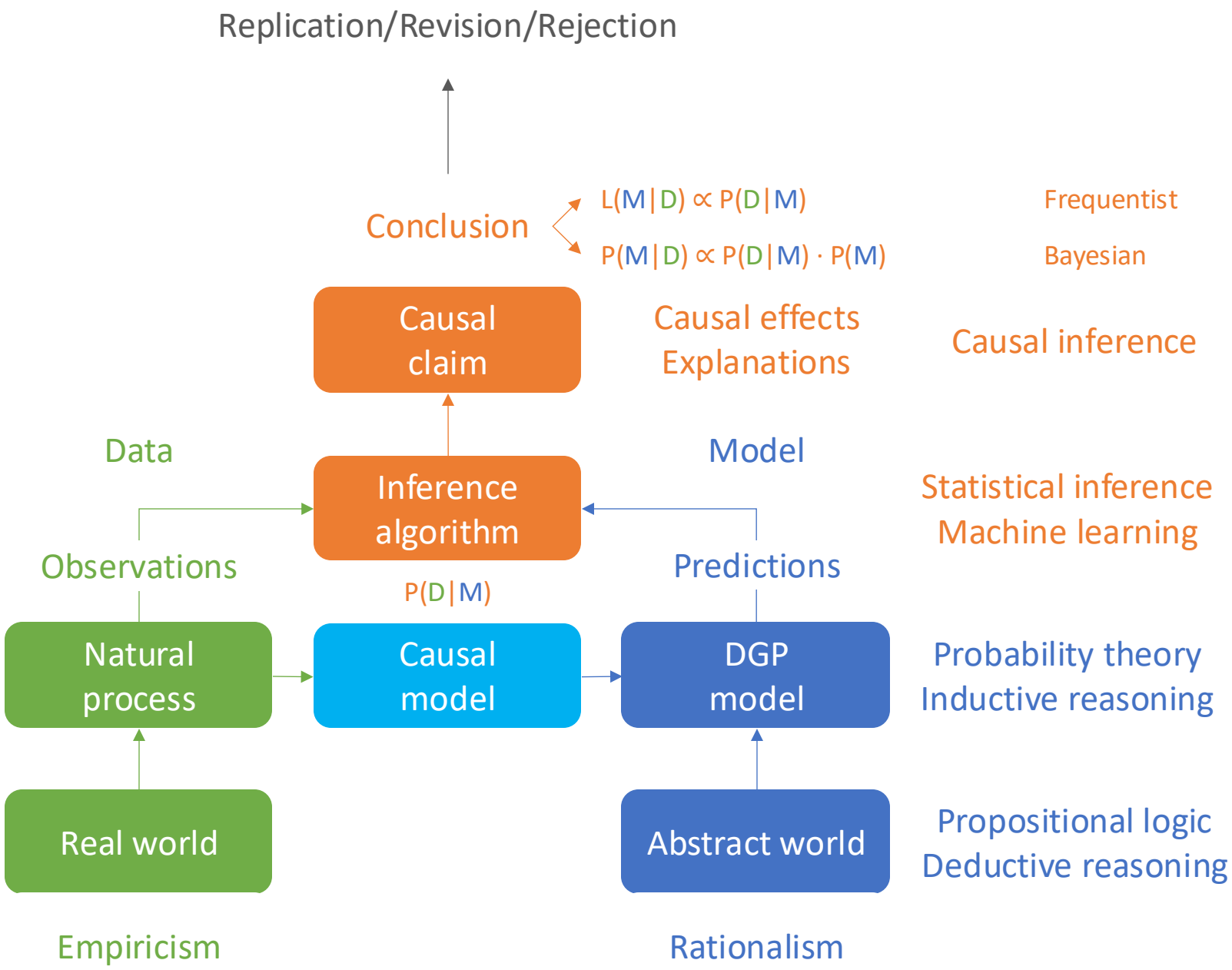
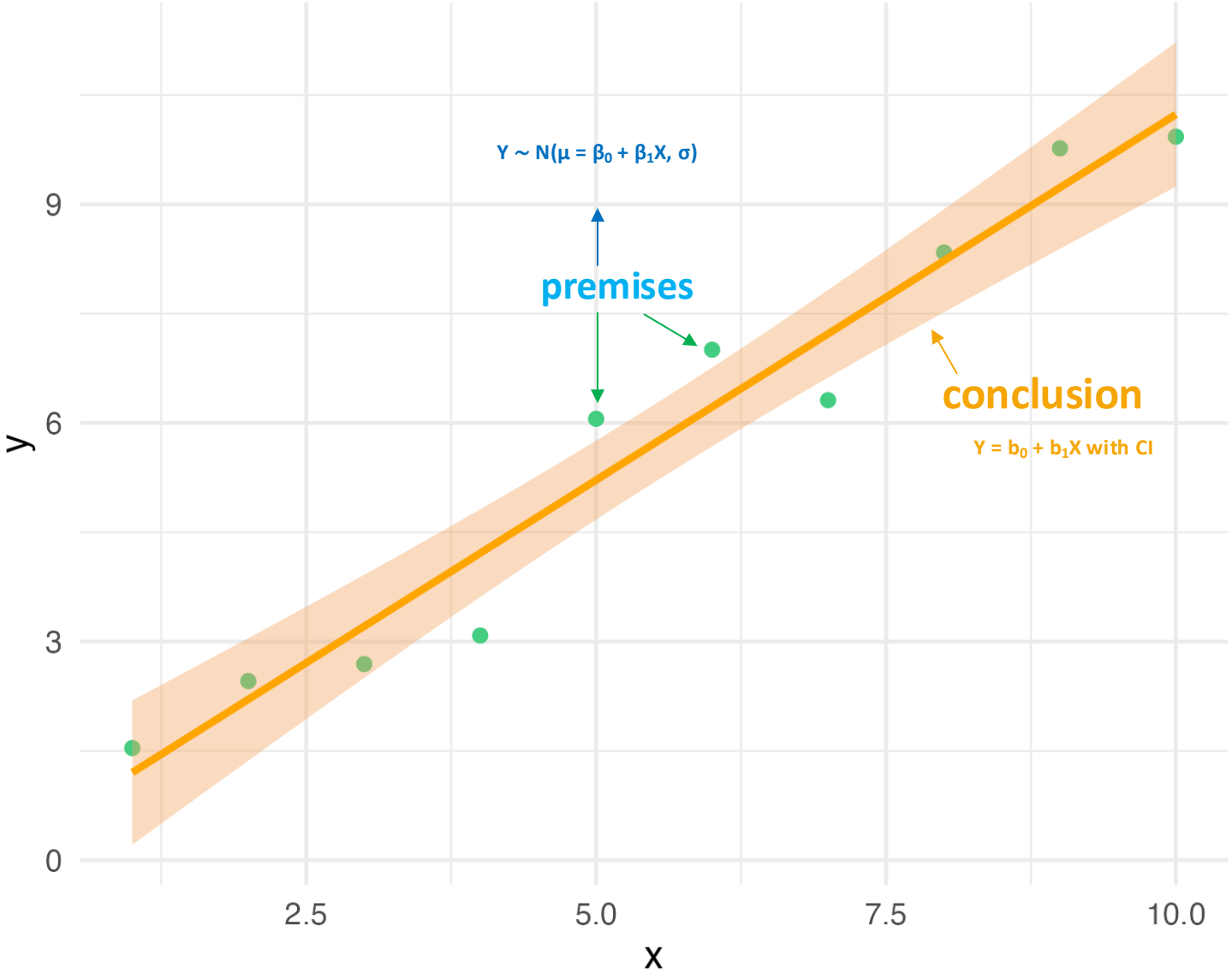
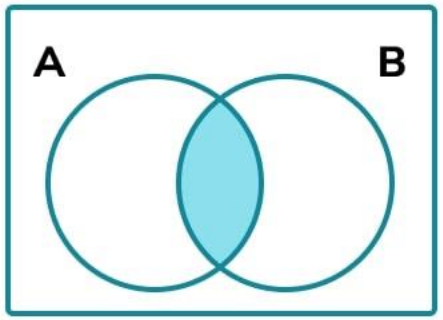
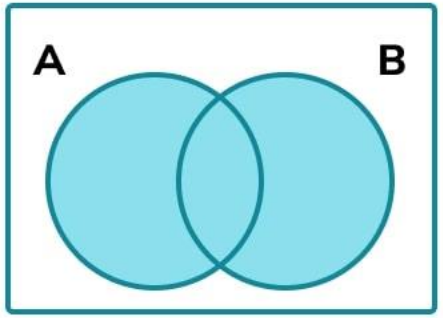
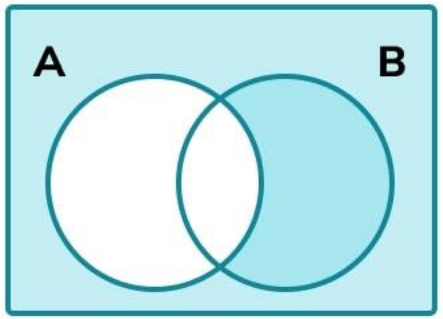
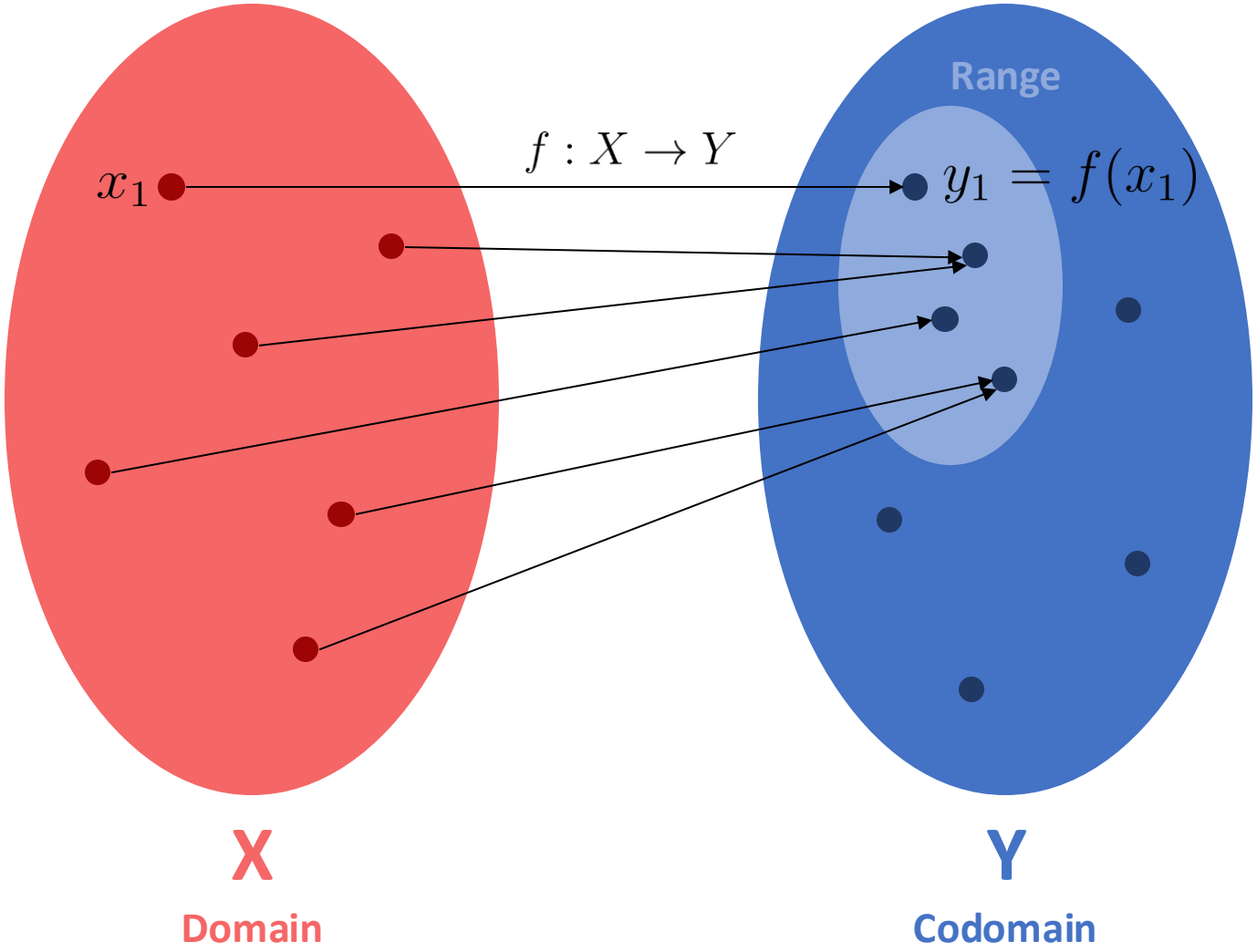
Galileo Galilei (1564-1642)
the “father of modern science”





