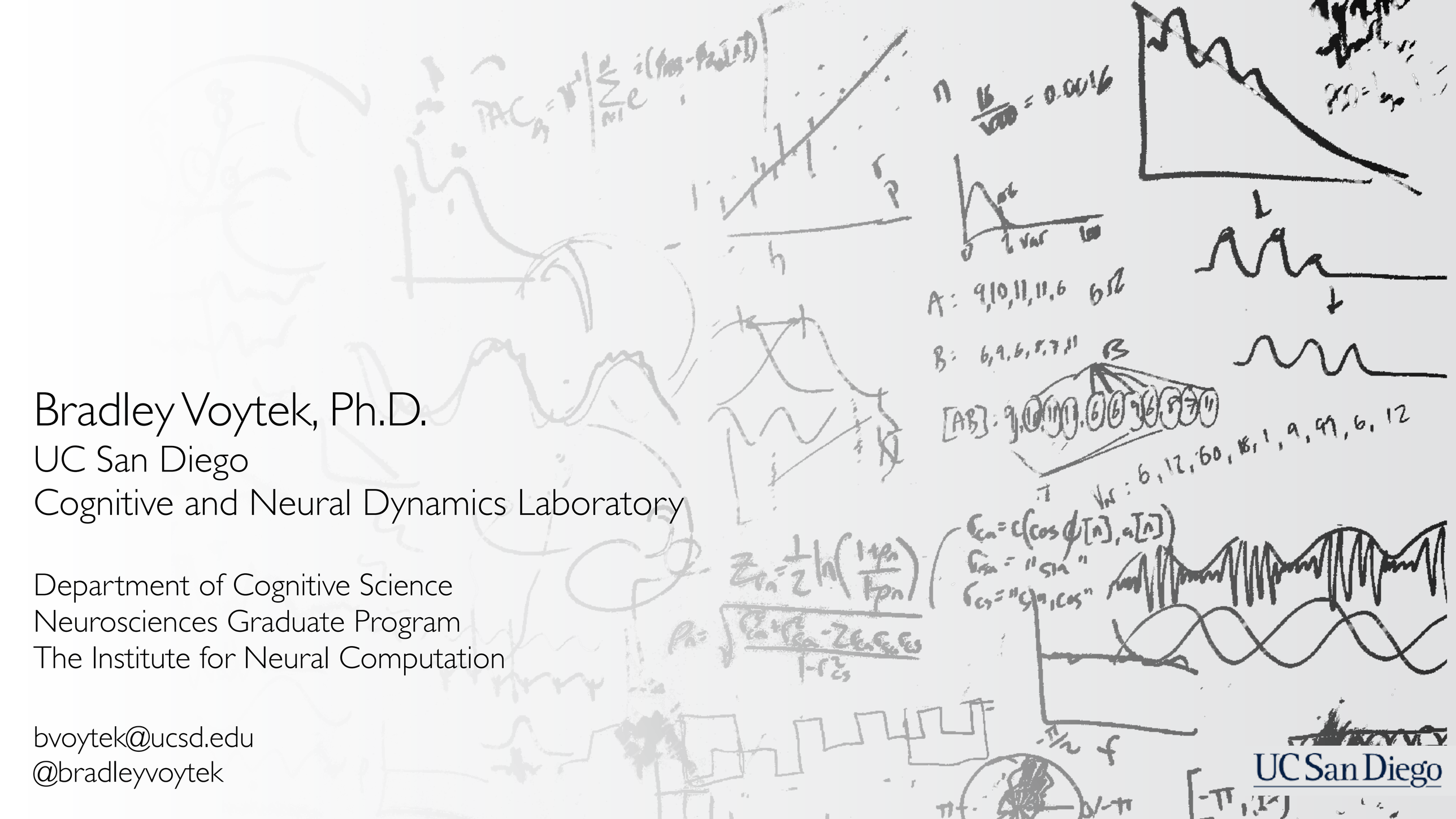


Bradley Voytek, Ph.D.

UC San Diego
Cognitive and Neural Dynamics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
The Institute for Neural Computation

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

COGS 108
Data Science in Practice

Why do we analyze data?

WAITLIST

- Only next 10 or so will get in for sure…
- …but there's an alternative in the works so please hold tight.

OTHER QUESTIONS

- Final Project and date
- Are the lab sections mandatory?
- Can I attend a different section other than the one I'm enrolled in?
- Piazza: Not listed there?
- Data Science major?

DATA SCIENCE MAJOR

- bit.ly/UCSD_DSlist

DS3

Data Science Student Society

Proposed course order

2. Why data analysis? (prediction and classification)
3. Python!
4. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
5. Data gathering (How do you find and clean data?)
6. Data wrangling (JSON, CSV, XML, SQL, APIs)
7. Data cleaning
8. Data privacy and HIPAA (anonymization)
9. Basic data visualization
10. Data intuition and the "sniff test" (Fermi estimation)
11. Linear modeling
12. OLS (optimization)
13. Distributions and outliers
14. Distributions and outliers: CDF, PDFs
15. Multiple linear regression and collinearities
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Feature selection
18. Dimensionality reduction (PCA)
19. Clustering (knn and k-means)
20. Classification (SVM)
21. Interpretability (trees!)
22. Non-parametric statistics
23. NLP and text-mining (tf-idf, sentiment analysis)
24. Geospatial analysis
25. Unsupervised learning (dbscan)

Proposed course order

2. Why data analysis? (prediction and classification)
3. Python!
4. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
5. Data gathering (How do you find and clean data?)
6. Data wrangling (JSON, CSV, XML, SQL, APIs)
7. Data cleaning
8. Data privacy and HIPAA (anonymization)
9. Basic data visualization
10. Data intuition and the "sniff test" (Fermi estimation)
11. Linear models
12. OLS (optimization)
13. Distributions and outliers
14. Hypothesis testing and A/B testing: ANOVA
15. Multiple linear regression and collinearities
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Feature selection
18. Dimensionality reduction (PCA)
19. Clustering (knn and k-means)
20. Classification (SVM)
21. Interpretability (trees!)
22. Non-parametric statistics
23. NLP and text-mining (tf-idf, sentiment analysis)
24. Geospatial analysis
25. Unsupervised learning (dbscan)

YOU CAN LITERALLY TAKE AN ENTIRE CLASS ON EACH OF THESE

UC San Diego

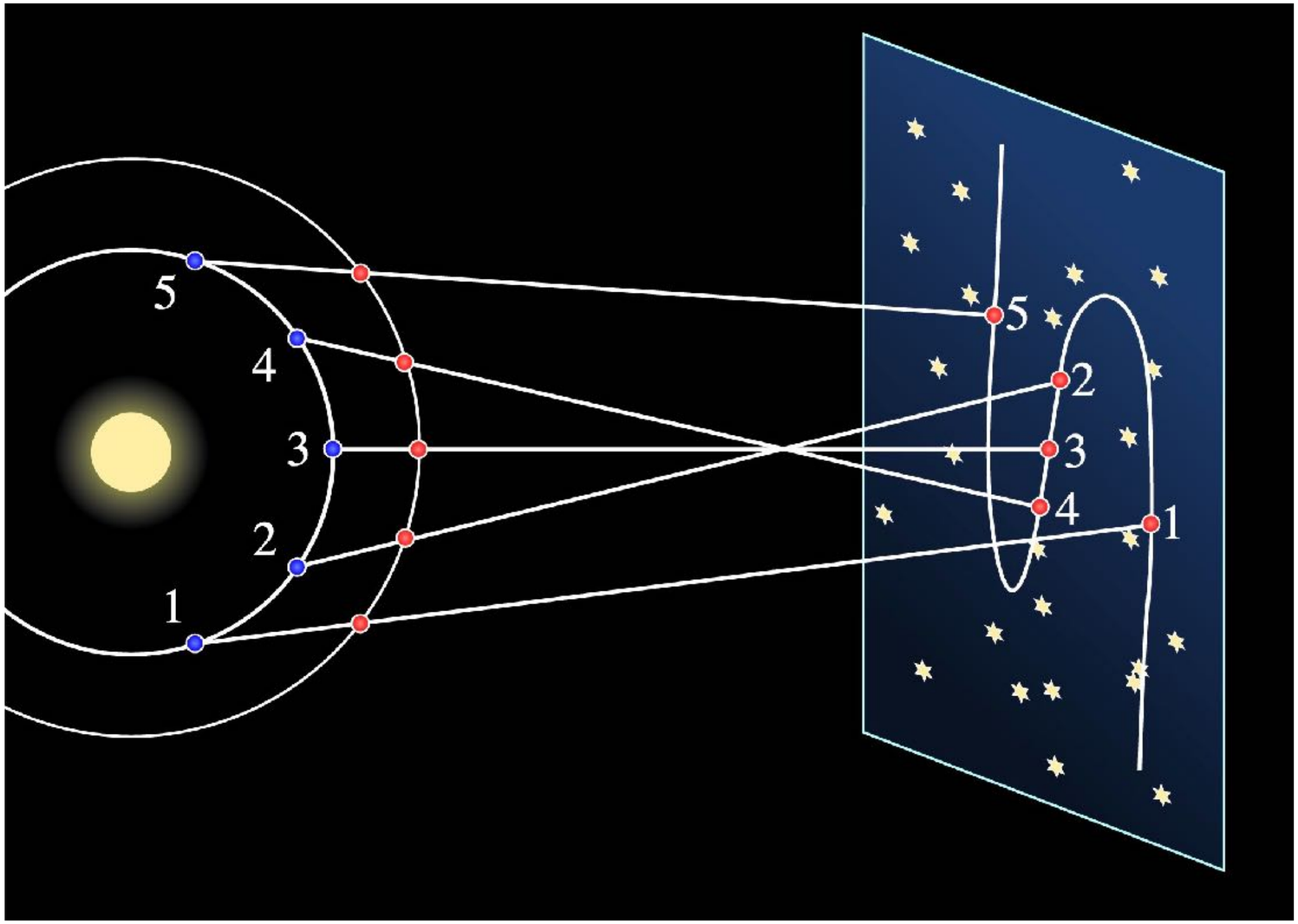
What is the point of data analysis?

- Prediction
- Classification
- Knowledge discovery?

What is the point of data analysis?

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

Prediction



Source: Wikipedia

ptolemys model of the universe

www.principiauniversi.com

Models

2.3 Parsimony

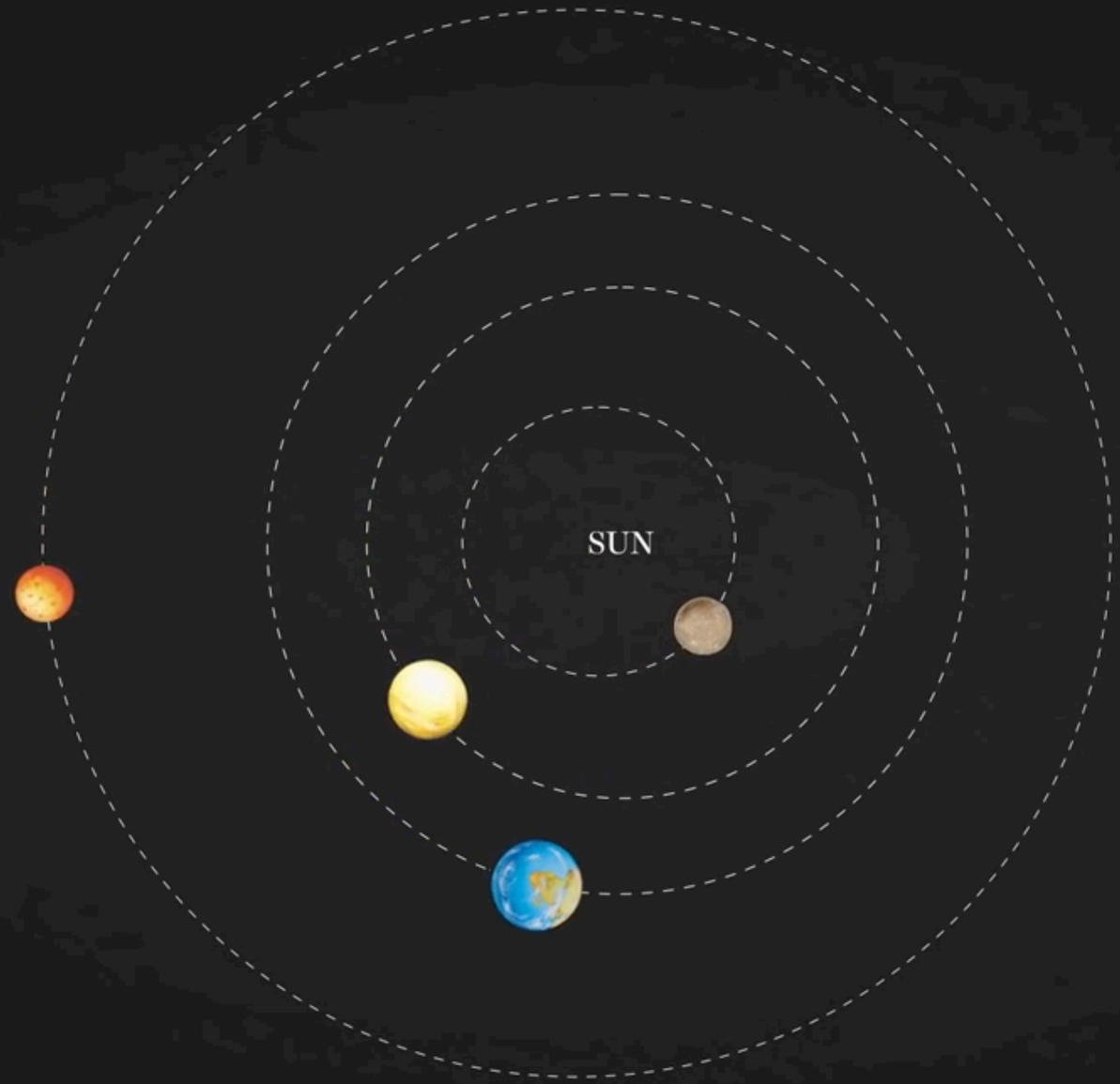
Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist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

2.4 Worrying Selectively

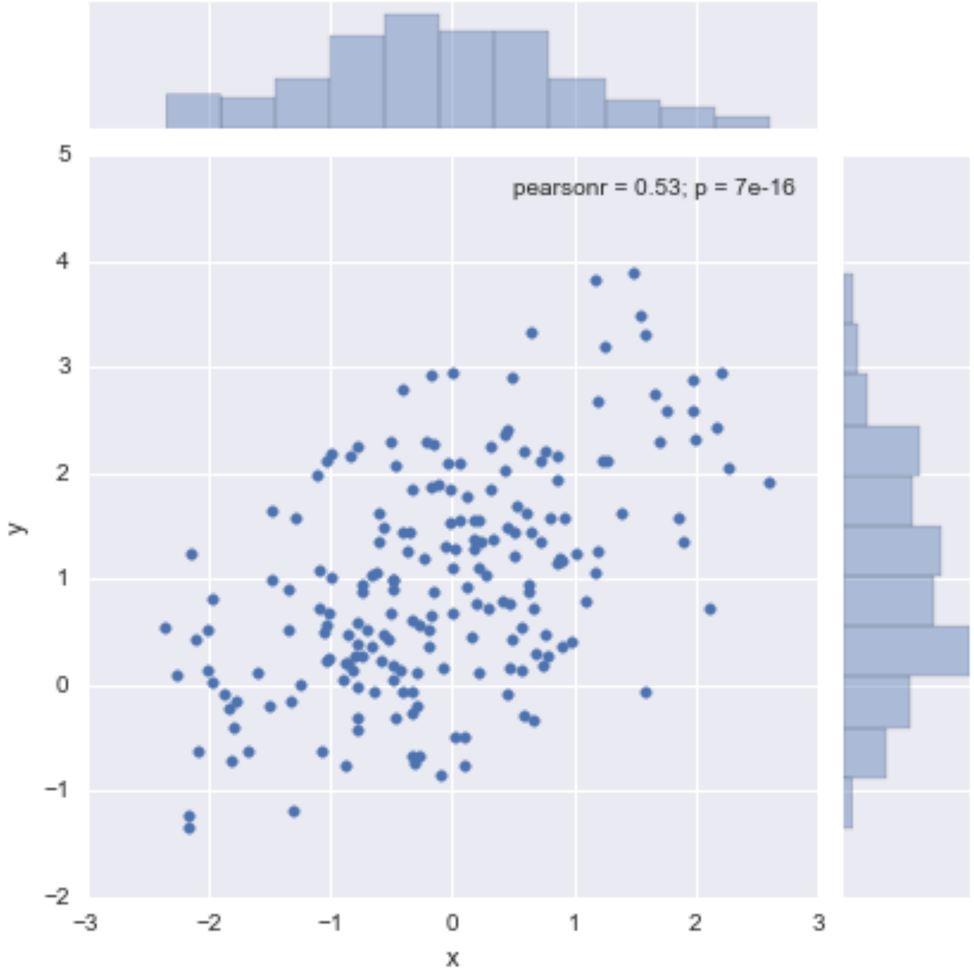
Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

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

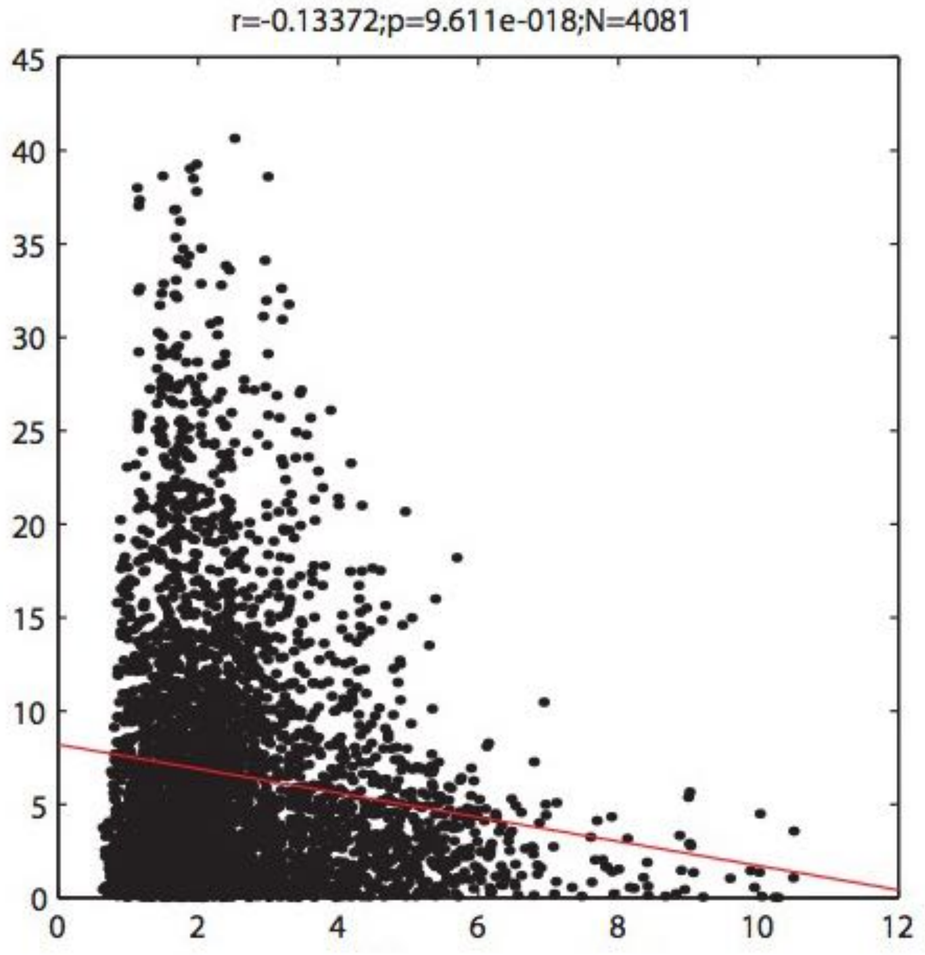
INNER SOLAR SYSTEM



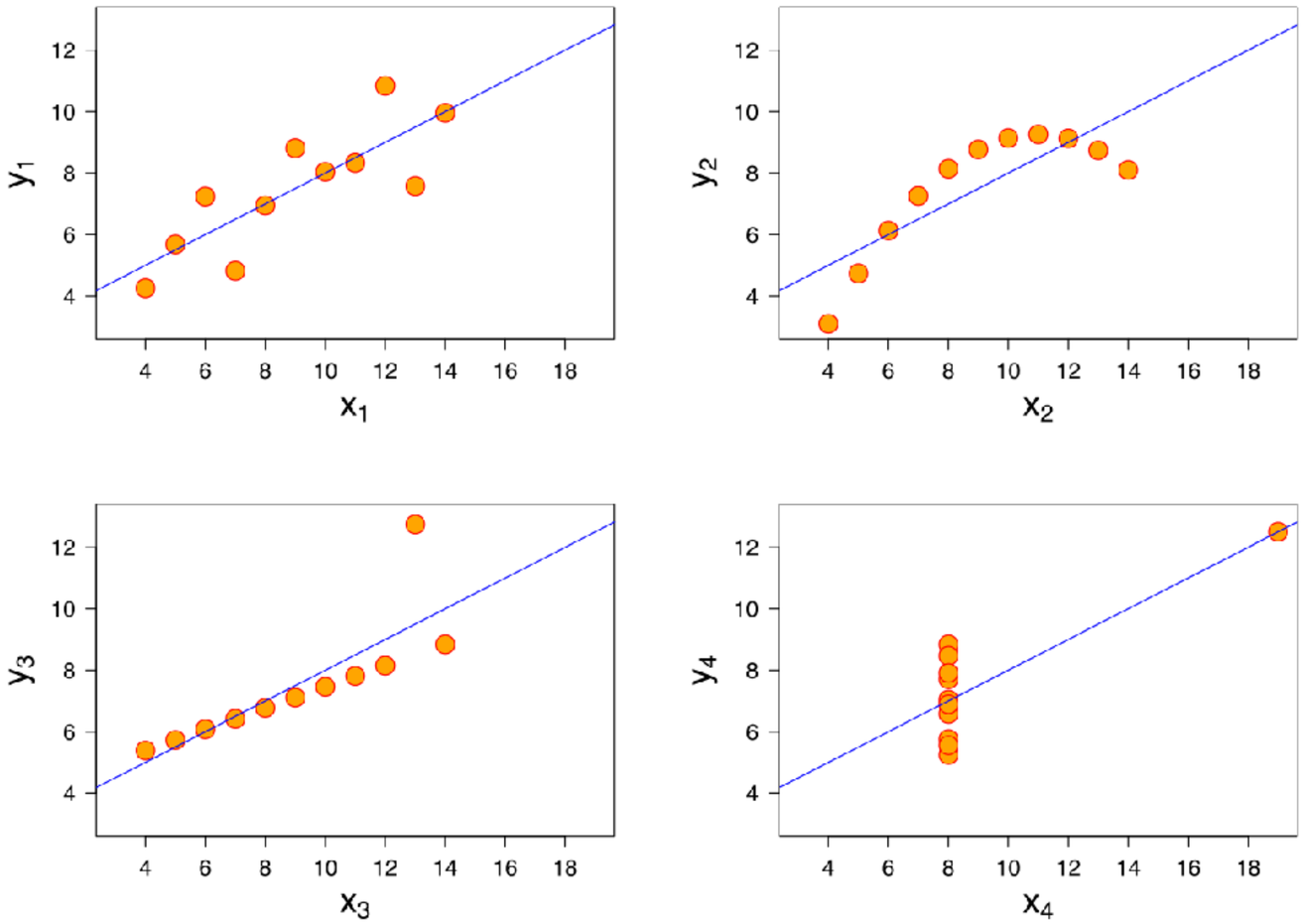
Prediction



Prediction

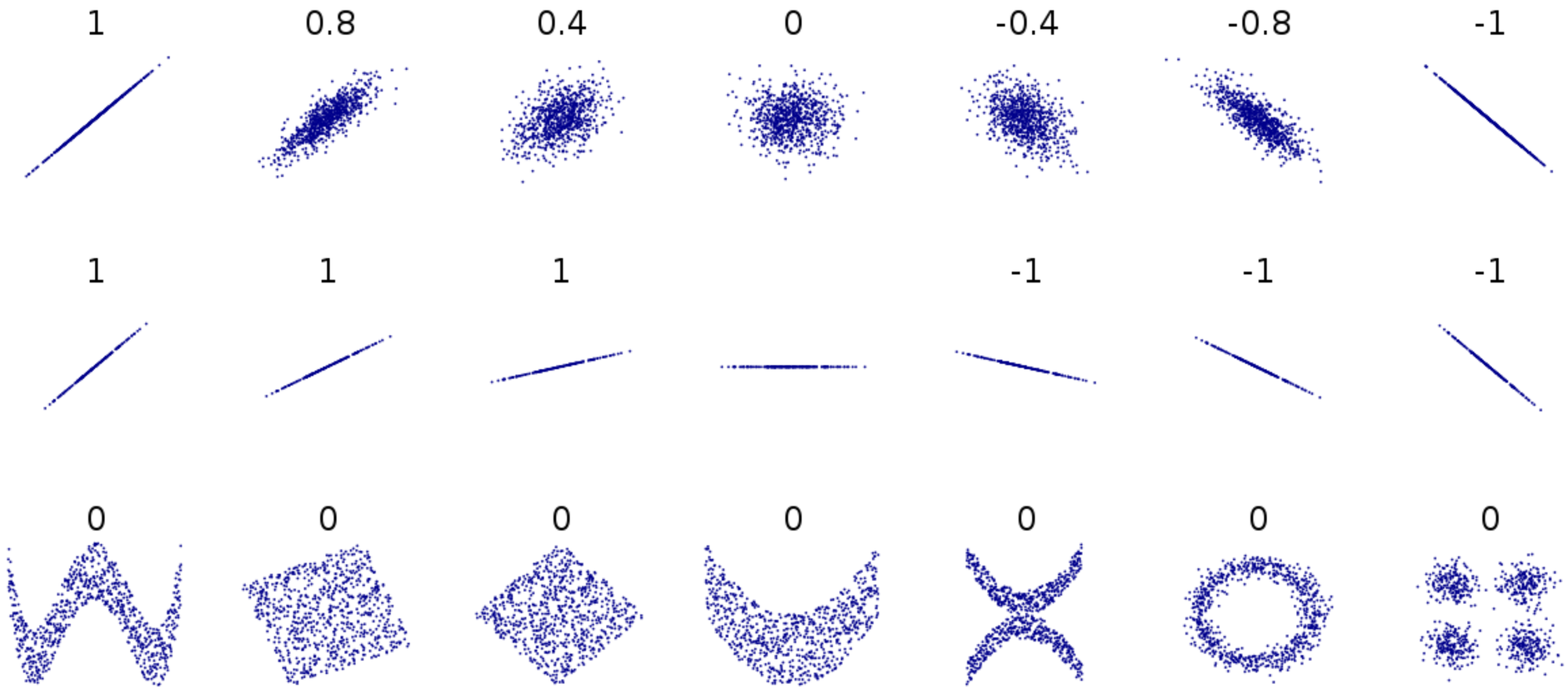


Prediction



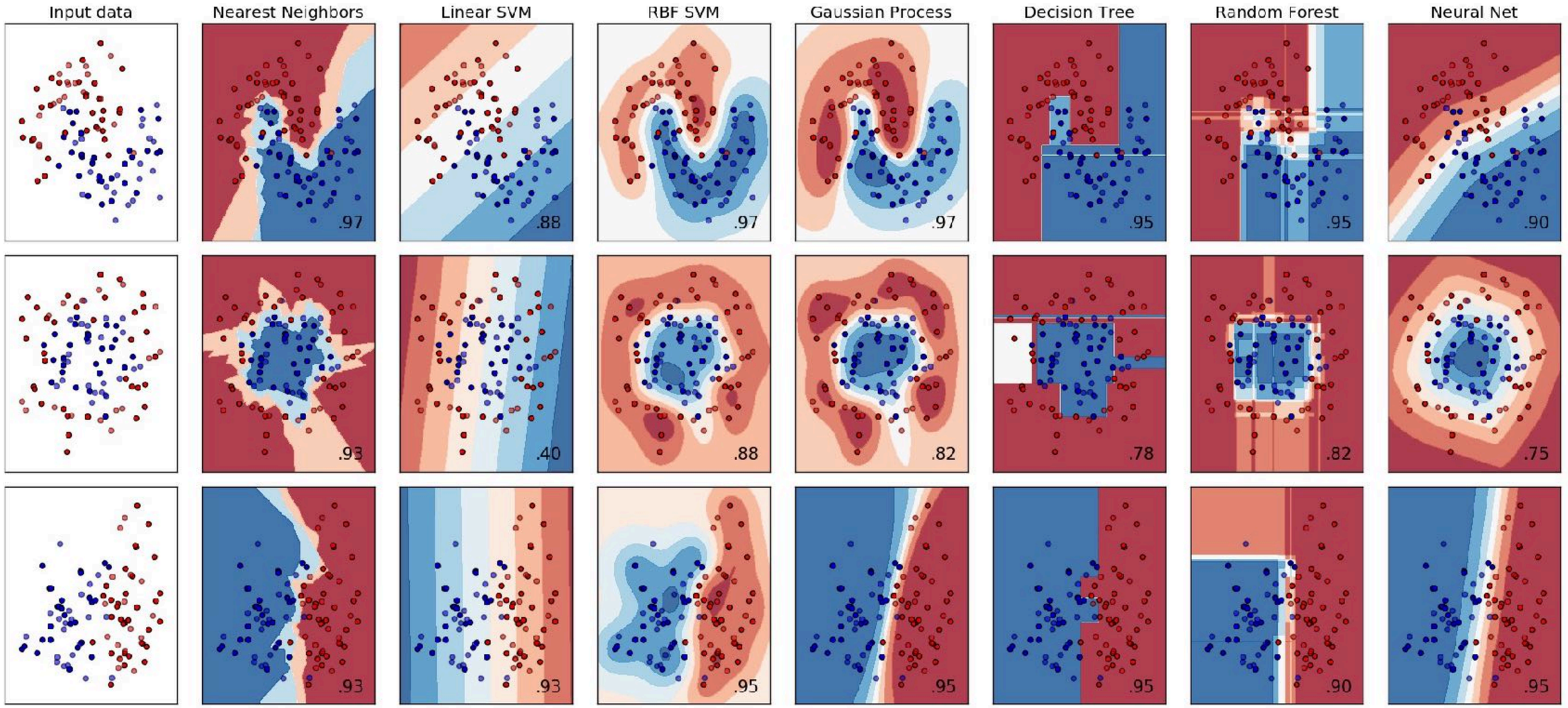
Source: Wikipedia

Prediction



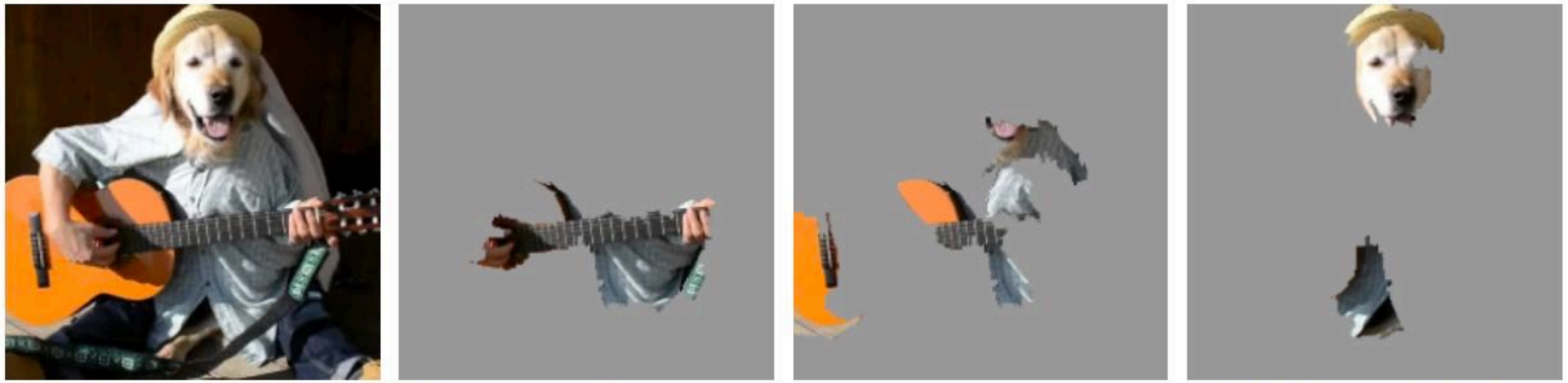
Source: Wikipedia

Classification



Source: scikit-learn

Classification

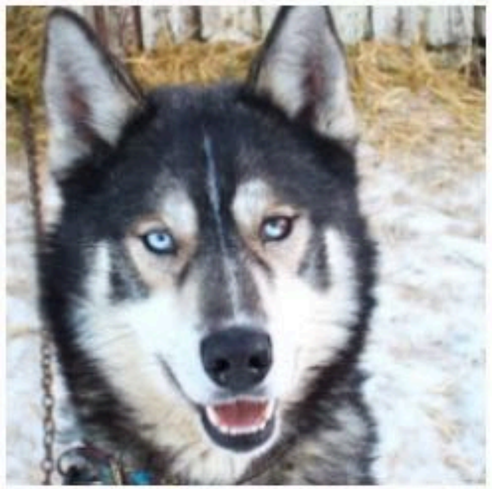


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Source: Ribeiro, Singh, Guestrin, *KDD* 2016

Classification



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Source: Ribeiro, Singh, Guestrin, *KDD* 2016

Classification

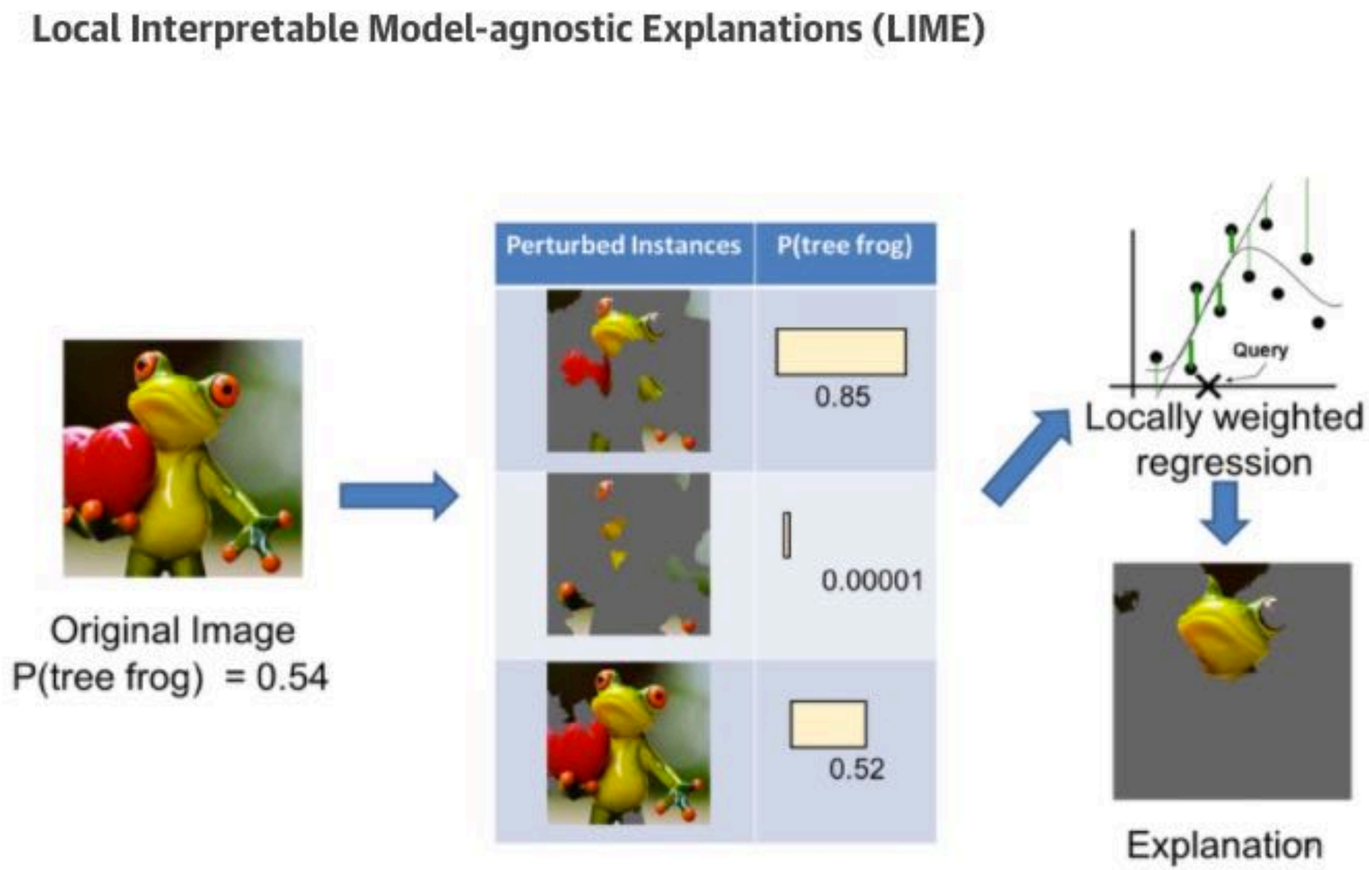
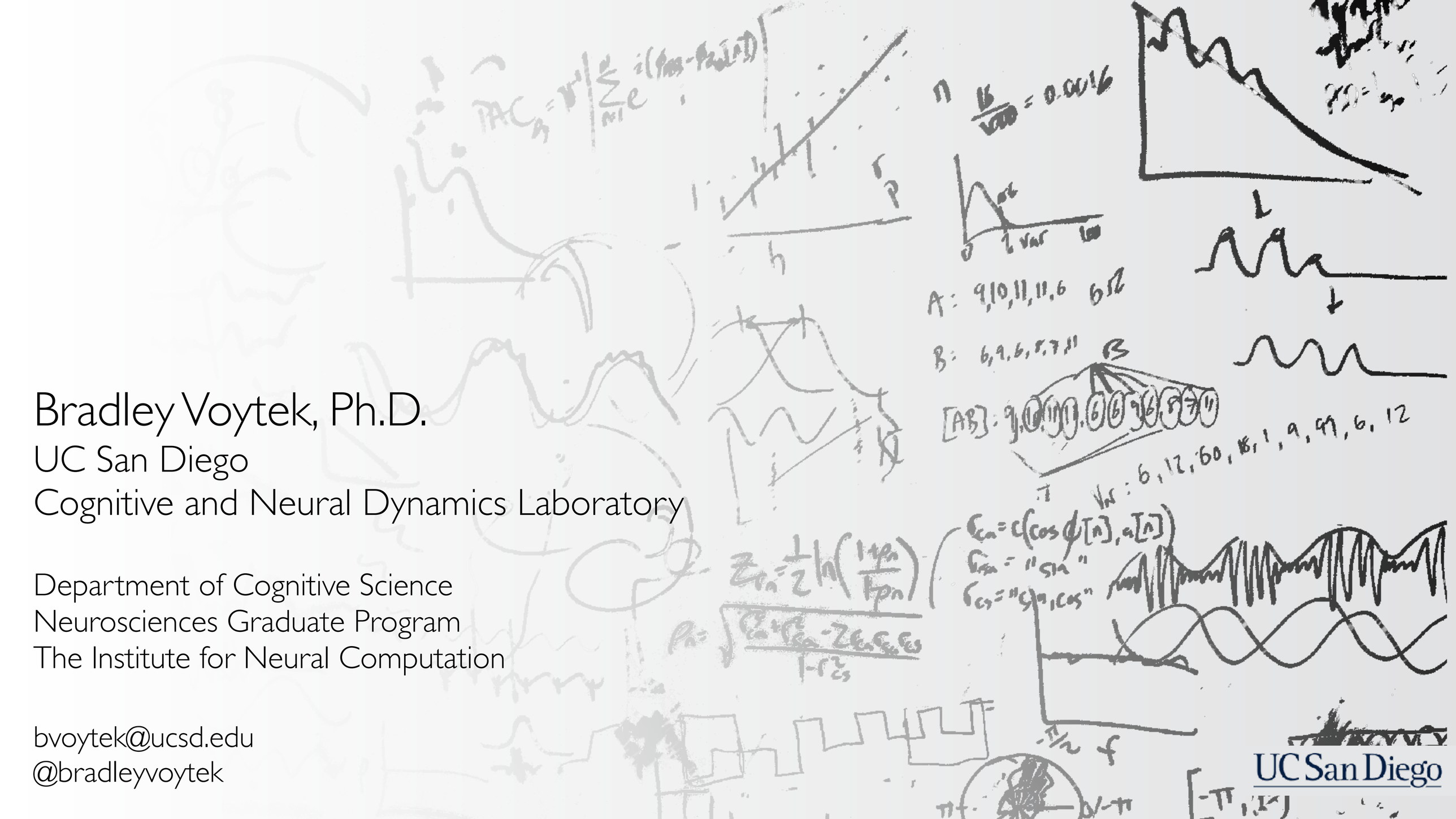


Figure 15. An illustration of the LIME process in which a weighted linear model is used to explain a single prediction from a complex neural network. Figure courtesy of Marco Tulio Ribeiro; image used with permission.

Source: Hall, Phan, Ambati, *O'Reilly Media* 2017



Bradley Voytek, Ph.D.

UC San Diego
Cognitive and Neural Dynamics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
The Institute for Neural Computation

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego