



Mathematics in Data Science

Mathematical Association of America
Distinguished Lecture Series

Nathan Carter

Wilder Teaching Professor
Mathematical Sciences Department
Bentley University
Waltham, MA

March 23, 2022

What is data science?



Quick Definition of Data Science

If you're using **mathematics or statistics**
powered by **modern computing tools**
to answer a **real world question**,
you are probably doing data science.



Example Data Science Questions

Data science = math/stats + computing + application

- ▶ **Social justice:** Do the U.S. government's records of mortgage applications reveal any patterns of discrimination?
- ▶ **Sports:** What baseball players offer the best value in my fantasy league draft?
- ▶ **Business:** How can shipping companies ensure every container on a ship gets filled?
- ▶ **Public health:** What percentage of the people in my state will eventually be vaccinated?



Mathematics in Data Science

Working with Data

- ▶
- ▶
- ▶
- ▶
- ▶
- ▶
- ▶
- ▶

Doing an Analysis

- ▶
- ▶
- ▶
- ▶
- ▶

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



Part 1

Working with Data



Example data: U.S. COVID vaccinations, 2021

Date	Location	Total vacc's	Total Distrib'd	People vacc'd	People vacc'd per hundred	...
2021-01-12	Alabama	78134.0	377025.0	70861.0	0.15	...
2021-01-13	Alabama	84040.0	378975.0	74792.0	0.19	...
2021-01-14	Alabama	92300.0	435350.0	80480.0	NaN	...
2021-01-15	Alabama	100567.0	444650.0	86956.0	0.28	...
2021-01-16	Alabama	NaN	NaN	NaN	NaN	...
2021-01-17	Alabama	NaN	NaN	NaN	NaN	...
2021-01-18	Alabama	NaN	NaN	NaN	NaN	...
2021-01-19	Alabama	130795.0	444650.0	114319.0	0.33	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-12-12	Wyoming	655151.0	840845.0	316311.0	46.42	...
2021-12-13	Wyoming	655179.0	840845.0	316355.0	46.42	...
2021-12-14	Wyoming	662705.0	845255.0	317645.0	46.72	...

Source: <https://ourworldindata.org/us-states-vaccinations>



Task #1: Lookup

How many people per hundred were vaccinated in Alabama as of January 19, 2021?

Date	Location	Total vacc's	Total Distrib'd	People vacc'd	People vacc'd per hundred	...
2021-01-12	Alabama	78134.0	377025.0	70861.0	0.15	...
2021-01-13	Alabama	84040.0	378975.0	74792.0	0.19	...
2021-01-14	Alabama	92300.0	435350.0	80480.0	NaN	...
2021-01-15	Alabama	100567.0	444650.0	86956.0	0.28	...
2021-01-16	Alabama	Nan	Nan	Nan	Nan	...
2021-01-17	Alabama	Nan	Nan	Nan	Nan	...
2021-01-18	Alabama	Nan	Nan	Nan	Nan	...
2021-01-19	Alabama	130795.0	444650.0	114319.0	0.33	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-12-12	Wyoming	655151.0	840845.0	316311.0	46.42	...
2021-12-13	Wyoming	655179.0	840845.0	316355.0	46.42	...
2021-12-14	Wyoming	662705.0	845255.0	317645.0	46.72	...



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶
- ▶
- ▶
- ▶
- ▶
- ▶

Doing an Analysis

- ▶
- ▶
- ▶
- ▶

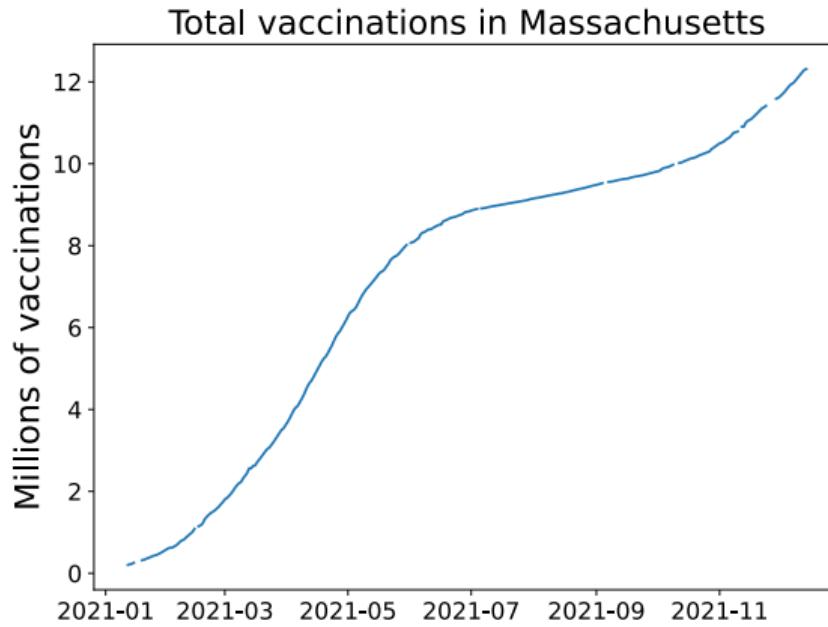
Unsolved Problems

- ▶
- ▶
- ▶
- ▶



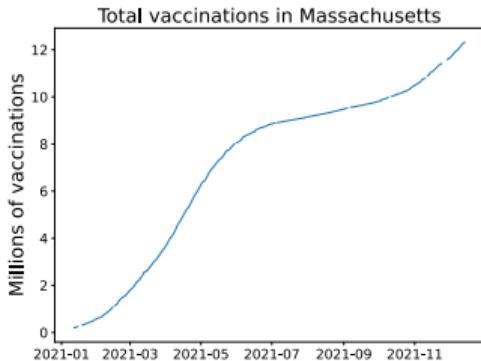
Task #2: Visualization

How has the vaccine distribution gone in my state?



Task #2: Visualization

How has the vaccine distribution gone in my state?



Data	Visualization
function on real numbers	line/curve plot
function from pairs to real numbers	surface plot/map
function with small (finite) domain	bar chart
subset of the plane	scatter plot
subset of the real line	histogram
matrix	heat map/graph



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶
- ▶
- ▶

Doing an Analysis

- ▶
- ▶
- ▶
- ▶

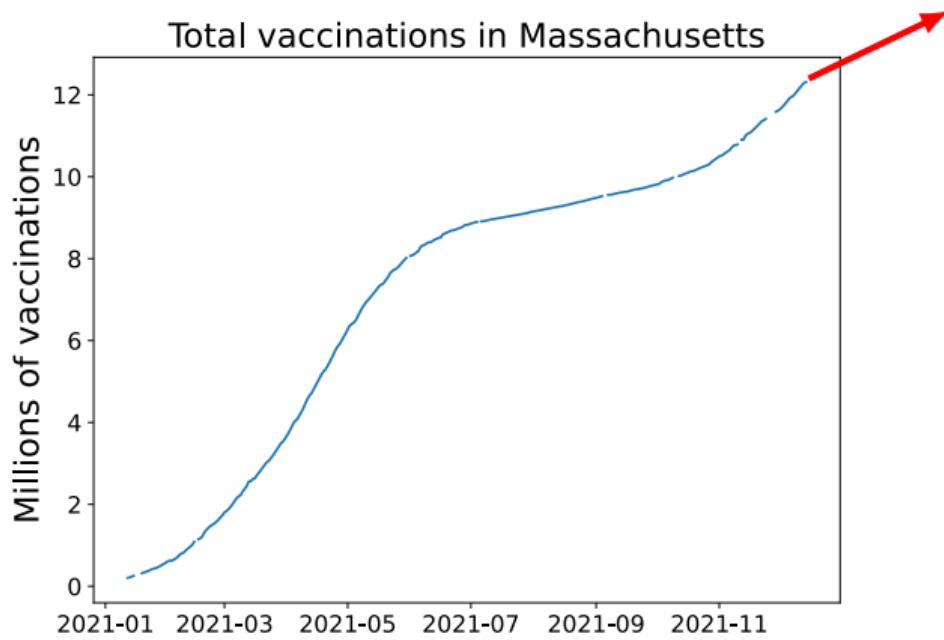
Unsolved Problems

- ▶
- ▶
- ▶
- ▶

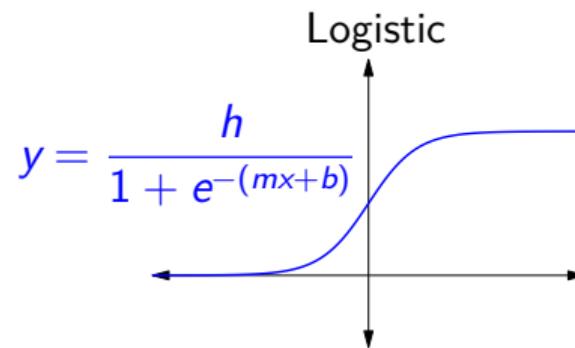
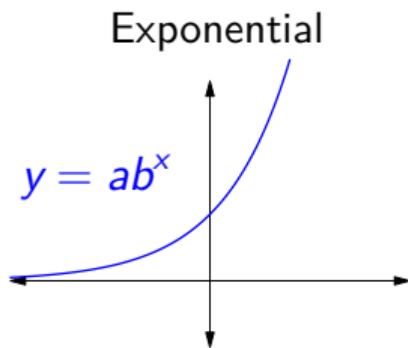
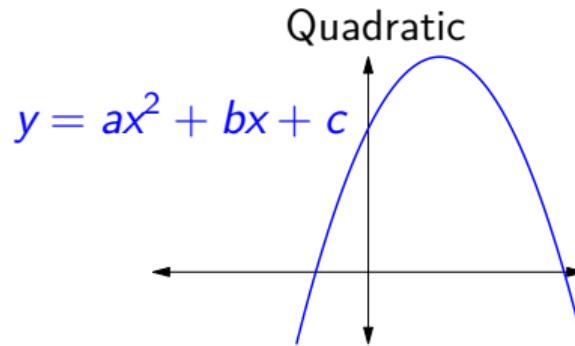
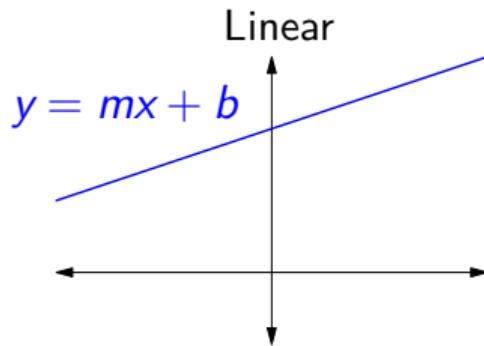


Task #3: Prediction

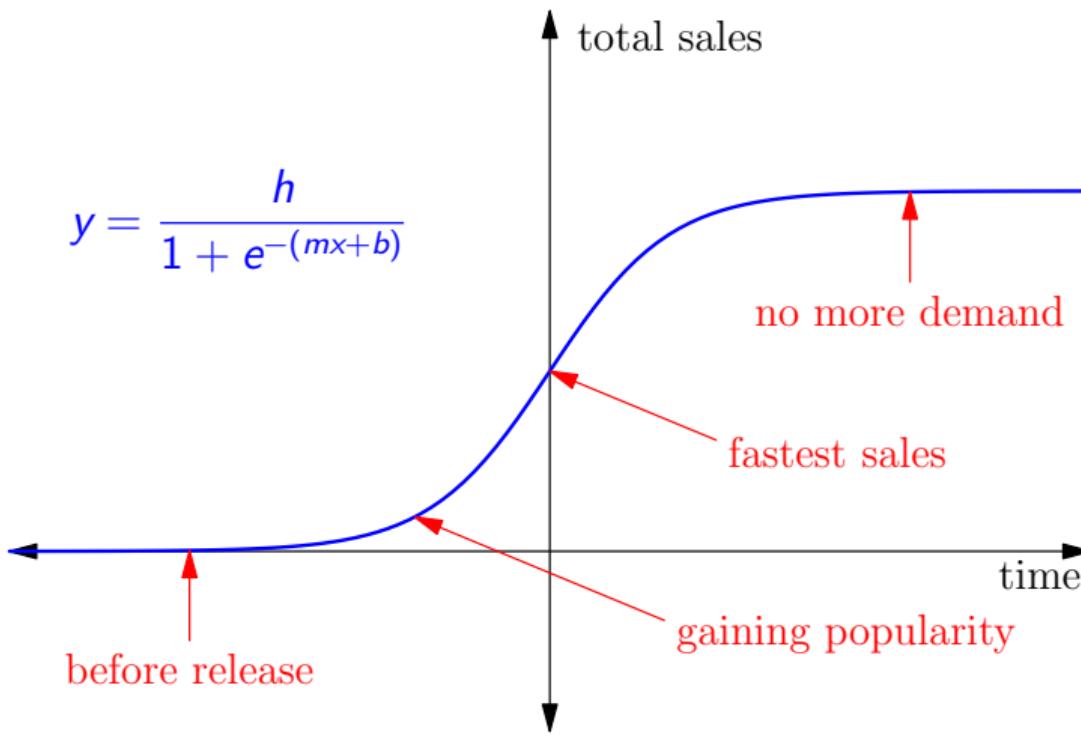
How many vaccinations will eventually be given in my state?



