Big Data and Economics

Causal Inference

Kyle Coombs
Bates College | ECON/DCS 368

Table of contents

Prologue

Prologue

- We see in the Opportunity Atlas that neighborhood income mobility is correlated with many outcomes

- But are any of these correlations causal?

- If so, we should be able to change neighborhood characteristics to change outcomes

Goals today

1. Separate causality and correlation
2. Discuss common challenges to establishing causality
3. Discuss approaches and assumptions to establish causality
   - Control for all unobserved variables correlated with treatment
   - Use treatment that is truly random
   - Something between these two

Warning

- This causality stuff is **really** tricky

- A causal paper may be intuitive -- that means it is a great paper, but finding your own intuitive causal relationship in the wild is hard

- Beyond intuition, the math and statistics are also hard

    - There are many interrelated frameworks to put some structure on the problem
    - Connections between frameworks can be hard to see and sometimes not particularly illuminating at first

- Be patient and comfortable with the fact that you won't understand everything at first, second, third, or even when you're trying to teach the material†

† cough, cough --- me

Attribution

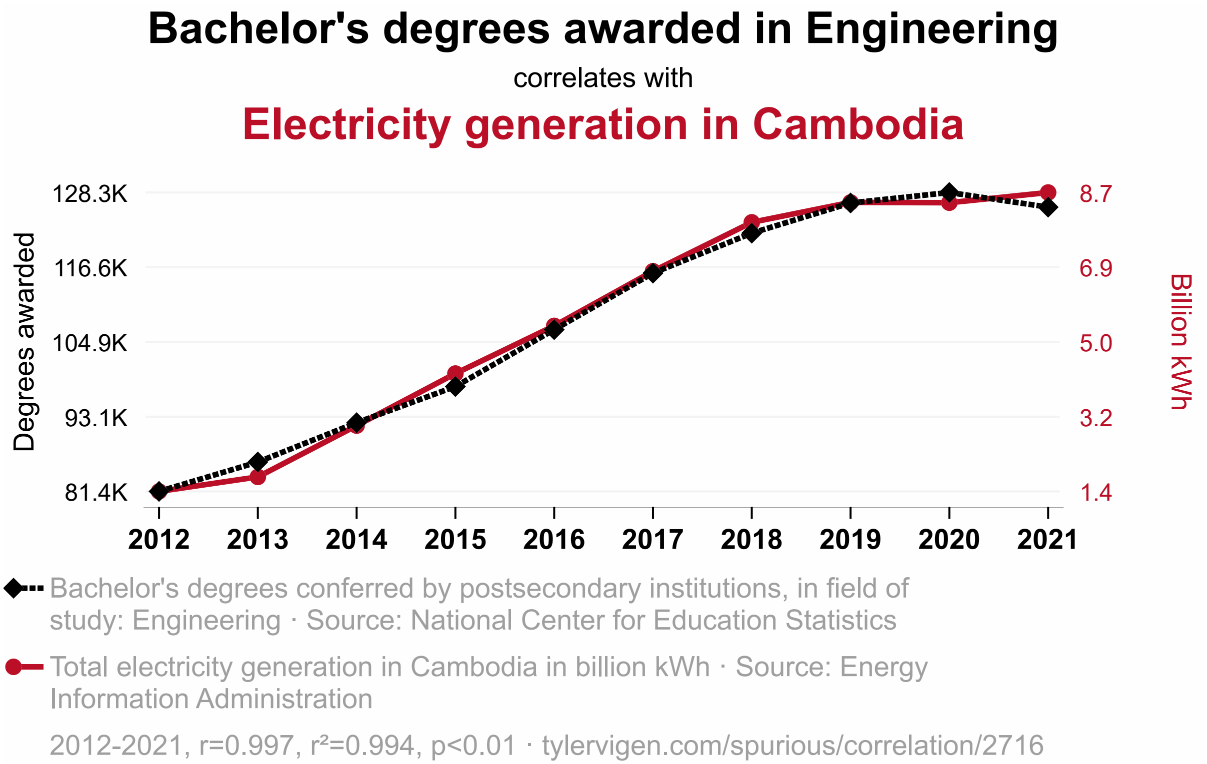
- These slides are adapted from work by Ed Rubin and Nick Huntington-Klein
- They're both superb econometric instructors and I highly recommend their work

Correlation vs. Causation

Spurious Correlations

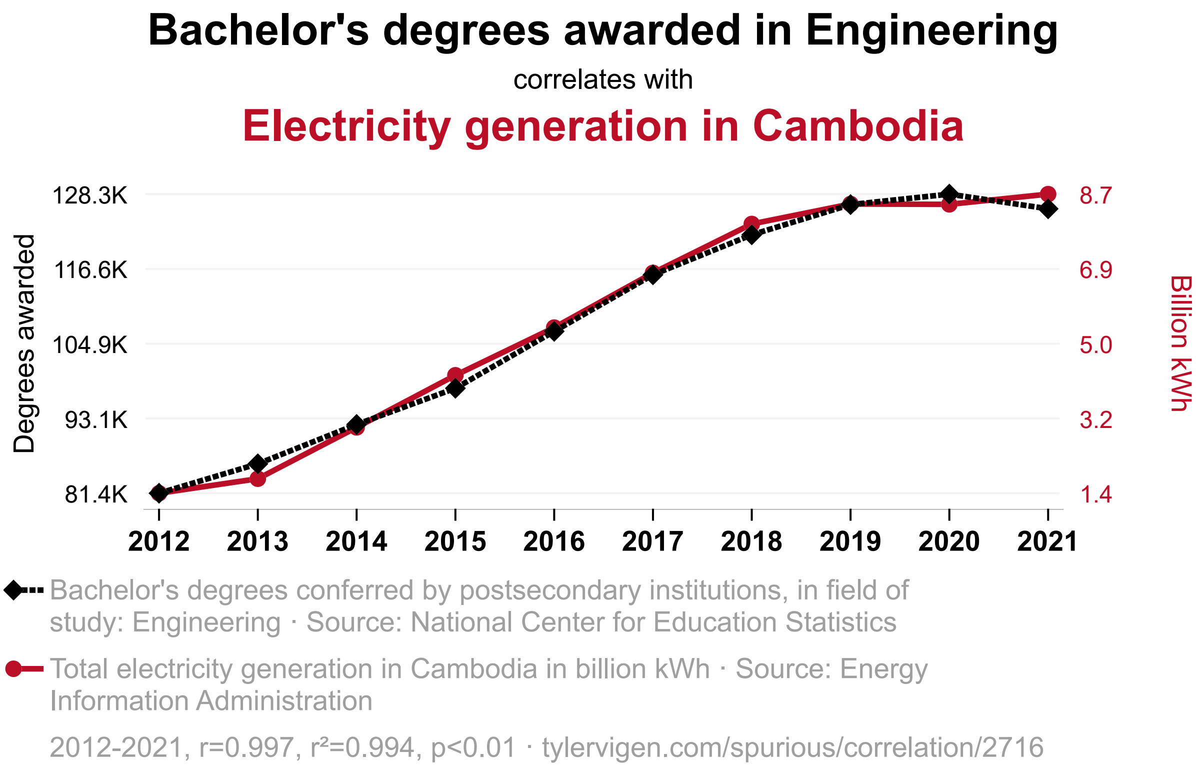
- Everyone submit a correlation that you know about in the world
- If you're not sure, please check out this delightful https://www.tylervigen.com/spurious-correlations

Spurious correlation and bad policy



Someone with this graph argues Cambodia should disincentivize engineering to fight climate change. Does that make sense?

Spurious correlation and bad policy



Someone with this graph argues Cambodia should disincentivize engineering to fight climate change. Does that make sense?

No! But this is why this matters. One nice-looking correlation plus a bad actor = very bad policy.

Correlation $\neq$ Causation

You've likely heard the saying

> Correlation is not causation.

- What does it mean?

Correlation $\neq$ Causation

You've likely heard the saying

> Correlation is not causation.

- What does it mean?

The saying is pointing out that there are violations of **exogeneity.**

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

- What does it mean?

The saying is pointing out that there are violations of **exogeneity.**

Although correlation is not causation, **causation (almost always) *requires* correlation**.

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

- What does it mean?

The saying is pointing out that there are violations of **exogeneity.**

Although correlation is not causation, **causation (almost always) *requires* correlation**.

New saying:

> Correlation plus **exogeneity** is causation.

- Today we're going to unpack this a bit to kick off a unit on causal inference methods

Causal questions

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

Causal questions

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- Did flipping the switch cause light to go on?
- How did this baby get here?

Causal questions

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- Did flipping the switch cause light to go on?
- How did this baby get here?

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

Causal questions

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- Did flipping the switch cause light to go on?
- How did this baby get here?

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

- Does job growth cause higher economic mobility?
- What caused the capital riot?
- How does the number of police officers affect crime?
- What is the effect of better air quality on test scores?
- Do tariffs reduce the amount of trade?
- How did cannabis legalization affect mental health/opioid addiction?

Non-causal correlations

Examples of non-zero *correlations* that are not *causal* (or may be causal in the other direction!)

Some obvious:

- People tend to wear shorts on days when ice cream trucks are out
- Rooster crowing sounds are followed closely by sunrise*

Some less obvious:

- Colds tend to clear up a few days after you take Emergen-C
- The performance of the economy tends to be lower or higher depending on the president's political party

Find more at https://www.tylervigen.com/spurious-correlations

*This case of mistaken causality is the basis of the film Rock-a-Doodle which I understand is extremely entertaining.

So what is causality?

- We say that `X` *causes* `Y` if...
- Were we to intervene and *change* the value of `X` without changing anything else...
- then `Y` would also change as a result

Important Note

- "X causes Y" *doesn't* mean that X is necessarily the *only* thing that causes Y
- And it *doesn't* mean that all Y must be X
- For example, using a light switch causes the light to go on
- But not if the bulb is burned out (no Y, despite X), or if the light was already on (Y without X), and it ALSO needs electricity (something else causes Y)
- But still we'd say that using the switch causes the light! The important thing is that X *changes the distribution* of Y, not that it necessarily makes it happen for certain

Prediction vs. causation

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Prediction vs. causation

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

Prediction vs. causation

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

2. **Causal inference:**[†] Estimate the true, population model that explains how y changes when we change x_j—focuses on β_j. Accuracy of $\hat{y}$ is not important. (So R^2 concerns can often take a hike.)

