

# EDCT GE2550: DATA SCIENCE IN EDUCATION

Big Data, Learning Analytics & The Information Age

4/8/16 10:42 AM

# In the news

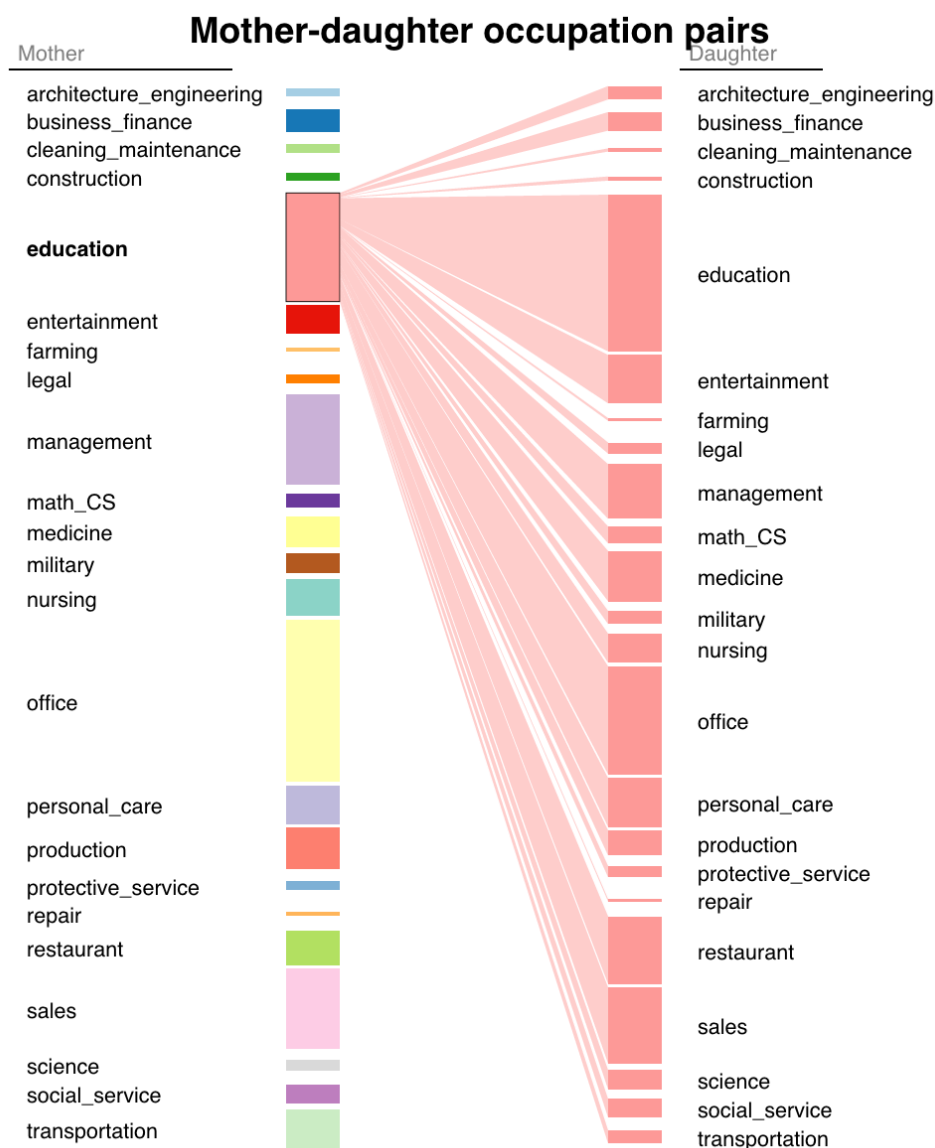


Research at Facebook

CRUNCH NETWORK

Your mobile phone bill is sparking an edtech renaissance

Posted Mar 15, 2016 by [Jeremy Friedman](#)



## Educators embrace technology at Google Summit

March 22, 2016 / by [Jeannette.Cruz](#) / 0 Comment

## EdTech: Mooc Sites Are Forcing Elite B-Schools To Embrace Digital Innovation

Silicon Valley start-ups have shaken up entire university model

Written by [Seb Murray](#) | MBA Distance Learning | Wednesday 23rd March 2016 00:19:00 GMT

<https://research.facebook.com/blog/do-jobs-run-in-families-/>

# Today

In the news

6:45 - 6:55

Quiz

6:55 - 7:05

Inference

7:05 - 7:15

Cross Validation

7:15 - 7:30

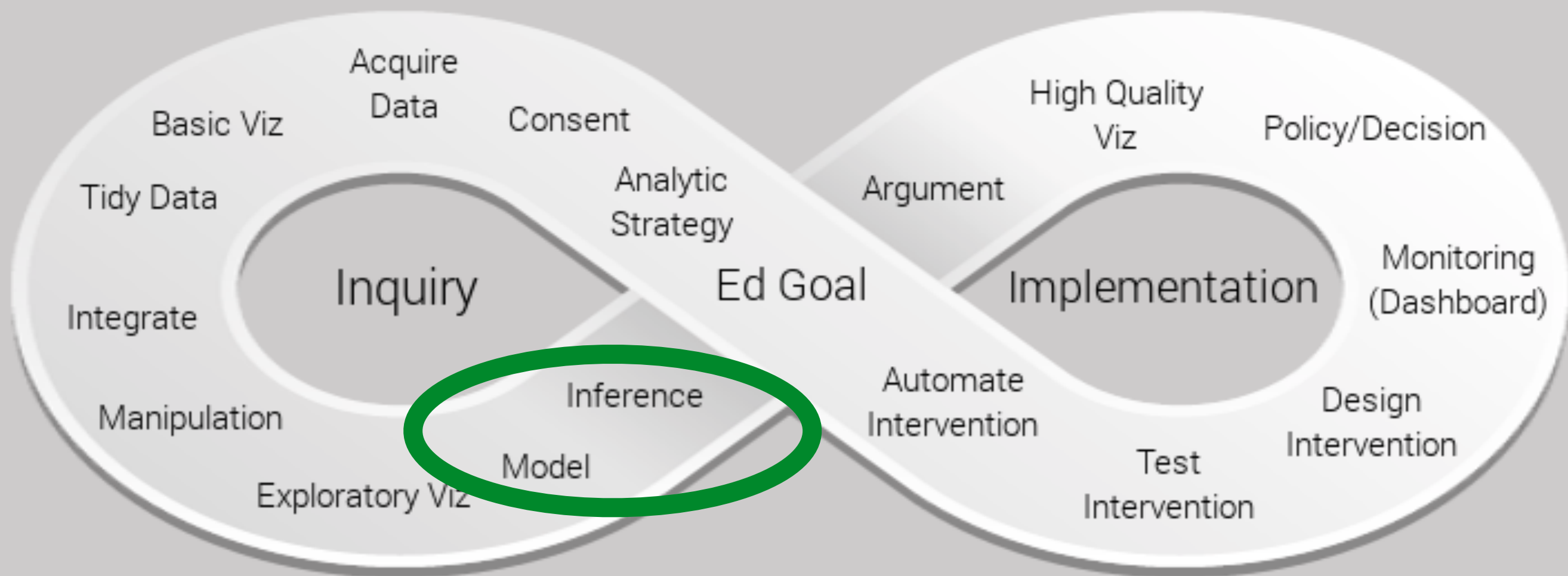
Decision Trees

7:30 - 7:40

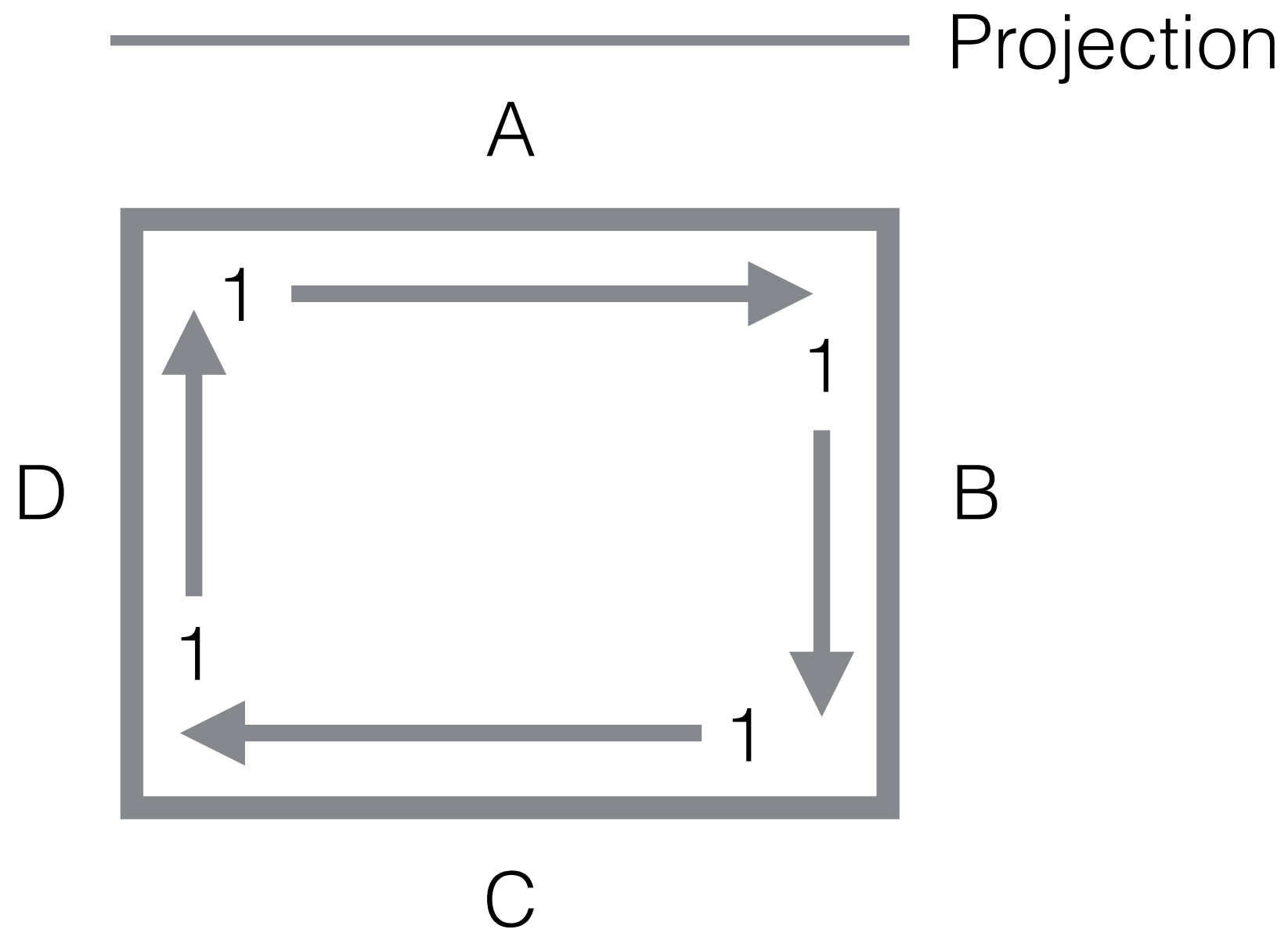
Twitter Game

7:40 - 8:00

# Ed Data Science Cycle



# Quiz



<http://bit.ly/1Udsi80>

# Inference

“A woman's guess is much more accurate than a  
man's certainty.”

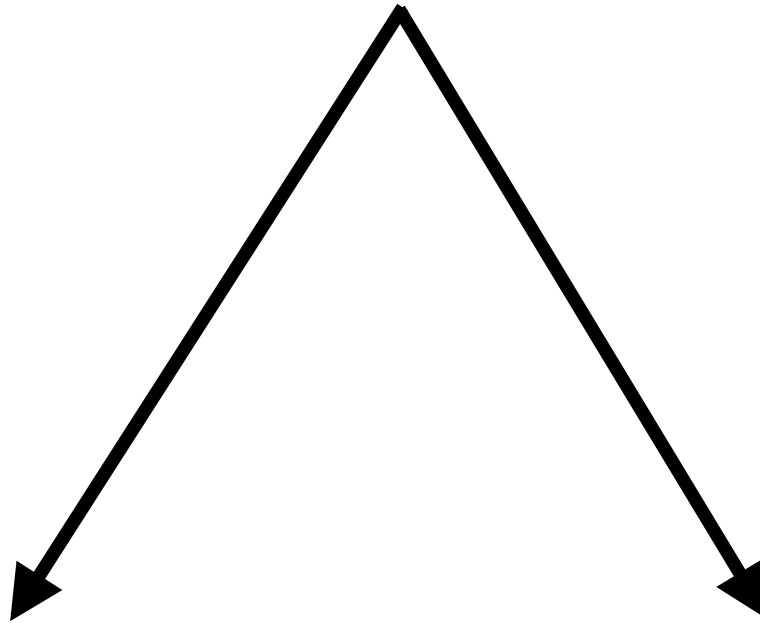
—Rudyard Kipling (Plain Tales from the Hills, 1888)



“Statistics is the study of uncertainty”

– LJ Savage, 1977

Variation



Patterns

Uncertainty

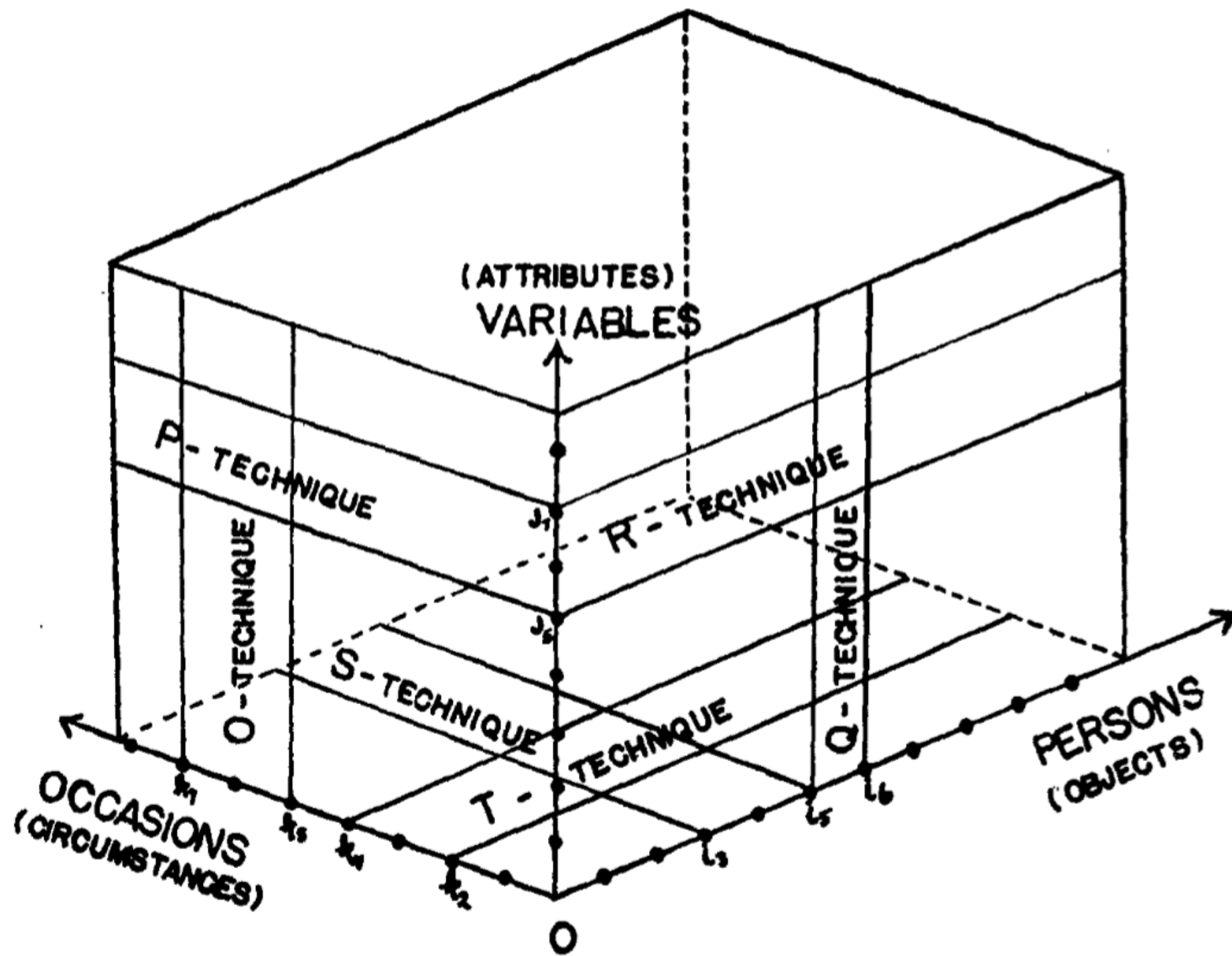
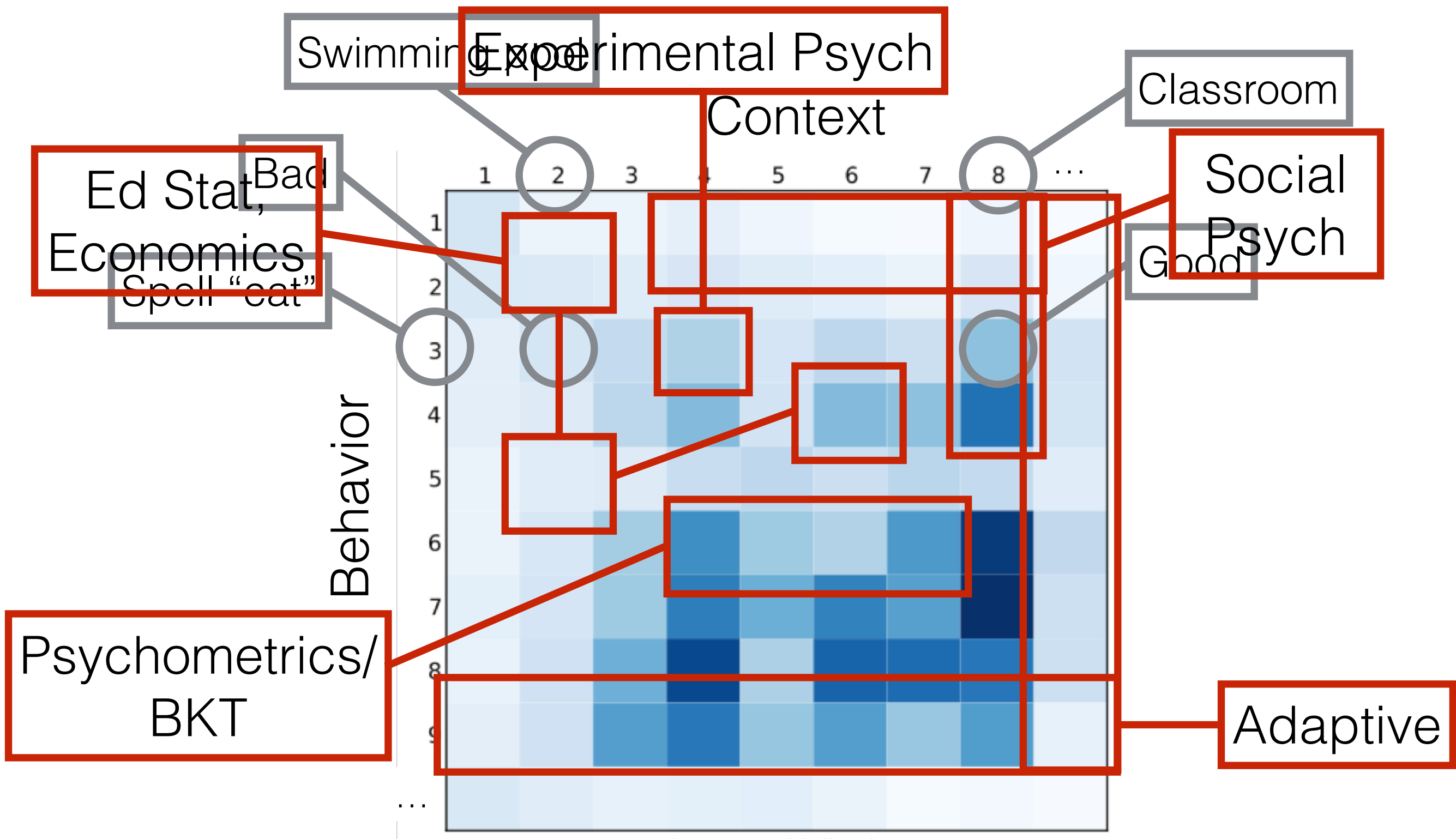
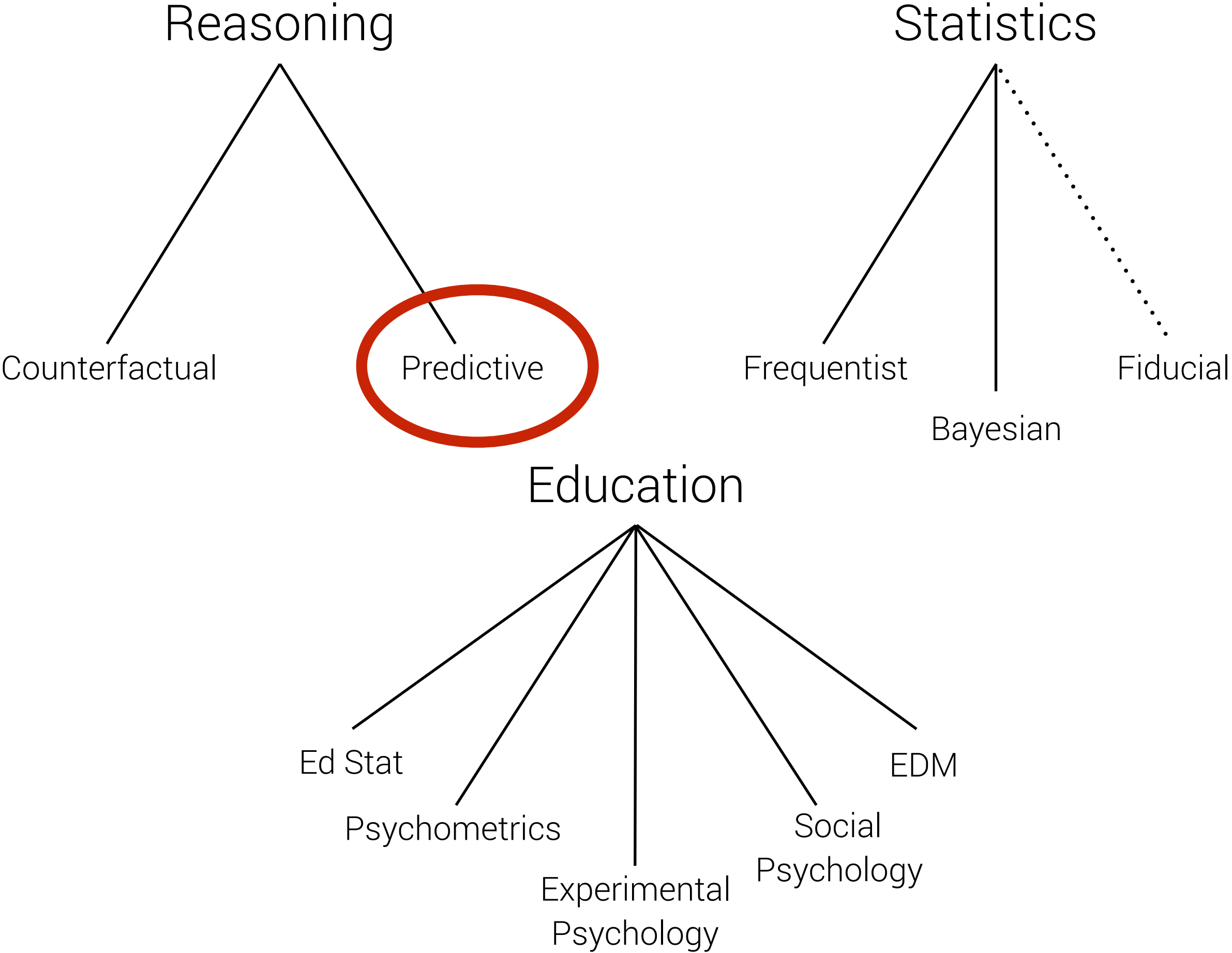
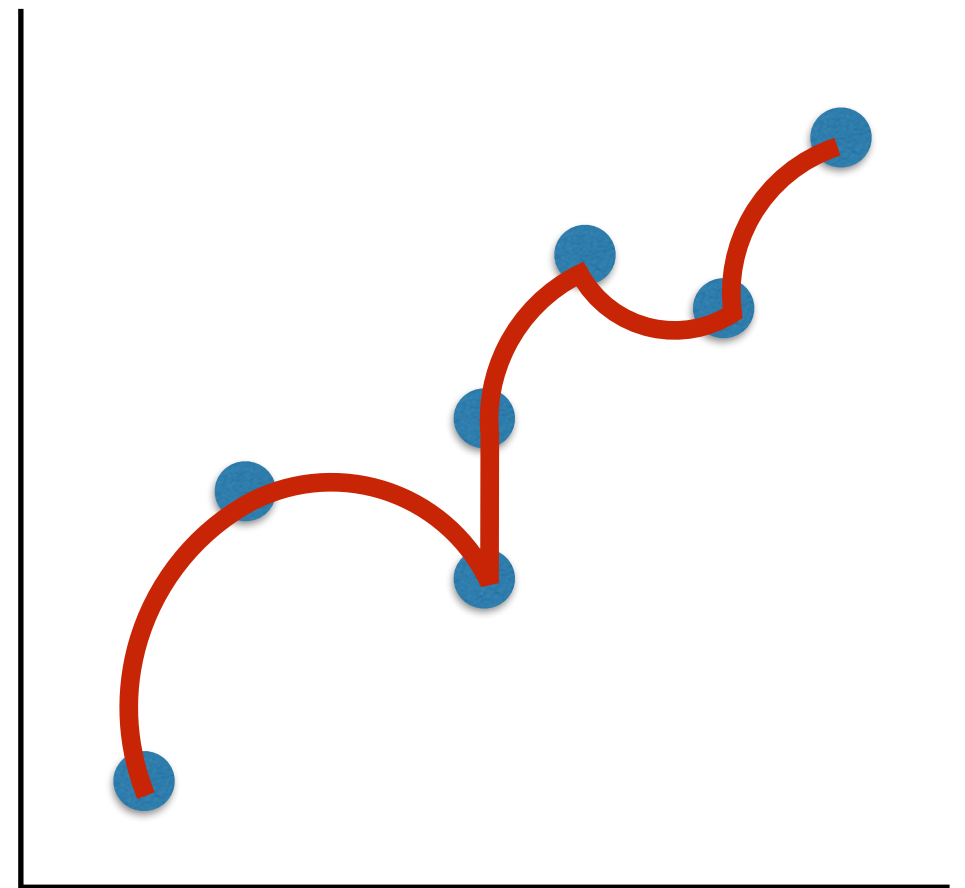
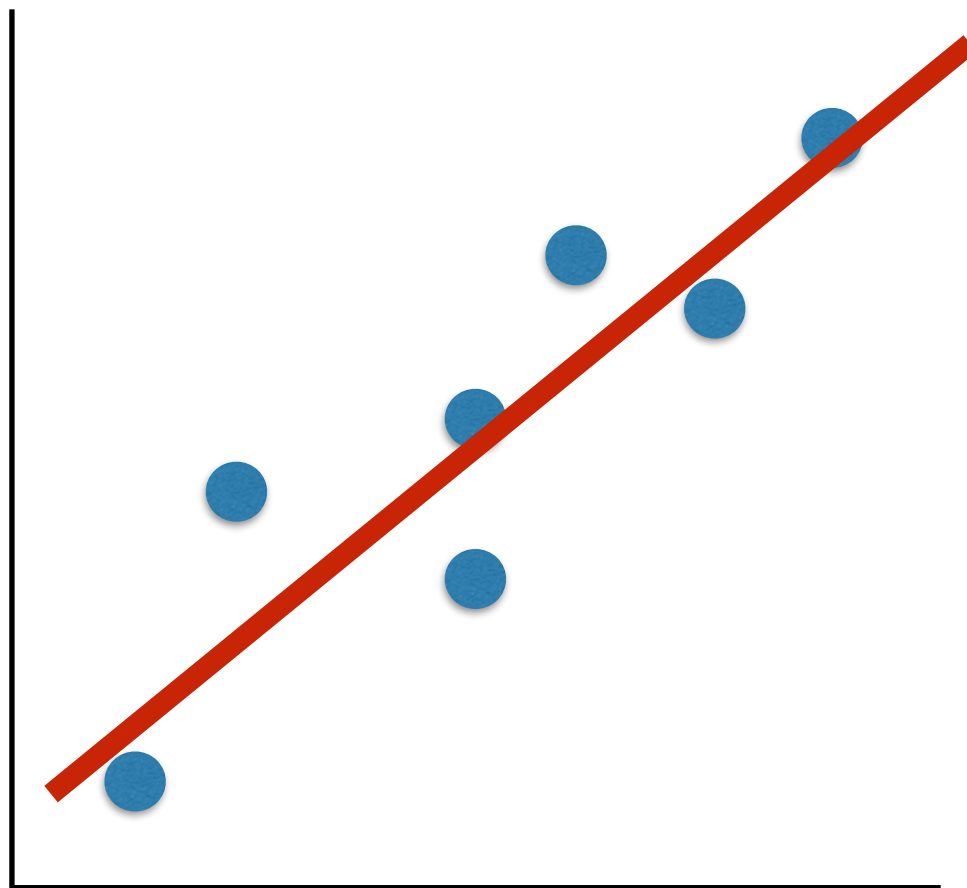


FIG. 1. THE COVARIATION CHART

Cattell, 1952







Which is more “accurate”?

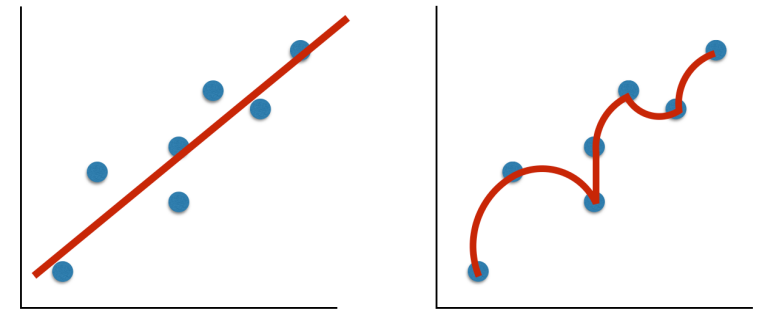
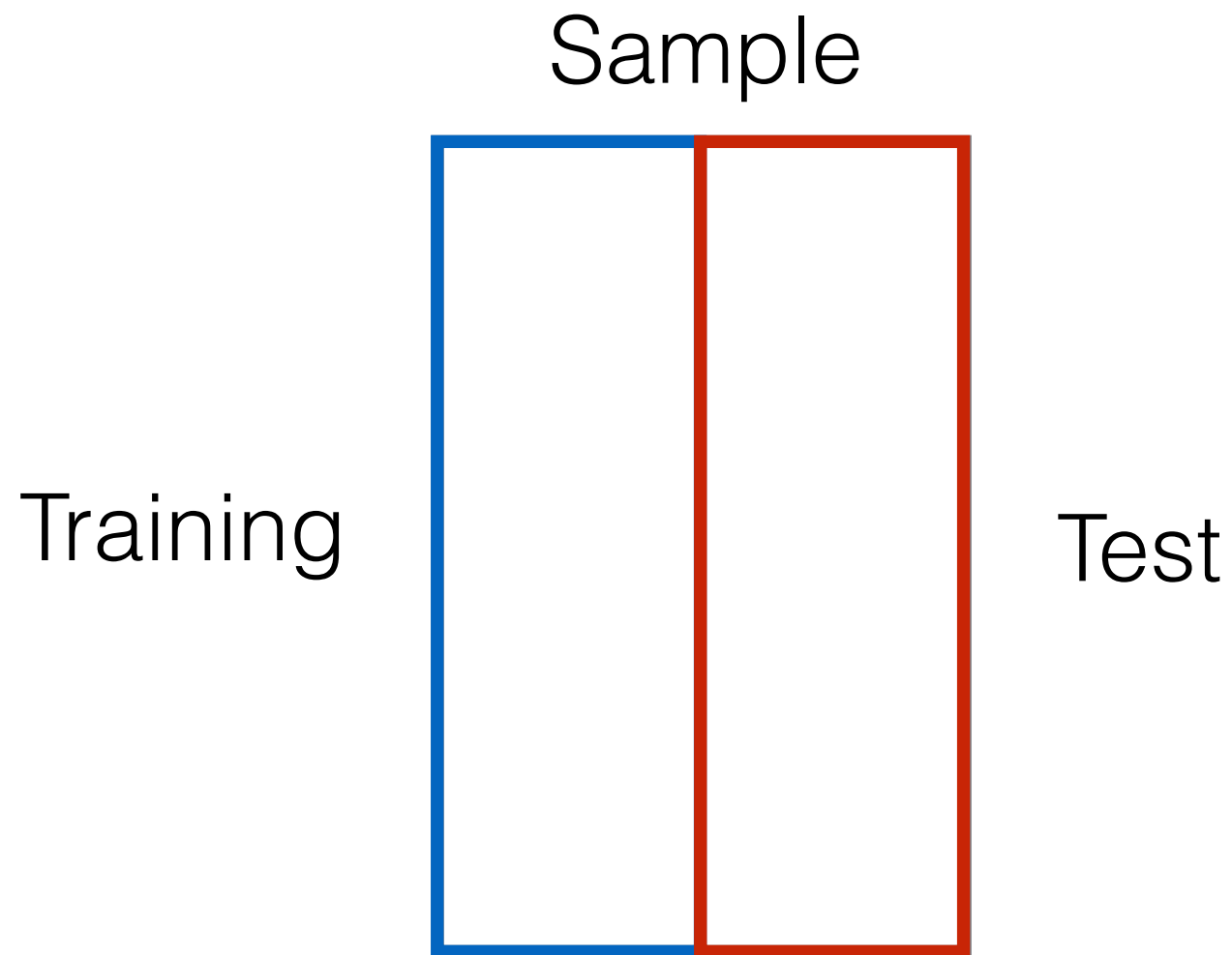
Which is more “useful”?

How can we tell?

# Cross Validation

- Estimate how accurately a predictive model will perform in practice
- Give an insight on how the model will generalize to an independent dataset

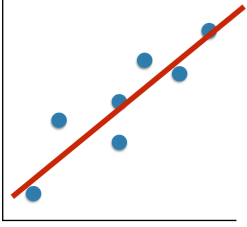
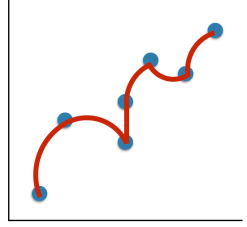
# Hold-out Validation



**Problem:** very dependent on which data are in each group



# K-Fold Cross Validation

	Sample			
Training 1		Test 1	5	2
Training 2		Test 2	4	2
Training 3		Test 3	3	1
Training 4		Test 4	5	4
Training 5		Test 5	4	2
			<hr/>	<hr/>
			4.2	2.2

Calculate how accurate we are in each “fold”  
and average the answer

# Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - a_i)^2}{n}}$$

Square root of the squared distance between the predicted and the observed value

# Activity

1. Everyone choose a preference: cats or dogs
2. Count how many in each category at your table
3. Write on the board the answer for each table
4. Table 1 is the prediction of Table 2, Table 3 is the prediction of Table 4, etc.
5. What is the RMSE of all the predictions?

Models: Decision Trees

[https://youtu.be/d-](https://youtu.be/d-umbokNLGQ)  
[umbokNLGQ](https://youtu.be/d-umbokNLGQ)



Frustrated Student

Does not ask  
for help

Asks for help



Frustrated Student



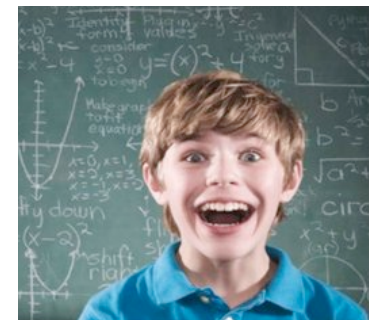
Teacher Responds

Frustration Unresolved

Frustration Resolved



Frustrated Student



Happy Student

1. Pick an attribute
2. Ask a question about it
3. Follow the answer path
4. Go back to 1 and repeat
5. Find climate answer or give up

# Pros & Cons

- Simple and easy to understand and communicate
- Can be used with little data
- Work across a range of domains
- Still useful with messy data
- Calculations get very complex very quickly
- Biased toward variables with many categories