Natural Analysts in Adaptive Data Analysis

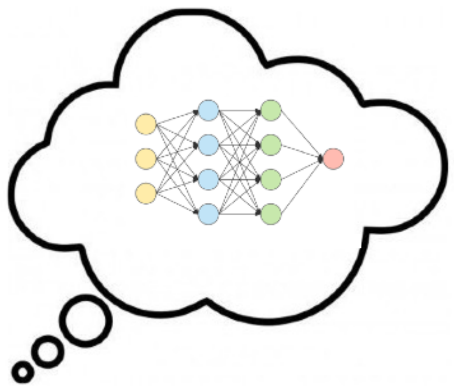
Tijana Zrnic

joint with Moritz Hardt

Berkeley
UNIVERSITY OF CALIFORNIA

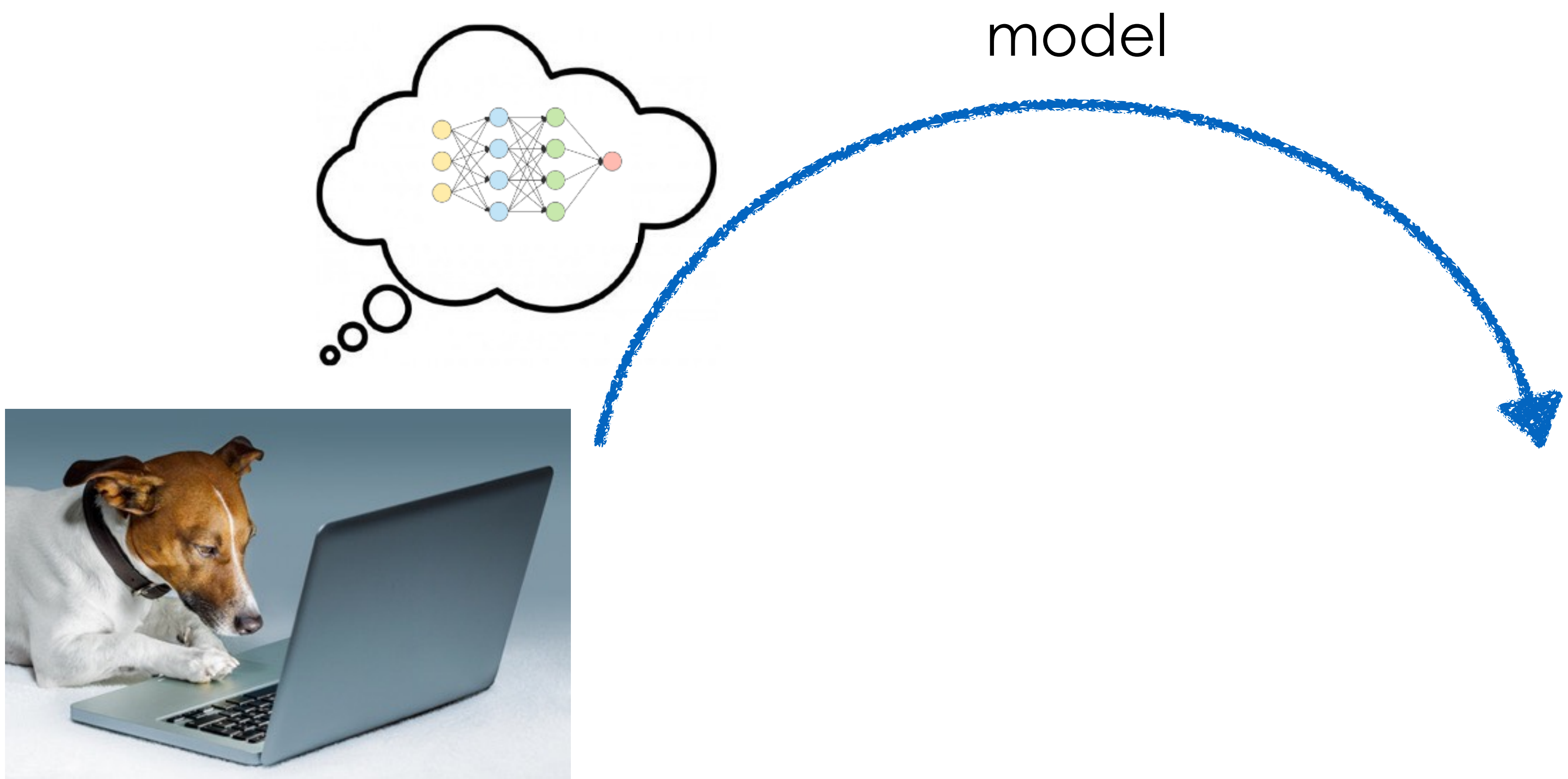
Adaptivity in Machine Learning

Adaptivity in Machine Learning



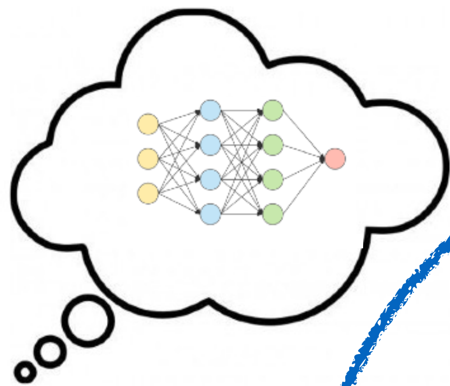
data analyst
with training data

Adaptivity in Machine Learning



data analyst
with training data

Adaptivity in Machine Learning