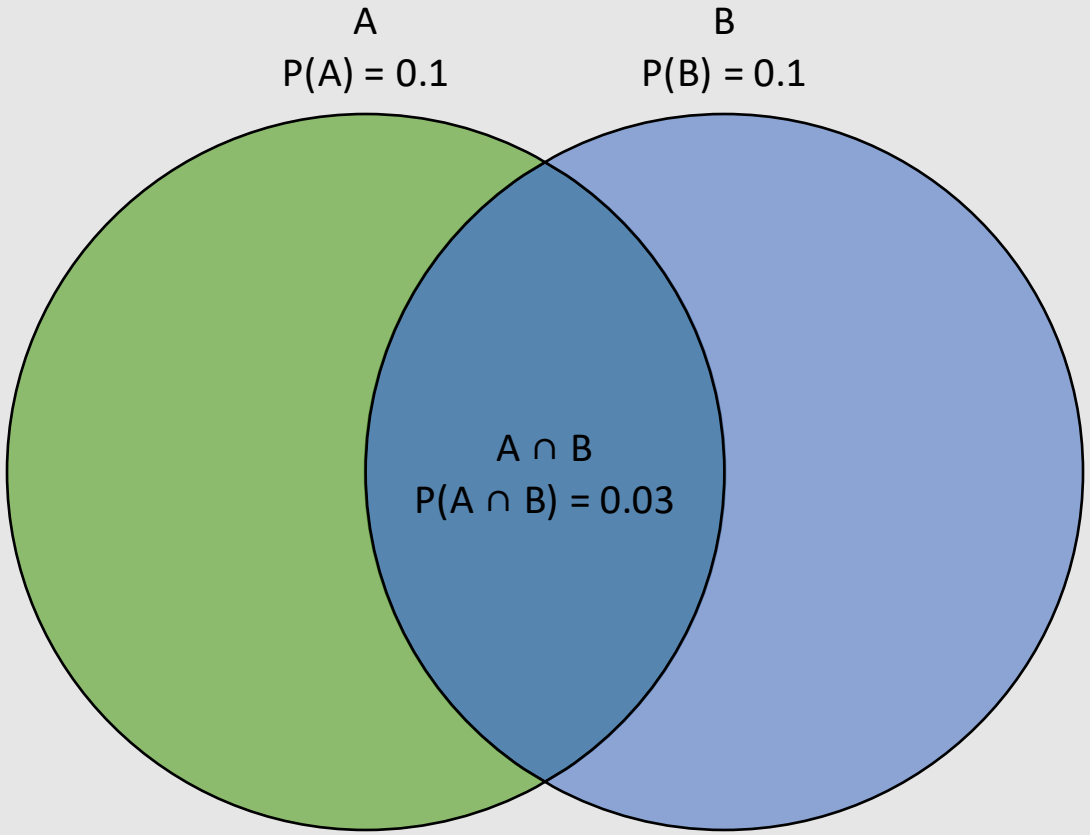
$$
\begin{array}{ll}
1. & \{(x, y)\} \\
2. & Y \sim N(\mu = \beta_0 + \beta_1 \cdot X, \sigma = \sigma) \\
\hline
\therefore & f(Y \sim N(\mu = b_0 + b_1 \cdot X, \sigma = s))
\end{array}
$$

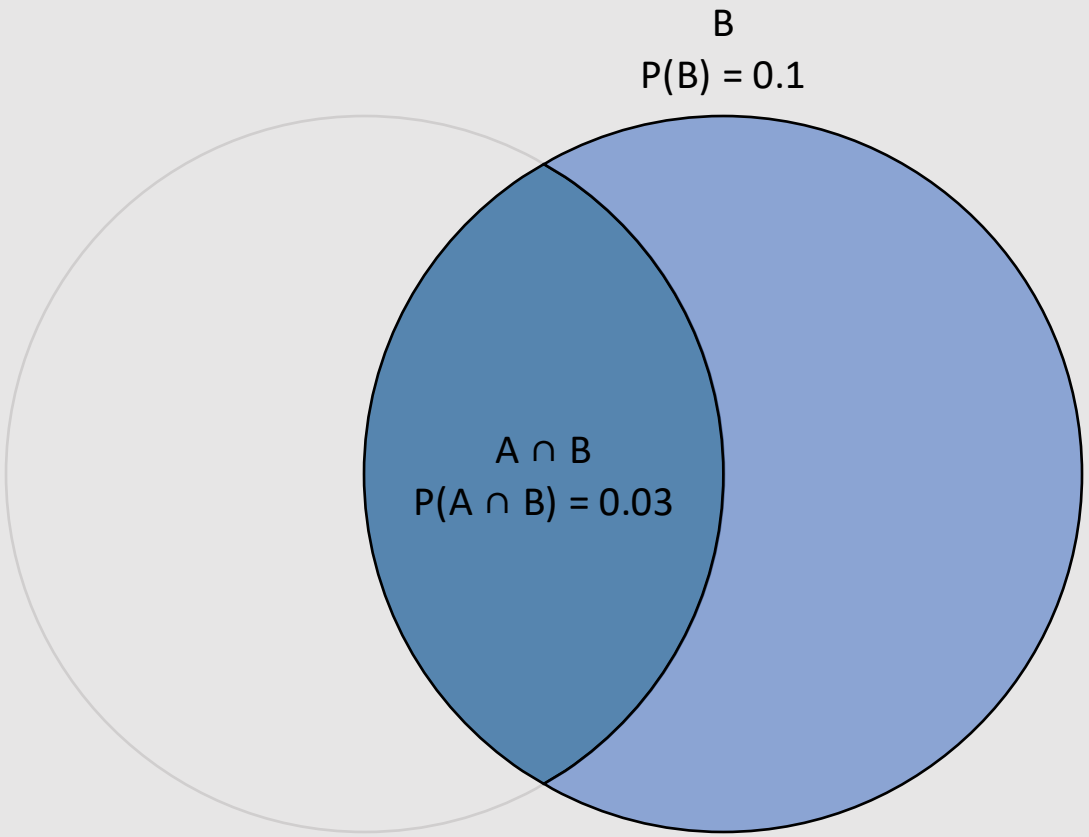
$A \cap B$	'A and B' The intersection of A and B. The elements in both sets A and B.	
$A \cup B$	'A or B' The union of A or B. Any element in set A or set B.	
A'	'Not A' The complement of A. Any element not in A.	

