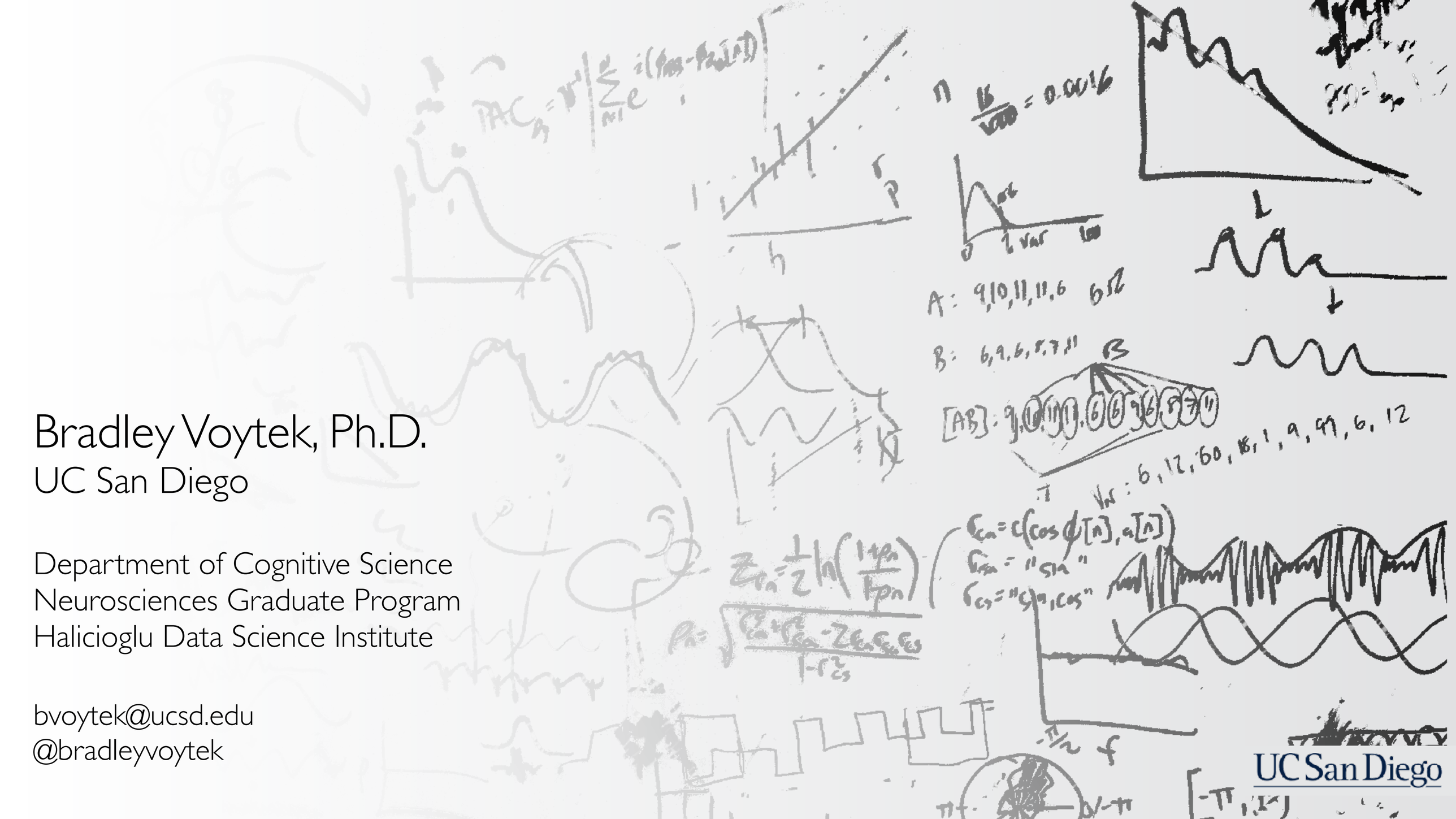


Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

WAITLIST

- Currently ~120 people deep.
- I can't manipulate the list to let anyone in, even if you're a graduating senior, etc.

OTHER QUESTIONS

- Piazza: Not listed there? Let me know!
- Groups: ONLY for Final Project, *not* home works!
  - 3-6 people; you choose your own groups

Sections and homeworks

- Sections are *not* mandatory
- But bring a computer if you can
- Python, Jupyter, and git *before sections*! (We'll post instructions)
- Sections: intend to go to the section for which you're registered, but you can attend another if needed
- Sections are a great place for assignment guidance
- TA office hours will be in the computer lab, giving you more time to work on assignments
- Assignments will be released Monday mornings
- Assignments will be due Sunday nights
- Assignment solutions will be posted one week after deadline
- Late assignments graded at 50% penalty, not late assignments allowed once solutions are posted

A + B
Data
Science
Student Society

December 2017

Harnessing Geospatial Open Data
Tuesday, December 12, 2017 at 6 PM
305 other guests

November 2017

Kaggle Team - Session 4 FA17
Tuesday, November 28, 2017 at 6 PM
282 other guests

Kaggle Team - Session 3 FA17
Tuesday, November 21, 2017 at 7:30 PM
285 other guests

ProductBIO

Making Climate Commitments Possible Thro...
Tuesday, November 21, 2017 at 2 PM
152 guests

COGS 9 - Uber Data Science Presentation
Friday, November 17, 2017 at 10 AM
314 other guests

ML Series: Deep Dreaming + Art Style Tra...
Wednesday, November 15, 2017 at 6:20 PM
307 other guests

Kaggle Team - Session 2 FA17
Tuesday, November 14, 2017 at 6 PM
294 other guests

Lab Open House
Tuesday, November 7, 2017 at 6:30 PM
216 other guests

What does a professor do?

Teaching

Teaching

Research

What does research look like?