Famous function families



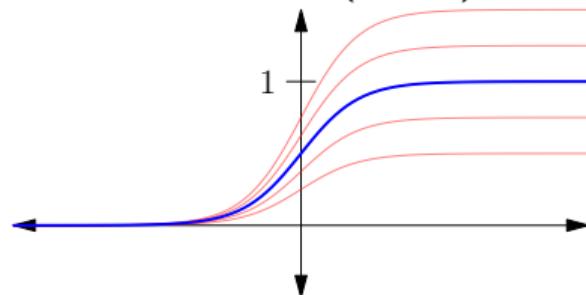
The logistic family: Rolling out a new product



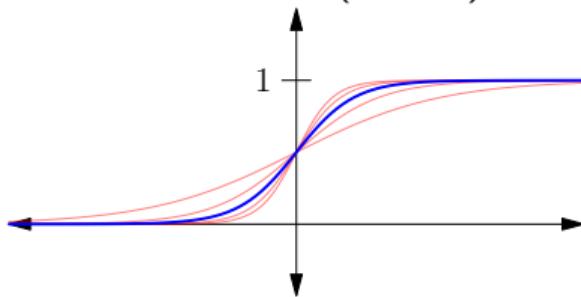
The logistic family: Three parameters

$$y = \frac{h}{1 + e^{-(mx+b)}}$$

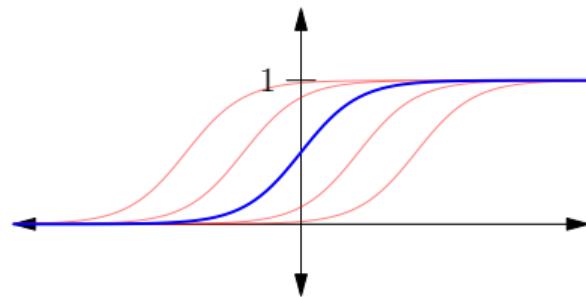
As h varies ($h > 0$)



As m varies ($m > 0$)

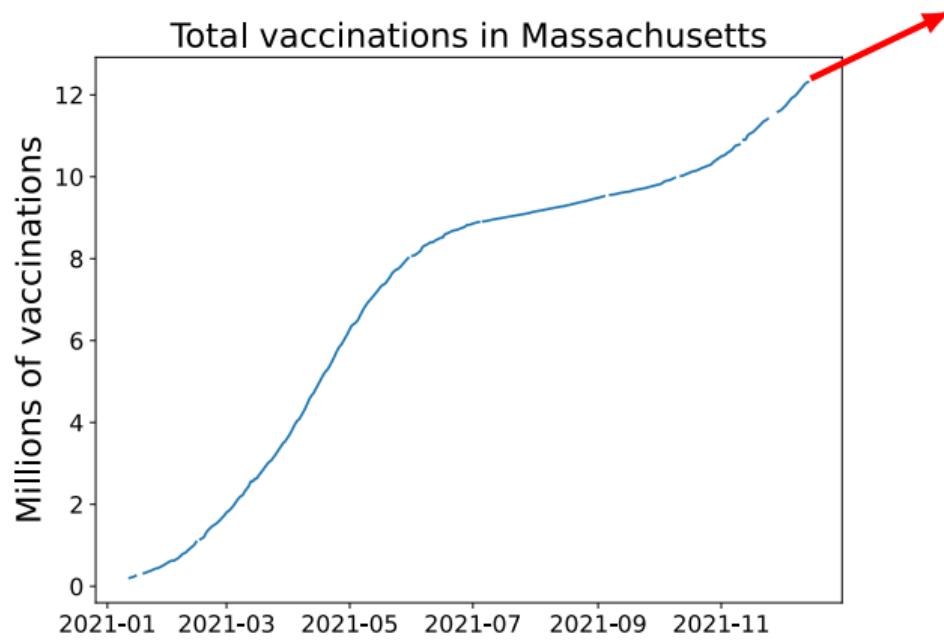


As b varies



Task #3: Prediction

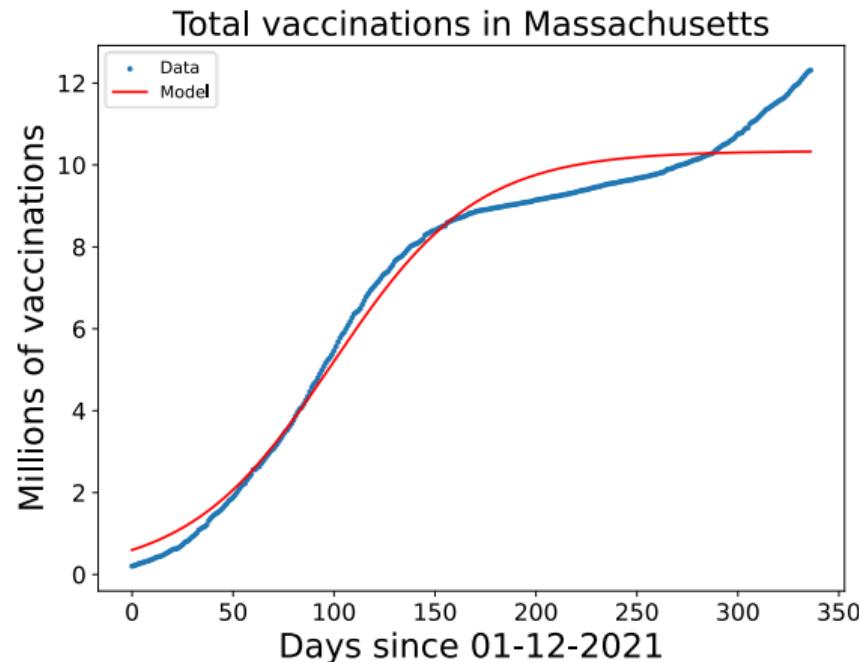
How many vaccinations will eventually be given in my state?



Task #3: Prediction

How many vaccinations will eventually be given in my state?

$$y = \frac{10.3}{1 + e^{-(0.028x - 2.79)}}$$



Task #3: Prediction

How many vaccinations will eventually be given in my state?

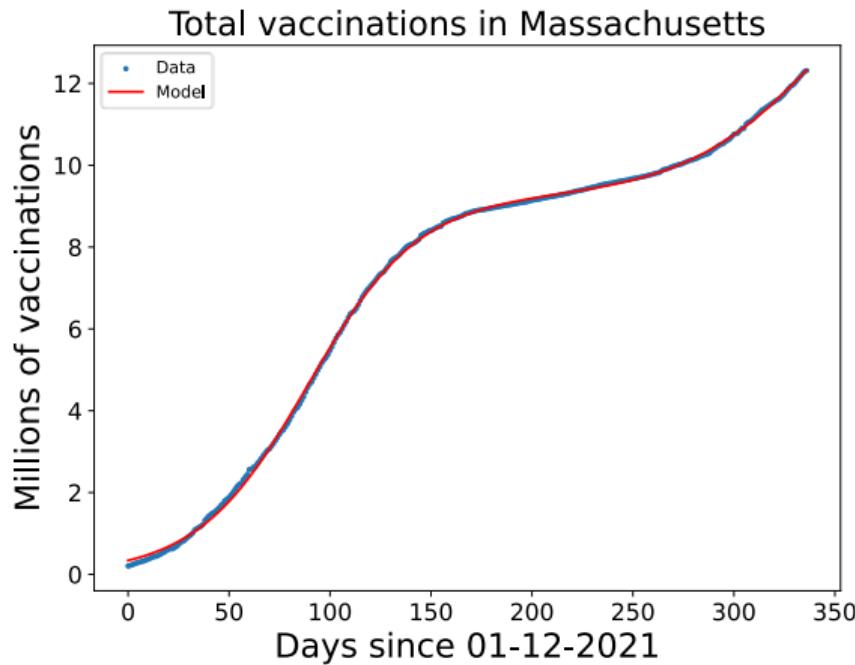
$$y = \frac{h_1}{1 + e^{-(m_1x+b_1)}} + \frac{h_2}{1 + e^{-(m_2x+b_2)}}$$



Task #3: Prediction

How many vaccinations will eventually be given in my state?

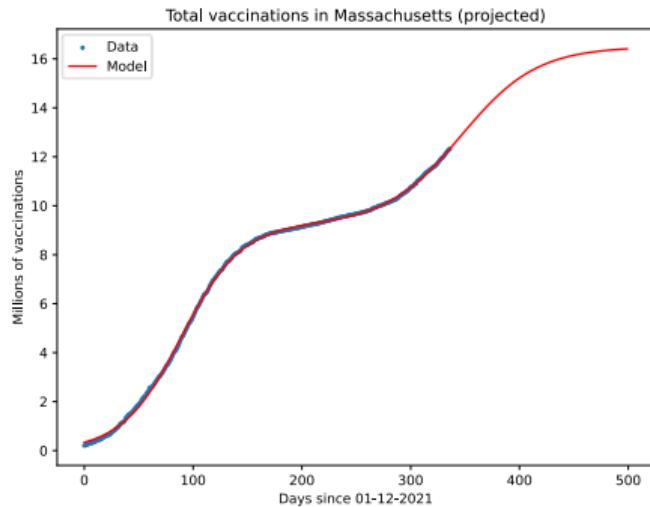
$$y = \frac{9.24}{e^{-(0.037x-3.27)}} + \frac{7.26}{e^{-(0.029x-9.97)}}$$



Task #3: Prediction

How many vaccinations will eventually be given in my state?

$$\lim_{t \rightarrow \infty} \left(\frac{h_1}{1 + e^{-(m_1 t + b_1)}} + \frac{h_2}{1 + e^{-(m_2 t + b_2)}} \right) = h_1 + h_2 \quad (\approx 16.5)$$



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶
- ▶
- ▶
- ▶

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶
- ▶
- ▶
- ▶

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



Part 2

Doing an Analysis



What do I mean by an “analysis?”

Example

Are vaccination rates in a state related to the state's politics?

- ▶ Find and load vaccination and voting data for each state.
- ▶ Perform the prediction we just saw on each state separately.
- ▶ Examine correlations between the predictions and voting data.
- ▶ Formulate those correlations as a hypothesis test.
- ▶ Summarize the results in a report with visualizations.



Example computational narrative (in Jupyter)

Vivaldi

deepnote.com/project/MA346-S22-Proj-1-Solution-Gw1XHlPtRuyL-CWpIW4Jw/%2Fnotebook.i...

