Misuse, Misreporting, Misinterpretation of Statistical Methods in Usable Privacy and Security Papers

Jenny Tang, Lujo Bauer, Nicolas Christin

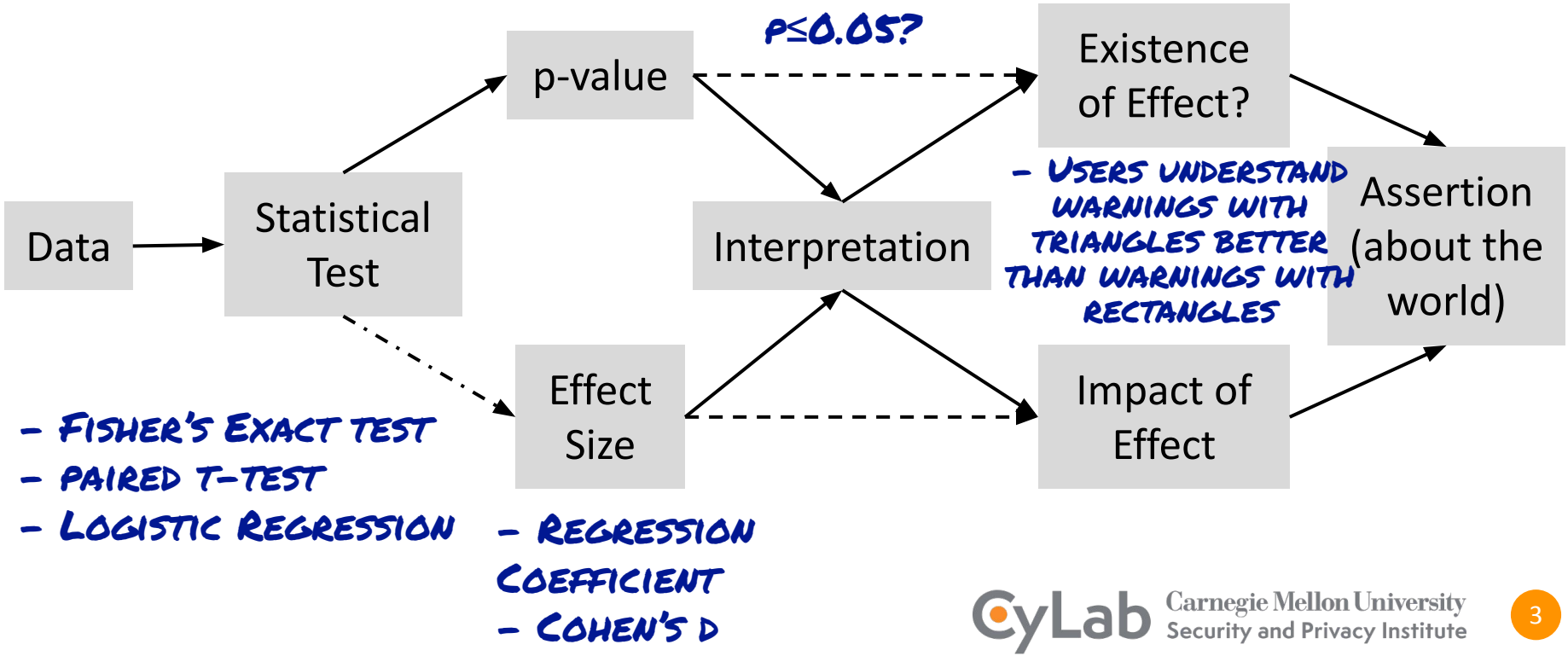
Carnegie Mellon University
Electrical & Computer Engineering

S3D Software and Societal Systems Department

CyLab Carnegie Mellon University Security and Privacy Institute

$$p \leq 0.05$$

Statistical Significance Testing Is Often Used as Support for Scientific Results



Let’s read a soup paper!

Let’s Read a Soup Paper!



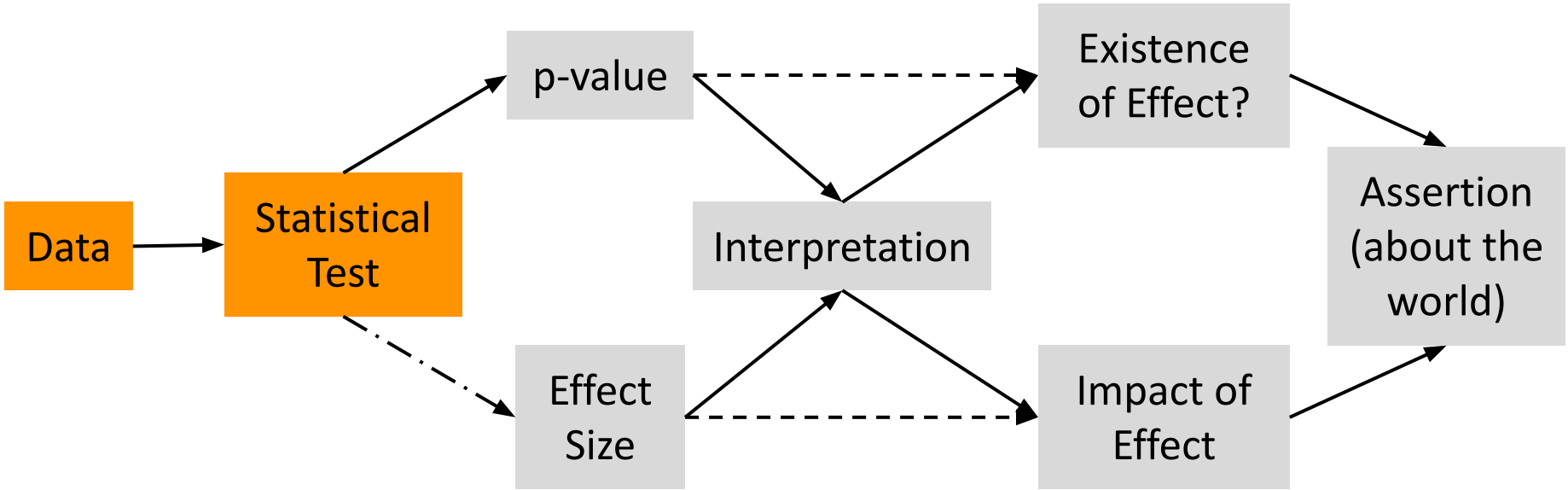
Miso Soup

VS

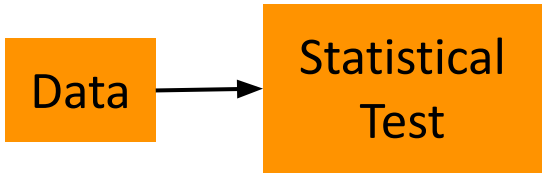


Clam Chowder

Misuse, Misreporting, Misinterpretation

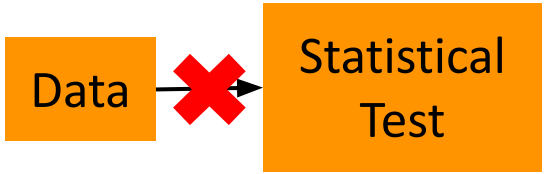


Misuse, Misreporting, Misinterpretation



Does the statistical test measure what it aims to investigate in the context of the data?

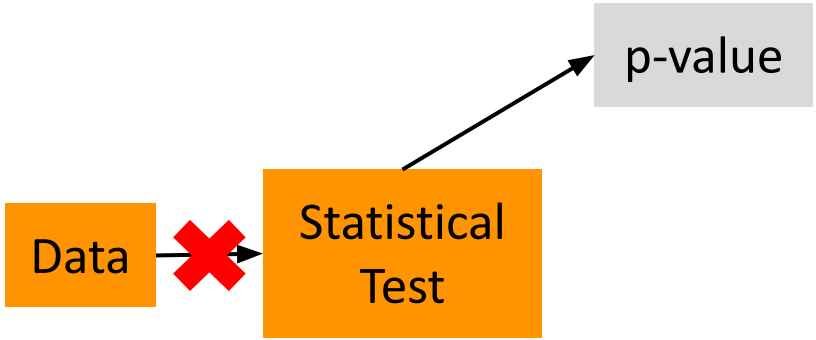
Misuse, Misreporting, Misinterpretation



Does the statistical test measure what it aims to investigate in the context of the data?

"We run a **paired t-test** on the **amount of money** users would be willing to spend on **clam chowder** compared to on **miso soup** …"

Misuse, Misreporting, Misinterpretation

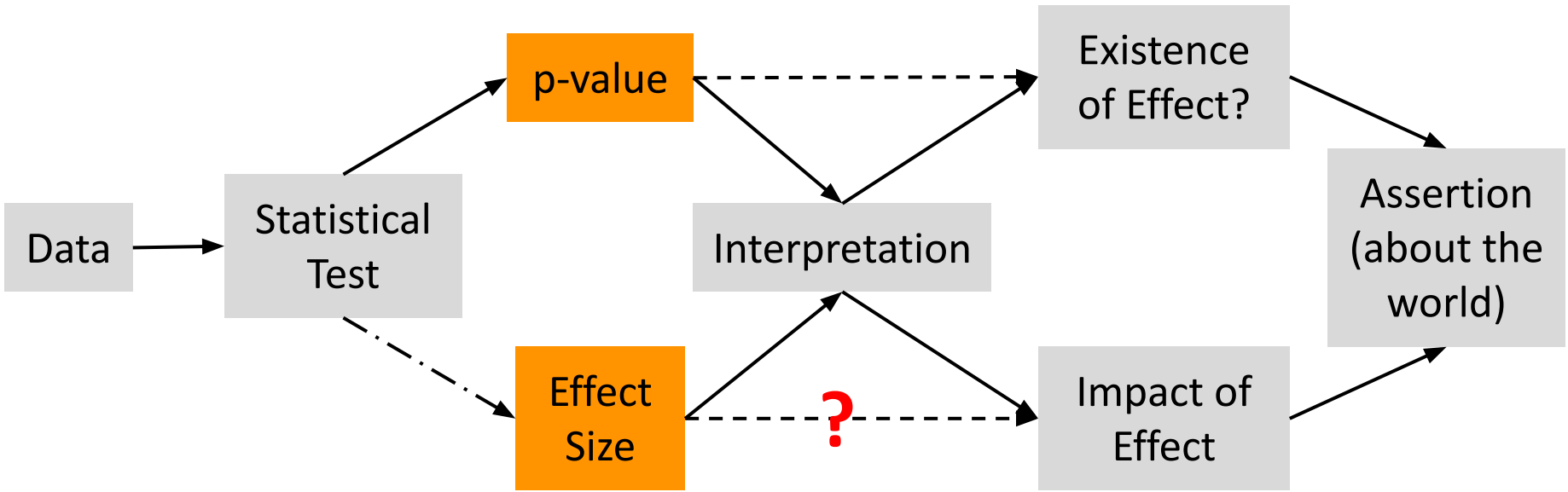


Does the statistical test measure what it aims to investigate in the context of the data?

TWO-SAMPLE T-TEST

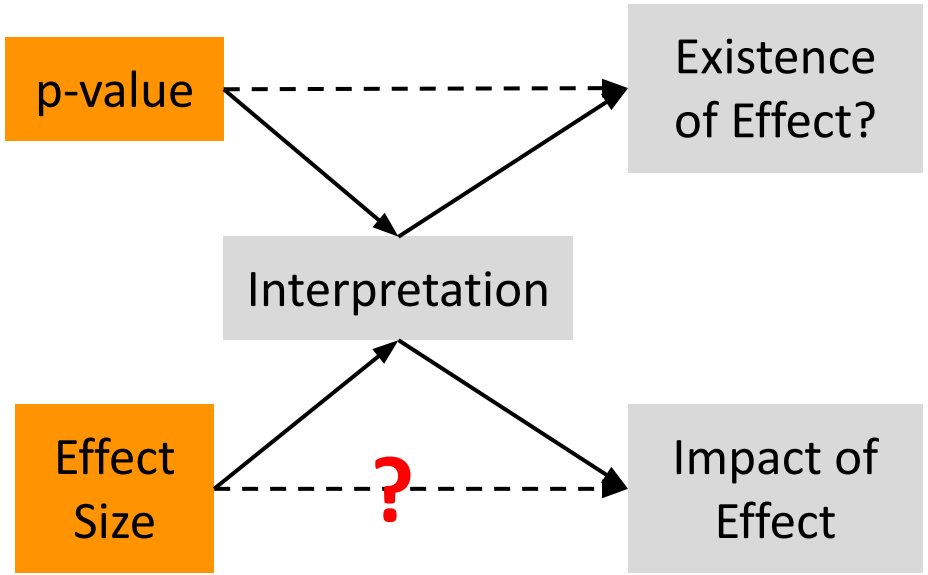
"We run a ~~**paired t test**~~ on the **amount of money** users would be willing to spend on **clam chowder** compared to on **miso soup** …"

Misuse, Misreporting, Misinterpretation



Misuse, Misreporting, Misinterpretation

Do the reported context and evidence provide enough information to evaluate the results from the statistical test?



Misuse, Misreporting, Misinterpretation

Do the reported context and evidence provide enough information to evaluate the results from the statistical test?

p-value

Effect Size

"…Users would **spend more money** on a bowl of clam chowder than on a bowl of miso soup (**p=.012**)."

Misuse, Misreporting, Misinterpretation

Do the reported context and evidence provide enough information to evaluate the results from the statistical test?

