To Explain
To Predict
or
To Describe?

Galit Shmueli 徐茉莉

Institute of Service Science

國立清華大學
NATIONAL TSING HUA UNIVERSITY

ISBIS 2019 Satellite Conference
August 15-16, 2019
Lanai Kijang, Kuala Lumpur, Malaysia

ISBIS: International Society for Business and Industrial Statistics
An Association of the International Statistical Institute

Today's Menu

1. **Definitions**
2. **Monopolies & confusion in academia & industry**
3. **Explanatory, predictive, descriptive modeling & evaluation are different**
   - Why?
   - Different **modeling paths**
   - Explanatory vs. predictive vs. descriptive **power**
4. **Where next?**

Definitions: Explain

Explanatory modeling
theory-based, statistical testing of causal hypotheses

Explanatory power
strength of relationship in statistical model

Definitions: Predict

Predictive modeling
empirical method for predicting new observations

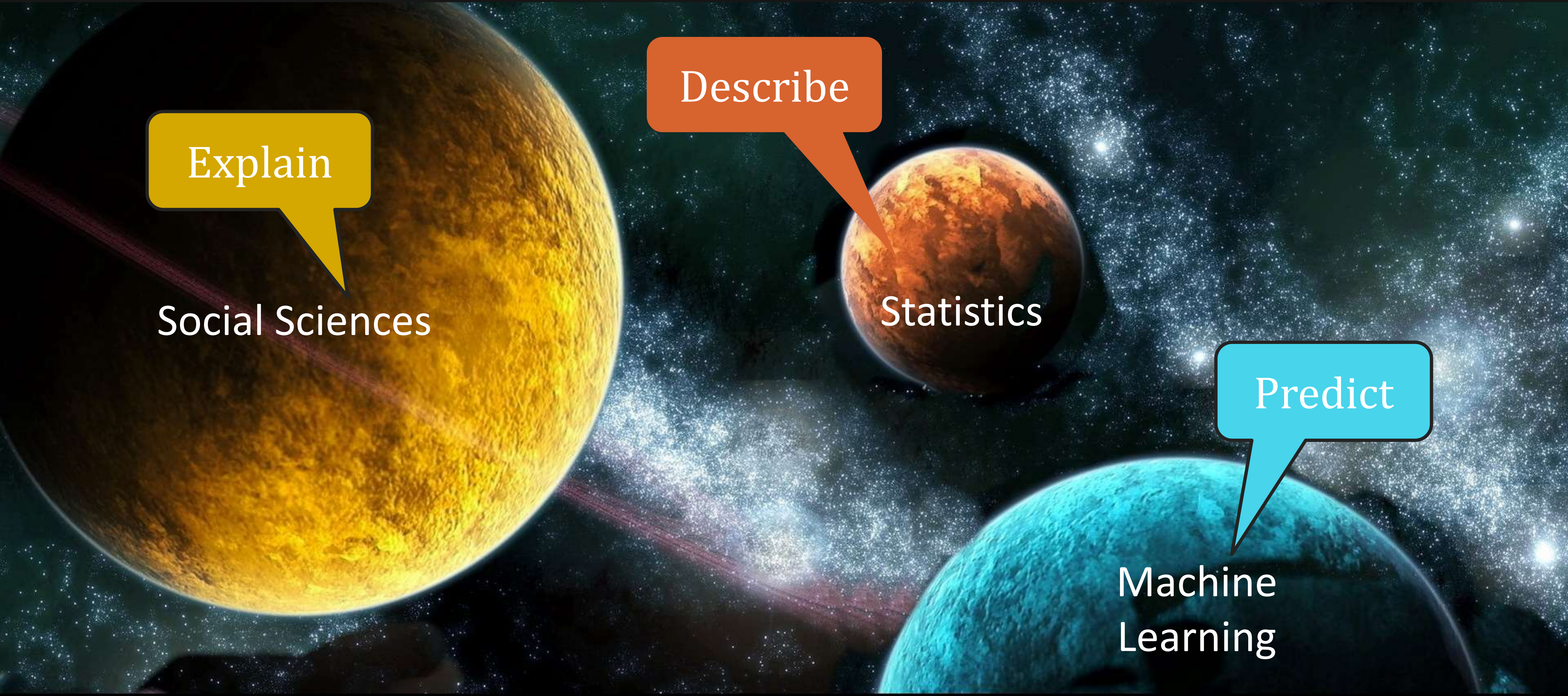
Predictive power
ability to accurately predict new observations

Definitions: Describe

Descriptive modeling
statistical model for approximating a distribution or relationship

Descriptive power
goodness of fit, generalizable to population

Monopolies in Different Fields



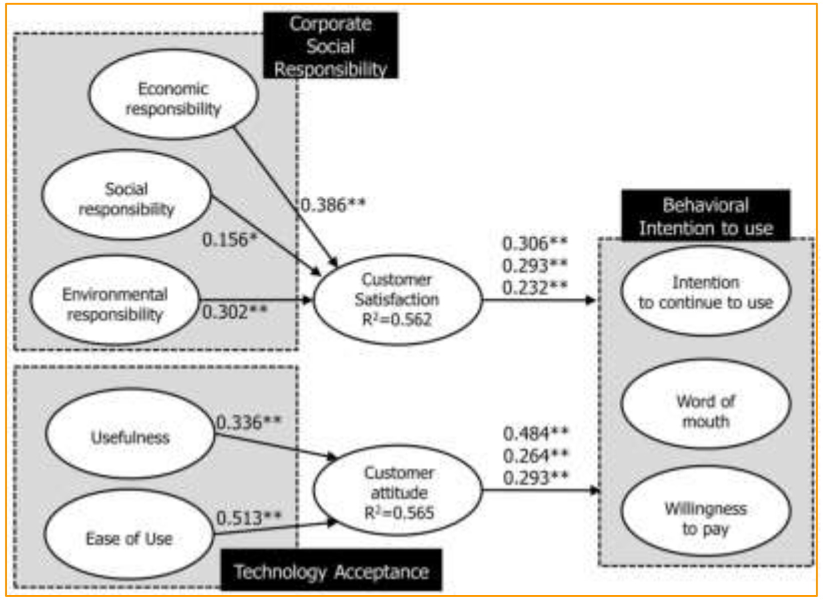
Social sciences & management research
Domination of ”Explain”

Purpose: test causal theory (“explain”)

Association-based statistical models

Prediction & description nearly absent

Classic journal paper



Start with a causal theory

Generate causal hypotheses on constructs

Operationalize constructs → measurable variables

Fit statistical model

Statistical inference → causal conclusions

Misconception #1:
The same model is best for explaining, describing, predicting

Social Sci & Mgmt: Build explanatory model and use it to ”predict”

“A good explanatory model will also predict well”

“You must understand the underlying causes in order to predict”

JOURNAL ARTICLE

Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior

Paul A. Pavlou and Mendel Fygenson

MIS Quarterly

Vol. 30, No. 1 (Mar., 2006), pp. 115-143

“To examine the **predictive** power of the proposed model, we compare it to four models in terms of **R^2 adjusted**”

HEALTH PSYCHOLOGY REVIEW

Taylor & Francis

Health Psychol Rev. 2016 Apr 2; 10(2): 148–167. PMCID

Published online 2014 Sep 17. doi: 10.1080/17437199.2014.947547

How well does the theory of planned behaviour predict alcohol consumption? A systematic review and meta-analysis

Richard Cooke,[a,*] Mary Dahdah,[a] Paul Norman,[b] and David P. French[c]

Journal of Applied Social Psychology

Explore this journal >

Predicting and Explaining Intentions and Behavior: How Well Are We Doing?

Stephen Sutton

First published: August 1998 Full publication history

DOI: 10.1111/j.1559-1816.1998.tb01679.x View/save citation

Cited by (CrossRef): 433 articles Check for updates | Citation tools ▾

View issue TOC

Volume 28, Issue 15

August 1998

Pages 1317–1338

Misconception #1:
The same model is best for explaining, describing, predicting

CS/eng/stat: Build a predictive model and use it to ”explain”

User Exercise Pattern Prediction through Mobile Sensing

Georgi Kotsev, Le T. Nguyen, Ming Zeng, and Joy Zhang
Carnegie Mellon University Silicon Valley
Moffet Field, California, USA
{georgi.kotsev, le.nguyen, ming.zeng, joy.zhang}@sv.cmu.edu

2014 6th International Conference on Mobile Computing, Applications and Services

Using Functio…

Shanshan Wang, Wolfgang Jank

Pages 144-160 | Published online: 01 Ja…

turnover by identifying empl… risk for leaving. These applic… including an individual's sala… results of their most recent p… amount of vacation time they… length of their commute. Fro… analytics programs generate… their likelihood of leaving dur… highlight the top factors influencing employees' interest in leaving.

On a Framework for the Prediction and Explanation of Changing Opinions

Eunice E. Santos*, Eugene Santos Jr.†, John T. Wilkinson†, Huadong Xia*
*Department of Computer Science
Virginia Polytechnic Institute and State University, Blacksburg, VA 24060
Email: santos@cs.vt.edu, xhd@vt.edu
†Thayer School of Engineering
Dartmouth College, Hanover, NH 03755
Email: {Eugene.Santos.Jr, John.T.Wilkinson}@dartmouth.edu

(Agent-based modeling using census data)
“our model is able to provide both **predictions** of how the population may vote and **why** they are voting this way”…

2009 IEEE International Conference on Systems, Man and Cybernetics

In this work, we present insights about user exercise pat-… pants. … user's … er the … Our …

- **Insights about users' exercise patterns**: We intro-… users. … actors …
- **Prediction**: We propose a … approach to predict the tendency of users' future number of exercises per week and compare the performance of different predictors and classifiers.

Misconception #2:
explain > predict or predict > explain

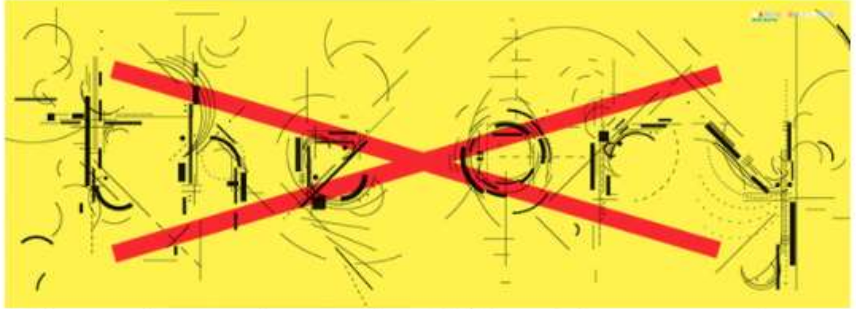
Emanuel Parzen, Comment on "Statistical Modeling: The Two Cultures"
***Statistical Science* 2001**

The two goals in analyzing data which Leo calls prediction and information I prefer to describe as "management" and "science." Management seeks *profit*, practical answers (predictions) useful for decision making in the short run. Science seeks *truth*, fundamental knowledge about nature which provides understanding and control in the long run.

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

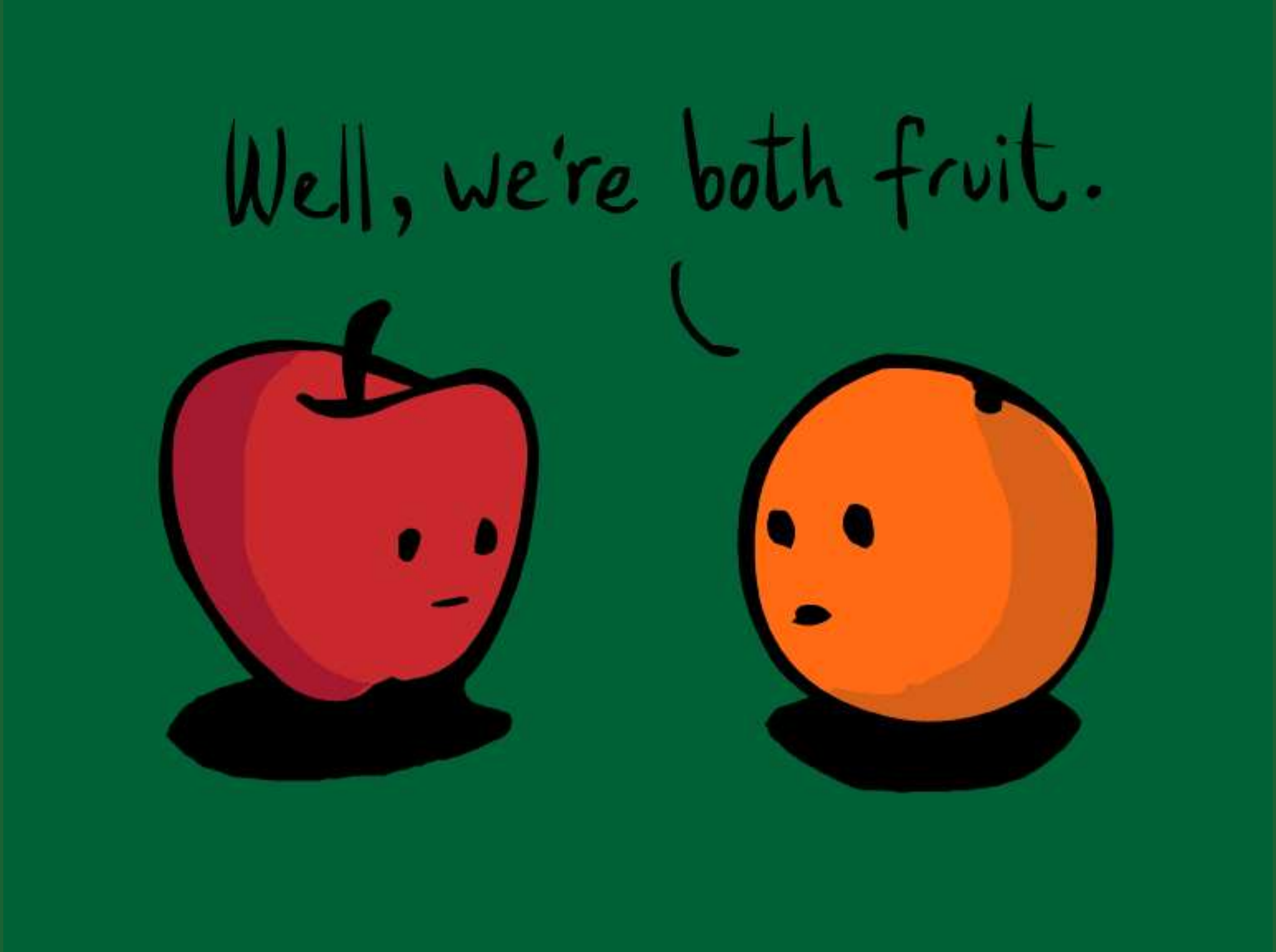
THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

*Chris Anderson is the editor in chief of Wired



* Illustration: Marian Bantjes * "**All models are wrong**, but some are useful."

"Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all"


Well, we're both fruit.

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

Why statistical

explanatory modeling

predictive modeling

descriptive modeling

are different

Different Scientific Goals

Different ***generalization***

Explanatory Model:

test/quantify causal effect between *constructs* for "average" unit in population

Descriptive Model:

test/quantify distribution or correlation structure for *measured* "average" unit in population

Predictive Model:

predict *values* for new/future individual units

Theory vs. its manifestation

Notation

Theoretical constructs: $\mathbf{X}$, $\mathbf{Y}$

Causal theoretical model: $\mathbf{Y=F(X)}$

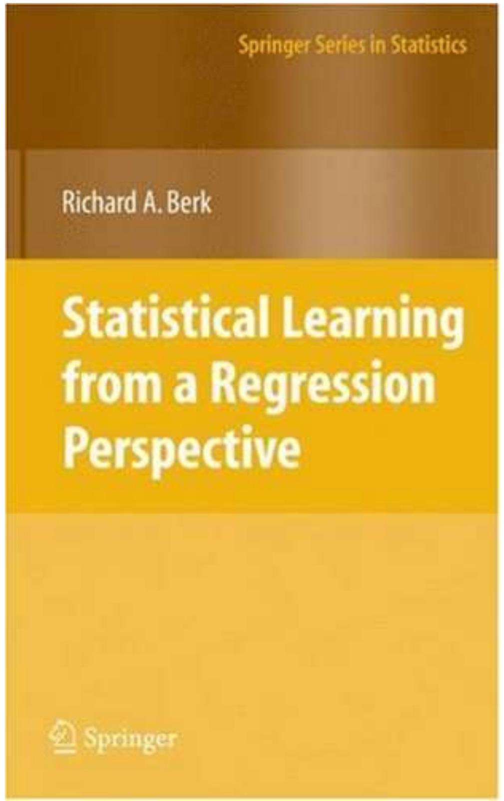
Measurable variables: *X*, *Y*

Statistical model: $E(y)=f(X)$

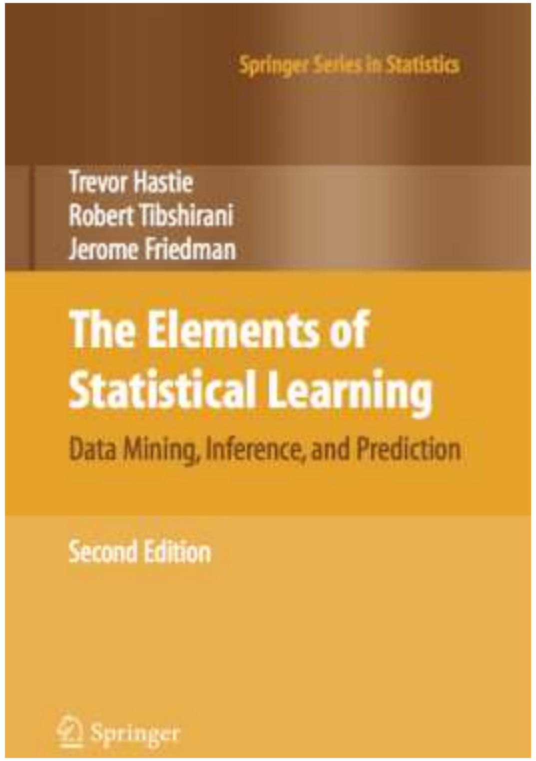
Breiman, "Statistical Modeling: The Two Cultures", *Stat Science*, 2001

5 aspects to consider

Theory	–	Data
Causation	–	Association
Retrospective	–	Prospective
Bias	–	Variance
Average Unit	–	Individual Unit



"The goal of finding models that are predictively accurate differs from the goal of finding models that are true."



$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \mathrm{Irreducible\ Error} + \mathrm{Bias}^2 + \mathrm{Variance}.
\end{aligned}$$

But there’s **more** than bias-variance

Example: Regression Model for Explanation

Underlying model: X $\rightarrow$ Y

Measures of
X, Y constructs

Danger:
endogeneity

$$y_i | x_i = \beta_0 + \beta_1 x_i + \beta_2\, x_{controls} + \varepsilon_i$$

parameter
of interest
(inference)

Chosen to avoid Omitted Var
Bias and backdoor/spurious
paths

Example: Regression Model for Description

All variables treated/interpreted as observable

Remain in model only if statistically significant

$$y_i | x_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

parameters of interest (inference)

Chosen b/c related to Y
Danger: multicollinearity

Residual analysis for GoF & test assumptions

Example: Regression Model for Prediction

All variables treated as observable,
available at time of prediction

Retain only if improve out-of-sample prediction

$$y_i | x_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

Quantity of interest for new i's (prediction)

Chosen b/c possibly correlated with Y
Danger: over-fitting

Evaluate overfitting (train vs holdout)

Point #1

best
explanatory
model

best
predictive
model

≠

best
descriptive
model

Predict ≠ Explain

*"we tried to benefit from an extensive set of attributes describing each of the movies in the dataset. Those attributes certainly carry a significant signal and can **explain some of the user behavior**. However… they could not help **at all** for improving the [predictive] accuracy."*

Bell et al., 2008

Predict ≠ Describe

Election Polls

"There is a subtle, but important, difference between reflecting current public sentiment and predicting the results of an election. Surveys have focused largely on the former… [as opposed to] survey based prediction models [that are] focused entirely on analysis and projection"

Kenett, Pfefferman & Steinberg (2017) "Election Polls – A Survey, A Critique, and Proposals", *Annual Rev of Stat & its Applications*

Goal Definition

Design & Collection

Data Preparation

EDA

Variables? Methods?

Evaluation, Validation & Model Selection

Model Use & Reporting

Study design
& data collection

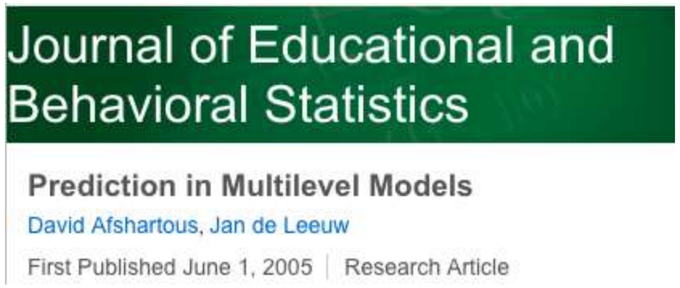
Observational or experiment?

Primary or secondary data?

Instrument (reliability+validity vs. measurement accuracy)

How much data?

How to sample?



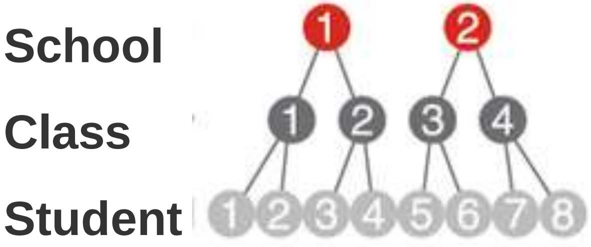
Journal of Educational and Behavioral Statistics

Prediction in Multilevel Models

David Afshartous, Jan de Leeuw

First Published June 1, 2005 | Research Article

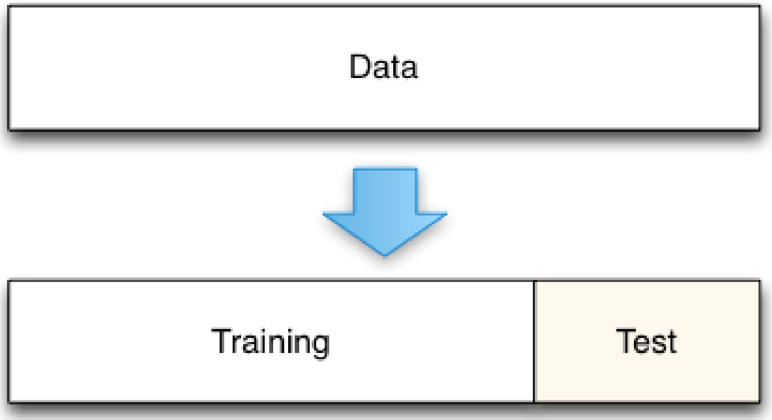
Multilevel (nested) data



predict: increase group size

explain/describe: increase #groups

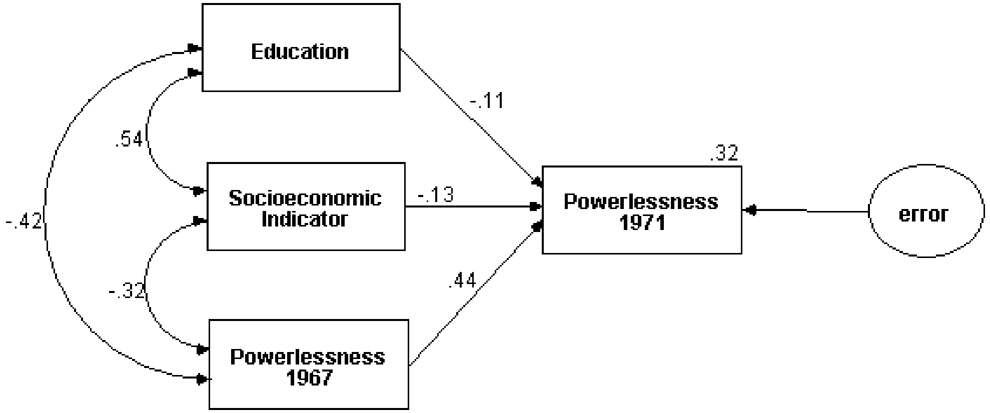
Data preprocessing

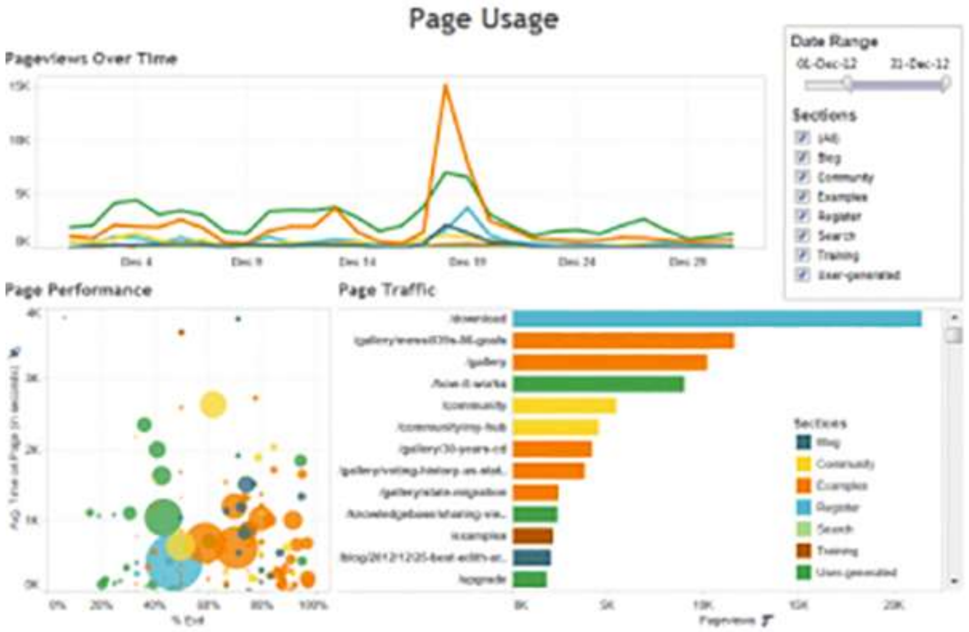


Reduced-Feature Models

Saar-Tsechansky & Provost, JMLR 2007

Data exploration, viz, reduction





Factor Analysis
(interpretable)

PCA

Dimension Reduction
(fast, small)

Which variables?

causal role vs. associations

endogeneity

ex-post

availability

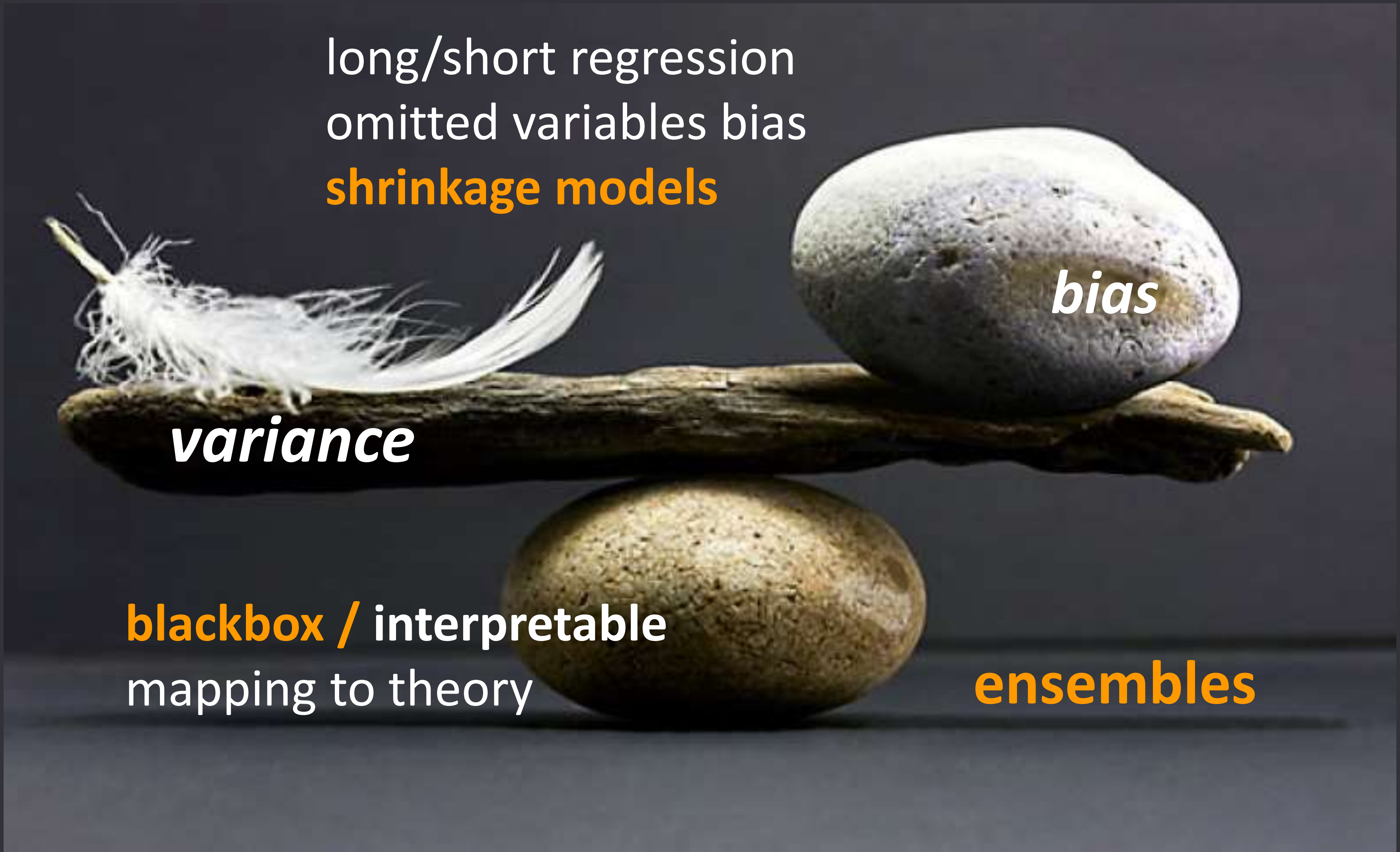
leading,
coincident,
lagging indicators

multicollinearity

identifiability

A, B, A*B

Methods / Models



Validation

Model fit ≠ Explanatory power

theoretical model ⟷ statistical model ⟷ Data

Evaluation, Validation & Model Selection

statistical model ⟷ training data

statistical model → holdout data

Predictive power

training data, holdout data } Over-fitting analysis

Point #2

explanatory power

predictive power ≠ descriptive power

Cannot infer one from the others

interpretation

out-of-sample

p-values

overall, specific

prediction accuracy

Performance Metrics

R^2

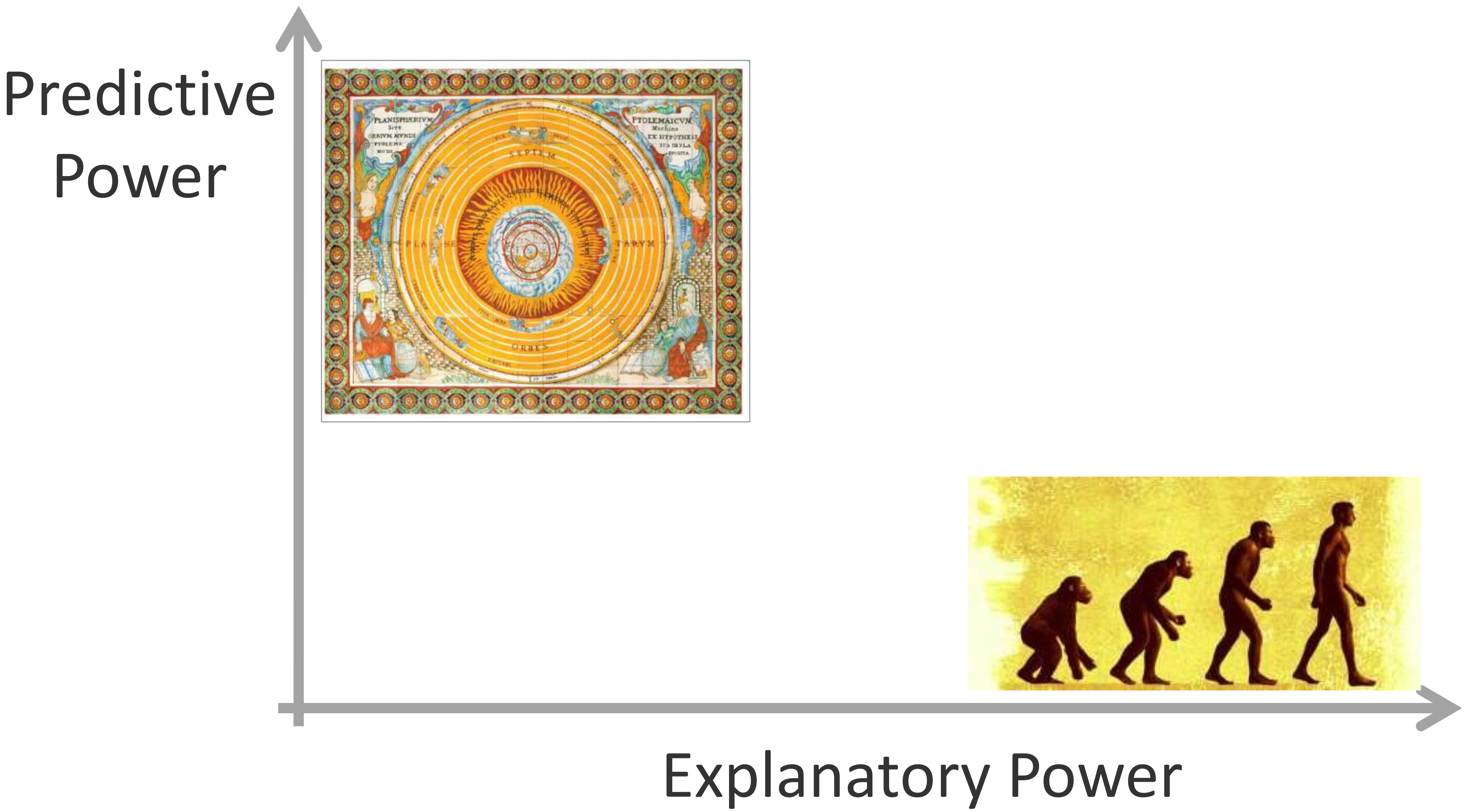
costs

goodness-of-fit

training vs holdout

type I,II errors

over-fitting


Predictive Power
Explanatory Power

Convinced?


Point #1
best explanatory model
best predictive model
best descriptive model


Point #2
explanatory power
predictive power
descriptive power
Cannot infer one from the others

WHERE NEXT

Currently in Academia

(social sciences, management)

- Theory-based explanatory modeling
- Prediction underappreciated
- Distinction blurred
- Unfamiliar with predictive modeling – getting better

How/why use prediction

(predictive models + evaluation)

for scientific research

beyond project-specific

solution/utility/profit?

The predictive power of an explanatory/descriptive model has important scientific value

relevance, reality check, predictability

Prediction for Scientific Research

- Generate new theory
- Develop measures
- Compare theories
- Improve theory
- Assess relevance
- Evaluate predictability

Shmueli & Koppius, “Predictive Analytics in Information Systems Research”
MIS Quarterly, 2011

Currently in Industry

(and machine learning)

- Data-driven predictive modeling
- Prediction over-appreciated
- Distinction blurred
- A-B testing
- Unfamiliar with theory-based
  explanatory modeling

Will the customer pay?

What **causes** non-payment?

Implications:

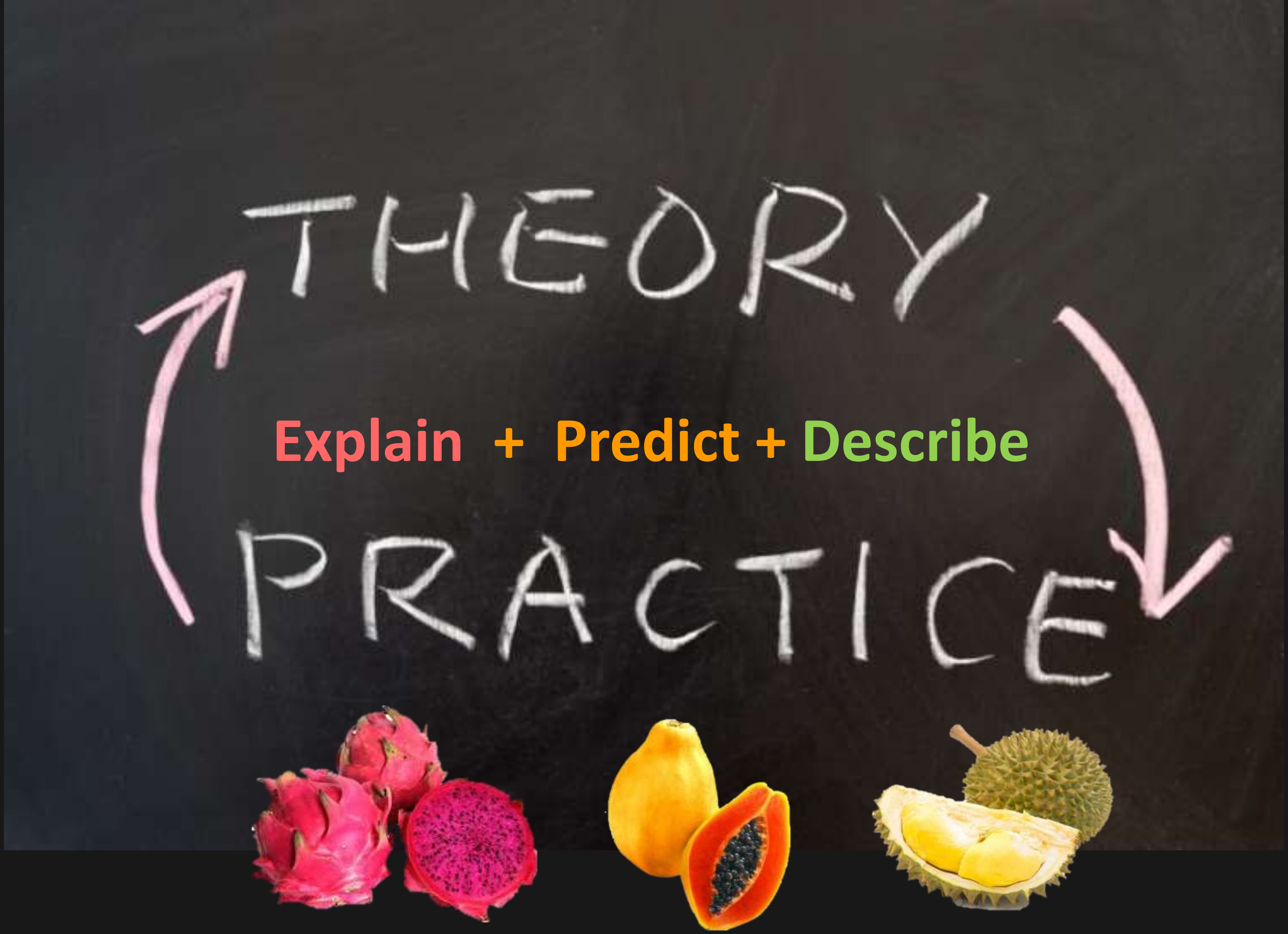
Short-term solutions

Shallow/no understanding

Ethical, social, human pitfalls

How to do theory-based explanatory modeling with Behavioral Big Data?

Shmueli (2017) "Research Dilemmas With Behavioral Big Data", *Big Data*, vol 5(2), pp. 98-119


THEORY
Explain + Predict + Describe
PRACTICE