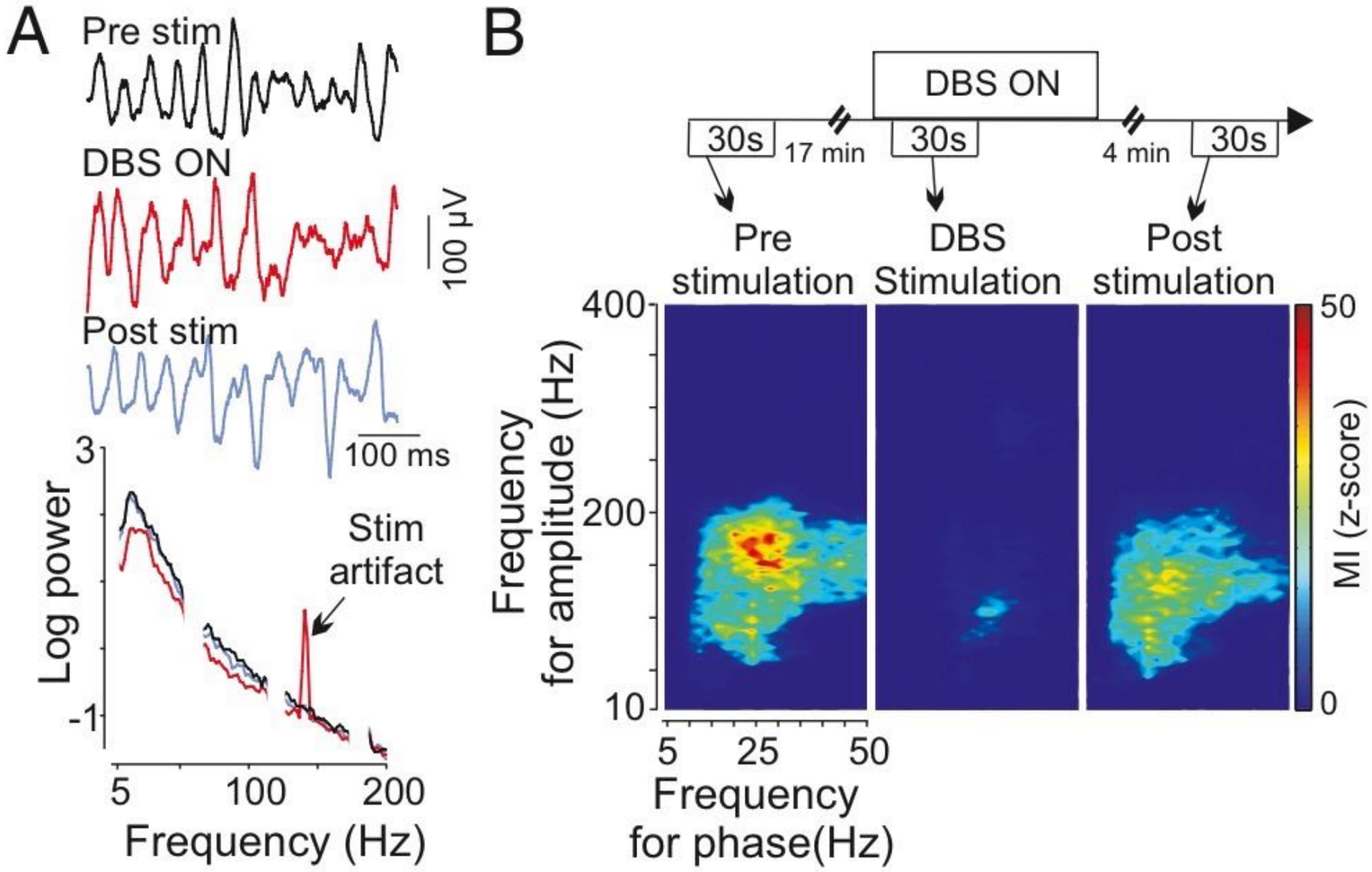
Deep brain stimulation

Deep brain stimulation

WARNING

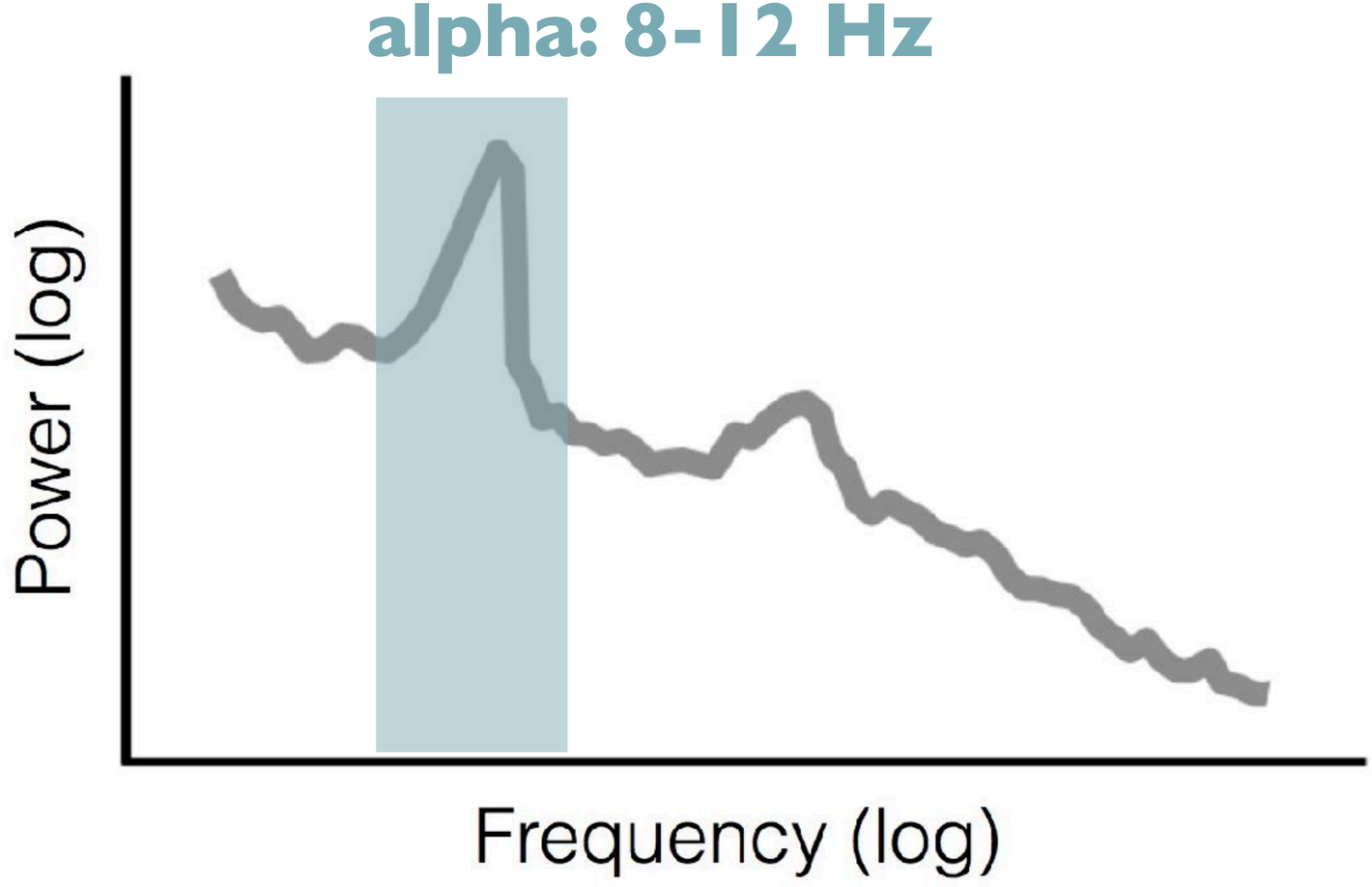
Electrocorticography

Oscillations as pathology

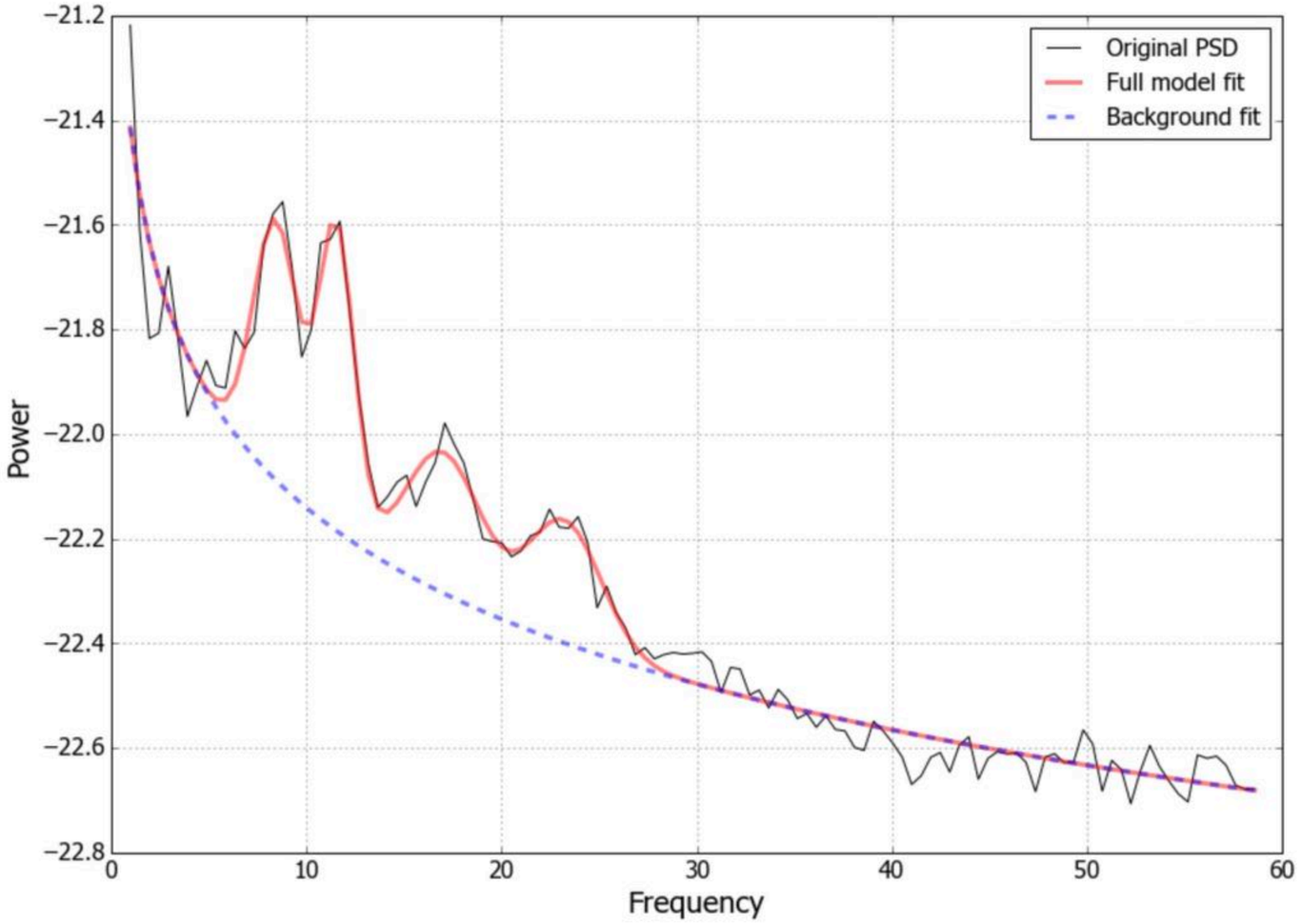


Source: de Hemptinne *et al.*, *PNAS* 2013; de Hemptinne *et al.*, *Nature Neurosci* 2015

