GENERAL ASSEMBLY

INTRODUCTION TO MODELING

Matt Brems

General Assembly, D.C.

AGENDA

1. Data Science Process & Modeling

2. Linear Regression

DATA SCIENCE PROCESS

1. Define problem.

2. Gather data.

3. Explore data.

4. Model with data.

5. Evaluate model.

6. Answer problem.

MODELING

- Modeling is something that we naturally do.

MODELING

- Modeling is something that we naturally do.

- A **model** is a simplification of reality.

MODELING

- Modeling is something that we naturally do.

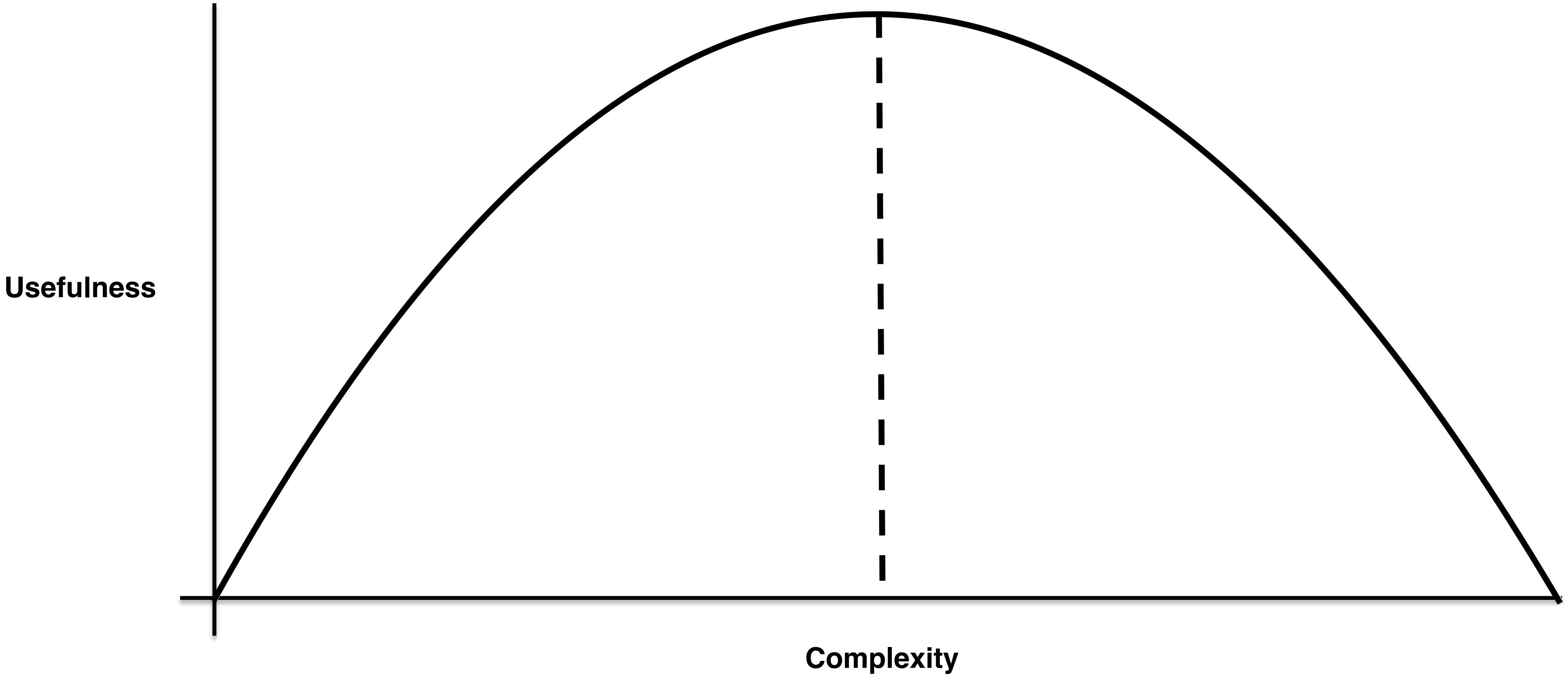
- A **model** is a simplification of reality.

  - How do we simplify?
    - Making assumptions about how things behave.
    - Taking into account only really important factors.

MODELING

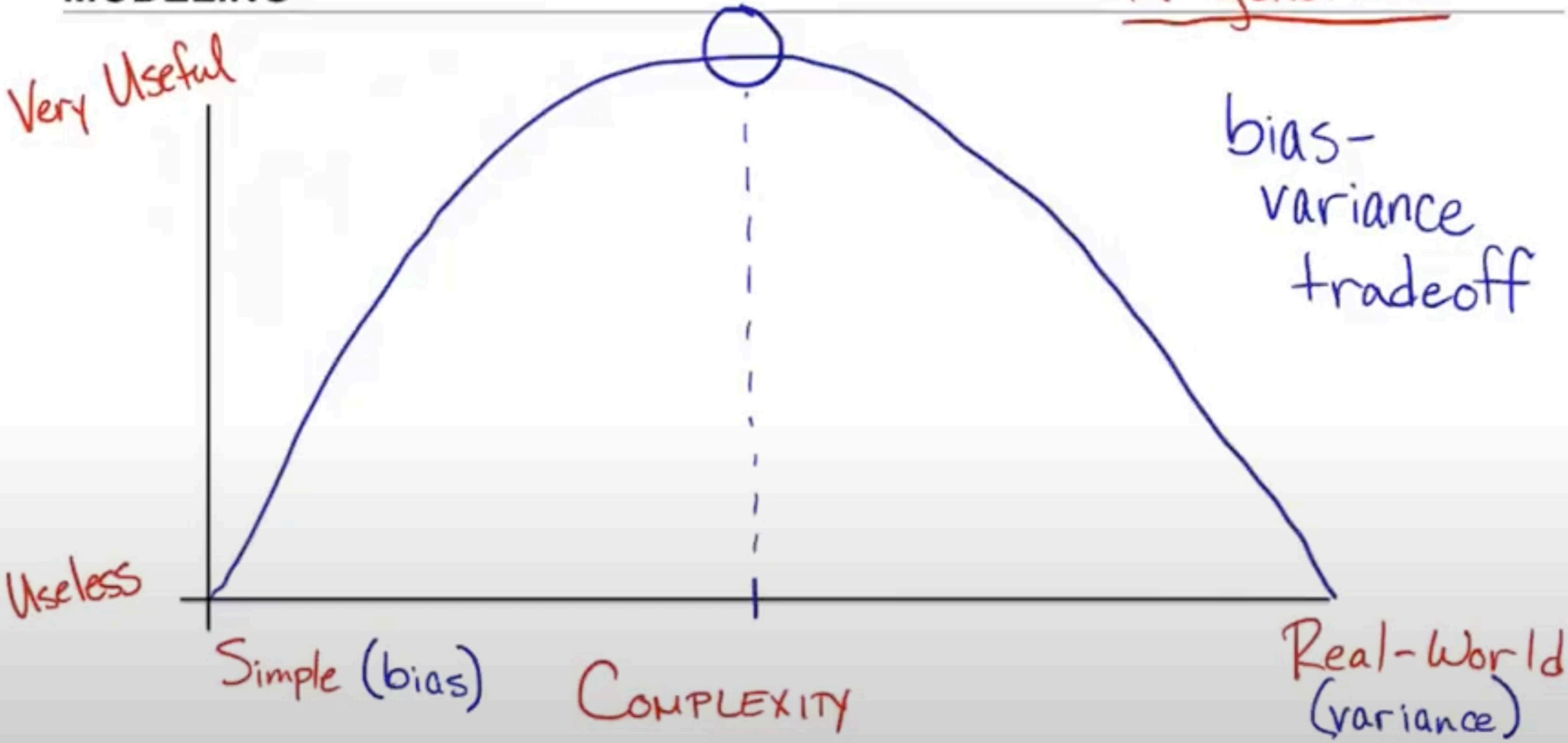
“Essentially, all models are wrong, but some are useful.”
– George Box, 1987

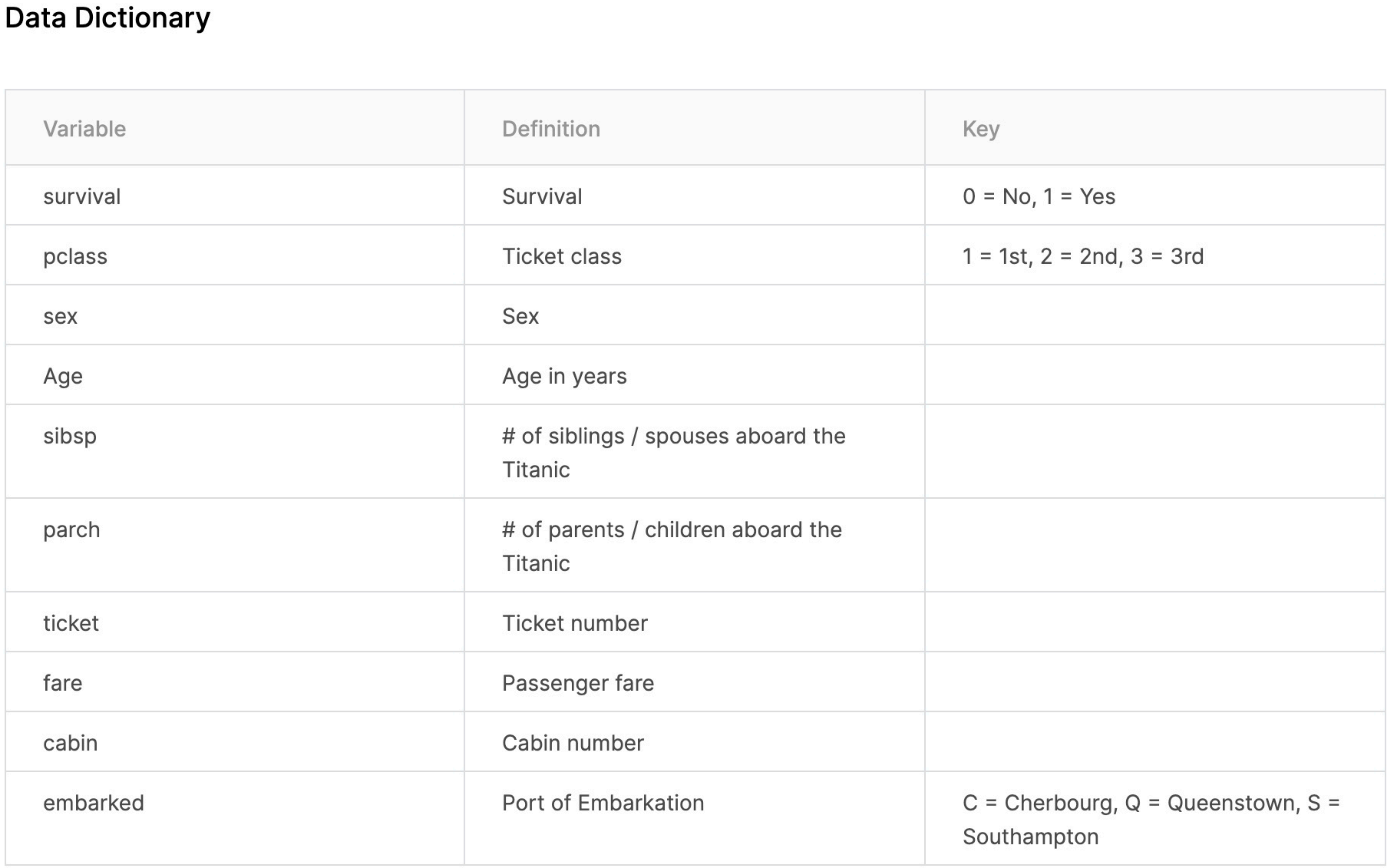
MODELING



MODELING

in general!





Data Dictionary

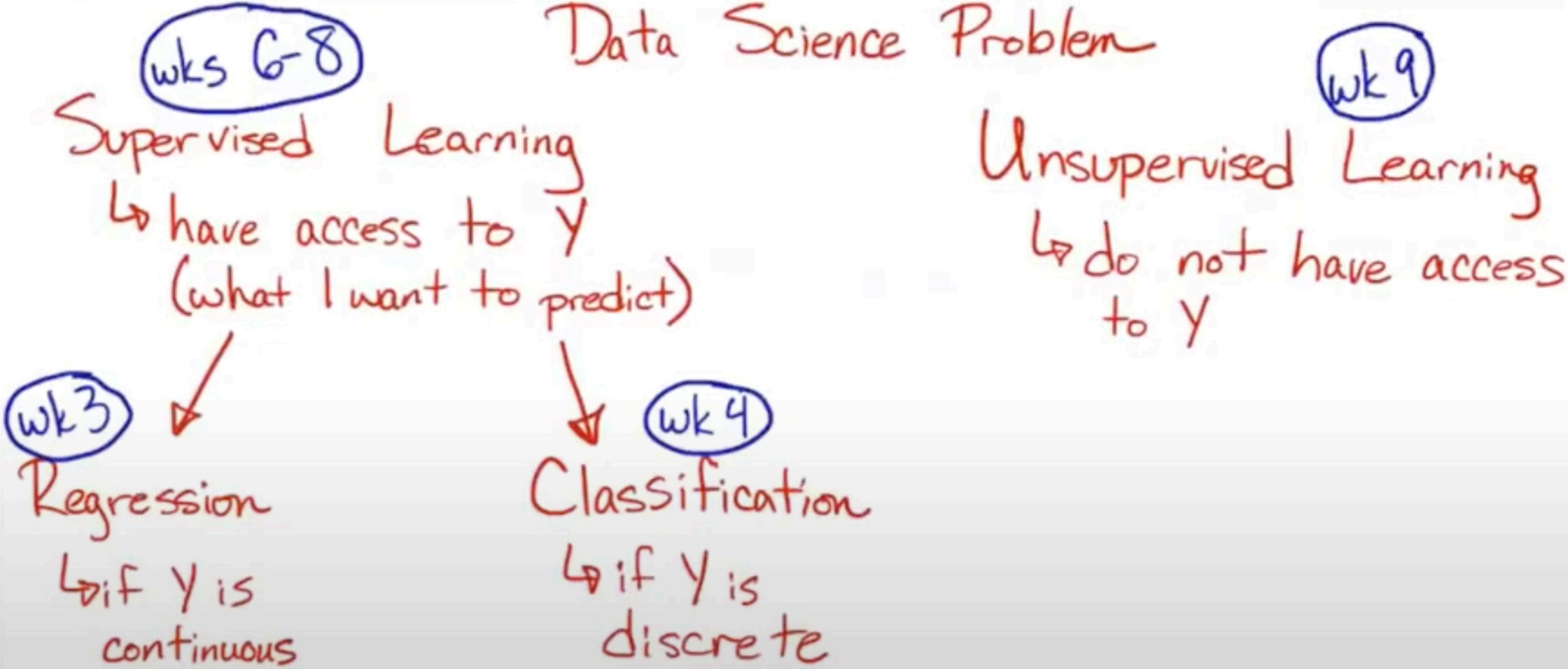
Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

WHY DO WE MODEL?

- Prediction
    - How long does it take me to get to work?
    - How much money is a 29-year-old DSI alum expected to make?

- Inference
    - What is the effect of sex on income?
    - How much more money can I be expected to make in a year?

MACHINE LEARNING ALGORITHMS



Data Science Problem

wks 6-8

Supervised Learning

↳ have access to y (what I want to predict)

wk 3

Regression

↳ if y is continuous

wk 4

Classification

↳ if y is discrete

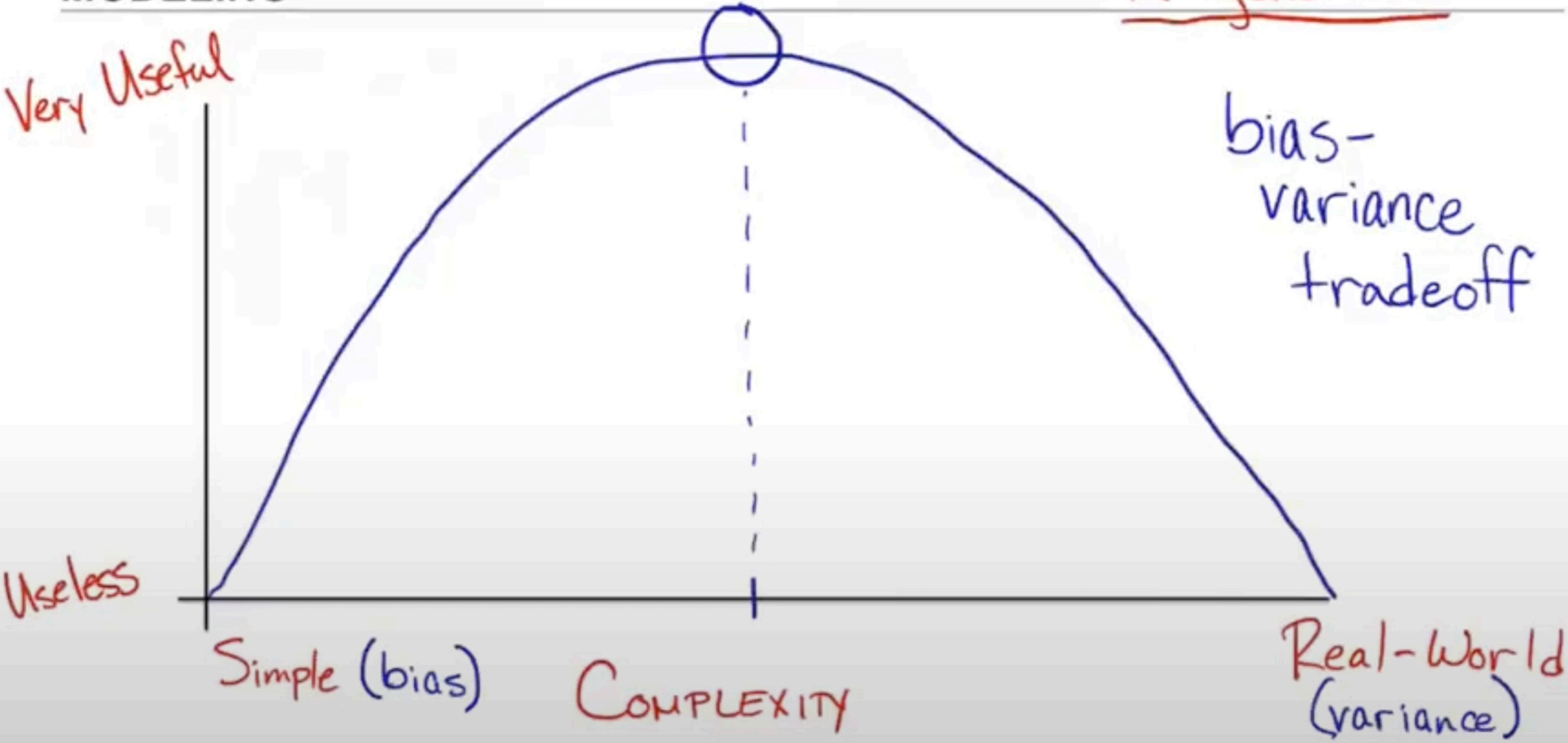
wk 9

Unsupervised Learning

↳ do not have access to y

MODELING

in general!



MACHINE LEARNING ALGORITHMS

TERMINOLOGY

- $\boldsymbol{X}$: our data, the independent/explanatory variables we use to predict $\boldsymbol{Y}$.
- $\boldsymbol{Y}$: our data, the dependent variable we want to predict.
- $\hat{\boldsymbol{Y}}$: our predicted values of $\boldsymbol{Y}$.

f: X -> Y

Input Output

MODELING GOALS

1. Use observed values of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to model relationship between them.

2. Build model that makes $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ as close as possible.

3. Use observed values of $\boldsymbol{X}$ and existing model to make predictions $\widehat{\boldsymbol{Y}}$.