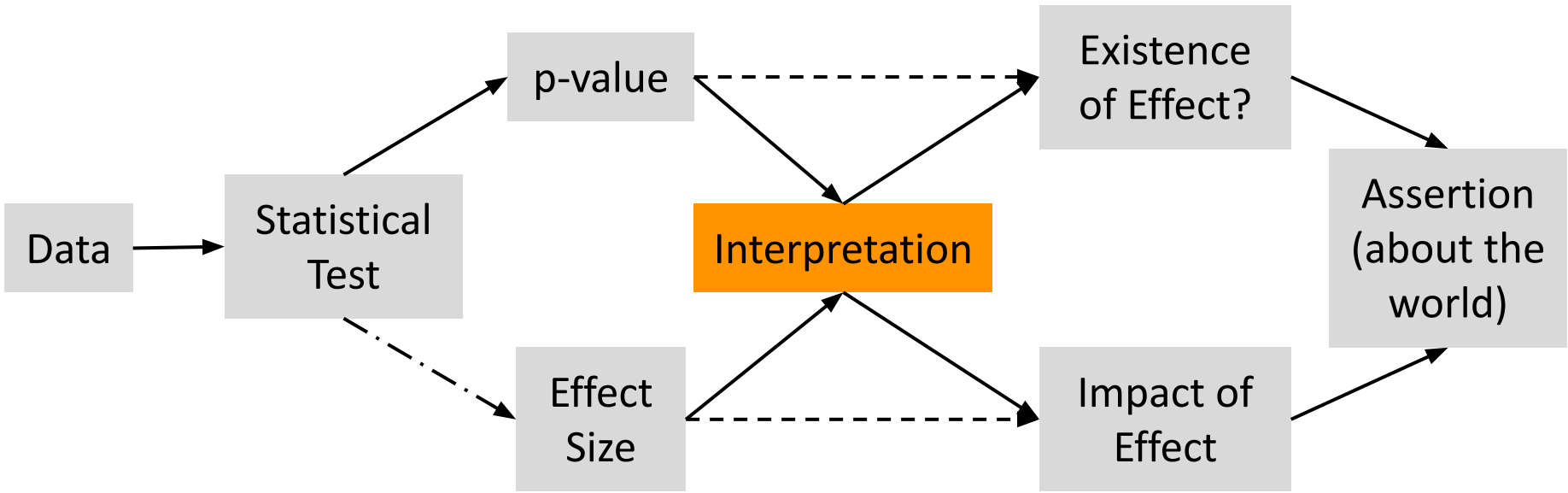
p-value

Effect Size



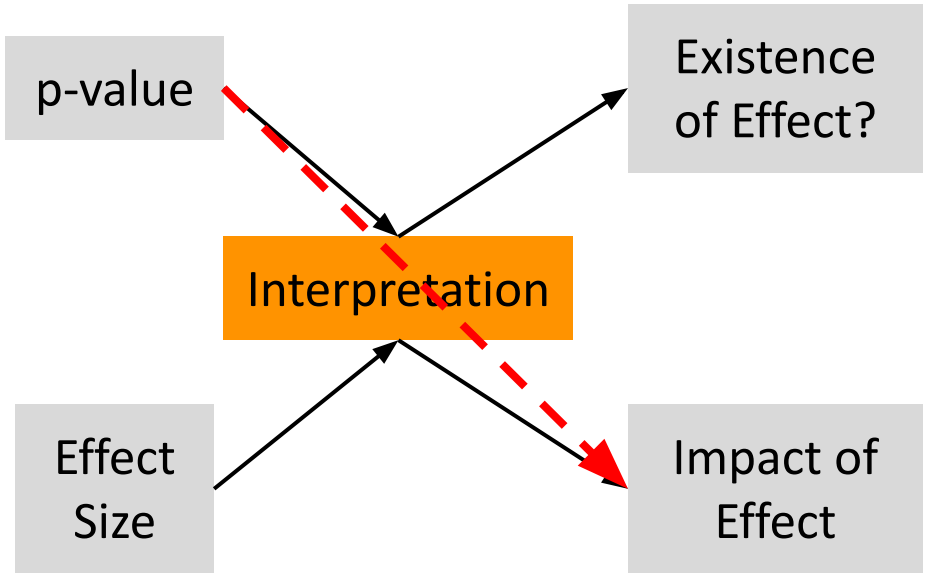
"…Users would **spend more money** on a bowl of clam chowder than on a bowl of miso soup (**p=.012**)."

Misuse, Misreporting, Misinterpretation



Misuse, Misreporting, Misinterpretation

Do the results from the statistical tests support the assertions made in the paper?



Misuse, Misreporting, Misinterpretation

"Users like clam chowder **much more** than miso soup (**p<.001**)."

Interpretation

Do the results from the statistical tests support the assertions made in the paper?

Misuse, Misreporting, Misinterpretation

"Users like clam chowder **much more** than miso soup **(p<.001)**."

SMALL P-VAL
!=
LARGE EFFECT

Interpretation

Do the results from the statistical tests support the assertions made in the paper?

What about SOUPS papers?

Soup Paper





SOUPS Papers

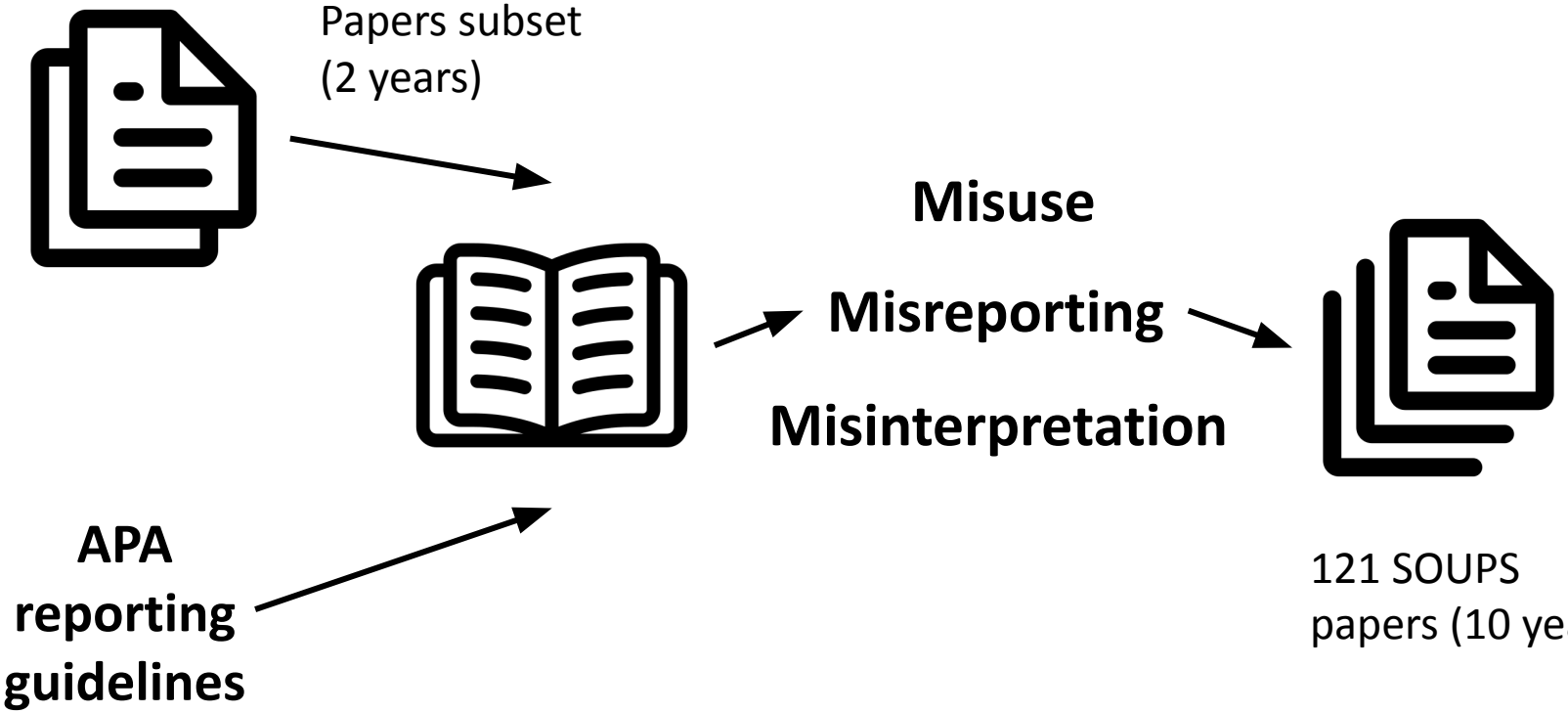


We analyzed 121 SOUPS papers that used statistical significance testing



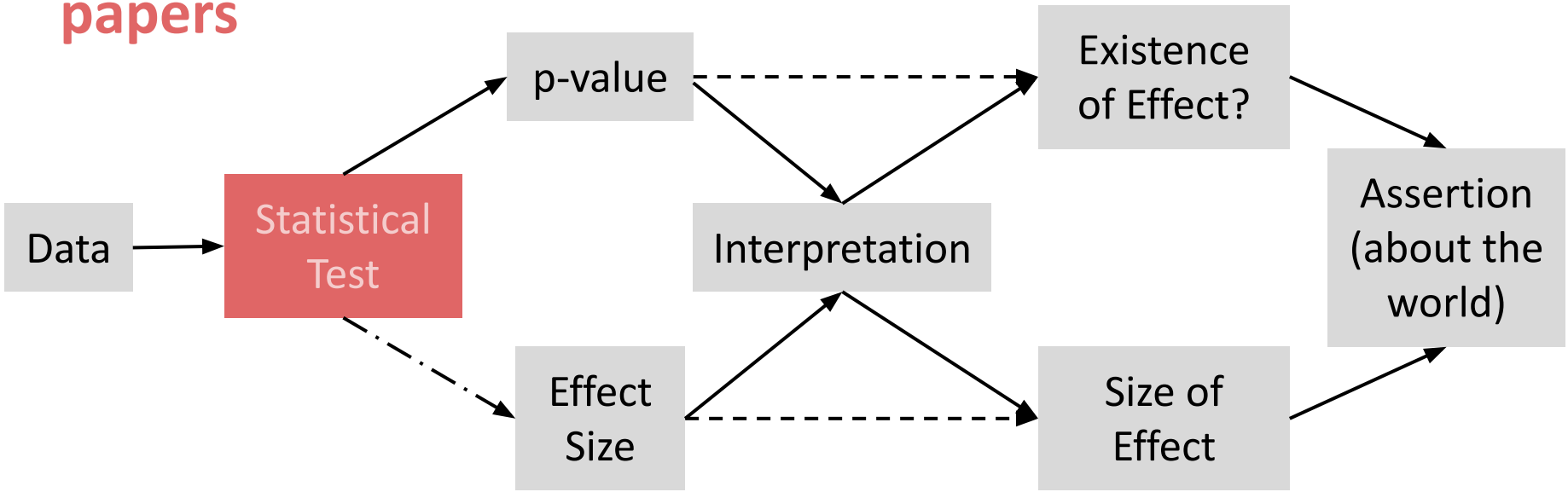
- Published in even-numbered years (2006-2024)
  - 223 papers
- Had at least one statistical test
  - 121 papers
  - 10 years

We Coded Each Instance of a Statistical Test



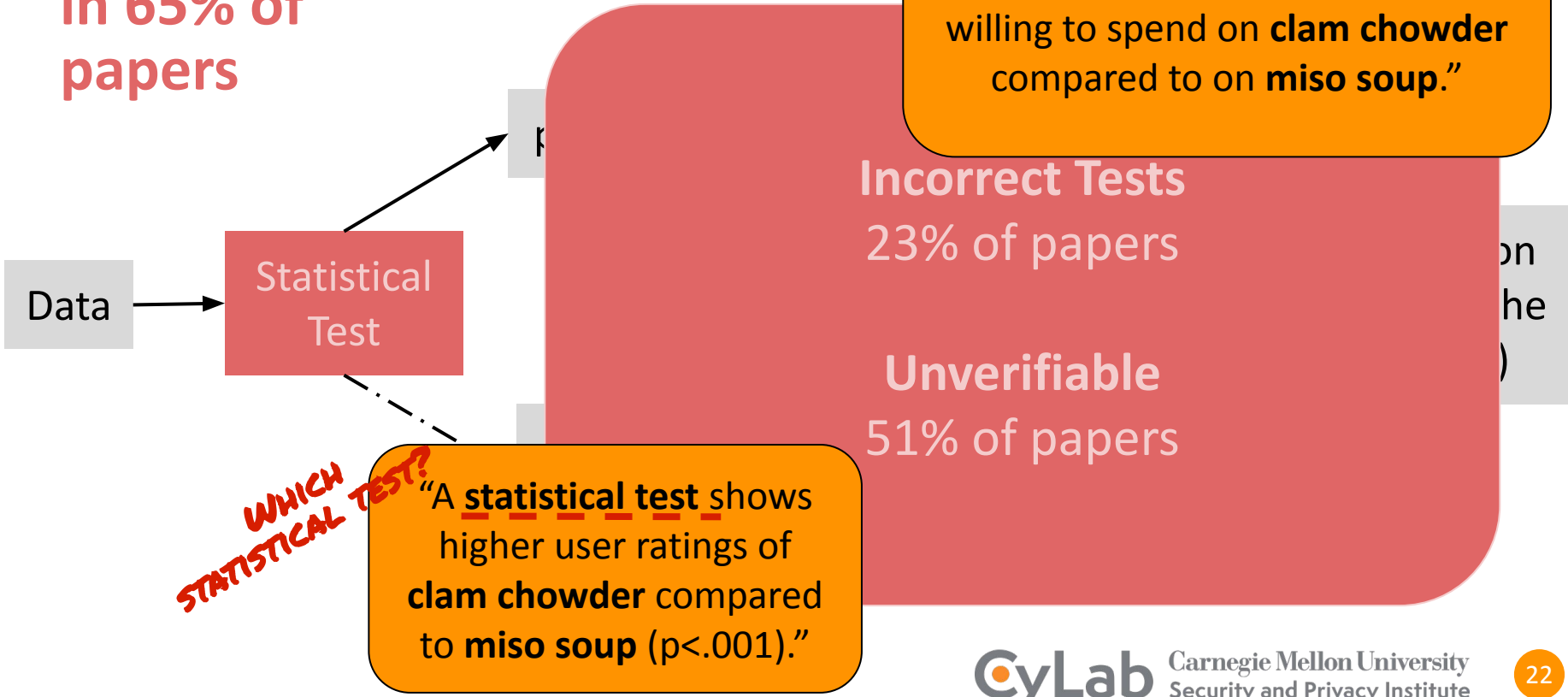
Misuse, Misreporting, Misinterpretation

in 65% of papers



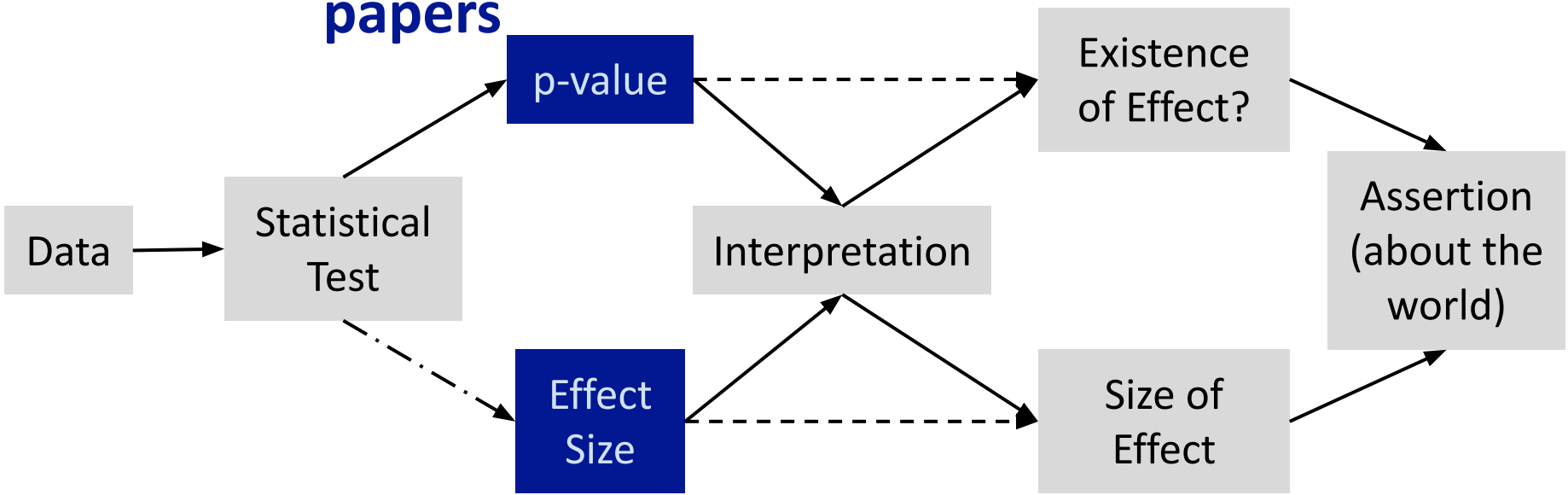
Misuse, Misreporting, Misint

in 65% of papers



Misuse, **Misreporting**, Misinterpretation

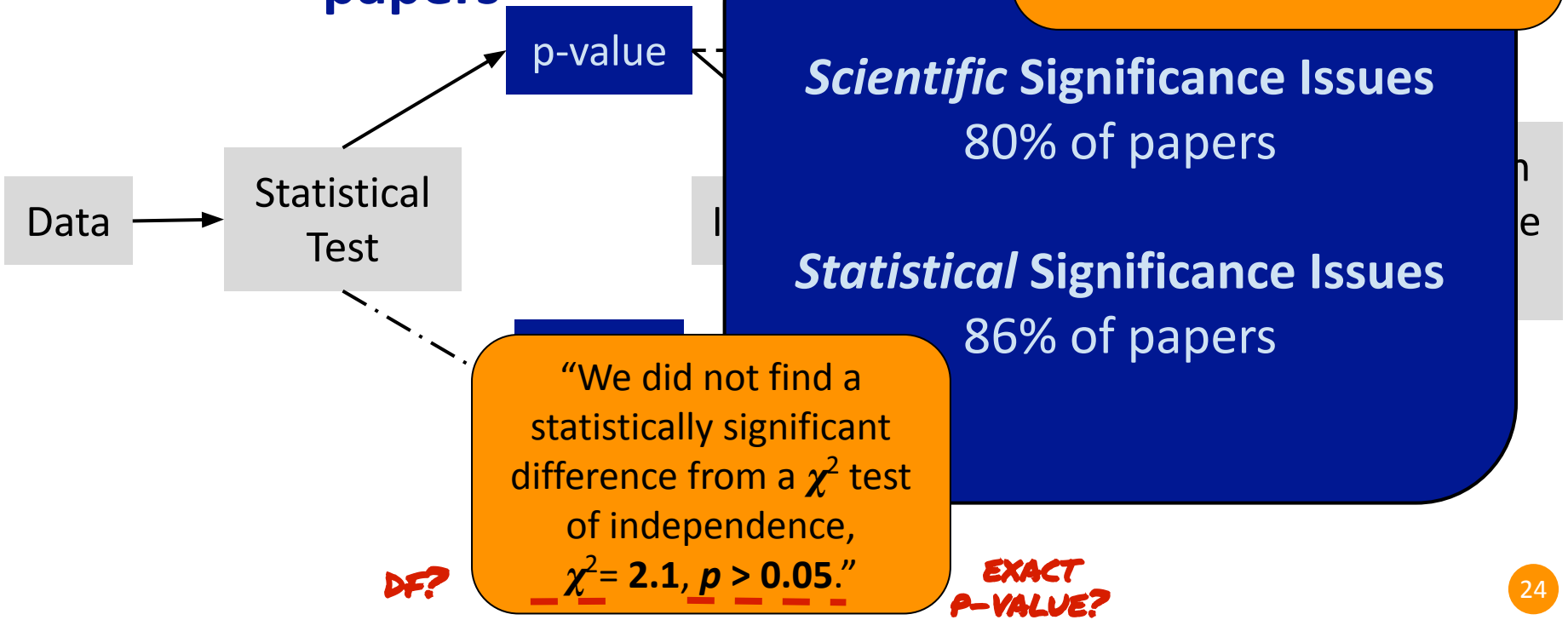
in 65% **in 95% of papers**



Misuse, **Misreporting**, Misinterp

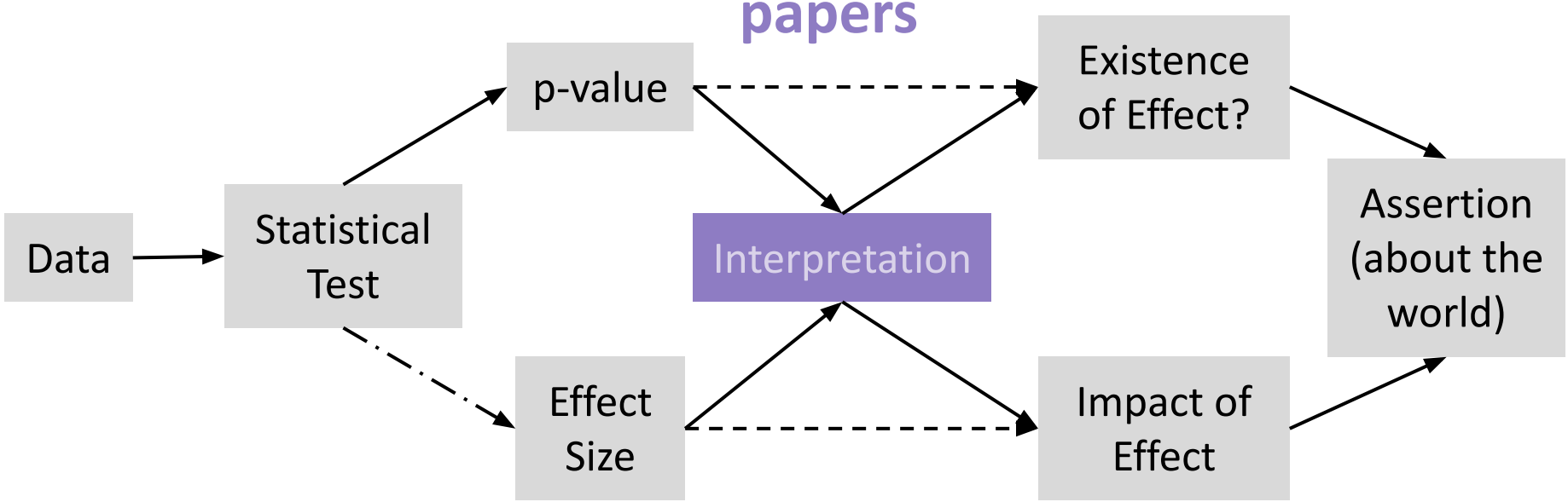
in 65%

in 95% of papers



Misuse, Misreporting, Misinterpretation

in 65% in 95% in 45% of papers



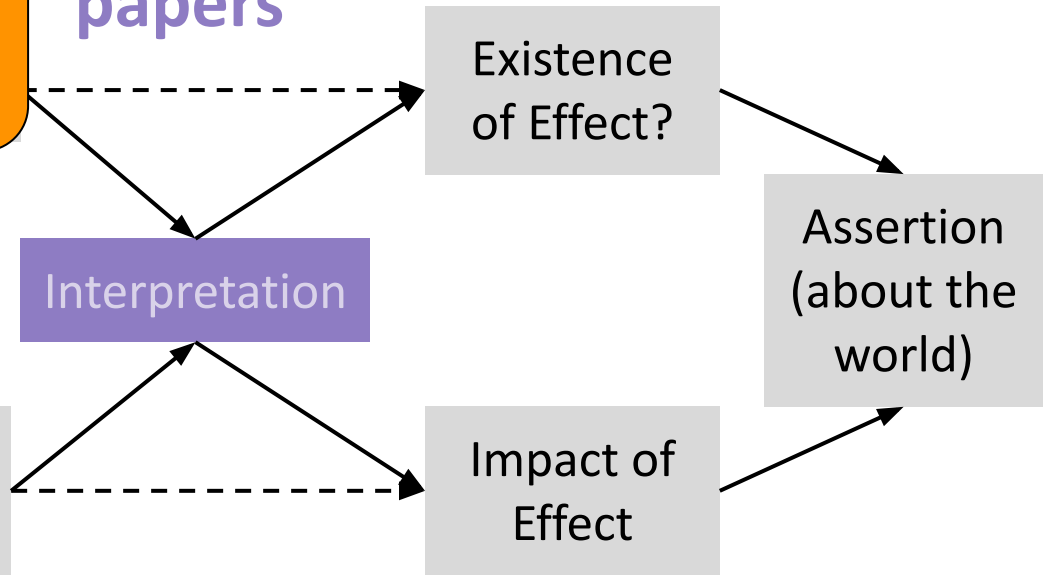
Misuse, Misreporting, Misinterpretation in 65% in 45% of papers

"Users like clam chowder **much more** than miso soup (**p<.001**)."

Improper Interpretation
27% of papers

No Interpretation
7% of papers

Regression table with no further explanation

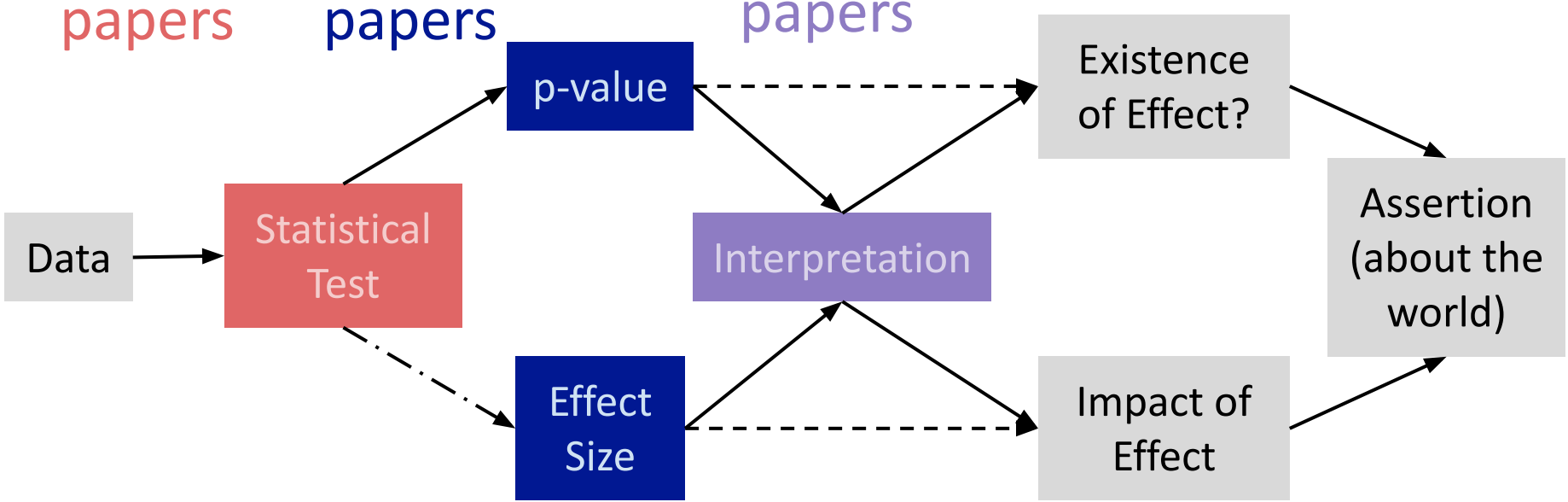


Misuse, Misreporting, Misinterpretation

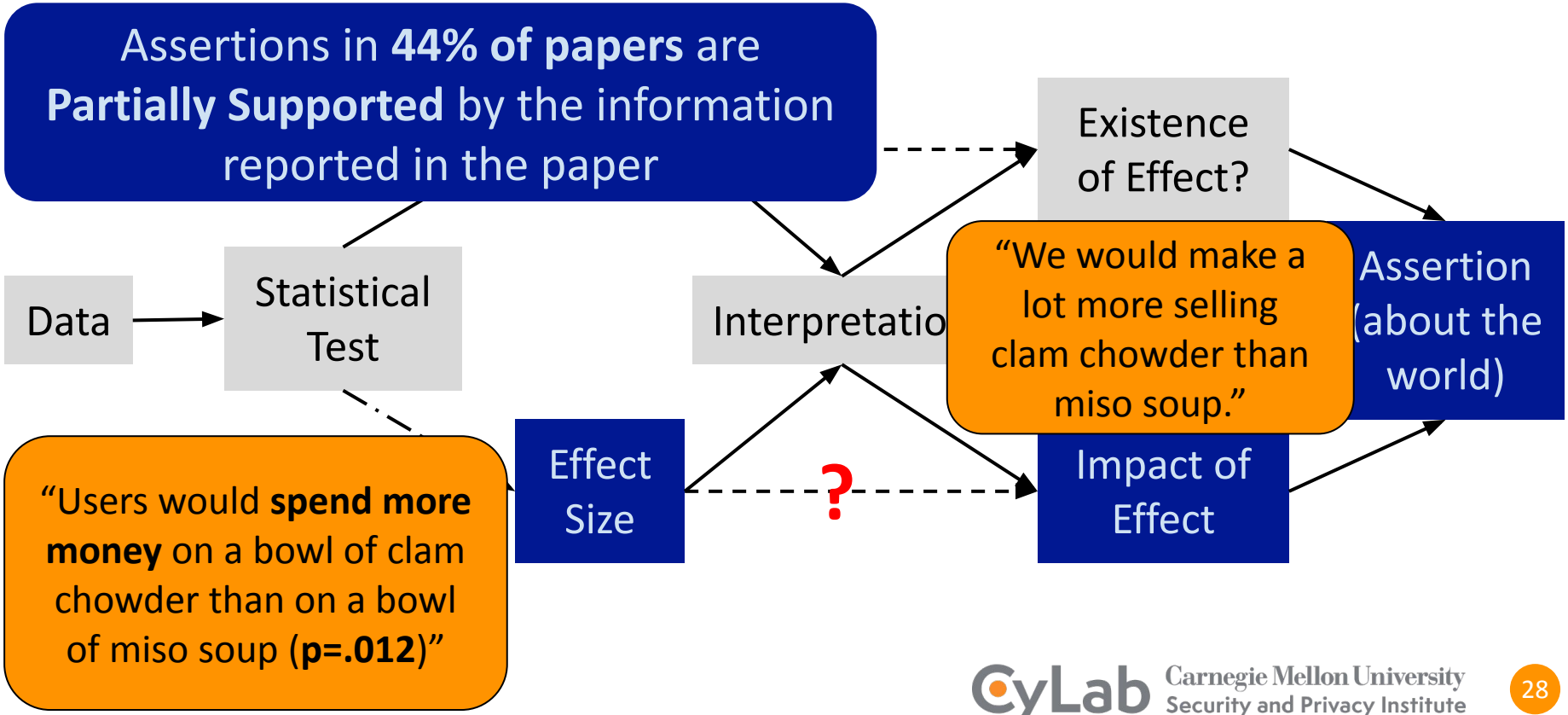
in 65% of papers

in 95% of papers

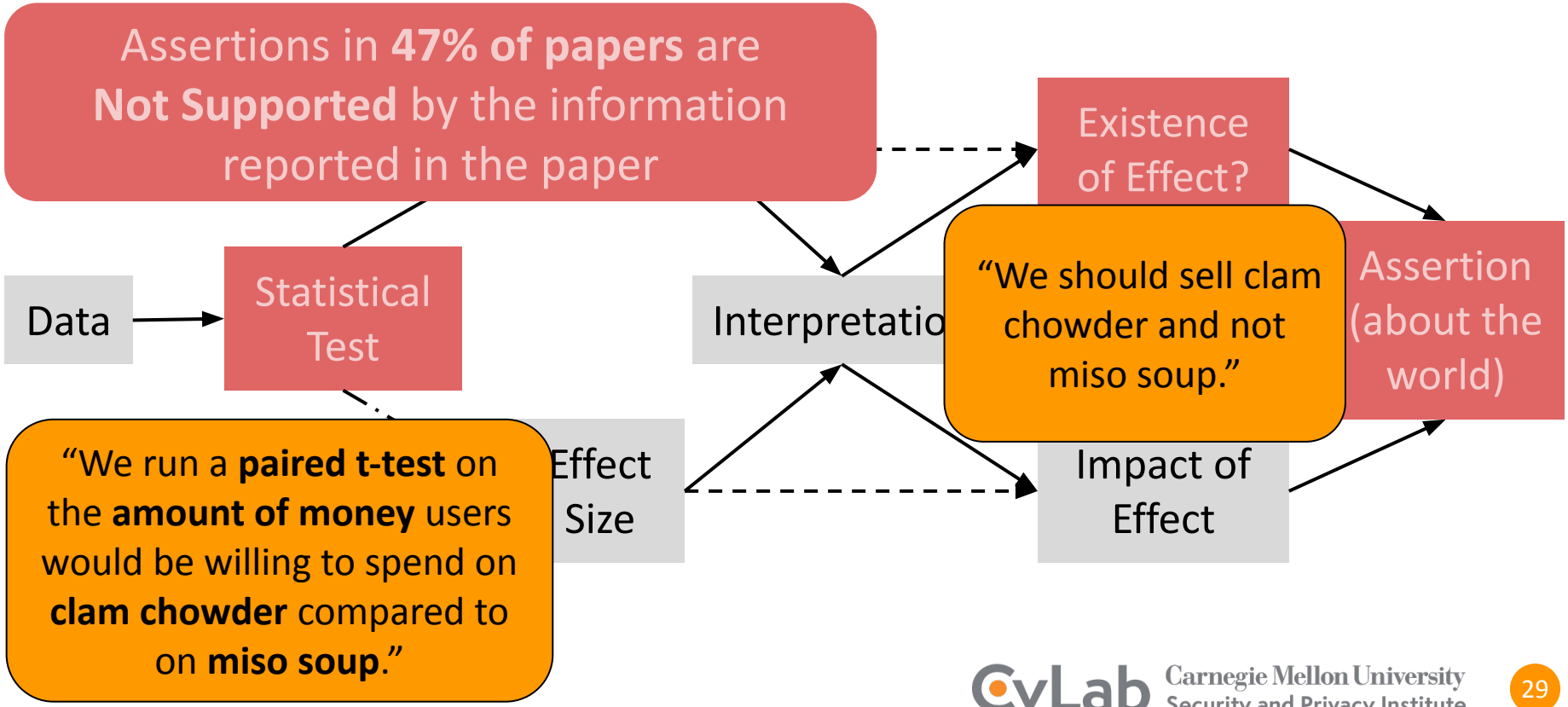
in 45% of papers



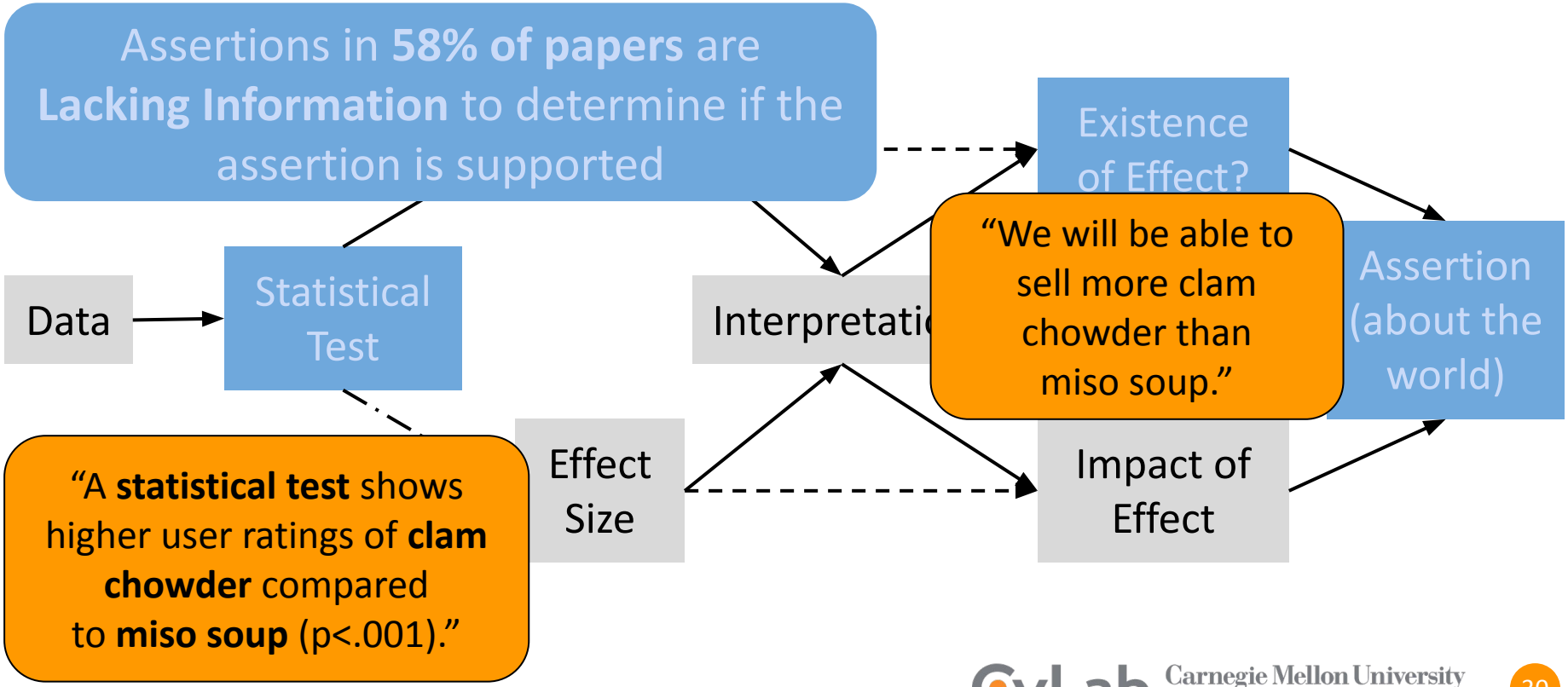
Some Issues Affect the Impact of the Assertion



Some Issues Affect the Validity of the Assertion



Some Issues Affect Evaluation of the Assertion



Statistical Testing Checklist
(for Authors, Reviewers, and Readers)

- ❑ Is statistical significance testing necessary (to provide relevant support for the assertions)?
- ❑ Is there information to understand and reconstruct the test?
  - ❑ Precise test name
  - ❑ Variables tested (independent and dependent)
- ❑ Is there information to evaluate results beyond statistical significance?
  - ❑ Magnitude and direction of effects
  - ❑ Model fit (for regressions)
- ❑ Are results interpreted for a non-statistical audience?
- ❑ Are results placed in field-specific context?

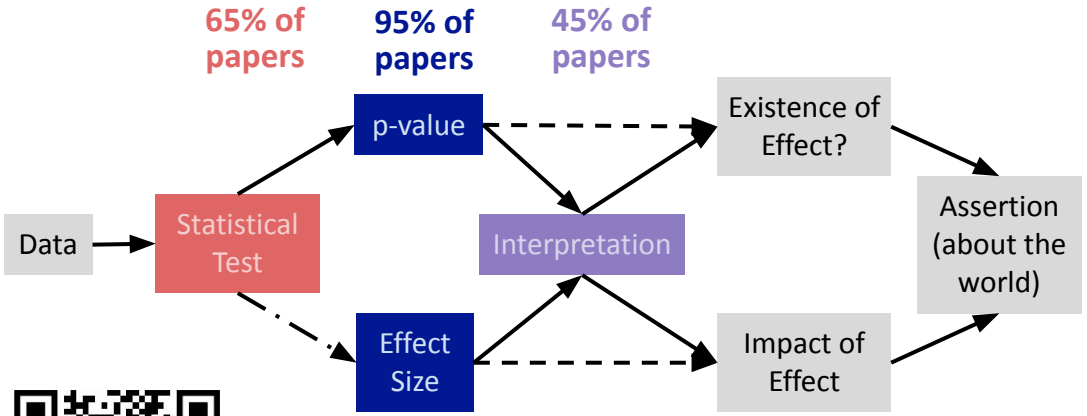
MISSING FOR ASSERTIONS IN 58% OF PAPERS

MISSING FOR ASSERTIONS IN 44% OF PAPERS

Misuse, Misreporting, Misinterpretation of Statistical Methods in Usable Privacy and Security Papers

Jenny Tang, Lujo Bauer, Nicolas Christin

Most UPS papers we analyzed have at least one assertion that is not fully supported by the information reported in the paper



Statistical significance is not sufficient in and of itself

- ❑ Examine effect sizes
- ❑ Report all necessary information to reconstruct, evaluate, and replicate results

← READ THE PAPER HERE!