Bayesian Statistics as an Alternative for Analyzing Data and Testing Hypotheses

Benjamin Scheibehenne

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Can Social Norms Increase Towel Reuse?

Standard Environmental Message

Descriptive Social Norm

Reuse:	35.1%	44.1%

$$X^2(1, N = 433) = 3.72; p = .05$$

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Goldstein, Cialdini, & Griskevicius, 2008, *Journal of Consumer Research*

The Null Ritual

1. Set up a statistical null hypothesis of “no mean difference” or “zero correlation.” Do not specify any substantive alternative hypotheses.
2. Calculate a p-value.
3. If p<.05, refer to the result as “statistically significant” and interpret it as if you obtained evidence for the alternative hypothesis.

Gigerenzer, Krauss, & Vitouch (2004)

What is Wrong with this Approach?

- p-values do not quantify evidence for your null or alternative hypothesis
- With high enough sample size, you are guaranteed to get $p<.05$
- p-values depend on the intention of the researcher (identical data can yield different p values)
- p-values are not consistent: Little can be learned from non-significant results

When H_0 is True, p-Values are not Consistent

Alternative hypothesis is true

More data → higher chance of identifying the true hypothesis

Null-Hypothesis is true

More data → chance of identifying the true (null) hypothesis is constant (at 1-α)

Under H_0, p-Values are uniformly distributed

p-Values Depend on the Intention of the Experimenter

- Default assumption: *N* is fixed

$$X^2(1, N = 433) = 3.72;\ p = .05$$

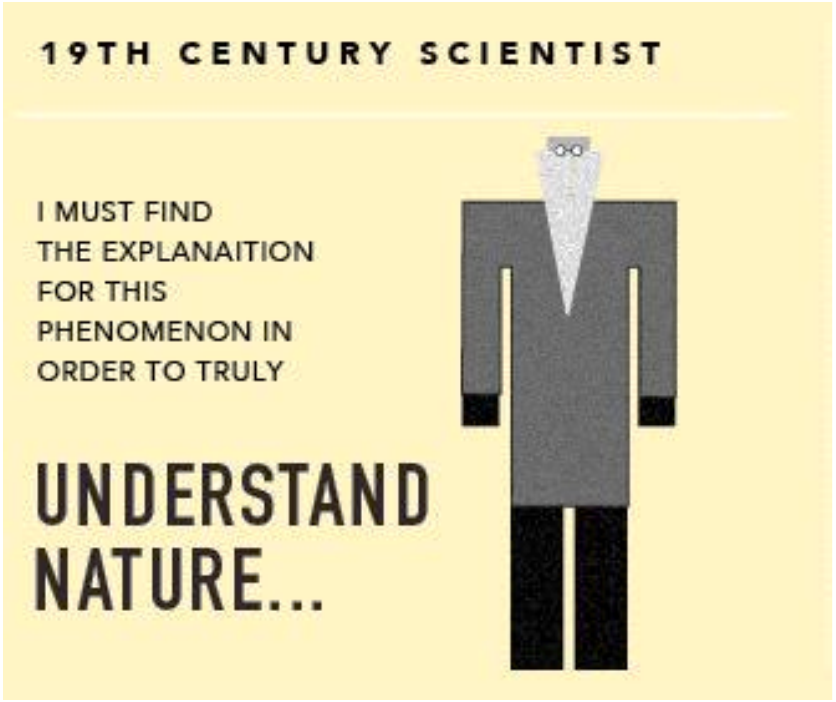
actually: $p = .053$

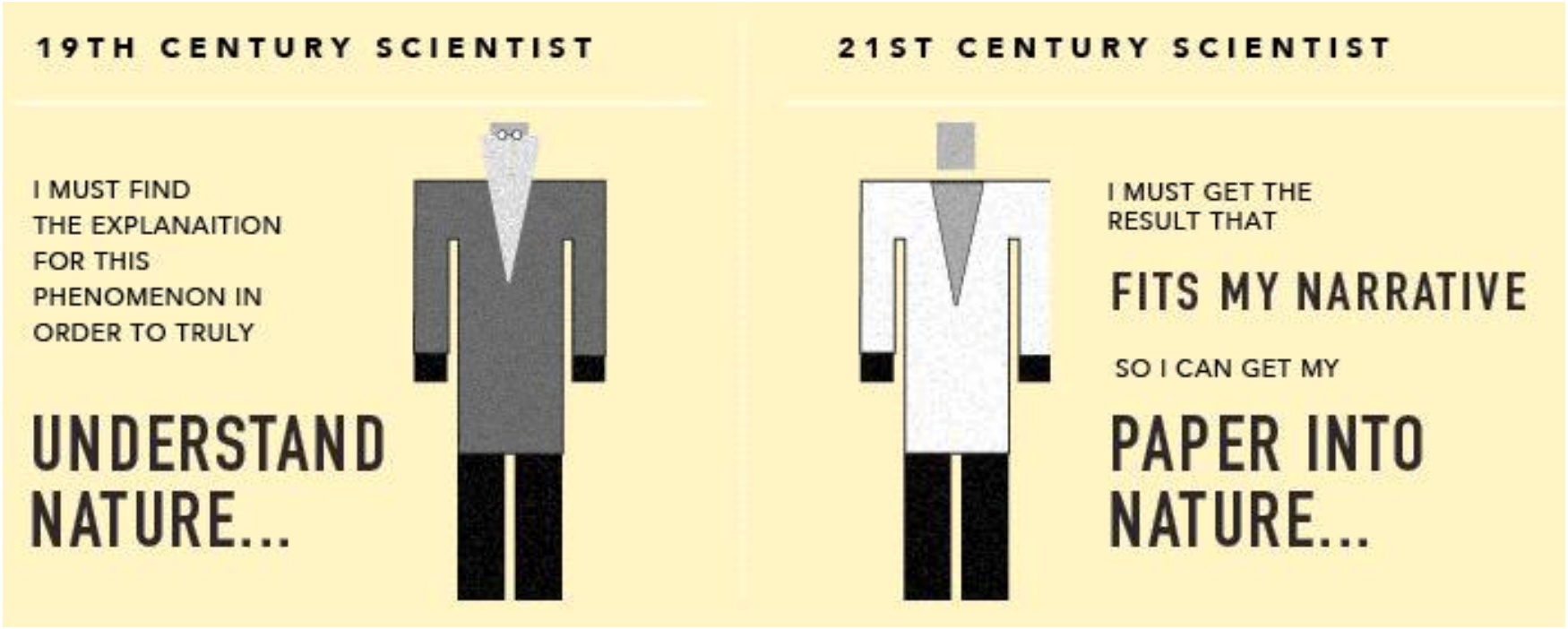
- What if the number of hotel guests is a random variable?

$N_{control} \sim \mu = 211 \quad \sigma = 30$

$N_{ex} \sim \mu = 222 \quad \sigma = 30$

$$X^2(1, N = 433) = 3.72;\ p = .06$$


19TH CENTURY SCIENTIST
I MUST FIND
THE EXPLANAITION
FOR THIS
PHENOMENON IN
ORDER TO TRULY
UNDERSTAND
NATURE...


19TH CENTURY SCIENTIST
I MUST FIND
THE EXPLANAITION
FOR THIS
PHENOMENON IN
ORDER TO TRULY
UNDERSTAND
NATURE...
21ST CENTURY SCIENTIST
I MUST GET THE
RESULT THAT
FITS MY NARRATIVE
SO I CAN GET MY
PAPER INTO
NATURE...

American Statistical Association Issues Warning

nature International weekly journal of science

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

"The widespread use of "statistical significance" (generally interpreted as "p ≤ 0.05") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today[1]. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied in ways that cast doubt on statistics generally, he adds.

Message is not new…

"To report a significant result and reject the null in favor of an alternative hypothesis is meaningless […]." (*Richard Feynman*)

"In the land of the blind, keep your eyes closed."

