JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

# Introduction to Data Science

Tamás Budavári
Dept of Applied Mathematics & Statistics
Dept of Computer Science
Dept of Physics & Astronomy

# About you

- PhD / Masters / Undergraduate?
- What major?

# About me

- Background in Physics: Stat/Bio/Astro
  - Astronomy surveys $\rightarrow$ Big Data
- Research interest

  Computational Statistics; Bayesian Inference;

  Statistical Learning; Scientific Databases;
- Office: Whitehead 212C

# About the course

- Introduction to data science
- Basic methods – used all the time
- Presentations + Codes
- Syllabus posted soon

# Grades

- 30% Homework 1 & 2
- 50% Midterm 1 & 2
- 20% Project

# Pre-requisites

- Linear algebra
- Intro to prob/stat?

# Sections

- Use your laptop
  - Also in class…

# Plan for the Timeline

- Homework 1 – graded in time for dropping
- Midterm 1
- Homework 2
- Midterm 2 – few weeks before end of semester
- Project – presentations

# Format of Lectures

- Alternating between
  - Presentations
  - Coding
- Everything is going to Blackboard

# Homework

- Data Science problems
- Much like the examples

# Unhomework

- Same but not graded

# Exams

- In class
- Coding

# What's coming?

# Statistical Learning

| Supervised | Unsupervised |
|---|---|

# Statistical Learning

| | Supervised | Unsupervised |
|---|---|---|
| Discrete | | |
| Continuous | | |

# Statistical Learning

| | Supervised | Unsupervised |
|---|---|---|
| **Discrete** | Classification | |
| **Continuous** | | |

# Statistical Learning

| | Supervised | Unsupervised |
|---|---|---|
| **Discrete** | Classification | |
| **Continuous** | Regression | |

# Statistical Learning

| | Supervised | Unsupervised |
|---|---|---|
| **Discrete** | Classification | Clustering |
| **Continuous** | Regression | |

# Statistical Learning

| | Supervised | Unsupervised |
|---|---|---|
| Discrete | Classification | Clustering |
| Continuous | Regression | Dimensionality Reduc’n |

# Topics

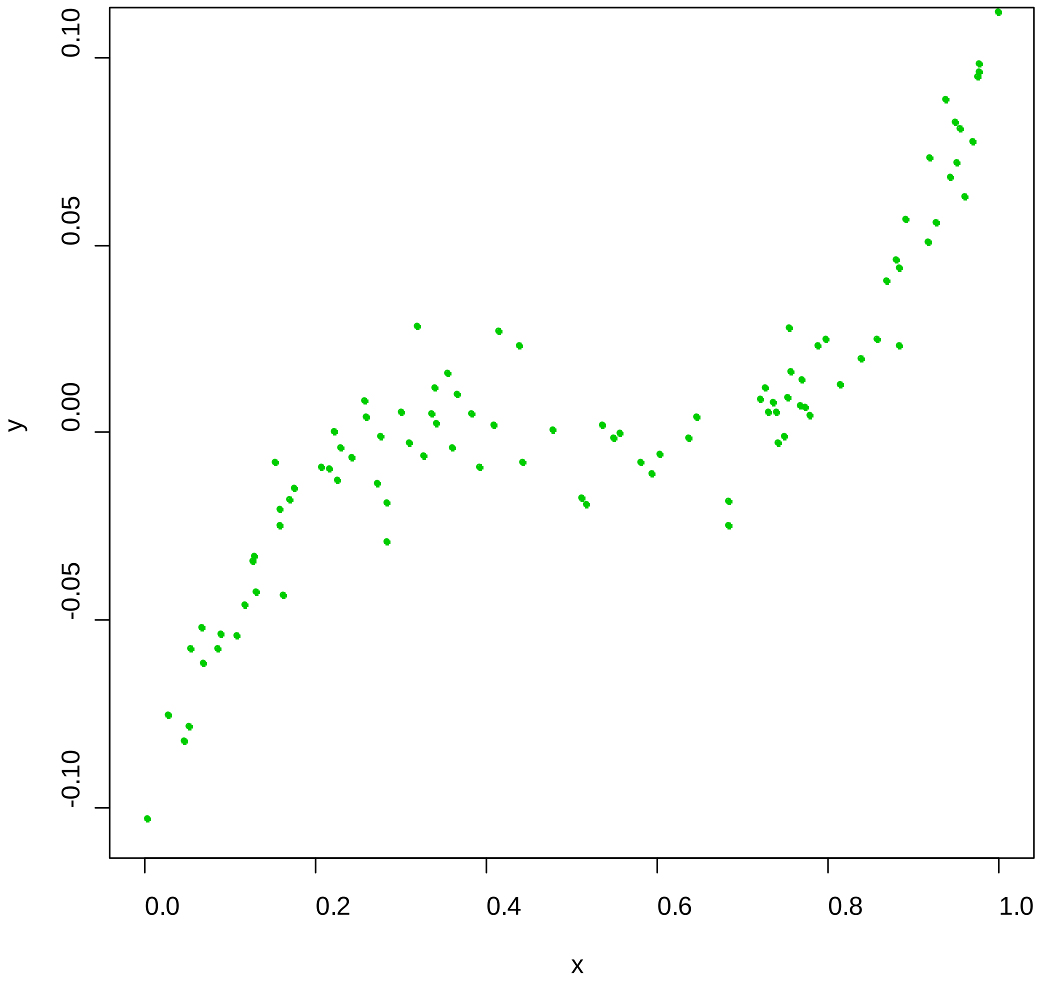
descriptive statistics – probabilistic density functions – normal distributions – regression – classification – nearest neighbors – bias-variance – Bayesian inference – robustness – regularization – support vector machines – decisions trees – clustering – principal component analysis – expectation maximization – neural networks – spectral embedding – …