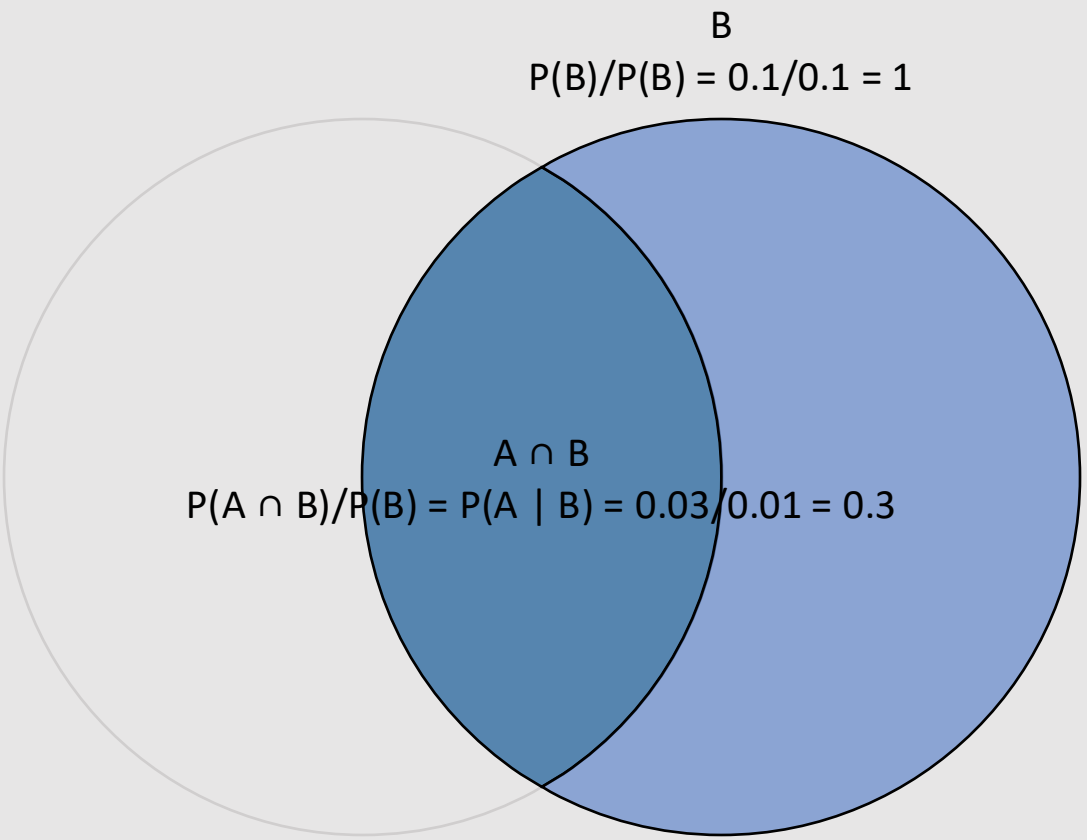

Range
$f : X \to Y$
x_1
$y_1 = f(x_1)$
X
Domain
Y
Codomain

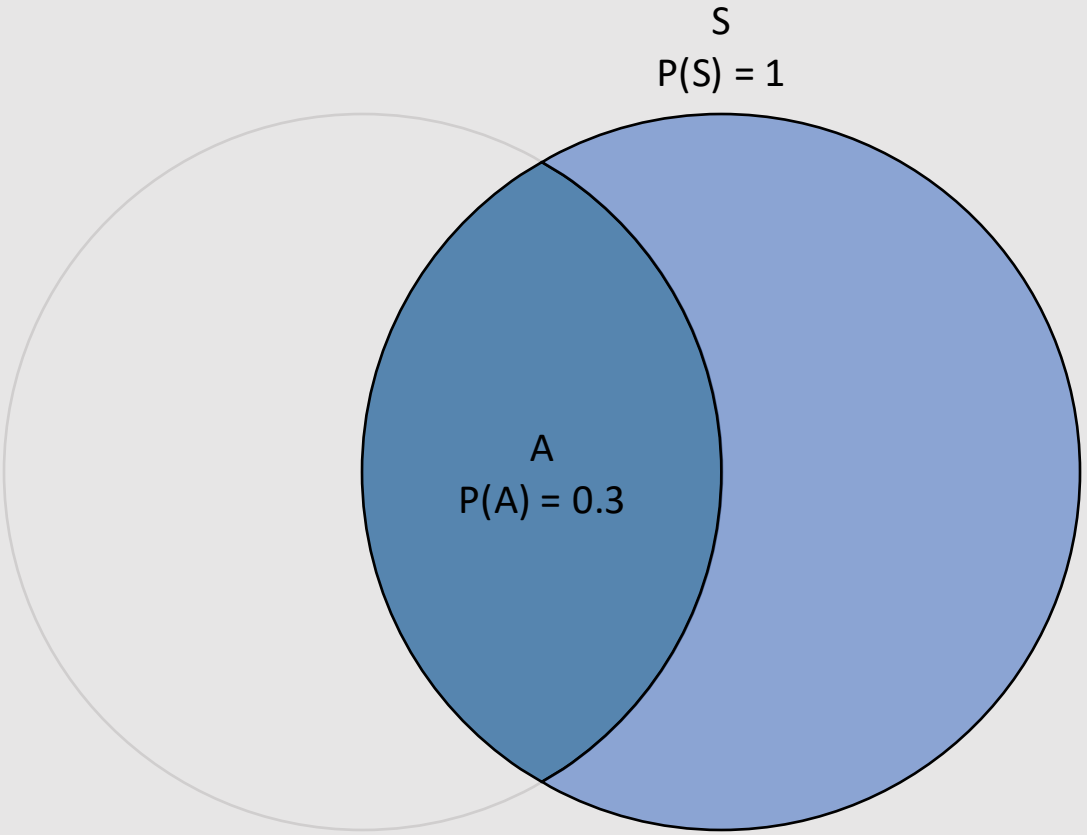
S

P(S) = 1



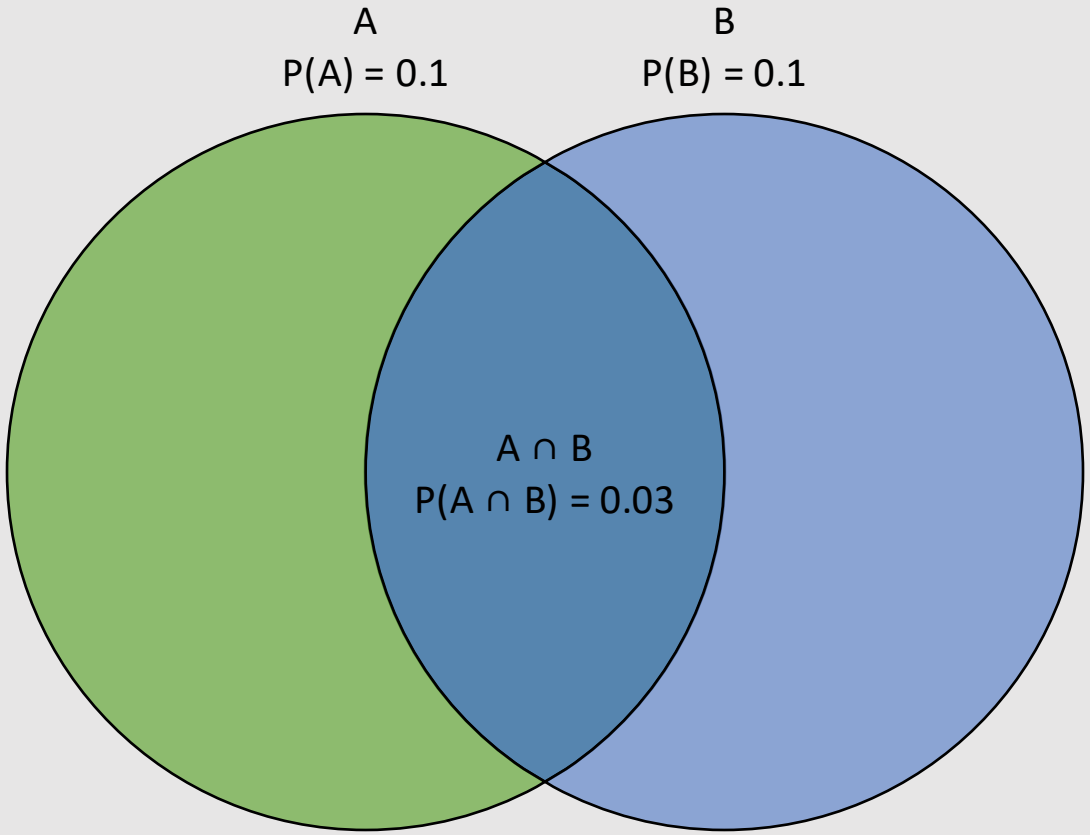

B
P(B) = 0.1
A ∩ B
P(A ∩ B) = 0.03


B
P(B)/P(B) = 0.1/0.1 = 1
A ∩ B
P(A ∩ B)/P(B) = P(A | B) = 0.03/0.01 = 0.3


S
P(S) = 1
A
P(A) = 0.3

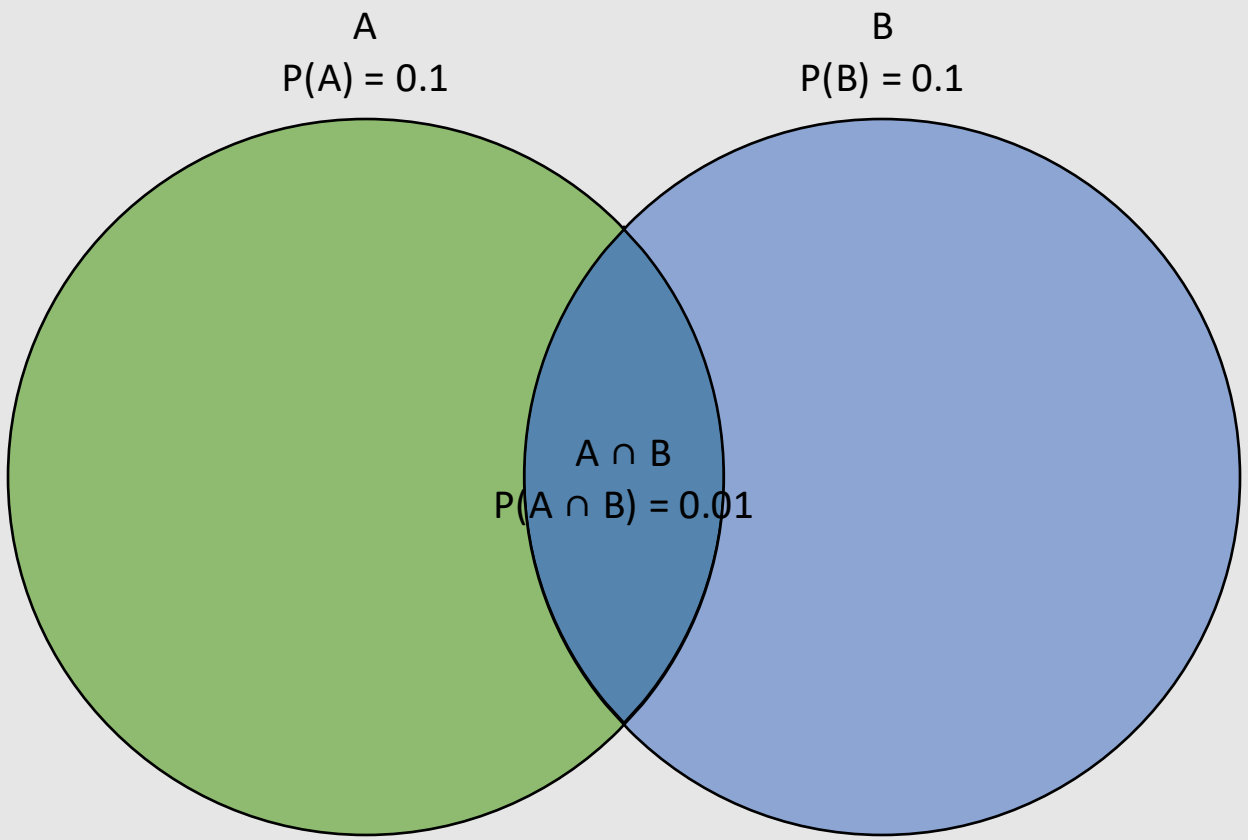
S

P(S) = 1



S

P(S) = 1

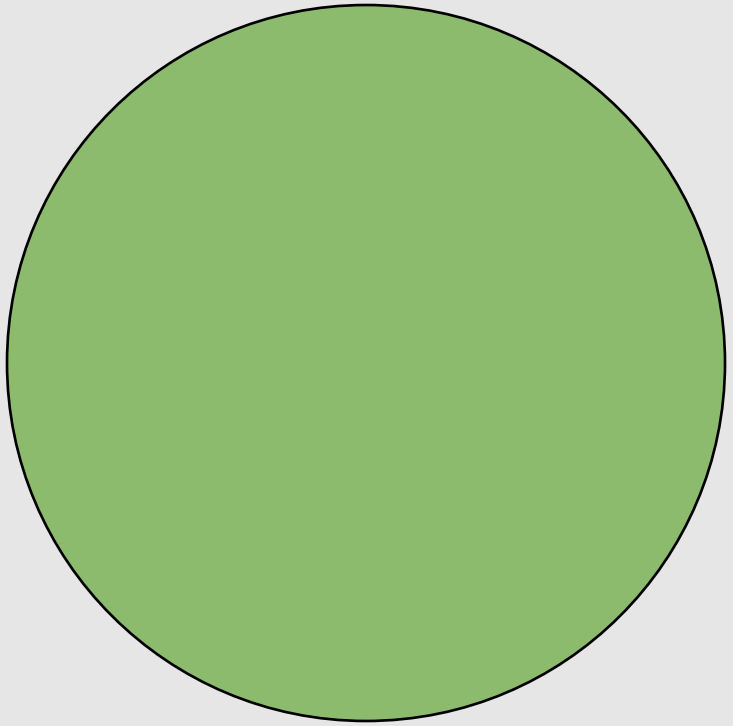