† Often called *causal identification*.

Why Causality?

- Many interesting questions to answer with data are causal
- Some are non-causal - for example, "how can we predict whether this photo is of a dog or a cat" is vital to how Google Images works, but it doesn't care what *caused* the photo to be of a dog or a cat
- Nearly every *why* question is causal and what we want to know!
- Also, this is economists' comparative advantage!
    - Plenty fields do statistics. But few make causal inference standard training for students
- This understanding of causality makes economists useful! *This* is one big reason why tech companies have whole economics departments

Fundamental Problem of Causal Inference

The challenges

Causal inference can be pretty difficult—both practically and econometrically.

The challenges

Causal inference can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

The challenges

Causal inference can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Selection bias
- Measurement error

The challenges

Causal inference can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Selection bias
- Measurement error

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$.

The challenges

Causal inference can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Selection bias
- Measurement error

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$. Causality requires us to **hold all else constant** (*ceterus paribus*) on average, i.e.

- The amount our model misses the mark (u) is equally likely to be positive as negative, or unbiased

Fund. Problem of Causal Inference

- The econometric problems largely fall under the umbrella problem that is fundamental to causal inference
- In short, it is impossible to observe a treated unit in the counterfactual world where they were not untreated
- Unless your name is Evelyn Quan, Marty McFly, Loki, or Miles Morales, this sort of multiversal experimentation is not possible
- You're stuck with the rest of in 2024, using an extremely clever, but limited causal infernce toolbox that relies on **exogeneity**

What is exogeneity?

- Can anyone tell me what exogeneity is?

What is exogeneity?

- Can anyone tell me what exogeneity is?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- Let's break this equation down into its component parts

What is exogeneity?

- Can anyone tell me what exogeneity is?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- Let's break this equation down into its component parts
    - y is the outcome/dependent variable
    - x_k are the independent/explanatory variables
    - β_k are the coefficients on the explanatory variables
    - u is the error term: anything else that affects y that we didn't/couldn't include
- Formally, exogeneity means $\mathbb{E}[u_i|X] = 0$: in expectation ("on average") the error term is zero after controlling for all the x variables
- Intuition: anything we left out that explains y is uncorrelated with our x variables

Causal inference approaches

- So how can we get $\mathbb{E}[u_i|X] = 0$ to make a causal claim?

1. **Random assignment**: Randomly assign units to treatment/control
   - The treatment is completely exogenous by design
2. **Conditional independence assumption (CIA)**: Control for everything that could possibly affect y that is related x
   - The treatment is then "as good as random," but you can't prove it
   - Sometimes called "selection on observables" and is often a tough sell
3. **Natural/quasi experiments**: A treatment is not randomly assigned, but due to something that "as good as random" with respect to treatment
   - This is the bread and butter of applied microeconomics

Assumptions

- All causal inference tools require an assumption about the world
- Your goal is to pick the least objectionable assumption possible
- You **cannot** prove these assumptions, that's why they're assumptions
- You can potentially see whether other patterns in the data are consistent with your assumption
    - e.g. Check placebo outcomes like parent's income for those who do/do not win a school lottery
    - These tests will change depending on your assumption/question/topic

Selection on observables

Prince Charles and Ozzy Osbourne are very similar. Source: Andrew Heiss' Mastodon

Causation

Effect of education on mobility

- Let's explore the three causal inference approaches with two simple examples

1. What is the effect of fertilizer on crop yield?

2. What is the effect of education on income mobility of those born at the 25th percentile of the income distribution?

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on crop yield.

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).

Violates *all else equal* (exogeneity). Biased and/or spurious results.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water). Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).

Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?

A: **Run an experiment!**

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on crop yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water). Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?

A: **Run an experiment!** 💩

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

By randomizing plots of land into **treatment** or **control**, we will, on average, include all kinds of land (soild, slope, water, *etc.*) in both groups.

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

By randomizing plots of land into **treatment** or **control**, we will, on average, include all kinds of land (soild, slope, water, *etc.*) in both groups.

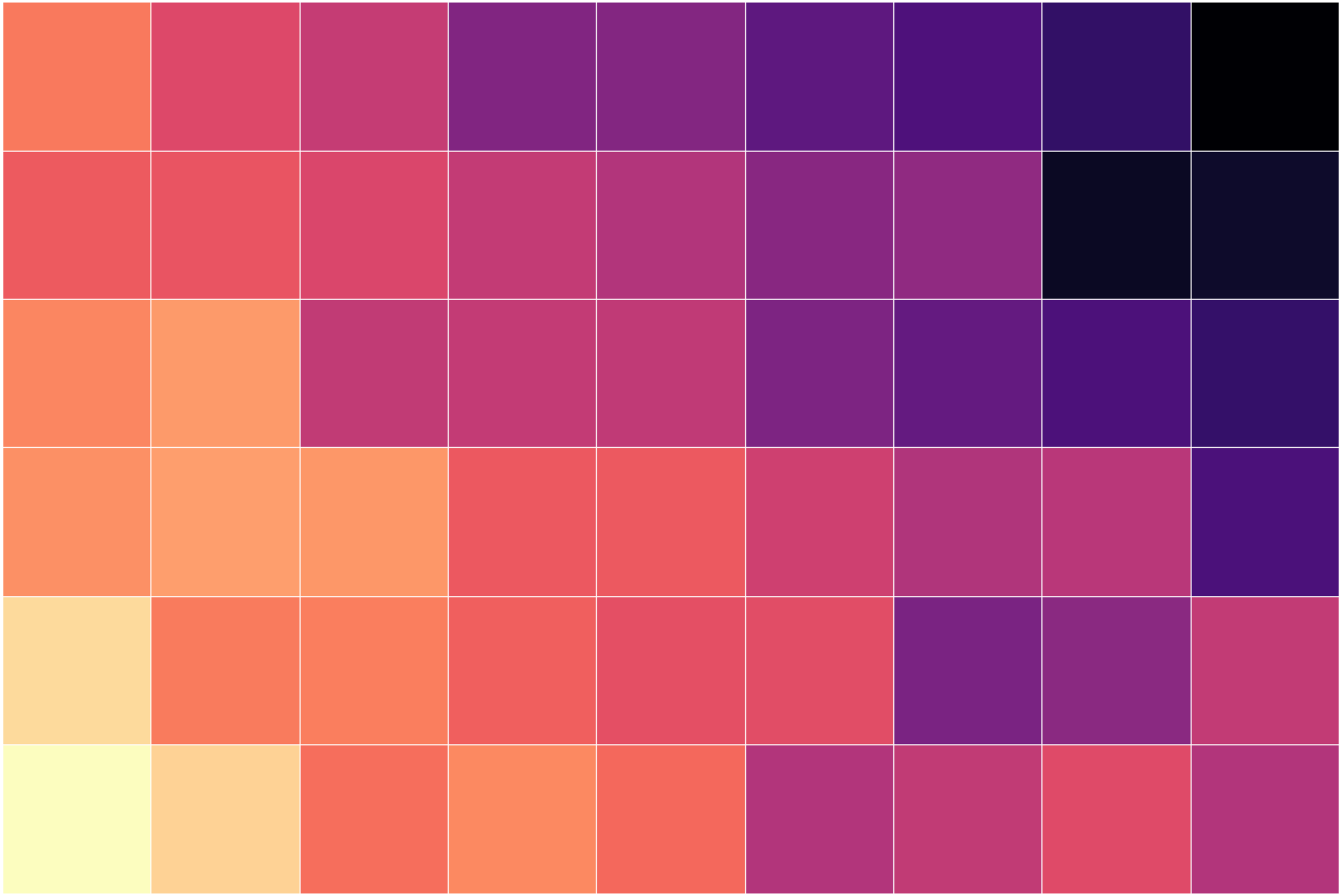
All else equal!

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

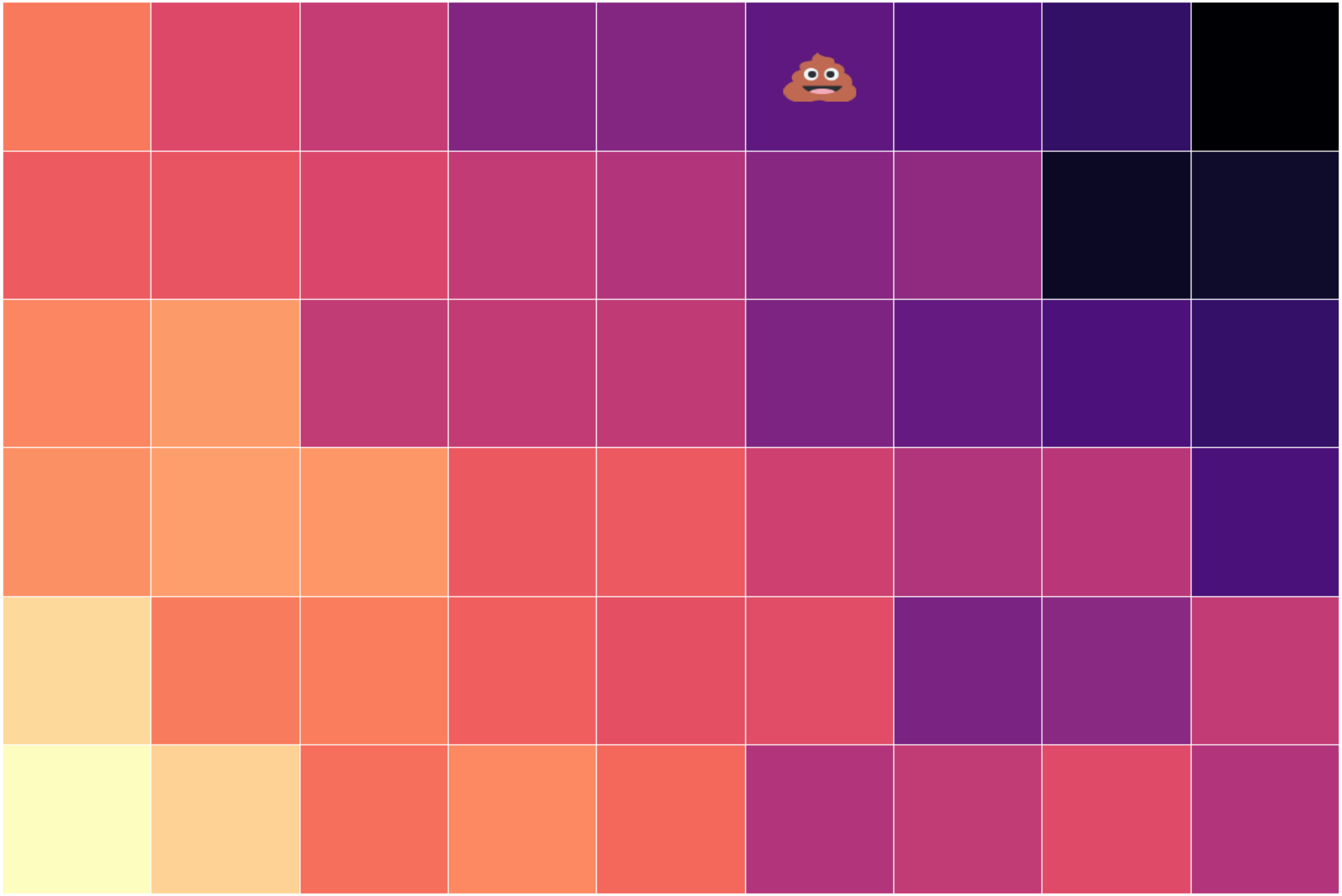
54 equal-sized plots

01	02	03	04	05	06	07	08	09
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54

