INFSCI 2595

Fall 2019

ROOM LOCATION

Lecture 01

Administrative items

- Instructor
  - Dr. Joseph P. Yurko
  - jyurko@pitt.edu
  - Office: 5419 Sennott Square
  - Office hours: Tuesday 11:00AM – 12:30PM & 2:30PM – 4:00PM

- TA
  - Leila Karimi, PhD Candidate
  - leila.karimi@pitt.edu
  - Office: LERSAIS Lab (Room 410) Information Science Building
  - Office hours: Mondays 2:00PM – 4:00PM

Important!!

- INFSCI 2595 has merged with CS 1675.
- Themes and goals of the course are the same.
- Content has been changed relative to previous years.

CONGRATULATIONS!!

You are my first course at Pitt!!

My background

- Dr. Joseph P. Yurko – School of Computing and Information

- Education
  - MIT – Aerospace Engineering, SB (2008)
  - MIT – Nuclear Engineering, SM (2010)
  - MIT – Nuclear Engineering, PhD (2014)

- Work experience
  - FPoliSolutions, LLC – consultant in the nuclear industry
  - Alcoa/Arconic – data scientist in the manufacturing industry

My background

- Dr. Joseph P. Yurko – School of Computing and Information

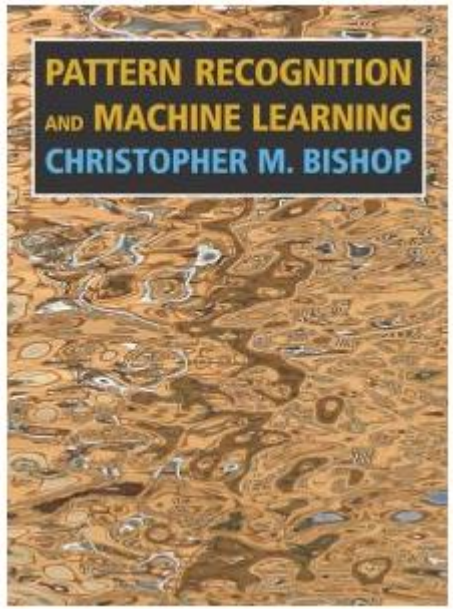
- Education
  - MIT – Aerospace Engineering, SB (2008)
  - MIT – Nuclear Engineering, SM (2010)
  - MIT – Nuclear Engineering, PhD (2014)

- Work experience
  - FPoliSolutions, LLC – consultant in the nuclear industry
  - Alcoa/Arconic – data scientist in the manufacturing industry

I transitioned from being an engineer to a **full-time data scientist**.

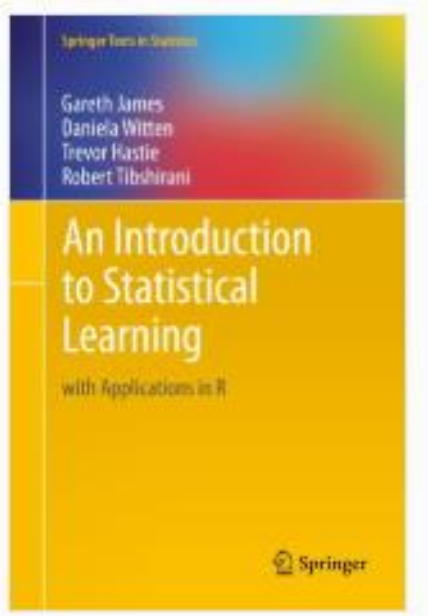
Began with wanting to apply **machine learning** methods to an engineering problem!

Text books - required



Available as a free download from Microsoft:

https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book



Free on the Pitt network via Springerlink:

https://link.springer.com/book/10.1007/978-1-4614-7138-7

Other useful books – free books in red

- Machine Learning: a Probabilistic Perspective – Kevin Murphy
  - https://www.cs.ubc.ca/~murphyk/MLbook/
- Applied Predictive Modeling – Max Kuhn & Kjell Johnson
  - https://link.springer.com/book/10.1007/978-1-4614-6849-3
  - `caret` package: http://topepo.github.io/caret/index.html
- R for Data Science – Hadley Wickham & Garret Grolemund
  - https://r4ds.had.co.nz/
- Elements of Statistical Learning
  - https://link.springer.com/book/10.1007/978-0-387-21606-5
- R for Everyone – Jared Lander
  - https://www.jaredlander.com/r-for-everyone/
- Bayesian Computation With R – Jim Albert
  - https://link.springer.com/book/10.1007/978-0-387-92298-0
- Introduction to Econometrics With R
  - https://www.econometrics-with-r.org/index.html
- Statistical Rethinking: A Bayesian Course With Examples in R and Stan – Richard McElreath
  - https://xcelab.net/rm/statistical-rethinking/

Software tools and programming languages

- Machine learning methods are available in many software tools
  - Microsoft Azure, SAS JMP, Matlab, Stata, and many more

- Great libraries and packages exist in a wide variety of programming languages
  - Python, Matlab, R, C/C++, Java, Julia and many other languages

- **In this course, we will focus on R.**

Grading

- Homework – 45%
  - 13 Homework assignments, **DROP two lowest grades**

- Exams – 45%
  - Midterm: 75 minutes in lecture, week of October 7.
  - Final: TBD
  - Both weighted equally

- Short quizzes in lecture – 10%

Homework submission

- Available each Wednesday AM – **due on FRIDAYS at 12PM NOON**
- We will use **R Markdown** to automatically render reproducible reports which include code, discussion, and figures.
  - See https://rmarkdown.rstudio.com/ for an overview.
- Assignments will be posted on CourseWeb and must be submitted via CourseWeb.
- If you want to also upload handwritten work:
  - Use an app on your smartphone to convert a picture to a PDF, upload the PDF.
  - If you do not have a smartphone, contact me ASAP to discuss options.

