PSYCH 201B

Statistical Intuitions for Social Scientists

What is statistics?

01/12/2025

Last time… course goals:

(a) Developing statistical intuitions

"**Statistical thinking** will one day be as necessary for efficient citizenship as the ability to read and write."

~ Samuel S. Wilks, 1951

of χ^2 *and the log-likelihood ratio!*

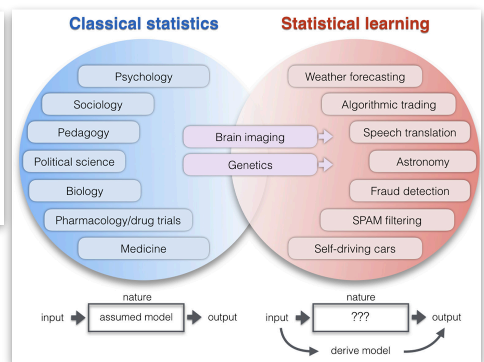
(b) Bridging cultural differences

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



(c) Programming as theory-building

Today’s Goals

(a) Developing statistical intuitions

“**Statistical thinking** will one day be as necessary for efficient citizenship as the ability to read and write.”

~ Samuel S. Wilks, 1951

of χ^2 and the log-likelihood ratio!

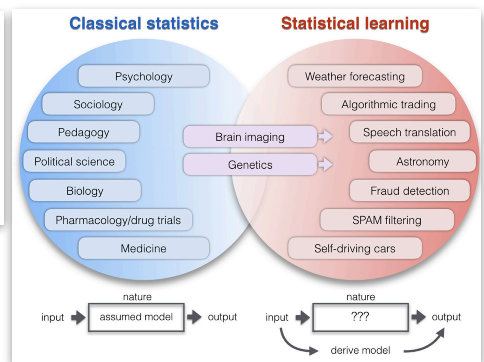
(b) Bridging cultural differences

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



(c) Programming as theory-building

Today's Goals

(a) Developing statistical intuitions

"**Statistical thinking** will one day be as necessary for efficient citizenship as the ability to read and write."

~ Samuel S. Wilks, 1951

of χ^2 and the log-likelihood ratio!

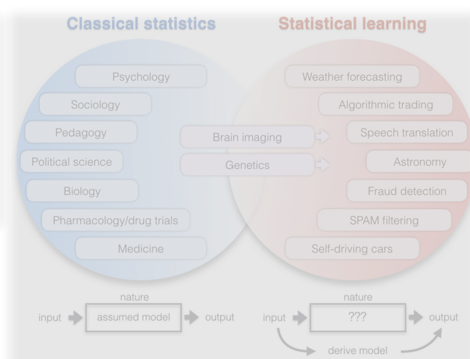
(b) Bridging cultural differences

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



(c) Programming as theory-building

What is statistical *thinking*?

(computational)

What is statistical *thinking*?

(computational)

statsthinking21

Main web site for Statistical Thinking for the 21st Century

View the Project on GitHub
statsthinking21/statsthinking21

Statistical Thinking for the 21st Century

An open source textbook for statistics, with companions for R and Python

Russell Poldrack

Stanford University

A commercially published version of this book (with an expanded version of Chapter 17) is now available from Princeton University Press: Statistical Thinking: Analyzing Data in an Uncertain World

- One of primary reference materials
- Full PDFs always available on course website
- Readings are *optional* – but please *skim* them if you can – then we can have a fruitful conversation in class!

What is statistical *thinking*?

(computational)

- A way of understanding the complexities of the world
- Describe them simpler terms that capture essential structural or functional aspects
- Provide with us some idea of our knowledge uncertainty
- Distinct from other forms of knowledge & thinking: intuition, heuristics, etc

Statistical thinking allows us to:

- Describe - simplify enough to understand & communicate
- Decide - choose actions, behaviors, policies based on data, particularly when we're *uncertain*
- Predict - make principled guesses about new situations given our knowledge of previous ones

Ok but what even *is* statistics?!

Ok but what even *is* statistics?!

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081

Taylor & Francis
Taylor & Francis Group

Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

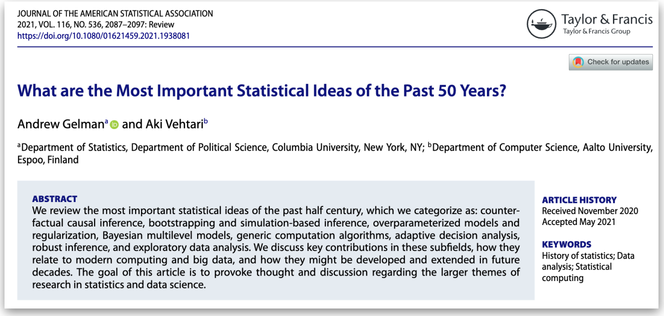
ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provide thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

The science of principled abstraction from data – with explicit acknowledgment of uncertainty and limitations

What Statistics **is** about


JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081

Taylor & Francis
Taylor & Francis Group

Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

Constructing **models** of the world

Models are *deliberate* simplifications

Every model is wrong – the question is whether the model is *useful*… did we *learn* something?

What Statistics is about


JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081
Taylor & Francis
Taylor & Francis Group
Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

Learning by confronting data

Data are not answers – they are checks on our *assumptions*

Learning happens when *predictions clash with observations*

What Statistics is about


JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081

Taylor & Francis
Taylor & Francis Group

Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

Quantifying uncertainty honestly

Precision without humility is *misleading*

Uncertainty reflects both *randomness* and *mistaken assumptions*

What Statistics **is** about


JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081

Taylor & Francis
Taylor & Francis Group

Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

Iterating not concluding

Statistics is a *process* – not a *verdict*

Modeling is a *workflow* – model, check, revise, recheck, etc

What Statistics is not

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 2087–2097: Review
https://doi.org/10.1080/01621459.2021.1938081

Taylor & Francis
Taylor & Francis Group

Check for updates

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT
We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

A way to certify *truth*

Just hypothesis testing

Machinery to justify decisions (“significant/not”)

Separate from your assumptions

Statistical thinking $\neq$ Statistical inference

- Inference is a *narrow step* within the broader framework of statistical thinking
  - *How uncertain am about this parameter estimate?*
  - *Under the null, how surprising is this estimate?*

- Statistical thinking is a mode of *reasoning*
  - *Is this model a sensible abstraction of the phenomenon?*
  - *What did we throw-away when we summarized this?*
  - *What assumptions are silently doing the work?*
  - *What would failure look like?*

So how do we *do* this?