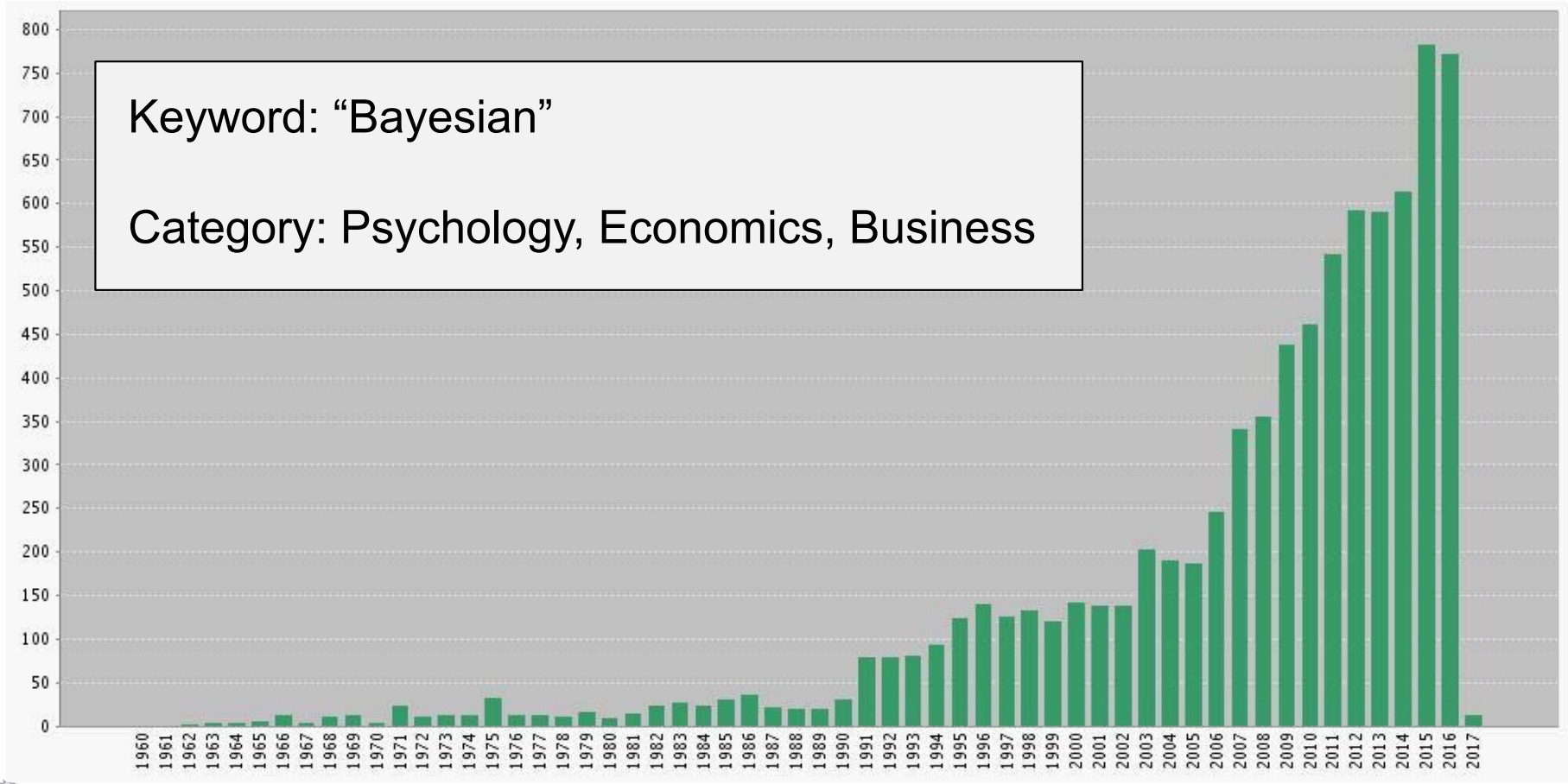
[proverb from Kurdistan]

Bayesian Approach Provides an Alternative

"Friends don't let friends compute p-values"

Bayesian Approach Increasingly Attractive

 # Articles in Web of Science



What Makes the Bayesian Approach Attractive?

New Results
- Evidence for Hypotheses (incl. H0)

Better Results
- More accurate estimates and predictions
- Intuitive interpretation

Compare different models and hypotheses
- Take model complexity into account

Broad application
- “One-stop shop“

What Makes the Bayesian Approach Attractive?

New Results
- Evidence for Hypotheses (incl. H0)

Better Results
- More accurate estimates and predictions
- Intuitive interpretation

Compare different models and hypotheses
- Take model complexity into account

Broad application
- “One-stop shop“

Can Social Norms Increase Towel Reuse?

Standard Message

Descriptive Social Norm

What is the evidence?

Ex1:	35.1%	44.1%	N = 433	p = .05
Ex2:	37.2%	44.5%	N = 1,595	p = .03

Bayesian Hypothesis Test

- Suppose we have two hypotheses: H1 and H0
- Which hypothesis is better supported by the data?
- The model that makes the best prediction for the data!
- Ratio of predictive performance = **Bayes Factor**

Bayes Factor

Prior odds

$$\frac{p(H_1)}{p(H_0)}$$

•

Bayes Factor

X

Posterior odds

$$= \frac{p(H_1 \mid D)}{p(H_0 \mid D)}$$

Change in probability of one hypothesis/ model over another after seeing the data

posterior probability of one model over the other

Bayes Factor

Prior odds

$$\frac{p(H_1)}{p(H_0)}$$

$\bullet$

Bayes Factor

$$\frac{\int p(D \mid H_1) \bullet p(H_1)}{\int p(D \mid H_0) \bullet p(H_0)}$$

$=$

Posterior odds

$$\frac{p(H_1 \mid D)}{p(H_0 \mid D)}$$

Change in probability of one hypothesis/ model over another after seeing the data

posterior probability of one model over the other

Bayes Factor

Prior odds | **Bayes Factor** | **Posterior odds**

$$\frac{p(H_1)}{p(H_0)} \bullet \frac{\int p(D \mid H_1) \bullet p(H_1)}{\int p(D \mid H_0) \bullet p(H_0)} = \frac{p(H_1 \mid D)}{p(H_0 \mid D)}$$

Change in probability of one hypothesis/ model over another after seeing the data

posterior probability of one model over the other

Simple Interpretation:

BF = 3: After seeing the data, H_1 became 3 times more probable than H_0

Bayes Factor

Prior odds | **Bayes Factor** | **Posterior odds**

$$\frac{p(H_1)}{p(H_0)} \bullet \frac{\int p(D \mid H_1) \bullet p(H_1)}{\int p(D \mid H_0) \bullet p(H_0)} = \frac{p(H_1 \mid D)}{p(H_0 \mid D)}$$

Change in probability of one hypothesis/ model over another after seeing the data

posterior probability of one model over the other

Simple Interpretation:

BF = 3: After seeing the data, H_1 became 3 times more probable than H_0

1 to 3	Anecdotal
3 to 10	Substantial
10 to 100	Strong
>100	Decisive

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Advantages of Bayes Factors

- Quantifies evidence instead of forcing an all-or-none decision
- Easy to interpret
- All assumptions are explicit
- Able to distinguish between “data support H_0” and “data are not diagnostic”

Bayes Factor for Towel Data

Ex1	Reuse	Throw
Standard message	74	137
Social norm	98	124

Gunel & Dickey (1974) *Biometrika*

Bayes Factor for Towel Data

Ex1	Reuse	Throw
Standard message	74	137
Social norm	98	124

Assumptions:

- Sample size in each condition (row) is fixed
- Probability of reuse = $\pi_{standard}$, $\pi_{social\ norm}$
- Prior: Each value for π is equally probable a-priori
- H0: $\pi_{standard} = \pi_{social\ norm}$
- H1: $\pi_{standard} < \pi_{social\ norm}$

Independent multinomial distribution

Bayes Factor for Towel Data

Ex1	Reuse	Throw
Standard message	74	137
Social norm	98	124

Assumptions:

- Sample size in each condition (row) is fixed
- Probability of reuse = $\pi_{standard}$, $\pi_{social\ norm}$
- Prior: Each value for π is equally probable a-priori
- H0: $\pi_{standard} = \pi_{social\ norm}$
- H1: $\pi_{standard} < \pi_{social\ norm}$

Independent multinomial distribution

Result:

- BF = 1.46
- BF for undirected hypothesis ($\pi_{standard} \neq \pi_{social\ norm}$) = 0.75

Can Social Norms Increase Towel Reuse?

Standard Message

Descriptive Social Norm

What is the evidence?

Ex1:	35.1%	44.1%	N = 433	p = .05	**BF = 1.46**
Ex2:	37.2%	44.5%	N = 1,595	p = .03	**BF = 2.03**

Can Social Norms Increase Towel Reuse?

Lots of failed replications…

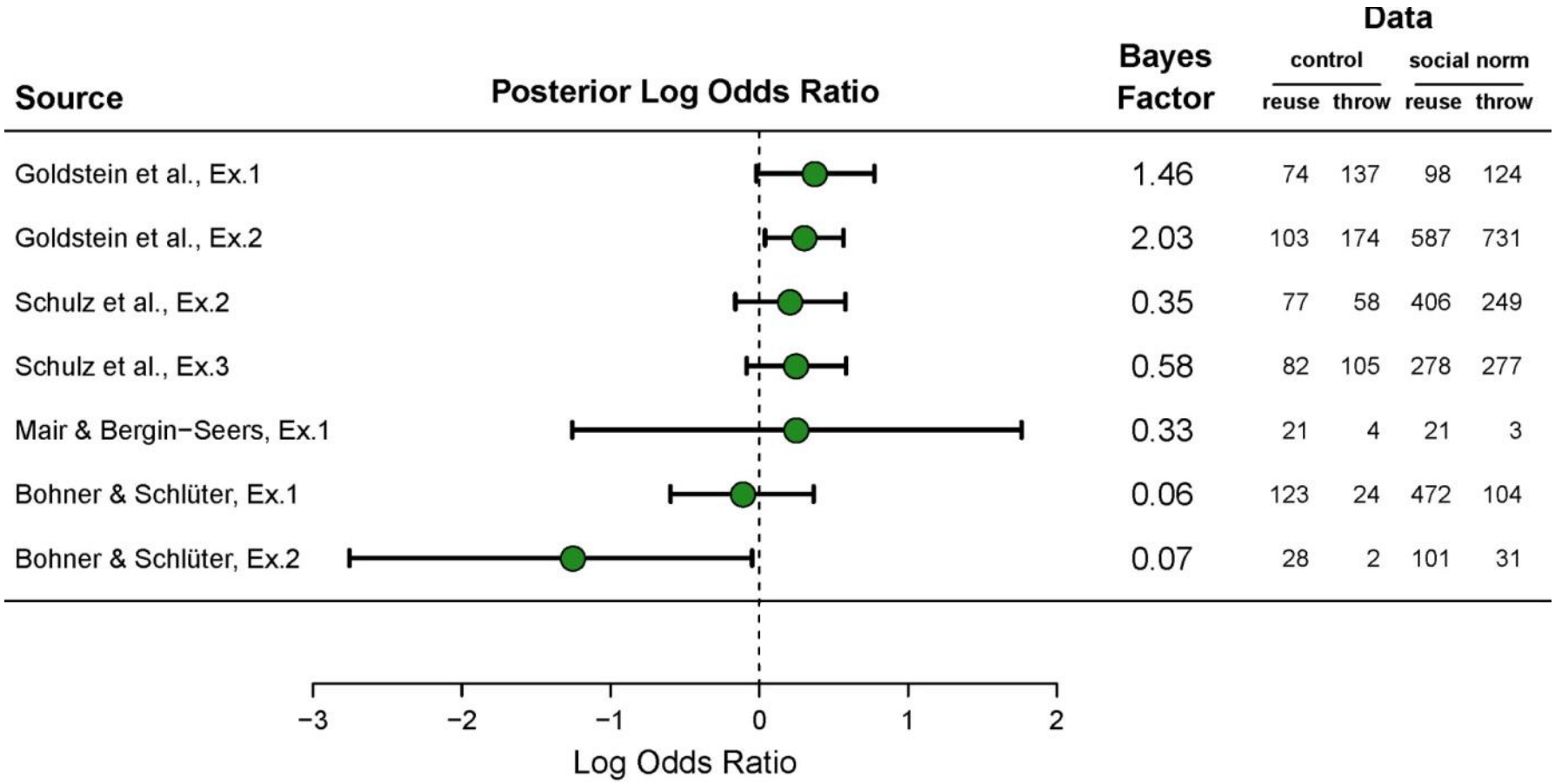
Schultz et al. (2007) *Social Influence*

Mair & Bergin-Seers (2010) *Tourism and Hospitality Research*

Bohner & Schlüter (2014) *PloS ONE*

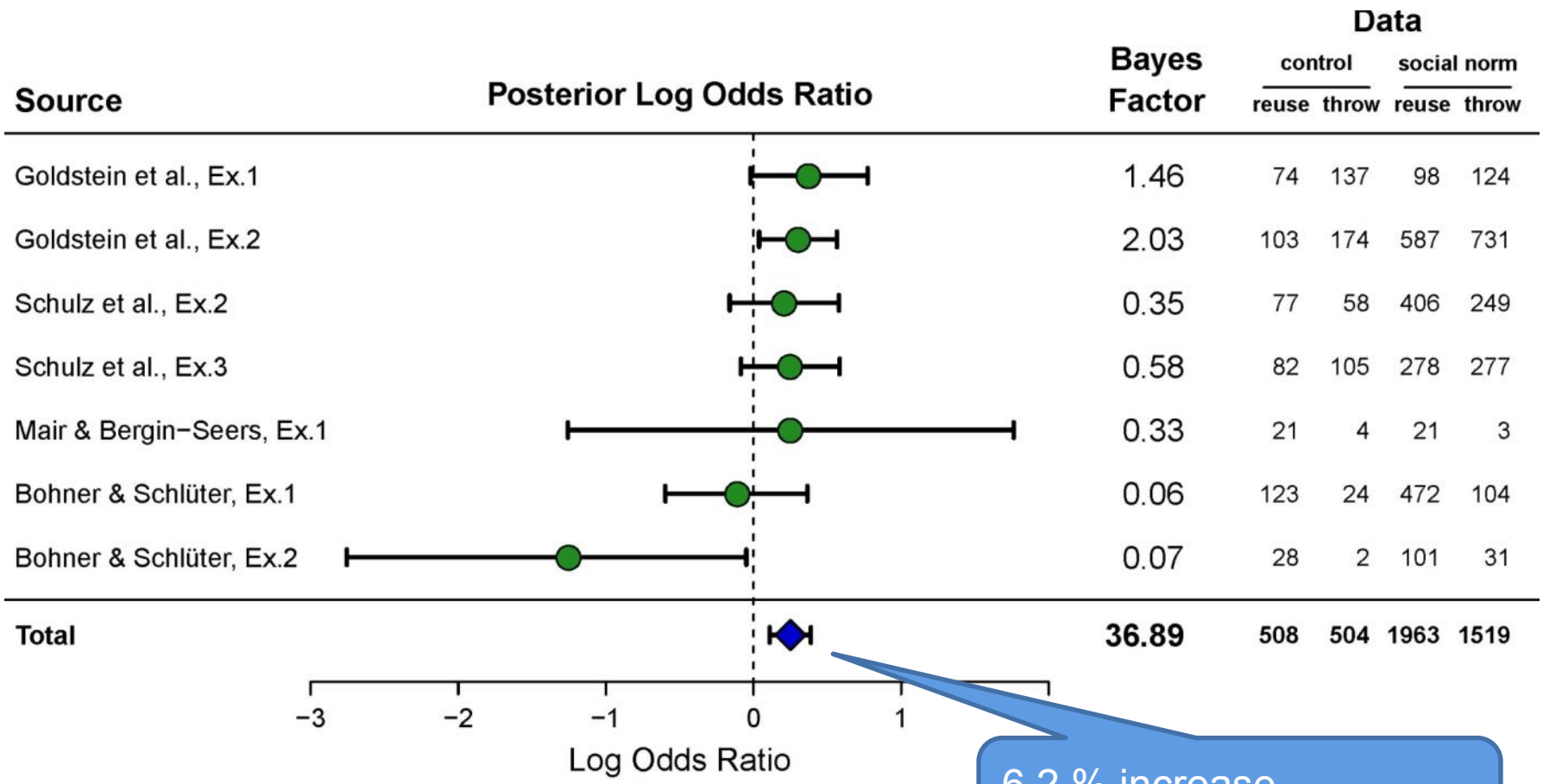
} all p's > .14

Can Social Norms Increase Towel Reuse?



UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

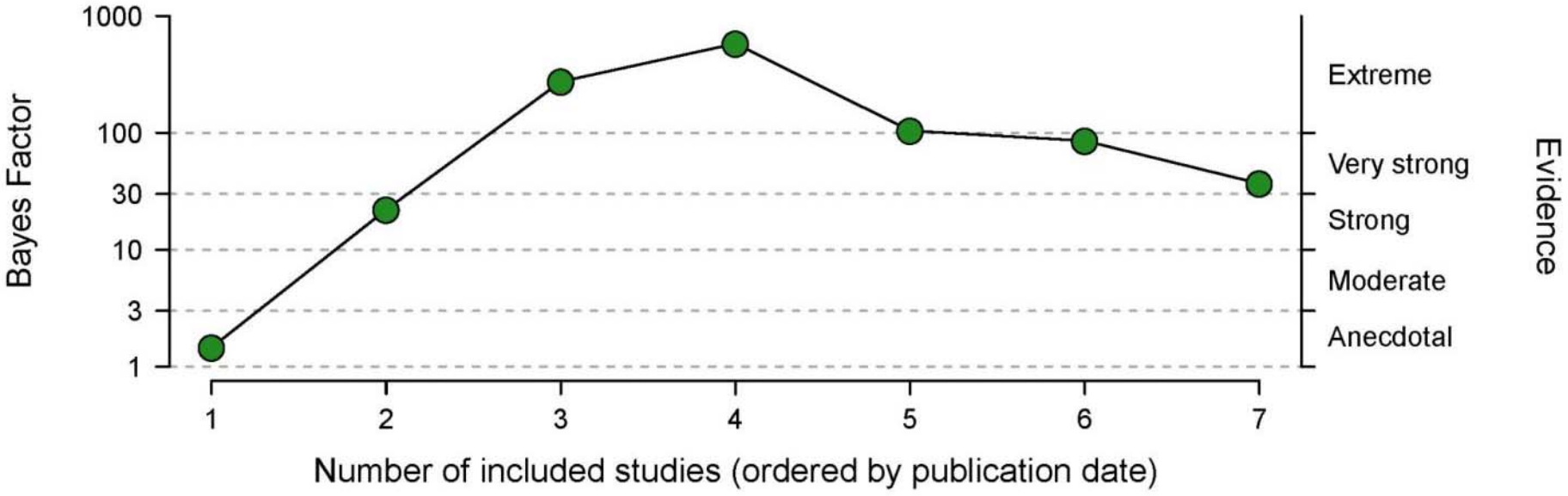
Can Social Norms Increase Towel Reuse?



UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Can Social Norms Increase Towel Reuse?

- Cumulative evidence “as the data comes in”
- Can be updated in the future



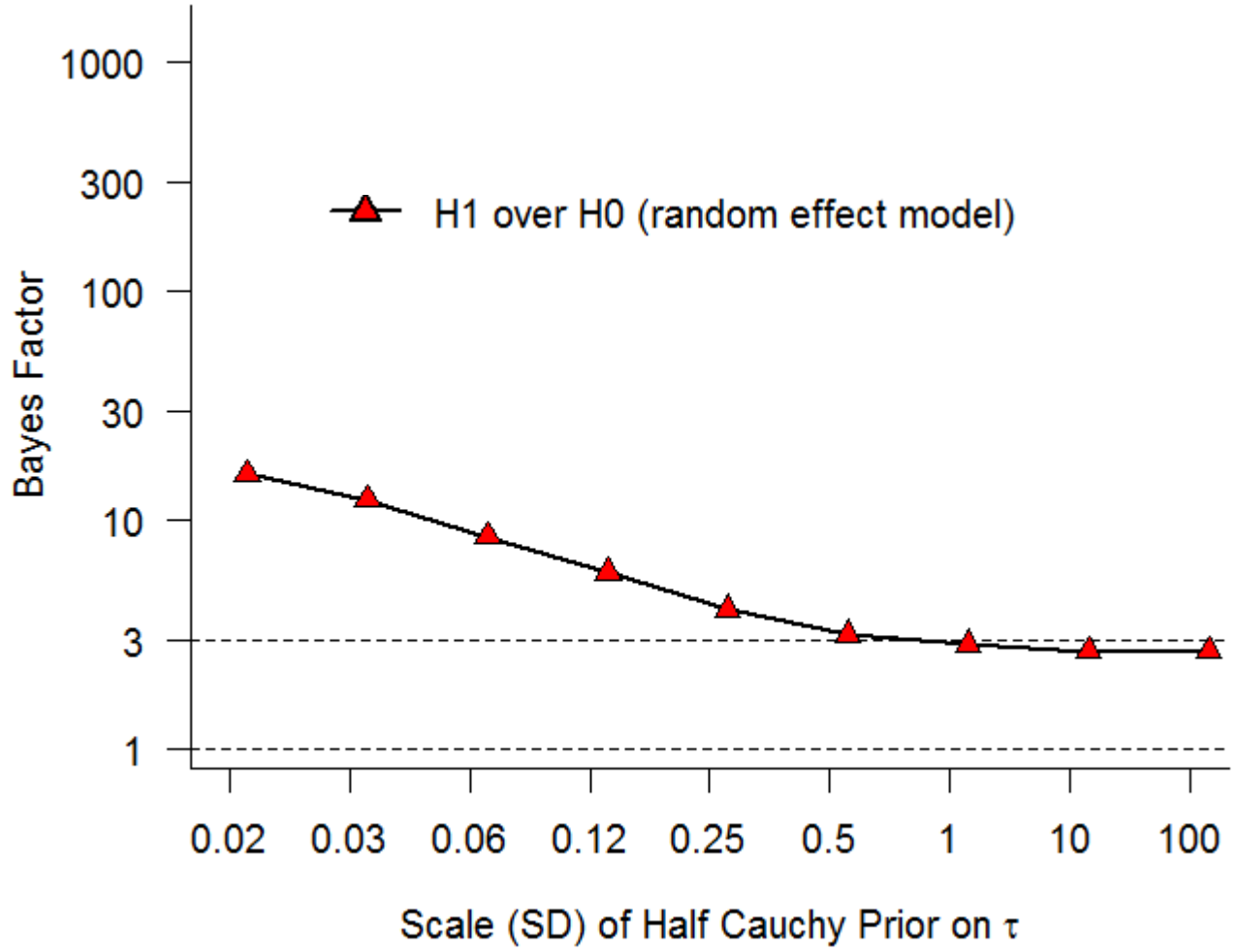
Repeated Testing is Allowed

- Multiple tests in NHST = risk of false-positive findings
- In a Bayesian framework, optional stopping does not distort the results
- No correction for repeated testing required

"It follows that any significance test based on Bayes's theorem does not depend on the sequential stopping rule used […]"

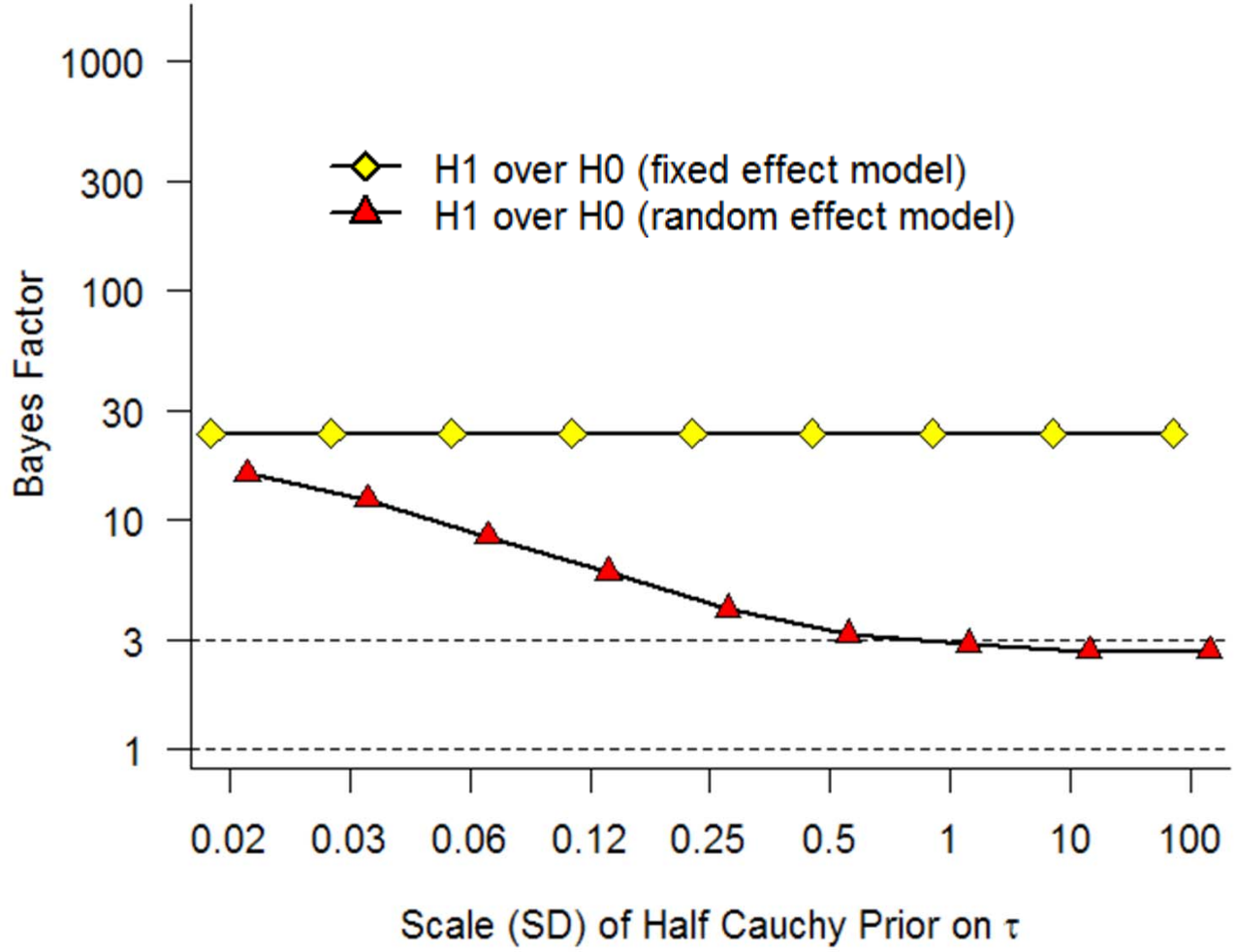
Dennis Lindley (1957)

Would a Random-Effect Model be Better Suited?



= Prior on the between-study variance

Would a Random-Effect Model be Better Suited?



= Prior on the between-study variance

What Makes the Bayesian Approach Attractive?

New Results
- Evidence for Hypotheses (incl. H0)

Better Results
- More accurate estimates and predictions
- Intuitive interpretation

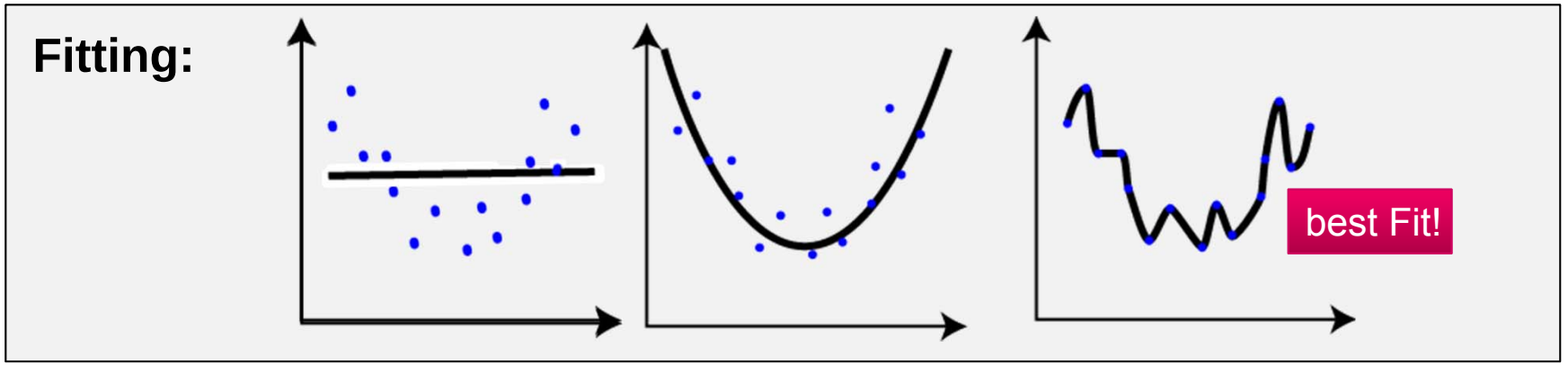
Compare different models and hypotheses
- Take model complexity into account

Broad application
- “One-stop shop“

How can we Compare Models?

How can we Compare Models?

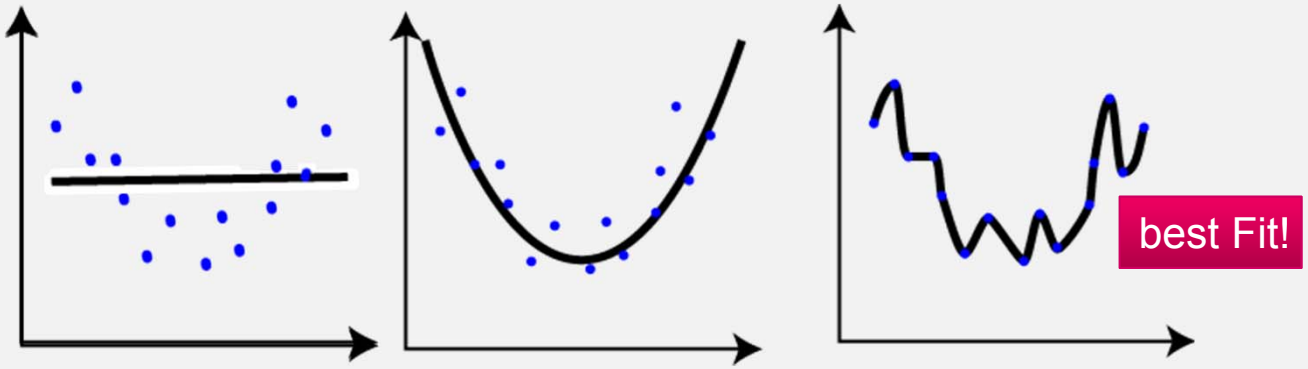
More complex models provide a better *description* of the data...



How can we Compare Models?

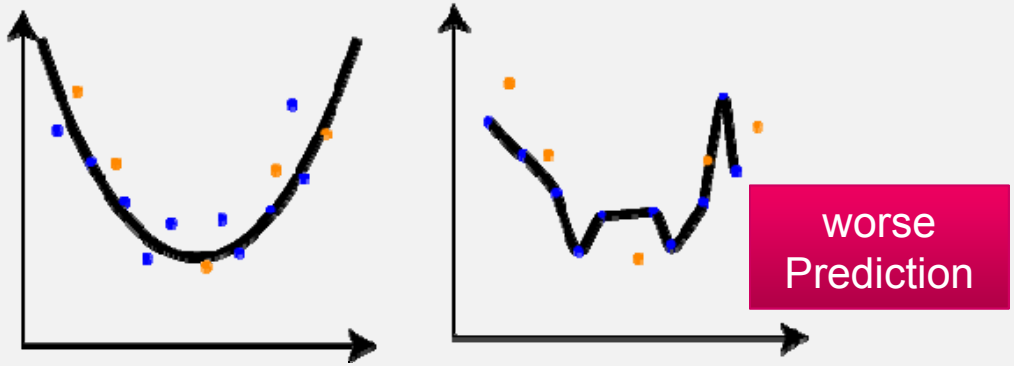
More complex models provide a better *description* of the data...

Fitting:



…but not necessarily a better *explanation*:

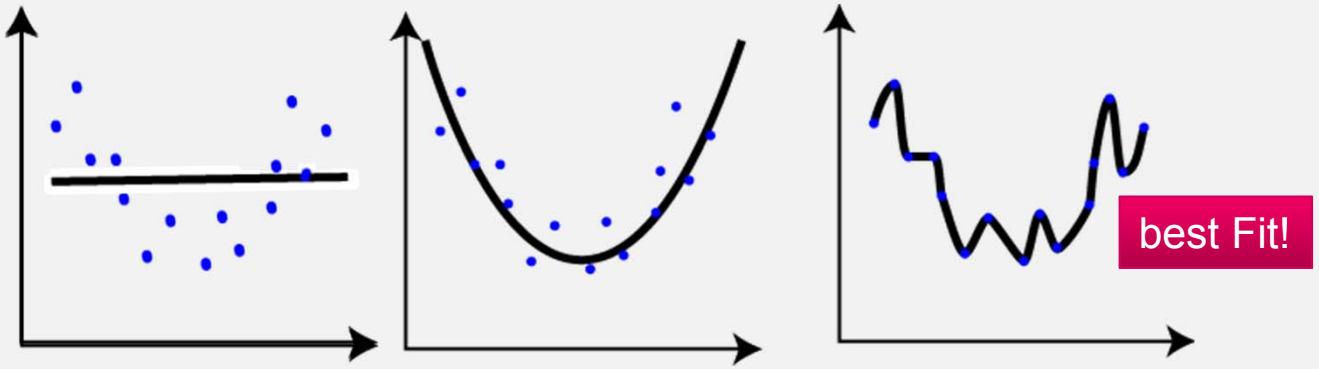
Prediction of new data:



How can we Compare Models?

