Features of Success

Finding Structure In High Dimensional Data

CHIDATA

Recap: Simulation v.s. Model Fitting

Fitting: Observe xi, yi, **infer theta**

Regression Problems -> Continuous Y

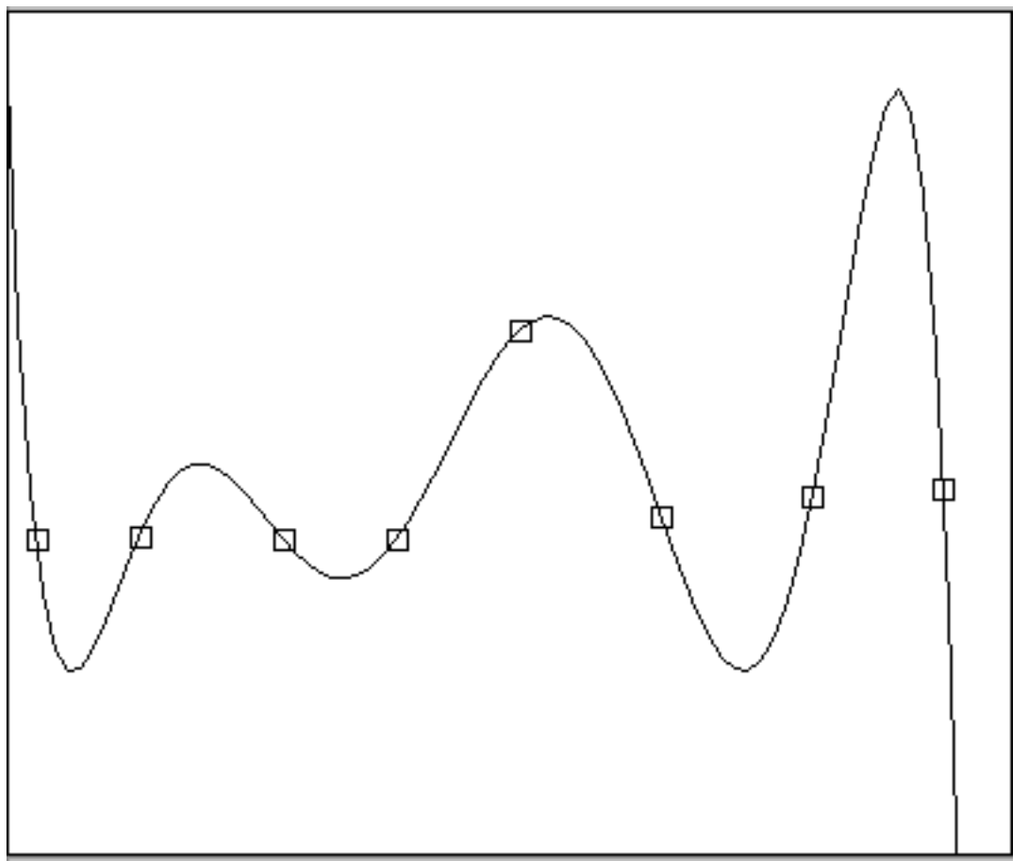
Classification Problems -> Discrete Y

CHIDATA

Recap: Remember This?

How many features do you need to guarantee 0 training error?

N+1 features



CHIDATA

Problem: Actually Pretty Common!

"More Observations Than Individuals"

Stocks: 1000s of securities, Millions of trades

Genomics: 6.4 bn base pairs, 100s Clinical Population

Natural Language: Inf possible word sequences, Millions of Documents

CHIDATA

Problem: Actually Pretty Common!