The four fundamental intuitions

1) Aggregation (compression)

- **What?**
  - The science of throwing away information — on purpose
- **Why?**
  - Too much complexity! Patterns become invisible and comparisons are impossible
  - Allows us to see structure and communicate about it
- **Misconceptions**
  - More aggregation = better data
  - The mean (simple average) has no assumptions
- **Consider**
  - What variation matters? What variation is getting ignored?

2) Sampling (generalization)

- **What?**
  - How we connect data to the world-at-large
- **Why?**
  - We never get to observe *everything* — all the people in the world. So we need to *generalize* from a dataset (sample) to some broader population
- **Misconceptions**
  - Random sampling is always required
  - Larger samples automatically generalize better
- **Consider**
  - Sampling assumptions define what claims you're allowed to make…*who* you sampled, under what *conditions*, using what *model*

3) Uncertainty (limits of inference)

- **What?**
  - Reflects what we do not know — not what went “wrong”
- **Why?**
  - Even with perfect measurements we are constrained by *finite samples* and *approximate* models. Therefore certainty is always *provisional.*
- **Misconceptions**
  - Using a model = correct/safe
  - $p < .05$ removes uncertainty
- **Consider**
  - Over-certainty often signals model *failure*…honest uncertainty quantifies *robustness*

4) Learning (modeling & updating)

- **What?**
  - Changing your mind in a principled way — iterative model comparison & refinement
- **Why?**
  - Data alone do not teach us anything. To learn we require: expectations, predictions, comparisons to reality.
- **Misconceptions**
  - Once a model is fit the work is done
  - Learning = finding the “right” model
- **Consider**
  - Prediction & explanation *constrain* each other — and our models evolve as we learn

The four fundamental intuitions

- **Aggregation**
  - Gives us a model
- **Sampling**
  - Tells us where it applies
- **Uncertainty**
  - Keeps us honest
- **Learning**
  - Forces iteration

Statistics is not about being right

It's the process of becoming less wrong over time

Today's Goals

(a) Developing statistical intuitions

"**Statistical thinking** will one day be as necessary for efficient citizenship as the ability to read and write."

~ Samuel S. Wilks, 1951

of χ^2 *and the log-likelihood ratio!*

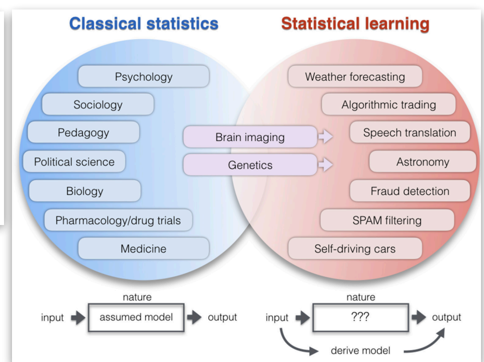
(b) Bridging cultural differences

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



(c) Programming as theory-building

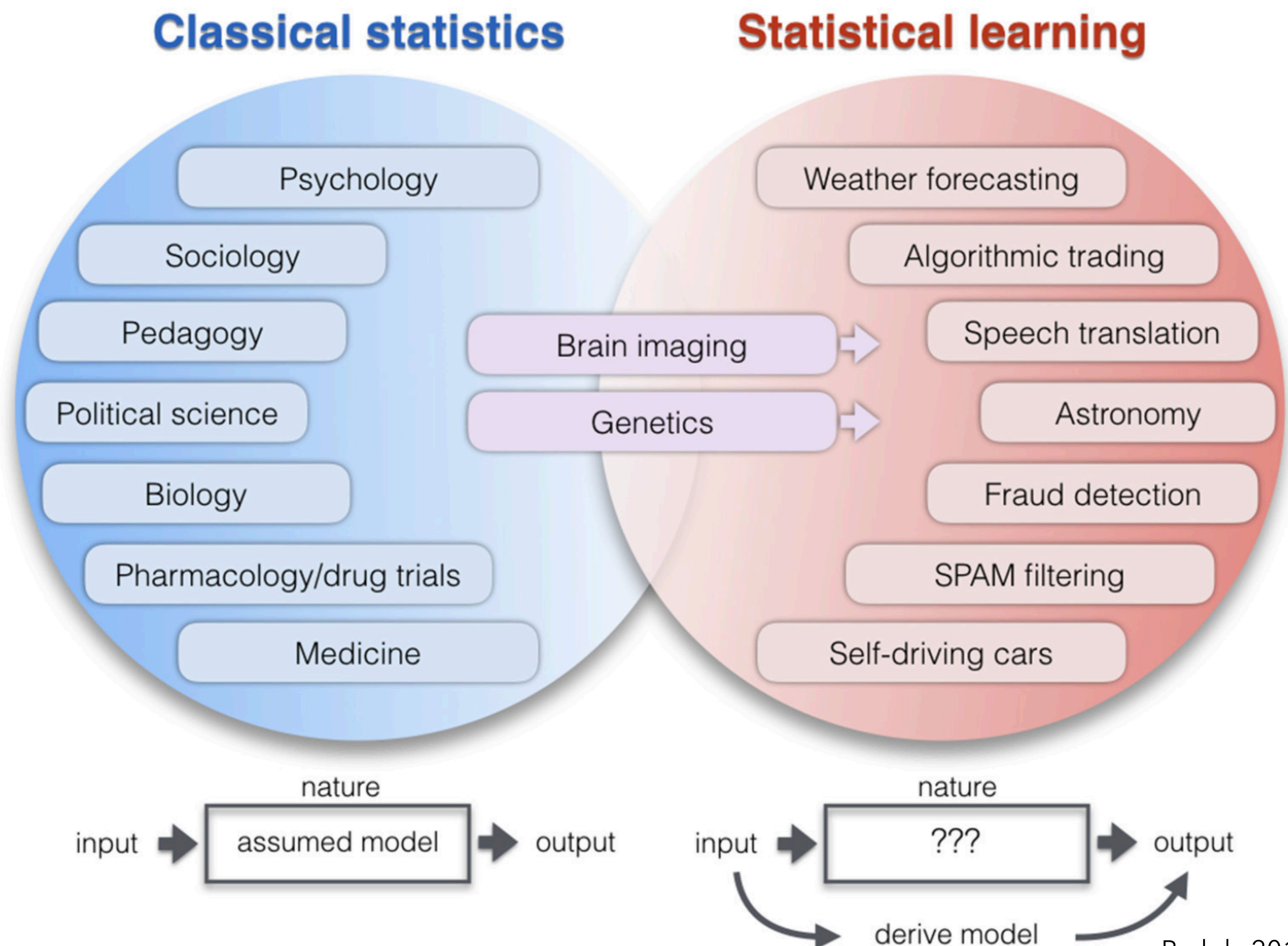
The two cultures of statistics

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

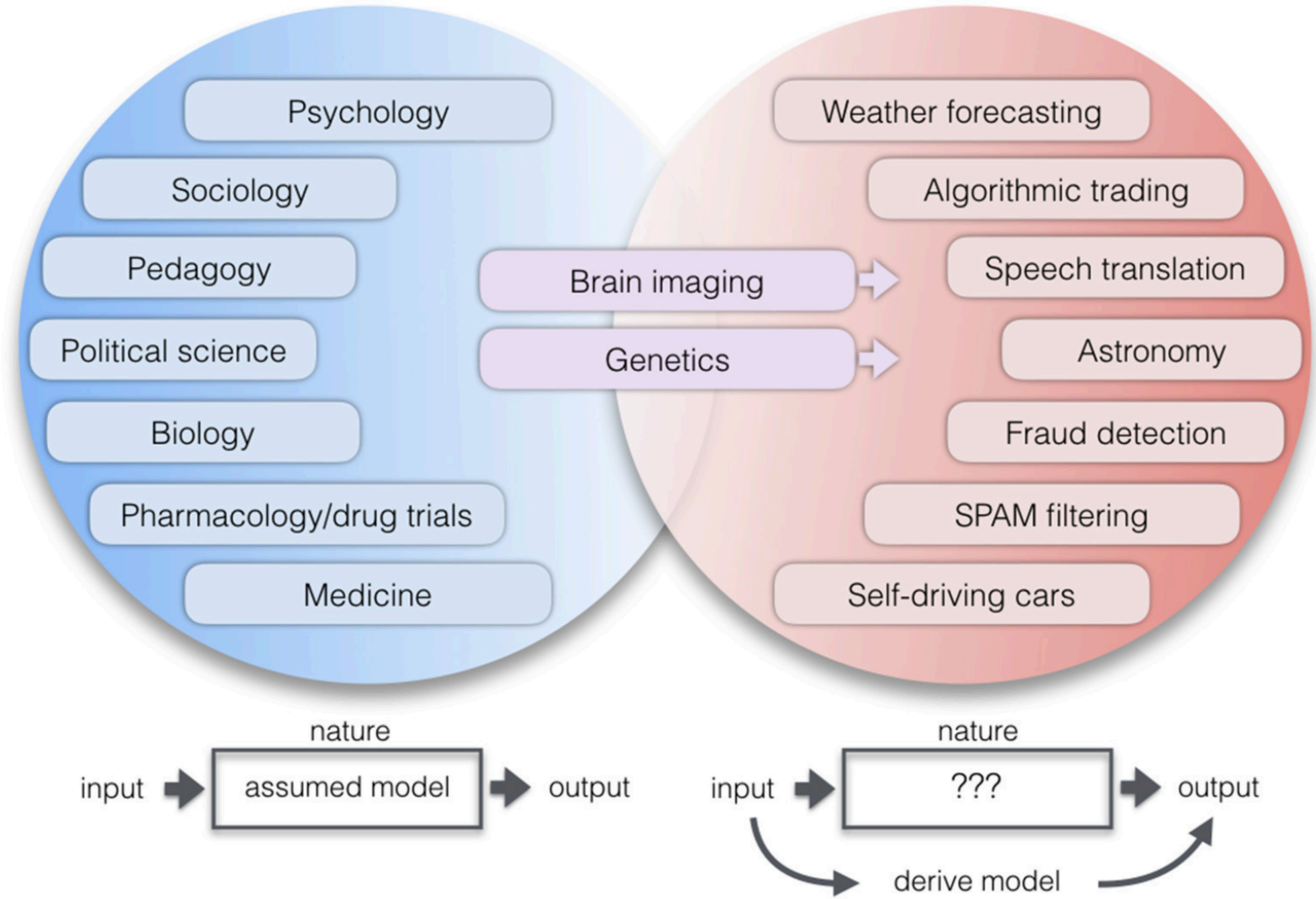
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.


Classical statistics
Statistical learning
Psychology
Sociology
Pedagogy
Political science
Biology
Pharmacology/drug trials
Medicine
Brain imaging
Genetics
Weather forecasting
Algorithmic trading
Speech translation
Astronomy
Fraud detection
SPAM filtering
Self-driving cars
nature
input
assumed model
output
nature
input
???
output
derive model

Bzdok, 2017

Explanation

Prediction



Explanation ⟷ Prediction

Classical inference

Theory		
null-hypothesis testing & multiple comparisons	bias-variance decomposition	Vapnik-Chervonenkis dimensions & curse of dimensionality
degrees of freedom df		hypothesis space $\mathcal{H}$
asymptotic consistency		finite-sample theorems
Invalidations of inferential process		
double dipping/circular analysis	post-selection inference	data snooping/peeking
Outcome metrics		
(in-sample) p values sensitivity/specificity effect size/power	explained variance metrics AUC/ROC curve confidence intervals	out-of-sample prediction accuracy/precision/recall/F1 scores learning curves certainty estimates via bootstrap
Representative methods		
Student's t-test F-test ANOVA Binomial test X^2-test linear regression	general(ized) linear model	support vector machines LASSO/ridge regression/elastic net logistic regression nearest neighbors random forests kernel methods („deep") neural networks

Out-of-sample generalization

Bzdok & Ionnadis, 2019

Explanation ↔ **Prediction**

Knowledge guided: pre-specified, simple but inflexible

Pattern guided: “late-commitment” , complex but flexible

Late commitment: Using theoretical assumptions to constrain analysis, not design

In the first step of the empirical cycle, we strive to minimize the theoretical assumptions built into the experimental design. This approach is motivated by the observation that designs, e.g., of fMRI experiments, can be made much more versatile (allowing us to address more neuroscientific questions) at moderate costs in terms of statistical efficiency (for addressing a given question). A general design that can address a 100 questions appears more useful than a restricted design that addresses a single question with slightly greater efficiency.

Statistical power is afforded by combining the evidence – usually by averaging. When we decide on a grouping of experimental events (e.g., for a block design), we commit to a particular way of combining the evidence and thus give up versatility. Ungrouped-events designs allow us to combine the evidence in many different ways *during analysis*. First, this approach allows for exploratory analyses, which can (1) test basic assumptions of a field, (2) usefully direct our attention to larger phenomena (in terms of explained variance), and (3) lead to unexpected discoveries. Second, ungrouped-events designs allow a broad set of theoretically constrained analyses to be performed on the same data. And third, as a consequence, such designs allow us to combine data across studies and research groups in order to address a particular question with a power otherwise unattainable. In the Appendix, we assess this third point, the poten-

Explanation ⟷ **Prediction**

Knowledge guided: pre-specified, simple but inflexible

Pattern guided: “late-commitment” , complex but flexible

Explainable narrative centered on inferences about single *parameters* “beta-hat” problems

Opaque “black box” centered on inferences about out-of-sample *prediction* “y-hat” problems

IV
(independent variable)
(manipulated)

$$\hat{Y} = a + bX + \varepsilon$$

Predicted DV
(dependent variable)
(measured)

Estimated parameter
(coefficient)
(relationship between Y~X)

IV
(independent variable)
(manipulated)

$$\hat{Y} = a + bX + \varepsilon$$

Prediction

Explanation

Explanation ⟷ Prediction

Explanation

Knowledge guided: pre-specified, simple but inflexible

Explainable narrative centered on inferences about single *parameters* "beta-hat" problems

Tools: p-values, ANOVA, t-test, Bayes posterior

Formally justified: rigorous mathematical and theoretical guarantees, well understood analytic routines and assumptions

Prediction

Pattern guided: "late-commitment" , complex but flexible

Opaque "black box" centered on inferences about out-of-sample *prediction* "y-hat" problems

Tools: cross-validation, regularization, model comparison

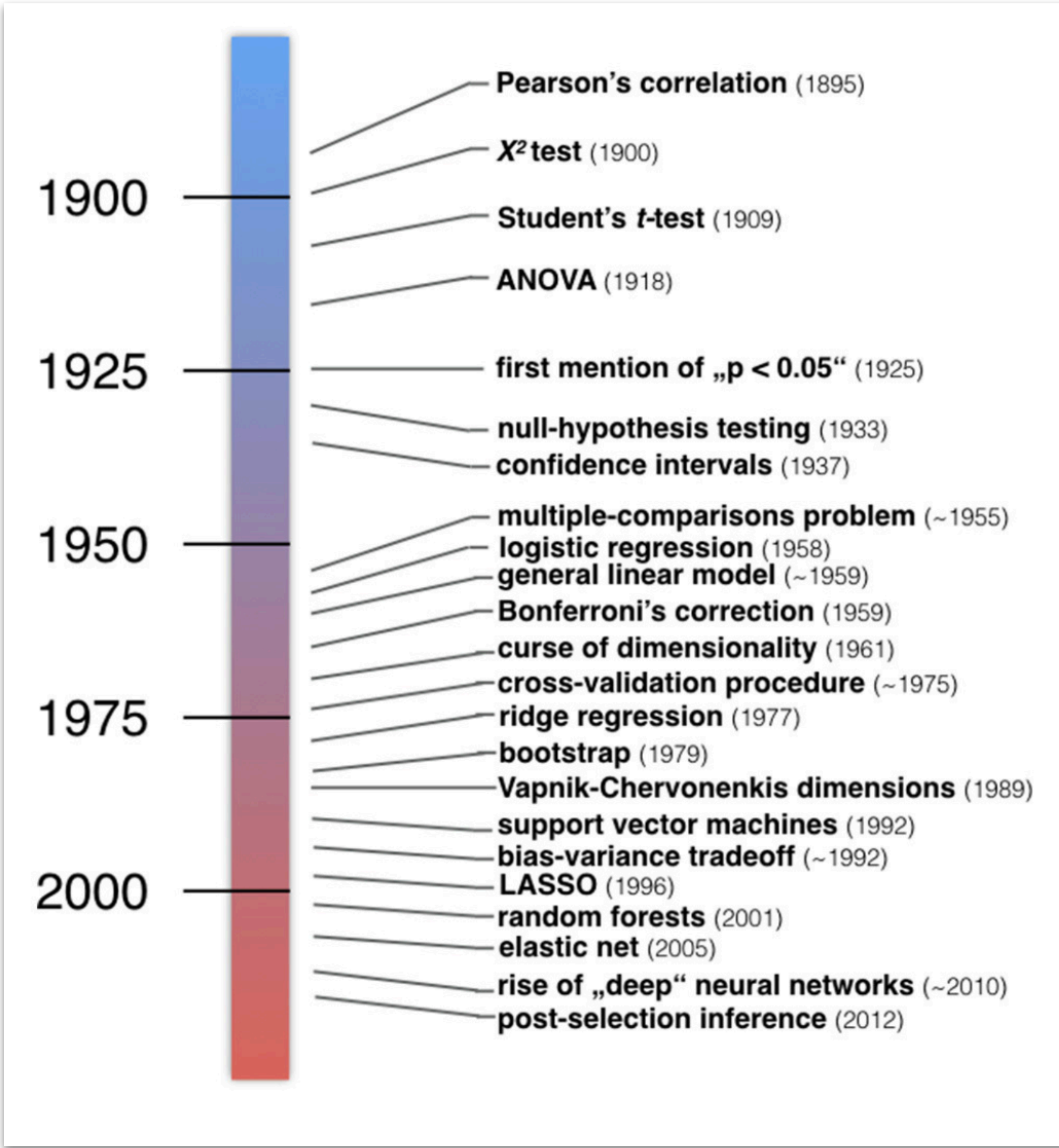
Empirically justified: informally validated via resampling and application to unseen data

Explanation ⟷ **Prediction**

Explanation	#	Prediction
„Truth“ is in the model	1	„Truth“ is in the data
„Long data“ (n samples > p variables)	2	„Wide data“ (n samples << p variables)
Make a set of judicious assumptions	3	Make least assumptions possible
Test a model for the data deductively	4	Learn a model from the data inductively
Impose mathematical rigor	5	Let the data speak for themselves
Choose a model before visiting the data	6	Choose a model as the data are visited
More confirmatory than exploratory	7	More exploratory than confirmatory
Model assumptions are largely explicit	8	Model assumptions are largely implicit
Data are often experimental	9	Data are often observational
Tractable models with few parameters	10	Expressive models with many parameters

Bzdok & Ionnadis, 2019

Why should you care?



← But tend to hang out here in practice

← We learn up to here

← They want/will make us irrelevant

Bzdok & Ionnadis, 2019

How can we do better?

Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning

Perspectives on Psychological Science
1–23

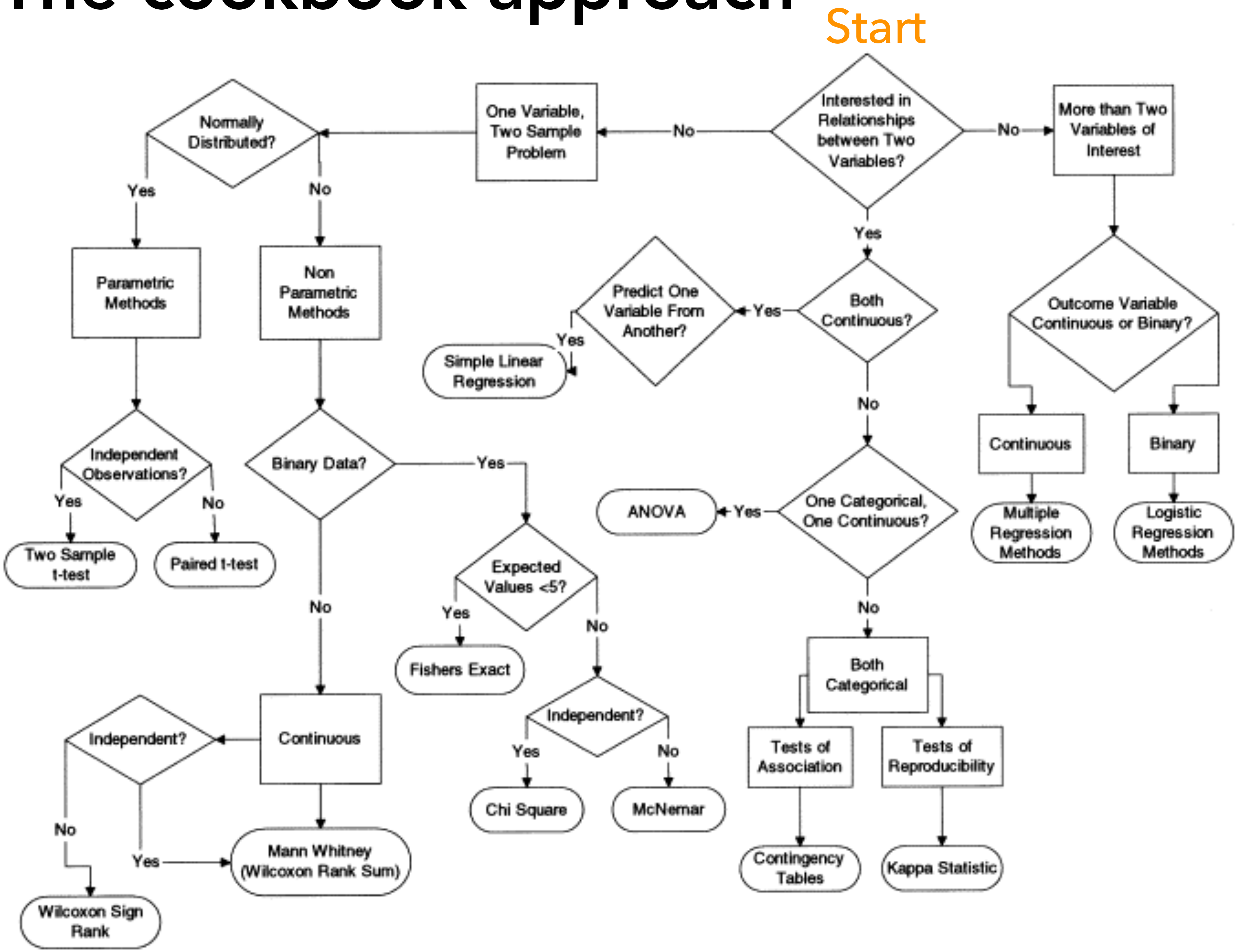
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691617693393
www.psychologicalscience.org/PPS
SAGE

Tal Yarkoni and Jacob Westfall
University of Texas at Austin

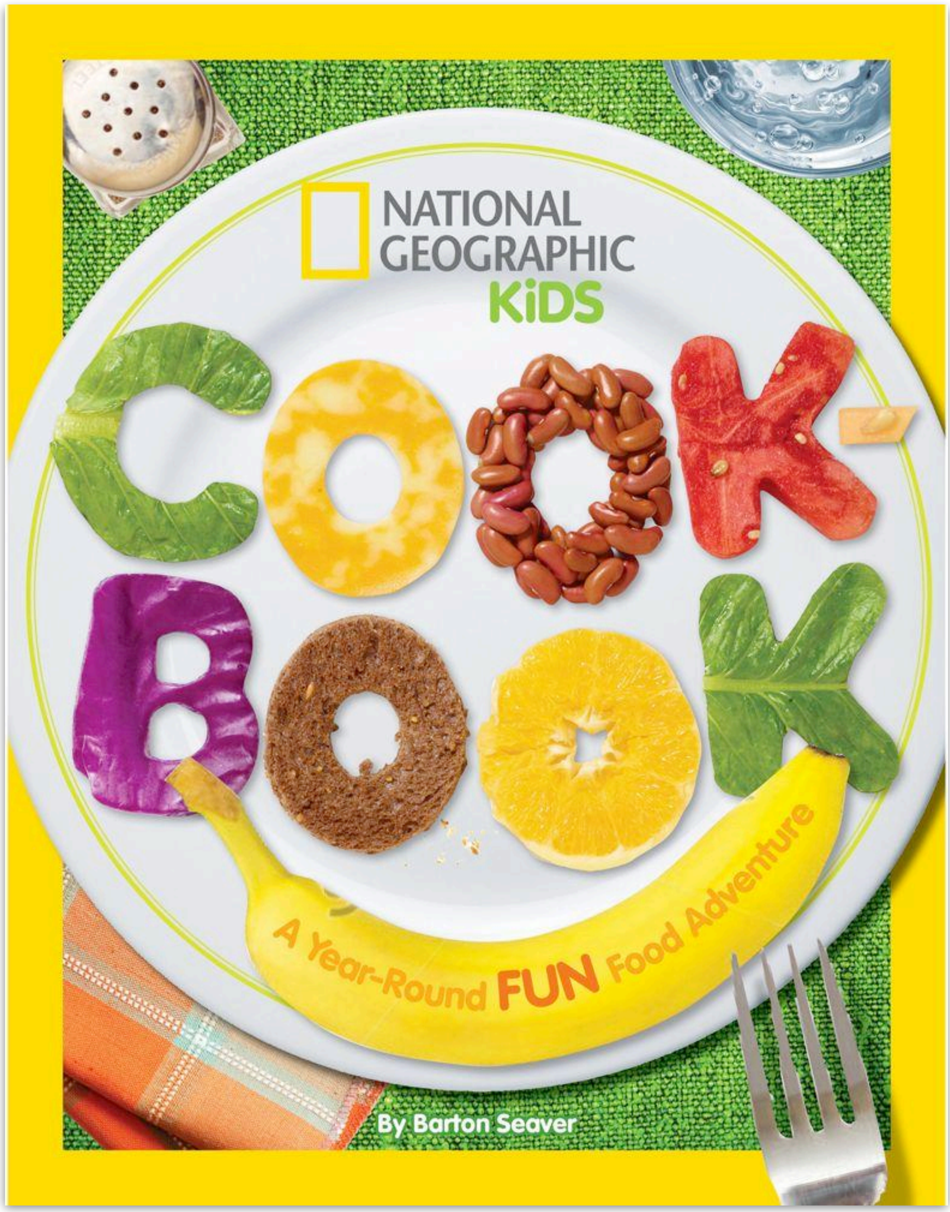
Abstract
Psychology has historically been concerned, first and foremost, with explaining the causal mechanisms that give rise to behavior. Randomized, tightly controlled experiments are enshrined as the gold standard of psychological research, and there are endless investigations of the various mediating and moderating variables that govern various behaviors. We argue that psychology's near-total focus on explaining the causes of behavior has led much of the field to be populated by research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors with any appreciable accuracy. We propose that principles and techniques from the field of machine learning can help psychology become a more predictive science. We review some of the fundamental concepts and tools of machine learning and point out examples where these concepts have been used to conduct interesting and important psychological research that focuses on predictive research questions. We suggest that an increased focus on prediction, rather than explanation, can ultimately lead us to greater understanding of behavior.

Cookbook Approach vs. Model Comparison

The cookbook approach

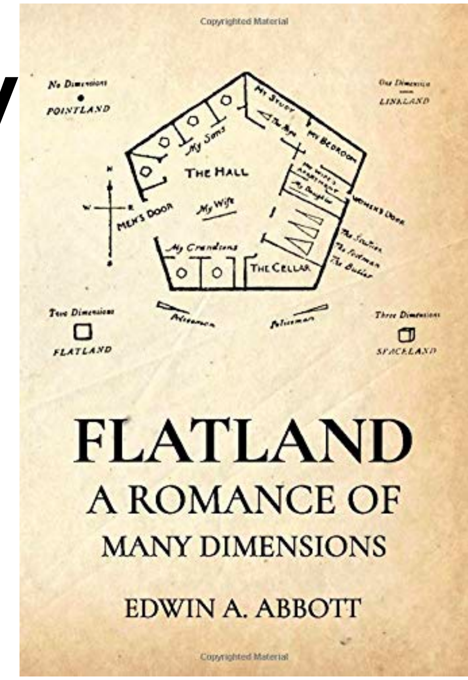


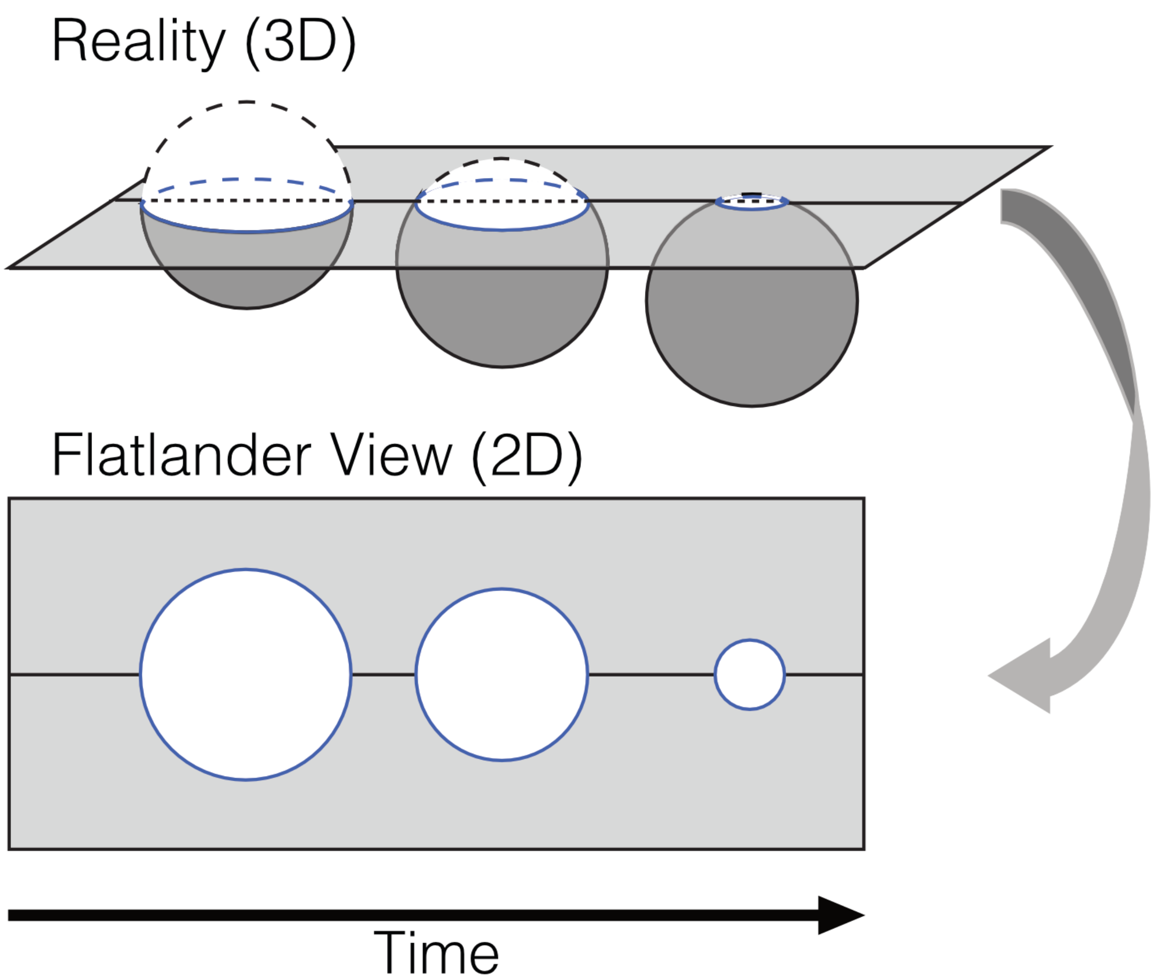
The cookbook approach



- Many statistics textbooks are organized in this way
- Works reasonably well if what we want to cook is in the book
- Leaves us with **no idea what to do if we can't find a recipe!**

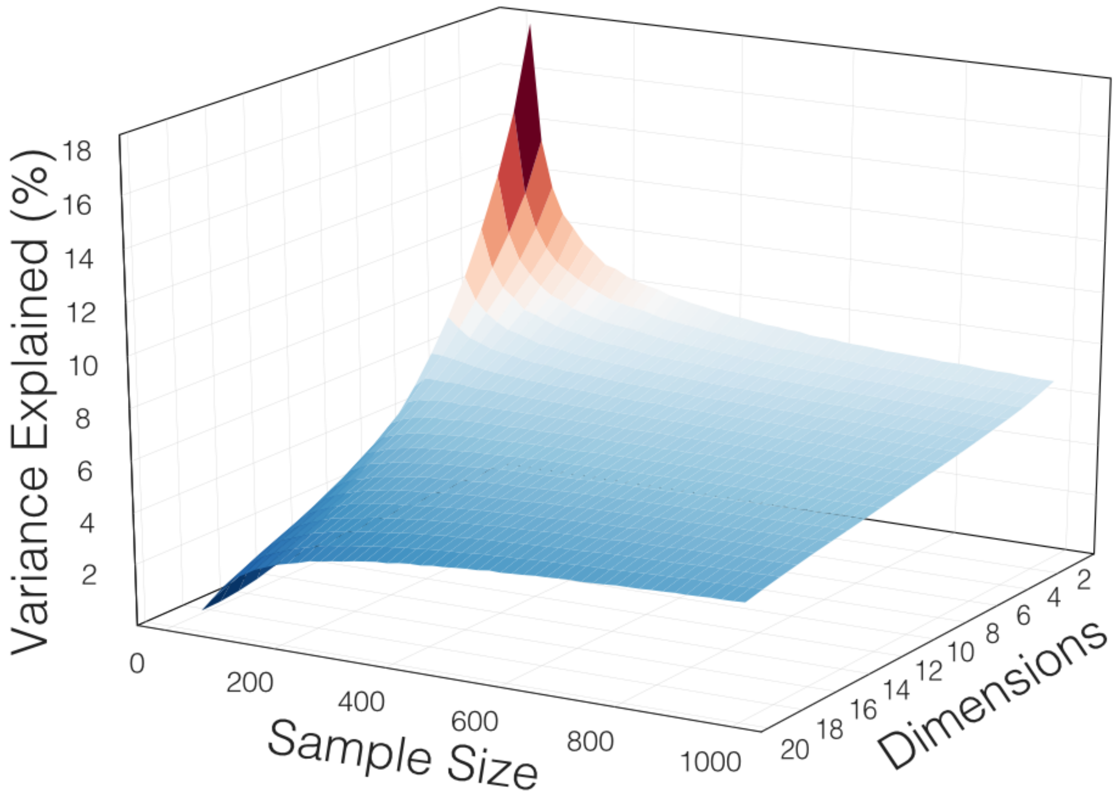
The problem: The Flatland Fallacy





The Flatland Fallacy

Inadequate Sampling Favors
Low Dimensional Explanations



"~~Ecological validity~~"—coined by Egon Brunswik in 1947

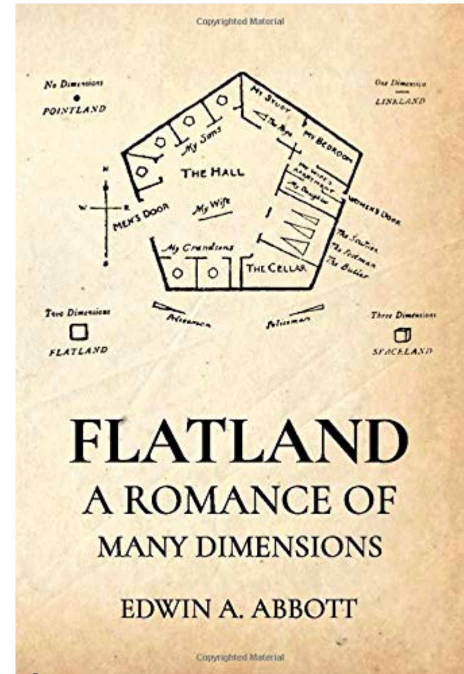
...to mean something else (¬_¬)

Representative design

Ecological generalizability demands a "representative sampling of situations" where "situational instances in an ecology are analogous to individuals in a population."

Jolly & Chang, 2019

The Flatland Fallacy



If we're **limited by recipes** we've memorized

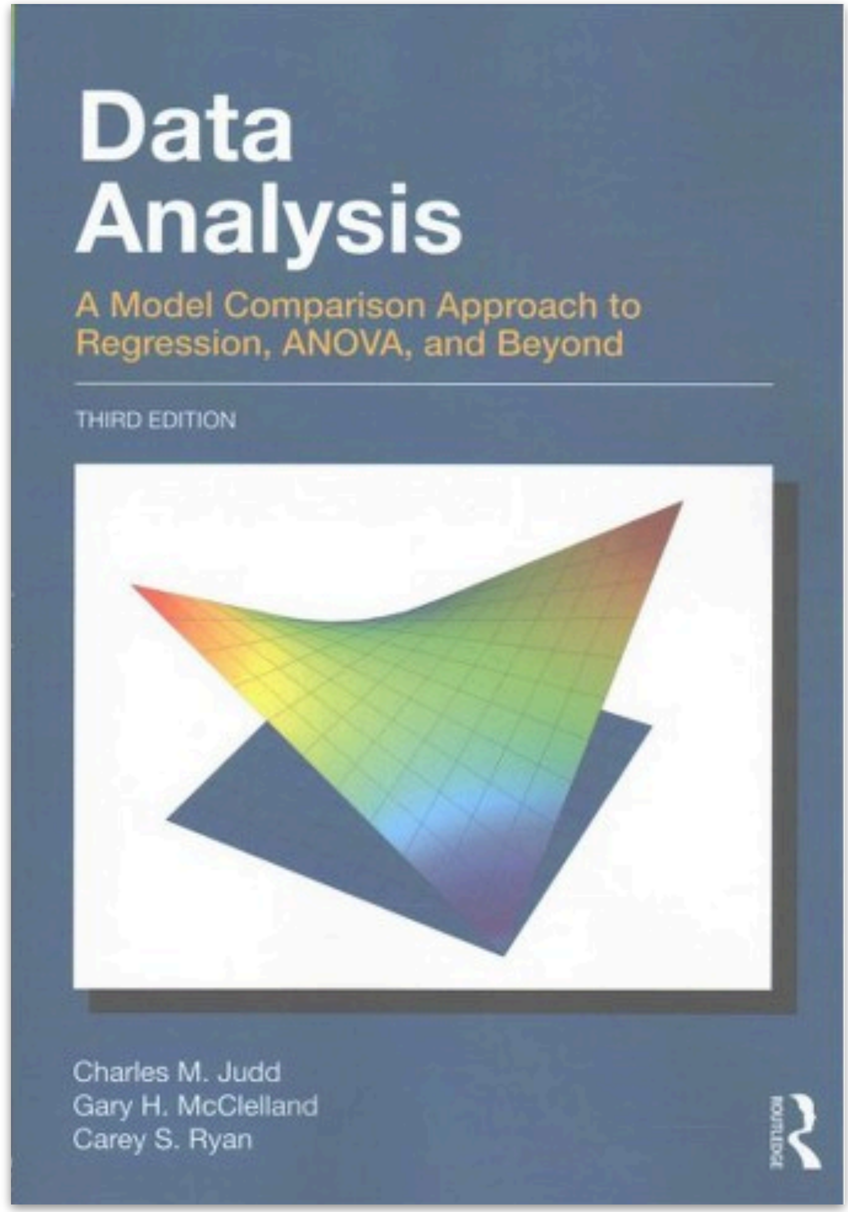
We're **limited** to designing **experiments** we can analyze

We're **limiting** the **answers** to our scientific questions

What are we even studying!?