S

$P(S) = 1$

A

$P(A) = 0.1$

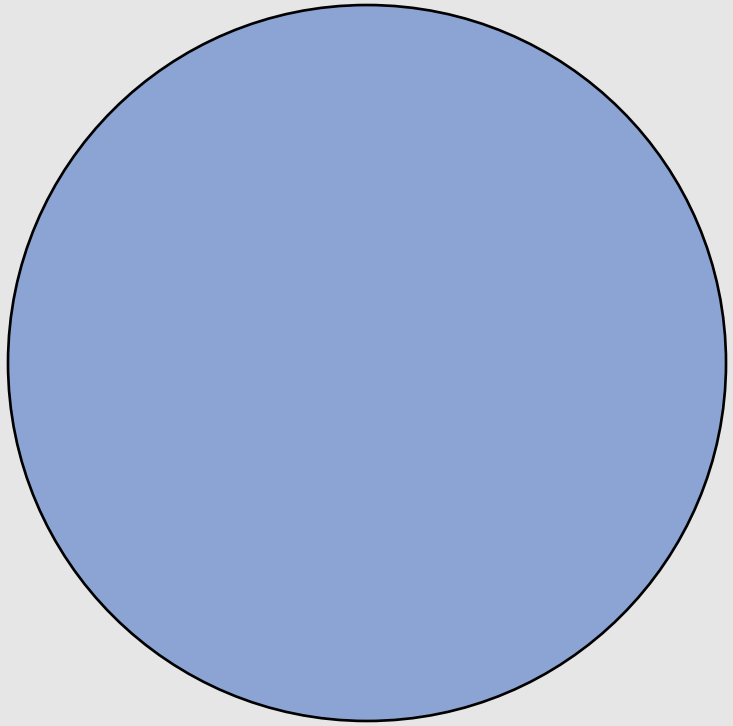


$A \cap B$

$P(A \cap B) = 0$

B

$P(B) = 0.1$

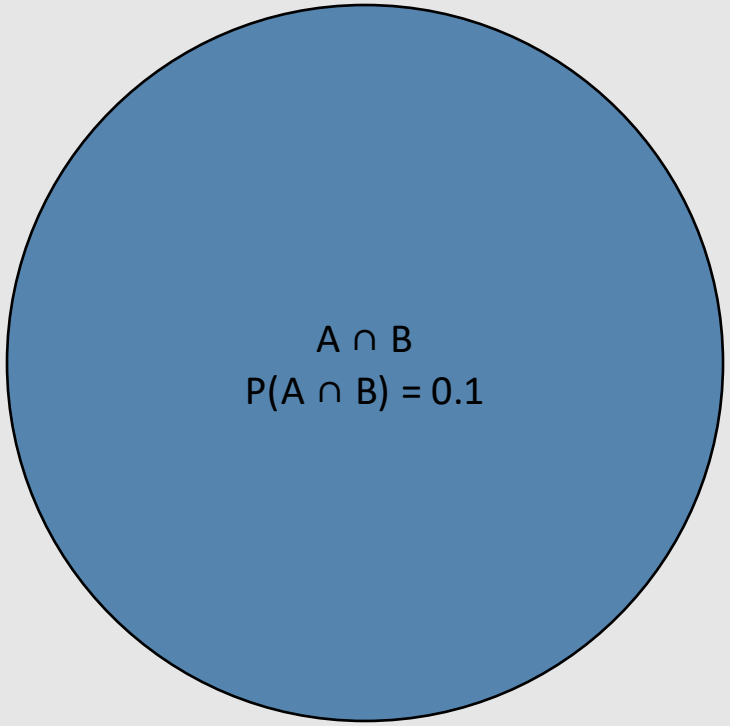


S

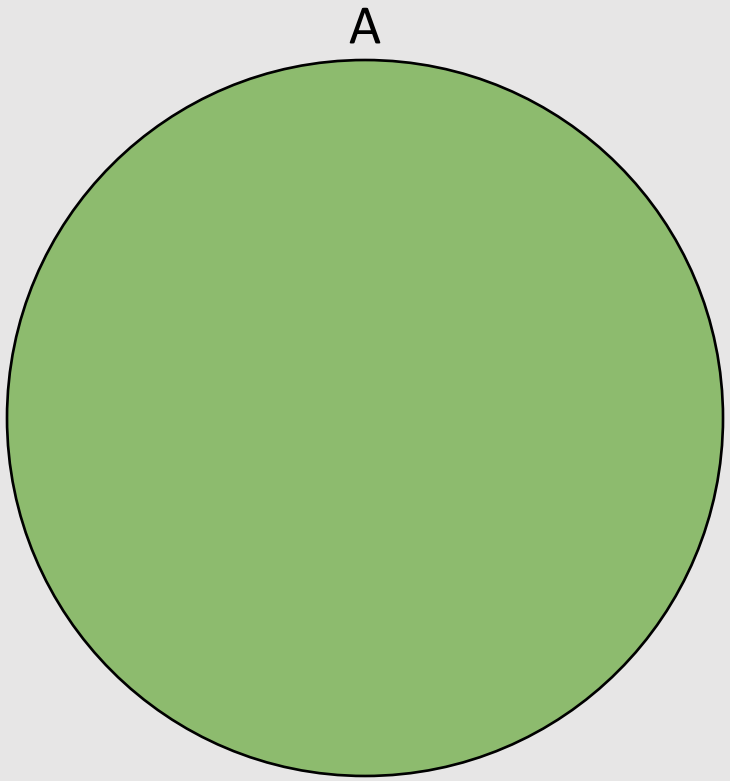
$P(S) = 1$

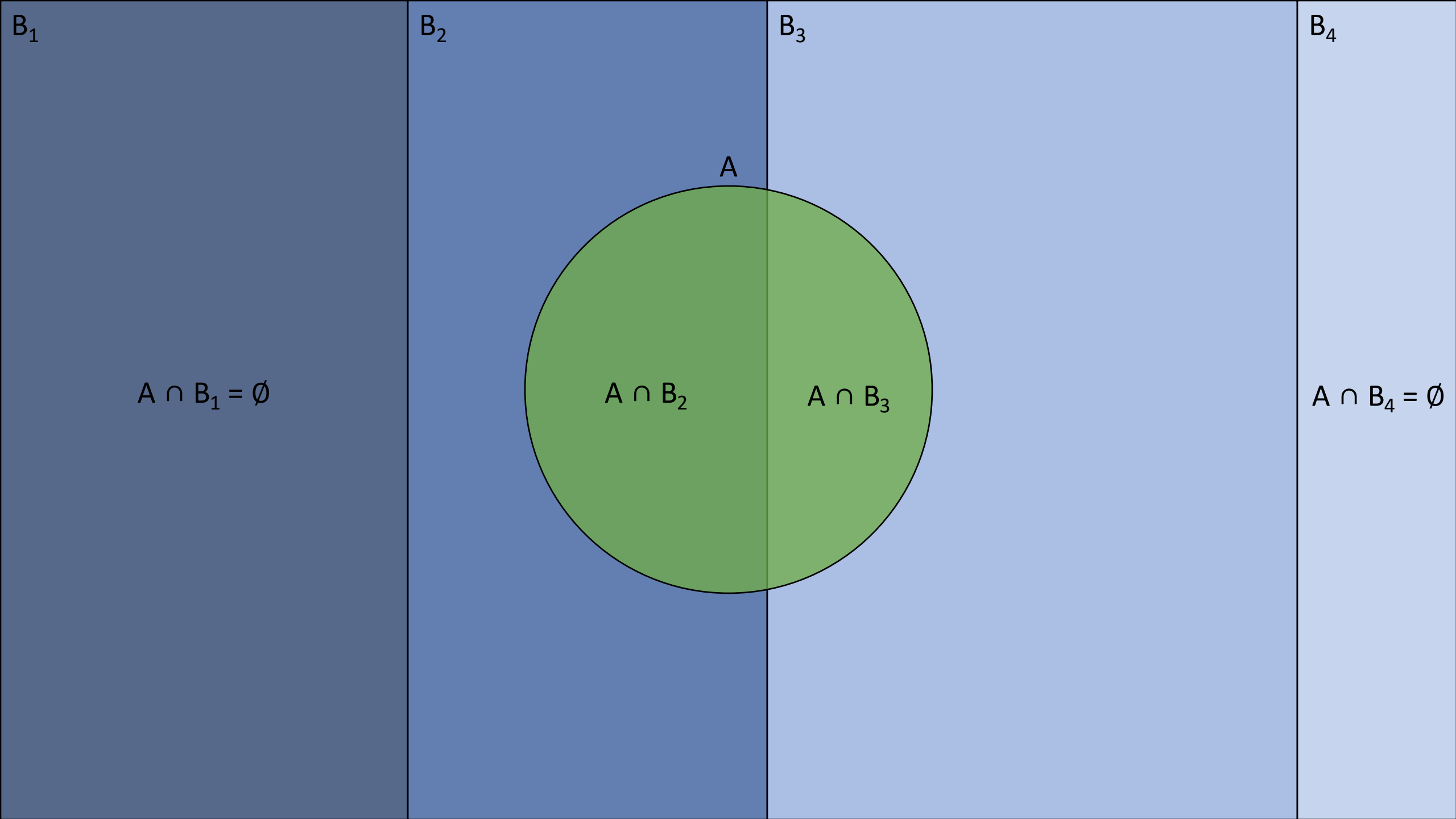
$A = B$

$P(A) = P(B) = 0.1$



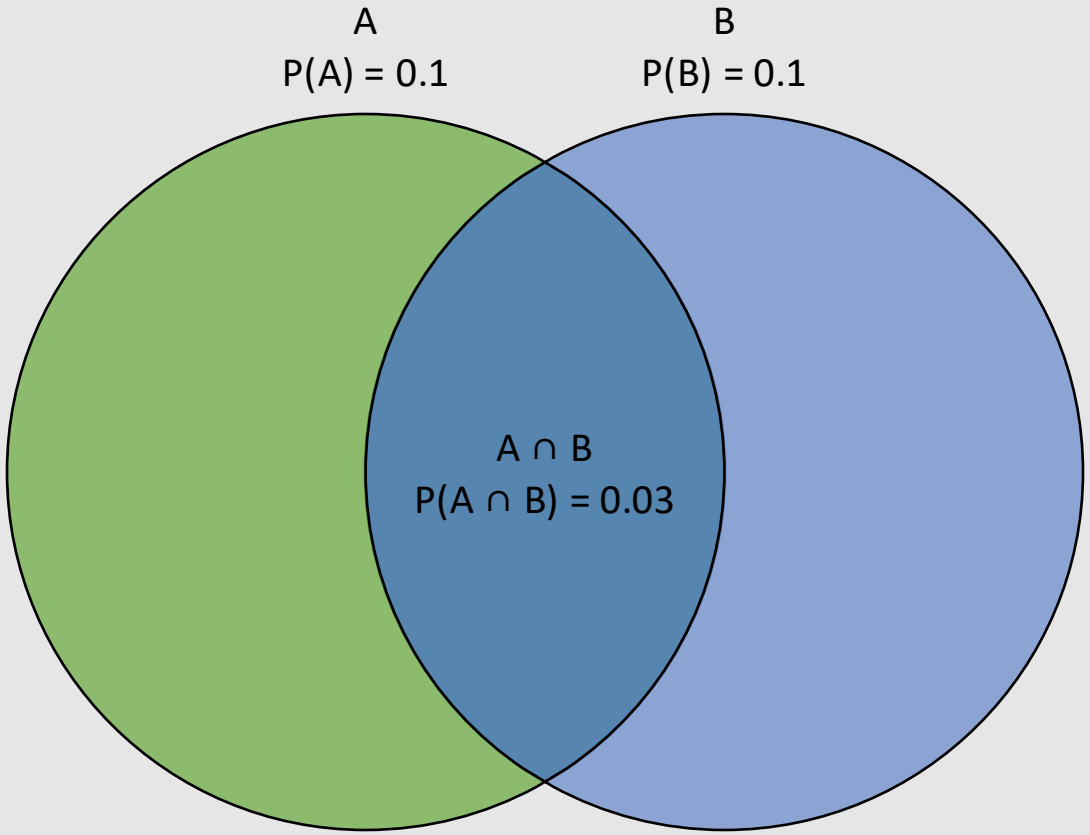
S




B_1
B_2
B_3
B_4
A
$A \cap B_1 = \emptyset$
$A \cap B_2$
$A \cap B_3$
$A \cap B_4 = \emptyset$

