INTRODUCTION TO STATISTICAL MODELING

Total and partial change

Interpreting effect size

Input X = 7 ⟶ Model M ⟶ Response Y = 3.2

Change in output

Input X = 8 ⟶ Model M ⟶ Response Y = 3.5

$$\frac{(3.5 - 3.2)}{(8 - 7)} = 0.3$$

Is this big or small?

It depends on the units!

Example: used car prices

- Car price influenced by mileage, age, condition, etc.
- Price goes down as mileage goes up
- Effect size has units (dollars per mile)

Modeling car prices

```
   Year Mileage Price Color Location Model Age
2  1994   94000  1988 white  Phoenix    GL  15
6  1996  115730  2199 beige  Phoenix    GL  13
7  1997   74564  2995 green  Phoenix    GL  12
8  1998  143000  1200  blue   Fresno    SE  11
11 1999   85000  2488 white  Phoenix    SE  10
12 2000   94727  3879  gray  Phoenix   SES   9
```

```
> ford_mod <- lm(Price ~ Mileage + Age + Model + Location,
               data = Used_Fords)

> effect_size(ford_mod, ~ Age)
      slope Age   to:Age Mileage Model  Location
1 -536.7337   3 6.144894 48333.5    SE Cambridge
> effect_size(ford_mod, ~ Mileage)
        slope Mileage to:Mileage Age Model  Location
1 -0.05467762 48333.5   82420.18   3    SE Cambridge
```

Comparing effect sizes

$$\frac{\text{dollars}}{\text{mile}} \neq \frac{\text{dollars}}{\text{year}}$$

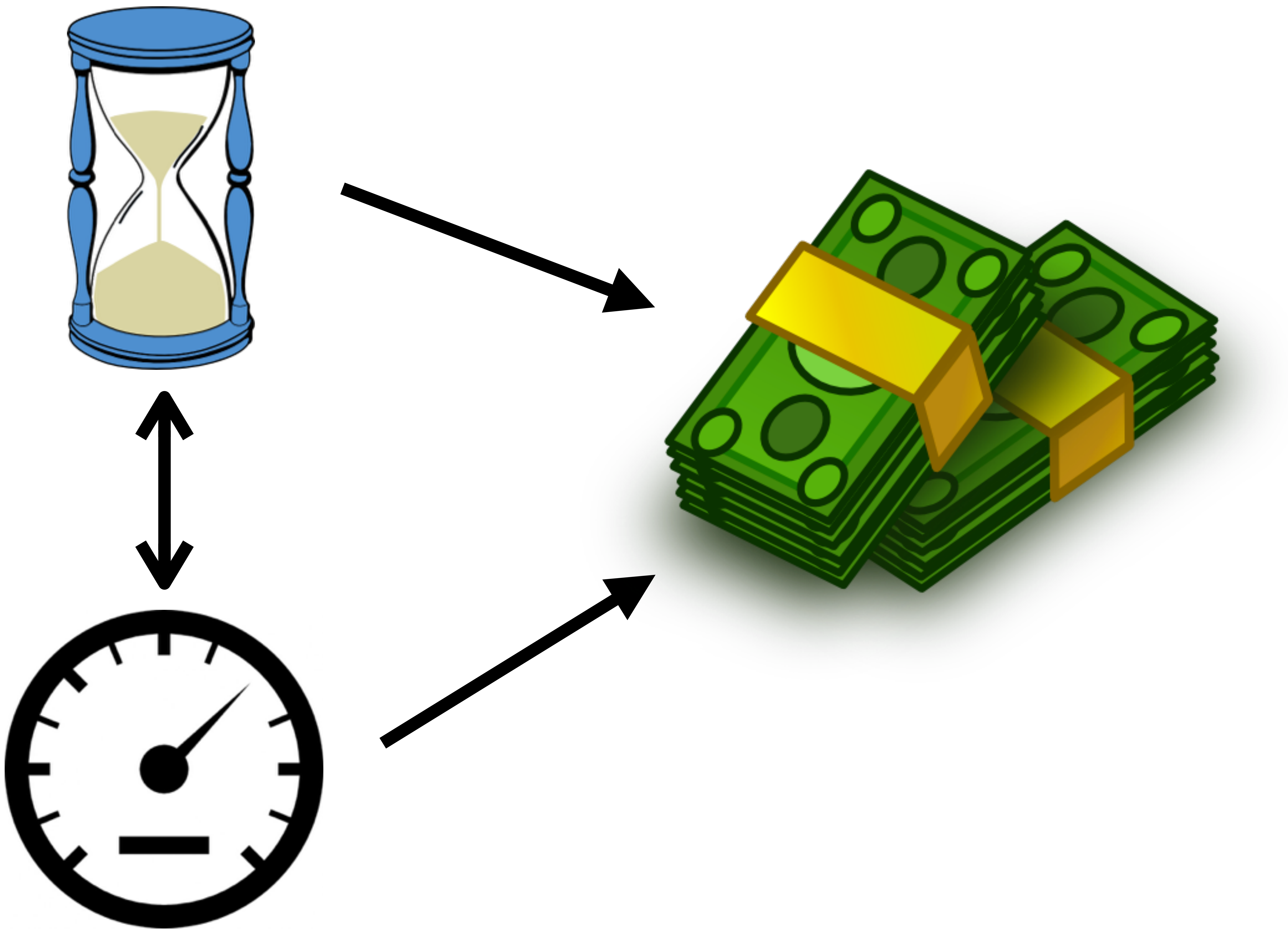
$$-0.055\frac{\text{dollars}}{\text{mile}} \times 10000\frac{\text{miles}}{\text{year}} = -550\frac{\text{dollars}}{\text{year}}$$

Compare to **-536 dollars/year** from last slide

Total vs. partial change

- **Partial** change: Impact on the response of changing one input *holding all other inputs constant*
- **Total** change: Impact on the response of changing one input *letting the others change as they will*

Total and partial change in car prices



Total and partial change in car prices

Question 1: How much resale value will I lose holding onto my car for another year?

Both mileage and age will change, so ***total change*** *is appropriate*

Question 2: I'm thinking of going on a quick 1000-mile trip. How much will that change the resale of my car?

Mileage changes a lot, but age hardly changes, so ***partial change*** *is appropriate*

Implications for model building

- **Partial change**: Include all covariates that you want to hold constant while varying the explanatory variable
- **Total change**: Exclude all covariates that you want to allow to change along with the explanatory variable

Partial change in used car prices

```
> ford_mod <- lm(Price ~ Mileage + Age + Model + Location,
                 data = Used_Fords)

# Partial effect size of Mileage
> effect_size(ford_mod, ~ Mileage)
        slope Mileage to:Mileage Age Model  Location
1 -0.05467762 48333.5   82420.18   3    SE Cambridge

# Partial effect size of Age
> effect_size(ford_mod, ~ Age)
      slope Age   to:Age Mileage Model  Location
1 -536.7337   3 6.144894 48333.5    SE Cambridge
```

Total change in used car prices

```
> price_vs_age <- lm(Price ~ Age, data = Used_Fords)

# Total effect size for Age: use model that excludes Mileage
> effect_size(price_vs_age, ~ Age)
      slope Age    to:Age
1 -1124.556   3 6.293838

# Total effect size for Mileage: use model that excludes Age
> price_vs_mileage <- lm(Price ~ Mileage, data = Used_Fords)
> effect_size(price_vs_mileage, ~ Mileage)
       slope Mileage to:Mileage
1 -0.1103443 49144.5   83384.38
```

Let's practice!

R-squared

Some notation

- Correlation: r ("little r")
- Coefficient of determination: R^2 ("R-squared")

The original publication...

SOCIETIES AND ACADEMIES.

LONDON.

Royal Society, December 20, 1888.—"Correlations and their Measurement, chiefly from Anthropometric Data." By Francis Galton, F.R.S.

Two organs are said to be co-related or correlated, when variations in the one are generally accompanied by variations in the other, in the same direction, while the closeness of the relation differs in different pairs of organs. All variations being due to the aggregate effect of many causes, the correlation is a consequence of a part of those causes having a common influence over both of the variables, and the larger the proportion of the common influences the closer will be the correlation. The length of the cubit is correlated with the stature, because a long cubit usually implies a tall man. If the correlation between them were very close, a very long cubit would usually imply a very tall stature, but if it were not very close, a

Little r and modeling

- Simple summary of a simple model: A ~ B
- A and B are both quantitative
- Sign indicates sign of effect size (positive or negative)
- Magnitude tells us...nothing about prediction error
- No physical units

R-squared and modeling

- A generalization of r to more complex modeling formulas: `A ~ B + C + ...`
- A is quantitative
- Always positive
- Magnitude tells us...still not much

Features of R-squared

- Number between 0 and 1
- Refers to a statistical model of data
- Fraction of the variance in the response variable accounted for by the model
- Bigger is not always better
- R-squared is about prediction, but that's not always the goal

Alternatives to R-squared

- **Predictive ability:** Cross validated prediction error
- **Mechanics of system:** Effect sizes and confidence intervals

Let's practice!

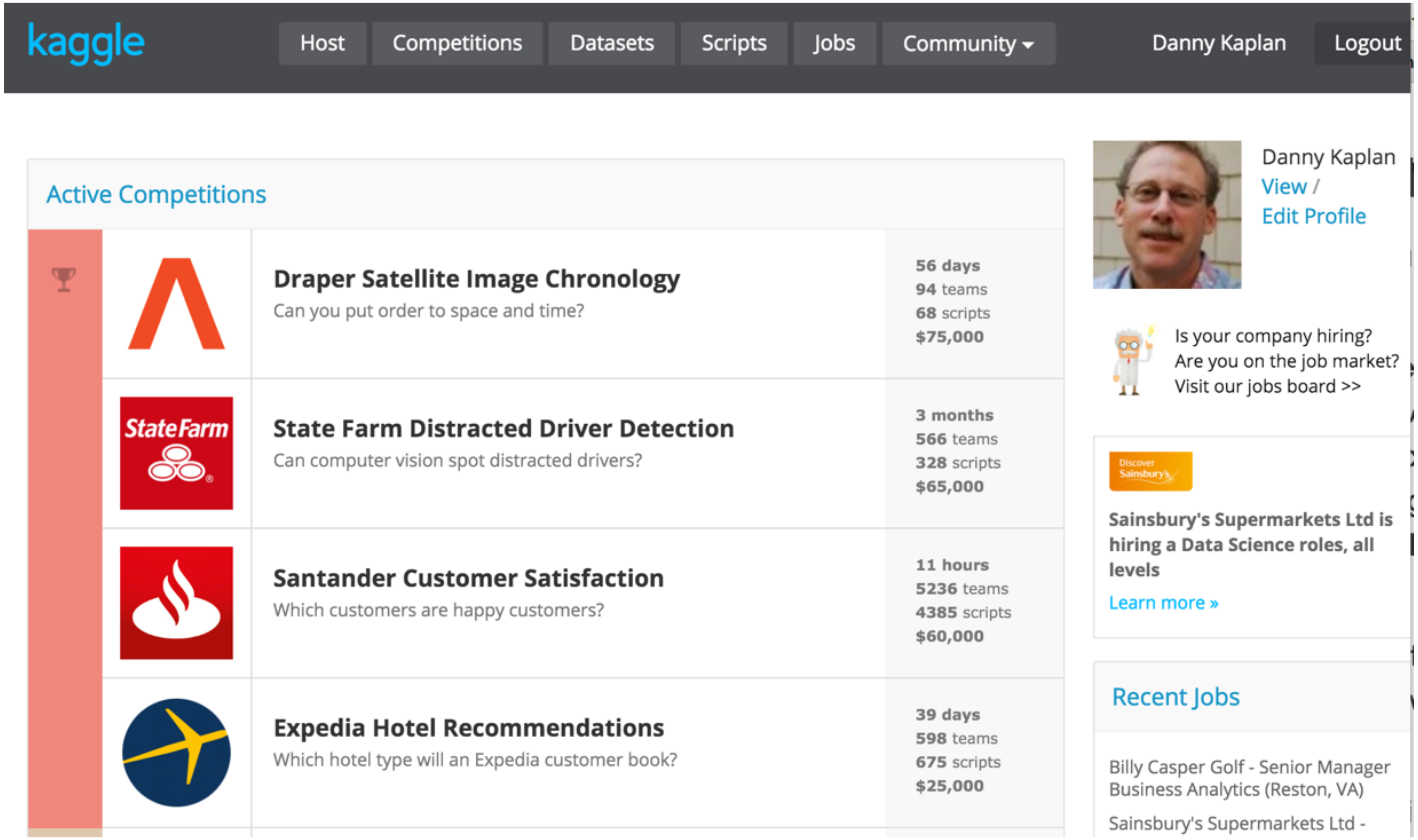
INTRODUCTION TO STATISTICAL MODELING

Degrees of freedom

You've seen...

- Model architectures: `lm()` and `rpart()`
- Explanatory and response variables
- Interactions between explanatory variables
- Prediction error and cross validation
- Covariates

Ready for Kaggle?



From a Kaggle competition...

TAB FOOD INVESTMENTS

Completed • $30,000 • 2,257 teams

Restaurant Revenue Prediction

Mon 23 Mar 2015 – Mon 4 May 2015 (12 months ago)

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - Timeline
- Forum

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
sampleSubmission	.csv (1.52 mb)
train.csv	.zip (4.54 kb)
test.csv	.zip (2.45 mb)

The restaurant data

```
          City Type revenue
1     İstanbul   IL 5653753
2       Ankara   FC 6923131
3   Diyarbakır   IL 2055379
4        Tokat   IL 2675511
5    Gaziantep   IL 4316715
...
```

```
> nrow(Revenue)
[1] 137
> names(Revenue)
 [1] "City"        "City.Group"  "Type"        "P1"          "P2"
 [6] "P3"          "P4"          "P5"          "P6"          "P7"
[11] "P8"          "P9"          "P10"         "P11"         "P12"
[16] "P13"         "P14"         "P15"         "P16"         "P17"
[21] "P18"         "P19"         "P20"         "P21"         "P22"
...
```

Modeling revenue

```
> mod_1 <- lm(revenue ~ City, data = Revenue)
> rsquared(mod_1)
[1] 0.25

> mod_2 <- lm(revenue ~ City * Type, data = Revenue)
> rsquared(mod_2)
[1] 0.32

> mod_3 <- lm(revenue ~ . , data = Revenue)
> rsquared(mod_3)
[1] 0.59

> mod_4 <- lm(revenue ~ City * Type *
                (P6 + P13 + P1 + P2 + P4 + P28 + P25),
              data = Revenue)
> rsquared(mod_4)
[1] 0.75
```

Analysis of variance (ANOVA)

```
> anova(mod_4)
Analysis of Variance Table

Response: revenue
                Df Sum Sq Mean Sq F value Pr(>F)
City            33   7.80   0.236    0.91   0.60
Type             2   1.13   0.564    2.18   0.13
P6               1   0.11   0.112    0.43   0.52
P13              1   0.36   0.365    1.41   0.25
P1               7   1.06   0.152    0.59   0.76
P2               1   0.00   0.000    0.00   0.98
P4               1   0.17   0.173    0.67   0.42
P28              1   0.61   0.611    2.36   0.14
P25              1   0.03   0.029    0.11   0.74
...
```

Let's practice!