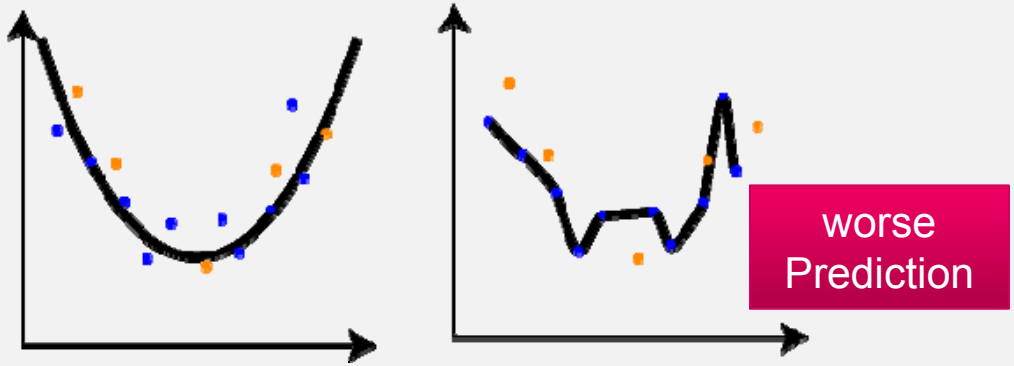
More complex models provide a better *description* of the data...

Fitting:



…but not necessarily a better *explanation*:

Prediction of new data:



UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

Trade-off between complexity and fit

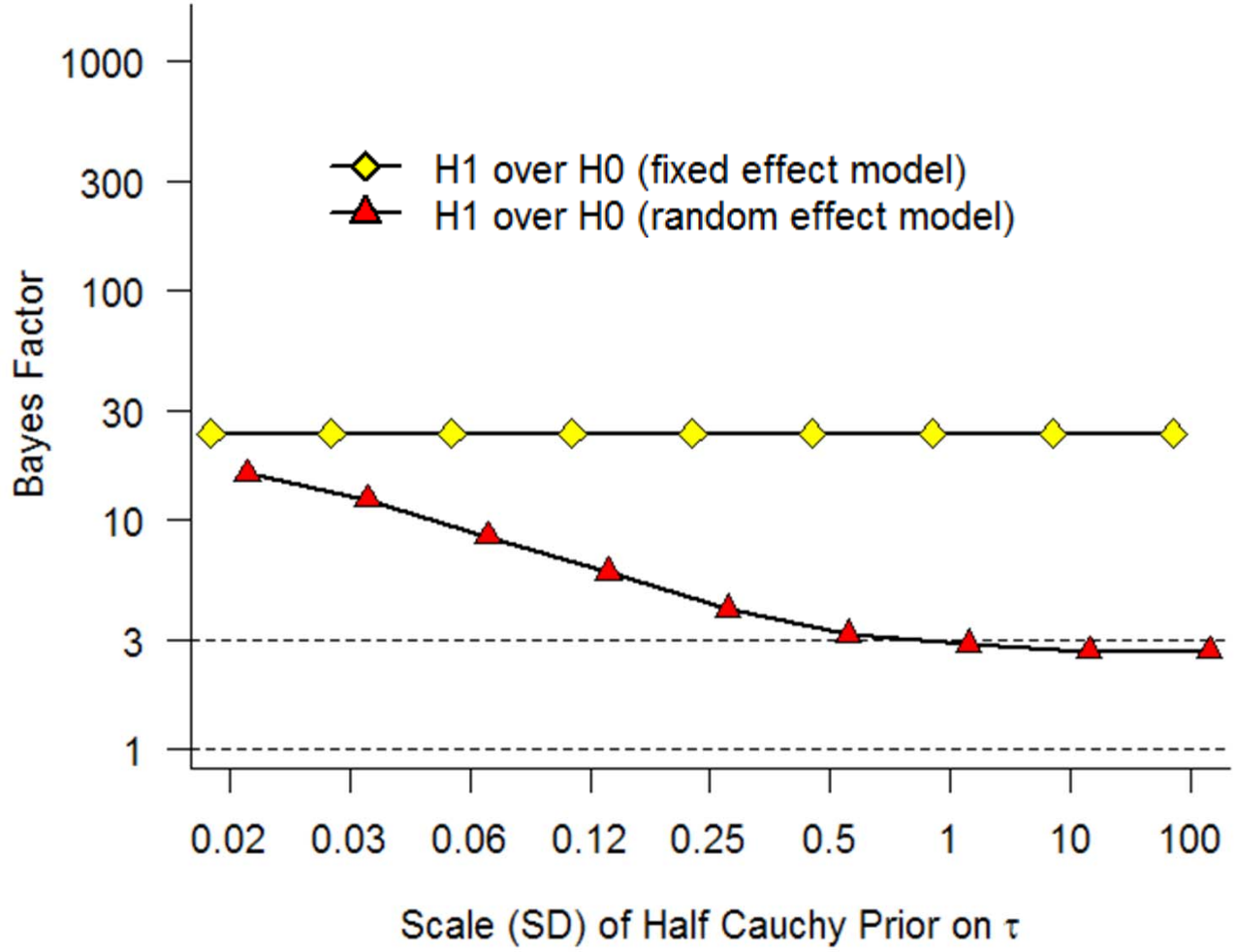
Bayes Factor

average probability of the observed data across all predictions the model makes

$$\frac{\int p(D|\theta_1, M_1) \cdot p(\theta_1, M_1)}{\int p(D|\theta_2, M_2) \cdot p(\theta_2, M_2)}$$

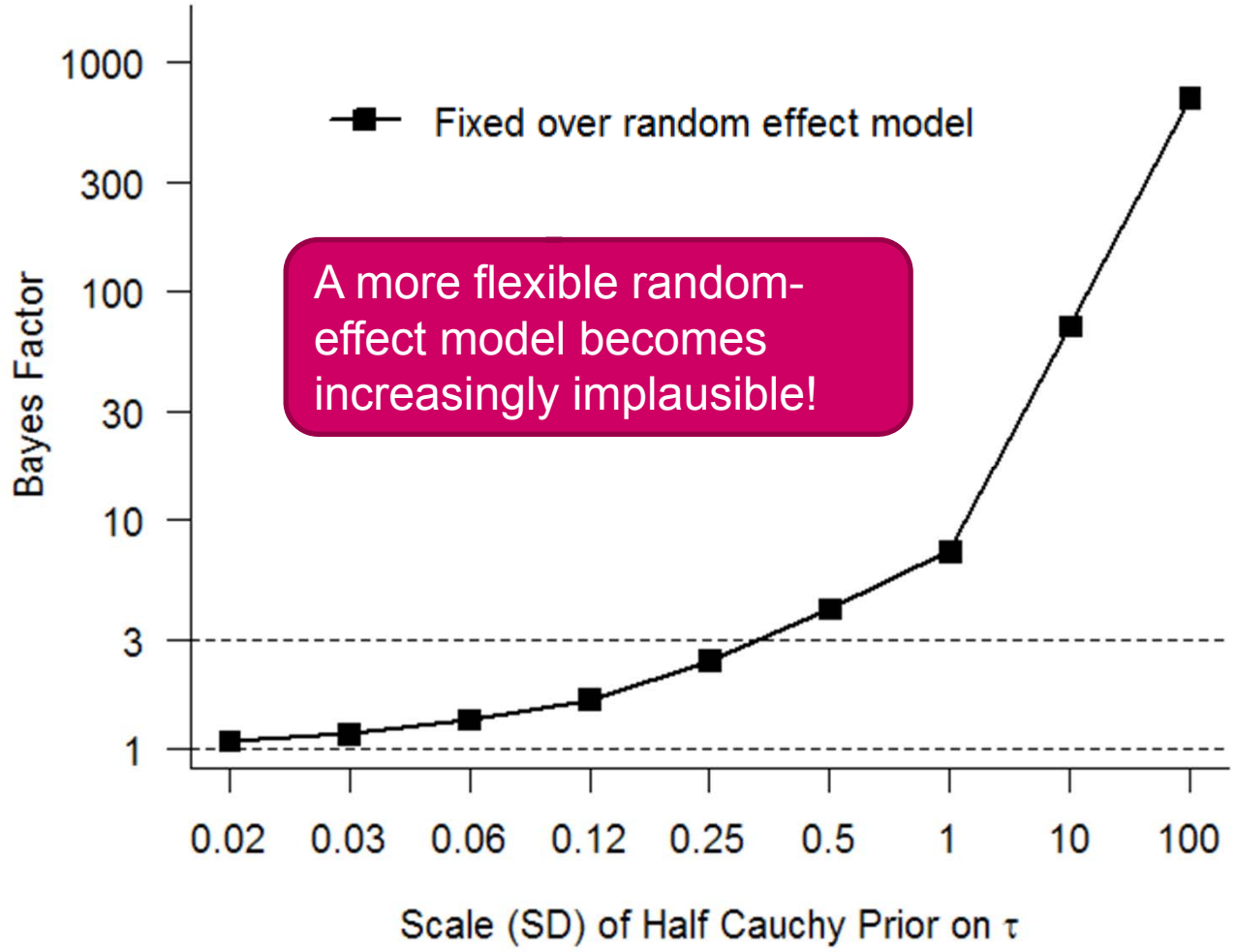
- Flexible models make many predictions
- If most of these predictions are wrong, this drives down the average probability

Would a Random-Effect Model be Better Suited?



= Prior on the between-study variance

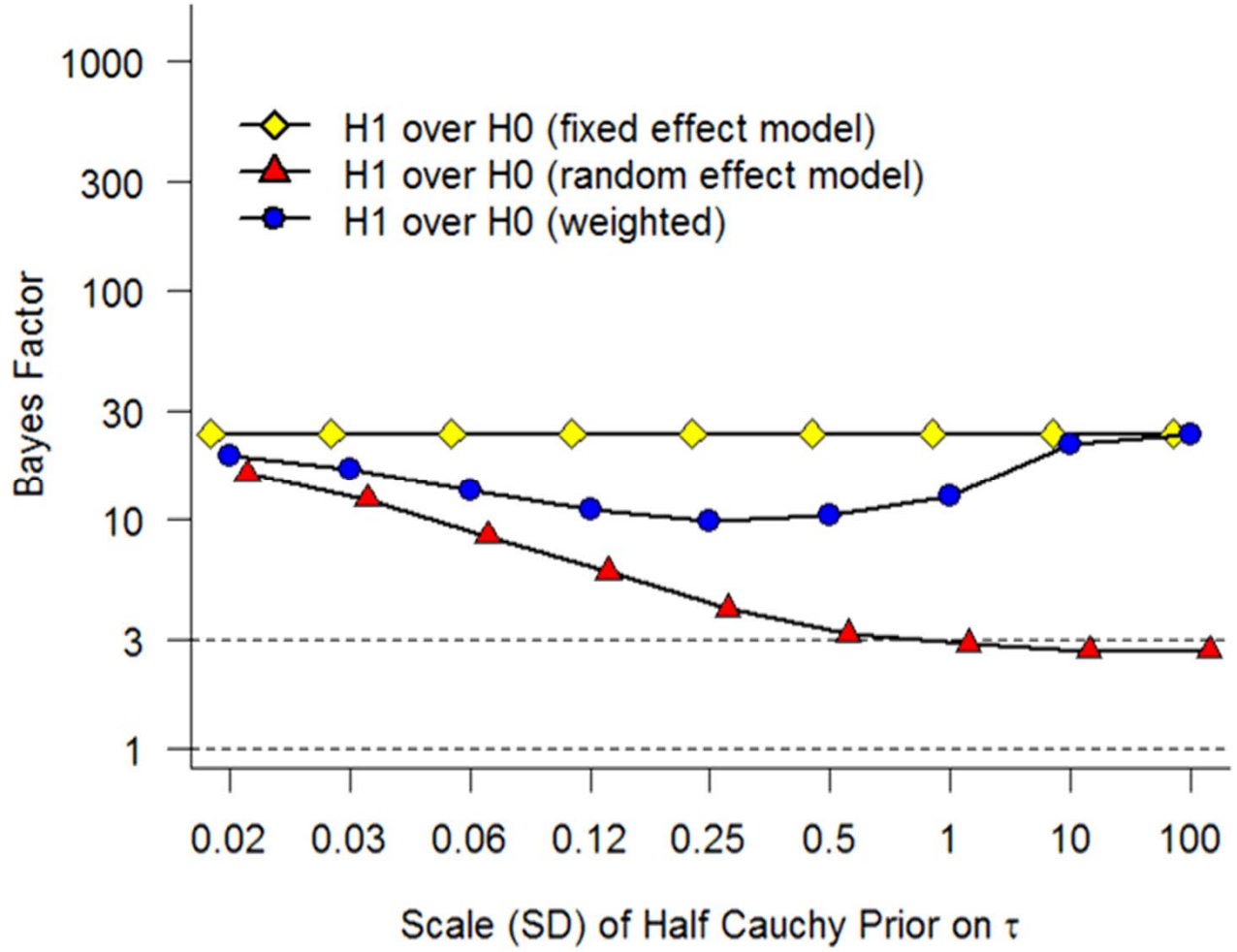
Would a Random-Effect Model be Better Suited?



= Prior on the between-study variance

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT

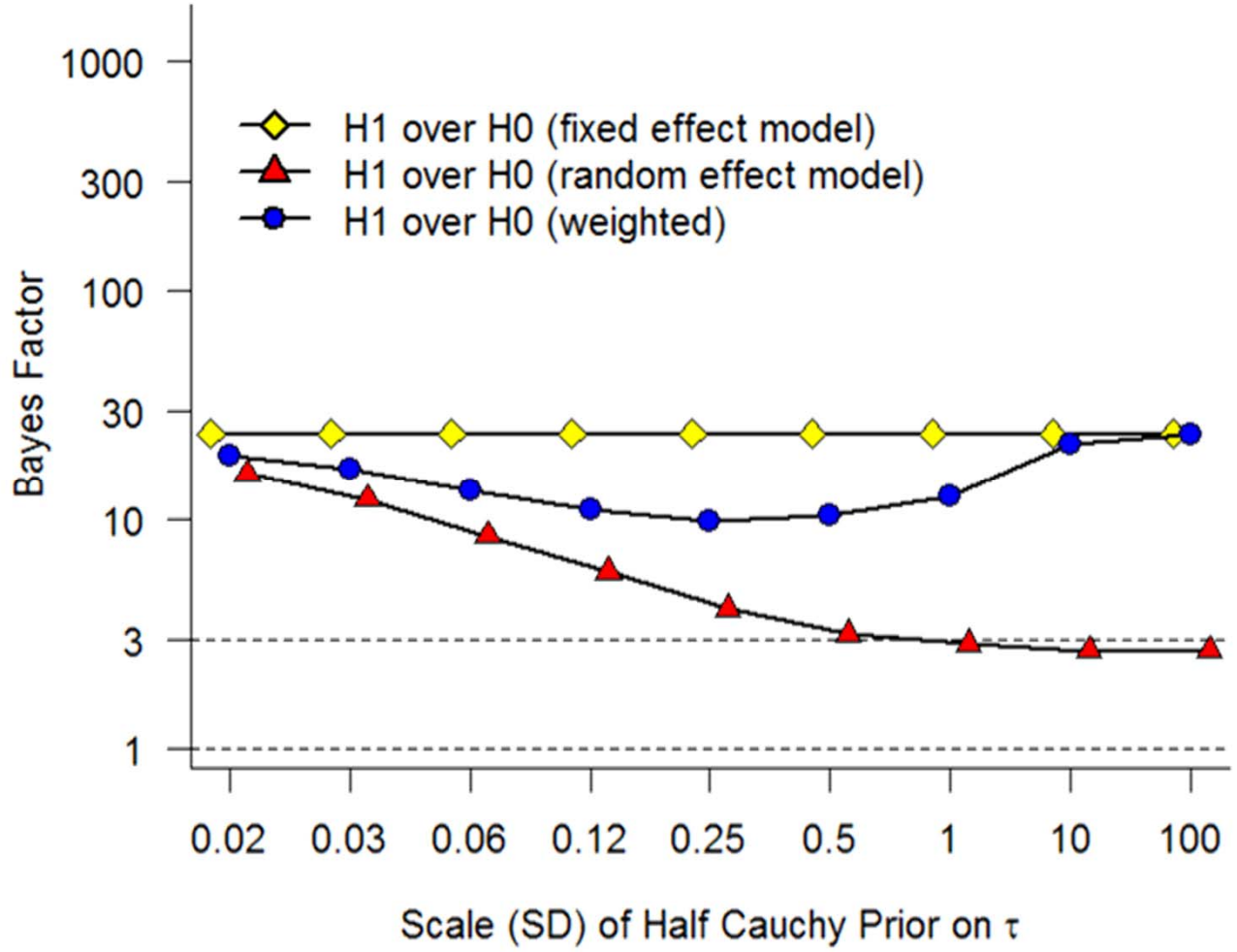
Resolution: Bayesian Model Averaging



= Prior on the between-study variance

Scheibehenne, et al. (under review) *Psychological Science*

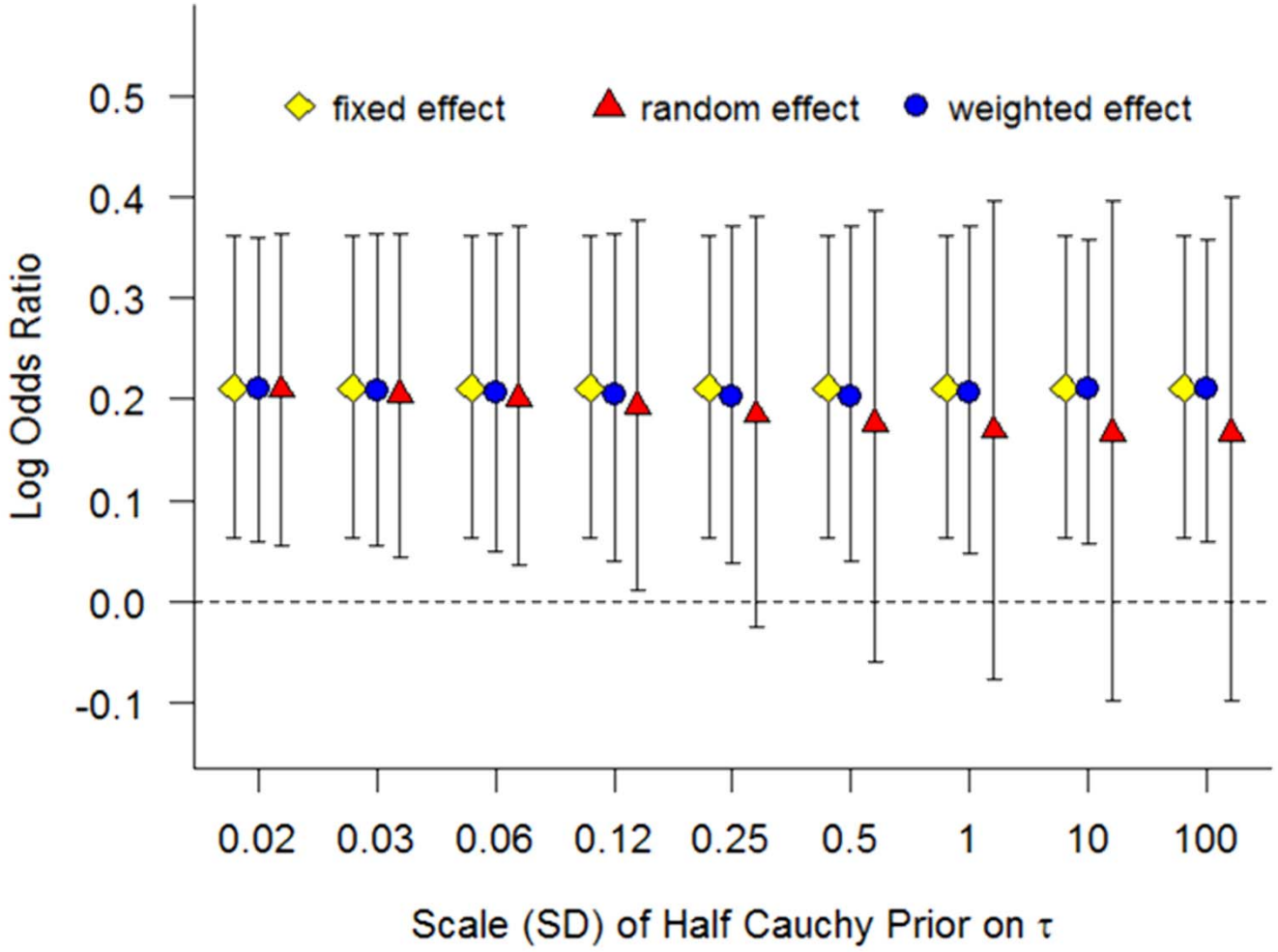
Resolution: Bayesian Model Averaging



= Prior on the between-study variance

Scheibehenne, et al. (under review) *Psychological Science*

Resolution: Bayesian Model Averaging



= Prior on the between-study variance

Scheibehenne, et al. (under review) *Psychological Science*

What Makes the Bayesian Approach Attractive?

New Results
– Evidence for Hypotheses (incl. H0)

Better Results
– More accurate estimates and predictions
– Intuitive interpretation

Compare different models and hypotheses
- Take model complexity into account

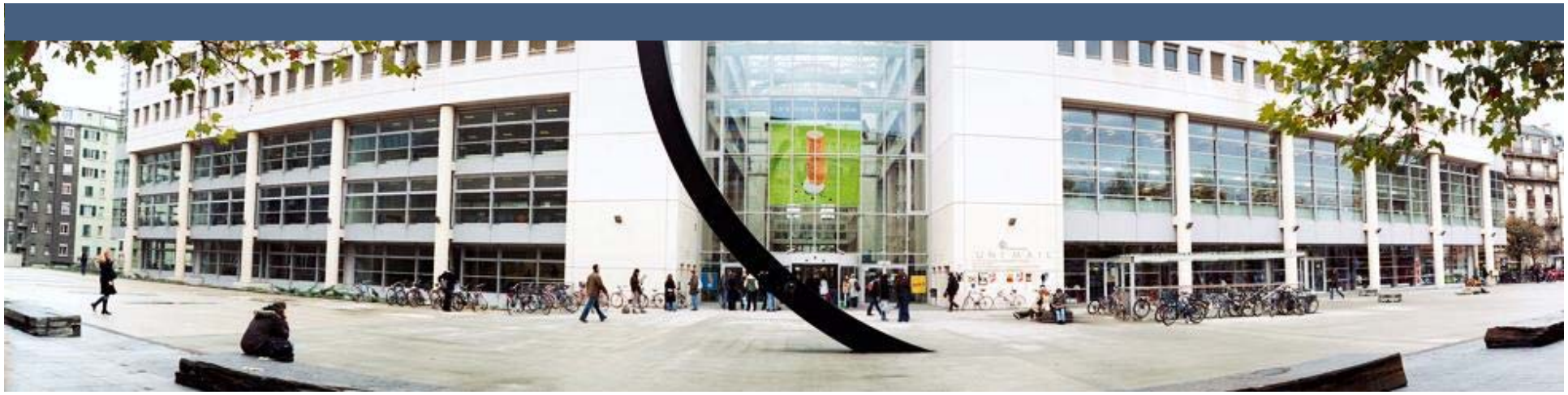
Broad application
- “One-stop shop“

Broad Application

- Well suited for estimating more complex models
- Lends itself to hierarchical designs

Advantages of Bayesian Approach

- Probabilities for hypotheses
- Able to distinguish b
  “data are not diagno
- Easy to interpret
- Can be updated as
- Possibility to include
- All assumptions are
- Model comparison:
- Well-suited for big d
- High "street credibili



Bayesian Statistics as an Alternative for Analyzing Data and Testing Hypotheses

Benjamin Scheibehenne

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF ECONOMICS AND MANAGEMENT