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. Data Science Problem
2. Gather data. .csv, database } 80%
3. Explore data. EDA
4. Model with data.
5. Evaluate model.
6. Answer problem.

MODELING

- Modeling is something that we naturally do.

Commute time

~5 minutes to get to Metro
~20 minutes on Metro (40) (30)
~5 minutes from Metro to GA

~30

Lyft: 20

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.

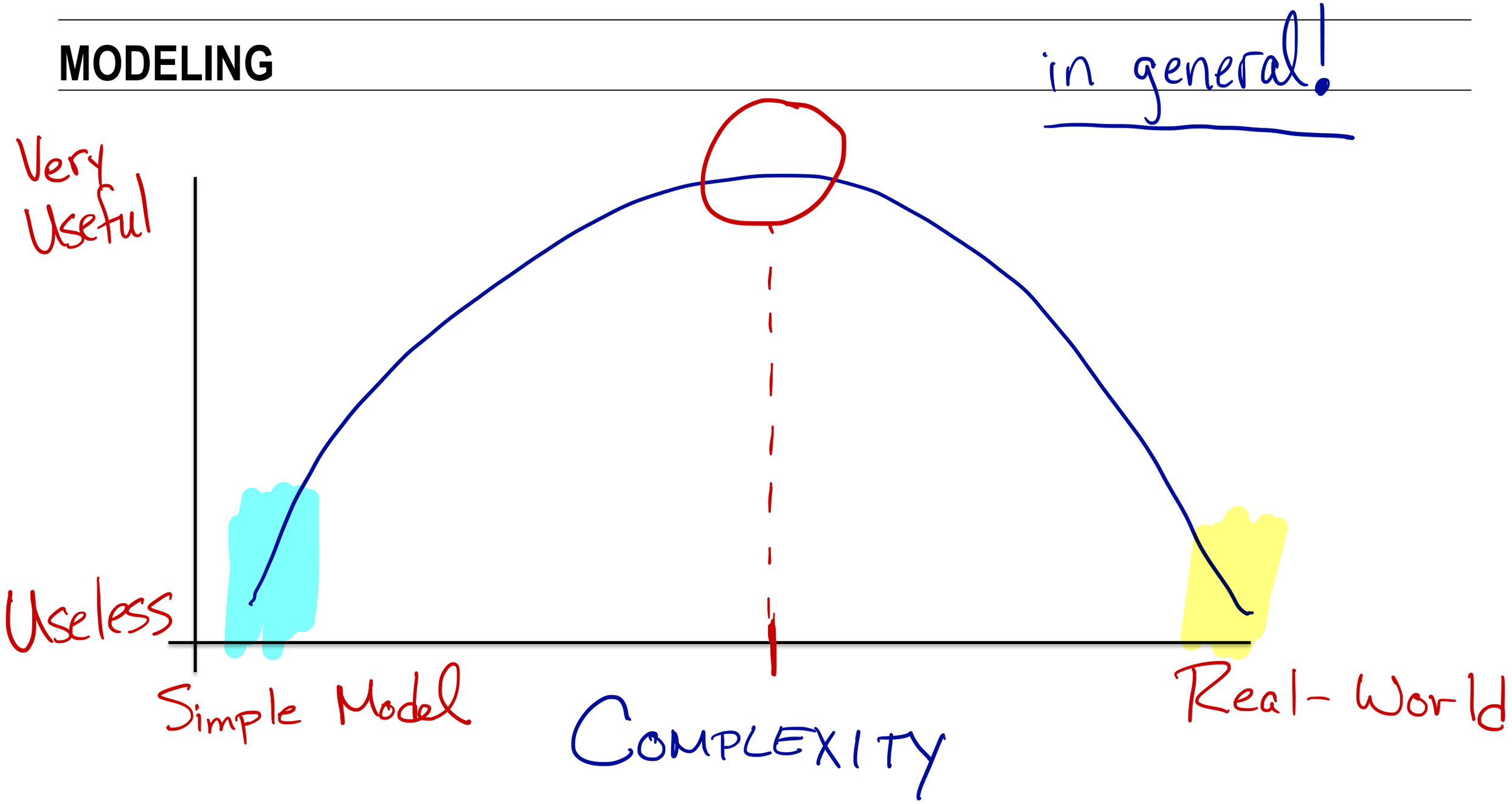
  - How do we simplify?
    - Making assumptions about how things behave.
    - Taking into account only really important factors.

MODELING

"Essentially, all models are wrong, but some are useful."

– George Box, 1987

MODELING


in general!
Very Useful
Useless
Simple Model
COMPLEXITY
Real-World

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?

stock prices

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

medical studies

MACHINE LEARNING ALGORITHMS

- **Machine learning** is a term we use to describe getting computers (machines) to learn without needing to be explicitly programmed.

- There are many different machine learning algorithms we'll cover in the class - from linear regression to neural networks!

MACHINE LEARNING ALGORITHMS

wk 6, 7

DATA SCIENCE PROBLEM

wk 8

Supervised Learning
↳ have access to our Y variable

- Regression (wk3)
  ↳ Y is continuous
- Classification (wk4)
  ↳ Y is discrete

Unsupervised Learning
↳ do not have access to output we want to predict.

↳ clustering

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.
- $\widehat{\boldsymbol{Y}}$: our predicted values of $\boldsymbol{Y}$.

↳ "Y-hat"

$f: X \xrightarrow{\text{relationship}} Y$

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 $\hat{\boldsymbol{Y}}$ as close as possible.

observed → predicted

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

f $\qquad$ $f(X) = \hat{y}$