

AI Agents for Automating Scientific Hypotheses Generation and Verification from Massive Offline Data

Ying Jin

*Department of Statistics and Data Science
The Wharton School, University of Pennsylvania*

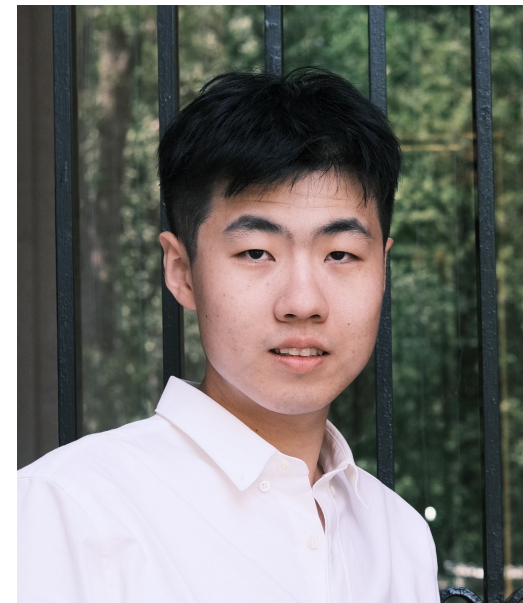
IMSI Workshop on Reinforcement Learning from Offline Data and Human Feedback, April 2026

Collaborators

Stanford | Department of Statistics
SCHOOL OF HUMANITIES AND SCIENCES



Fred Zhang
UPenn Wharton



Kexin Huang
Stanford CS → Phylo



Ryan Li
Stanford CS



Michael Li
Stanford CS



Emmanuel Candès
Stanford Stats & Math



Jure Leskovec
Stanford CS

AI for scientific discovery

Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuanqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#),
[Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P.](#)
[Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Debora](#)
[Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#)  [+ Show authors](#)

[Nature](#) **620**, 47–60 (2023) | [Cite this article](#)

151k Accesses | **588** Citations | **743** Altmetric | [Metrics](#)

AI for scientific discovery

Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuanqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#),
[Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P.](#)
[Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Debora](#)
[Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#)  [+ Show authors](#)

[Nature](#) **620**, 47–60 (2023) | [Cite this article](#)

151k Accesses | **588** Citations | **743** Altmetric | [Metrics](#)

Recent emerging paradigm

AI as prediction engines



AI as **automation agents**

AI for scientific discovery

Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuanqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#),
[Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P.](#)
[Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Debora](#)
[Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#) ✉ [+ Show authors](#)

[Nature](#) **620**, 47–60 (2023) | [Cite this article](#)

151k Accesses | **588** Citations | **743** Altmetric | [Metrics](#)

Recent emerging paradigm

AI as prediction engines



AI as **automation agents**

Predict drug binding affinities

Predict protein structures

Simulate particle interactions

AI for scientific discovery

Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuanqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#),
[Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P.](#)
[Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Debora](#)
[Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#)  [+ Show authors](#)

[Nature](#) **620**, 47–60 (2023) | [Cite this article](#)

151k Accesses | **588** Citations | **743** Altmetric | [Metrics](#)

Recent emerging paradigm

AI as prediction engines



AI as **automation agents**

Predict drug binding affinities
Predict protein structures
Simulate particle interactions

Write and execute codes
Analyze massive datasets
Design experiments (wet/dry lab)

AI for scientific discovery

Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuanqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#), [Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P. Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Debora Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#) ✉ [+ Show authors](#)

[Nature](#) **620**, 47–60 (2023) | [Cite this article](#)

151k Accesses | **588** Citations | **743** Altmetric | [Metrics](#)

Recent emerging paradigm

AI as prediction engines



AI as **automation agents**

Accelerating scientific breakthroughs with an AI co-scientist

[Google research]

February 19, 2025 · Juraj Gottweis, Google Fellow, and Vivek Natarajan, Research Lead

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

August 13, 2024

Article | [Open access](#) | Published: 25 March 2026

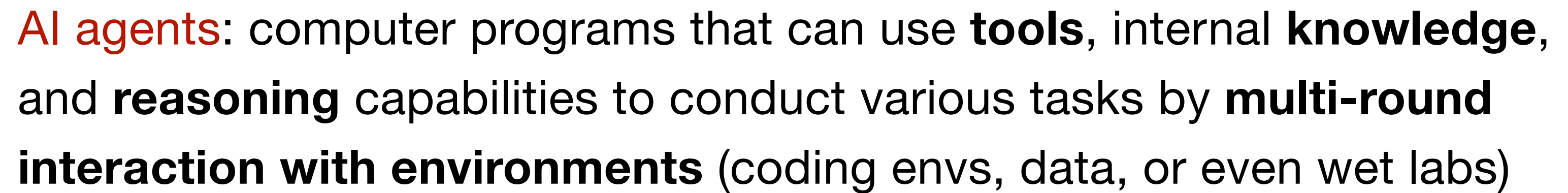
Towards end-to-end automation of AI research

[Chris Lu](#), [Cong Lu](#), [Robert Tjarko Lange](#), [Yutaro Yamada](#) ✉, [Shengran Hu](#), [Jakob Foerster](#), [David Ha](#) ✉ & [Jeff Clune](#) ✉

[Nature](#) **651**, 914–919 (2026) | [Cite this article](#)

[Sakana.ai]

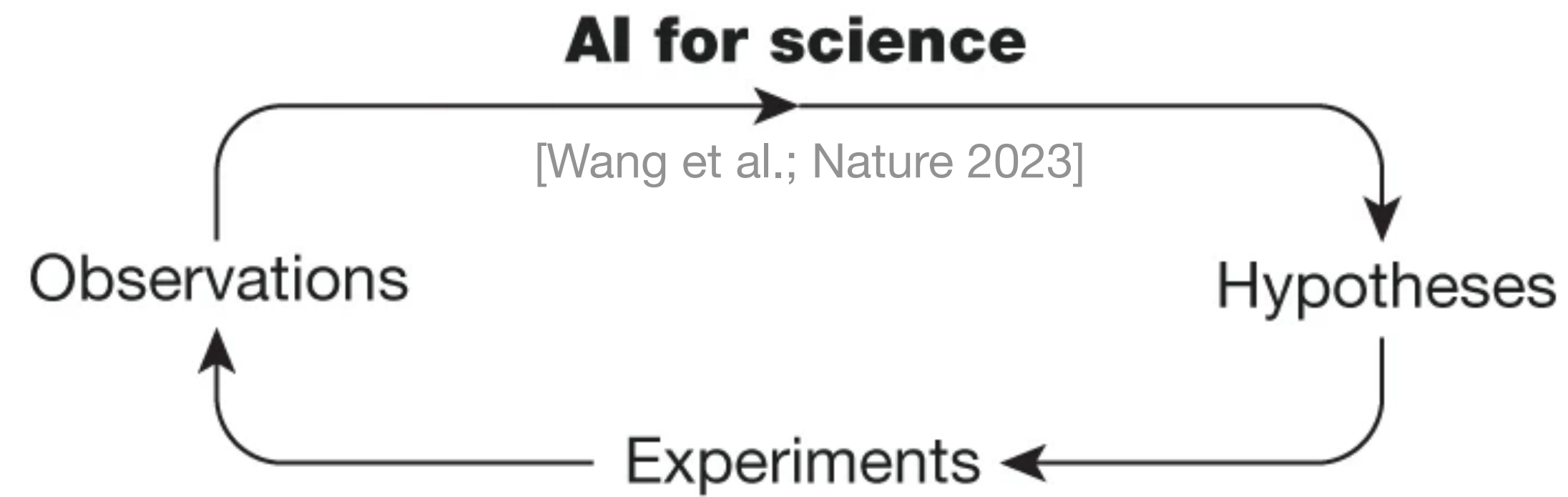
[Cell]



2

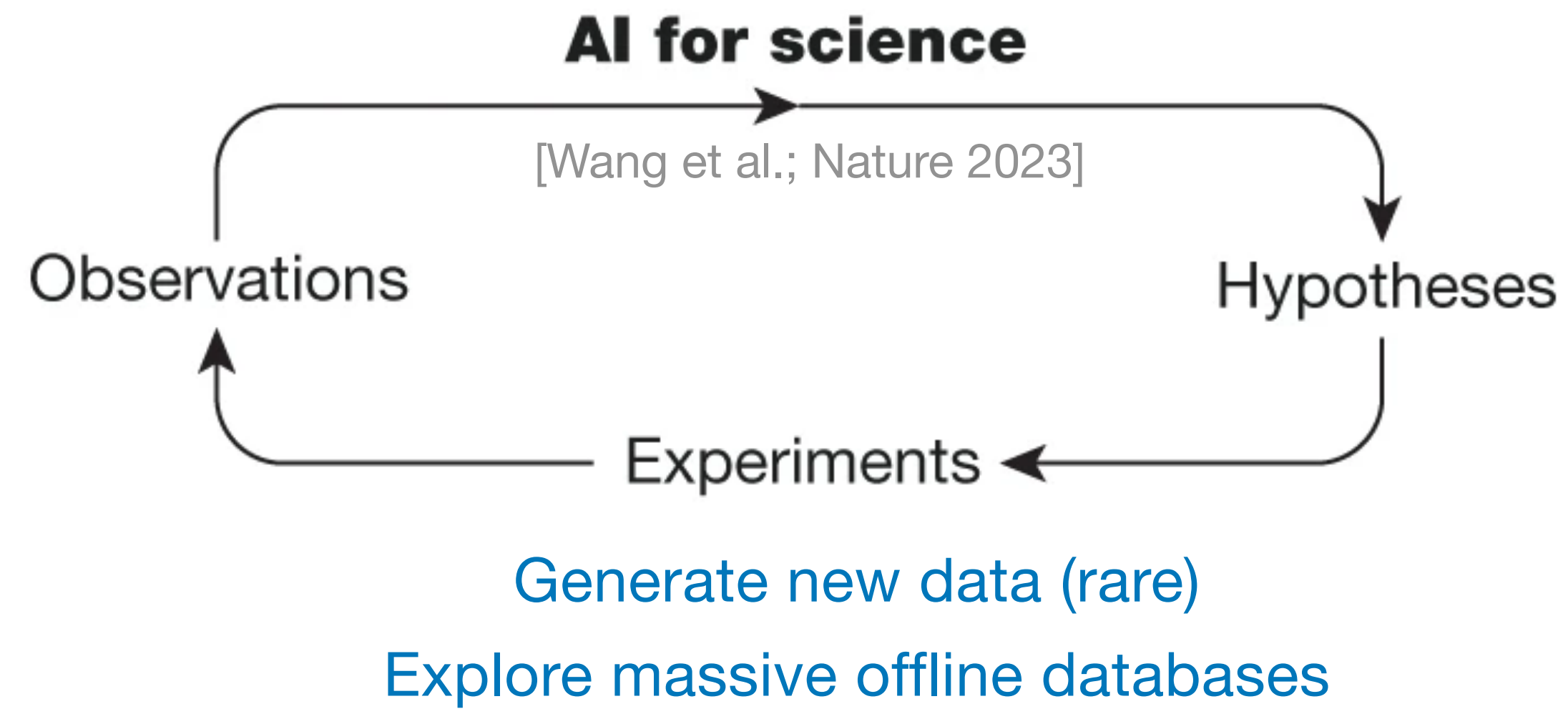
False discoveries?

Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



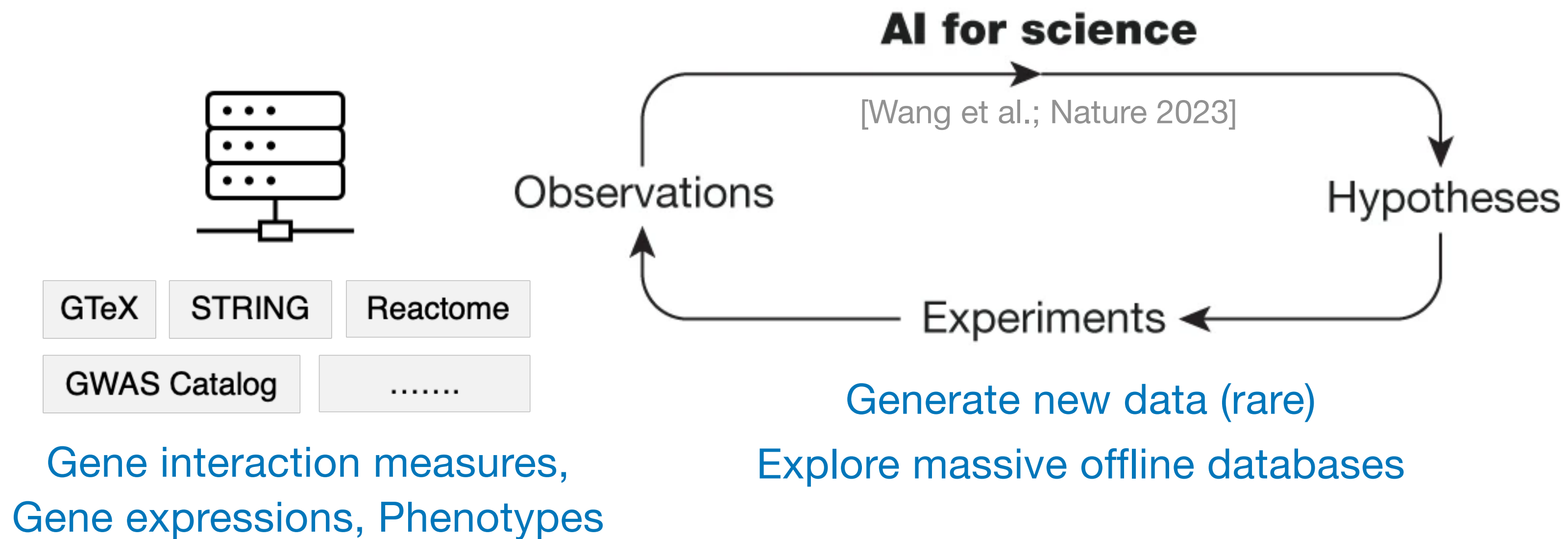
False discoveries?

Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



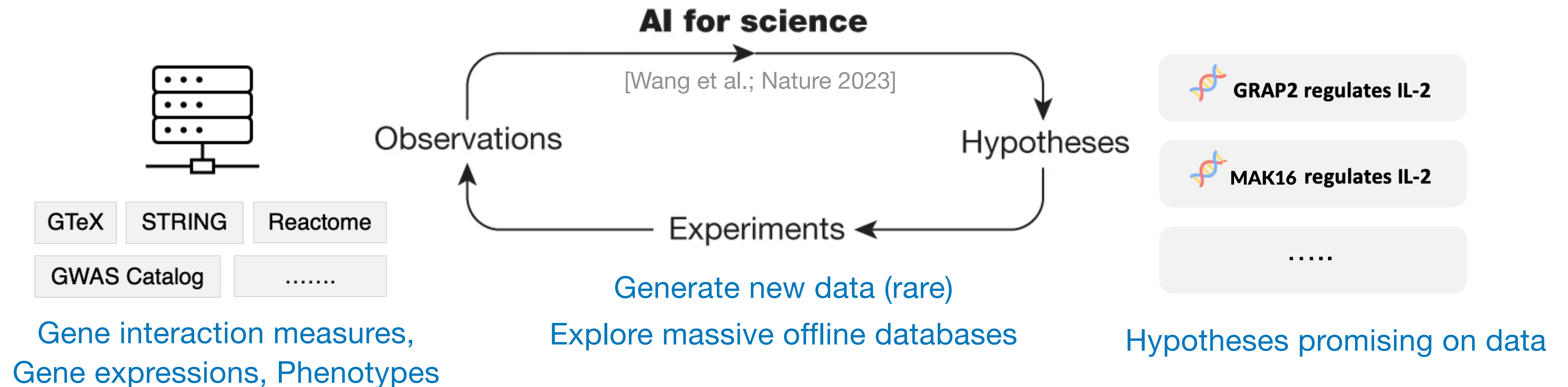
False discoveries?

Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



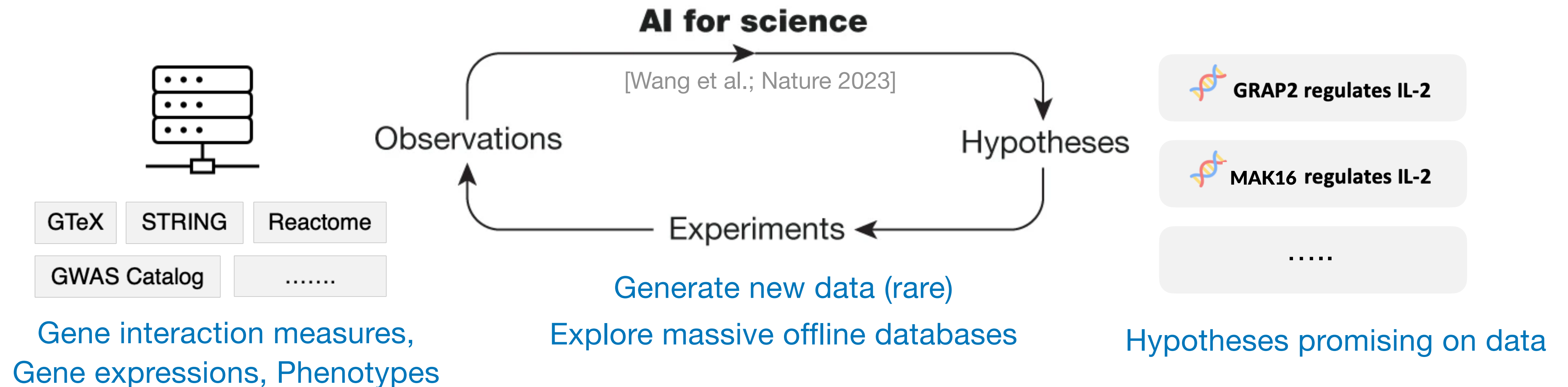
False discoveries?

Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



False discoveries?

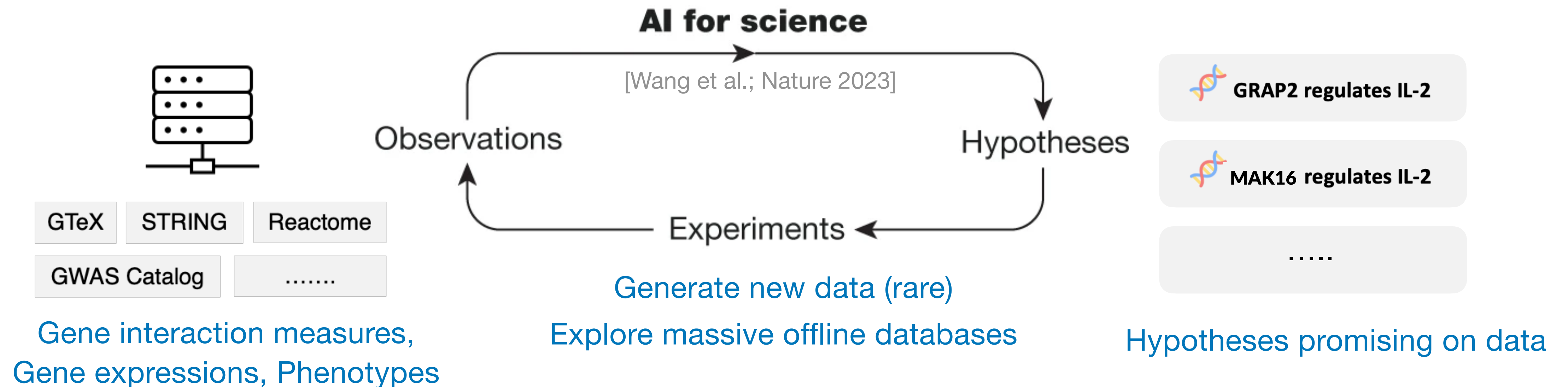
Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



The “Reward” problem: Do AI-generated hypotheses deserve costly followup?

False discoveries?

Vision: AI agents automatically explore the environment (labs, data, analyses) to produce scientific discoveries



The “Reward” problem: Do AI-generated hypotheses deserve costly followup?

This is important but hard to answer because

- ▶ AI agents generate many hypotheses at scale $\leadsto$ nearly impossible to manually follow-up on each of them
- ▶ Agents surface the most striking patterns in data $\leadsto$ selection bias, cannot fully trust
- ▶ LLM backbones $\leadsto$ hallucinations, random generations, various errors

Scaling statistical evidence for automated scientific discovery

Statistical evidence is a unified language for scientific credibility



GRAP2 regulates IL-2



MAK16 regulates IL-2

.....

Hypotheses proposed

Scaling statistical evidence for automated scientific discovery

Statistical evidence is a unified language for scientific credibility

 GRAP2 regulates IL-2

 MAK16 regulates IL-2

.....

Hypotheses proposed

+

p-value: 0.001

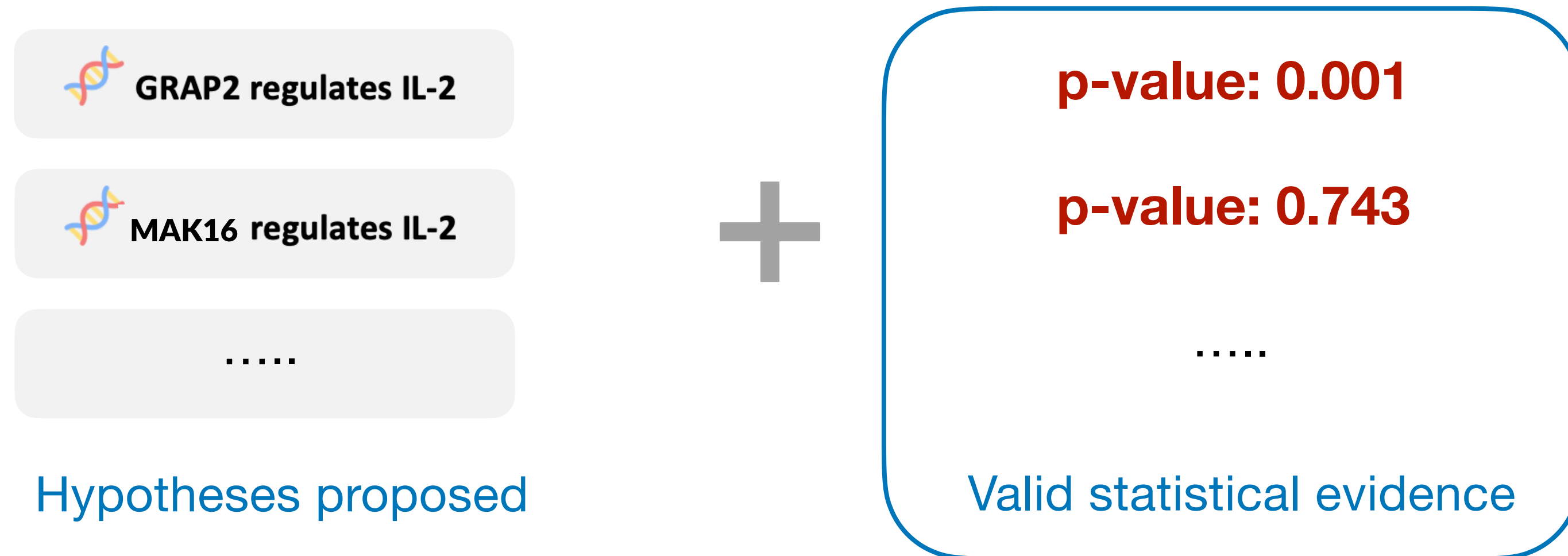
p-value: 0.743

.....

Valid statistical evidence

Scaling statistical evidence for automated scientific discovery

Statistical evidence is a unified language for scientific credibility

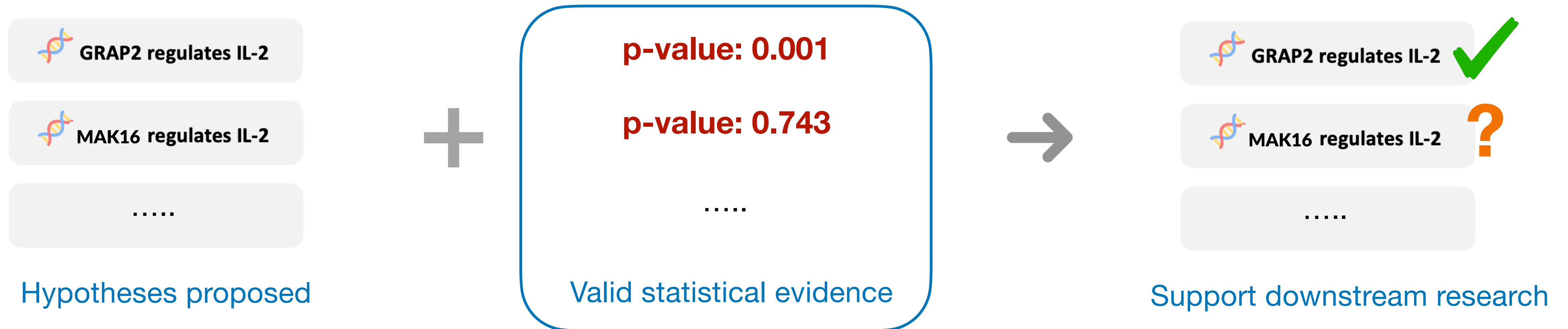


Why (error-controlled, theoretically justified) statistical evidence?

- P-values/e-values $\leadsto$ Interpretable, unified measure of strength of support
- Type-I error control $\leadsto$ rigorous basis for follow-up investigations with limited waste

Scaling statistical evidence for automated scientific discovery

Statistical evidence is a unified language for scientific credibility

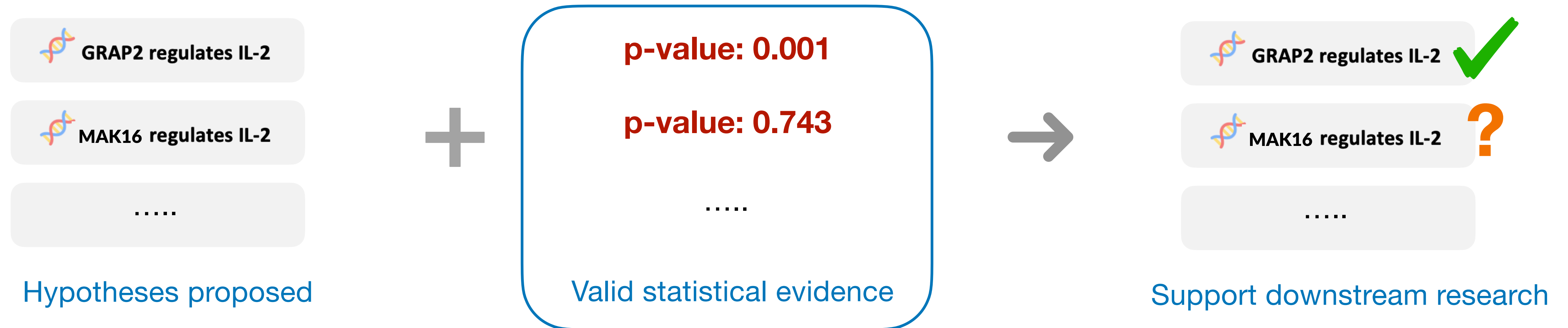


Why (error-controlled, theoretically justified) statistical evidence?

- ▶ P-values/e-values $\leadsto$ Interpretable, unified measure of strength of support
- ▶ Type-I error control $\leadsto$ rigorous basis for follow-up investigations with limited waste

Scaling statistical evidence for automated scientific discovery

Statistical evidence is a unified language for scientific credibility



Why (error-controlled, theoretically justified) statistical evidence?

- P-values/e-values $\leadsto$ Interpretable, unified measure of strength of support
- Type-I error control $\leadsto$ rigorous basis for follow-up investigations with limited waste

This talk: Scaling statistical evidence generation with & for AI agents

Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)



Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Hypothesis generation: ongoing (20%)



Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Hypothesis generation: ongoing (20%)

Challenge: Reusing data to compute evidence leads to selection bias (overestimate confidence)



Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Hypothesis generation: ongoing (20%)

Challenge: Reusing data to compute evidence leads to selection bias (overestimate confidence)



Both require agents to **interact rigorously with offline data** for statistical validity

Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data

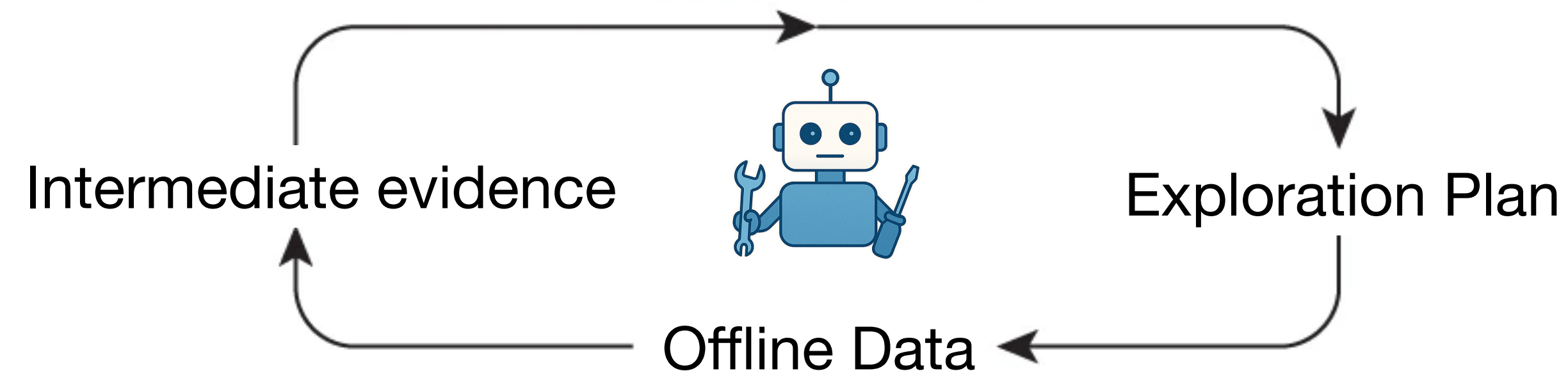
Hypothesis generation: ongoing (20%)

Challenge: Reusing data to compute evidence leads to selection bias (overestimate confidence)

Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Key: agent sequentially interacts with offline data to accumulate evidence

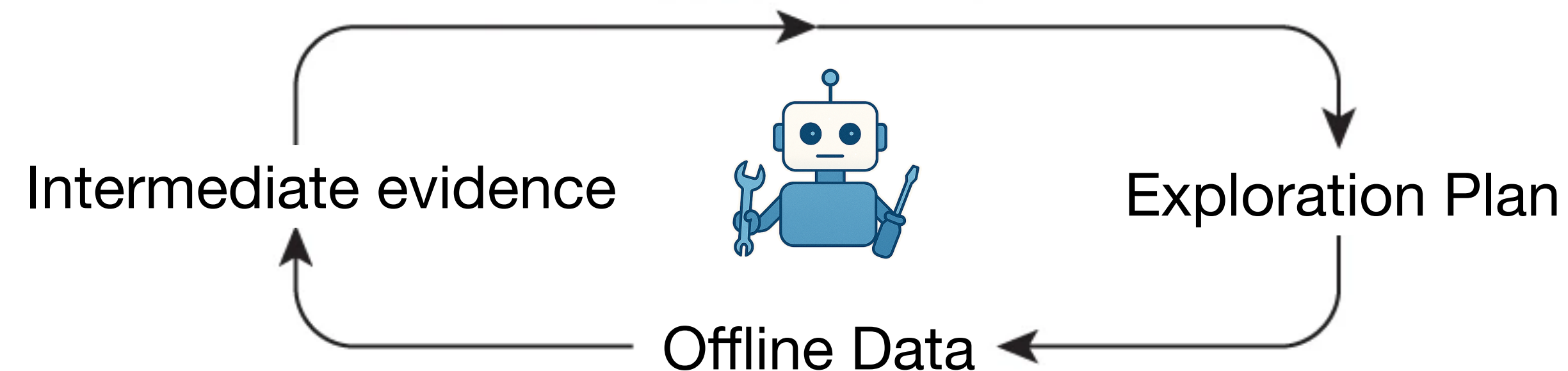
Hypothesis generation: ongoing (20%)

Challenge: Reusing data to compute evidence leads to selection bias (overestimate confidence)

Two stories: hypothesis testing and generation

Hypothesis testing: POPPER (80%)

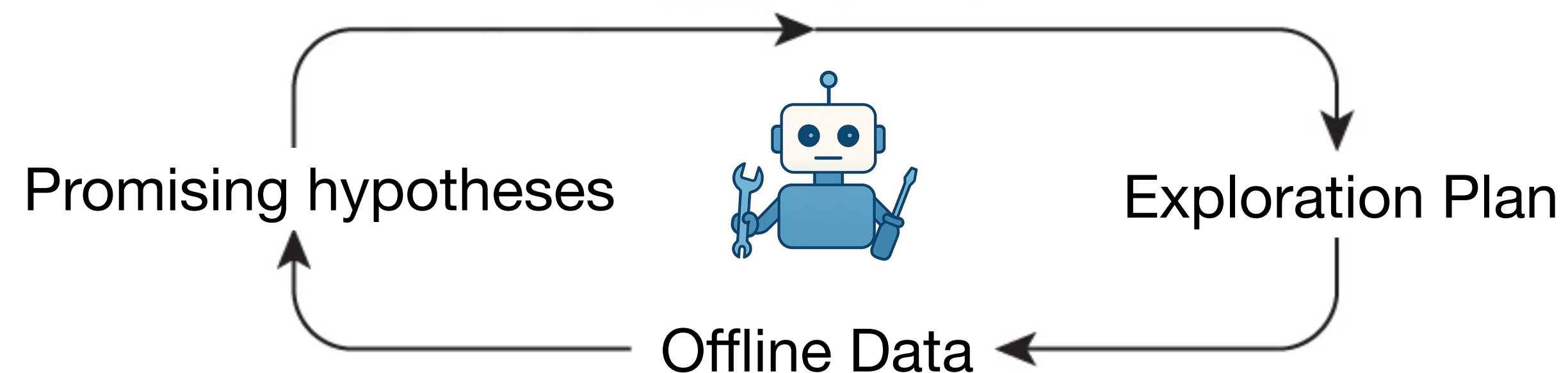
Challenge: Automation + free-form, abstract scientific hypotheses + hard to directly test with data



Key: agent sequentially interacts with offline data to accumulate evidence

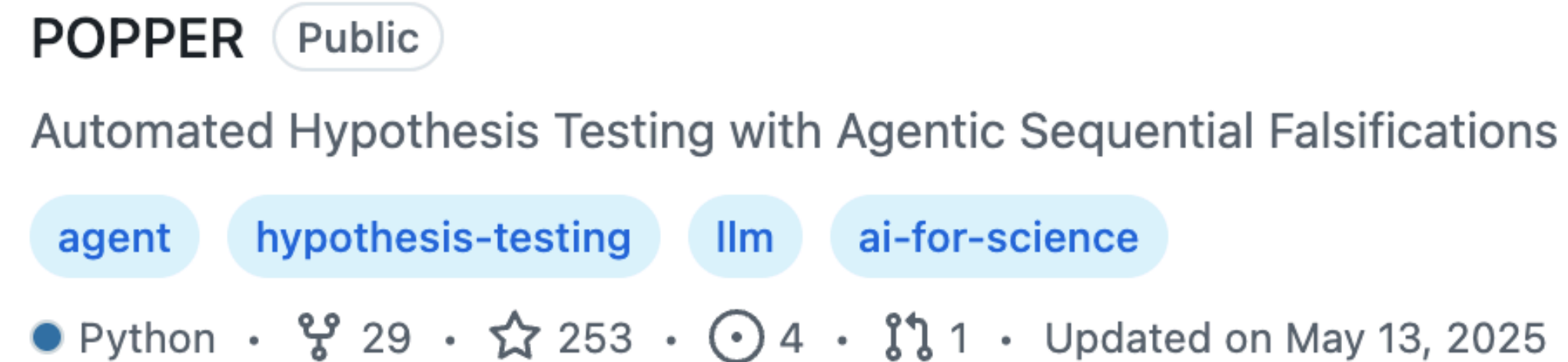
Hypothesis generation: ongoing (20%)

Challenge: Reusing data to compute evidence leads to selection bias (overestimate confidence)

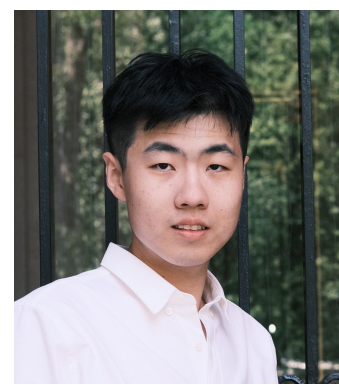


Key: agent sequentially interacts with offline data to generate hypotheses, then compute statistical evidence with separate data

POPPER: Automated hypothesis validation with agentic sequential falsifications



Huang, K.^{*}, Jin, Y.^{*}, Li, R.^{*}, Li, M. Y., Candès, E., & Leskovec, J. (2025).
Automated Hypothesis Validation with Agentic Sequential Falsifications. ICML 2025.



Kexin
Huang



Ryan Li



Michael Li

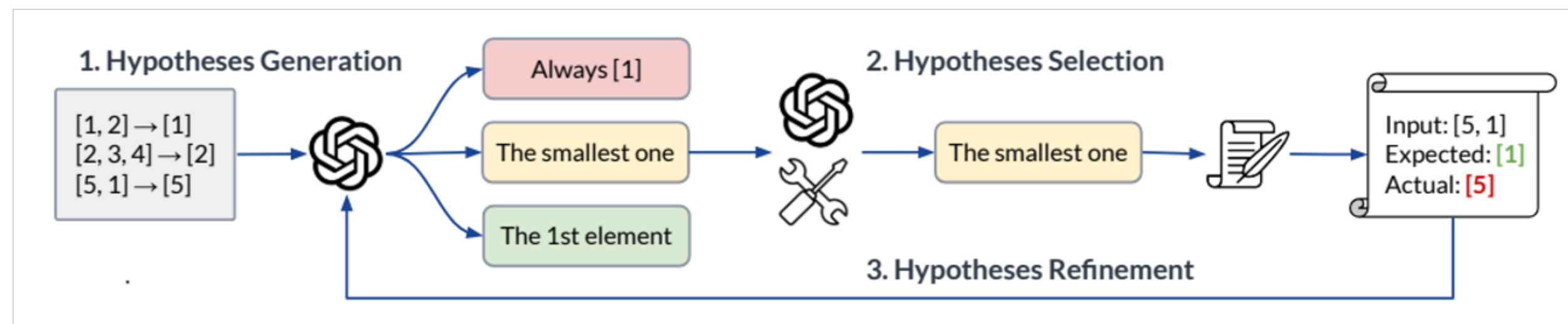


Emmanuel
Candès



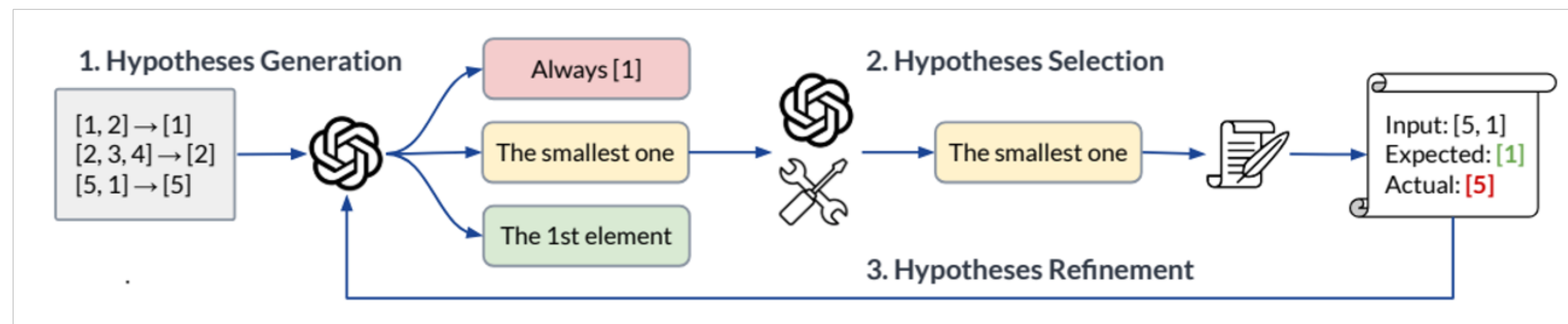
Jure
Leskovec

Hypothesis validation with AI agents



Common type of “hypothesis” in the literature

Hypothesis validation with AI agents

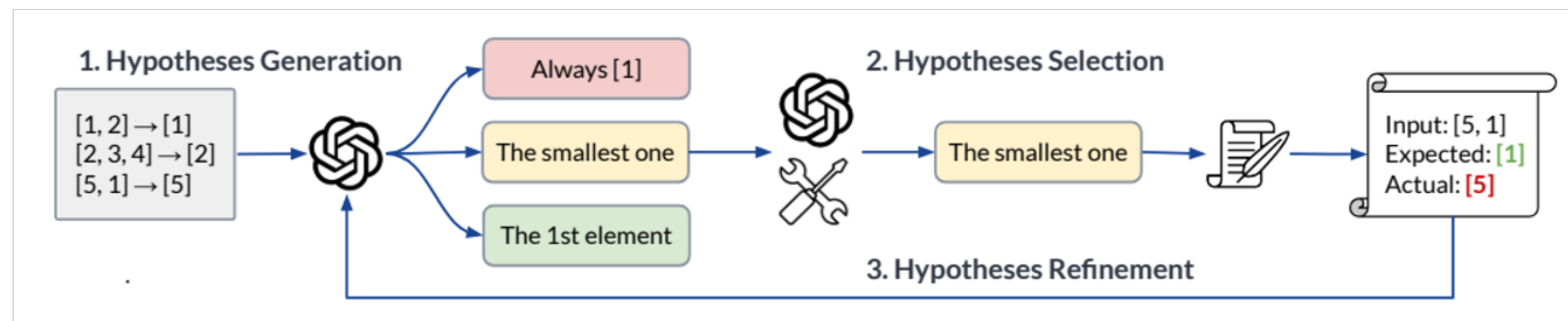


Common type of “hypothesis” in the literature

Challenge: Scientific hypotheses can be free-form, abstract, hard to directly test with data

- E.g. LRRC32 is a drug target for Ulcerative colitis (UC)

Hypothesis validation with AI agents



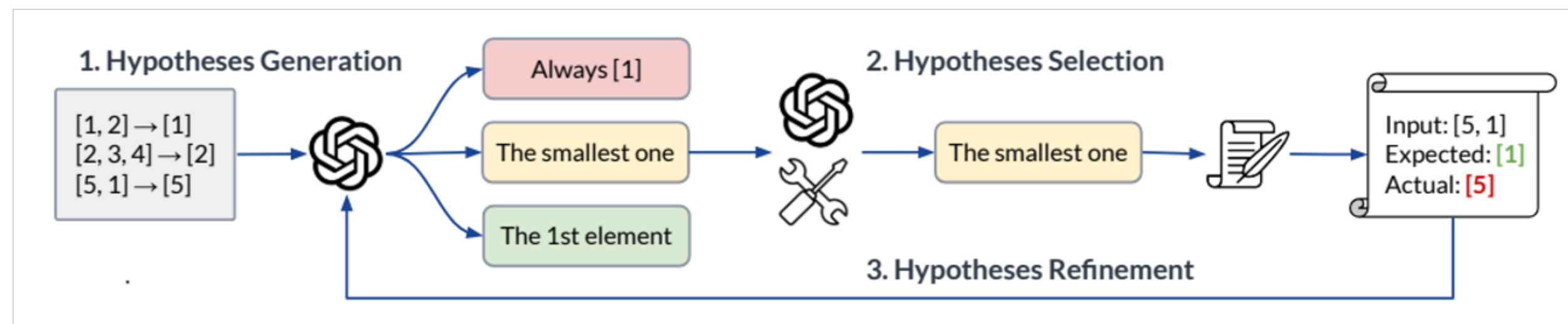
Common type of “hypothesis” in the literature

How to **rigorously** validate **free-form** hypothesis **at scale**?

Challenge: Scientific hypotheses can be free-form, abstract, hard to directly test with data

- E.g. LRRC32 is a drug target for Ulcerative colitis (UC)

Hypothesis validation with AI agents



Common type of “hypothesis” in the literature

How to **rigorously** validate **free-form** hypothesis at scale?

Challenge: Scientific hypotheses can be free-form, abstract, hard to directly test with data

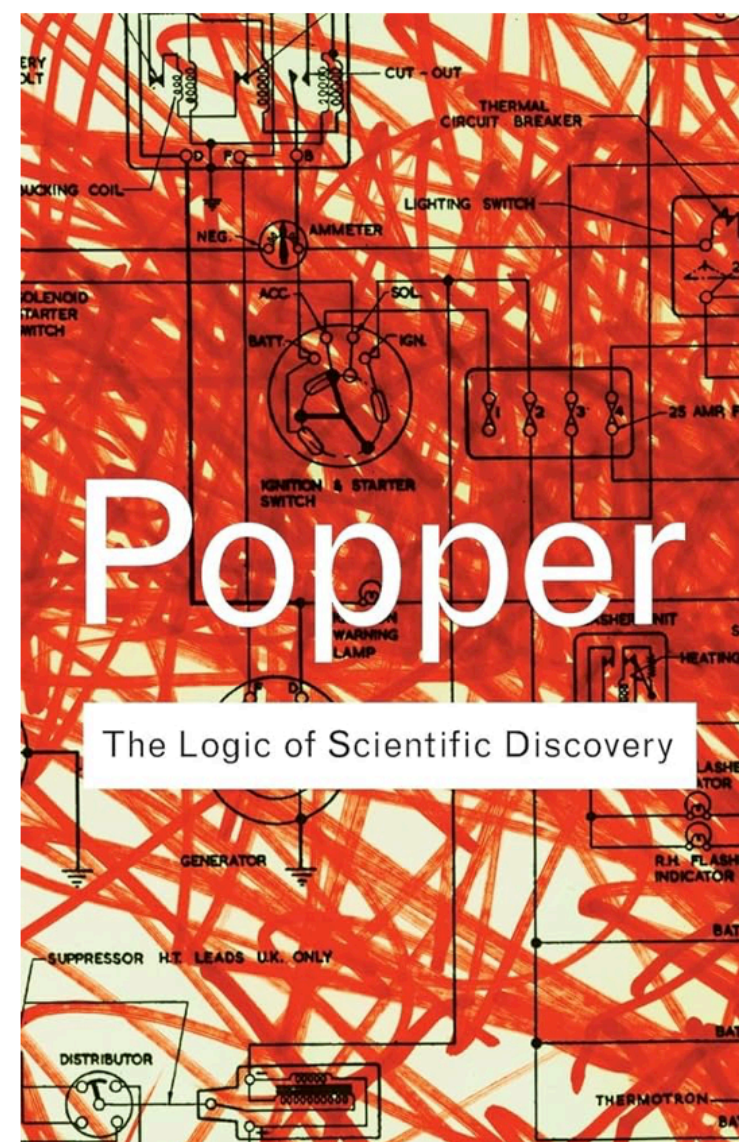
- E.g. LRRC32 is a drug target for Ulcerative colitis (UC)



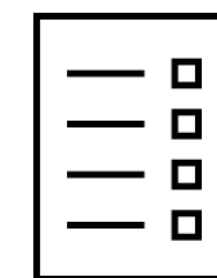
Same as classical hypothesis testing, except that it's done by AI agents.

How does a scientist validate a hypothesis?

Popper: testing by falsification*



?
Hypothesis



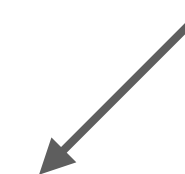
Falsification test
“Find refutable
evidence”



Falsified



Keep testing



Conduct experiments
Collect data
Analyze data

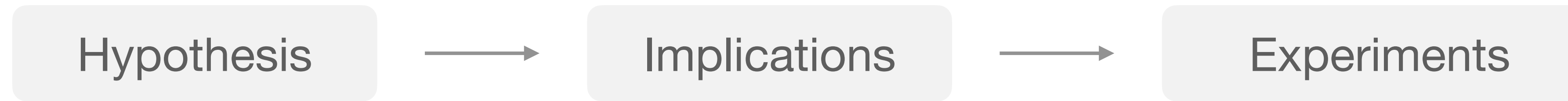
* We take Popper's perspective on the falsification process (nature of scientific testing), but not necessarily more broadly his view of what is scientific or not.

Why falsification tests?

Lots of scientific hypotheses cannot be directly validated with data, but their implications can

- Hypothesis (null): LRRC32 is not a drug target for Ulcerative colitis (UC)
- Implications of the null hypothesis
 - LRRC32 expression should not be significantly high at UC-relevant cell types
- Falsification experiment
 - Null: LRRC32 expression is **not high** in UC cell types
 - Alternative: LRRC32 expression is **high** in UC cell types

Building an agent by imitating the falsification process



Building an agent by imitating the falsification process

Hypothesis



Implications

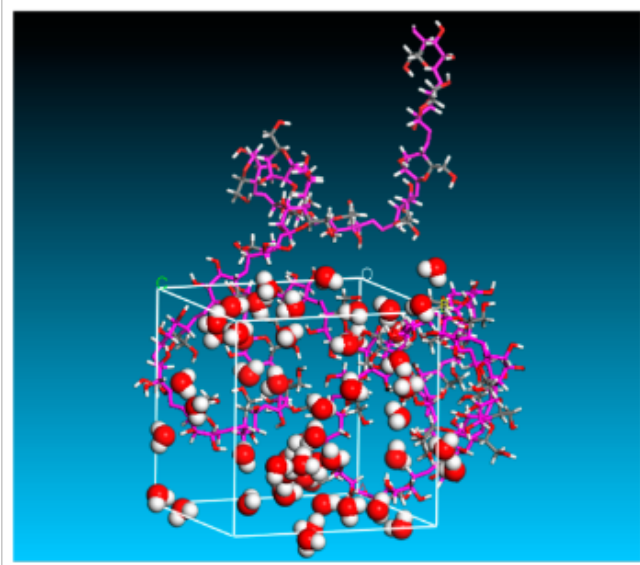


Experiments

Wet-lab



Dry-lab



Building an agent by imitating the falsification process

Hypothesis



Implications

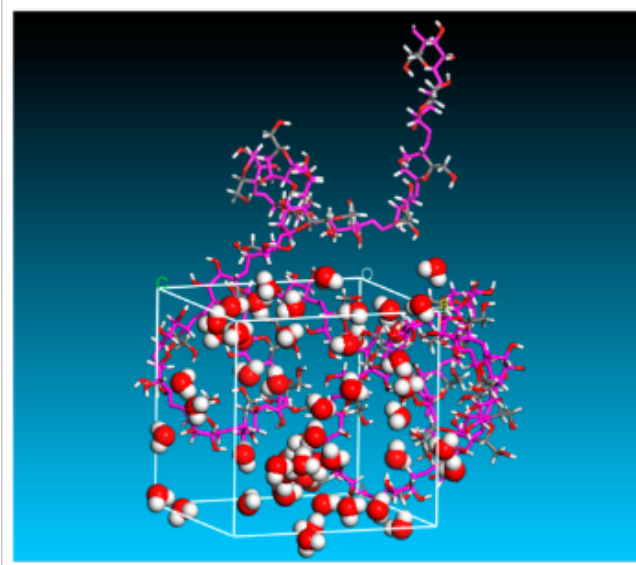


Experiments

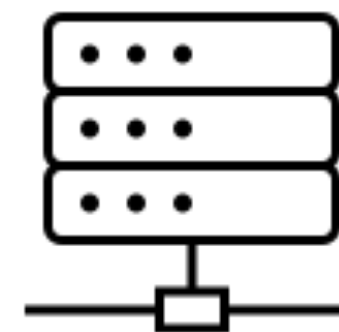
Wet-lab



Dry-lab



Generate



Data

Building an agent by imitating the falsification process

Hypothesis



Implications

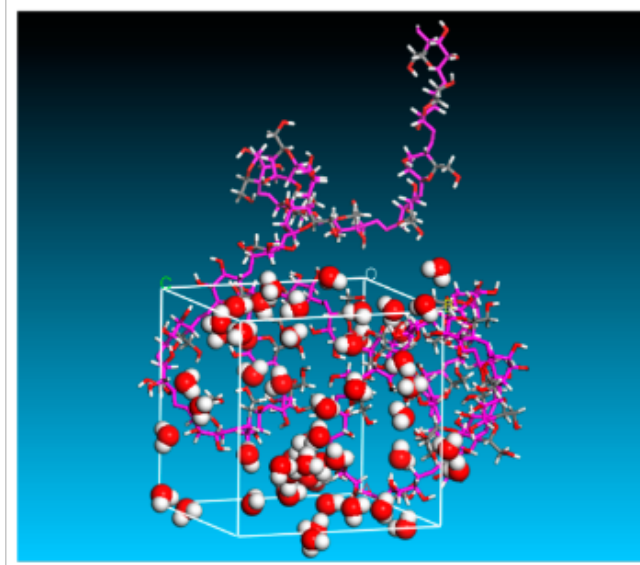


Experiments

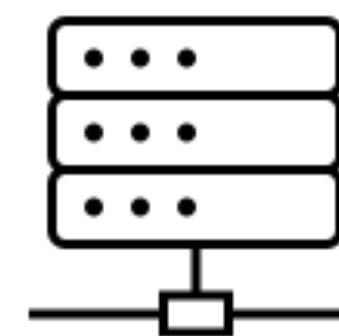
Wet-lab



Dry-lab



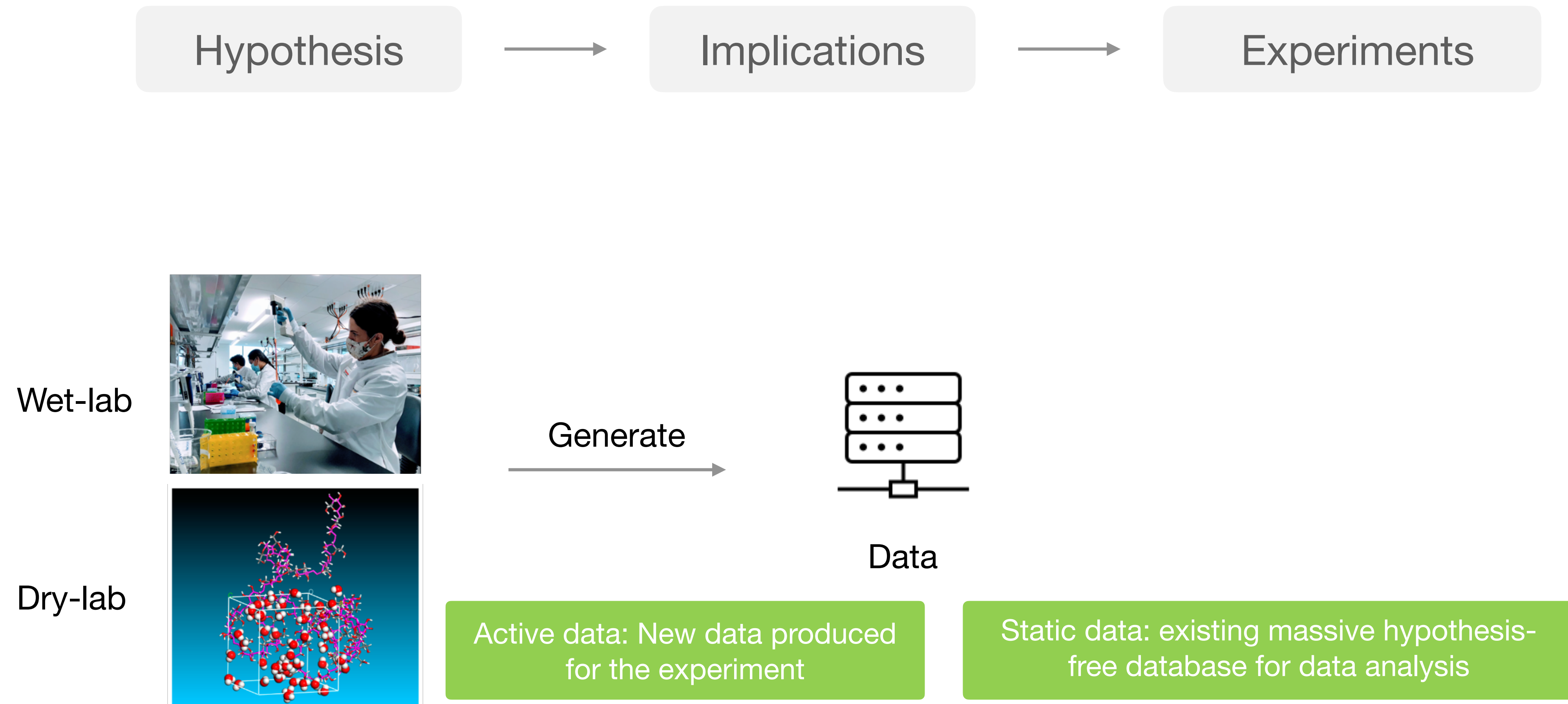
Generate



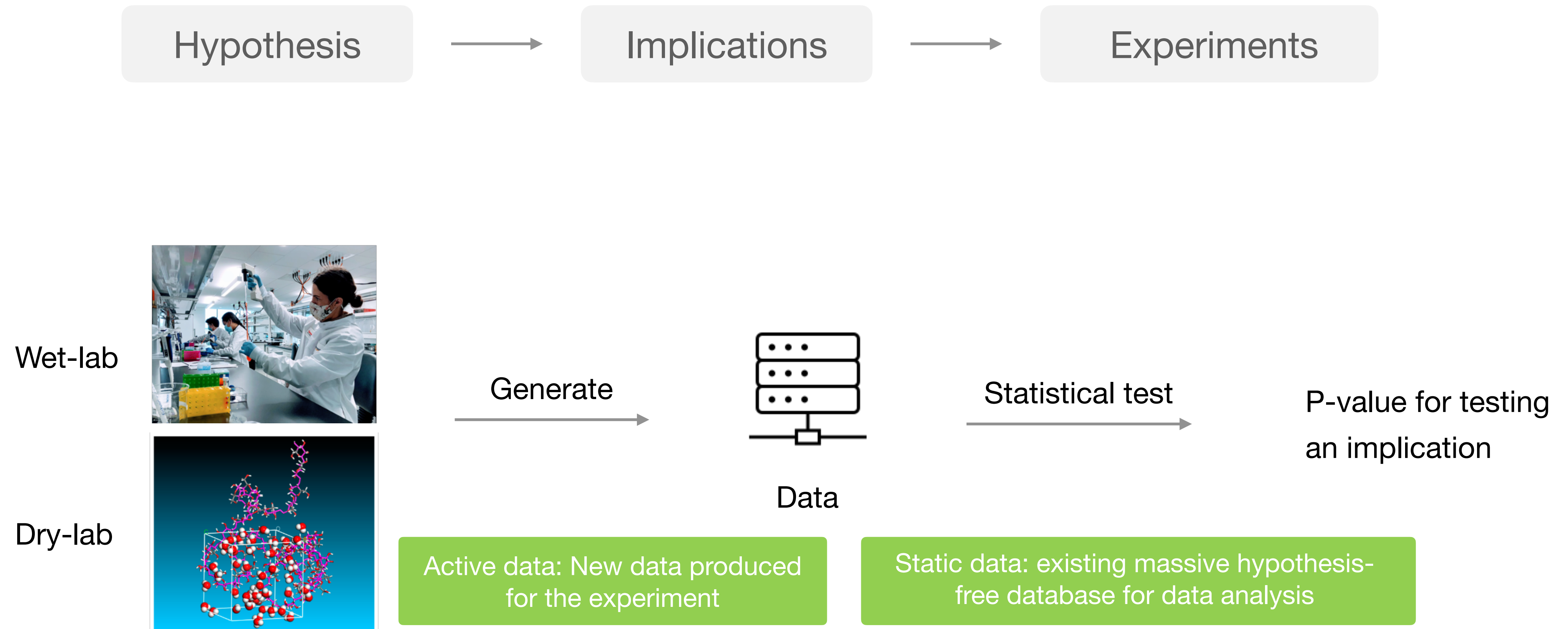
Data

Active data: New data produced for the experiment

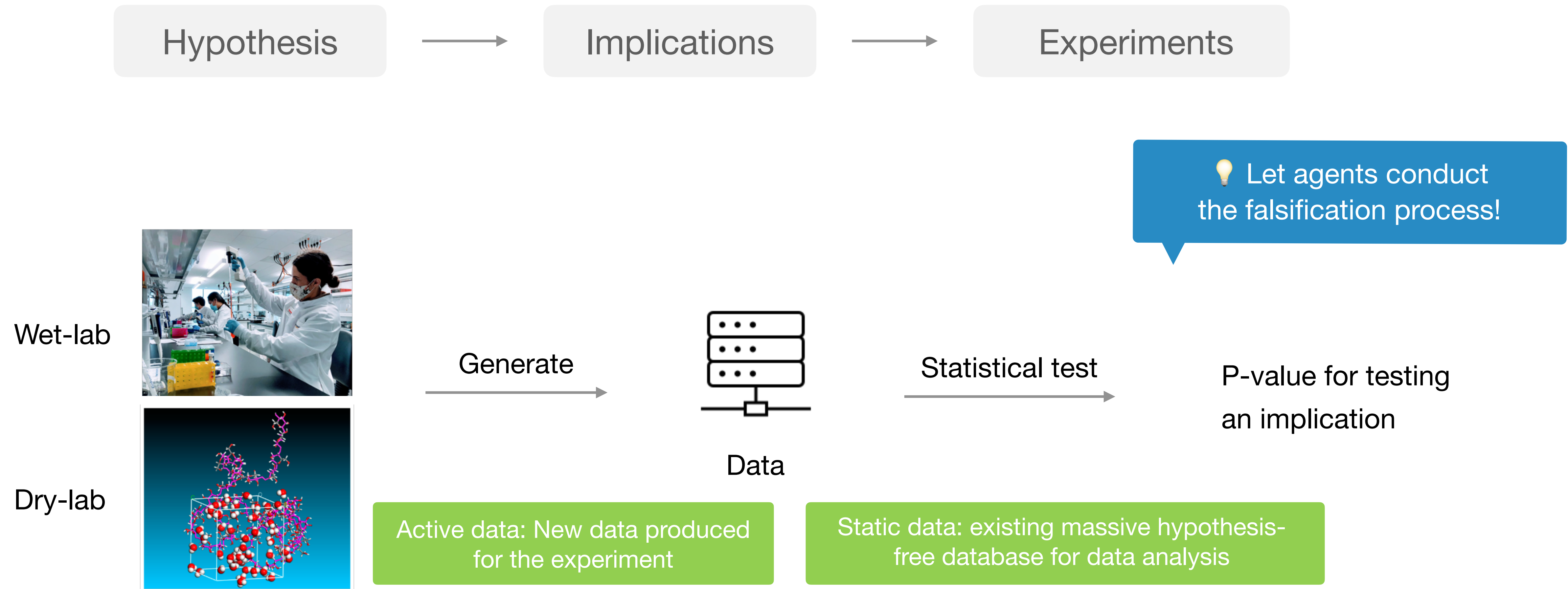
Building an agent by imitating the falsification process



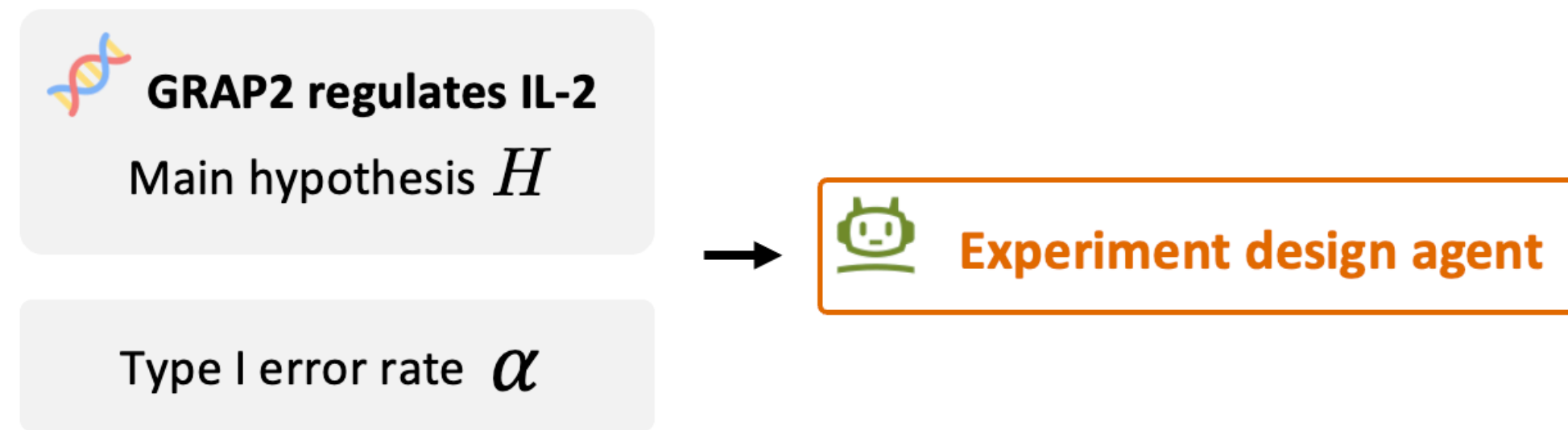
Building an agent by imitating the falsification process



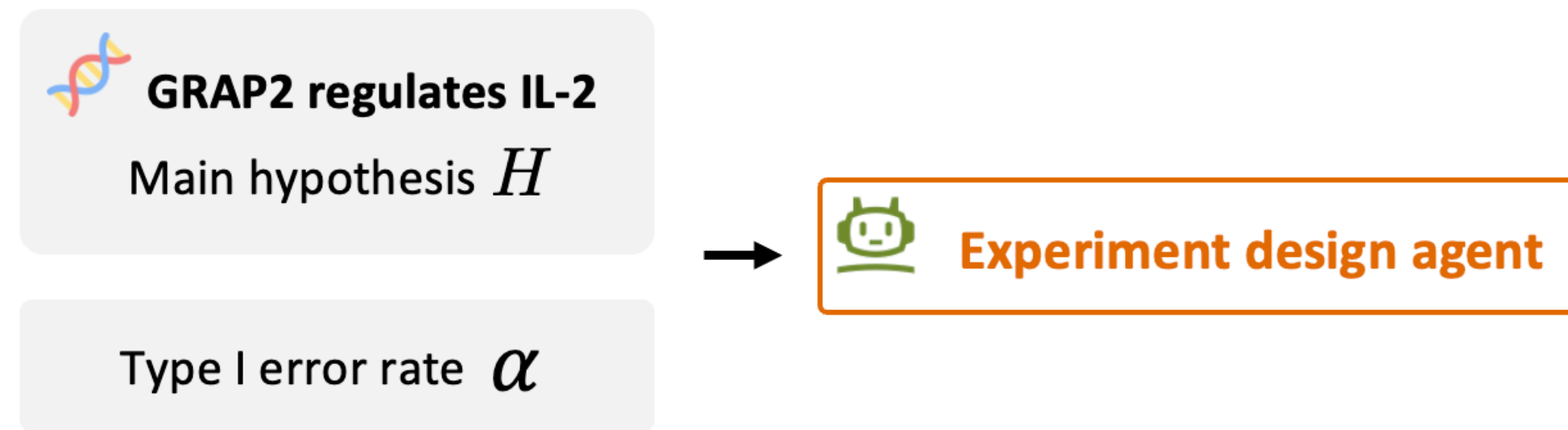
Building an agent by imitating the falsification process



Building an agent by imitating the falsification process



Building an agent by imitating the falsification process

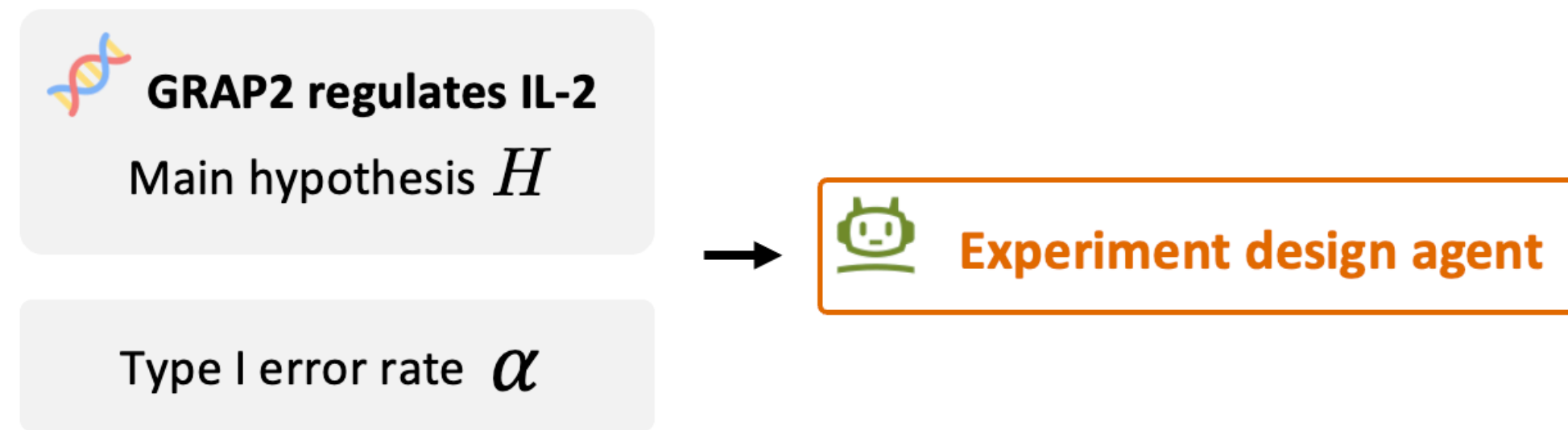


Designing a falsification experiment requires

- ▶ Reasoning
- ▶ Domain knowledge
- ▶ Creativity

We use an LLM as **experiment design agent**

Building an agent by imitating the falsification process



Designing a falsification experiment requires

- ▶ Reasoning
- ▶ Domain knowledge
- ▶ Creativity

We use an LLM as **experiment design agent**

- Input: the main hypothesis H
- Output: a designed experiment
 - ▶ Implied sub-hypothesis
 - ▶ The null and alternative
 - ▶ Experimental protocol (what data to collect/analyze)

Building an agent by imitating the falsification process



Designing a falsification experiment requires

- ▶ Reasoning
- ▶ Domain knowledge
- ▶ Creativity

We use an LLM as **experiment design agent**

Execution agent implements the experiment

- ▶ Data analysis
- ▶ Wet-lab data collection (dream)
- ▶ Computer simulations

Basically, another LLM

- Input: the main hypothesis H
- Output: a designed experiment
 - ▶ Implied sub-hypothesis
 - ▶ The null and alternative
 - ▶ Experimental protocol (what data to collect/analyze)

Building an agent by imitating the falsification process



Designing a falsification experiment requires

- ▶ Reasoning
- ▶ Domain knowledge
- ▶ Creativity

We use an LLM as **experiment design agent**

- Input: the main hypothesis H
- Output: a designed experiment
 - ▶ Implied sub-hypothesis
 - ▶ The null and alternative
 - ▶ Experimental protocol (what data to collect/analyze)

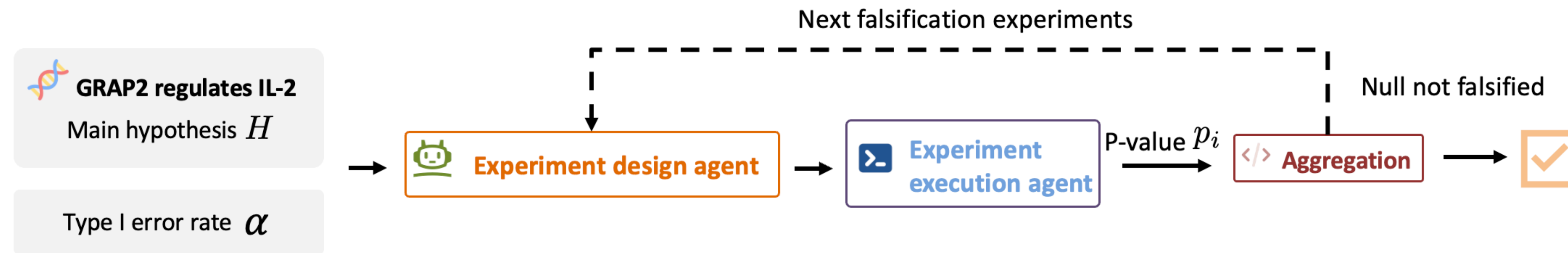
Execution agent implements the experiment

- ▶ Data analysis
- ▶ Wet-lab data collection (dream)
- ▶ Computer simulations

Basically, another LLM

- Input: the experimental design
- Output: p-values for testing the sub-hypothesis

Building an agent by imitating the falsification process



Designing a falsification experiment requires

- ▶ Reasoning
- ▶ Domain knowledge
- ▶ Creativity

We use an LLM as **experiment design agent**

- Input: the main hypothesis H
- Output: a designed experiment
 - ▶ Implied sub-hypothesis
 - ▶ The null and alternative
 - ▶ Experimental protocol (what data to collect/analyze)

Execution agent implements the experiment

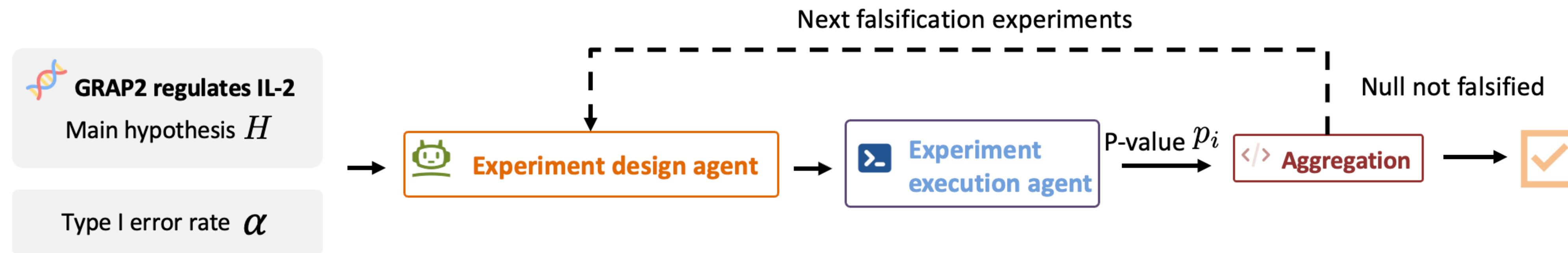
- ▶ Data analysis
- ▶ Wet-lab data collection (dream)
- ▶ Computer simulations

Basically, another LLM

- Input: the experimental design
- Output: p-values for testing the sub-hypothesis

Finally, continue the process if one round of falsification isn't enough: Aggregation!

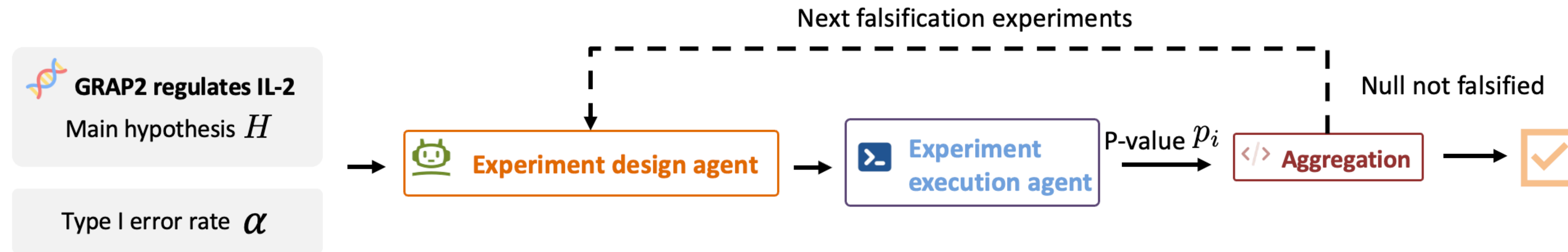
Sequential safe testing with error control



Continue the process if one round of falsification isn't enough. Ideally,

- ▶ Adaptively determine whether to continue
- ▶ Combine diverse evidence from tests for different sub-hypotheses
- ▶ Maintain statistical rigor for final decision

Sequential safe testing with error control



Continue the process if one round of falsification isn't enough. Ideally,

- ▶ Adaptively determine whether to continue
- ▶ Combine diverse evidence from tests for different sub-hypotheses
- ▶ Maintain statistical rigor for final decision

JOURNAL ARTICLE

Safe testing FREE

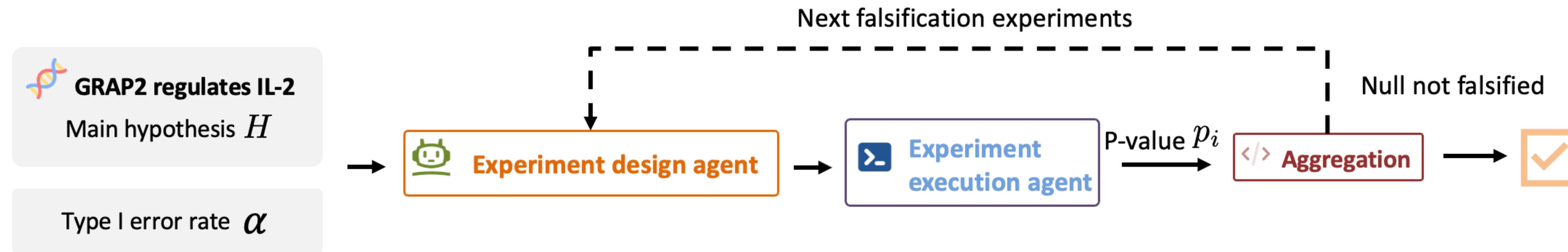
Peter Grünwald , Rianne de Heide, Wouter Koolen [Author Notes](#)

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 86, Issue 5, November 2024, Pages 1091–1128, <https://doi.org/10.1093/jrsssb/qkae011>

Published: 07 March 2024 **Article history** ▼

Sequential testing with e-values!

Sequential safe testing with error control



Continue the process if one round of falsification isn't enough. Ideally,

- ▶ Adaptively determine whether to continue
- ▶ Combine diverse evidence from tests for different sub-hypotheses
- ▶ Maintain statistical rigor for final decision

JOURNAL ARTICLE

Safe testing FREE

Peter Grünwald , Rianne de Heide, Wouter Koolen [Author Notes](#)

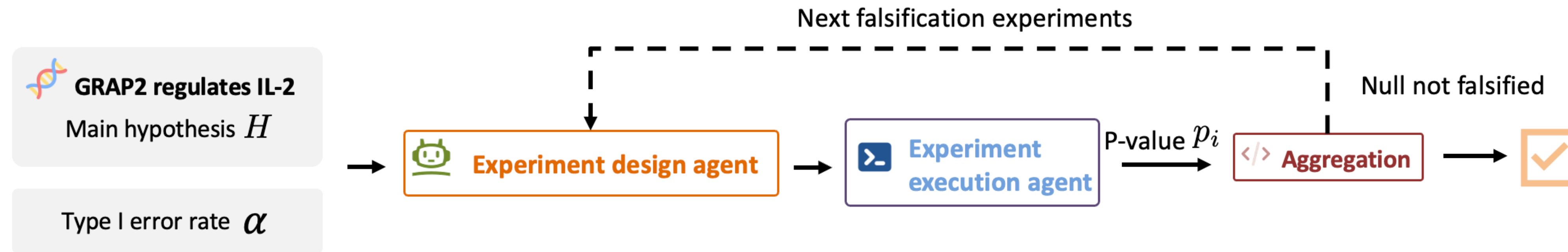
Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 86, Issue 5, November 2024, Pages 1091–1128, <https://doi.org/10.1093/jrsssb/qkae011>

Published: 07 March 2024 [Article history](#) ▼

E-value: non-negative random variable whose expectation under the null is at most 1. [Vovk and Wang, 2021]

Sequential testing with e-values!

Sequential safe testing with error control



Continue the process if one round of falsification isn't enough. Ideally,

- ▶ Adaptively determine whether to continue
- ▶ Combine diverse evidence from tests for different sub-hypotheses
- ▶ Maintain statistical rigor for final decision

JOURNAL ARTICLE

Safe testing FREE

Peter Grünwald , Rianne de Heide, Wouter Koolen [Author Notes](#)

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 86, Issue 5, November 2024, Pages 1091–1128, <https://doi.org/10.1093/jrsssb/qkae011>

Published: 07 March 2024 **Article history** ▼

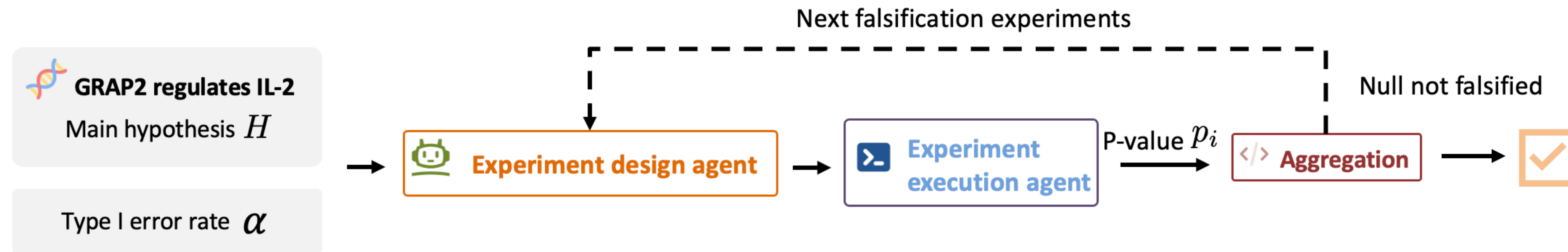
E-value: non-negative random variable whose expectation under the null is at most 1. [Vovk and Wang, 2021]

Benefits of e-values:

easy to combine, flexible to construct, optional stopping, valid testing, interpretable as “wealth”...

Sequential testing with e-values!

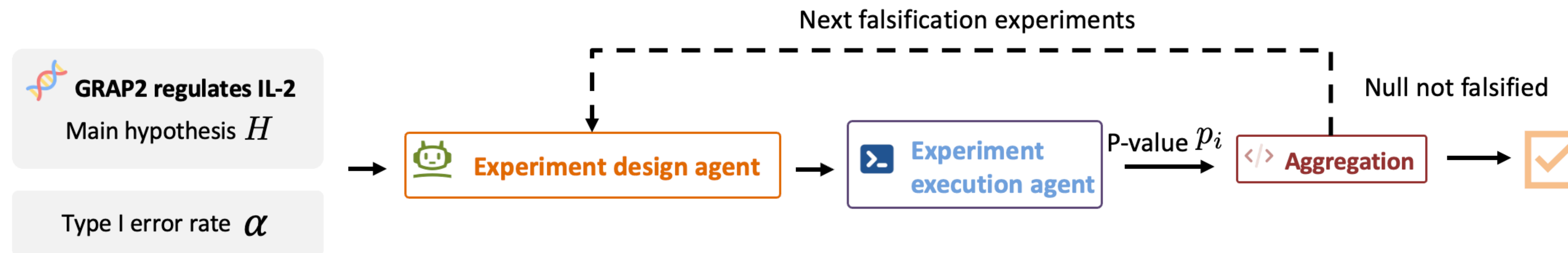
POPPER: theoretical framework



Goal: valid type-I error control $\mathbb{P}(\text{reject}) \leq \alpha$ if the main null hypothesis H_0 is true

- ▶ Rigorous error control for scientific discovery
- ▶ Quantify strength of evidence even for early-stage exploration

POPPER: theoretical framework



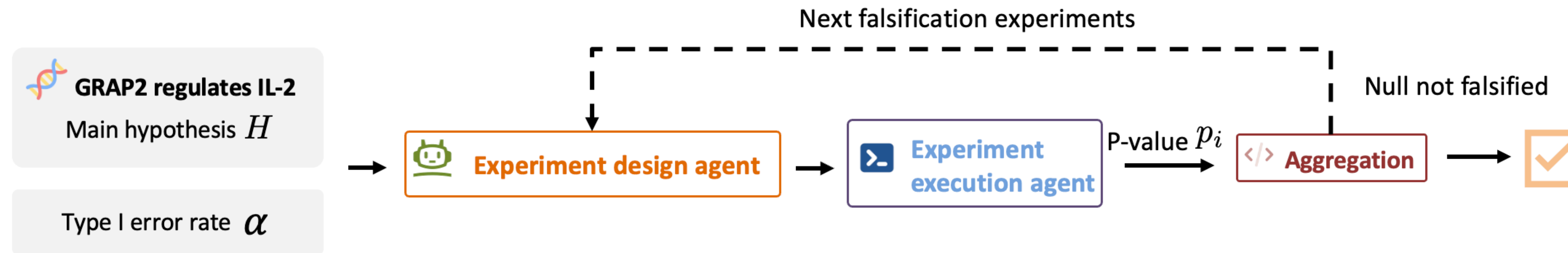
Goal: valid type-I error control $\mathbb{P}(\text{reject}) \leq \alpha$ if the main null hypothesis H_0 is true

- ▶ Rigorous error control for scientific discovery
- ▶ Quantify strength of evidence even for early-stage exploration

Three key conditions to guarantee type-I error control

1. Implication
2. E-value validity
3. Optional stopping

POPPER: theoretical framework



Goal: valid type-I error control $\mathbb{P}(\text{reject}) \leq \alpha$ if the main null hypothesis H_0 is true

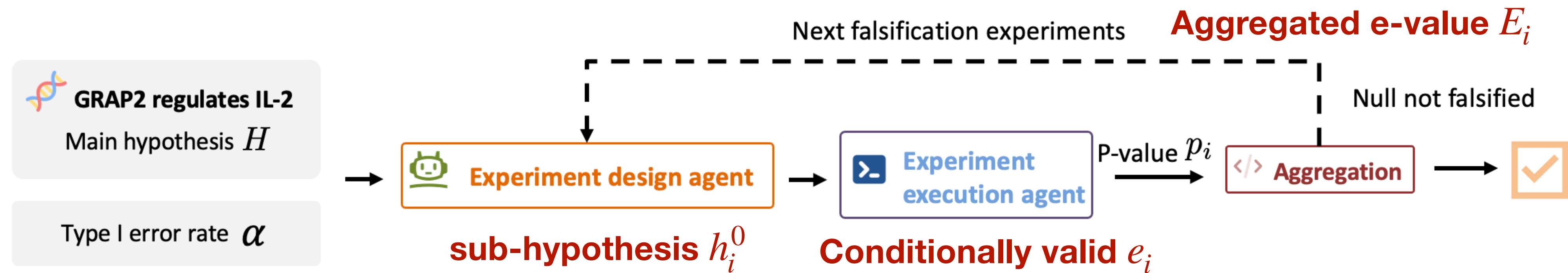
- ▶ Rigorous error control for scientific discovery
- ▶ Quantify strength of evidence even for early-stage exploration

Three key conditions to guarantee type-I error control

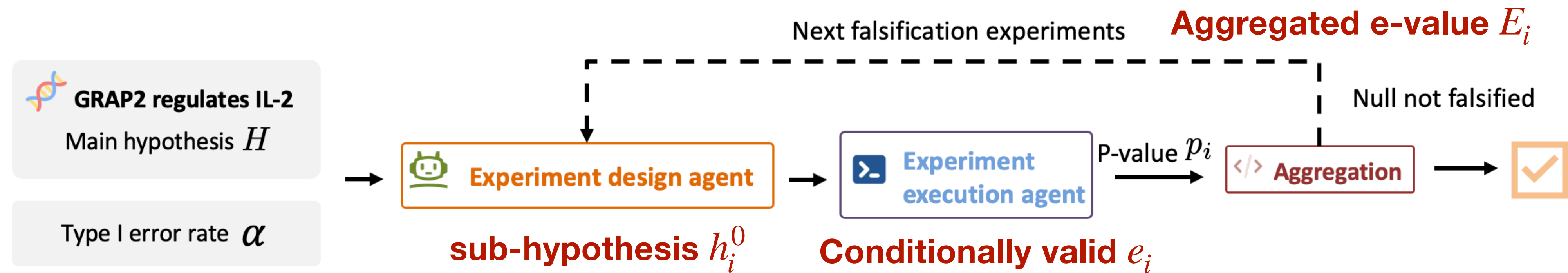
1. **Implication**
2. **E-value validity**
3. **Optional stopping**

Then, aggregating the e-values in existing rounds leads to a valid e-value for testing H_0

POPPER: theoretical framework

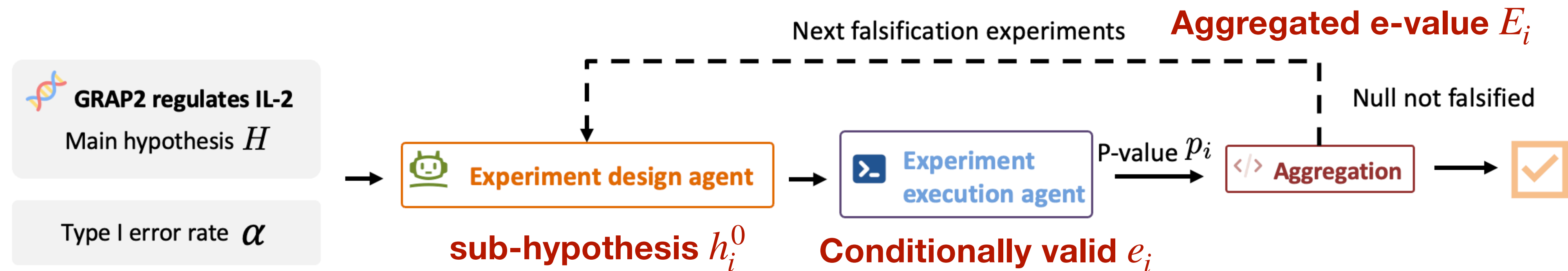


POPPER: theoretical framework



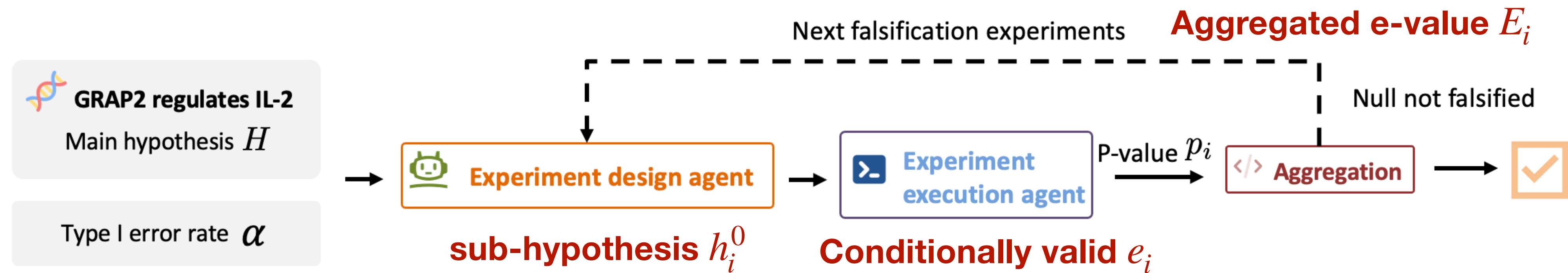
- 1. Implication:** In each round i , the design agent proposes a sub-hypothesis h_i^0 such that H_0 implies h_i^0
 $\leadsto$ agent self-refinement + external valid checker (another LLM)

POPPER: theoretical framework



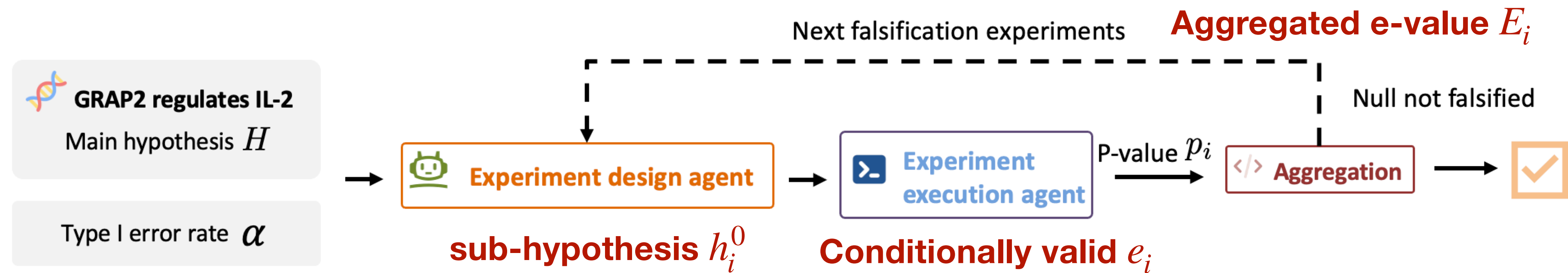
- 1. Implication:** In each round i , the design agent proposes a sub-hypothesis h_i^0 such that H_0 implies h_i^0
 $\leadsto$ agent self-refinement + external valid checker (another LLM)
- 2. E-value validity:** Let $\mathcal{D}_i$ be the data used in all rounds before i , then the execution agent in round i produces a nonnegative random variable e_i obeying $\mathbb{E}[e_i \mid \mathcal{D}_i] \leq 1$ under the sub-null h_i^0
 $\leadsto$ use new dataset + well-executed statistical test at each step

POPPER: theoretical framework



- 1. Implication:** In each round i , the design agent proposes a sub-hypothesis h_i^0 such that H_0 implies h_i^0
 $\leadsto$ agent self-refinement + external valid checker (another LLM)
- 2. E-value validity:** Let $\mathcal{D}_i$ be the data used in all rounds before i , then the execution agent in round i produces a nonnegative random variable e_i obeying $\mathbb{E}[e_i \mid \mathcal{D}_i] \leq 1$ under the sub-null h_i^0
 $\leadsto$ use new dataset + well-executed statistical test at each step
- 3. Optional stopping:** Define aggregated e-value as $E_i = \prod_{t \leq i} e_t$. We stop if and only if $i \geq T_0$ or $E_i \geq 1/\alpha$

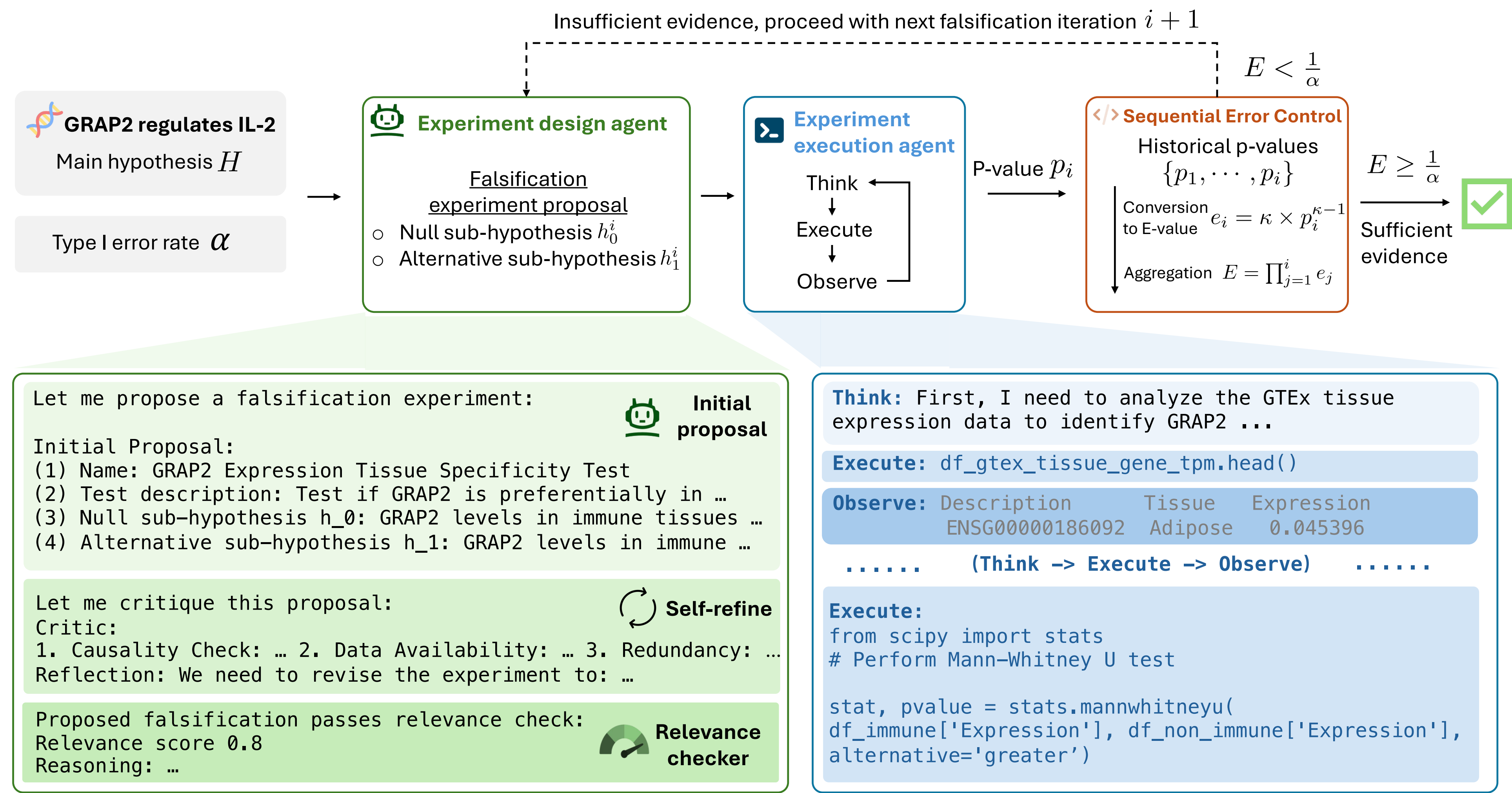
POPPER: theoretical framework



- 1. Implication:** In each round i , the design agent proposes a sub-hypothesis h_i^0 such that H_0 implies h_i^0
 $\leadsto$ agent self-refinement + external valid checker (another LLM)
- 2. E-value validity:** Let $\mathcal{D}_i$ be the data used in all rounds before i , then the execution agent in round i produces a nonnegative random variable e_i obeying $\mathbb{E}[e_i \mid \mathcal{D}_i] \leq 1$ under the sub-null h_i^0
 $\leadsto$ use new dataset + well-executed statistical test at each step
- 3. Optional stopping:** Define aggregated e-value as $E_i = \prod_{t \leq i} e_t$. We stop if and only if $i \geq T_0$ or $E_i \geq 1/\alpha$

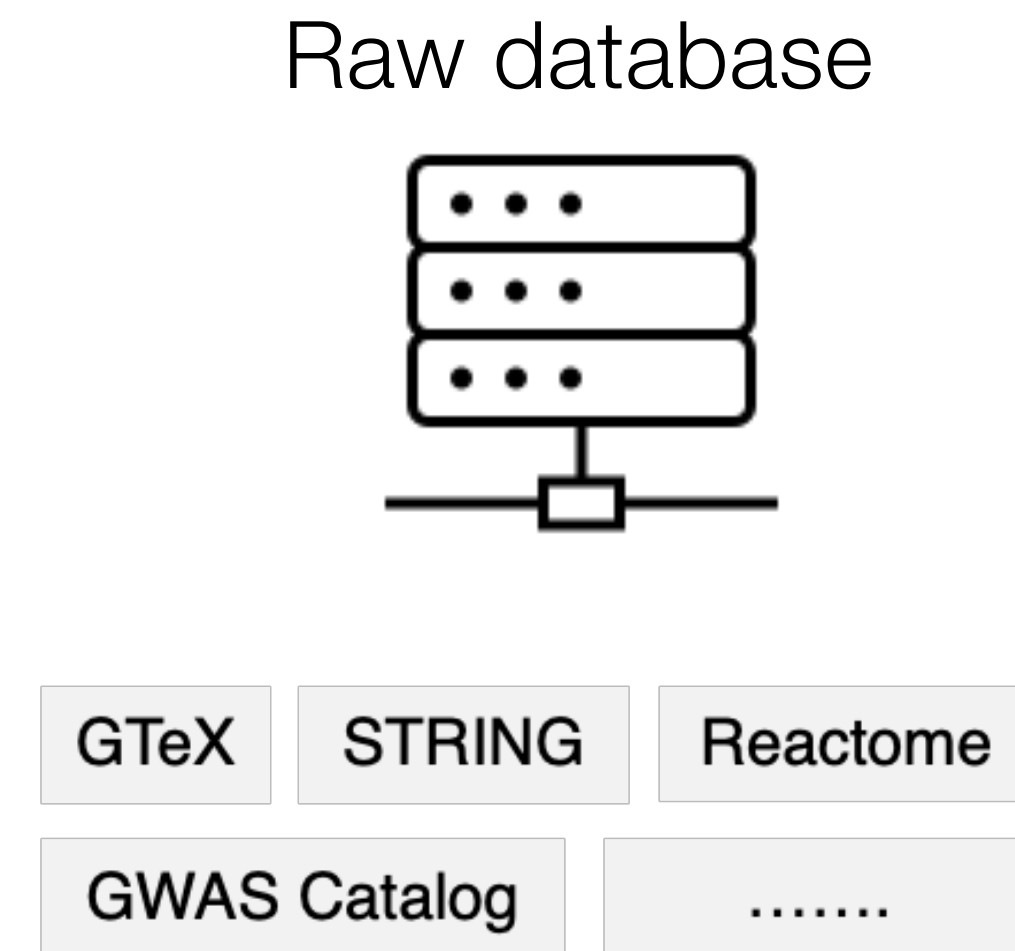
The terminal e-value E_τ obeys $\mathbb{E}[E_\tau] \leq 1$ under the global null H_0 ,
hence the test $\mathbf{1}\{E_\tau > 1/\alpha\}$ controls the type-I error

POPPER



Instantiation of POPPER with static data

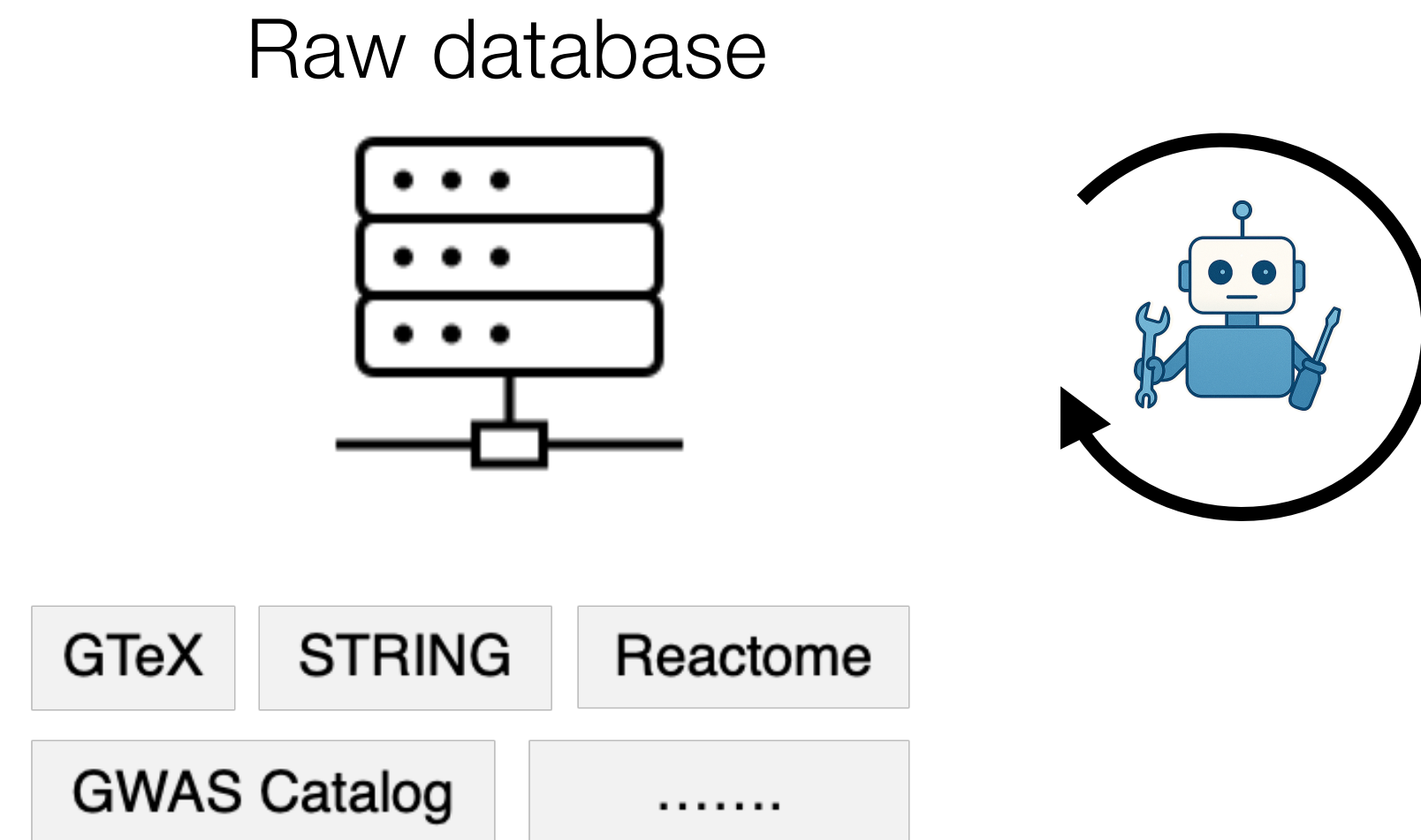
Offline data: massive atlas of hypothesis-free diverse multi-scale and multi-modal datasets



Experiment execution: python coding environment

Instantiation of POPPER with static data

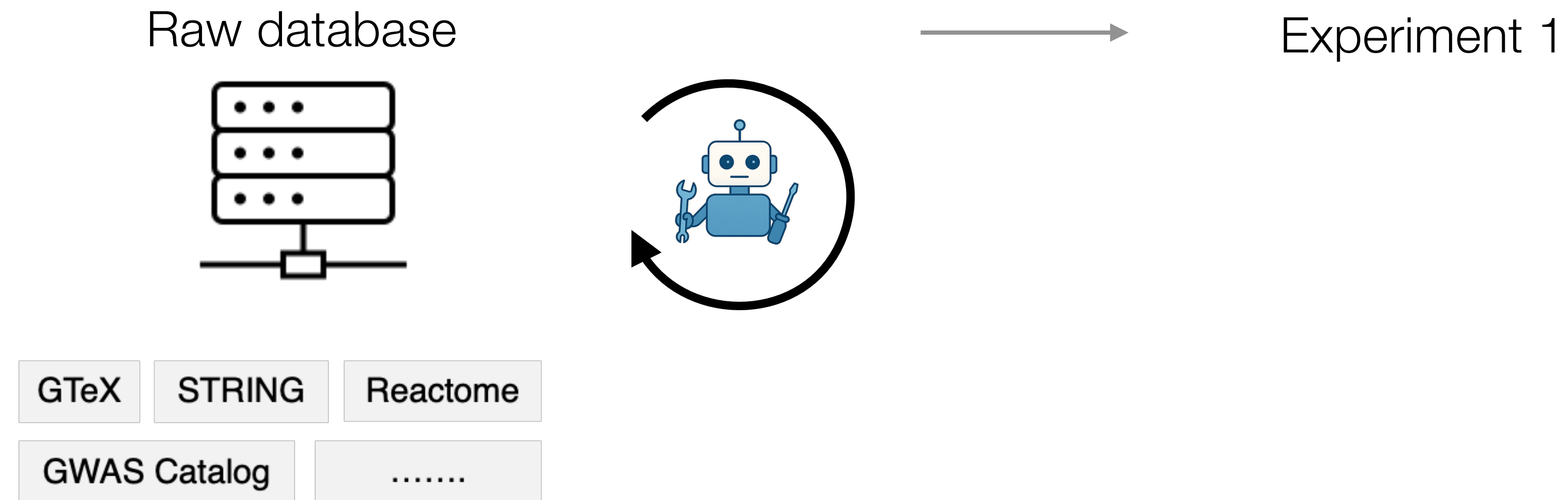
Offline data: massive atlas of hypothesis-free diverse multi-scale and multi-modal datasets



Experiment execution: python coding environment

Instantiation of POPPER with static data

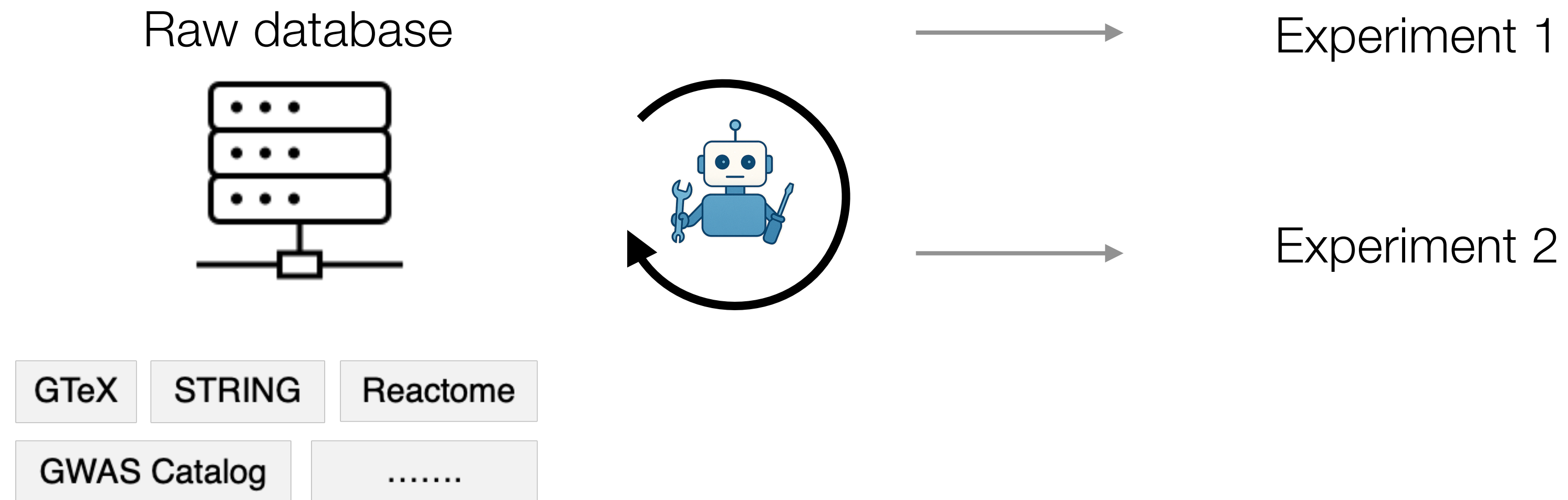
Offline data: massive atlas of hypothesis-free diverse multi-scale and multi-modal datasets



Experiment execution: python coding environment

Instantiation of POPPER with static data

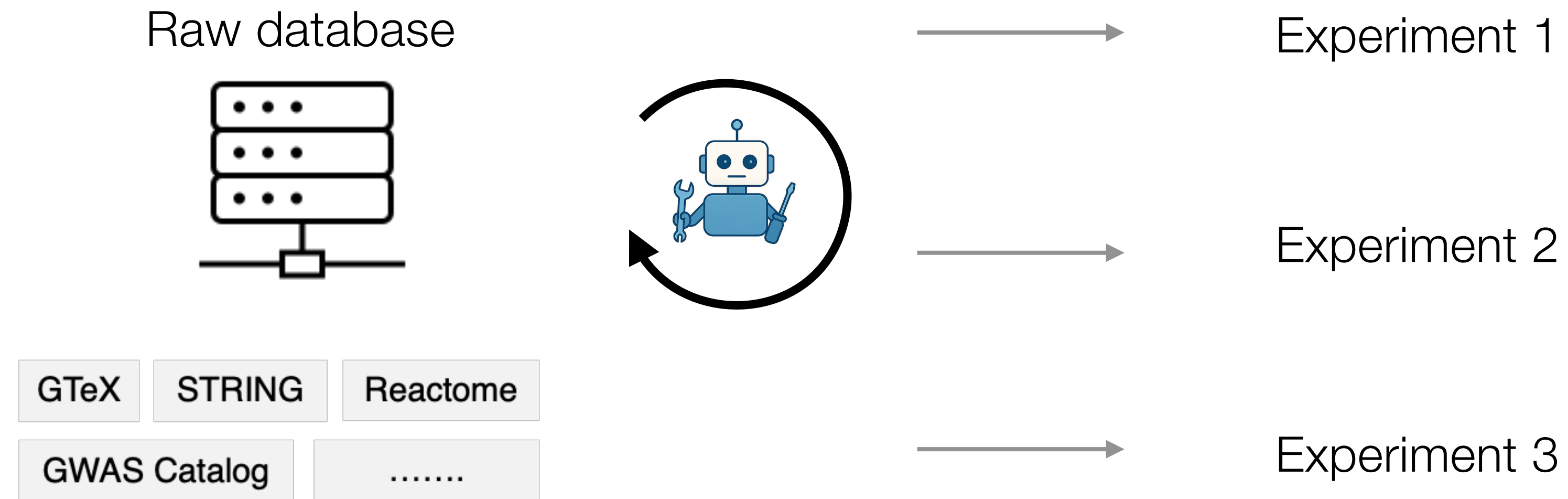
Offline data: massive atlas of hypothesis-free diverse multi-scale and multi-modal datasets



Experiment execution: python coding environment

Instantiation of POPPER with static data

Offline data: massive atlas of hypothesis-free diverse multi-scale and multi-modal datasets



Experiment execution: python coding environment

Evaluation domain

- ▶ Target validation in biology (TargetVal):
 - ▶ A key task in drug discovery
 - ▶ Validating hypothesis of the form “Gene A regulates Phenotype B”
 - ▶ Ground truth: expensive CRISPR experiment
- ▶ Discovery Bench: 86 non-null hypotheses mined from literature with paired data
 - ▶ Sociology, biology, humanities, economics, engineering, and meta-science
 - ▶ “From 1700 BCE onwards, the use of hatchets and swords increased while the use of daggers decreased”

DISCOVERYBENCH: Towards Data-Driven Discovery with Large Language Models

Bodhisattwa Prasad Majumder^{*1} Harshit Surana^{*12} Dhruv Agarwal^{*3}
Bhavana Dalvi Mishra^{*1} Abhijeetsingh Meena² Aryan Prakhar² Tirth Vora²
Tushar Khot¹ Ashish Sabharwal¹ Peter Clark¹

¹Allen Institute for AI ²OpenLocus ³University of Massachusetts Amherst

Website: <https://github.com/allenai/discoverybench>
 <https://huggingface.co/datasets/allenai/discoverybench>
^{*}equal contributions

Curated static database for biology

```
df_gtex_tissue_gene_tpm:
{'Description': 'ENSG00000186092', 'Tissue': 'Adipose - Subcutaneous', 'Expression': 0.0453961, 'Gene': 'OR4F5'}

df_gwas_catalog:
{'DATE ADDED TO CATALOG': '2018-07-30', 'PUBMEDID': 27618447, 'FIRST AUTHOR': 'Surendran P', 'DATE': '2016-10-01', 'JOURNAL': 'Nat Genet', 'LINK': 'www.ncbi.nlm.nih.gov/pubmed/27618447', 'DISEASE/TRAIT': 'systolic blood pressure', 'INITIAL SAMPLE SIZE': 'up to 165,276 European ancestry individuals, up to 27,487 South Asian ancestry individuals, 22,077 African American individuals', 'REGION': '3q26.2', 'CHR_ID': '3', 'CHR_POS': '169383111', 'REPORTED GENE(S)': 'MECOM', 'MAPPED_GENE': 'MECOM', 'DOWNSTREAM_GENE_DISTANCE': nan, 'STRONGEST SNP-RISK ALLELE': 'rs448378-A', 'SNPS': 'rs448378', 'MERGED': 0, 'SNP_ID_CURRENT': 448378.0, 'CONTEXT': 'intron_variant', 'BETA': 0.0196, '95% CI (TEXT)': 'mmHg decrease', 'PLATFORM [SNPS PASSING QC]': 'Illumina [242296]', 'CNV': 'N'}

df_gene_info:
{'gene_id': 'ENSG00000228037', 'transcript_id': 'ENST00000424215', 'chr': '1', 'gene_start': 2581560, 'gene_end': 2584533, 'strand': 1, 'transcript_start': 2581560, 'gene_type': 'lncRNA'}

df_genetic_interaction:
{'interaction_id': 206363, 'gene_a_id': 'YCR011C', 'gene_b_id': 'YCL025C', 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:16269340', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_genebase_synonymous_filtered:
{'annotation': 'synonymous', 'Pvalue': 0.72199, 'Pvalue_Burden': 0.85198, 'Pvalue_SKAT': 0.52215, 'BETA_Burden': 0.0034703, 'SE_Burden': 0.018598, 'gene': 'TSPAN6', 'phenotype': 'TSPAN6'}

df_genebase_missense_LC_filtered:
{'annotation': 'missense_LC', 'Pvalue': 0.85434, 'Pvalue_Burden': 0.69637, 'Pvalue_SKAT': 0.87135, 'BETA_Burden': -0.0046974, 'SE_Burden': 0.012037, 'gene': 'TSPAN6', 'phenotype': 'TSPAN6'}

df_genebase_pLoF_filtered:
{'annotation': 'pLoF', 'Pvalue': 0.59145, 'Pvalue_Burden': 0.96398, 'Pvalue_SKAT': 0.42053, 'BETA_Burden': -0.00053797, 'SE_Burden': 0.011913, 'gene': 'DFP1', 'phenotype': 'DFP1'}

df_affinity_capture_ns:
{'interaction_id': 241421, 'gene_a_id': 'ENSG00000169217', 'gene_b_id': 'ENSG00000144028', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:17353931', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_two_hybrid:
{'interaction_id': 269293, 'gene_a_id': 'ENSG00000114395', 'gene_b_id': 'ENSG00000109103', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:16169070', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_synthetic_growth_defect:
{'interaction_id': 299191, 'gene_a_id': 'SPBC30D10.04', 'gene_b_id': 'SPAC14C4.13', 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:18931302', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_affinity_capture_rna:
{'interaction_id': 637509, 'gene_a_id': 'YDR515W', 'gene_b_id': 'YAL030W', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:22271760', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_co_fractionation:
{'interaction_id': 740709, 'gene_a_id': 'ENSG00000160201', 'gene_b_id': 'ENSG00000063244', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:22939629', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_synthetic_lethality:
{'interaction_id': 818564, 'gene_a_id': 'YLR410C', 'gene_b_id': 'YLR103C', 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:23390603', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_dosage_growth_defect:
{'interaction_id': 908565, 'gene_a_id': 'YBL009W', 'gene_b_id': 'YPR050W', 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:22282571', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_proximity_label_ns:
{'interaction_id': 913336, 'gene_a_id': 'ENSG00000206560', 'gene_b_id': 'ENSG00000103194', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:24255178', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_synthetic_rescue:
{'interaction_id': 986367, 'gene_a_id': 'YJR017C', 'gene_b_id': 'YHR114W', 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:24470217', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_reconstituted_complex:
{'interaction_id': 2841862, 'gene_a_id': 'ENSG00000169083', 'gene_b_id': 'ENSG00000097007', 'experimental_system_type': 'physical', 'pubmed_id': 'PUBMED:24728074', 'organism_id_a': '559896', 'organism_id_b': '559896'}

df_eqtl_ukbb:
{'rs_id': 'rs111972115', 'gene_id': 'ENSG00000134250', 'cell_type_name': 'CEDAR-MONOCYTE_CD14', 'qtl_score': 4.851273983601853, 'gene_name': 'NOTCH2'}

df_pqtl_ukbb:
{'rs_id': 'rs41265213', 'gene_id': 'ENSG00000160712', 'cell_type_name': 'SCALLOP_2020-UBERON_0001969', 'qtl_score': 5.13715334001706, 'gene_name': 'IL6R'}

df_sqtl_ukbb:
{'rs_id': 'rs12756687', 'gene_id': 'ENSG00000265491', 'cell_type_name': 'GTEx-sQTL-Brain_Caudate_basal_ganglia', 'qtl_score': 3.19846622732464, 'gene_name': 'RNF115'}

df_variant_table:
{'RS': 'rs116587930', 'ID': '1:727841_G_A', 'CHR': 1, 'POS': 727841, 'A1': 'G', 'A2': 'A', 'MAF': 0.0507035}

df_trait:
{'pheno_name': '120_1', 'description': 'Birth weight known; Yes - pounds and ounces'}
```

We curated a database of 85 million observations (rows) across 22 sources in biology for our experiment.

Typical biostatisticians work with a few datasets only; the agent can perform joint analysis across massive datasets from different modalities and sub-fields of biology.

This is a difficult task that requires expertise and familiarity with many cross-disciplinary skills.

POPPER achieves type-I error control & improved power

- ▶ Evaluating type-I error
 - ▶ Randomly permute relevant columns in the databases so all hypotheses are null
- ▶ Evaluating power
 - ▶ Use the original database, but test on hypotheses known to be non-null

Methods using LLM capabilities without the “statistical layer”

Method	Type I Error ($\alpha = 0.1$)			Power		
	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG
CodeGen	0.145±0.031✗	0.020±0.014✓	0.004±0.009✓	0.378±0.066	0.140±0.022	0.040±0.042
CodeGen (o1)	0.248±0.015✗	0.013±0.012✓	0.000±0.000✓	0.419±0.028	0.250±0.100	0.183±0.076
ReAct	0.078±0.061✓	0.000±0.000✓	0.000±0.000✓	0.383±0.017	0.010±0.022	0.020±0.045
Self-Refine	0.117±0.028✗	0.100±0.069✓	0.067±0.064✓	0.476±0.066	0.183±0.029	0.067±0.064

Methods using other aggregation of p-values / construction of e-values

Fisher Combined Test	0.311±0.040✗	0.264±0.083✗	0.173±0.023✗	0.741±0.058	0.800±0.071	0.650±0.050
LLM-Likelihood ratio	0.152±0.031✗	0.016±0.014✓	0.180±0.028✗	0.428±0.034	0.185±0.074	0.357±0.132
POPPER (Ours)	0.103±0.020✓	0.082±0.046✓	0.085±0.028✓	0.638* ±0.066	0.580* ±0.125	0.591* ±0.069

POPPER achieves type-I error control & improved power

- ▶ Evaluating type-I error
 - ▶ Randomly permute relevant columns in the databases so all hypotheses are null
- ▶ Evaluating power
 - ▶ Use the original database, but test on hypotheses known to be non-null

Methods using LLM capabilities without the “statistical layer”

Method	Type I Error ($\alpha = 0.1$)			Power		
	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG
CodeGen	0.145±0.031✗	0.020±0.014✓	0.004±0.009✓	0.378±0.066	0.140±0.022	0.040±0.042
CodeGen (o1)	0.248±0.015✗	0.013±0.012✓	0.000±0.000✓	0.419±0.028	0.250±0.100	0.183±0.076
ReAct	0.078±0.061✓	0.000±0.000✓	0.000±0.000✓	0.383±0.017	0.010±0.022	0.020±0.045
Self-Refine	0.117±0.028✗	0.100±0.069✓	0.067±0.064✓	0.476±0.066	0.183±0.029	0.067±0.064

Methods using other aggregation of p-values / construction of e-values

Fisher Combined Test	0.311±0.040✗	0.264±0.083✗	0.173±0.023✗	0.741±0.058	0.800±0.071	0.650±0.050
LLM-Likelihood ratio	0.152±0.031✗	0.016±0.014✓	0.180±0.028✗	0.428±0.034	0.185±0.074	0.357±0.132
POPPER (Ours)	0.103±0.020✓	0.082±0.046✓	0.085±0.028✓	0.638* ±0.066	0.580* ±0.125	0.591* ±0.069

- ▶ Controlling how agents interact with the offline data is key to error control

POPPER achieves type-I error control & improved power

Method	Type I Error ($\alpha = 0.1$)			Power		
	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG	DiscoveryBench	TargetVal-IL2	TargetVal-IFNG
POPPER-NoReleCheck	0.134 $\pm$ 0.021✗	0.340 $\pm$ 0.139✗	0.300 $\pm$ 0.113✗	0.610 $\pm$ 0.042	0.897 $\pm$ 0.004	0.717 $\pm$ 0.126
POPPER-CodeGen	0.140 $\pm$ 0.022✗	0.105 $\pm$ 0.017✓	0.090 $\pm$ 0.045✓	0.544 $\pm$ 0.032	0.526 $\pm$ 0.133	0.450 $\pm$ 0.079
POPPER (Ours)	0.103 $\pm$ 0.020✓	0.082 $\pm$ 0.046✓	0.085 $\pm$ 0.028✓	0.638* $\pm$ 0.066	0.580* $\pm$ 0.125	0.591* $\pm$ 0.069

Compared to NoReleCheck

↪ relevant checker is a key factor in controlling type-I error (assumption 1)

Compared to CodeGen

↪ reasoning in execution agent improves power

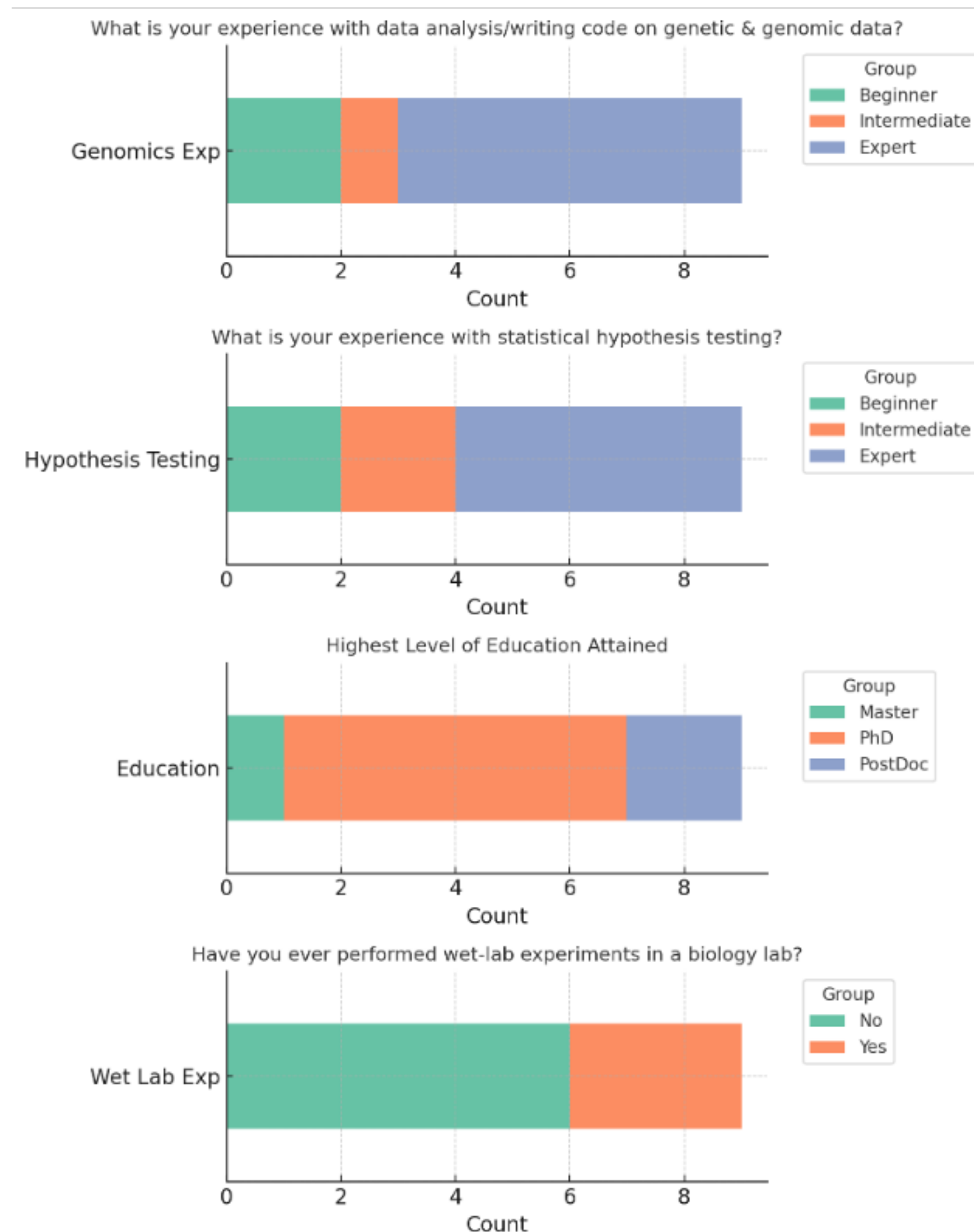
POPPER with various LLM backbones

Method	Type I Error ($\alpha = 0.1$)		Power	
	DiscoveryBench	TargetVal-IL2	DiscoveryBench	TargetVal-IL2
Claude-Haiku-3.5	0.230 $\pm$ 0.079	0.780 $\pm$ 0.120	0.844 $\pm$ 0.017	0.835 $\pm$ 0.113
Llama 3.3 70B	0.147 $\pm$ 0.036	0.116 $\pm$ 0.020	0.690 $\pm$ 0.027	0.515 $\pm$ 0.078
GPT-4o	0.143 $\pm$ 0.039	0.096 $\pm$ 0.043	0.730 $\pm$ 0.054	0.385 $\pm$ 0.102
Claude-Sonnet-3.5	0.103 $\pm$ 0.020	0.082 $\pm$ 0.046	0.638 $\pm$ 0.066	0.580* $\pm$ 0.125
o1	0.091* $\pm$ 0.015	0.031* $\pm$ 0.015	0.654* $\pm$ 0.019	0.336 $\pm$ 0.121

- ▶ The success of POPPER relies heavily on the reasoning capabilities of LLMs
 - Need to create correct implied sub-hypotheses
 - Need to generate reasonable experimental designs
 - Need to coherently generate code (tool use)
 - Need to faithfully execute the experimental plan without hallucinations

Comparing POPPER with human experts

- ▶ We recruited 9 $\geq$ graduate-level participants to conduct 2 hypothesis testing tasks from TargetVal-IL2, with access to the same databases (18 tasks are randomly assigned)



Once you open the notebook, run the following cell to start the time clock

```
In [1]: import time
start_time = time.time()
```

Instructions

Given a biology hypothesis "**Gene MAK16 regulates the production of Interleukin-2 (IL-2).**", your task is to validate it using the given raw databases by performing relevant data analysis, formulating statistical tests, and implementing them. The validation should be purely data-driven, not literature-driven. For statistical test, use significance level of $\alpha=0.1$.

Output (1) If the hypothesis is valid or not given the data (2) relevant statistics (e.g. p-value, etc)

IMPORTANT

- You must only use the `database` folder in the current task folder to perform the analysis. DO NOT use the data from the other task or any external data.
- You must NOT use LLMs or internet about the direct answer to the biological hypothesis.
- You can use internet/LLMs if you are not sure about the code syntax or library usage or statistical tests or have biological questions in general.
- You can use any python library to perform the analysis.
- The tasks are randomly sampled and may be one true & one false / all true / all false

Here are the list of available data sources with columns and example rows:

```
df_gtex_tissue_gene_tpm:
{'Description': 'ENSG00000186092', 'Tissue': 'Adipose - Subcutaneous',
 'Expression': 0.0453961, 'Gene': 'OR4F5'}
```

```
df_gene_info:
{'gene_id': 'ENSG00000228037', 'transcript_id': 'ENST00000424215', 'chr': '1',
 'gene_start': 2581560, 'gene_end': 2584533, 'strand': 1, 'transcript_start':
 2581560, 'transcript_end': 2584533, 'tss': 2581560, 'transcript_is_canonical':
 1.0, 'gene_name': nan, 'percentage_gene_gc_content': 51.11, 'gene_type':
 'lncRNA'}
```

```
df_genetic_interaction:
{'interaction_id': 206363, 'gene_a_id': 'YCR011C', 'gene_b_id': 'YCL025C',
 'experimental_system_type': 'genetic', 'pubmed_id': 'PUBMED:16269340',
 'organism_id_a': 559292, 'organism_id_b': 559292, 'throughput_type': 'High
 Throughput', 'experimental_score': '-5.6431'}
```

some dataframes are omitted for presentation purposes

```
In [3]: ## loading the datasets
import pandas as pd
import glob

database = {}
```

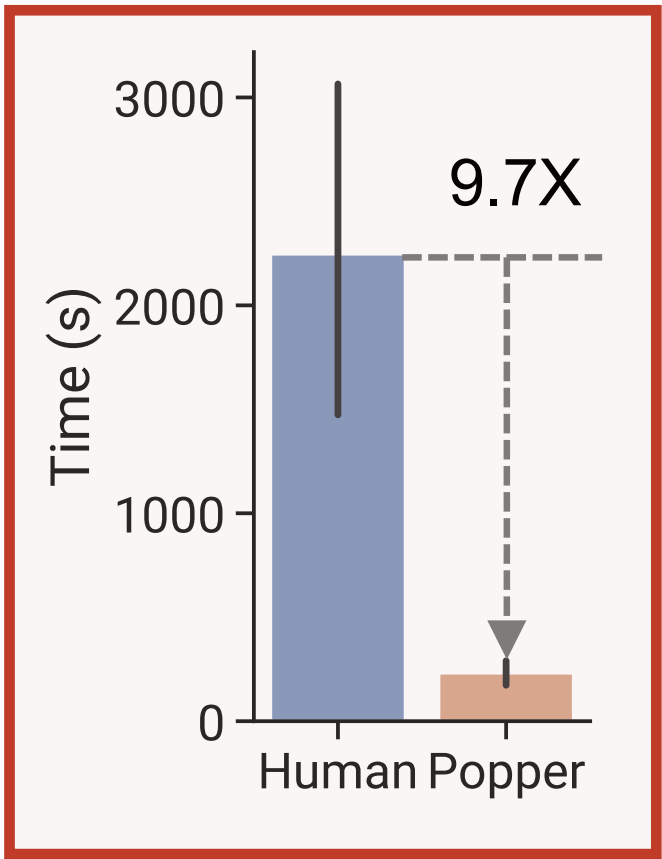
Comparing POPPER with human experts



Comparing POPPER with human experts



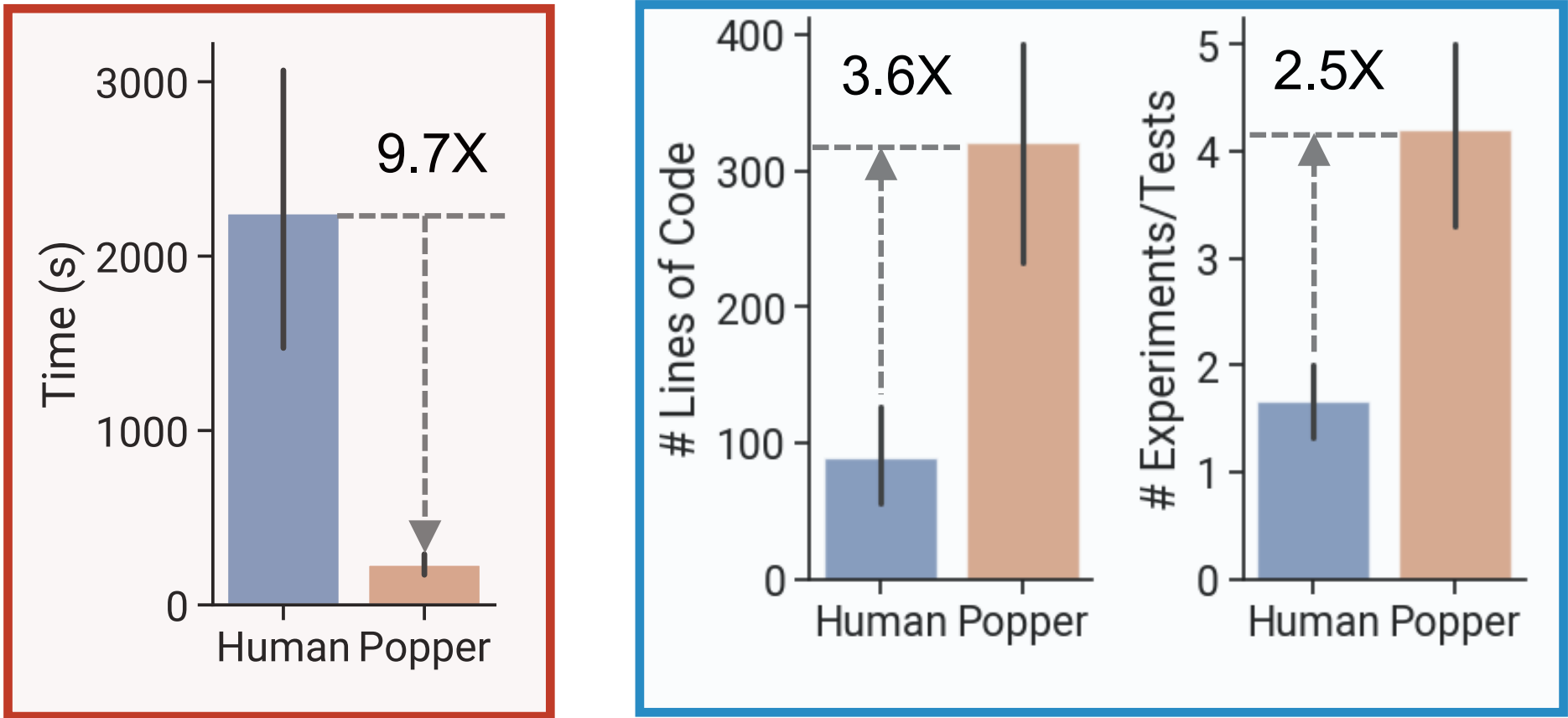
POPPER attains similar accuracy while greatly reducing the time