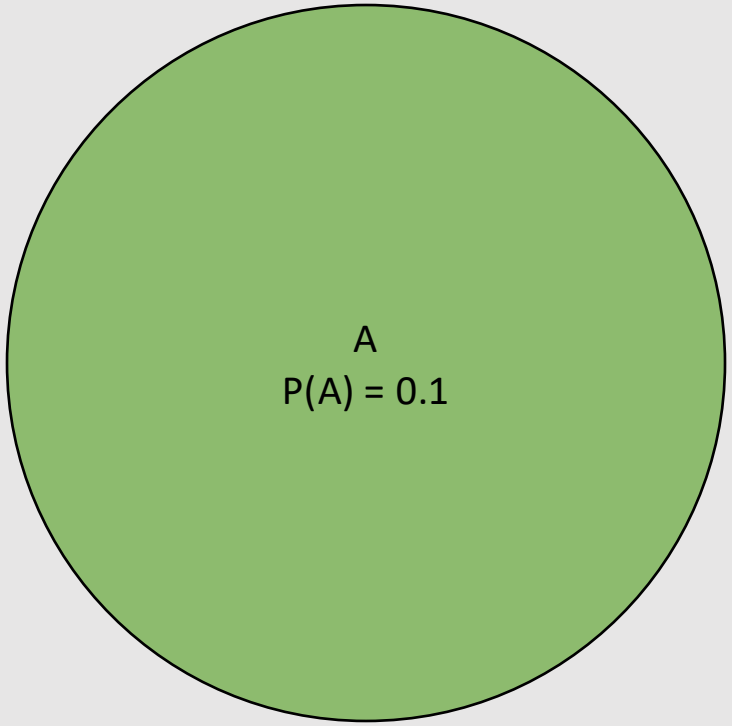
S

P(S) = 1



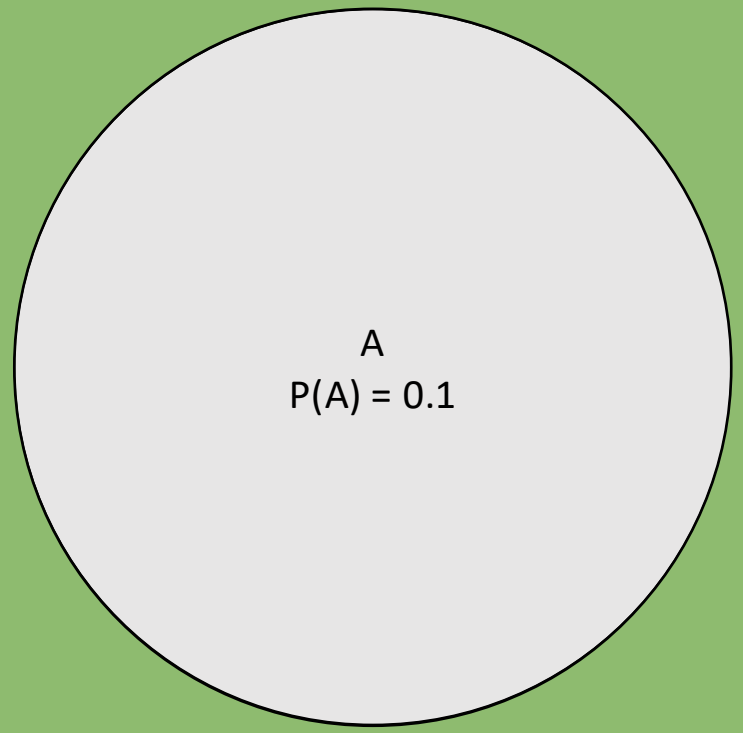
S

P(S) = 1



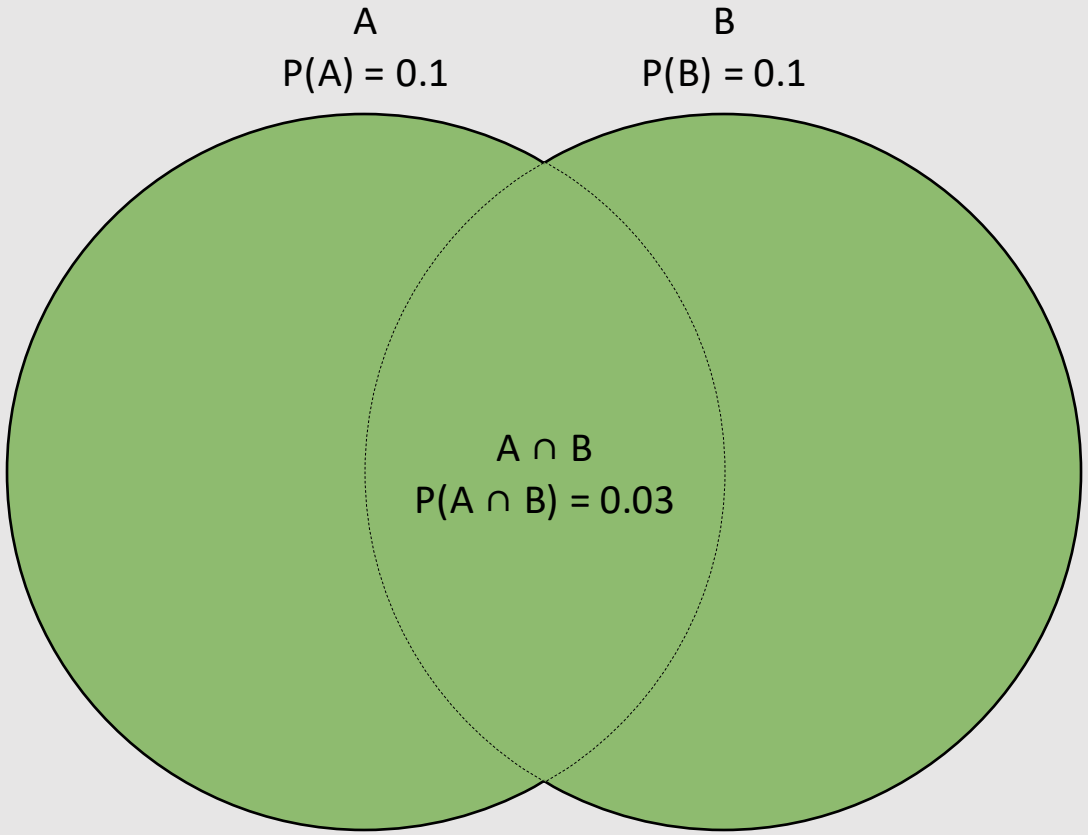
S

P(S) = 1



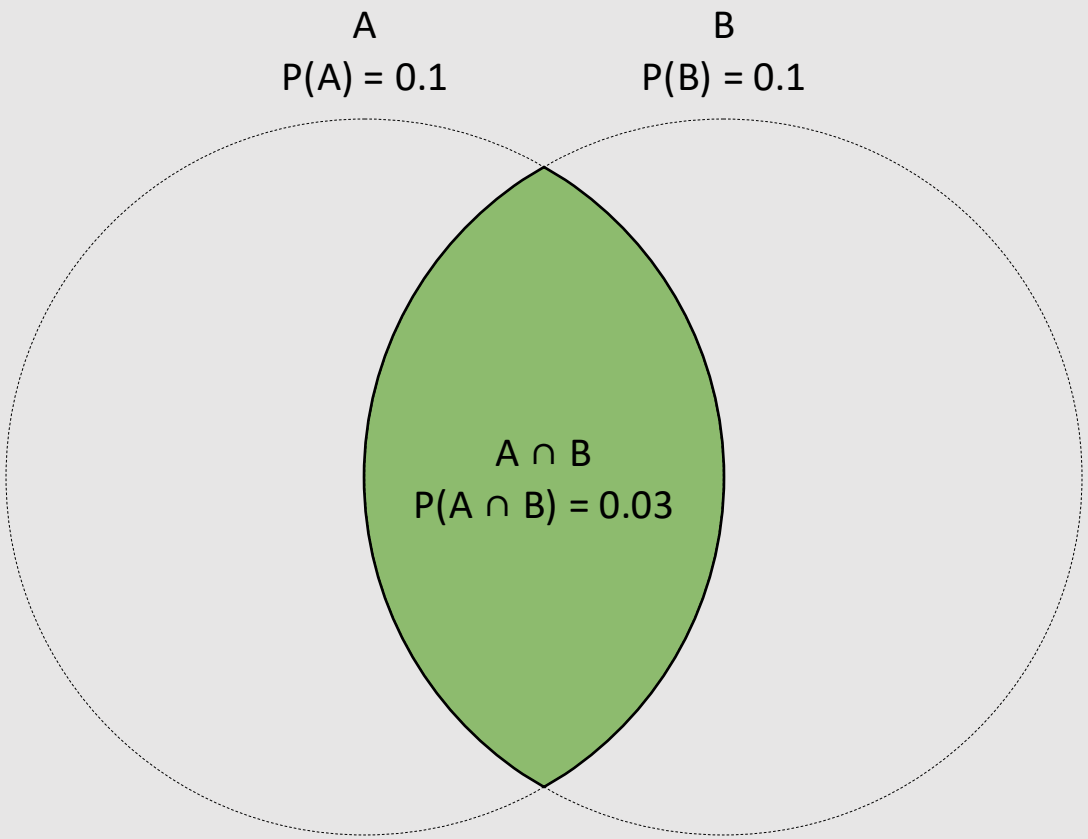
S

P(S) = 1



S

P(S) = 1



$$\underbrace{P(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{P(H)}^{\text{PRIOR}} \cdot \overbrace{P(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{P(E)}_{\text{EVIDENCE}}}$$

