INFSCI 2595 - Introduction to Machine Learning

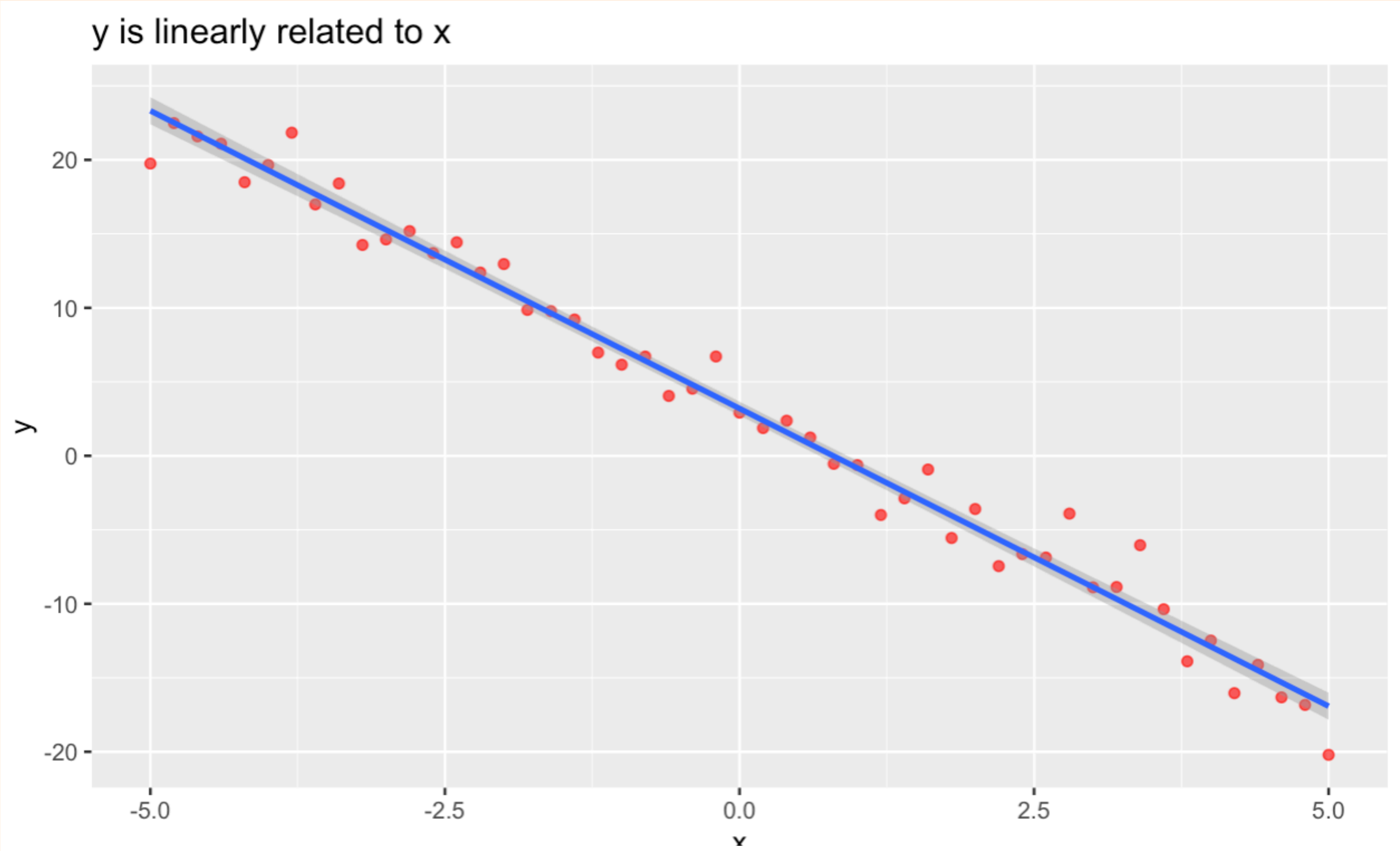
Lecture 1 - Introduction, Administration, and Overview of Course

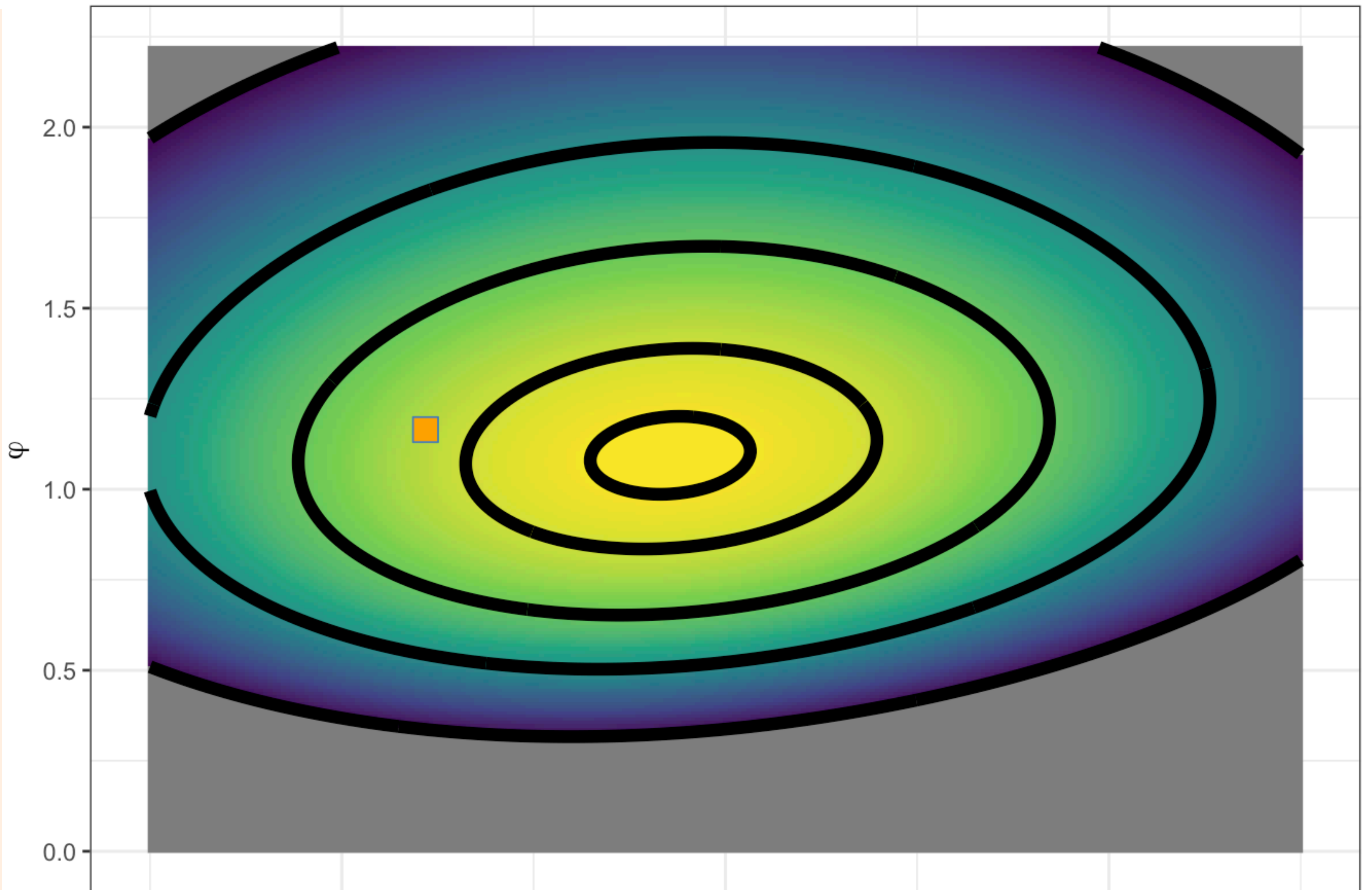
Spring 2025

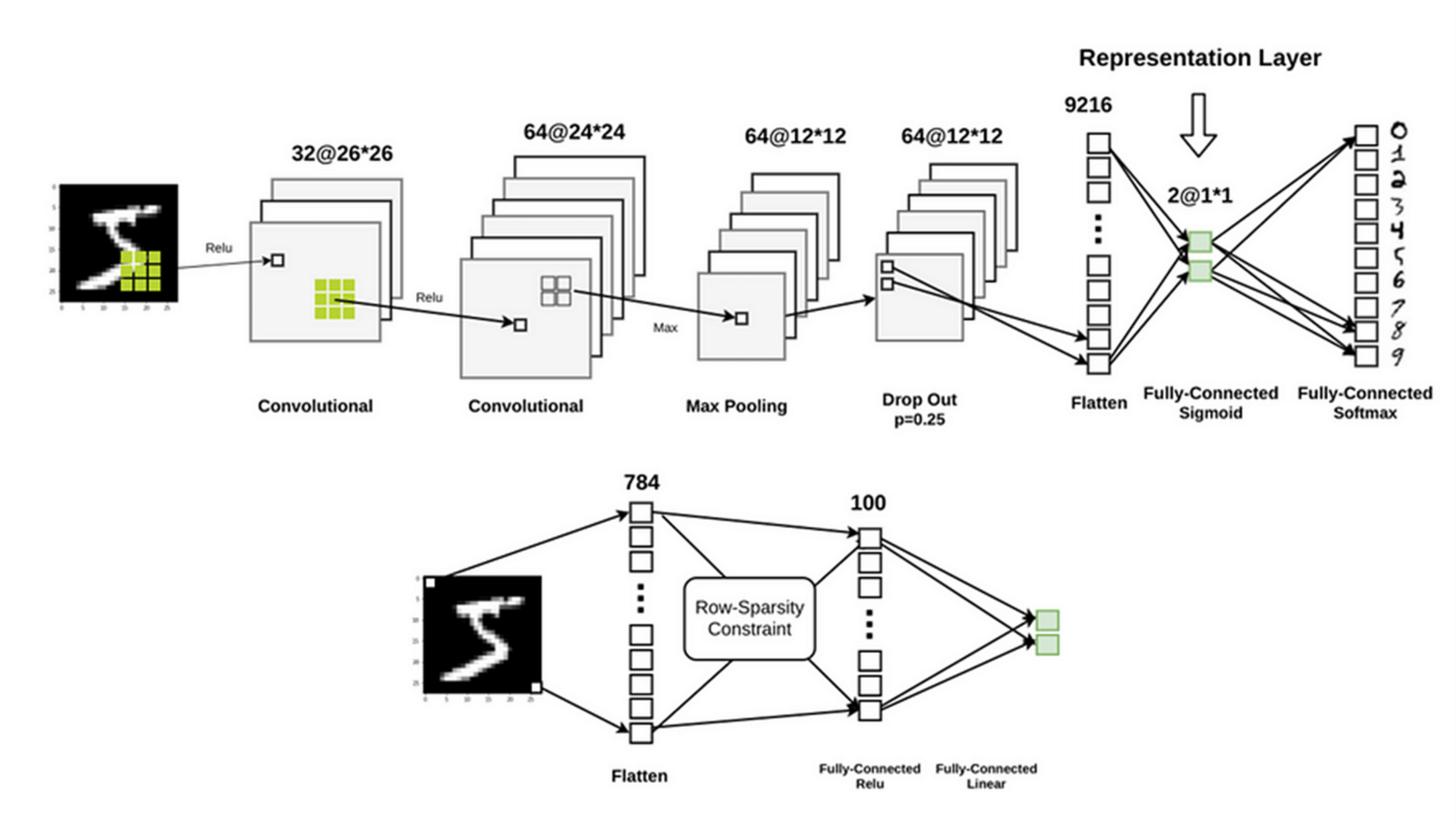
Instructor: Dr. Patrick Skeba

Machine learning is a set of tools for recognizing patterns in data

…essentially, applied linear algebra and probability!







Even the most advanced “artificial intelligence” is based in basic principles of mathematics and statistics

- The objective/loss function: all ML algorithms require the programmer to specify *what is to be learned.* For example, if you were designing the Youtube algorithm, you might want to optimize the time spent on the platform by a user.
- Every ML model has a set of tuneable parameters which govern which types of patterns they are capable of identifying in the data.
- All machine learning seeks to minimize the defined cost function by varying the tuneable parameters.
- Machine learning requires training data to learn the optimal parameters, and testing data to evaluate whether the model can generalize to new samples.
- Understanding the assumptions that each model makes about the inputs it receives and their relationships to the output allows us to interpret the trained model.

Learning a transformation from the input to desired output

Course Objectives

What you’ll learn from this course

Algorithms and mathematical foundations of important supervised learning models:

- Linear models
- Generalized linear models (logistic regression)
- Bayesian linear models
- Regularized regression techniques
- Support vector machines (SVMs)
- Neural networks
- Decision trees, random forests, and boosting algorithms
- Unsupervised techniques: PCA and clustering

What you'll learn from this course

Data processing and preparation for ML models

- Utilize R's powerful data visualization tools to explore datasets, identify relationships between inputs and outputs, and make decisions about further processing steps.
- Clean data and deal with missing or malformed inputs
- Standardize, lump, and normalize data inputs/features to reduce the influence of outliers
- Decide which inputs to use, which to ignore, and which to combine or transform to make new features (PCA, feature selection, non-linear basis functions)

In general, you'll learn what steps to take to prepare datasets for ML modeling. A big component of this is Exploratory Data Analysis (EDA) where you create figures and describe the data to motivate further analysis

What you'll learn from this course

Essential machine learning and data science concepts

- Evaluate how well a model fits its training data (e.g. accuracy, mean squared error)
- Assess how ML models perform on new data (held-out sets, cross-validation, bootstrapping). You'll also understand why it is necessary to test models on unseen data to give a realistic measure of performance
- Bias-variance tradeoff: balancing fit goodness with predictive performance
- Quantifying and understanding uncertainty
- Fitting different functions and shapes to training data, and finding the optimal parameters of those functions with:
    - Maximum likelihood estimation (frequentist statistics)
    - Maximum a posteriori estimation (Bayesian inference)

Course Administration

Course Administration: Canvas

- All material for this course will be available on Canvas
- The module for each week will contain the lecture slides, homework assignments and reading assignments, and additional material to help with math and programming.
- All assignments will be submitted through Canvas. Weekly quizzes and the final exam will also be administered there.
- Class announcements will also be made on Canvas, so make sure to check it regularly

Textbooks and other resources

- Several (free) textbooks are linked to on Canvas. There will be assigned readings each week, and also additional readings that will help your understanding of the coding and ML concepts.

- These textbooks are excellent resources, even beyond this class. Beyond the required texts, dozens of other resources are also provided on Canvas that you should bookmark for future use.

- I will often post samples of code that demonstrate important techniques and concepts from class. Occasionally I will also publish detailed derivations or proofs from class - could optionally share the Latex source.

- I also like to share Youtube videos and other online resources that give exceptionally intuitive or visual explanations of topics from class.

Additional Help

- Prof Skeba’s office hours: TBD (I’ll send a link after class)
    - Office: Sennott Square 5409
- I strongly encourage you to ask for help freely and early. I enjoy holding office hours and find them very helpful for students.

Accommodations

- If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact me and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term.
- DRS will verify your disability and determine reasonable accommodations for this course.
- Accommodations cannot be retroactively applied, please reach out as early as possible.

Collaboration and Cheating

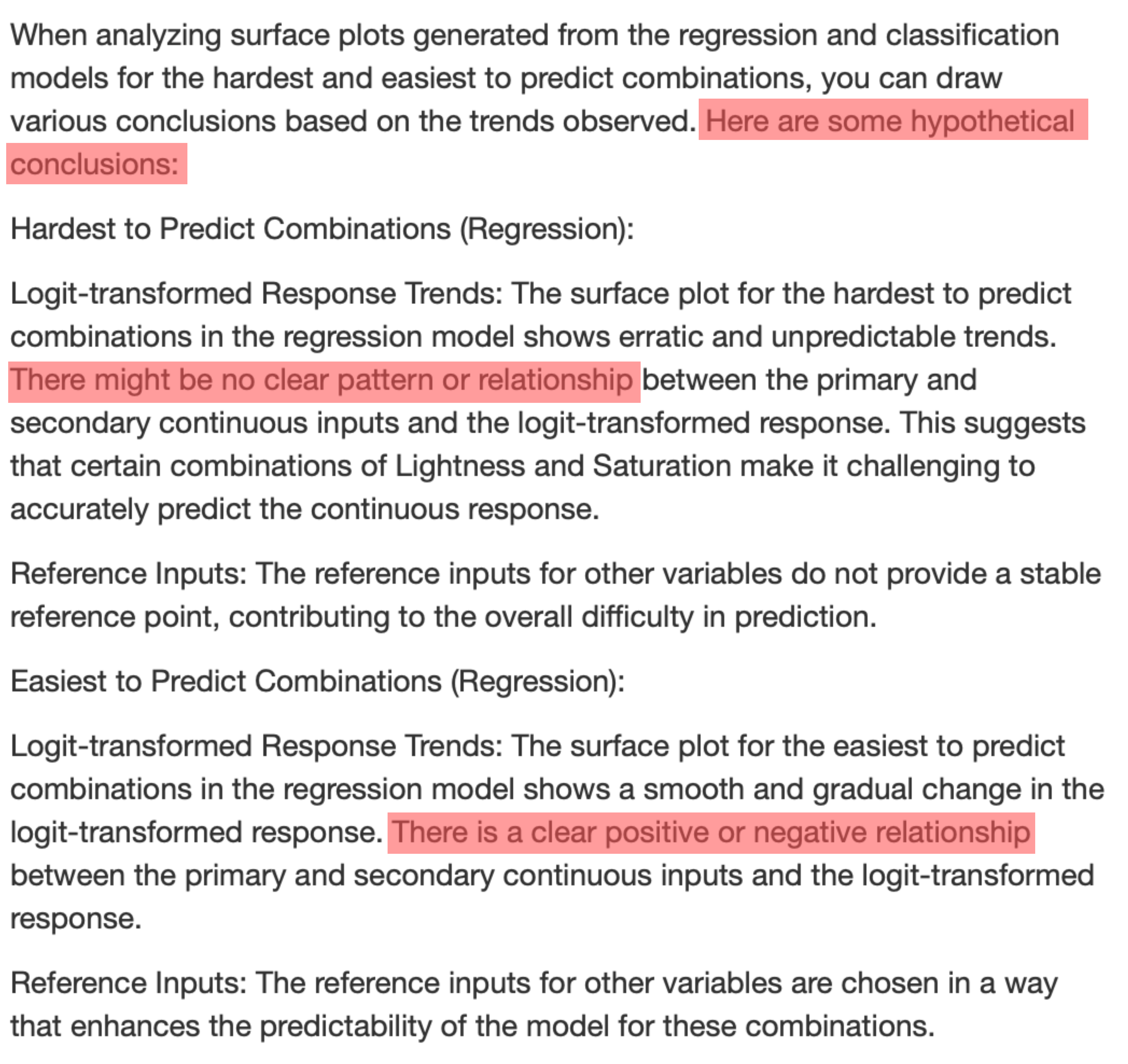
- All work submitted by you **must be your own.**
- Plagiarism or copying somebody else's work will result in a 0 on the assignment for all parties involved. A second violation will force me to give you an F in the course.
- **You may collaborate on homework assignments, and the final project.**
  - Please put the name of your collaborators at the top of your document
  - You should not submit **identical** files for any homework or project. The written answers should express your own thoughts, even if you formed them with the help of someone else.

A note on Chat-GPT

- Chat-GPT is pretty dumb, and there’s a very good chance it gives you the wrong answer to a question.
- In my opinion, such LLMs are hardly useful for coding unless you already have enough knowledge to verify its results.
- I will consider it cheating if you’ve clearly copied AI-generated results and passed it off as your own work. But this isn’t a good idea anyway!

A note on

- Chat-GPT is pr … s you the wrong answer
- In my opinion, … u already have enough k
- I will consider i … sults and passed it off as … y!



When analyzing surface plots generated from the regression and classification models for the hardest and easiest to predict combinations, you can draw various conclusions based on the trends observed. Here are some hypothetical conclusions:

Hardest to Predict Combinations (Regression):

Logit-transformed Response Trends: The surface plot for the hardest to predict combinations in the regression model shows erratic and unpredictable trends. There might be no clear pattern or relationship between the primary and secondary continuous inputs and the logit-transformed response. This suggests that certain combinations of Lightness and Saturation make it challenging to accurately predict the continuous response.

Reference Inputs: The reference inputs for other variables do not provide a stable reference point, contributing to the overall difficulty in prediction.

Easiest to Predict Combinations (Regression):

Logit-transformed Response Trends: The surface plot for the easiest to predict combinations in the regression model shows a smooth and gradual change in the logit-transformed response. There is a clear positive or negative relationship between the primary and secondary continuous inputs and the logit-transformed response.

Reference Inputs: The reference inputs for other variables are chosen in a way that enhances the predictability of the model for these combinations.