Notion Carter MK346 S22 Proj 1 Solution

Vaccinations and Politics

Load the vaccination data

The data were downloaded from GitHub [here](#), as specified in the original project assignment.

```
import pandas as pd
df = pd.read_csv('us_state_vaccinations.csv')
df
```

date object	location object	total_vaccines	total_distrib_...	people_vaccin...	people_fully_v...	total_vaccine...	people_fully_v...	people_vaccin...
2021-01-12	0.3%	1.4K	476.0 - 5308293...	8988.5 - 166319...				
2021-01-13	0.3%	Alabama	1.5K	8988.5 - 166319...				
488 others	99.5%	63 others	99.5%					
8 2021-01-12	Alabama	78134	377825	78881	0.15	1.59	7278	1.45
1 2021-01-13	Alabama	846460	378975	74792	0.19	1.71	9245	1.53
2 2021-01-14	Alabama	923099	433398	88489	#NA	1.88	#NA	1.64
3 2021-01-15	Alabama	1005507	444959	86956	0.28	2.85	13488	1.77
4 2021-01-16	Alabama	nan	nan	nan	#NA	#NA	nan	nan
5 2021-01-17	Alabama	nan	nan	nan	#NA	#NA	nan	nan
6 2021-01-18	Alabama	nan	nan	nan	#NA	#NA	nan	nan
7 2021-01-19	Alabama	1387795	444698	114319	0.33	2.87	16348	2.33
8 2021-01-20	Alabama	1392090	483275	121113	0.37	2.84	17958	2.47
9 2021-01-21	Alabama	1659119	493125	144429	0.44	3.38	21345	2.95

240000 rows, showing 10 per page 0 < Page 1 of 2487 >

The vaccination data for all states and territories has been loaded and can be browsed above.



Example computational narrative (in Jupyter)

Vivaldi

deepnote.com/project/MA346-S22-Proj-1-Solution-Gw1XHlPtRuyL-CWpIW4Jw/%2Fnotebook.i...

Notion Carter MK346 S22 Proj 1 Solution

Vaccinations and Politics ← Headsings

Load the vaccination data

The data were downloaded from GitHub [here](#), as specified in the original project assignment.

```
import pandas as pd
df = pd.read_csv('us_state_vaccinations.csv')
df
```

	date object	location object	total_vaccines	total_distrib_...	people_vaccina...	people_fully_r...	total_vaccine...	people_fully_r...	people_vaccin...
0	2021-01-12	United Sta...	1.45	476.0 - 5308293...	8908.5 - 1663113...				
1	2021-01-13	Alabama	0.3%	53 others	59.8%				
2	2021-01-13	Alabama	846960	377825	78861	0.15	1.59	7278	1.45
3	2021-01-15	Alabama	92399	433398	88489	1.88	1.88	1.88	1.88
4	2021-01-16	Alabama	nan	nan	nan	nan	nan	nan	nan
5	2021-01-17	Alabama	nan	nan	nan	nan	nan	nan	nan
6	2021-01-18	Alabama	nan	nan	nan	nan	nan	nan	nan
7	2021-01-19	Alabama	138799	444698	114319	0.33	2.87	16348	2.33
8	2021-01-20	Alabama	139298	483275	121113	0.37	2.84	17958	2.47
9	2021-01-21	Alabama	165919	49125	144429	0.44	3.38	21345	2.95

← Tables

The vaccination data for all states and territories has been loaded and can be browsed above.



Example computational narrative (in Jupyter)

The six values shown above are for β_1 through β_6 , so the model for Massachusetts would be as follows (parameters rounded for brevity).

$$\hat{y} = \frac{134.7}{1 + e^{0.017(98.3-t)}} + \frac{78.6}{1 + e^{0.018(228.3-t)}}$$

Comparing model to data

We can plot the data and the model on the same curve to visualize how similar or different they are. (There are quantitative measures for this as well, but we will just do a visual check here.) The following function defines how to do this, and we run it on Massachusetts data again, as an example.

```
def model.compare(state):
    betas = get_state.betas(state)
    model = lambda t: model.logistic(t, *betas)
    data = off.state.get(state)
    predictions = [model(i) for i in range(len(data))]
    plt.plot(data.date, predictions, label='Model')
    plt.plot(data.date, data.vax, label='Data')
    plt.xlabel('Date')
    plt.ylabel('Vaccinations per Hundred People')
    plt.legend()
    return plt.show()

model.compare('Massachusetts')
```

Modeling Vaccinations per Hundred in Massachusetts

Table of model parameters by state

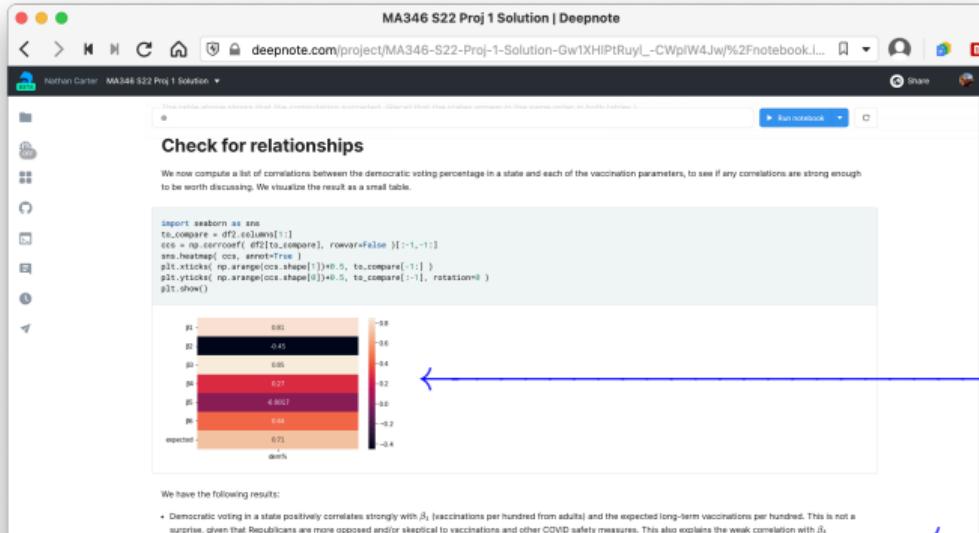
Math

Code

Graphs



Example computational narrative (in Jupyter)



Visualizations

Text



Computational narratives are like proofs



Computational narratives are like proofs

Vaccination data timeframe

What time period does the vaccination data cover?

```
df.Date.iloc[0], df.Date.iloc[-1]
```

[8]

```
(Timestamp('2021-01-12 00:00:00'), Timestamp('2021-12-14 00:00:00'))
```

The minimum and maximum dates in the dataset are therefore:

- January 12, 2021
- December 14, 2021



Computational narratives are like proofs

Vaccination data timeframe

What time period does the vaccination data cover?

```
df.Date.min(), df.Date.max()
```

[7]

```
(Timestamp('2021-01-12 00:00:00'), Timestamp('2021-12-14 00:00:00'))
```

The minimum and maximum dates in the dataset are therefore:

- January 12, 2021
- December 14, 2021



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶
- ▶

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



Great Expectations, from Superconductive

great_expectations

COMMUNITY EXPECTATIONS DOCUMENTATION CASE STUDIES BLOG

Greetings! We now have a cloud offering [Learn More](#)

Case studies from Great Expectations

Stay up to date on our releases and comments on the current state of data management



Great Expectations Case Study: Vimeo

March 24, 2021

About Vimeo Vimeo is one of the world's leading service providers of HD streaming video, allowing over 200M professionals, teams, and...



Great Expectations Case Study: Heineken

November 16, 2020

Check out the video of the HEINEKEN team presenting at our first Great Expectations Community Show & Tell in October 2020! About HEINEKEN...



Great Expectations Case Study: Avanade

October 08, 2020

About Avanade Avanade is a global professional services company providing IT consulting and services focused on the Microsoft platform.

Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$.

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p .

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p . □



Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$. ← Point A

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p .

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p . □



Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$. \leftarrow Point A

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p . \leftarrow Point B

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p . \square



Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$. ← Reader's knowledge at A

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p . ← Reader's knowledge at B

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p . □



Computational narratives are step-by-step constructions

← Computer's memory at A

Vaccination data timeframe

What time period does the vaccination data cover?

```
df.Date.min(), df.Date.max()
```

[7]

```
(Timestamp('2021-01-12 00:00:00'), Timestamp('2021-12-14 00:00:00'))
```

The minimum and maximum dates in the dataset are therefore:

- January 12, 2021
- December 14, 2021

← Computer's memory at B



Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$.

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p .

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p .



Don't actually read this proof

Theorem

Every prime $p > 3$ is adjacent to a multiple of 6.

Proof.

Let $p > 3$ be prime and therefore divisible by neither 2 nor 3. So for some k we have $p = 2k + 1$ and for some m we have either $p = 3m + 1$ or $p = 3m + 2$. **We consider each case separately.**

Case 1: When $p = 3m + 1$, $p - 1 = 3m$ and $p - 1$ is divisible by 3. But $p - 1 = 2k$ as well, so $p - 1$ is also divisible by 2. So $p - 1$ is divisible by 6 and adjacent to p .

Case 2: When $p = 3m + 2$, $p + 1 = 3(m + 1)$ and $p + 1$ is divisible by 3. But $p + 1 = 2(k + 1)$ as well, so $p + 1$ is also divisible by 2. So $p + 1$ is divisible by 6 and adjacent to p .

Since the theorem holds in both cases, it holds in general. □



Computational narratives are *narratives*

Vaccination data timeframe ← Reader's intuition and comfort at A

What time period does the vaccination data cover?

```
df.Date.min(), df.Date.max()
```



```
(Timestamp('2021-01-12 00:00:00'), Timestamp('2021-12-14 00:00:00'))
```

The minimum and maximum dates in the dataset are therefore: ← Reader's intuition and comfort at B

- January 12, 2021
- December 14, 2021



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶ **Step-by-step constructions**
- ▶ **Supplying intuition**

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



Which advertisements to show?

All News Images Videos Shopping More Tools

About 18,200,000 results (0.58 seconds)

Ads · Vehicles for sale



2020 Hyundai Ioniq Hybrid Blue
\$23,990
Used - 33k mi
Carvana
📍 Danvers



2020 Hyundai Ioniq Hybrid Blue
\$24,990
Used - 10k mi
Carvana



2019 Hyundai Ioniq Hybrid SEL
\$22,300
Used - 21k mi
Grieco Hyundai
📍 Johnston

Tax, title, and processing fees may apply

<https://www.hyundaiusa.com/vehicles/2022-ioniq-5>

2022 IONIQ 5 | Electric SUV, Global Reveal | Hyundai USA

A breakthrough SUV that's the first model we've ever built to be exclusively electric. Its arrival is the exciting beginning of our new series of all-electric ...

You've visited this page 3 times. Last visit: 1/14/22



People also ask

How much will the Ioniq 5 cost?

Is the Ioniq 5 available?

Is the Hyundai Ioniq 5 electric?

How big is the Hyundai Ioniq 5?



