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 next few weeks, 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 ≠ 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 ≠ 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.

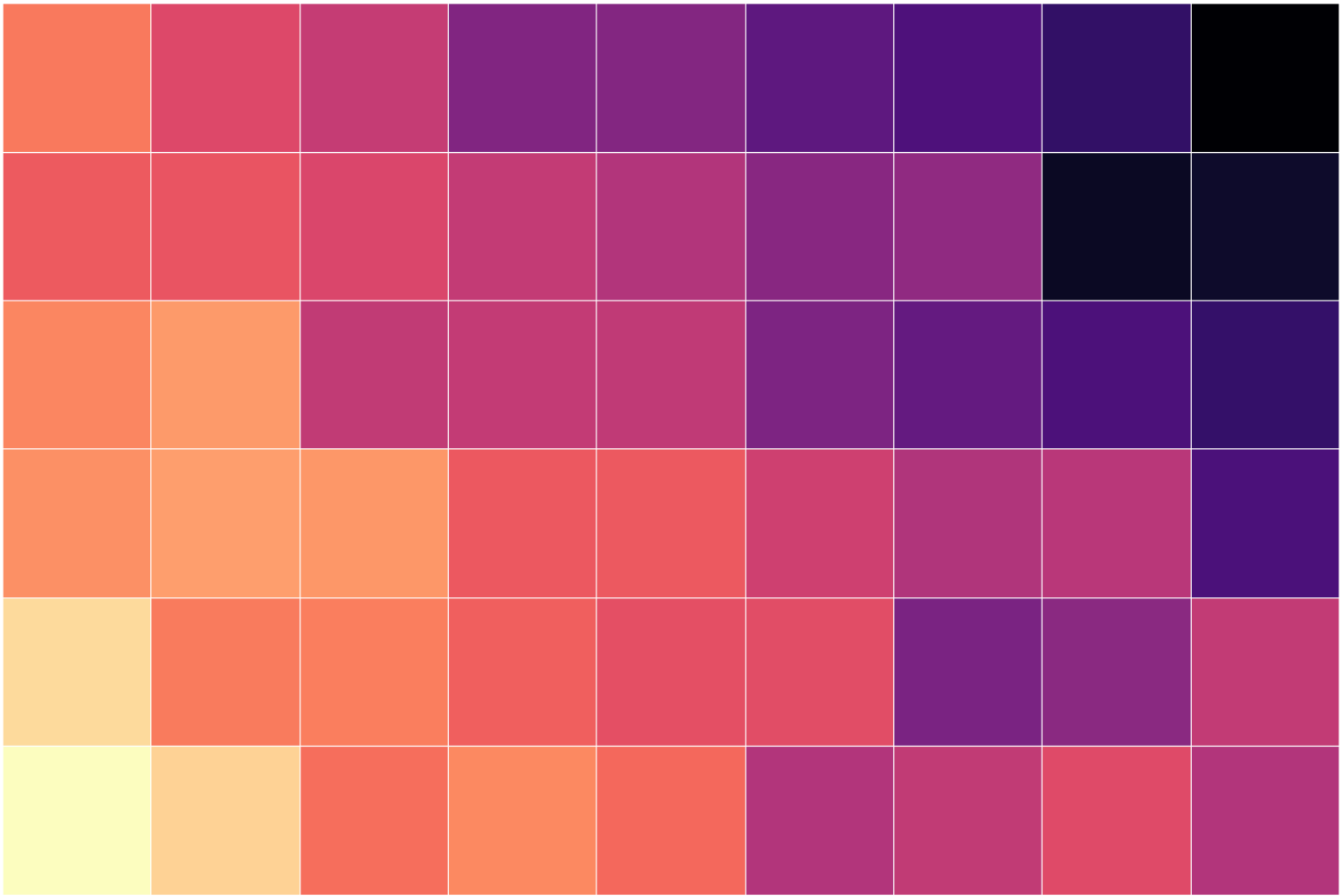
All else equal!

