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. Real-world 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 mins. to Metro

~20 mins on Metro (40) (30)

~5 mins from Metro to GA

Lyft: 15 minute

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

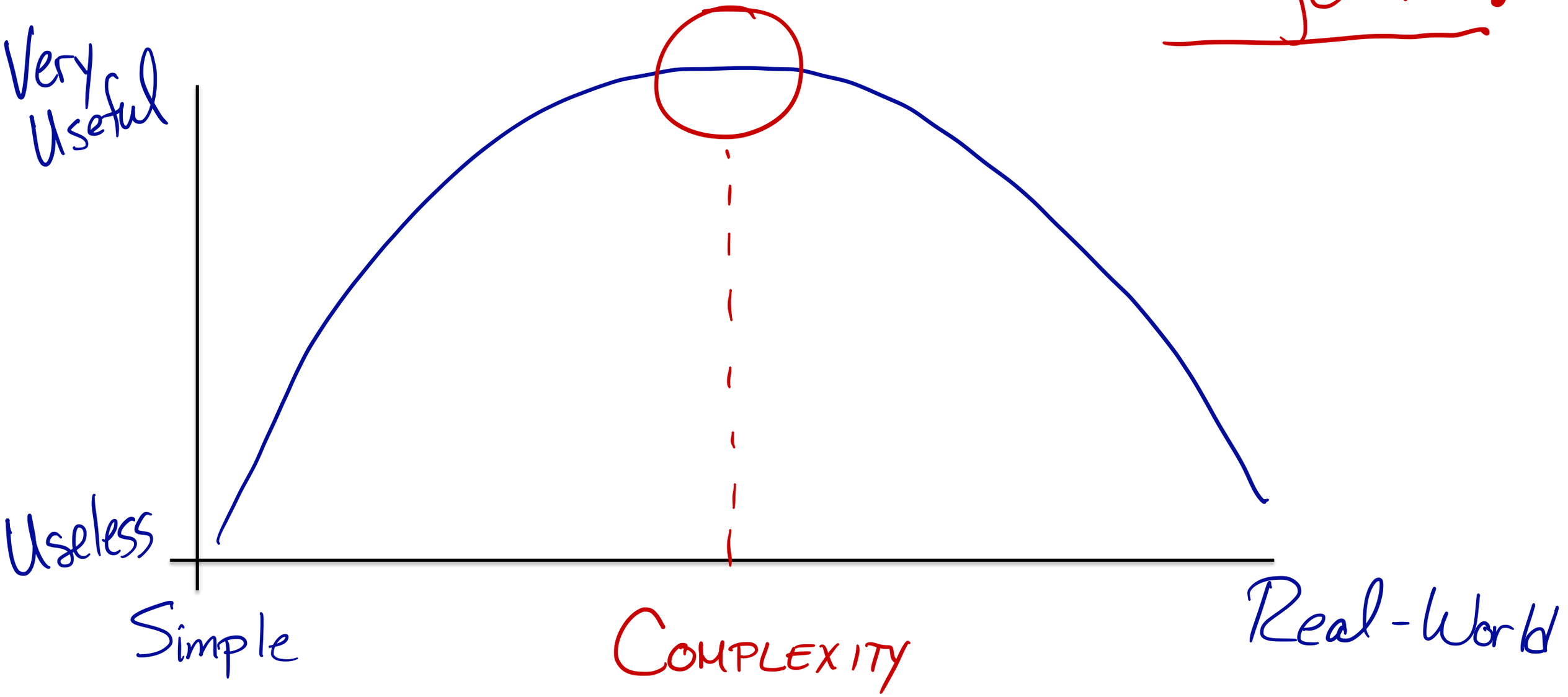
Is my model useful?

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

– George Box, 1987

MODELING

in general!



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?

stock prices

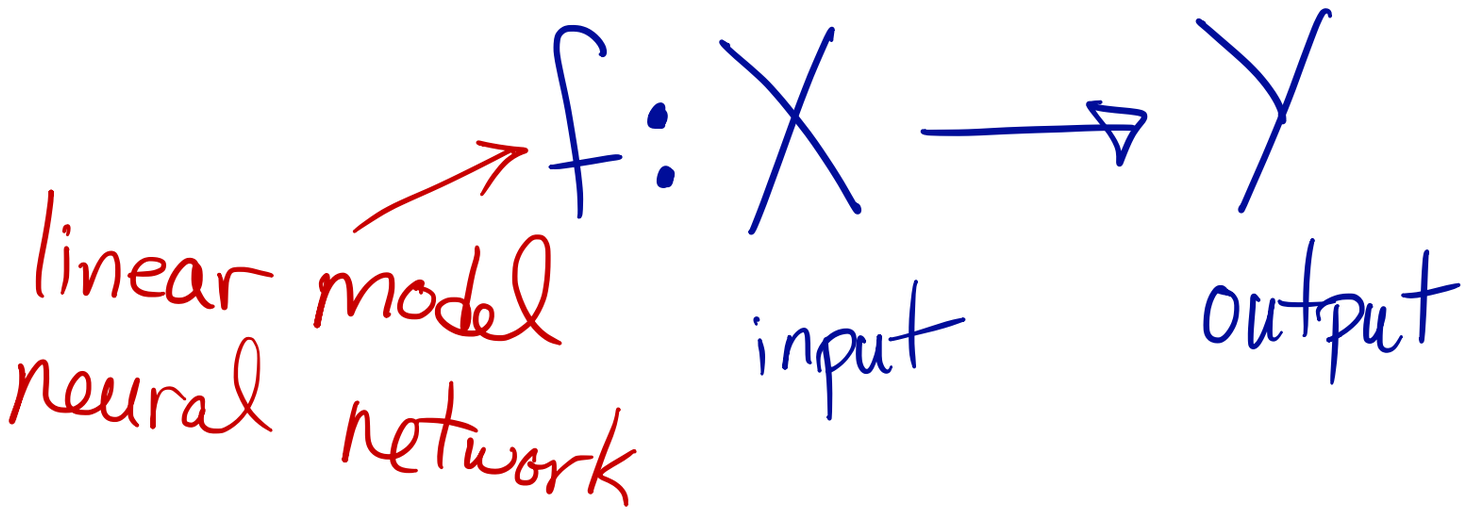
I do want to know "why."

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 Y (what I want to predict)

Regression wk3
↳ continuous Y

Classification wk4
↳ discrete Y

Unsupervised Learning
↳ do not have access to Y
↳ clustering
↳ PCA

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 $\boldsymbol{Y}$.

$f(X) = \hat{Y}$

$f: X \rightarrow Y$

input output

MODELING GOALS

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

$f: X \rightarrow Y$

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

$Y \approx \hat{Y}$

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

$f(X) = \hat{Y}$