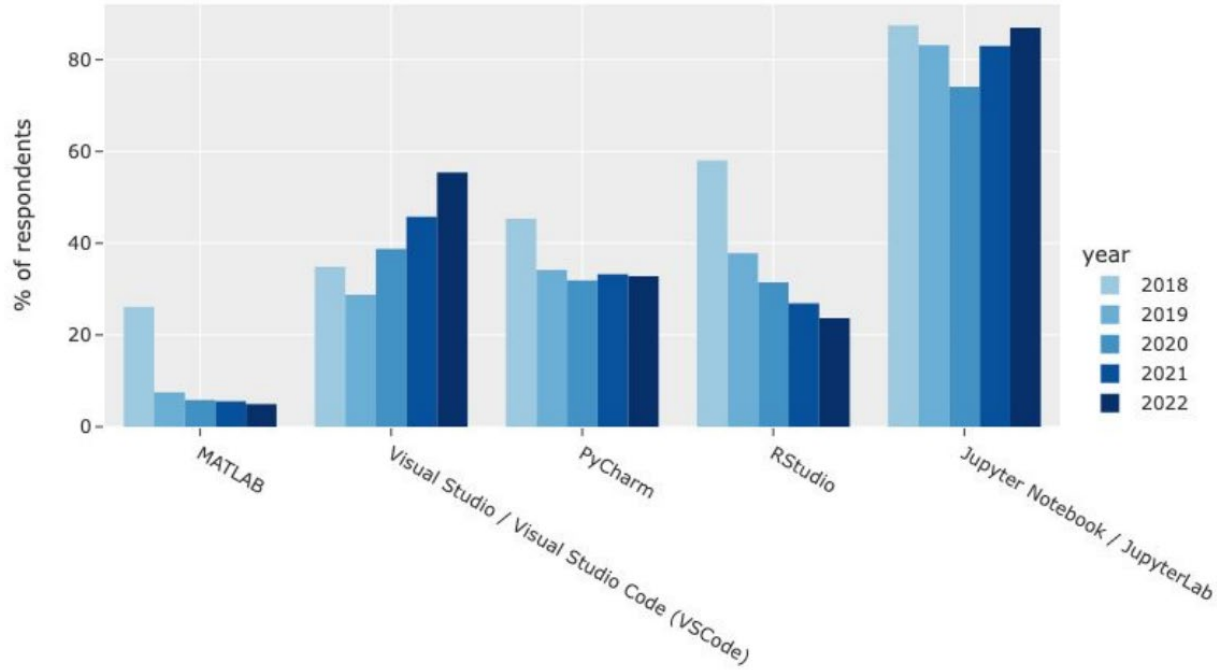
Part 3

TOOLS

Kaggle: Making Data Science a Sport

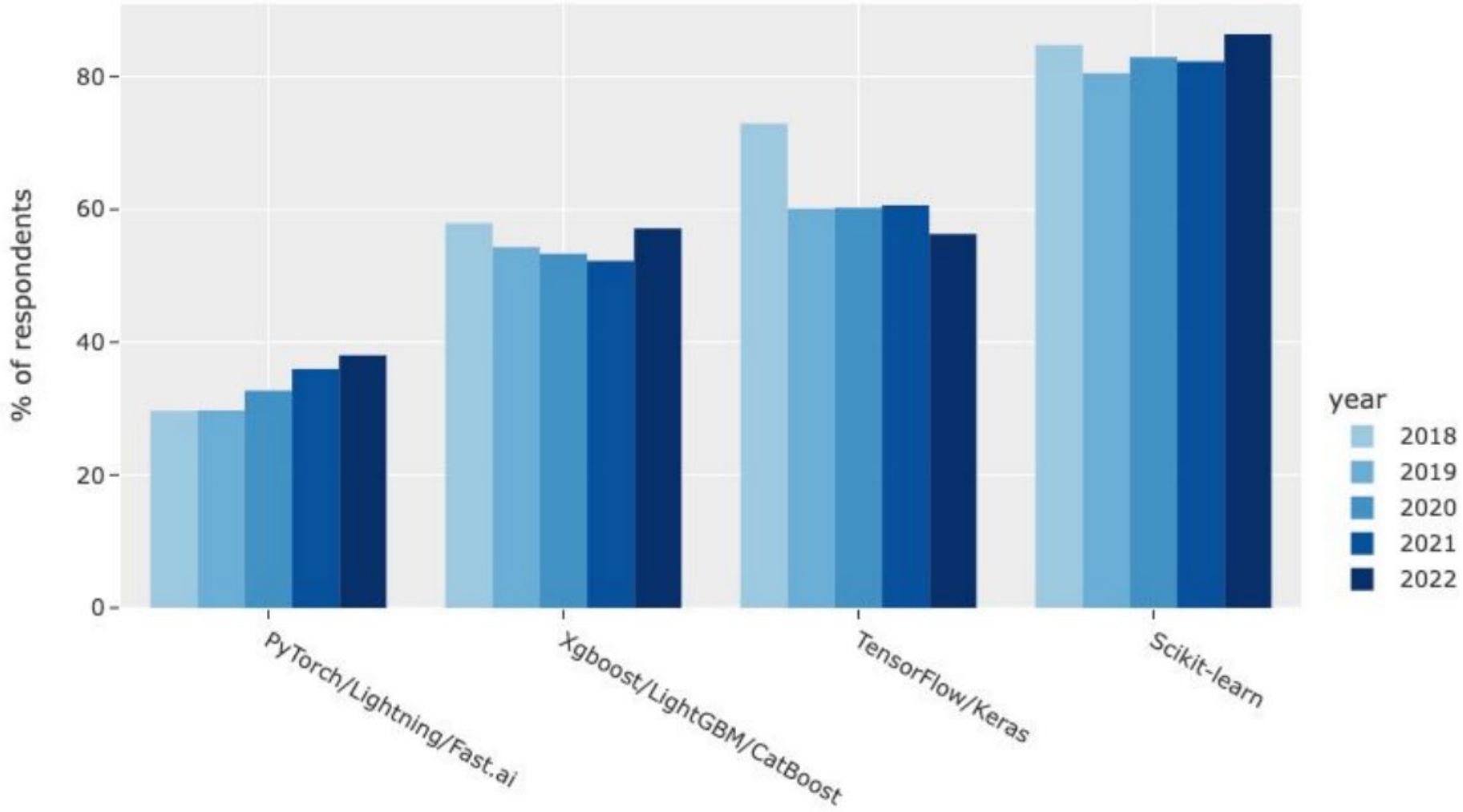
https://www.kaggle.com/kaggle-survey-2022





Kaggle: Making Data Science a Sport

https://www.kaggle.com/kaggle-survey-2022

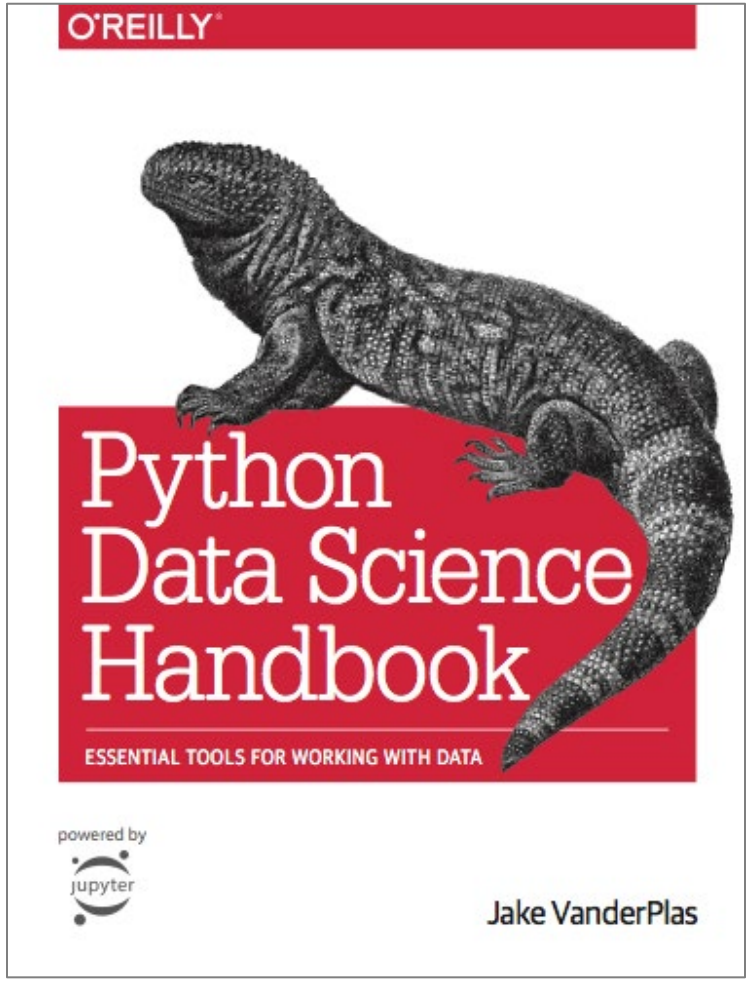


Getting Started

Great python resources:

A Whirlwind Tour of Python by Jake VanderPlas.
Freely available online at:
https://jakevdp.github.io/WhirlwindTourOfPython/

Python Data Science Handbook by Jake VanderPlas.
Freely available online at:
https://jakevdp.github.io/PythonDataScienceHandbook/



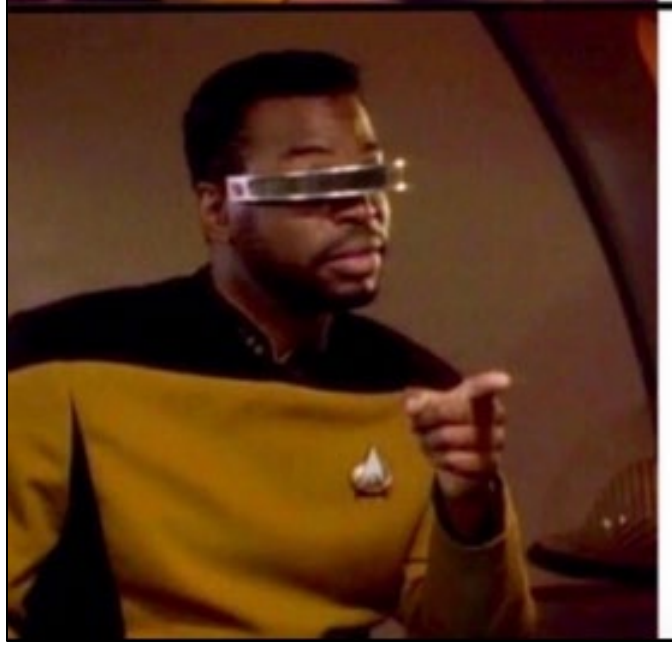
Part N-1

POLICIES

Doctors: Googling stuff online does not make you a doctor.

Programmers:

Not sharing your source code out of greed



Not sharing your source code out of shame