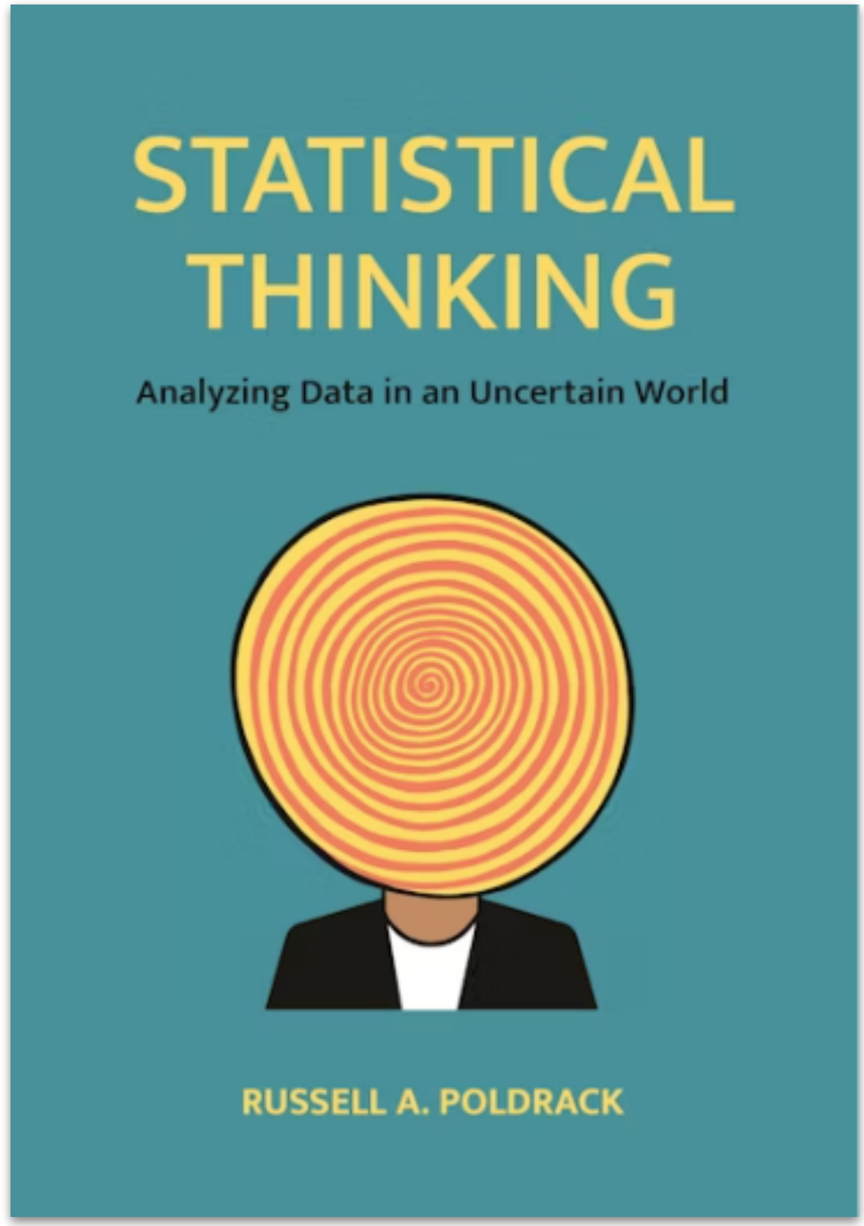
Jolly & Chang, 2019

Model comparison approach



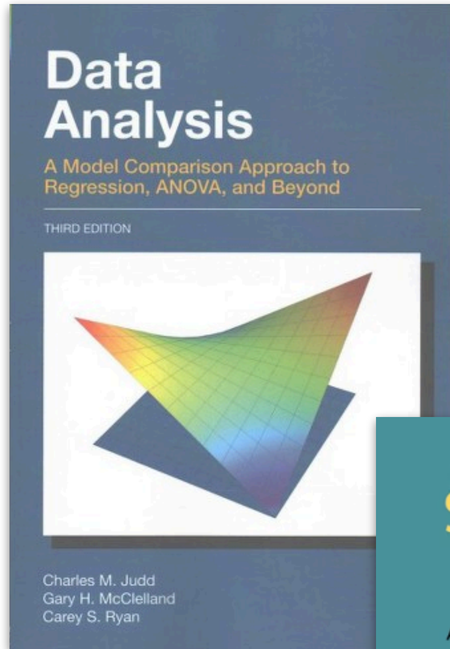
Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

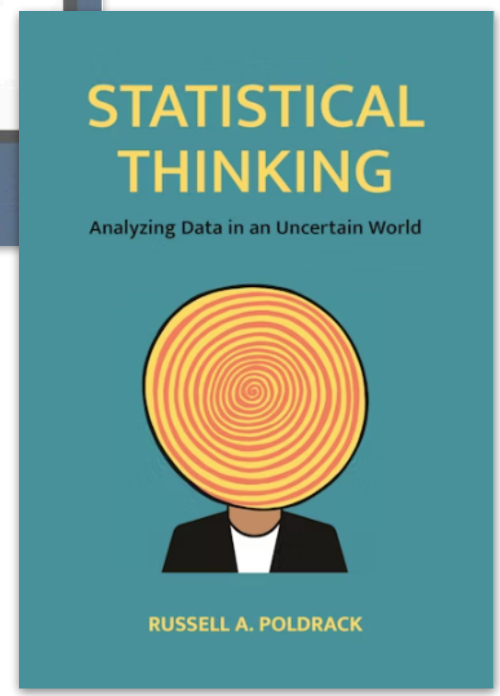
Model comparison approach



Model comparison approach

- Covers *all* statistical analyses
- Generates better insights and understanding
- Smoother transition into more advanced topics (e.g. Bayesian, deep-learning)





Questions?

See you tomorrow @2pm for Lab 2 (data visualization)!