

Explanatory vs. Predictive Statistical Modeling



Galit Shmuéli



UNIVERSITY OF
MARYLAND

ROBERT H. SMITH
SCHOOL OF BUSINESS



2008 Statistics Workshop
United States Military Academy

What is R² ?

R-square	Adj R-Sqr	Probability	# of Obs
0.269	0.266	0.000	506

REGRESSION COEFFICIENTS

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance
Intercept	0.193182	0.000000	0.999
Group 1	-0.050325	-0.036319	0.393
Group 2	-0.014610	-0.010544	0.487
Group 3	0.806818	0.512045	0.000

A black and white photograph of a forest path. The path is a light-colored dirt or gravel trail that curves through a dense forest of trees and bushes. The lighting is dramatic, with strong highlights on the path and deep shadows in the surrounding foliage. The text 'From Goals' is overlaid in white, bold, sans-serif font on the left side of the image.

From Goals

To Models

Goal: Explain

theory

causality

retrospective

Goal: Predict

utility

relevance

new theory

association

prospective

Philosophy of Science

**“Explanation and prediction
have the same logical structure”**

Hempel & Oppenheim, 1948

**“It becomes pertinent to investigate the
possibilities of predictive procedures
autonomous of those used for explanation”**

Helmer & Rescher, 1959

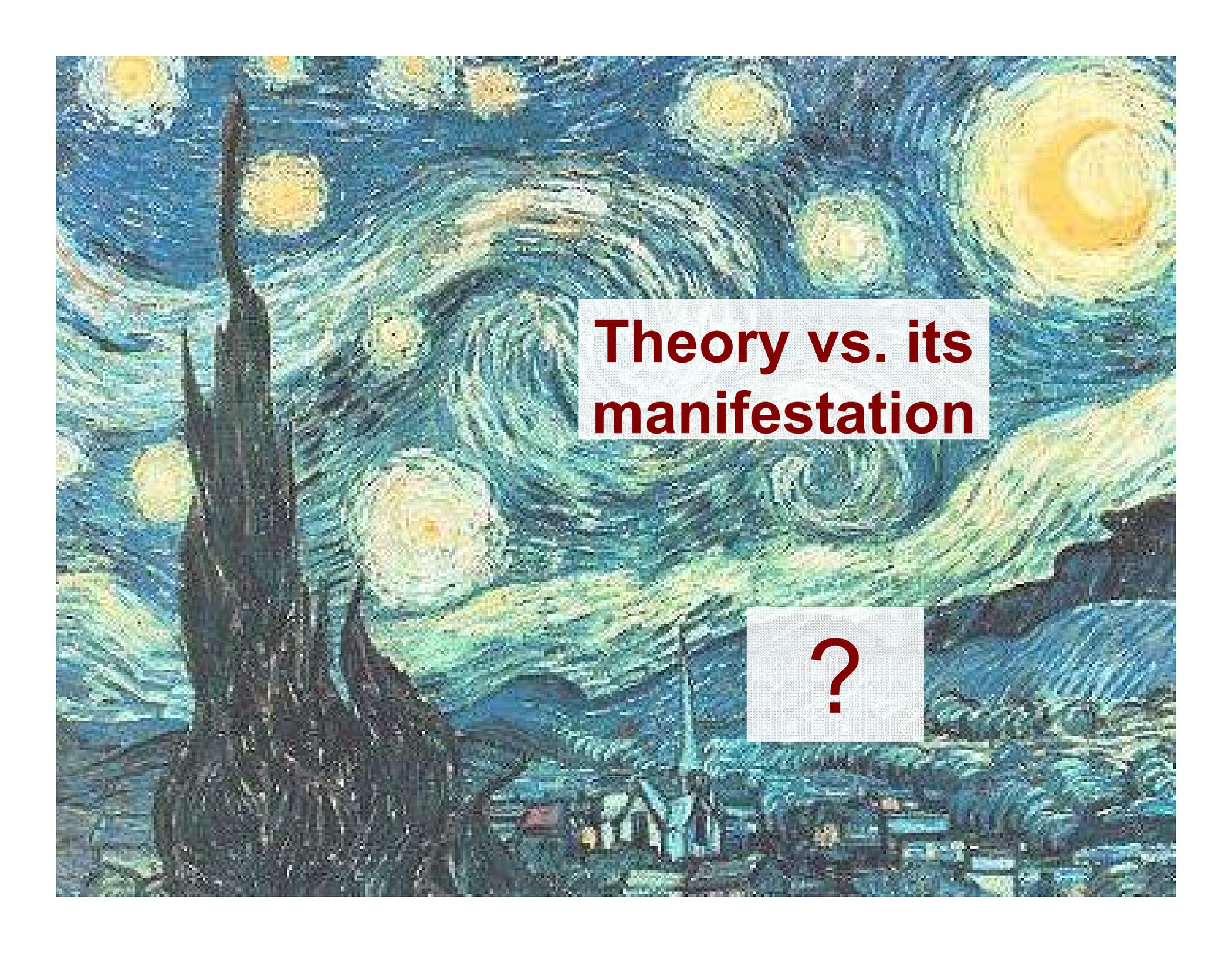
**“Theories of social and human behavior
address themselves to two distinct goals of
science: (1) prediction and (2) understanding”**

Dubin, *Theory Building*, 1969

Best explanatory model

≠

Best predictive model

The background of the slide is a reproduction of the painting 'The Starry Night' by the Dutch Impressionist painter J.M.W. Turner. The painting depicts a night scene with a dark, turbulent sky filled with swirling, luminous clouds and numerous bright, glowing stars. In the foreground, a dark, silhouetted landscape features a prominent, tall, dark structure on the left, possibly a church spire or a tower, and a small, brightly lit building in the distance. The overall mood is one of awe and wonder, capturing the vastness and beauty of the night sky.

**Theory vs. its
manifestation**

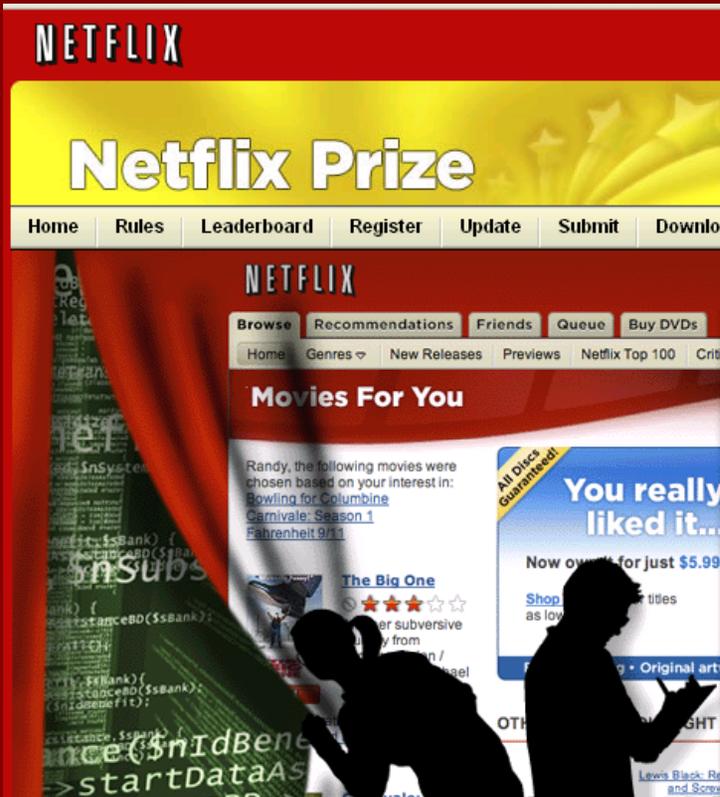
?

A composite image of Earth and the Moon in space. The Earth is in the upper left, showing continents and oceans. The Moon is in the lower center, showing its craters. The background is black space.

Statistics = Explain

Data Mining = Predict

Predict \neq Explain



+



?

Predict \neq Explain

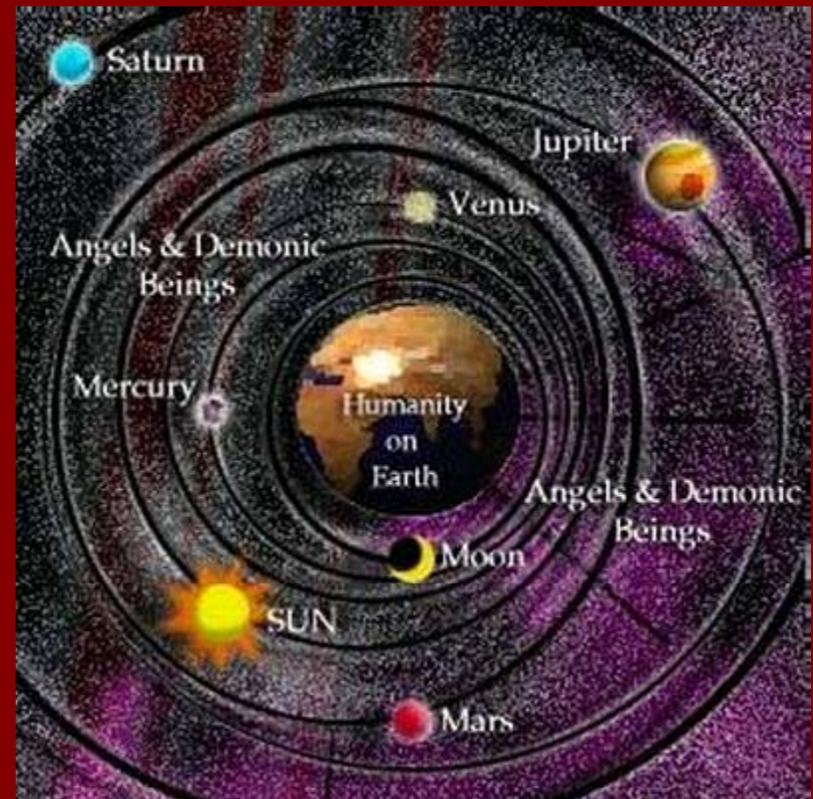
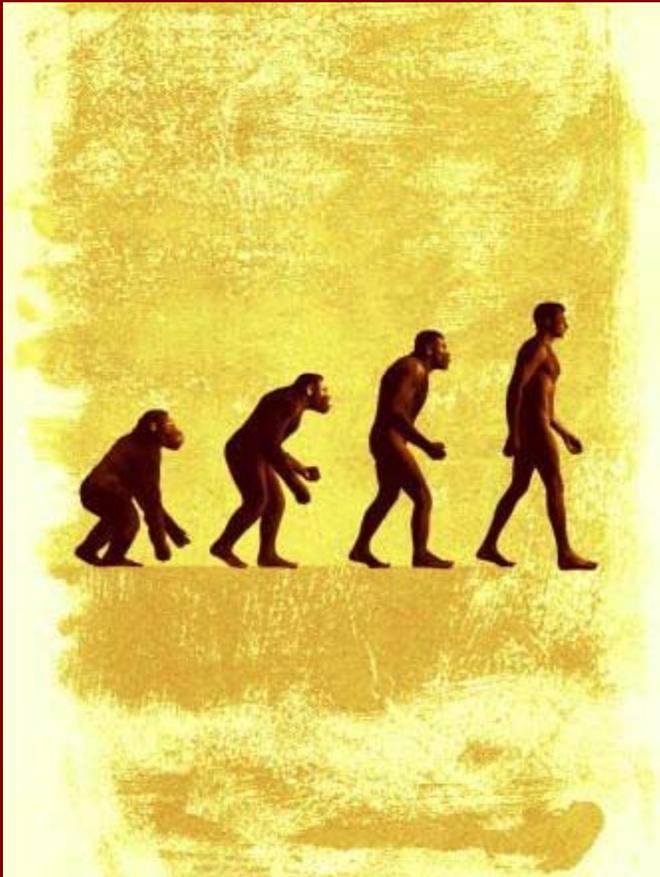
The FDA considers two products **bioequivalent** if the 90% CI of the relative mean of the generic to brand formulation is within 80%-125%



“We are planning to... develop predictive models for bioavailability and bioequivalence”

Lester M. Crawford, 2005
Acting Commissioner of Food & Drugs

Controversy



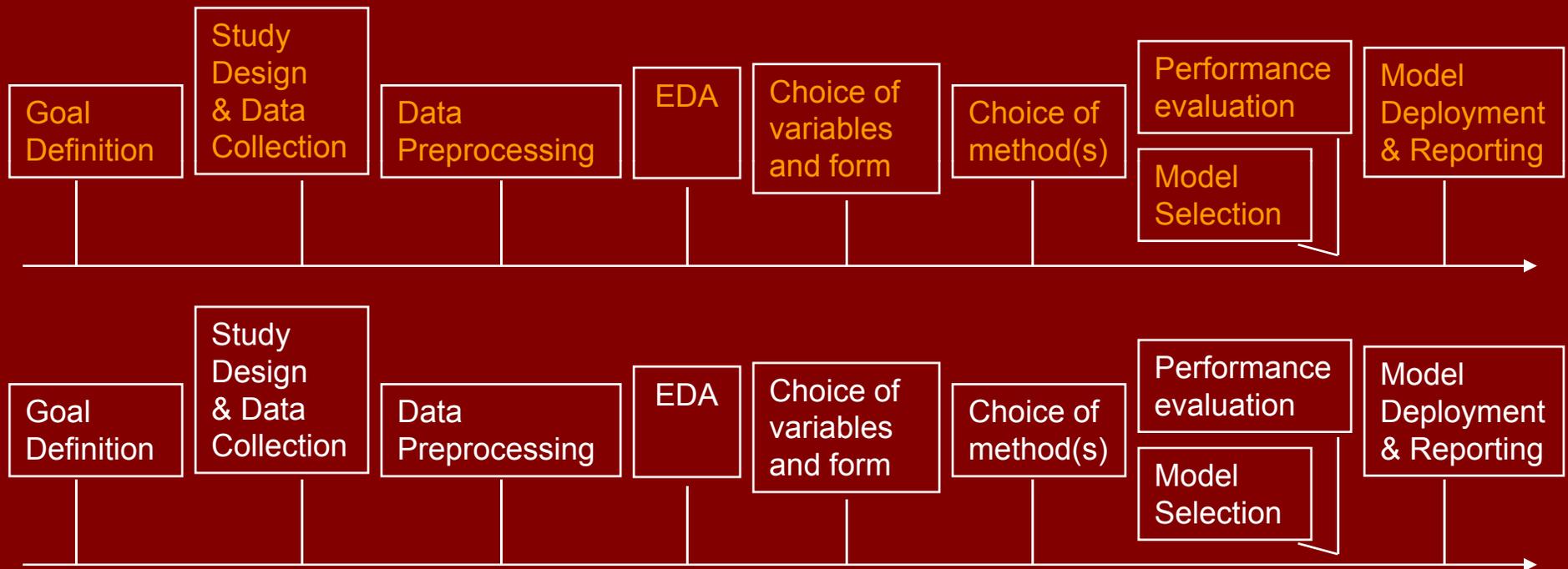
Ptolemaic astronomy

China's Diverging Paths, photo by Clark Smith

Two statistical modeling paths



Find the Differences



Model Deployment & Reporting

explain phenomena

predict new data

support / refute theory

reality check

evaluate predictability

lead to new theory

interpretation

out-of-sample

p-values

**prediction
accuracy**

R^2

**Performance
Metrics**

costs

goodness-of-fit

run time

type I,II errors

over-fitting

What is Optimized?

Bias or Prediction MSE

$$\begin{aligned}MSE &= E(Y - \hat{Y}|x)^2 = E(Y - \hat{f}^*(x))^2 \\&= E(Y - f(x) + f(x) - f^*(x) + f^*(x) - \hat{f}^*(x))^2 \\&= E(Y - f(x))^2 + (f^*(x) - f(x))^2 + E(\hat{f}^*(x) - f^*(x))^2\end{aligned}$$

$Var(Y) =$
uncontrollable

$bias^2 =$ model
misspecification

estimation
(sampling variance)

Shrinkage

variance

bias

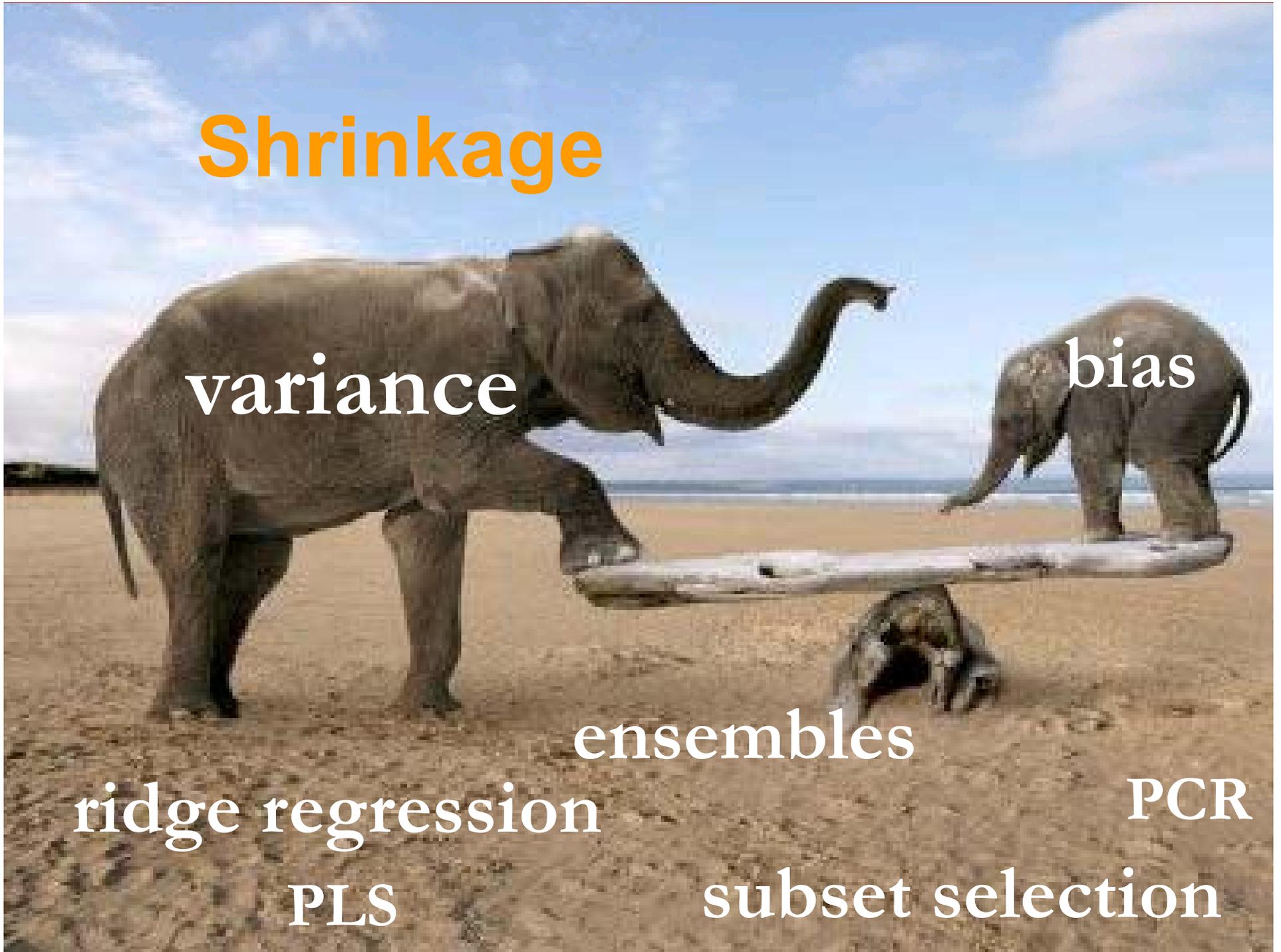
ensembles

ridge regression

PCR

PLS

subset selection



Which Variables?



Multicollinearity?

A, B, A*B?

theory associations

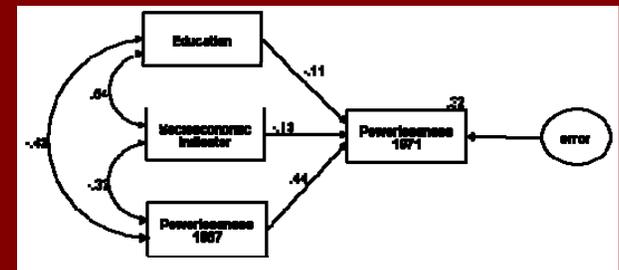
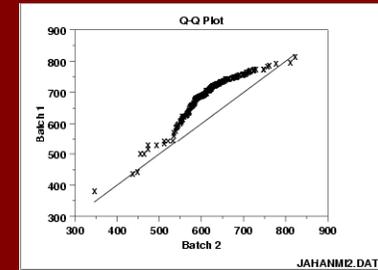
ex-post
availability

summary stats **plots**

Explore!



trends **outliers**



PCA
SVD

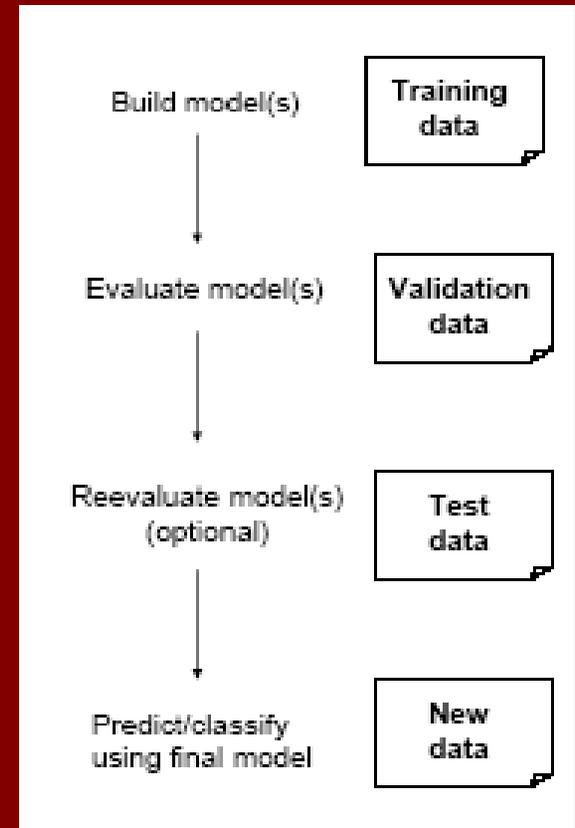
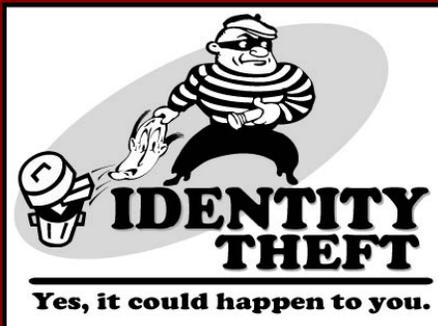
Data Preprocessing



missing



**reduced-
feature
models**



partitioning

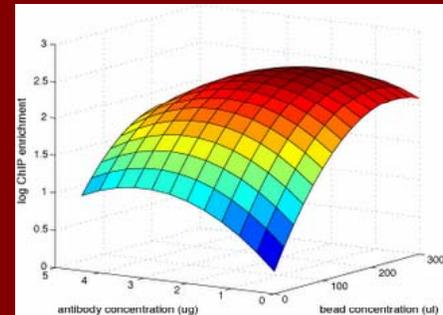
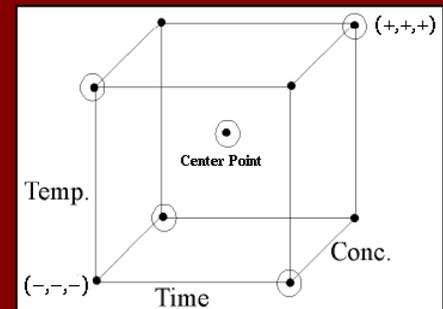
Study Design & Data Collection

How much?

accuracy **power** **cost**



Fixed vs. random effects



DoX

Three Current Problems

Predictive role underappreciated

Distinction blurred

Inappropriate modeling

“While the value of scientific prediction... is beyond question... the inexact sciences [do not] have...the use of predictive expertise well in hand.”

Helmer & Rescher, 1959



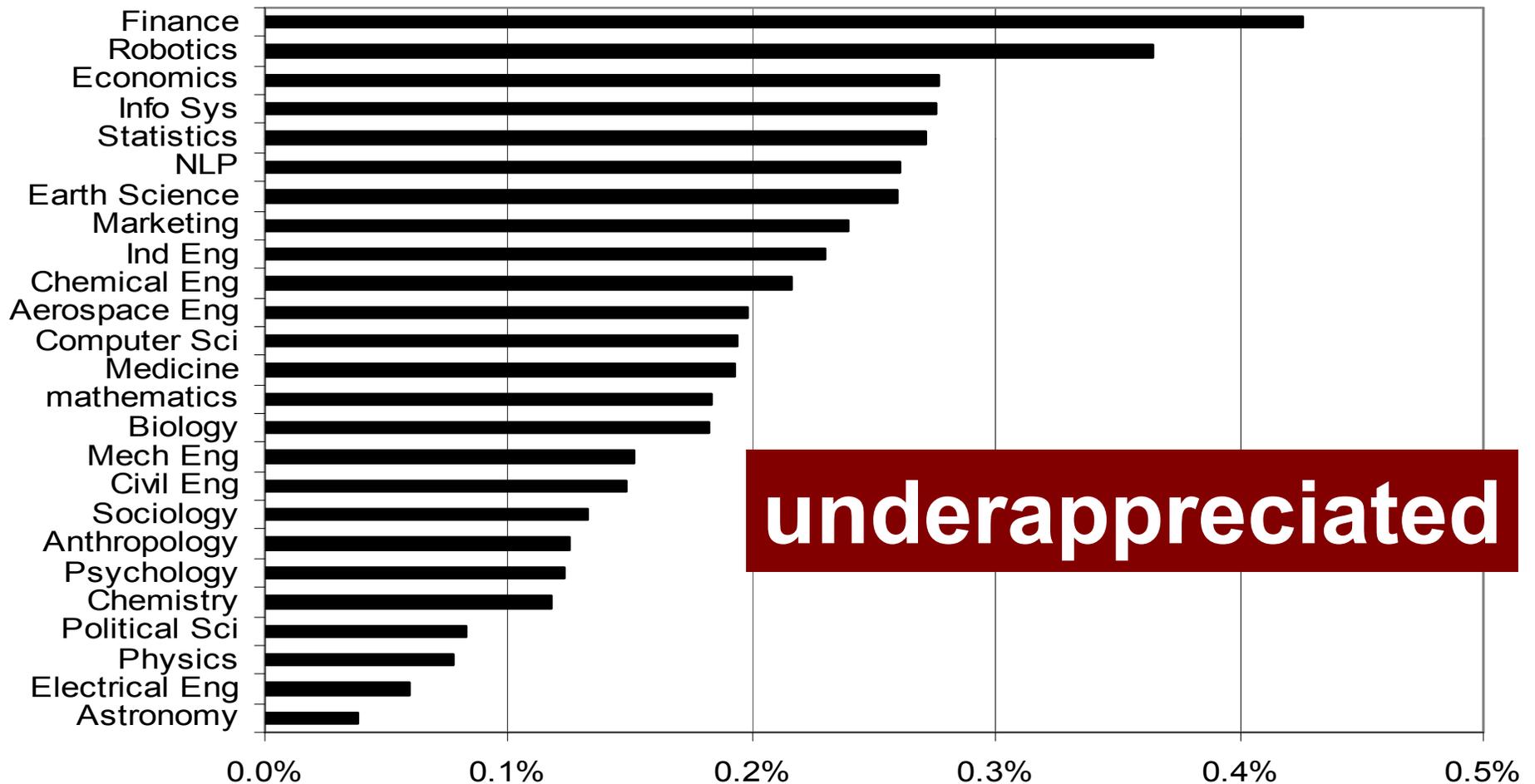
[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

[Advanced Scholar Search](#)
[Scholar Preferences](#)
[Scholar Help](#)

Scholar All articles - [Recent articles](#)

Results 1 - 10 of about 20,600 for "statistics" "predictive model".

% of "Predictive Model" hits on Google Scholar



EBSCOhost
Research Databases

[Basic Search](#) | [Advanced Search](#) | [Choose Database](#) | [Select another EBSCO service](#)

[Sign In to My EBSCOhost](#) | [Keyword Search](#) | [Publication Search](#) | [Subject Terms](#) | [Indexes](#)

Find: in
or in
and in
in [Info](#) (new v

Refine Search | [Search History / Alerts](#) | Results

Limit your results:

Full Text

References Available

Scholarly (Peer Reviewed) Journals

Published Date from Yr: to Yr:

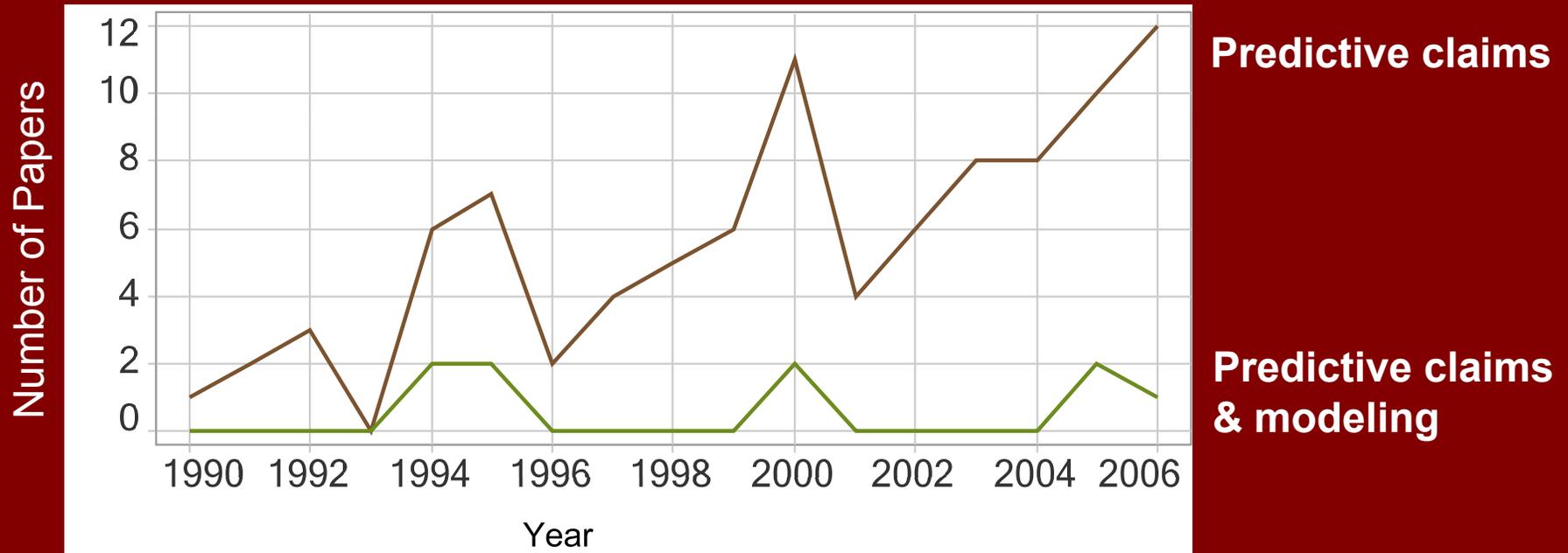
Publication



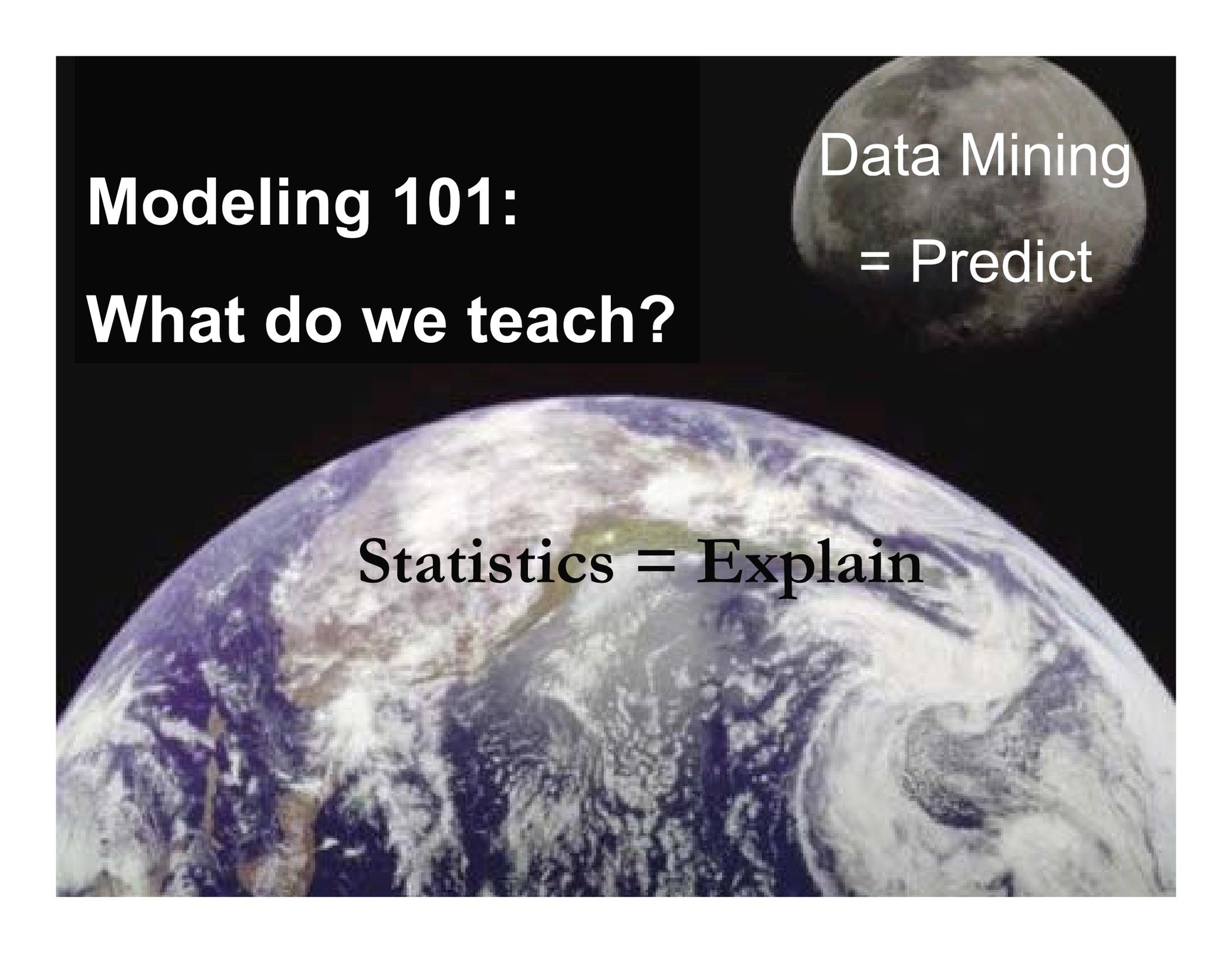
Predictive goal stated?
Predictive modeling followed?

“Examples of [predictive] theory in IS do not come readily to hand, suggesting that they are not common”

Gregor, MISQ 2006



Sound predictive modeling is rare



Modeling 101: What do we teach?

Data Mining
= Predict

Statistics = Explain

Bottom Line: Acknowledge!

Philosophy:
explaining and **predicting**
are different, necessary

Statistical Modeling

distinction blurred
↓
incorrect modeling