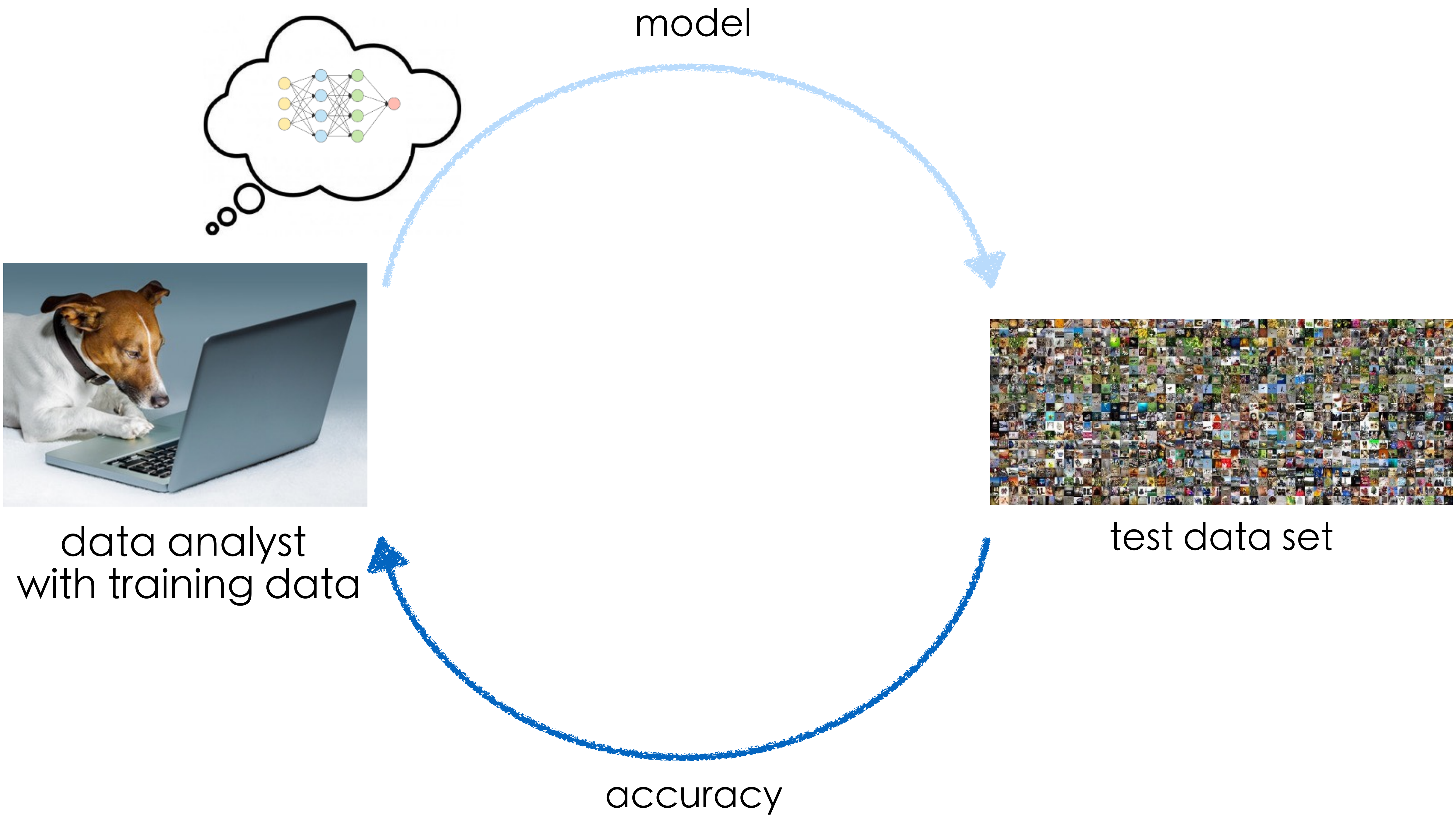


model

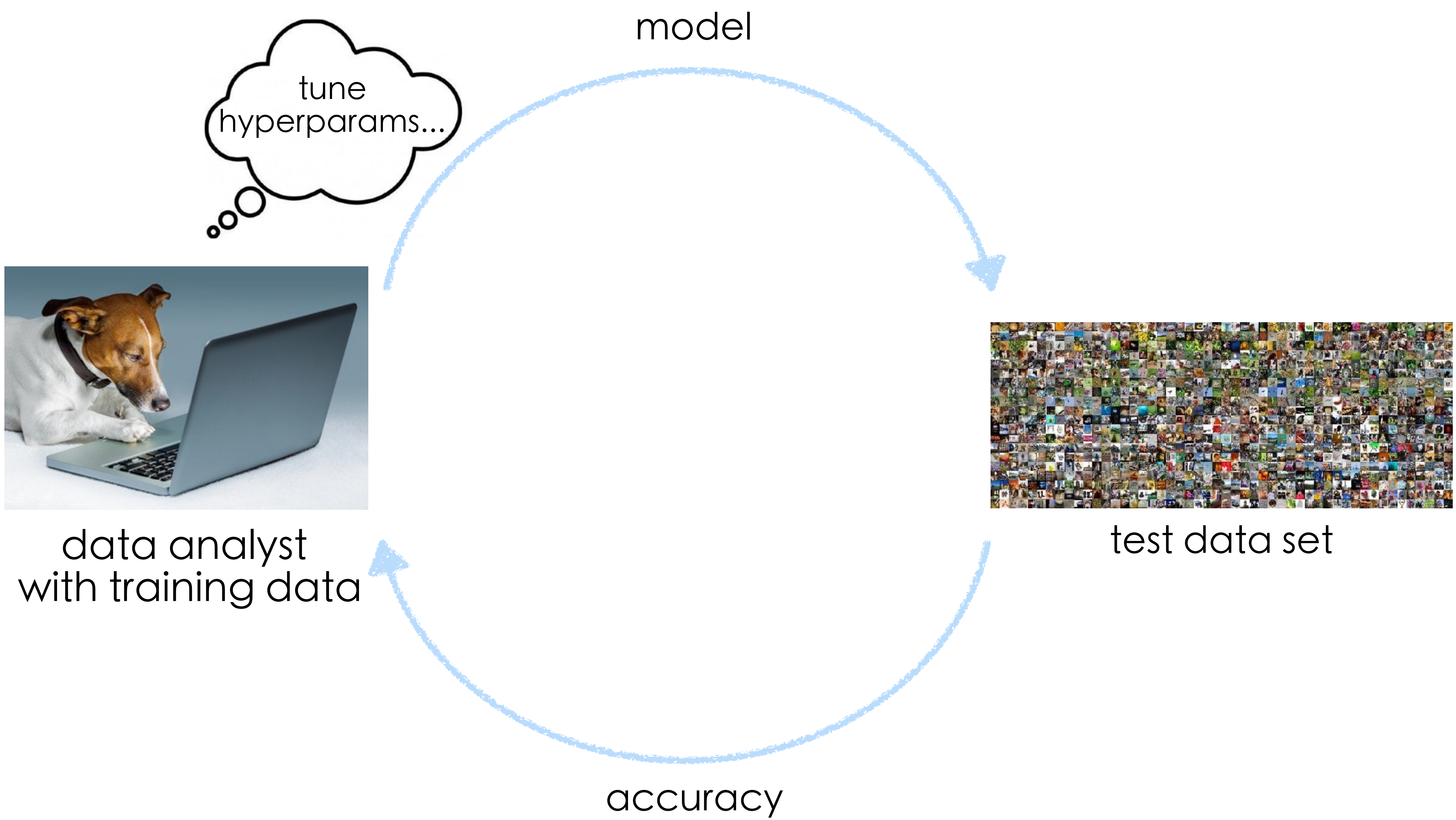
data analyst
with training data

test data set

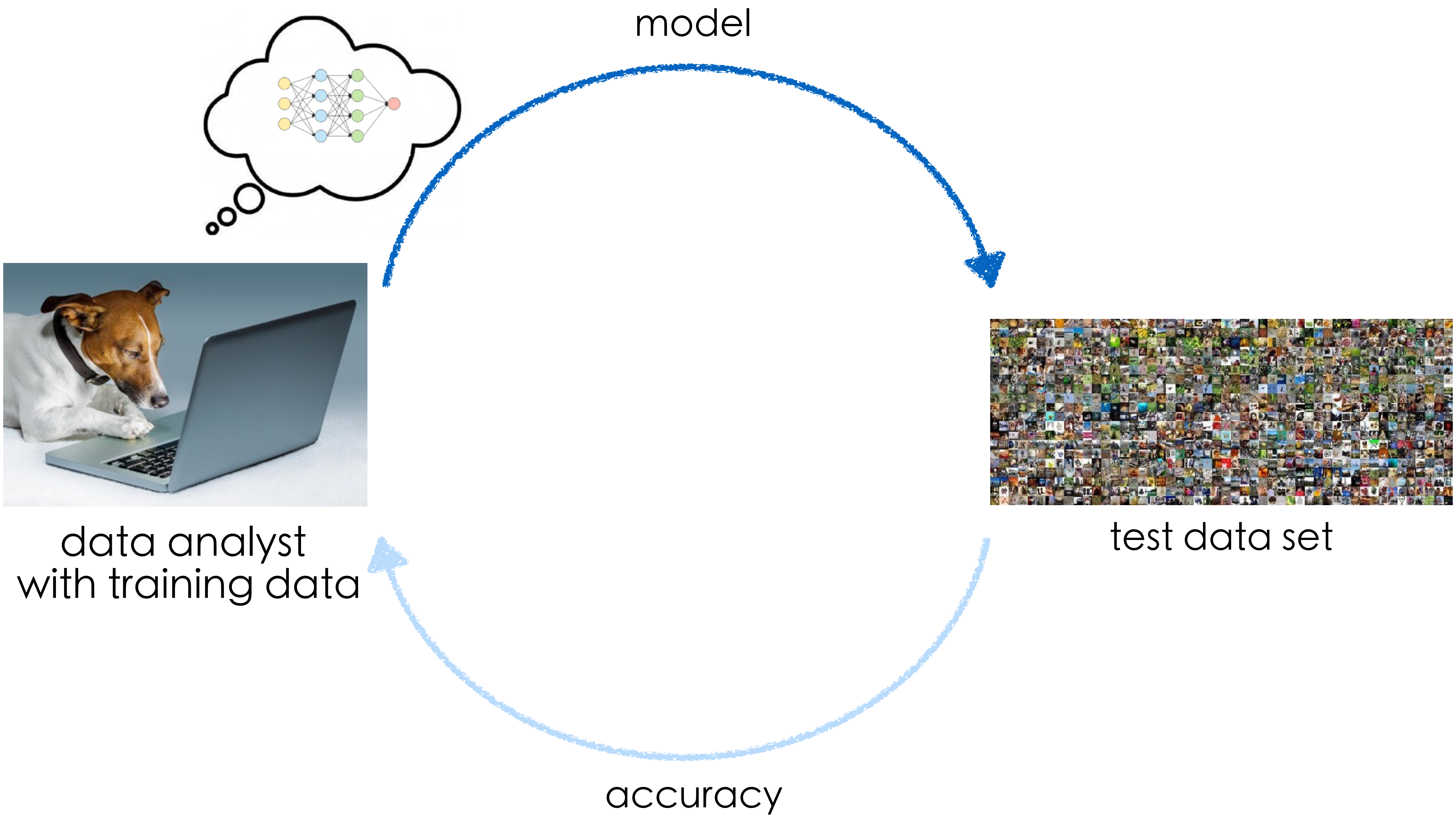
Adaptivity in Machine Learning



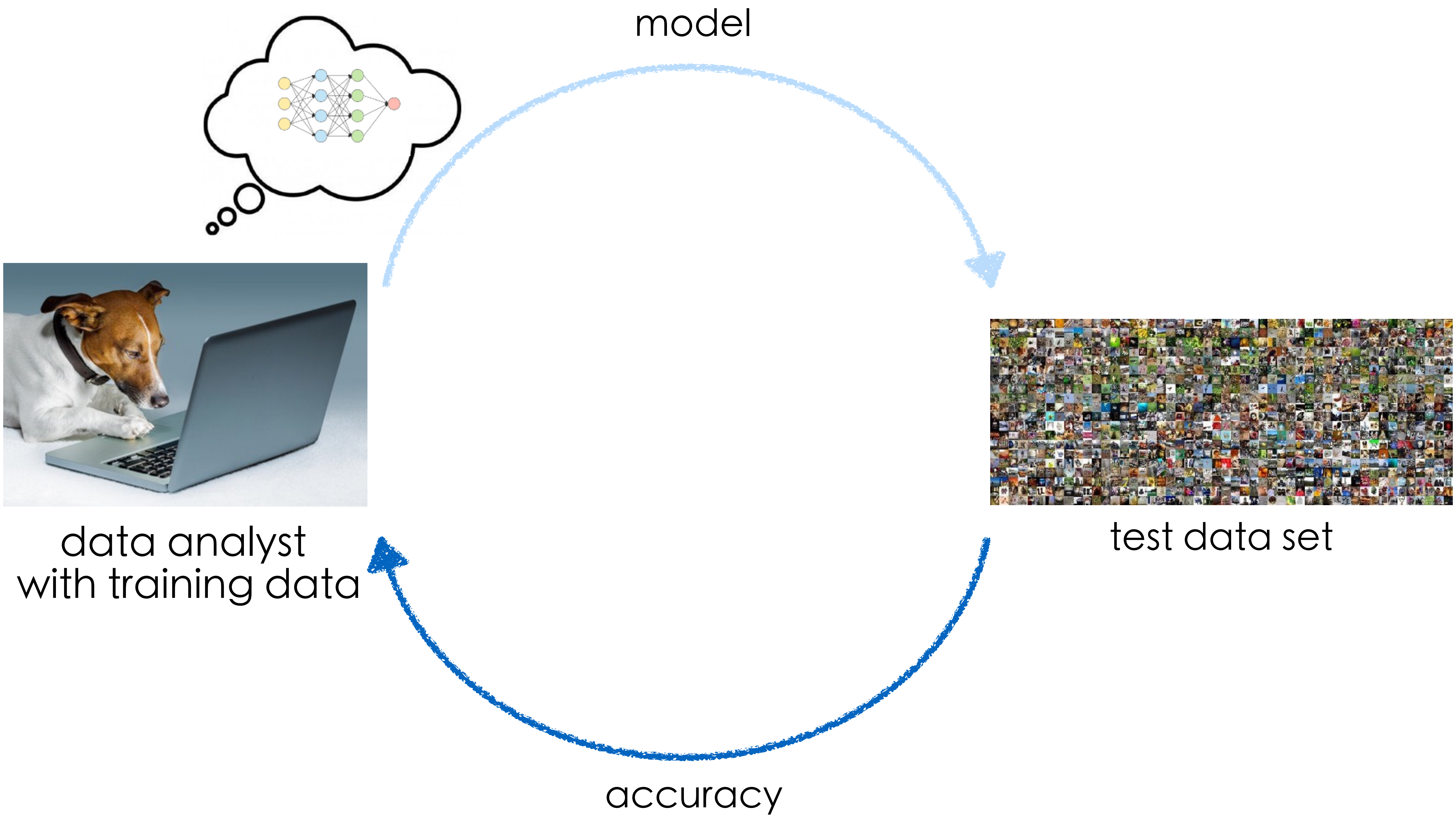
Adaptivity in Machine Learning



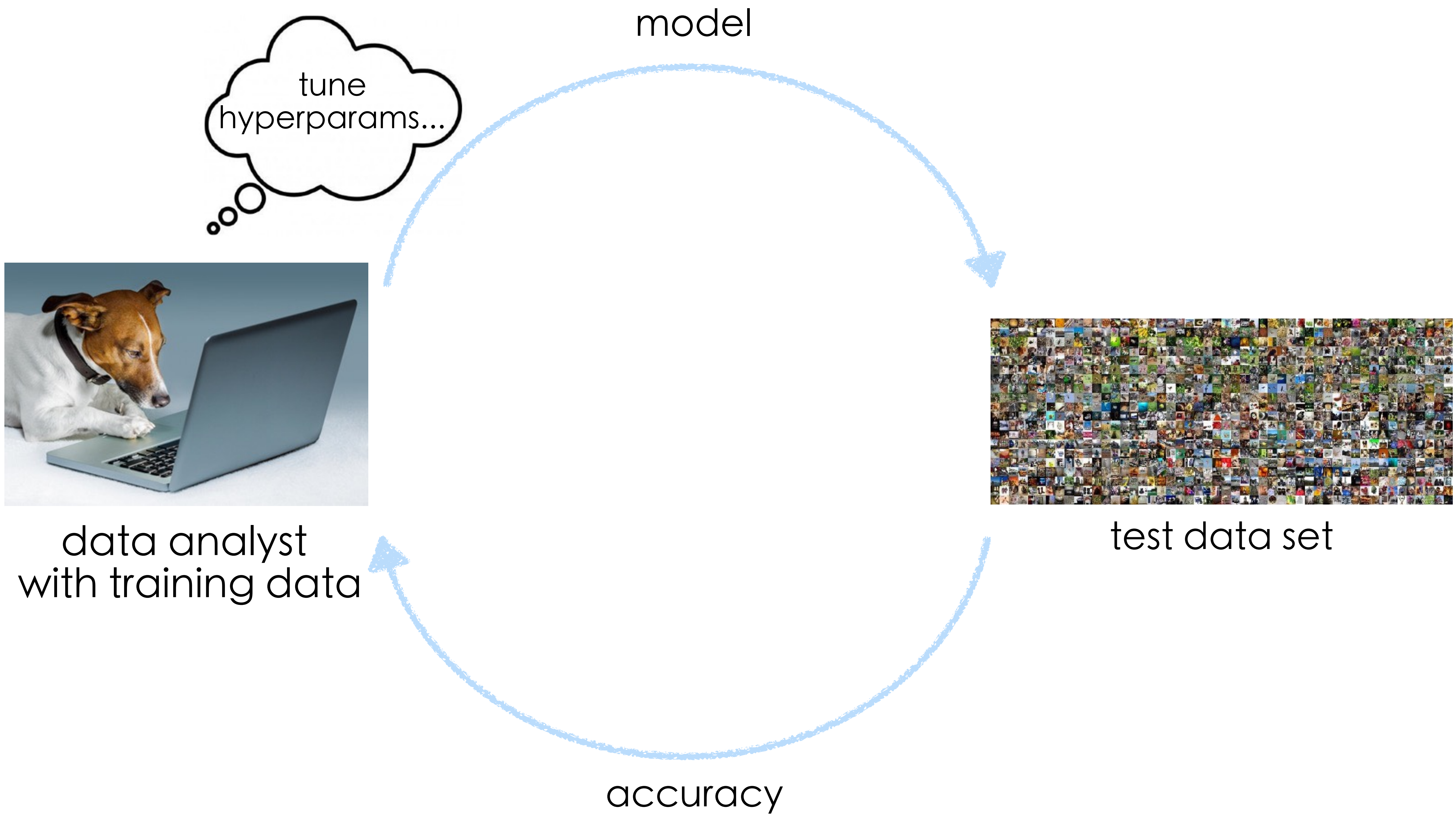
Adaptivity in Machine Learning



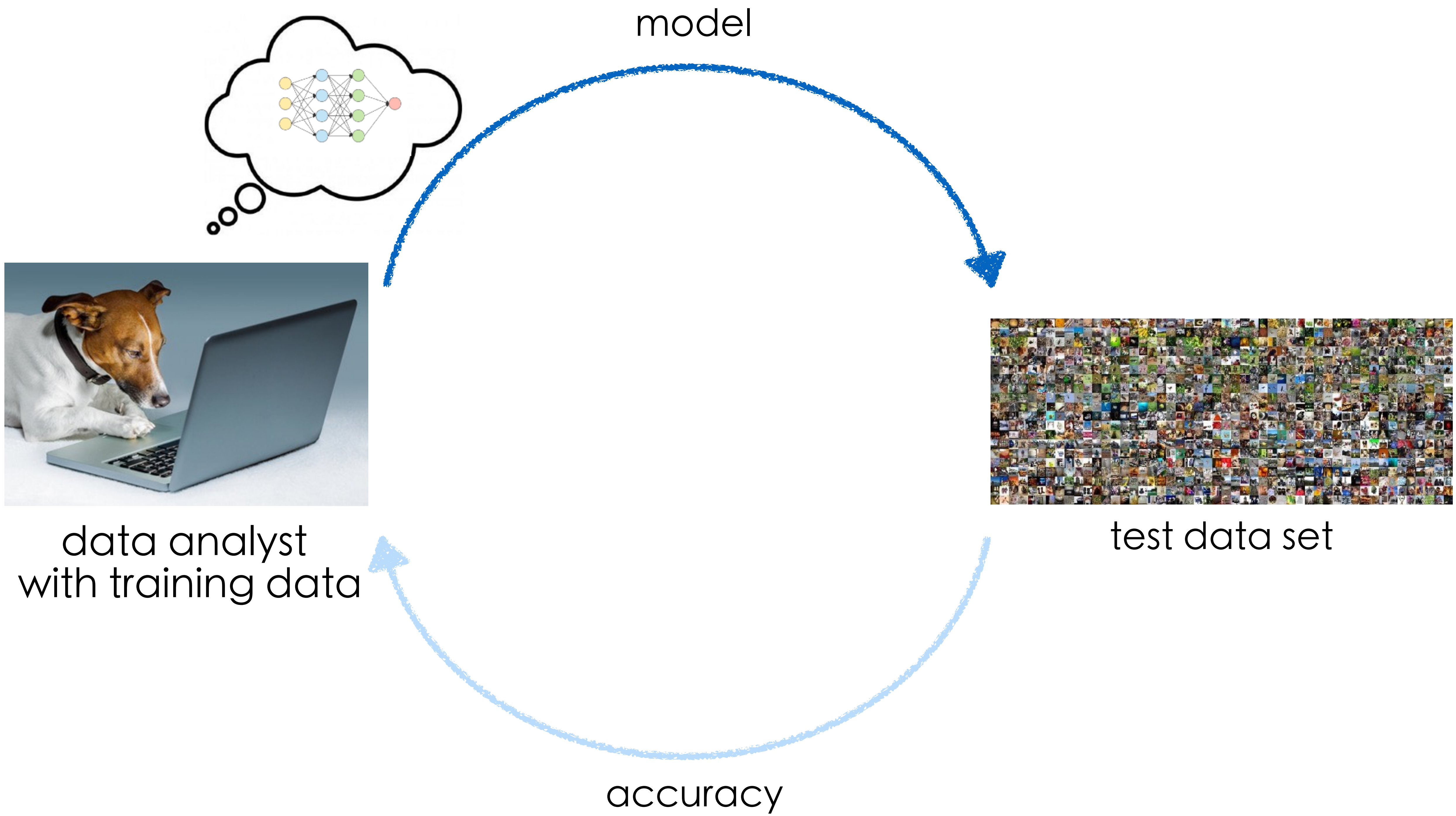
Adaptivity in Machine Learning



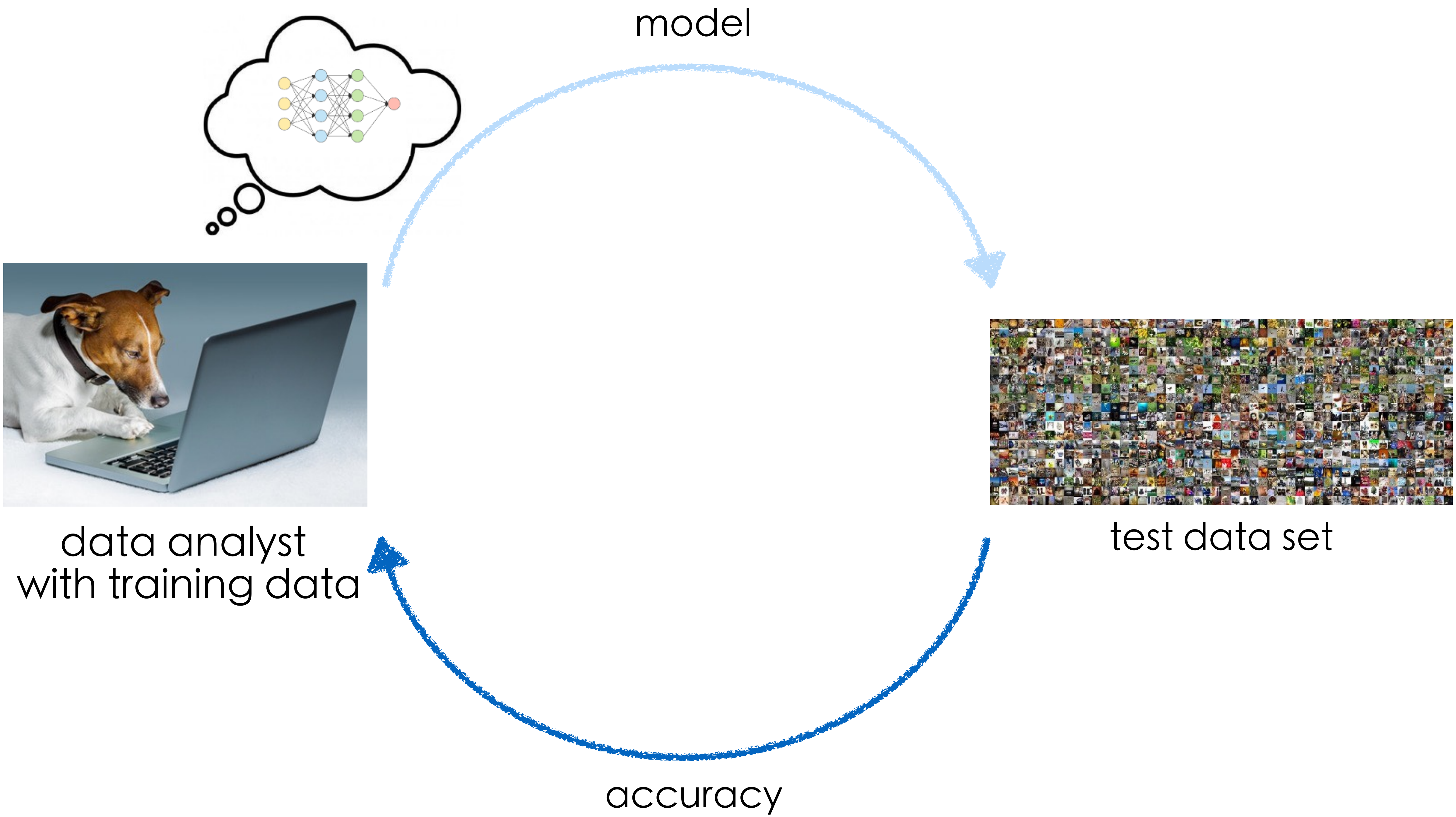
Adaptivity in Machine Learning



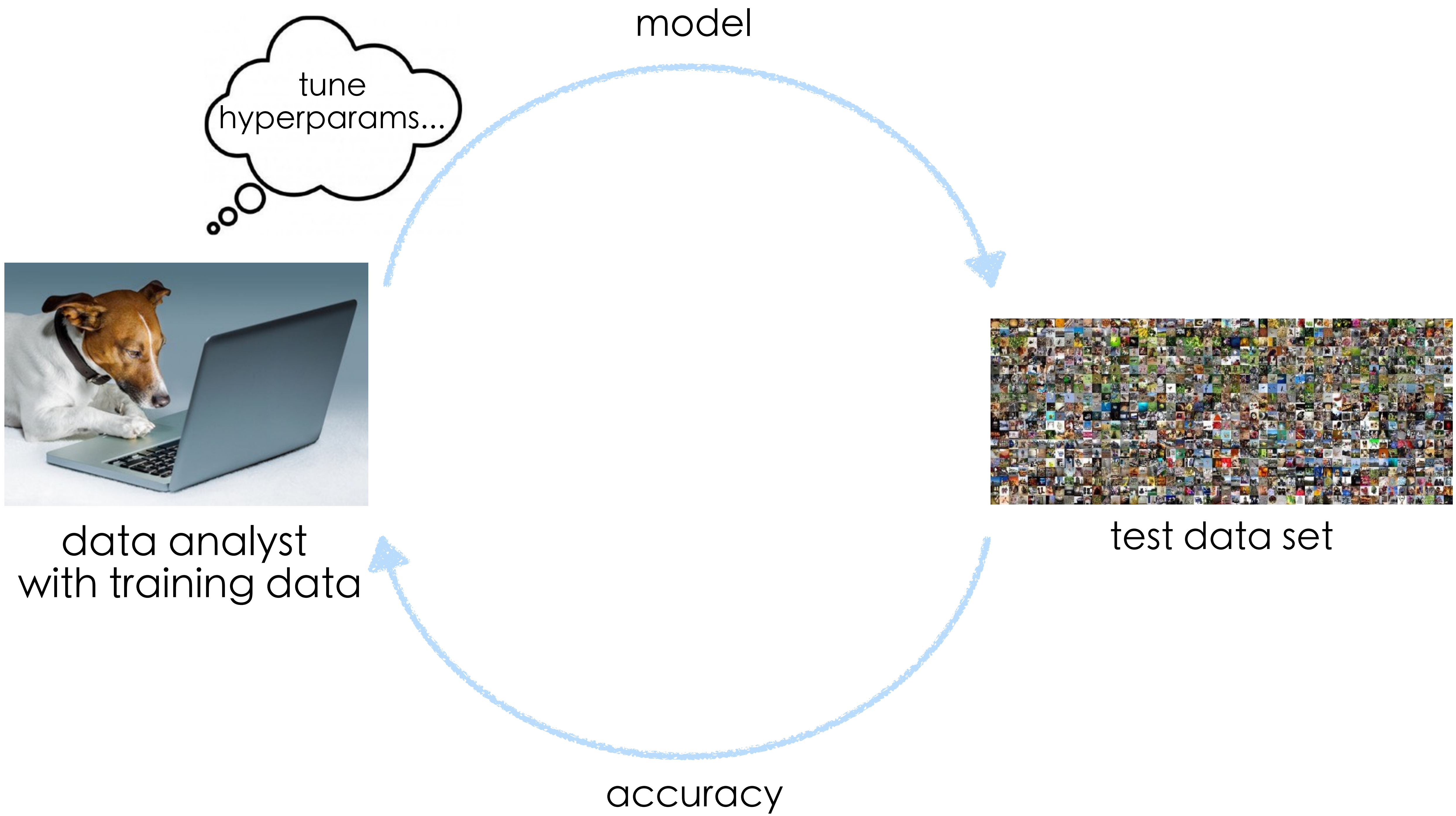
Adaptivity in Machine Learning



Adaptivity in Machine Learning



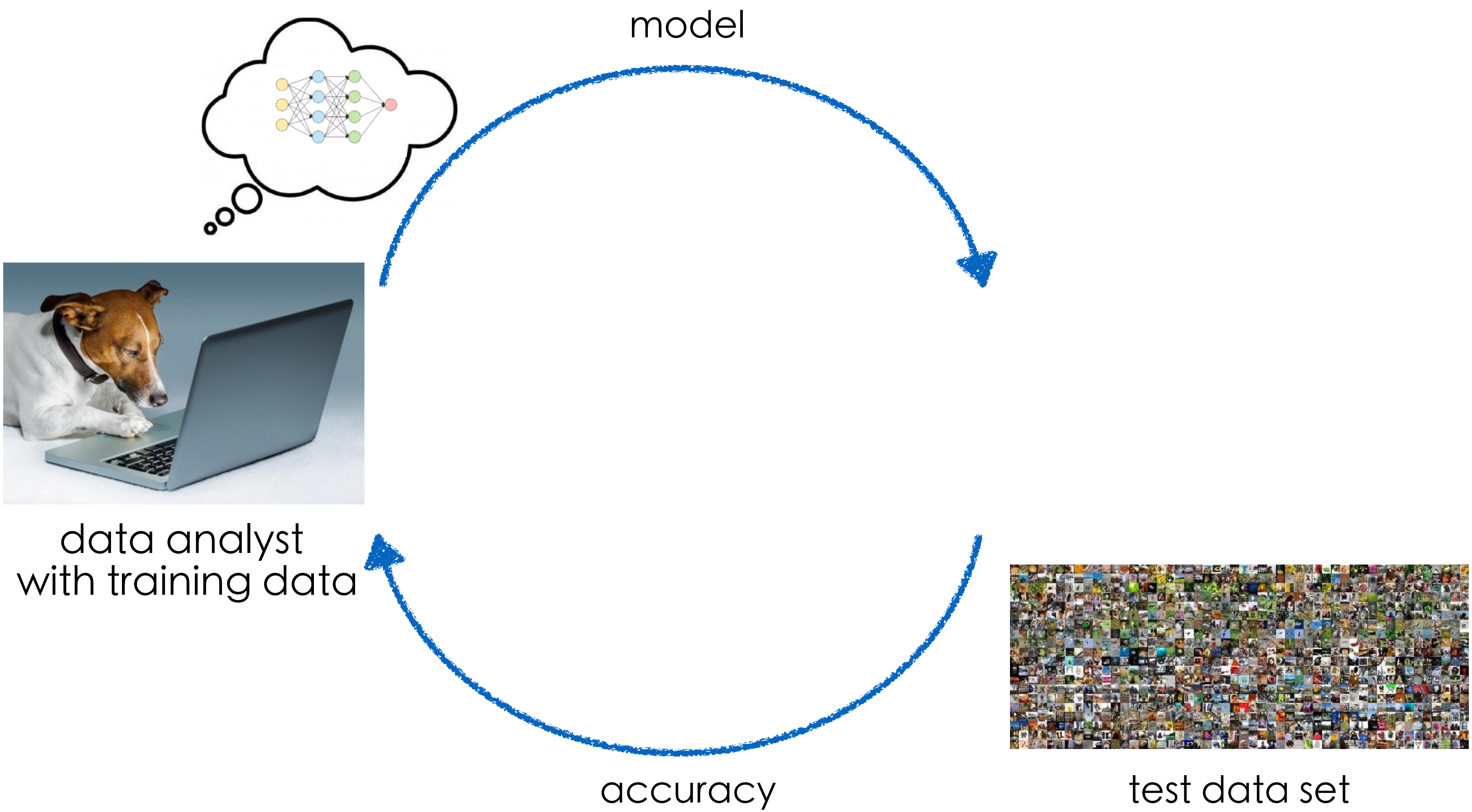
Adaptivity in Machine Learning



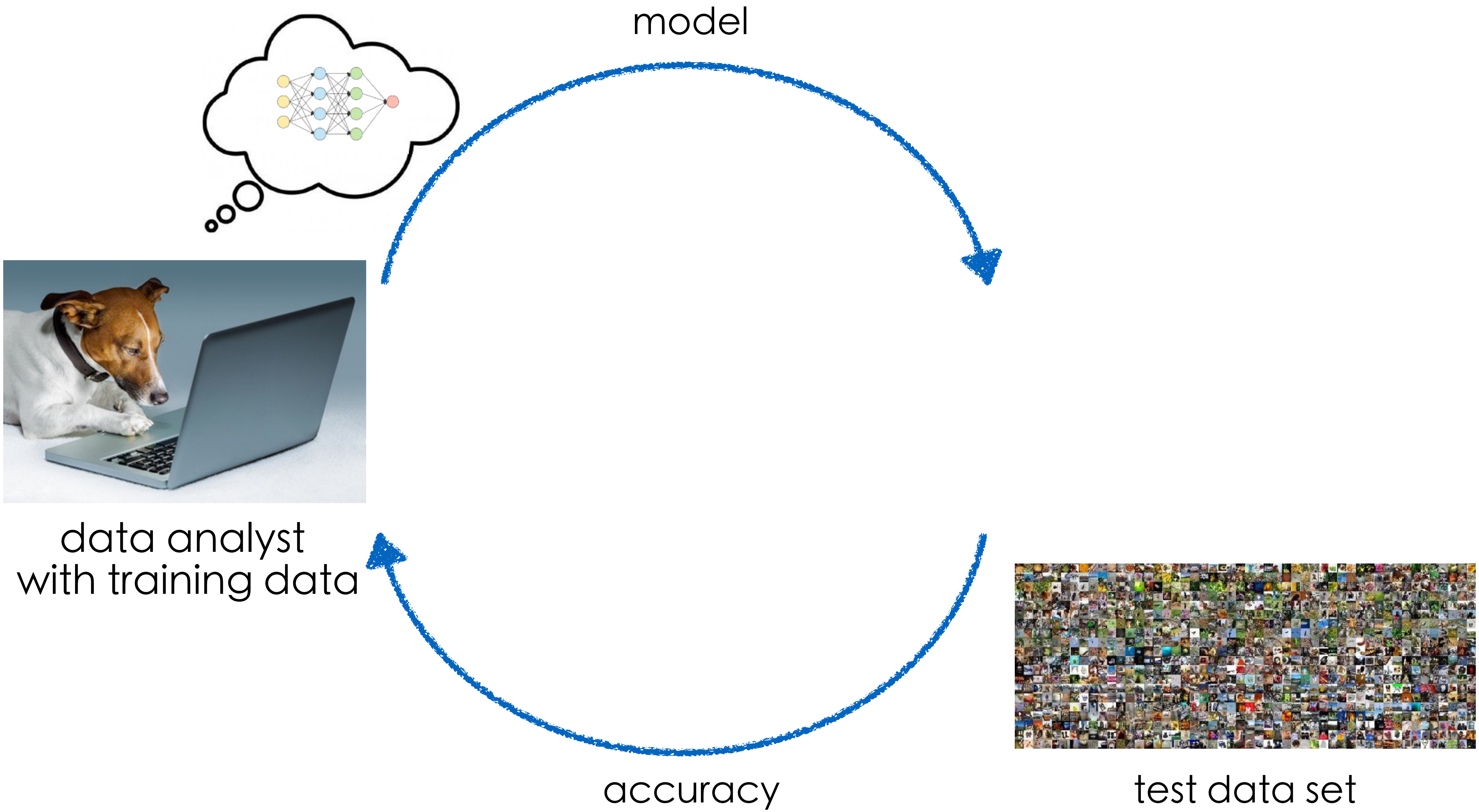
Adaptivity in Machine Learning

- After t tested models, how well does the final model generalize?
  - Depends on how the accuracies are computed

Classical Holdout vs Response Mechanism

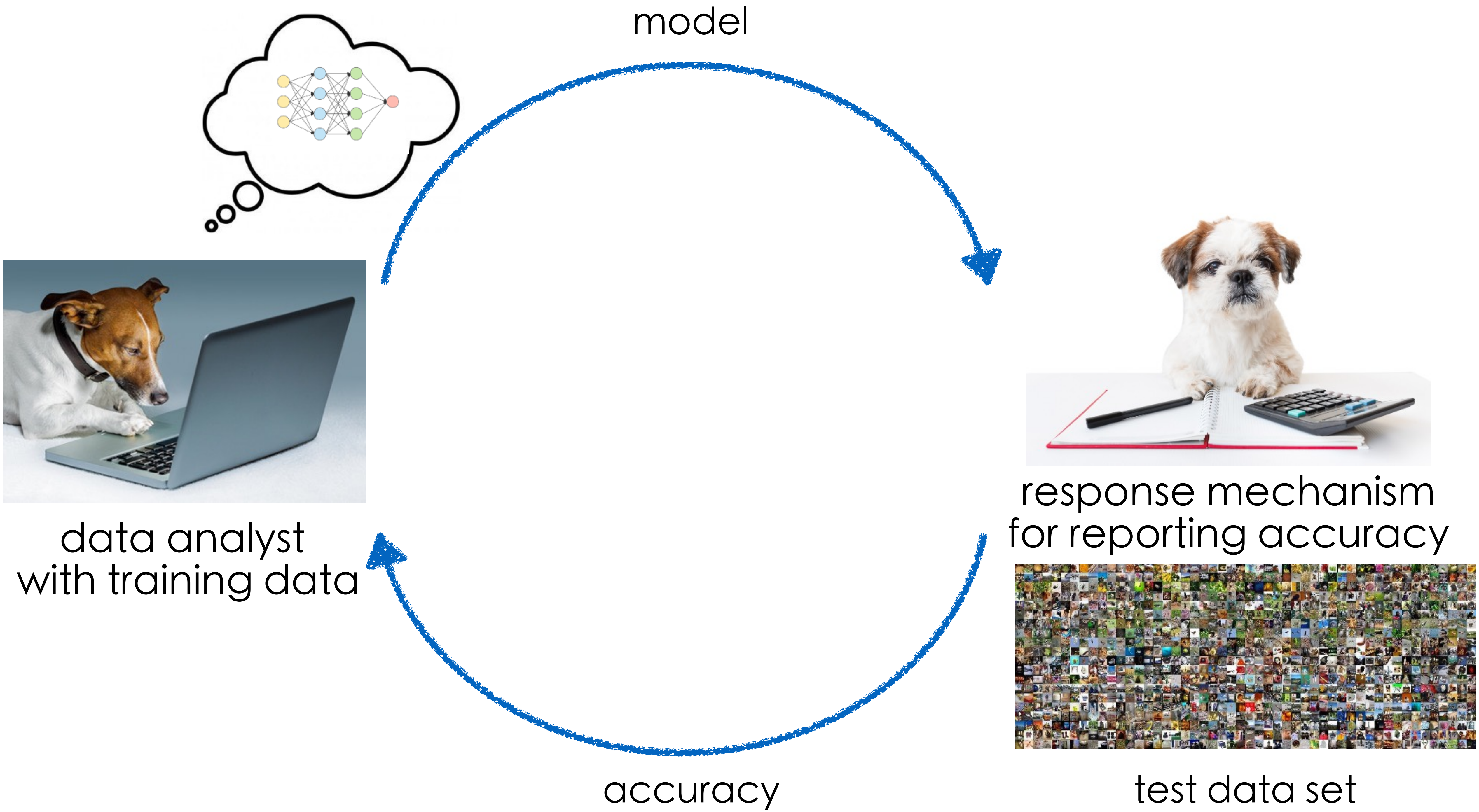


Classical Holdout vs Response Mechanism



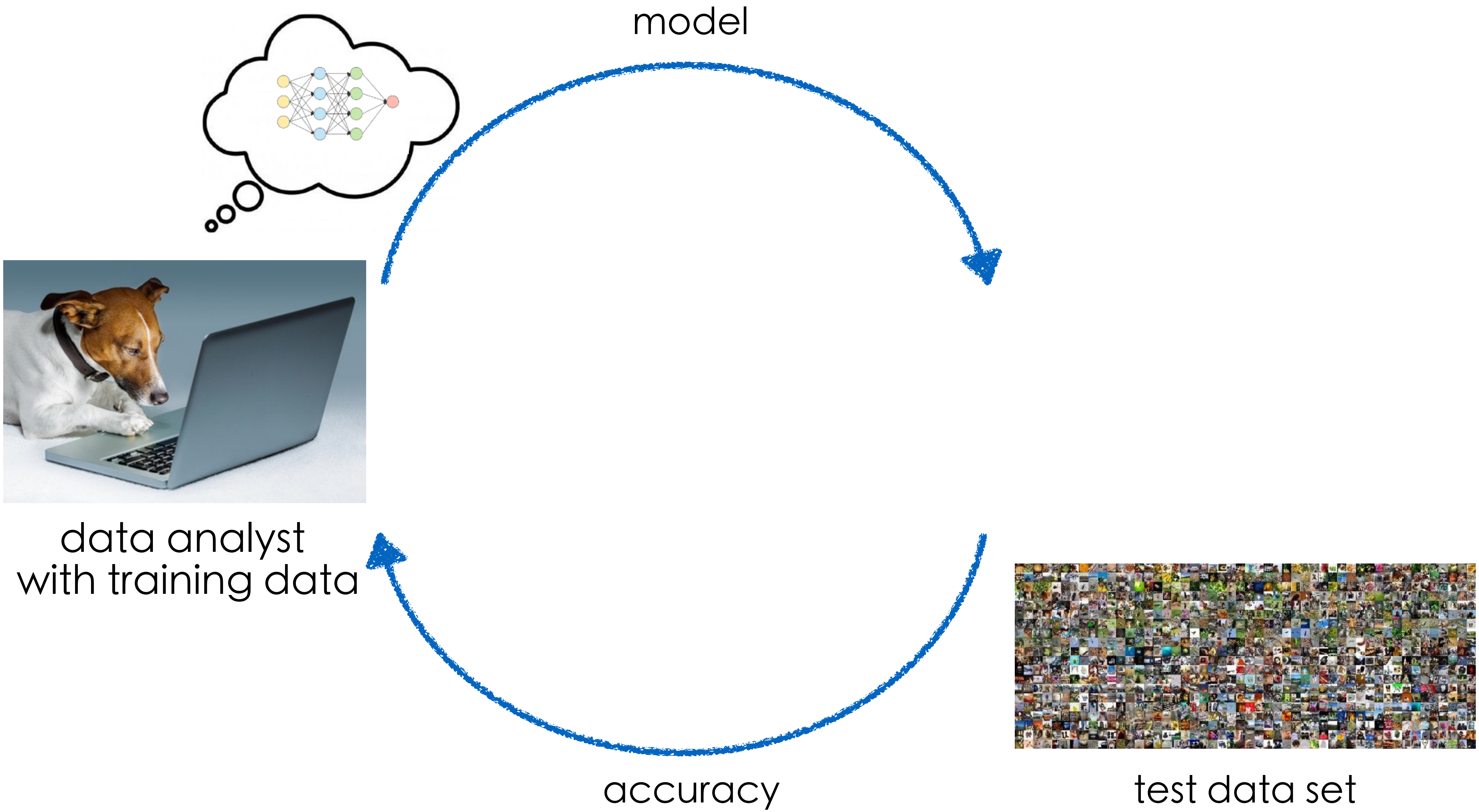
- Reporting exact sample accuracy allows for great overfitting

Classical Holdout vs Response Mechanism

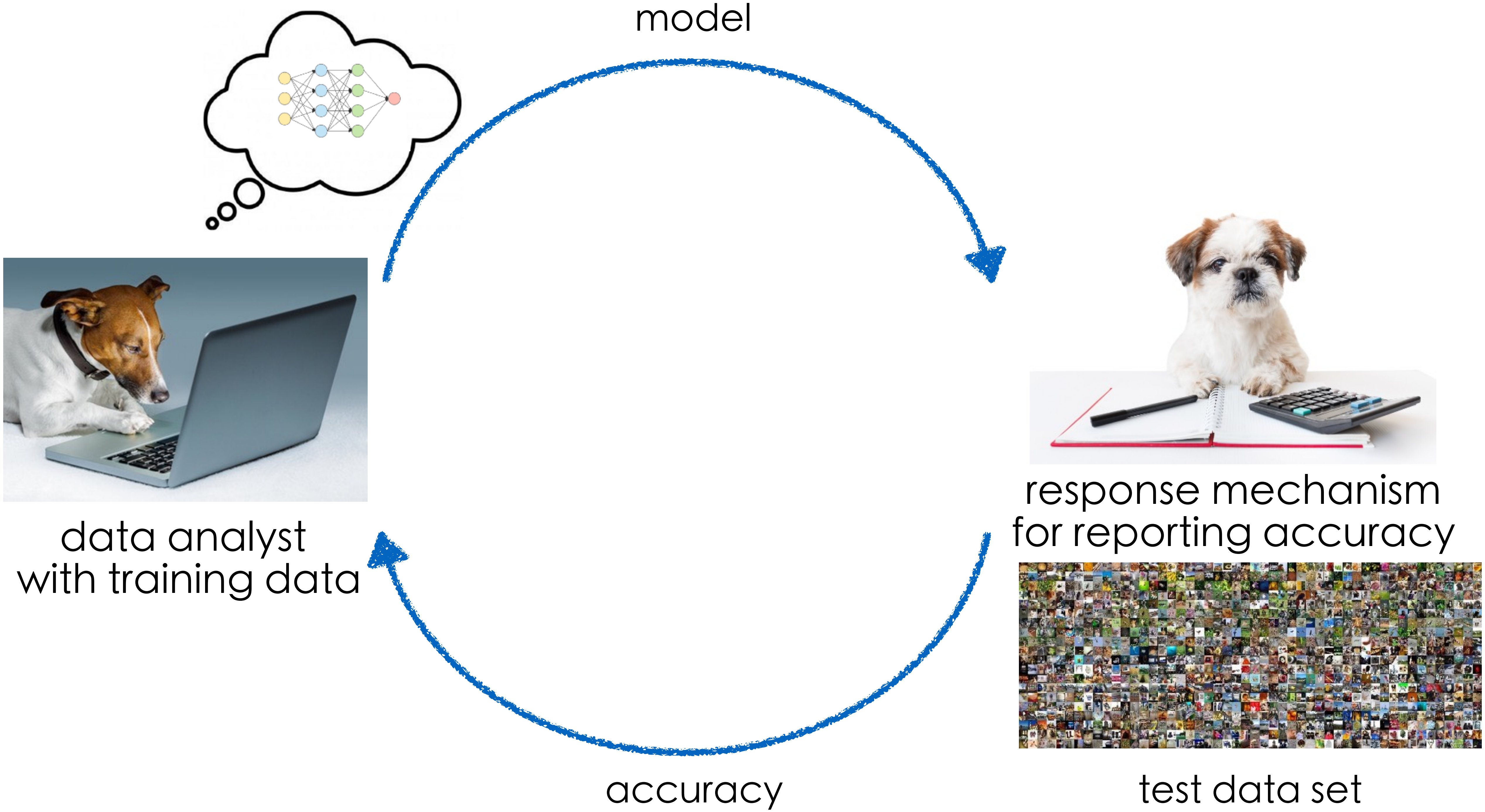


- Reporting exact sample accuracy allows for great overfitting
- Better bounds can be obtained by having a non-trivial response mechanism in charge of reporting accuracy on the test data

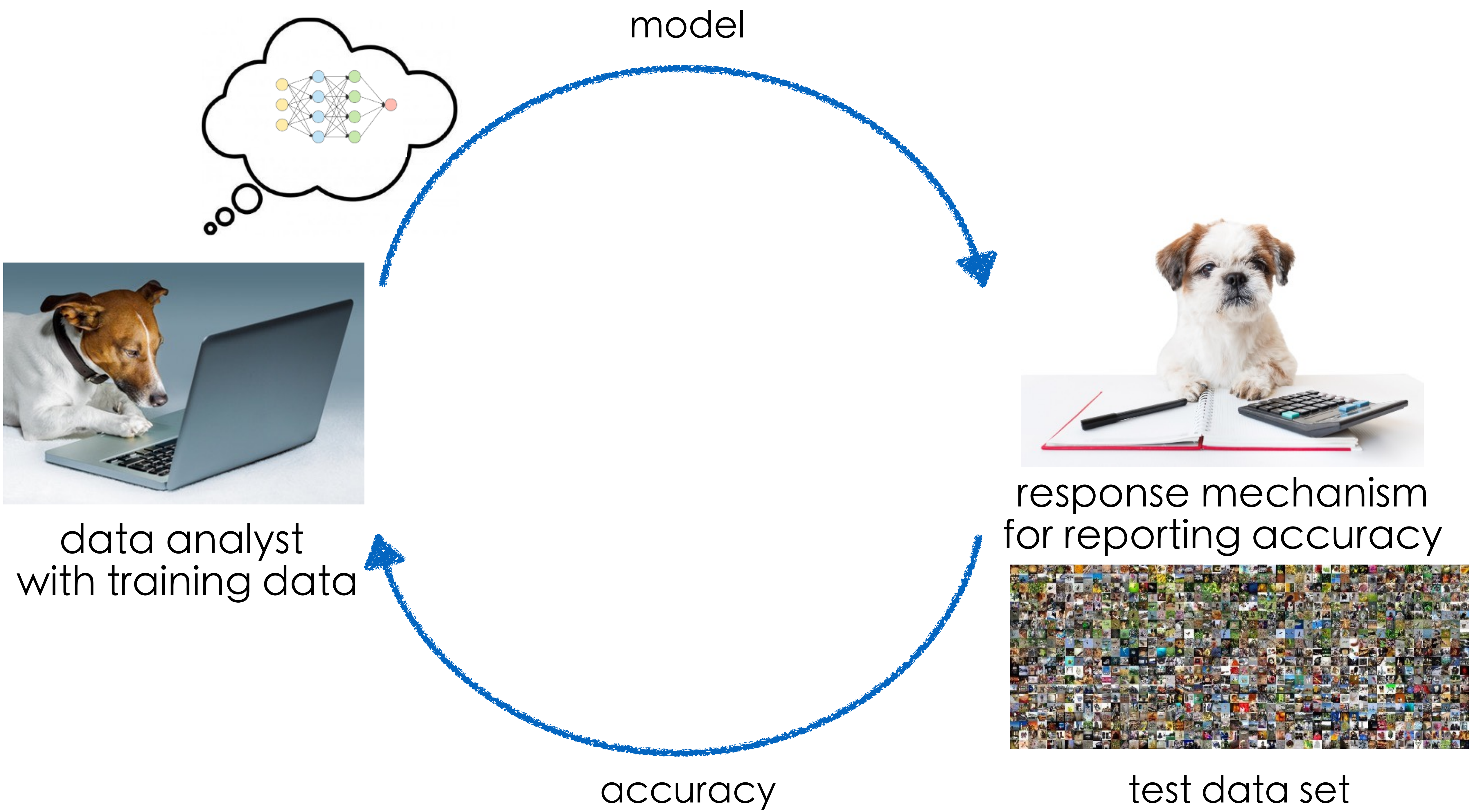
Main Questions



Main Questions

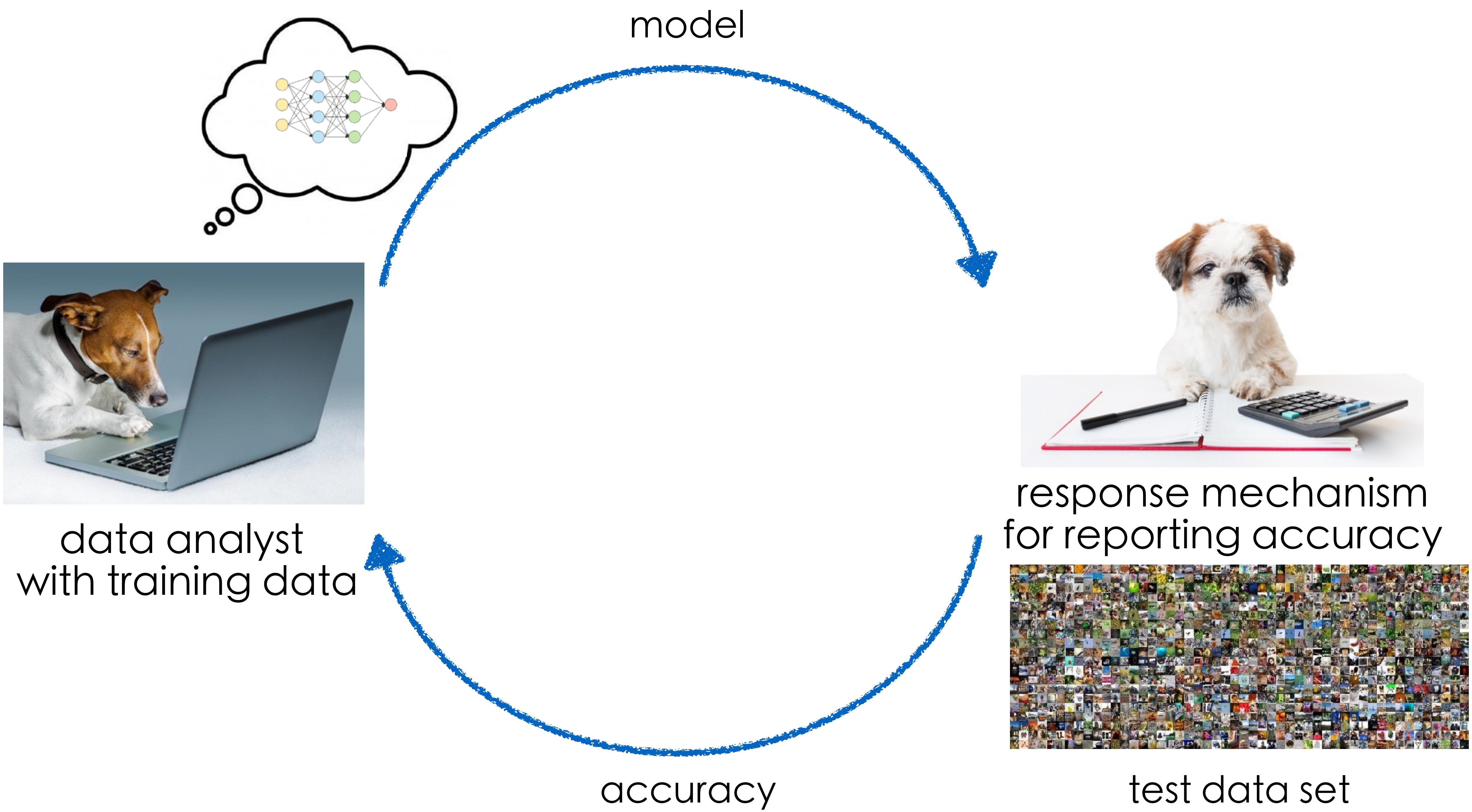


Main Questions



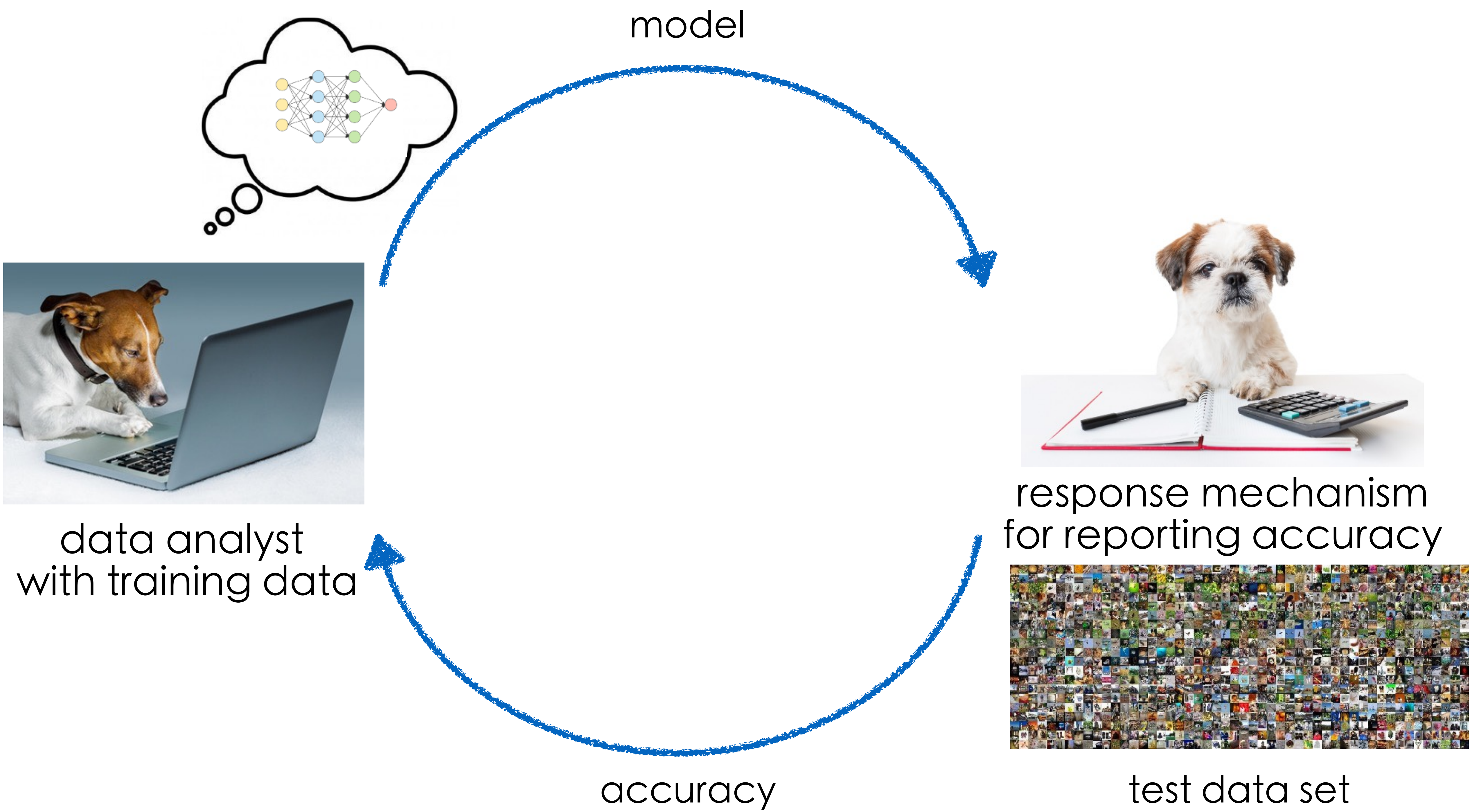
- How do we construct a mechanism such that its responses generalize to the population?

Main Questions



- How do we construct a mechanism such that its responses generalize to the population?
  - want 95% reported accuracy on test data $\approx$ 95% accuracy on fresh data from same population

Main Questions



- How do we construct a mechanism such that its responses generalize to the population?
  - want 95% reported accuracy on test data $\approx$ 95% accuracy on fresh data from same population
- For such a good mechanism, how much does a possibly adversarial analyst overfit?

Abstraction via Adaptive Data Analysis

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)

analyst

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)

Abstraction via Adaptive Data Analysis

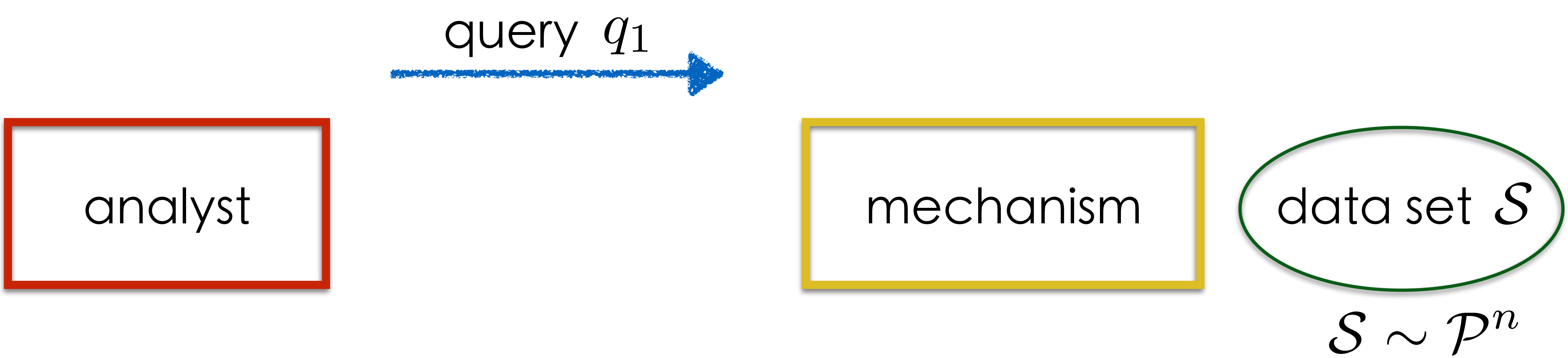
Framework of Dwork et al. (2015)

$\mathcal{P}$ - population distribution

n - sample size

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



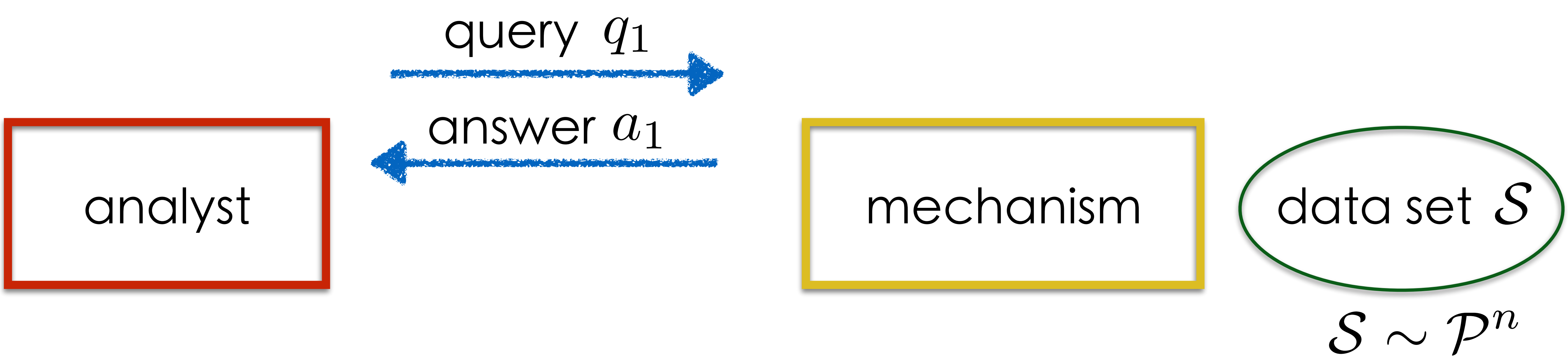
$\mathcal{P}$ - population distribution

n - sample size

$q_i : \mathrm{supp}(\mathcal{P}) \to [0,1]^d$ - queries posed by analyst

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



$\mathcal{P}$ - population distribution

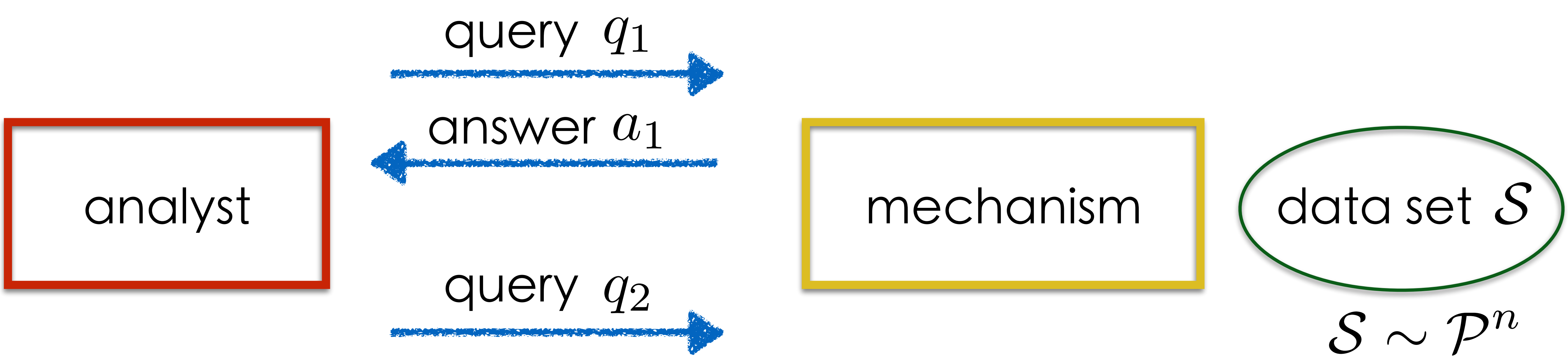
n - sample size

$q_i : \mathrm{supp}(\mathcal{P}) \to [0,1]^d$ - queries posed by analyst

$a_i \in \mathbb{R}^d$ - answers given by mechanism

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



$\mathcal{P}$ - population distribution

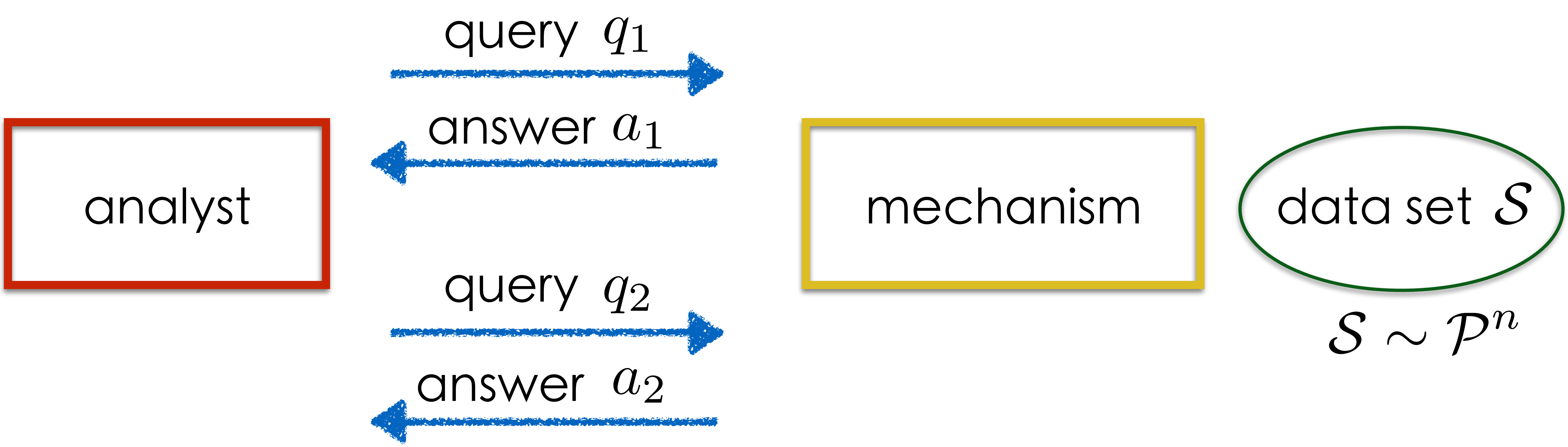
n - sample size

$q_i : \mathrm{supp}(\mathcal{P}) \to [0,1]^d$ - queries posed by analyst

$a_i \in \mathbb{R}^d$ - answers given by mechanism

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



$\mathcal{P}$- population distribution

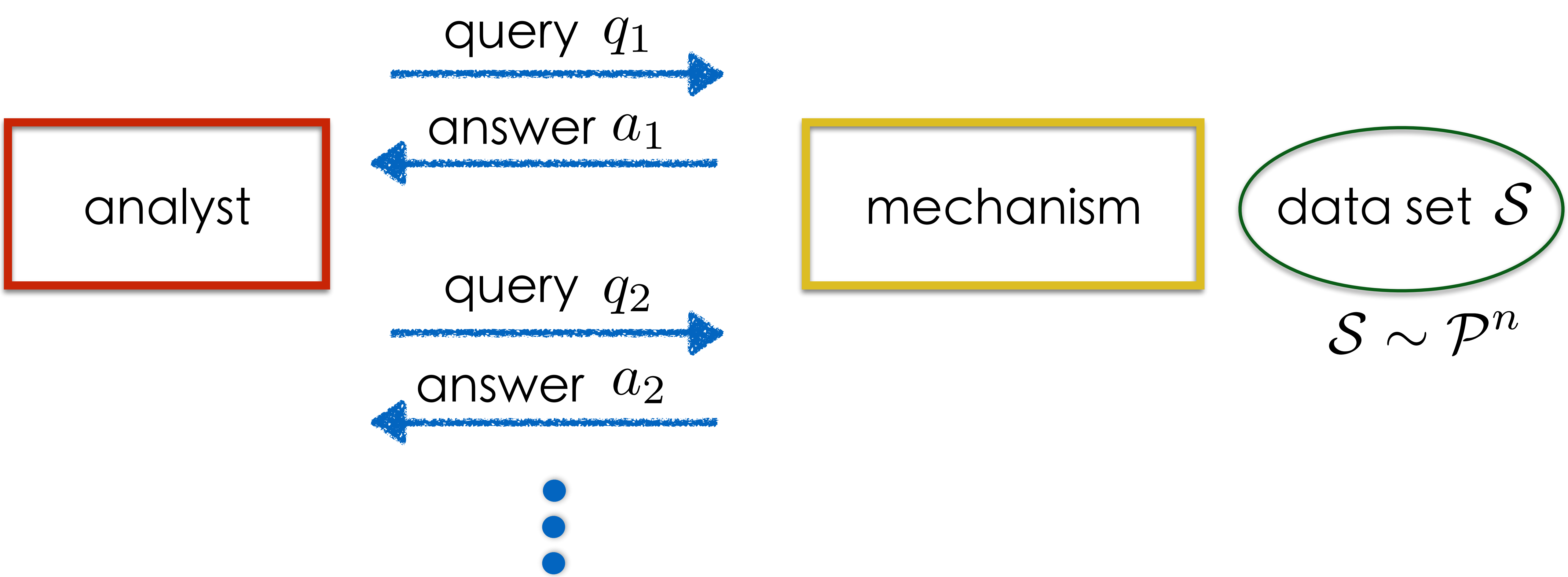
n - sample size

$q_i : \mathrm{supp}(\mathcal{P}) \to [0,1]^d$- queries posed by analyst

$a_i \in \mathbb{R}^d$ - answers given by mechanism

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



$\mathcal{P}$ - population distribution

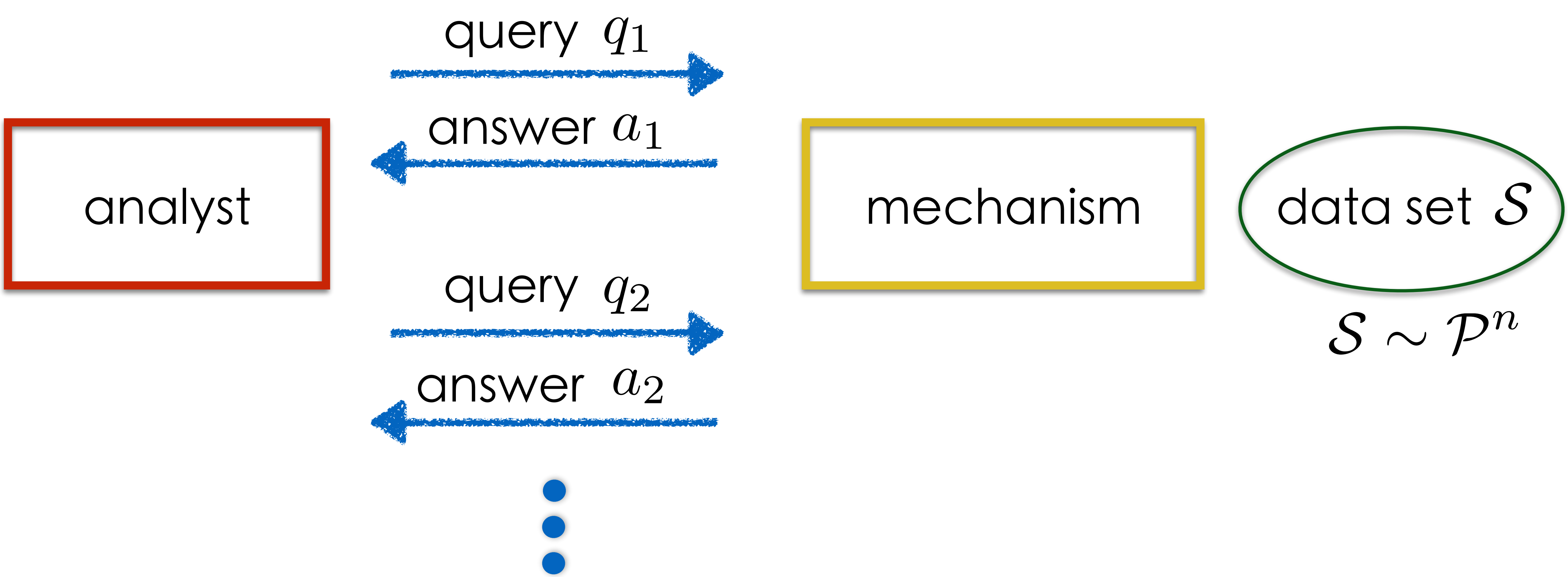
n - sample size

$q_i : \mathrm{supp}(\mathcal{P}) \to [0,1]^d$ - queries posed by analyst

$a_i \in \mathbb{R}^d$ - answers given by mechanism

Abstraction via Adaptive Data Analysis

Framework of Dwork et al. (2015)



$\mathcal{P}$- population distribution

n - sample size

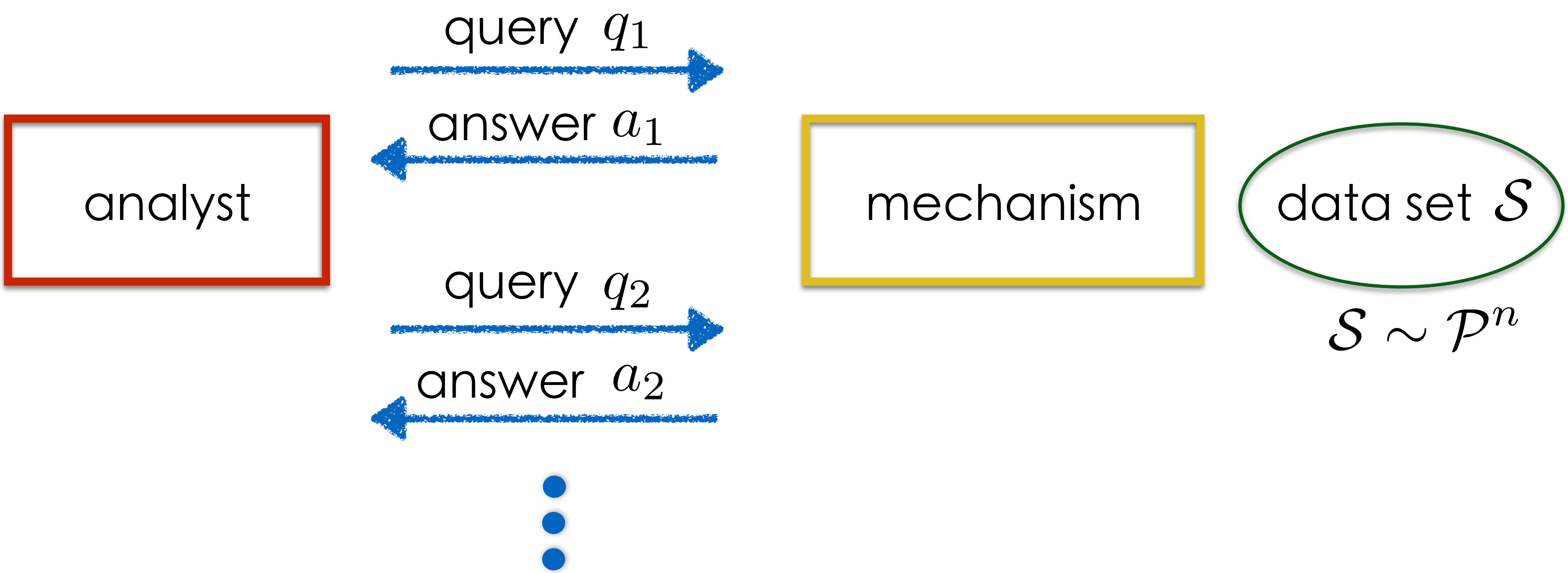
$\mathcal{S} = \{X_1, \ldots, X_n\}$ - data set

$q_i : \text{supp}(\mathcal{P}) \to [0,1]^d$- queries posed by analyst

$a_i \in \mathbb{R}^d$ - answers given by mechanism

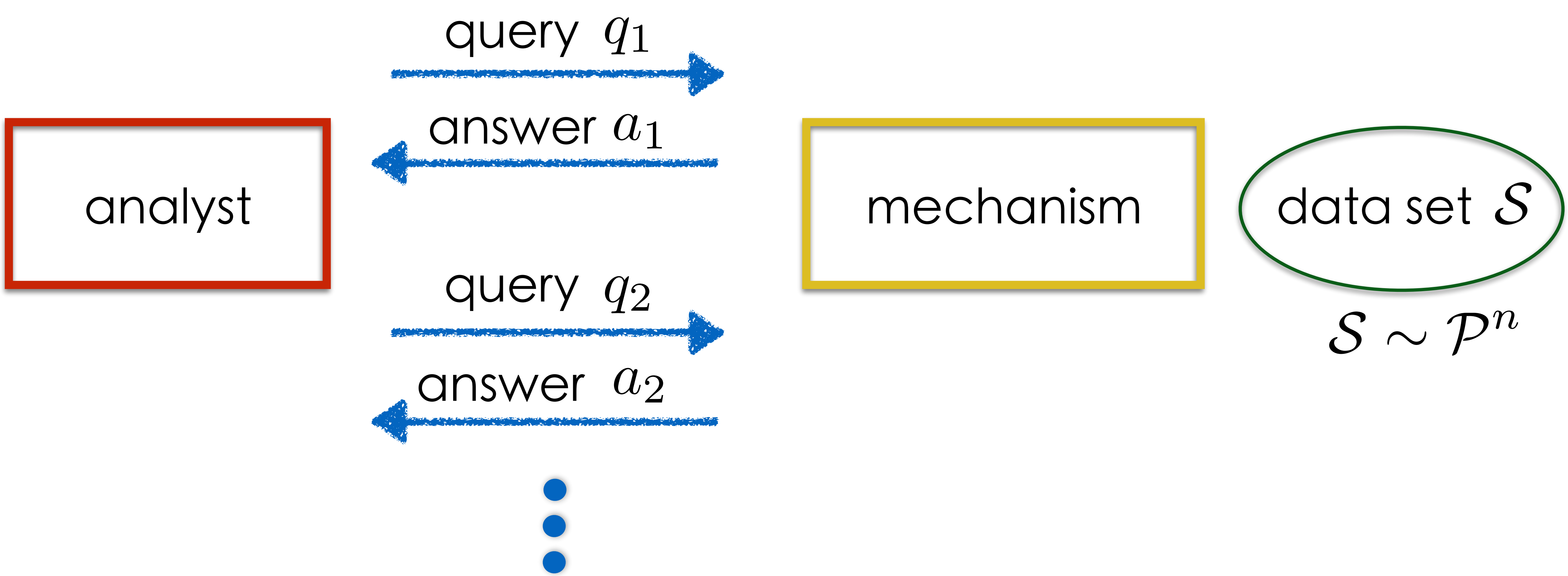
Generalization Error

Framework of Dwork et al. (2015)



Generalization Error

Framework of Dwork et al. (2015)



Goal: design a mechanism such that, for t queries posed by the analyst, generalization error is at most ϵ:

$$\max_{1 \leq i \leq t} \|\mathbb{E}_{X \sim \mathcal{P}}[q_i(X)] - a_i\|_\infty \leq \epsilon \text{ with high probability.}$$

examples of analysts	
human analyst	iterative algorithms, e.g. gradient descent
q_i- classification error of i-th classifier on data set $\mathcal{S}$	q_i - gradient of the empirical risk on data set $\mathcal{S}$

examples of mechanisms		
empirical mechanism	Gaussian mechanism	truncation to a fixed number of bits
$a_i = \frac{1}{n}\sum_{j=1}^{n} q_i(X_j)$	$a_i = \frac{1}{n}\sum_{j=1}^{n} q_i(X_j) + \xi_i,$ $\xi_i \sim N(0, \sigma^2 I_d)$	$a_i = \mathrm{trunc}\left(\frac{1}{n}\sum_{j=1}^{n} q_i(X_j)\right)$

Known bounds on the generalization error after t queries:

Known bounds on the generalization error after t queries:

non-adaptive analyst

fully adaptive analyst

Known bounds on the generalization error after t queries:

generalization error

$$O\left(\sqrt{\frac{\log(td)}{n}}\right)$$

generalization error

$$\tilde{O}\left(\frac{(td)^{1/4}}{\sqrt{n}}\right)$$

Known bounds on the generalization error after t queries:

non-adaptive analyst

fully adaptive analyst

generalization error

$$O\left(\sqrt{\frac{\log(td)}{n}}\right)$$

tight

generalization error

$$\tilde{O}\left(\frac{(td)^{1/4}}{\sqrt{n}}\right)$$

tight*

*tight for a broad class of mechanisms, believed to be tight in general

Known bounds on the generalization error after t queries:

overly optimistic

non-adaptive analyst

overly pessimistic

fully adaptive analyst

generalization error

$$O\left(\sqrt{\frac{\log(td)}{n}}\right)$$

tight

generalization error

$$\tilde{O}\left(\frac{(td)^{1/4}}{\sqrt{n}}\right)$$

tight*

*tight for a broad class of mechanisms, believed to be tight in general

Known bounds on the generalization error after t queries:

overly optimistic

overly pessimistic

non-adaptive analyst

natural analysts?

fully adaptive analyst

generalization error

$$O\left(\sqrt{\frac{\log(td)}{n}}\right)$$

generalization error

$$\tilde{O}\left(\frac{(td)^{1/4}}{\sqrt{n}}\right)$$

Are there natural categories of analysts which interpolate between logarithmic and polynomial error?

Analysts as Dynamical Systems

Analysts as Dynamical Systems

We model the data analyst as a dynamical system:

Analysts as Dynamical Systems

We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function

Analysts as Dynamical Systems

We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function

Analysts as Dynamical Systems

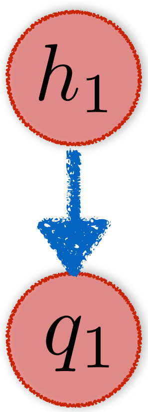
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

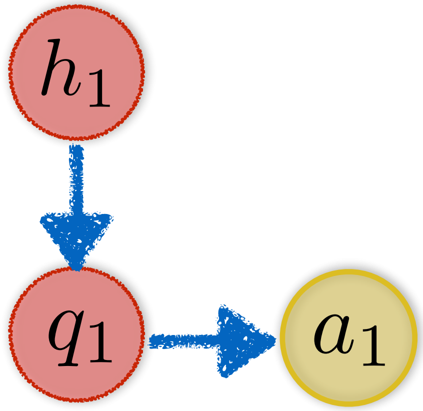
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

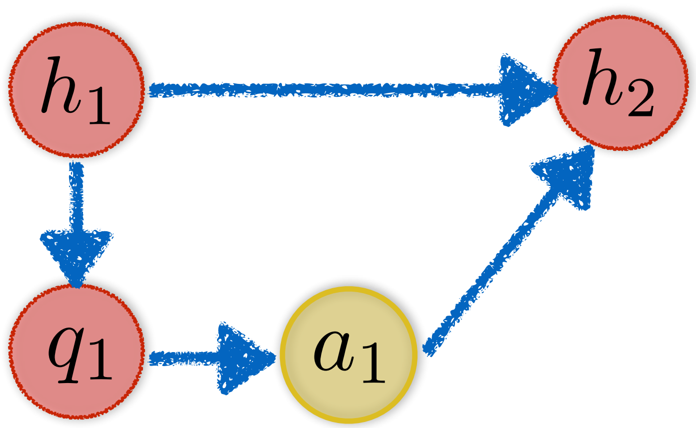
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

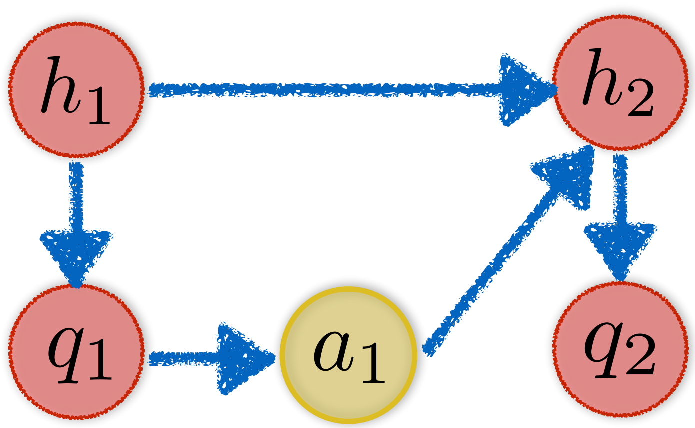
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

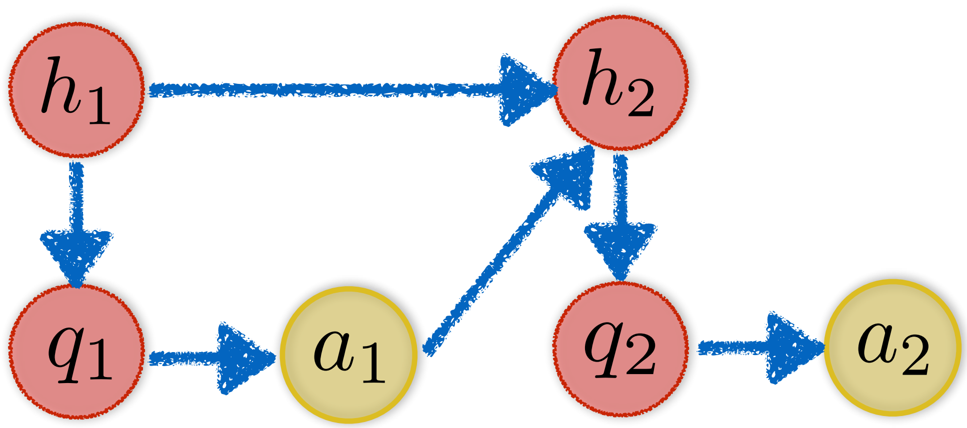
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

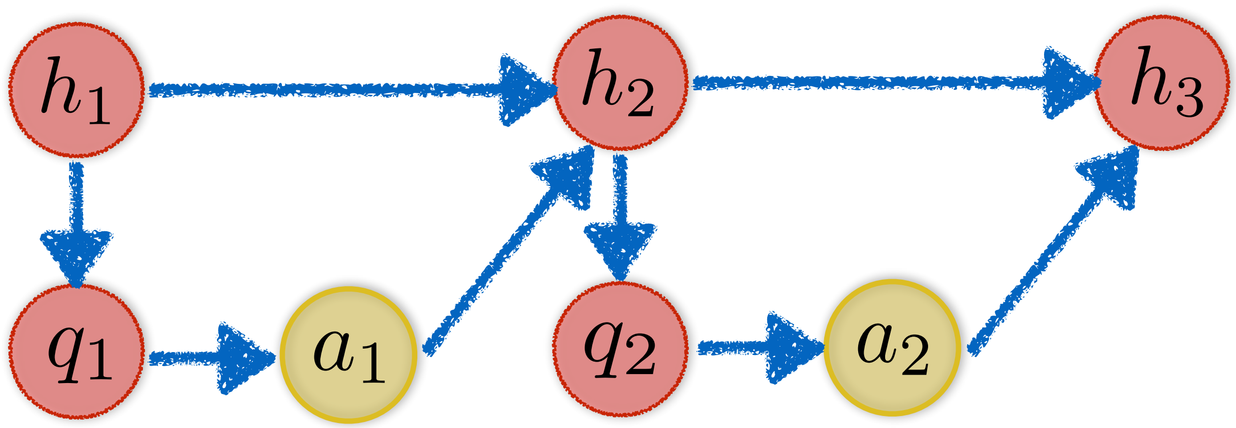
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

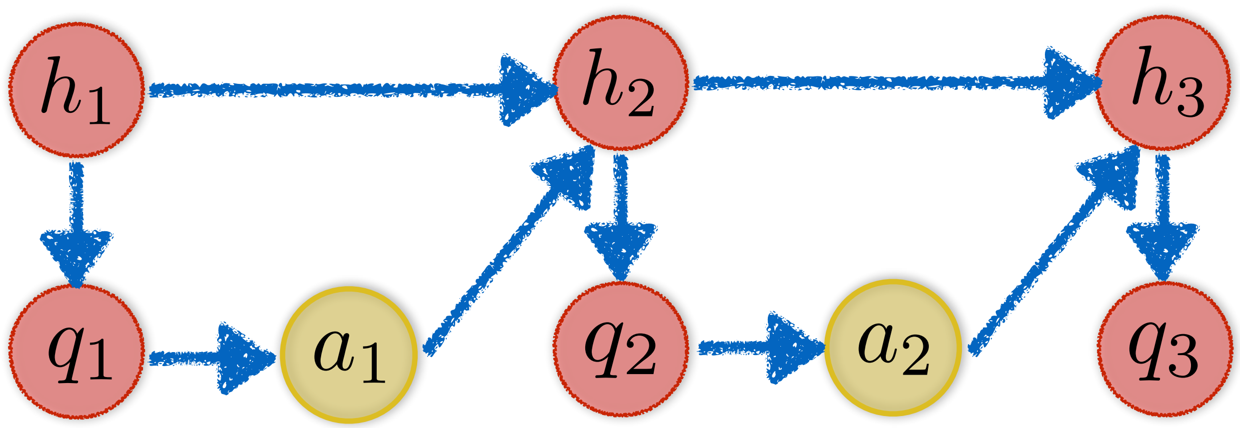
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

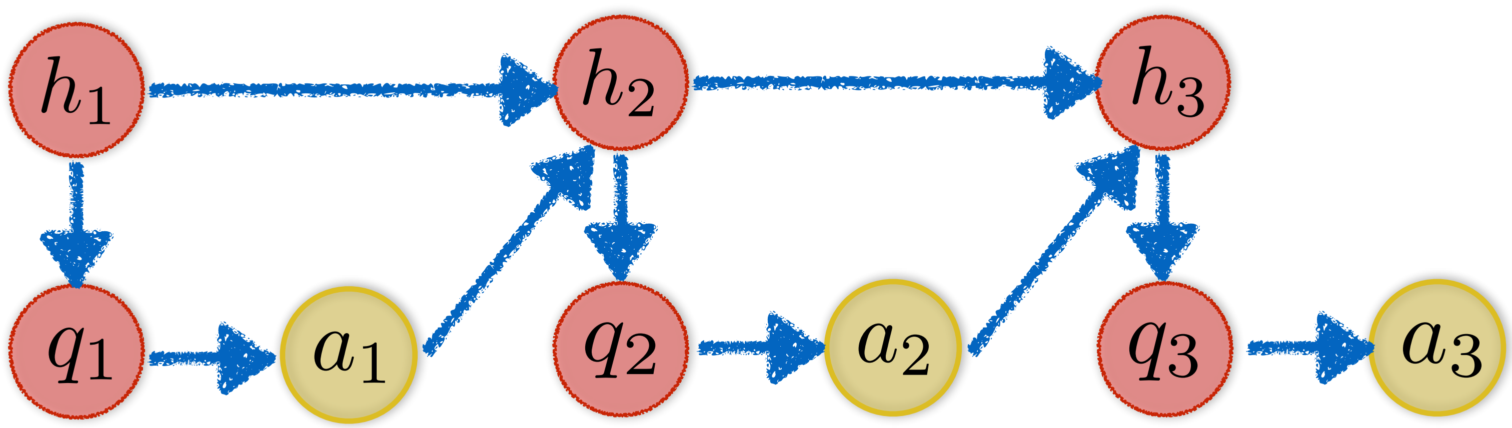
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

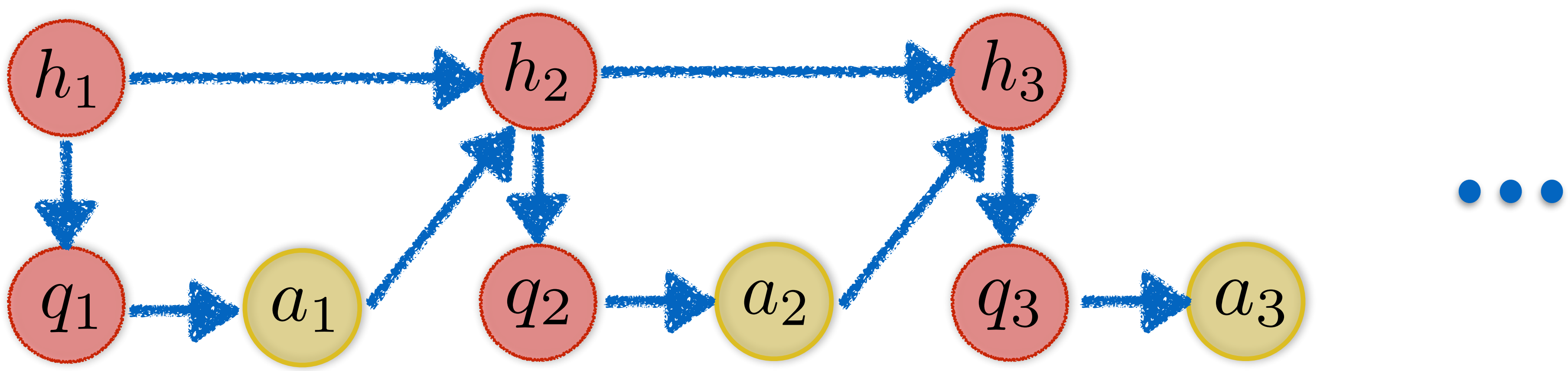
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



Analysts as Dynamical Systems

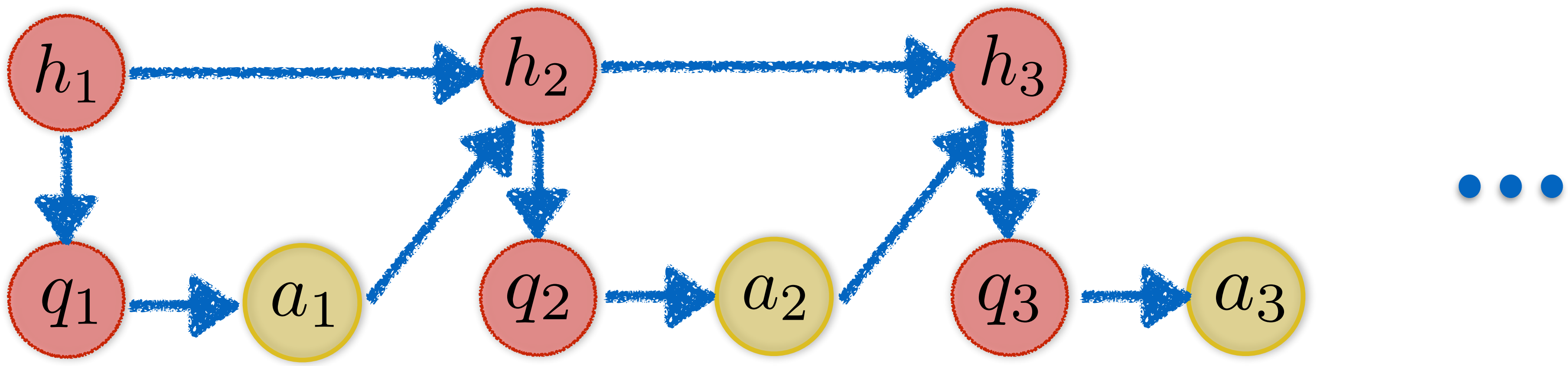
We model the data analyst as a dynamical system:

$$h_t = \psi_t(h_{t-1}, a_{t-1})$$
$$q_t = f_t(h_t)$$

h_t - history, i.e. encoding of past interactions

ψ_t - arbitrary transition map

f_t - arbitrary function



With no restriction on the transition map, this representation captures an arbitrary adaptive analyst

Stability of Natural Analysts

Stability of Natural Analysts

- Stability of dynamical systems makes analysts natural

Stability of Natural Analysts

- Stability of dynamical systems makes analysts natural
- Stability ensures convergence of algorithms, encodes different human biases, like sensitivity to interactions far enough in the past, etc.

