COMP 330: Over-Fitting and Regularization

Chris Jermaine and Kia Teymourian

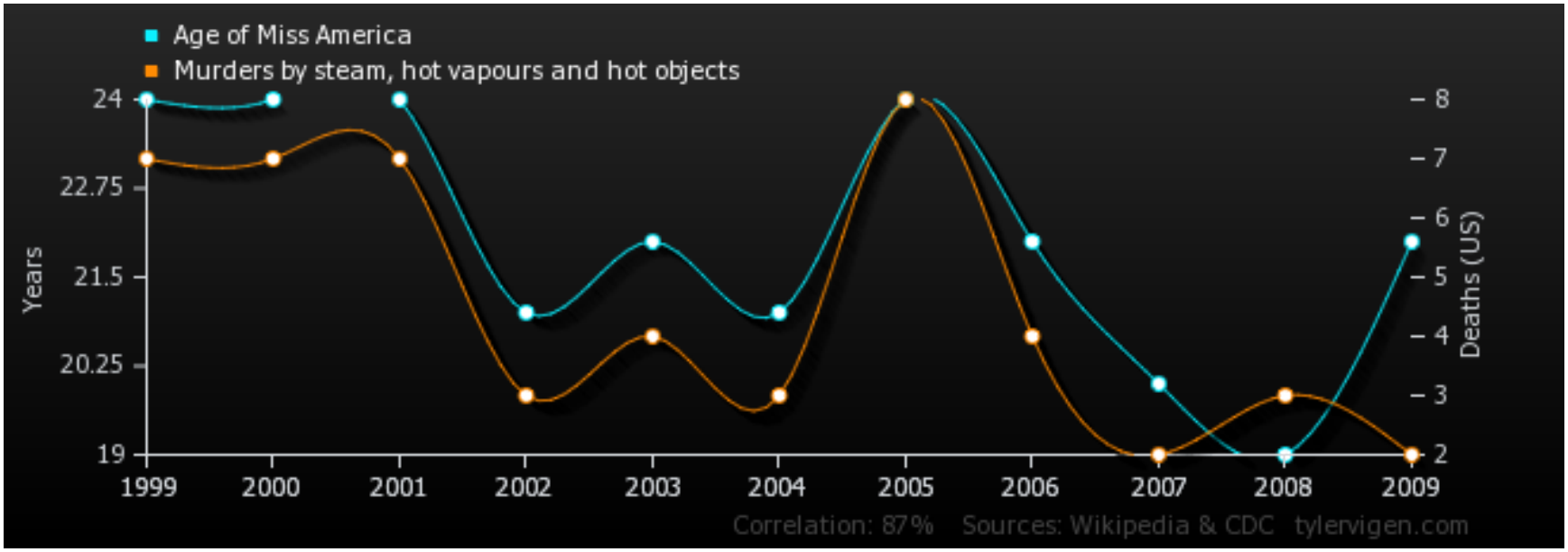
Rice University

Over-Fitting

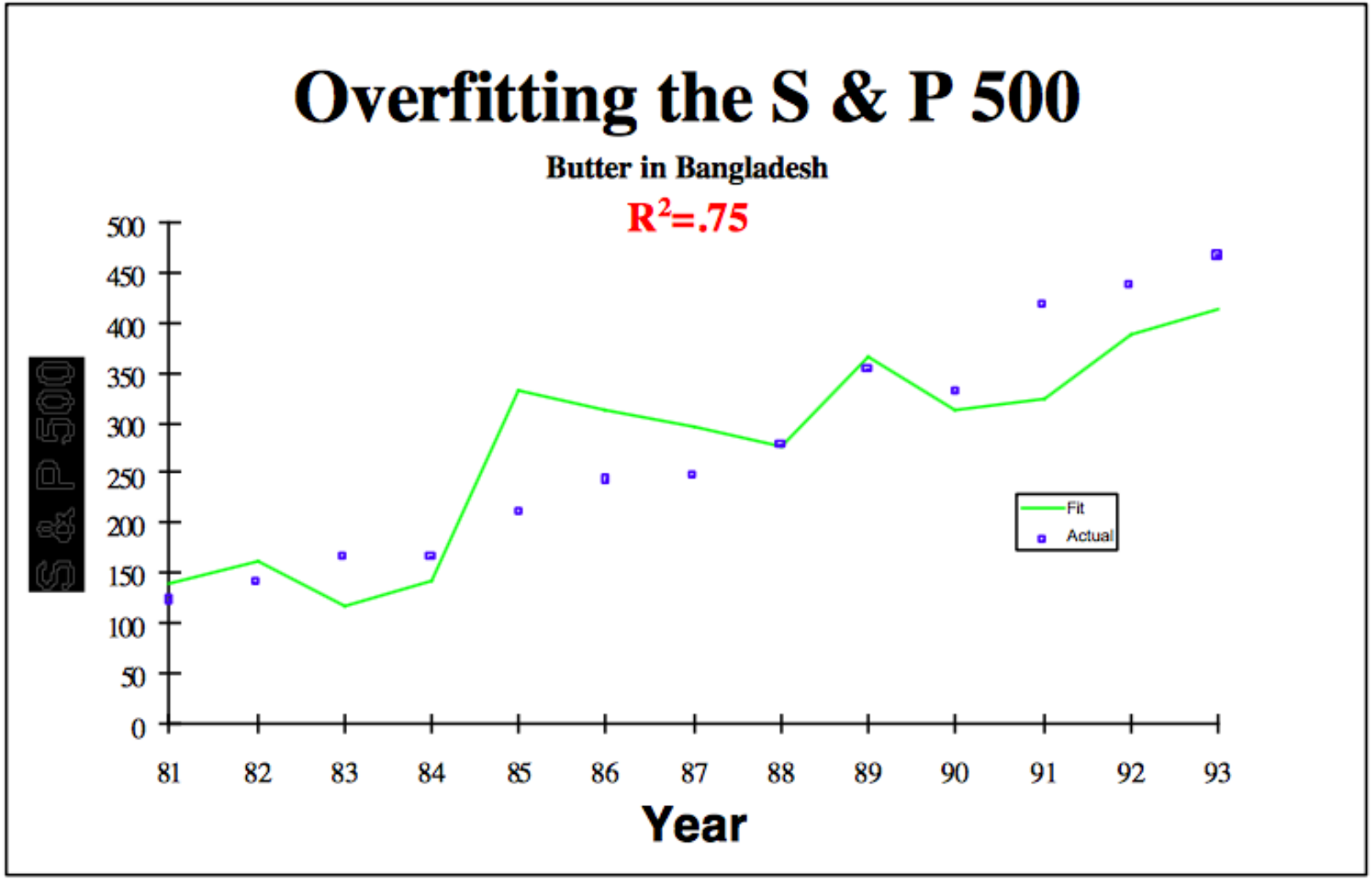
Fundamental problem in data science

- Given enough hypotheses to check...
- One of them is bound to be true

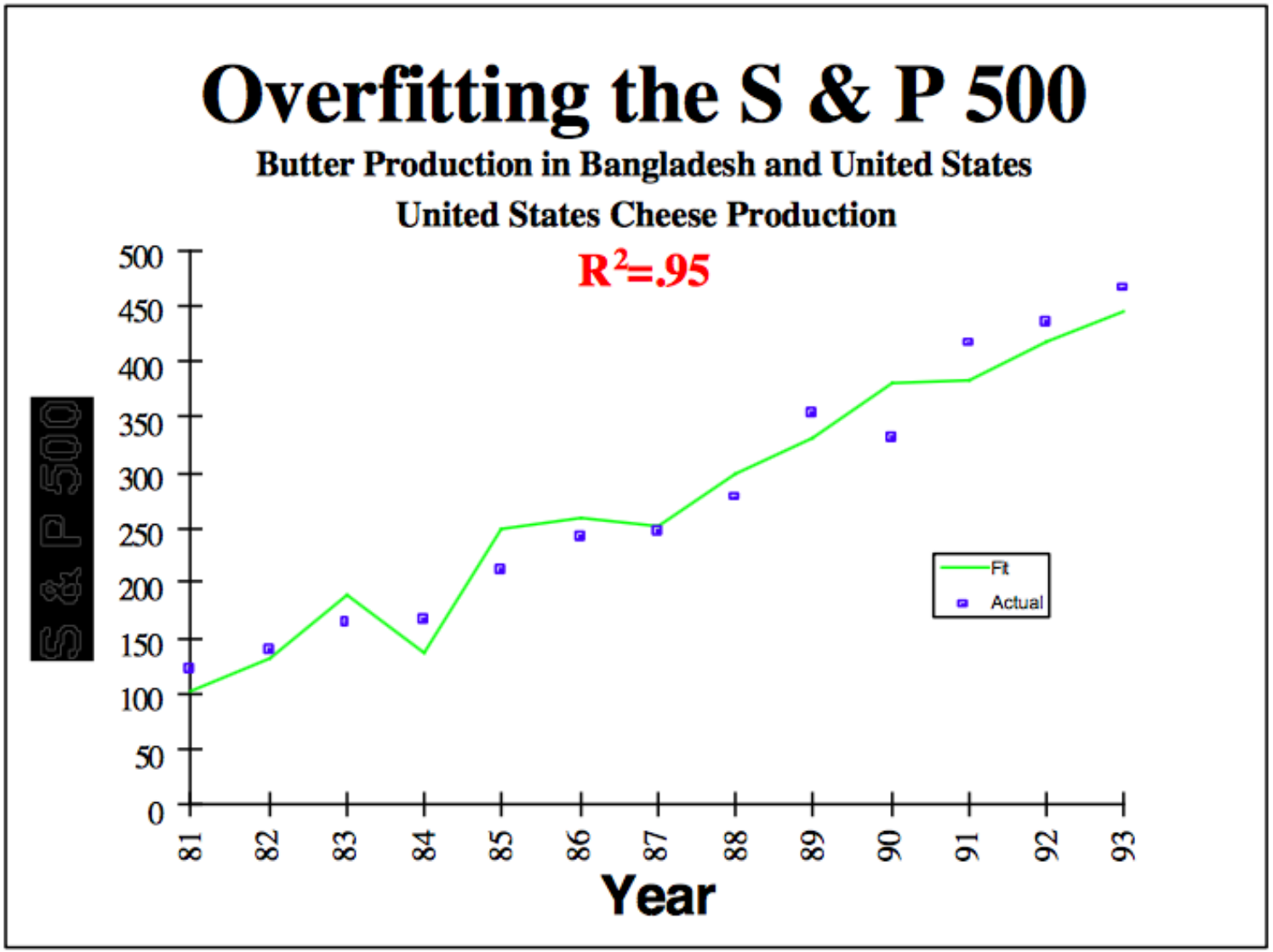
Miss America and Murder-By-Steam



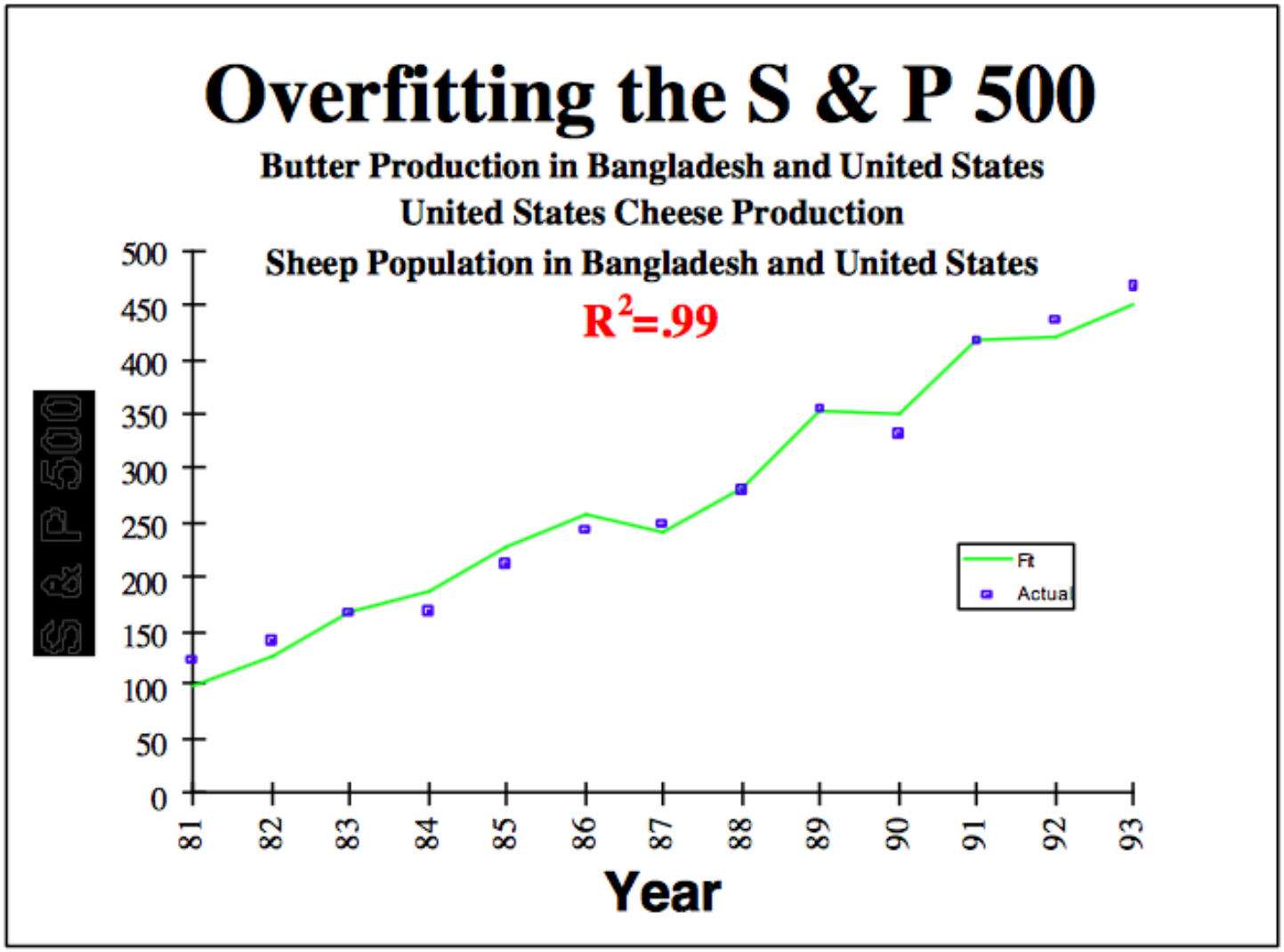
Predicting the S&P 500



Predicting the S&P 500



Predicting the S&P 500



None of these Models Likely to Generalize

That means:

- They've learned the input data
- Not any underlying truth
- When deployed in the field, likely to fail

"Data Mining"

Was originally a derogitory term used by stats

- Meant that you could always find something if you look hard enough

Detecting Over-Fitting

Detection method number 1: sniff test

Detection method number 2: independent validation and test sets

- Avoid temptation!

Detection important...

- Avoiding altgother is as well!

Avoiding Over-Fitting: Occam's Razor

The Razor stated simply: When you have many hypotheses that match observd facts equally well, the simplest one is preferred.

- Been around for a long time!
- Credited to William of Ockham (died 1347)
- First stated explicitly by John Punch, 1639: "Entities must not be multiplied beyond necessity"

Avoiding Over-Fitting: Occam's Razor

Of course, not that simple

- You never have a large number of equally good hypotheses
- Best you can do: have a bias towards simple models...
- Under the assumption they will generalize well

The Bias-Variance Trade-Off

In a nutshell, we have two main sources of error in learning

- "Bias": error from incorrect model assumptions
- "Variance": sensitivity of model to bad data

Understanding the Trade-Off

Expected squared error or any prediciton is:

$$E[(Y - \hat{f}(X))^2]$$

- Here:
- Y is the output we are trying to predict
- $\hat{f}(.)$ is the model we are learning (is a random variable!)
- X is the observed data (ex: set of regressors)
- Y is the value we are trying to preict from X

Understanding the Trade-Off

Expected squared error of any prediciton is:

$$\begin{aligned} E[(Y - \hat{f}(X))^2] &= E[Y^2 + \hat{f}^2(X) - 2Y\hat{f}(X) \\ &= E[Y^2] + E[\hat{f}^2(X)] + E[2Y\hat{f}(X)] \\ &= Var(Y) + Var(\hat{f}^2(X)) + (\hat{f}^2(X) - E[\hat{f}(X)])^2 \\ &= Var(Y) + Var(\hat{f}^2(X)) + Bias(\hat{f}(X))^2 \end{aligned}$$

Means error is a sum of:

- "Looseness" of relationship between X and Y
- Sensitivity of the learner to variability of the training data (variance)
- Inability of the learner $\hat{f}$ to learn the relationship between X and Y (bias)

It is the second one (variance) that leads to over-fitting

Ideally, Reduce Both Bias and Variance!

Unfortunately, not possible

"In real life"

- We don't know the real model–bias is guaranteed
- Since we ARE wrong, choose a general model and lots of features
- Hence variance (over-fitting) is also guaranteed
- Best we can do: choose sweet spot where error is minimized

Regularization

Massively important idea in data science

- In a nutshell:
- Give learning algorithm ability to choose complexity of model
- Automatically choose the correct trade-off between bias and variance

Done by adding a penality term to objective function

- Penalizes model for complexity

Regularization

Typically, penality is a norm computed over the model

Recall, l_p norm of a vector $\langle x_1, x_2, ... \rangle$ computed as:

$$\left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Common penalties are l_1, l_2

Example: Logistic Regression

Standard objective function is:

$$\sum_i y_i\theta_i - \log(1 + e^{\theta_i})$$

where θ_i is $\sum_j x_{i,j} r_j$

Change objective function to:

$$\sum_i y_i\theta_i - \log(1 + e^{\theta_i}) + \lambda||r||_p$$

here $||r||_p$ is the l_p norm of the regression coeficients

- If $p = 1$ have "the lasso"
- If $p = 2$ have "ridge regression"
- λ controls the magnitude of the penalty
- Typically, try different values of λ during validation

Closing Remarks

When regularizing, important to normalize data

- That is, transform so mean, variance are one
- Why?

Bayesians argue they are protected from over-fitting

- A good prior protects against complex models
- In fact, "the lasso" closely related to BLR with Laplace prior on r
- Ridge regression closely related to BLR with Normal prior on r

Questions?