Oscillations are *bumps*



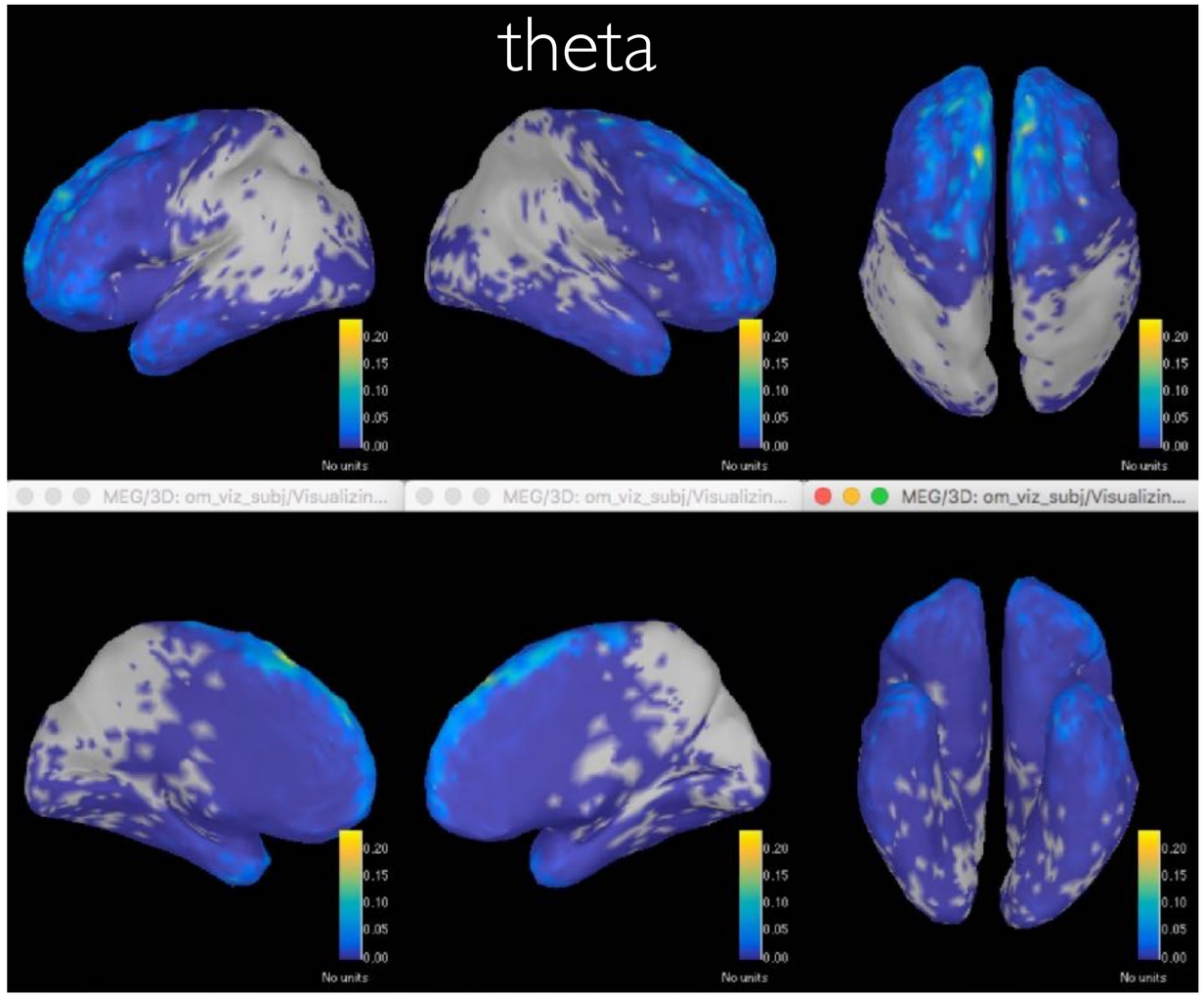
Oscillations are variable

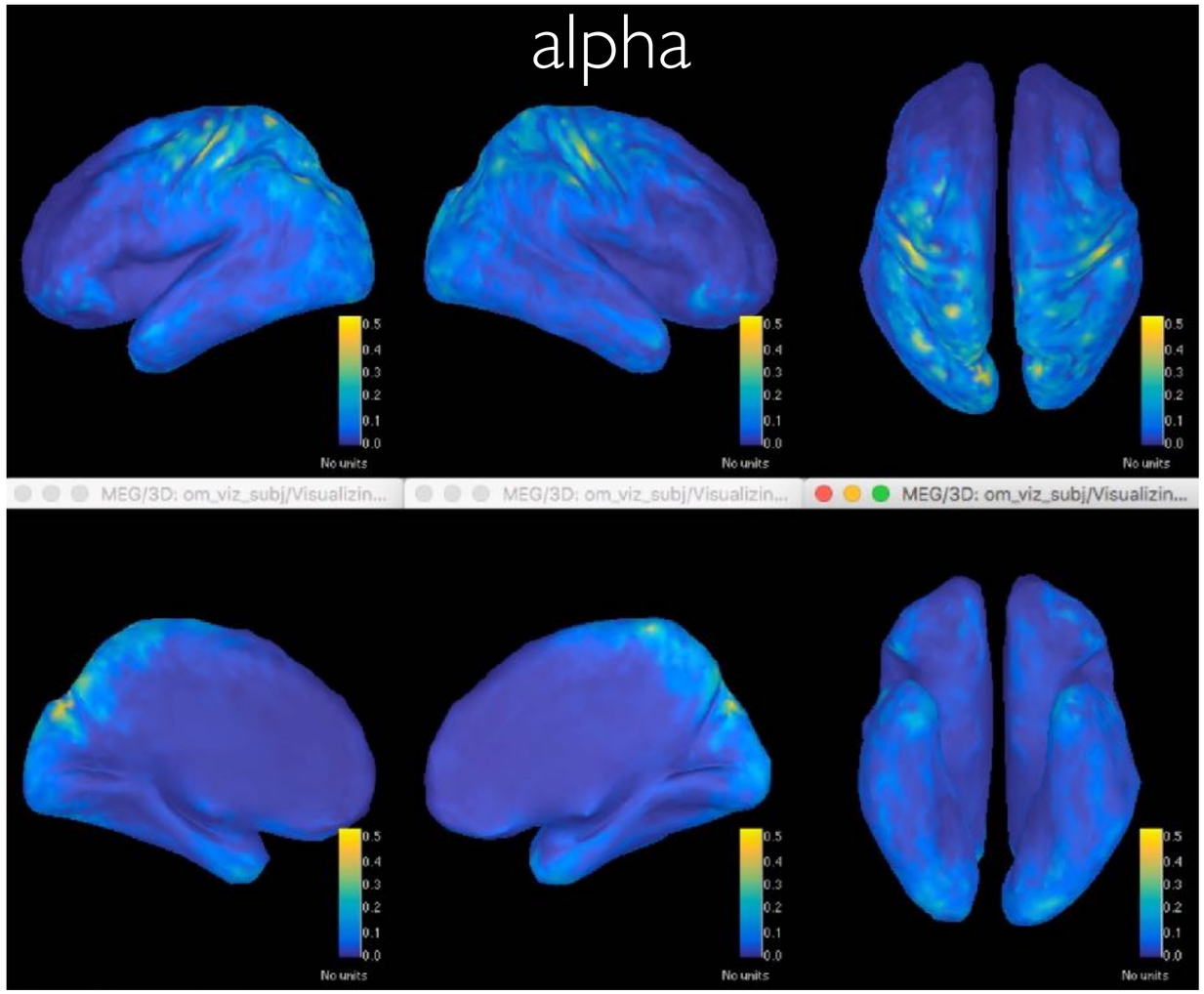



```
==============================================================

                    FOOOF - PSD MODEL

The input PSD was modelled in the frequency range 0.5-59 Hz
                Frequency Resolution is 0.49 Hz

                  Background Parameters:
               -21.4220, -0.7274, 8.53e-03

                  4 oscillations were found:
            CF:   8.46, Amp:  0.495, BW:  2.32
            CF:  11.54, Amp:  0.582, BW:  1.80
            CF:  17.05, Amp:  0.270, BW:  3.89
            CF:  23.24, Amp:  0.234, BW:  3.76

                  R^2 of model fit is 0.9824

          Root mean squared error of model fit is 0.0328

==============================================================
```

Source: Davis, Donoghue, Peterson, *et al.*, (in preparation)

Oscillations are variable

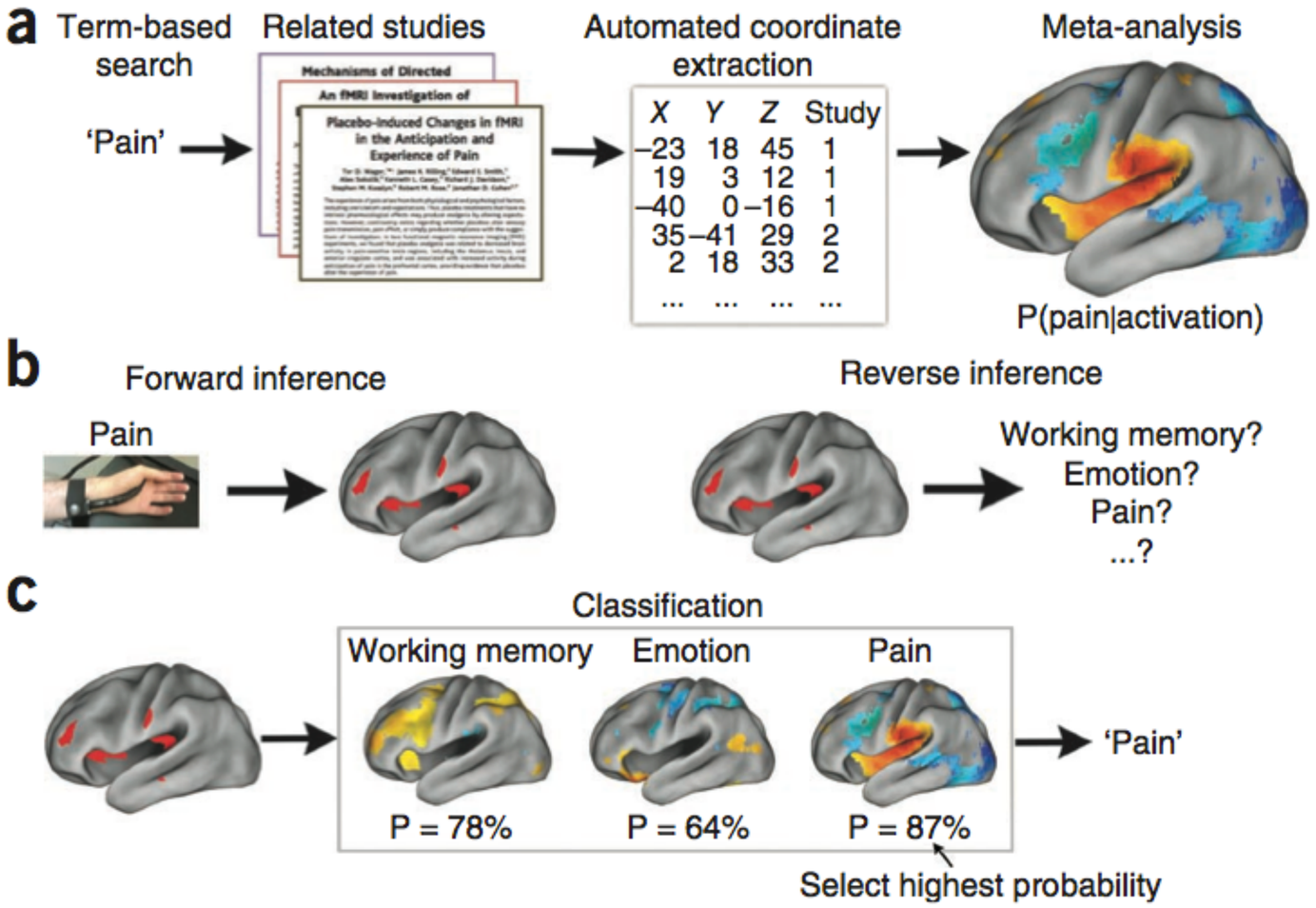




Source: Davis, Donoghue, Peterson *et al.* (in prep); Donoghue *et al.* (in preparation)

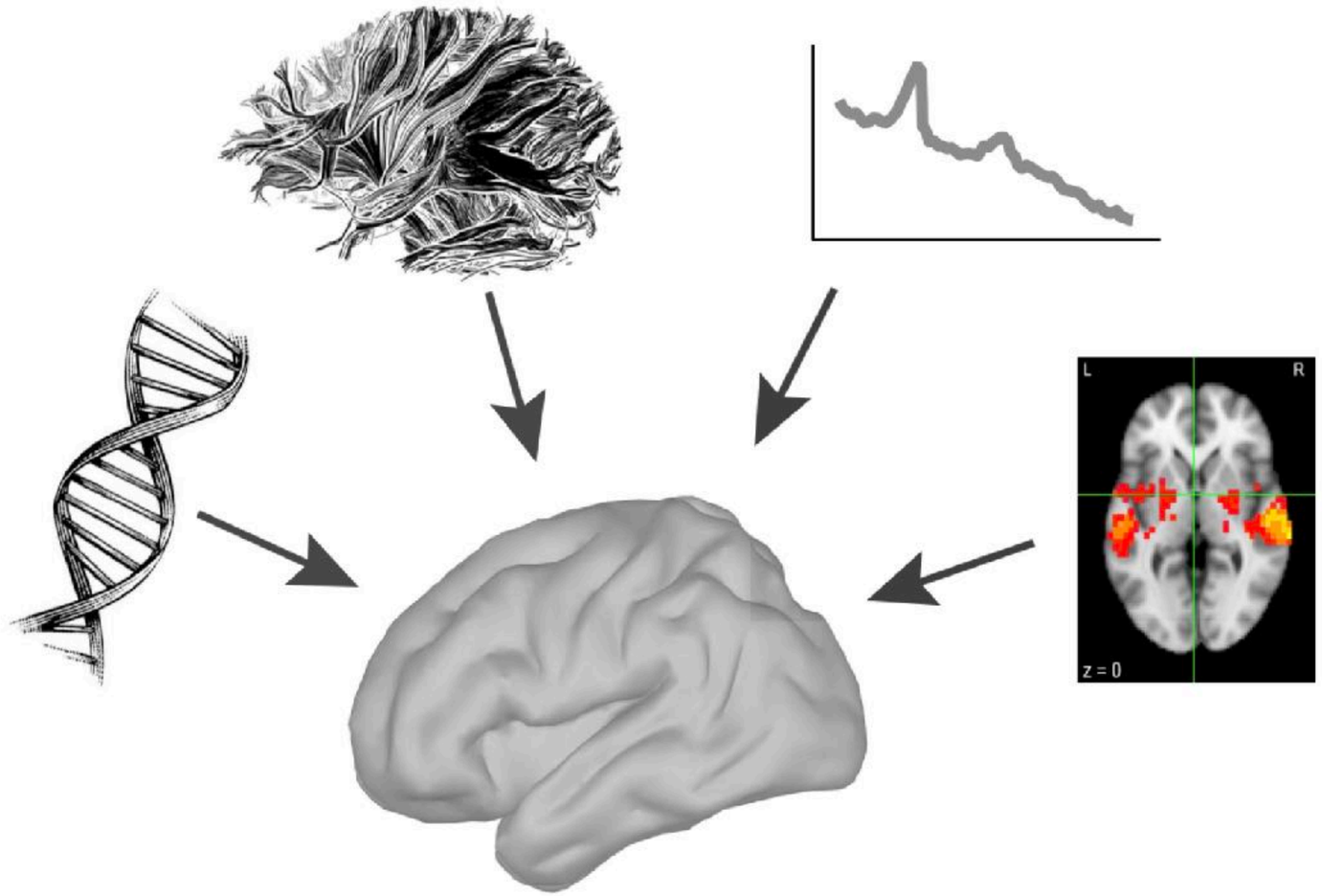
UC San Diego

NeuroSynth



Source: Yarkoni *et al.*, Nature Methods 2011

Multimodal comparisons



Source: Donoghue *et al.* (in preparation)

COGS 108
Data Science in Practice

Why and how do we analyze data?

UC San Diego

Proposed course order

1. Introduction: Why data analysis? (prediction and classification)
2. Python!
3. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
4. Data gathering, wrangling, and cleaning (How do you find and clean data? (JSON, CSV, XML, SQL, APIs)
5. Data privacy, ethics, and HIPAA (anonymization)
6. **Jan 25** Guest lecture: Kevin Novak: Chief Data Officer, Tala (Formerly: Head of Data & Engineering, *Uber*)
7. Basic data visualization
8. Data intuition and the "sniff test" (Fermi estimation; distributions and outliers: histograms, CDF, PDFs)
9. Non-parametric statistics
10. Linear modeling
11. **Feb 13** Ilkay Altintas, PhD: Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
12. NO CLASS!
13. OLS (optimization)
14. Multiple linear regression and collinearities
15. **Feb 27** Josh Wills: Director of Data Engineering, Slack (Formerly: Director of Data Science, Cloudera; Analytics, Google)
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Dimensionality reduction (PCA); clustering and classification (k-means, knn, SVM)
18. Feature selection
19. NLP and text-mining (bag of words, tf-idf, sentiment analysis)
20. Geospatial analysis

Proposed course order

1. Introduction: Why data analysis? (prediction and classification)
2. Python!
3. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
4. Data gathering, wrangling, and cleaning (How do you find and clean data? (JSON, CSV, XML, SQL, APIs)
5. Data privacy, ethics, and HIPAA (anonymization)
6. **Jan 25** Guest lecture: Kevin Novak: Chief Data Officer, Tala (Formerly: Head of Data & Engineering, *Uber*)
7. Basic data visualization
8. Data intuition and the "shift test" (Form estimation; distribution and outliers; histograms, CDF, PDFs)
9. Non-parametric statistics
10. Linear modeling
11. **Feb 13** Ilkay Altintas, PhD: Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
12. NO CLASS!
13. OLS (optimization)
14. Multiple linear regression and collinearities
15. **Feb 27** Josh Wills: Director of Data Engineering, Slack (Formerly: Director of Data Science, Cloudera; Analytics, Google)
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Dimensionality reduction (PCA); clustering and classification (k-means, knn, SVM)
18. Feature selection
19. NLP and text-mining (bag of words, tf-idf, sentiment analysis)
20. Geospatial analysis

YOU CAN LITERALLY TAKE AN ENTIRE CLASS ON EACH OF THESE

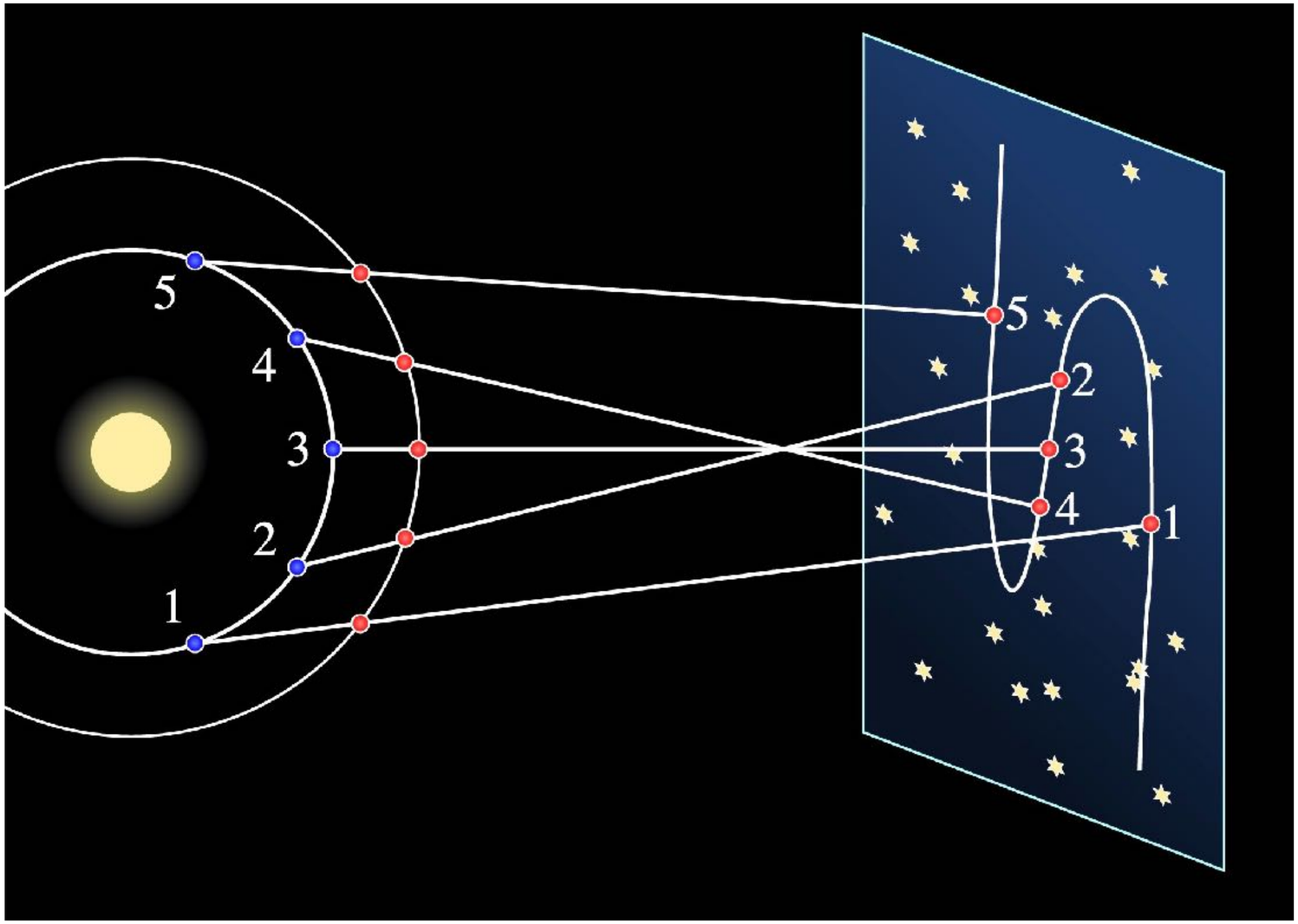
What is the point of data analysis?

- Prediction
- Classification
- Knowledge discovery?

What is the point of data analysis?

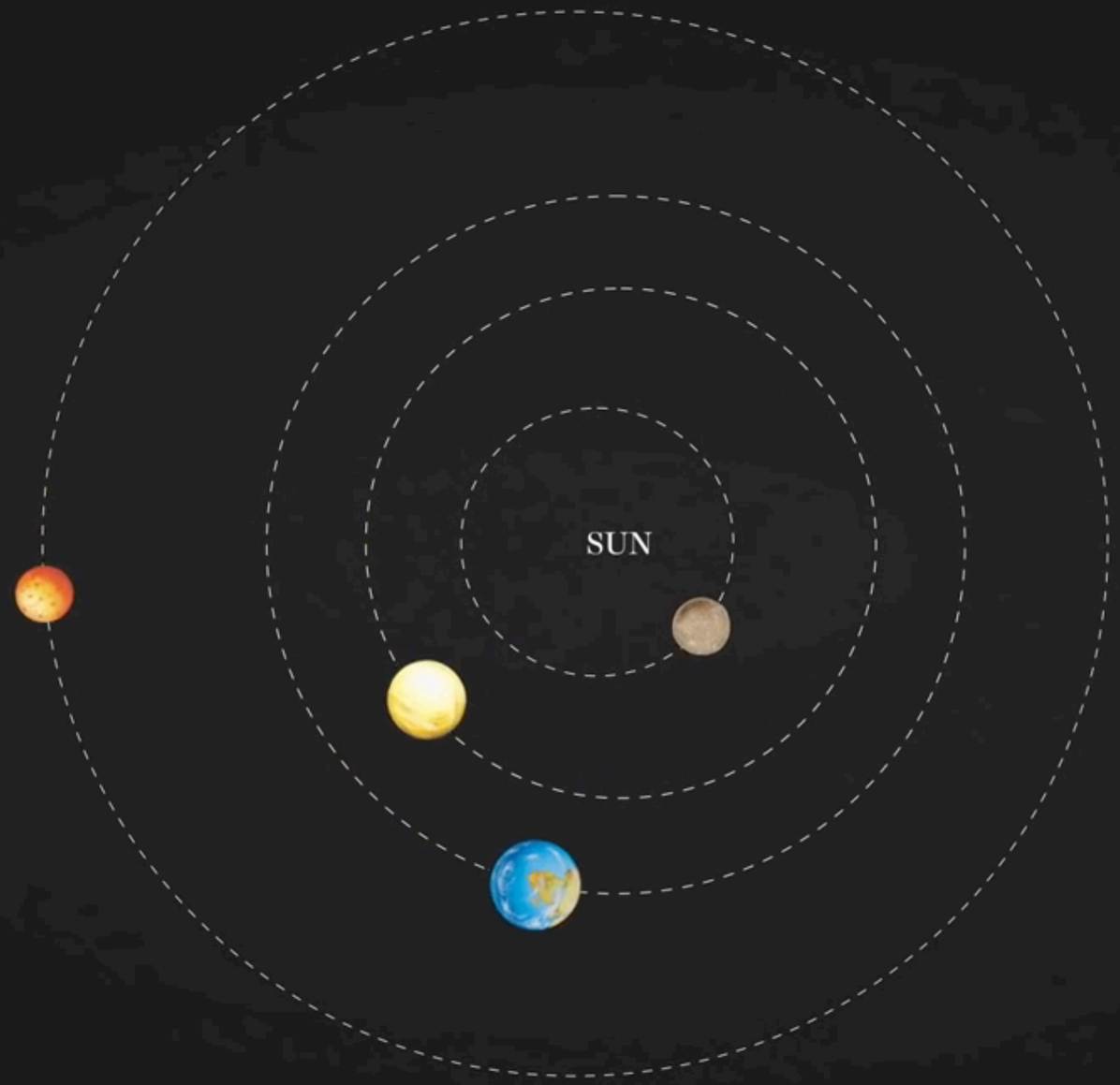
- Prediction
- Classification
- Knowledge discovery?
- **DOING USEFUL SHIT**