Collaboration and Cheating

- Your work must be your own.
- **Collaboration with other students on homework is allowed.**
- Include the names of the student(s) you worked with on your submitted homework assignments.

Collaboration and Cheating

- Copying or plagiarism are **NOT** allowed.
  - Penalty applied to **ALL** parties involved – the copier and the copied-from.
  - First offense: 0 on the assignment.
  - Second offense: F for the course.
- Late submissions penalized 20% for every day after the due date.
  - If you require an extension, contact me **BEFORE** the due date to discuss.
  - Examples: family emergency, conference presentation

Students with disabilities

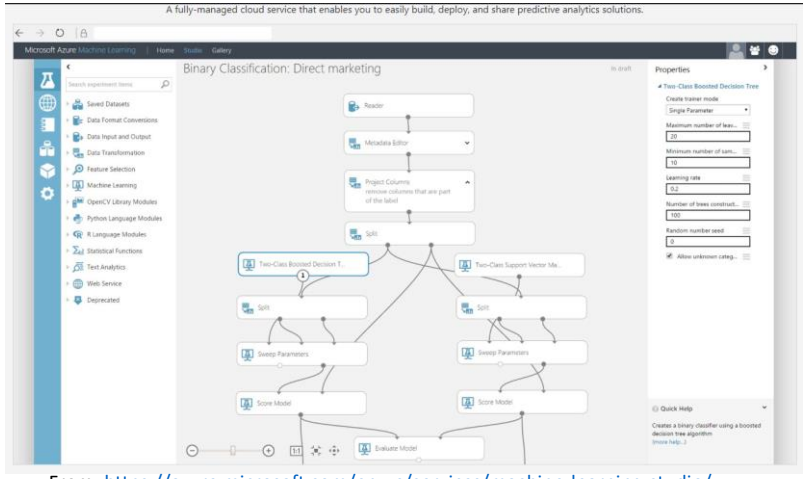
- If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact me and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, **as early as possible in the term**.

- DRS will verify your disability and determine reasonable accommodations for this course.

Why machine learning?

It is EVERYWHERE!!!

Microsoft Azure



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

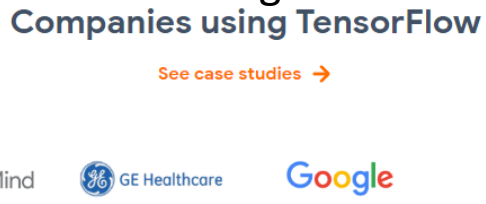
GE Predix

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

Amazon Web Services (AWS)

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

Google



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

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class on:

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class on:
  - Probability and statistics

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class on:
  - Probability and statistics
  - Linear Algebra

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class which covered:

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class which covered:
  - Ordinary Least Squares (OLS)

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class which covered:
  - Ordinary Least Squares (OLS)
  - Binomial distribution

Today, it's easy to say **why**...so let's focus on *how*...

- Ask yourself the following questions, no need to answer out loud.

- Have you taken a class which covered:
  - Ordinary Least Squares (OLS)
  - Binomial distribution
  - Matrix factorization/decomposition

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?

Today, it's easy to say **why**…so let's focus on *how*…

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?
  - Overfitting

Today, it's easy to say **why**...so let's focus on *how*...

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?
  - Overfitting
  - Resampling

Today, it's easy to say **why**...so let's focus on *how*...

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?
  - Overfitting
  - Resampling
  - Neural networks

Today, it's easy to say **why**...so let's focus on *how*...

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?
  - Overfitting
  - Resampling
  - Neural networks
  - Hyperparameters

Today, it's easy to say **why**...so let's focus on *how*...

- Ask yourself the following questions, no need to answer out loud.

- How familiar are you with the following topics?
  - Overfitting
  - Resampling
  - Neural networks
  - Hyperparameters
  - Bayesian statistics

Today, it's easy to say why machine learning…but the important questions are…

Why should you learn the math?

Why should you learn the details?

We go into the math behind the methods in order to:

- Understand their **assumptions** because assumptions control their behavior.
- Identify their strengths and weaknesses.
- Understand when a method is **NOT** appropriate.

We go into the math behind the methods in order to:

- Understand their **assumptions** because assumptions control their behavior.
- Identify their strengths and weaknesses.
- Understand when a method is **NOT** appropriate.

REALITY CHECK

Most of the time, you can just apply methods without understanding how they work.

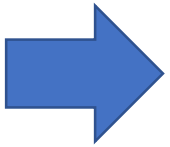
The concern is not for those situations when it's clear they're not working.

The concern is for when it is NOT obvious the methods are struggling.

Course overview

Our roadmap – the course is broken up into two primary pieces

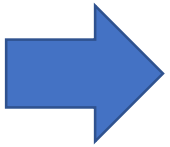
FOUNDATIONS



APPLICATIONS

Our roadmap – each piece can be subdivided into:

FOUNDATIONS
Descriptive
Model/Explain



APPLICATIONS
Predictive

Our roadmap – each piece can be subdivided into:

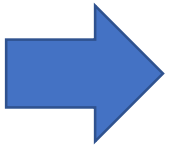
FOUNDATIONS

Descriptive

- Distribution fitting

Model/Explain

- linear models (lm)
- generalized linear models (glm)



APPLICATIONS

Predictive

- Support Vector Machines
- Neural Networks
- Bayesian Belief Networks
- Decision trees
- Ensemble methods

By focusing on being *predictive*, we will be focusing on **SUPERVISED LEARNING** methods

- In supervised learning, we observe INPUTS and RESPONSES
- Inputs are typically denoted as x (scalar), $\mathbf{x}$ (vector), and $\mathbf{X}$ (matrix)
- Responses or Outputs are typically denoted as y ($\mathbf{y}$, $\mathbf{Y}$) or t ($\mathbf{t}$)

By focusing on being *predictive*, we will be focusing on **SUPERVISED LEARNING** methods

- Our goal in supervised learning is to learn the mapping (or relationship) between the output and the inputs

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

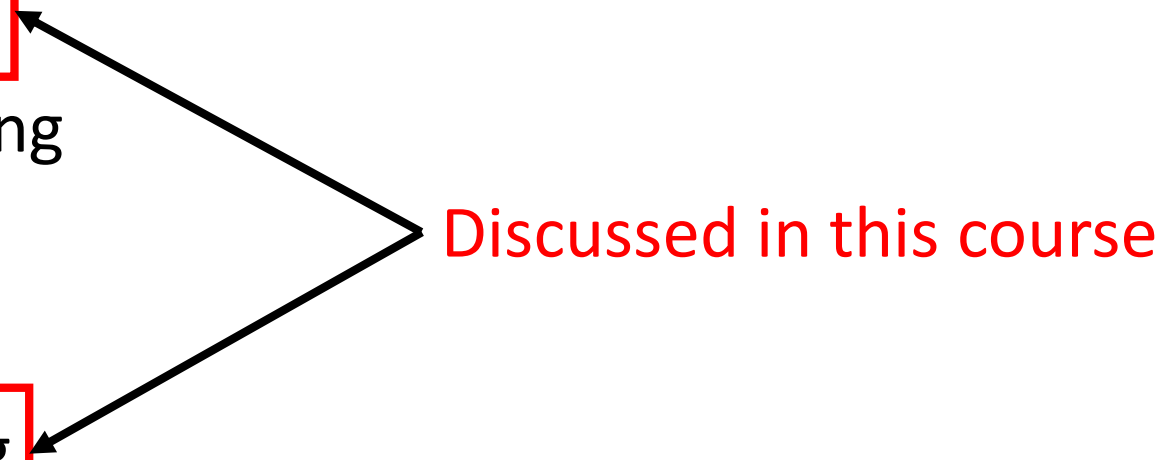
- Supervised learning divided into types based on type of response
  - **Regression** – continuous response
  - **Classification** – discrete or categorical response

Other types of learning paradigms

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

Other types of learning paradigms

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

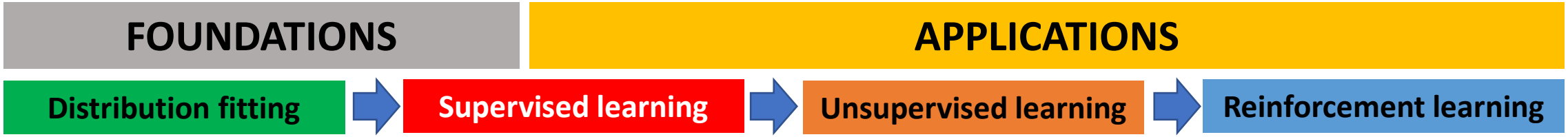


Other types of learning paradigms

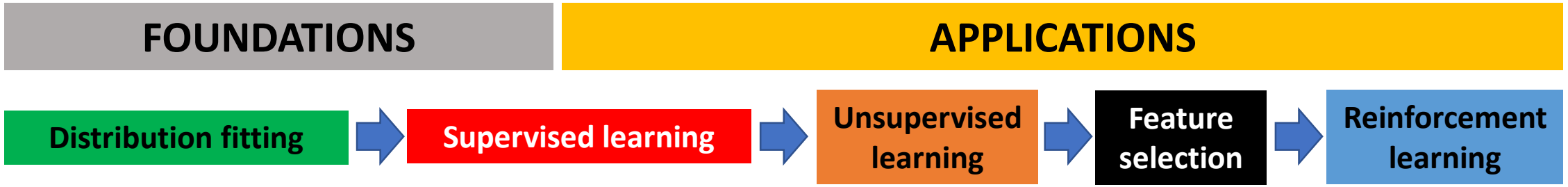
- Unsupervised learning or “Data discovery”
  - Observe variables without distinction between inputs and responses
  - Identify patterns in the data
  - Find relationships between variables and between observations
  - Useful in high-dimensionality situations

- Reinforcement learning
  - Learn the best actions to take based on observed rewards and punishments

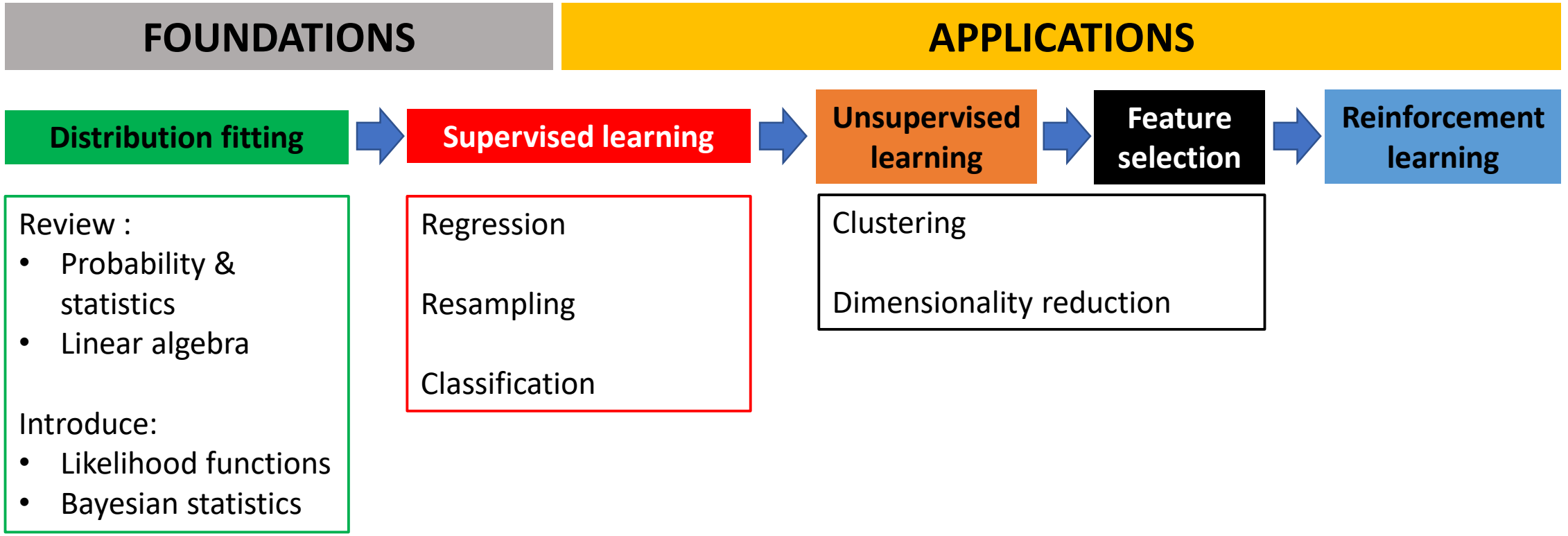
The course overview in terms of learning paradigms



We will also include a section on **Feature selection**



Key topics in the course



Calendar – **MIDTERM** and **FINAL WEEK** shown by **RED FILLED** calendar dates

August

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	**28**	**29**	30	31

September

S	M	T	W	T	F	S
1	2	3	**4**	**5**	6	7
8	9	10	**11**	**12**	13	14
15	16	17	**18**	**19**	20	21
22	23	24	**25**	**26**	27	28
29	30					

October

S	M	T	W	T	F	S
		1	**2**	**3**	4	5
6	7	8	**9**	**10**	11	12
13	14	15	**16**	**17**	18	19
20	21	22	**23**	**24**	25	26
27	28	29	**30**	**31**		

November

S	M	T	W	T	F	S
					1	2
3	4	5	**6**	**7**	8	9
10	11	12	**13**	**14**	15	16
17	18	19	**20**	**21**	22	23
24	25	26	27	28	29	30

December

S	M	T	W	T	F	S
1	2	3	**4**	**5**	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Supervised learning process

Collect data consisting of INPUTS and RESPONSES

Observation Number	Input 1	Input 2	Input 3	…	Response 1	Response 2	…
1							
2							
3							
⋮							

One row is one observation

Collect data consisting of INPUTS and RESPONSES

Observation Number	Input 1	Input 2	Input 3	…	Response 1	Response 2	…
1							
2							
3							
⋮							

One row is one observation

This is the ideal state!

Data are typically spread across multiple **data sources**…and are NOT organized in a consistent manner

Source 1

Key	A	B

Source 2

Key	C	D	E	F	G

Source 3

Key	H	I

…

The data contained within the various data sources are usually referred to as “raw data”.

We will refer to obtaining the “raw data” as the data access or query step.

Use Exploratory Data Analysis (EDA) to help understand the various data sources.

Data are typically spread across multiple **data sources**…and are NOT organized in a consistent manner

Source 1

Key	A	B

Source 2

Key	C	D	E	F	G

Source 3

Key	H	I

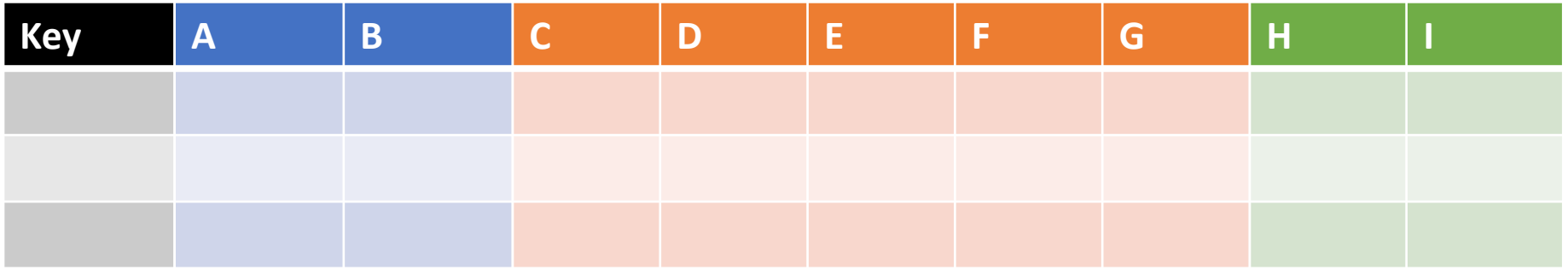
…

The different sources must be merged through **contextualization**:

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.

The contextualized dataset is **model** or **analytics ready**!



Use EDA to assess the impact of the contextualization process.

Clean the contextualized dataset

- 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 therefore work with "complete" datasets.

Identify performance metrics

- Before we build/train models, we need to decide how we will assess the performance of the models.

- Response type (categorical vs continuous response) usually dictates the style of metrics to use.

- Select multiple metrics appropriate for the response type.
  - Different metrics give us different clues about the strengths and weaknesses of a model at approximating the dataset under consideration.

Identify candidate set of models and *preprocessing* techniques

- Selecting a set of candidate models is usually straightforward
  - Want to try out varying levels of complexity (simple to very complex).
- Preprocessing refers to the addition, deletion, or transformation of the data used to build the model.
- Usually, preprocessing and cleaning are discussed together.
- I prefer to think of preprocessing with modeling because preprocessing impacts:
  - Interpretation of the terms within the model (for parametric models).
  - Efficiency of the model building algorithm (most models).

Generalization and overfitting

- When we build/train a model, we are trying to “fit” that model to the data.
  - The dataset we use to build or fit the model with is referred to as the **training set**.

- But, does being accurate on the training dataset represent the model will be accurate on a **new**, yet unseen, dataset?

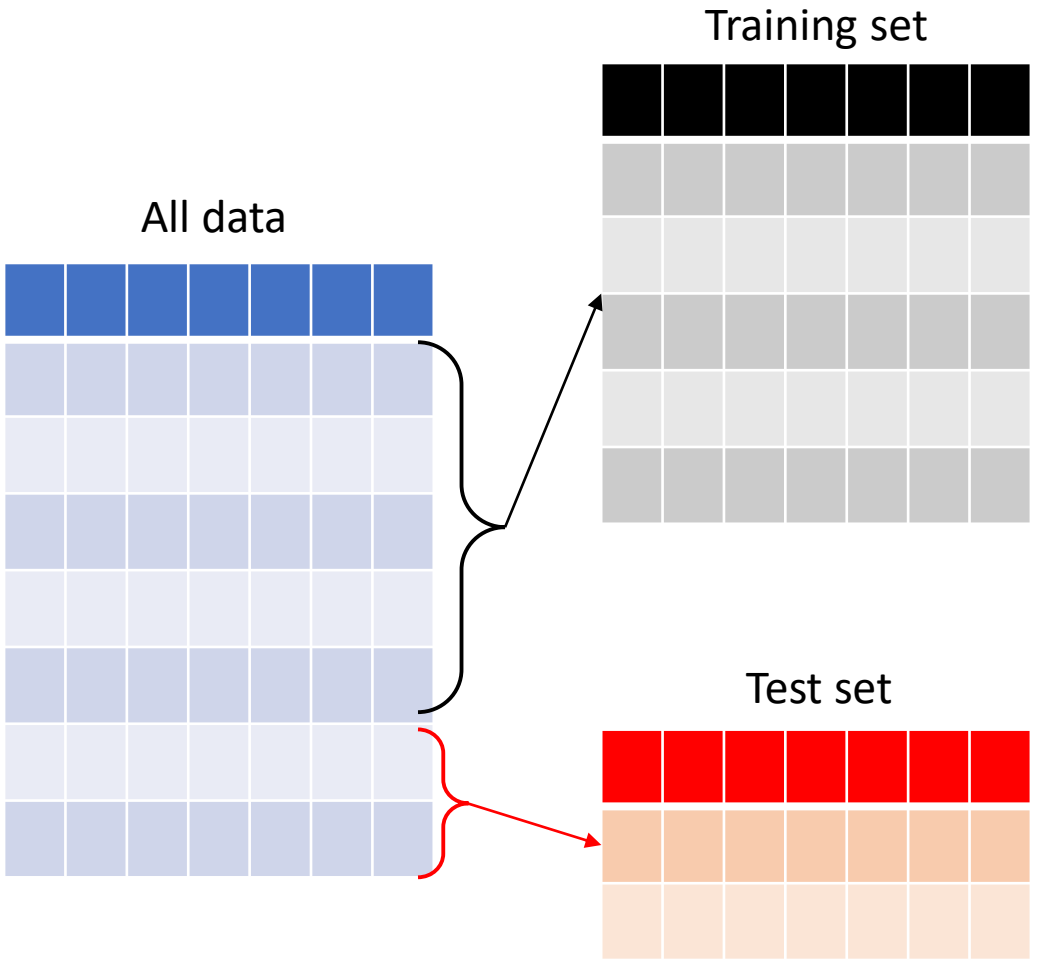
Generalization and overfitting

- Concern: A model *only* explains the observed data, or in other words, is **overfit** to the training data.
  - Can you think of what would cause this to happen?

- Key question is: How well does the model **generalize**?

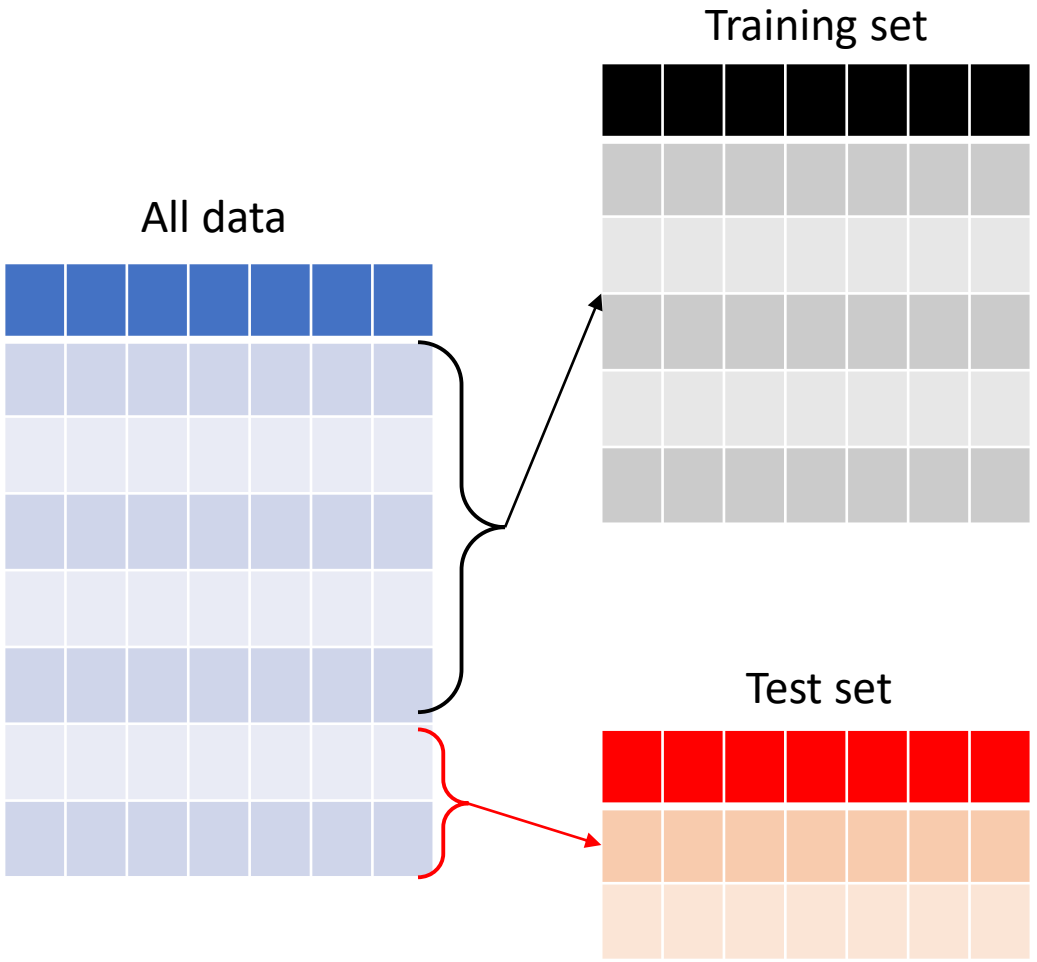
How can we guard against overfitting?

- If we have enough data, let's split the data into a training and test set.
- train/fit the models with the training set, and then evaluate model performance on the **hold-out test** set.
- We are in effect treating the hold-out test set as "new" data.



How can we guard against overfitting?

- Simple way to do this, is just to randomly partition the data into the training and test sets.
- What if though…we get “lucky”…in the training and test split?



Instead of using a single training and test split, we will perform the operation multiple times.

- This strategy is known as resampling.

- For each training/test split, train the model and evaluate the model performance on the test set.

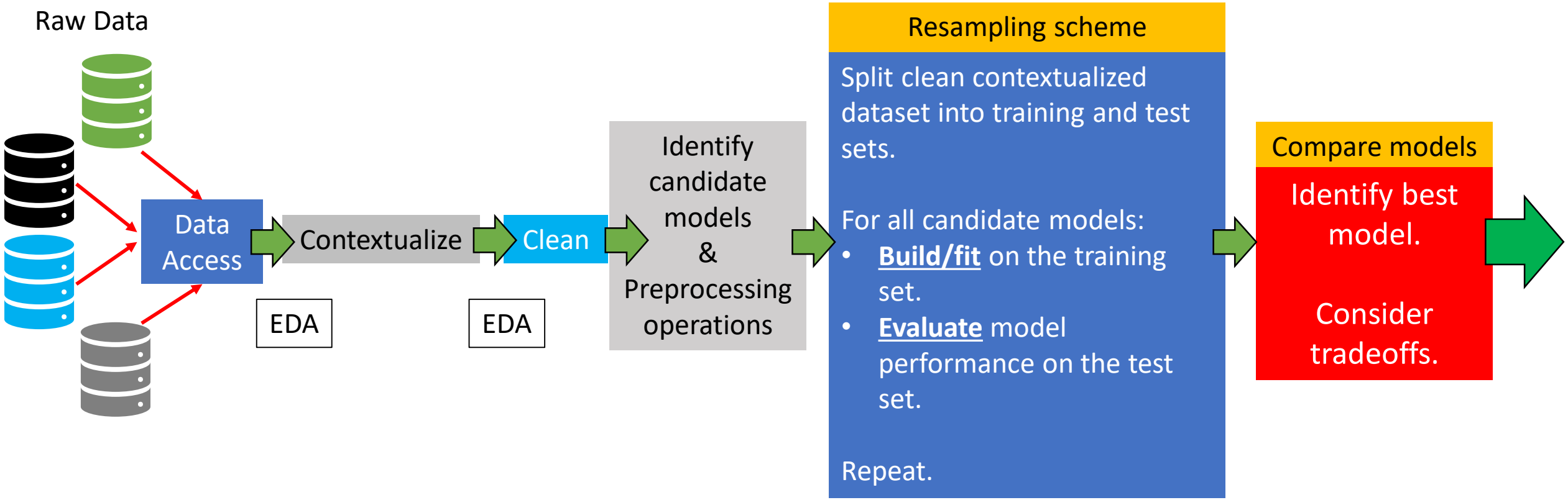
- An individual model's performance is then **summarized** based on the performance across the different test sets.

Common types of resampling approaches

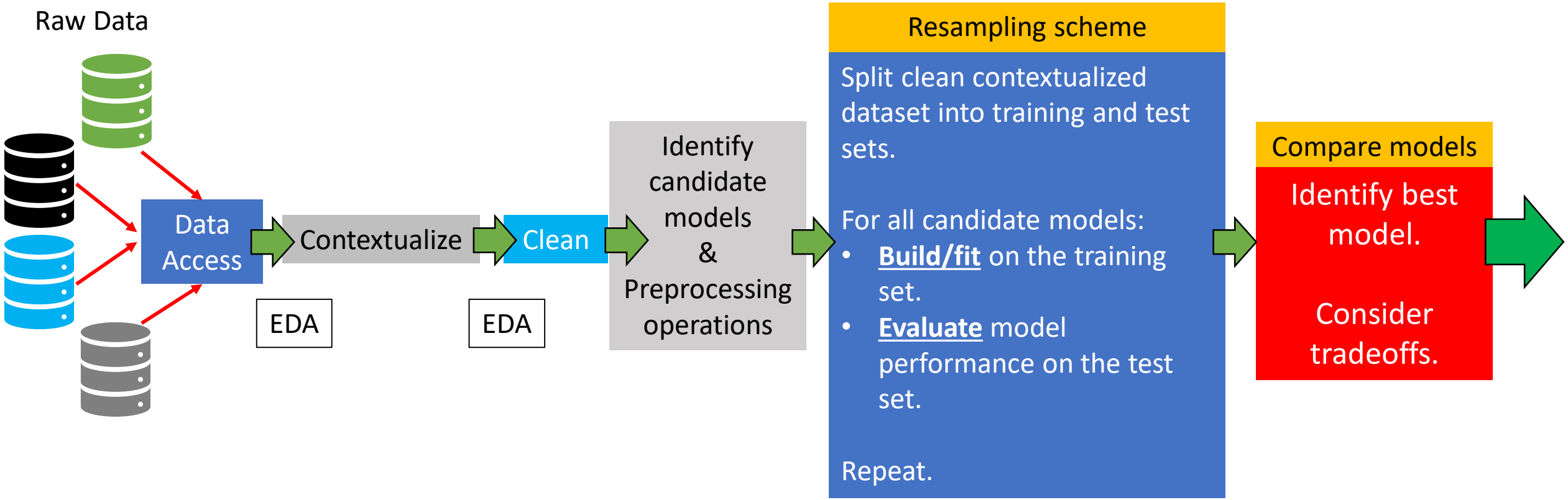
- Random sub-sampling
- Bootstrap
- k-fold Cross-validation -> Leave one-out cross-validation

- Requires building the model multiple times.
  - If the model is very complex, and time consuming to fit, resampling can be very time consuming!

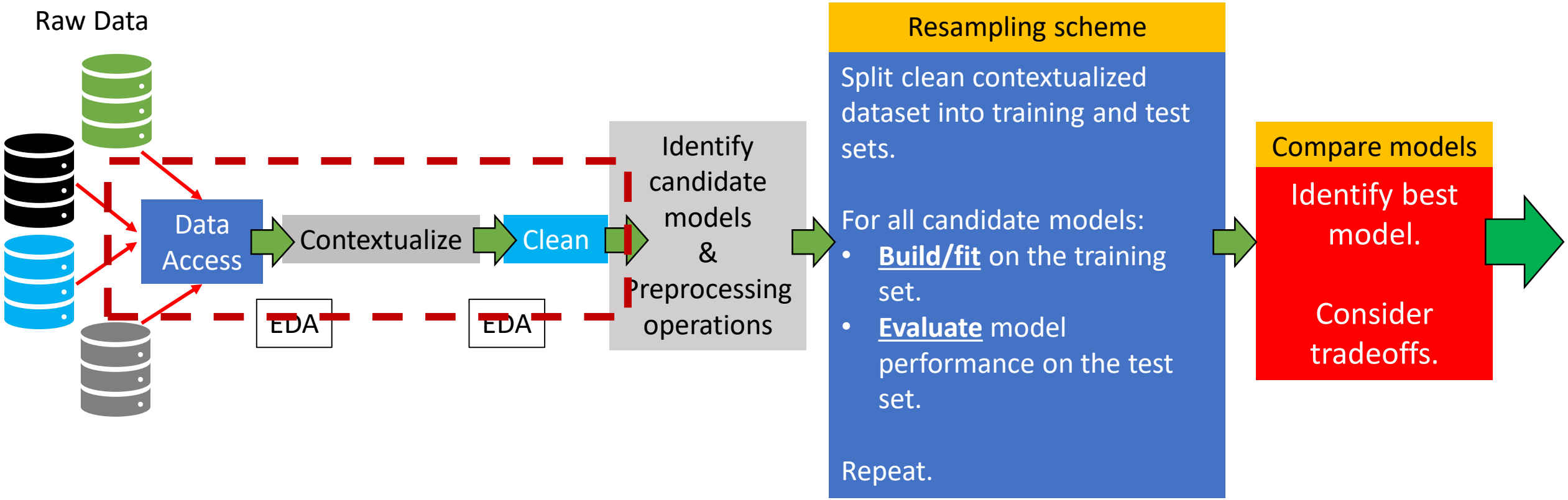
Complete supervised learning workflow



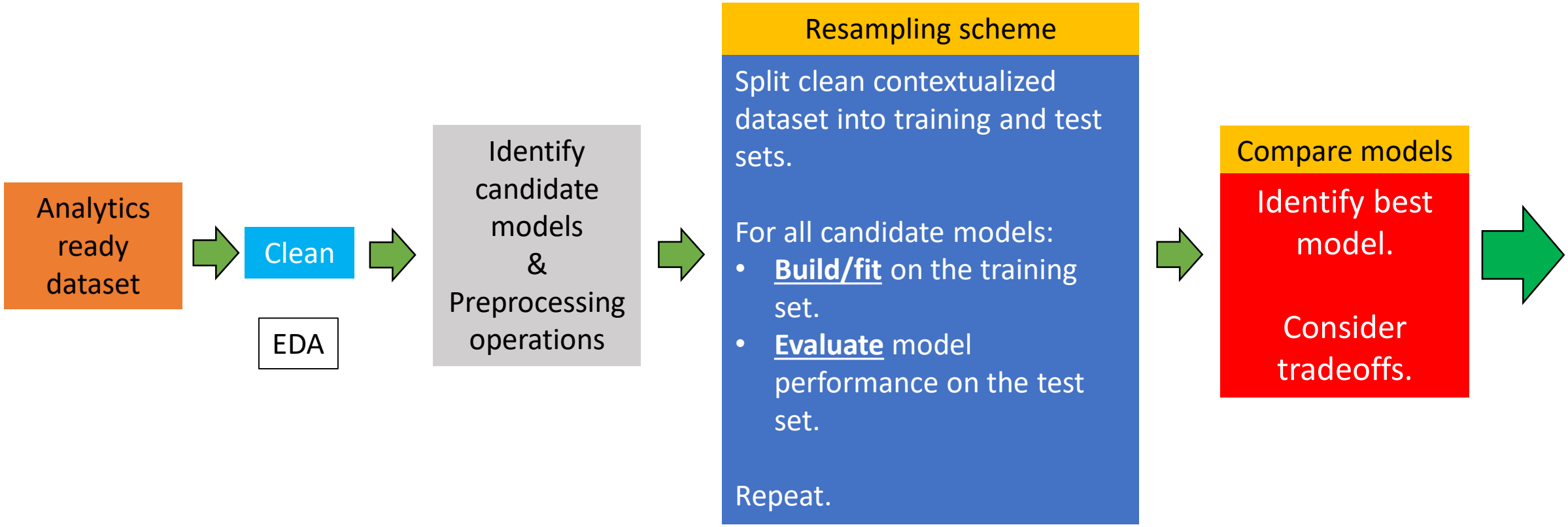
Which portion is the most time consuming?



Access, Contextualization, and Cleaning can take up 60% to 80% of a project!!



In this course, we will work with datasets that have already been contextualized and are mostly clean.



Let’s see this in action!

Demo with the `BostonHousing` dataset