Comparing POPPER with human experts

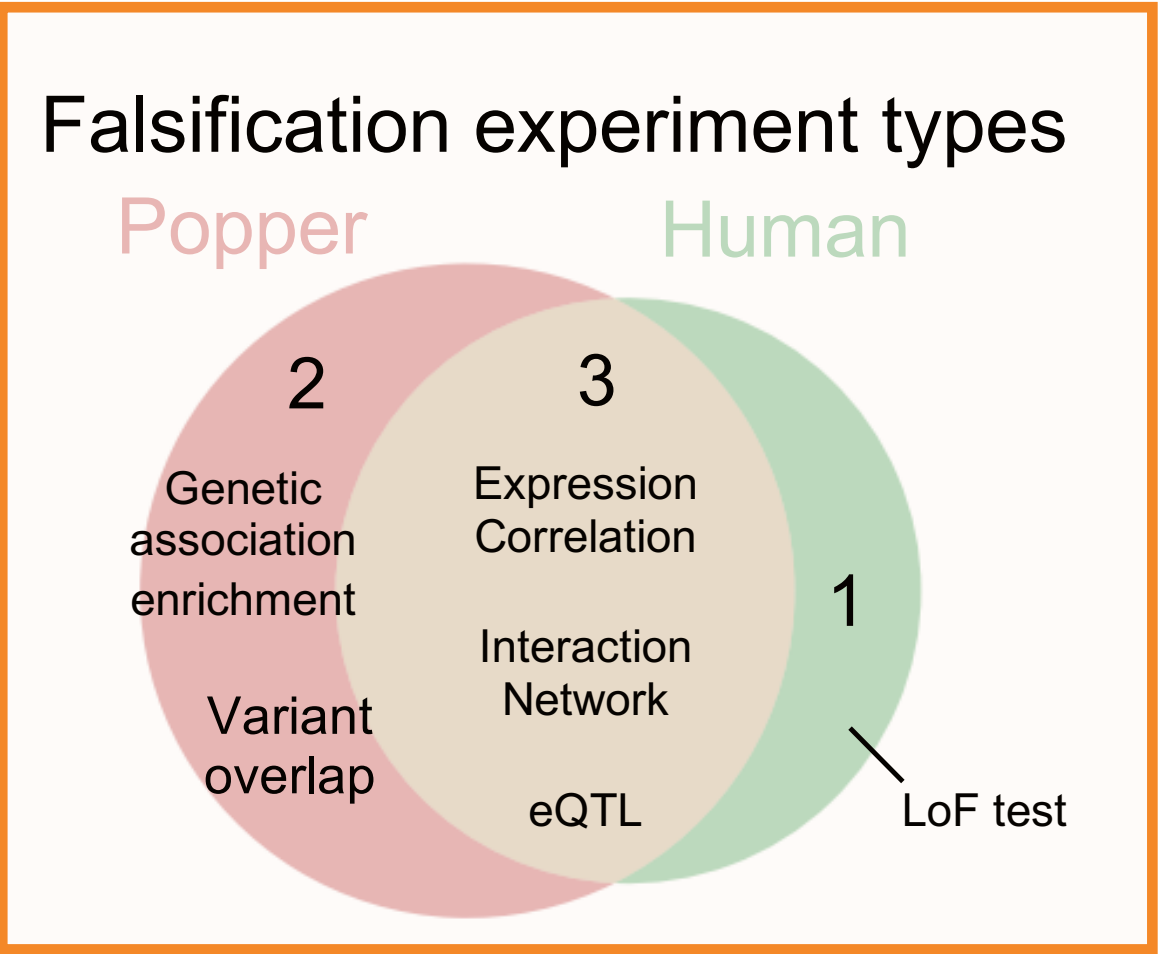


POPPER attains similar accuracy while greatly reducing the time



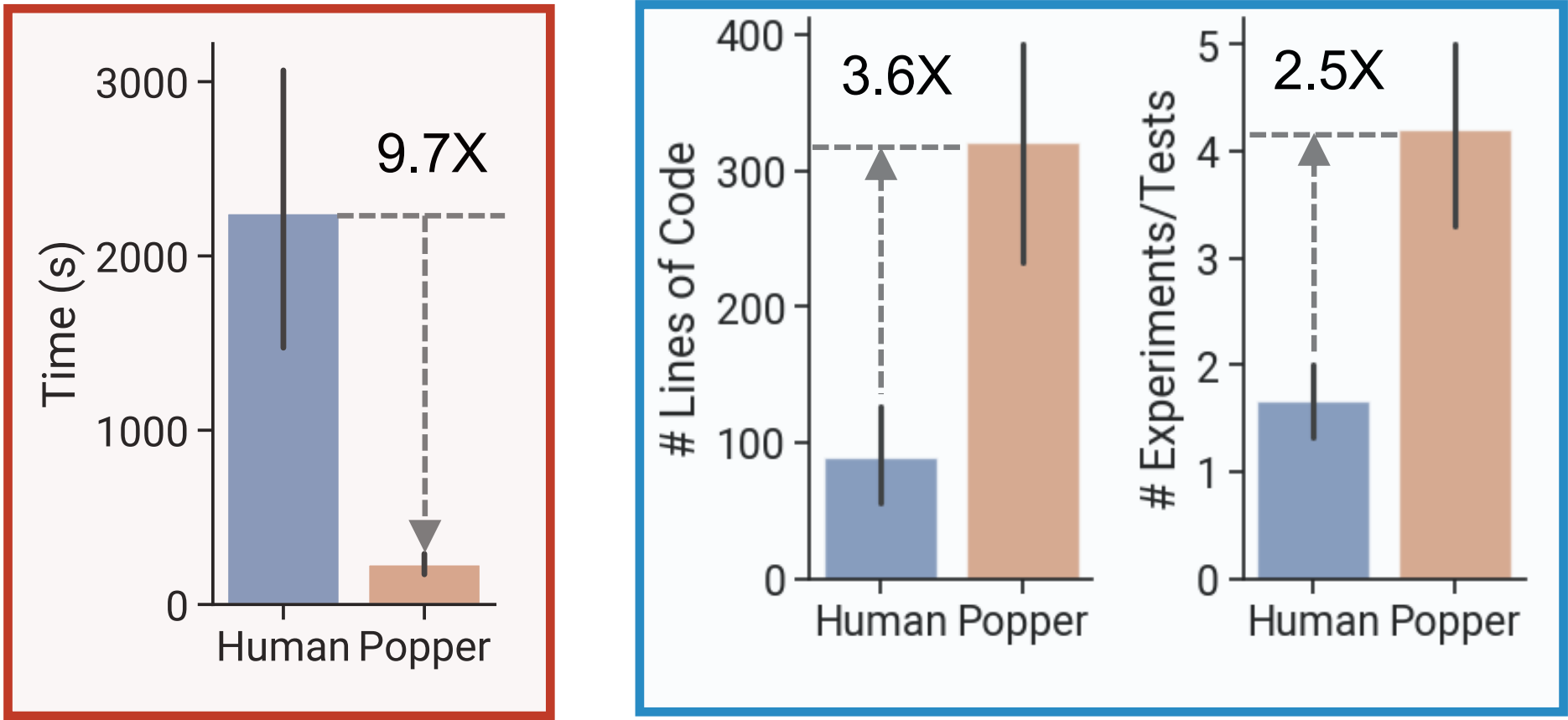
POPPER conducts more tests (more labors)

Comparing POPPER with human experts



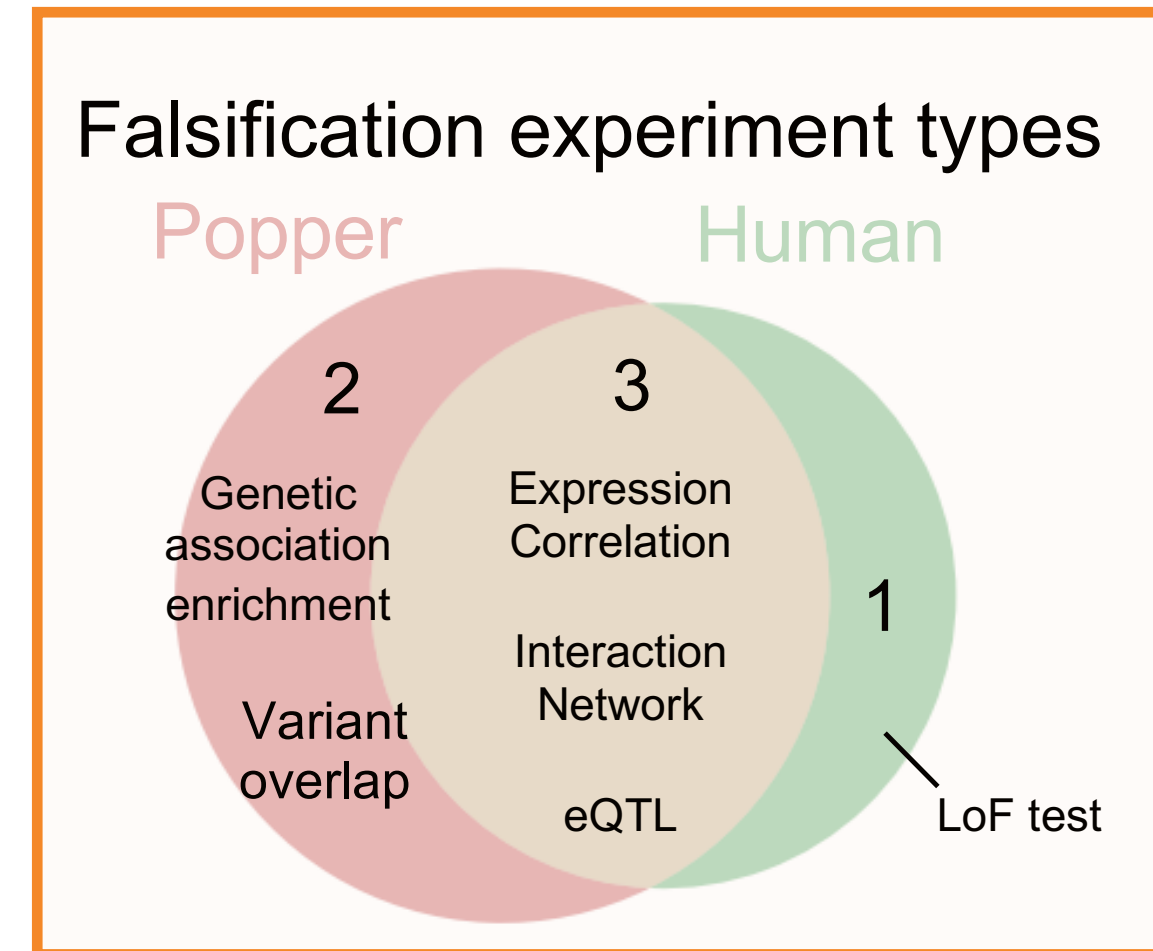
Overlap in the type of tests conducted, POPPER slightly more diverse

POPPER attains similar accuracy while greatly reducing the time



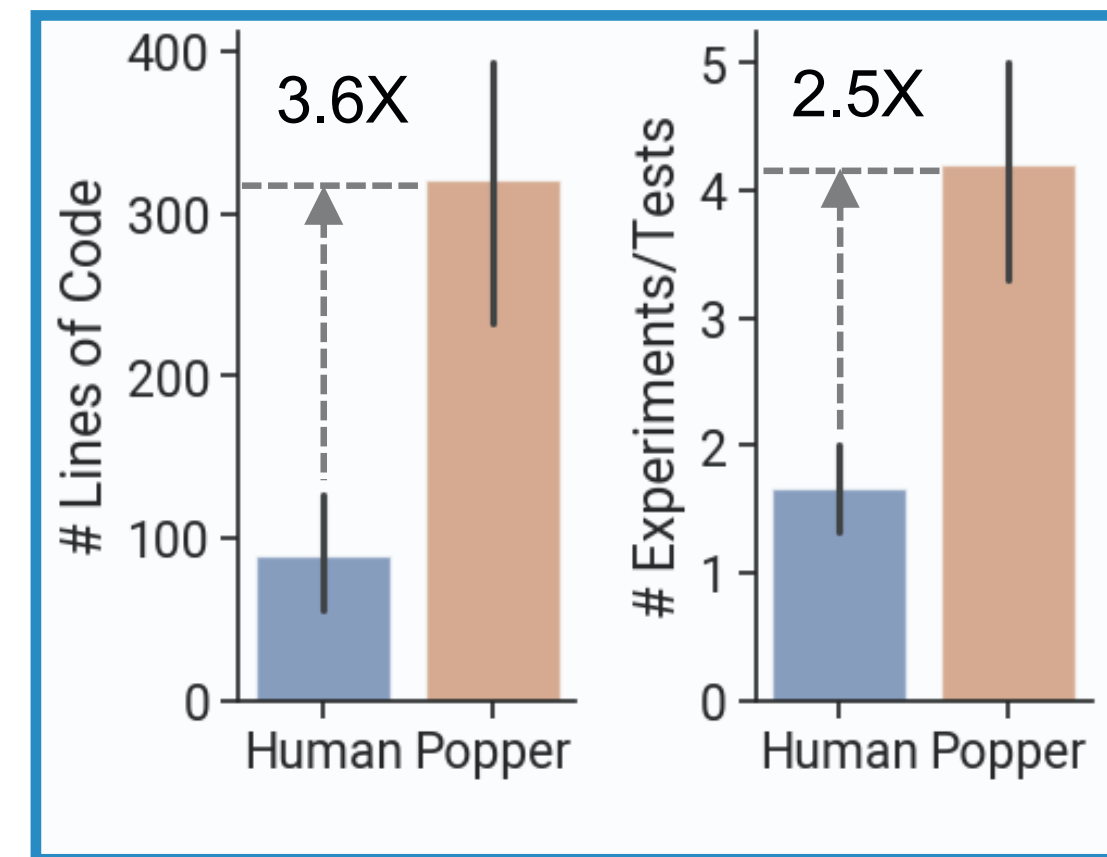
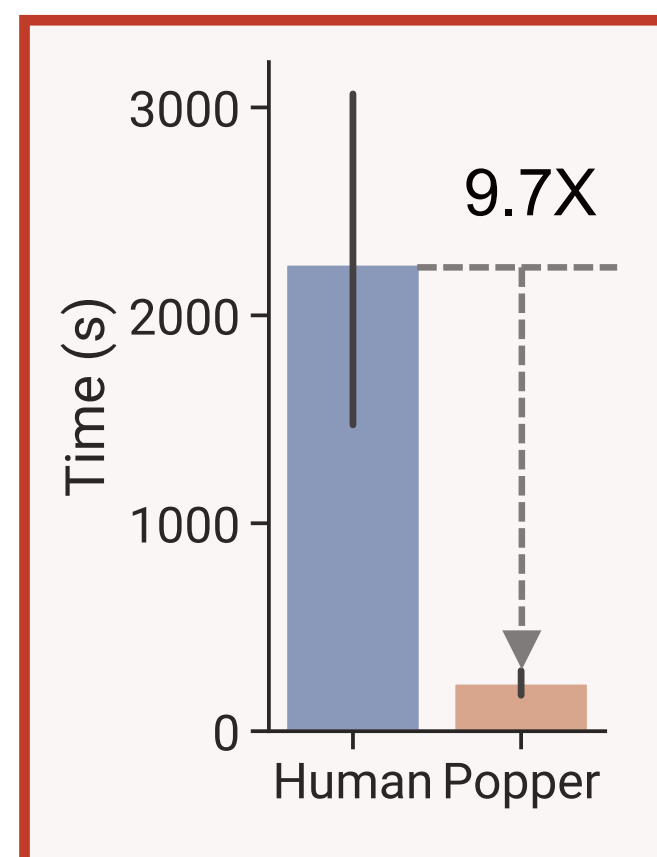
POPPER conducts more tests (more labors)

Comparing POPPER with human experts

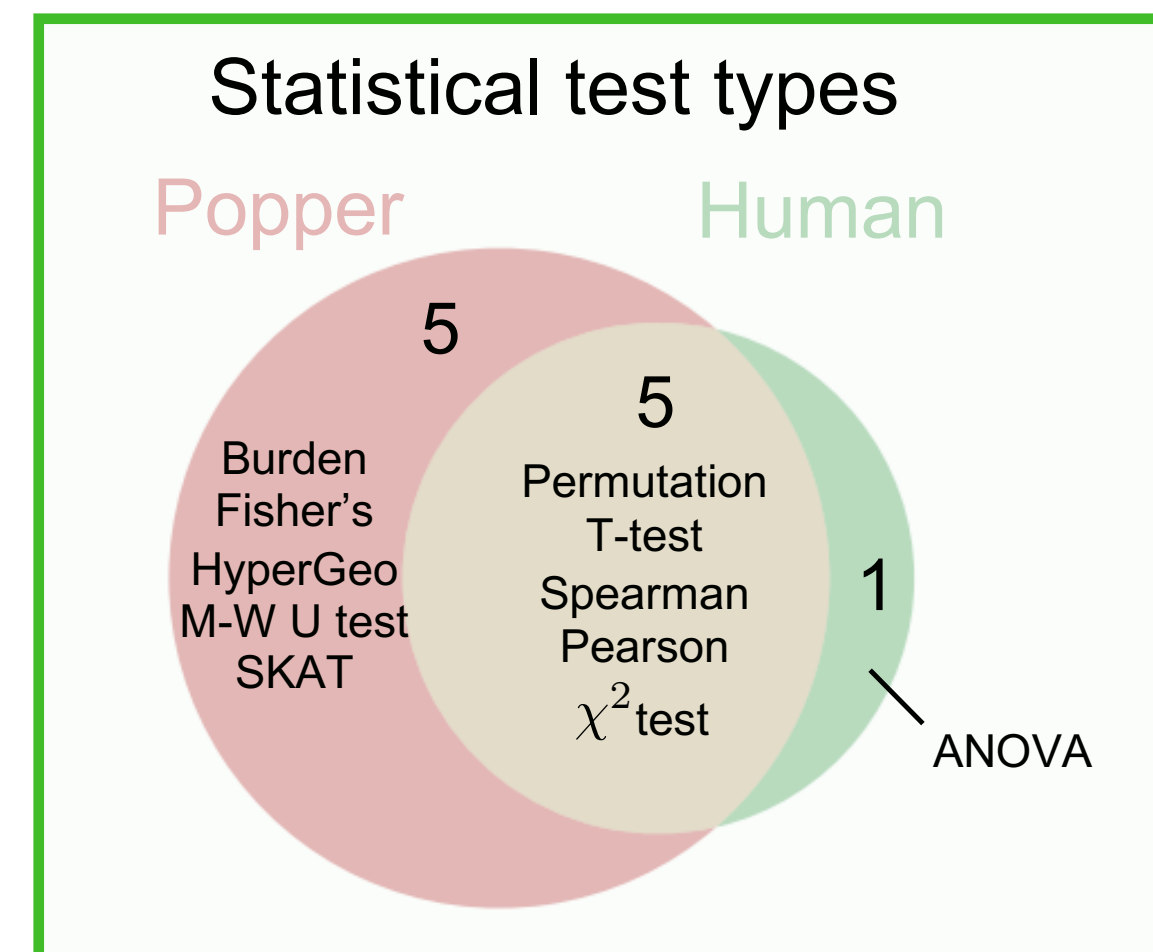


Overlap in the type of tests conducted, POPPER slightly more diverse

POPPER attains similar accuracy while greatly reducing the time



POPPER conducts more tests (more labors)



POPPER conducts more diverse tests than human scientists

POPPER: Summary

- ▶ POPPER is an agentic framework to validate a free-form hypothesis
 - ▶ Convert the main hypothesis to various implied, testable sub-hypotheses
 - ▶ Accumulate evidence to falsify the main hypothesis
 - ▶ E-value sequential safe testing allows adaptive continuation
- ▶ Unleashing the power of LLMs/agents while maintaining statistical rigor
 - ▶ A “minimal” statistical framework to regulate scientific discovery by LLM agents
 - ▶ Safely exploit the knowledge and reasoning abilities of LLM agents

Huang, K., Jin, Y., Li, R., Li, M. Y., Candès, E., & Leskovec, J. (2025). *Automated Hypothesis Validation with Agentic Sequential Falsifications*. ICML 2025.



POPPER Public

Automated Hypothesis Testing with Agentic Sequential Falsifications

agent

hypothesis-testing

llm

ai-for-science

Python • 23 • 216 • 3 • 1 • Updated on May 14



```
from popper import Popper

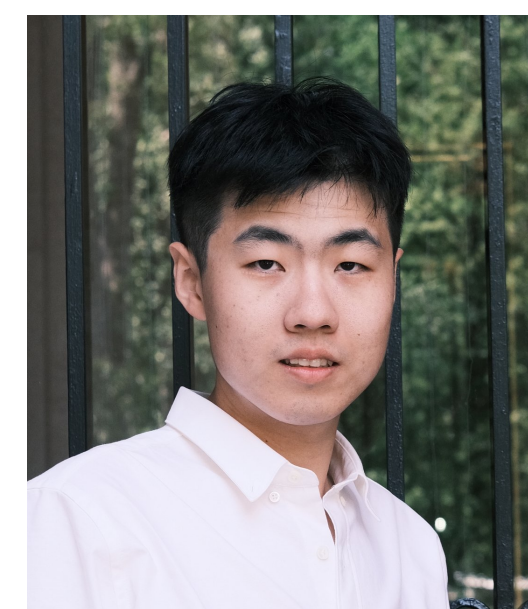
agent = Popper(llm="claude-3-5-sonnet-20240620")
agent.register_data(data_path='path/to/data', loader_type='bio')
agent.configure(alpha=0.1)
results = agent.validate(hypothesis="Your hypothesis here")
```


HypoOrchestra: Generating hypotheses with supporting statistical evidence

(Ongoing joint work with Fred Zhang and Kexin Huang)

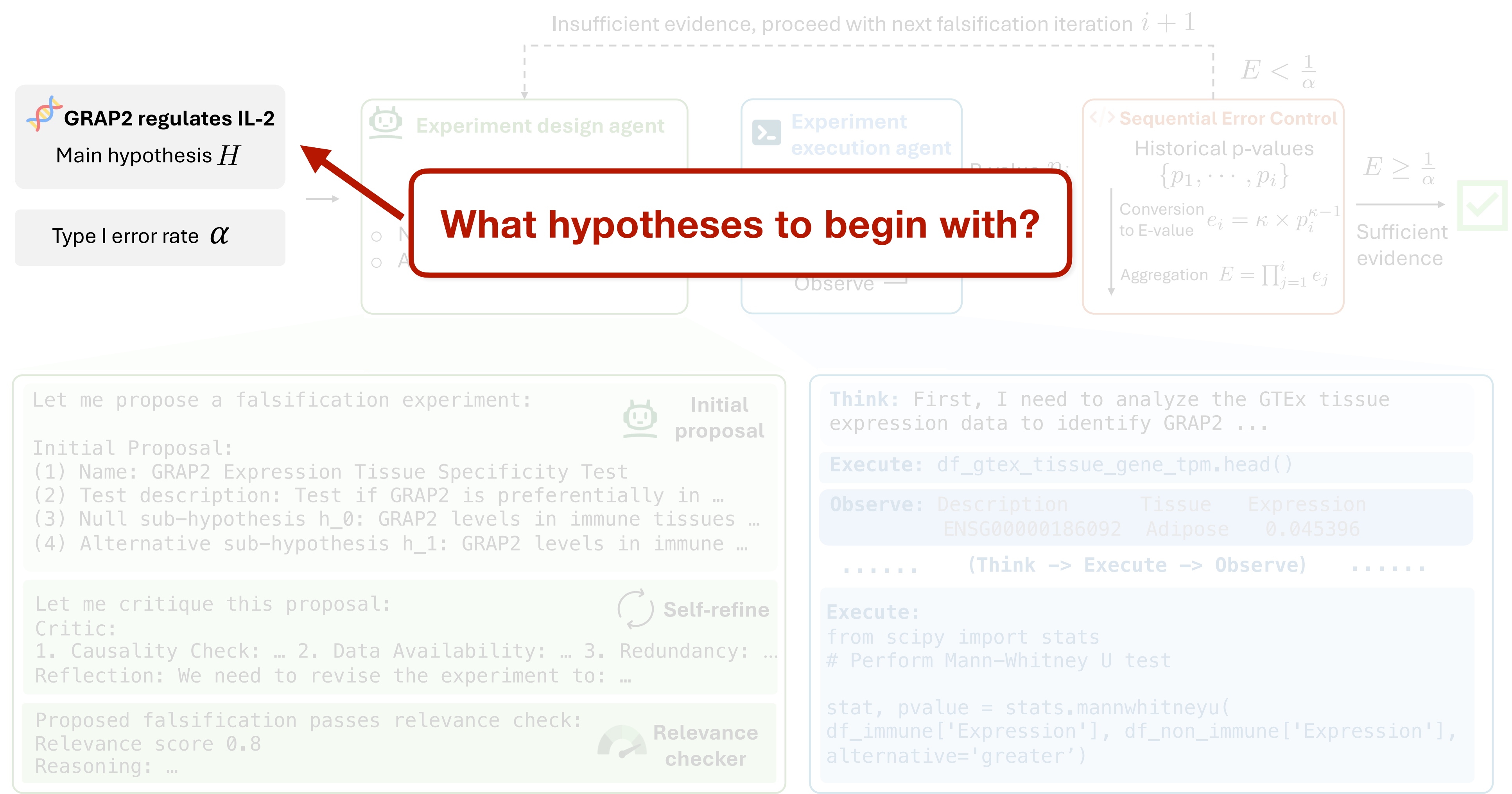


Fred Zhang
UPenn Wharton

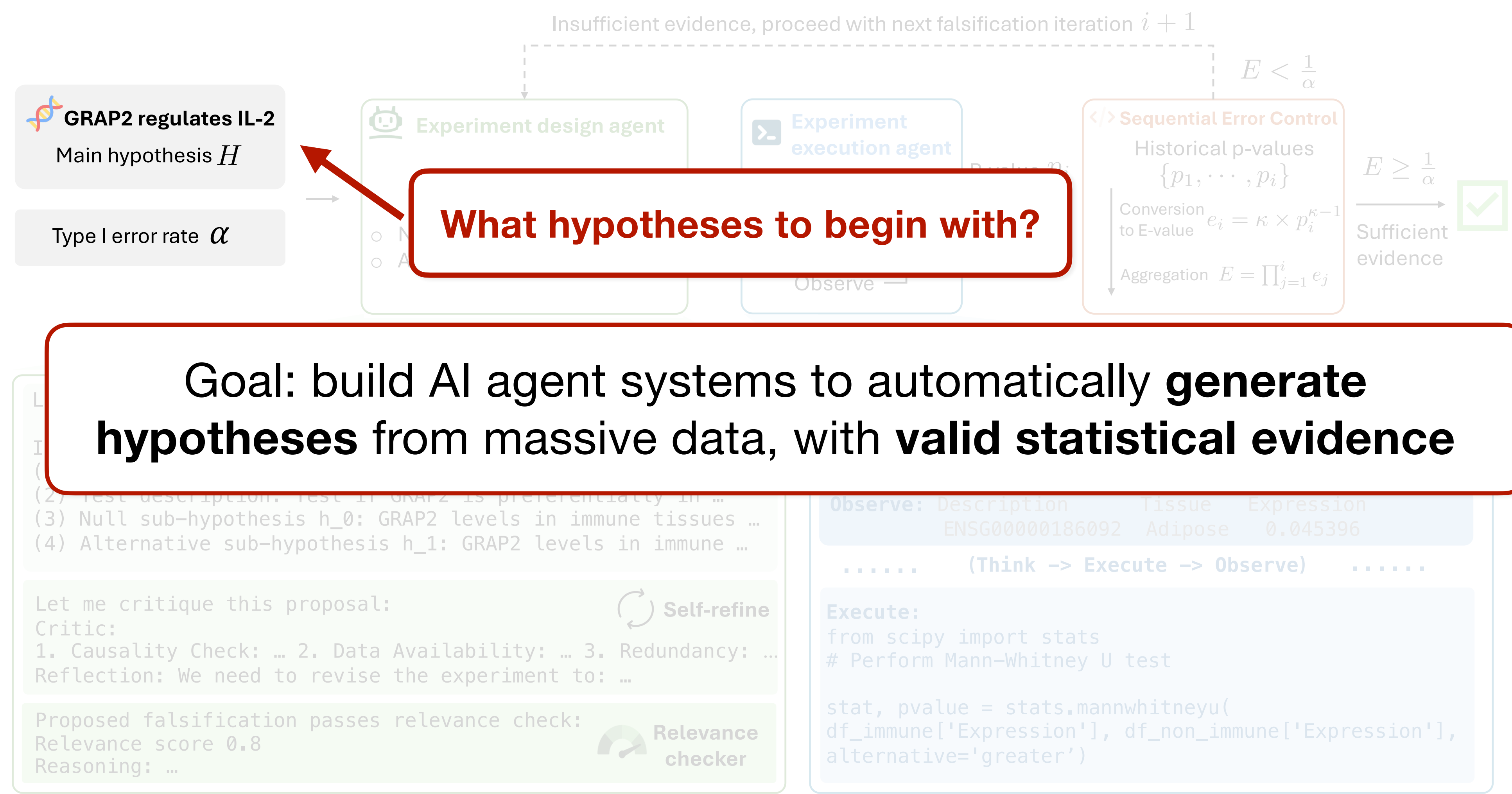


Kexin Huang
Stanford CS → Phylo

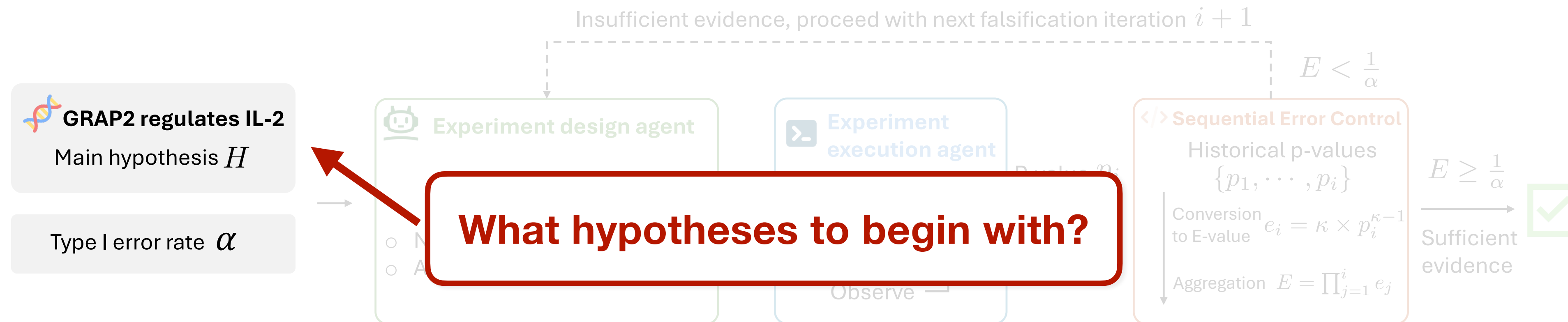
Generating evidence-supported hypotheses with AI agents