Prediction



Source: Wikipedia

INNER SOLAR SYSTEM



ptolemys model of the universe

www.principiauniversi.com

Models

2.3 Parsimony

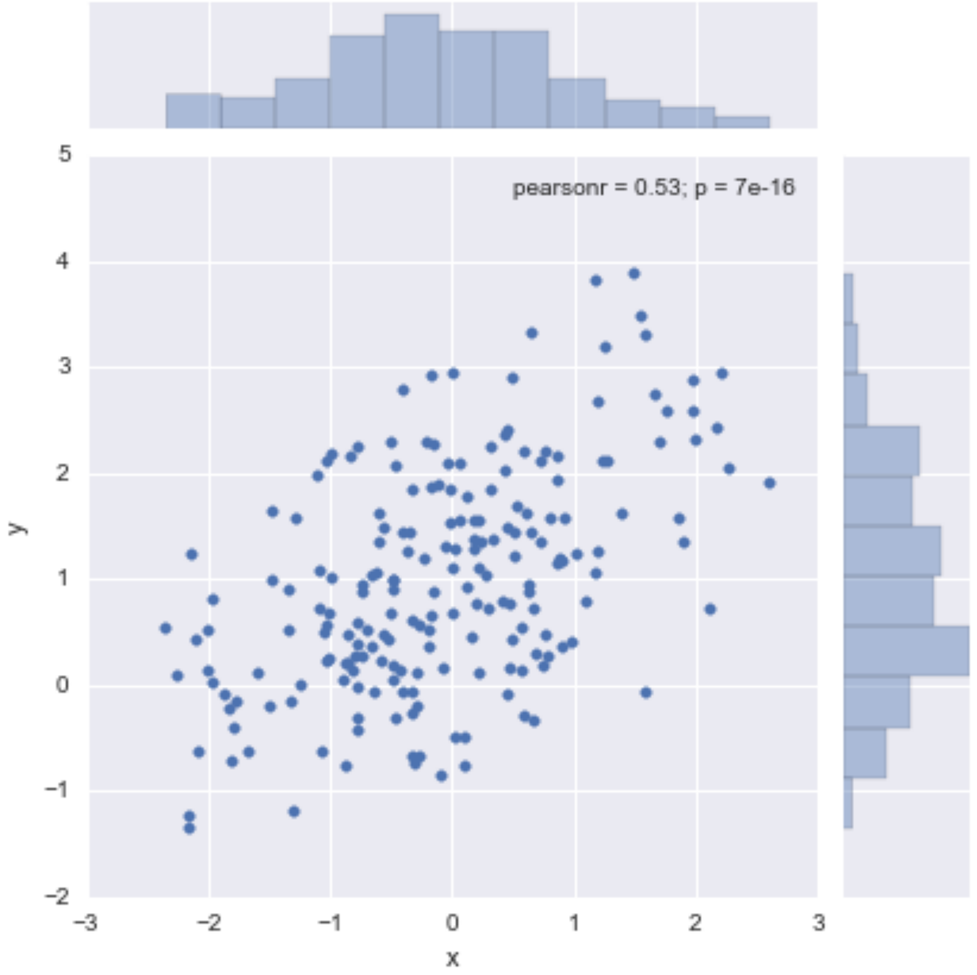
Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

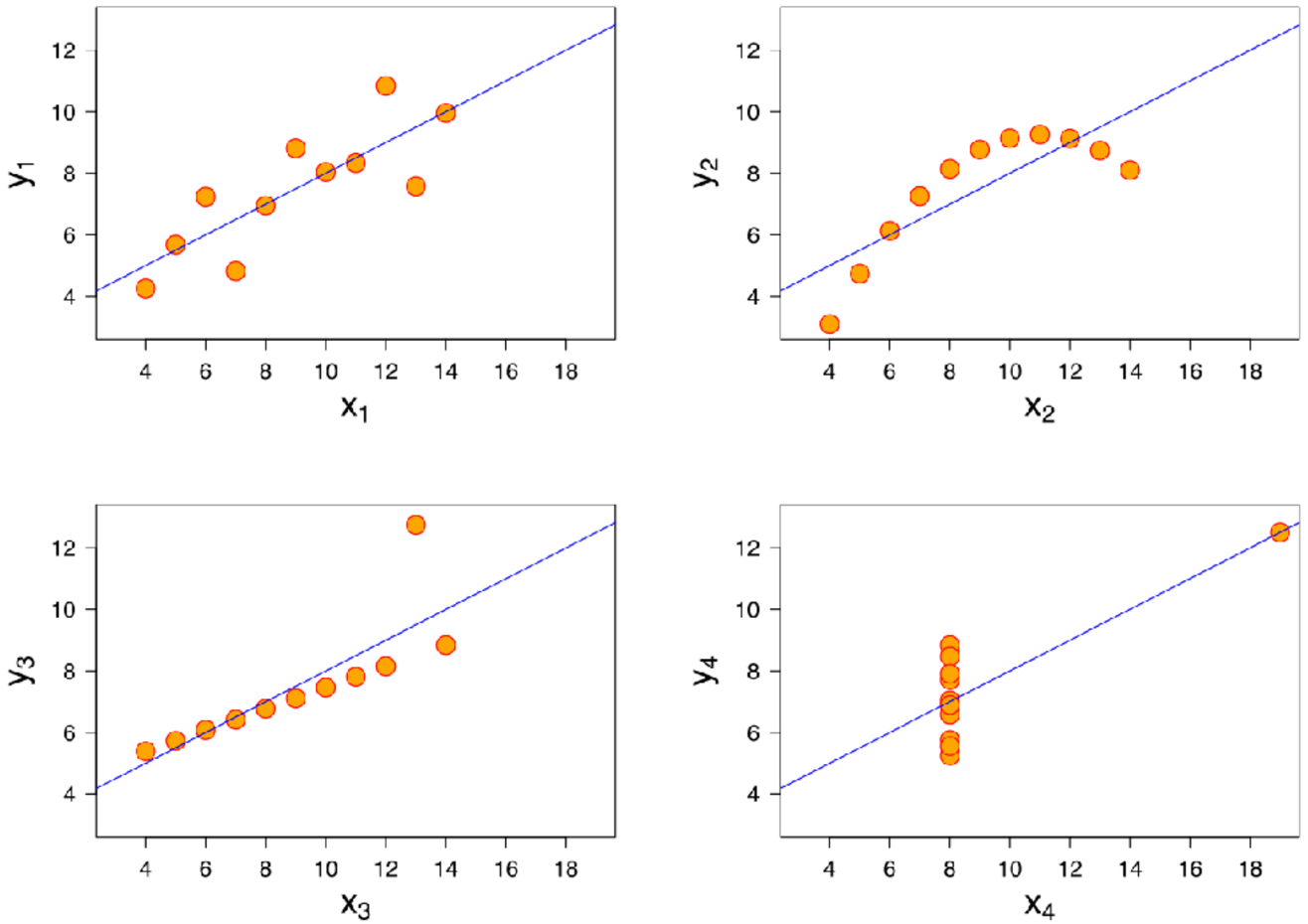
Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