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

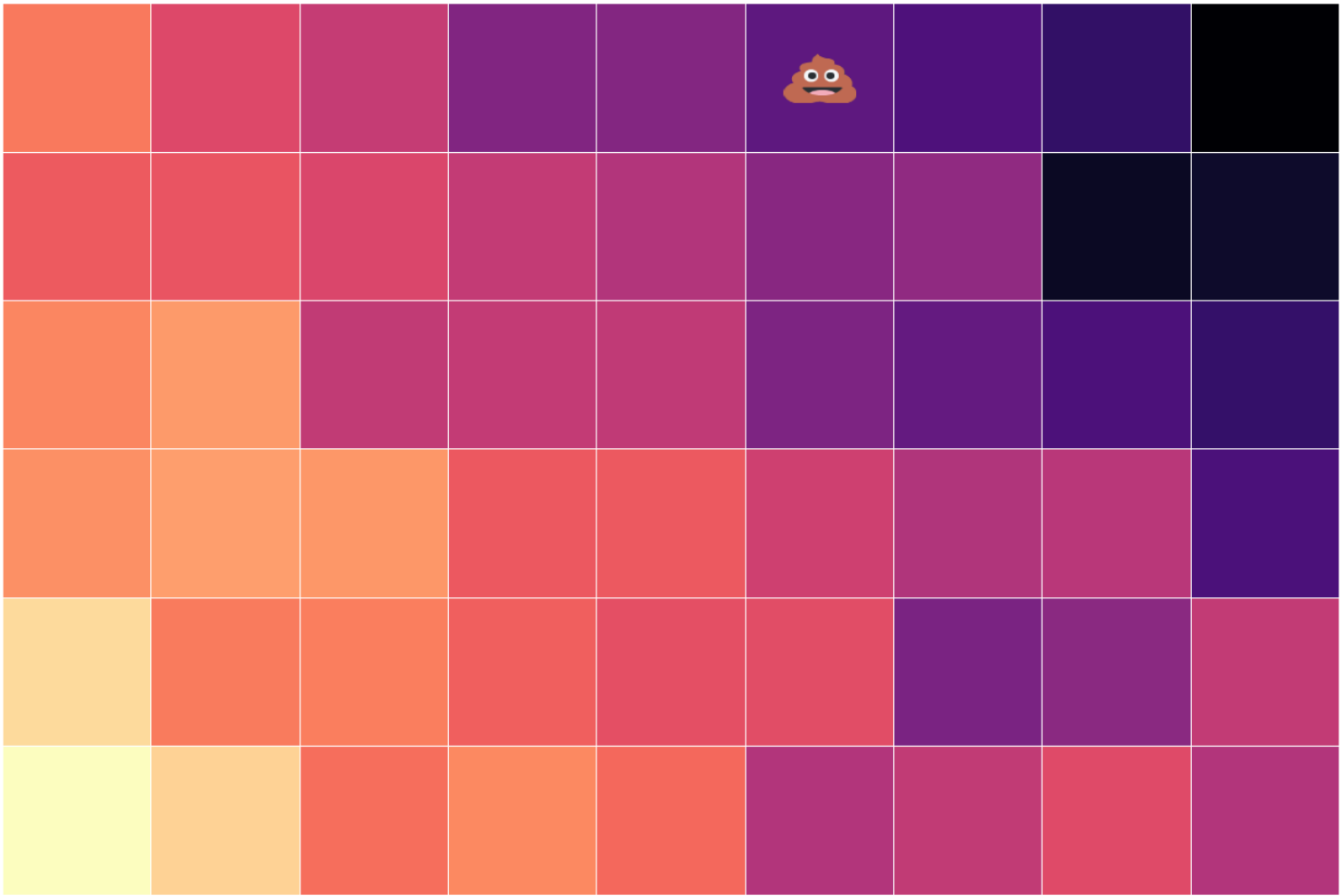
54 equal-sized plots

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

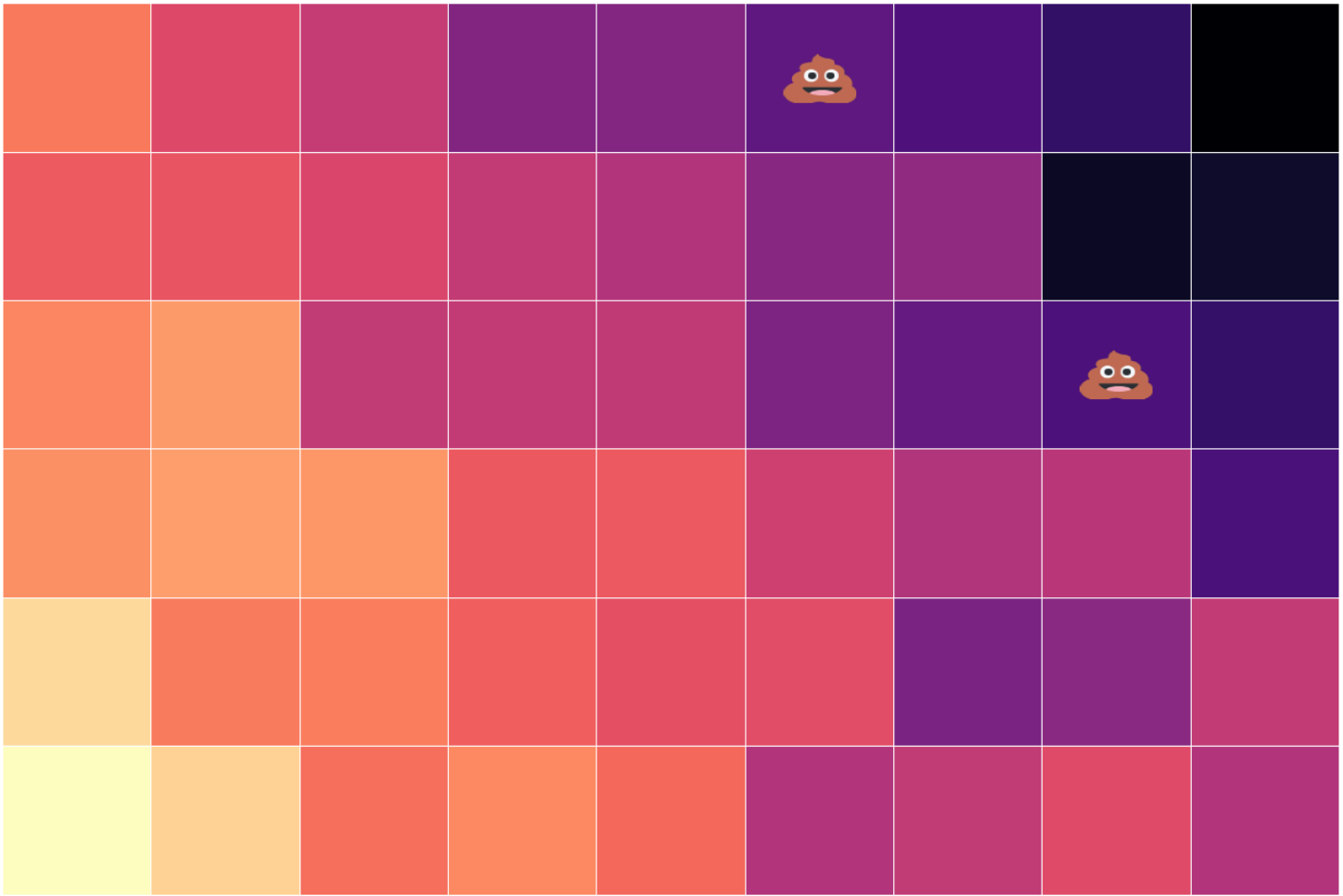
54 equal-sized plots of varying quality



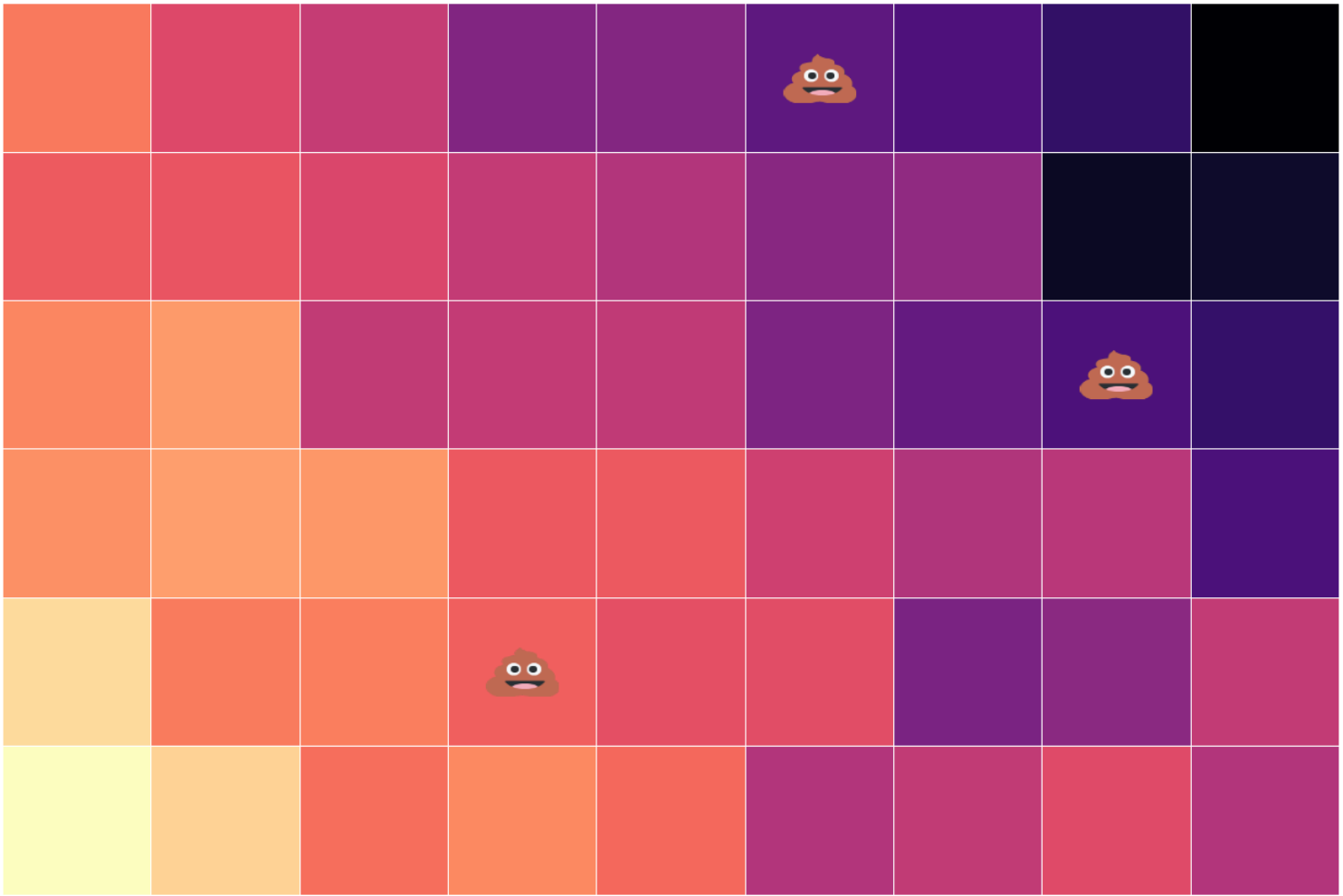
54 equal-sized plots of varying quality plus randomly assigned treatment



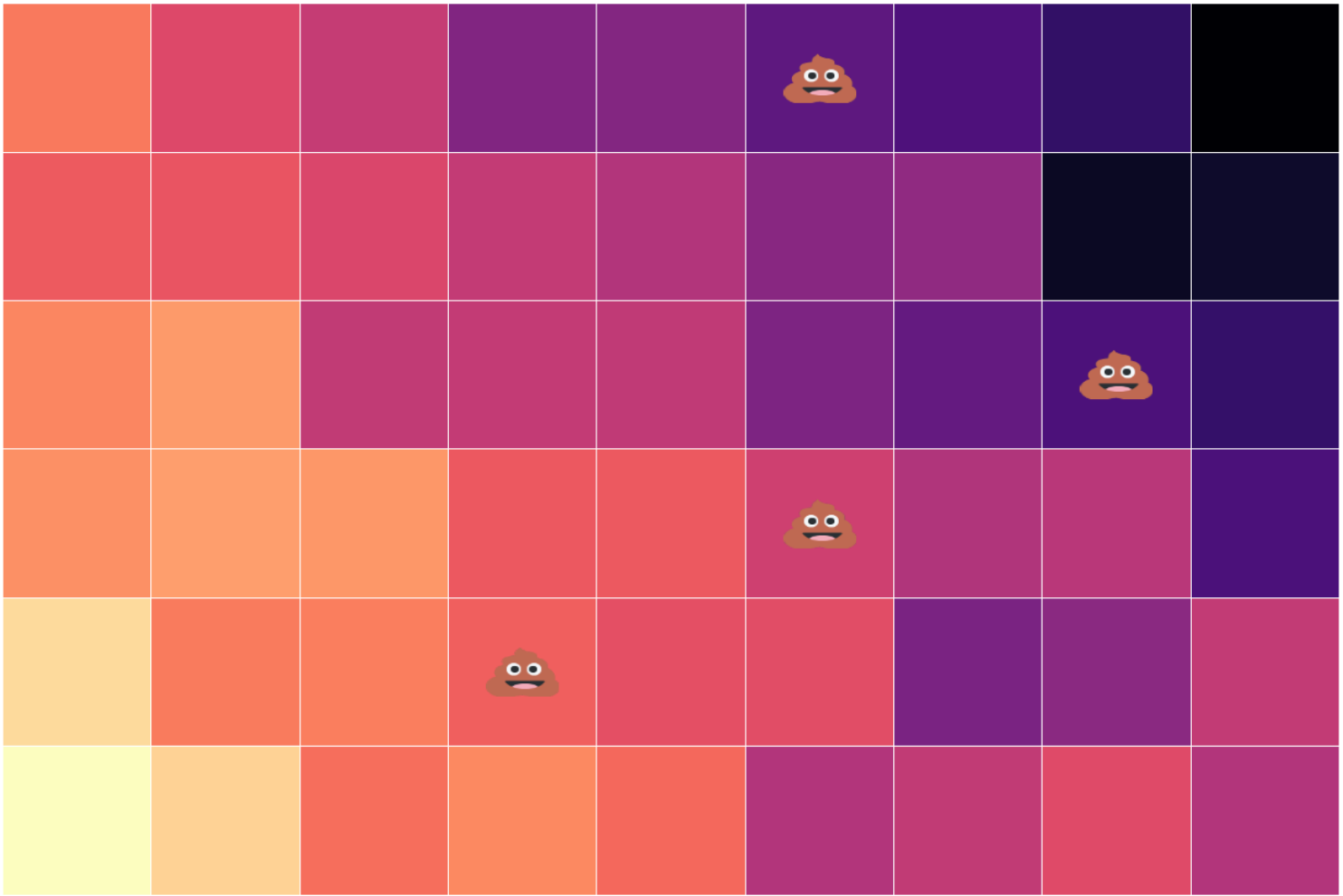
54 equal-sized plots of varying quality plus randomly assigned treatment