More images

2022 Hyundai IONIQ 5

Sport utility vehicle

8.5/10 Car and Driver	3/5 Kelley Blue Book
--------------------------	-------------------------

MSRP: From \$43,650

Range: 256 to 303 mi battery-only

Dimensions: 183" L x 74" W x 63" H

Cargo volume: 27.2 ft³, 59.3 ft³ with seat area

Battery: 77.4 kWh 697 V lithium polymer

MPGe: Up to 132 city / 98 highway

Configurations

SE From \$43,650

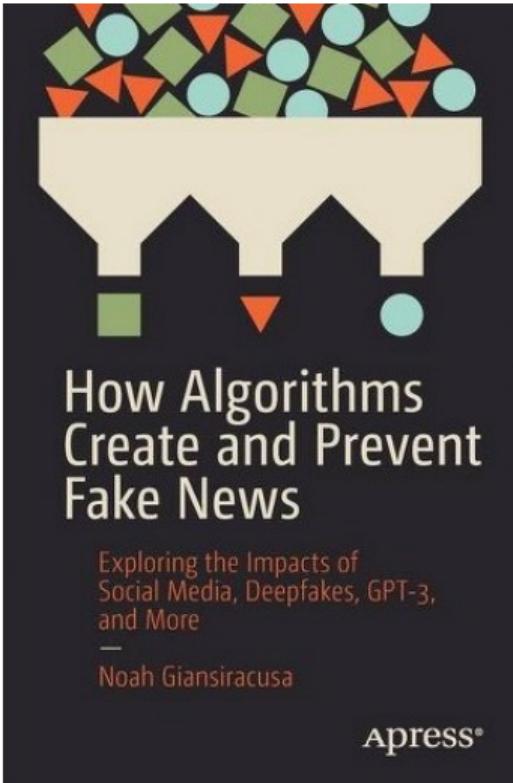
Which advertisements to show?

Google's assumption

If we let the behavior of people on our site determine which ads should show up for *everyone* that will be good, in all cases.



Which advertisements to show?



SEARCH ENGINES

acmqueue Discrimination in Online Ad Delivery

Google ads, black names and white names, racial discrimination, and click advertising

Latanya Sweeney

Do online ads suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names? What is a black-sounding name or white-sounding name, anyway? How many more times would an ad have to appear adversely affecting one racial group for it to be considered discrimination? Is online activity so ubiquitous that computer scientists have to think about societal consequences such as structural racism in technology design? If so, how is this technology to be built? Let's take a scientific dive into online ad delivery to find answers.

"Have you ever been arrested?" Imagine this question appearing wherever someone enters your name in a search engine. Perhaps you are in competition for an award, a scholarship, an appointment, a promotion, or a new job, or maybe you are in a position of trust, such as a professor, a physician, a banker, a judge, a manager, or a volunteer. Perhaps you are completing a rental application, selling goods, applying for a loan, joining a social club, making new friends, dating, or engaged in any one of hundreds of circumstances for which someone wants to learn more about you online. Appearing alongside your list of accomplishments is an advertisement implying you may have a criminal record, whether you actually have one or not. Worse, the ads may not appear for your competitors.

Job applications frequently include questions such as: Have you ever been arrested? Have you ever been charged with a crime? Other than a traffic ticket, have you been convicted of a crime? Employers ask these questions to establish trustworthiness. Because others often equate a criminal record with not being reliable or honest, protections exist for those having criminal records.

If an employer disqualifies a job applicant based solely upon information indicating an arrest record, the company may face legal consequences. The U.S. EEOC (Equal Employment Opportunity Commission) is the federal agency charged with enforcing Title VII of the Civil Rights Act of 1964, a law that applies to most employers, prohibiting employment discrimination based on race, color, religion, sex, or national origin. Guidance issued in 1973 extended protections to people with criminal records.¹⁴ Title VII does not prohibit employers from obtaining criminal background information. Certain uses of criminal information, however, such as a blanket policy or practice of excluding applicants or disqualifying employees based solely upon information indicating an arrest record, can result in a charge of discrimination.

To make a determination, the EEOC uses an adverse impact test that measures whether certain practices, intentional or not, have a disproportionate effect on a group of people whose defining characteristics are covered by Title VII. To decide, you calculate the percentage of people affected in each group and then divide the smaller value by the larger to get the ratio and compare the result to 80. For example, suppose a company laid off comparable black and white workers at the same rate—25 percent of blacks and 25 percent of whites—then the ratio, 25 divided by 25, would be 100 percent. If the ratio is less than 80 percent, then the EEOC considers the effect disproportionate and may hold the employer responsible for discrimination.¹⁵

1

A set of small, light-colored navigation icons typically used in presentation software like Beamer. They include symbols for back, forward, search, and other document-related functions.

Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶ Step-by-step constructions
- ▶ Supplying intuition

Unsolved Problems

- ▶
- ▶
- ▶
- ▶



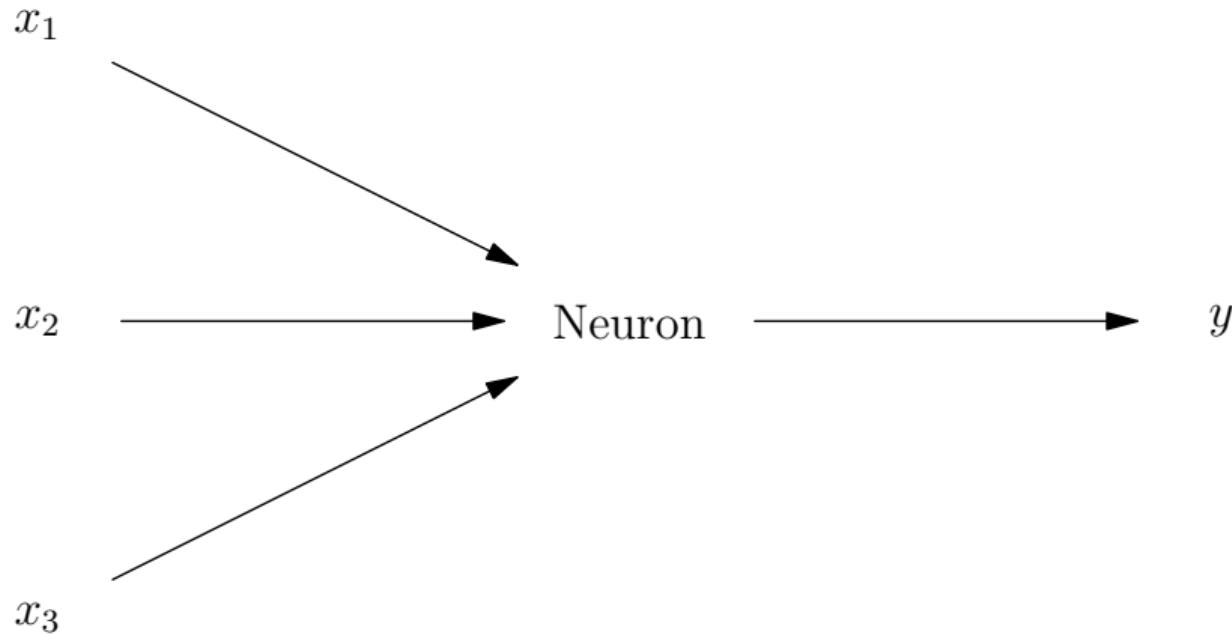
Part 3

Unsolved Problems



Neural Networks

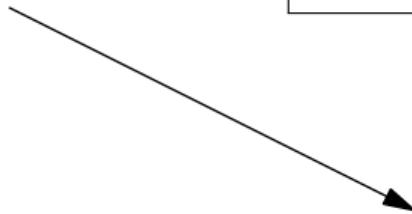
Inspiration



Neural Networks

Inspiration

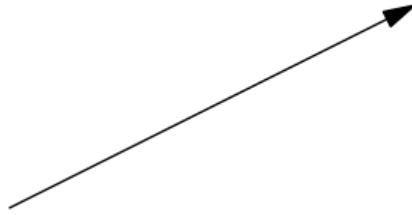
x_1



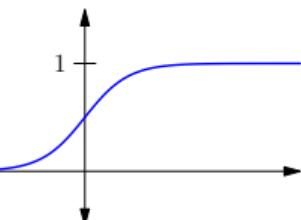
x_2



x_3



$$y = \frac{1}{1+e^{-(mx+b)}}$$



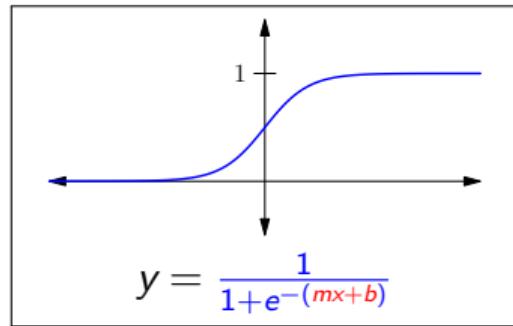
Neural Networks

Inspiration

x_1

x_2

x_3



$$mx + b$$

Neuron

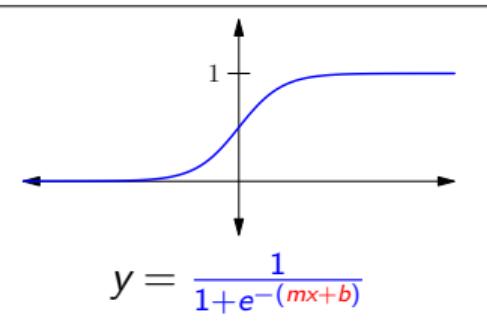
$$\frac{1}{1 + e^{-x}}$$

y



Neural Networks

Inspiration



x_1

x_2

x_3

$$m_1x_1 + b_1$$

$$m_2x_2 + b_2$$

$$m_3x_3 + b_3$$

Neuron

$$\frac{1}{1 + e^{-S}}$$

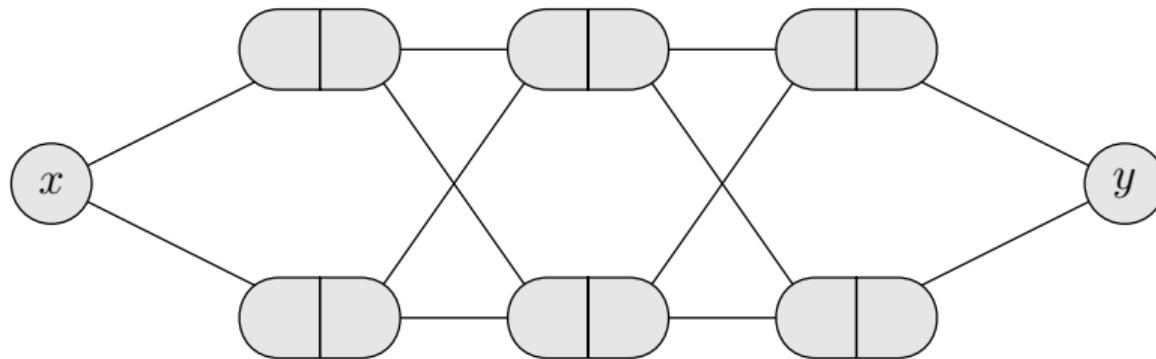
"sigmoid"

$$S = \sum(m_i x_i + b_i)$$



Neural Networks

Example network

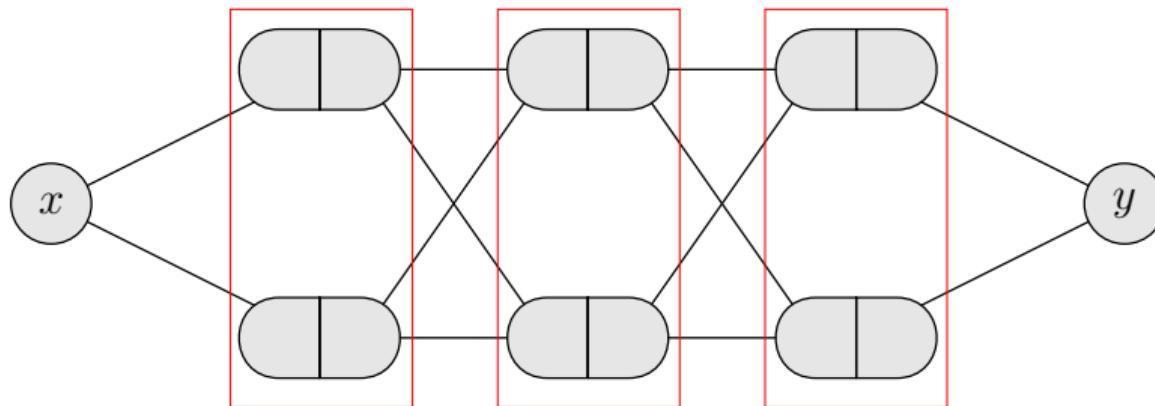


with sigmoid activation function ($\frac{1}{1+e^{-x}}$)



