Math 650.4 Data Science / Machine Learning 1/28

- Def: Models are "approximations" / "abstractions" to "reality" / "truth" / "phenomenon"

Model	Reality
Model airplane	Airplane
Map	Roads in a city
wind tunnel	Fast moving air in the sky
"Early to bed, early to rise makes a man wealthy"	Success of a person

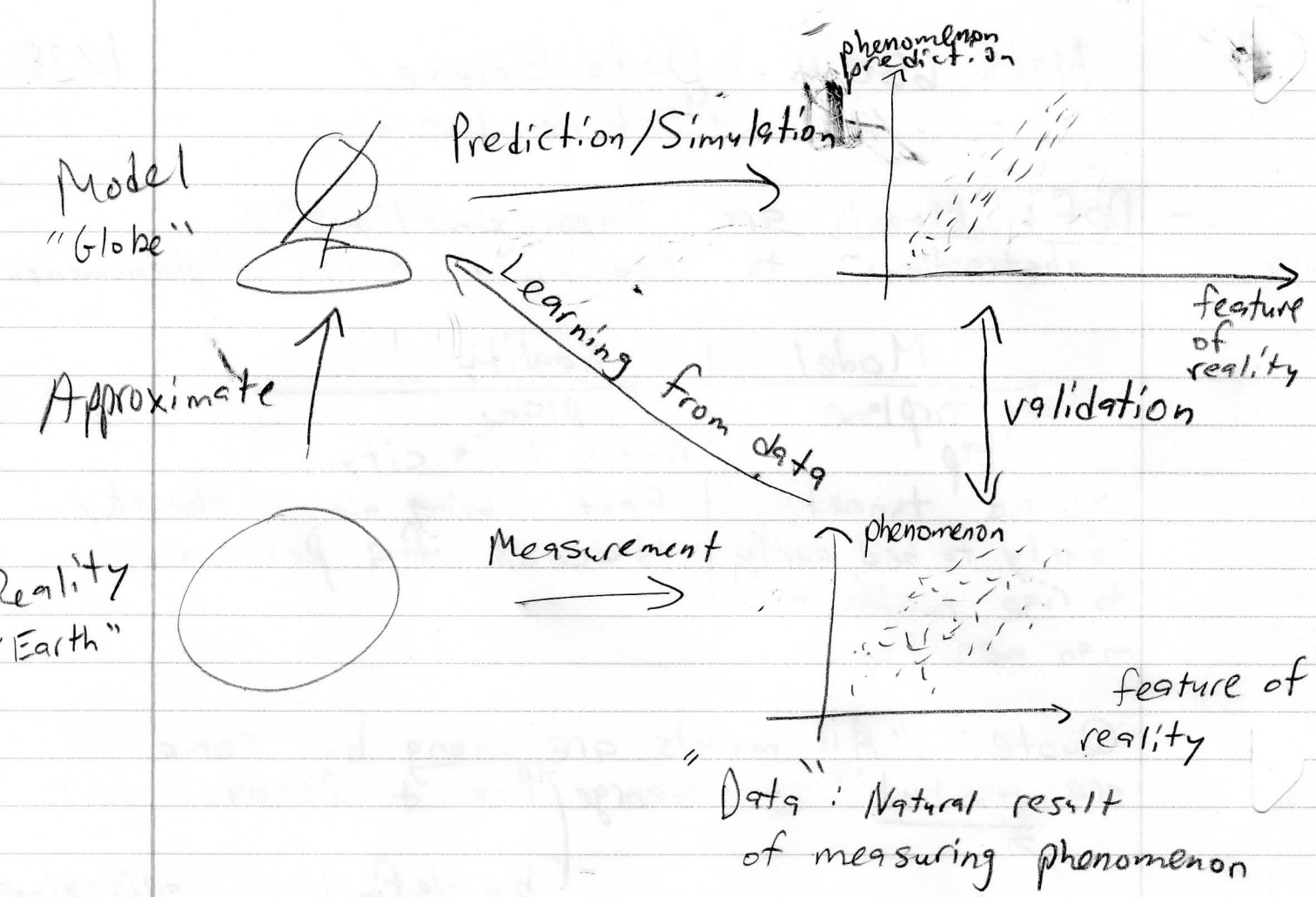
Quote: "All models are wrong but some are useful" - George Box & Draper

useful: you can use to your advantage

wrong: by definition, approximate

- Main Goals

1. Prediction: Can the model predict a future event of the phenomenon under examination?
2. Explanation: How does the phenomenon work? (science)



"Data": Natural result of measuring phenomenon

- "Early to bed, early to rise makes a man healthy, wealthy, & wise."

- What is this modeling?
  Health, wealth, wisdom.
  3 different phenomena
  Ambiguous model.

  To make this model concrete, we need numerical definitions, i.e "metrics."
  Metrics define how to measure both phenomena & features of reality.

  What are the features? (Inputs)
  1. Bed time
  2. wake time

  What are the predictions? (Output)
  1. Health
  2. wealth
  3. Wise

  Metric for bedtime (b): Avg. bedtime in 24-hr time
  Metric Evaluation
  1. Does it capture the feature / phenomenon? Yes
  2. Is it easily readable & unambiguous? Yes
  3. Good resolution? Yes
  4. Monotonic?

- wake time (w)
- health measured by longevity (l)
- wealth measured by networth at 65 (m)
- wisdom measured by philosophy exam (s)

We want to estimate f where

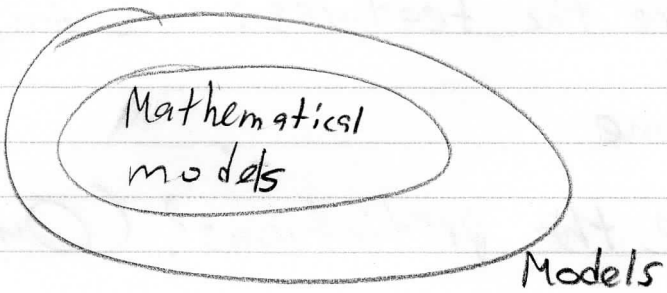
outputs

$$\begin{bmatrix} l \\ m \\ s \end{bmatrix} = f(b, w)$$

inputs

prediction of phenomena

"mathematical model" are ideas & abstractions, not physical entities



Mathematical models are at least 4,000 years old

Examples:

$$a = \frac{F}{m} = f(m, F)$$

$$E = mc^2$$

deterministic

$y = t(z_1, z_2, \ldots, z_t)$ ← true causal inputs

↖ true function (unknown) that combines z_i's

↖ phenomenon / response / outcome / endpoint / dependent variable (one-dimensional)

Example

Phenomenon is pay back mortgage ($y=1$) or not pay back mortgage ($y=0$).

$y \in \{0, 1\} = Y$ (output space), which is binary

↑ "positive class"

What are the causal inputs?

z_1: has the money $\in \{0, 1\}$ at payback time

z_2: unforseen emergency $\in \{0, 1\}$

z_3: criminal intent $\in \{0, 1\}$

ex: $t(z_1, z_2, z_3) = z_1(1 - z_2)(1 - z_3)$

- Fundamental Modeling Problem: You don't know the z's or t.

Next best thing is to obtain measurements & approximate the z's. Call these measurements x's.

x_1 : credit score $\in \mathbb{R}^+$
x_2 : salary based on tax return $\in \mathbb{R}^+$
x_3 : missed loan previously $\in \{0, 1\}$
x_4 : crime in past $\in \{0, 1\}$