Cost-effective data annotation with Bayesian experimental design

Quan Nguyen

PyData 2024

Who I am

- **Quan Nguyen**
- Postdoc researcher in Bayesian machine learning, decision-making under uncertainty

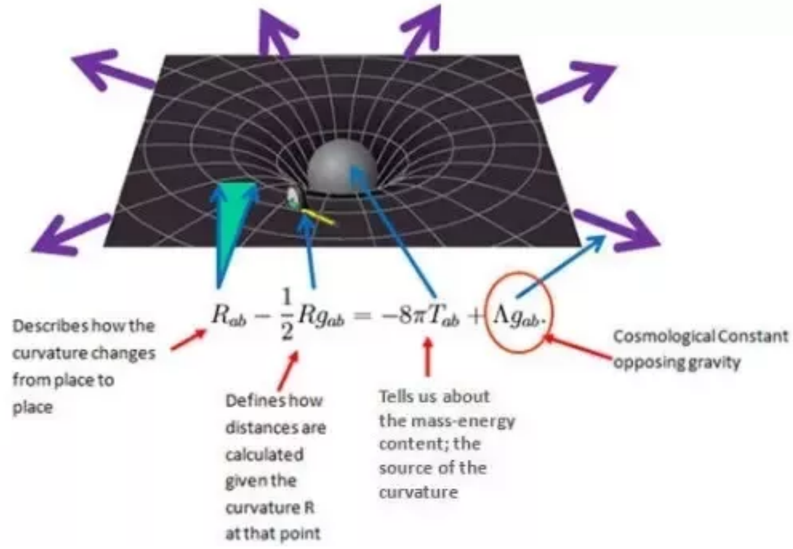
krisnguyen135.github.io

github.com/KrisNguyen135/Talks

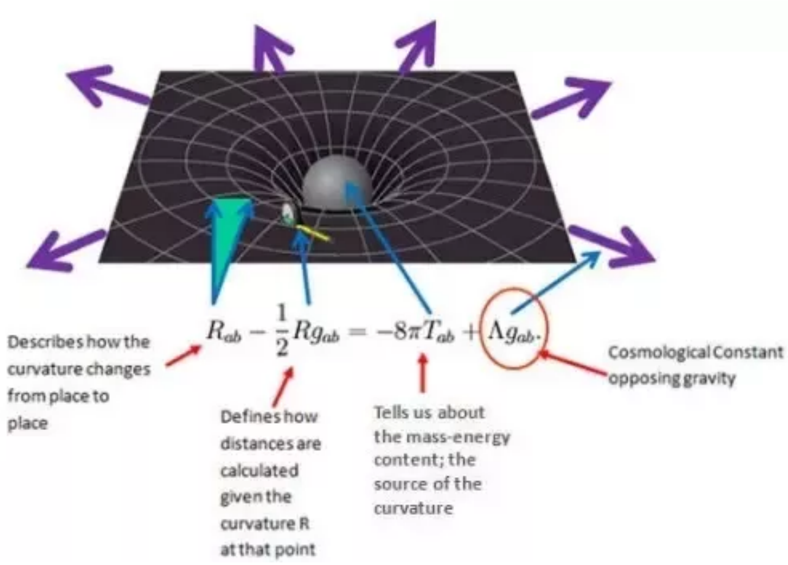
@the_subtrahend


B C G J K L N
P Q S V X Y Z
I U
Wheel
gamesbytim.com Settings Clear Board End Game
$650 * 2 = $1,300
WHEE O
ORT E
FUN & GAMES
Solve
Normal Round
3 Players
Jake
Carissa
Jamie
Round
$ 1,000
$ 950
$ 600
Totals
$ 0
$ 0
$ 0


B C G J K L N
P Q S V X Y Z
I U
Wheel
gamesbytim.com Settings Clear Board End Game
$650 * 2 = $1,300
WHEE O
ORT E
FUN & GAMES
Solve
Normal Round
3 Players
Jake
Carissa
Jamie
Round
$ 1,000
$ 950
$ 600
Totals
$ 0
$ 0
$ 0


$R_{ab} - \frac{1}{2}Rg_{ab} = -8\pi T_{ab} + \Lambda g_{ab}$
Describes how the curvature changes from place to place
Defines how distances are calculated given the curvature R at that point
Tells us about the mass-energy content; the source of the curvature
Cosmological Constant opposing gravity


B C G J K L N
P Q S V X Y Z
I U
Wheel
gamesbytim.com Settings Clear Board End Game
$650 * 2 = $1,300
WHEE O
ORT E
FUN & GAMES
Solve
Normal Round
3 Players
Jake
Carissa
Jamie
Round
Totals
$1,000
$950
$600
$0
$0
$0


$R_{ab} - \frac{1}{2}Rg_{ab} = -8\pi T_{ab} + \Lambda g_{ab}$
Describes how the curvature changes from place to place
Defines how distances are calculated given the curvature R at that point
Tells us about the mass-energy content; the source of the curvature
Cosmological Constant opposing gravity


Yes
No
Yes
No
Yes
No
Yes
No
occurred, and

problem

experiment

utility

“F”

+100

“Z”

-10

problem	experiment	utility
	Who did you vote for in 2020?	+100
	Were you born after 1800?	-10

What do these examples have in common?

What do these examples have in common?

- We query an oracle (a human, some process of interest, etc.) for labels 👍

What do these examples have in common?

- We query an oracle (a human, some process of interest, etc.) for labels 👍
  - but queries are “expensive” (time, money, environment, user experience, etc.) 👎

What do these examples have in common?

- We query an oracle (a human, some process of interest, etc.) for labels 👍
  - but queries are “expensive” (time, money, environment, user experience, etc.) 👎
- Idea: target labels that yield the most information 🎯

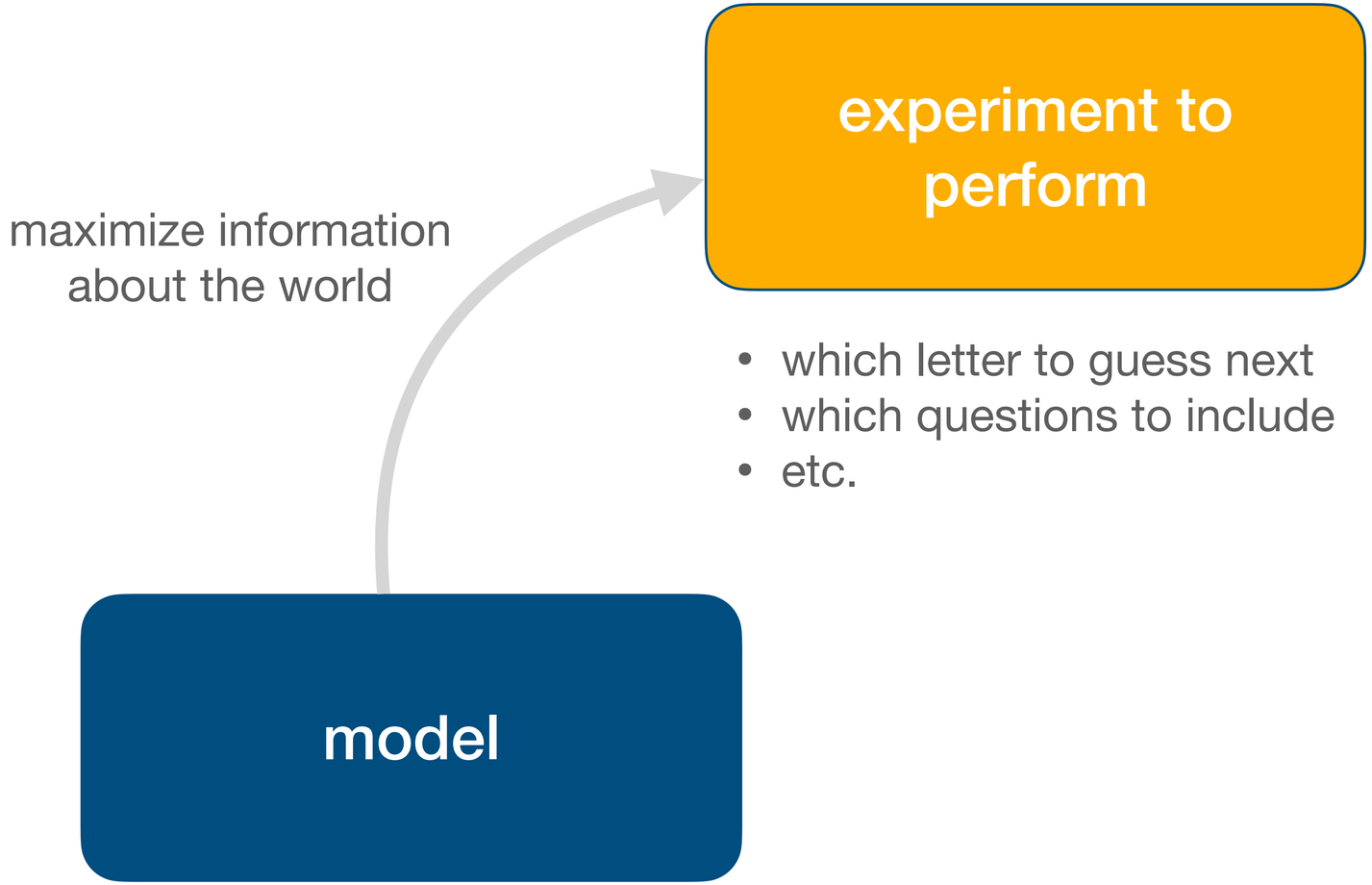
Bayesian experimental design

model

- what possible words it could be
- how people generally think
- etc.

Rainforth. *Bayesian Experimental Design and Active Learning*. 2022.

Bayesian experimental design

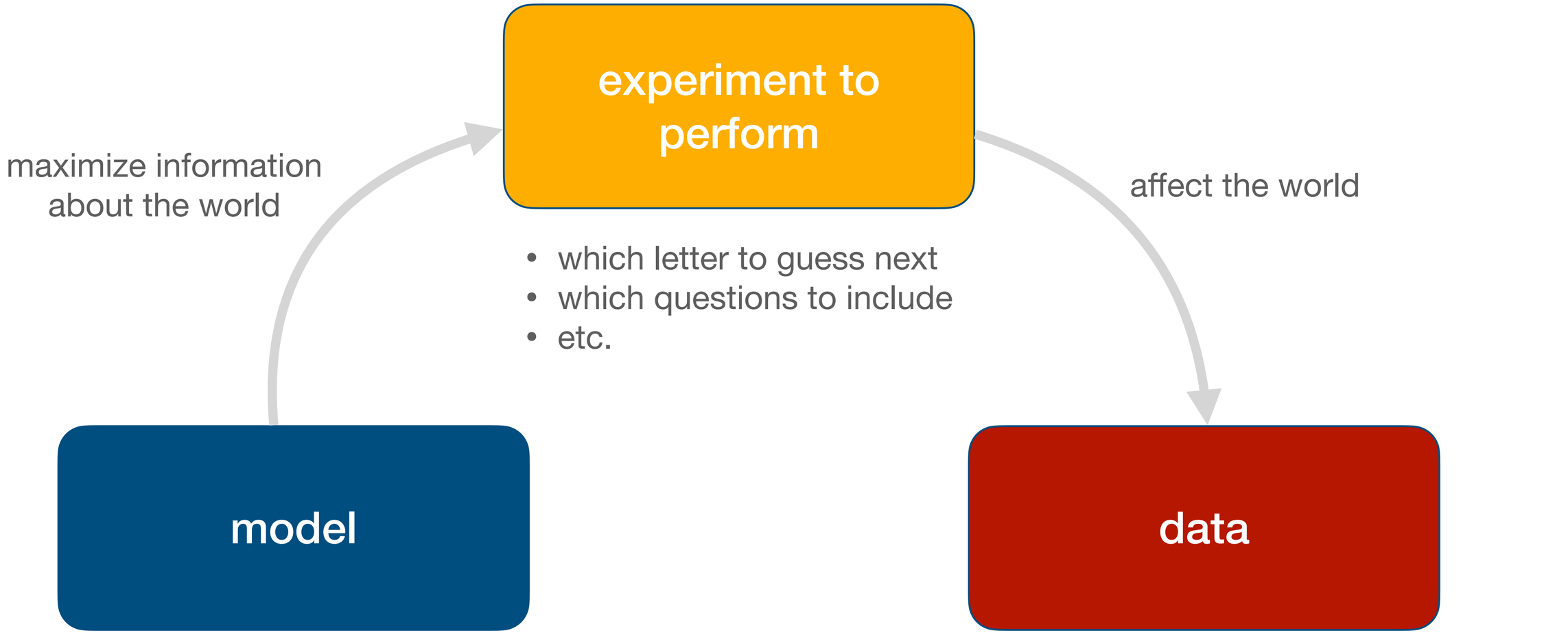


- which letter to guess next
- which questions to include
- etc.

- what possible words it could be
- how people generally think
- etc.

Rainforth. *Bayesian Experimental Design and Active Learning*. 2022.

Bayesian experimental design



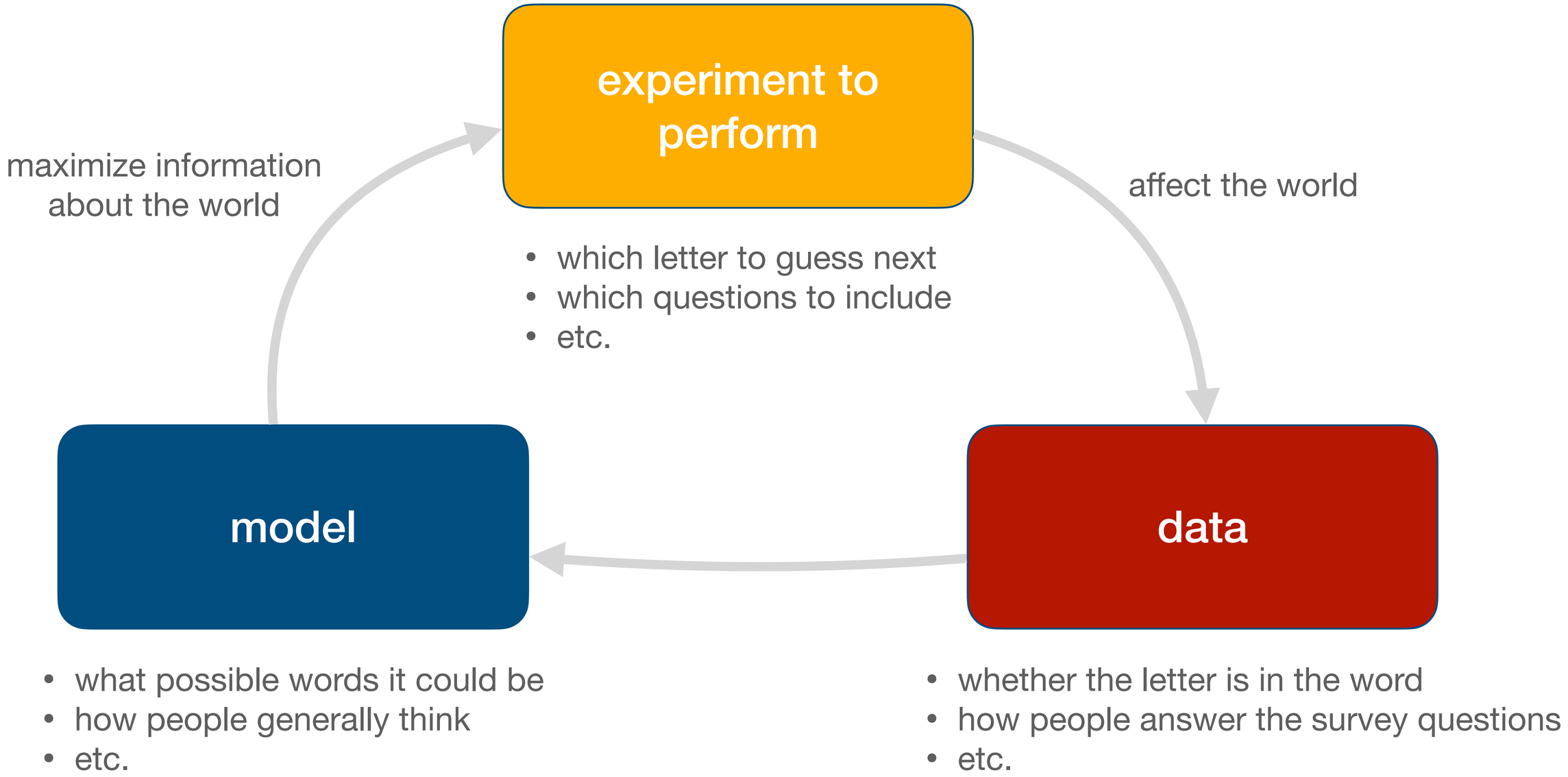
- which letter to guess next
- which questions to include
- etc.

- what possible words it could be
- how people generally think
- etc.

- whether the letter is in the word
- how people answer the survey questions
- etc.

Rainforth. *Bayesian Experimental Design and Active Learning*. 2022.

Bayesian experimental design



- which letter to guess next
- which questions to include
- etc.

- whether the letter is in the word
- how people answer the survey questions
- etc.

- what possible words it could be
- how people generally think
- etc.

Rainforth. *Bayesian Experimental Design and Active Learning*. 2022.

Let’s talk specifics: The (Bayesian) model

Let’s talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^{n}$

Let's talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^n$

1. Place a prior $p(\theta)$ on θ

Let's talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^n$

1. Place a prior $p(\theta)$ on θ
2. Assume an observation model $y = f(x; \theta)$

Let's talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^n$

1. Place a prior $p(\theta)$ on θ
2. Assume an observation model $y = f(x; \theta)$
    - Derive likelihoods $p(y_i \mid x_i, \theta)$

Let's talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^n$

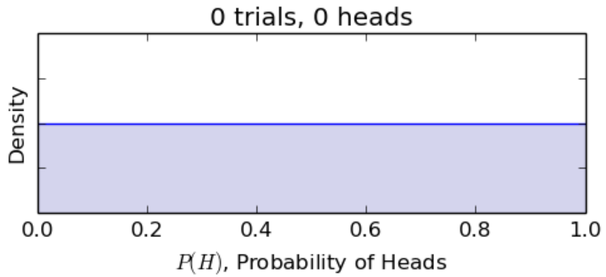
1. Place a prior $p(\theta)$ on θ
2. Assume an observation model $y = f(x; \theta)$
    - Derive likelihoods $p(y_i \mid x_i, \theta)$
3. Compute posterior distribution $p(\theta \mid D)$

Let's talk specifics: The (Bayesian) model

Given: unknown quantity of interest θ and observed data points $D = \{x_i, y_i\}_{i=1}^n$

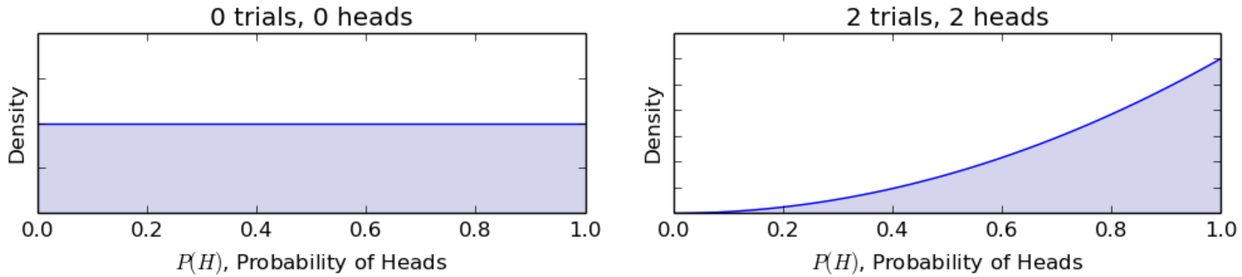
1. Place a prior $p(\theta)$ on θ
2. Assume an observation model $y = f(x; \theta)$
    - Derive likelihoods $p(y_i \mid x_i, \theta)$
3. Compute posterior distribution $p(\theta \mid D)$
    - probabilistically belief about θ conditioned on data

Bayesian model updating: An example



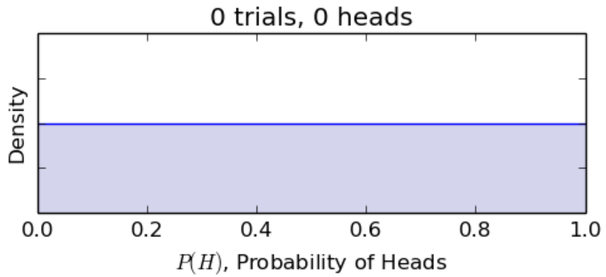
Bayesian Statistics: A Beginner's Guide, Quantstart

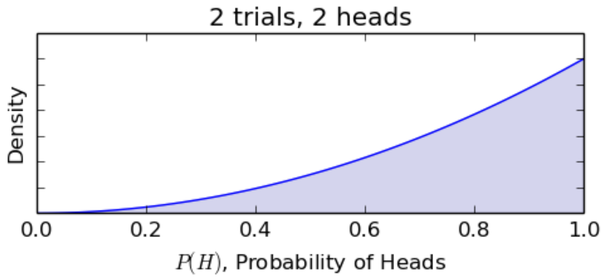
Bayesian model updating: An example

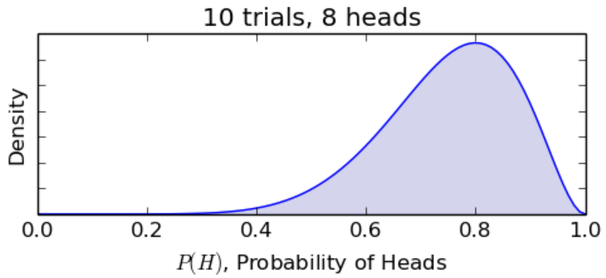


Bayesian Statistics: A Beginner's Guide, Quantstart

Bayesian model updating: An example

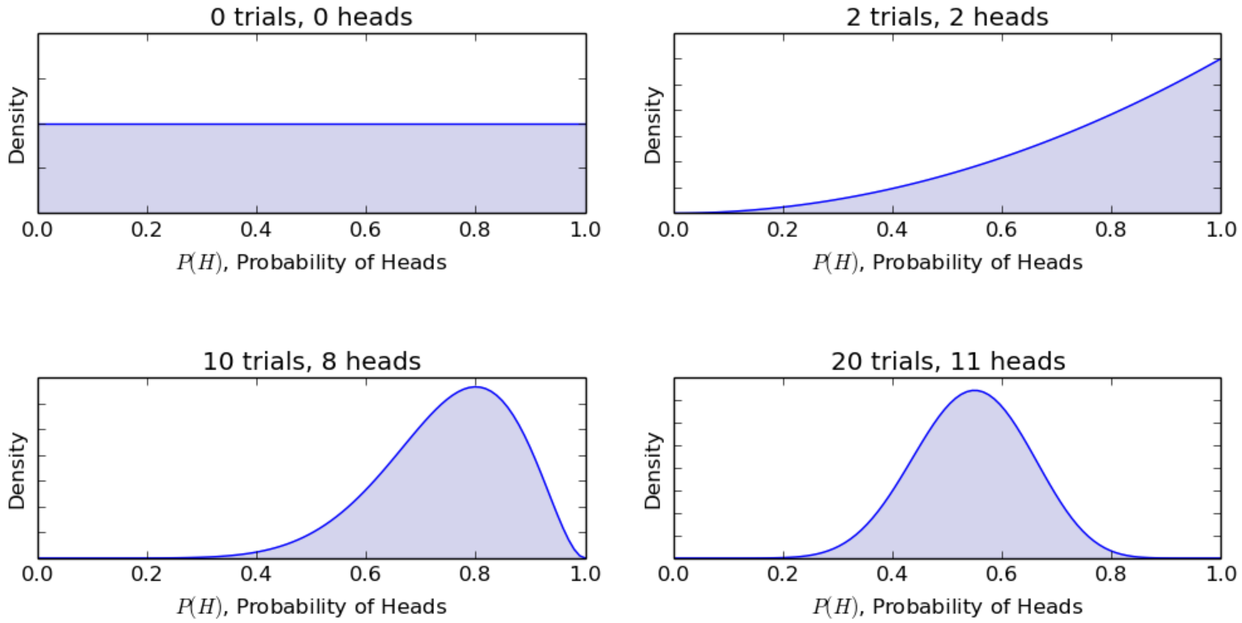






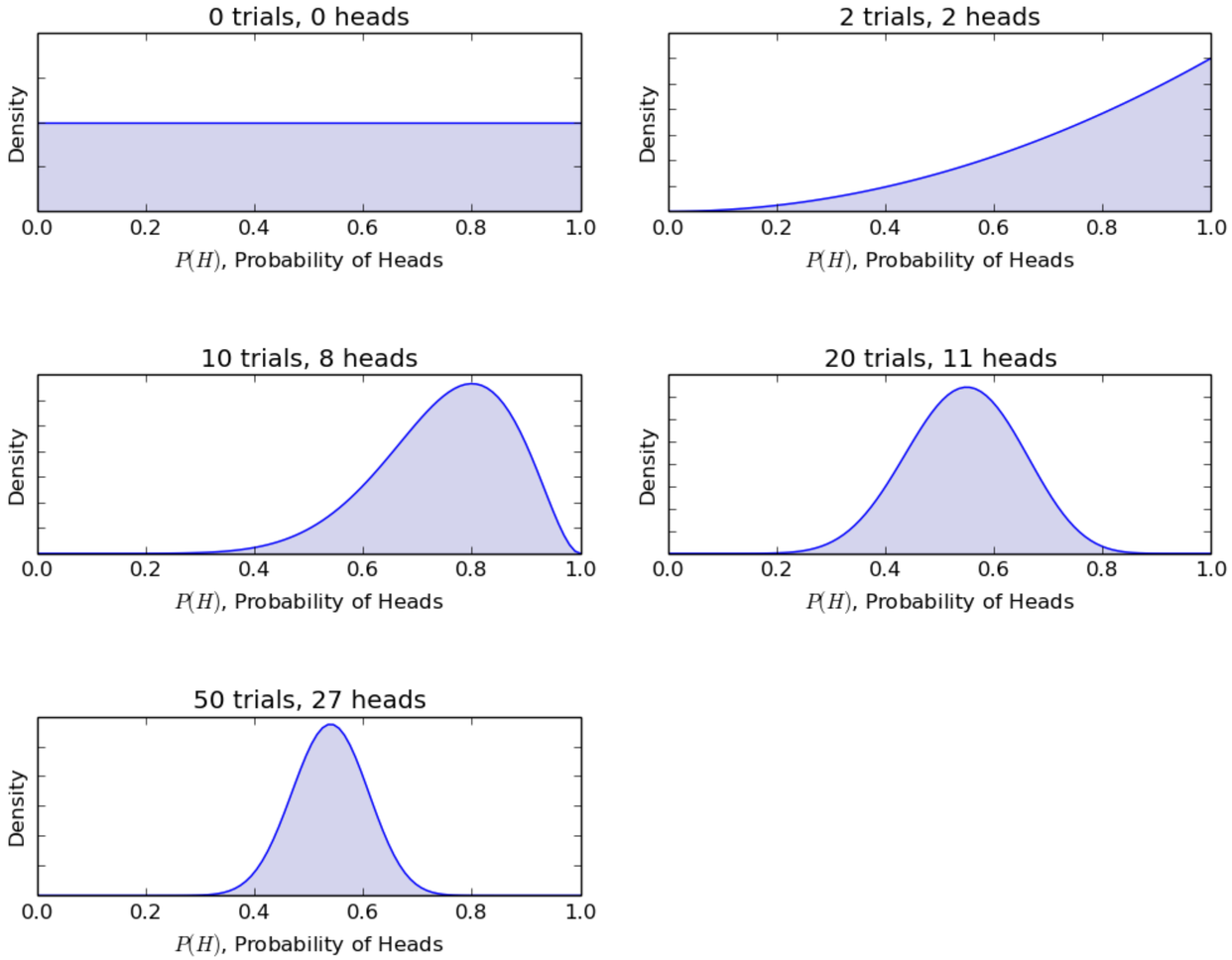
Bayesian Statistics: A Beginner's Guide, Quantstart

Bayesian model updating: An example



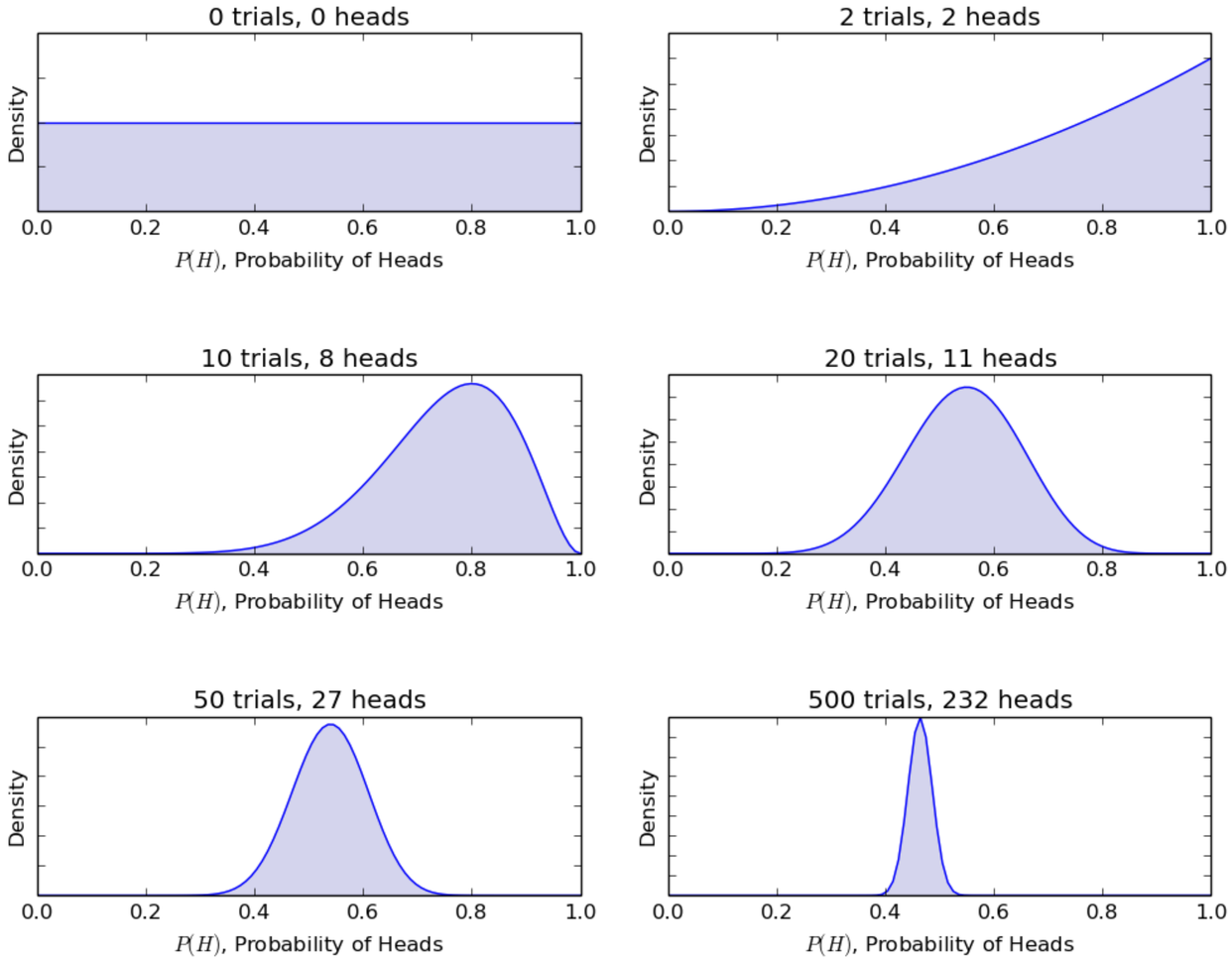
Bayesian Statistics: A Beginner's Guide, Quantstart

Bayesian model updating: An example



Bayesian Statistics: A Beginner's Guide, Quantstart

Bayesian model updating: An example



Bayesian Statistics: A Beginner's Guide, Quantstart

Let’s talk specifics: Maximizing information

Let’s talk specifics: Maximizing information

What do we really want?

Let’s talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ

Let’s talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ
- to decrease the uncertainty in our belief about θ

Let’s talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ
- to decrease the uncertainty in our belief about θ
  - Shannon entropy $H(\theta) = -\sum_i p(\theta_i) \log p(\theta_i)$

Let’s talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ
- to decrease the uncertainty in our belief about θ
  - Shannon entropy $H(\theta) = -\sum_i p(\theta_i) \log p(\theta_i)$
- but $p(\theta \mid D \cup \{x, y\})$ is not known until we observe $\{x, y\}$

Let’s talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ
- to decrease the uncertainty in our belief about θ
  - Shannon entropy $H(\theta) = -\sum_i p(\theta_i) \log p(\theta_i)$
- but $p(\theta \mid D \cup \{x, y\})$ is not known until we observe $\{x, y\}$
  - settle for the **expected value**

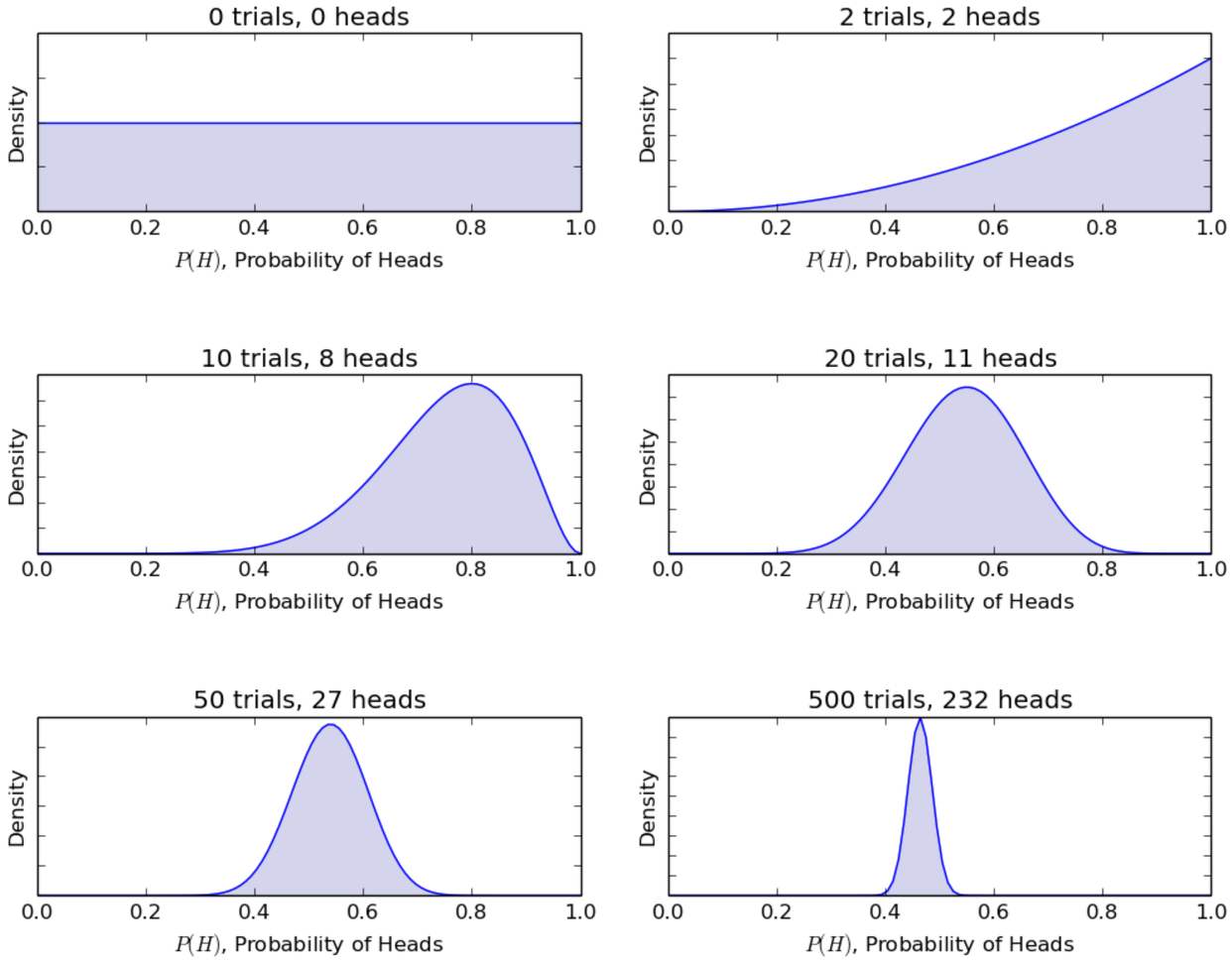
Let's talk specifics: Maximizing information

What do we really want?

- to know as much as possible about θ
- to decrease the uncertainty in our belief about θ
  - Shannon entropy $H(\theta) = -\sum_i p(\theta_i) \log p(\theta_i)$
- but $p(\theta \mid D \cup \{x, y\})$ is not known until we observe $\{x, y\}$
  - settle for the **expected value**

$$x_* = \arg\min_x \mathbb{E}_y \Big[H\big[\theta \mid D \cup \{x, y\}\big] \Big]$$

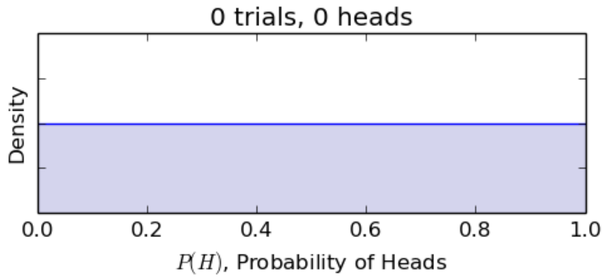
Measuring uncertainty: An example

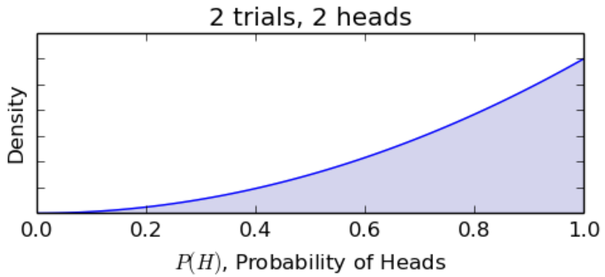


Bayesian Statistics: A Beginner's Guide, Quantstart

Measuring uncertainty: An example

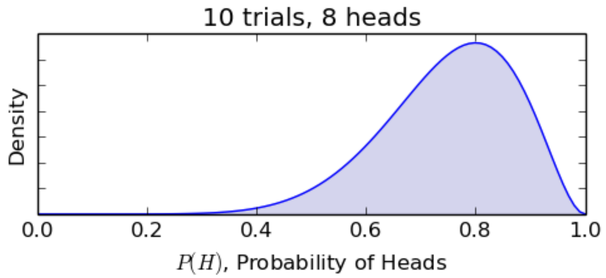
$H = 0$

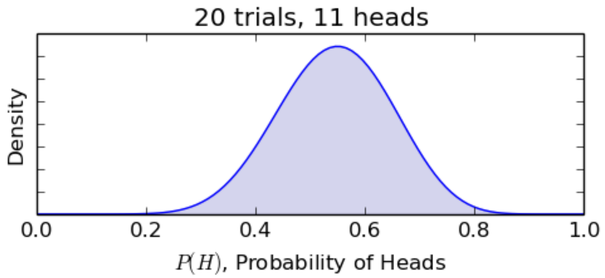




$H \approx -0.4$

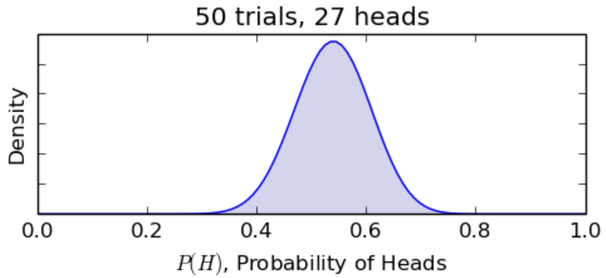
$H \approx -0.6$

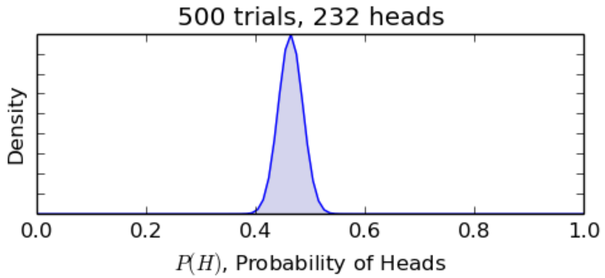




$H \approx -0.9$

$H \approx -1.3$





$H \approx -2.4$

Bayesian Statistics: A Beginner's Guide, Quantstart

Do-it-yourself experimental design

Given a sorted array of numbers
$X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$
at unknown index, find this index by
comparing z against elements in X.

Do-it-yourself experimental design

Given a sorted array of numbers
$X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$
at unknown index, find this index by
comparing z against elements in X.

- sound suspiciously familiar?

Do-it-yourself experimental design

Given a sorted array of numbers
$X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$
at unknown index, find this index by
comparing z against elements in X.

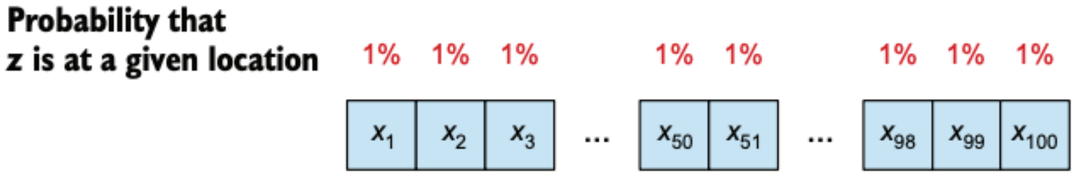
- sound suspiciously familiar?

- place a uniform prior on the index of z

Do-it-yourself experimental design

Given a sorted array of numbers
$X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$
at unknown index, find this index by
comparing z against elements in X.

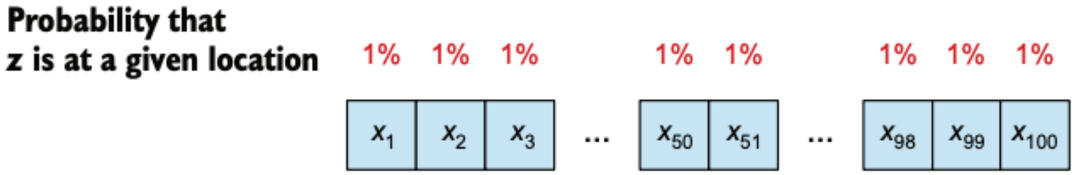
- sound suspiciously familiar?
- place a uniform prior on the index of z



Do-it-yourself experimental design

Given a sorted array of numbers
$X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$
at unknown index, find this index by
comparing z against elements in X.

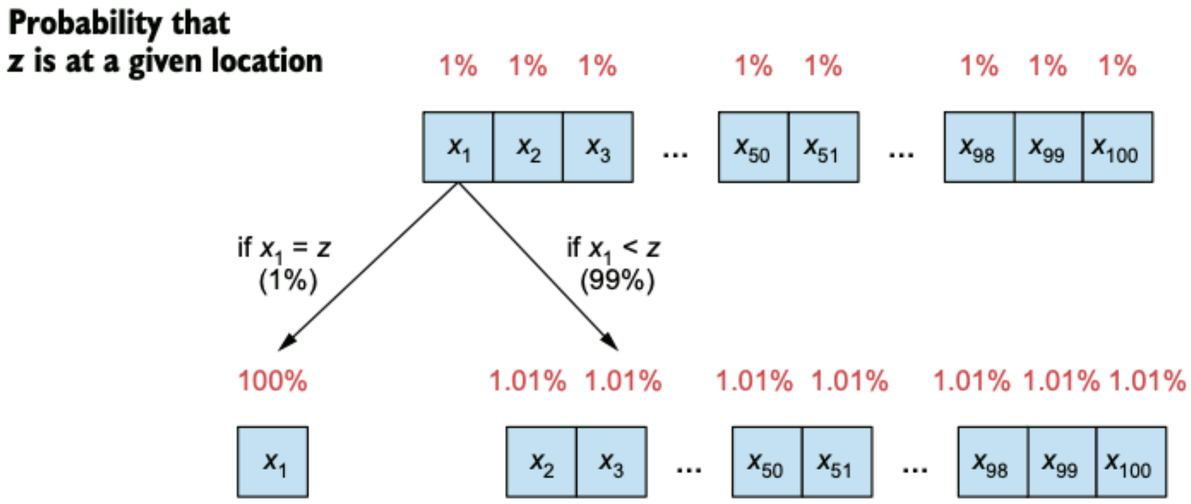
- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

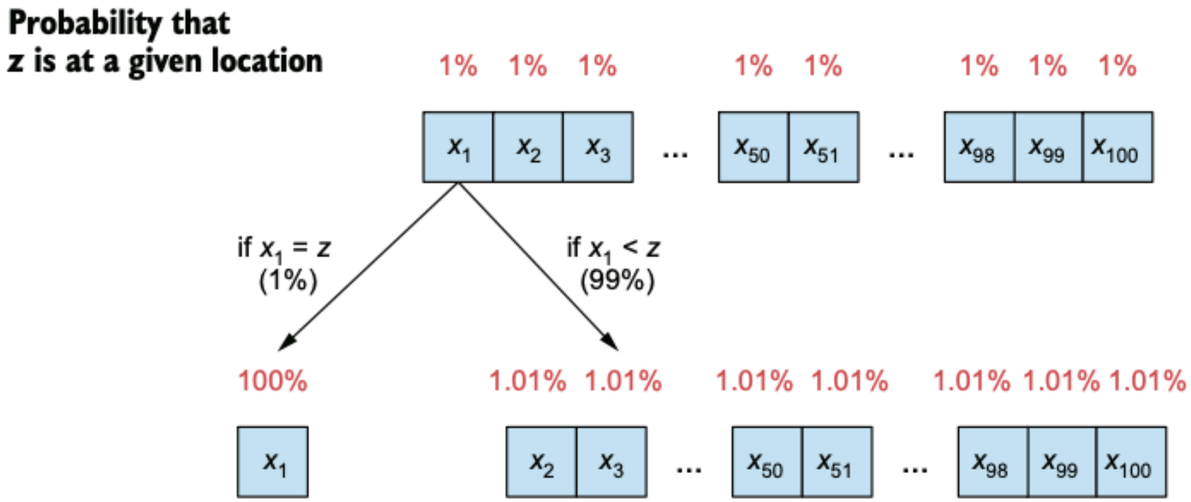
- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison

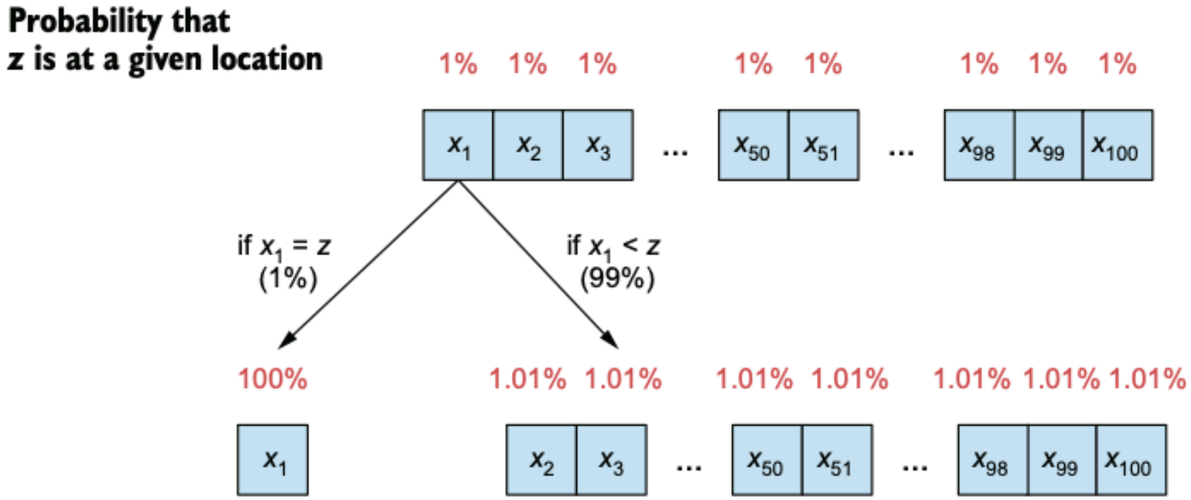


$$H = 0$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison

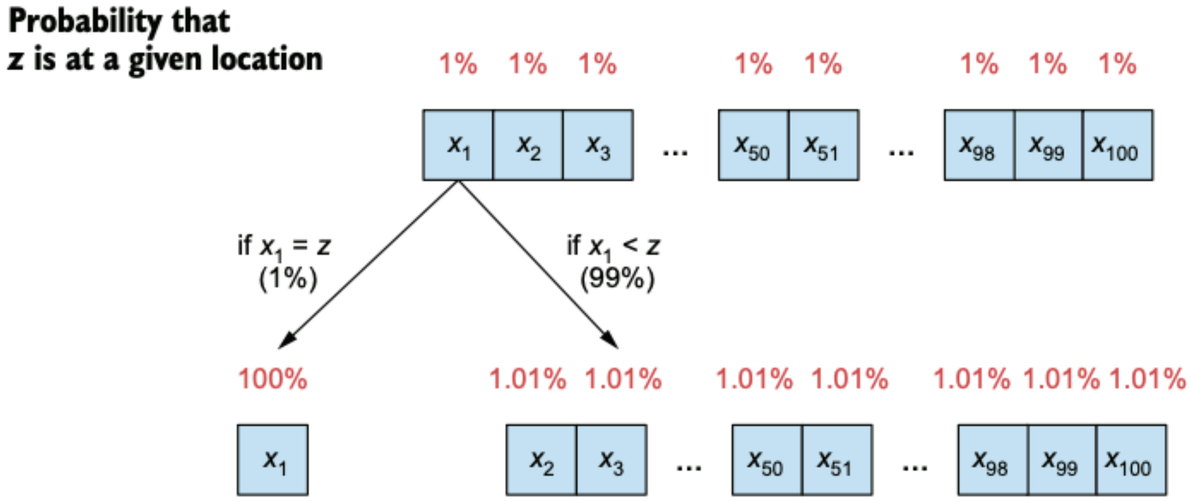


$$H = 0 \qquad H = -\sum_{i=2}^{100} 1.01\% \ \log(1.01\%) \approx 6.63$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



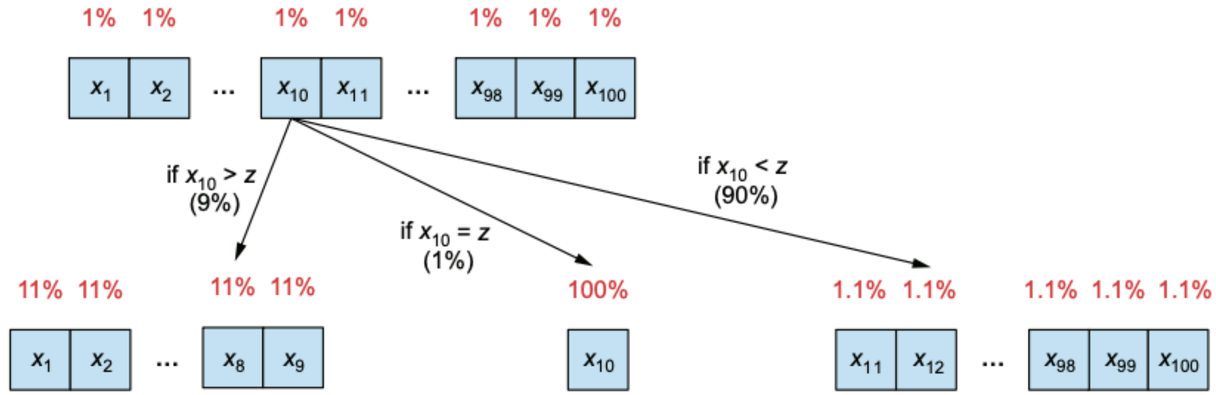
$$H = 0 \qquad H = -\sum_{i=2}^{100} 1.01\% \ \log(1.01\%) \approx 6.63$$

$$\text{average } H = 0.01\ (0) + 0.99\ (6.63) = 6.56$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

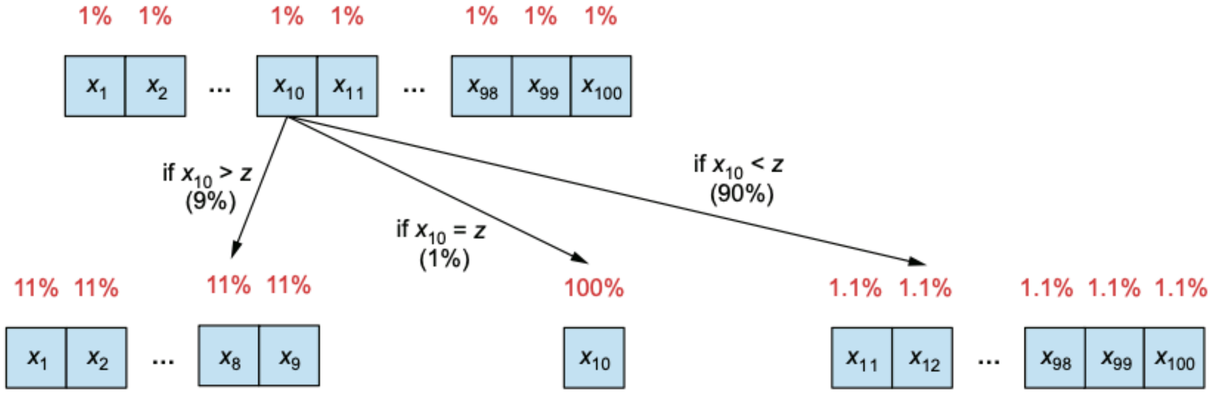
- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison

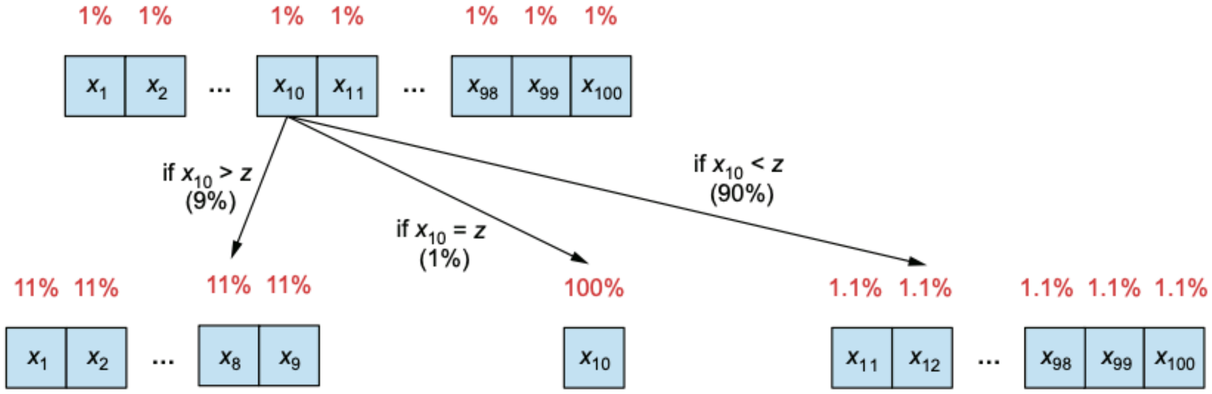


$$H = -\sum_{i=1}^{9} 11\% \ \log(11\%) \approx 3.17$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison

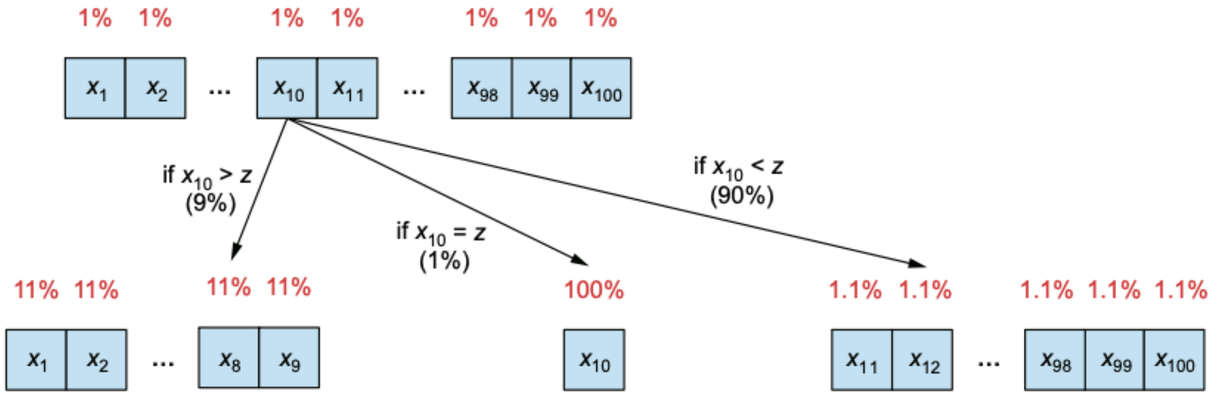


$$H = -\sum_{i=1}^{9} 11\% \ \log(11\%) \approx 3.17 \qquad H = 0$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison

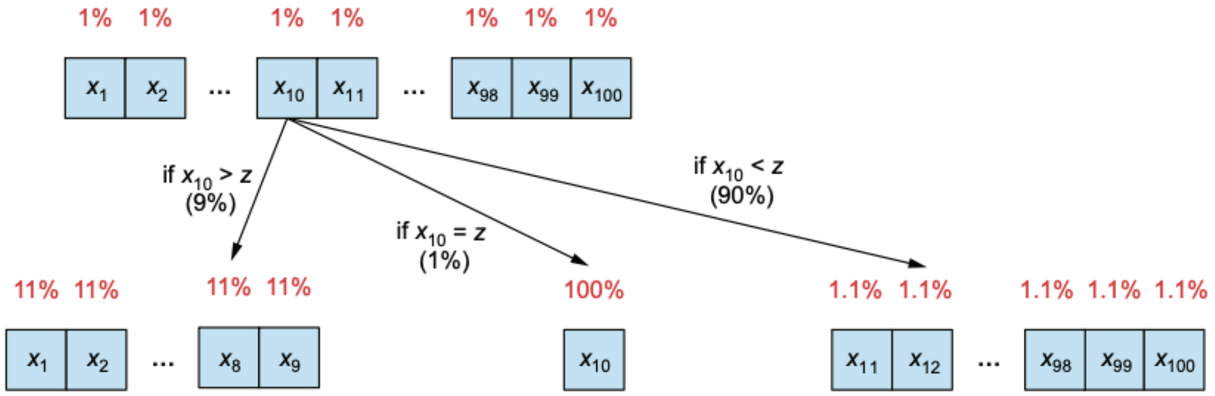


$$H = -\sum_{i=1}^{9} 11\% \ \log(11\%) \approx 3.17 \qquad H = 0 \qquad H = -\sum_{i=11}^{100} 1.1\% \ \log(1.1\%) \approx 6.49$$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



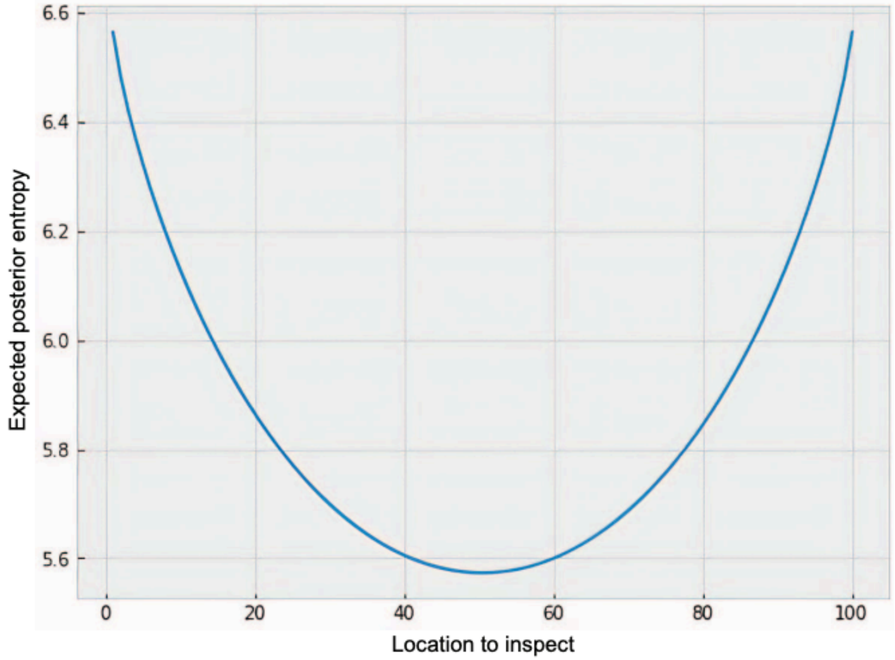
$$H = -\sum_{i=1}^{9} 11\% \ \log(11\%) \approx 3.17 \qquad H = 0 \qquad H = -\sum_{i=11}^{100} 1.1\% \ \log(1.1\%) \approx 6.49$$

average $H = 0.09\ (3.17) + 0.01\ (0) + 0.9\ (6.49) = 6.13$

Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

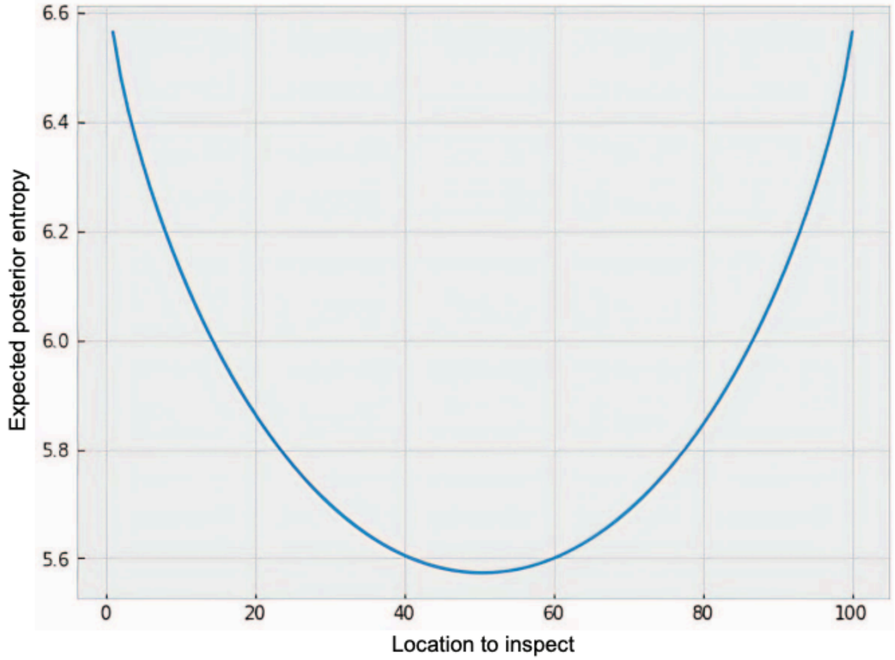
- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison



Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison
- pick the comparison with the lowest expected entropy

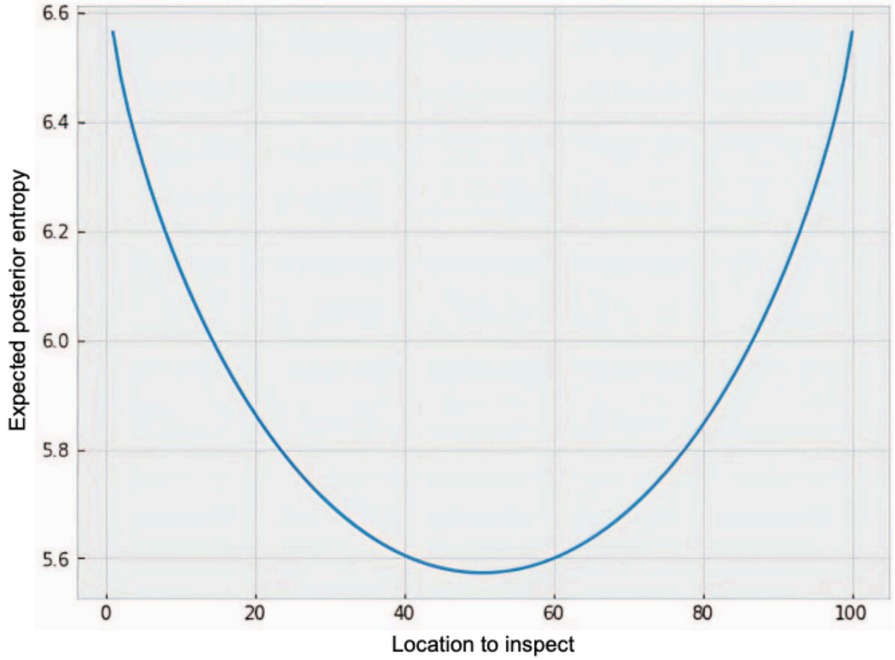


Do-it-yourself experimental design

Given a sorted array of numbers $X = \{x_1, x_2, \ldots, x_{100}\}$ and a target $z \in X$ at unknown index, find this index by comparing z against elements in X.

- sound suspiciously familiar?
- place a uniform prior on the index of z
- compute the expected entropy after each possible comparison
- pick the comparison with the lowest expected entropy

It's binary search!!



Bayesian experimental design in action

Bayesian experimental design in action

- **model**: people *randomly* become infected from time 0 to time T at some unknown transmission rate τ

Bayesian experimental design in action

- **model**: people *randomly* become infected from time 0 to time T at some unknown transmission rate τ

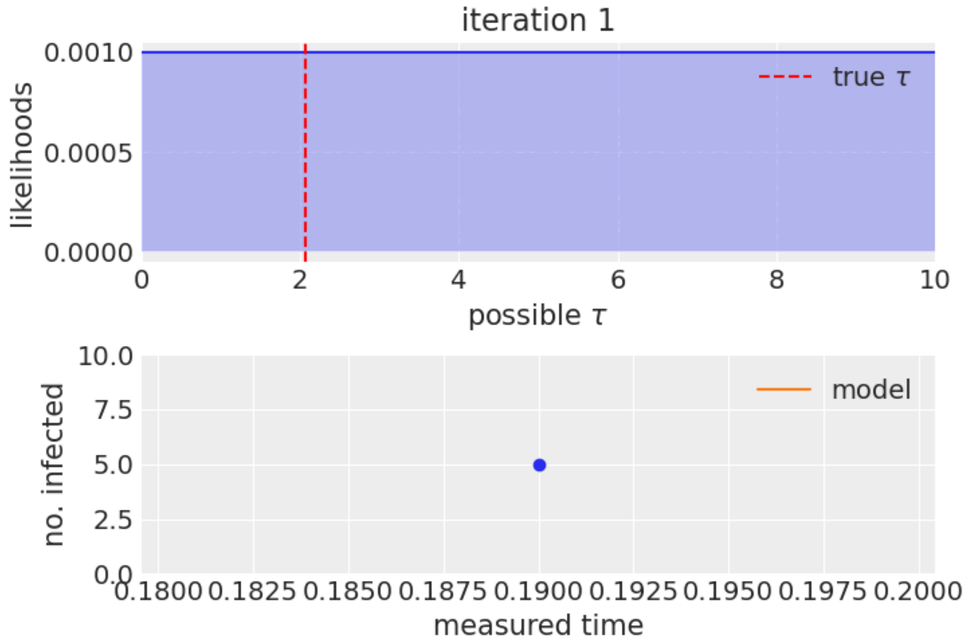
- **experiment**: time t to inspect how many people are infected

Bayesian experimental design in action

- **model**: people *randomly* become infected from time 0 to time T at some unknown transmission rate τ
- **experiment**: time t to inspect how many people are infected
- **data**: *roughly* how many people are infected at surveyed time t

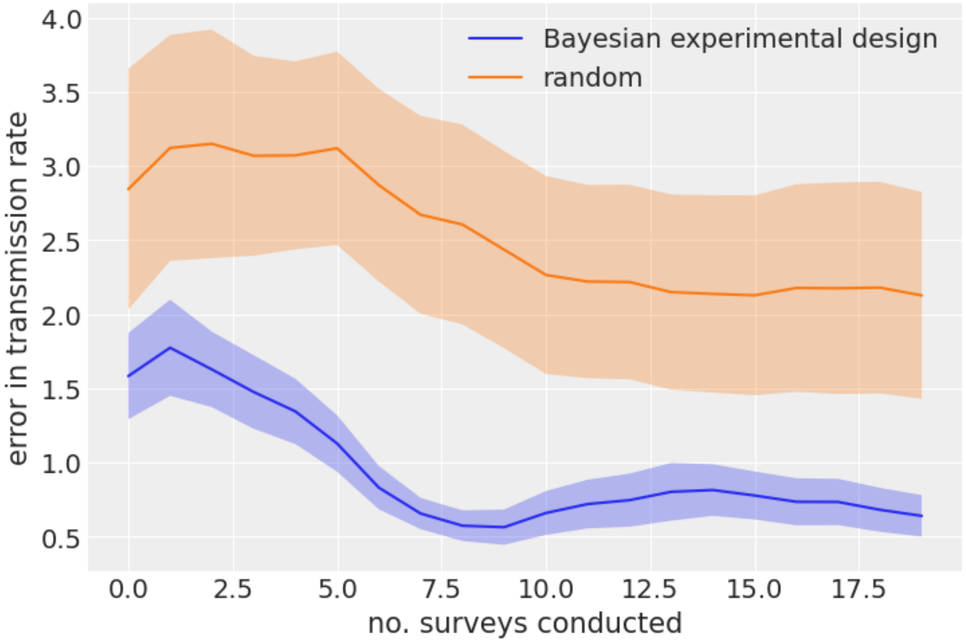
Bayesian experimental design in action

- **model**: people *randomly* become infected from time 0 to time T at some unknown transmission rate τ
- **experiment**: time t to inspect how many people are infected
- **data**: *roughly* how many people are infected at surveyed time t

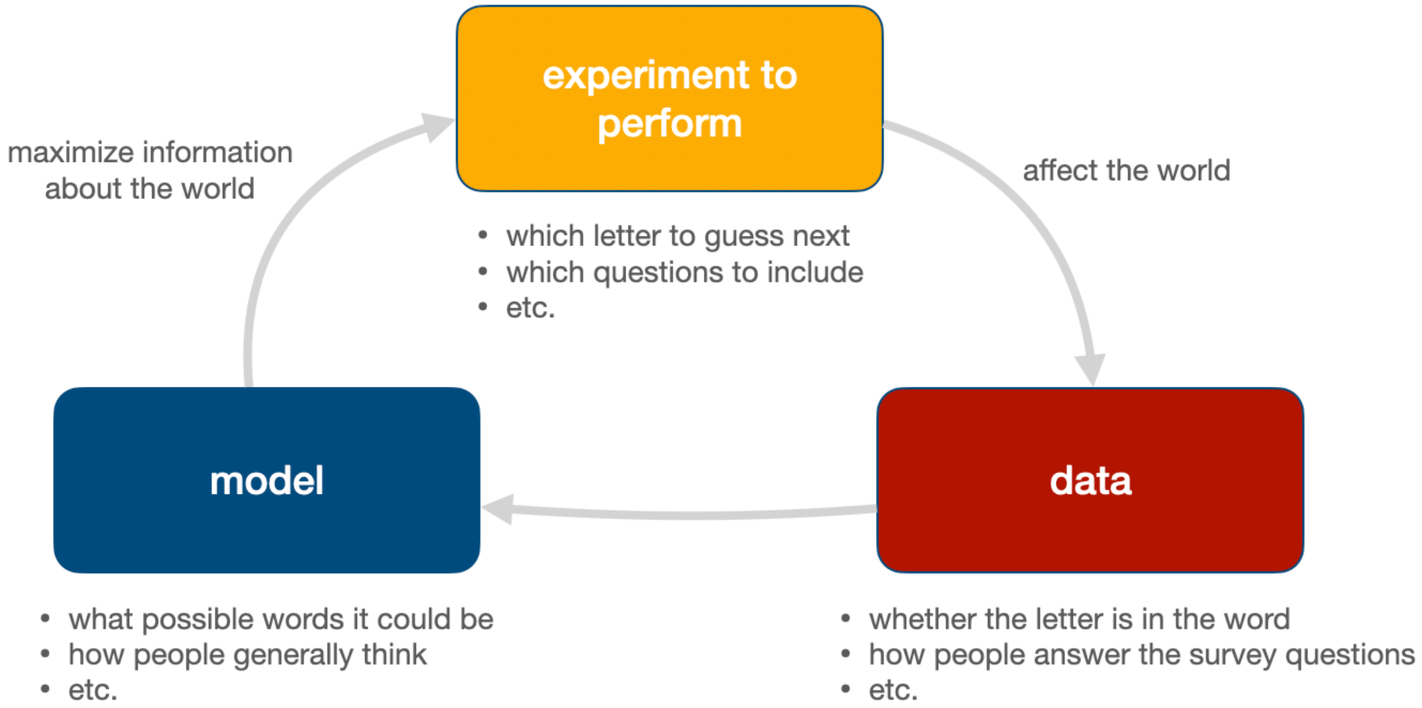


Bayesian experimental design in action

- **model**: people *randomly* become infected from time 0 to time T at some unknown transmission rate τ
- **experiment**: time t to inspect how many people are infected
- **data**: *roughly* how many people are infected at surveyed time t

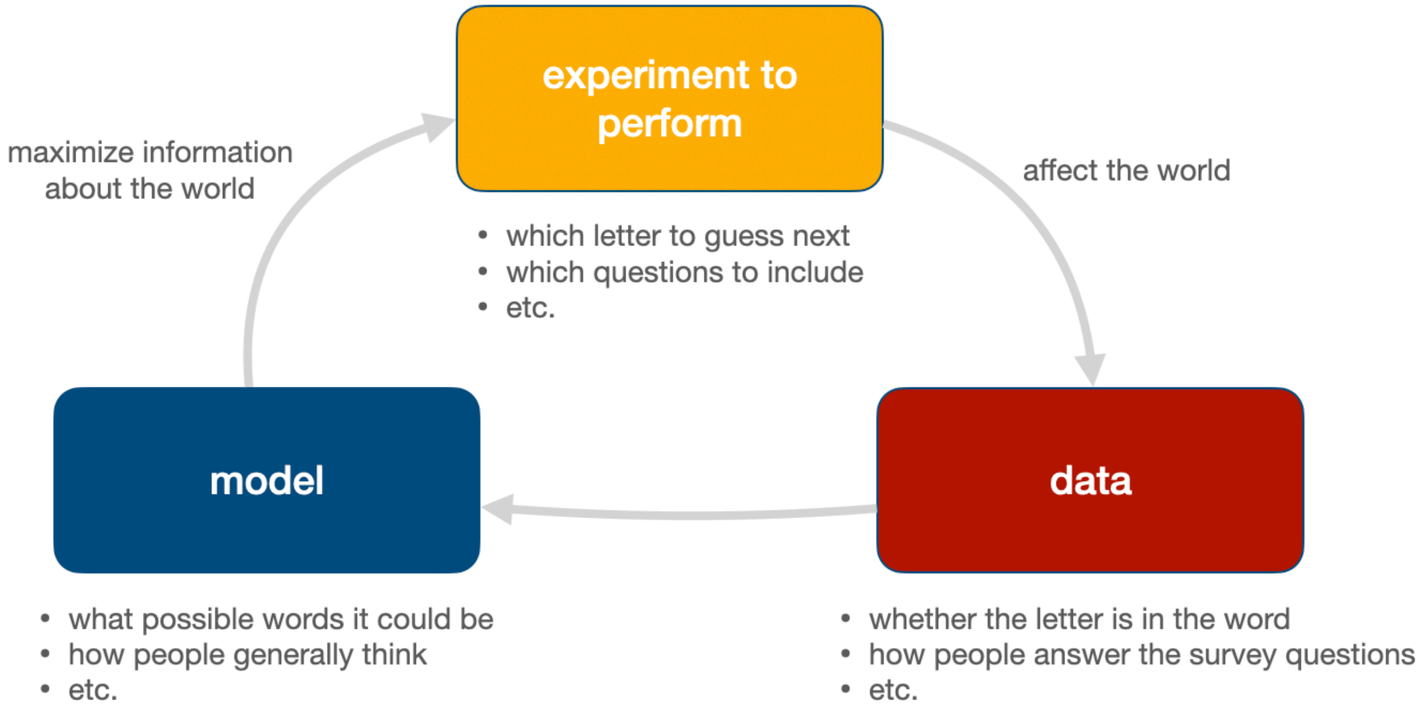


Completing the experimental design loop

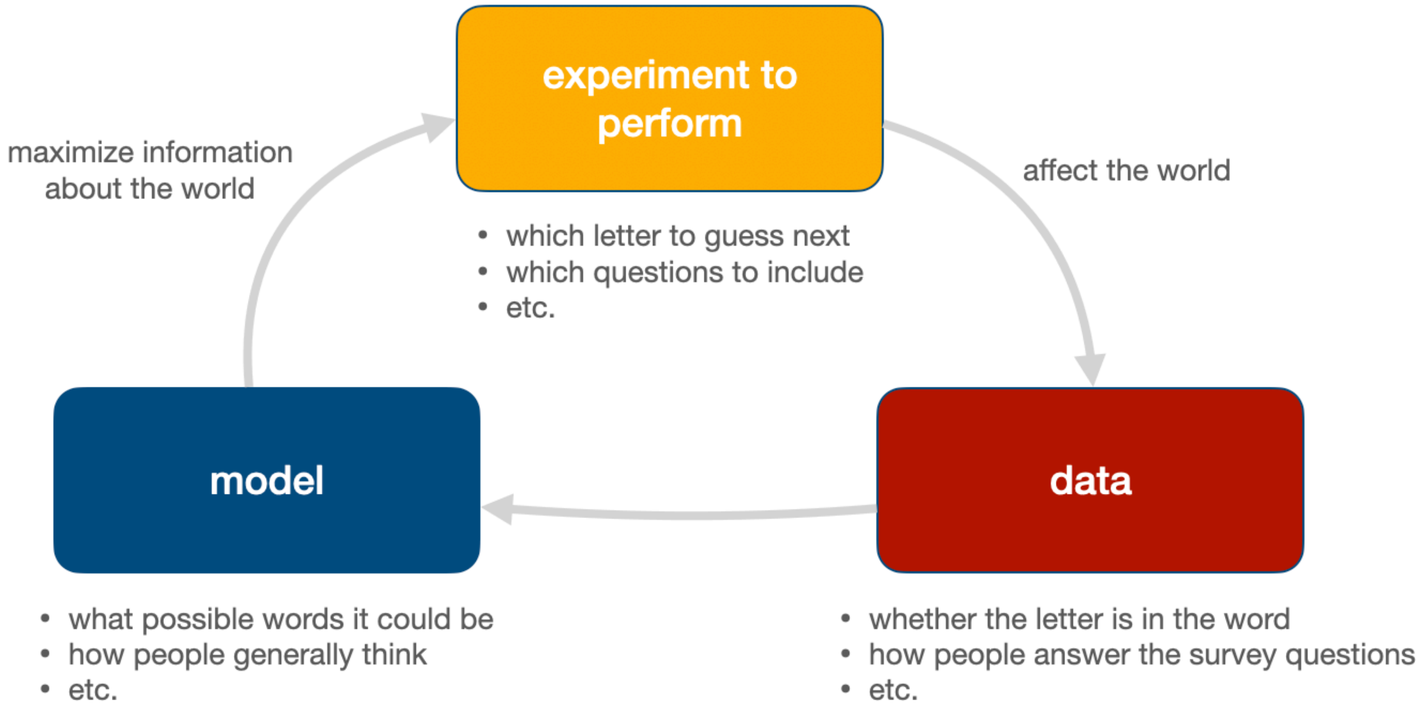


Completing the experimental design loop

A detective's search for clues



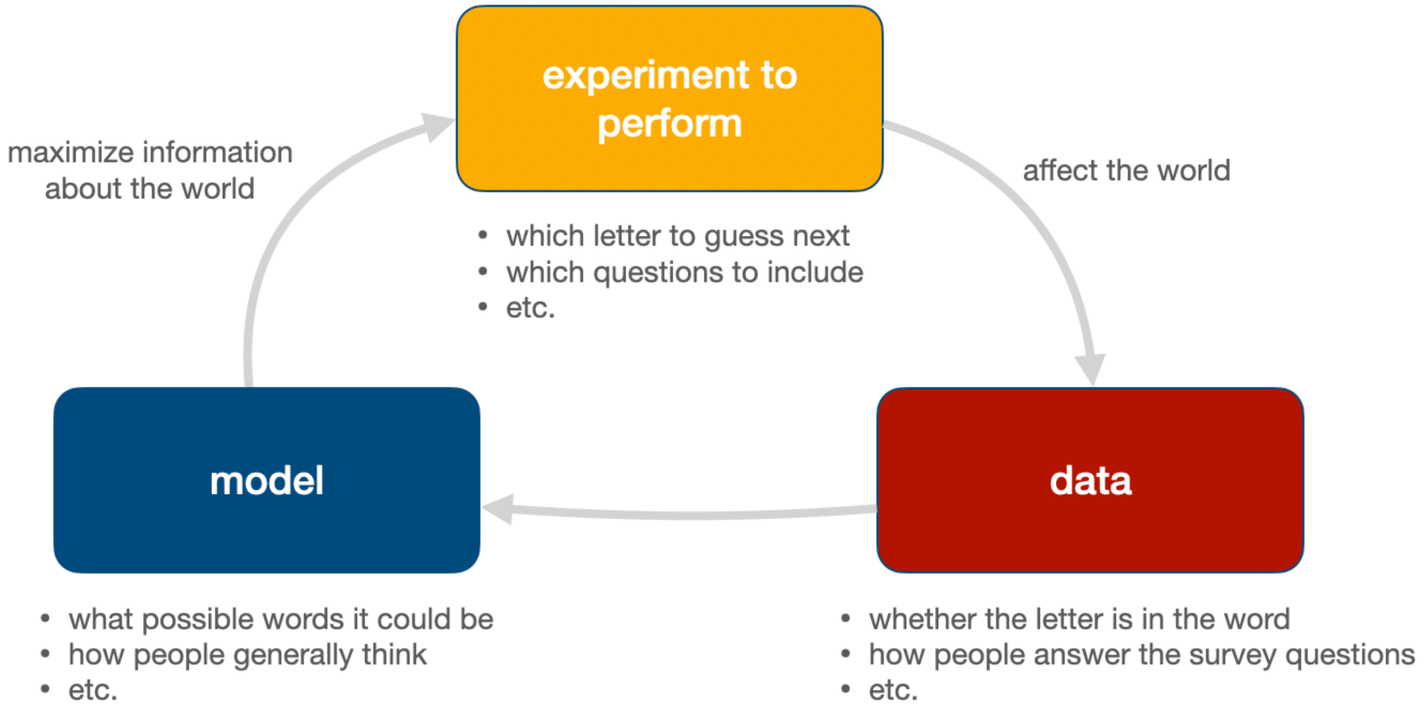
Completing the experimental design loop



A detective's search for clues

- experimental design

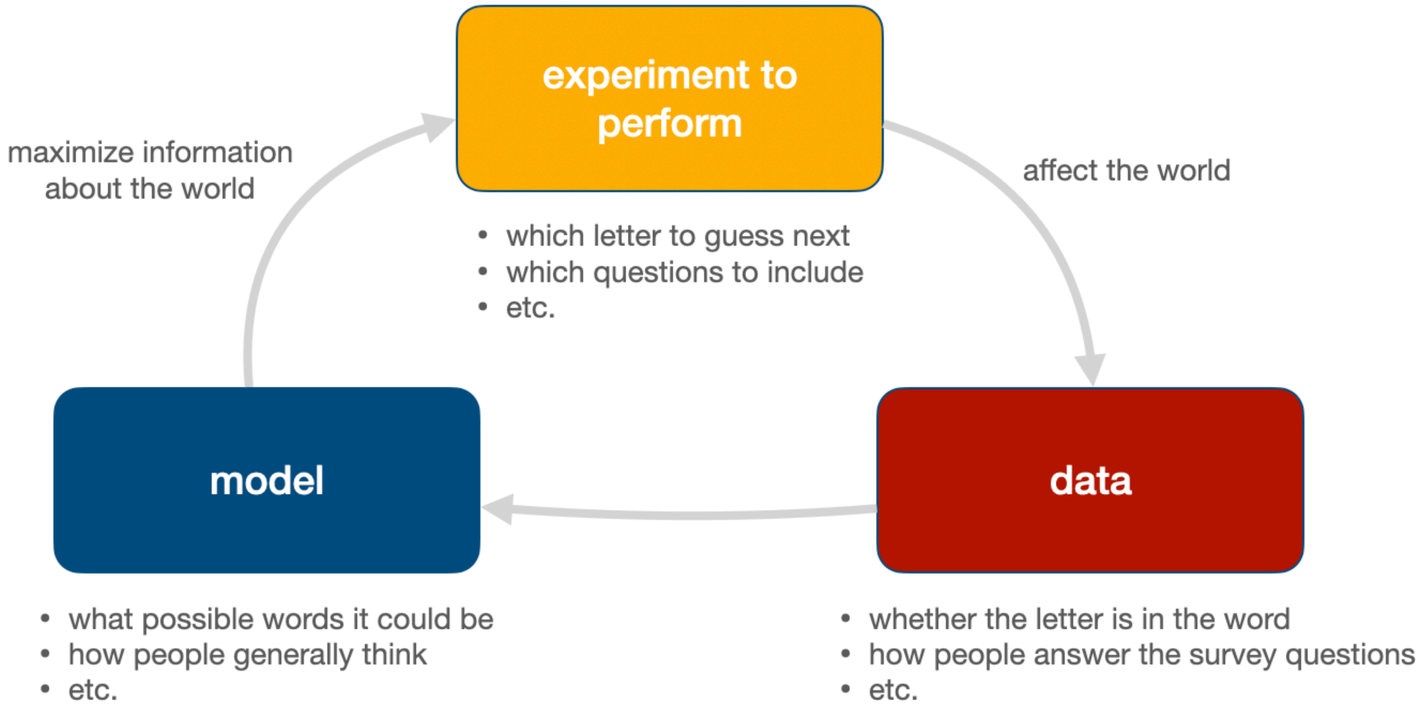
Completing the experimental design loop



A detective's search for clues

- experimental design
- Bayesian experimental design

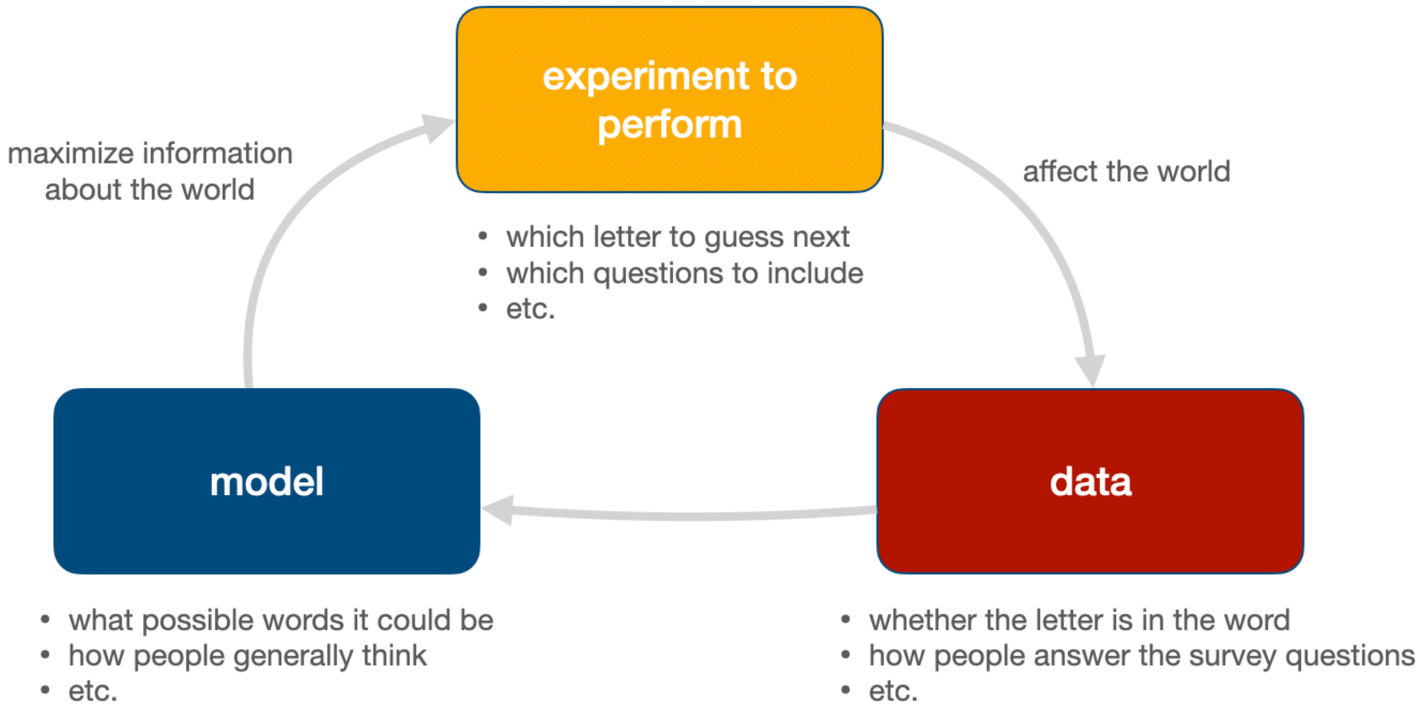
Completing the experimental design loop



A detective's search for clues

- experimental design
- Bayesian experimental design
- optimal experimental design

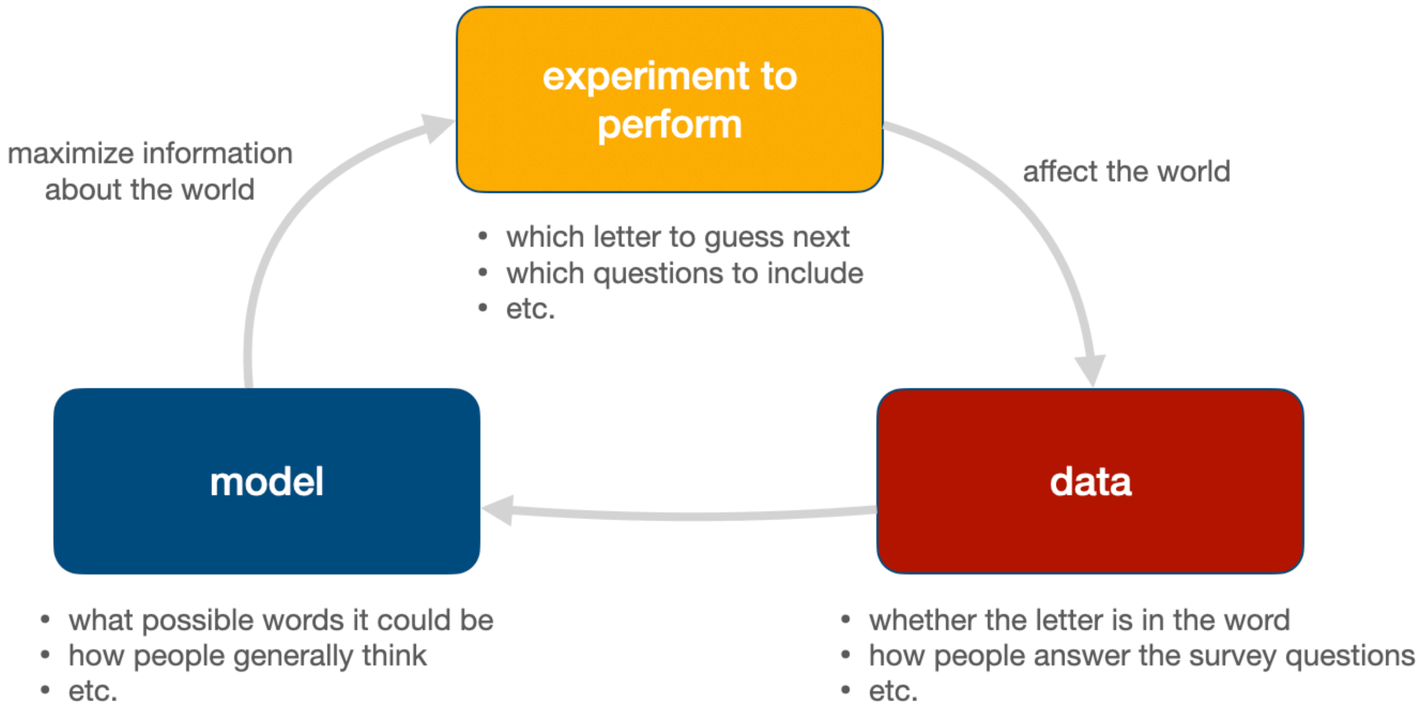
Completing the experimental design loop



A detective's search for clues

- experimental design
- Bayesian experimental design
- optimal experimental design
- design of experiments (DoE)

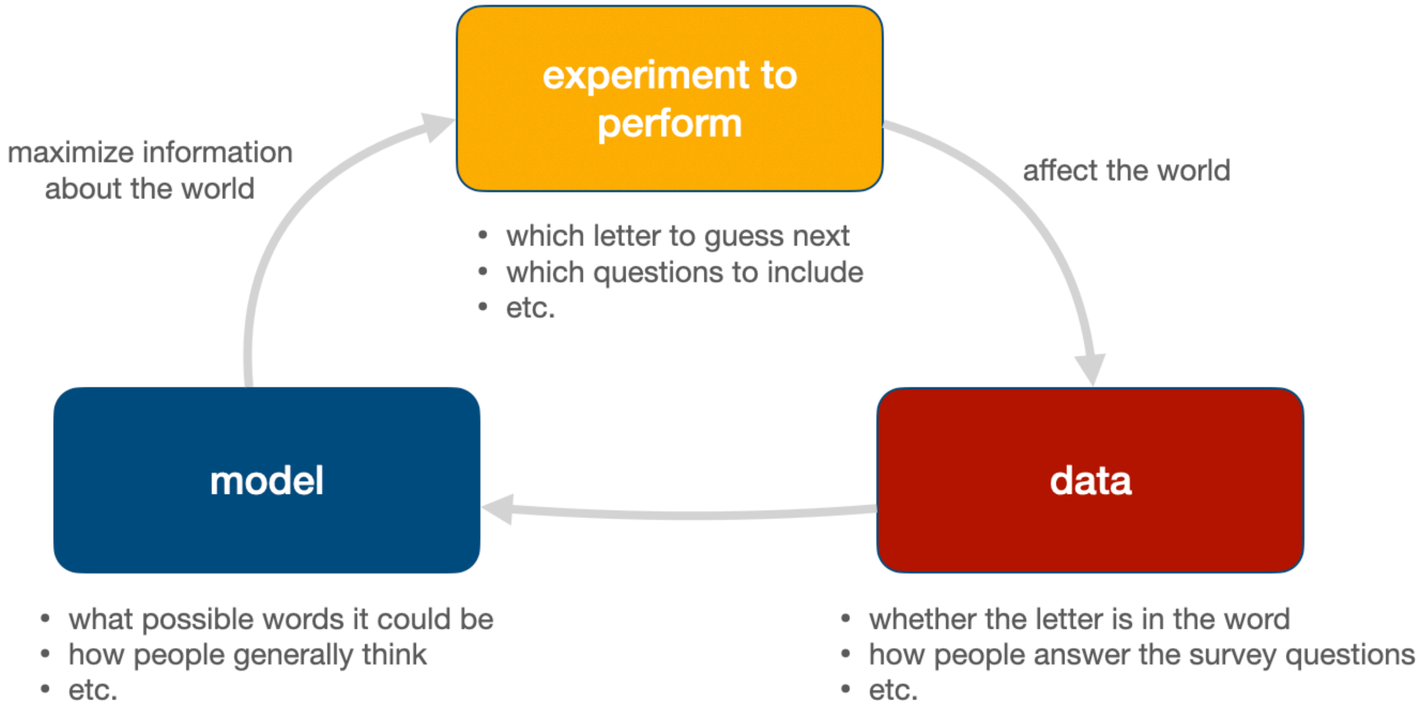
Completing the experimental design loop



A detective's search for clues

- experimental design
- Bayesian experimental design
- optimal experimental design
- design of experiments (DoE)
- active learning

Completing the experimental design loop



A detective's search for clues

- experimental design
- Bayesian experimental design
- optimal experimental design
- design of experiments (DoE)
- active learning
- adaptive experimentation
- etc.

