Big Data and Economics

Causal Effects of Neighborhoods

Kyle Coombs
Bates College | ECON/DCS 368

Table of contents

Prologue

Prologue

- We saw 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
- How do we know if a correlation is causal?

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 estimation:**† Estimate the actual data-generating process—learning about 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.

† Often called *causal identification*.

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 estimation:**† Estimate the actual data-generating process—learning about 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.

For the rest of the term, we will focus on **causally estimating** β_j.

† Often called *causal identification*.

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model 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

As you saw in the data-analysis exercise, determining and estimating the true model 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
- Measurement error
- How precise can/must we be?

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model 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
- Measurement error
- How precise can/must we be?

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

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model 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
- Measurement error
- How precise can/must we be?

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*).

It's complicated

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

It's complicated

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

- What caused the forest fire?
- How did this baby get here?

It's complicated

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

- What caused the forest fire?
- How did this baby get here?

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

It's complicated

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

- What caused the forest fire?
- How did this baby get here?

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

- What causes some countries to grow and others to decline?
- What caused the capital riot?
- Did lax regulation cause Texas's recent energy problems?
- How does the number of police officers affect crime?
- What is the effect of better air quality on test scores?
- Do longer prison sentences decrease crime?
- How did cannabis legalization affect mental health/opioid addiction?

Correlation $\neq$ Causation

You've likely heard the saying

> Correlation is not causation.

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

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

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

Although correlation is not causation, **causation *requires* correlation**.

Correlation $\neq$ Causation

You've likely heard the saying

> Correlation is not causation.

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

Although correlation is not causation, **causation *requires* correlation**.

New saying:

> Correlation plus exogeneity is causation.

Let's work through a few examples.

Causation

Example: The causal effect of fertilizer†

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

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

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

Q: Could we simply regress yield on fertilizer?

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

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

Q: Could we simply regress yield on fertilizer?

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

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer[†]

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

Q: Could we simply regress yield on fertilizer?

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

Q: Why not?

[†] Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer[†]

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

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on corn 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?

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn 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!**

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn 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!** 💩

† Many of the early statistical and econometric studies involved agricultural field trials.

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.

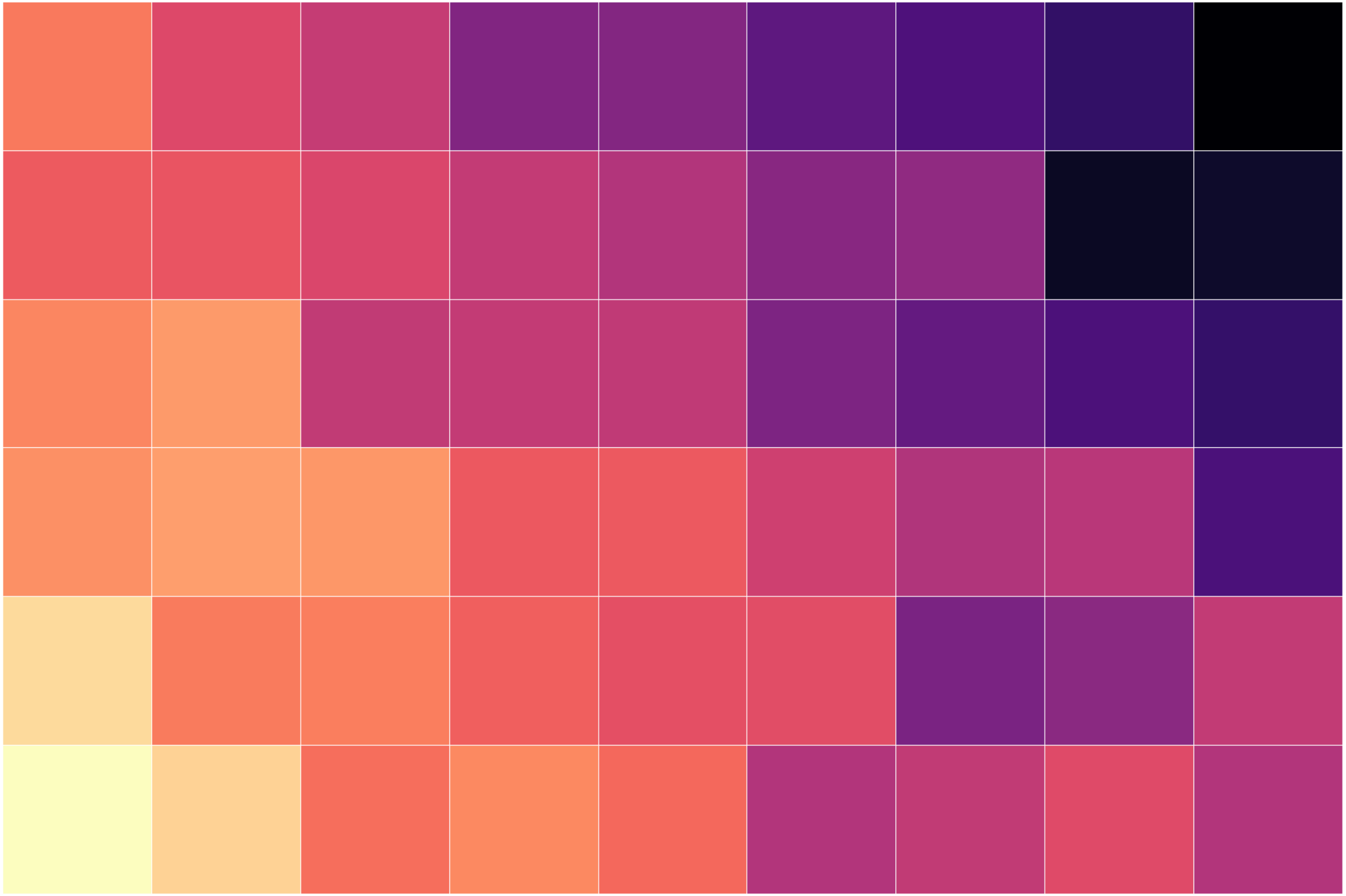
All else equal!