$$\underbrace{P(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{P(H)}^{\text{PRIOR}} \cdot \overbrace{P(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{P(H) \cdot P(E \mid H) + P(H') \cdot P(E \mid H')}_{\text{EVIDENCE}}}$$

$$\underbrace{P(H_j \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{P(H_j)}^{\text{PRIOR}} \cdot \overbrace{P(E \mid H_j)}^{\text{LIKELIHOOD}}}{\underbrace{\sum_{i=1}^{n} P(H_i) \cdot P(E \mid H_i)}_{\text{EVIDENCE}}}$$

$$\{H_1, H_2, \ldots, H_n\} \text{ is a partition of H}$$

POSTERIOR PRIOR LIKELIHOOD

$$P(H \mid E) \propto P(H) \cdot P(E \mid H)$$

$$\underbrace{P(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{P(H)}^{\text{PRIOR}} \cdot \overbrace{P(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{P(E)}_{\text{EVIDENCE}}}$$

$$\underbrace{p(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{p(H)}^{\text{PRIOR}} \cdot \overbrace{p(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{p(E)}_{\text{EVIDENCE}}}$$

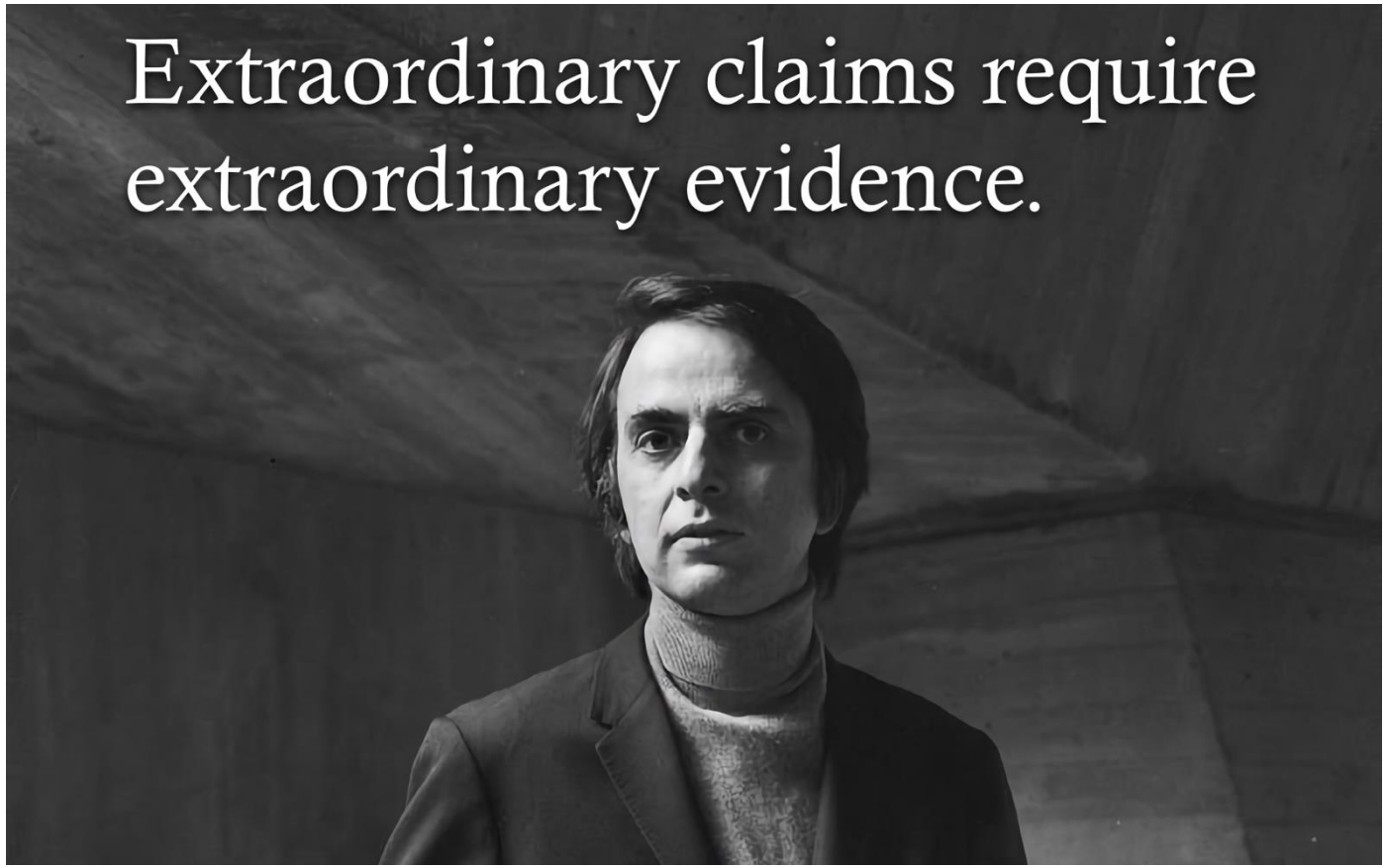
$$\underbrace{p(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{p(H)}^{\text{PRIOR}} \cdot \overbrace{p(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{k}_{\text{KNOWN NORMALIZING CONSTANT}}}$$

POSTERIOR PRIOR LIKELIHOOD

$$p(H \mid E) \propto p(H) \cdot p(E \mid H)$$

POSTERIOR PRIOR LIKELIHOOD

$$p(H \mid E) \propto p(H) \cdot p(E \mid H)$$



$$\underbrace{p(H \mid E)}_{\text{POSTERIOR}} = \frac{\overbrace{p(H)}^{\text{PRIOR}} \cdot \overbrace{p(E \mid H)}^{\text{LIKELIHOOD}}}{\underbrace{p(E)}_{\text{EVIDENCE}}}$$

LIKELIHOOD

$$p(H \mid E) = \frac{p(H) \cdot p(E \mid H)}{p(E)}$$

EVIDENCE

LIKELIHOOD

$$p(H \mid E) = \frac{p(H) \cdot p(E \mid H)}{p(E)}$$

UNKNOWN CONSTANT

LIKELIHOOD

$$p(H \mid E) = k(E) \cdot p(E \mid H)$$

UNKNOWN CONSTANT

LIKELIHOOD FUNCTION

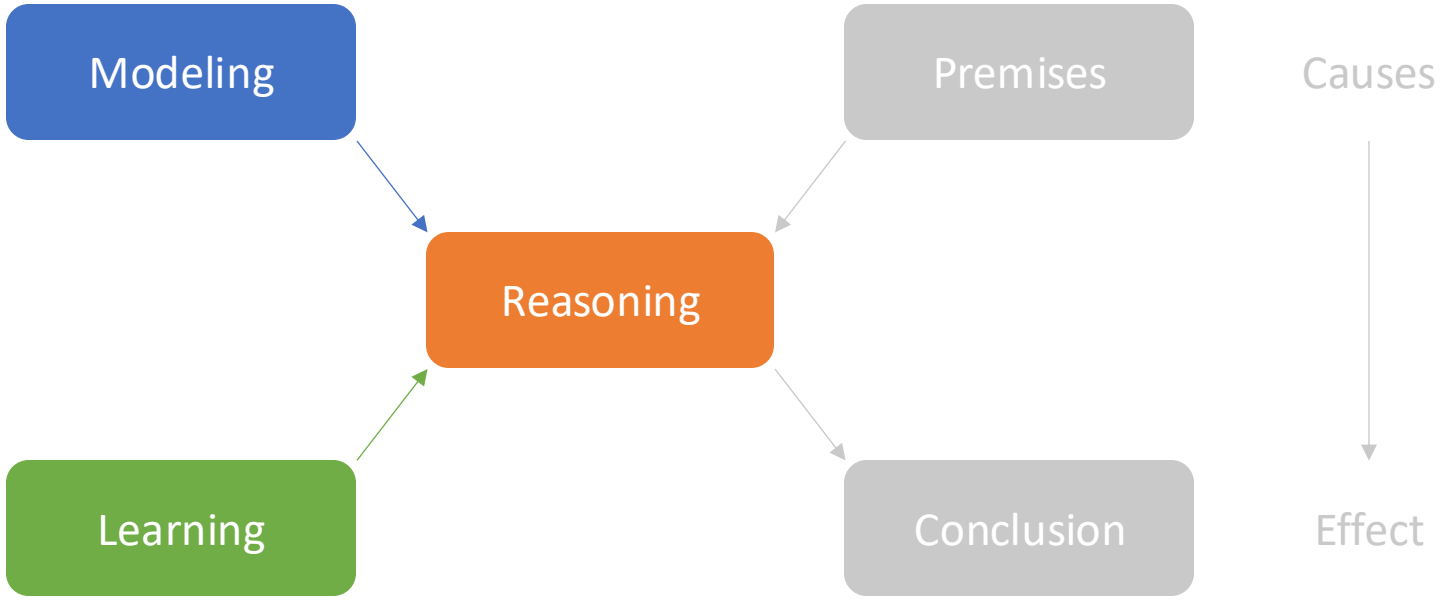
LIKELIHOOD

$$\mathcal{L}(H \mid E) = k(E) \cdot p(E \mid H)$$

UNKNOWN CONSTANT

LIKELIHOOD FUNCTION LIKELIHOOD

$$\mathcal{L}(H \mid E) \propto p(E \mid H)$$


Modeling
Premises
Causes
Reasoning
Learning
Conclusion
Effect

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

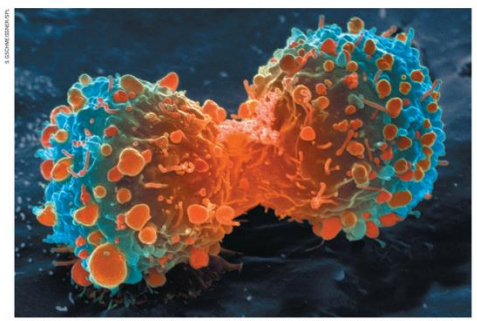
Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