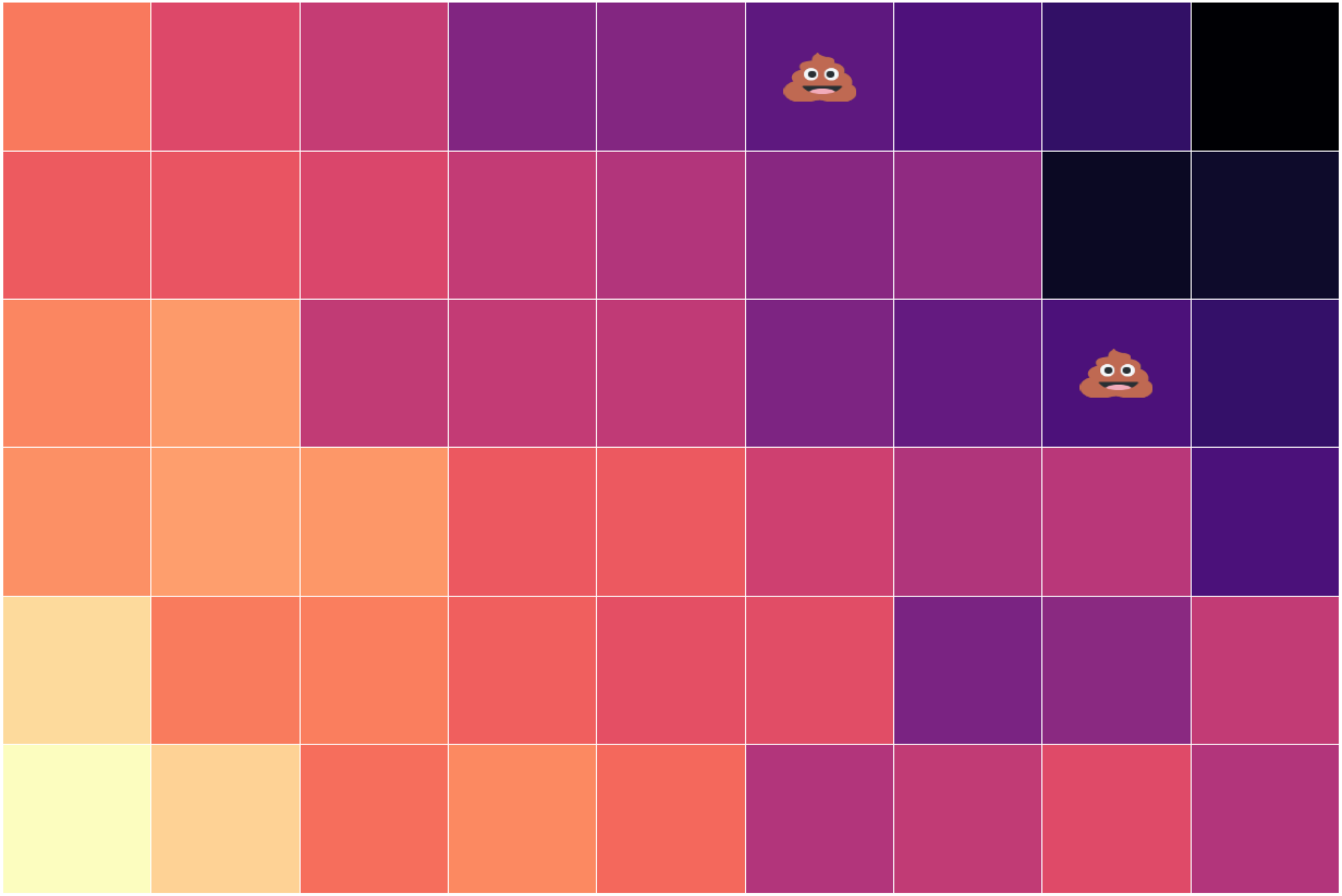
54 equal-sized plots of varying quality



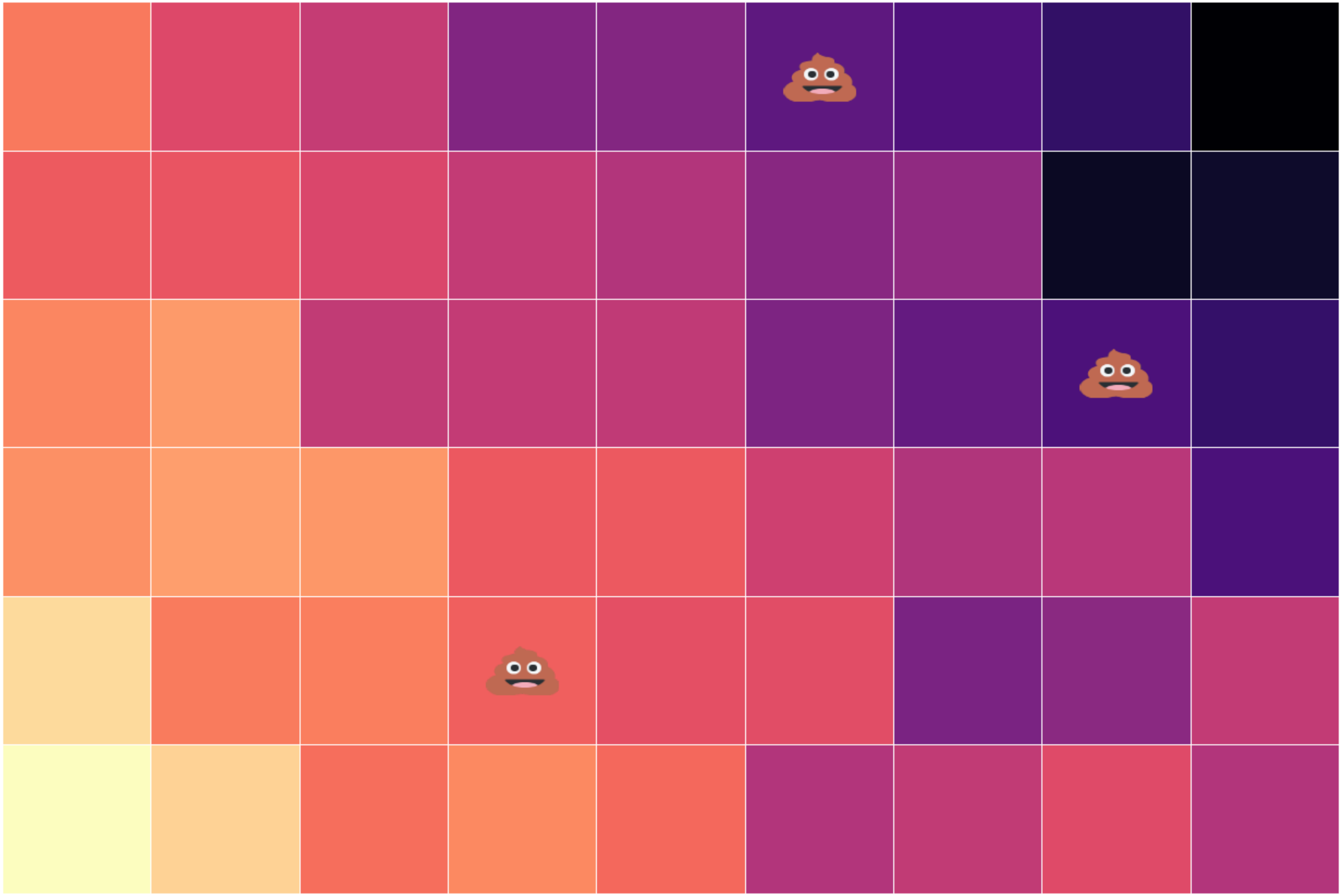
54 equal-sized plots of varying quality plus randomly assigned treatment



