L15: Cross-Validation & p-values

Jeff M. Phillips

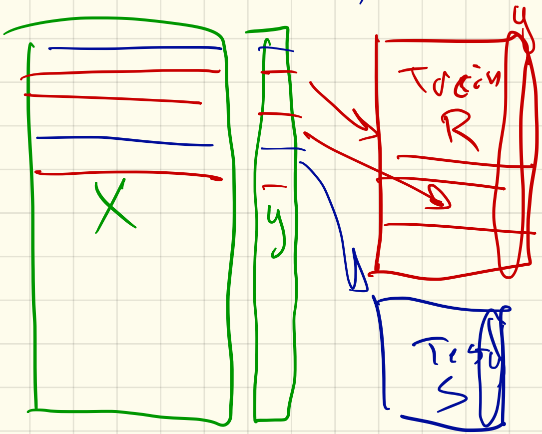
March 2, 2020

Cross-Validation

- How to choose parameters.
- Predict Generalization

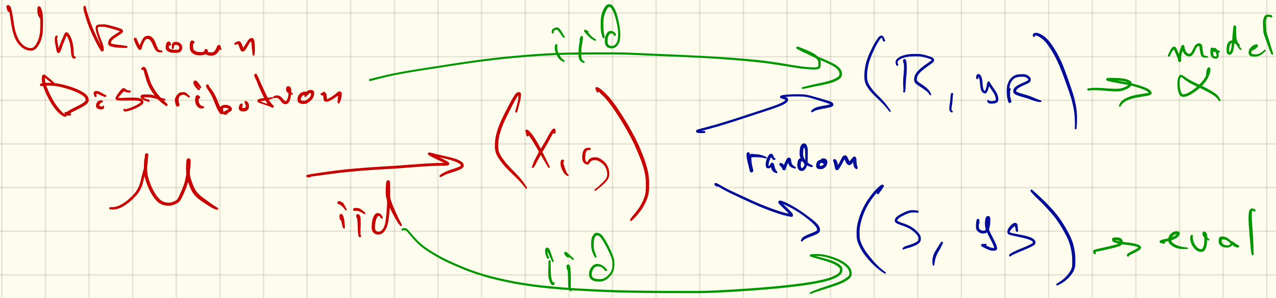
Cost function $C((X, \alpha, s))$

$$C((X,y), \alpha, s) = \sum_{i=1}^{n} \left(y_i - \langle \alpha, x_i \rangle\right)^2 + s\|\alpha\|_2^2$$



1. Split $(X,y) \to R, S$
2. Build models
   $\alpha_{s_1} \leftarrow (R^T R + s_1 I)^{-1} R^T y_R$
   $\alpha_{s_2} \leftarrow \ldots$
3. Evaluate Model
   $C_{s_1} = \sum_{x_i \in S} \left(y_i - \langle x_i, \alpha_{s_1} \rangle\right)^2$

Why Does C-V Make Sense?



What should Train / Test split be?

70% / 30% 90% / 10% 99% / 1%

↳ As you get more data

→ build more complex model.

Aim for $|S| \gtrsim 1000$, more if evaluating a lot of parameters.

Cross-Validation on Small Data

"Artisinal"

size of data $n = 20$

Leave-one out (LOO) CV

1. Splits n different ways

$R_1 = \{x_2, x_3 \dots x_n\}$ $S_1 = \{x_1\}$

$R_i = \{x_1, x_2 \dots x_{i-1}, x_{i+1}, \dots x_n\}$ $S_i = \{x_i\}$

2. Build n models $\alpha_1(s) \leftarrow \text{Train}(R_1, y_{R_1})$

$\vdots$

$\alpha_i(s) \leftarrow \text{Train}(R_i, y_{R_i})$

3. Eval $\text{Ave}(\text{Cost}(S_1, \alpha_1), \dots \text{Cost}(S_i, \alpha_i) \dots \}$

Choose param s, smallest $\rightarrow$ Rebuild on all of (X, y)

2 Uses

1. Choose param ← so far
2. Eval model

If you want to do both:

Split 3 ways X → R train
→ S test (of param)
→ E evaluates generalization

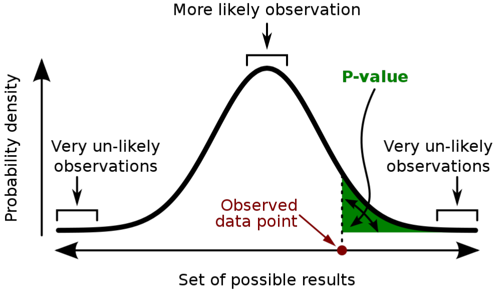
p-values

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.**



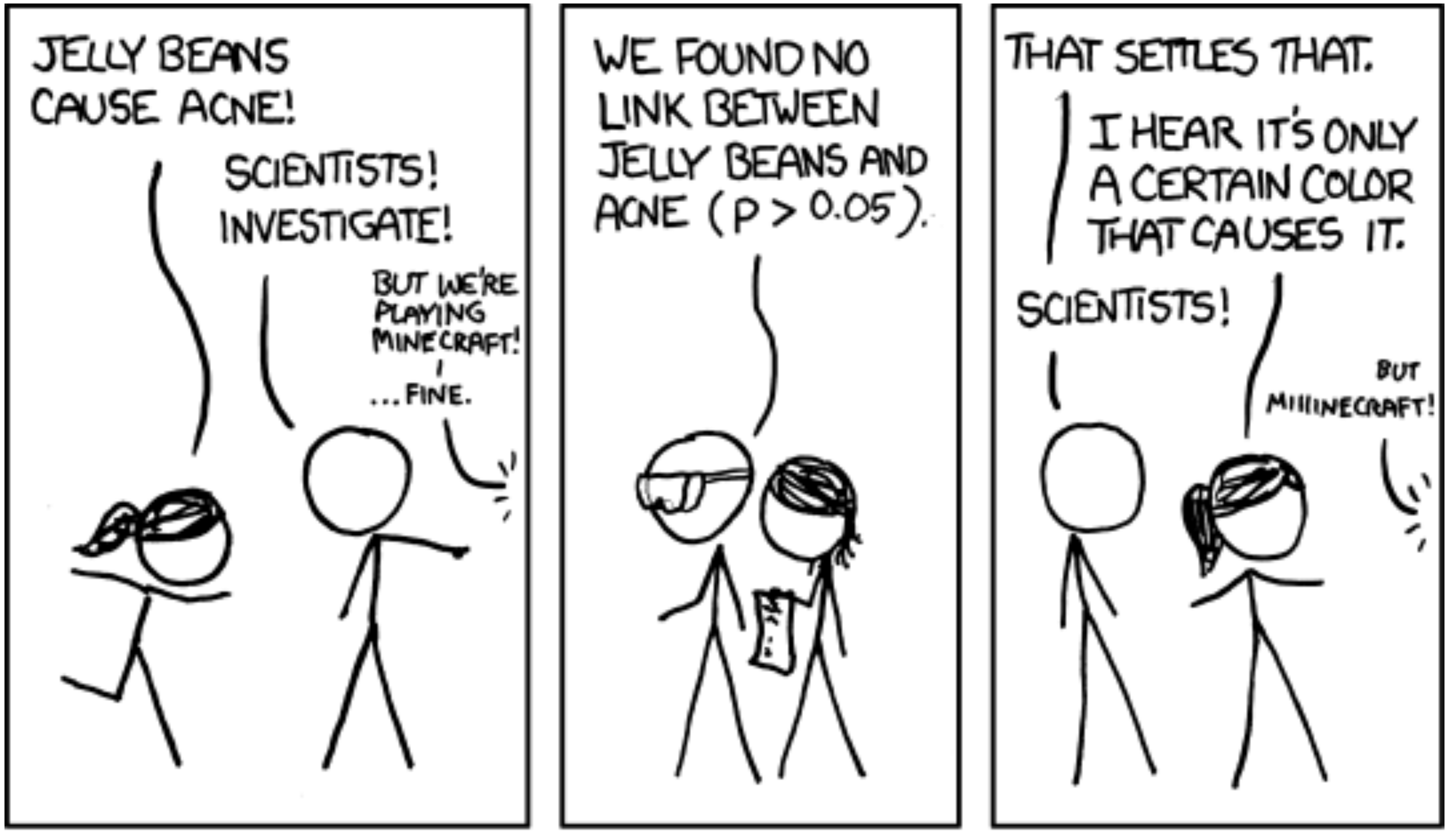
A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