Neural Networks

Example network

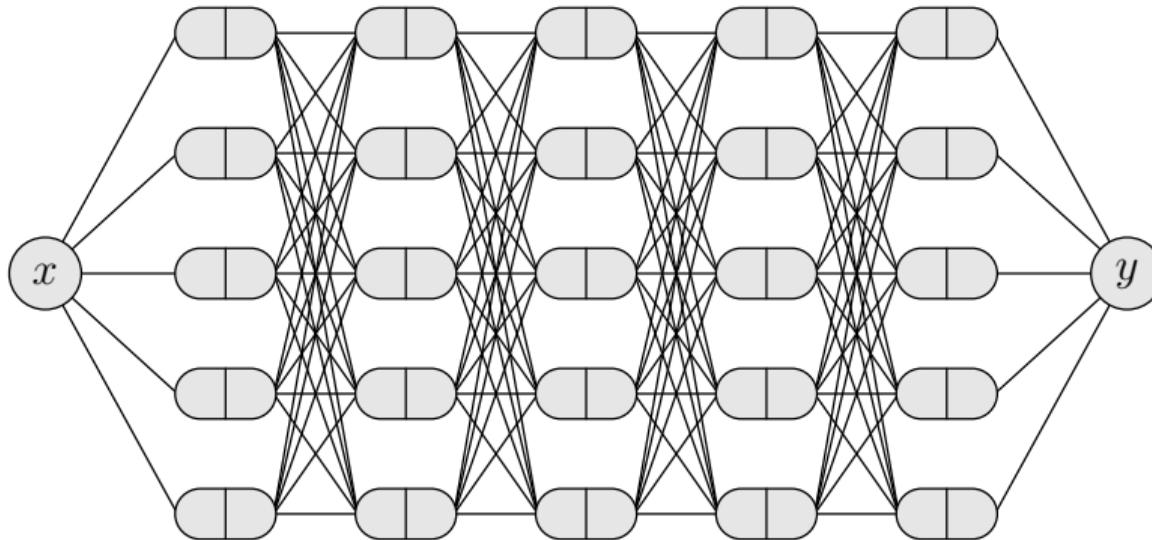


with sigmoid activation function ($\frac{1}{1+e^{-x}}$)

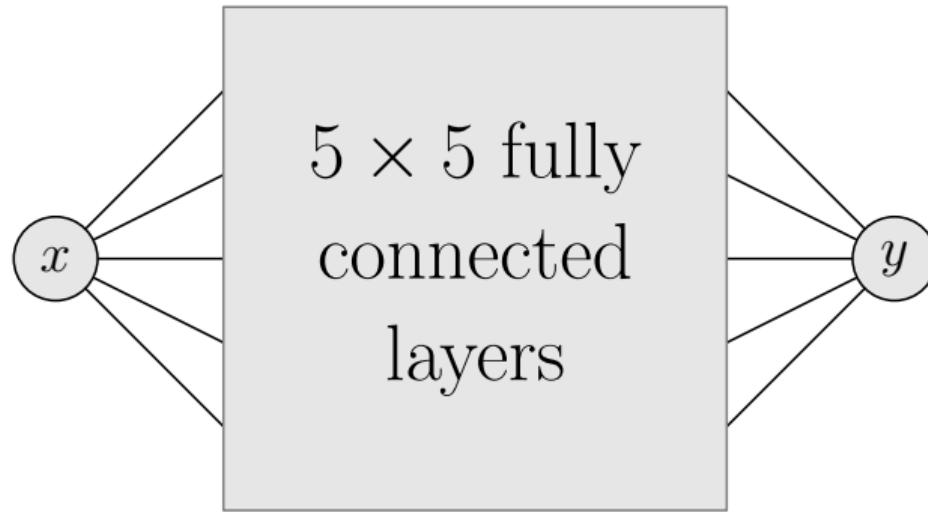


Neural Networks

Network with 5 dense layers of width 5

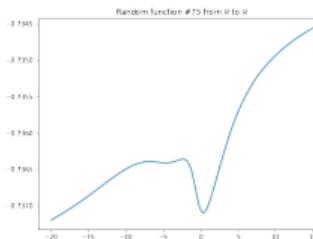
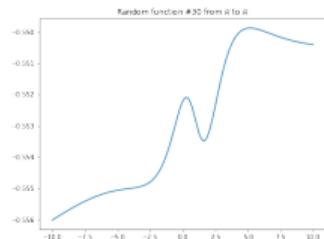
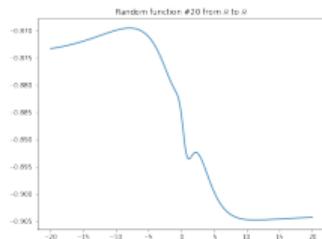
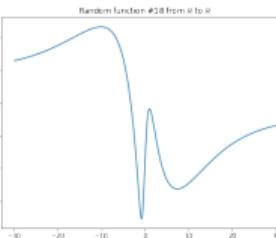
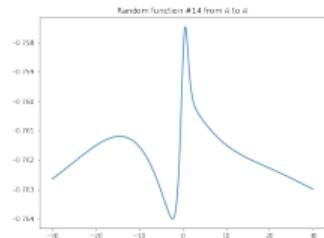
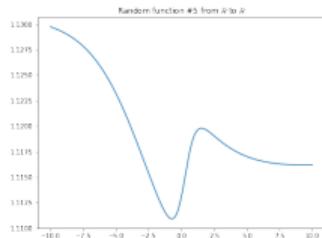


Neural Networks

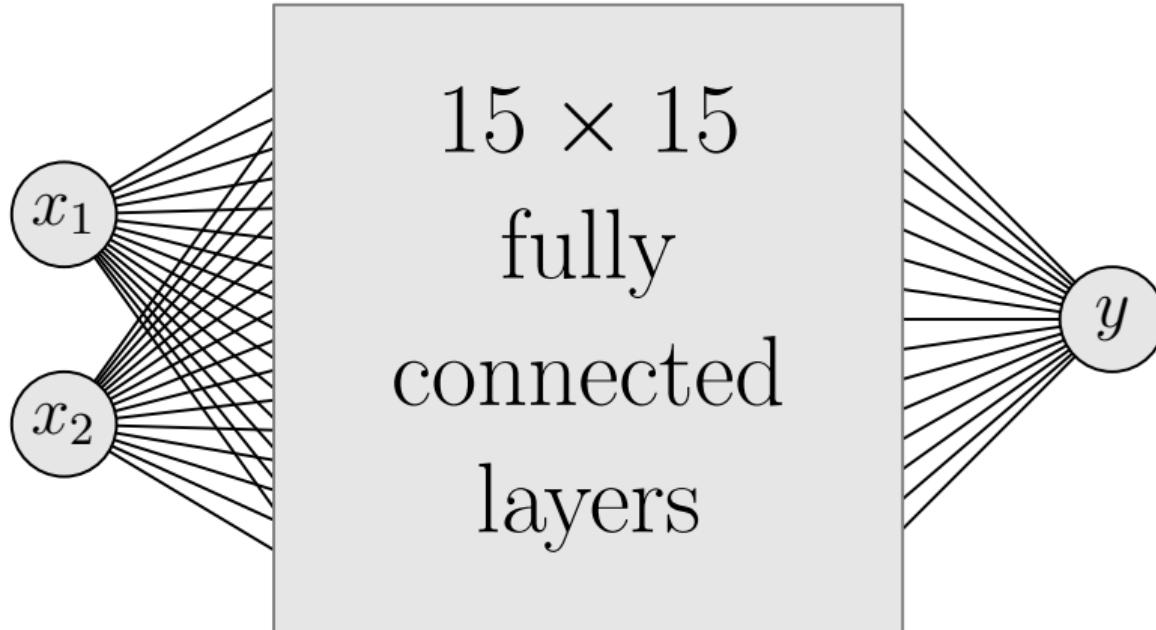


Example Neural Network functions $f: \mathbb{R} \rightarrow \mathbb{R}$

5 dense layers of width 5 with sigmoid activation function

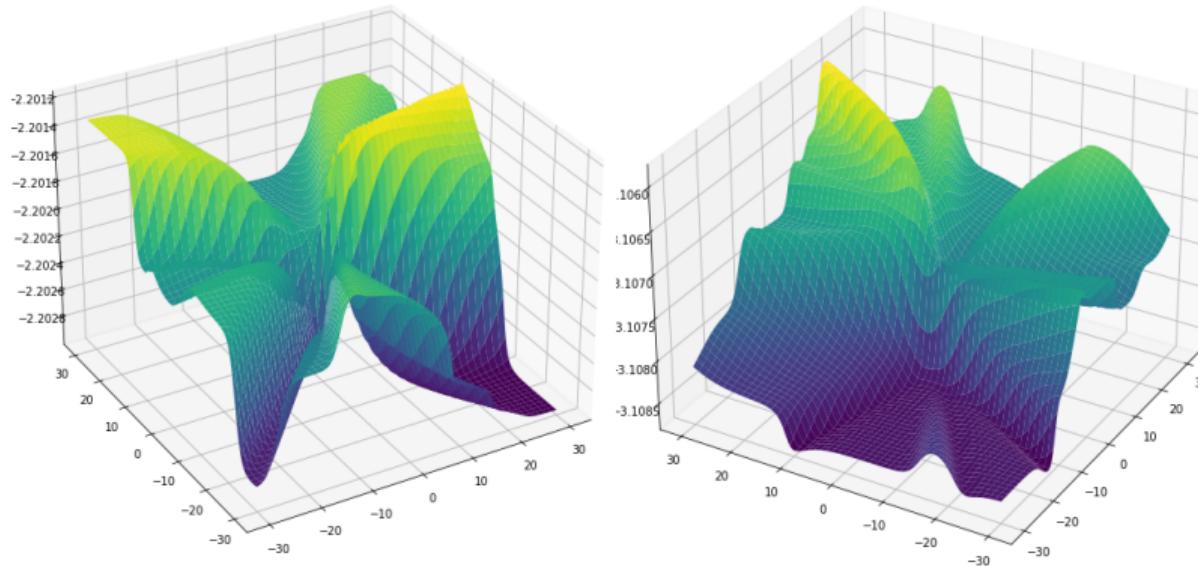


Neural Networks



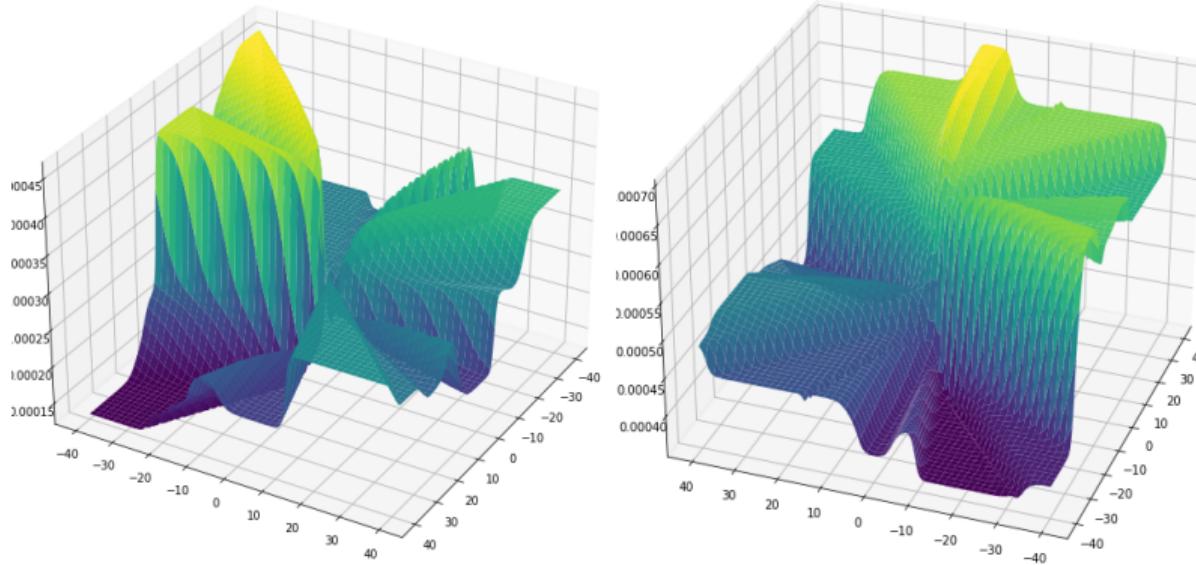
Example Neural Network functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}$