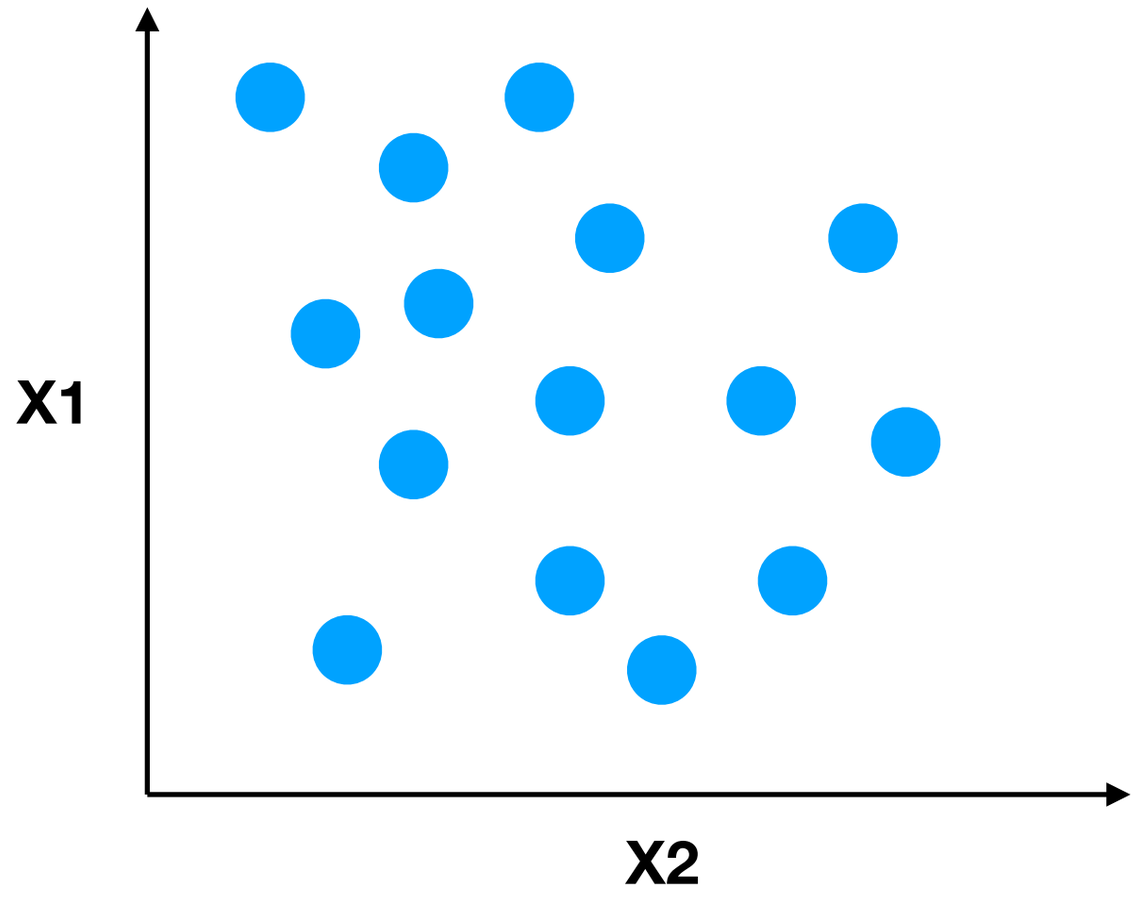
Have to *assume* structure in the data

Low dimensional structure: Large numbers of features are rarely completely independent from each other

CHIDATA

Two Independent Features

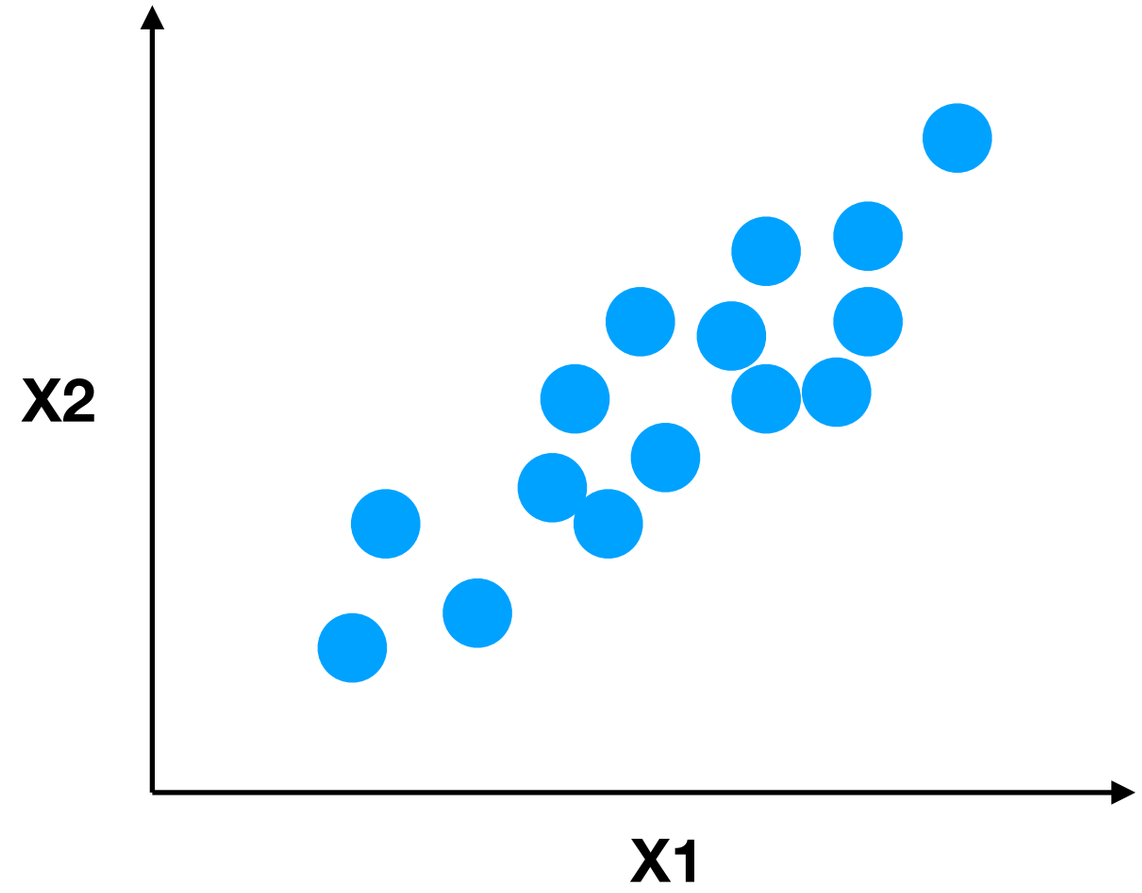
“2 dimensions worth of information”



CHIDATA

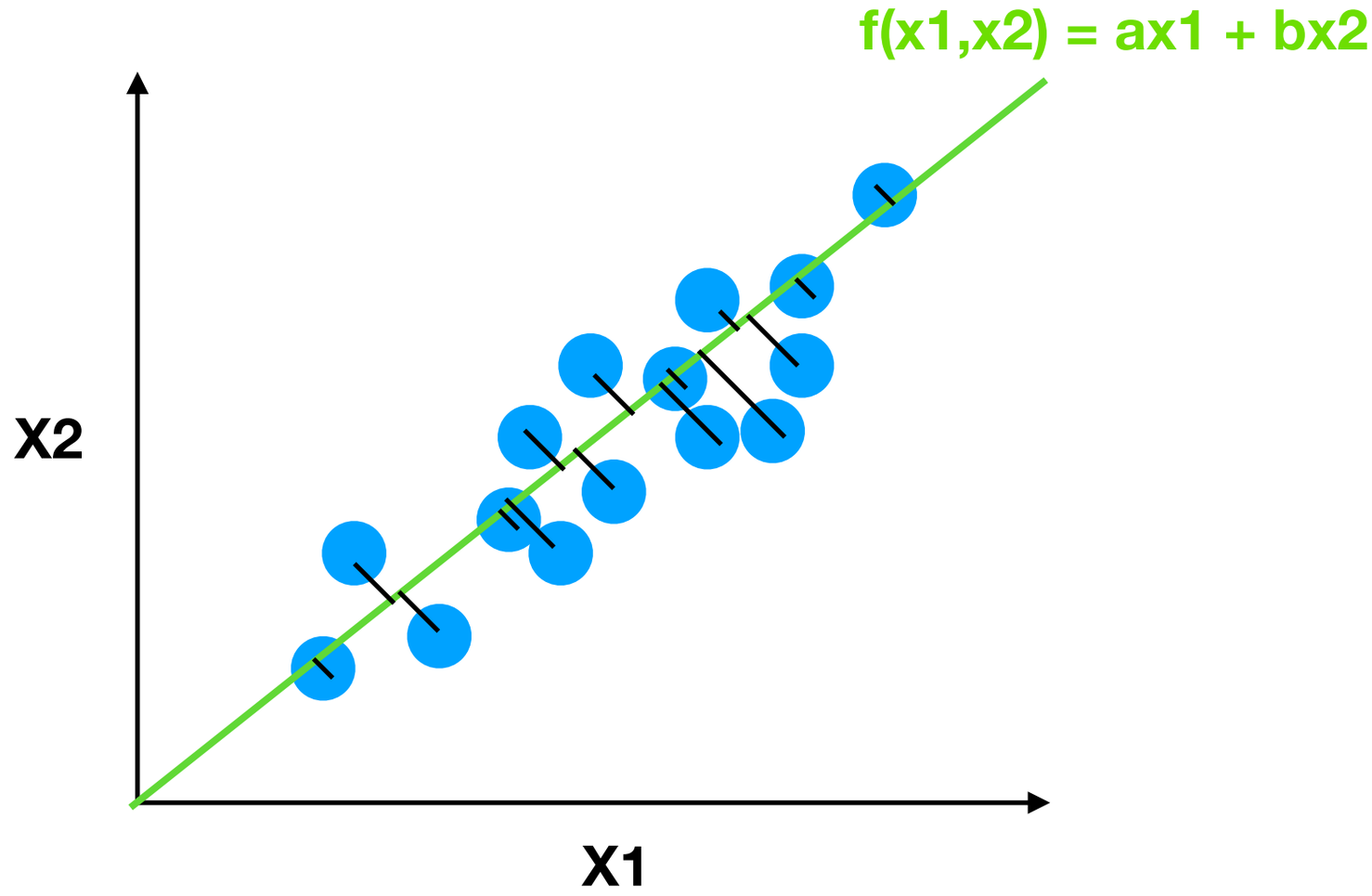
Two Dependent Features

Effectively 1 dimension of information

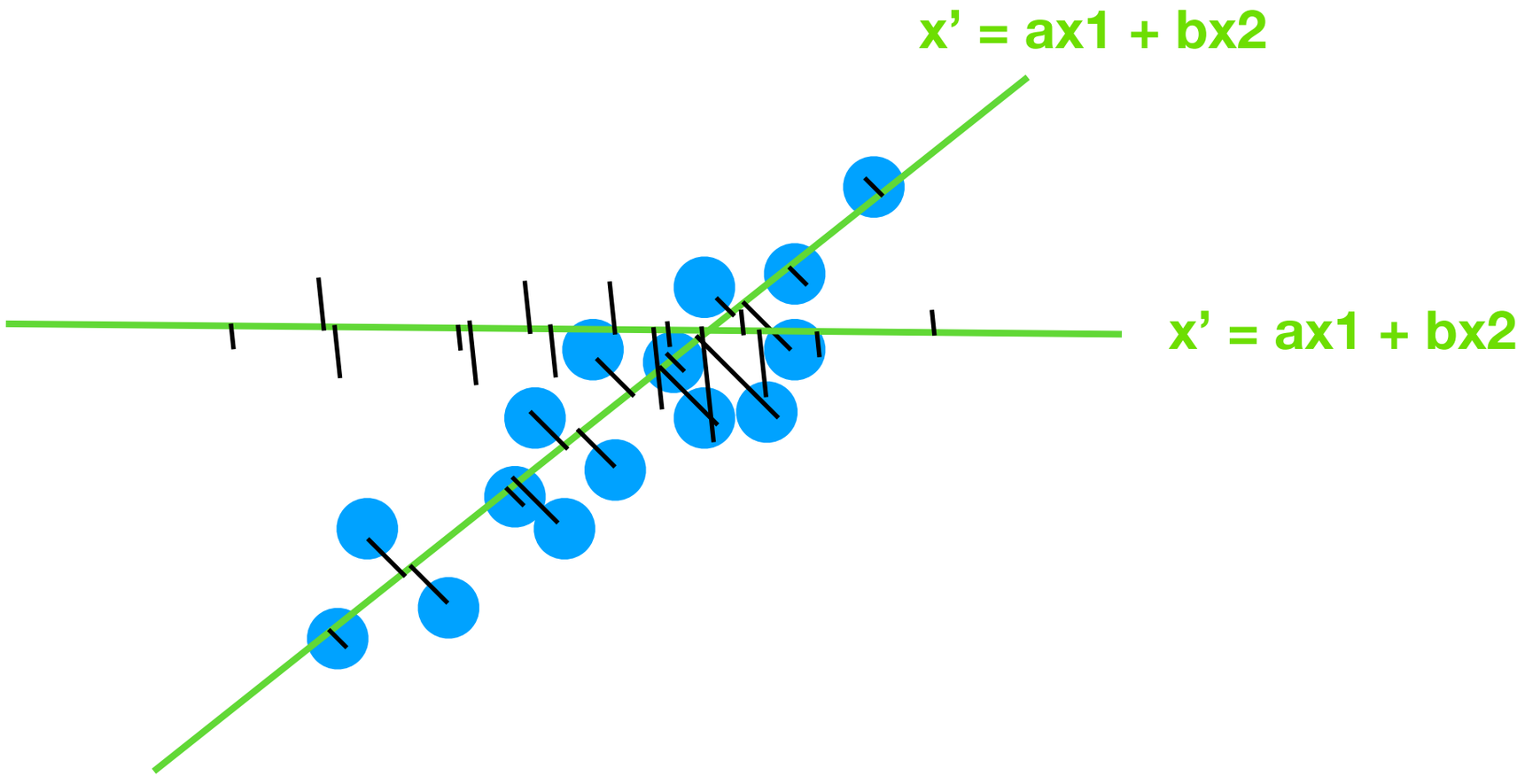


Intuition: you could use x1 to estimate x2

Two Dependent Features



Two Dependent Features

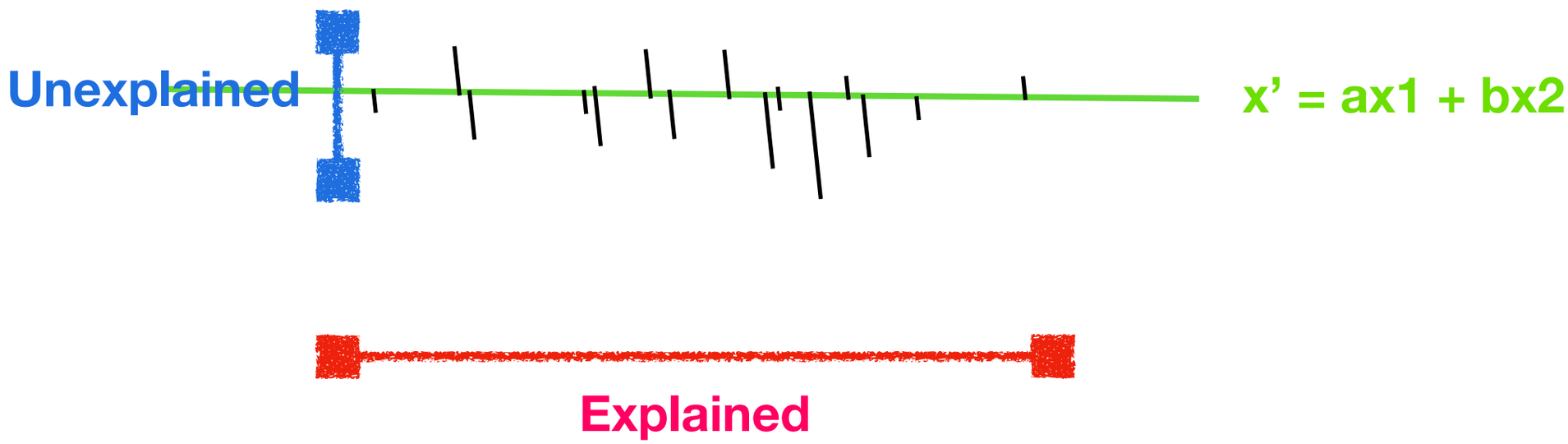


A new 1-d feature that approximately captures both x1 and x2

CHIDATA

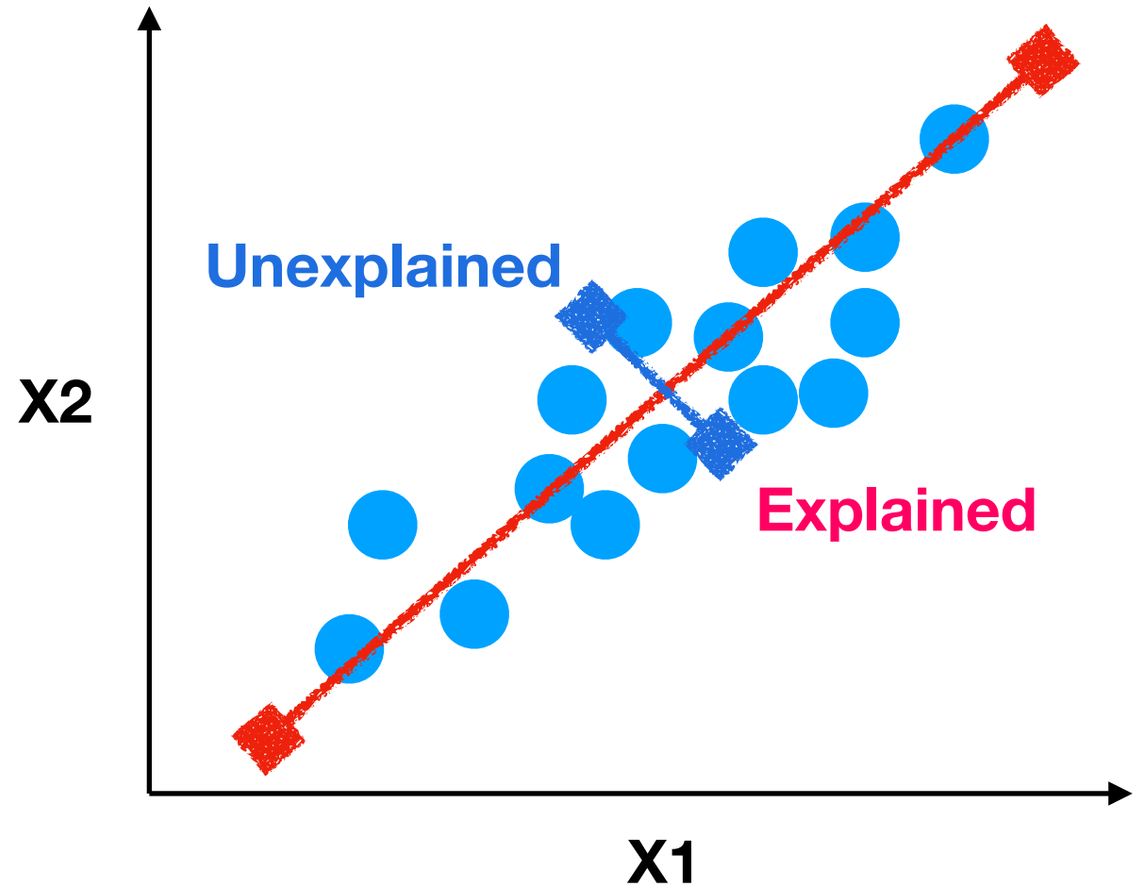
Approximation Quality

Explained Variation: How much variation of the data does the new 1-d feature explain.



CHIDATA

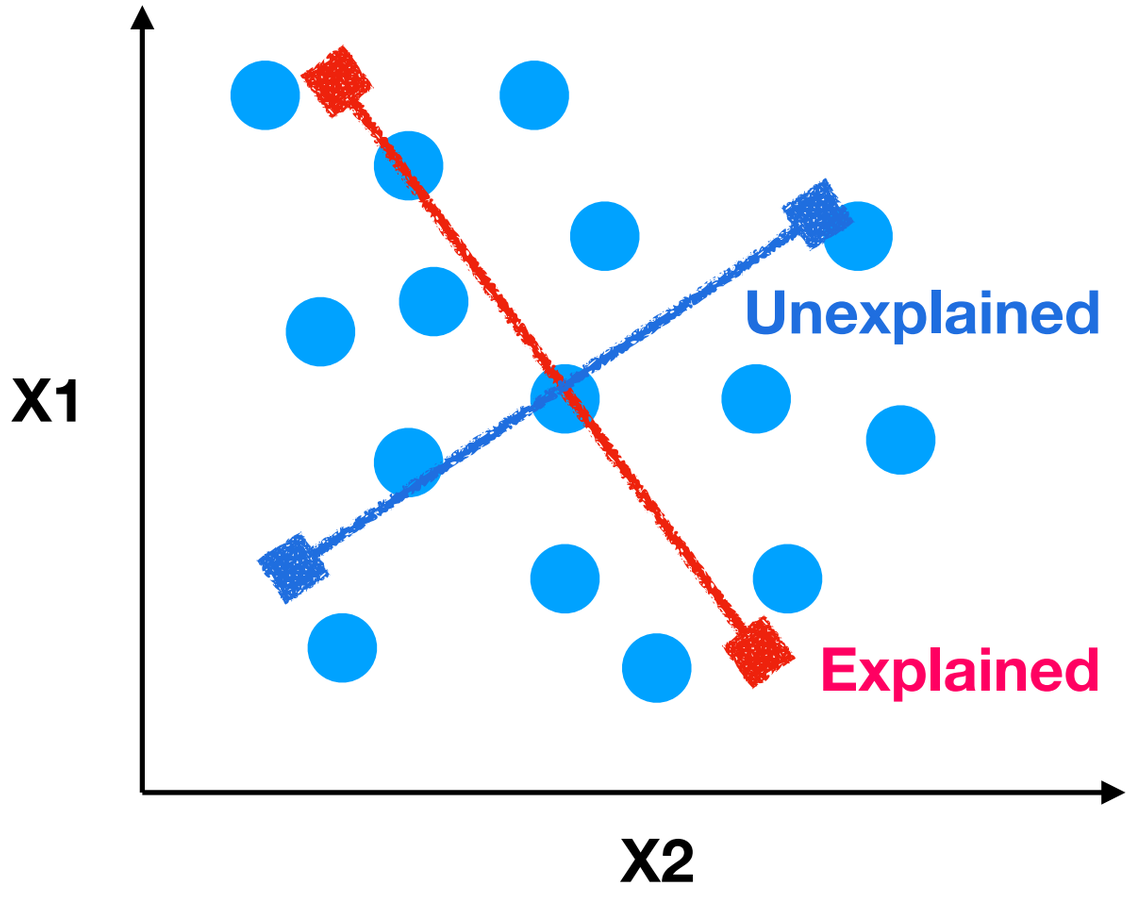
Two Dependent Features



CHIDATA

Two Independent Features

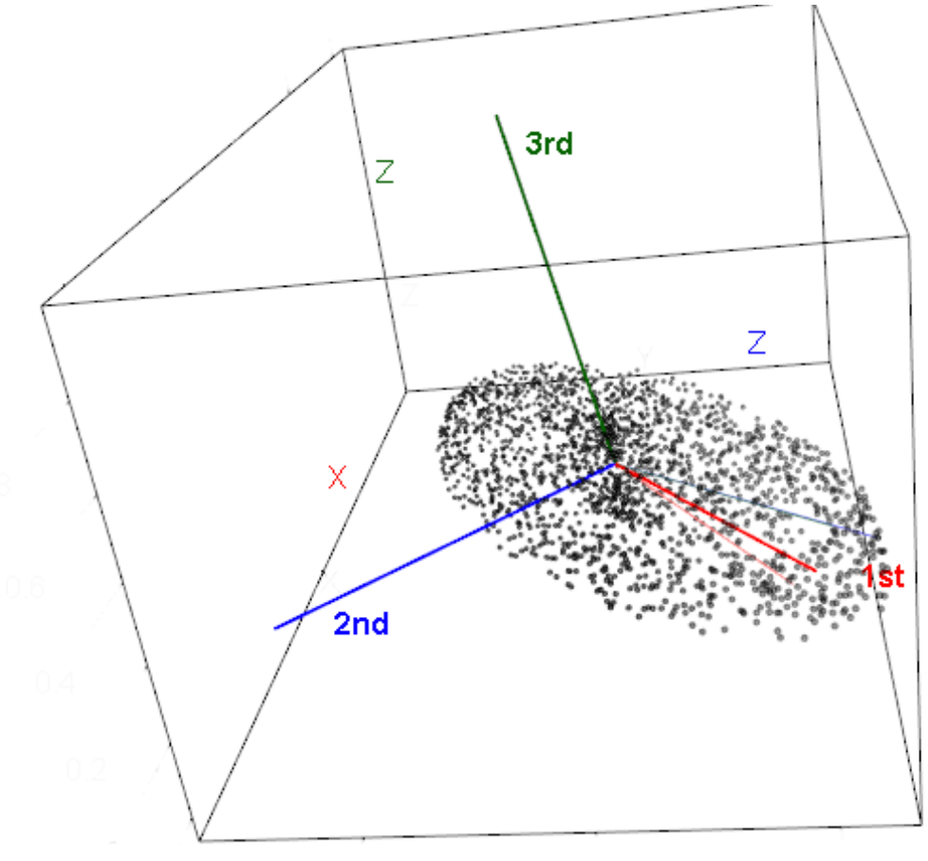
High unexplained variation



CHIDATA

Principal Component Analysis

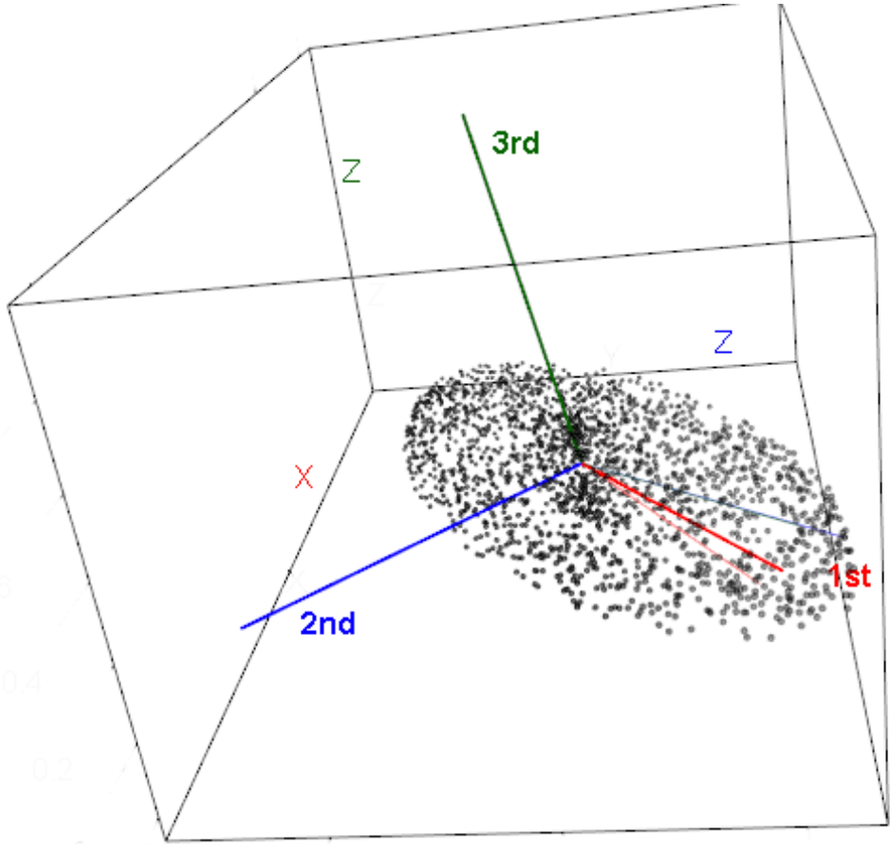
"PCA": Project features onto a d-dimension plane that explains as much variation as possible.

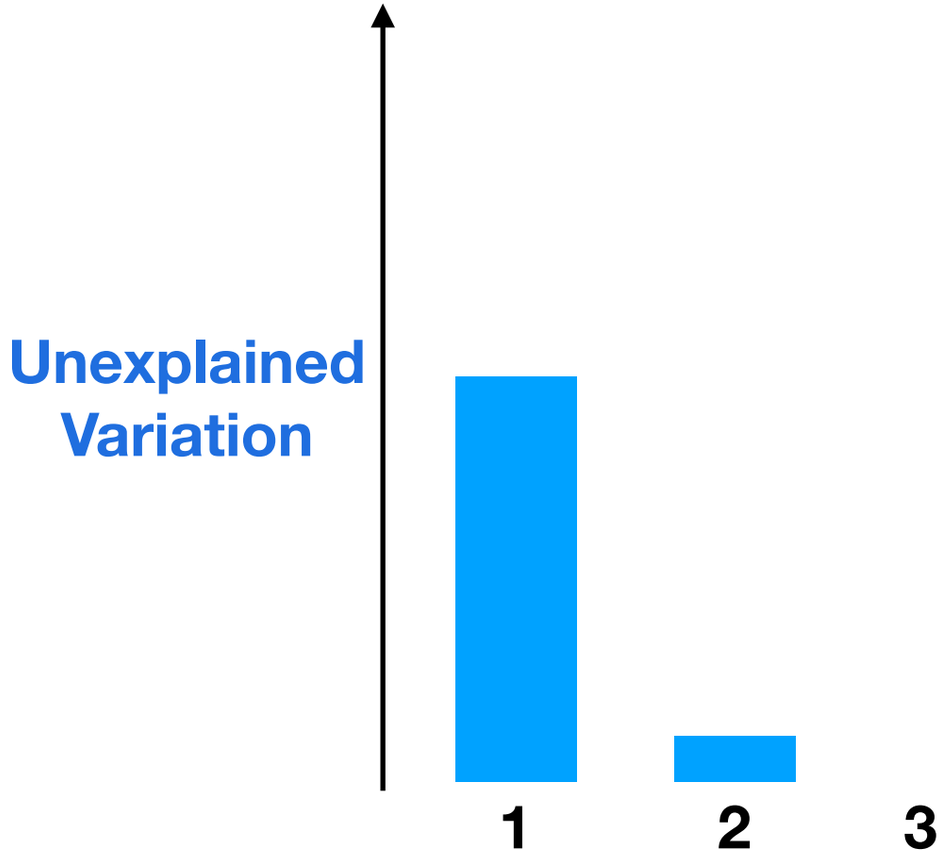


CHIDATA

Principal Component Analysis

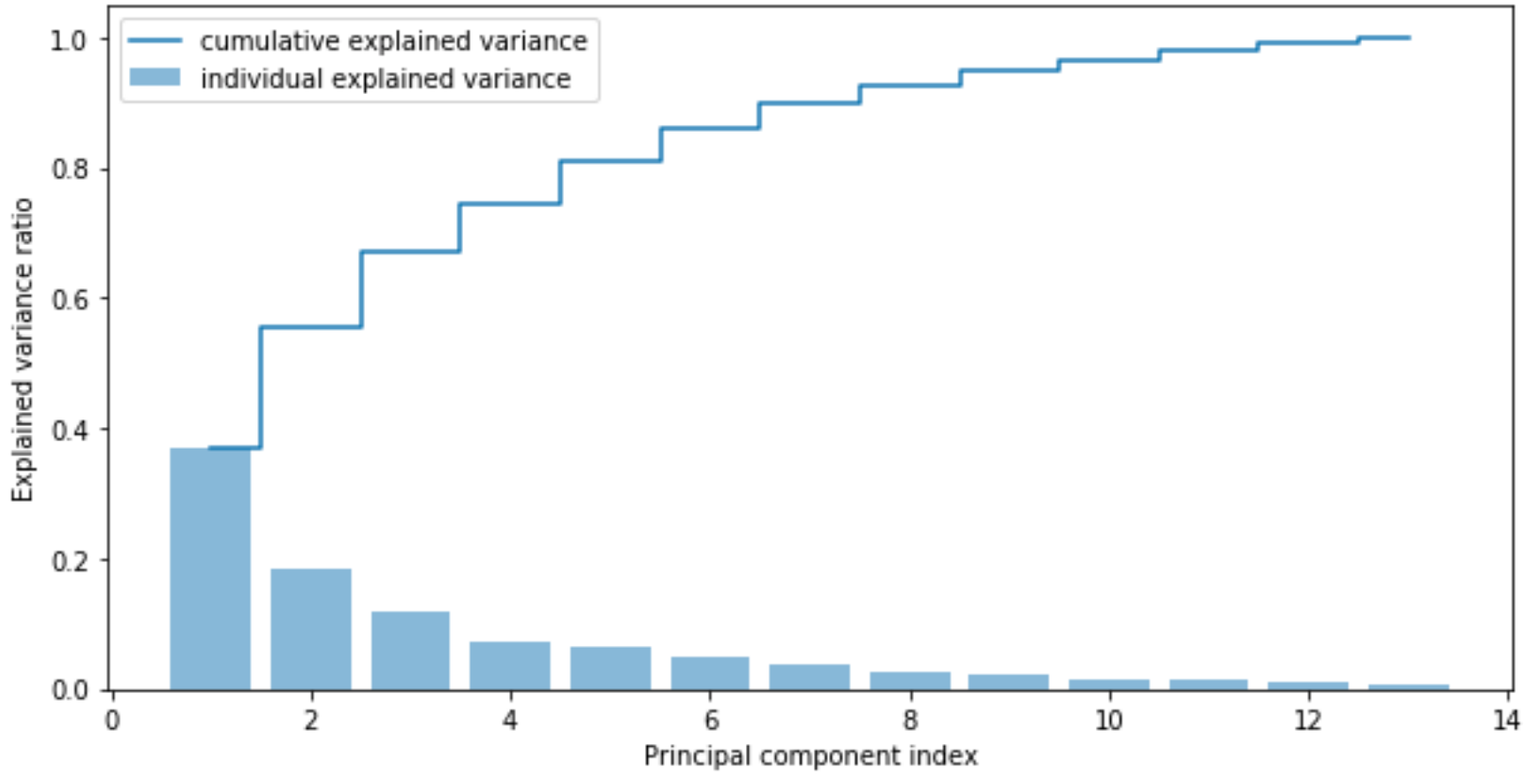
"PCA": Project features onto a d-dimension plane that explains as much variation as possible.





Principal Component Analysis

"PCA": Project features onto a d-dimension plane that explains as much variation as possible.



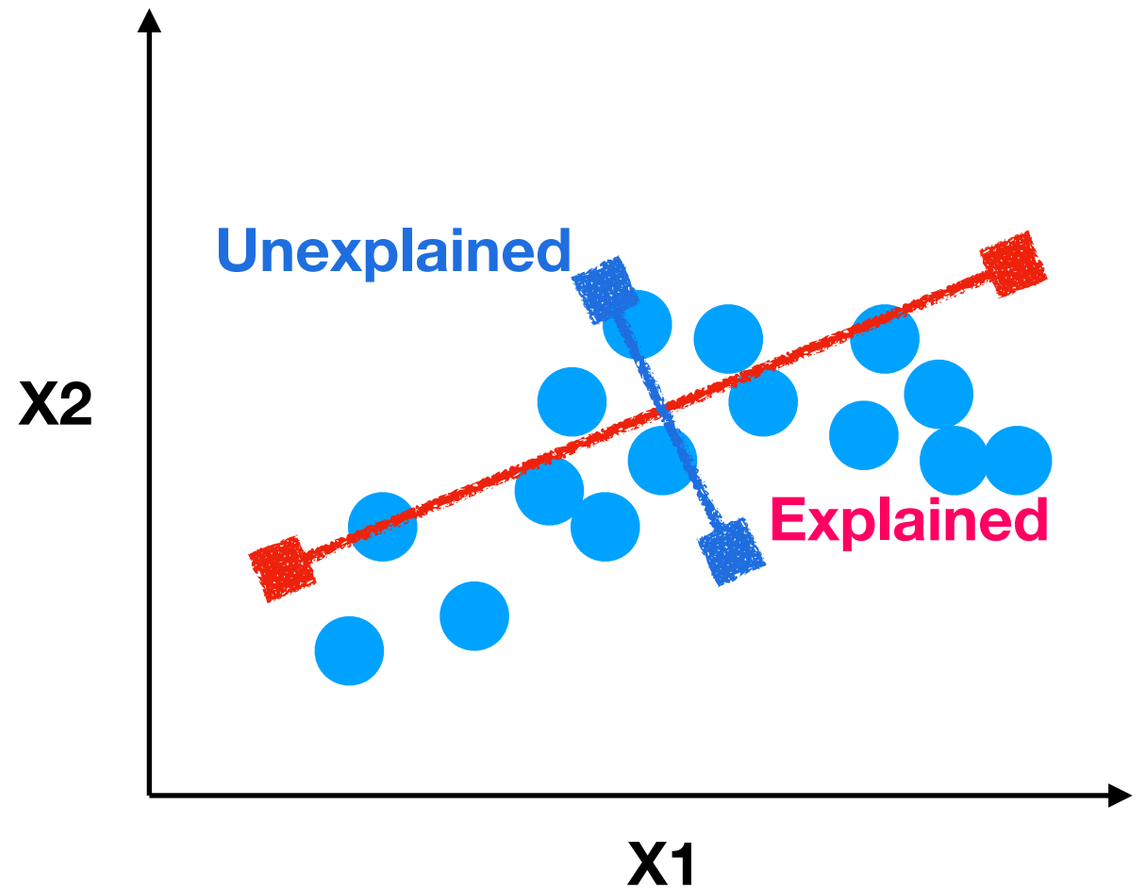
Demo!

194 countries, 21 flag features

CHIDATA

PCA

Have to *assume* structure in the data

Low dimensional structure: Is mostly “linear”

Two Dependent Features