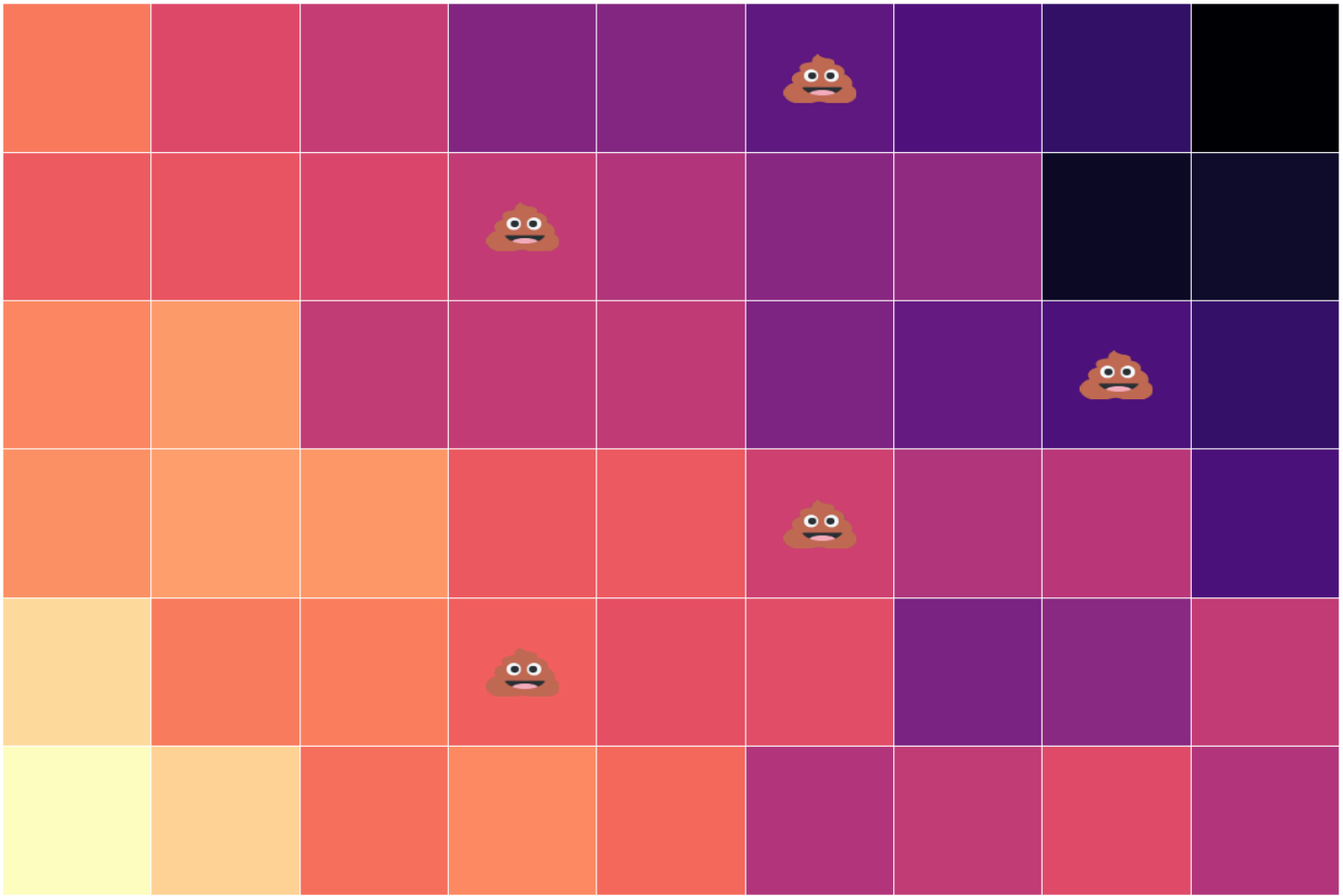
54 equal-sized plots of varying quality plus randomly assigned treatment



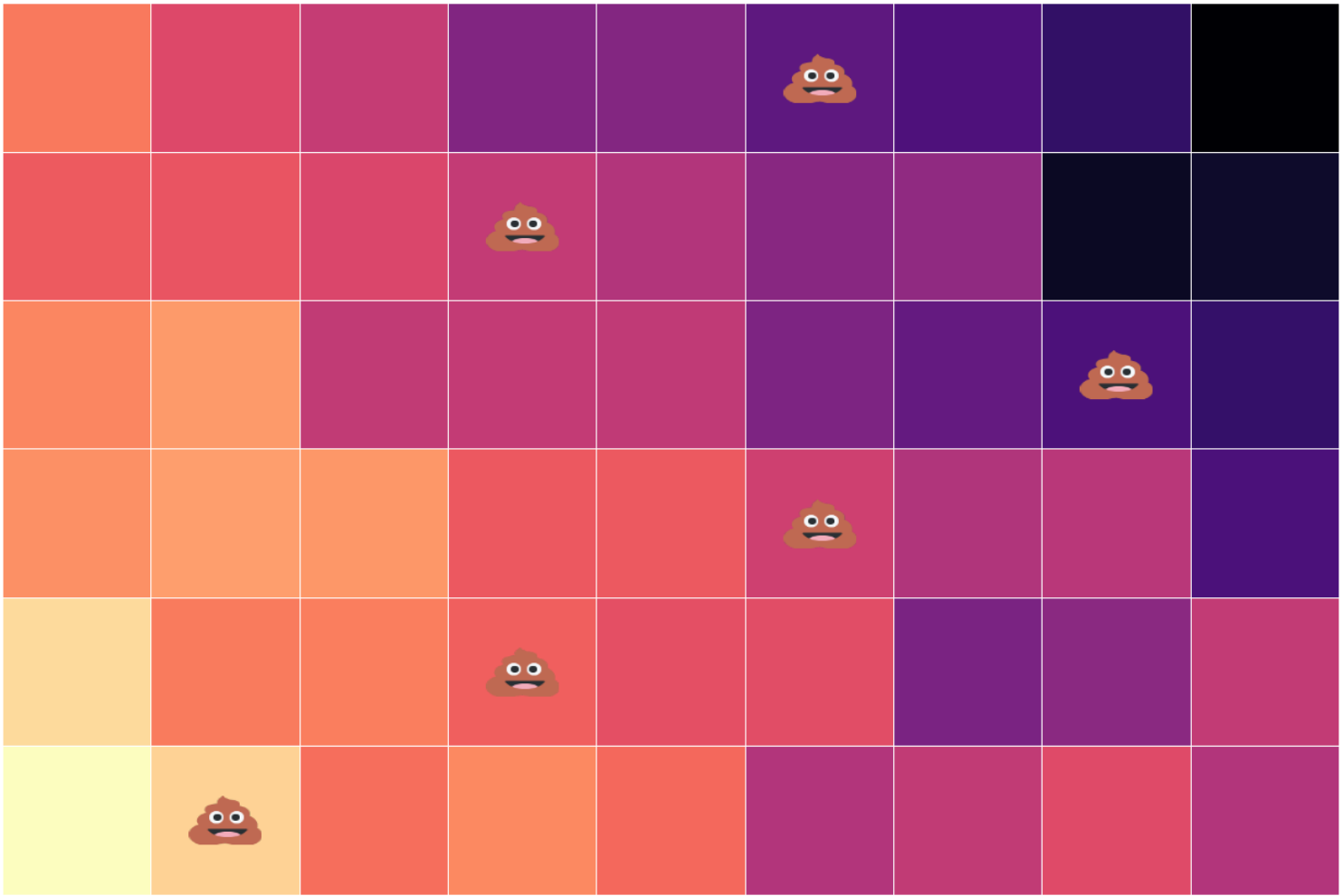
54 equal-sized plots of varying quality plus randomly assigned treatment



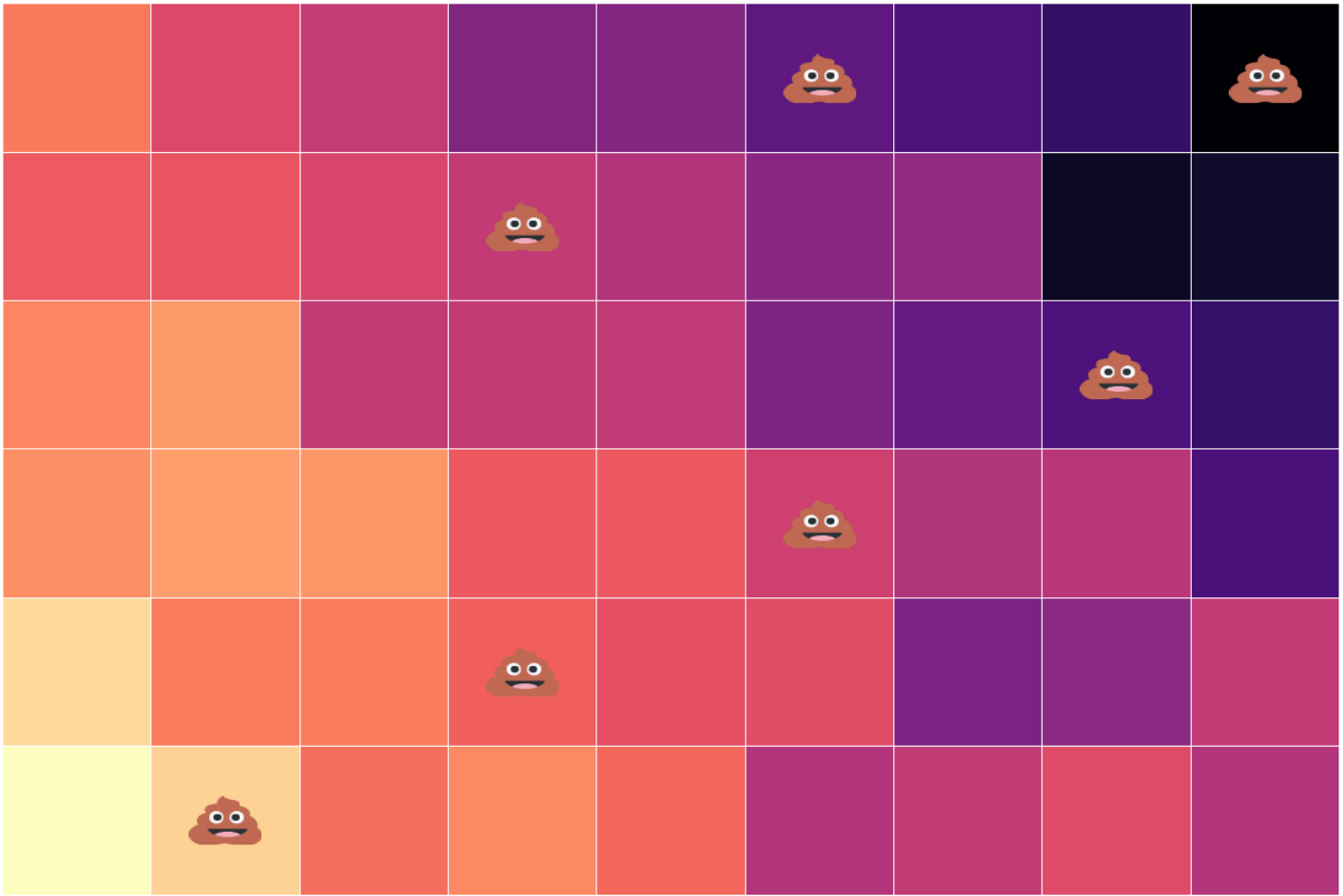
54 equal-sized plots of varying quality plus randomly assigned treatment



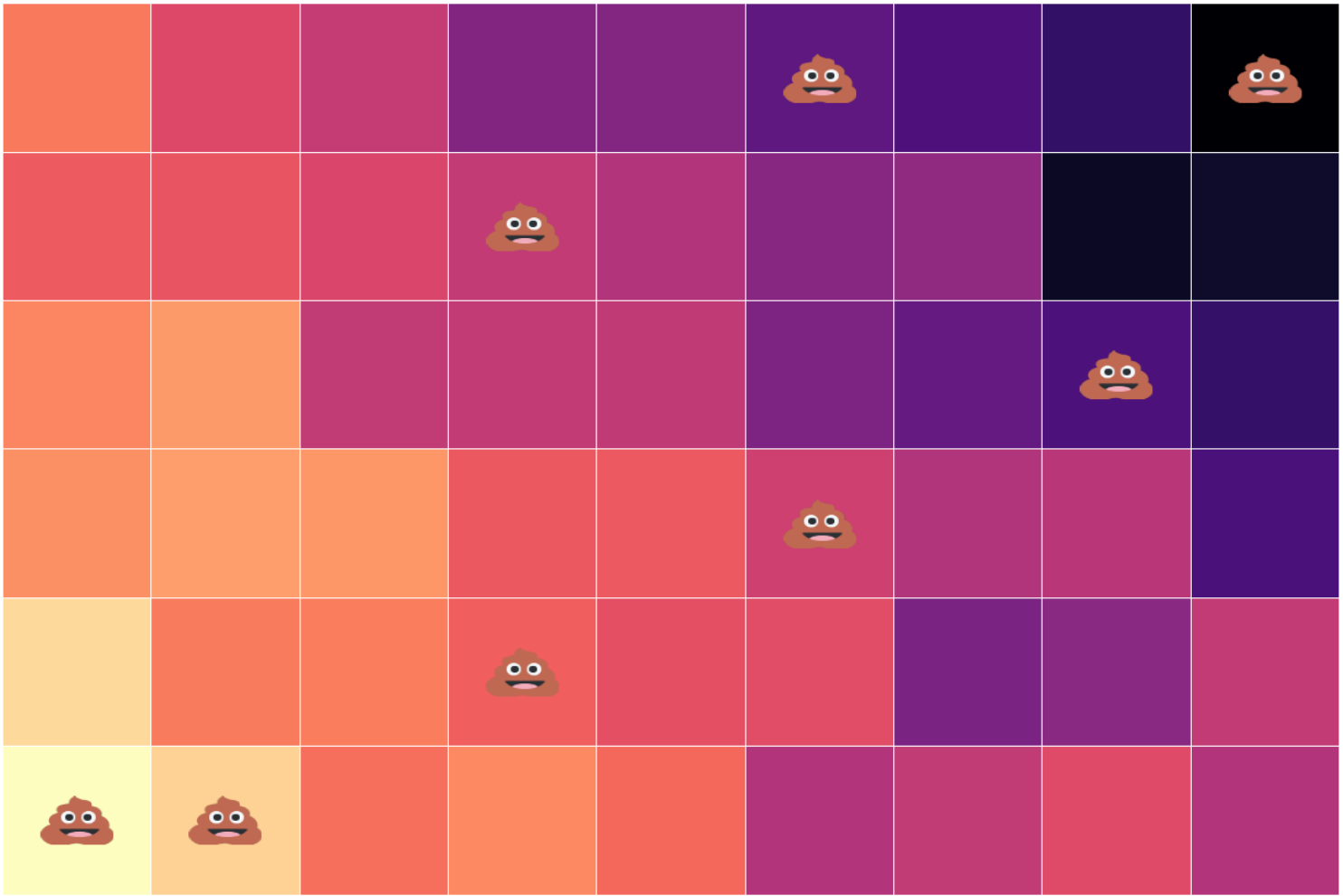
54 equal-sized plots of varying quality plus randomly assigned treatment



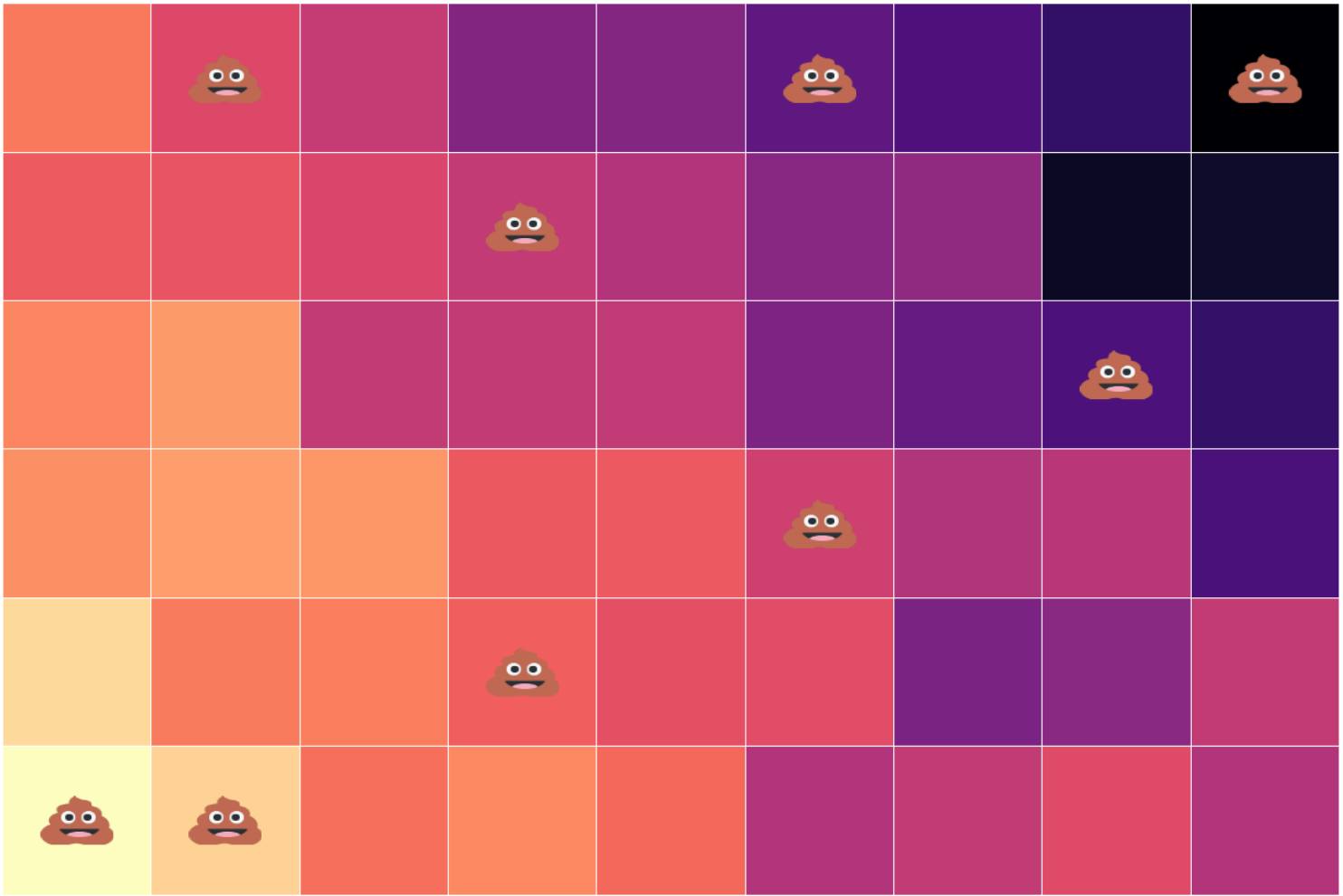
54 equal-sized plots of varying quality plus randomly assigned treatment



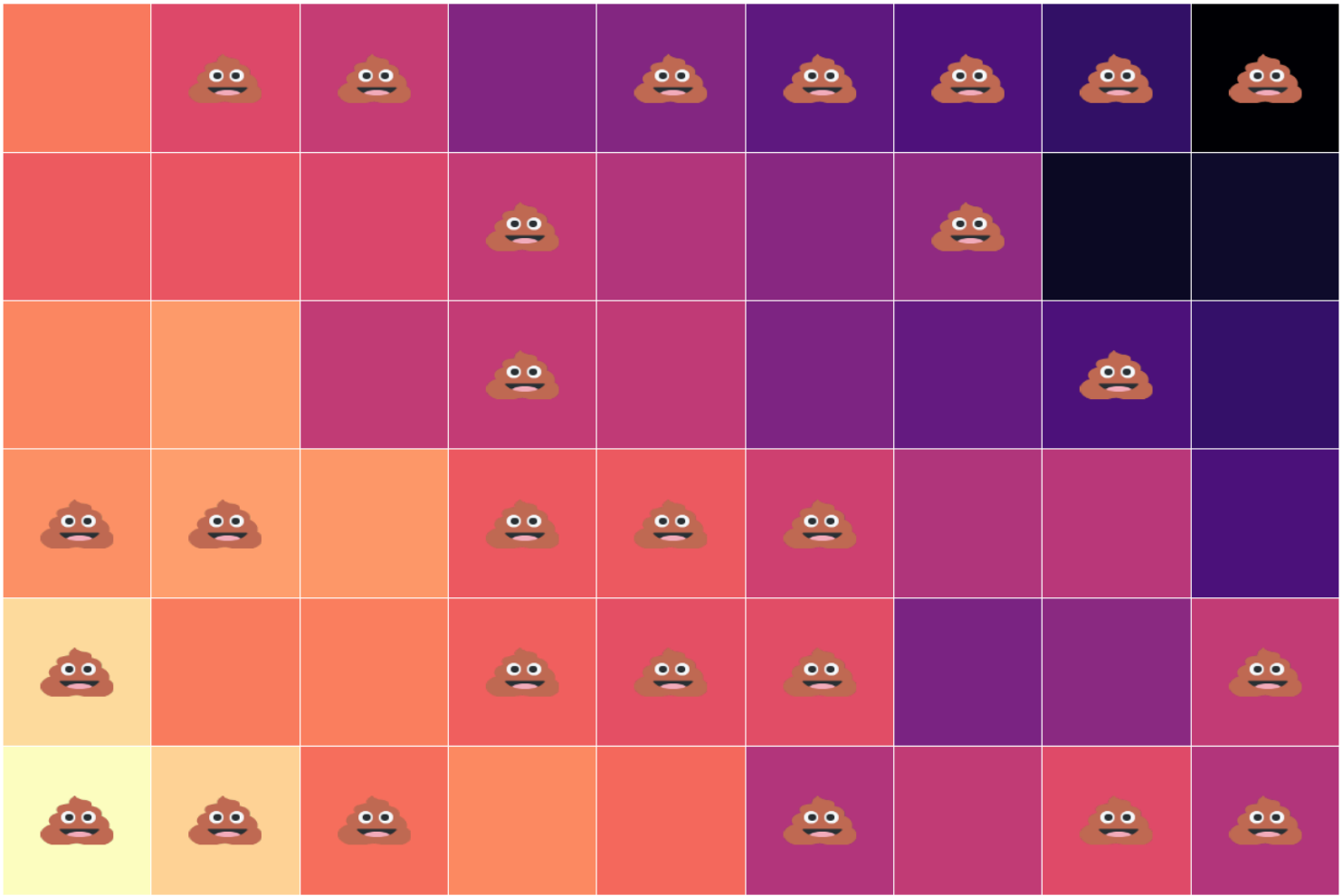
54 equal-sized plots of varying quality plus randomly assigned treatment