Source: Box GEP, *J Am Stat Assoc*, 1976

Prediction

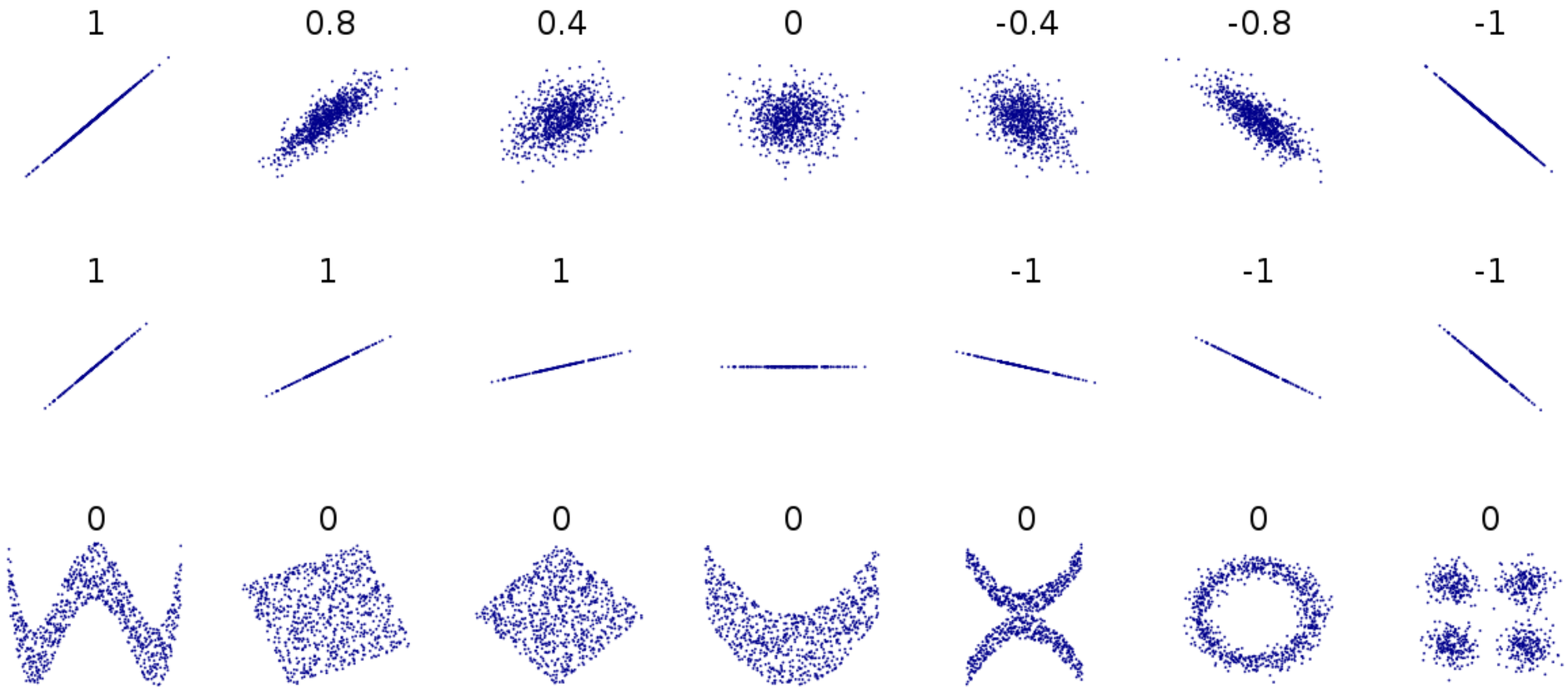


Prediction

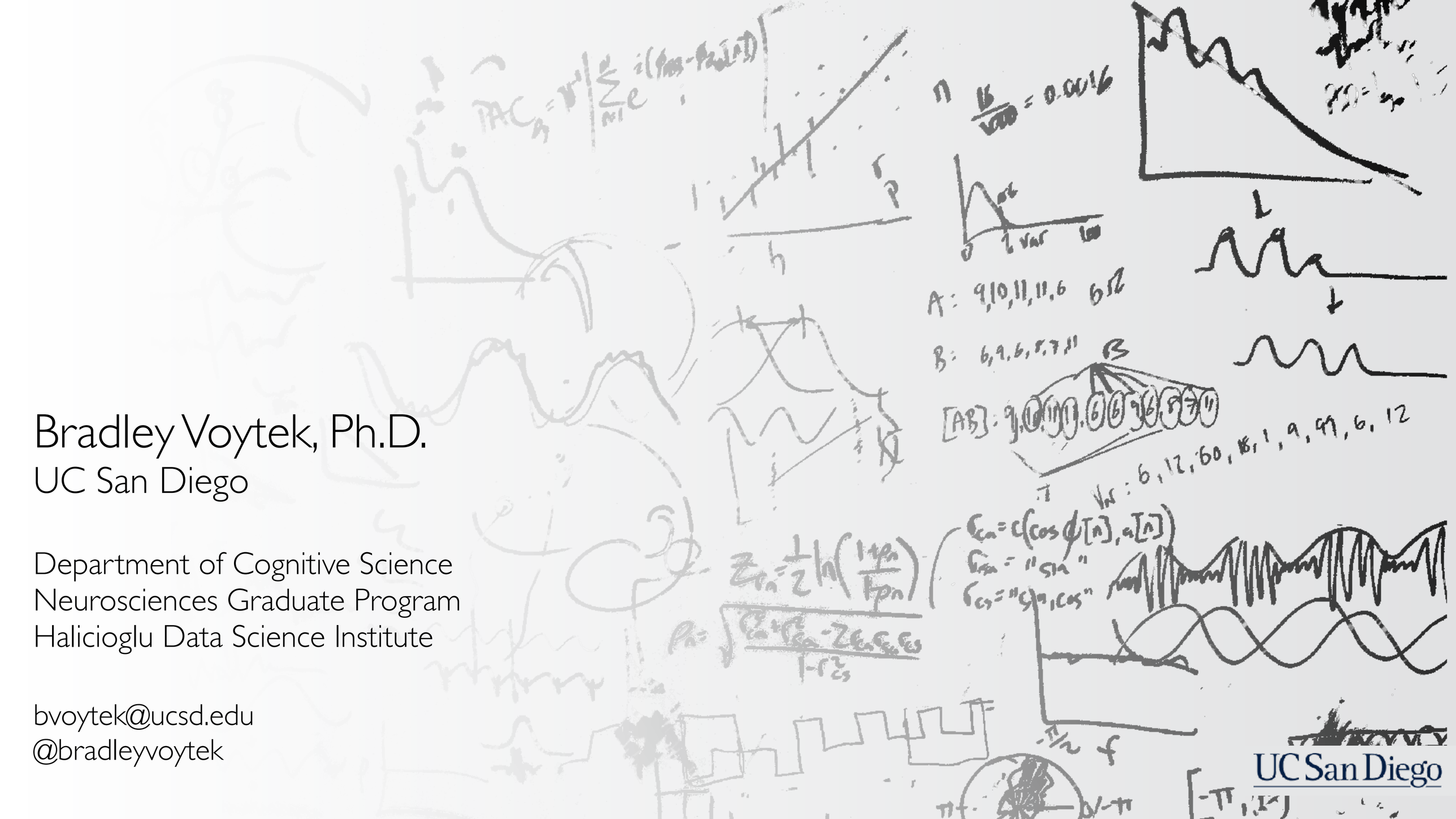


Source: Wikipedia

Prediction



Source: Wikipedia



Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego