



**190F**  
Fall 2018

# Foundations of Data Science

## Lecture 27

---

Correlation

**Prediction**

# Guessing the Future

---

- Based on incomplete information
  - One way of making predictions:
    - To predict an outcome for an individual,
    - find others who are like that individual
    - and whose outcomes you know.
    - Use those outcomes as the basis of your prediction.
-

# Association

# Two Numerical Variables

---

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

**Visualize, then quantify**

(Demo)

---

# Correlation Coefficient

# Definition of $r$

---

**Correlation Coefficient ( $r$ ) =**

average of	product of	x in standard units	and	y in standard units
---------------	------------	---------------------------	-----	---------------------------

Measures how clustered the scatter is around a straight line

---

# The Correlation Coefficient $r$

---

- Measures **linear** association
- Based on standard units
- $-1 \leq r \leq 1$ 
  - $r = 1$ : scatter is perfect straight line sloping up
  - $r = -1$ : scatter is perfect straight line sloping down
- $r = 0$ : No linear association; *uncorrelated*

(Demo)

---



# Watch Out For ...

---

- Nonlinearity
  - Outliers
  - Ecological Correlations
-

# Discussion Question

For each pair, which one will have a higher value of  $r$ ?