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

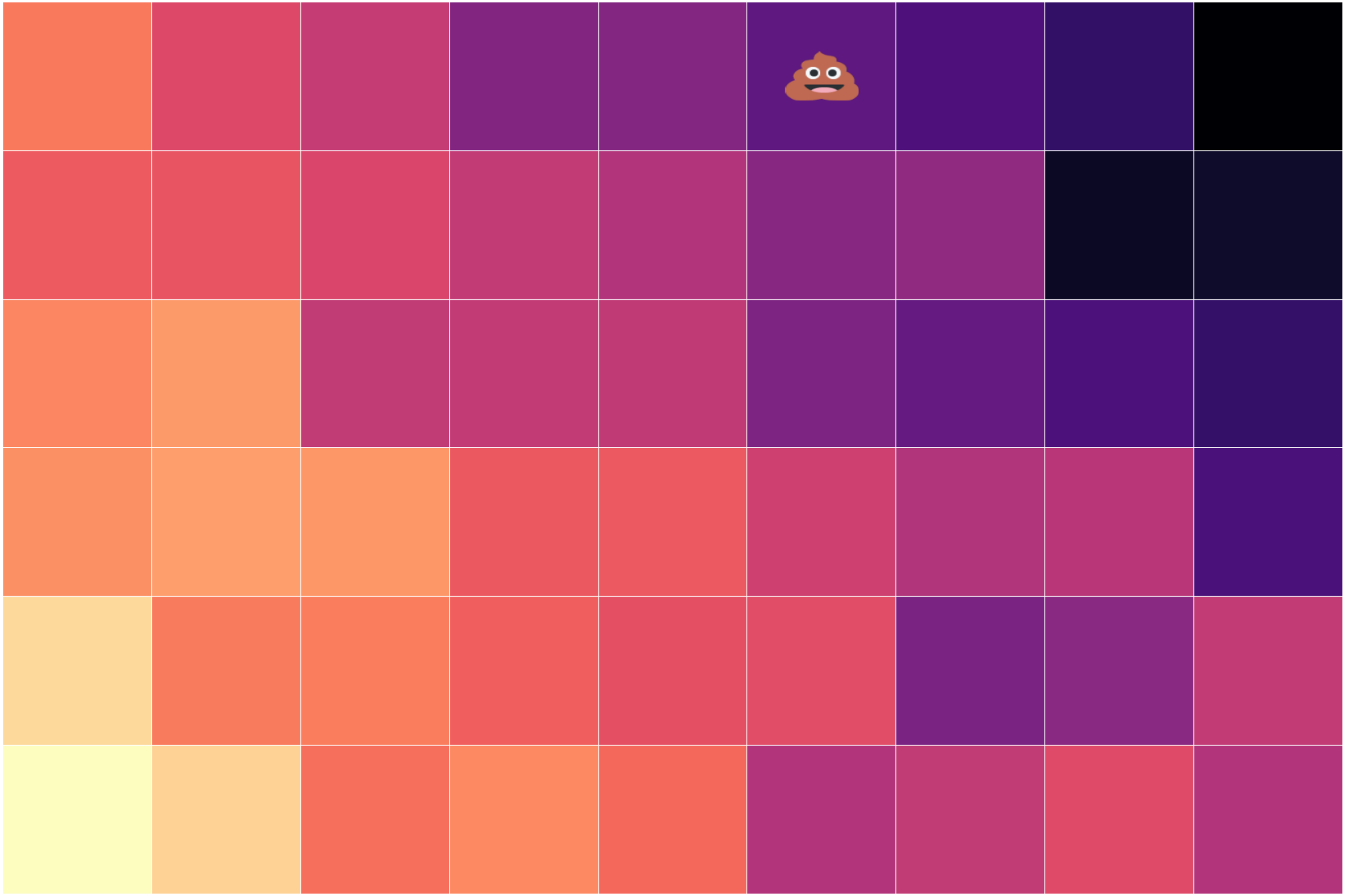
54 equal-sized plots

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

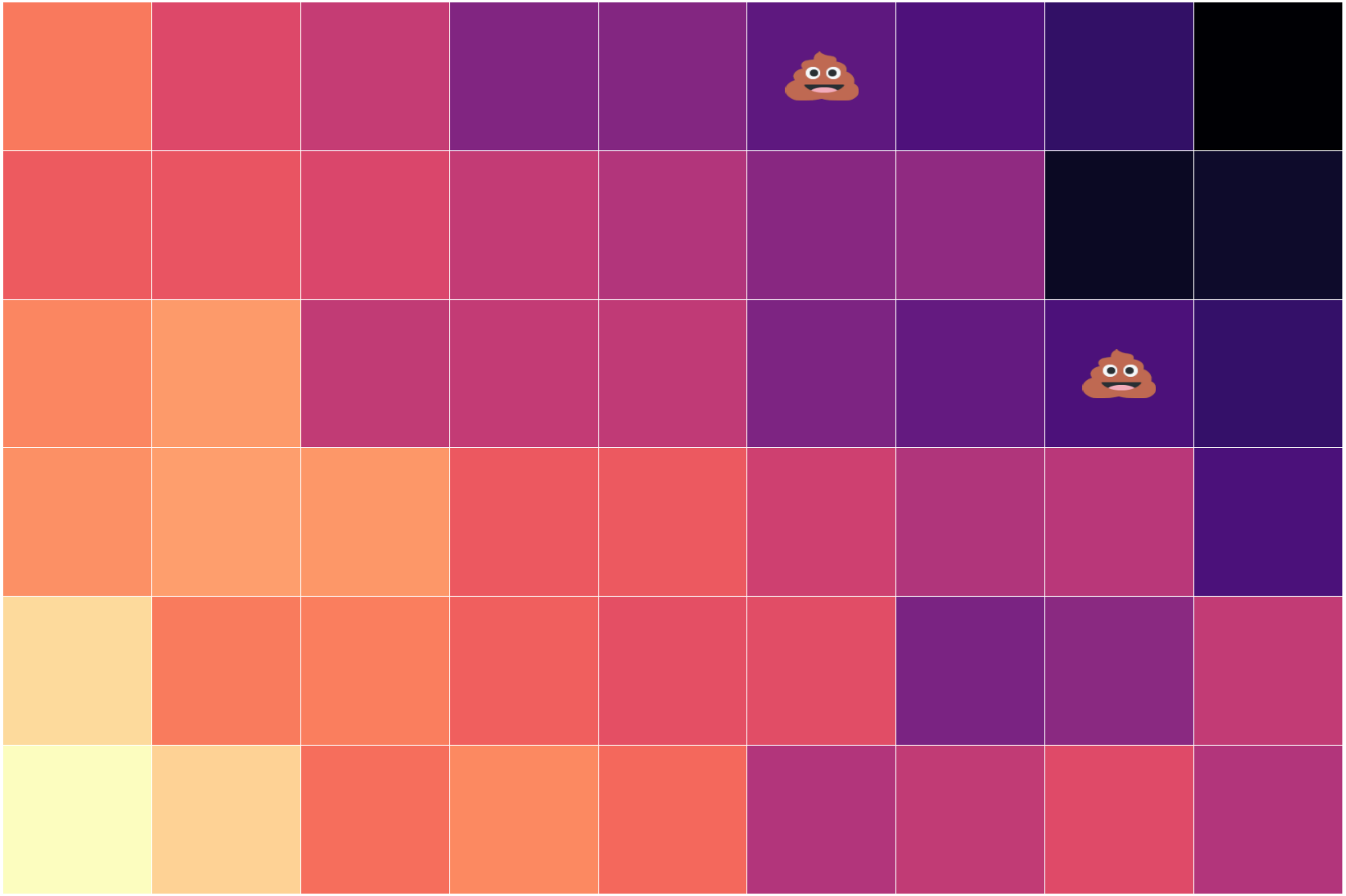
54 equal-sized plots of varying quality



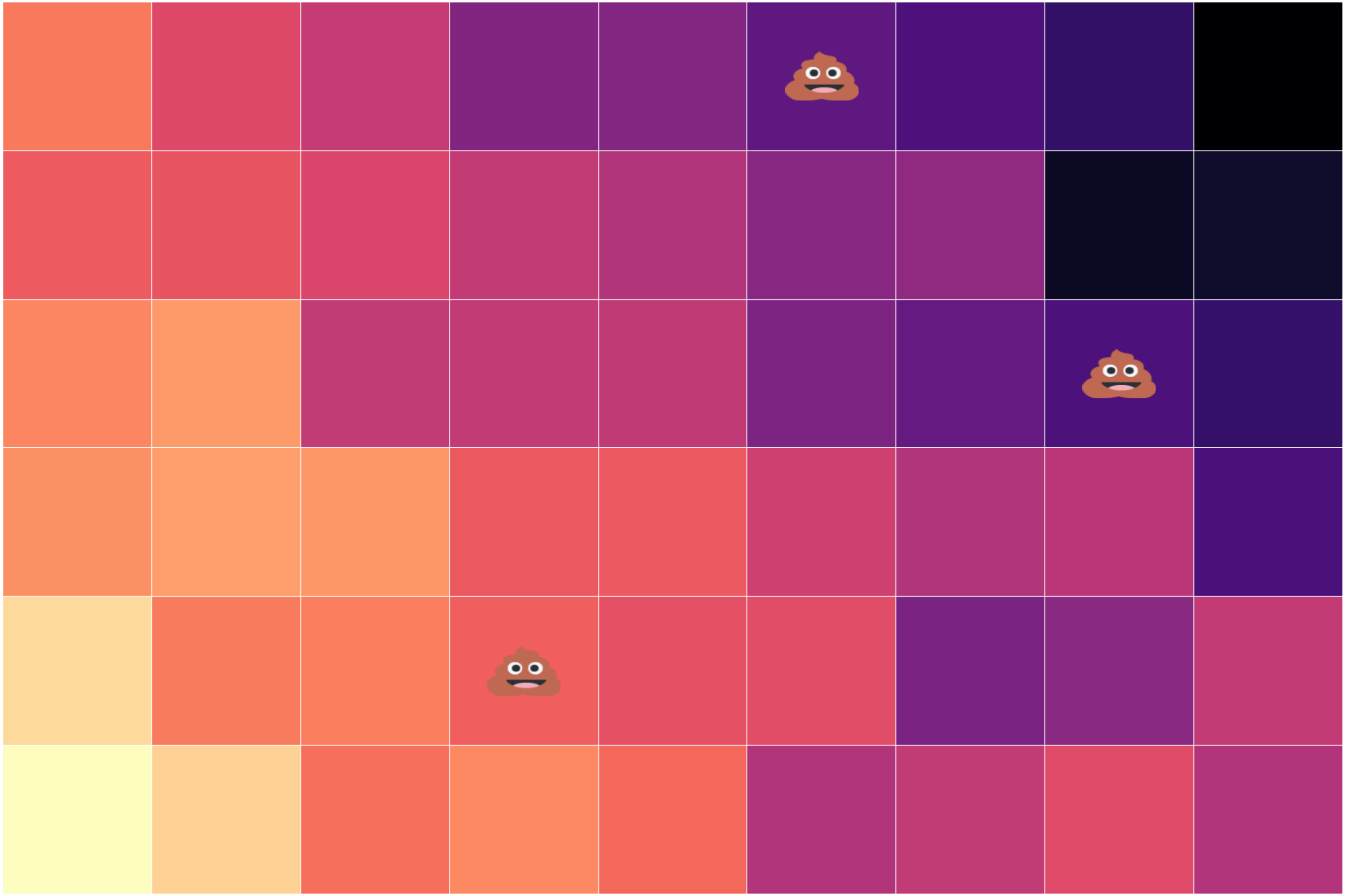
54 equal-sized plots of varying quality plus randomly assigned treatment



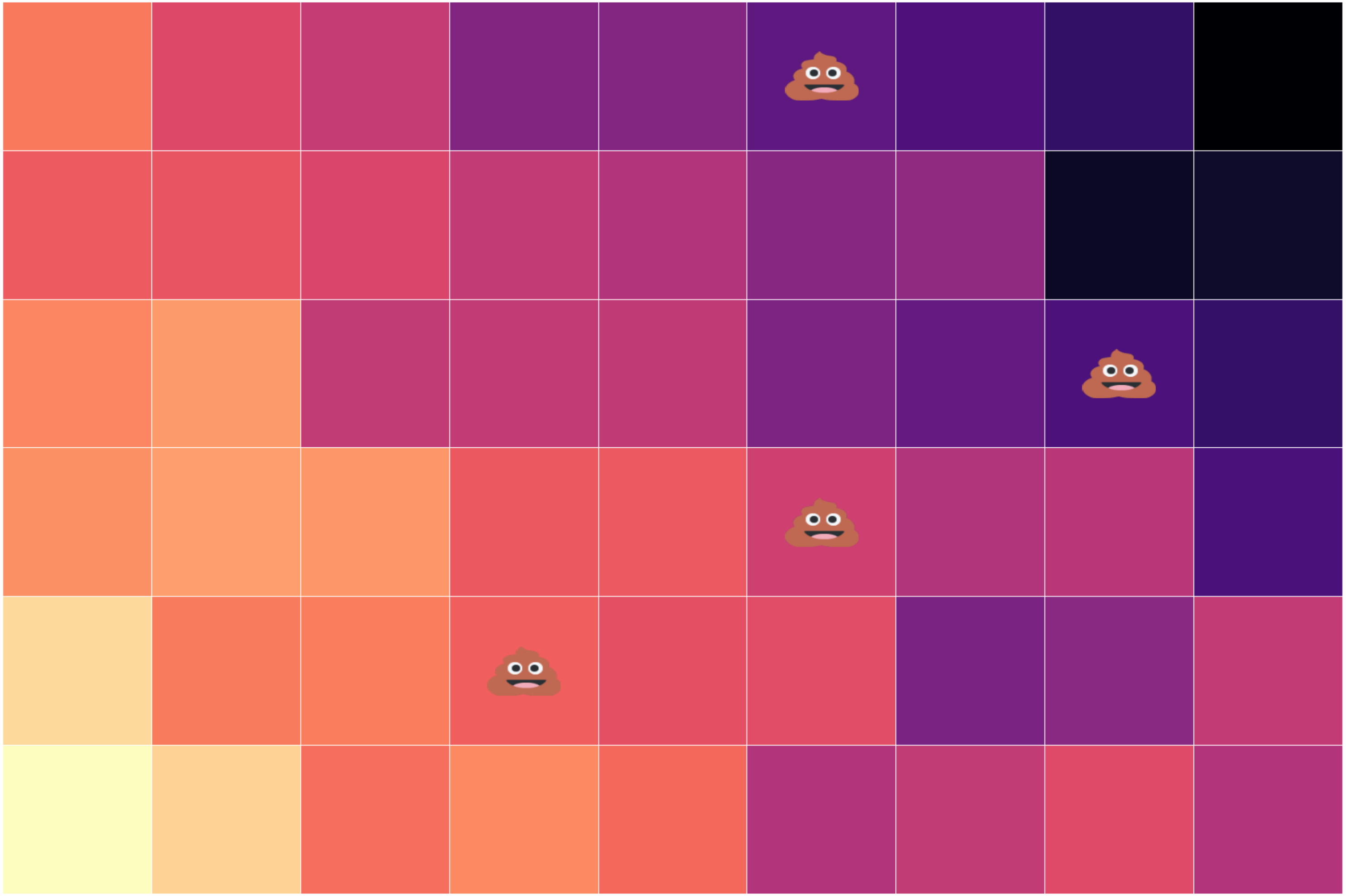
54 equal-sized plots of varying quality plus randomly assigned treatment



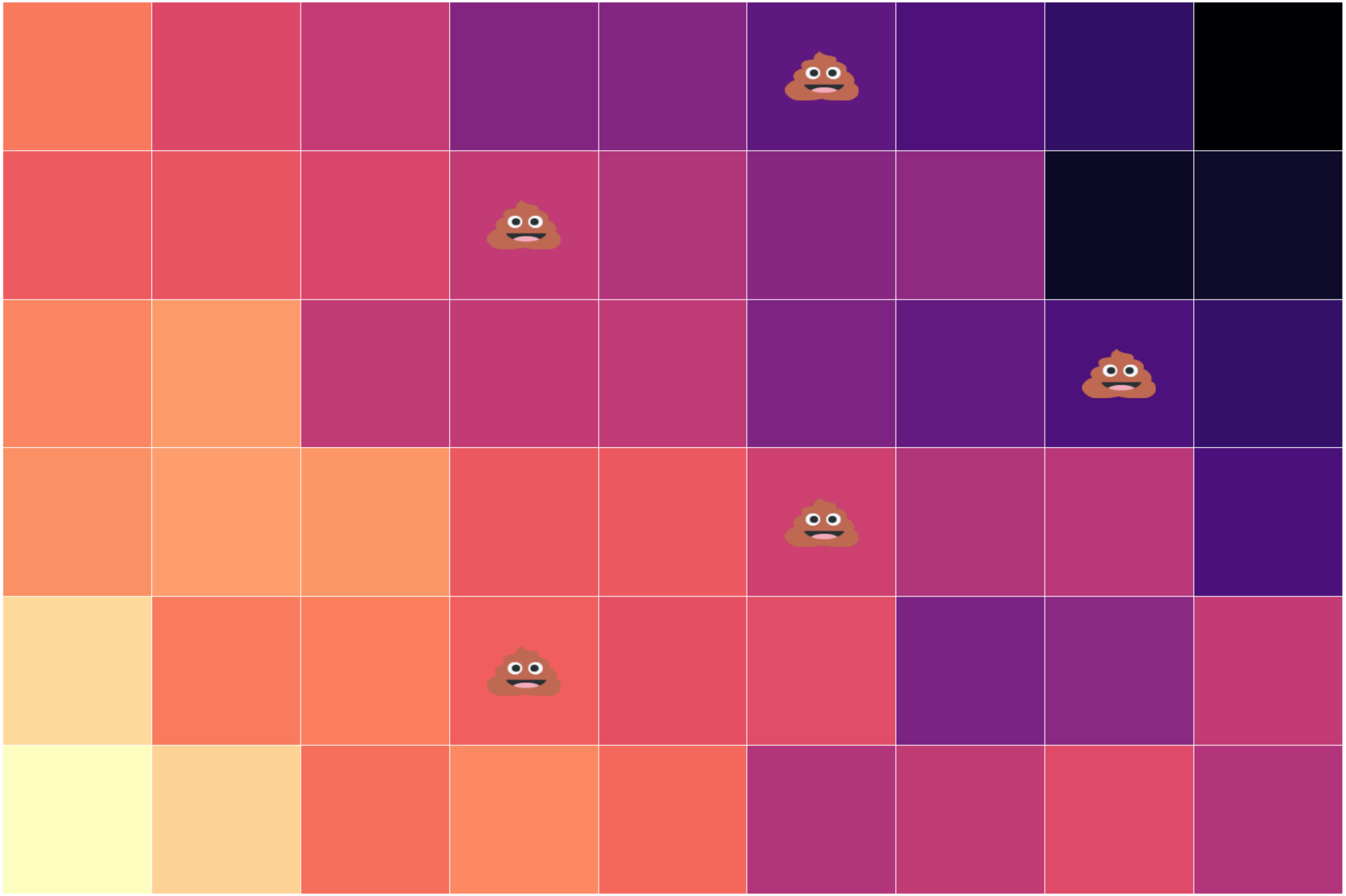
54 equal-sized plots of varying quality plus randomly assigned treatment



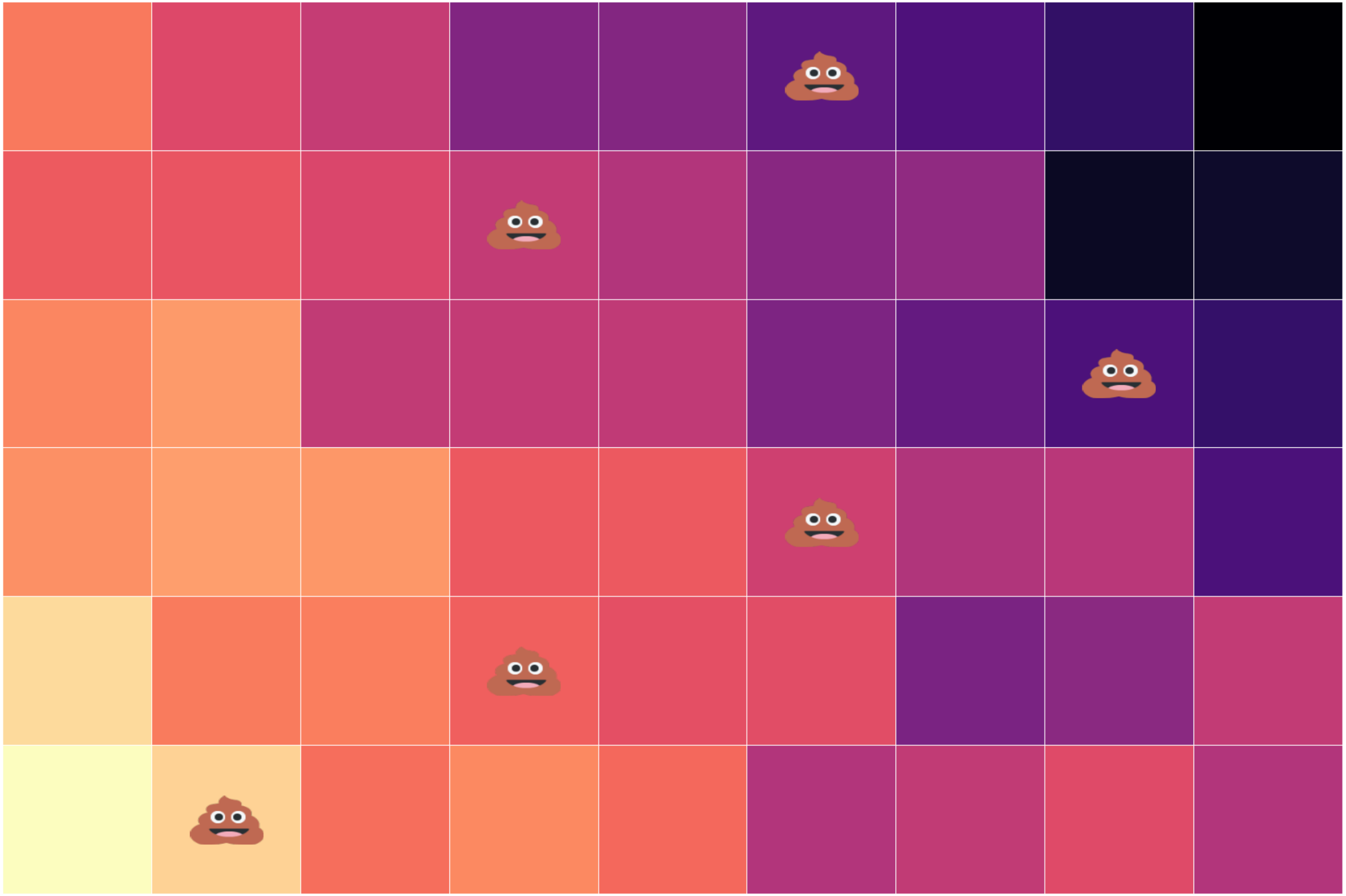
54 equal-sized plots of varying quality plus randomly assigned treatment



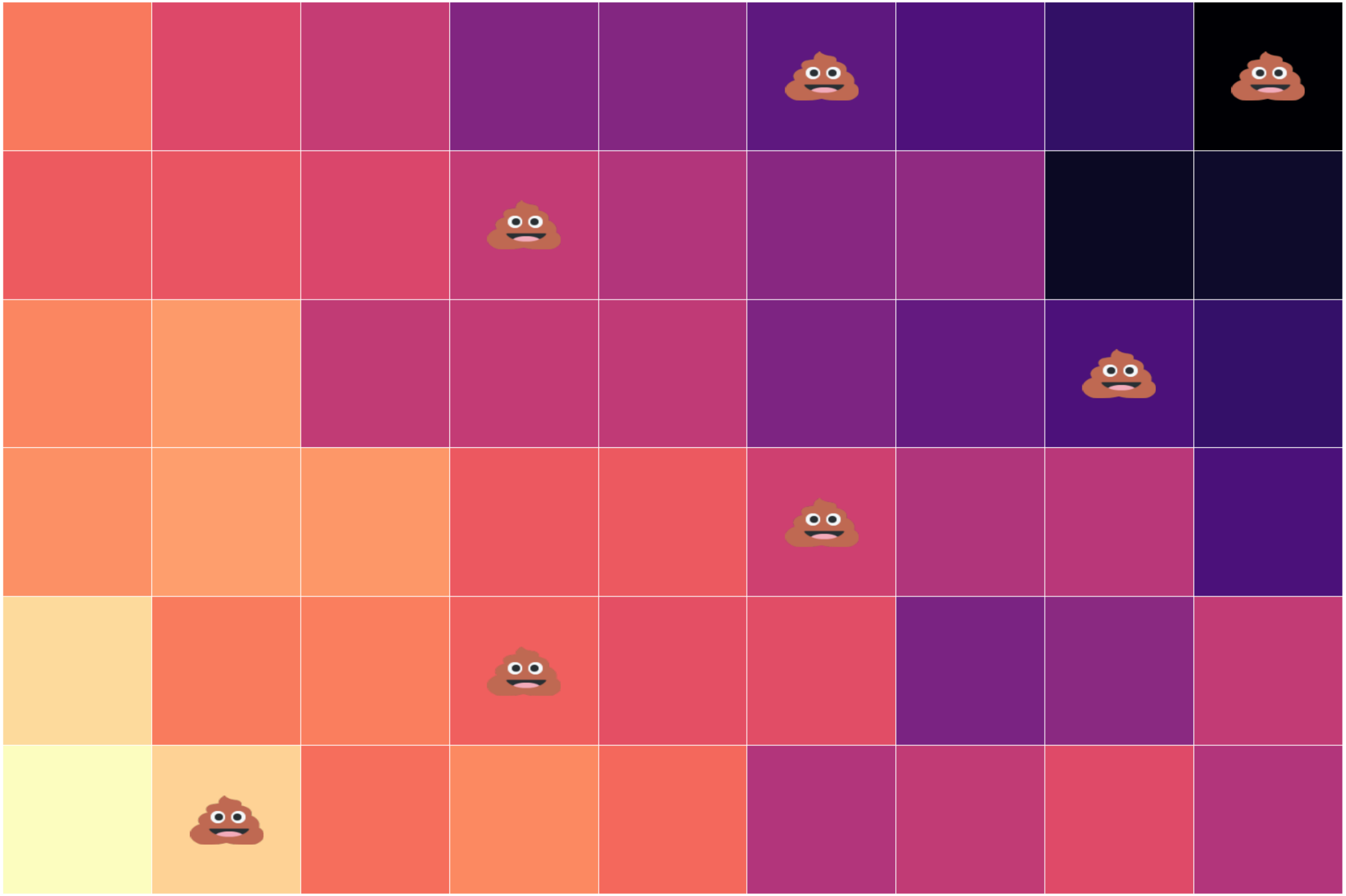
54 equal-sized plots of varying quality plus randomly assigned treatment



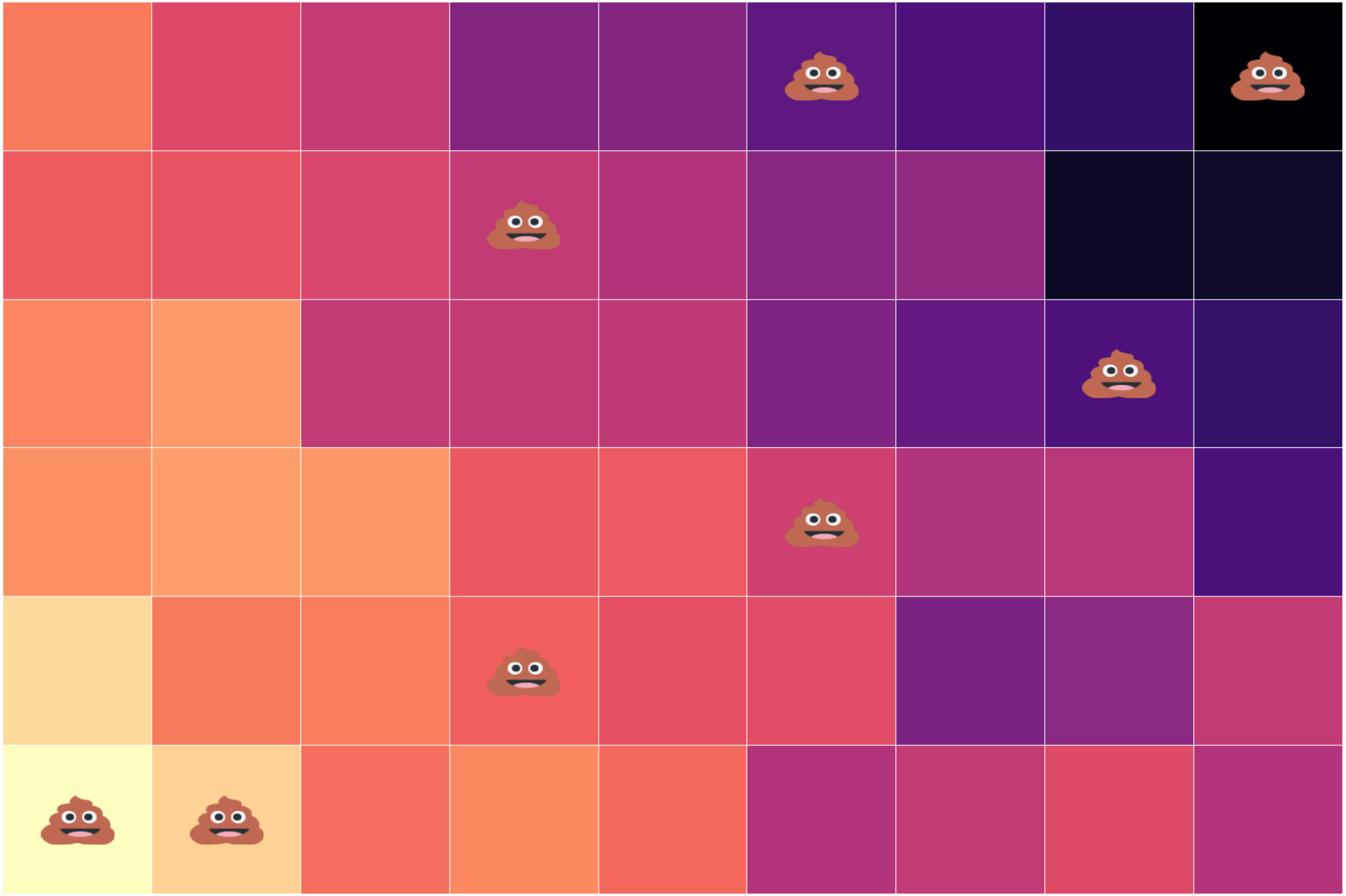
54 equal-sized plots of varying quality plus randomly assigned treatment



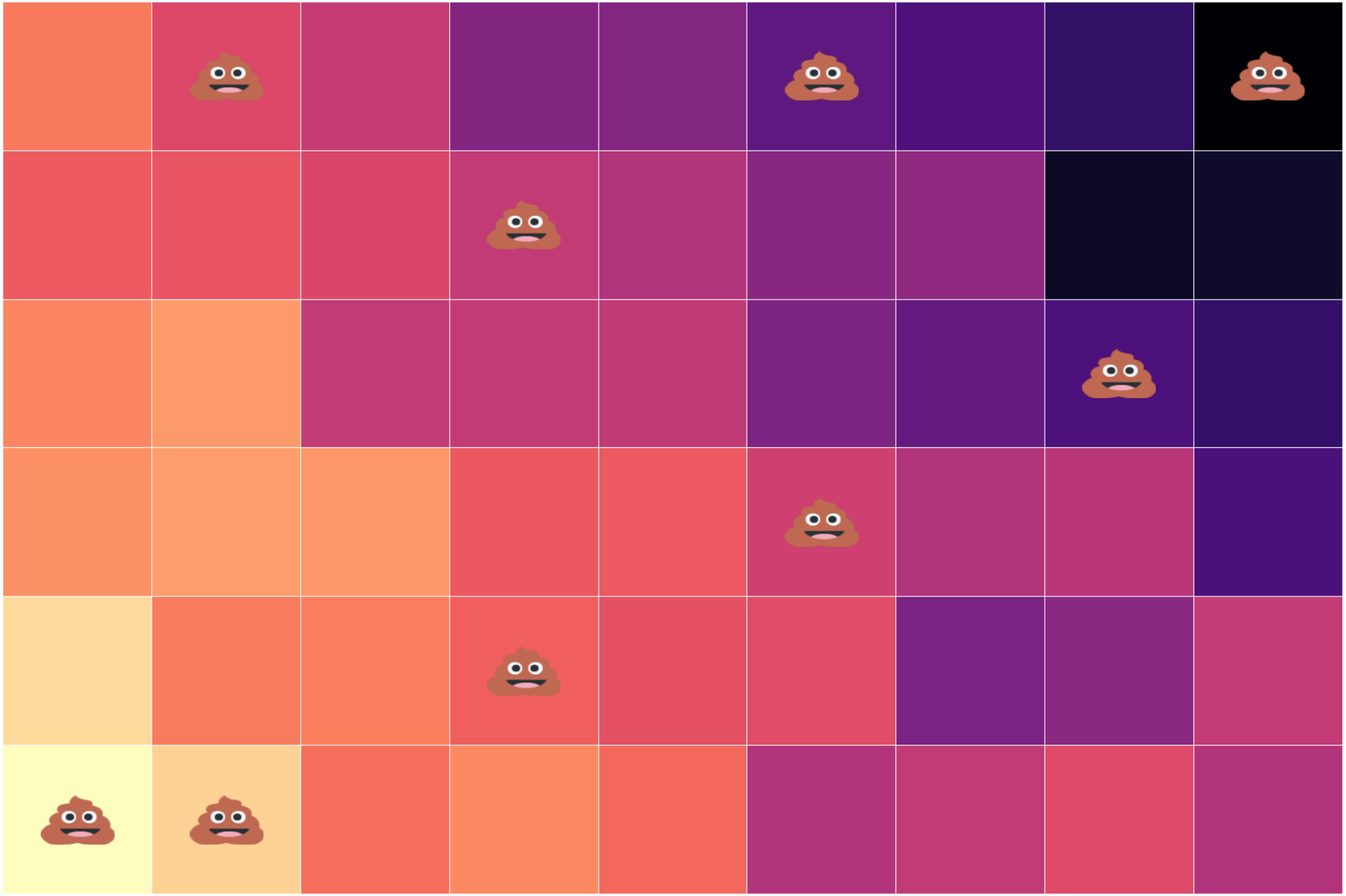
54 equal-sized plots of varying quality plus randomly assigned treatment



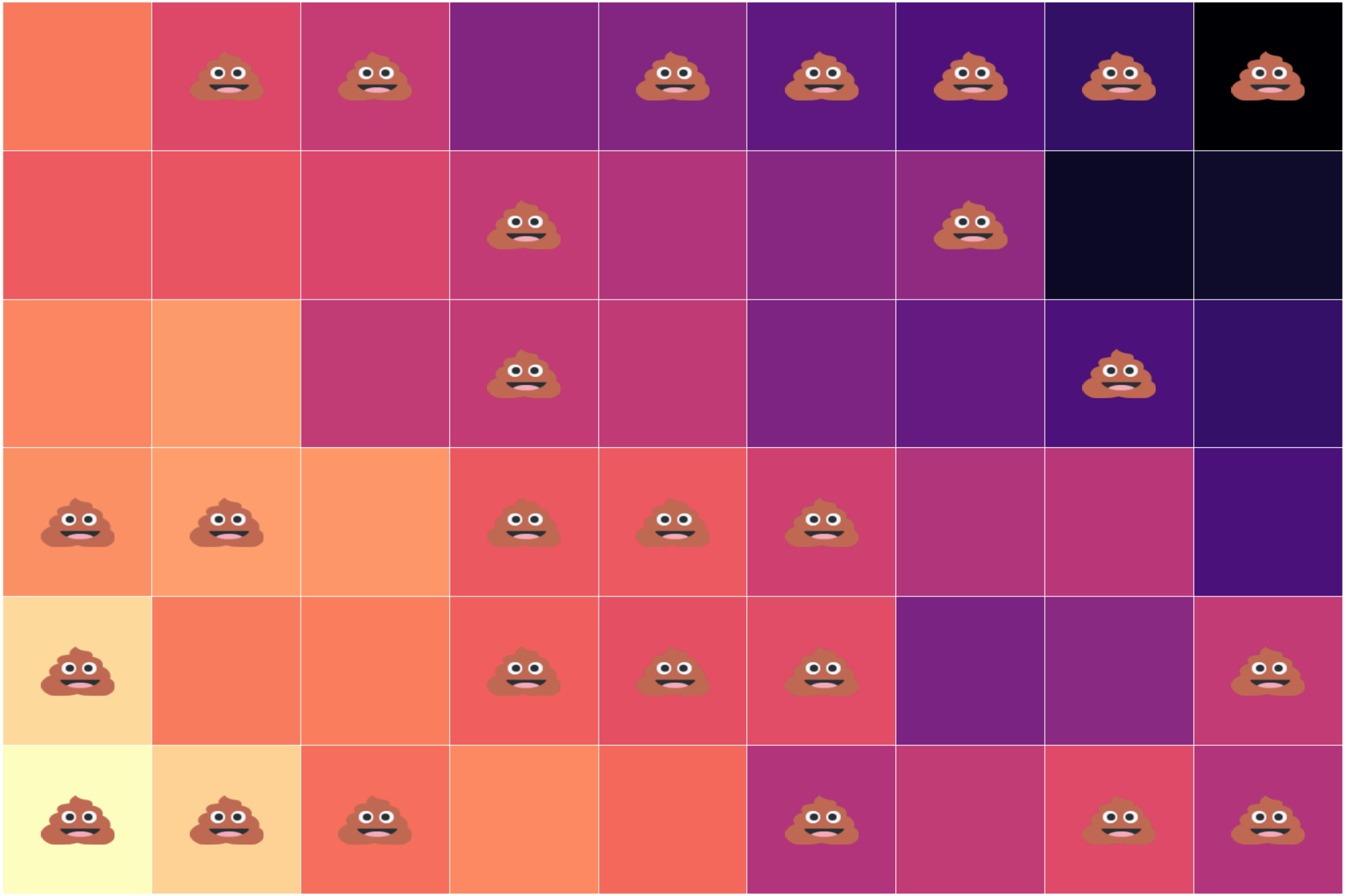
54 equal-sized plots of varying quality plus randomly assigned treatment



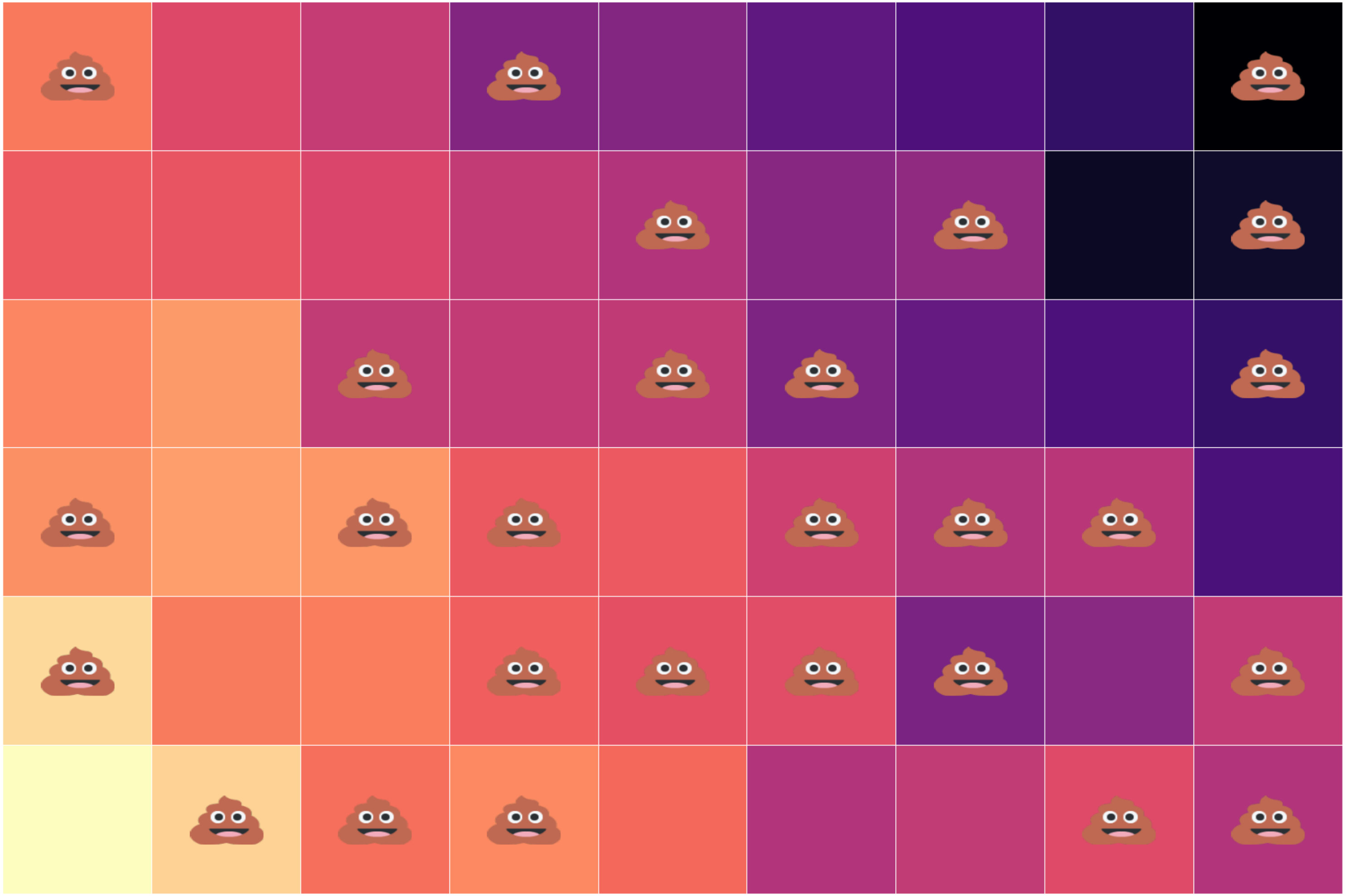
54 equal-sized plots of varying quality plus randomly assigned treatment



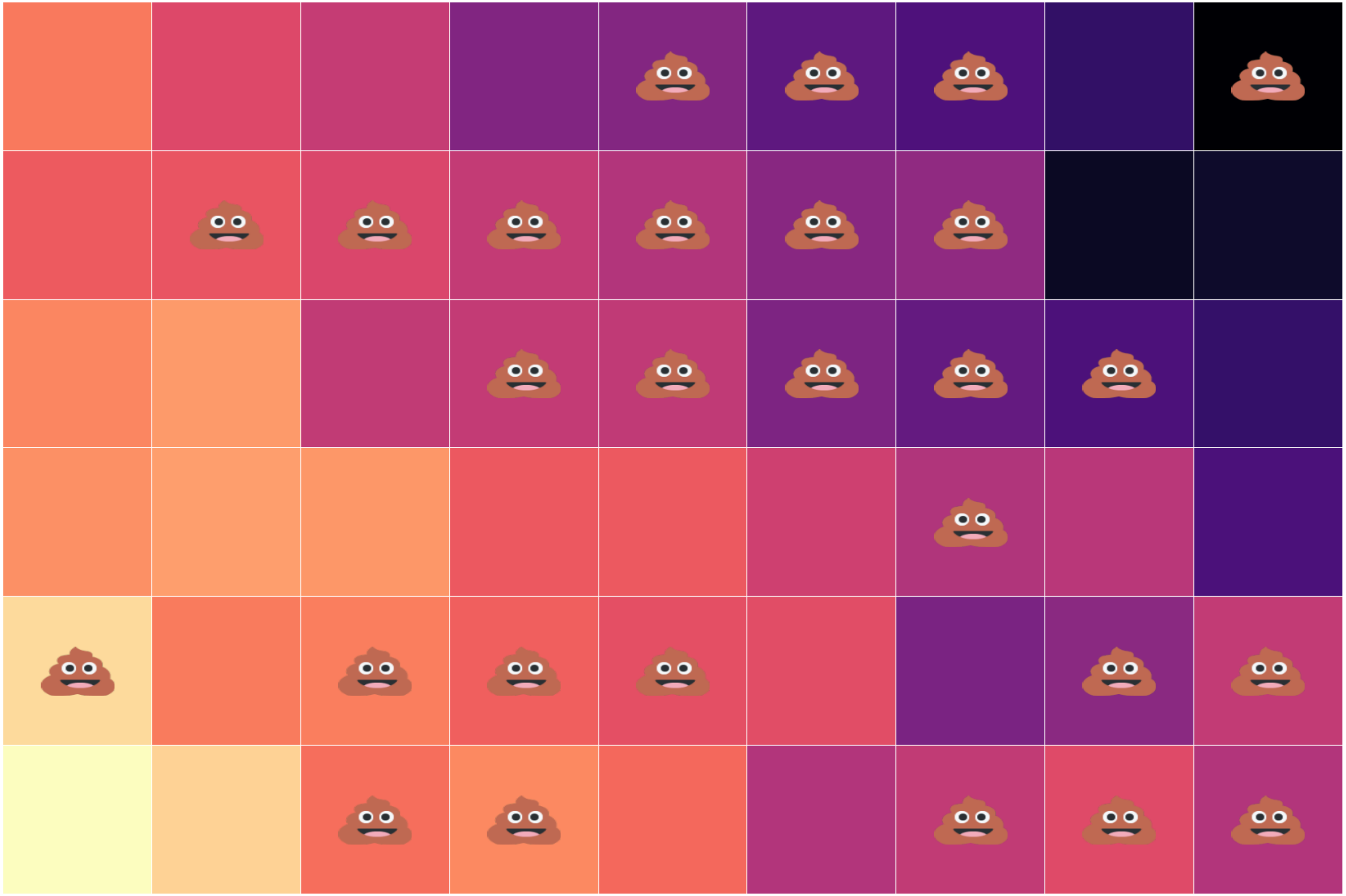
54 equal-sized plots of varying quality plus randomly assigned treatment



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 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

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 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

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 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

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).

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 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

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?

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 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

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).

Causal Effects of Neighborhoods vs. Sorting

- Two very different explanations for variation in children's outcomes across areas

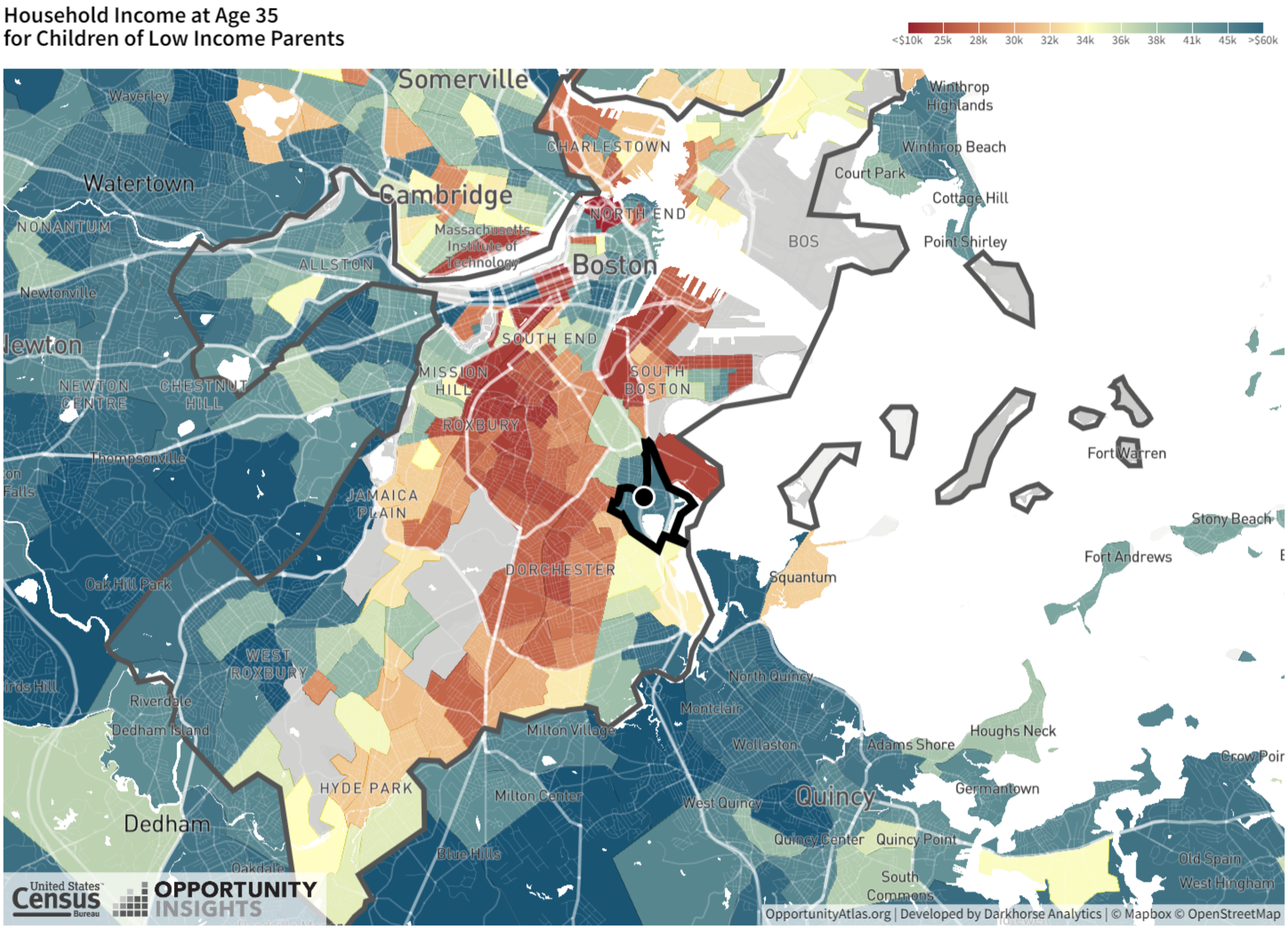
  1. Sorting: different people live in different places

  2. Causal effects: places have a causal effect on upward mobility for a given person

Identifying Causal Effects of Neighborhoods

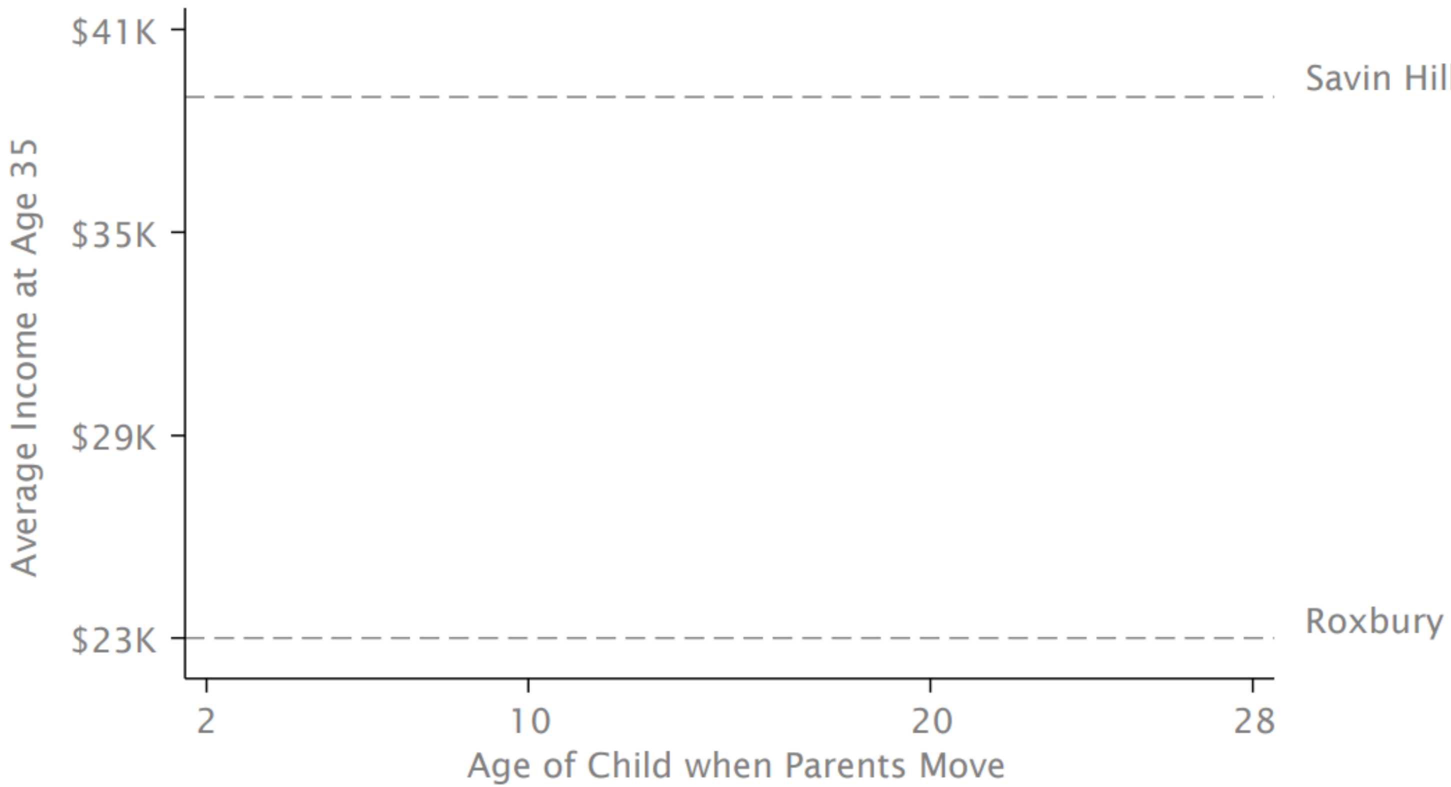
- Ideal experiment: randomly assign children to neighborhoods and compare outcomes in adulthood
  - Any issues with this?
- How can we approximate this same thing?
- Chetty and Hendren (2018) use a **quasi-experimental** design:
  - Sample of 3 million families that move across Census tracts
  - Key idea: exploit variation in the *age of child* when the family moves to identify causal effects of neighborhood

Moving a short distance in Boston



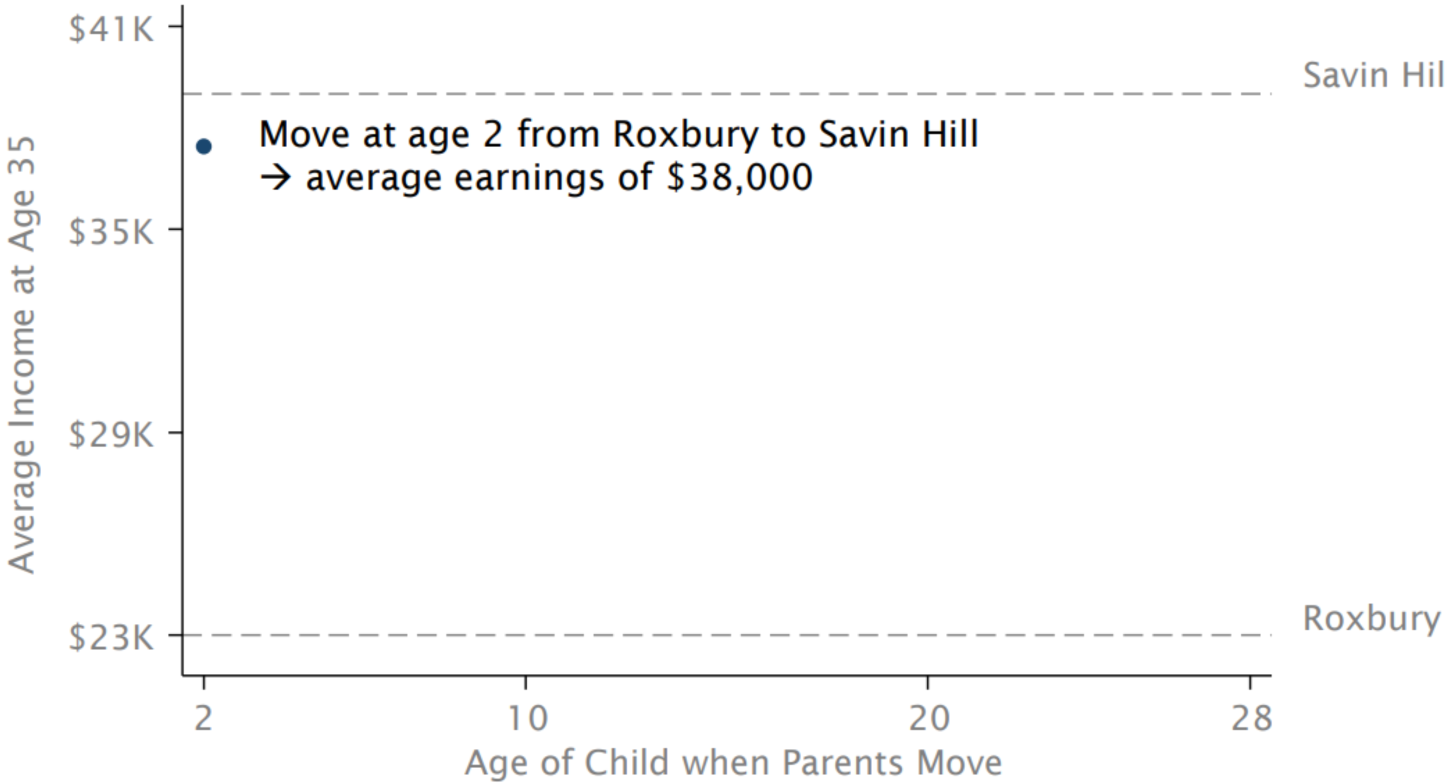
Opportunity Atlas of MA: Savin Hill outlined, Roxbury nextdoor.