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

Building your own Bayesian experimental design application

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

Building your own Bayesian experimental design application

- specify the observation model

```
def my_model_function(settings, parameters, constants):
    """Example model function

    The ``(settings, parameters, constants)`` argument structure is required
    Args:
        settings (tuple or tuple of array(s)): knob settings
        parameters (tuple of arrays, or tuple): parameter distribution sample(s)
        constants (tuple): infrequently changed values

    Returns: a noise-free model value
    """
    # Unpack the arguments.  See the "Specify ..." sections in the text.
    knob, = settings
    phase, delay = parameters
    temperature, = constants

    # This is where the model calculation goes.  It could be defined as a separate
    # function as suggested here, or the raw math expressions could go here.
    model_result = my_model_calculation(knob, phase, delay, temperature)

    return model_result
```

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

Building your own Bayesian experimental design application

- specify the observation model
- specify the search space

```
knob = np.linspace(0, 11, 111)
setting_values = (knob, )
```

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

Building your own Bayesian experimental design application

- specify the observation model
- specify the search space
- specify the prior

```
n_samples = 50000
phase = np.random.uniform(-np.pi/2, np.pi/2, n_samples)
```

Bayesian experimental design in Python

Python library: pages.nist.gov/optbayesexpt

- extensive documentation and examples

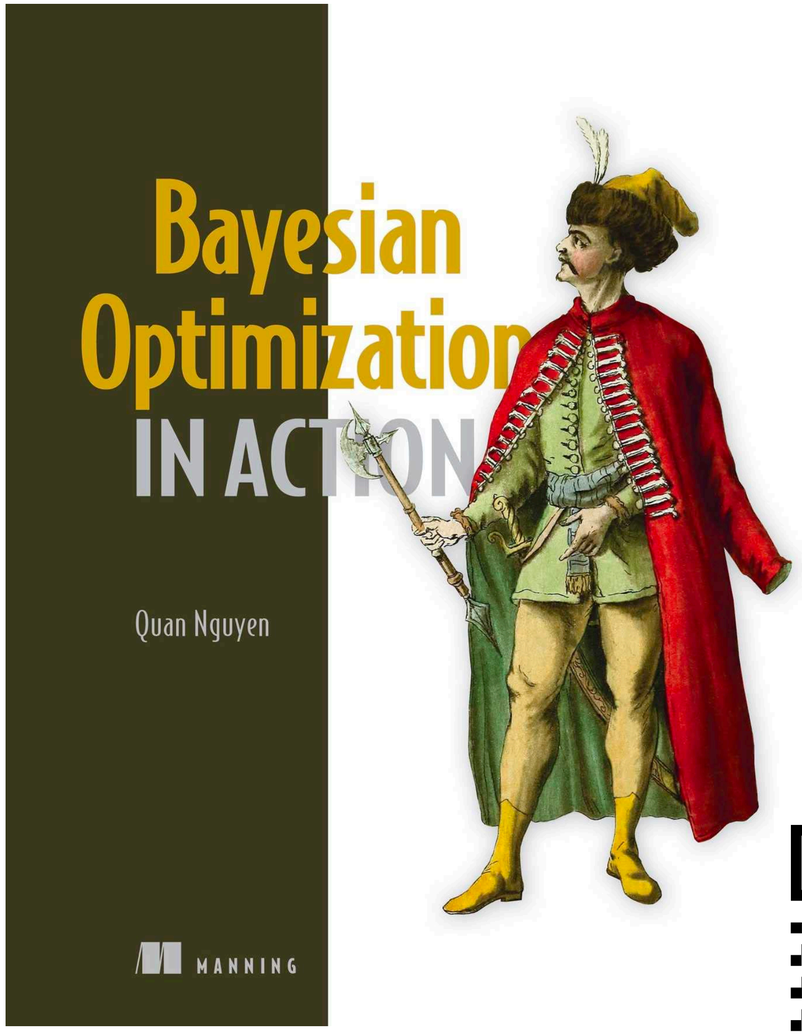
Building your own Bayesian experimental design application

- specify the observation model
- specify the search space
- specify the prior
- run!

```
while still_measuring:

    # (1) my_obe picks a single combination of settings - there's a choice of methods.
    # settings = my_obe.opt_setting()
    #    --  or --
    settings = my_obe.good_setting(pickiness=a_value_between_1_and_10)

    # The experiment makes a measurement using settings and returns a result
    # (Machine goes "bing!")
    # measurement results are reported as tuples
    measurement = (actual_settings, result, uncertainty)
    # (2) report the measurement
    my_obe.pdf_update(measurement)

# end while loop

# get results from the parameter distribution
#
mean_values = my_obe.mean()
std_deviaion_values = my_obe.std()
covariance_matrix = my_obe.covariance()
```

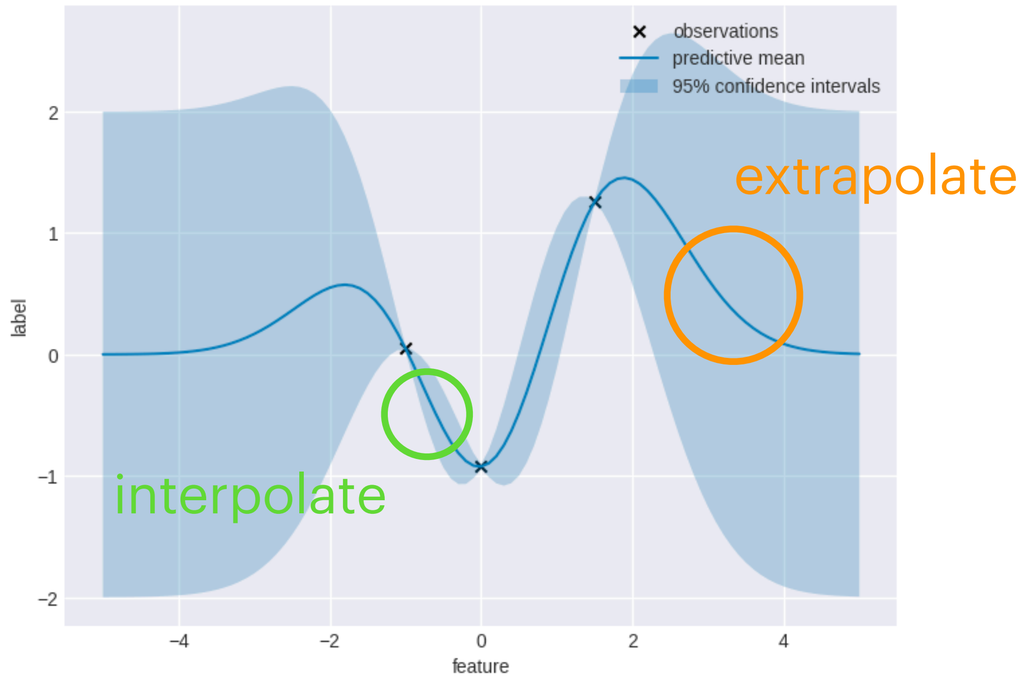
Bonus: Gaussian process & Bayesian optimization

- Uncertainty quantification for **optimizing a black-box function**
- Topics: Gaussian process, multi-armed bandits, BED for optimization
- PyData talks last year and the year before
- https://mng.bz/lY92
- 5 free ebooks: email nguyenminhquan135@gmail.com
- Discount code: **pyglobaldata45**



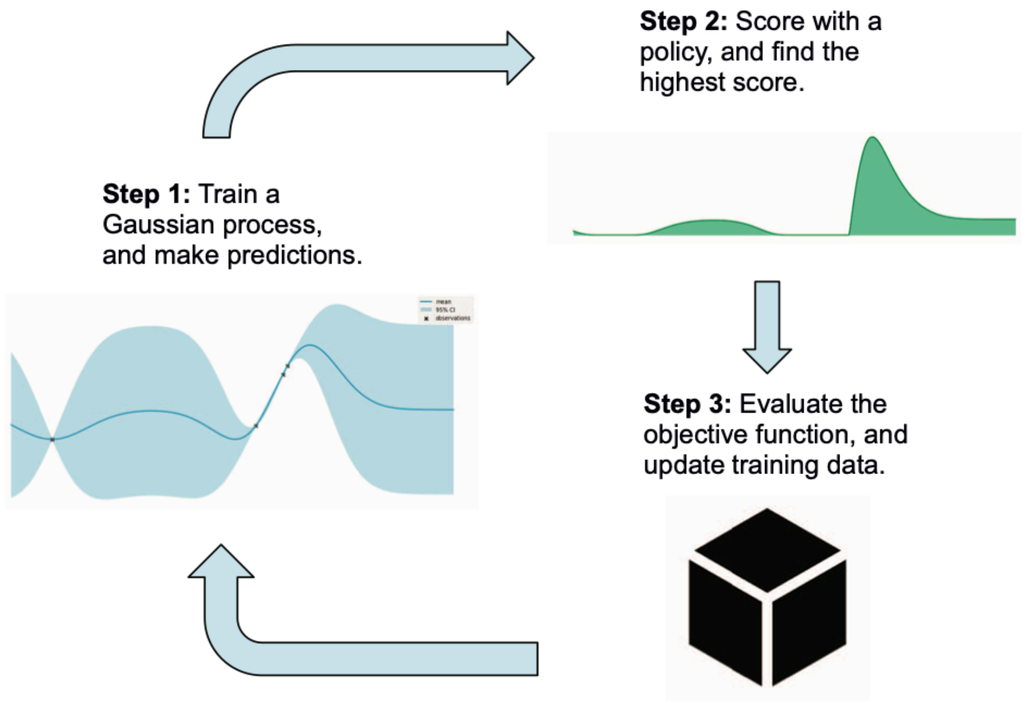
Bonus: Gaussian process & Bayesian optimization

- Uncertainty quantification for **optimizing a black-box function**
- Topics: Gaussian process, multi-armed bandits, BED for optimization
- PyData talks last year and the year before
- https://mng.bz/lY92
- 5 free ebooks: email nguyenminhquan135@gmail.com
- Discount code: **pyglobaldata45**



Bonus: Gaussian process & Bayesian optimization

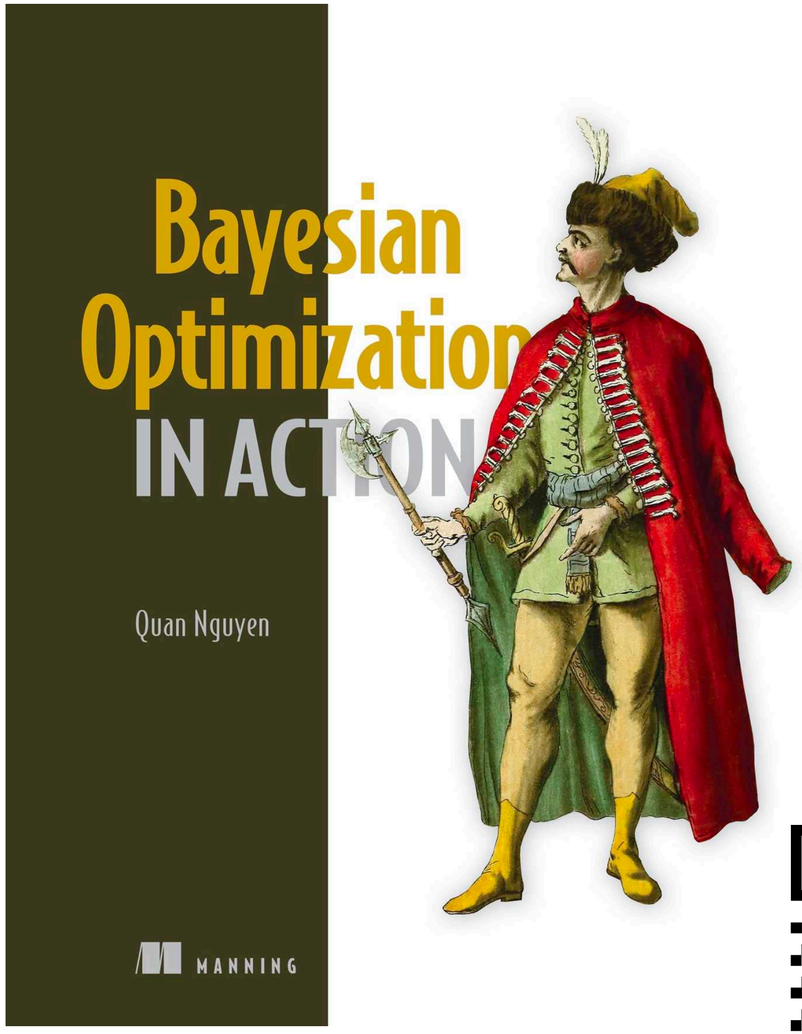
- Uncertainty quantification for **optimizing a black-box function**
- Topics: Gaussian process, multi-armed bandits, BED for optimization
- PyData talks last year and the year before
- https://mng.bz/lY92
- 5 free ebooks: email nguyenminhquan135@gmail.com
- Discount code: **pyglobaldata45**



The Bayesian optimization loop, which combines a Gaussian process for making predictions with calibrated uncertainty (chapters 2 and 3) and a policy for decision making regarding how to evaluate the objective function next (chapters 4 through 6)

Bonus: Gaussian process & Bayesian optimization

- Uncertainty quantification for **optimizing a black-box function**
- Topics: Gaussian process, multi-armed bandits, BED for optimization
- PyData talks last year and the year before
- https://mng.bz/lY92
- 5 free ebooks: email nguyenminhquan135@gmail.com
- Discount code: **pyglobaldata45**



Last name *et al*., *Title*, Venue & year.

Last name *et al*., *Title*, Venue & year.