worlds hypothesis H_0

Alternate Hypothesis H_1

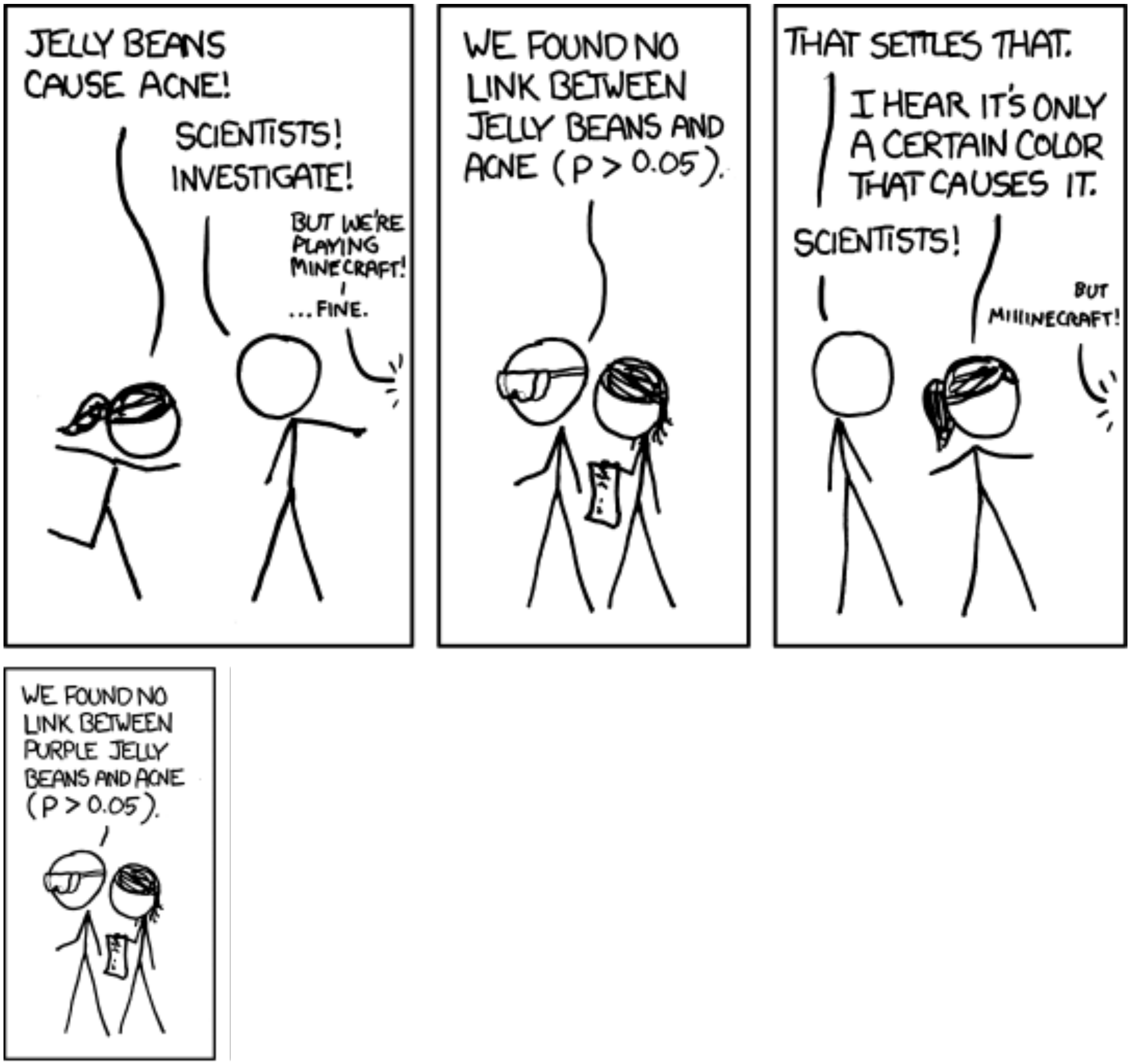
1. Multiple Hypothesis Testing

https://xkcd.com/882/



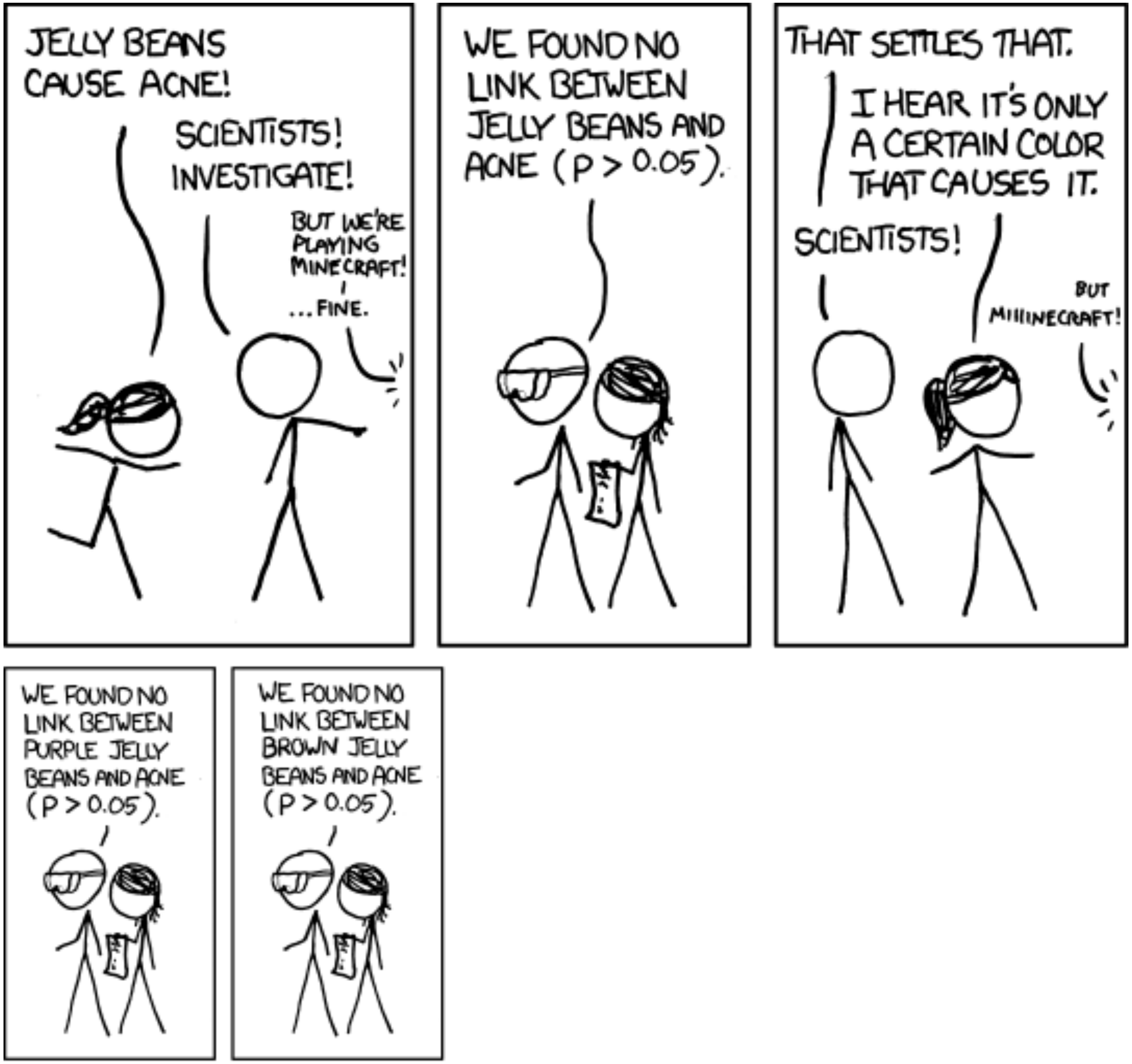
1. Multiple Hypothesis Testing

https://xkcd.com/882/



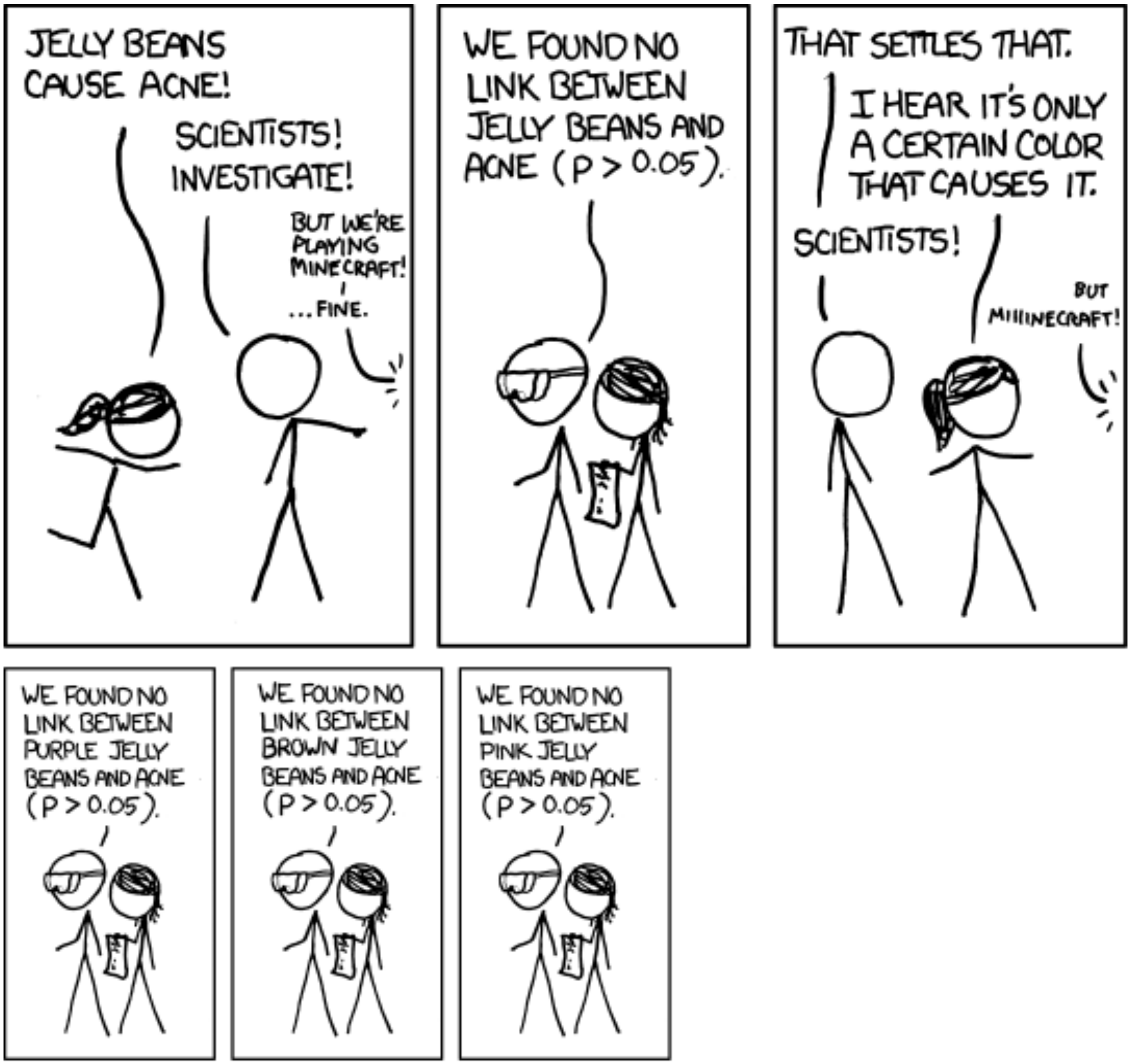
1. Multiple Hypothesis Testing

https://xkcd.com/882/



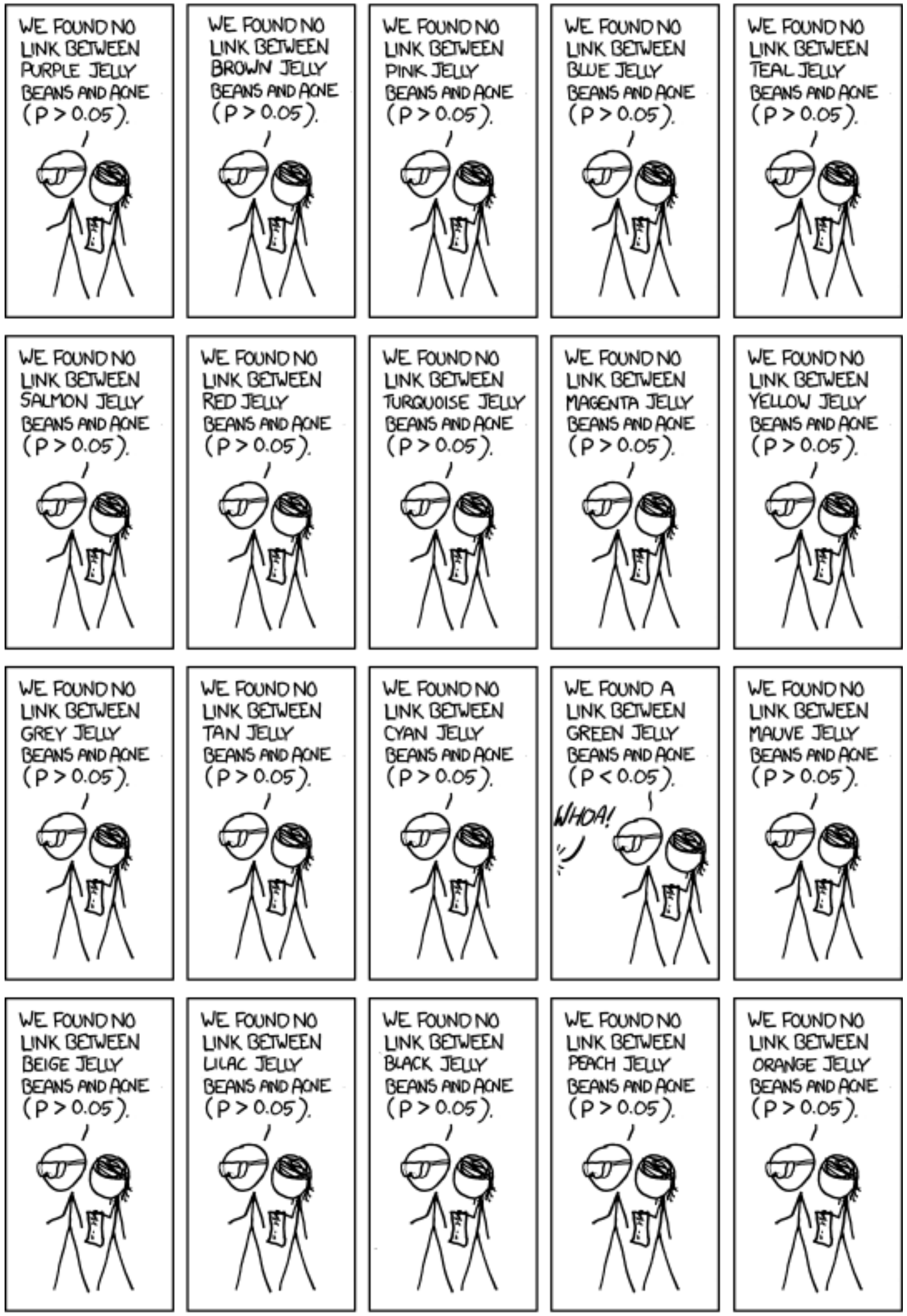
1. Multiple Hypothesis Testing

https://xkcd.com/882/



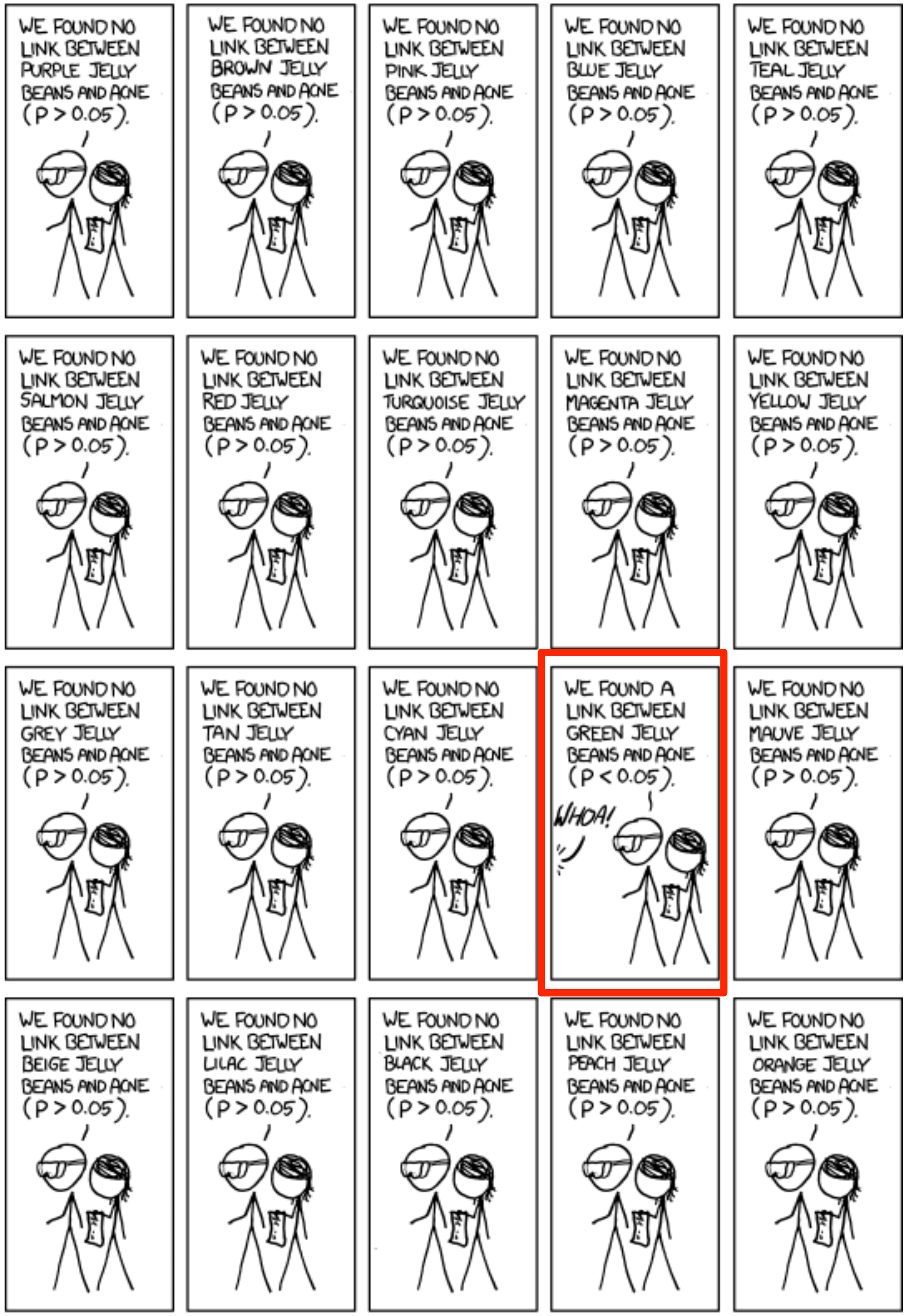
1. Multiple Hypothesis Testing

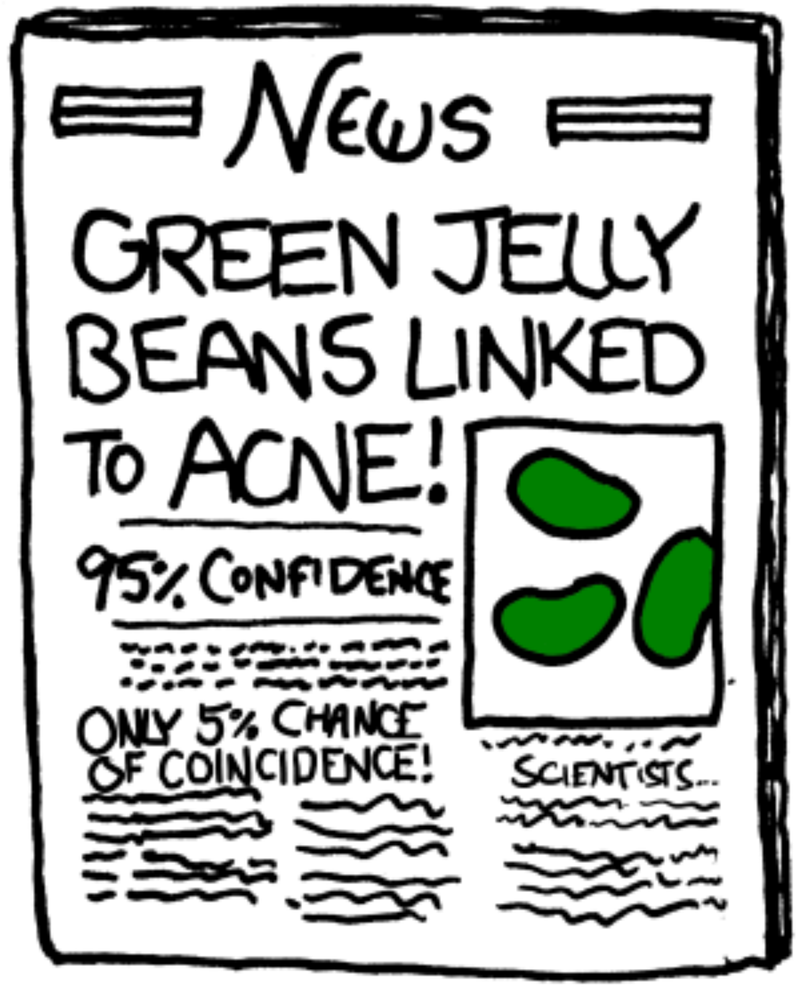
https://xkcd.com/882/



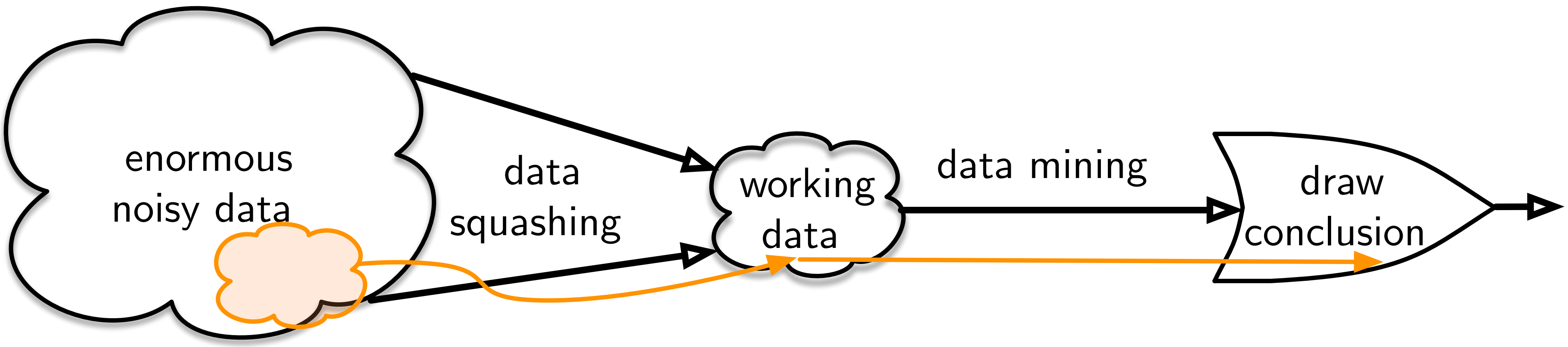
1. Multiple Hypothesis Testing

https://xkcd.com/882/

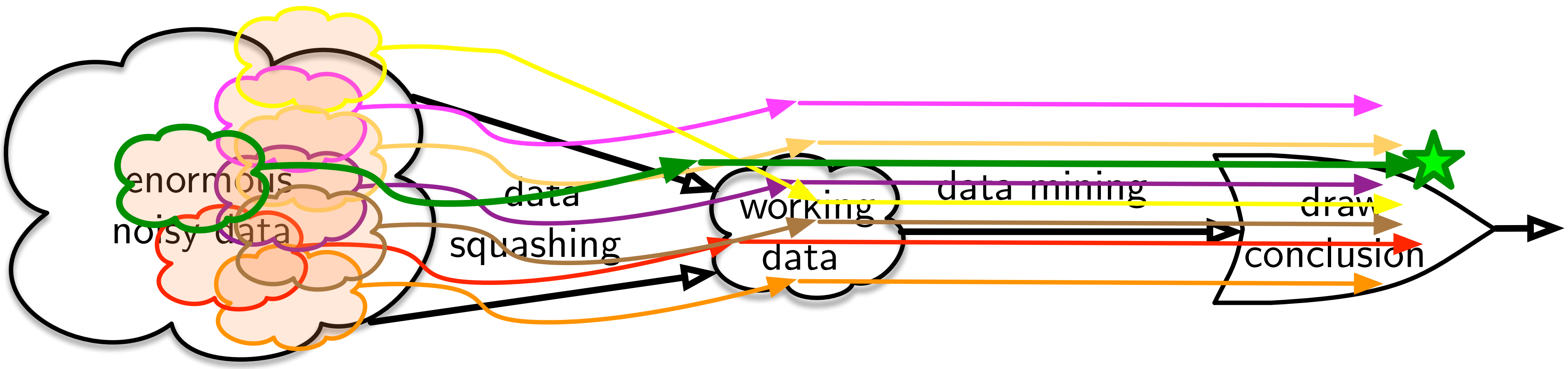




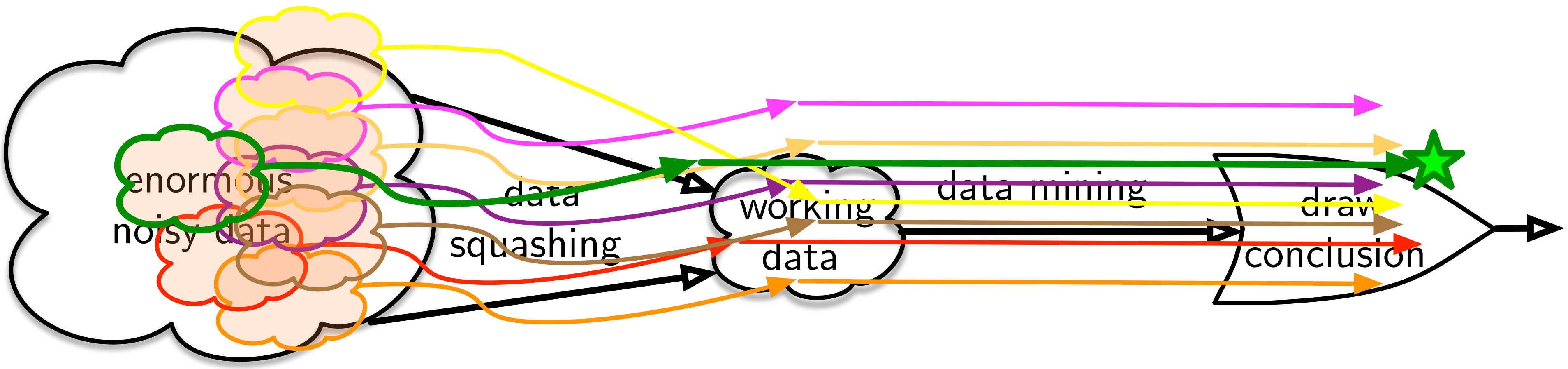
1. Multiple Hypothesis Testing



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