Moving to a Higher Mobility Neighborhood and Income



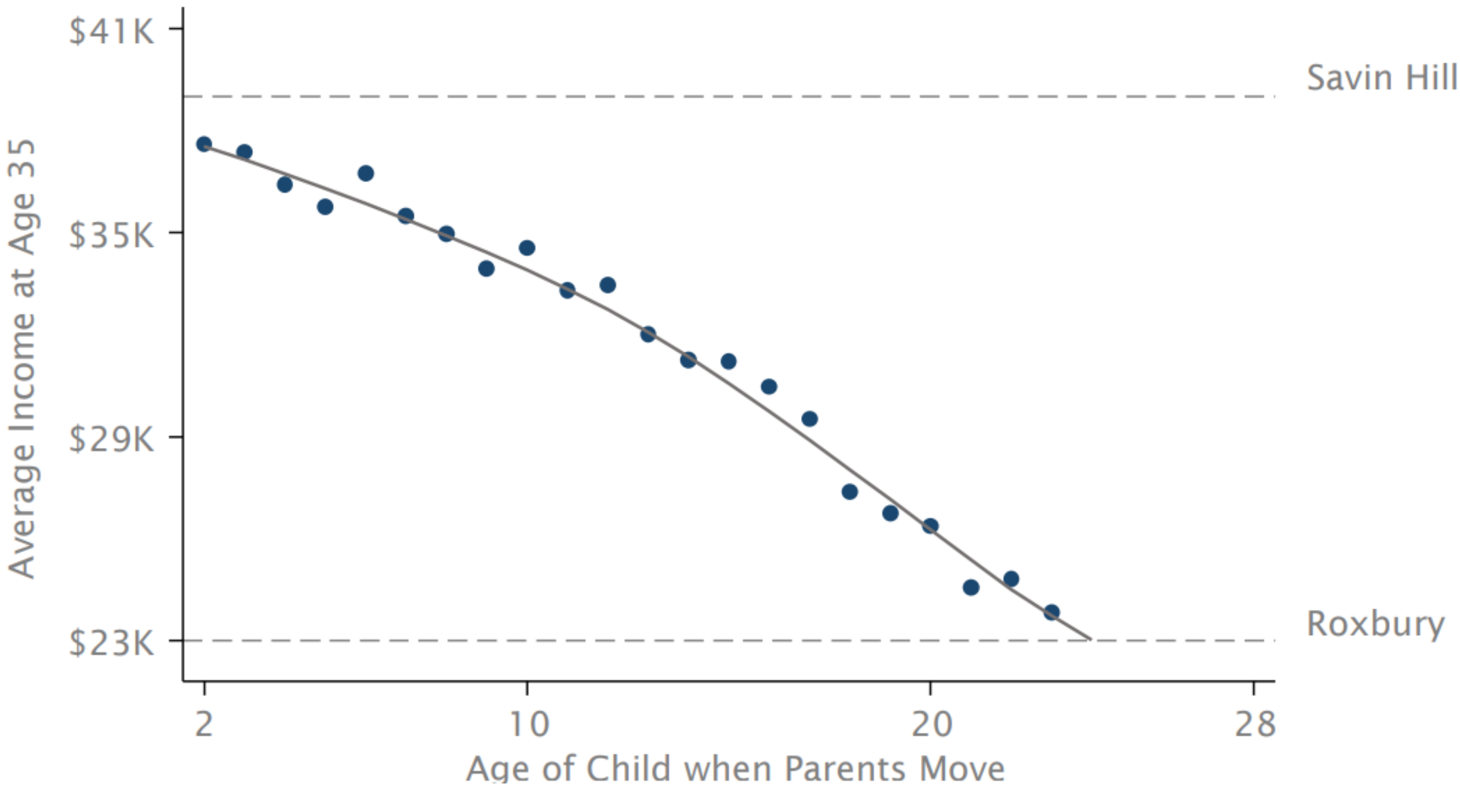
Chetty and Hendren (2018).

Moving to a Higher Mobility Neighborhood and Income



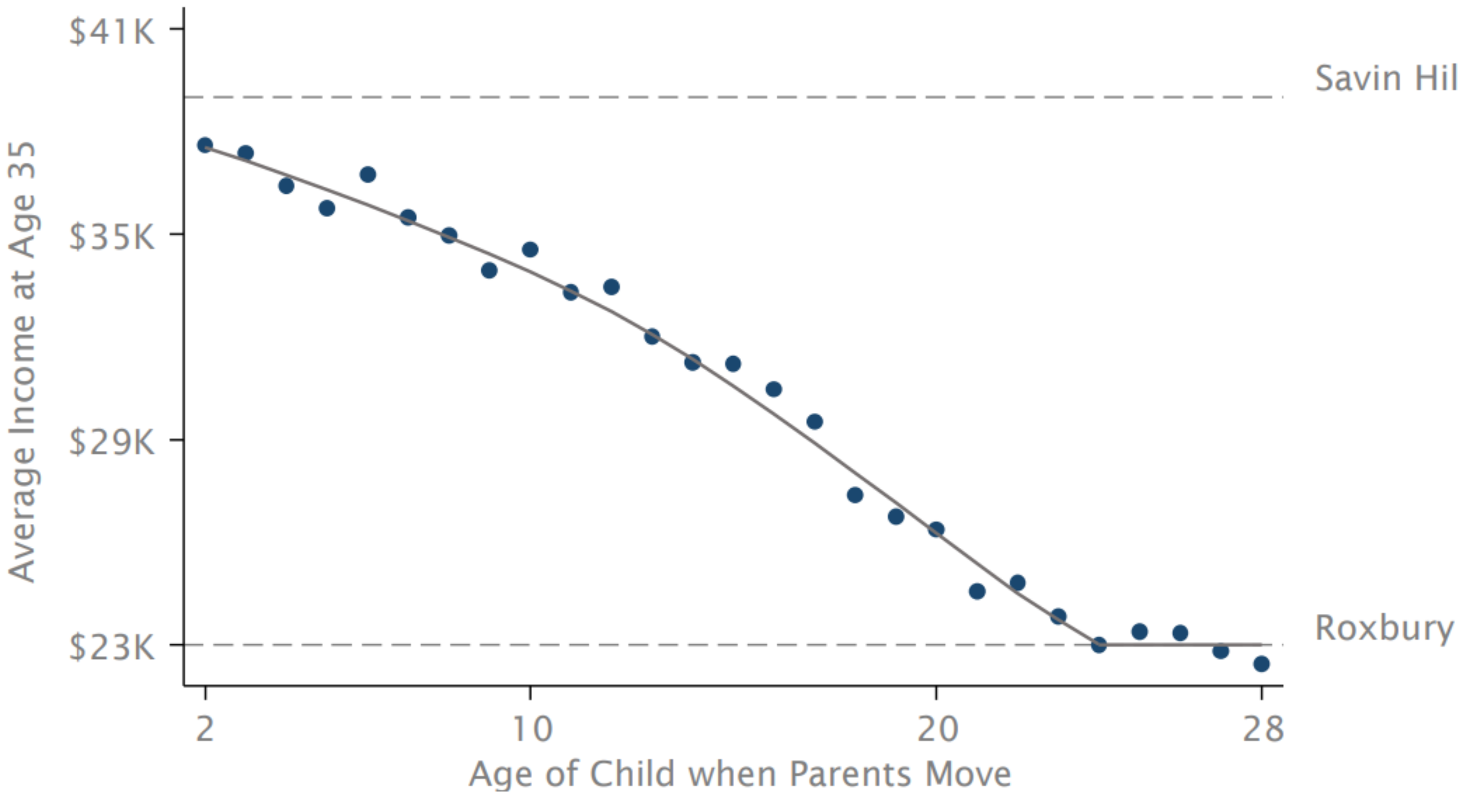
Chetty and Hendren (2018).

Moving to a Higher Mobility Neighborhood and Income



Chetty and Hendren (2018).

Moving to a Higher Mobility Neighborhood and Income



Chetty and Hendren (2018).

We have to make some assumptions

- All causal work requires assumptions
- Key assumption: *timing* of moves between areas is unrelated to other determinants of a child's outcomes
- Why might this not hold?

We have to make some assumptions

- All causal work requires assumptions
- Key assumption: *timing* of moves between areas is unrelated to other determinants of a child's outcomes
- Why might this not hold?

    1. Parents who move to good areas when their children ar young might be different from those who move later
    2. Moving may be related to other factors (e.g., change in parents' job) that affect children directly

"Testing" assumptions

- You cannot fully test assumptions, but you can look for evidence they are violated

- Two approaches to evaluate validity of timing of move assumption:

  1. Compare siblings' outcomes to control for family "fixed" effects

  2. Use differences in neighborhood effects across subgroups to implement "placebo" tests

  - Ex: some places (e.g. low-crime areas) have better outcomes for boys than girls

  - Move to place where boys have higher earnings --> son improves in proportion to exposure, but not daughter

- Conclude that ~2/3 of variation in upward mobility across areas is due to causal effects of neighborhoods

Next lecture: Fixed effects and difference-in-differences