# Supervised Learning

# Learning

- Model
  - Unknown function
  - Random noise

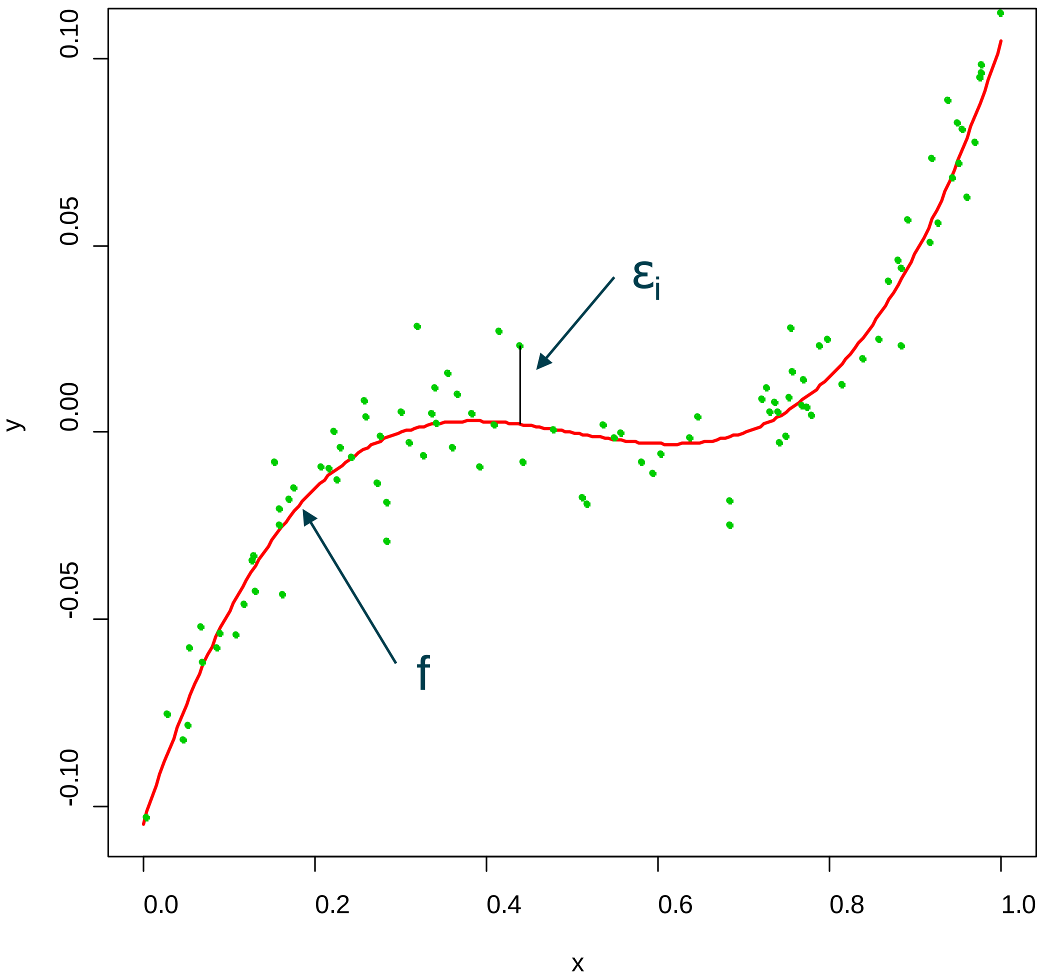
$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$



# Learning

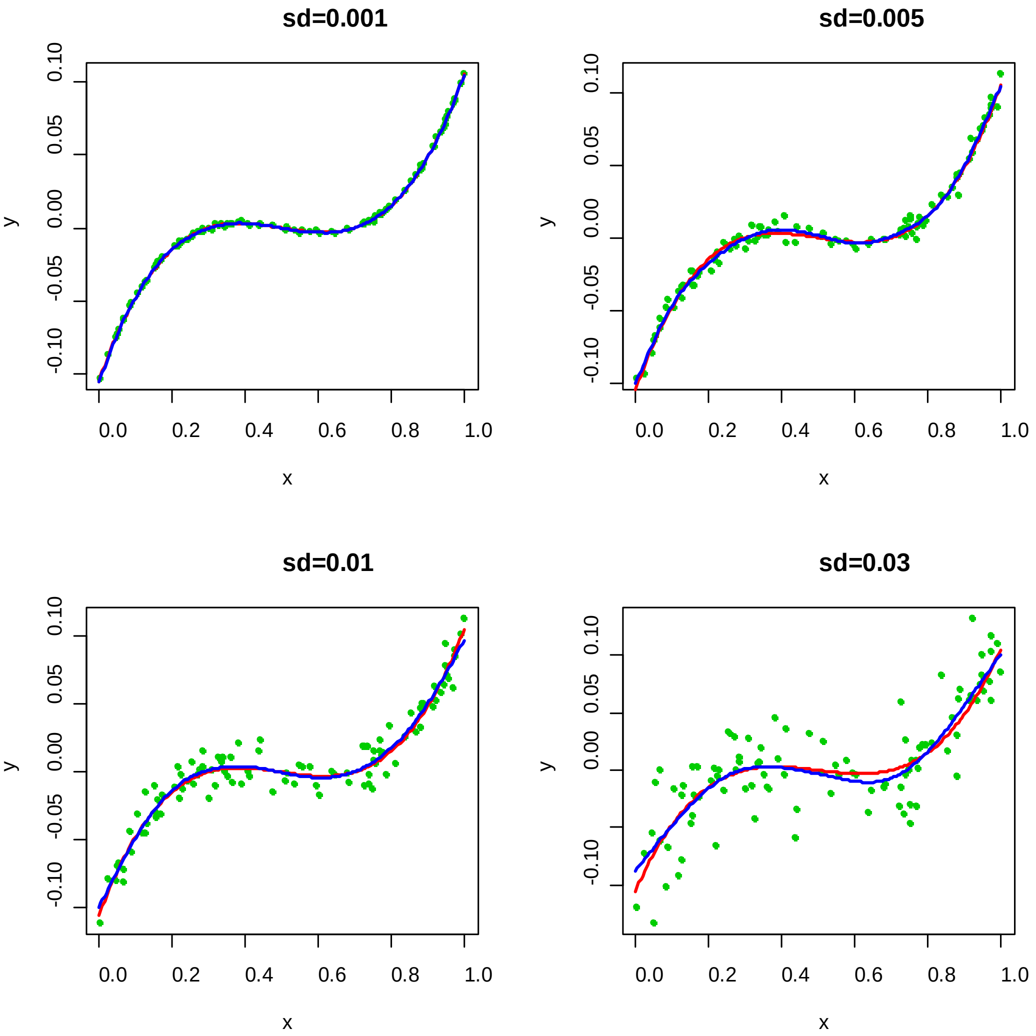
- Model
  - Unknown function
  - Random noise

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$



# Noise!

- Different scatter
- Different solutions



# Why learn $f(x)$ ?

- Inference
  - Relation of variables to target
- Prediction
  - Estimate $y$ for a new $x$

# How to estimate *f*(*x*) ?

- Using a training set with both
  - Input
  - Output

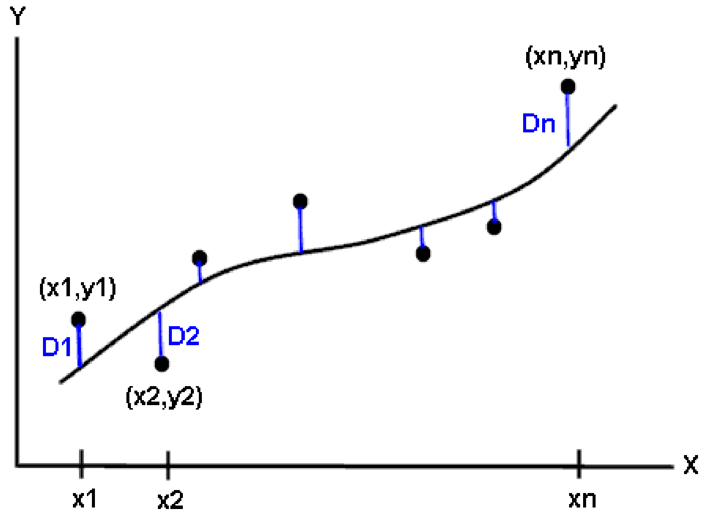
$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$

- For example, assuming a linear model

$$f(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$
$$f(\boldsymbol{x}_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d}$$

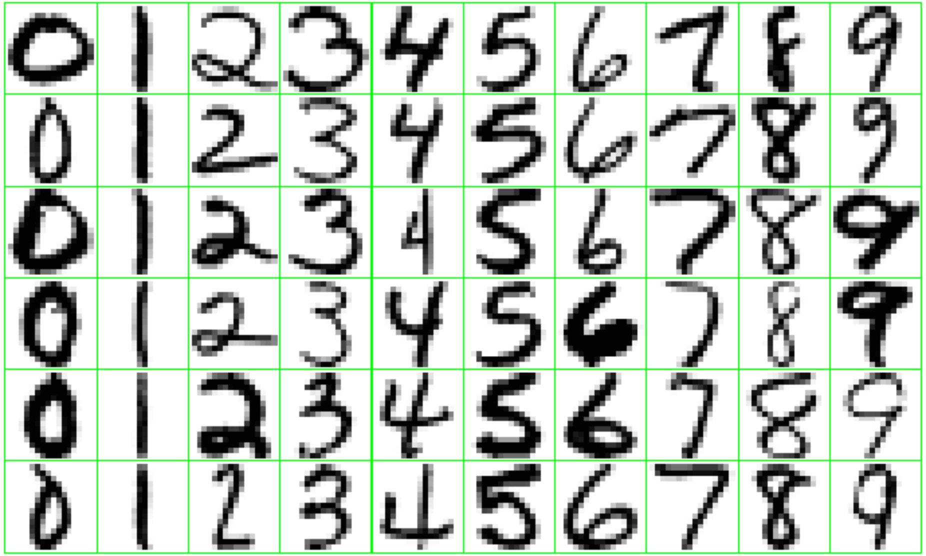
# How to estimate $f(x)$ ?

- One way is the method of least squares
  - Form differences of $Y_i$ and $f(\boldsymbol{X}_i)$
  - Minimize the sum of squares



# Handwritten Numbers?

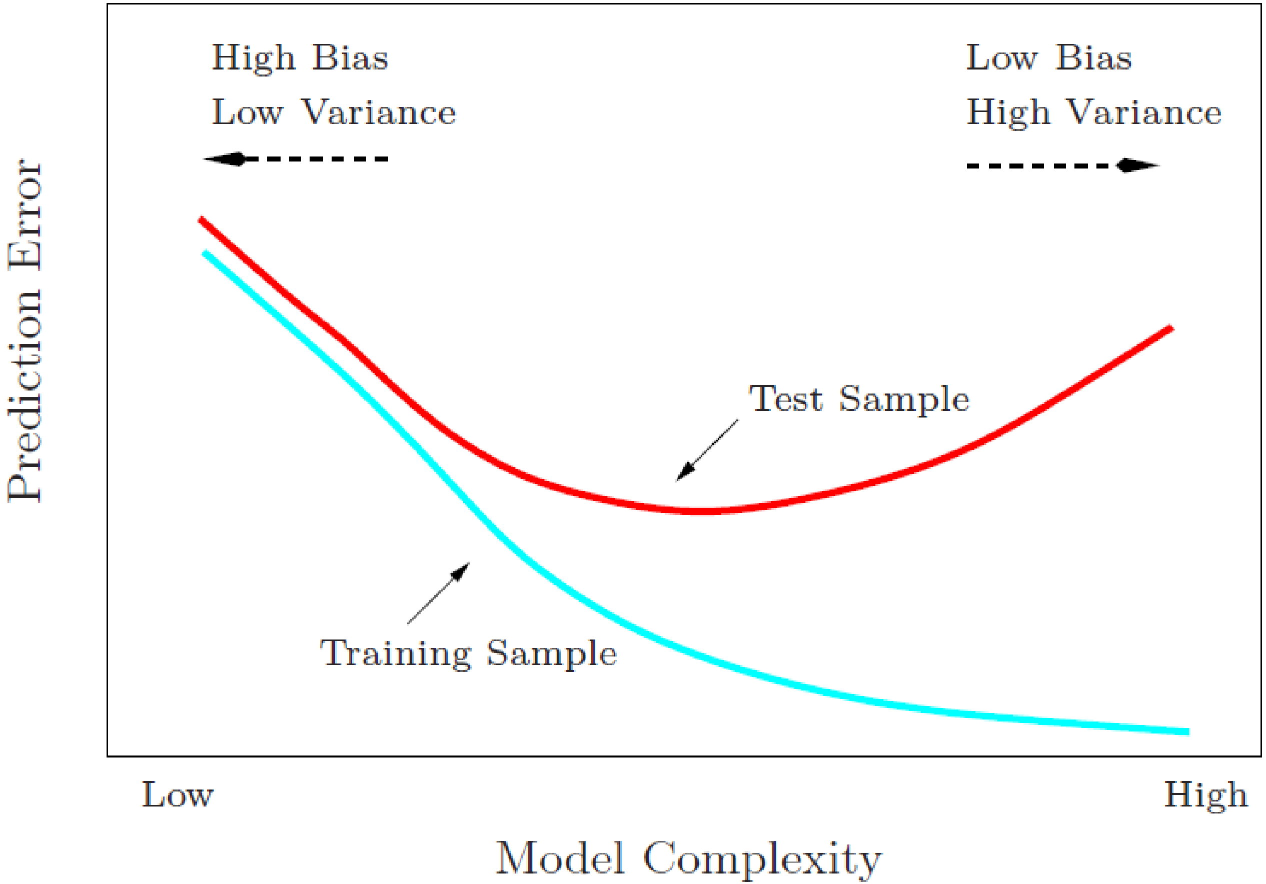
- Which digit?
- Classification!
  - Training set



# Complexity

Complicated models can better fit the data but harder to interpret and understand

- Too simple: underfitting
  - Bad fit on training & test sets
- Too complex: overfitting
  - Better on training but worse on test set


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

# Interpretation

*There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.*

–Andreas Buja

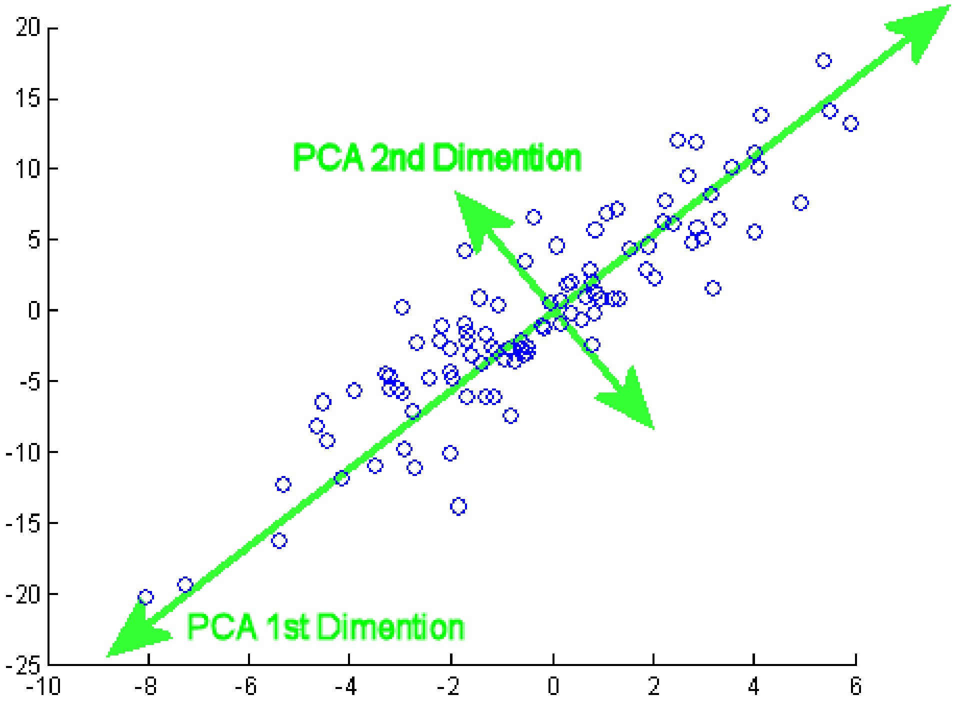
# Unsupervised Learning

# Clustering

- If no labels are provided
- We learn the clusters

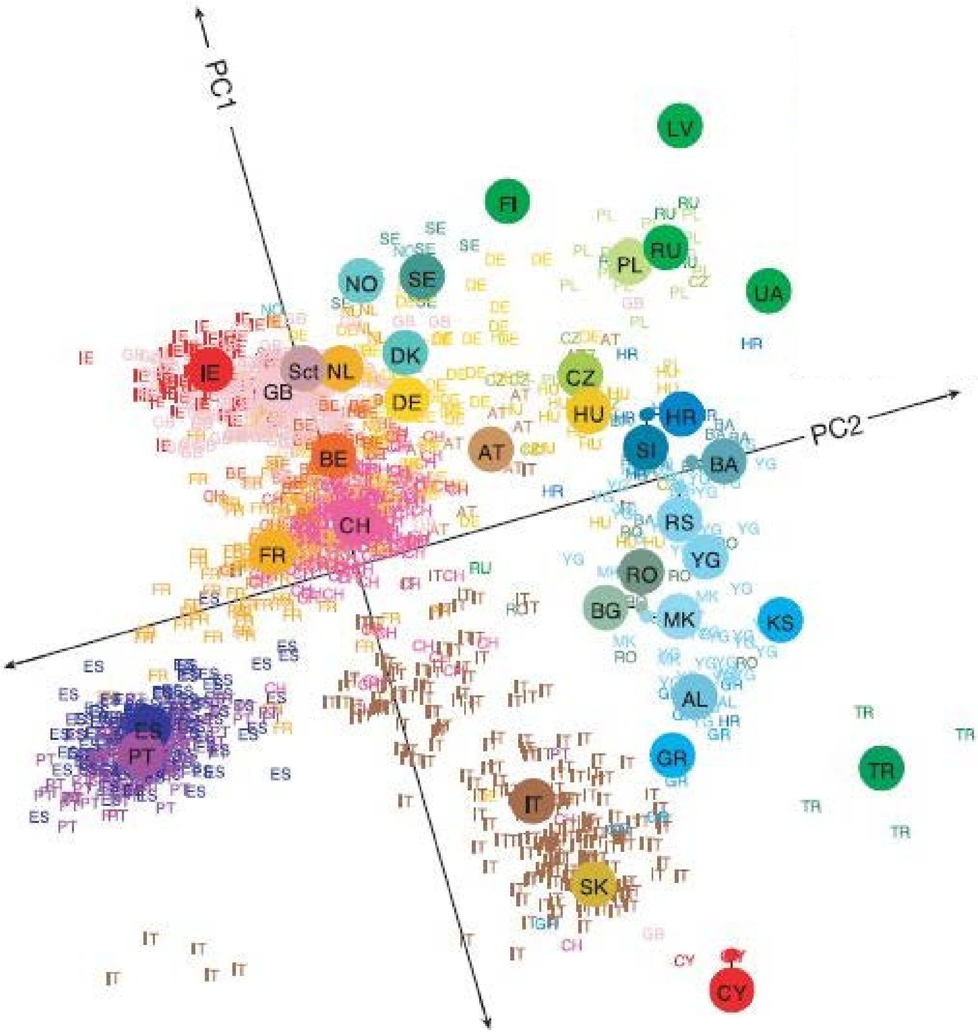
# Principal Component Analysis

- Our model:
  - Direction of largest variation is relevant
  - The rest is “noise”



# Genes

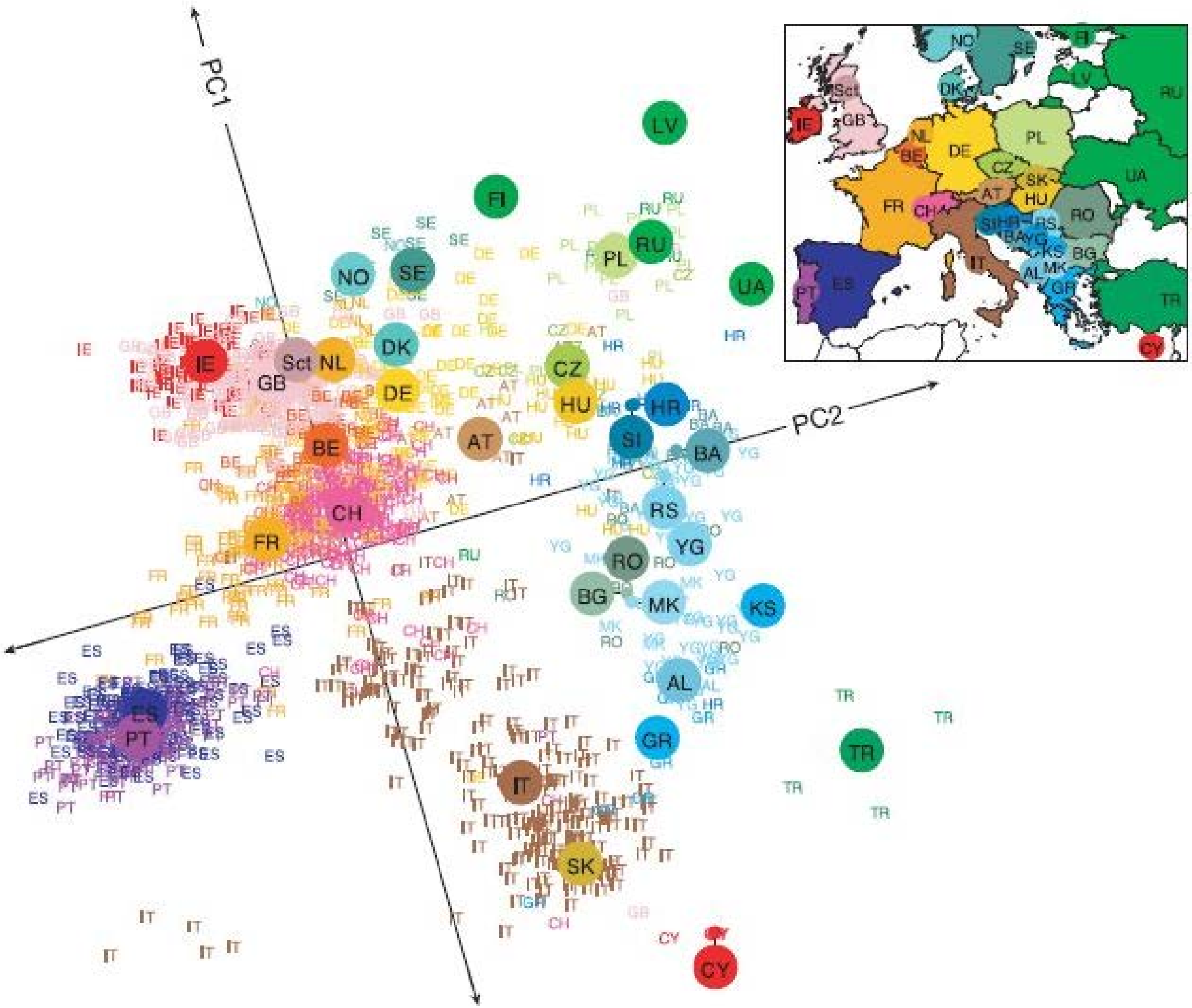
- PCA



PC1
PC2
LV
FI
RU
UA
PL
NO
SE
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
BE
AT
CH
FR
RS
YG
RO
BG
MK
KS
AL
ES
PT
GR
TR
IT
SK
CY

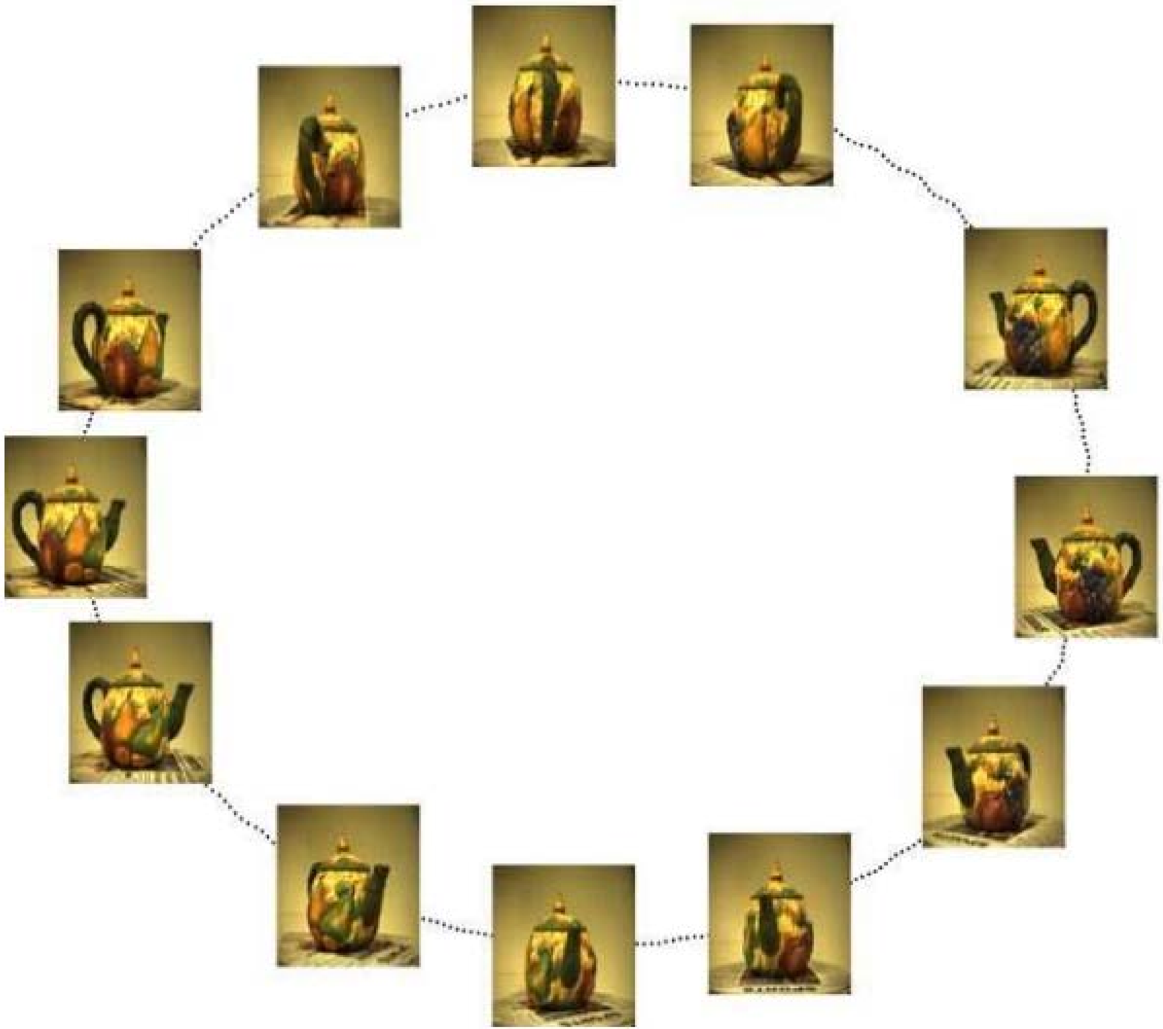
# Genes

- PCA
- Map



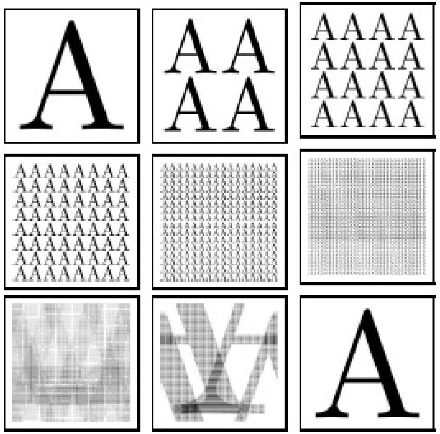
# Nonlinear

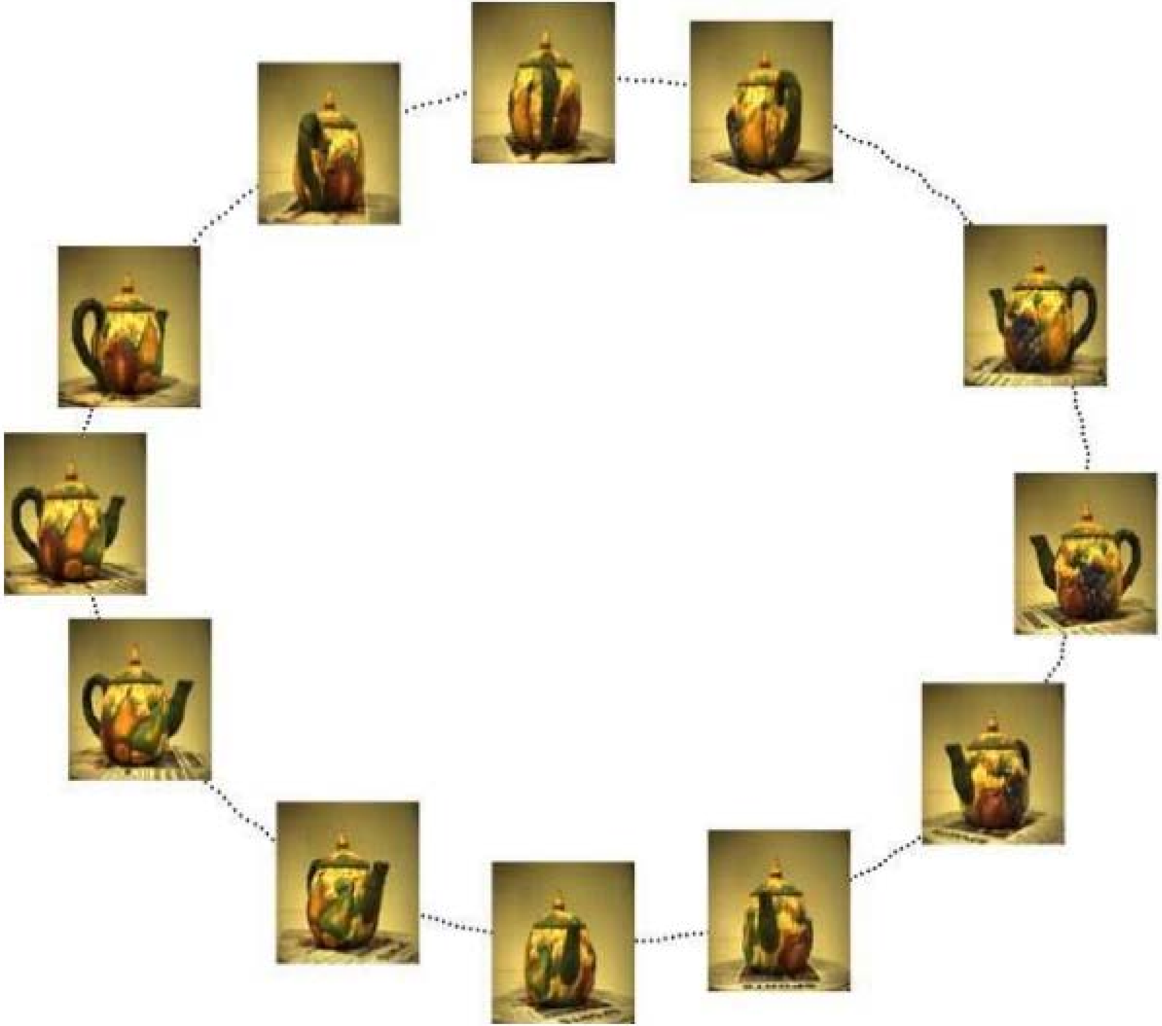
- It's a rotation!



# Nonlinear

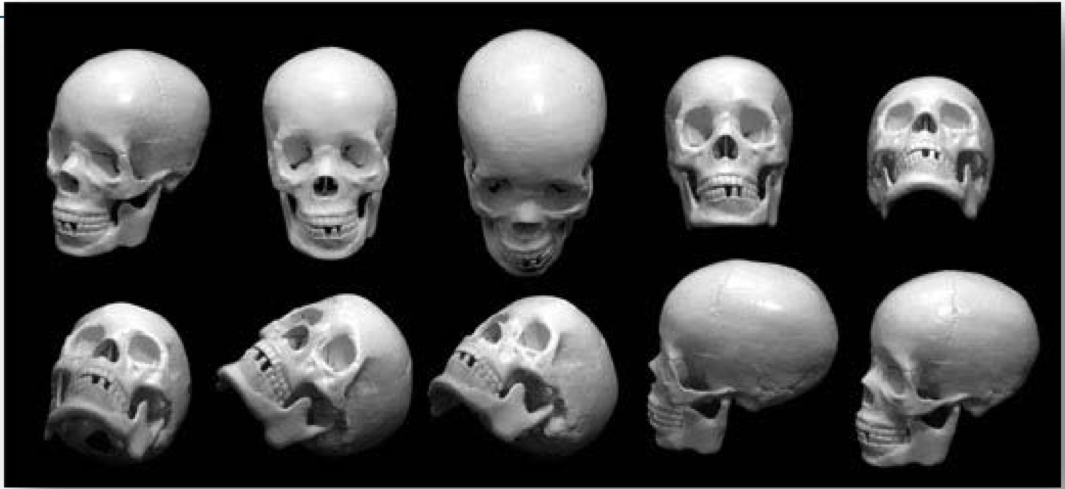
- It’s a rotation!
  - Even if pixels are shuffled!





# More Parameters

- How many?



# More Parameters

- How many?

# Jupyter Notebook

# Python

- General programming language
  - For scripting and prototyping
- Modules for everything
  - Including numerical & statistical packages

# Jupyter

- Interactive analysis
  - Easy to use
  - Web interface
  - Smart rendering

JOHNS HOPKINS
WHITING SCHOOL
*of* ENGINEERING