Late Submission Policy

- Late penalties will be applied automatically by Canvas
  - Deductions are rolling, so the exact amount will depend on how many hours late your final submission is made.
  - For extraordinary circumstances, please talk to me.
- Homework assignments and projects: 20% per 24 hours after deadline
- Quizzes: flat 50% penalty (max 3 days)

Grading Breakdown

- **Homework Assignments (9): 35% - drop lowest score**
- **Weekly Quizzes (10): 10%**
- **Midterm Exam: 15%**
- **Final Exam: 15%**
- **Final Project: 20%**
- **In-class Participation (Tophat surveys): 5%**

Homework Assignments

- 9 HWs, submit R Markdown file and rendered HTML. Assigned/due on Wednesdays
- Mostly focused on understanding the statistical fundamentals of machine learning
- You will use R to directly implement the mathematical equations and linear algebra expressions that drive ML
- You will also learn how to manipulate and clean data, apply resampling schemes, train and test models, and visualize the models' performance and learned behavior, using popular R packages.
- Homeworks (and posted solutions) also act as tutorials. Not only a big portion of your grade, but the best resource to learn class material

Weekly Quizzes

- The course will be rather fast paced, and each week builds off material from the previous week
- The weekly quizzes act as a short review of important concepts from the week
- Questions will be derived from the lecture material, but the weekly readings will make them easier

Midterm Project

- **Take-home programming project, but no collaboration**
- **This midterm tests your understanding of the concepts, math, and programming required to learn distributions from data.**
- **You are required to perform a mixture of derivations and programming to solve the questions on the exam.**
- **Assigned the 6th week of the semester, and due one week later**

Final Project

- Full report on the analysis of a realistic dataset (TBD)
- Emphasis will be on not only performing the analyses and ML experiments, but also clearly communicating your process and findings.
- More details in weeks to come…

Final Exam

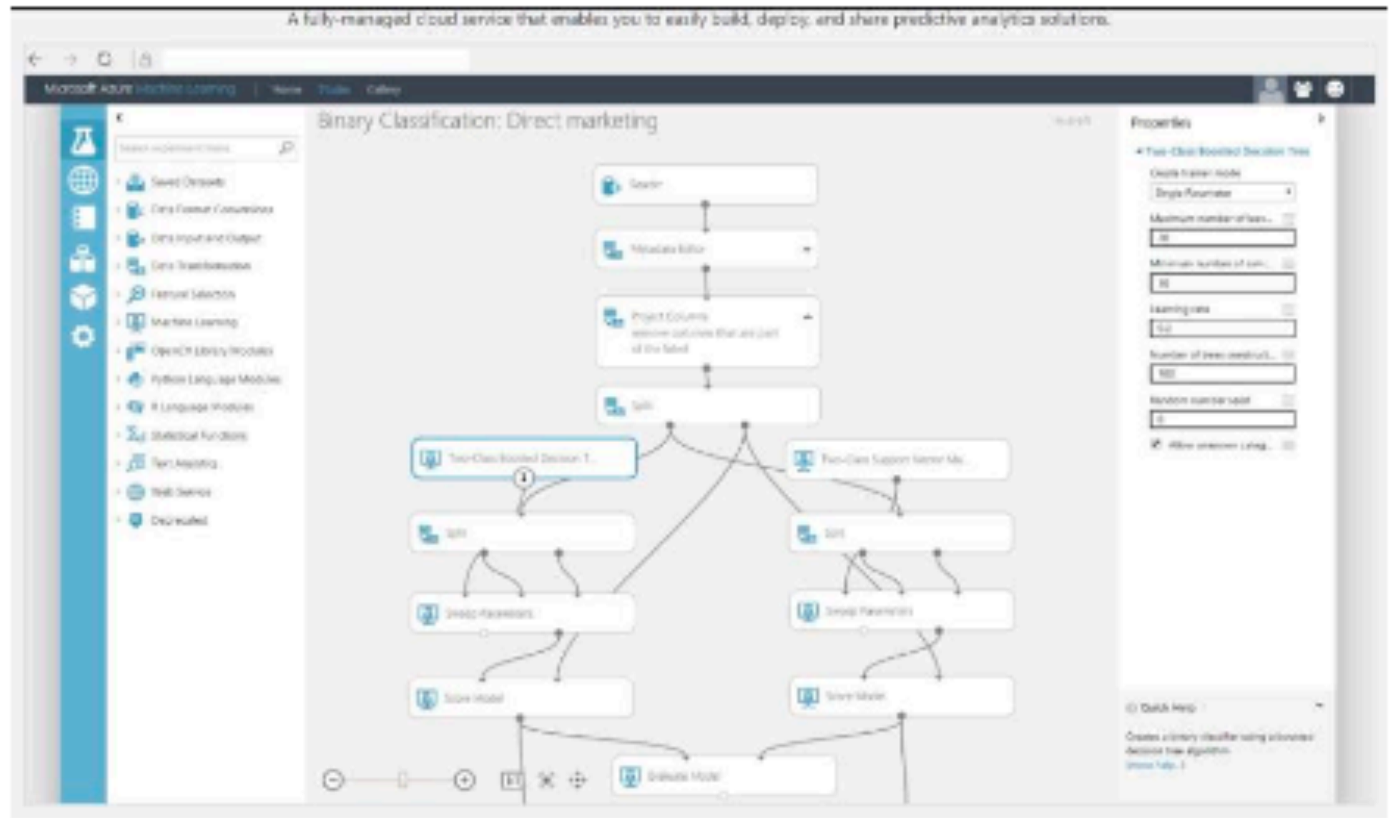
- Conducted online, assigned during the final week of class
- Multiple choice
- Tests major concepts from the semester

Machine Learning in Context

Machine learning is impacting nearly every field in one way or another

Microsoft Azure

A fully-managed cloud service that enables you to easily build, deploy, and share predictive analytics solutions.

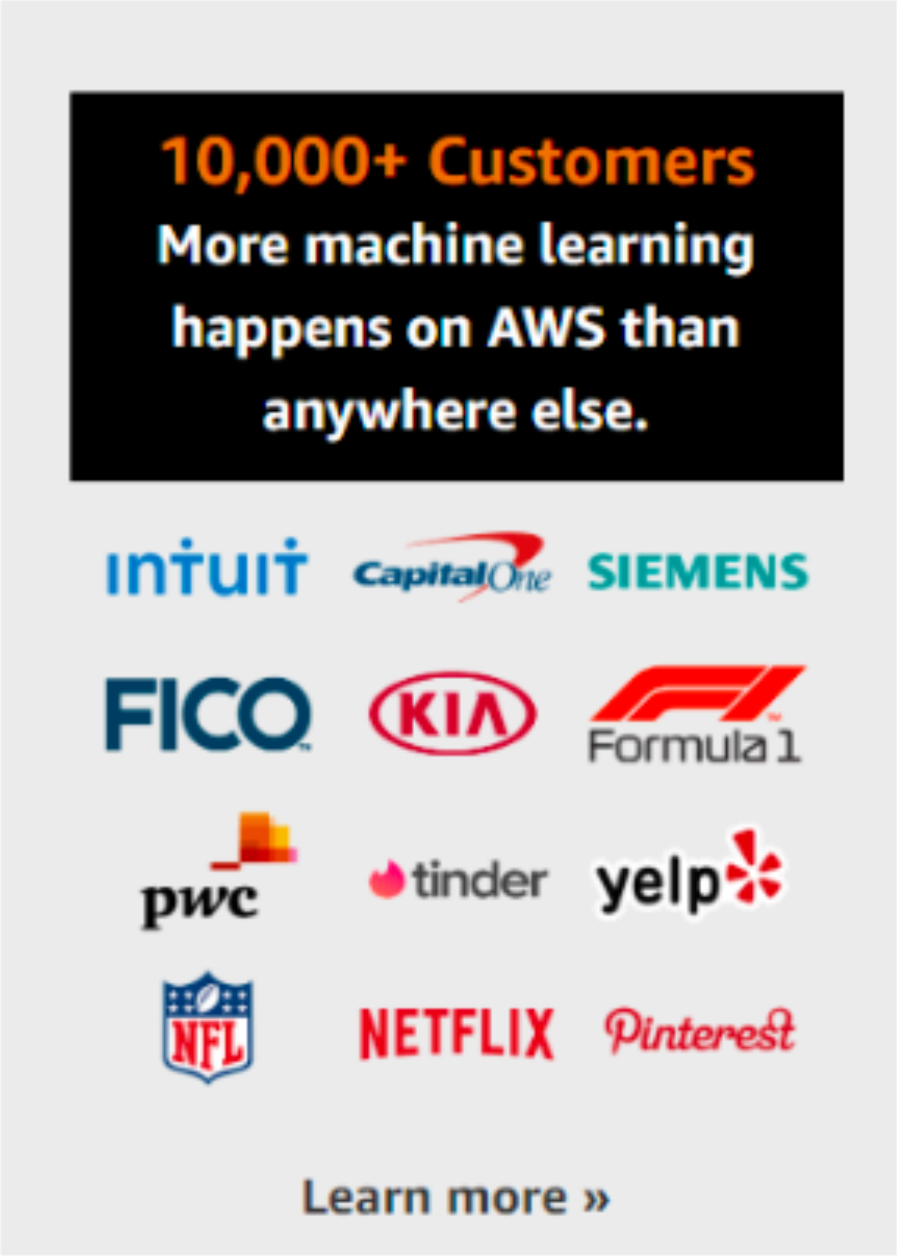


From: https://azure.microsoft.com/en-us/services/machine-learning-studio/

GE Predix

From: https://www.informationweek.com/iot/ge-uses-machine-learning-to-restore-italian-power-plant/d/d-id/1325918

Amazon Web Services (AWS)

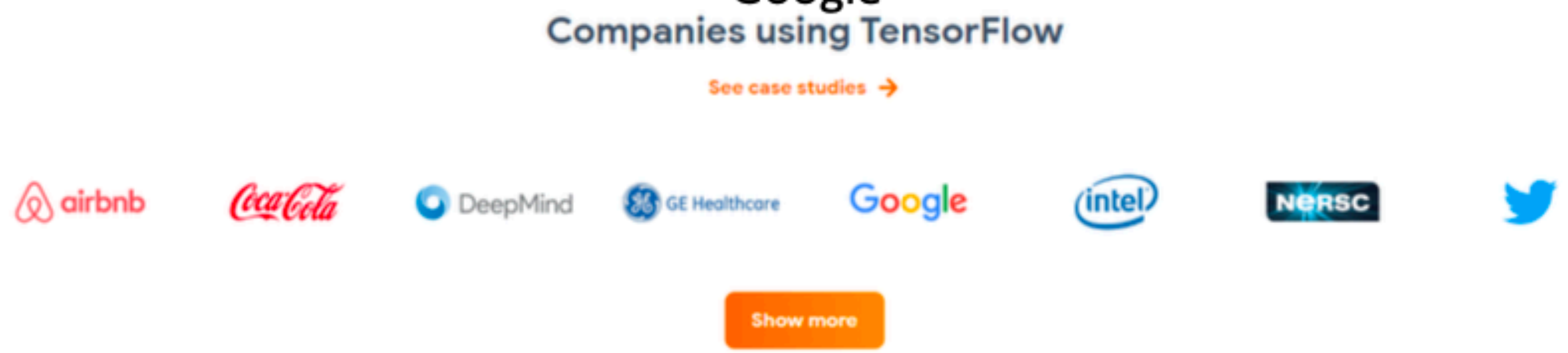


10,000+ Customers
More machine learning happens on AWS than anywhere else.