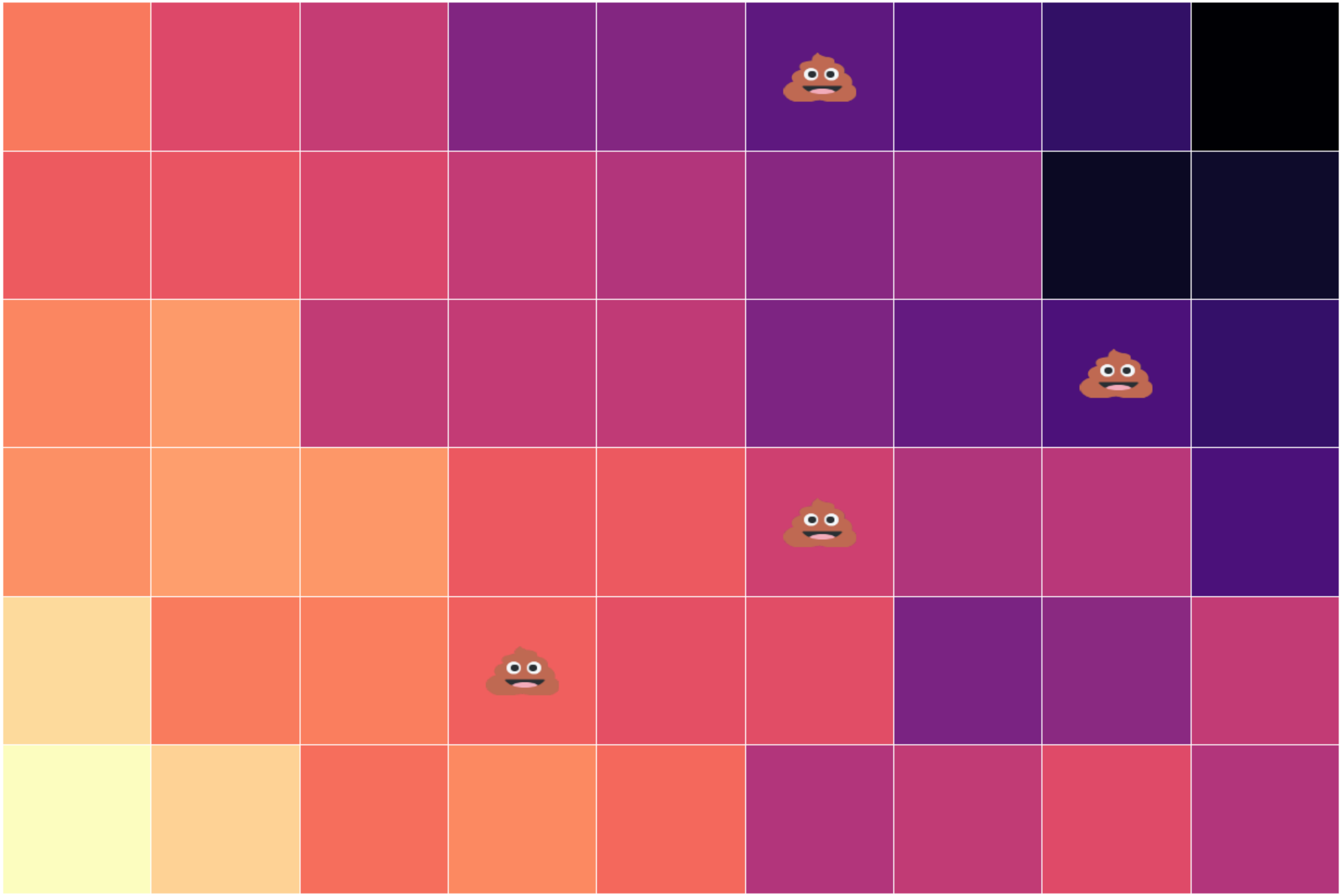
54 equal-sized plots of varying quality plus randomly assigned treatment



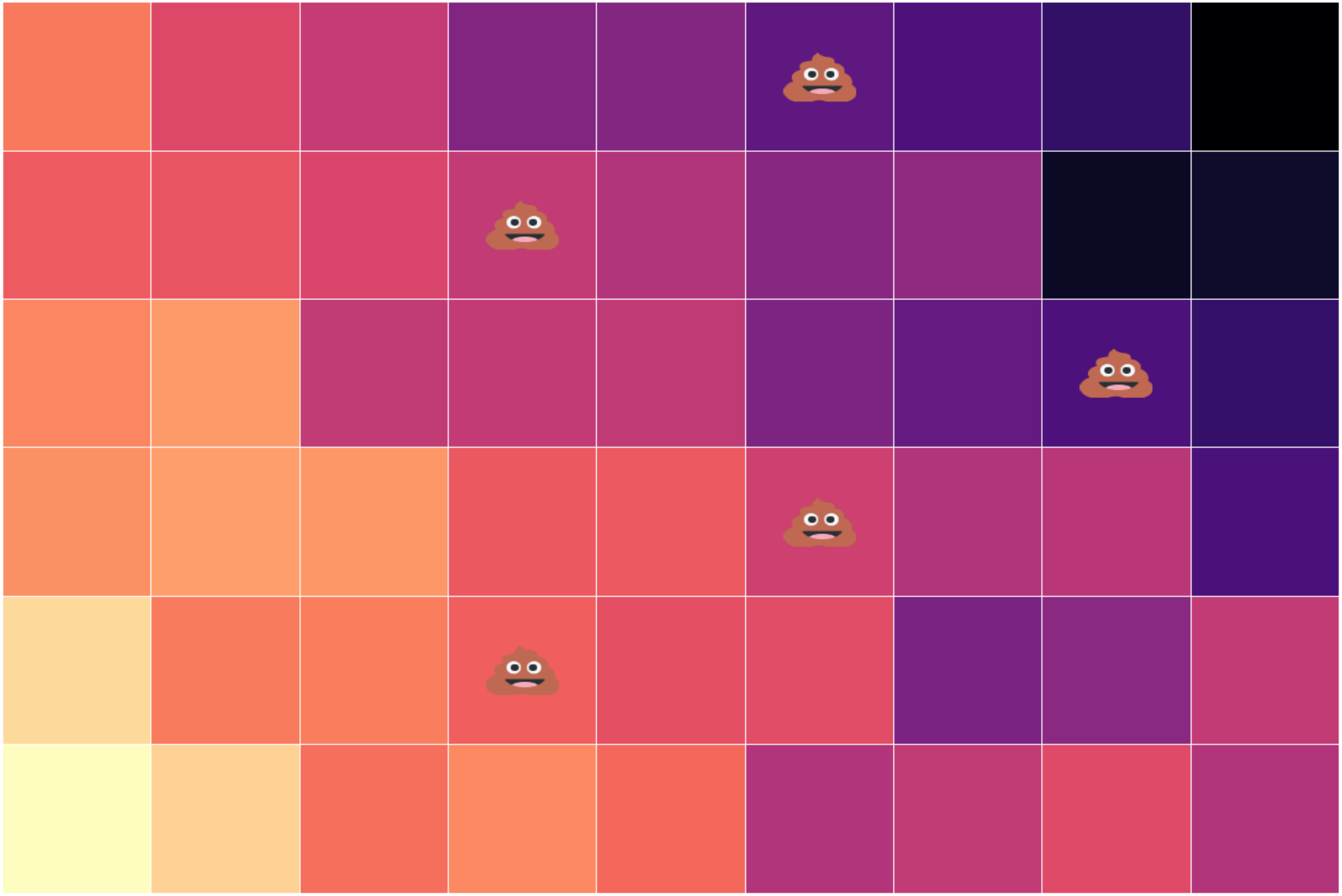
54 equal-sized plots of varying quality plus randomly assigned treatment



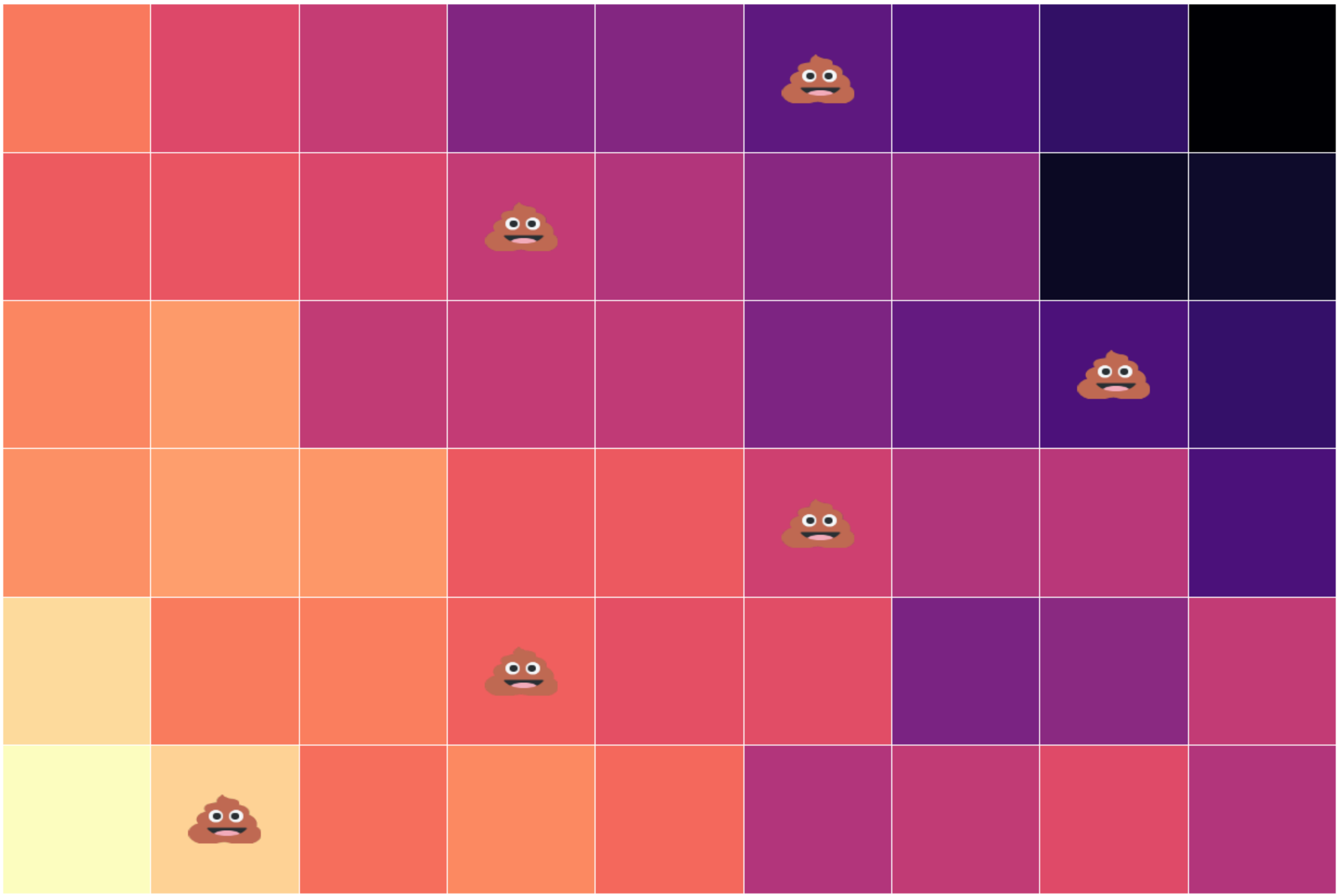
54 equal-sized plots of varying quality plus randomly assigned treatment