Stability of Natural Analysts

- Stability of dynamical systems makes analysts natural
- Stability ensures convergence of algorithms, encodes different human biases, like sensitivity to interactions far enough in the past, etc.
- Encoding different stability notions, we introduce two main classes of natural analysts: progressive and conservative

Progressive Analysts

Progressive Analysts

Progressive analysts contract their history as:

$$\|\psi_t(h, a) - \psi_t(h', a)\| \leq \lambda \|h - h'\|, \ \forall h, h', a$$

for some $\lambda \in (0, 1)$.

Progressive Analysts

Progressive analysts contract their history as:

$$\|\psi_t(h,a) - \psi_t(h',a)\| \leq \lambda \|h - h'\|, \ \forall h, h', a$$

for some $\lambda \in (0,1)$.

The parameter λ encodes how fast past interactions with the mechanism are forgotten;
$\lambda \approx 0$ is minimal adaptivity, while $\lambda \approx 1$ implies full adaptivity.

Progressive Analysts

Progressive analysts contract their history as:

$$\|\psi_t(h,a) - \psi_t(h',a)\| \leq \lambda \|h - h'\|, \ \forall h, h', a$$

for some $\lambda \in (0,1)$.

The parameter λ encodes how fast past interactions with the mechanism are forgotten;
$\lambda \approx 0$ is minimal adaptivity, while $\lambda \approx 1$ implies full adaptivity.

examples of progressive analysts	
human analysts	algorithmic analysts
analysts with recency bias	stable RNNs, Bellman operator

Main Theorem for Progressive Analysts

Main Theorem for Progressive Analysts

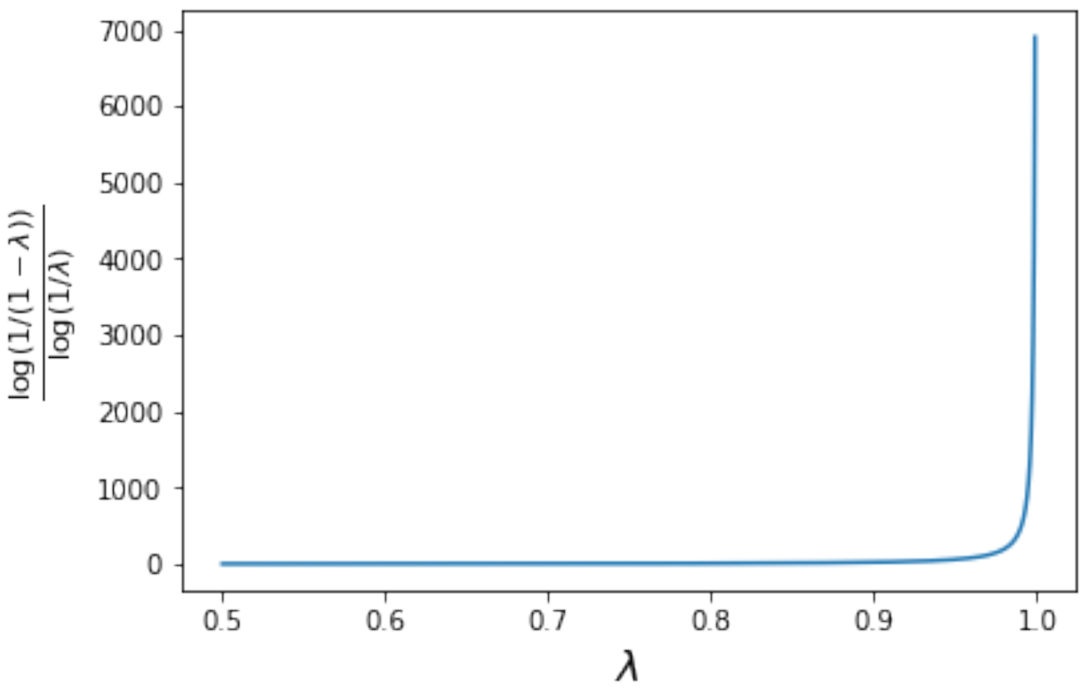
There exists a computationally efficient mechanism for answering t queries chosen adaptively by a progressive analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\sqrt{\frac{\log(1/(1-\lambda))\log(t)d}{\log(1/\lambda)n}}\right)$$

Main Theorem for Progressive Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a progressive analyst such that the overall generalization error is at most

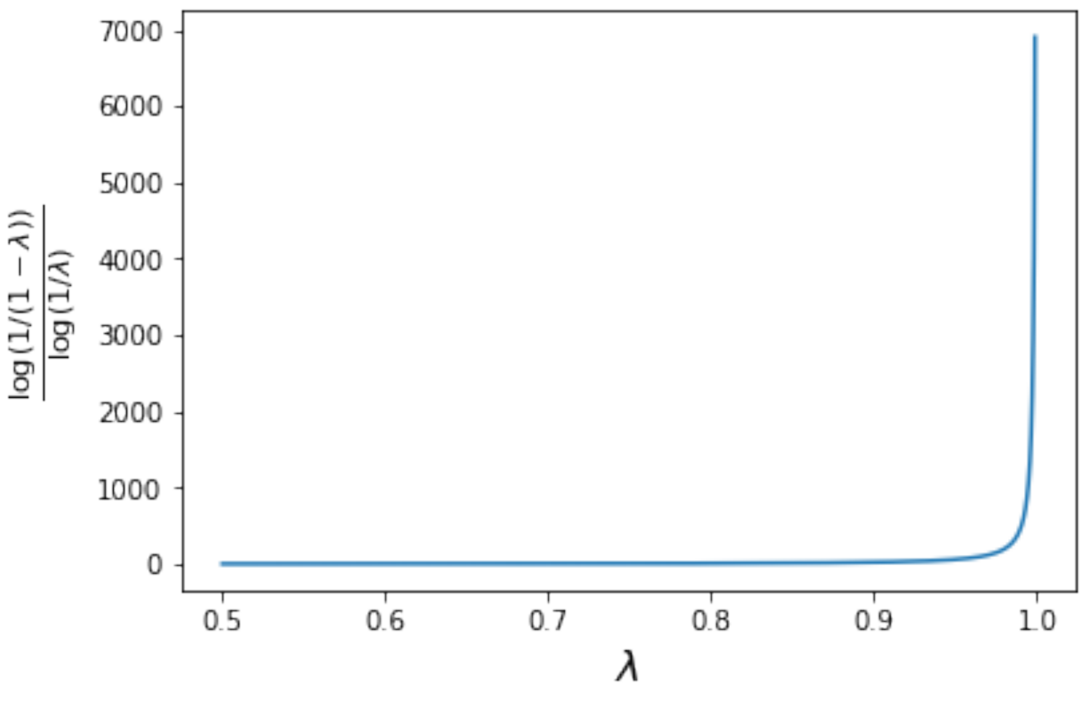
$$\approx \tilde{O}\left(\sqrt{\frac{\log(1/(1-\lambda))\log(t)d}{\log(1/\lambda)n}}\right)$$



Main Theorem for Progressive Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a progressive analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\sqrt{\frac{\log(1/(1-\lambda))\log(t)d}{\log(1/\lambda)n}}\right)$$



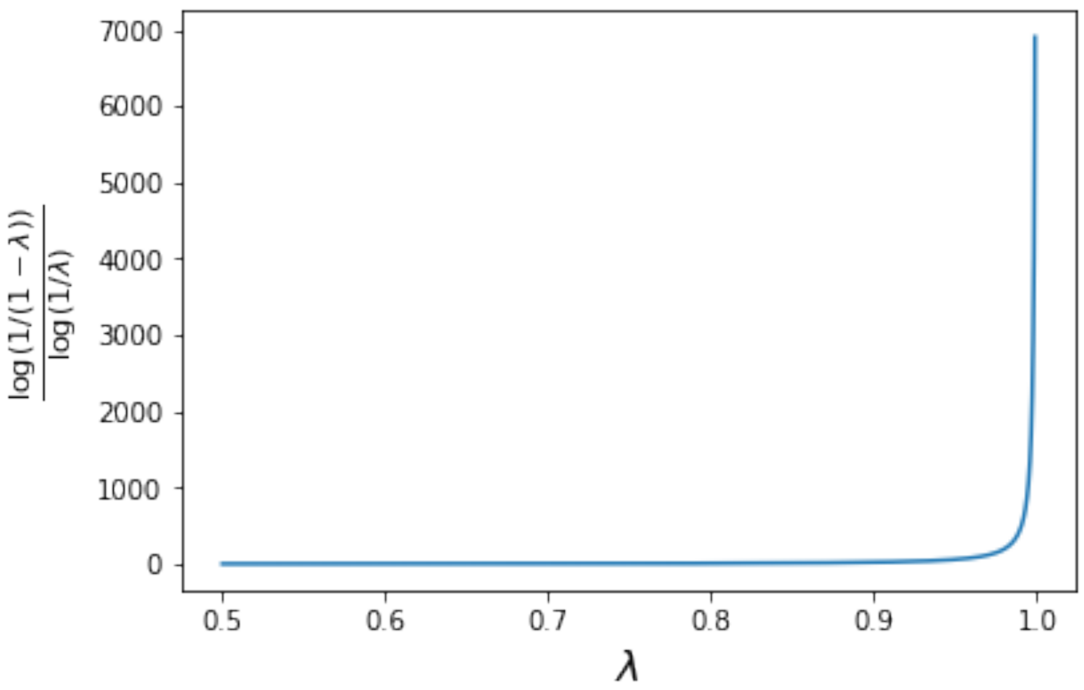
For $\lambda = 1 - 1/t$ the analyst is fully adaptive and we recover a (suboptimal) fully adaptive bound

$$\tilde{O}(\sqrt{td/n})$$

Main Theorem for Progressive Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a progressive analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\sqrt{\frac{\log(1/(1-\lambda))\log(t)d}{\log(1/\lambda)n}}\right)$$



For $\lambda = 1 - 1/t$ the analyst is fully adaptive and we recover a (suboptimal) fully adaptive bound

$$\tilde{O}(\sqrt{td/n})$$

For $\lambda = 0$ the analyst can only adapt to the last answer and we have

$$\tilde{O}(\sqrt{\log(t)d/n})$$

Conservative Analysts

Conservative Analysts

Conservative analysts contract new evidence* as:

$$\|\psi_t(h,a) - \psi_t(h,a')\| \leq \eta_t \|a - a'\|, \ \forall h, a, a'$$

for some sequence $\{\eta_t\}$ such that $\lim_{t \to \infty} \eta_t = 0.$

*alternate condition for conservative analysts given in paper

Conservative Analysts

Conservative analysts contract new evidence* as:

$$\|\psi_t(h, a) - \psi_t(h, a')\| \leq \eta_t \|a - a'\|, \ \forall h, a, a'$$

for some sequence $\{\eta_t\}$ such that $\lim_{t \to \infty} \eta_t = 0.$

The sequence $\{\eta_t\}$ encodes how fast the knowledge of the analyst saturates.

*alternate condition for conservative analysts given in paper

Conservative Analysts

Conservative analysts contract new evidence* as:

$$\|\psi_t(h,a) - \psi_t(h,a')\| \leq \eta_t \|a - a'\|, \ \forall h, a, a'$$

for some sequence $\{\eta_t\}$ such that $\lim_{t\to\infty} \eta_t = 0.$

The sequence $\{\eta_t\}$ encodes how fast the knowledge of the analyst saturates.

examples of conservative analysts	
human analysts	algorithmic analysts
analysts with anchoring bias	optimization algorithms, e.g. gradient descent

*alternate condition for conservative analysts given in paper

Main Theorem for Conservative Analysts

Main Theorem for Conservative Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a conservative analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\frac{(\min\{t, K(\eta_t)\} d \log(t))^{1/4}}{\sqrt{n}}\right), \; K(\eta_t) = \min\{t : \eta_t \leq C/\sqrt{d}\}$$

for some constant C.

Main Theorem for Conservative Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a conservative analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\frac{(\min\{t, K(\eta_t)\}d\log(t))^{1/4}}{\sqrt{n}}\right), \quad K(\eta_t) = \min\{t : \eta_t \leq C/\sqrt{d}\}$$

for some constant C.

If $\eta_t \approx 0, \forall t$ we recover the non-adaptive bound $\tilde{O}(\sqrt{\log(td)/n})$

Main Theorem for Conservative Analysts

There exists a computationally efficient mechanism for answering t queries chosen adaptively by a conservative analyst such that the overall generalization error is at most

$$\approx \tilde{O}\left(\frac{(\min\{t, K(\eta_t)\}d\log(t))^{1/4}}{\sqrt{n}}\right), \quad K(\eta_t) = \min\{t : \eta_t \leq C/\sqrt{d}\}$$

for some constant C.

If $\eta_t \approx 0, \forall t$ we recover the non-adaptive bound $\tilde{O}(\sqrt{\log(td)/n})$

If $\{\eta_t\}$ has a slow decay, we recover the (tight) bound under full adaptivity $\tilde{O}((td)^{1/4}/\sqrt{n})$

Summary

- Generalization bounds in adaptive data analysis show a wide gap due to considering only overly optimistic or overly pessimistic settings

- In our work, we smoothly interpolate between the two by using stability parameters as a knob

Future Directions

- Empirical evaluation of patterns of human adaptivity

- Preventing the analyst from knowing the distribution of the data

- Limiting query complexity

Thank you