Learn more »

From: https://aws.amazon.com/machine-learning/

Google



Companies using TensorFlow

See case studies →

Show more

From: https://www.tensorflow.org/

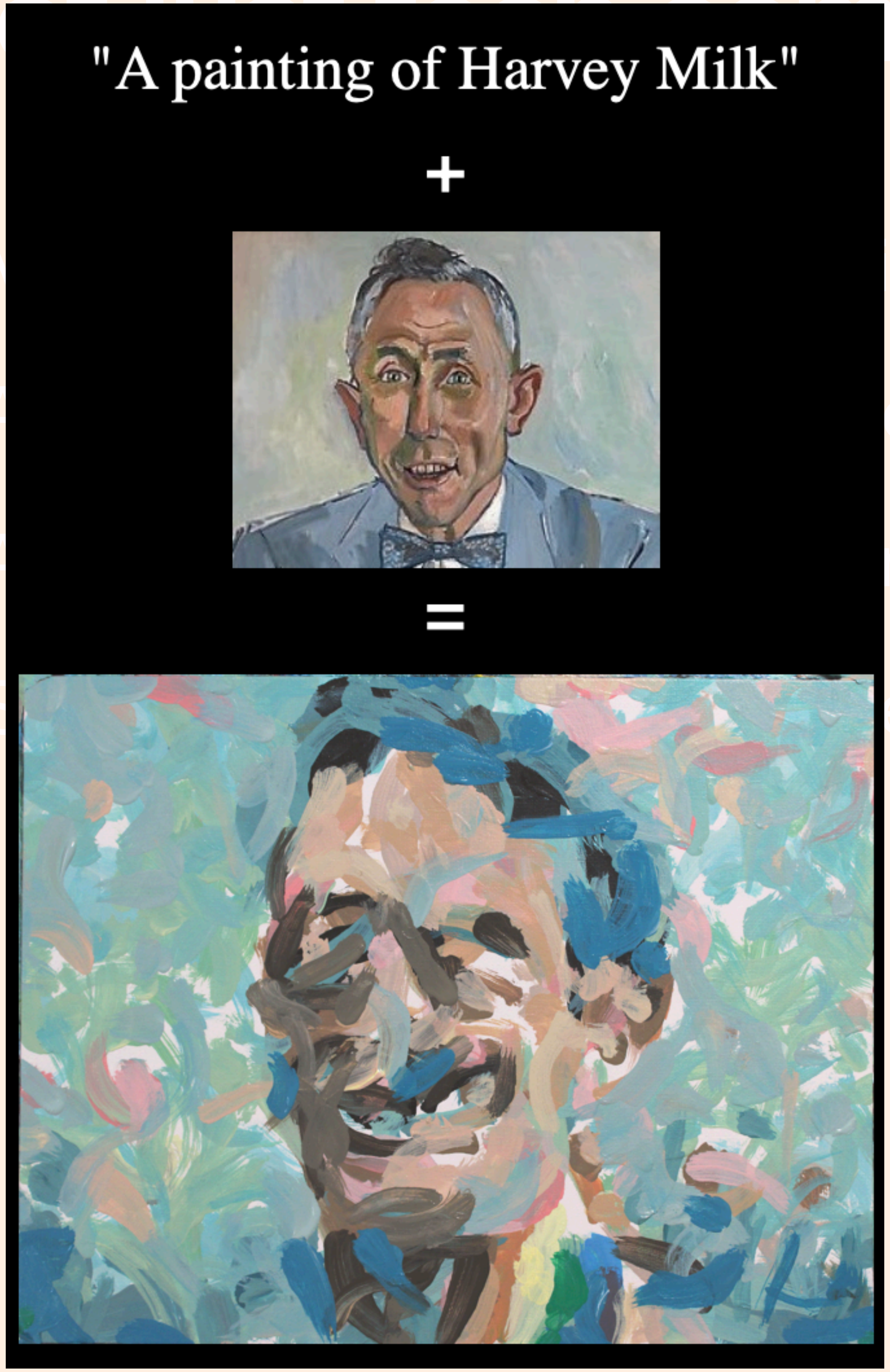
3

Not just tech companies and business either…

- Medical research and practice
- Analyze text corpora for literary and legal studies
- Art and “content” generation
- Social science and public health

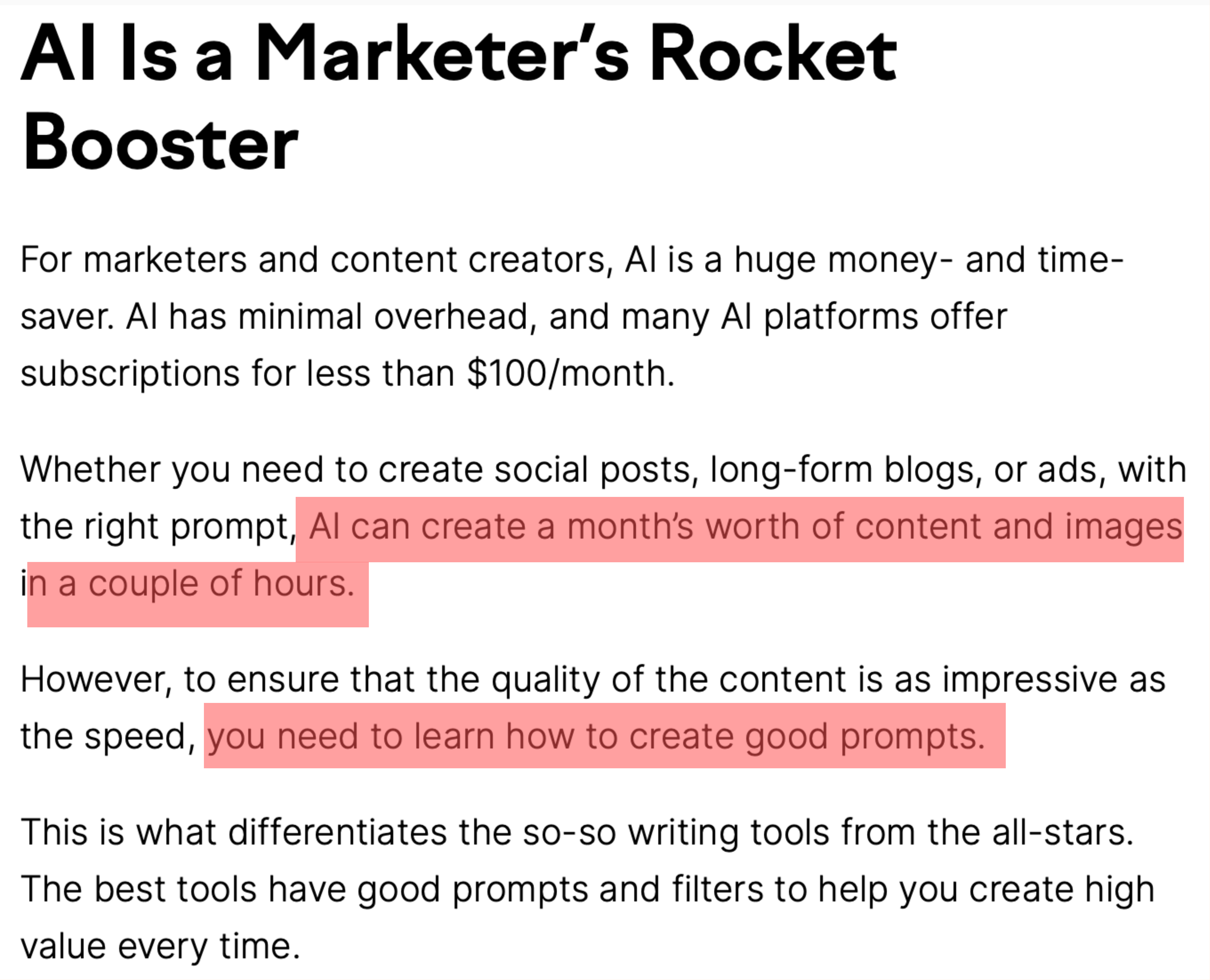
N panies and business either…

- M ce
- A or literary a
- A n
- S ealth



https://pschaldenbrand.github.io/frida/

VS



AI Is a Marketer's Rocket Booster

For marketers and content creators, AI is a huge money- and time-saver. AI has minimal overhead, and many AI platforms offer subscriptions for less than $100/month.

Whether you need to create social posts, long-form blogs, or ads, with the right prompt, AI can create a month's worth of content and images in a couple of hours.

However, to ensure that the quality of the content is as impressive as the speed, you need to learn how to create good prompts.

This is what differentiates the so-so writing tools from the all-stars. The best tools have good prompts and filters to help you create high value every time.

Enshitification (hyperlink)

Why is ML becoming so popular?

What does it promise?

- Uncover hidden patterns in data automatically
- Explain relationships between observed measurements/traits/categories and some outcome (e.g. purchases, clicks, patient health, manufacturing capacity)
- Predict outcomes of unseen samples - e.g. how likely is a new customer to buy an Amazon cordless drill based on other customers “like them”
- Adapt software to its environment and user feedback
- Generative AI: create new output based on some class of input (e.g. chatbot responses, AI art)

Why is ML becoming so popular?

What has enabled it?

- Large-scale, ubiquitous digital surveillance
- Massive datacenters and supercomputers capable of processing all the data created by internet-connected devices
- Rise in popularity of ML and data science programs in universities
- Development of powerful algorithms that effectively process data from a variety of sources

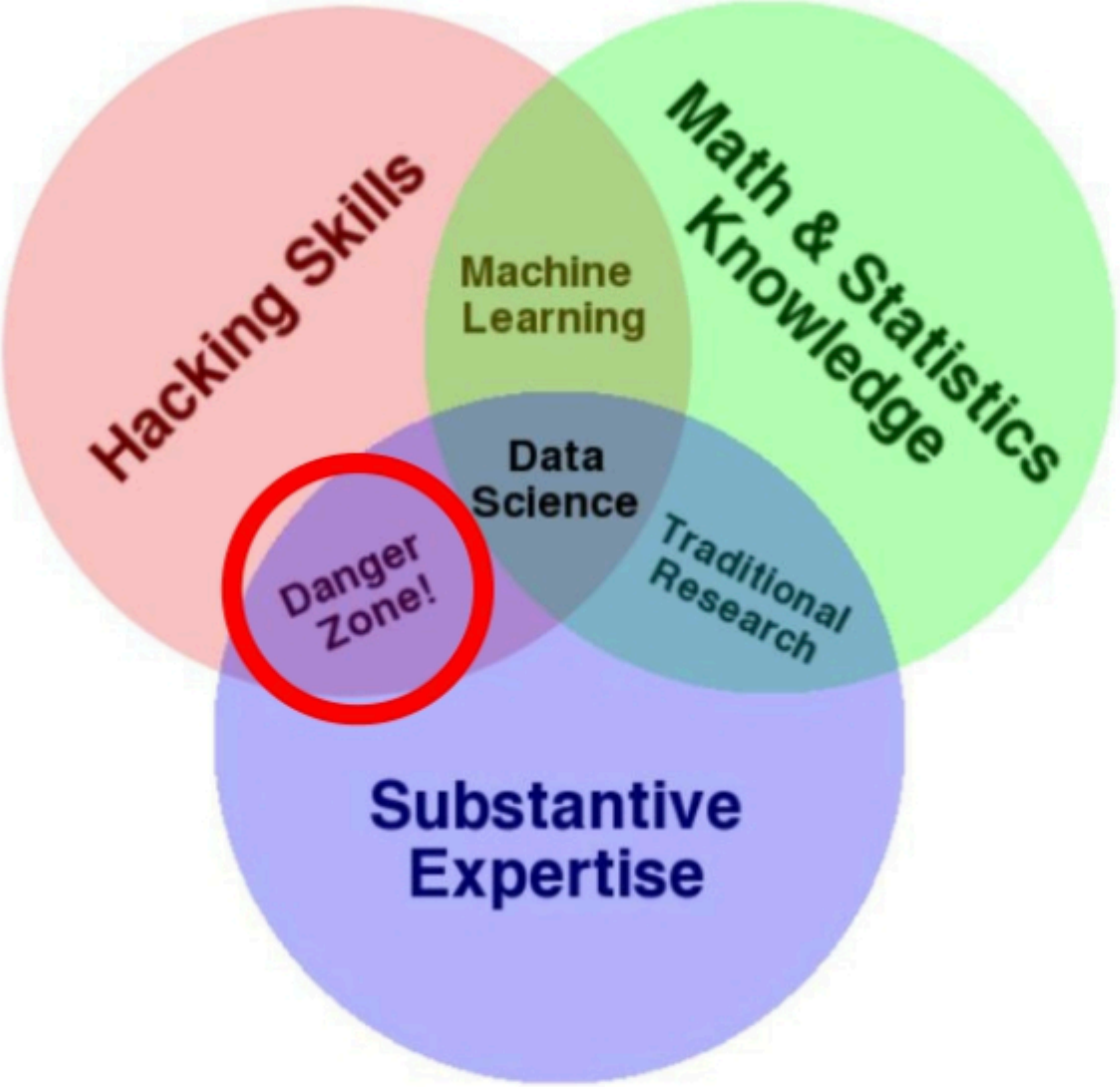
How will we learn ML in this class?

- Programming is an important aspect of machine learning, but not the most important.
- Software, languages, and specific techniques are constantly changing - with the right foundation, you can adapt easily to these
- The most important part of machine learning is understanding the assumptions behind the model.
- Fair amount of mathematics, but I will emphasize the generalizable concepts. I occasionally recommend Youtube videos from channels like 3Blue1Brown and StatsQuest because of their visual, intuitive explanations of math subjects - they’re often much easier to understand than it might seem at first!

If we understand the assumptions, we understand what controls the model behavior

- **Allows us to identify the strengths and weaknesses of various methods**
- **We can understand when a method is NOT appropriate**
- **Reality check: it’s possible to apply ML methods without understanding how they work, and achieve good results. However, when issues arise that break your routine or address novel subjects, strong background knowledge will allow you to troubleshoot and innovate**

Understanding the math and statistics is the best way to avoid the “danger zone”



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

Ultimately, you will be able to answer the following kinds of questions (and more):

- **WHY did it work?**
- **WHY didn’t some other method work?**
- **WHY did you set it up that way?**
- **I tried that, WHY are my results different?**
- **WHY does that matter?**
- **I already knew that, WHY should I care?**

The best way to understand the assumptions is to dive into the math, statistics, and probability

- Weekly quizzes will check your conceptual understanding.
- The weekly homework assignments will assess if you can derive aspects of the statistical fundamentals and manipulate the key equations with linear algebra associated with the methods.
- Weekly textbook readings will greatly strengthen your understanding of the mathematical and programming material.

If you understand the math and algorithms then you have the foundation to work in any programming language (once you know the syntax)!

Homework assignments will show how you can implement the statistical and linear algebra operations programmatically

- This class uses the R statistical programming language because our emphasis is on the statistics.
- Many packages are included to build popular ML models, apply statistical tests, and summarize and visualize data
- You will also learn tidy data principles and and how to create visually appealing and informative figures with ggplot
- Link and instructions on Canvas (Resources and references module) to install R and the Rstudio IDE

Minimum required R libraries for the course

- **Use** `install.packages(“tidyverse”)` **function to download and install:**
  - `tidyverse`
  - `tidymodels`
  - `ggthemes`
  - `caret`
  - `coefplot`
  - `factoextra`

- **Make sure you can create an Markdown file and render an HTML document**
  - **File -> New File -> R Markdown -> HTML**

We will focus on predictive modeling, predictive analytics, supervised learning.

- In supervised learning, we make a distinction between INPUTS and RESPONSES (aka OUTPUTS, TARGETS, or CLASSES)

- Inputs are typically denoted as x (scalar), $\mathbf{x}$ (vector), or $\mathbf{X}$ (matrix)

- Responses are typically denoted as $y\ (\mathbf{y}, \mathbf{Y})$ or $t\ (\mathbf{t}, \mathbf{T})$

The goal in supervised learning is to learn a relationship between the output and the input

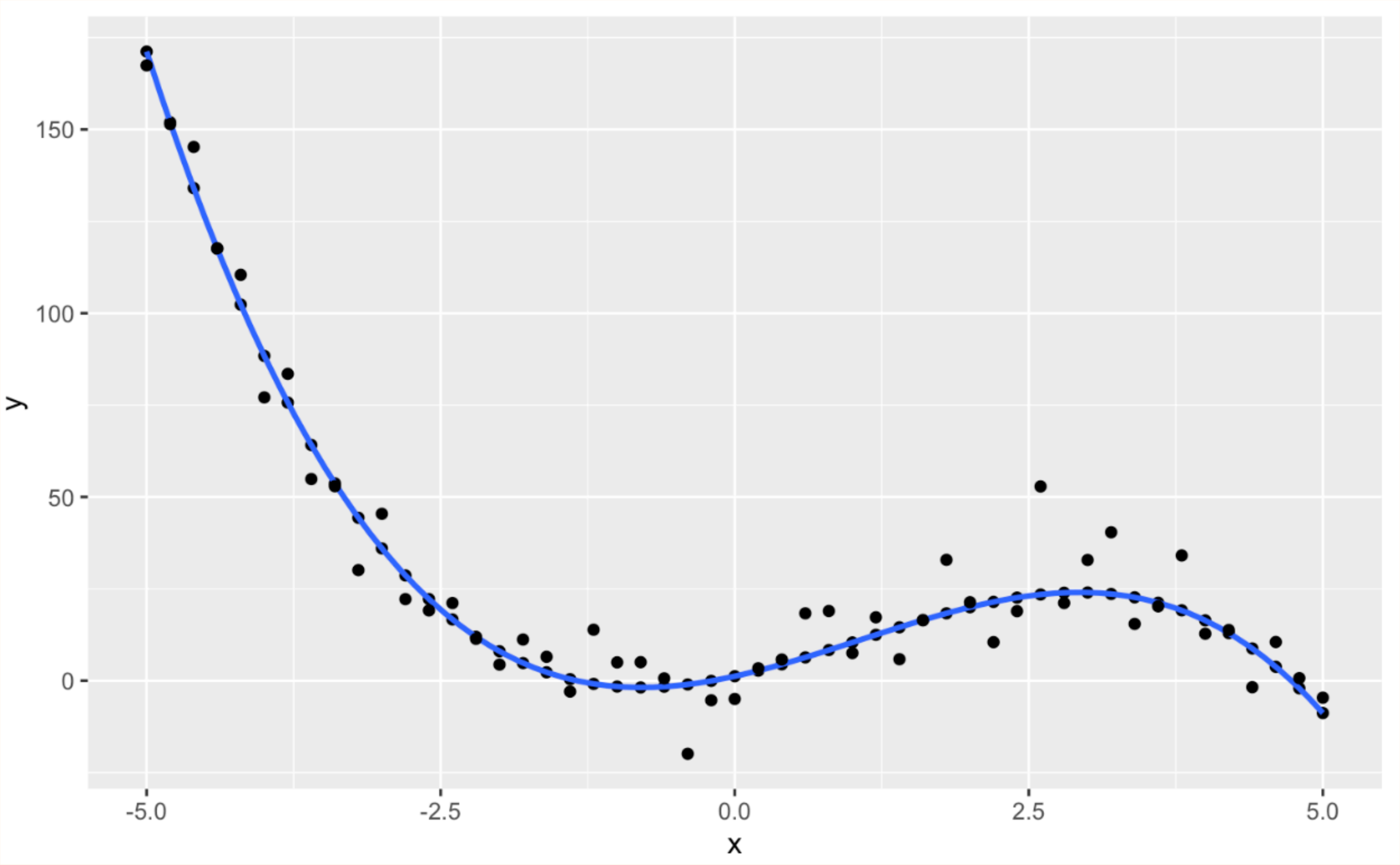
- In other words, we want to find a function that maps each input to its corresponding output. Or, alternatively, find a transformation from input to output.
- Assume that the output, y, can be expressed as some function of the input x. Since machine learning deals with data, and not purely mathematical objects, there always exists some error between the expected value of the function and the observed data itself

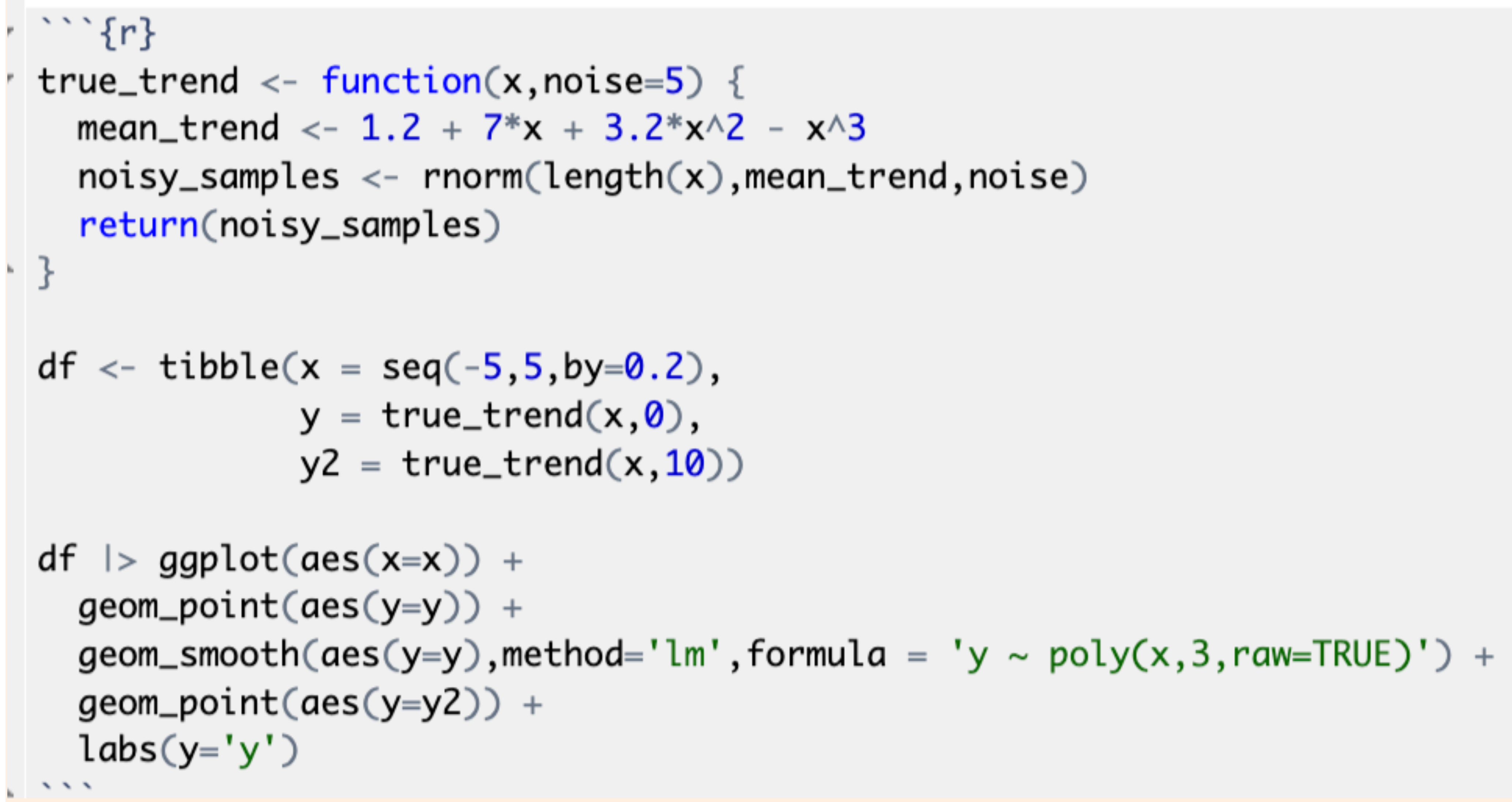
$$y \approx f(\mathbf{x})$$

$$y = f(\mathbf{x}) + \texttt{error}$$

Data is based on measurements, and all measurements have some degree of error

Notice how the observed labels differ from the smooth curve of f(x)





~~~
```{r}
true_trend <- function(x,noise=5) {
  mean_trend <- 1.2 + 7*x + 3.2*x^2 - x^3
  noisy_samples <- rnorm(length(x),mean_trend,noise)
  return(noisy_samples)
}

df <- tibble(x = seq(-5,5,by=0.2),
             y = true_trend(x,0),
             y2 = true_trend(x,10))

df |> ggplot(aes(x=x)) +
  geom_point(aes(y=y)) +
  geom_smooth(aes(y=y),method='lm',formula = 'y ~ poly(x,3,raw=TRUE)') +
  geom_point(aes(y=y2)) +
  labs(y='y')
```
~~~

The approximations allow us to MODEL the real world relationship

- Since all models are approximations, all models have ERROR

$$y \approx f(\mathbf{x}) + \text{error}$$

- All models have parameters or coefficients we need to learn as part of the model *fitting, training,* or *building* process

- We learn those parameters by minimizing the model error (sometimes called the loss function)

- This type of learning is supervised because the outputs guide, or supervise, the learning of the unknown model parameters

Loss functions are related to probability distributions

- We will learn about the probability distributions behind popular loss functions
- Understanding the models **probabilistically** allows us to comprehend more about the data generating process.
- Understanding the data generating process helps us to comprehend what the model THINKS can happen.

Supervised learning is divided into categories based on the response type

- Regression - continuous response
  - Predict the sale price of a house/stock price
  - Predict the temperature next week
  - Predict the expected views on an internet video

- Classification - discrete or categorical response
  - Predict if a mortgage will default or not
  - Predict if Tesla's stock will crash below a threshold
  - Predict which song a Spotify user would like to hear next

Classification can be further divided based on the number of classes, categories, or labels to predict

- Binary classification - 2 possible classes
  - Yes/no, true/false, pass/fail, etc…
- Multiclass classification - 3 or more possible classes
  - Does an image contain a human, horse, fire hydrant, bike, etc…
  - Which team will win the Super Bowl?

- We will primarily focus on binary classification in this class, but we will cover multiclass as well - textbooks will frequently include the multiclass generalizations anyway

Other response types exist too:

- **Counts** - integer counts of an occurrence over ad defined interval of time or space
  - Number of calls to a center per hour
  - Number of goals scored in a soccer match
  - Number of COVID-19 cases identified per day

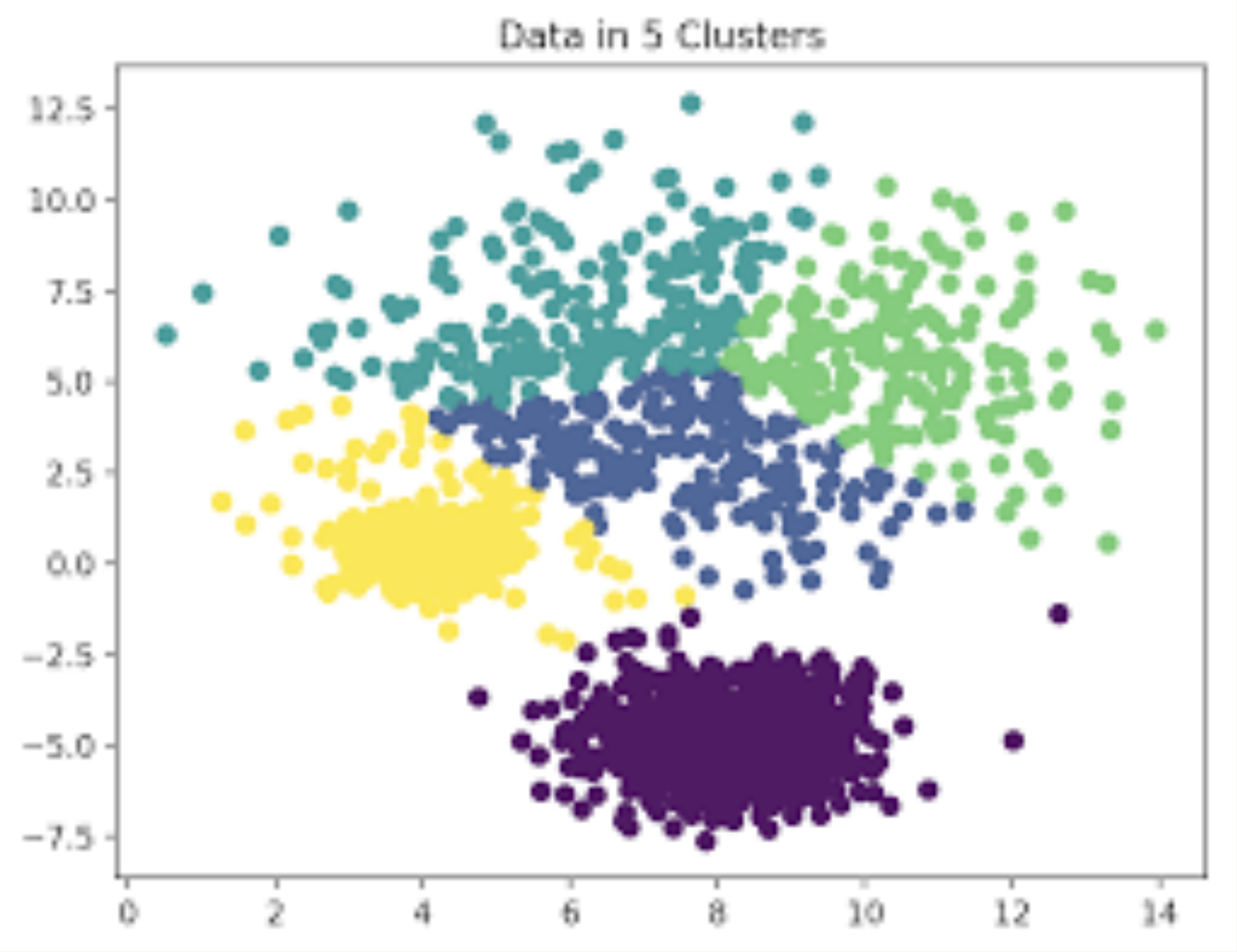
- **Hazard/survival/reliability analysis** - probability of surviving given survival up to that point in time
  - Very important in the insurance industry, engineering and manufacturing, and medical testing

Other types of learning besides supervised:

- Unsupervised learning
- Semi-supervised learning
- Transfer learning
- Deep learning
- Reinforcement learning

Unsupervised learning or “Data discovery”

- Observe variables without distinction between inputs and responses. That is, plot the inputs without regard for their corresponding label (if one even exists)
- Goal: identify interesting patterns in the data
  - Find relationships between variables
  - Find relationships between observations
- Useful in high-dimensional situations
- We will use cluster analysis to discover patterns

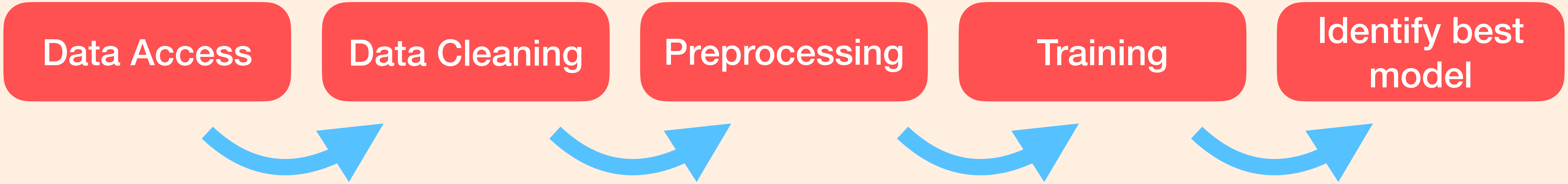


Machine learning pipeline

The machine learning workflow describes the process from collection of data, to presentation of a final model

- The Airfoil noise example on Canvas depicts the multiple stages in the pipeline

- Though it's often depicted in a linear fashion, you will often iterate between steps in the process of data analysis to ultimately decide on the best model.



Data Access

Data Cleaning

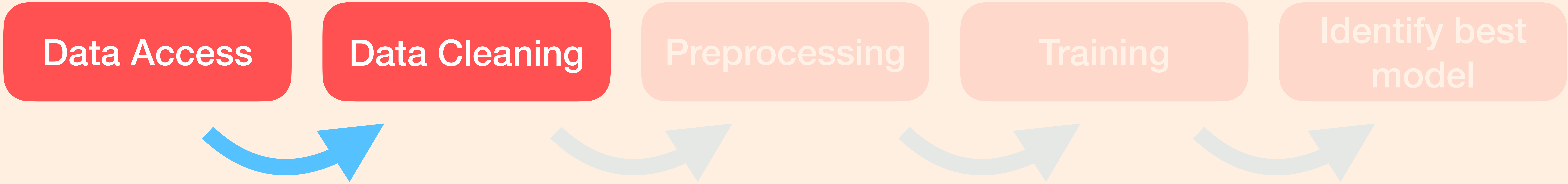
Preprocessing

Training

Identify best model

In the real world, a significant amount of time (perhaps the majority) will be spent collecting data and wrangling it into a form that is convenient for machine learning

- **Possibly 60-80% of a data project could be spent sourcing data, gathering it together, and cleaning it before you can even train a ML model.**
- **The data used to train ML models is more important than the particular algorithm being applied. It can be difficult to decide which data you need to collect, and to navigate institutional and ethical challenges related to that collection.**
- **Real data is messy, incomplete, and scattered across multiple sources.**



Data Access

Data Cleaning

Preprocessing

Training

Identify best model

…in this class, we'll be using mostly clean, convenient data to focus on the ML concepts

Data is typically stored across multiple data sources, and are not organized in a convenient manner

Key	A	B
1	…	…
2	…	…
3	…	…
4	…	…
…		

Key	C	D
1	…	…
2	…	…
3	…	…
…		

Key	E	F	G
1	…	…	…
2	…	…	…
3	…	…	…
…			

- When data is initially collected, we refer to it as “raw data”
- We will refer to obtaining the “raw data” as the data access or query step .
- Need Exploratory Data Analysis (EDA) to understand the various data sources.

Data sources need to be merged together using data contextualization

Key	A	B	C	D	E	F	G
1	…	…	…	…	…	…	…
2	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…

- Aggregate them together – identify the common “keys” or unique identifiers shared across the data sources.
- Align them to the same basis – rows must represent the same thing.
- Contextualized data is **model or analytics-ready**

The airfoil example works with inputs and responses stored together in rectangular (flat) structure

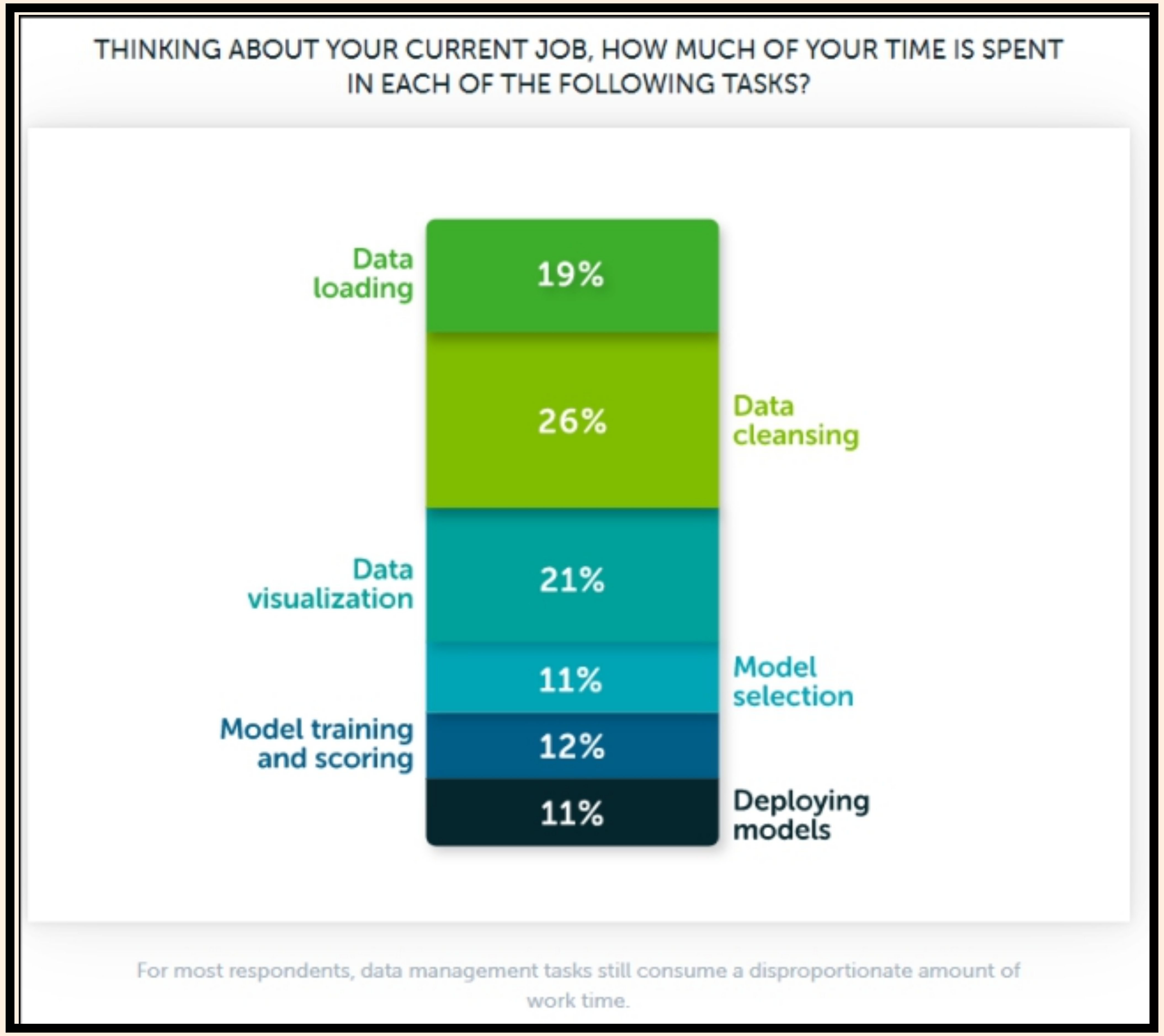
	Input 1	Input 2	Response 1	Response 2
Observation 1	3.5	green	43	TRUE
Observation 2	5.6	green	57	FALSE
Observation 3	1.2	yellow	4	FALSE
Observation 4	13.8	blue	129	FALSE
...				

- Each row represents **one** observation, or sample of the data. Each column is a separate input feature, or response value
- Both inputs and responses are stored in each line
- This is the ideal data format - easy to read and manipulate

The contextualized data can then be cleaned

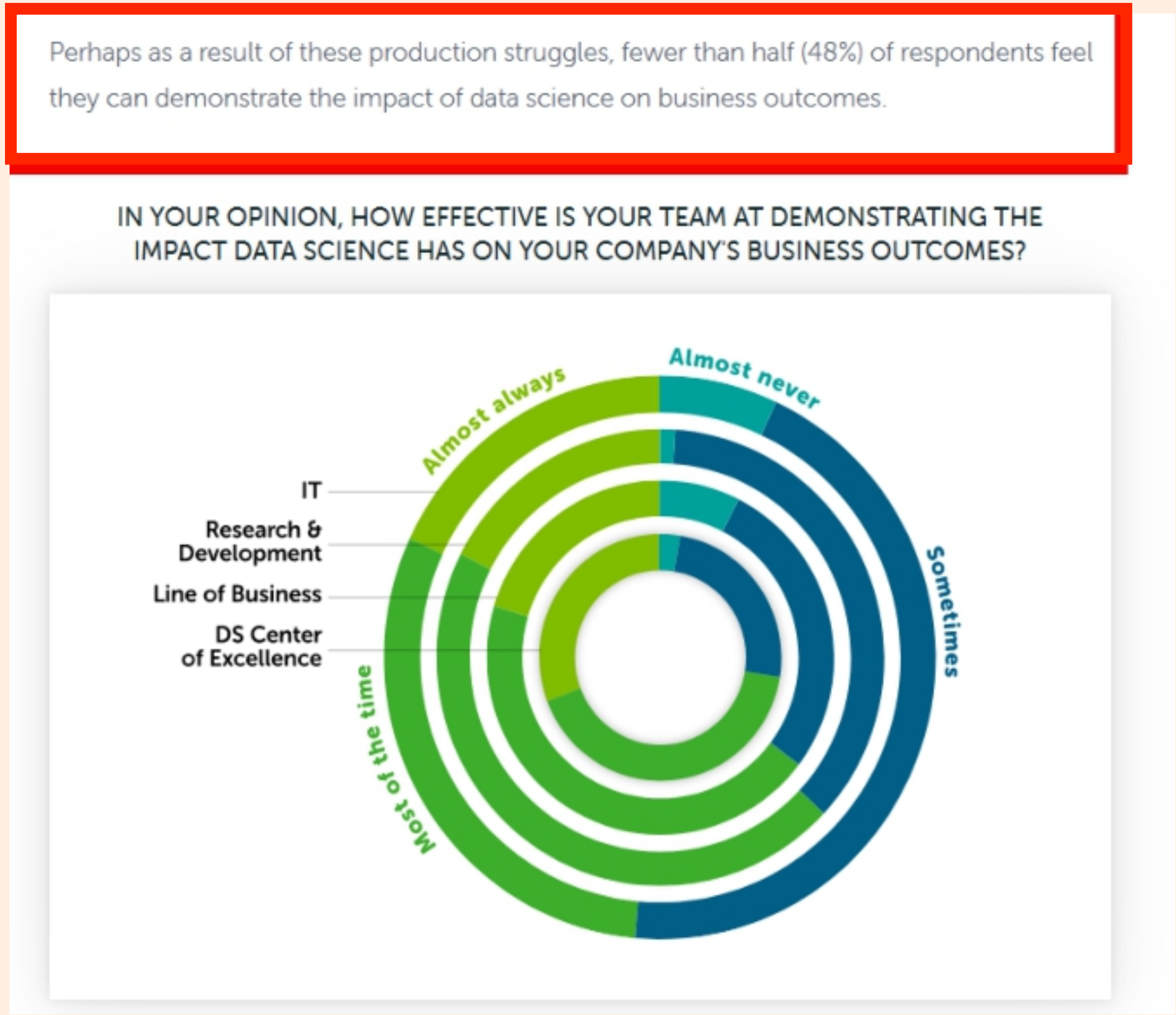
- Remove duplicate rows
  - One row represents one observation, so we don’t want to “double count”.
  - Benefit of unique identifiers or “keys” to help prevent/identify repeat rows.
- Remove incorrect values
  - Sensor errors
  - Human entry errors
- Missing values
  - Most methods require missing values to be removed.
  - We will discuss missing values in more detail later in the semester.
  - But if you have a lot of missing values…things will be challenging!

Data scientists agree!