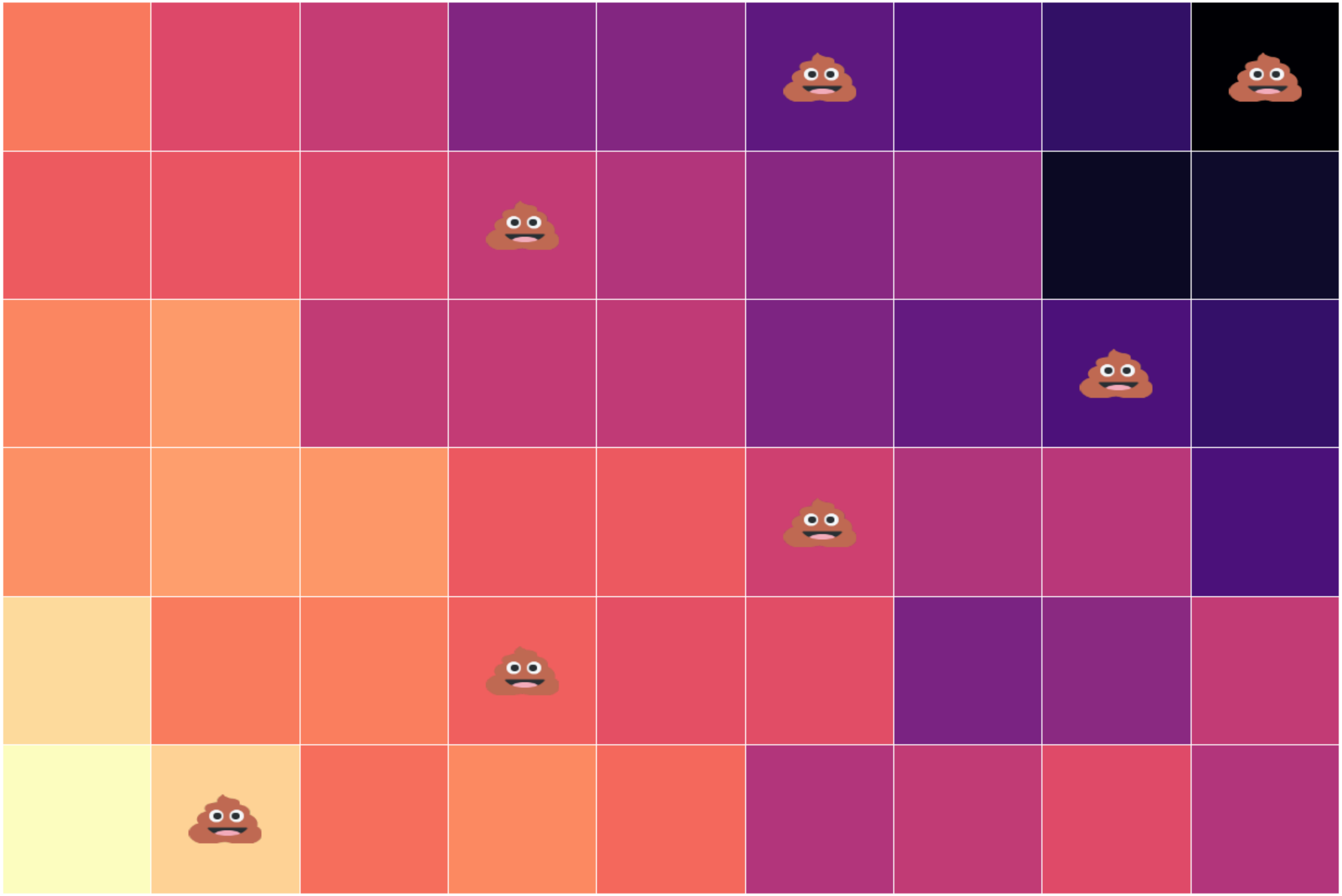
54 equal-sized plots of varying quality plus randomly assigned treatment



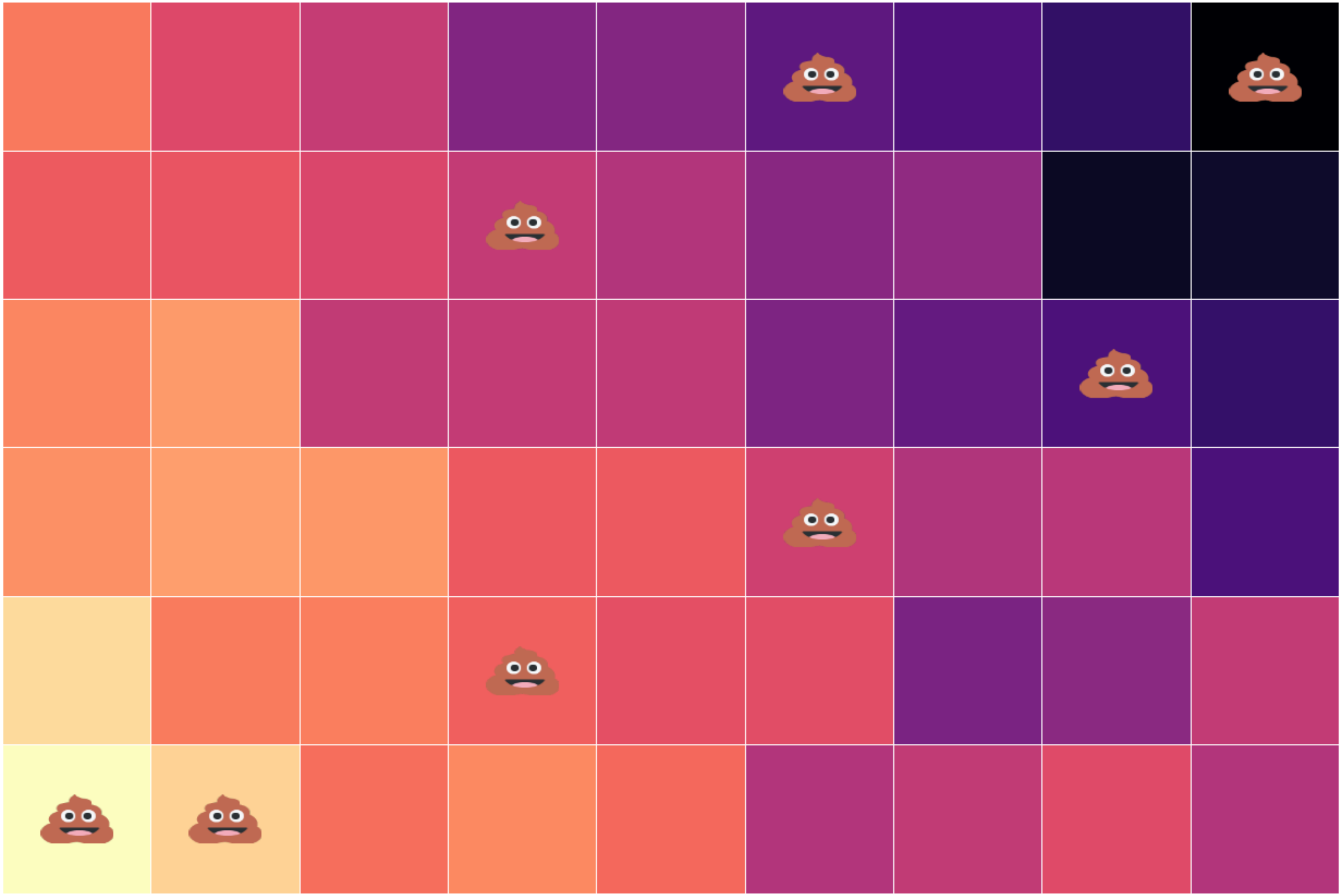
54 equal-sized plots of varying quality plus randomly assigned treatment



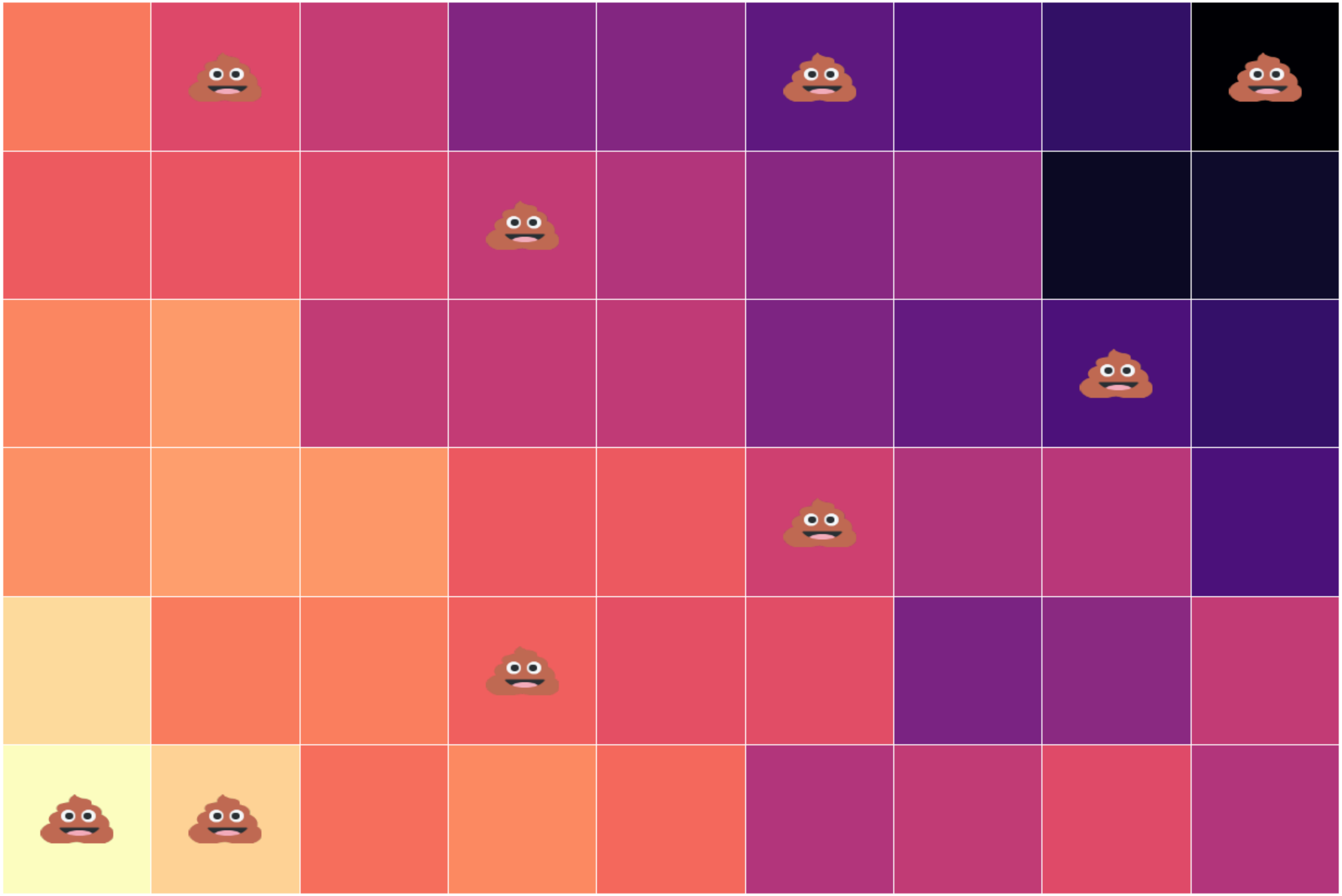
54 equal-sized plots of varying quality plus randomly assigned treatment



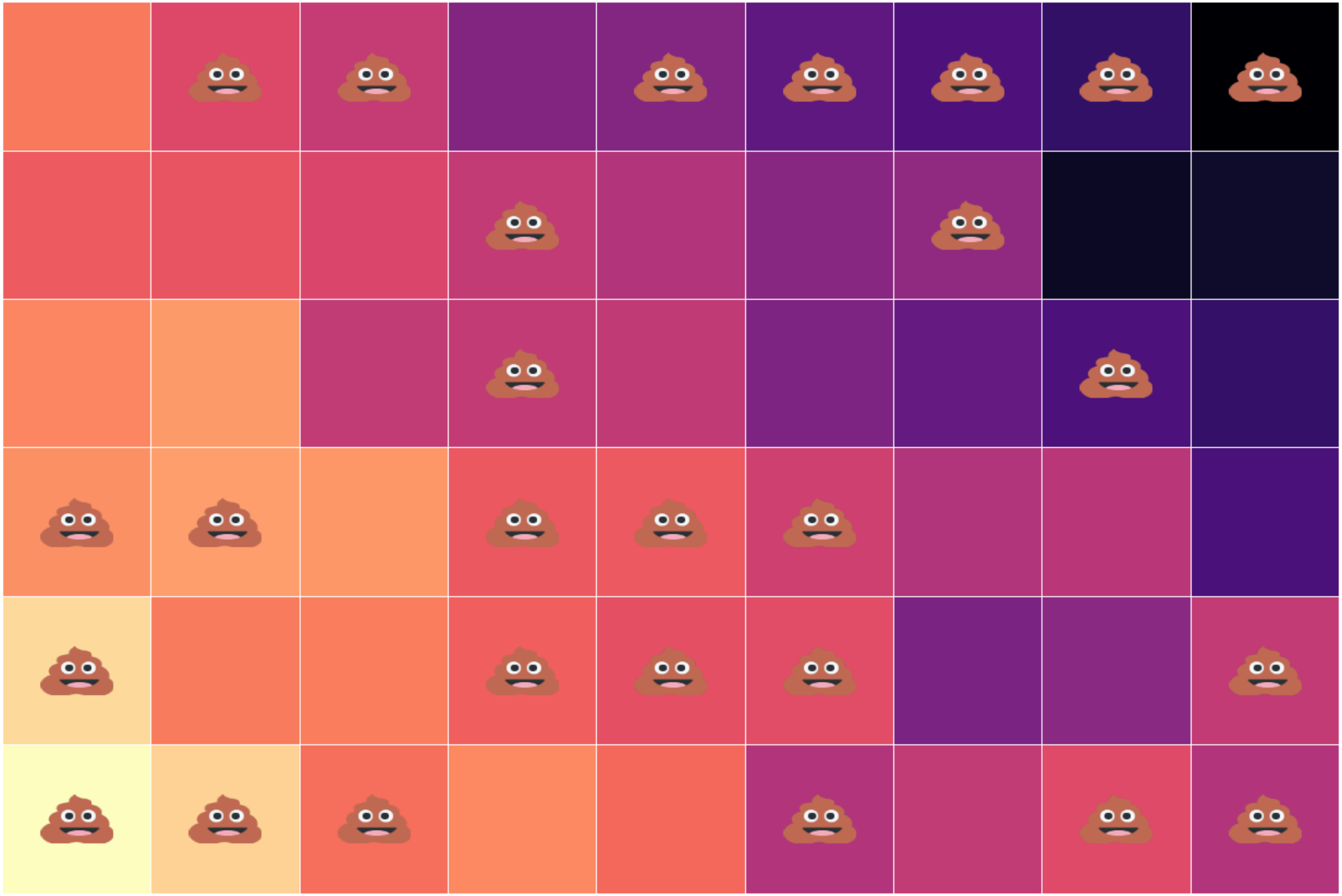
54 equal-sized plots of varying quality plus randomly assigned treatment



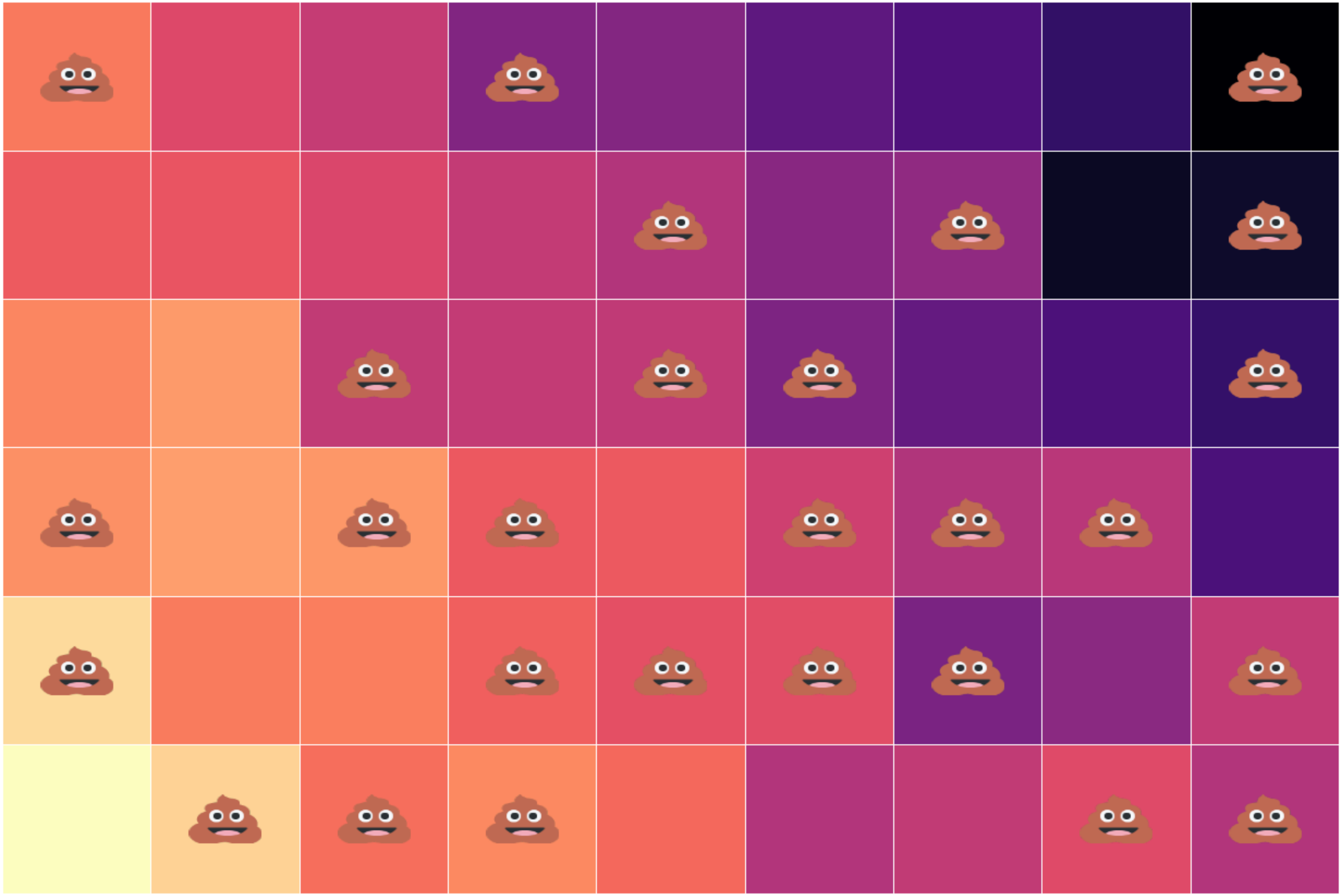
54 equal-sized plots of varying quality plus randomly assigned treatment



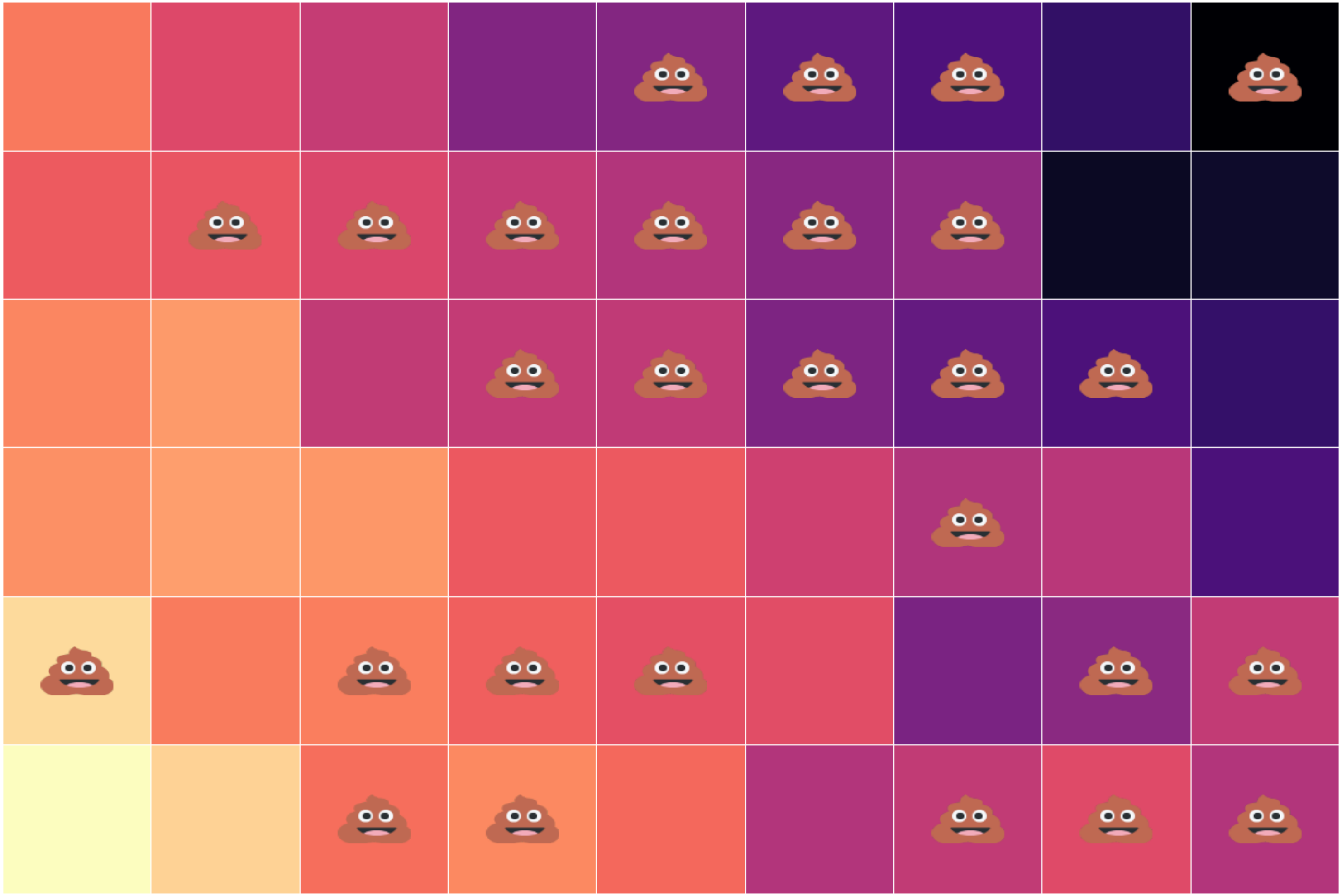
54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment



Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$ \begin{align} \overline{\text{Yield}}*text{Treatment}* - *overline{\text{Yield}}*\text{Control} \end{align} $$

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

\$\$ \begin{align} \overline{\text{Yield}}*text{Treatment}* - *overline{\text{Yield}}*\text{Control} \end{align} \$\$

Alternatively, we can use the regression

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$ \begin{align} \overline{\text{Yield}}*text{Treatment} - \overline{\text{Yield}}*\text{Control} \end{align} $$

Alternatively, we can use the regression

$$\mathbf{Yield}_i = \beta_0 + \beta_1 \mathbf{Trt}_i + u_i \tag{1}$$

where $\mathbf{Trt}_i$ is a binary variable (=1 if plot i received the fertilizer treatment).

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$ \begin{align} \overline{\text{Yield}}\text{Treatment} - \overline{\text{Yield}}\text{Control} \end{align} $$

Alternatively, we can use the regression

$$\mathbf{Yield}_i = \beta_0 + \beta_1 \mathbf{Trt}_i + u_i \tag{1}$$

where $\mathbf{Trt}_i$ is a binary variable (=1 if plot i received the fertilizer treatment).

Q: Should we expect (1) to satisfy exogeneity? Why?

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$ \begin{align} \overline{\text{Yield}}*text{Treatment} - \overline{\text{Yield}}*\text{Control} \end{align} $$

Alternatively, we can use the regression

$$\text{Yield}_i = \beta_0 + \beta_1 \text{Trt}_i + u_i \tag{1}$$

where Trt_i is a binary variable (=1 if plot i received the fertilizer treatment).

Q: Should we expect (1) to satisfy exogeneity? Why?

A: On average, **randomly assigning treatment should balance** trt. and control across the other dimensions that affect yield (soil, slope, water).

Example: Returns to education

Labor economists, policy makers, parents, and students are all interested in the (monetary) *return to education*.

Example: Returns to education

Labor economists, policy makers, parents, and students are all interested in the (monetary) *return to education.*

Thought experiment:

- Randomly select an individual.
- Give her an additional year of education.
- How much do her earnings increase?

This change in earnings gives the **causal effect** of education on earnings.

Example: Returns to education

Q: Could we simply regress earnings on education?

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

1. People *choose* education based upon many factors, *e.g.*, ability.
2. Education likely reduces experience (time out of the workforce).
3. Education is ***endogenous*** (violates *exogeneity*).

Causality

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

1. People *choose* education based upon many factors, *e.g.*, ability.
2. Education likely reduces experience (time out of the workforce).
3. Education is ***endogenous*** (violates *exogeneity*).

The point (2) above also illustrates the difficulty in learning about educations while *holding all else constant*.

Many important variables have the same challenge—gender, race, income.

Example: Returns to education

Q: So how can we estimate the returns to education?

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

Causality

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Option 2: Control for all that **unobserved variation** that affects both education and earnings. (CIA)

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Option 2: Control for all that **unobserved variation** that affects both education and earnings. (CIA)

Option 3: Look for a ***natural experiment***—a policy or accident in society that arbitrarily increased education for one subset of people.

Example: Returns to education

- Let's try controlling for every variable that affects both education and earnings, under CIA it should work!

$$\mathbf{Earnings} = \beta_0 + \beta_1 \mathbf{Edu} + \beta_2 \mathbf{Race} + \beta_3 \mathbf{Gender} + \beta_4 \mathbf{Ability} + \cdots + u$$

- Anyone see any problems?

Example: Returns to education

- Let's try controlling for every variable that affects both education and earnings, under CIA it should work!

$$\textbf{Earnings} = \beta_0 + \beta_1 \textbf{Edu} + \beta_2 \textbf{Race} + \beta_3 \textbf{Gender} + \beta_4 \textbf{Ability} + \cdots + u$$

- Anyone see any problems?
- Should race and gender be interacted? Race or gender and education?
- How do we measure ability? Specialized tests? Do those tests capture everything?
- Should we control for experience in a job?
- Uh oh, this is getting complicated and I'm not even sure we learn much

Causality

Example: Returns to education

- Natural experiment approach: what policies arbitrarily increase education for a subset of people?

Causality

Example: Returns to education

- Natural experiment approach: what policies arbitrarily increase education for a subset of people?
- Admissions cutoffs: people around the cutoff are similar, but above gets more education
    - Regression discontinuity
- Lottery enrollment and/or capacity constraints: people who get in get more education
    - Instrumental variables
- New school built: people near school get more education
    - Difference-in-differences

Causality

Real-world experiments

RCTs and certain policy changes yield **real experiments** to isolate causal effects.

Characteristics

- Feasible—we can actually (potentially) run the experiment.
- Compare individuals randomized into treatment against individuals randomized into control.
- Require "good" randomization to get *all else equal* (exogeneity).

Real-world experiments

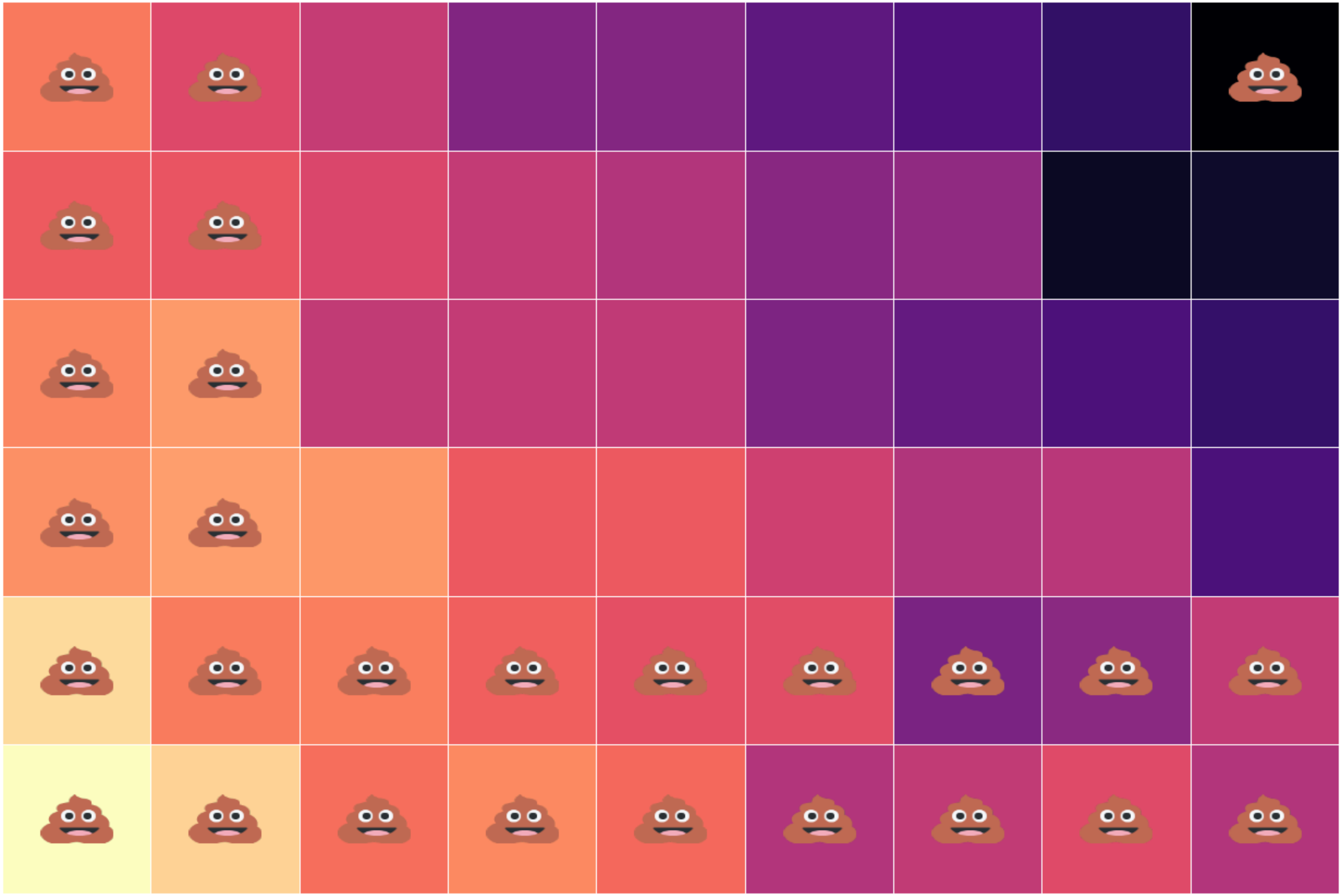
RCTs and certain policy changes yield **real experiments** to isolate causal effects.

Characteristics

- Feasible—we can actually (potentially) run the experiment.
- Compare individuals randomized into treatment against individuals randomized into control.
- Require "good" randomization to get *all else equal* (exogeneity).

Note: Your experiment's results are only as good as your randomization.

Unfortunate randomization



Causality

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

Causality

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$ \begin{align} y_{\text{Treatment},i} - y_{\text{Control},i} \end{align} $$

Causality

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$ \begin{align} y{*text{Treatment},i} - y*{\text{Control},i} \end{align} $$

which we will write (for simplicity) as

$$ \begin{align} y{*1,i} - y*{0,i} \end{align} $$

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

`$$ \begin{align} y{\text{Treatment},i} - y{\text{Control},i} \end{align} $$`

which we will write (for simplicity) as

`$$ \begin{align} y{1,i} - y{0,i} \end{align} $$`

This *ideal experiment* is clearly infeasible†, but it creates nice notation for causality (the Rubin causal model/Neyman potential outcomes framework).

† Without (1) God-like abilities and multiple universes or (2) a time machine.

The ideal experiment

The *ideal* data for 10 people

```
##      i trt  y1i  y0i
## 1    1   1 5.01 2.56
## 2    2   1 8.85 2.53
## 3    3   1 6.31 2.67
## 4    4   1 5.97 2.79
## 5    5   1 7.61 4.34
## 6    6   0 7.63 4.15
## 7    7   0 4.75 0.56
## 8    8   0 5.77 3.52
## 9    9   0 7.47 4.49
## 10  10   0 7.79 1.40
```

The ideal experiment

The *ideal* data for 10 people

```
##     i trt  y1i  y0i
## 1   1   1 5.01 2.56
## 2   2   1 8.85 2.53
## 3   3   1 6.31 2.67
## 4   4   1 5.97 2.79
## 5   5   1 7.61 4.34
## 6   6   0 7.63 4.15
## 7   7   0 4.75 0.56
## 8   8   0 5.77 3.52
## 9   9   0 7.47 4.49
## 10 10   0 7.79 1.40
```

Calculate the causal effect of trt. $$ \begin{align} \tau*i* = *y*{1,i} - y_{0,i} \end{align} $$ for each individual \(i\).

The ideal experiment

The *ideal* data for 10 people

```
##      i trt  y1i  y0i effect_i
## 1    1   1 5.01 2.56     2.45
## 2    2   1 8.85 2.53     6.32
## 3    3   1 6.31 2.67     3.64
## 4    4   1 5.97 2.79     3.18
## 5    5   1 7.61 4.34     3.27
## 6    6   0 7.63 4.15     3.48
## 7    7   0 4.75 0.56     4.19
## 8    8   0 5.77 3.52     2.25
## 9    9   0 7.47 4.49     2.98
## 10  10   0 7.79 1.40     6.39
```

Calculate the causal effect of trt. $$ \begin{align} \tau*i* = *y*{1,i} - y_{0,i} \end{align} $$ for each individual \(i\).

Causality

The ideal experiment

The *ideal* data for 10 people

```
##      i trt  y1i  y0i effect_i
## 1    1   1 5.01 2.56     2.45
## 2    2   1 8.85 2.53     6.32
## 3    3   1 6.31 2.67     3.64
## 4    4   1 5.97 2.79     3.18
## 5    5   1 7.61 4.34     3.27
## 6    6   0 7.63 4.15     3.48
## 7    7   0 4.75 0.56     4.19
## 8    8   0 5.77 3.52     2.25
## 9    9   0 7.47 4.49     2.98
## 10  10   0 7.79 1.40     6.39
```

Calculate the causal effect of trt. $$ \begin{align} \tau*i* = *y*{1,i} - y_{0,i} \end{align} $$ for each individual \(i\).

The mean of τ_i is the **average treatment effect** (ATE).

Thus, $\bar{\tau} = 3.82$

The ideal experiment

This model highlights the fundamental problem of causal inference. $$ \begin{align} \tau*i = \color{#e64173}{y*{1,i}} &- \color{#6A5ACD}{y_{0,i}} \end{align} $$

Causality

The ideal experiment

This model highlights the fundamental problem of causal inference. $$ \begin{align} \tau*i = \color{#e64173}{y*{1,i}} &- \color{#6A5ACD}{y_{0,i}} \end{align} $$

The challenge:

If we observe $y_{1,i}$, then we cannot observe $y_{0,i}$.
If we observe $y_{0,i}$, then we cannot observe $y_{1,i}$.

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
##      i trt  y1i  y0i
## 1    1   1 5.01   NA
## 2    2   1 8.85   NA
## 3    3   1 6.31   NA
## 4    4   1 5.97   NA
## 5    5   1 7.61   NA
## 6    6   0   NA 4.15
## 7    7   0   NA 0.56
## 8    8   0   NA 3.52
## 9    9   0   NA 4.49
## 10 10   0   NA 1.40
```

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
##     i trt  y1i  y0i
## 1   1   1 5.01   NA
## 2   2   1 8.85   NA
## 3   3   1 6.31   NA
## 4   4   1 5.97   NA
## 5   5   1 7.61   NA
## 6   6   0   NA 4.15
## 7   7   0   NA 0.56
## 8   8   0   NA 3.52
## 9   9   0   NA 4.49
## 10 10   0   NA 1.40
```

We can't observe $y_{1,i}$ and $y_{0,i}$.

But, we do observe

- $y_{1,i}$ for i in 1, 2, 3, 4, 5
- $y_{0,j}$ for j in 6, 7, 8, 9, 10

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
##     i trt  y1i  y0i
## 1   1   1 5.01   NA
## 2   2   1 8.85   NA
## 3   3   1 6.31   NA
## 4   4   1 5.97   NA
## 5   5   1 7.61   NA
## 6   6   0   NA 4.15
## 7   7   0   NA 0.56
## 8   8   0   NA 3.52
## 9   9   0   NA 4.49
## 10 10   0   NA 1.40
```

We can't observe $y_{1,i}$ and $y_{0,i}$.

But, we do observe

- $y_{1,i}$ for i in 1, 2, 3, 4, 5
- $y_{0,j}$ for j in 6, 7, 8, 9, 10

Q: How do we "fill in" the `NA`s and estimate $\bar{\tau}$?

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Then, rephrasing the previous slide,

- We only observe $y_{1,i}$ when $D_i = 1$.
- We only observe $y_{0,i}$ when $D_i = 0$.

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Then, rephrasing the previous slide,

- We only observe $y_{1,i}$ when $D_i = 1$.
- We only observe $y_{0,i}$ when $D_i = 0$.

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Q$_{2.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for $\overline{\tau}$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Q$_{2.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for $\overline{\tau}$?

Time for math! 🎉

Causality

Causally estimating the treatment effect

Assumption: Let $\tau_i = \tau$ for all i.

This assumption says that the treatment effect is equal (constant) across all individuals i.

Causality

Causally estimating the treatment effect

Assumption: Let $\tau_i = \tau$ for all i.

This assumption says that the treatment effect is equal (constant) across all individuals i.

Note: We defined

$$ \begin{align} \tau*i = \tau = \color{#e64173}{y*{1,i}} - \color{#6A5ACD}{y_{0,i}} \end{align} $$

which implies

$$ \begin{align} \color{#e64173}{y*{1,i}} = \color{#6A5ACD}{y*{0,i}} + \tau \end{align} $$

$Q_{3.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Q$_{3.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$\mathbf{Q_{3.0}}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$\mathbf{Q_{3.0}}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$$= E(y_{1,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$Q_{3.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$$= E(y_{1,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= E(\tau + y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$\mathbf{Q_{3.0}}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$$= E(y_{1,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= E(\tau + y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= \tau + E(y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

Q$_{3.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$$= E(y_{1,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= E(\tau + y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= \tau + E(y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= \text{Average causal effect} + \text{Selection bias}$$

Q$_{3.0}$: Is $E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0)$$

$$= E(y_{1,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= E(\tau + y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= \tau + E(y_{0,i} \mid D_i = 1) - E(y_{0,i} \mid D_i = 0)$$

$$= \text{Average causal effect} + \text{Selection bias}$$

So our proposed group-difference estimator give us the sum of

1. τ, the **causal, average treatment effect** that we want
2. **Selection bias:** How much trt. and control groups differ (on average).

Inference: Did we just get lucky?

Inference: Did we just get lucky?

- Most of today's lecture covered causal identification
- That's how you know whether the average treatment effect is causal
- But the other key part is inference: how do you know whether the average treatment effect is *statistically* different from zero?
- That's where "inference" comes in
- Inference is the practice of determining how special your results are.
- Generally you get a confidence interval and p-value (except Bayesian inference)

Types of inference

1. **Asymptotic** inference: what you saw in econometrics
    - Under a few assumptions, you can make inferences
2. **Randomization**: maybe you saw it?
    - Assign placebo treatments to see if results are unique
    - Are your results are driven by something about the treated group?
3. **Bootstrapping**: maybe you saw it?
    - Resample data to see if your results are sensitive to the sample
    - Are your results are driven by something about the sample?
4. **Bayesian**: I doubt you've seen this
    - Assume a prior distribution for β and update it
    - Generates a "credibility" interval

Next lecture: Regression analysis