DS-UA 112
Introduction to Data Science

Lecture 5

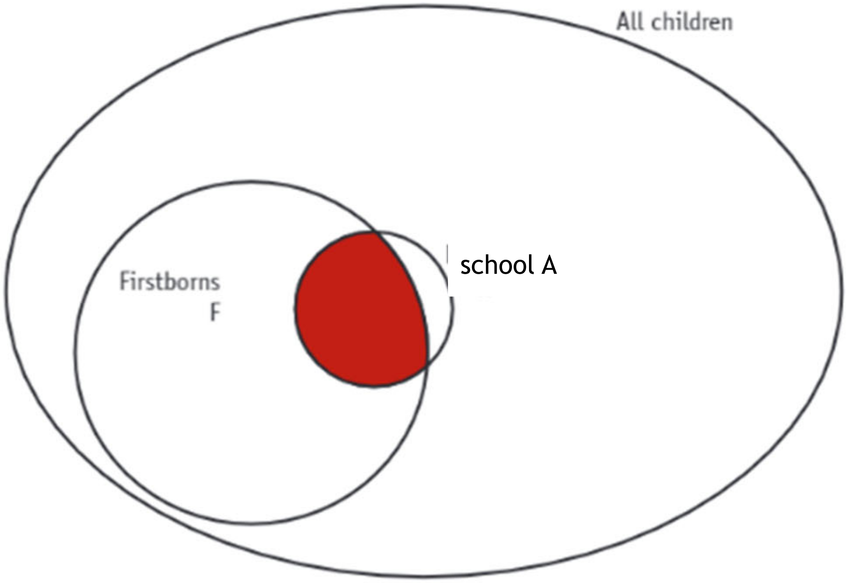
Agenda

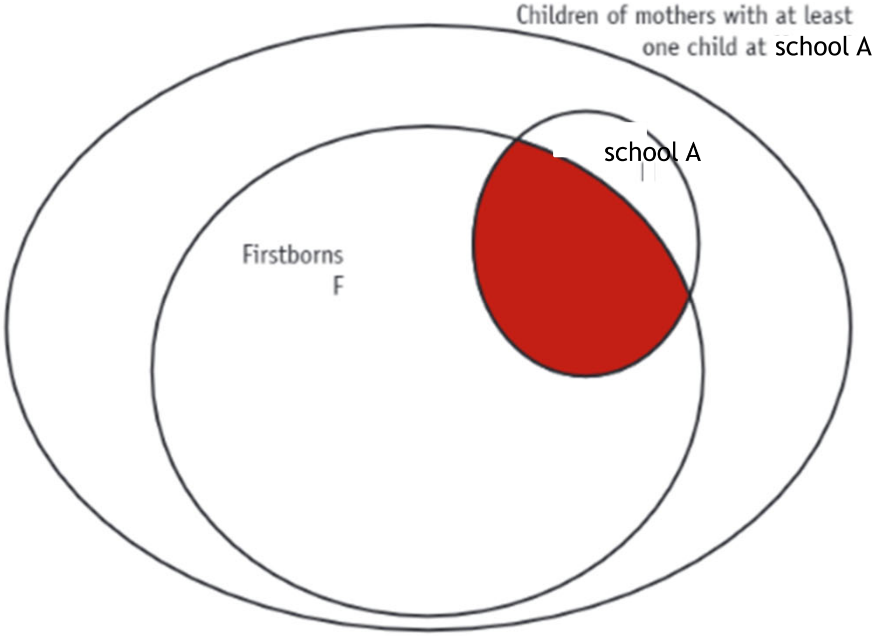
- Review
- Lesson
- Demo

Reminders

- Announcement
  - Section
  - Office Hours
- Survey 2
- Homework
- Lecture
  - Links to Agenda
  - Forum

Review


All children
Firstborns
F
school A

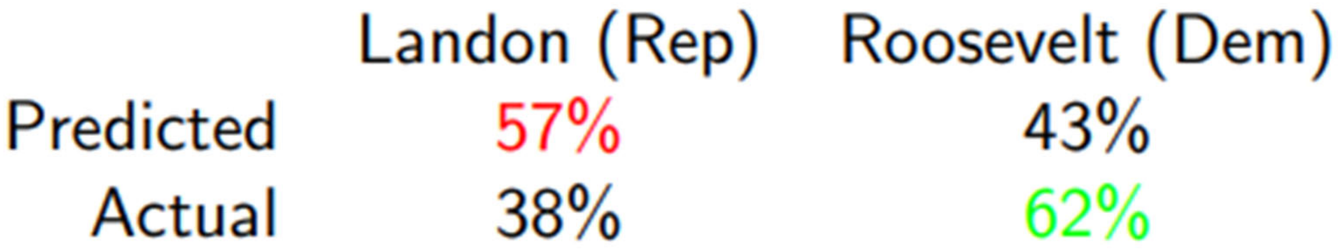

Children of mothers with at least
one child at school A
school A
Firstborns
F

Review

$$
\begin{aligned}
\frac{P\left(A|F\right)}{P\left(A|\text{not }F\right)} &= \frac{P\left(F|A\right)P(A)}{P\left(F\right)}\cdot\left[\frac{P\left(\text{not }F|A\right)P(A)}{P\left(\text{not }F\right)}\right]^{-1}\\
&= \frac{P\left(F|A\right)}{P\left(\text{not }F|A\right)}\cdot\frac{P\left(\text{not }F\right)}{P\left(F\right)}\\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)}\cdot\frac{1-P\left(F\right)}{P\left(F\right)}\\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)}\cdot\left(\frac{1}{P\left(F\right)}-1\right)\\
&= \frac{P\left(F|A\right)}{1-P\left(F|A\right)}\cdot\left(\underset{\uparrow}{\lambda}-1\right)
\end{aligned}
$$

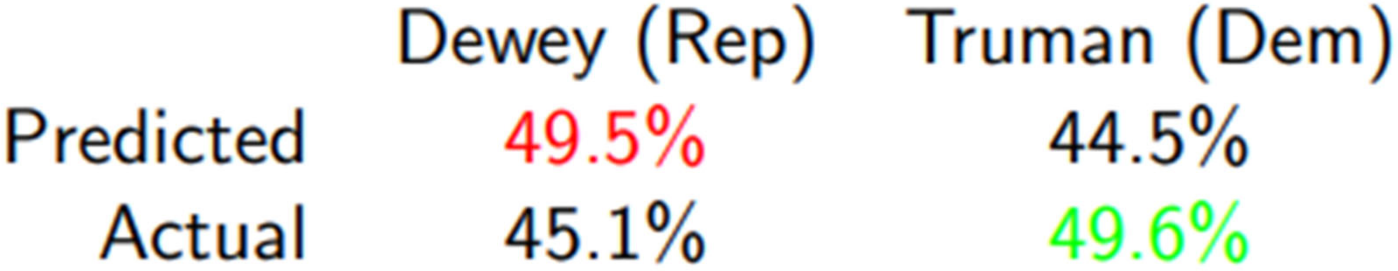
Fertility Rate

Review



	Landon (Rep)	Roosevelt (Dem)
Predicted	57%	43%
Actual	38%	62%

Review



	Dewey (Rep)	Truman (Dem)
Predicted	49.5%	44.5%
Actual	45.1%	49.6%

Review

- Self-selected sample.
  - Sample is whoever chooses to answer.
- Convenience sample
  - Sample is whomever/whatever is convenient for investigator.
- Judgment sample
  - Sample is whomever/whatever investigator deliberately selects

Review

- Probability sample
  - Sample is selected based on probabilistic procedure.
  - Assigns precise probability to the event that each particular sample is drawn from the population
  - This allows to quantify uncertainty/confidence about a prediction

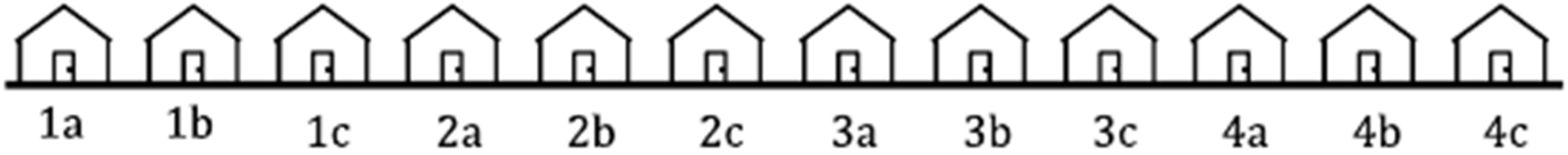
Review

- Probability sample
  - Simple Random Sample

1a
1b
1c
2a
2b
2c
3a
3b
3c
4a
4b
4c

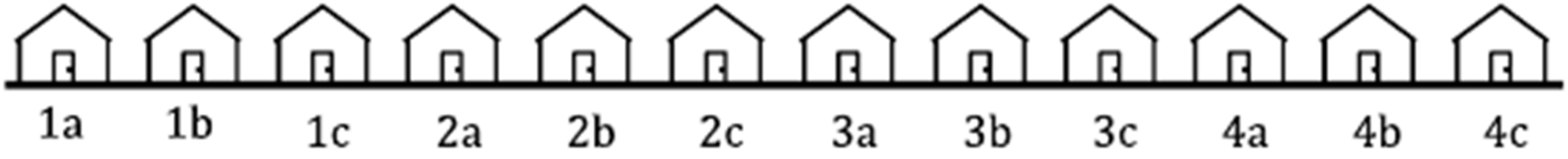
Review

- Probability sample
  - Simple Random Sample
  - Cluster Sample



Review

- Probability sample
  - Simple Random Sample
  - Cluster Sample
  - Stratified Sample



Lesson

Every analysis starts by drawing a data sample ***S*** from a population ***D.***

Lesson

Every analysis starts by drawing a data sample S from a population D.

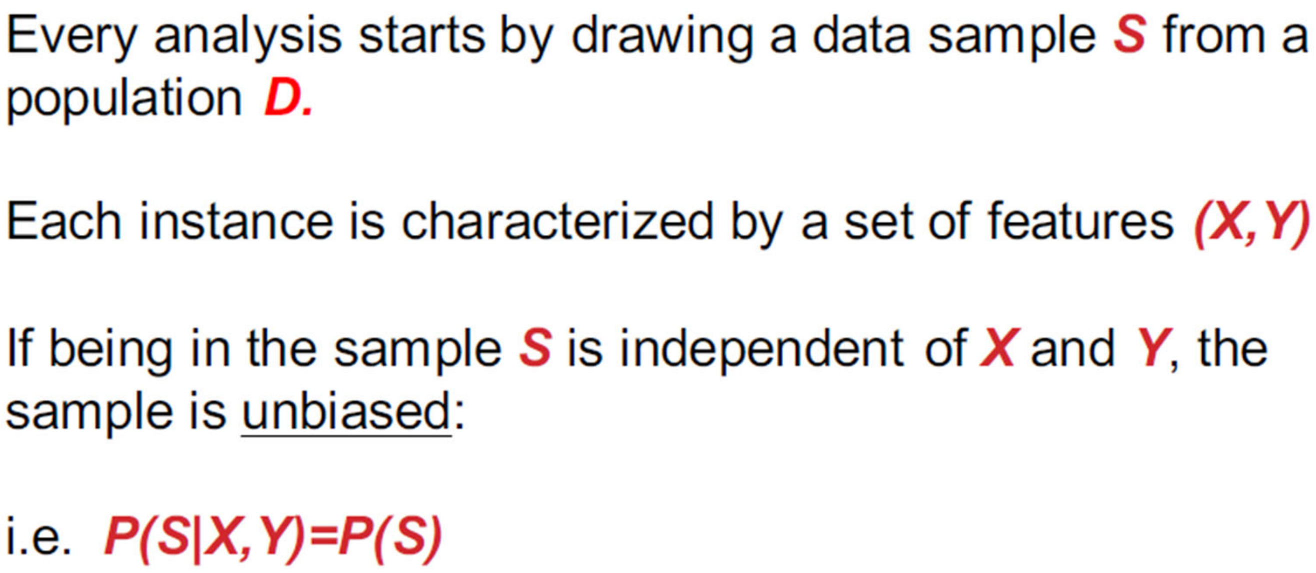
Each instance is characterized by a set of features (X,Y)

Lesson

Every analysis starts by drawing a data sample ***S*** from a population ***D.***

Each instance is characterized by a set of features ***(X,Y)***

Lesson



Every analysis starts by drawing a data sample S from a population D.

Each instance is characterized by a set of features (X,Y)

If being in the sample S is independent of X and Y, the sample is unbiased:

i.e. $P(S|X,Y)=P(S)$

Lesson

Every analysis starts by drawing a data sample S from a population D.

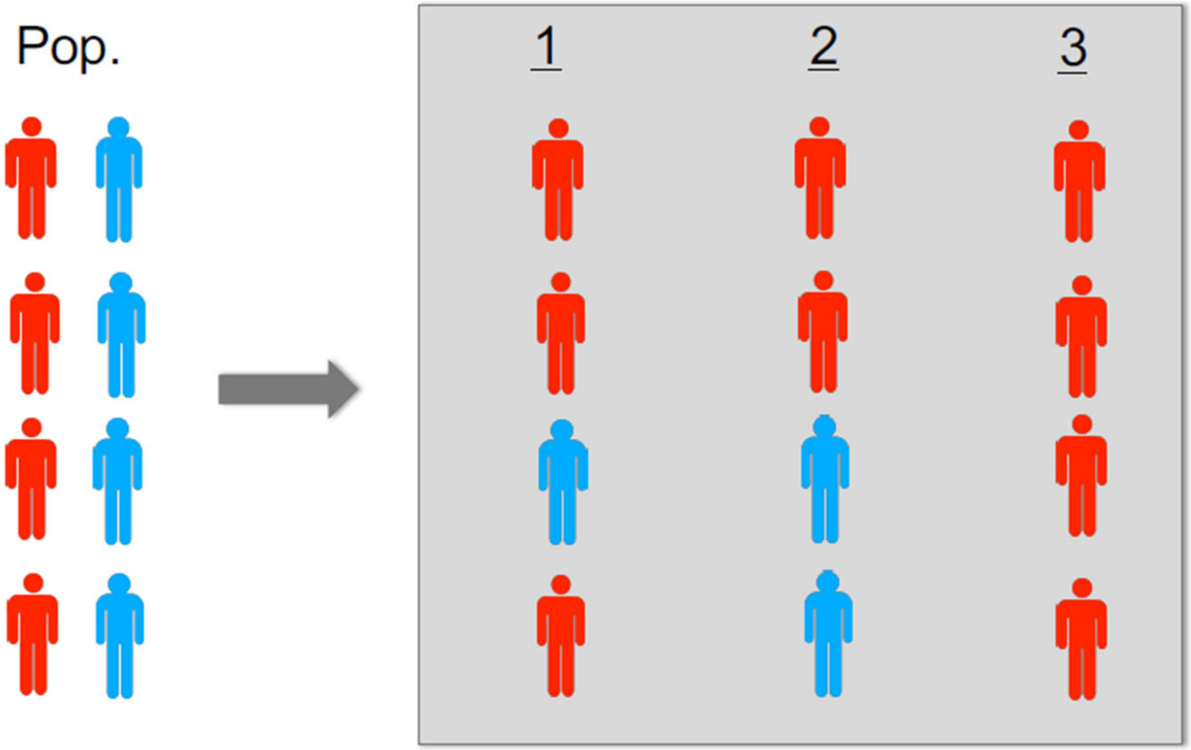
Each instance is characterized by a set of features (X,Y)

If being in the sample S is independent of X and Y, the sample is unbiased:

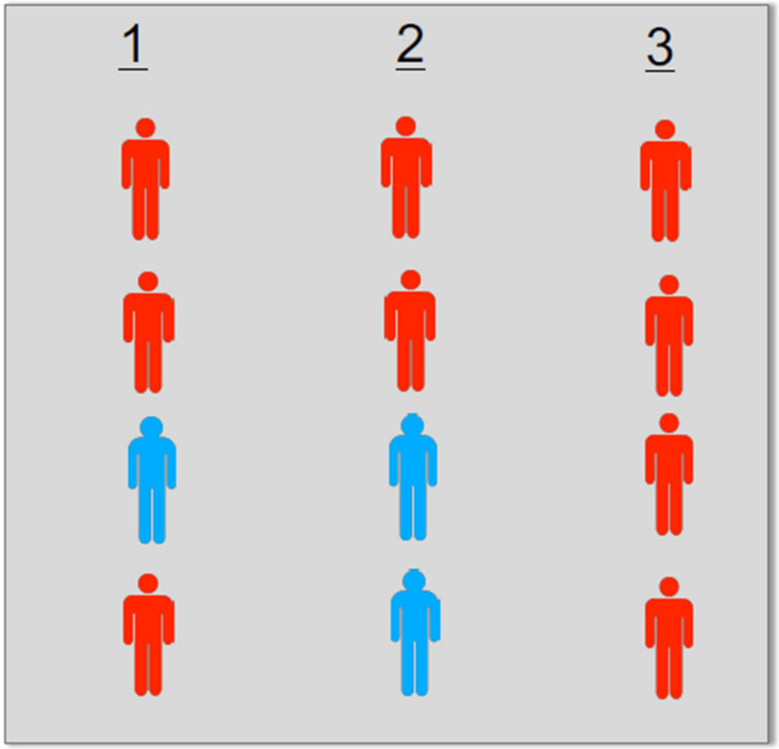
i.e. $P(S|X,Y)=P(S)$

Else the sample is biased: i.e. $P(S|X,Y)\neq P(S)$

Lesson



Lesson



$P(S1) = 0.5$
$P(S1|R) = 0.75$
$P(S1|B) = 0.25$ ✗

$P(S2) = 0.5$
$P(S2|R) = 0.5$
$P(S2|B) = 0.5$ ✓

$P(S3) = 0.5$
$P(S3|R) = 1$
$P(S3|B) = 0$ ✗

Lesson

- What to do about bias?
  - Avoid it

Lesson

- What to do about bias?
  - Avoid it
  - Adjust it

Lesson

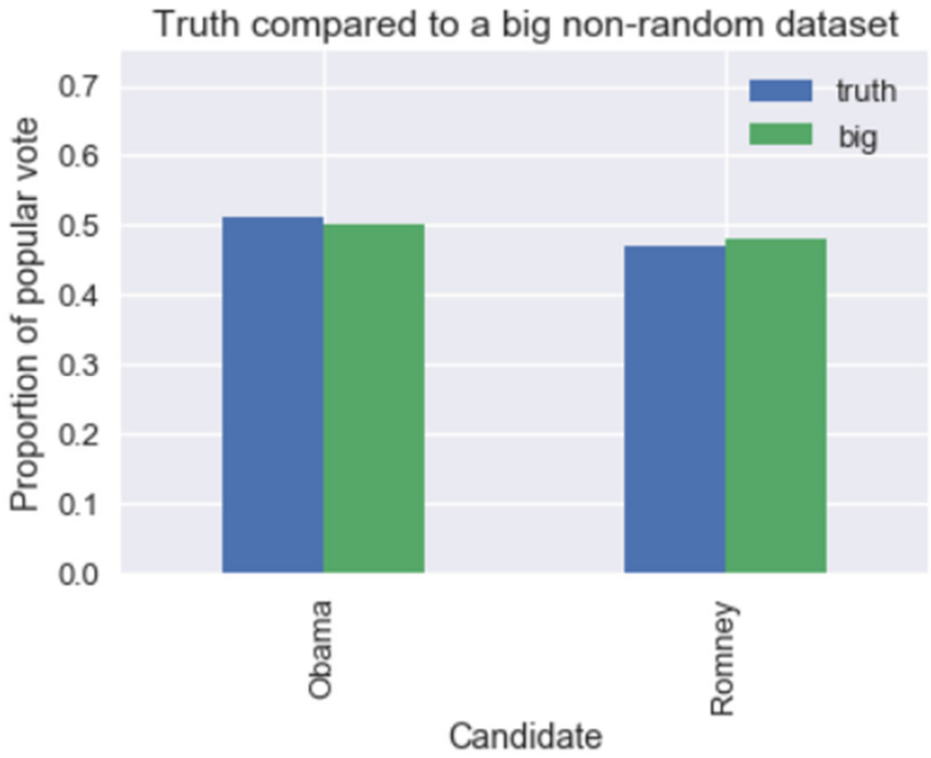
- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it

Lesson

- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it
    - Generalizability
    - Identifiability

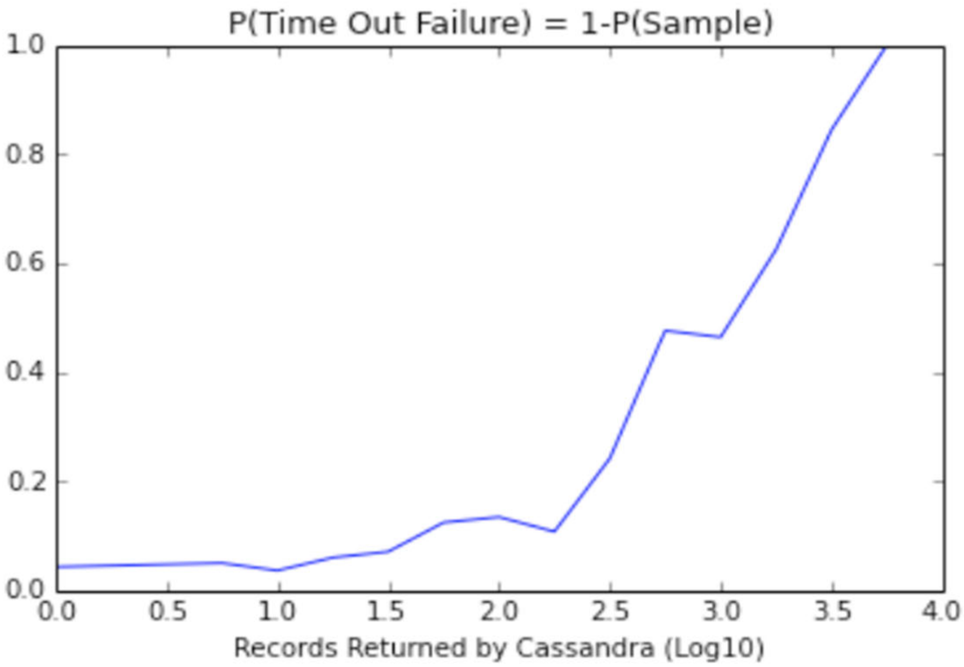
Lesson

- How to Adjust It?
  - Can large amounts of data correct for bias?



Lesson

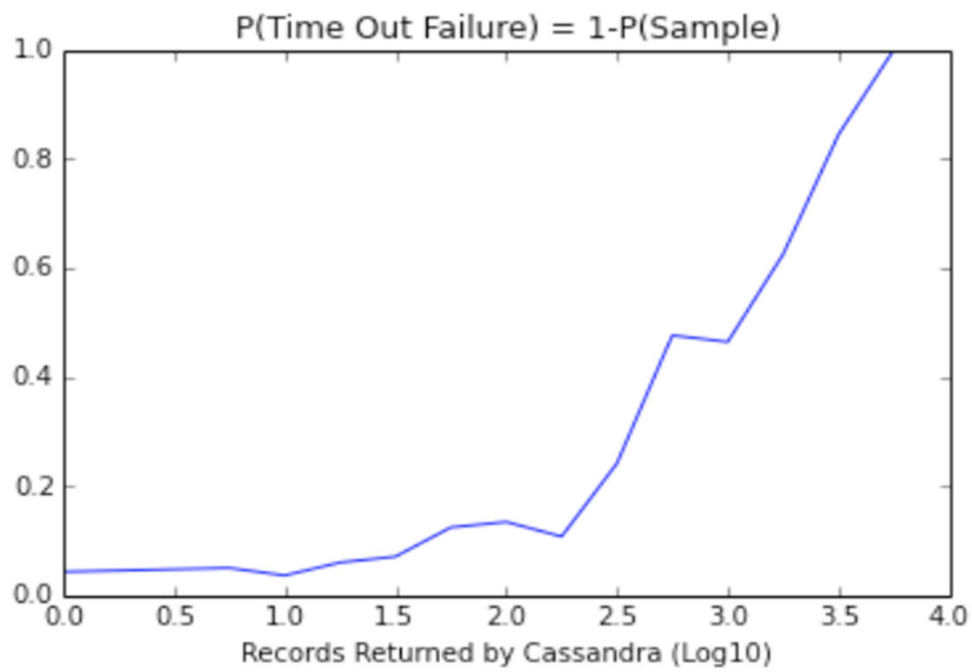
- How to Adjust It?
  - Can we “rescale” the probabilities?



Lesson

- How to Adjust It?
  - Can we “rescale” the probabilities?

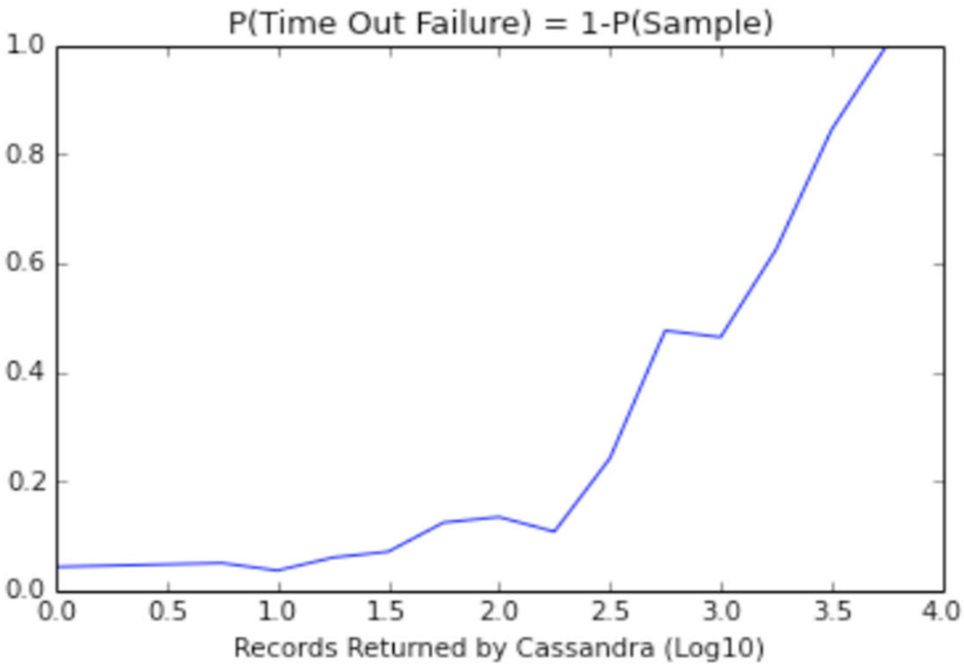
Why would predictions about behavior be more accurate for newer customers?



Lesson

- How to Adjust It?
  - Can we “rescale” the probabilities?

Why would predictions about behavior be more accurate for newer customers?

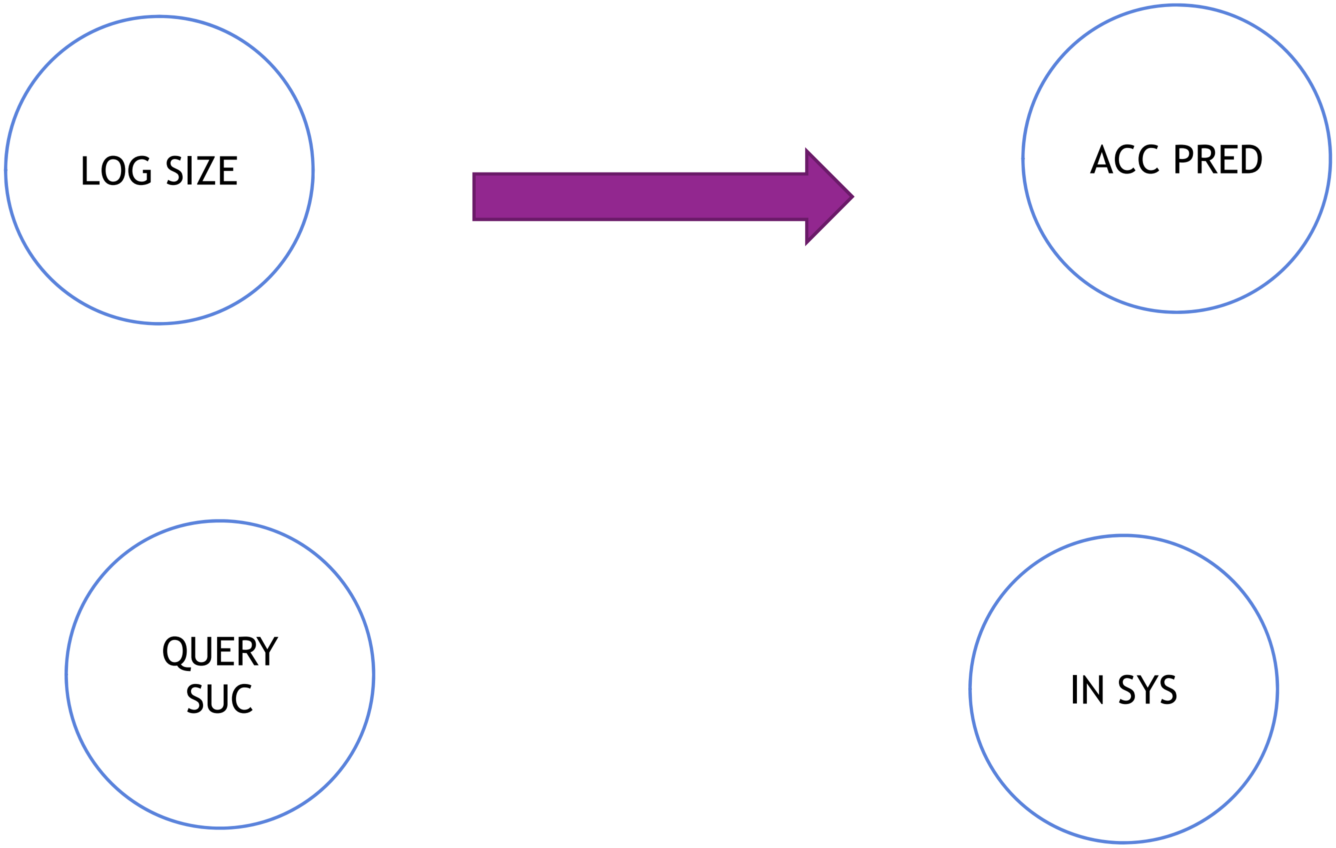


Its pretty clear that $P(S|X) \neq P(S)$, where X is the number of records attached to the user.

Lesson

- ACC PRED
  - Accurate Prediction of Customer Behavior
- LOG SIZE
  - Size of Database Entry
- QUERY SUC
  - Whether Researcher waited for Query Results
- IN SYS
  - Whether Researcher entered Database Entry into Prediction System

Lesson



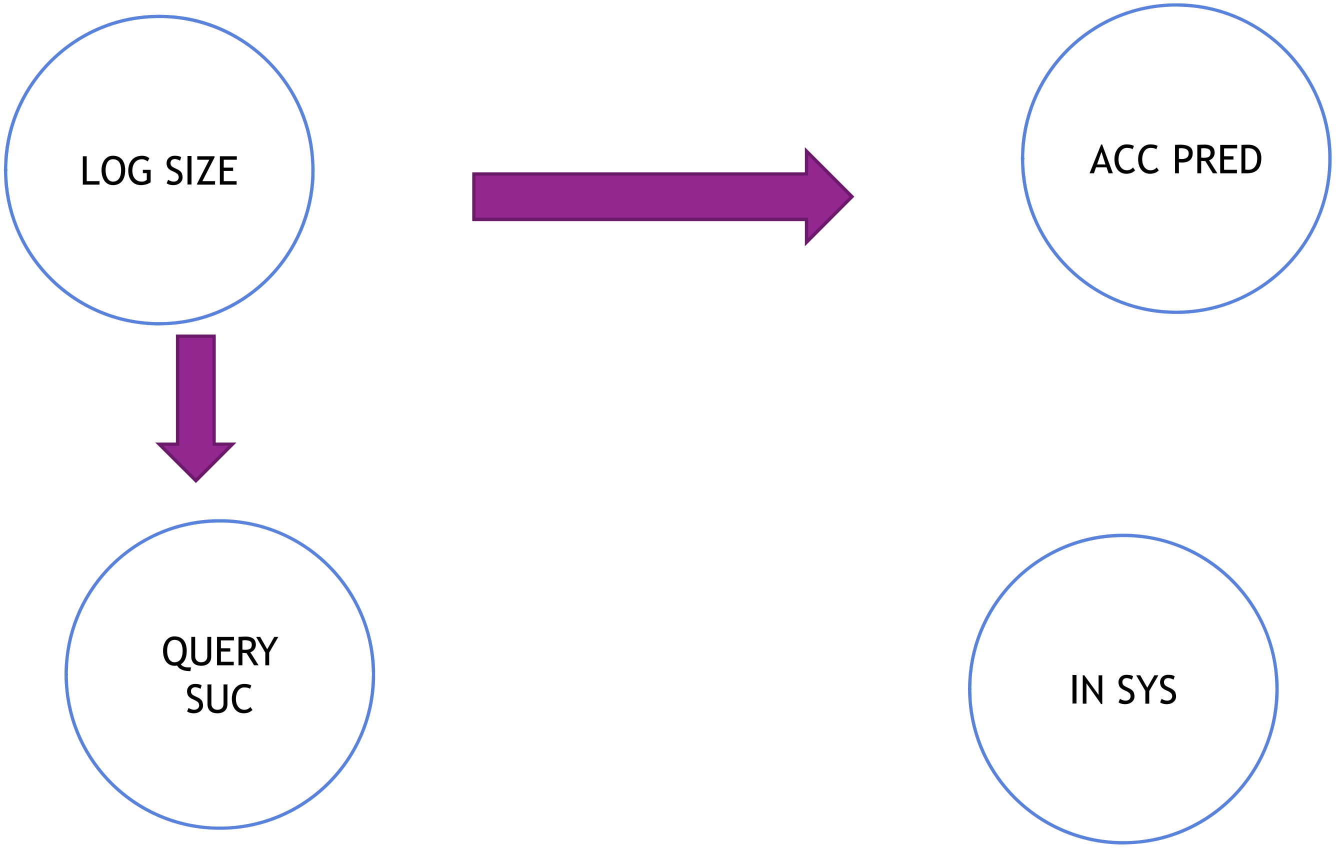
LOG SIZE

ACC PRED

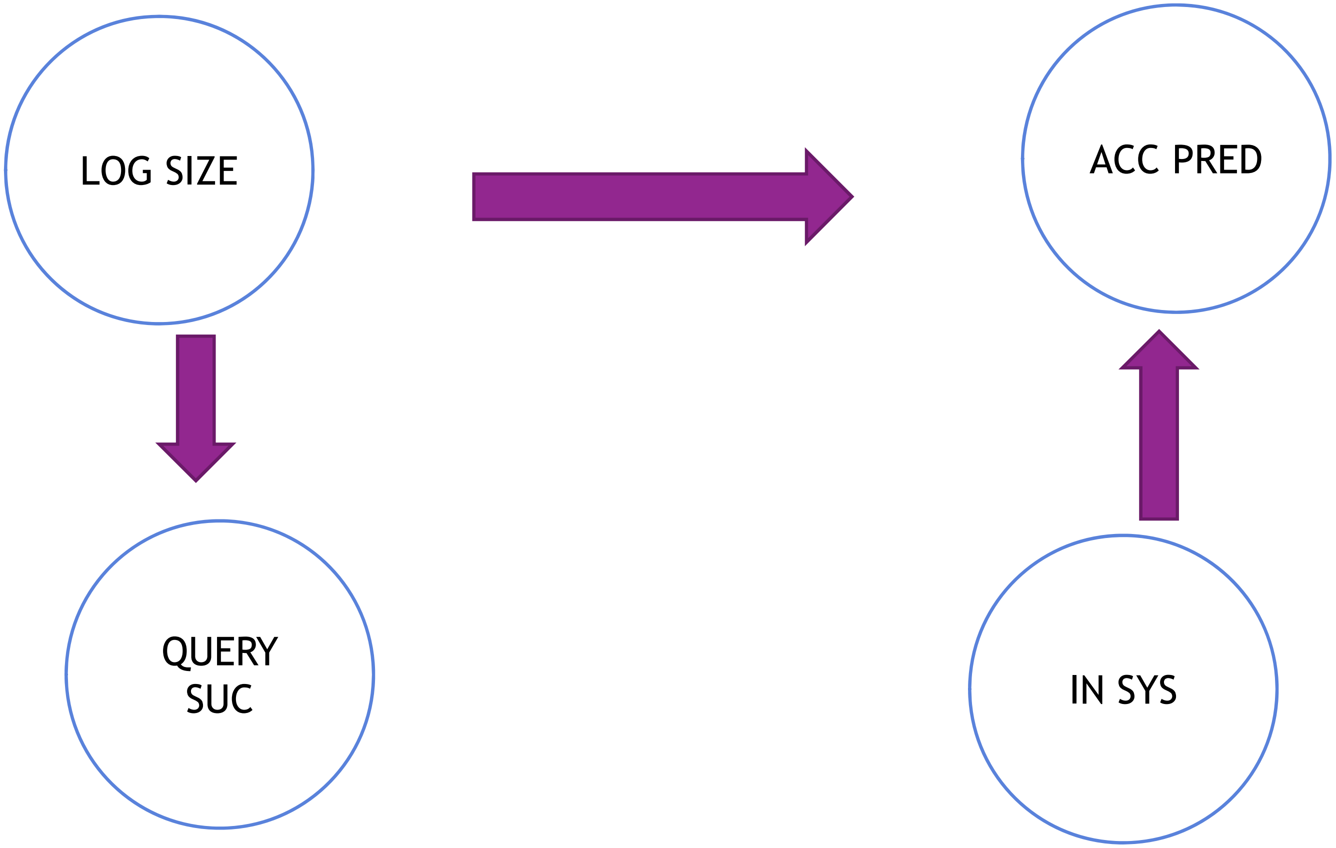
QUERY SUC

IN SYS

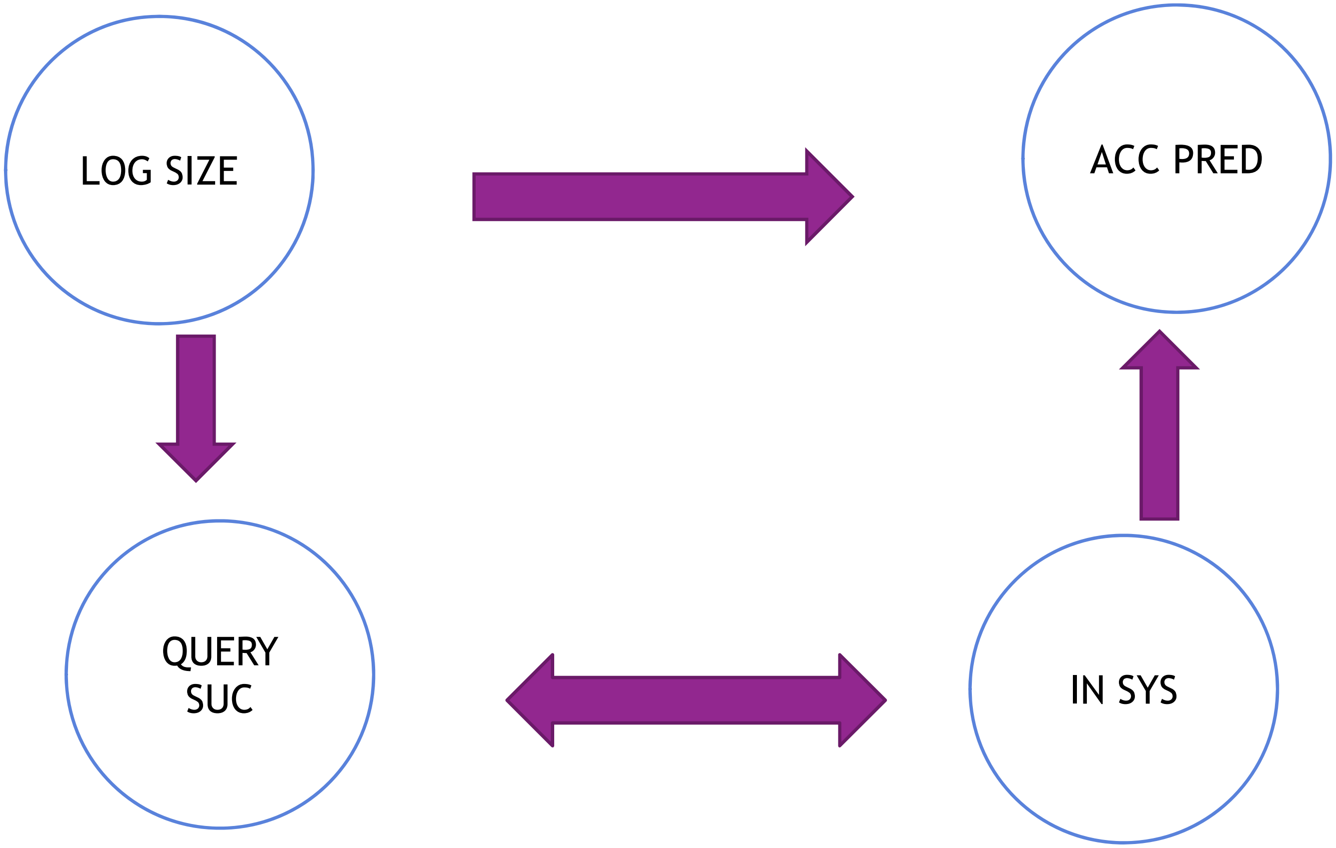
Lesson



Lesson



Lesson



Lesson

P(ACC PRED | LOG SIZE) =

Lesson

P(ACC PRED | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE,NOT QUERY SUC)P(NOT QUERY SUC | LOG SIZE) =

Lesson

P(ACC PRED | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE,NOT QUERY SUC)P(NOT QUERY SUC | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE, NOT IN SYS)P(NOT QUERY SUC | LOG SIZE) =

Lesson

P(ACC PRED | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE,NOT QUERY SUC)P(NOT QUERY SUC | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

(0) P(NOT QUERY SUC | LOG SIZE)

Lesson

$$P(\text{ACC PRED} \mid \text{LOG SIZE}) =$$

$$P(\text{ACC PRED} \mid \text{LOG SIZE, QUERY SUC})P(\text{QUERY SUC} \mid \text{LOG SIZE})$$

$$+$$

$$P(\text{ACC PRED} \mid \text{LOG SIZE,NOT QUERY SUC})P(\text{NOT QUERY SUC} \mid \text{LOG SIZE}) =$$

$$P(\text{ACC PRED} \mid \text{LOG SIZE, QUERY SUC})P(\text{QUERY SUC} \mid \text{LOG SIZE})$$

$$+$$

$$0$$

Lesson

$$P(\text{ACC PRED} \mid \text{LOG SIZE})$$

$$\neq$$

$$P(\text{ACC PRED} \mid \text{LOG SIZE, QUERY SUC})$$

Lesson

$$P(\text{ACC PRED} \mid \text{LOG SIZE})$$

$$=$$

$$P(\text{ACC PRED} \mid \text{LOG SIZE, QUERY SUC})P(\text{QUERY SUC} \mid \text{LOG SIZE})$$

Lesson

$$\frac{P(\text{ACC PRED} \mid \text{LOG SIZE})}{P(\text{QUERY SUC} \mid \text{LOG SIZE})} = P(\text{ACC PRED} \mid \text{LOG SIZE, QUERY SUC})$$