THINKING ABOUT YOUR CURRENT JOB, HOW MUCH OF YOUR TIME IS SPENT IN EACH OF THE FOLLOWING TASKS?

For most respondents, data management tasks still consume a disproportionate amount of work time.



Perhaps as a result of these production struggles, fewer than half (48%) of respondents feel they can demonstrate the impact of data science on business outcomes.

IN YOUR OPINION, HOW EFFECTIVE IS YOUR TEAM AT DEMONSTRATING THE IMPACT DATA SCIENCE HAS ON YOUR COMPANY'S BUSINESS OUTCOMES?

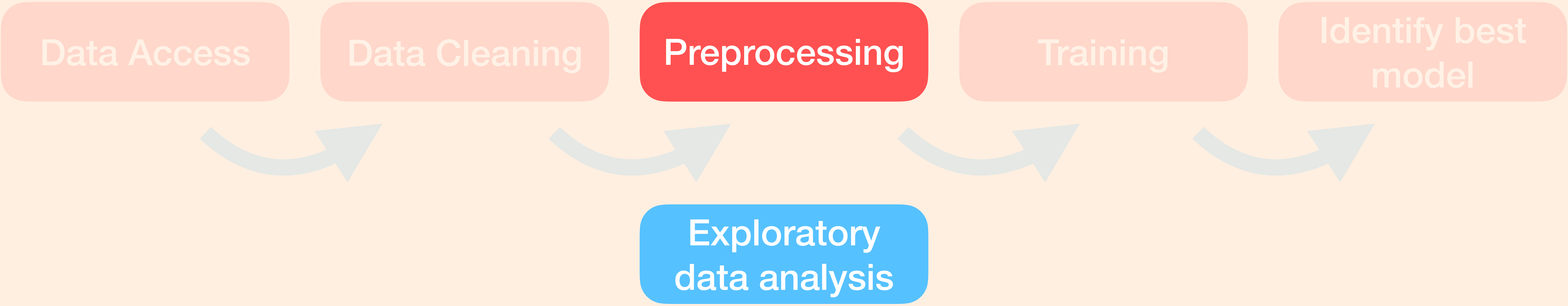
From: https://www.anaconda.com/state -of-data -science-2020?utm_medium=resources&utm_source=anaconda&utm_campaign=sods - 2020&utm_content= report

Once the dataset has been assembled and cleaned, you must explore the data and process the inputs and outputs to prepare them for

- **Correlations between features, outlier values, and other abnormalities can cause problems for certain algorithms**

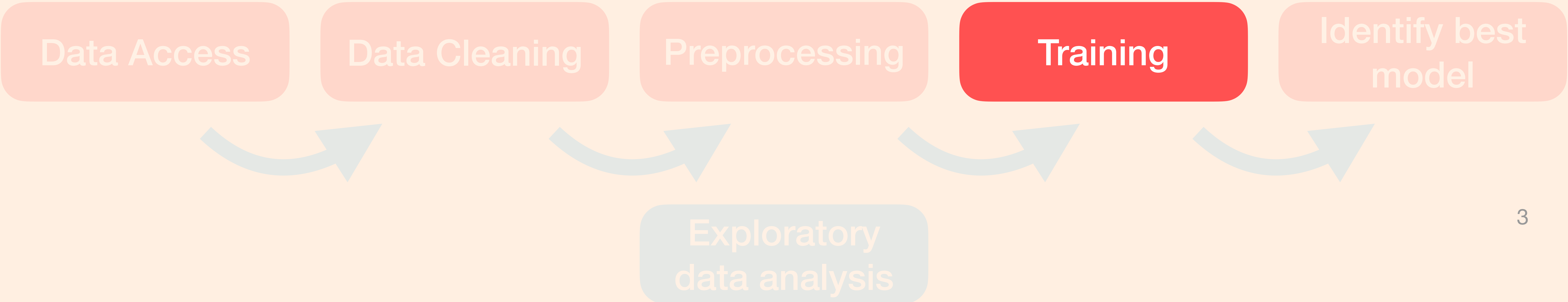
- **At this point, it's important to begin visualizing and summarizing the data to determine what preprocessing is necessary**

- **Use the visualizations to decide which models and feature representations might best model the relationship between input and output**



Train models

- At this stage, the preprocessed training data is passed into the machine learning algorithm, which then finds the optimal set of parameters according to the objective function.
- This is the flashy and exciting part of ML to many people!
- Many different procedures to learn a relationship between input and output (or between inputs), each with their own assumptions and set of available hypotheses.
- From the EDA, you can decide which models seem appropriate. It's often prudent and fairly easy to test out a number of different models



Data Access

Data Cleaning

Preprocessing

Training

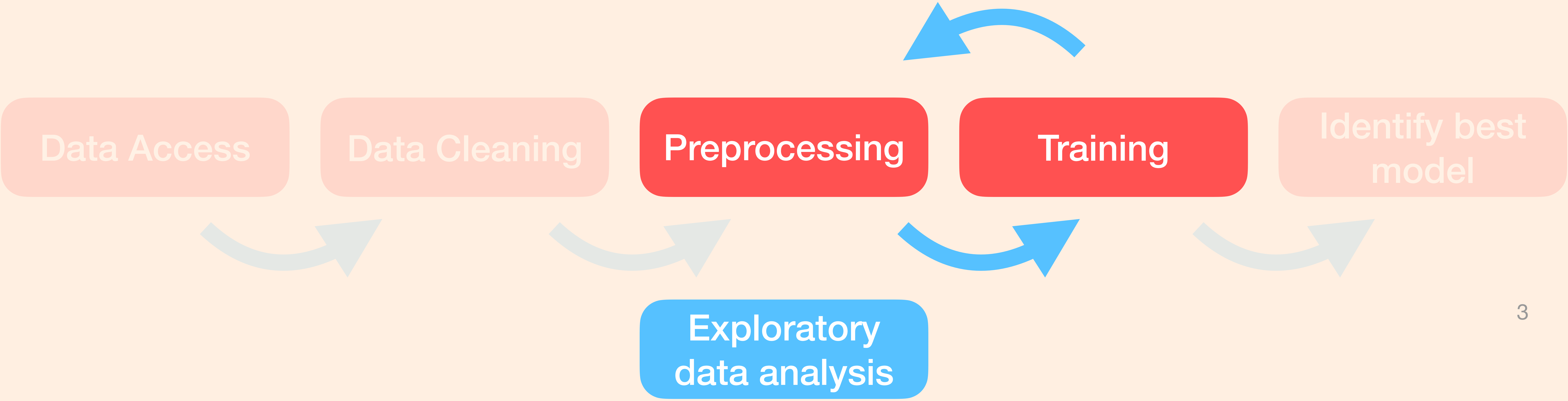
Identify best model

Exploratory data analysis

Preprocessing, data exploration, and model training should be iterative and adaptive

- Once you’ve trained some models and evaluated their performance, you can and should consider whether other models or pre-processing steps might lead to improvements

- Different models have different preprocessing requirements, and these can affect the results

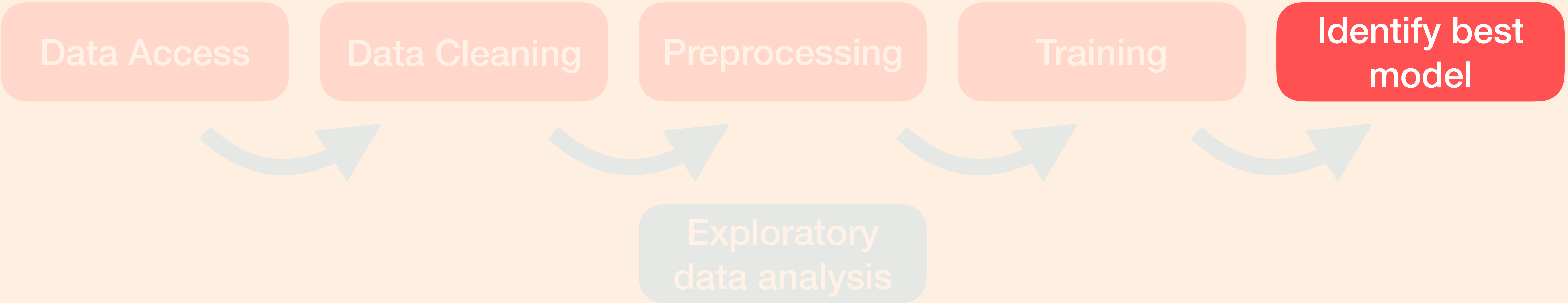


Model evaluation is extremely important and requires consideration of multiple factors

- Goodness-of-fit (that is, training performance) is important, but can be deceiving. It’s possible to overfit the training data, which leads to poor performance when predicting new data.

- The risk of overfitting is related to the complexity of the ML model and its number of tuneable parameters (degrees of freedom), as well as the size and distribution of the dataset.

- A significant amount of this course will deal with addressing the bias-variance tradeoff. We will discuss models that directly control for overfitting, as well as specialized performance metrics and train/test procedures that account for this risk.



Data Access

Data Cleaning

Preprocessing

Training

Identify best model

Exploratory data analysis

Truthfully, the majority of ML projects are not successful

- Even though AI/ML is receiving a lot of attention, in reality around 60-80% of “big data” projects fail
- Only about 10% of proof-of-concept projects make it into production
- Here are some references:
  - The ’20/80 Rule of Big Data’ has huge implications for IoT tech
  - Why do 87% of data science projects never make it into production?
- You should not be discouraged though! Many tasks and types of data are just not amenable to ML analysis, or shortcuts were taken during development that you’ll learn to avoid

Then why is this class so focused on the math?

- As a data scientist, your value comes from applying the models to various use cases, and/or finding insights about key features and factors driving behavior
- In research (and ideally in business), data acquisition is an important and involved process that can lead to findings all on its own
- With time and (interdisciplinary) experience, you'll learn the best practices for a given application. No two data science projects are the same, and you need domain knowledge to be truly effective
- The math, though, is always the same and enables you to select the best models for an application and understand your findings
- Whether for research, business, activism, etc, you must be able to clearly present and explain your data analysis.

The course is divided into 4 primary areas

Applied machine learning

Distribution fitting

Supervised learning deep dive

Unsupervised learning

The course is divided into 4 primary areas

Applied machine learning

Distribution fitting

Supervised learning deep dive

Unsupervised learning

- Performance metrics
- Overfitting
- Resampling methods
- Model selection

The course is divided into 4 primary areas

Applied machine learning

Distribution fitting

- Descriptive statistics
- Likelihood functions
- Optimization
- Bayesian statistics

Supervised learning deep dive

Unsupervised learning

The course is divided into 4 primary areas

Applied machine learning

Distribution fitting

Supervised learning deep dive

Derive how models are fit and how their assumptions impact performance

Unsupervised learning

The course is divided into 4 primary areas

Applied machine learning

Distribution fitting

Supervised learning deep dive

Unsupervised learning

- K Means
- Hierarchical clustering
- PCA