Generating evidence-supported hypotheses with AI agents



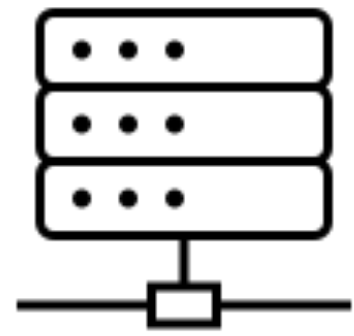
Generating evidence-supported hypotheses with AI agents



Goal: build AI agent systems to automatically **generate hypotheses** from massive data, with **valid statistical evidence**

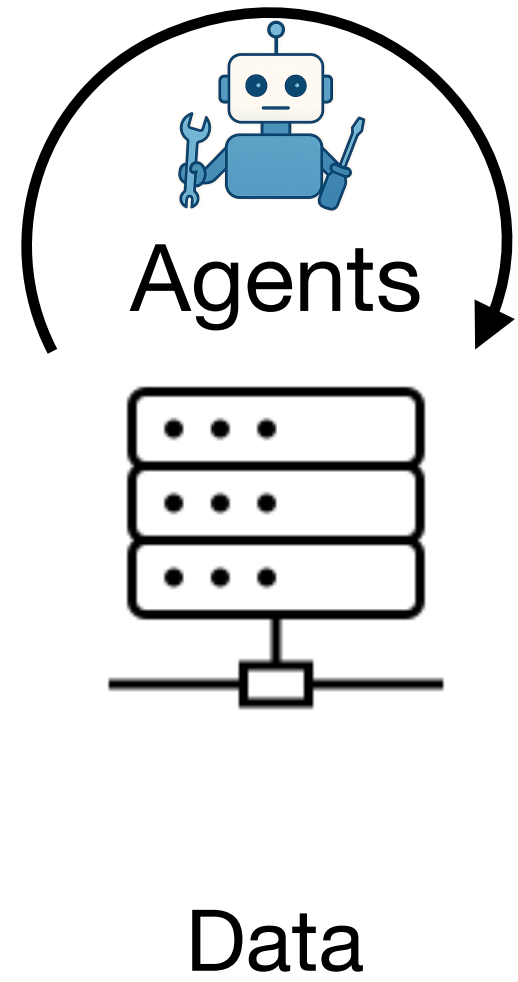


Challenge: selection bias from data reuse

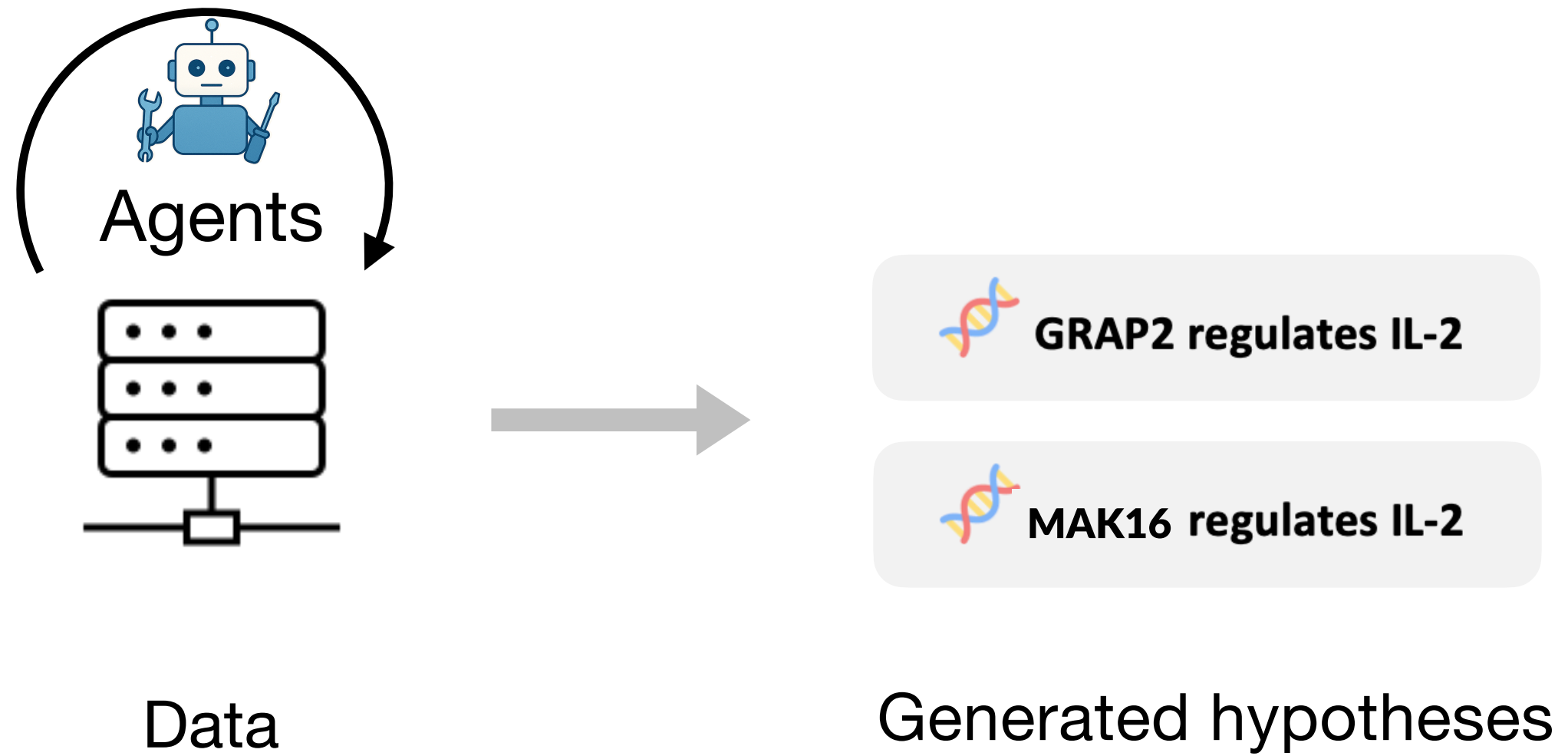


Data

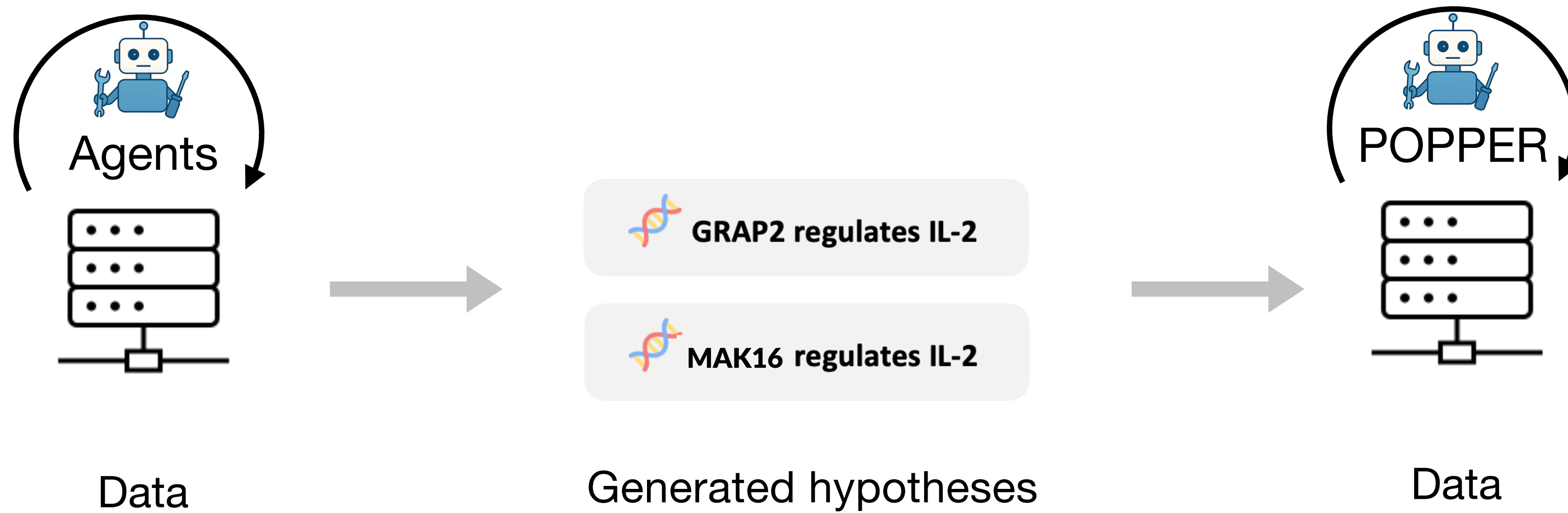
Challenge: selection bias from data reuse



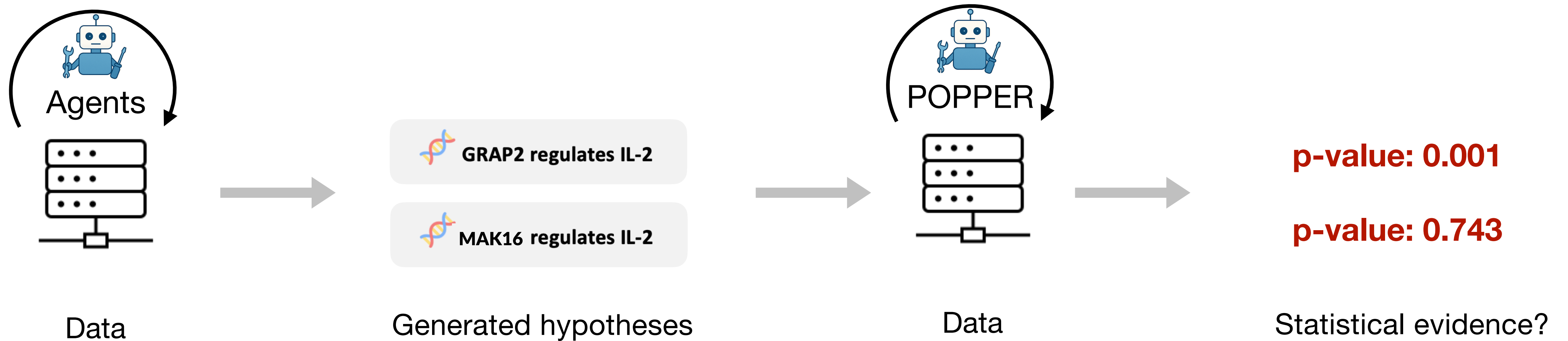
Challenge: selection bias from data reuse



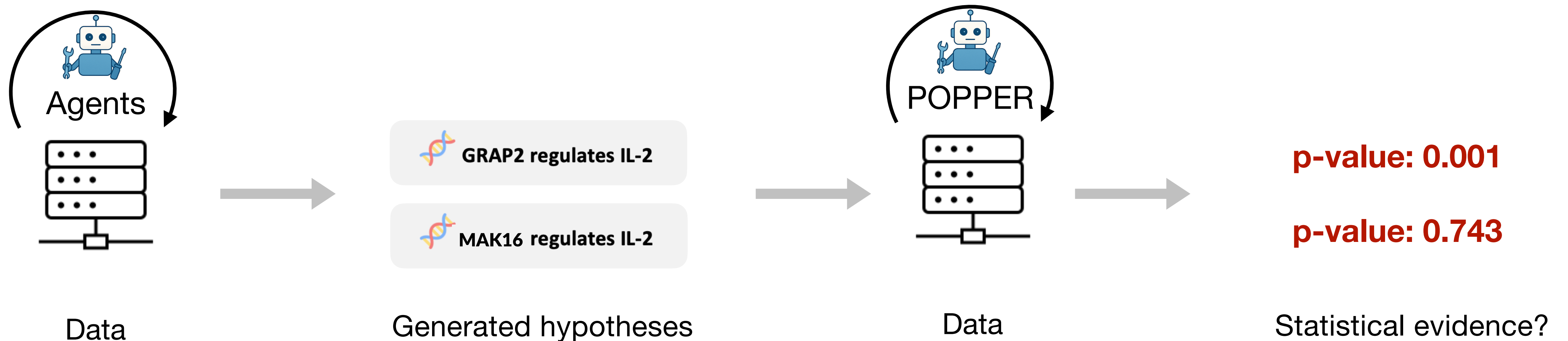
Challenge: selection bias from data reuse



Challenge: selection bias from data reuse



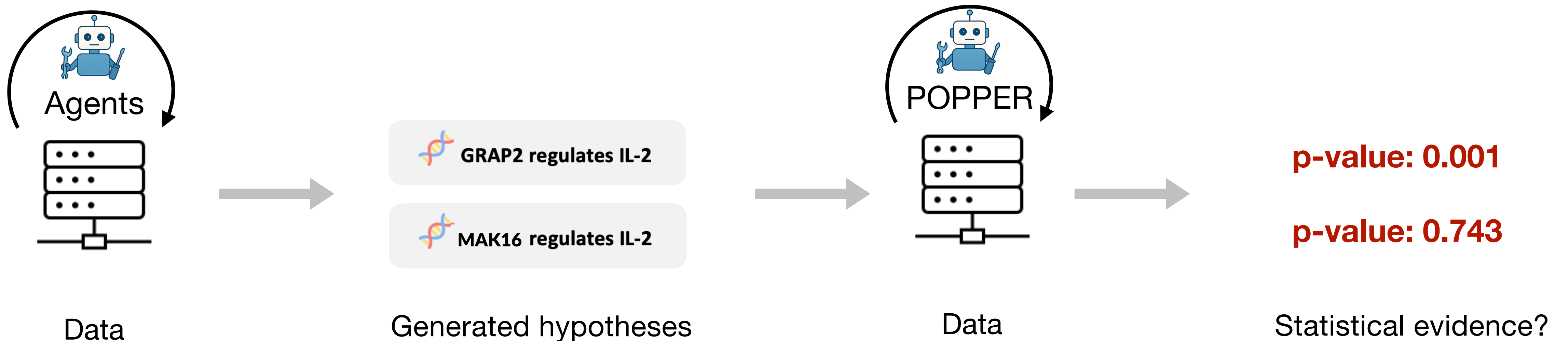
Challenge: selection bias from data reuse



Lesson from selective inference: Winner's curse!

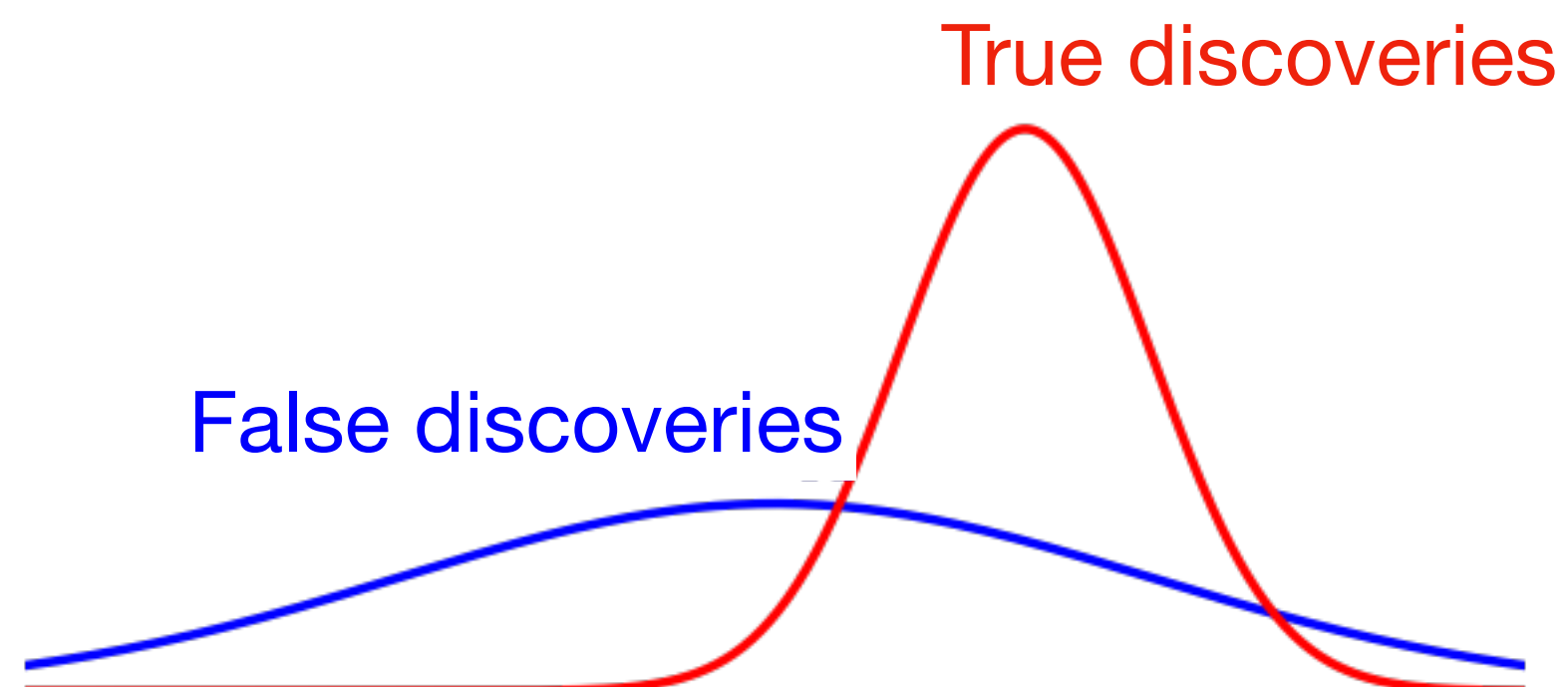
[Soric, 1989, Benjamini and Yekutieli, 2005, Berk et al., 2013]

Challenge: selection bias from data reuse

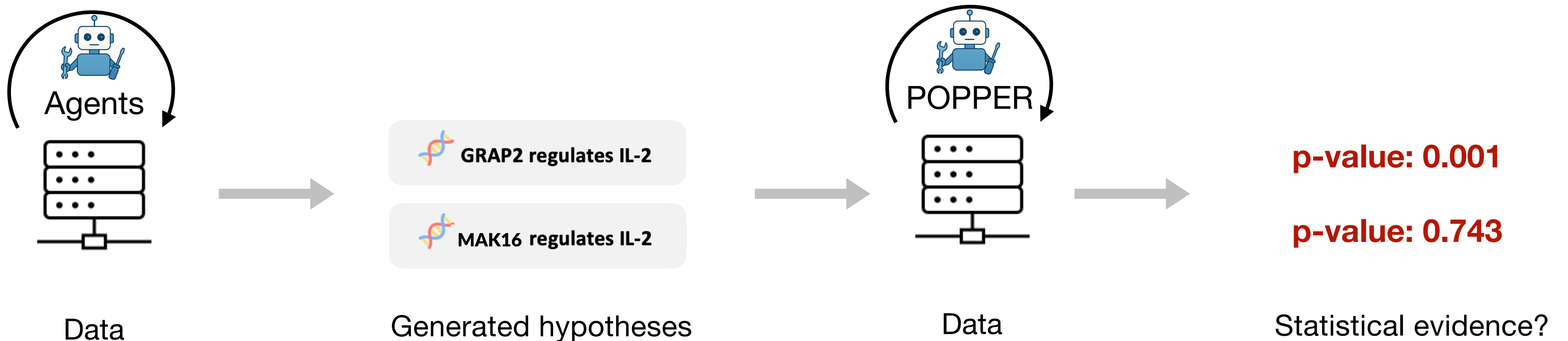


Lesson from selective inference: Winner's curse!

[Soric, 1989, Benjamini and Yekutieli, 2005, Berk et al., 2013]

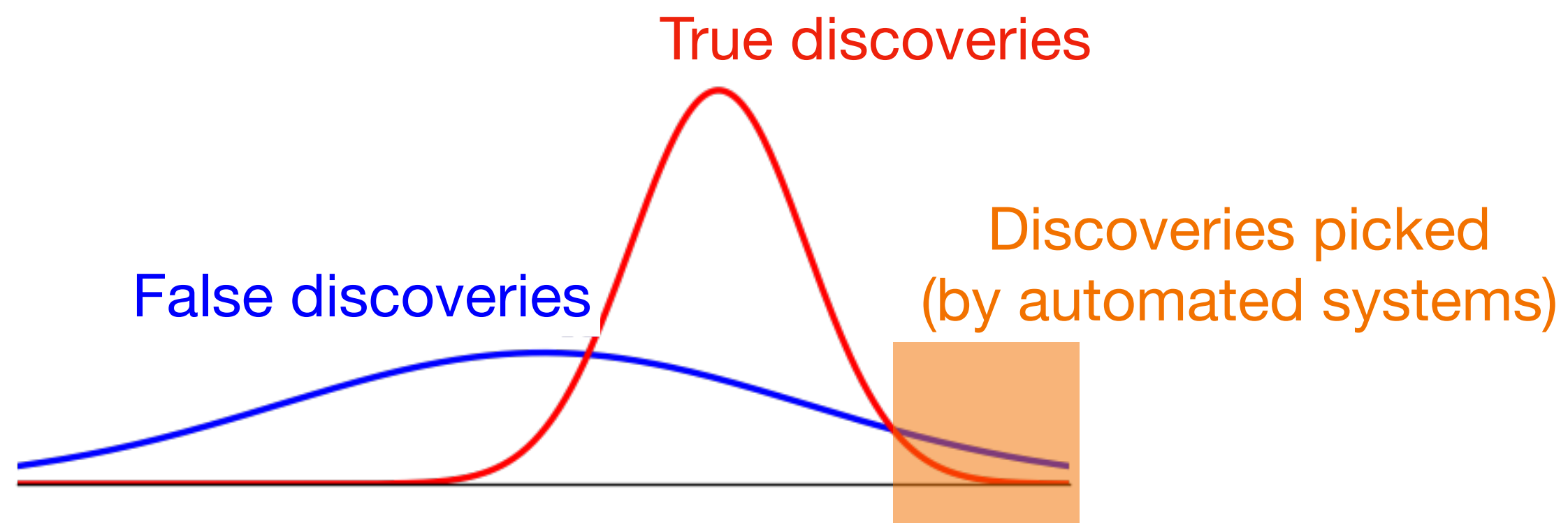


Challenge: selection bias from data reuse

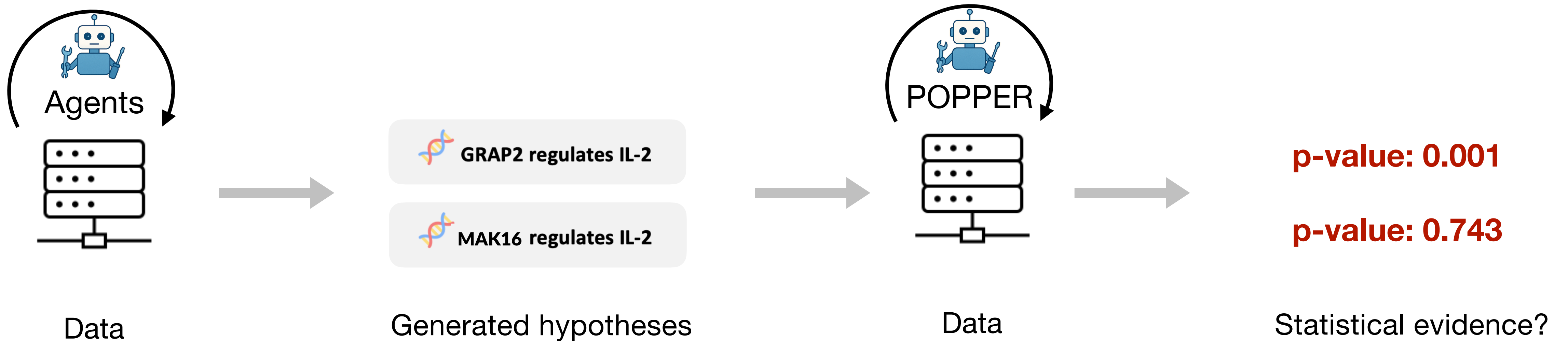


Lesson from selective inference: Winner's curse!

[Soric, 1989, Benjamini and Yekutieli, 2005, Berk et al., 2013]

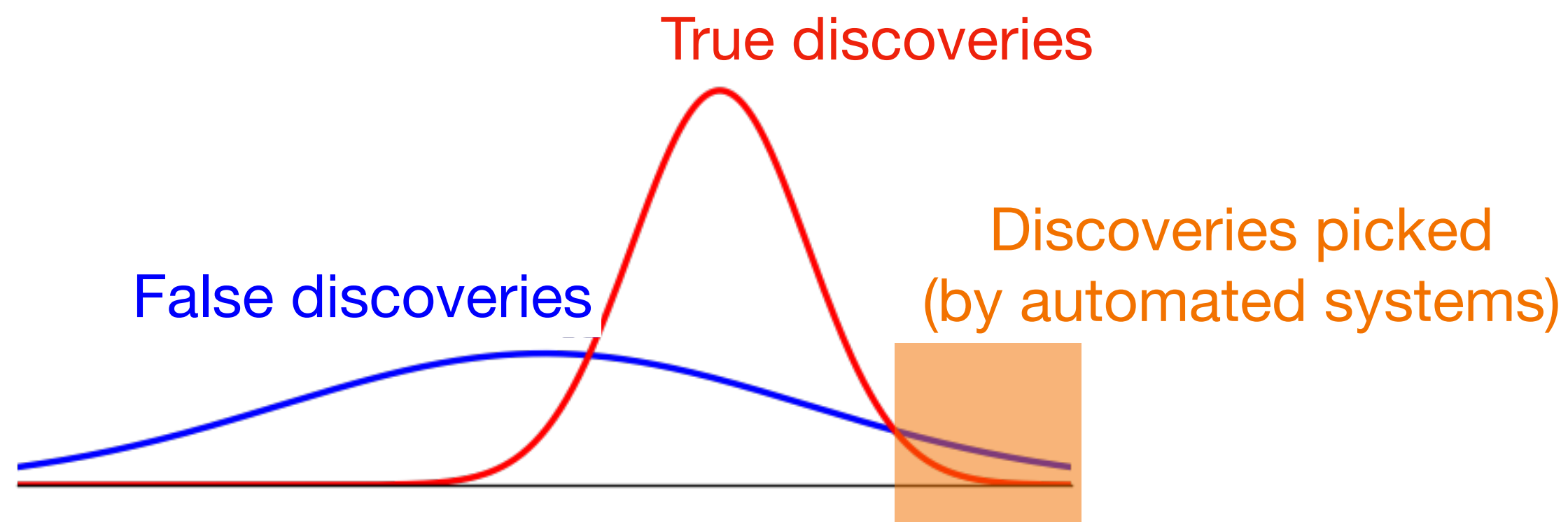


Challenge: selection bias from data reuse



Lesson from selective inference: Winner's curse!

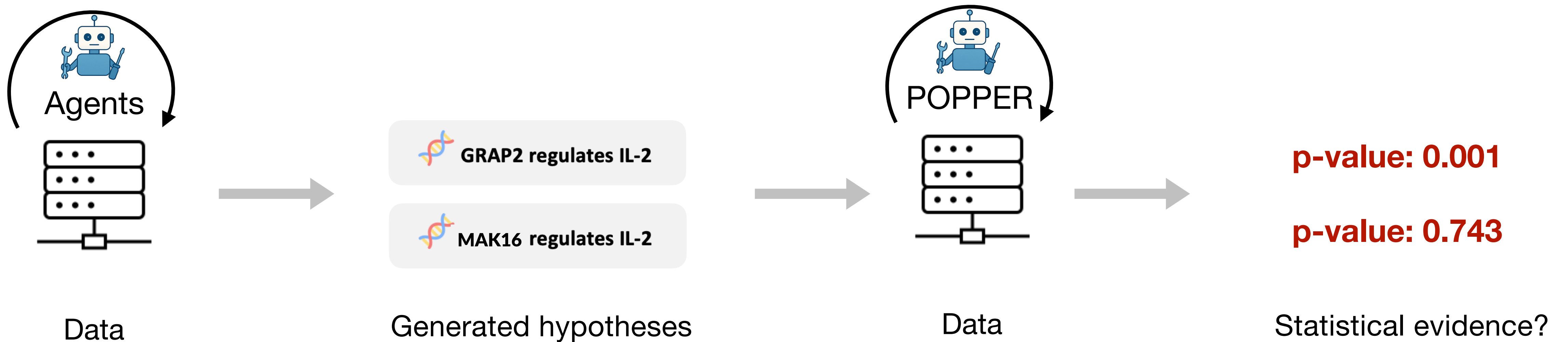
[Soric, 1989, Benjamini and Yekutieli, 2005, Berk et al., 2013]



Generation:

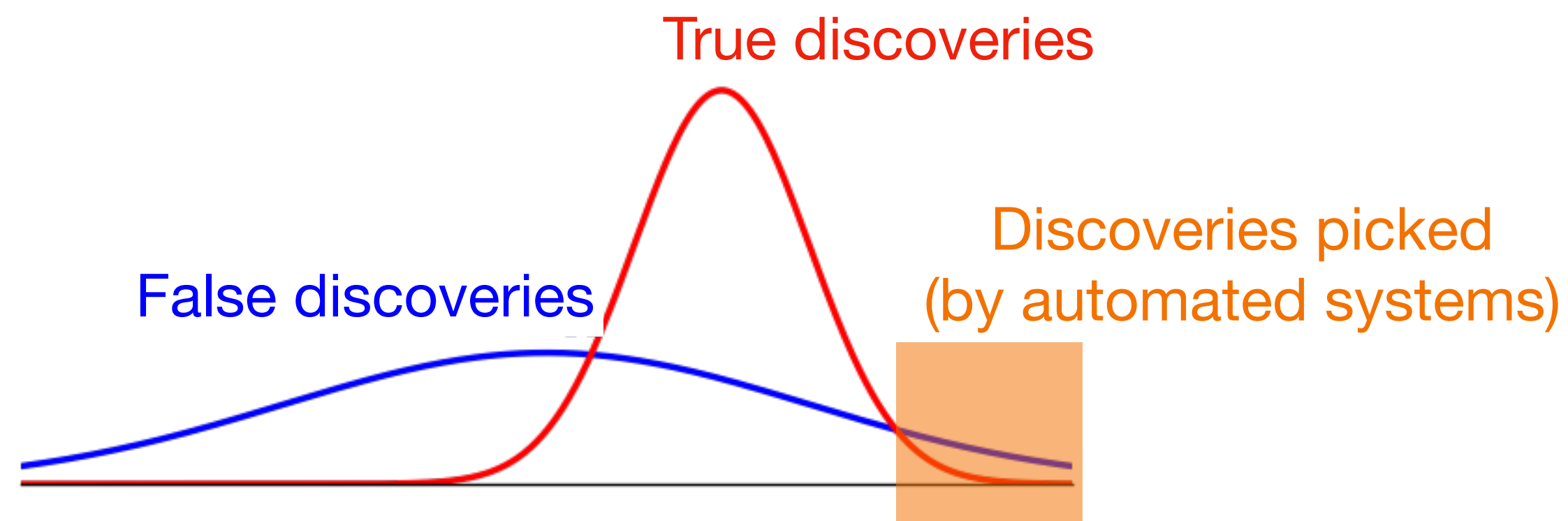
- ~ compute 100 test statistics
- ~ pick smallest p-values
- ~ generated hypotheses

Challenge: selection bias from data reuse



Lesson from selective inference: Winner's curse!

[Soric, 1989, Benjamini and Yekutieli, 2005, Berk et al., 2013]



Generation:

- ~ compute 100 test statistics
- ~ pick smallest p-values
- ~ generated hypotheses

Verification:

- ~ compute the same test statistics
- ~ **All Strong Evidence? Invalid!**

Orchestrating generation and validation via information control

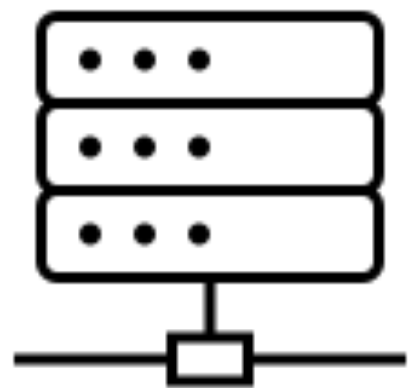
High-level idea:

- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge

Orchestrating generation and validation via information control

High-level idea:

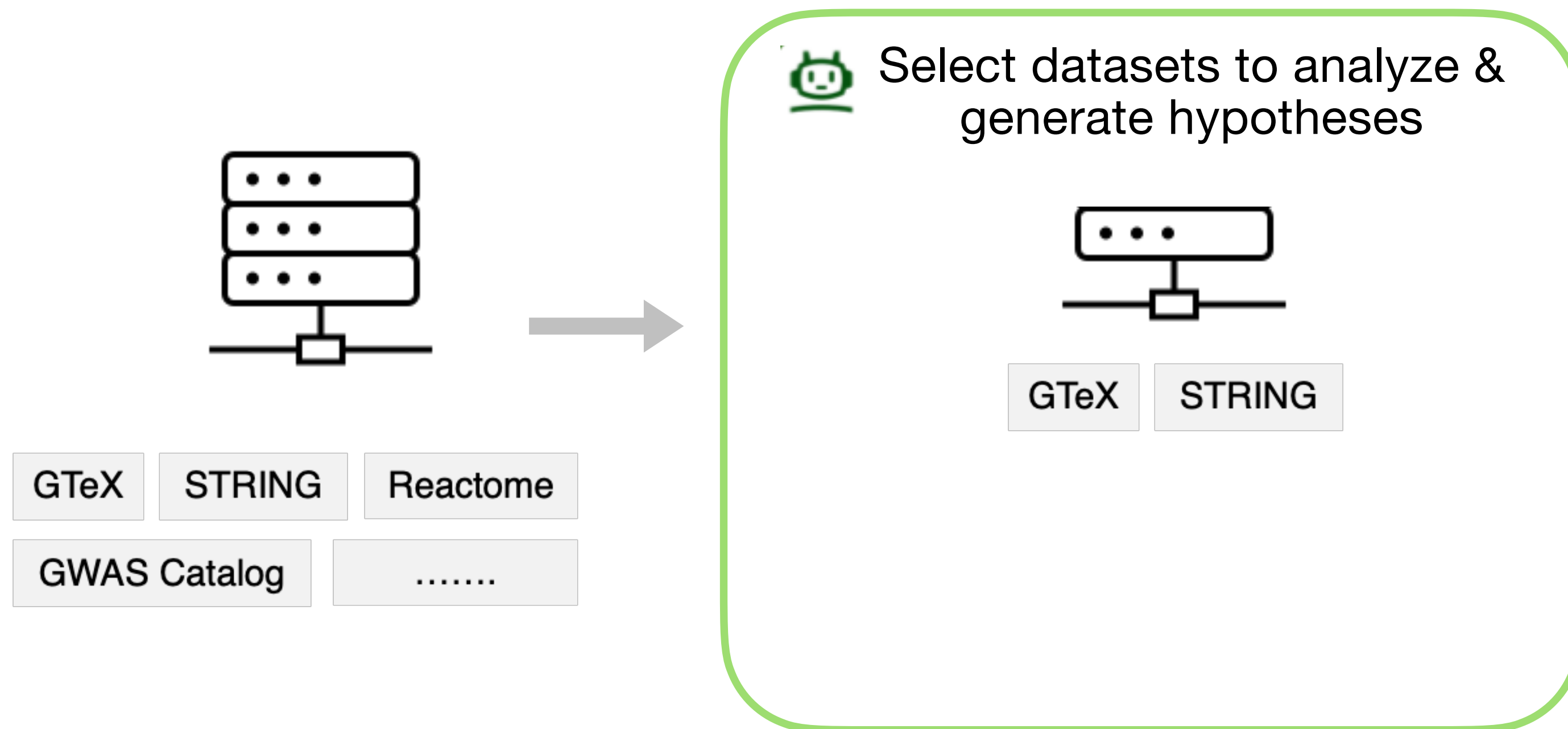
- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge



Orchestrating generation and validation via information control

High-level idea:

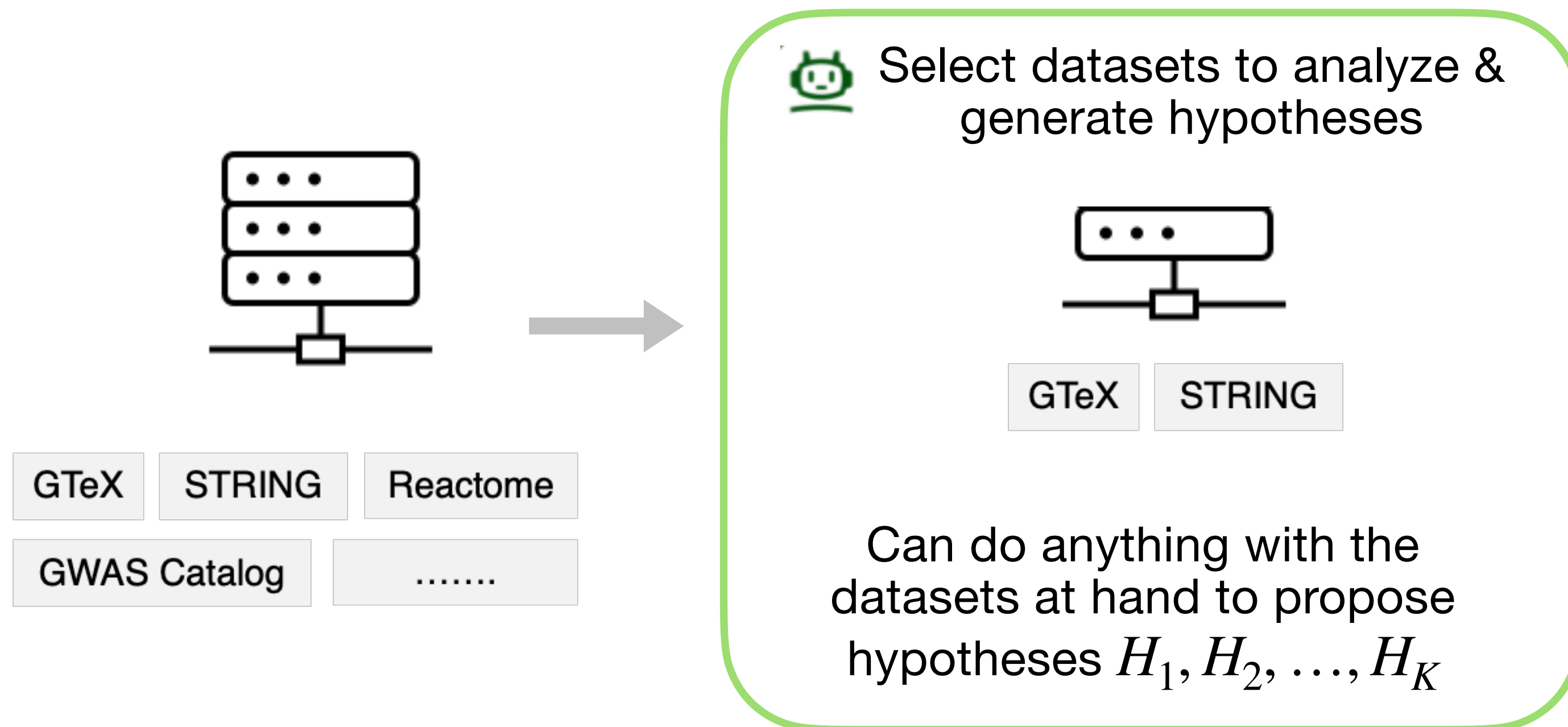
- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge



Orchestrating generation and validation via information control

High-level idea:

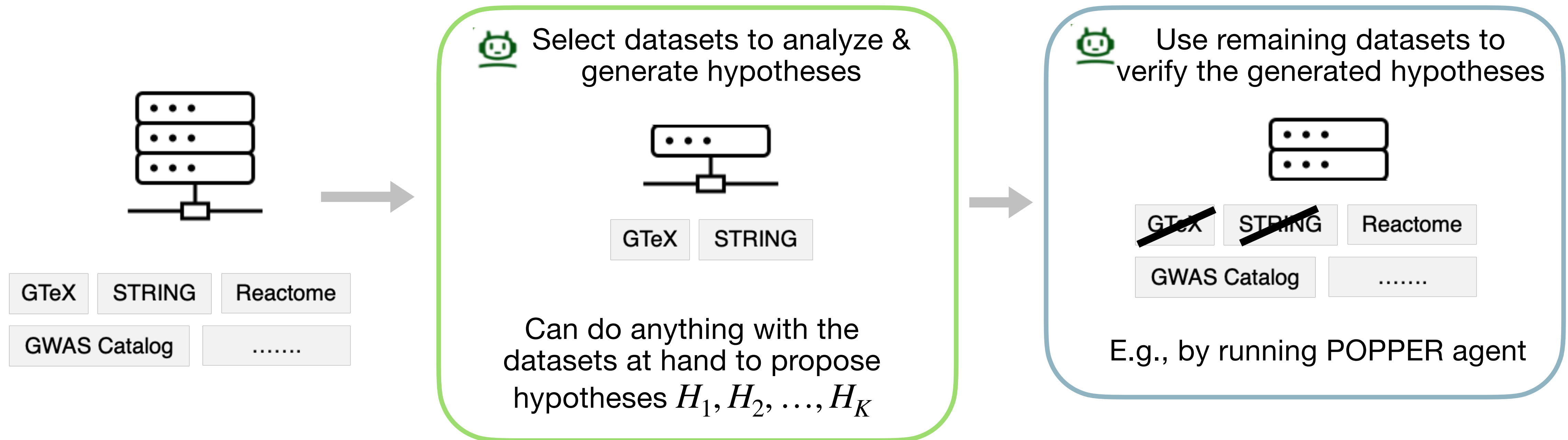
- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge



Orchestrating generation and validation via information control

High-level idea:

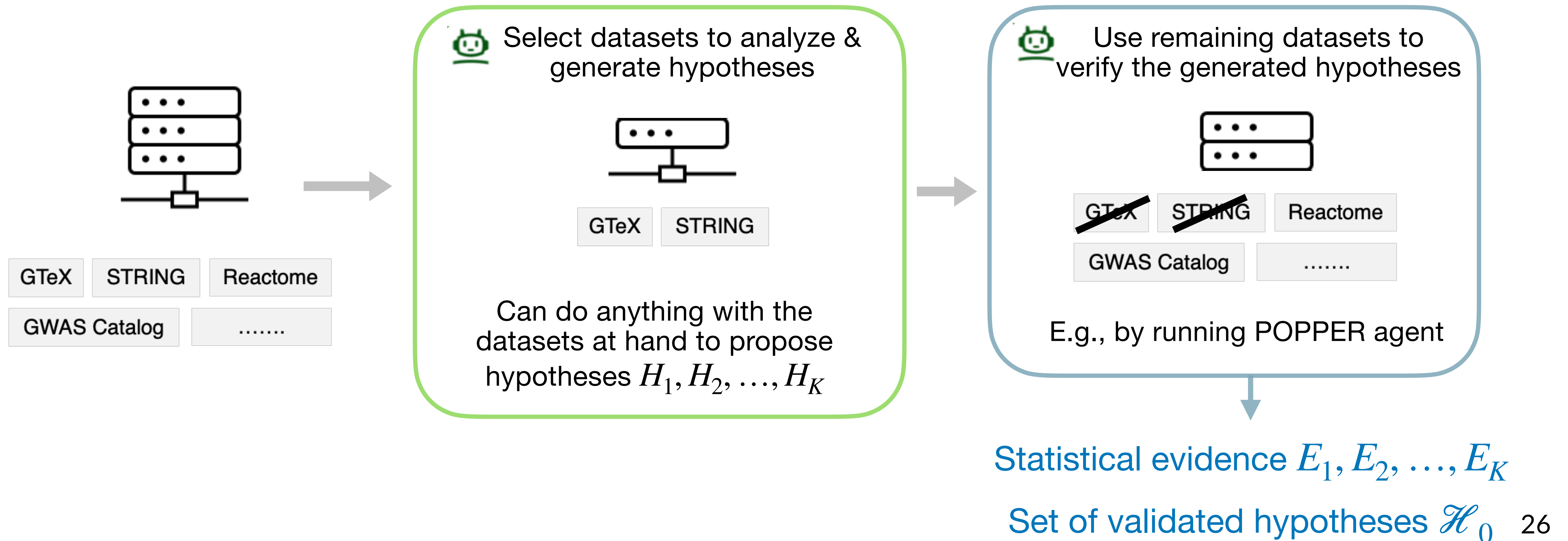
- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge



Orchestrating generation and validation via information control

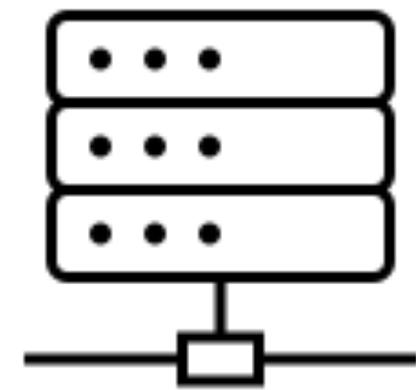
High-level idea:

- ▶ Keep $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{val}}$ separate & independent $\leadsto$ statistical validity
- ▶ While allowing agents to adaptively explore $\leadsto$ creativity, diversity, knowledge



Successful examples

Raw database



GTEx

STRING

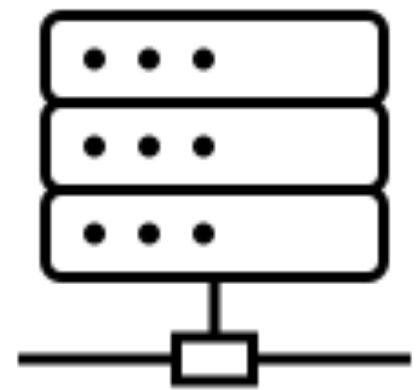
Reactome

GWAS Catalog

.....

Successful examples

Raw database



GTEx

STRING

Reactome

GWAS Catalog

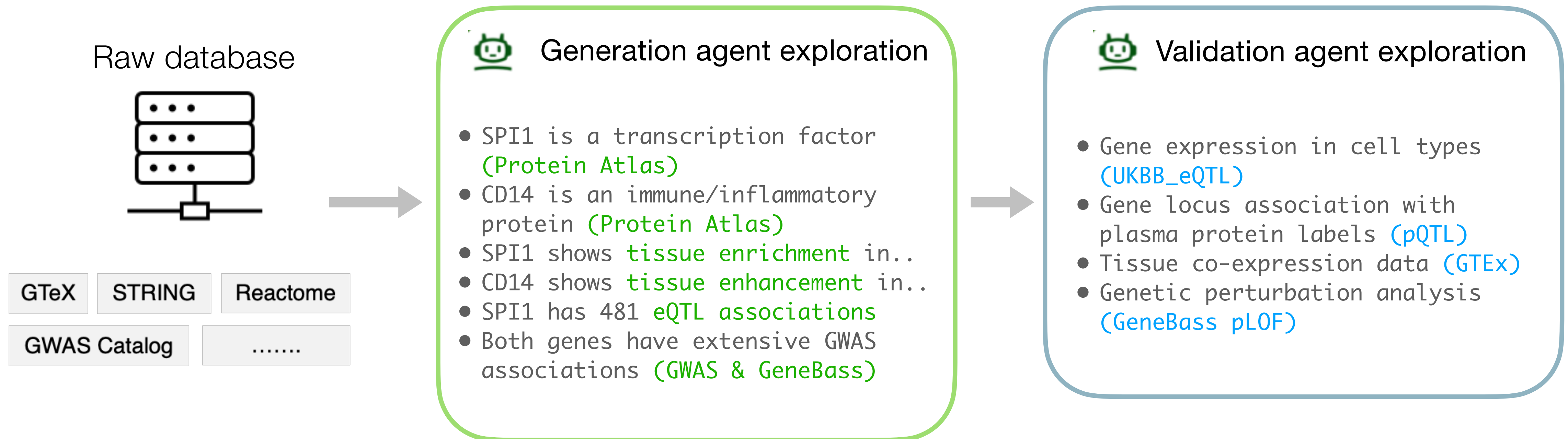
.....



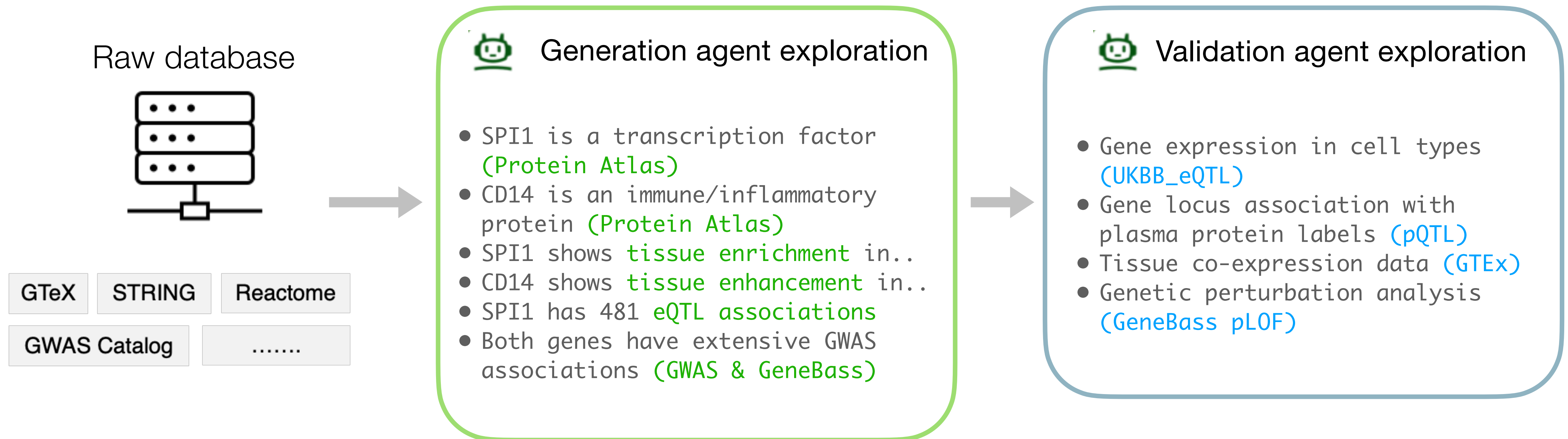
Generation agent exploration

- SPI1 is a transcription factor (Protein Atlas)
- CD14 is an immune/inflammatory protein (Protein Atlas)
- SPI1 shows tissue enrichment in..
- CD14 shows tissue enhancement in..
- SPI1 has 481 eQTL associations
- Both genes have extensive GWAS associations (GWAS & GeneBass)

Successful examples



Successful examples



Key challenges:

- ▶ Smartly choose what data to use for generation while ensuring validation is feasible
- ▶ Diverse hypotheses generation operations (data analysis, data combination, reasoning)
- ▶ Creating independent data blocks from dependent summary datasets

Summary

- ▶ Unleashing the power of agents for scientific discovery while maintaining statistical rigor
 - ▶ A “minimal” statistical framework for how LLM agents interact with “environment”
 - ▶ Safely exploit the knowledge and reasoning abilities of LLM agents
- ▶ **Hypothesis validation:** sequential search for supporting/refutation evidence
- ▶ **Hypothesis generation with evidence:** careful information control with adaptivity

Summary

- ▶ Unleashing the power of agents for scientific discovery while maintaining statistical rigor
 - ▶ A “minimal” statistical framework for how LLM agents interact with “environment”
 - ▶ Safely exploit the knowledge and reasoning abilities of LLM agents
- ▶ **Hypothesis validation:** sequential search for supporting/refutation evidence
- ▶ **Hypothesis generation with evidence:** careful information control with adaptivity

Turning offline data into a rigorous environment for scientific AI agents

Summary

- ▶ Unleashing the power of agents for scientific discovery while maintaining statistical rigor
 - ▶ A “minimal” statistical framework for how LLM agents interact with “environment”
 - ▶ Safely exploit the knowledge and reasoning abilities of LLM agents
- ▶ **Hypothesis validation:** sequential search for supporting/refutation evidence
- ▶ **Hypothesis generation with evidence:** careful information control with adaptivity

Turning offline data into a rigorous environment for scientific AI agents

Huang, K., Jin, Y., Li, R., Li, M. Y., Candès, E., & Leskovec, J. (2025). *Automated Hypothesis Validation with Agentic Sequential Falsifications*. ICML 2025.