15 dense layers of width 15 with sigmoid activation function



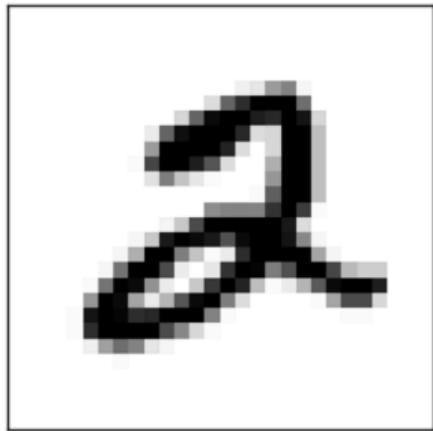
Example Neural Network functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}$

15 dense layers of width 15 with sigmoid activation function



What can neural networks do?

Example: Classify handwritten digits



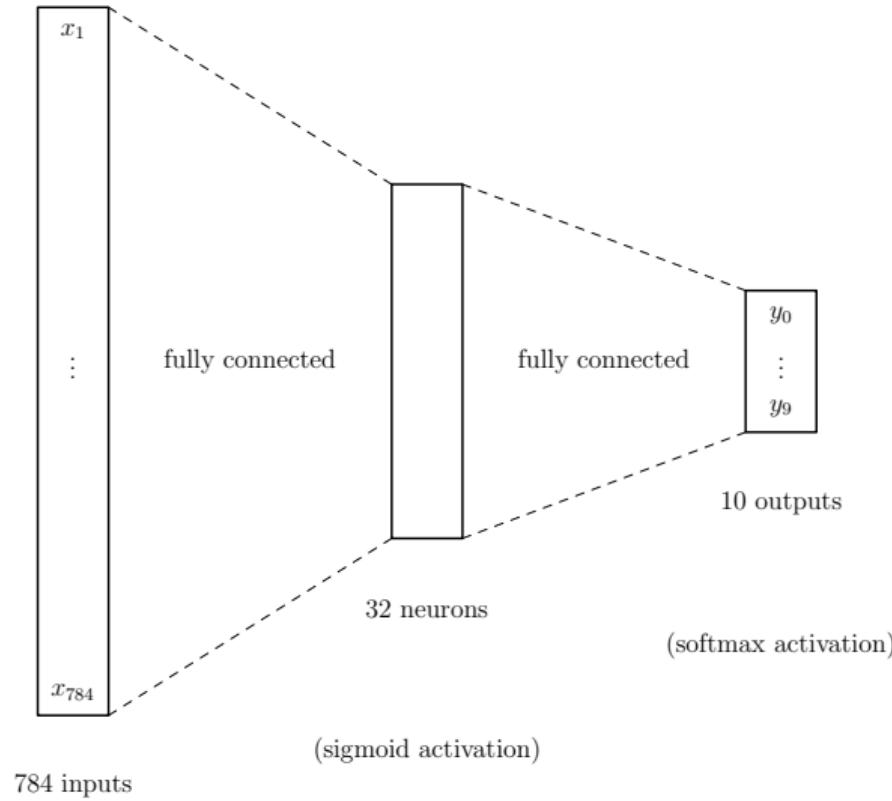
2

Grayscale image, 28×28
 $x_1, x_2, \dots, x_{784} \in [0, 1]$

Predicted digit
 $y \in \{0, \dots, 9\}$

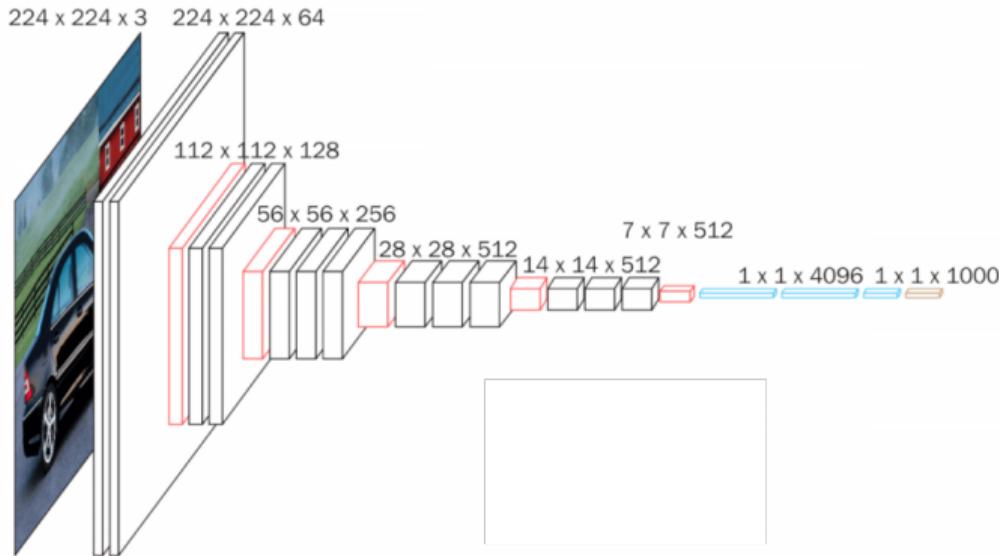


Architecture to classify handwritten digits



VGG16 Architecture

Authors: Karen Simonyan and Andrew Zisserman, 2014



(about 14.7 million parameters)

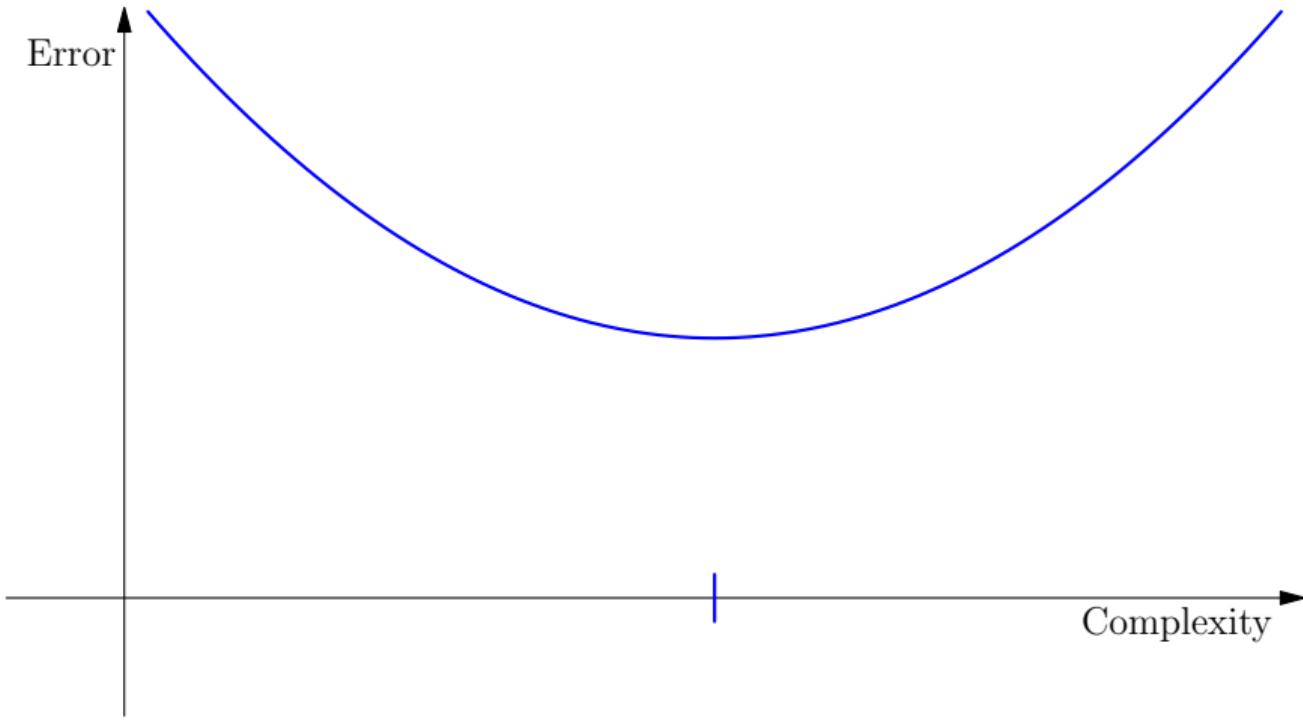


Two Unsolved Problems in Neural Networks

1. Are models with millions/billions of parameters *too* flexible?
2. What do all of those parameters mean (if anything)?

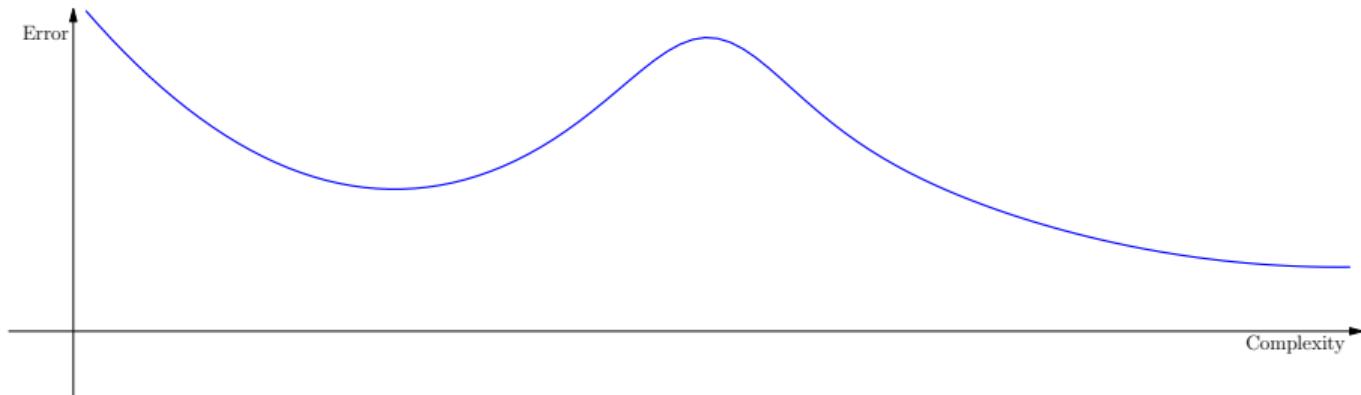


Traditional Wisdom in Mathematical Modeling



Double descent

(Made famous in 2018 by Belkin, Hsu, Ma, and Mandal)



What causes double descent?