COMMENT



S. SOCHMAN-SENAGOR/CPA

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▶

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

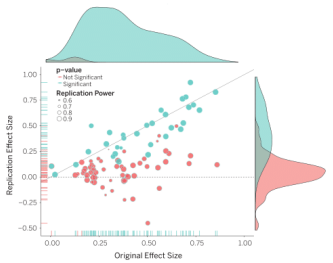
RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (*r*) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know. ■

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: nosek@virginia.edu

Cite this article as Open Science Collaboration, *Science* 349, aac4716 (2015). DOI: 10.1126/science.aac4716

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

It has been claimed and demonstrated that many (and possibly most) of the conclusions drawn from biomedical research are probably false[1]. A central cause for this important problem is that researchers must publish in order to succeed, and publishing is a highly competitive enterprise, with certain kinds of findings more likely to be published than others. Research that produces novel results, statistically significant results (that is, typically $p < 0.05$) and seemingly 'clean' results is more likely to be published[2,3]. As a consequence, researchers have strong incentives to engage in research practices that make their findings publishable quickly, even if those practices reduce the likelihood that the findings reflect a true effect (that is, non-null) effect[4]. Such practices include using flexible study designs and flexible statistical analyses and running small studies with low statistical power[1,5]. A simulation of genetic association studies showed that a typical dataset would generate at least one false positive result almost 97% of the time[6], and two efforts to replicate promising findings in biomedicine reveal replication rates of 25% or less[7,8]. Given that these publishing biases are pervasive across scientific practice, it is possible that false positives heavily contaminate the neuroscience literature as well, and this problem may affect at least as much, if not even more so, the most prominent journals[9,10].

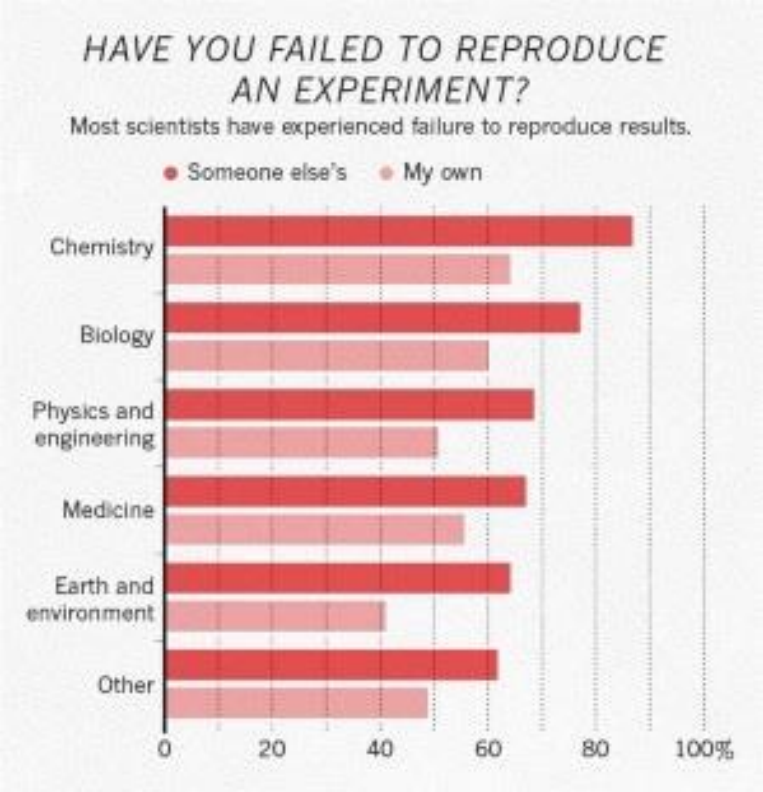
Here, we focus on one major aspect of the problem: low statistical power. The relationship between study power and the veracity of the resulting finding is under-appreciated. Low statistical power (because of low sample size of studies, small effects or both) negatively affects the likelihood that a nominally statistically significant finding actually reflects a true effect. We discuss the problems that arise when low-powered research designs are pervasive. In general, these problems can be divided into two categories. The first concerns problems that are mathematically expected to arise even if the research conducted is otherwise perfect: in other words, when there are no biases that tend to create statistically significant (that is, 'positive') results that are spurious. The second category concerns problems that reflect biases that tend to co-occur with studies of low power or that become worse in small, underpowered studies. We next empirically show that statistical power is typically low in the field of neuroscience by using evidence from a range of subfields within the neuroscience literature. We illustrate that low statistical power is an endemic problem in neuroscience and discuss the implications of this for interpreting the results of individual studies.

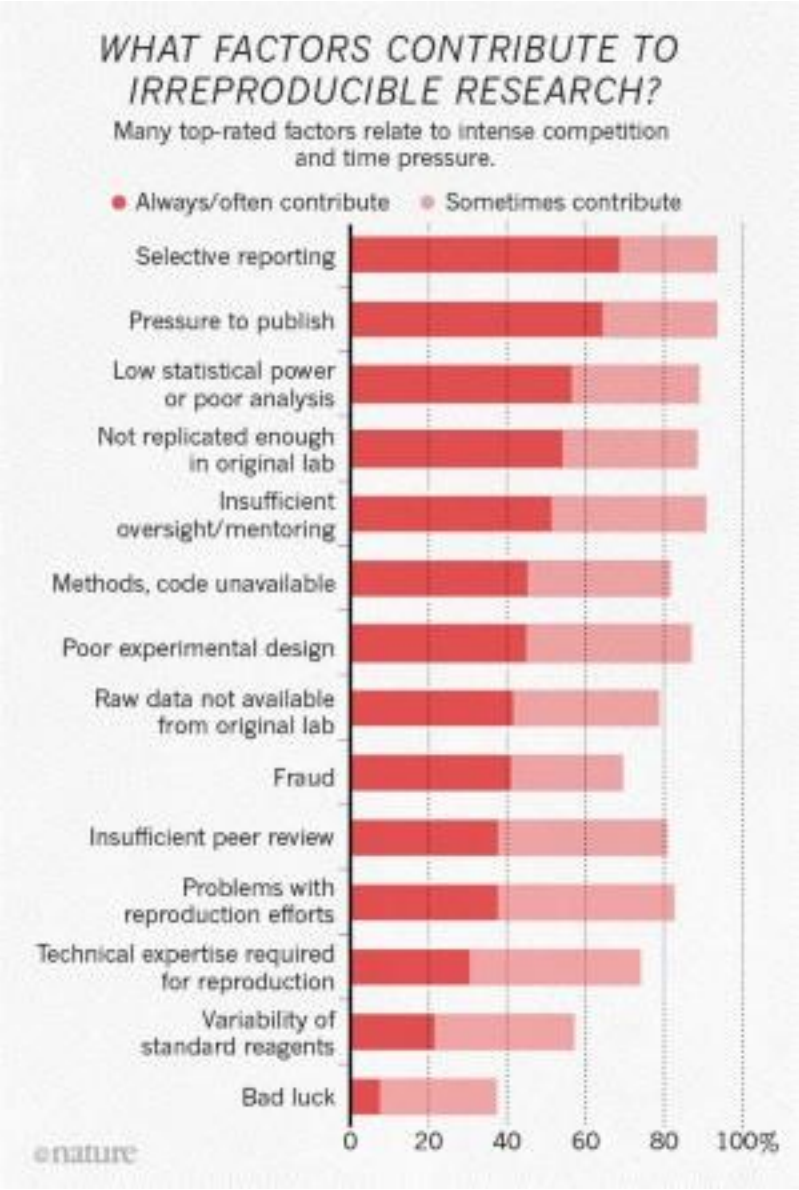
Low power in the absence of other biases

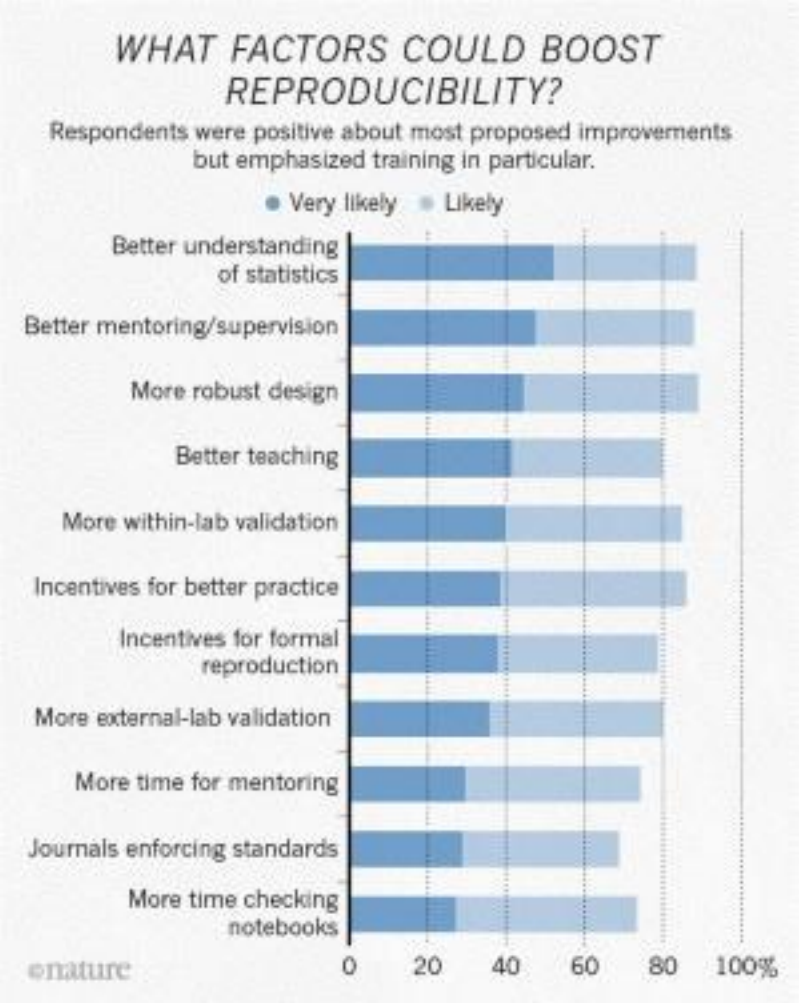
Three main problems contribute to producing unreliable findings in studies with low power, even when all other research practices are ideal. They are: the low probability of finding true effects; the low positive predictive value (PPV); see BOX 1 for definitions of key statistical terms) when an effect is claimed; and an exaggerated estimate of the magnitude of the effect when a true effect is discovered. Here, we discuss these problems in more detail.

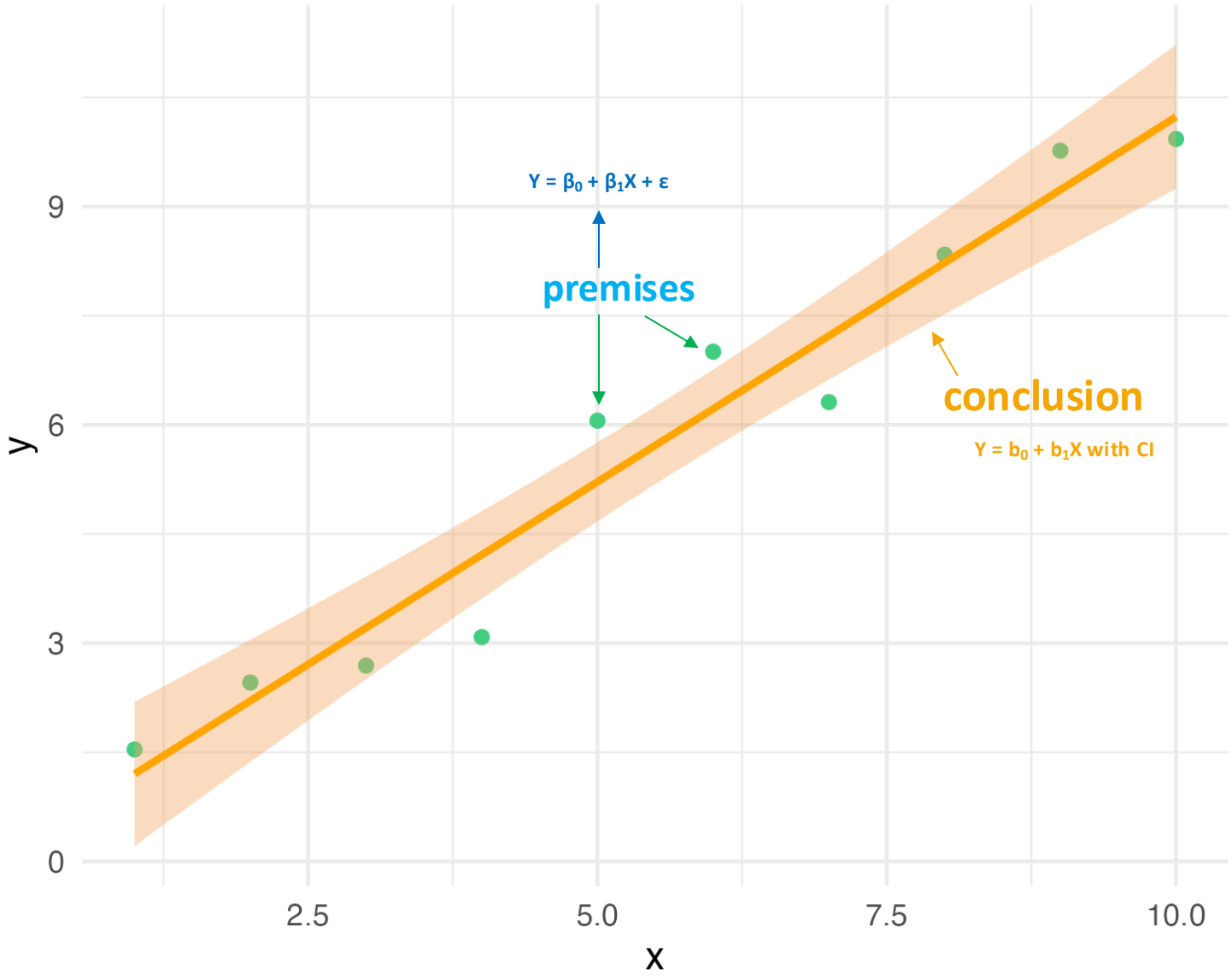
[1]School of Experimental Psychology, University of Bristol, Bristol, BS8 1TU, UK. [2]School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK. [3]Stanford University School of Medicine, Stanford, California 94305, USA. [4]Department of Psychology, University of Virginia, Charlottesville, Virginia 22904, USA. [5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK. [6]School of Physiology and Pharmacology, University of Bristol, Bristol, BS8 1TD, UK. Correspondence to M.R.M. e-mail: marcus.munafo@bristol.ac.uk

doi:10.1038/nrn3475
Published online 10 April 2013
Corrected online 15 April 2013


HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.
Someone else's
My own
Chemistry
Biology
Physics and engineering
Medicine
Earth and environment
Other
0
20
40
60
80
100%


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
20
40
60
80
100%
©nature


WHAT FACTORS COULD BOOST REPRODUCIBILITY?
Respondents were positive about most proposed improvements but emphasized training in particular.
Very likely
Likely
Better understanding of statistics
Better mentoring/supervision
More robust design
Better teaching
More within-lab validation
Incentives for better practice
Incentives for formal reproduction
More external-lab validation
More time for mentoring
Journals enforcing standards
More time checking notebooks
0
20
40
60
80
100%
©nature


Y = β0 + β1X + ε
premises
conclusion
Y = b0 + b1X with CI
9
6
3
0
y
2.5
5.0
7.5
10.0
x

		Deductive reasoning	Inductive reasoning
P_1 P_2 $\vdots$ P_n — C	E — H	$E \models H$ $E \not\models H$	$P(H \mid E)$

$$E \equiv P_1 \wedge P_2 \wedge \ldots \wedge P_n$$

		Deductive reasoning	Inductive reasoning
P_1 P_2 $\vdots$ P_n — C	E — H	$E \models H$ $E \not\models H$	$P(H \mid E)$

$$E \equiv P_1 \wedge P_2 \wedge \ldots \wedge P_n$$

$$E \Rightarrow H$$

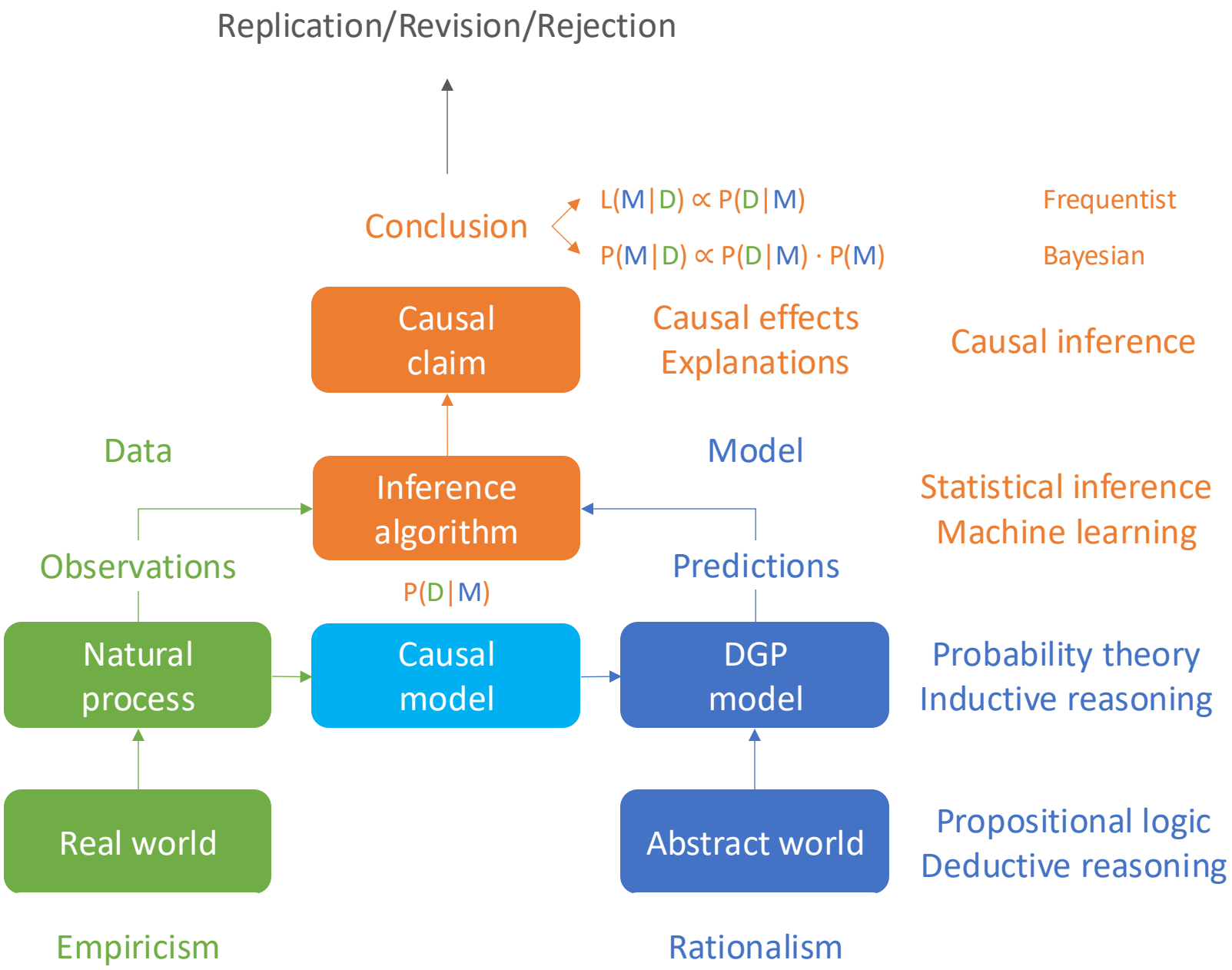
		Deductive reasoning	Inductive reasoning
P_1 P_2 $\vdots$ P_n — C	E — H	$V(H \mid E)$	$P(H \mid E)$

$$E \equiv P_1 \wedge P_2 \wedge \ldots \wedge P_n$$

$$E \Rightarrow H$$


gettyimages
Credit: chang
157512755


gettyimages
Credit: chang
157512755


Replication/Revision/Rejection
Conclusion
L(M|D) ∝ P(D|M)
P(M|D) ∝ P(D|M) · P(M)
Frequentist
Bayesian
Causal claim
Causal effects
Explanations
Causal inference
Data
Model
Inference algorithm
Statistical inference
Machine learning
Observations
Predictions
P(D|M)
Natural process
Causal model
DGP model
Probability theory
Inductive reasoning
Real world
Abstract world
Propositional logic
Deductive reasoning
Empiricism
Rationalism