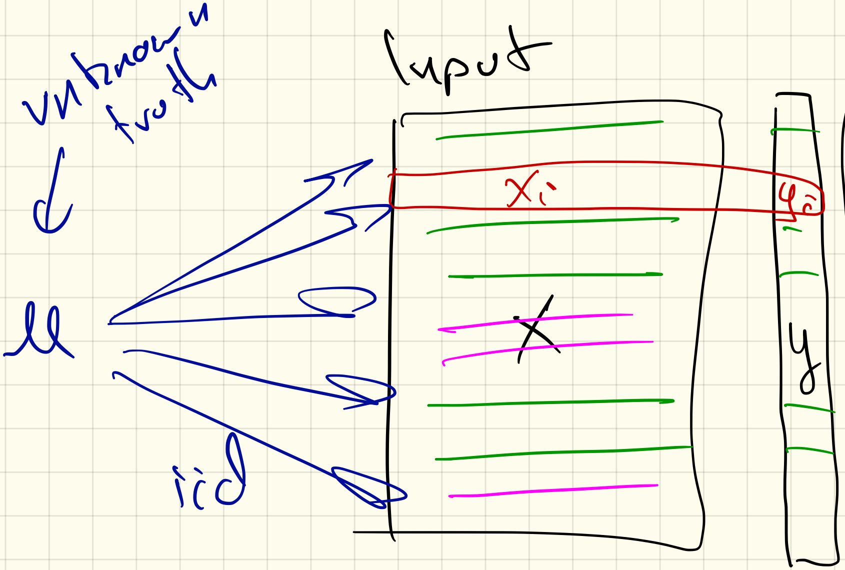
L15: Cross-Validation & p-values

Jeff M. Phillips

March 4, 2019



Split $(X, y) \rightarrow$ (random)

$\text{Test}(X, y) \cup \text{Train}(X, y)$

1. Build Model α on train

$$\alpha^*_{s,train}(s) = (X_{train}^T X_{train} + sI)^{-1} X_{train}^T y_{train}$$

2. Evaluate on Test

$$L_2(X_{test}, y_{test}, \alpha^*_{train}(s))$$

$$\|X_{test}\alpha^*_{train}(s) - y_{test}\|_2^2$$

Model : Ridge Regression

$$L_{RS}(X, y, \alpha) = \|X\alpha - y\|_2^2 + s\|\alpha\|_2^2$$

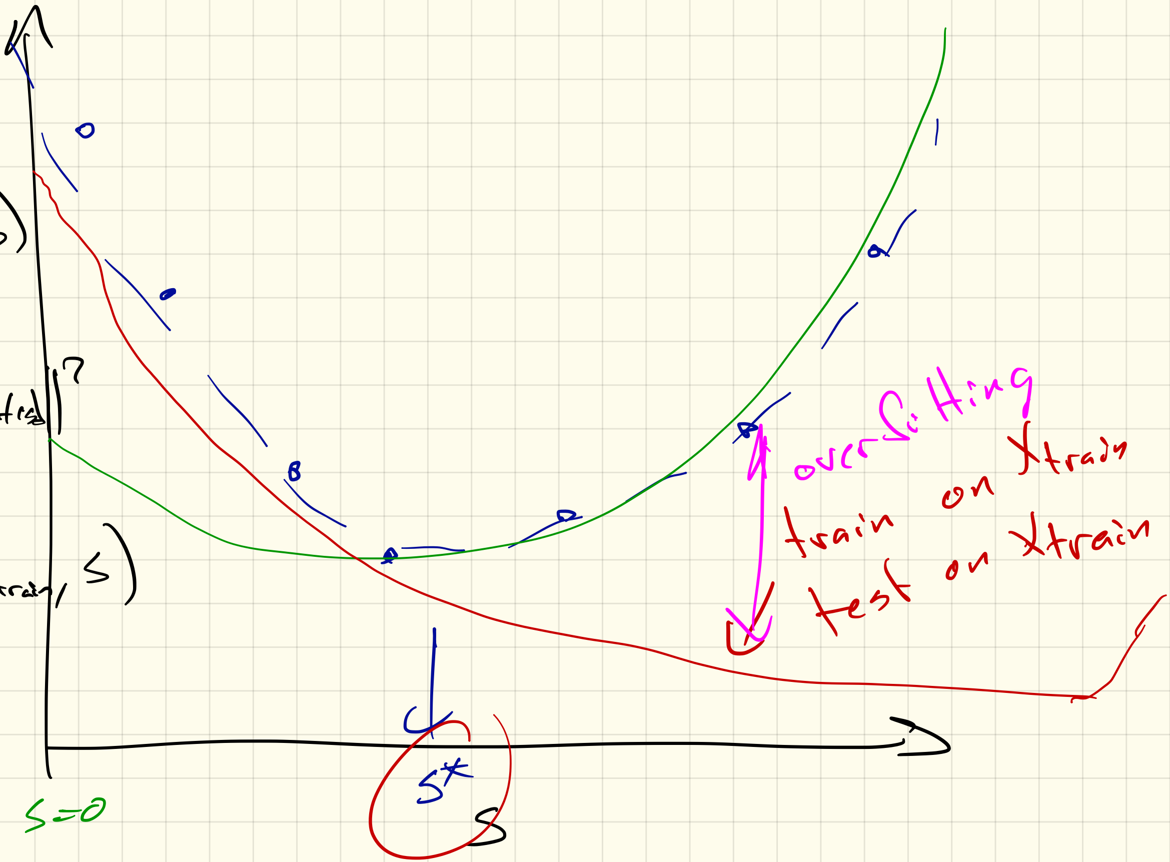
Cross-Validation

- Choose parameter (e.g. s)
- Measure Generalization

$$\mathrm{Cost}(X_{test}, s)$$

$$\|X_{test}\,\alpha^*(s) - y_{test}\|^2$$

$$\alpha^*(s) = \mathrm{soln}(X_{train}, s)$$



Generalization

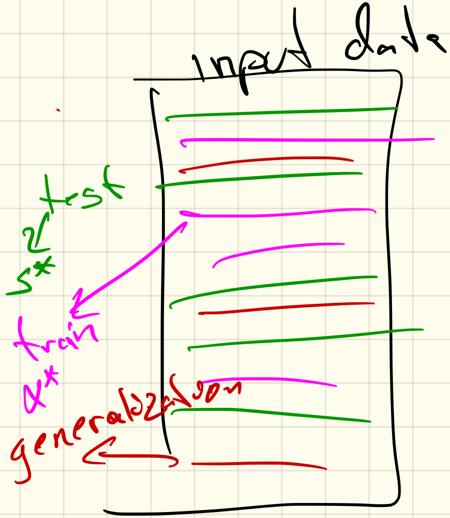
How well will Model α do on new data?

$$\frac{1}{|X_{test}|} \left\| X_{test}\, \alpha_{train}(s^*) - y_{test} \right\|_2^2$$

comes from test

$$E_{x' \sim \mu} \left[\left(y' - \underbrace{M(\alpha^*_{train}(s^*), x')}_{\hat{y}'} \right)^2 \right]$$

learned w/ X_{train} + X_{test}



How to split data

- 70% train 30% test
- 90% train 10% test

More → allows more complex model

better estimation CI bound

k-fold Cross-Validation

$X \rightarrow X_1, X_2, \ldots X_k$

for $j=1$ to k

train on $(X_1, X_2 \ldots X_{j-1}, X_{j+1} \ldots X_k)$

test on X_j

leave-one-out

$k = n = |X|$

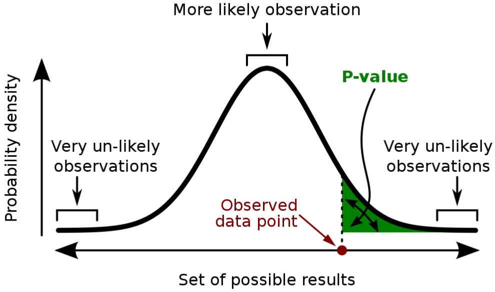
p-values

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

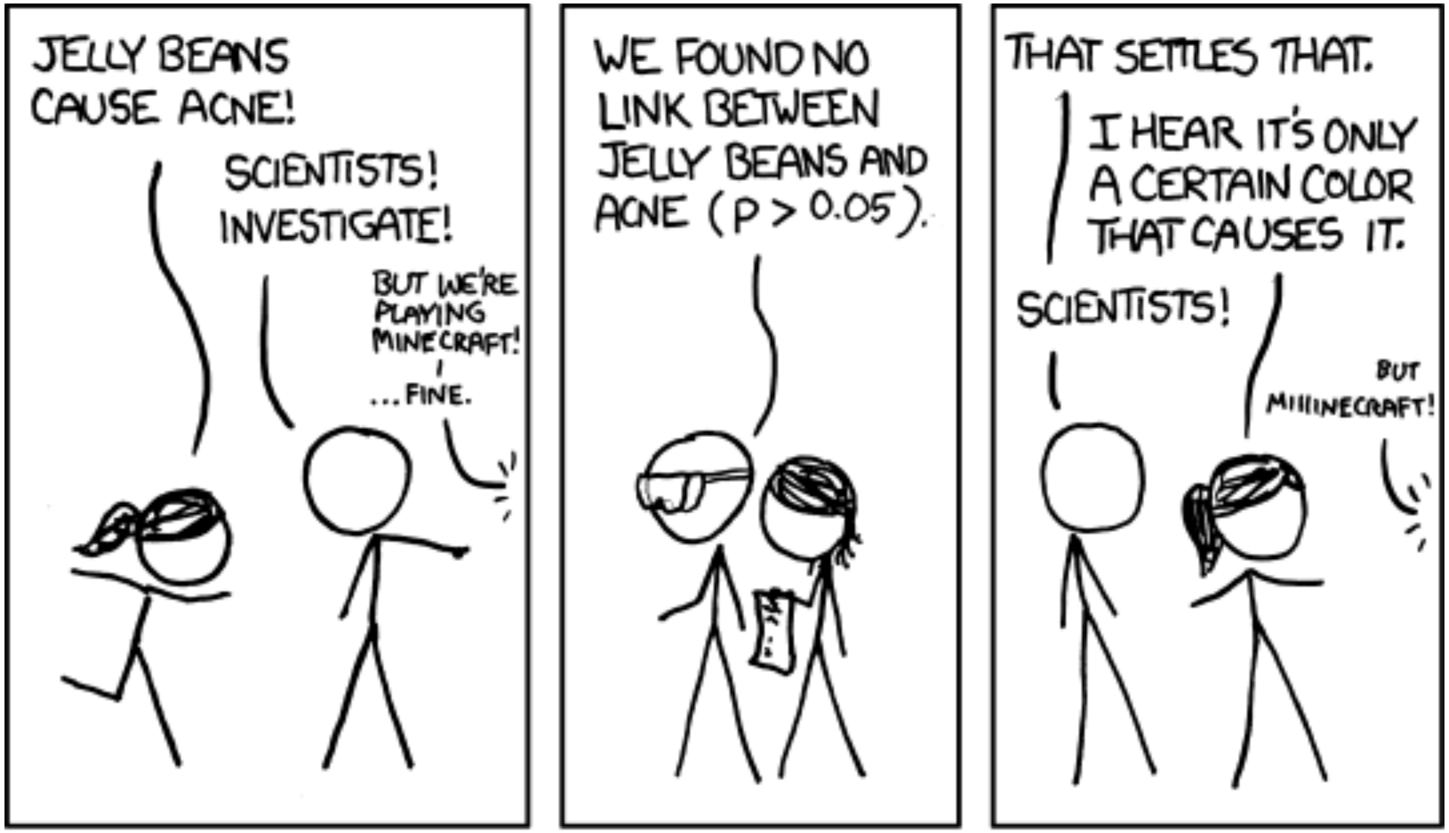
Using the p-value as a “score” is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

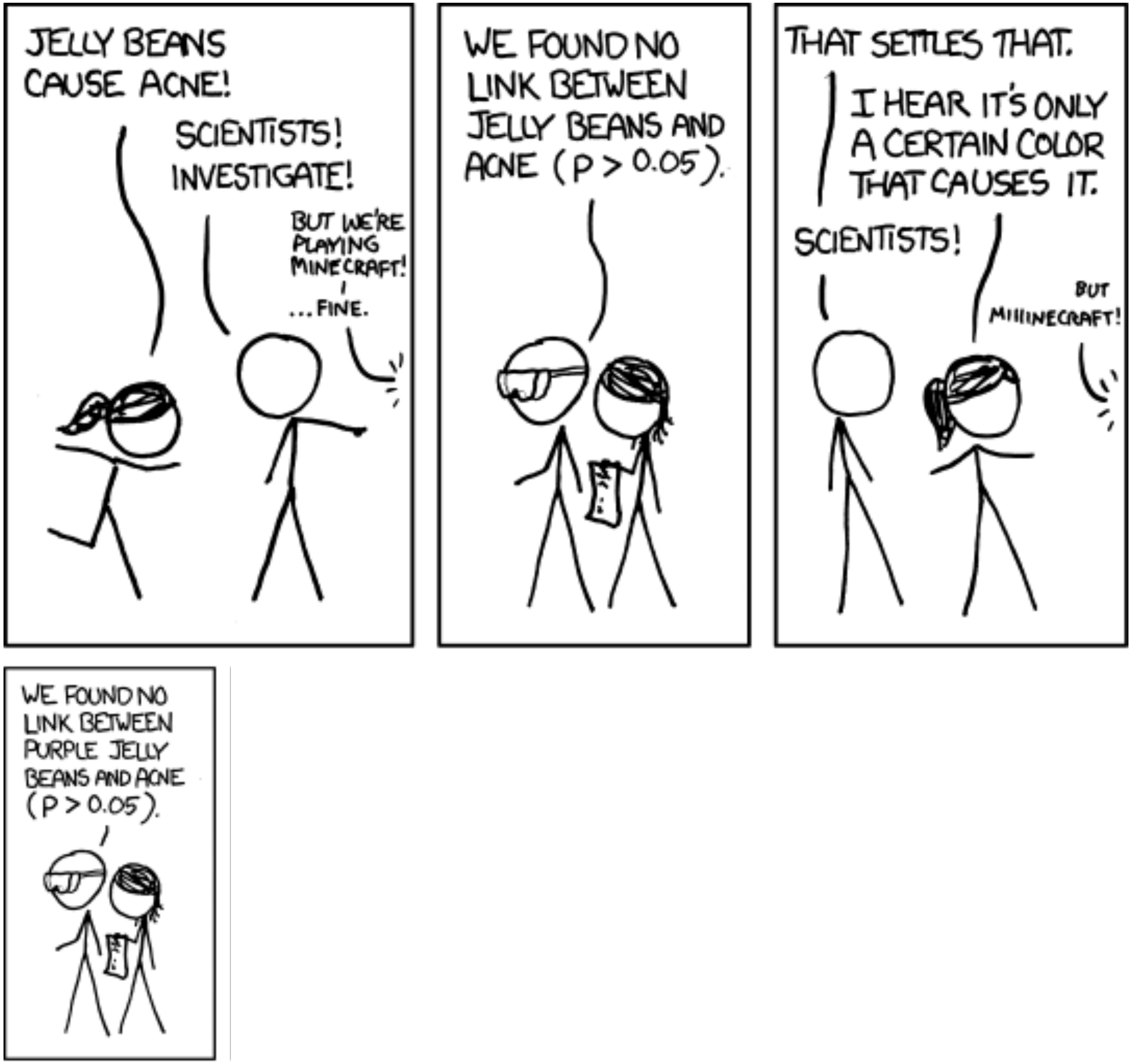
1. Multiple Hypothesis Testing

https://xkcd.com/882/



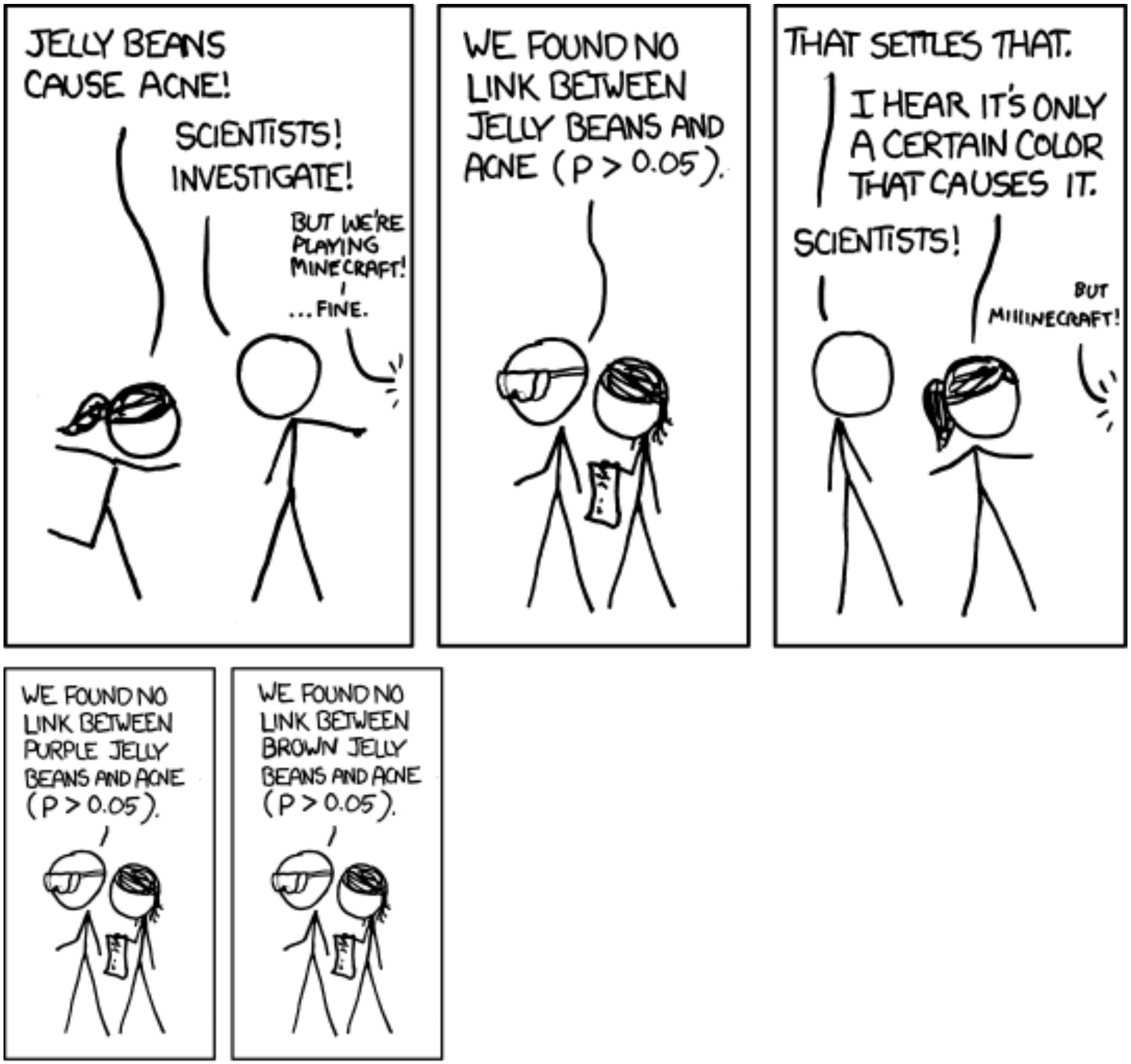
1. Multiple Hypothesis Testing

https://xkcd.com/882/



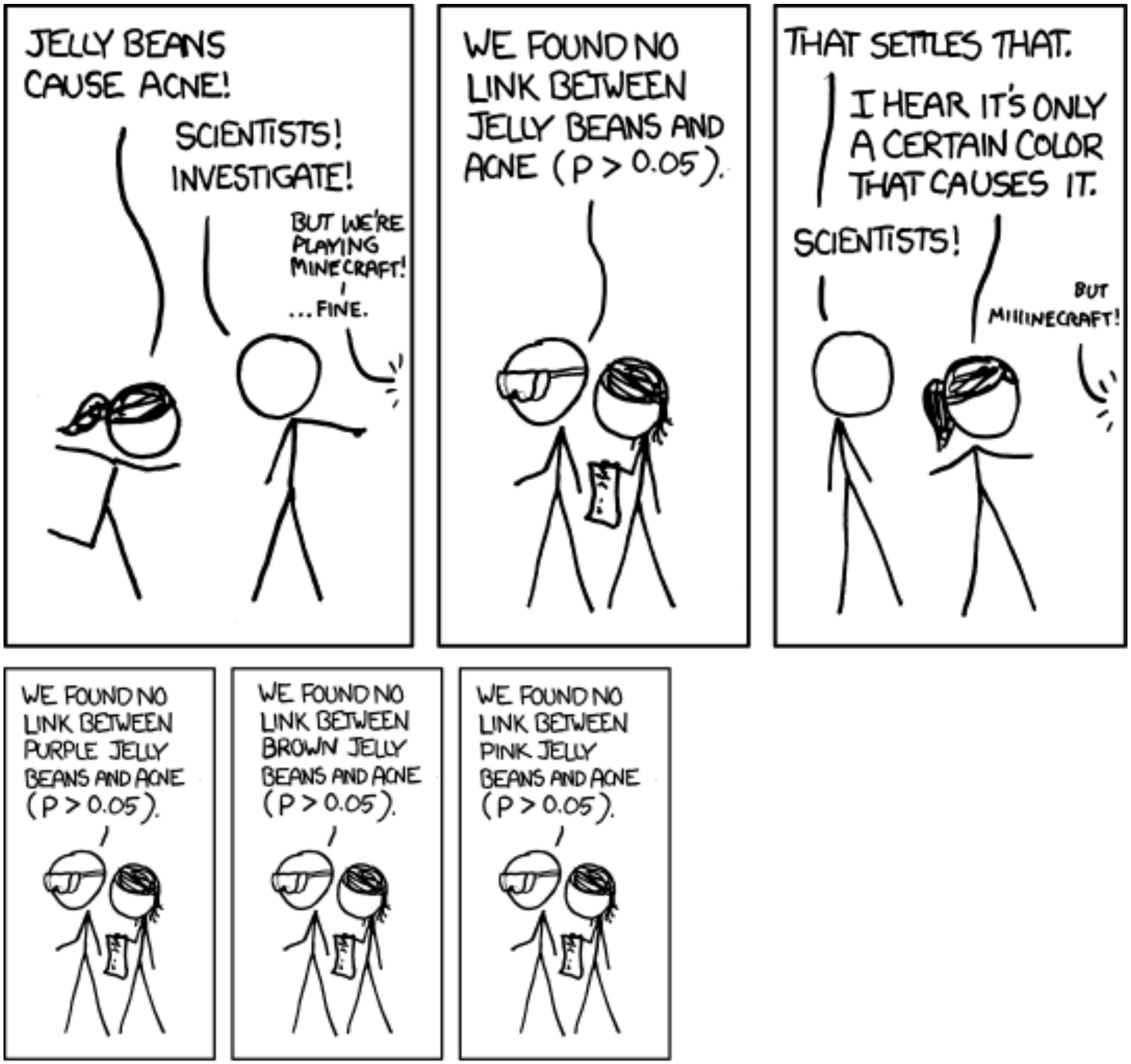
1. Multiple Hypothesis Testing

https://xkcd.com/882/



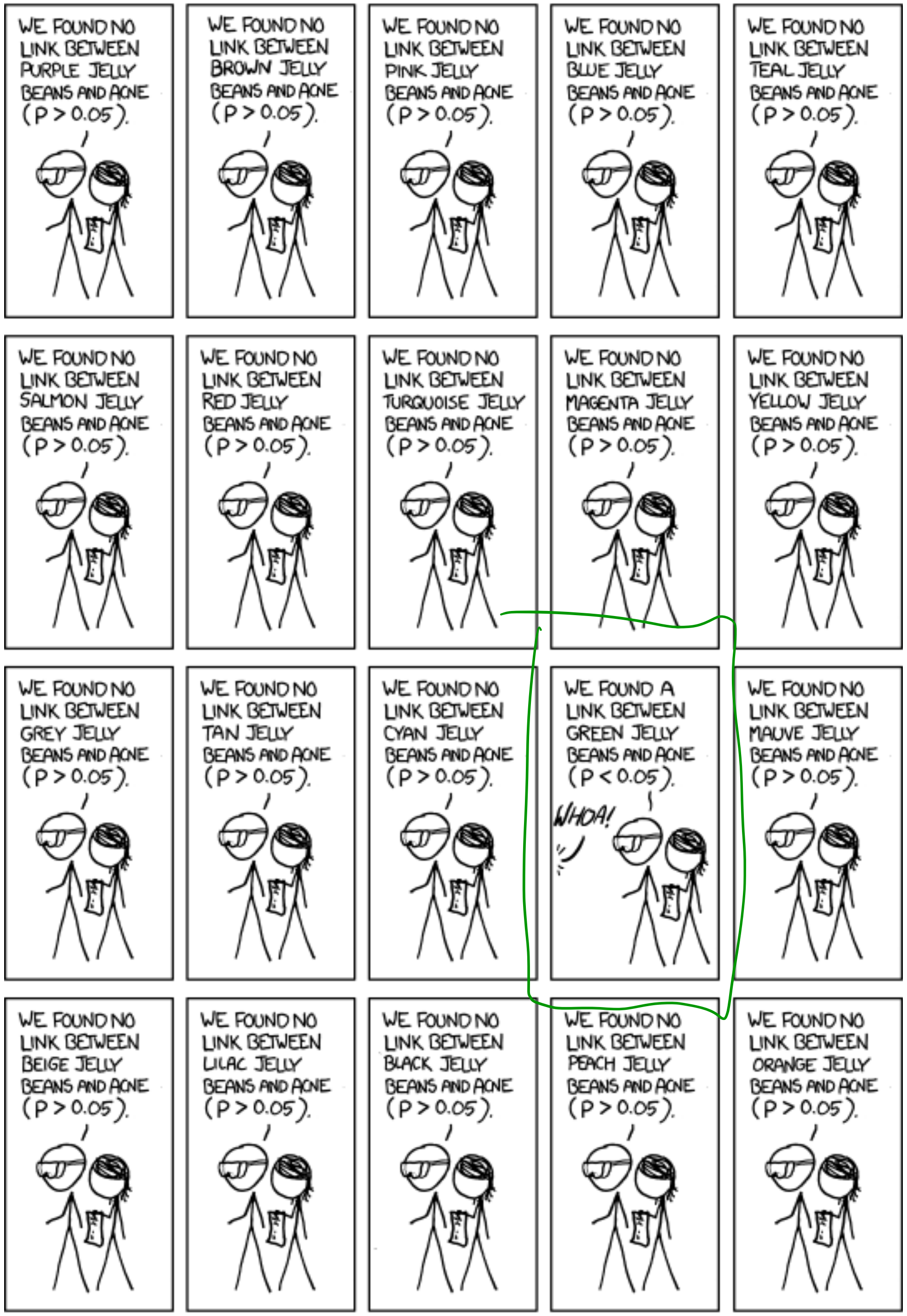
1. Multiple Hypothesis Testing

https://xkcd.com/882/



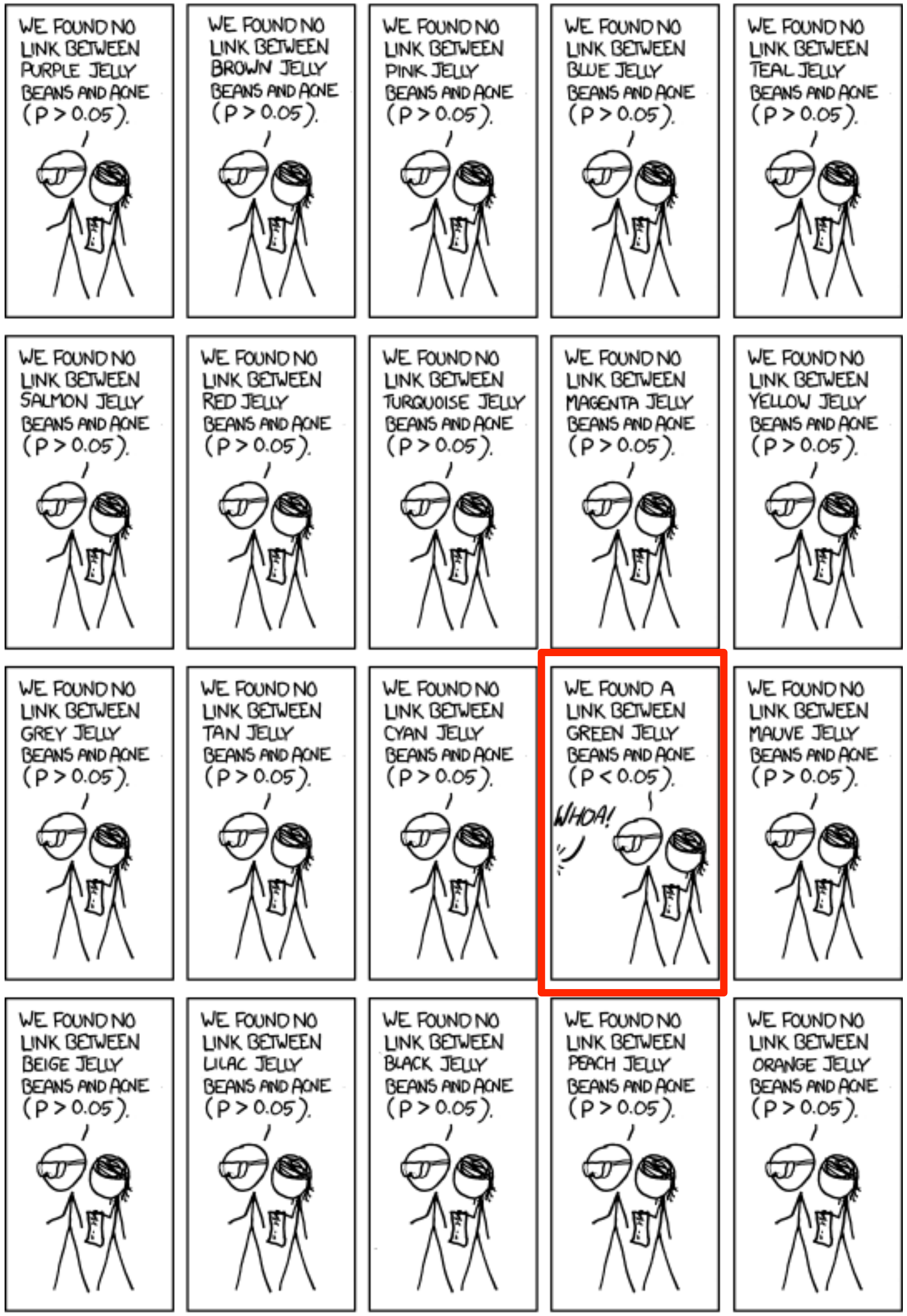
1. Multiple Hypothesis Testing

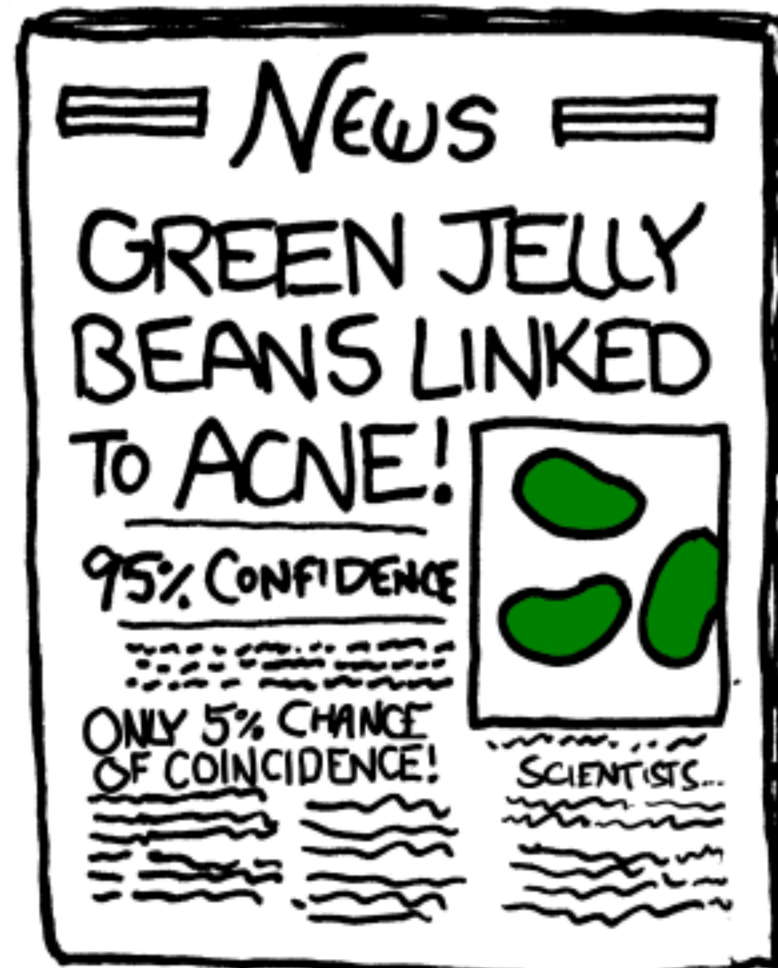
https://xkcd.com/882/



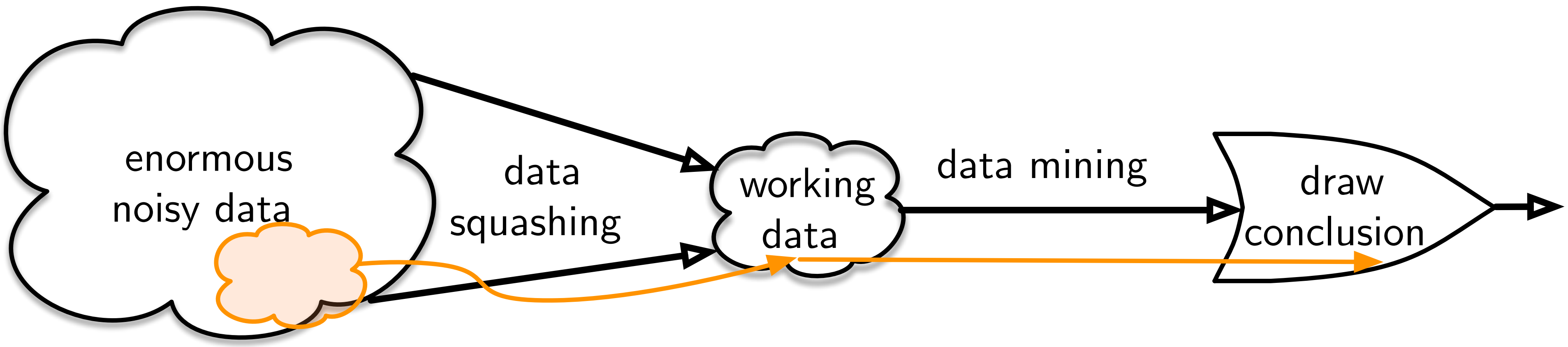
1. Multiple Hypothesis Testing

https://xkcd.com/882/

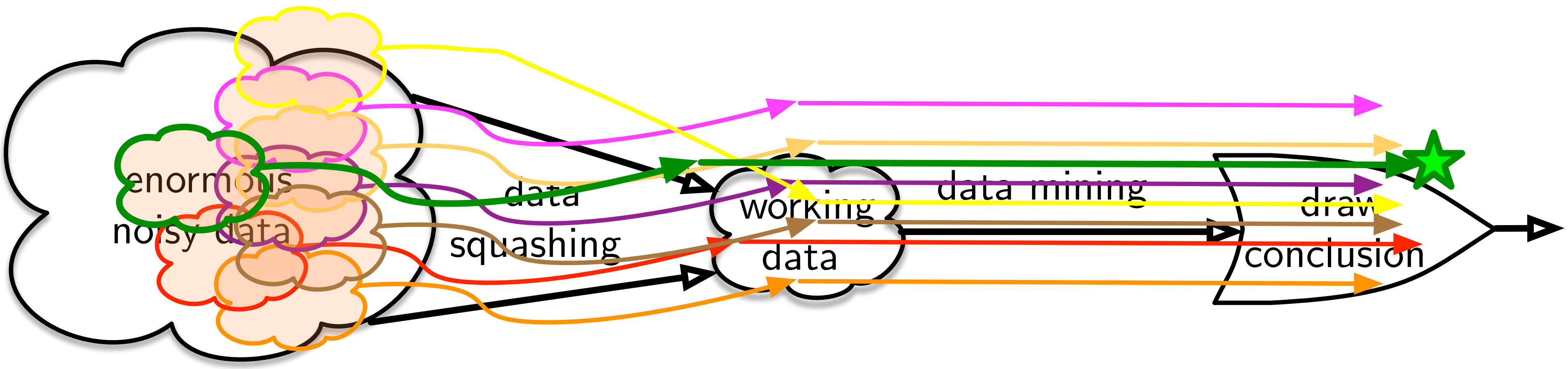




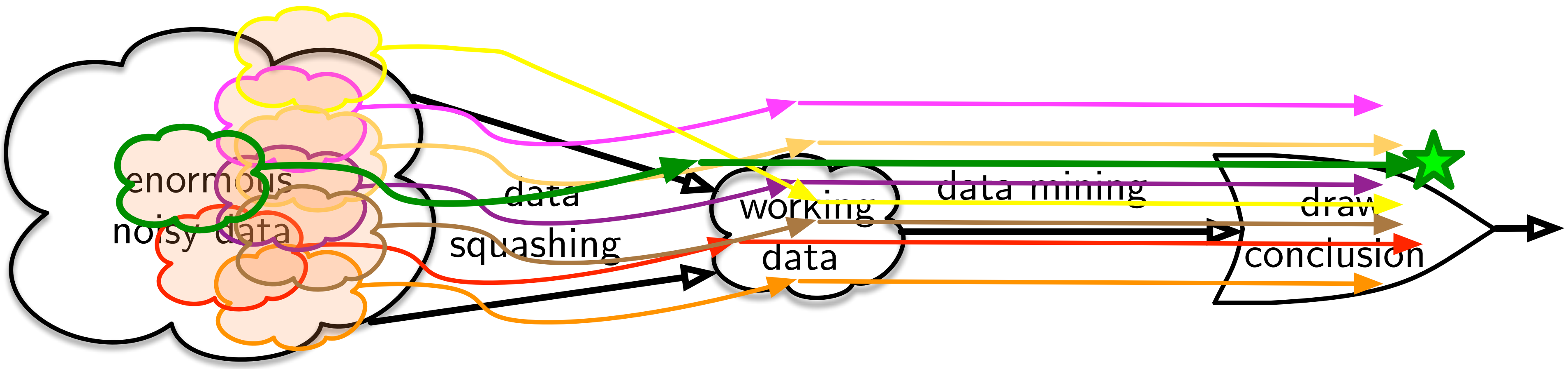
1. Multiple Hypothesis Testing



1. Multiple Hypothesis Testing



1. Multiple Hypothesis Testing



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis PLOS 2:8, 2005

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

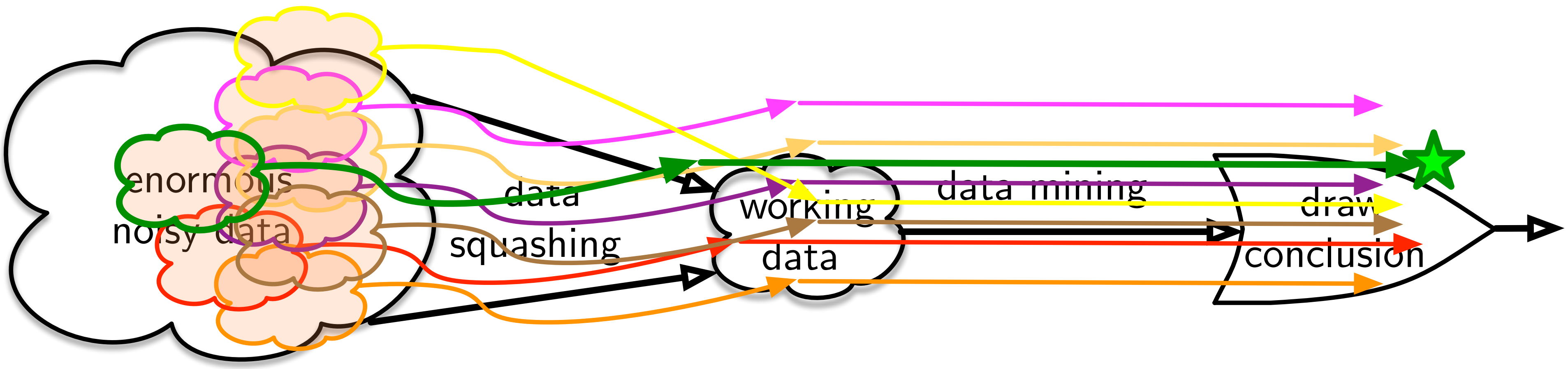
factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out

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

1. Multiple Hypothesis Testing



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis PLOS 2:8, 2005

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

factors that influence this problem and some corollaries thereof.

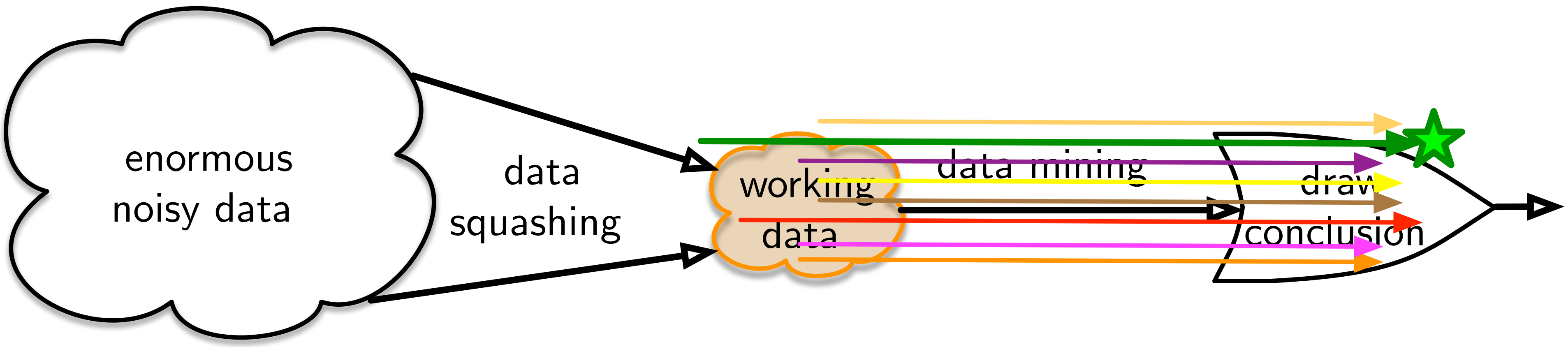
Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11

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 th

Bonferroni Correction?

1. Multiple Hypothesis Testing




Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis PLOS 2:8, 2005

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

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out

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

Bonferroni Correction?

Garden of Forking Paths [Gelman + Loken 2013]

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.

Garden of Forking Paths [Gelman + Loken 2013]

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.
2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

Garden of Forking Paths [Gelman + Loken 2013]

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.
2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.
3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.

Garden of Forking Paths [Gelman + Loken 2013]

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.
2. Classical test pre-chosen from a set of possible tests: thus, $T(y;\phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.
3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y;\phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.
4. "Fishing": computing $T(y;\phi_j)$ for $j = 1, \ldots, J$: that is, performing J tests and then reporting the best result given the data, thus $T(y;\phi^{\text{best}}(y))$.