Tea: A High-level Language and Runtime System for Automating Statistical Analysis

Eunice Jun

with Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke

To appear at ACM UIST in October 2019.

tea-lang.org

Posted by u/WestEnd89 1 hour ago

Question **[Q] What percentage of scores fall below below one standard deviation above the mean?**

Hi all,

I'm having a bit of trouble working out answers to percentage questions about normal distributions.

Posted by u/ice_shadow 6 hours ago

Question **[Q] Is classification in ML the opposite of ANOVA in classical stats?**

So ANOVA and Mixed Models tell you whether a certain factor had a significant effect on the response and whether levels of a factor had a significantly different effect on the response.

From what I understand, things like Logistic Regression, discriminant analysis, kNN, SVM etc seem to use the response to try to predict the classes the data points belong to.

So are these approaches basically opposites of each other?

If the ANOVA contrasts is significant, would one of the classification approaches also be expected to perform well?

And if a classification approach has high accuracy, sensitivity, specificity then can you

Posted by u/WestEnd89 1 hour ago

Question [Q] What percentage of scores fall below below one standard deviation above the mean?

Hi all,

I'm having a bit of trouble working out answers to percentage questions about normal distributions.

Posted by u/ice_shadow 6 hours ago

Question [Q] Is classification in ML the opposite of ANOVA in classical stats?

So ANOVA and Mixed Models tell you whether a certain factor had a significant effect on the response and whether levels of a factor had a significantly different effect on the response.

From what I understand, things like Logistic Regression, discriminant analysis, k
etc seem to use the response to try to predict the classes the data points belong

Posted by u/IzzyBee1 14 hours ago

Question [Q] What statistical test should I use?

Hi all, I'm looking for guidance from someone who knows what
sure don't! I have a semester of stats 101 under my belt (that I
college) and that's more or less the extent of my knowledge.

Project background: my workplace has a moth problem. We ha
building and I check them once a month. Each trap location is i
I also have a combined data set with monthly moth catch of all
~1.5 years' worth of data, but since a I check monthly, it's a rela
Also, since I'm recording moth catch, the data is relatively skew
me if I'm wrong) because many traps have caught 0 moths dur

Posted by u/banannah09 1 day ago

Question [Q] Can I do anything with this data?

Hello everyone! I've been reviewing some data for parents and children who received therapy. The way their mental health is measured is with 2 tests, so both parents and children should complete both of these tests before and after the treatment.

However... Even though 20 children received therapy, there are few cases where there is both pre and post treatment data (between 5-8 for both tests for both parents and children). I had many ideas for how I could analyse this data before, but now I'm not sure I can do anything with this aside from a few graphs (which I've already done)?

Posted by u/eddyks 19 hours ago

Question [Q] What is the best way to analyze my dataset?

Hi there,

Statistics is not my strongest point, so I was wondering if some of you could help me out a little bit.

So far I have almost 200 respondents who filled in my survey. They were given 6 sets, each having 3 statements, prior to being asked where they would buy a certain product (e.g. offline or online), which my moderator being the price of the product (high priced vs. low priced). Each set measures certain characteristics (e.g. price-conscious, time-conscious etc.). Now I want to test my hypotheses that price-conscious consumers buy high priced products rather offline than online.

What would be the best way to do this in SPSS?

Statistics is hard

Posted by u/WestEnd89 1 hour ago

Question [Q] What percentage of scores fall below below one standard deviation above the mean?

Hi all,

I'm having a bit of tro

distributions.

Posted by u/ice_shadow 6

Question [Q] Is clas

stats?

So ANOVA and Mixed

the response and whe

response.

From what I understa

etc seem to use the response

Posted by u/IzzyBee1 14 hou

Question [Q] What st

Hi all, I'm looking for gu

sure don't! I have a sem

college) and that's more

Project background: my

building and I check the

I also have a combined

~1.5 years' worth of dat

Also, since I'm recordin

me if I'm wrong) becaus

Posted by u/banannah09 1 day ago

Question [Q] Can I do anything with this data?

en who received

oth parents and

atment.

ases where there is

h parents and

but now I'm not

already done)?

aset?

of you could help me out

hey were given 6 sets,

ild buy a certain product

he product (high priced

rice-conscious, time-

nscious consumers buy

Which statistical test should I use?

Does an optimization make my program run faster?

H_1: Optimized code runs faster

H_0: Difference between run times due to chance

Which statistical test should I use?

Does an optimization make my program run faster?

Pearson's r
Pointbiserial
Kendall's T
Spearman's p
Student's t-test
Paired t-test
Mann-Whitney U
Wilcoxon signed rank

Welch's
F-test
Repeated measures
one-way ANOVA
Factorial ANOVA
Two-way ANOVA
Kruskal Wallis
Friedman

Fisher's Exact
Linear regression
Logistic regression
MANOVA
ANCOVA
MANCOVA
McNemar
Chi Square

H_1: Optimized code runs faster
H_0: Difference between run times due to chance

Which statistical test should I use?

Does an optimization make my program run faster?

Pearson's r Welch's Fisher's Exact

It depends!

Wilcoxon signed rank Friedman Chi Square

H_1: Optimized code runs faster

H_0: Difference between run times due to chance

Which statistical test should I use?

How do financial incentives affect users' performance?

H_1: Higher financial incentives, better user performance
H_0: Difference in performance due to chance

Which statistical test should I use?

How do financial incentives affect users' performance?

Pearson's r
Pointbiserial
Kendall's T
Spearman's p
Student's t-test
Paired t-test
Mann-Whitney U
Wilcoxon signed rank

Welch's
F-test
Repeated measures
one-way ANOVA
Factorial ANOVA
Two-way ANOVA
Kruskal Wallis
Friedman

Fisher's Exact
Linear regression
Logistic regression
MANOVA
ANCOVA
MANCOVA
McNemar
Chi Square

H_1: Higher financial incentives, better user performance
H_0: Difference in performance due to chance

Which statistical test should I use?

How do financial incentives affect users' performance?

Pearson's r | Welch's | Fisher's Exact

It depends!

Wilcoxon signed rank | Friedman | Chi Square

H_1: Higher financial incentives, better user performance

H_0: Difference in performance due to chance

Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

H_1: Tea first tastes better

H_0: Difference in taste due to chance

Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

Pearson's r
Pointbiserial
Kendall's T
Spearman's p
Student's t-test
Paired t-test
Mann-Whitney U
Wilcoxon signed rank

Welch's
F-test
Repeated measures
one-way ANOVA
Factorial ANOVA
Two-way ANOVA
Kruskal Wallis
Friedman

Fisher's Exact
Linear regression
Logistic regression
MANOVA
ANCOVA
MANCOVA
McNemar
Chi Square

H_1: Tea first tastes better

H_0: Difference in taste due to chance

Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

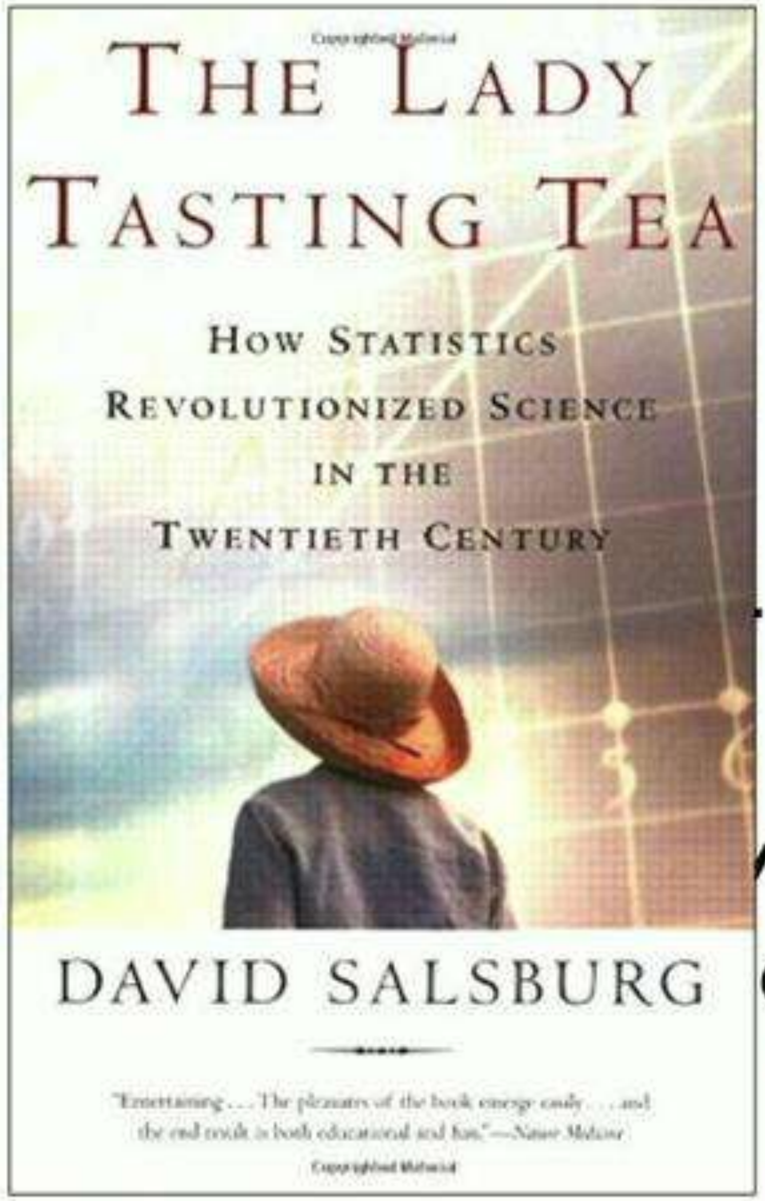
Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

H_1: Tea first tastes better

H_0: Difference in taste due to chance

Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?



- Fisher's Exact
- Linear regression
- Logistic regression
- MANOVA
- ANCOVA
- MANCOVA
- McNemar
- Chi Square

H_1: Tea first tastes better

H_0: Difference in taste due to chance

tea

tea

EASY {

Does this optimization make my program execute faster?
How do financial incentives affect users' performance on a task?
Does tea taste better with milk poured first then tea or tea first then milk?

tea

EASY

Does this optimization make my program execute faster?
How do financial incentives affect users' performance on a task?
Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

EASY

- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

EASY

- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Tea eliminates this problem entirely.

EASY
- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Tea eliminates this problem entirely.

EASY

Does this optimization make my program execute faster?
How do financial incentives affect users' performance on a task?
Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Each participant contributes **exactly** **one** data point

Tea eliminates this problem entirely.

EASY
- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Each participant contributes **exactly one** data point

Each participant contributes **exactly two** data points

Tea eliminates this problem entirely.

EASY
- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Each participant contributes **exactly one** data point

Each participant contributes **exactly two** data points

VERY BAD to mix up

Tea eliminates this problem entirely.

EASY
- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Each participant contributes **exactly one** data point

Each participant contributes **exactly two** data points

VERY BAD to mix up

Violate study design

Tea eliminates this problem entirely.

EASY
- Does this optimization make my program execute faster?
- How do financial incentives affect users' performance on a task?
- Does tea taste better with milk poured first then tea or tea first then milk?

HARD

Pearson's r	Welch's	Fisher's Exact
Pointbiserial	F-test	Linear regression
Kendall's T	Repeated measures	Logistic regression
Spearman's p	one-way ANOVA	MANOVA
Student's t-test	Factorial ANOVA	ANCOVA
Paired t-test	Two-way ANOVA	MANCOVA
Mann-Whitney U	Kruskal Wallis	McNemar
Wilcoxon signed rank	Friedman	Chi Square

```
t.test(x y = NULL
       alternative = c("two.sided" "less" "greater")
       mu = 0 paired = FALSE var.equal = FALSE
       conf.level = 0.95 …)
```

Difference between Student's t-test and Paired t-test

Each participant contributes **exactly one** data point

Each participant contributes **exactly two** data points

VERY BAD to mix up

Violate study design

Returns **wrong** statistical results!

Tea eliminates this problem entirely.

Stats is better with Tea

🧮 🍵 😍

Tea is correct by construction.

Tea is high-level.

Tea infers tests.

Tea *improves* upon *expert choices, prevents common mistakes.*

Home | Tea Demo

localhost:8888/notebooks/Tea Demo.ipynb

jupyter Tea Demo Last Checkpoint: 07/31/2019 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [ ]:
```
import tea
```

In [ ]:
```
# Load data
tea.data("./datasets/UScrime.csv")
```

In [ ]:
```
# Declare and annotate the variables of interest
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]

tea.define_variables(variables)
```

In [ ]:
```
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05,
}
```

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                    'study type': 'observational study',
                    'contributor variables': 'So',
                    'outcome variables': 'Prob',
                }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

data

variables

study design

assertions

hypothesis

* * NO STATISTICAL TEST * *

```
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
```

```
variables = [
    {
    ⟶ 'name' : 'So',
      'data type' : 'nominal',
      'categories' : ['0', '1']
    },
    {
    ⟶ 'name' : 'Prob',
      'data type' : 'ratio',
      'range' : [0,1]
    }
]
tea.define_variables(variables)
```

```
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
```

Nominal

Ordinal

Interval

Ratio

```
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
```

Nominal

Ordinal

Interval

Ratio

```
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
```

```
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
    --> 'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
    --> 'range' : [0,1]
    }
]
tea.define_variables(variables)
```

```
study_design = {
            'study type': 'observational study',
            'contributor variables': 'So',
            'outcome variables': 'Prob',
            }
    tea.define_study_design(study_design)
```

```
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
              }
    tea.define_study_design(study_design)
```

```
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
               }
    tea.define_study_design(study_design)
```

```
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
```

```
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'],hypothesis)
```

```
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'],hypothesis)
```

Nominal,Ordinal:

Chocolate > Mint

Grade 1 < Grade 2

Ordinal,Ratio,Interval:

Grade ~ Temperature

Time of day ~ - Temperature

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                    'study type': 'observational study',
                    'contributor variables': 'So',
                    'outcome variables': 'Prob',
               }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

data

variables

study design

assertions

hypothesis

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓completeness
✓syntax
✓well-formed hypotheses

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
               }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

→

logical constraints

$$continuous(x) \wedge \neg categorical(x)$$

$$normal(x) \rightarrow \neg categorical(x)$$

...

✓completeness
✓syntax
✓well-formed hypotheses

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                    'study type': 'observational study',
                    'contributor variables': 'So',
                    'outcome variables': 'Prob',
                }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓completeness
✓syntax
✓well-formed hypotheses

→

logical constraints

$$continuous(x) \land \neg categorical(x)$$
$$normal(x) \rightarrow \neg categorical(x)$$

...

↓

MaxSat

Z3

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                    'study type': 'observational study',
                    'contributor variables': 'So',
                    'outcome variables': 'Prob',
                }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓completeness
✓syntax
✓well-formed hypotheses

→ logical constraints

$$continuous(x) \wedge \neg categorical(x)$$

$$normal(x) \rightarrow \neg categorical(x)$$

...

↓

MaxSat

Z3 → {valid statistical tests}

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                  'study type': 'observational study',
                  'contributor variables': 'So',
                  'outcome variables': 'Prob',
              }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓completeness
✓syntax
✓well-formed hypotheses

→

logical constraints

$$continuous(x) \wedge \neg categorical(x)$$
$$normal(x) \rightarrow \neg categorical(x)$$

...

↓

MaxSat

Z3

→

{valid statistical tests}

↑

```
Test: students_t
***Test assumptions:
Exactly two variables involved in analysis: So, Prob
Exactly one explanatory variable: So
Exactly one explained variable: Prob
Independent (not paired) observations: So
Variable is categorical: So
Variable has two categories: So
Continuous (not categorical) data: Prob
Equal variance: So, Prob
Groups are normally distributed: So, Prob

***Test results:
name = Student's T Test
test_statistic = 4.202130736875173
p_value = 0.00012364897266532775
adjusted_p_value = 6.182448633266387e-05
alpha = 0.05
dof = 45
Effect size:
Cohen's d = 1.2426167296374897
A12 = 0.8366935483870968
Null hypothesis = There is no difference in means between 0 and 1 on Prob.
Interpretation = t(45) = 4.202130736875173, 6.182448633266387e-05. Reject the null
hypothesis at alpha = 0.05. The mean of Prob for So = 1 is significantly greater than
the mean for So = 0. The effect size is {"Cohen's d": 1.2426167296374897, 'A12':
0.8366935483870968}. The effect size is the magnitude of the difference, which gives
a holistic view of the results [1].
[1] Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not
enough. Journal of Graduate Medical Education, 4(3), 279-282.
```

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓ completeness
✓ syntax
✓ well-formed hypotheses

logical constraints

$$continuous(x) \land \neg categorical(x)$$

$$normal(x) \rightarrow \neg categorical(x)$$

...

MaxSat

Z3

```
Test: students_t
***Test assumptions:
Exactly two variables involved in analysis: So, Prob
Exactly one explanatory variable: So
Exactly one explained variable: Prob
Independent (not paired) observations: So
Variable is categorical: So
Variable has two categories: So
Continuous (not categorical) data: Prob
Equal variance: So, Prob
Groups are normally distributed: So, Prob

***Test results:
name = Student's T Test
test_statistic = 4.202130736875173
p_value = 0.00012364897266532775
adjusted_p_value = 6.182448633266387e-05
alpha = 0.05
dof = 45
Effect size:
Cohen's d = 1.2426167296374897
A12 = 0.8366935483870968
Null hypothesis = There is no difference in means between 0 and 1 on Prob.
Interpretation = t(45) = 4.202130736875173, 6.182448633266387e-05. Reject the null hypothesis at alpha = 0.05. The mean of Prob for So = 1 is significantly greater than the mean for So = 0. The effect size is {"Cohen's d": 1.2426167296374897, 'A12': 0.8366935483870968}. The effect size is the magnitude of the difference, which gives a holistic view of the results [1].
[1] Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of Graduate Medical Education, 4(3), 279-282.
```

{valid statistical tests}

How do we logically represent statistical knowledge?

Statistical test applies iff all preconditions apply

Student's t-test ↔

```
bivariate
one_x_variable
one_y_variable
independent_obs
categorical
two_categories
continuous
equal_variance
groups_normal
```

Rules formulated from Scott Klemmer and Jacob Wobbrock. 2019. Designing Running and Analyzing Experiments. (2019). https://www.coursera.org/learn/designexperiments; Andy Field Jeremy Miles and Zoë Field. 2012. Discovering statistics using R. Sage publications; J. Bruin. 2019. Choosing the Correct Statistical Test in SAS Stata SPSS and R. (2019). https://stats.idre.ucla.edu/other/mult-pkg/whatstat/; Kent State University Libraries. 2019. SPSS Tutorials: Analyzing Data. (2019). https://libguides.library.kent.edu/SPSS/AnalyzeData

Statistical test applies	iff	all preconditions apply

Student's t-test	↔	**test properties**
		`bivariate`
		`one_x_variable`
		`one_y_variable`
		`independent_obs`
		
		variable properties
		`categorical`
		`two_categories`
		`continuous`
		`equal_variance`
		`groups_normal`

Rules formulated from Scott Klemmer and Jacob Wobbrock. 2019. Designing Running and Analyzing Experiments. (2019). https://www.coursera.org/learn/designexperiments; Andy Field Jeremy Miles and Zoë Field. 2012. Discovering statistics using R. Sage publications; J. Bruin. 2019. Choosing the Correct Statistical Test in SAS Stata SPSS and R. (2019). https://stats.idre.ucla.edu/other/mult-pkg/whatstat/; Kent State University Libraries. 2019. SPSS Tutorials: Analyzing Data. (2019). https://libguides.library.kent.edu/SPSS/AnalyzeData

Statistical test applies iff all preconditions apply

Student's t-test $\leftrightarrow$

test properties

```
bivariate(xy)
one_x_variable(xy)
one_y_variable(xy)
independent_obs(xy)
```

variable properties

```
categorical(x)
two_categories(x)
continuous(y)
equal_variance(xy)
groups_normal(xy)
```

Rules formulated from Scott Klemmer and Jacob Wobbrock. 2019. Designing Running and Analyzing Experiments. (2019). https://www.coursera.org/learn/designexperiments; Andy Field Jeremy Miles and Zoë Field. 2012. Discovering statistics using R. Sage publications; J. Bruin. 2019. Choosing the Correct Statistical Test in SAS Stata SPSS and R. (2019). https://stats.idre.ucla.edu/other/mult-pkg/whatstat/; Kent State University Libraries. 2019. SPSS Tutorials: Analyzing Data. (2019). https://libguides.library.kent.edu/SPSS/AnalyzeData

Statistical test applies iff all preconditions apply

Student's t-test ↔

test properties

```
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
```

variable properties

```
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
```

{ Pearson's r, Pointbiserial, Kendall's T, Spearman's p, Student's t-test, Paired t-test, Mann-Whitney U, Wilcoxon signed rank, Welch's, F-test, Repeated measures, one-way ANOVA, Factorial ANOVA, Two-way ANOVA, Kruskal Wallis, Friedman, Chi Square, Fisher's Exact, Bootstrapping } Null Hypothesis Significance Testing Module

Rules formulated from Scott Klemmer and Jacob Wobbrock. 2019. Designing Running and Analyzing Experiments. (2019). https://www.coursera.org/learn/designexperiments; Andy Field Jeremy Miles and Zoë Field. 2012. Discovering statistics using R. Sage publications; J. Bruin. 2019. Choosing the Correct Statistical Test in SAS Stata SPSS and R. (2019). https://stats.idre.ucla.edu/other/mult-pkg/whatstat/; Kent State University Libraries. 2019. SPSS Tutorials: Analyzing Data. (2019). https://libguides.library.kent.edu/SPSS/AnalyzeData

Statistical test selection as constraint satisfaction

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

✓completeness
✓syntax
✓well-formed hypotheses

→ logical constraints

$$continuous(x) \land \neg categorical(x)$$

$$normal(x) \rightarrow \neg categorical(x)$$

...

↓

MaxSat

Z3

→ {valid statistical tests}

↑

```
Test: students_t
***Test assumptions:
Exactly two variables involved in analysis: So, Prob
Exactly one explanatory variable: So
Exactly one explained variable: Prob
Independent (not paired) observations: So
Variable is categorical: So
Variable has two categories: So
Continuous (not categorical) data: Prob
Equal variance: So, Prob
Groups are normally distributed: So, Prob

***Test results:
name = Student's T-Test
test_statistic = 4.20213073687517З
p_value = 0.00012364897266522775
adjusted_p_value = 6.182448633266387e-05
alpha = 0.05
dof = 45
Effect size:
Cohen's d = 1.2426167296374897
A12 = 0.8366935483870968
Null hypothesis = There is no difference in means between 0 and 1 on Prob.
Interpretation = t(45) = 4.202130736875173, 6.182448633266387e-05. Reject the null hypothesis at alpha = 0.05. The mean of Prob for So = 1 is significantly greater than the mean for So = 0. The effect size is {"Cohen's d": 1.2426167296374897, 'A12': 0.8366935483870968}. The effect size is the magnitude of the difference, which gives a holistic view of the results [1].
[1] Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of Graduate Medical Education, 4(3), 279-282.
```

How do we formulate a MaxSat problem?

Z3

Z3

Satisfiability of logical formulas

Z3

Satisfiability of logical formulas

boolean, real number, integer, uninterpreted functions

Z3

Test clauses

```
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
```

•••

Other data/statistical analysis constraints

$$(continuous(x) \lor categorical(x)) \land \neg(continuous(x) \land categorical(x))$$

$$normal(x) \rightarrow \neg categorical(x)$$

$$continuous(x) \lor ordinal(x) \rightarrow continuous(x)$$

•••

Test clauses

```
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
```

•••

Z3

Other data/statistical analysis constraints

$$(continuous(x) \vee categorical(x)) \wedge \neg(continuous(x) \wedge categorical(x))$$

$$normal(x) \rightarrow \neg categorical(x)$$

$$continuous(x) \vee ordinal(x) \rightarrow continuous(x)$$

•••

User assumptions

```
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
```

Z3

Test clauses

```
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
```

•••

Other data/statistical analysis constraints

$$(continuous(x) \lor categorical(x)) \land \neg(continuous(x) \land categorical(x))$$

$$normal(x) \rightarrow \neg categorical(x)$$

$$continuous(x) \lor ordinal(x) \rightarrow continuous(x)$$

...

User assumptions

```
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
```

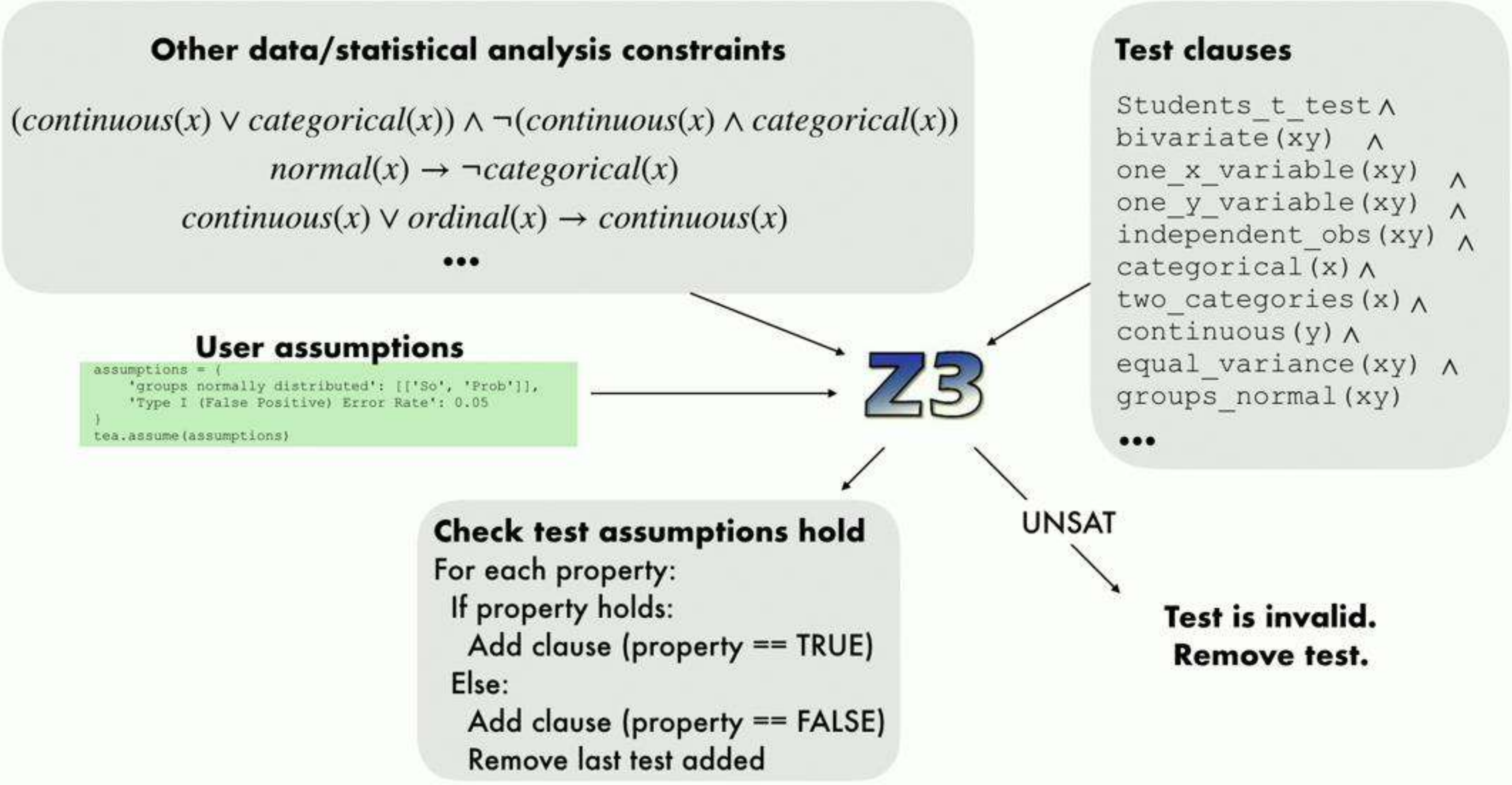
Z3

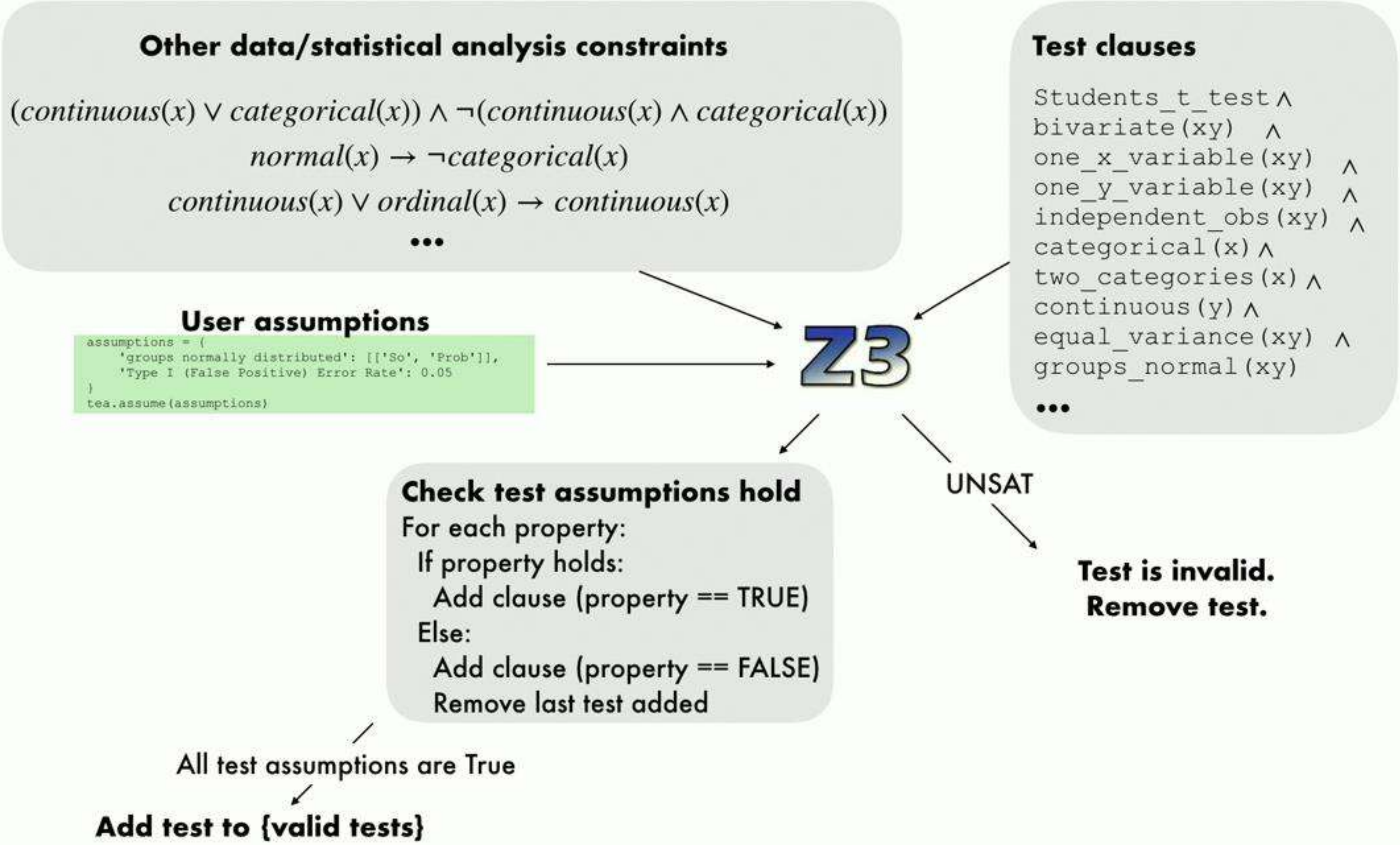
Test clauses

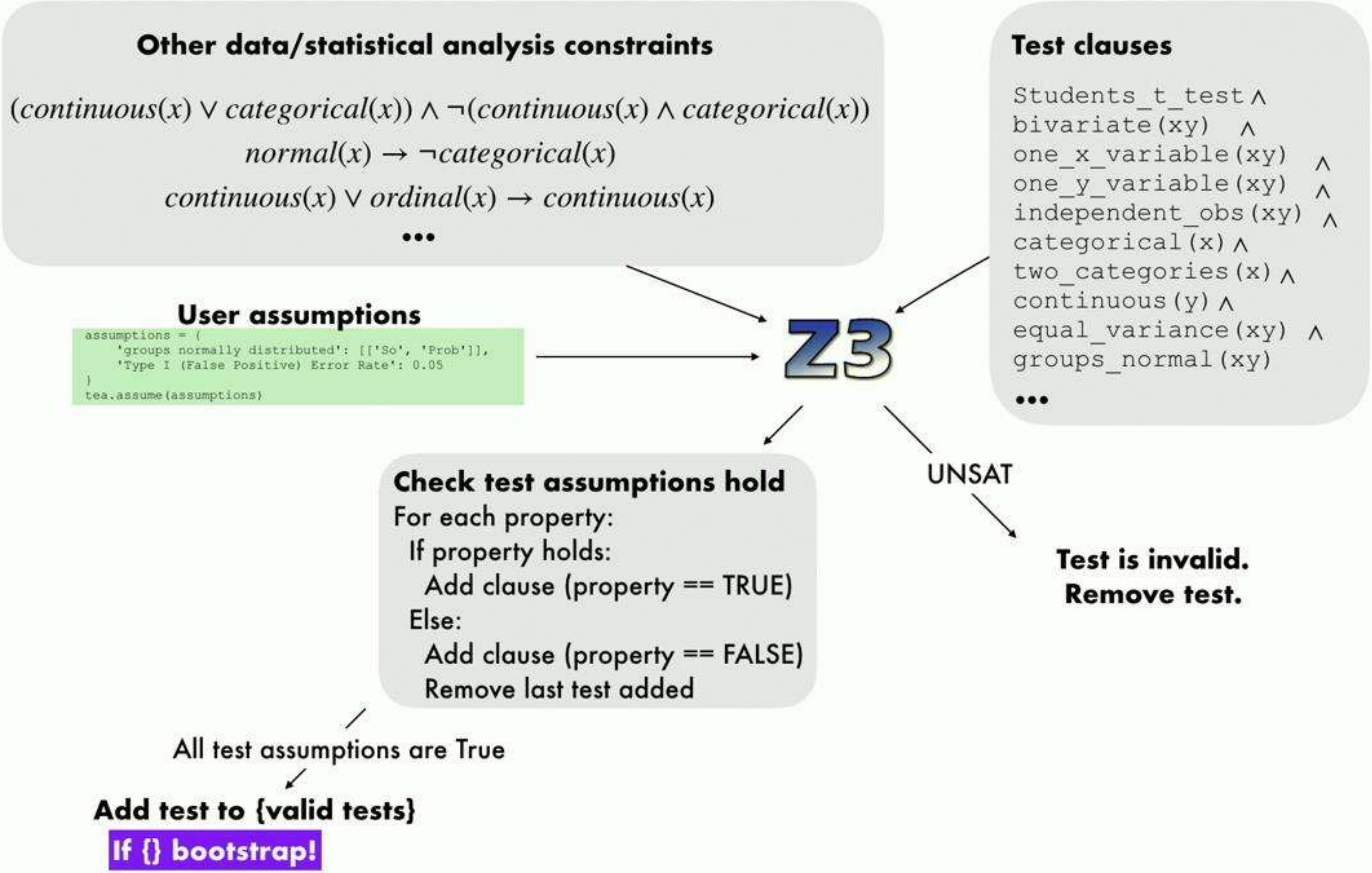
```
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
...
```

UNSAT

Test is invalid.
Remove test.


Other data/statistical analysis constraints
(continuous(x) ∨ categorical(x)) ∧ ¬(continuous(x) ∧ categorical(x))
normal(x) → ¬categorical(x)
continuous(x) ∨ ordinal(x) → continuous(x)
...
Test clauses
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
...
User assumptions
assumptions = {
'groups normally distributed': [['So', 'Prob']],
'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
Z3
Check test assumptions hold
For each property:
If property holds:
Add clause (property == TRUE)
Else:
Add clause (property == FALSE)
Remove last test added
UNSAT
Test is invalid.
Remove test.


Other data/statistical analysis constraints
(continuous(x) ∨ categorical(x)) ∧ ¬(continuous(x) ∧ categorical(x))
normal(x) → ¬categorical(x)
continuous(x) ∨ ordinal(x) → continuous(x)
...
Test clauses
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
...
User assumptions
assumptions = {
'groups normally distributed': [['So', 'Prob']],
'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
Z3
UNSAT
Test is invalid.
Remove test.
Check test assumptions hold
For each property:
If property holds:
Add clause (property == TRUE)
Else:
Add clause (property == FALSE)
Remove last test added
All test assumptions are True
Add test to {valid tests}


Other data/statistical analysis constraints
(continuous(x) ∨ categorical(x)) ∧ ¬(continuous(x) ∧ categorical(x))
normal(x) → ¬categorical(x)
continuous(x) ∨ ordinal(x) → continuous(x)
...
Test clauses
Students_t_test ∧
bivariate(xy) ∧
one_x_variable(xy) ∧
one_y_variable(xy) ∧
independent_obs(xy) ∧
categorical(x) ∧
two_categories(x) ∧
continuous(y) ∧
equal_variance(xy) ∧
groups_normal(xy)
...
User assumptions
assumptions = {
'groups normally distributed': [['So', 'Prob']],
'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
Z3
UNSAT
Test is invalid.
Remove test.
Check test assumptions hold
For each property:
If property holds:
Add clause (property == TRUE)
Else:
Add clause (property == FALSE)
Remove last test added
All test assumptions are True
Add test to {valid tests}
If {} bootstrap!

Tea Output

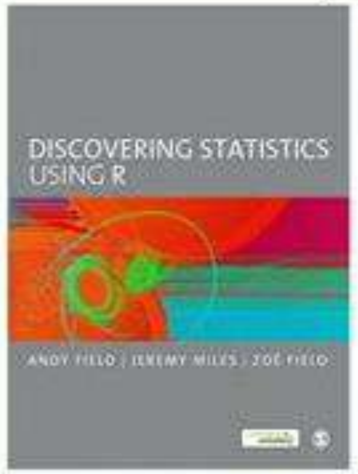
```
Test: students_t
***Test assumptions:
Exactly two variables involved in analysis: So Prob
Exactly one explanatory variable: So
Exactly one explained variable: Prob
Independent (not paired) observations: So
Variable is categorical: So
Variable has two categories: So
Continuous (not categorical) data: Prob
Equal variance: So Prob
Groups are normally distributed: So Prob: NormalTest(W=0.8997463583946228
p_value=0.07962072640657425)

***Test results:
name = Student's T Test
test_statistic = 4.20213
p_value = 0.00012
adjusted_p_value = 0.00006
alpha = 0.05
dof = 45
Effect size:
Cohen's d = 1.24262
A12 = 0.83669
Null hypothesis = There is no difference in means between So = 0 and So = 1 on Prob.
Interpretation = t(45) = 4.20213 p = 0.00006. Reject the null hypothesis at alpha =
0.05. The mean of Prob for So = 1 (M=0.06371 SD=0.02251) is significantly greater than
the mean for So = 0 (M=0.03851 SD=0.01778). The effect size is Cohen's d = 1.24262 A12 =
0.83669. The effect size is the magnitude of the difference which gives a holistic view
of the results [1].
[1] Sullivan G. M. & Feinn R. (2012). Using effect size—or why the P value is not
enough. Journal of graduate medical education 4(3) 279-282.
```

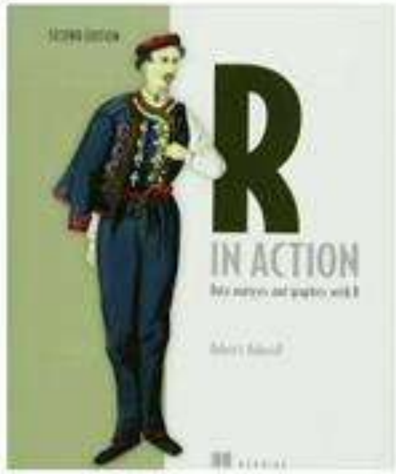
Explain rationale for test selection.

Contextualize results for accurate interpretation.

Evaluation

12 tutorials
code snippets + text



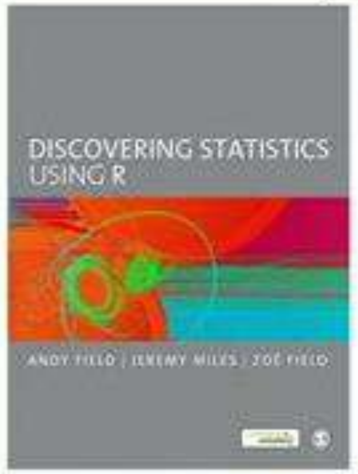


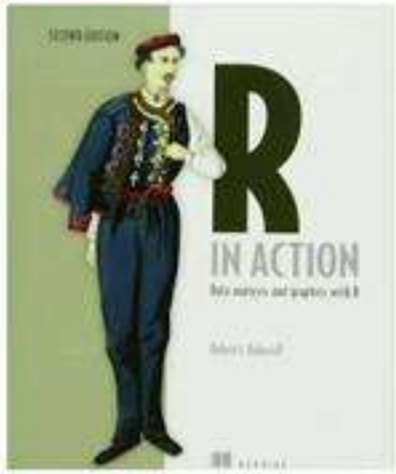
→

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

Evaluation

12 tutorials
code snippets + text





→

```python
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

I. How does Tea compare to experts?

Evaluation

12 tutorials
code snippets + text

→

```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                    'study type': 'observational study',
                    'contributor variables': 'So',
                    'outcome variables': 'Prob',
                }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

I. How does Tea compare to experts?

Replicate, even improve upon expert choices

Evaluation

12 tutorials
code snippets + text


```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

I. How does Tea compare to experts?

Replicate, even improve upon expert choices

II. How does Tea compare to novices?

Evaluation

12 tutorials
code snippets + text

→

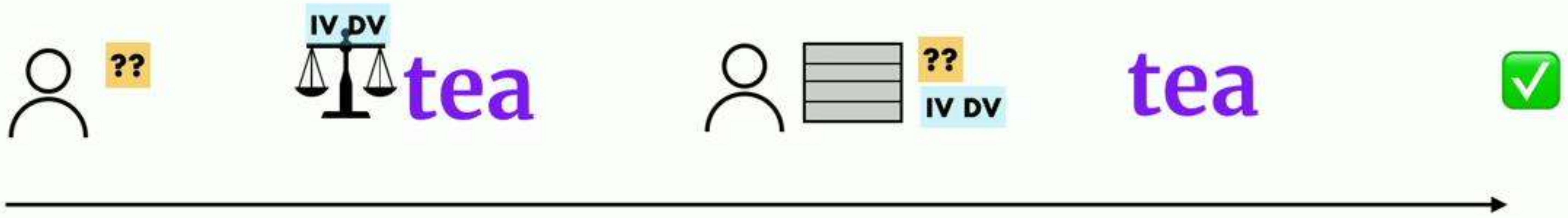
```
import tea
tea.data('UScrime.csv')
variables = [
    {
        'name' : 'So',
        'data type' : 'nominal',
        'categories' : ['0', '1']
    },
    {
        'name' : 'Prob',
        'data type' : 'ratio',
        'range' : [0,1]
    }
]
tea.define_variables(variables)
study_design = {
                'study type': 'observational study',
                'contributor variables': 'So',
                'outcome variables': 'Prob',
            }
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)
```

I. How does Tea compare to experts?

Replicate, even improve upon expert choices

II. How does Tea compare to novices?

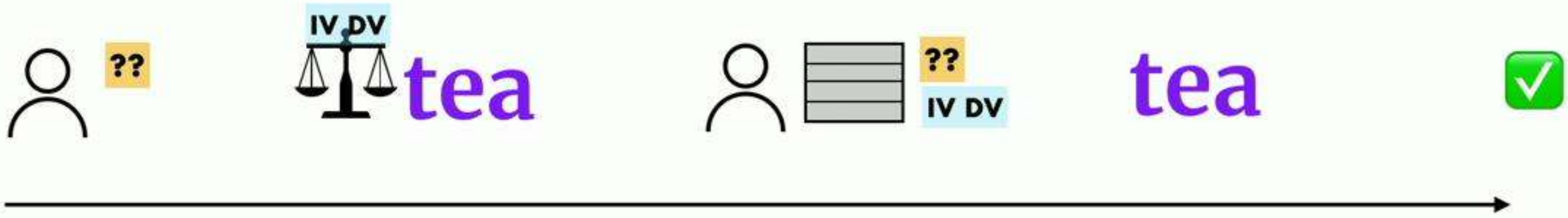
Avoid common mistakes and false conclusions



Tea automates statistical test selection and execution.

Tea can aid with experimental design.

Tea programs can act as a format for pre-registration.

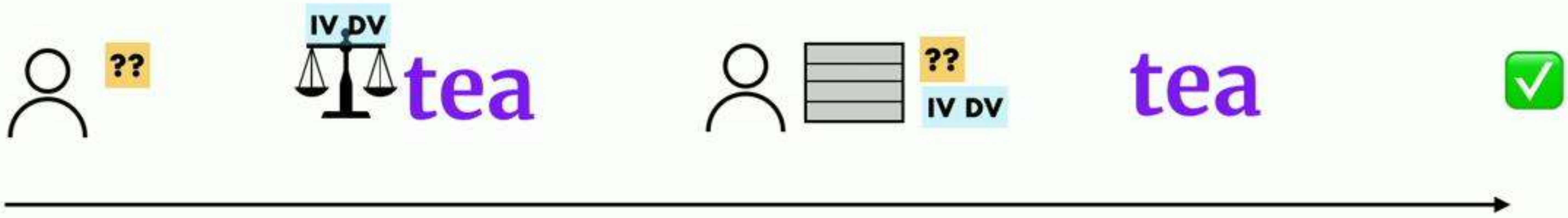


Tea automates statistical test selection and execution.

Tea can aid with experimental design.

Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.



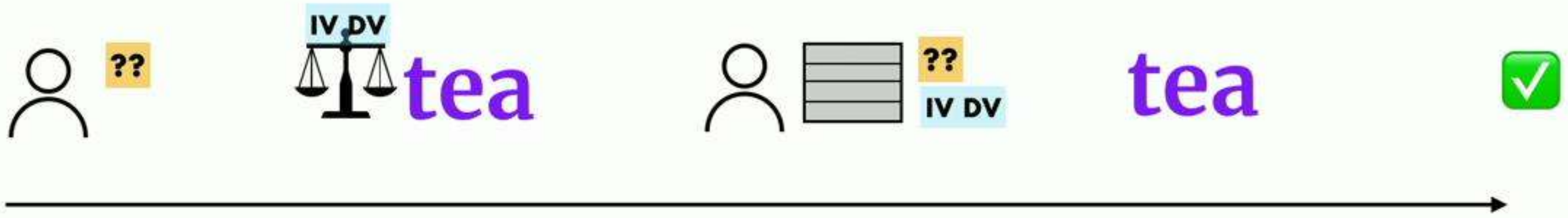
Tea automates statistical test selection and execution.

Tea can aid with experimental design.

Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.

Internal validity!



Tea automates statistical test selection and execution.

Tea can aid with experimental design.

Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.

Internal validity!

`pip install tealang`

`tea-lang.org`

Ongoing work

Field deployment, user testing

Future development

- linear modeling

tea

scone

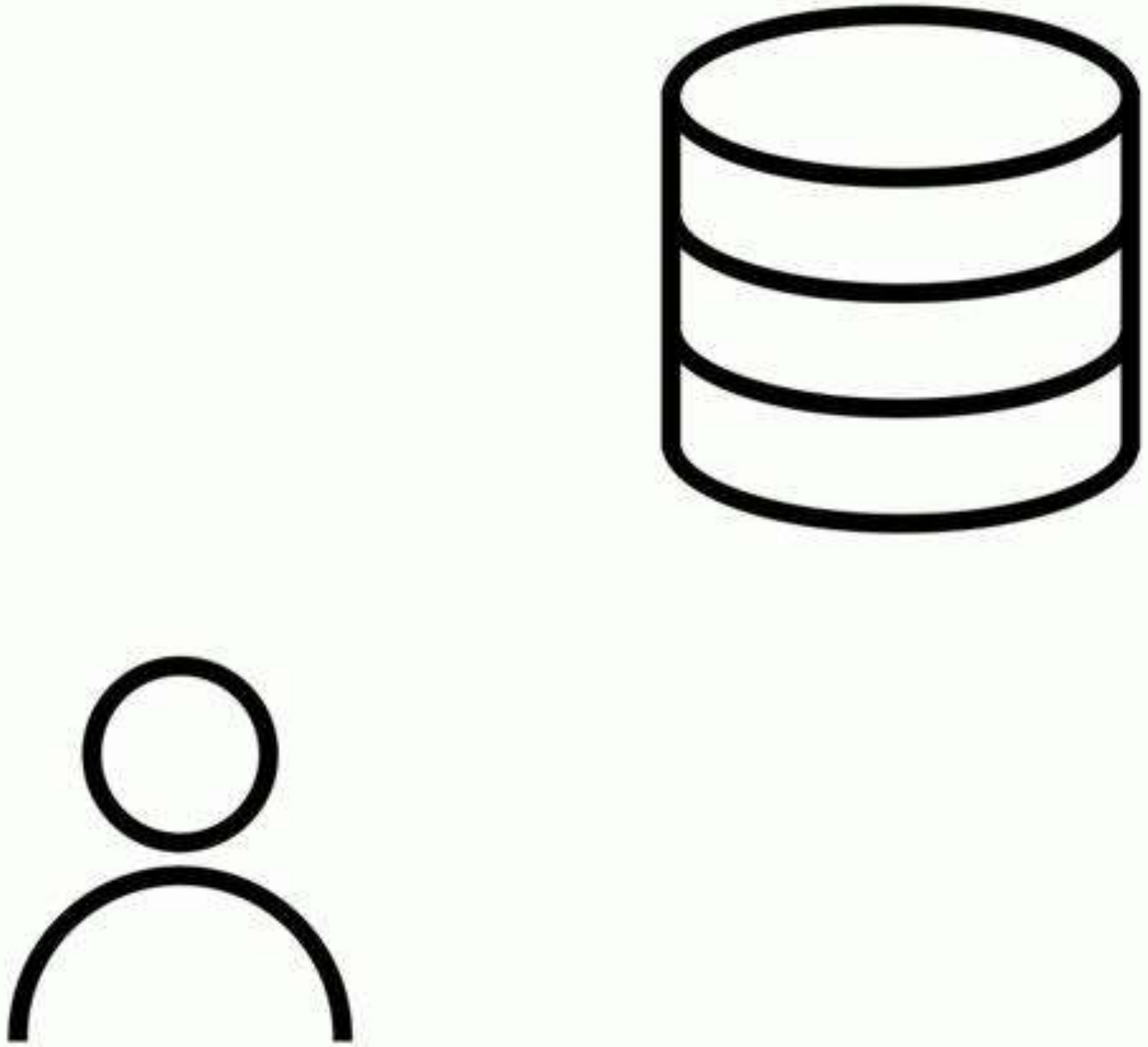
(behind the tea cup)

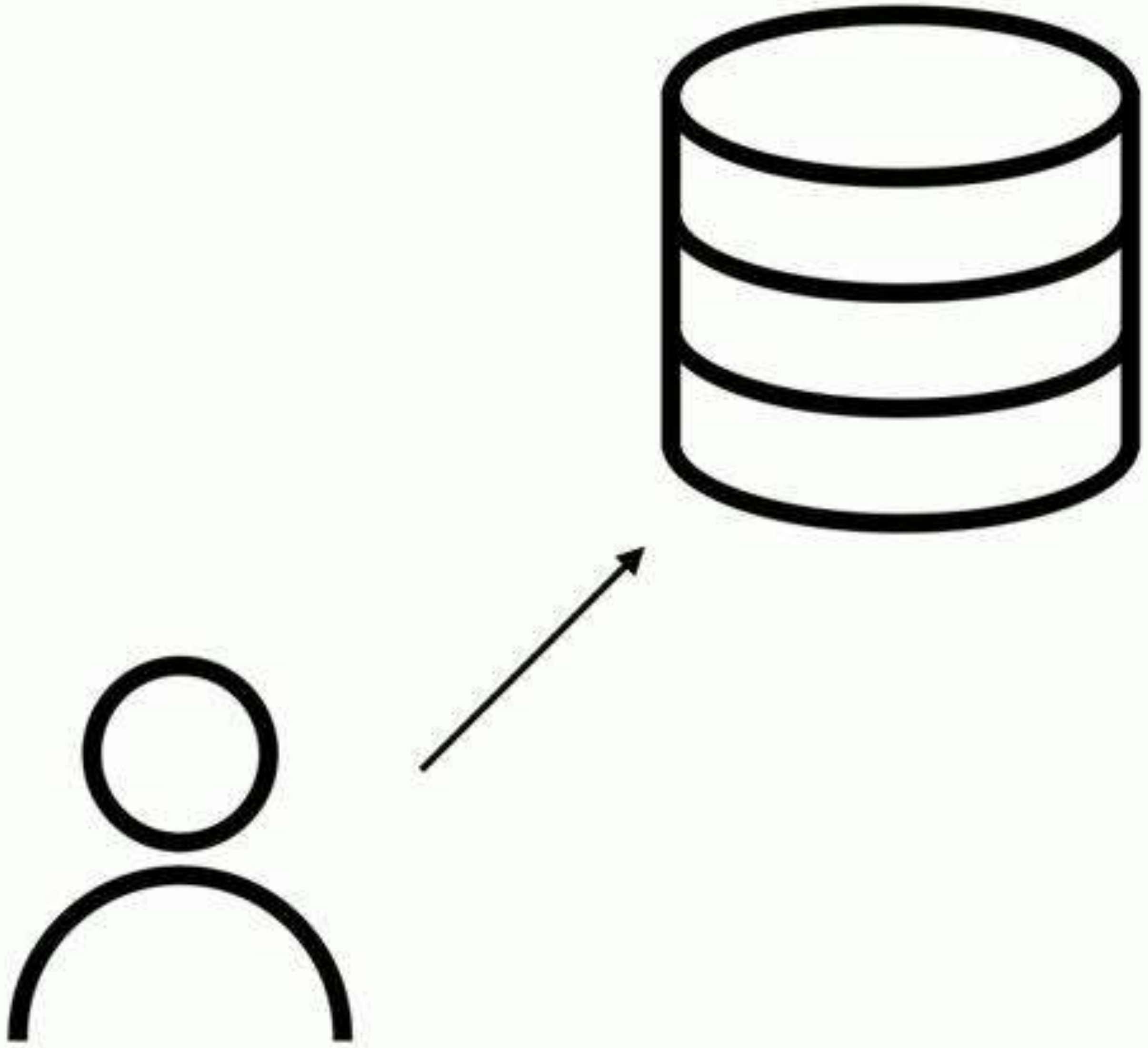
tea

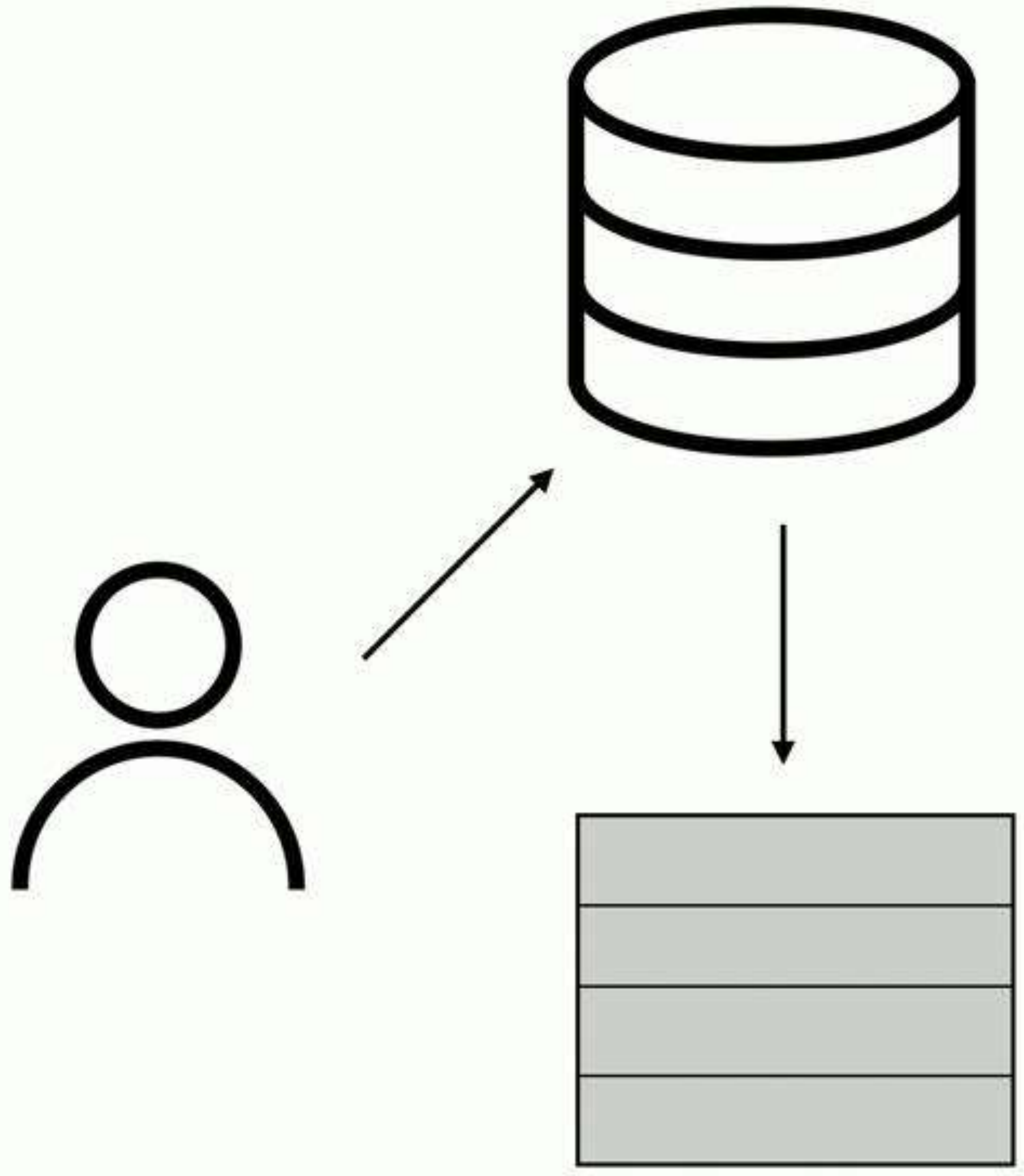
Scone: Smart Sampling for Smarter Statistics

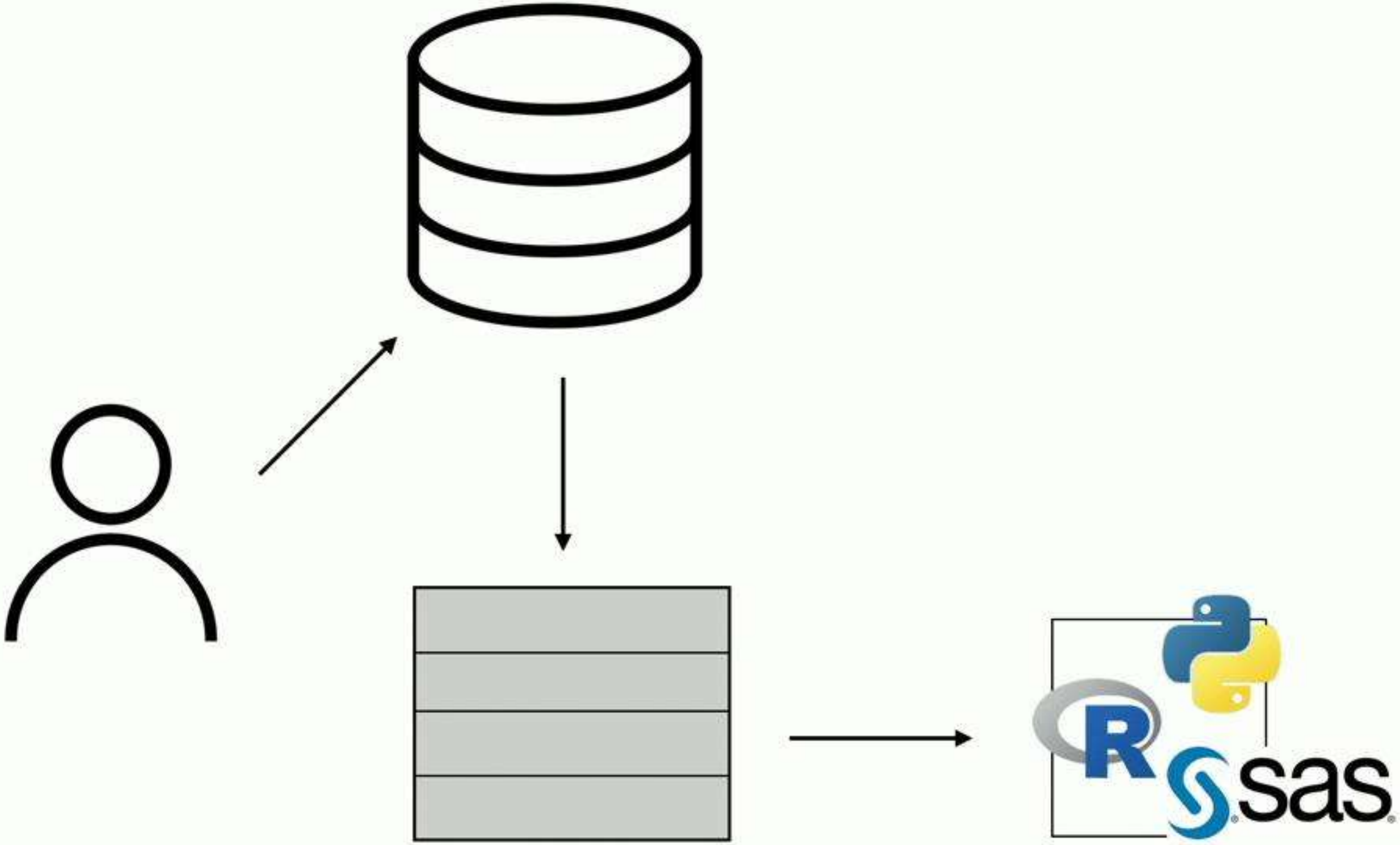
with Laurel Orr, Emery Berger, and Ben Zorn

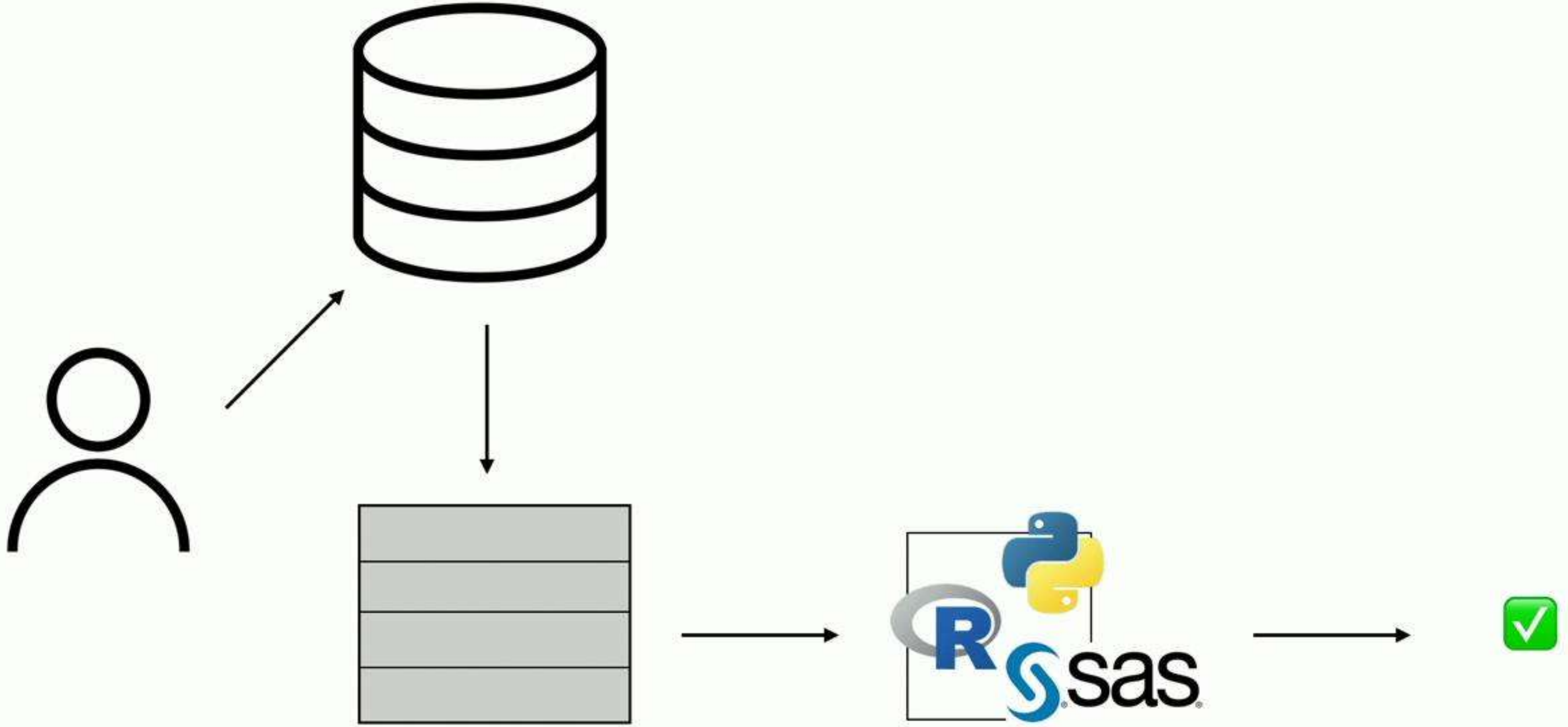
ongoing summer project

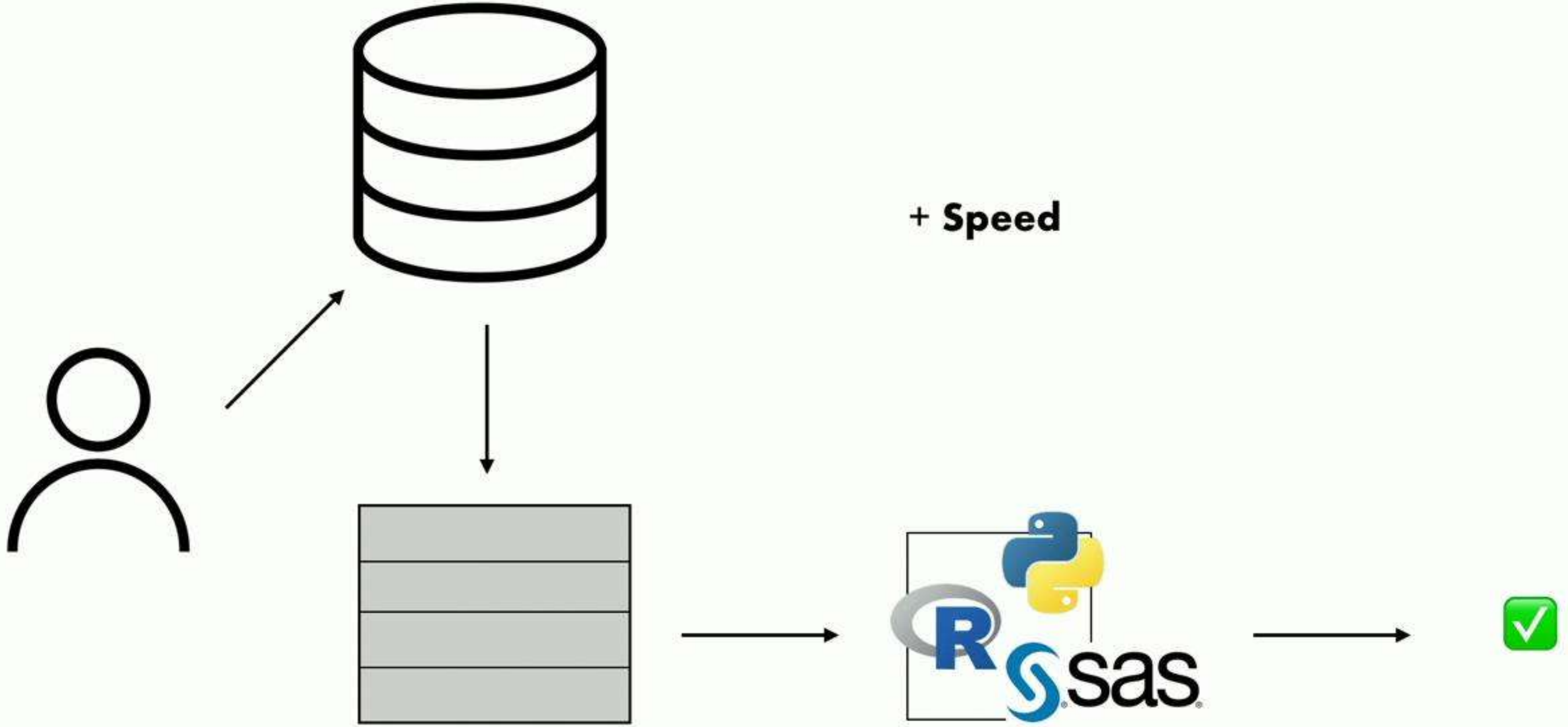


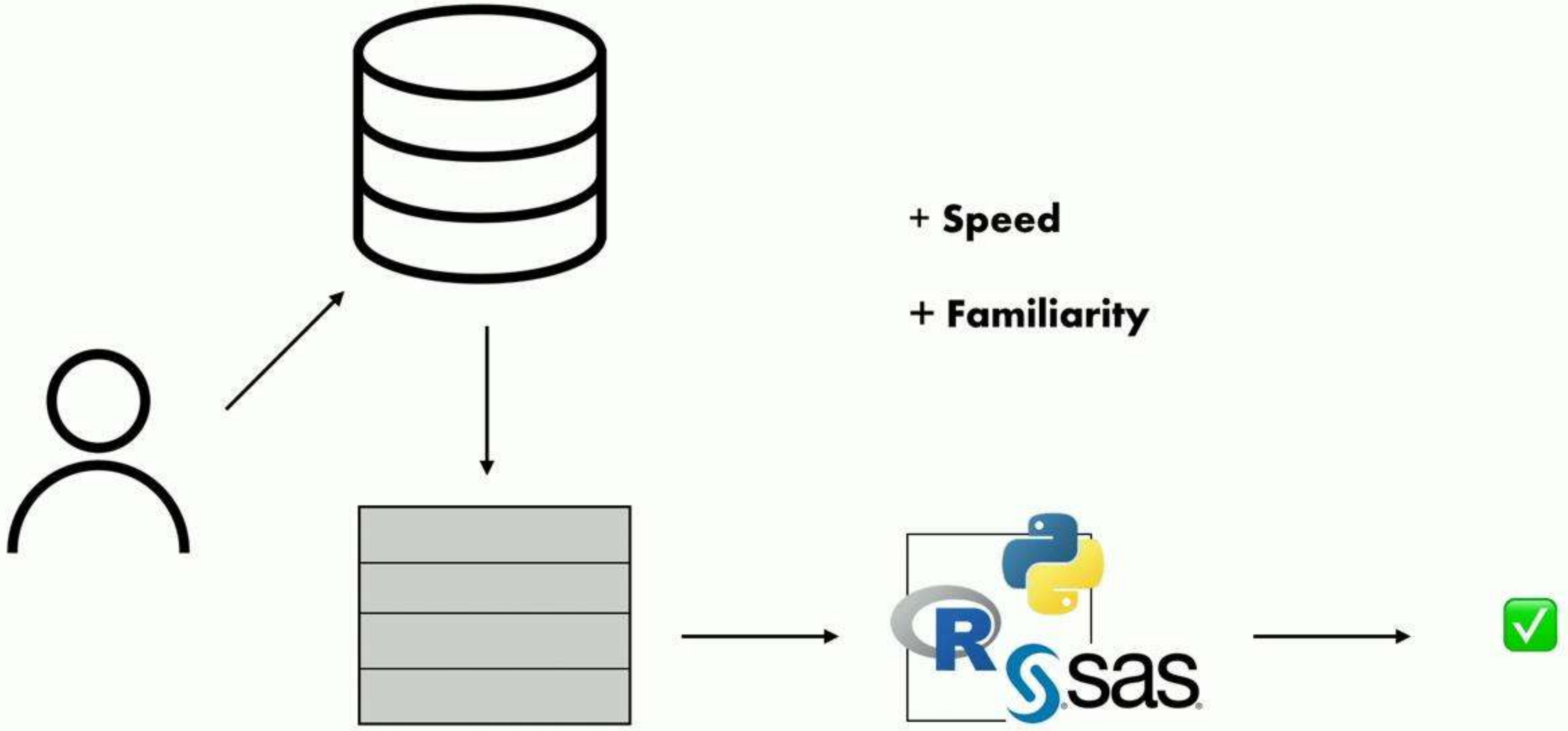


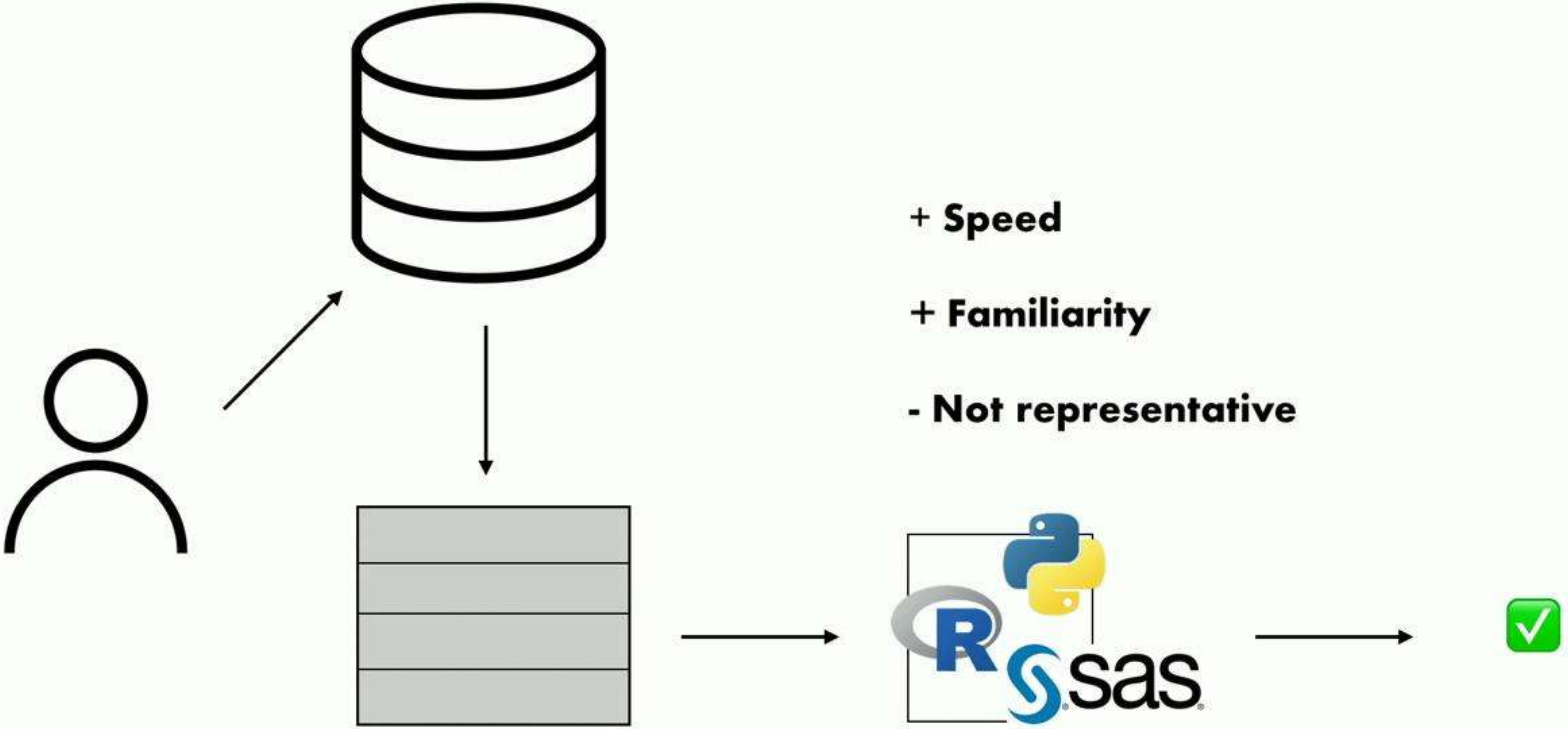