Theories and partial answers:

- ▶ The way we train neural networks ensures we get “smooth” ones. (Belkin et al., 2019)
- ▶ The parameters might have **much less freedom** than we think they do, because they outnumber the data. (Advani et al., 2019)
- ▶ The **number of parameters required** to achieve smoothness **is much greater** than a naive estimate would suggest ($\propto n \times d$). (Bubeck and Sellke, Dec.2021)



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶ Step-by-step constructions
- ▶ Supplying intuition

Unsolved Problems

- ▶ Probability
- ▶ Real analysis
- ▶
- ▶



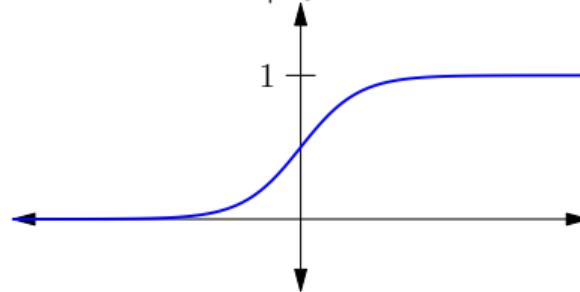
Two Unsolved Problems in Neural Networks

1. Are models with millions/billions of parameters *too* flexible?
2. What do all of those parameters mean (if anything)?

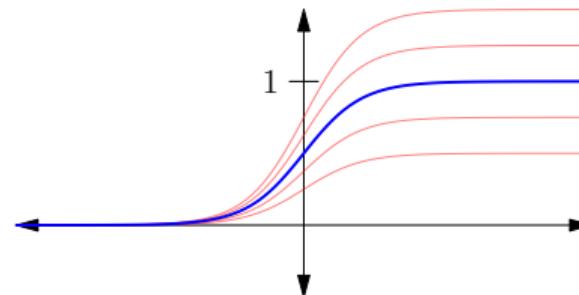


Interpreting Model Parameters

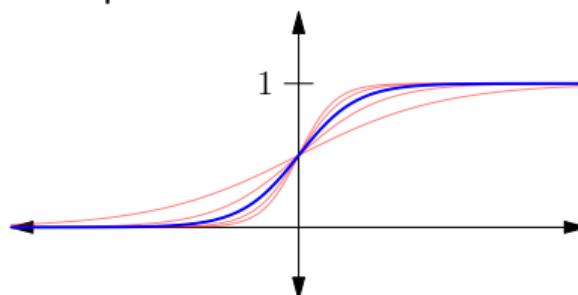
$$y = \frac{h}{1 + e^{-(mx+b)}}$$



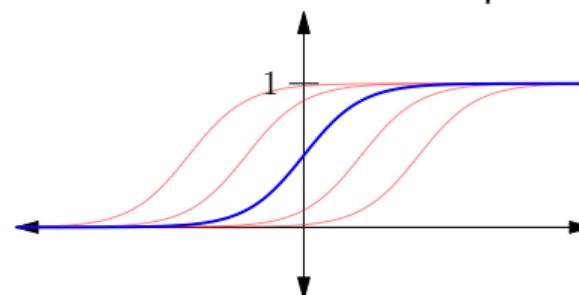
h = total vaccinations long-term



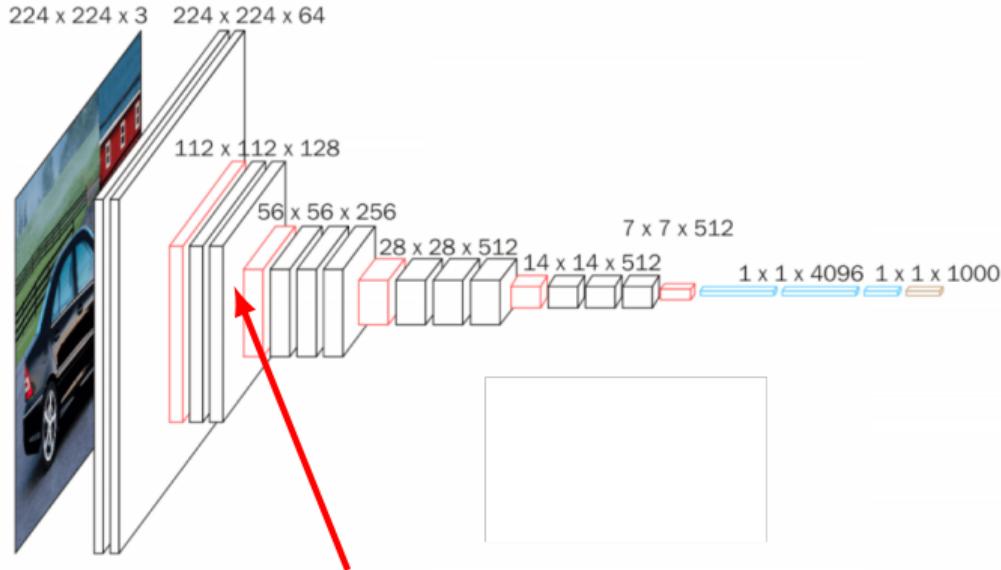
m ≈ speed of vaccine distribution



b = time when half complete



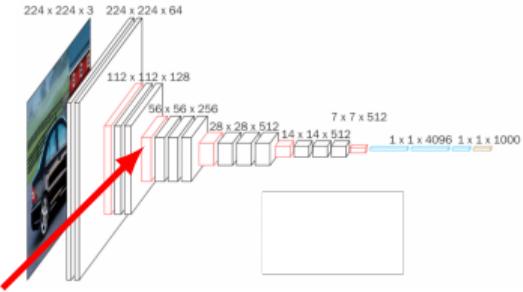
Interpretability



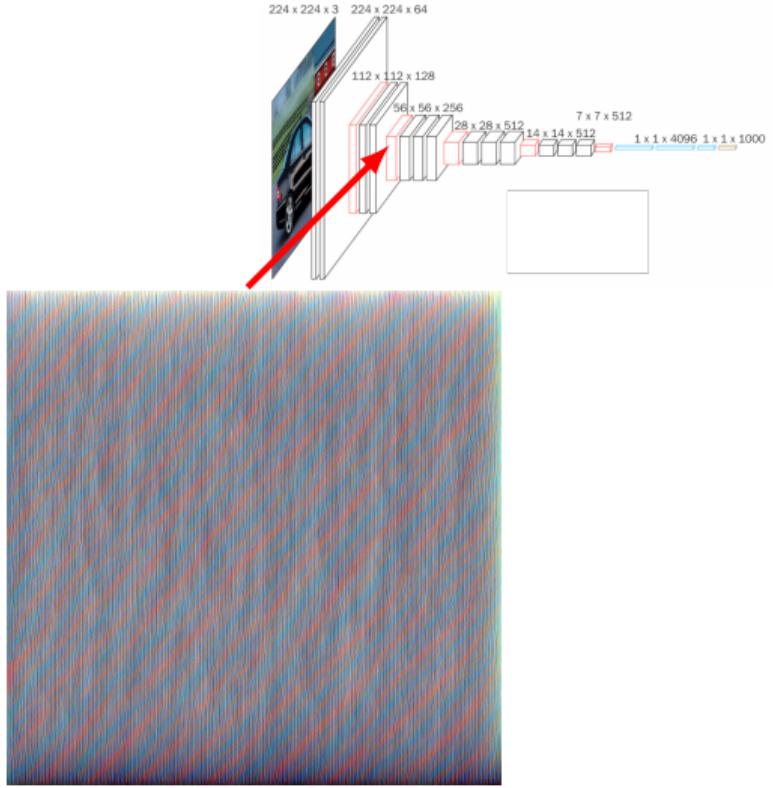
What does parameter $m_{5,903,646}$ mean?



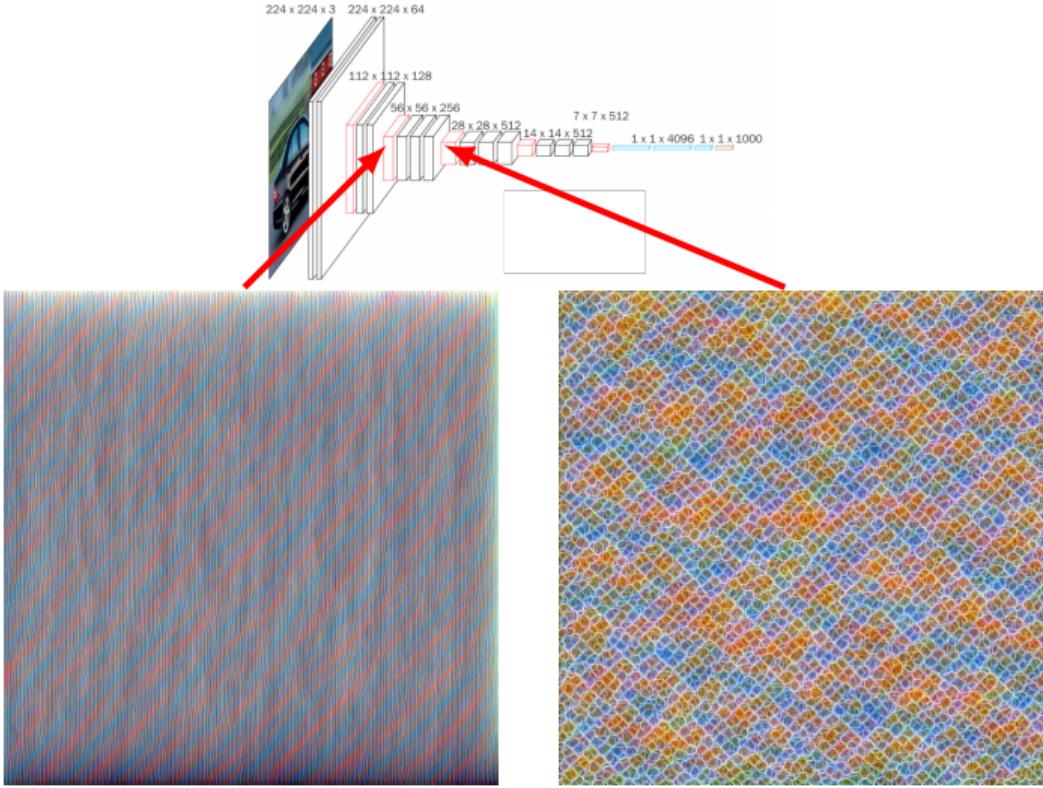
Interpreting Model Parameters



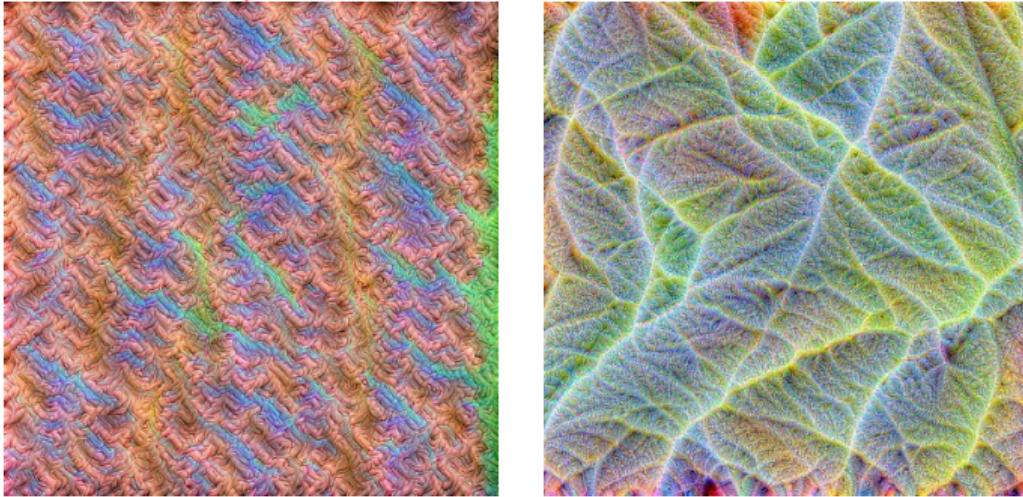
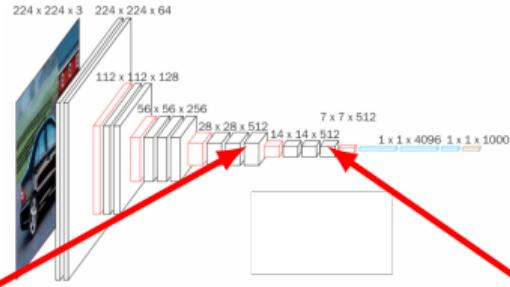
Interpreting Model Parameters



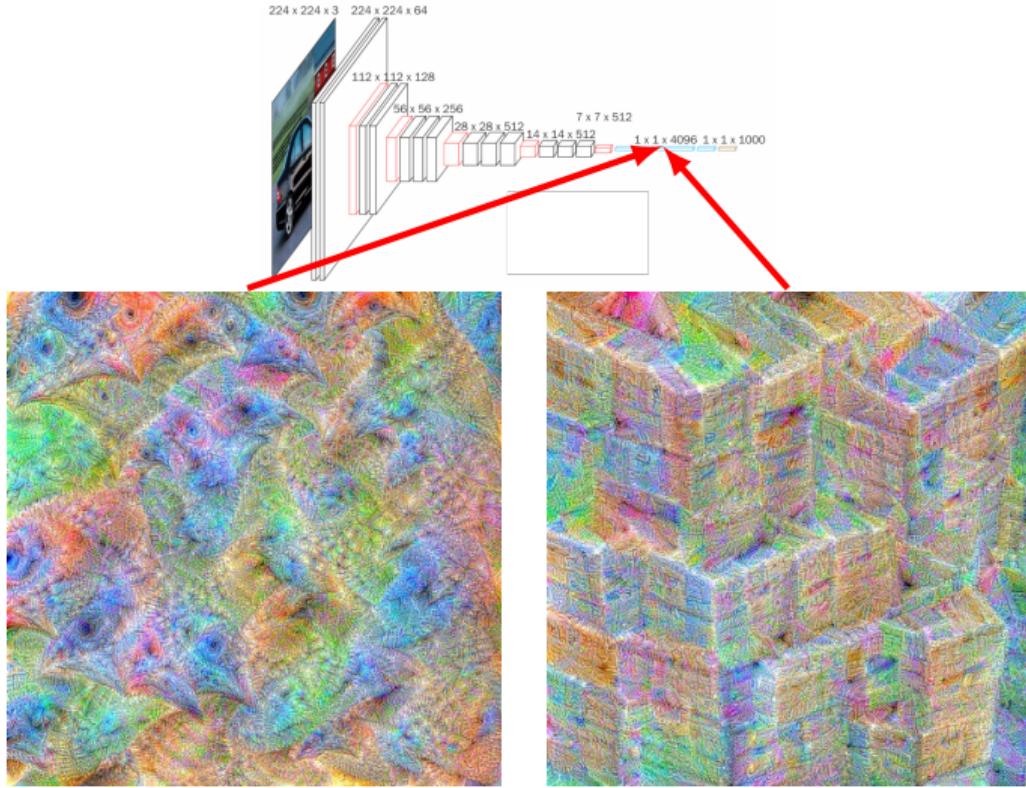
Interpreting Model Parameters



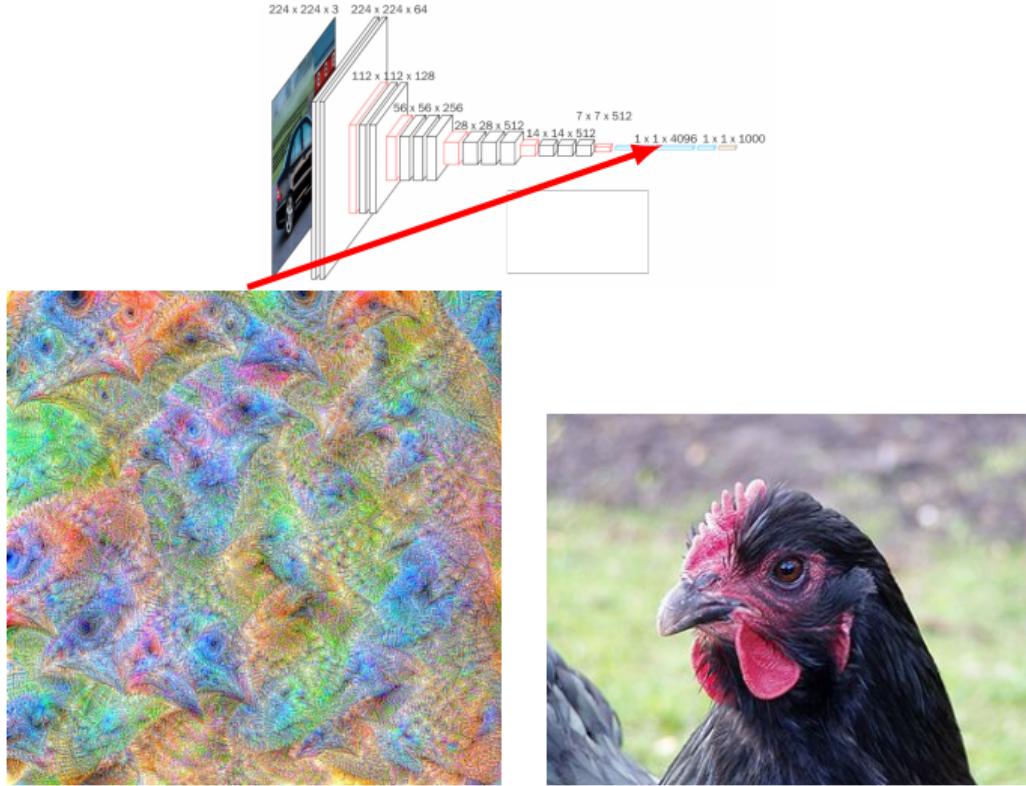
Interpreting Model Parameters



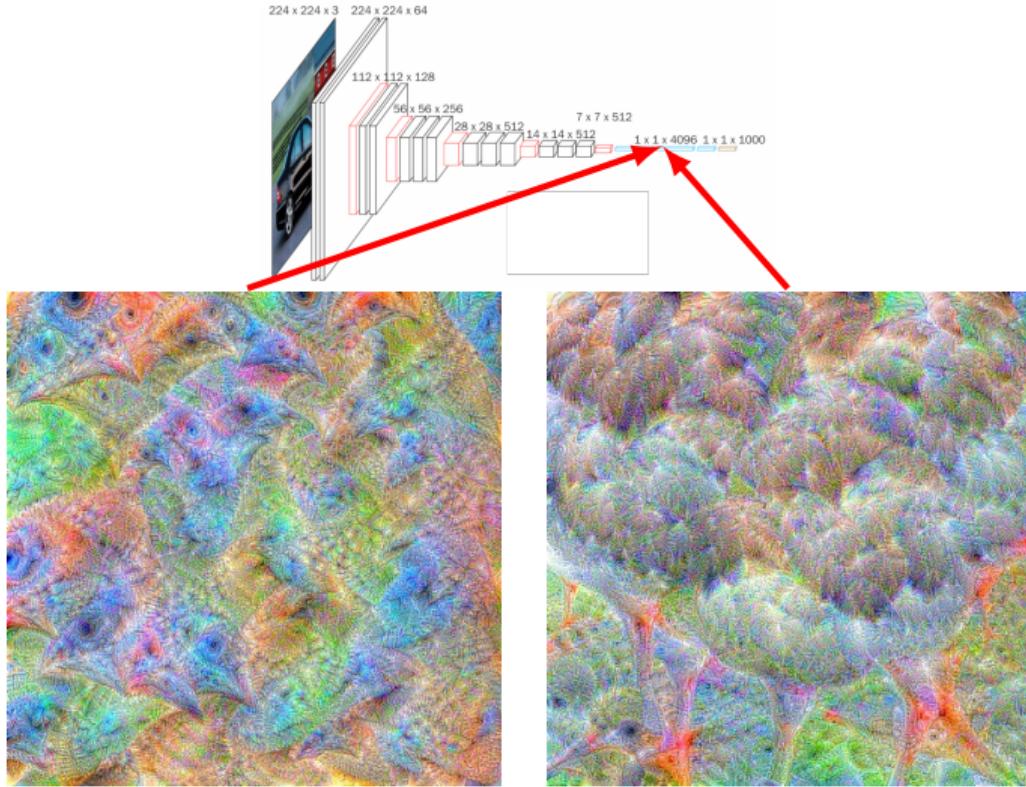
Interpreting Model Parameters



Interpreting Model Parameters



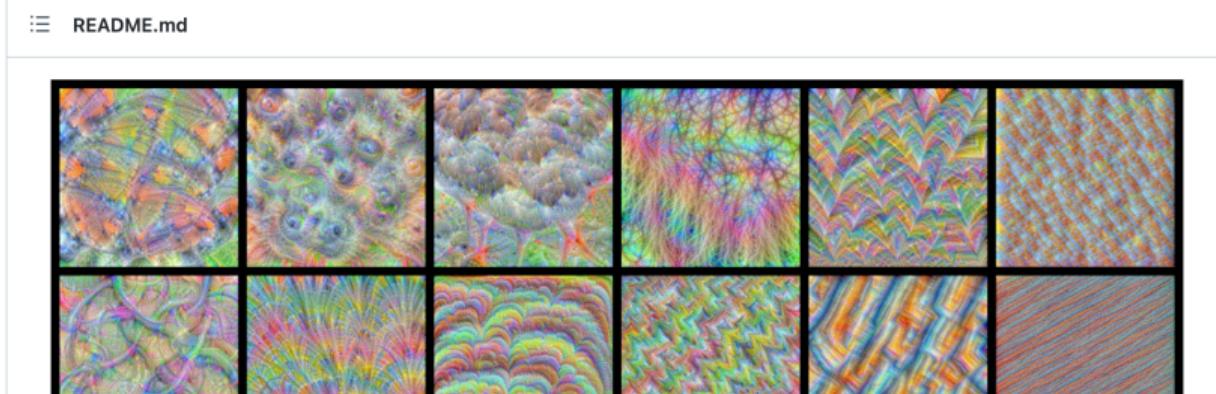
Interpreting Model Parameters



Interpreting Model Parameters

Author: Fabio Graetz, <https://github.com/fg91/visualizing-cnn-feature-maps>

 fg91 Remove broken link to colab notebook	cb245ed on Jul 21, 2019	22 commits
 experiments added resnet34 part 2	3 years ago	
 pictures added cover	3 years ago	
 Calculate_mean_activation_per_filt... cleaned up code	3 years ago	
 README.md Remove broken link to colab notebook	2 years ago	
 env.yml Describe how to setup virtual env	3 years ago	
 filter_visualizer.ipynb cleaned up code	3 years ago	



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶ Step-by-step constructions
- ▶ Supplying intuition

Unsolved Problems

- ▶ Probability
- ▶ Real analysis
- ▶ Multivariate calculus
- ▶ Linear algebra



Mathematics in Data Science

Working with Data

- ▶ Functions and relations
- ▶ Domain and range
- ▶ Famous sets (reals, etc.)
- ▶ Sets, subsets, finite sets
- ▶ Asymptotes/limits
- ▶ Famous function families
- ▶ Algebraic transformations

Doing an Analysis

- ▶ Questioning assumptions
- ▶ Using official definitions
- ▶ Step-by-step constructions
- ▶ Supplying intuition

Unsolved Problems

- ▶ Probability
- ▶ Real analysis
- ▶ Multivariate calculus
- ▶ Linear algebra



Thank you for joining us for this MAA
Distinguished Lecture Series presentation.



MATHEMATICAL ASSOCIATION OF AMERICA