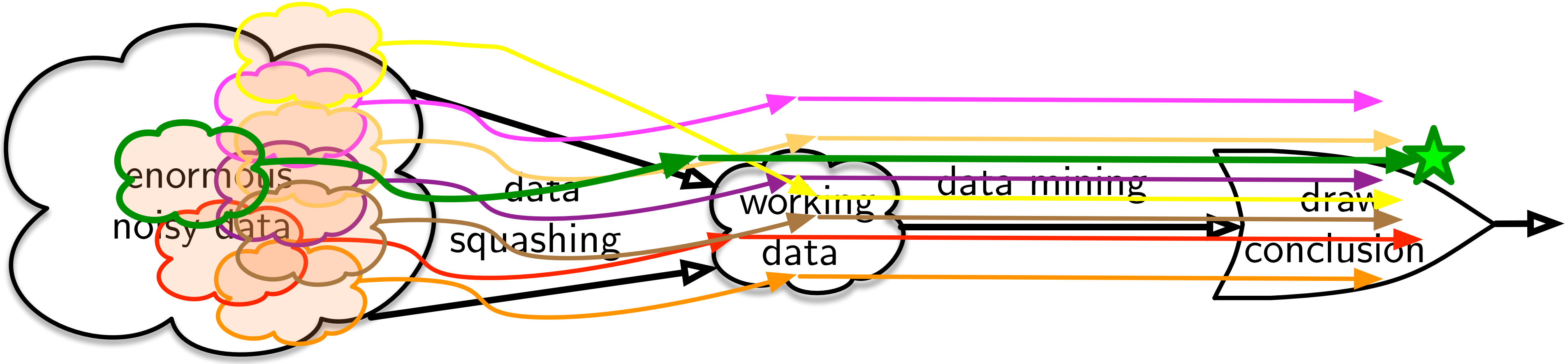
factors that influence this problem and some corollaries thereof.

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.

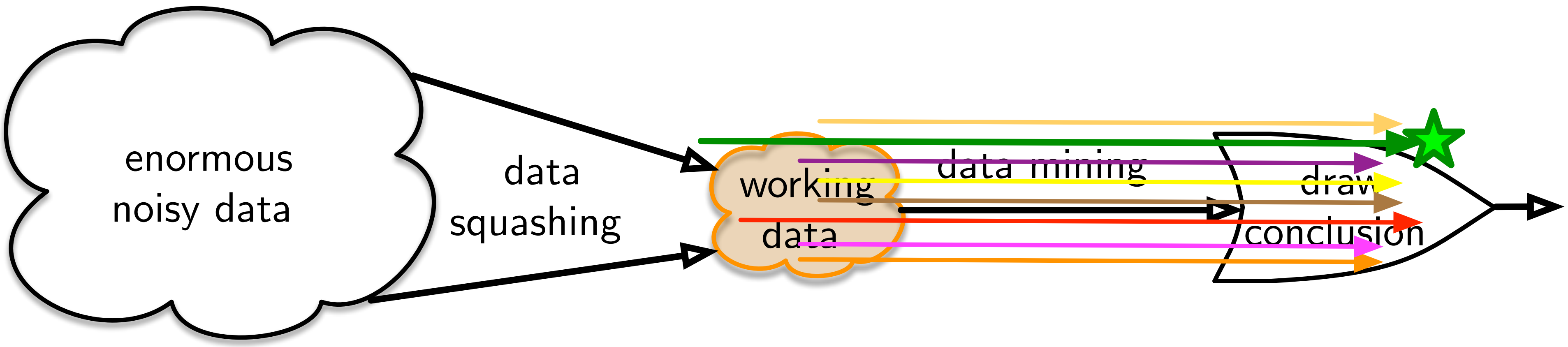
Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9,

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?

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 [9–1

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

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 $T(y)$.

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 $T(y; \phi^{\text{best}}(y))$.

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.

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 $T(y;\phi^{\text{best}}(y))$.

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 $T(y; \phi^{\text{best}}(y))$.