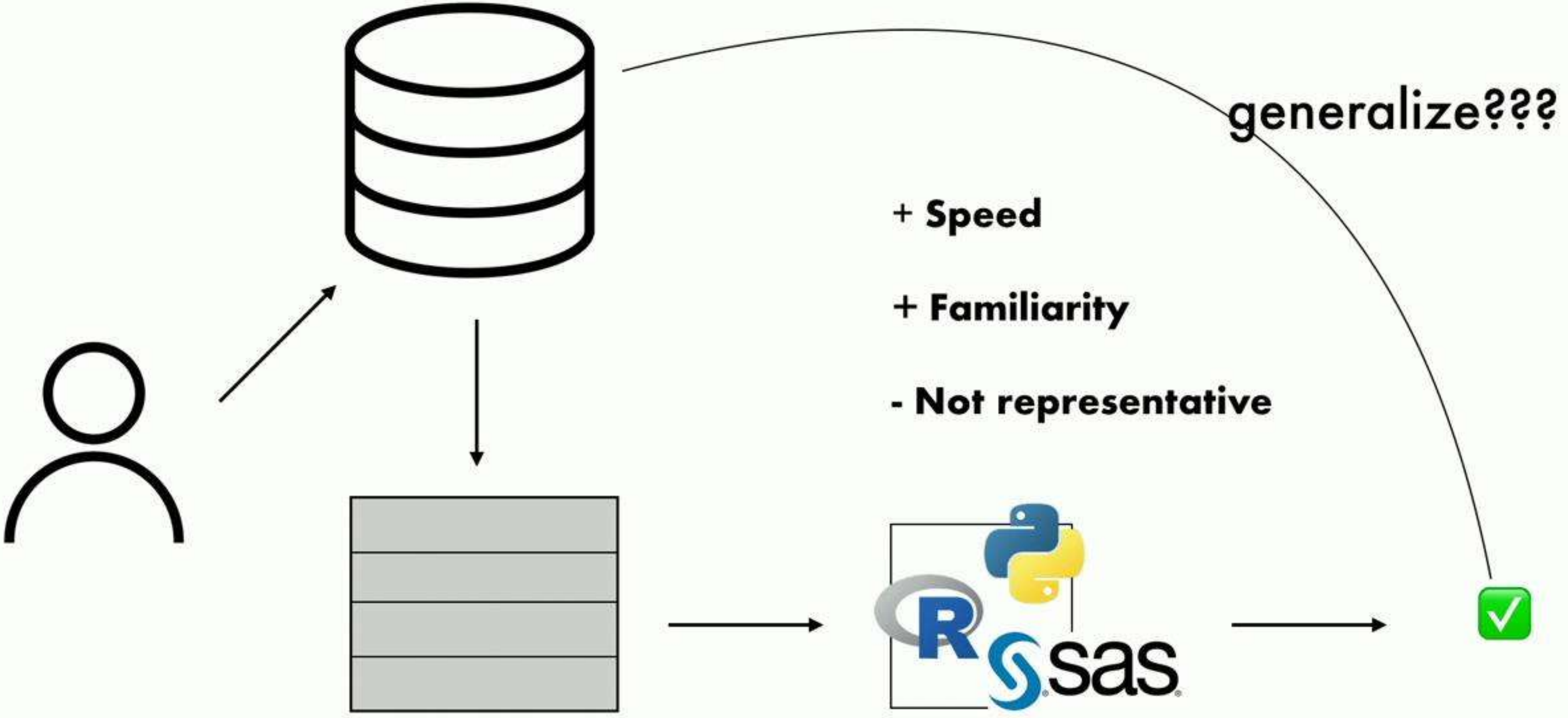


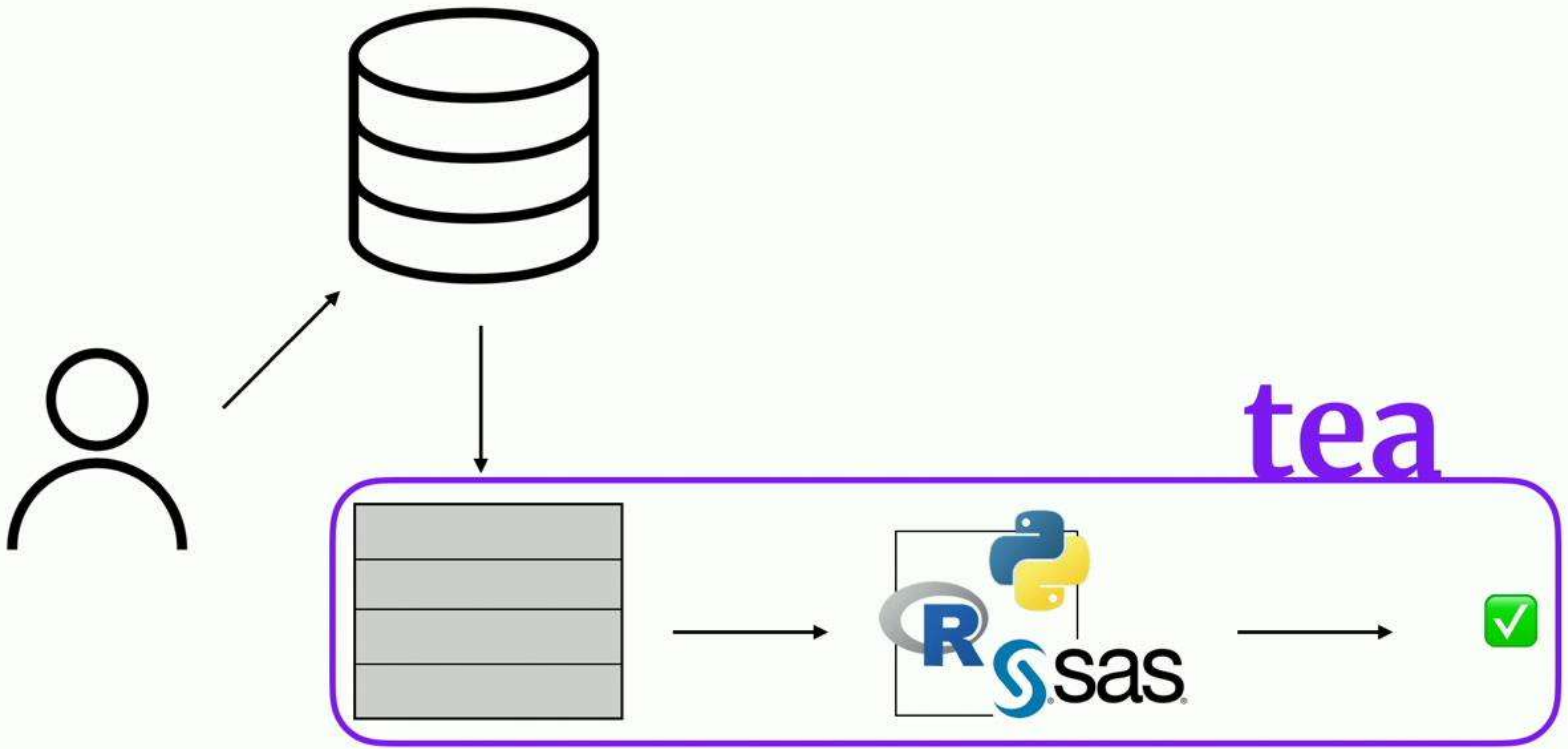


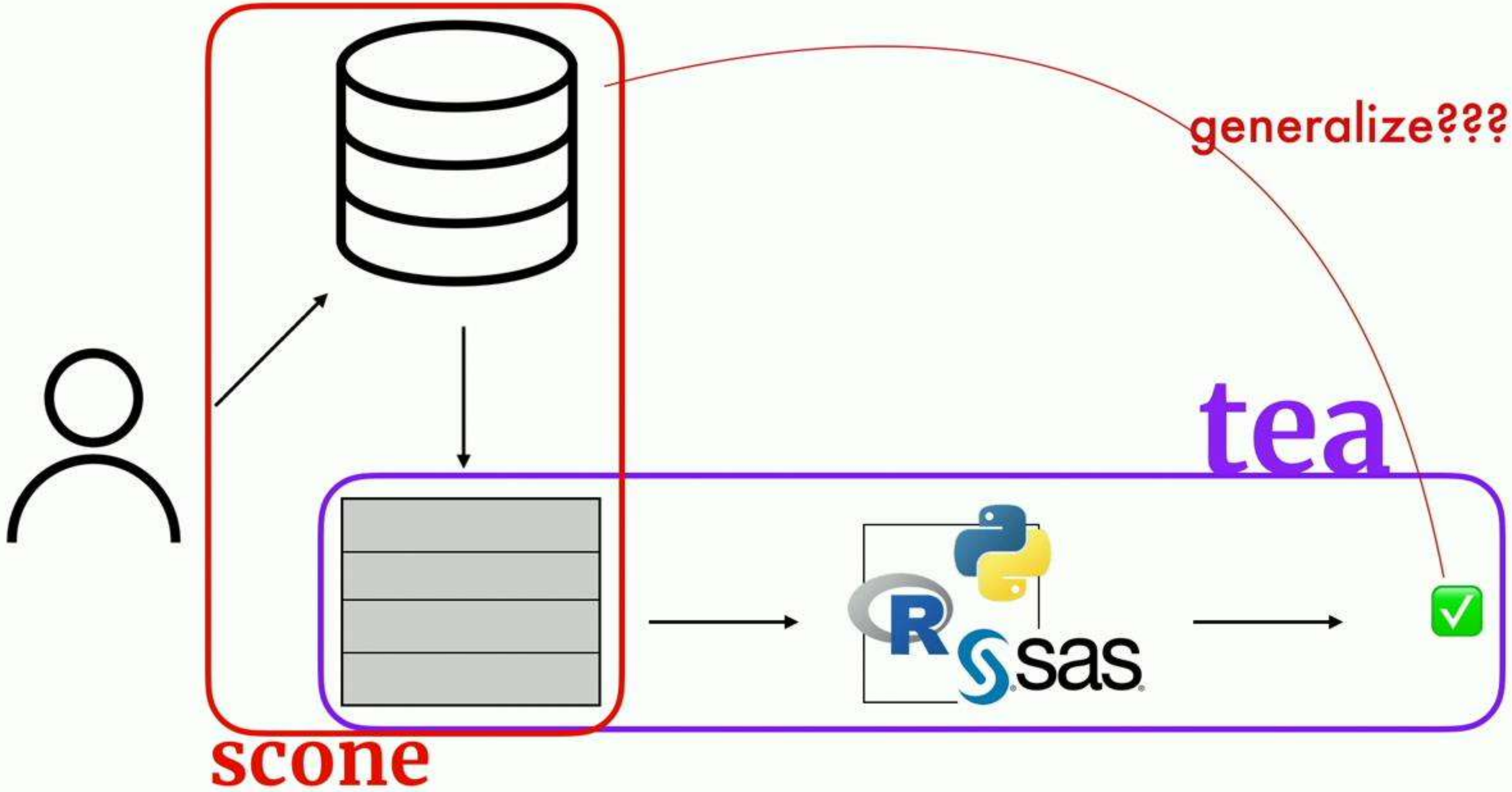

+ Speed

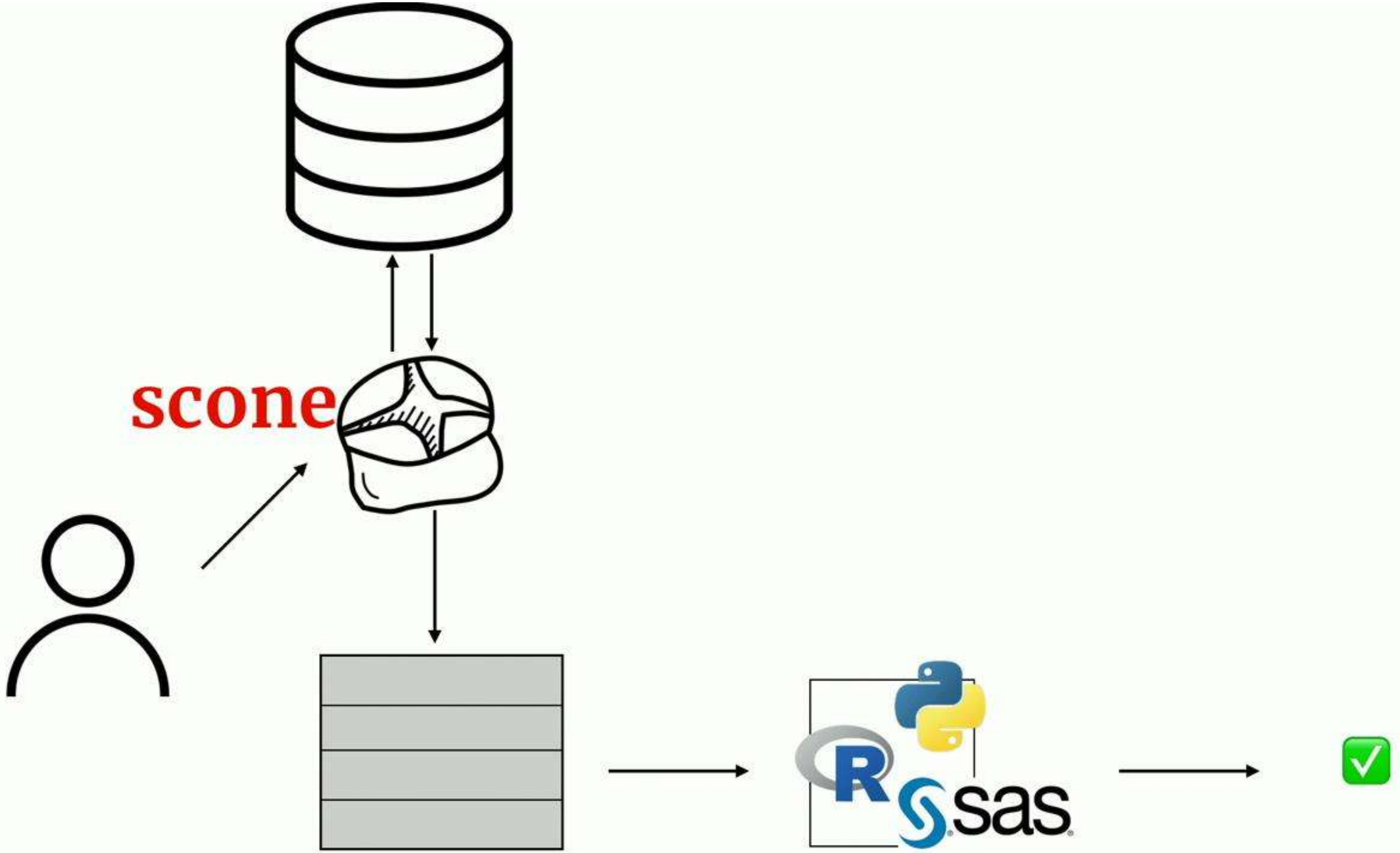

+ Speed
+ Familiarity

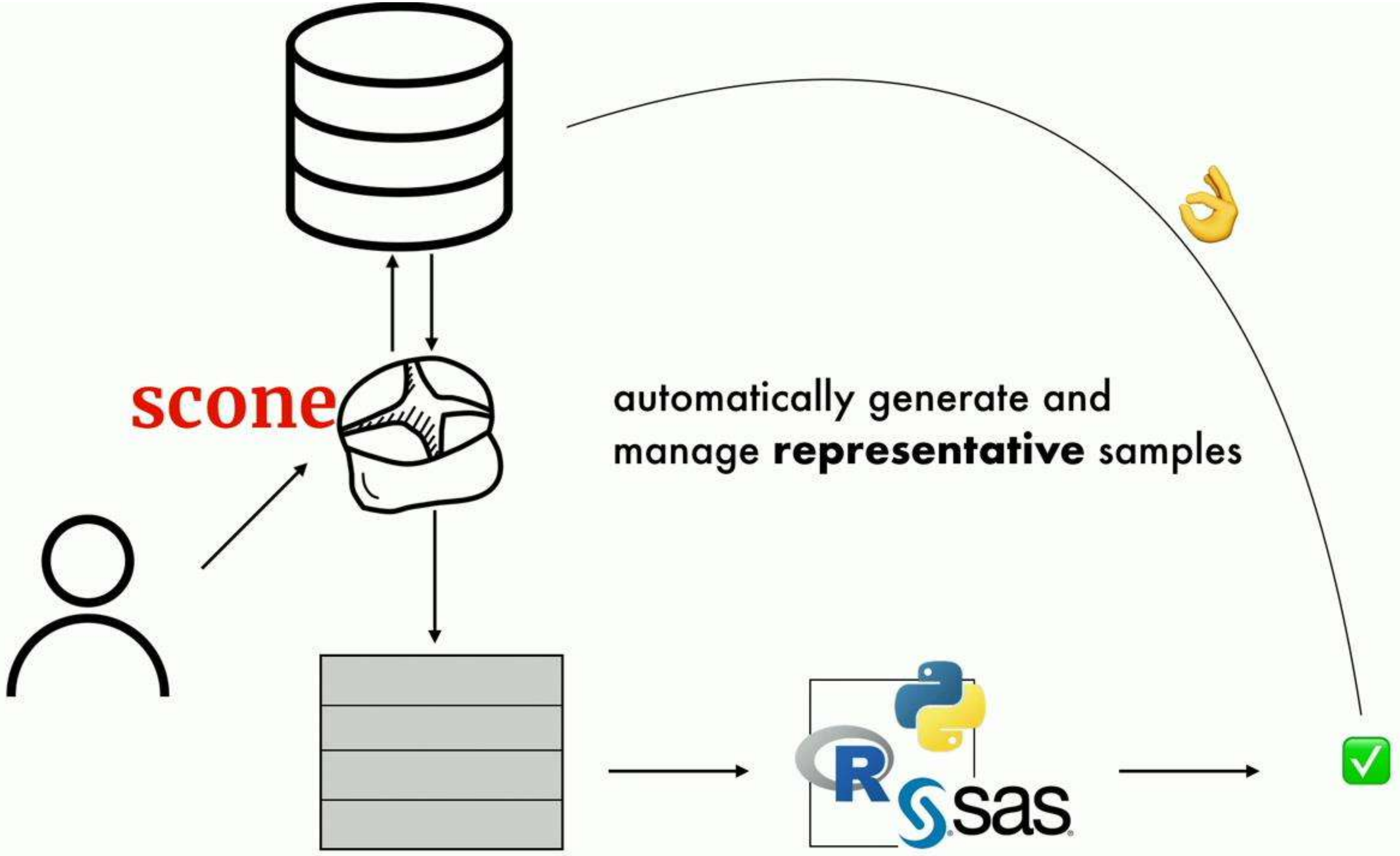

+ Speed
+ Familiarity
- Not representative


generalize???
+ Speed
+ Familiarity
- Not representative
R
sas


tea
R
sas


generalize???
tea
sas
scone


scone
sas


scone
automatically generate and manage **representative** samples

tea

Automated statistical analyses

Internal validity

```
pip install tealang
tea-lang.org
```

scone

Automated sampling

External validity

```
Stay tuned!
```

tea

Automated statistical analyses

Internal validity

```
pip install tealang
tea-lang.org
```

scone

Automated sampling

External validity

```
Stay tuned!
```

COLLABORATION, USERS, FEEDBACK