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

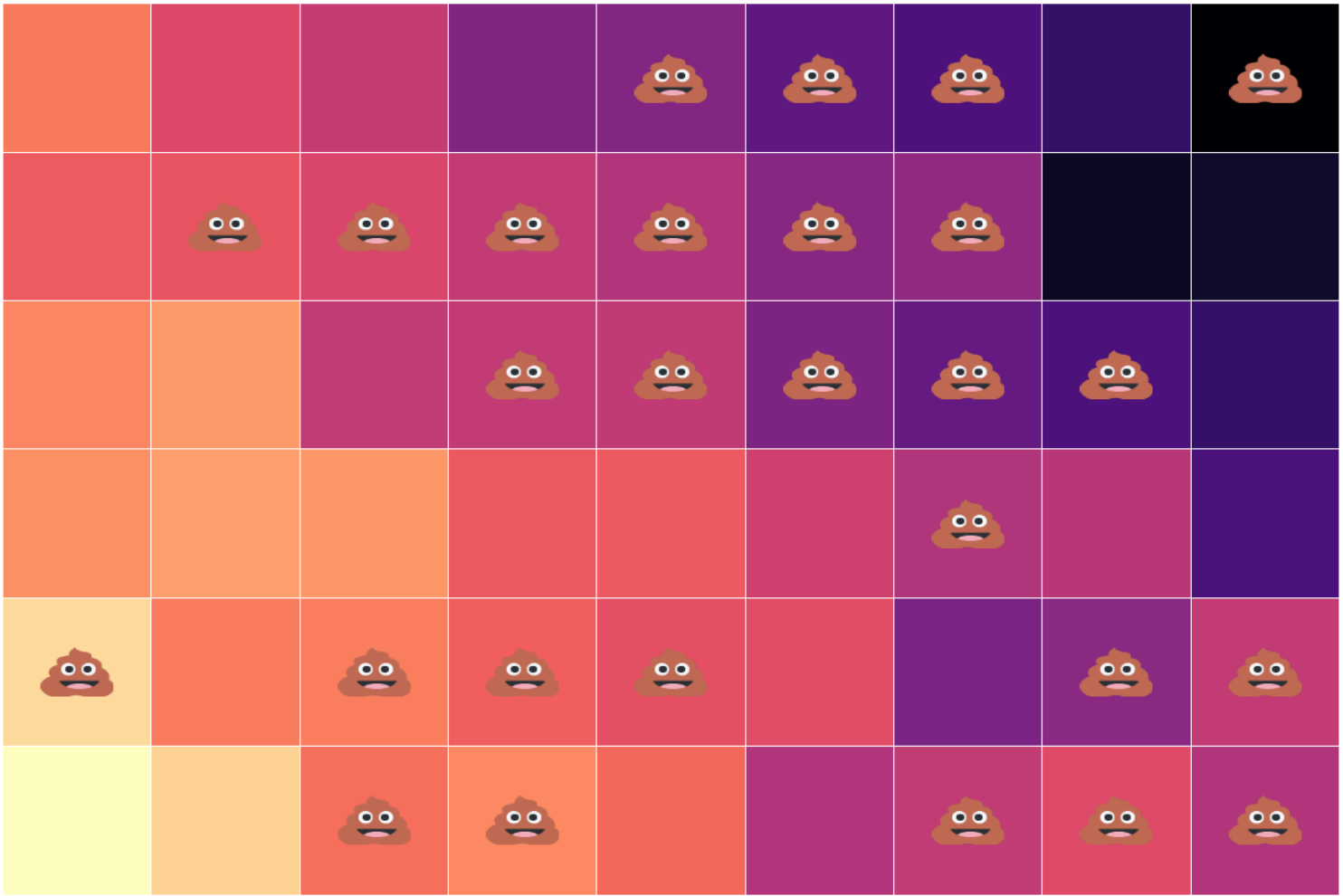


54 equal-sized plots of varying quality plus randomly assigned treatment



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 💩).

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

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 💩).

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

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 💩).

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

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).

Example 2: The Causal Effect of Neighborhoods

- Fertilizer is relatively easy to randomize and run an RCT
- But most policy questions are not so easy to answer with an RCT
- Why?
  - "Noisy" economic outcomes require larger sample sizes, but it is infeasible to randomize at scale
  - Also, unethical to randomize certain types of characteristics
- When we cannot experiment happens, we look for **quasi-experimental** designs

Quasi-experimental designs

- Quasi-experimental designs are not true experiments, but they can be used to estimate causal effects
- Key idea: **exploit** some kind of **exogenous** variation in the explanatory variable of interest
    - If the variation is exogenous, then we can use it to estimate causal effects
    - Why? It is uncorrelated with the error term in the regression
    - We have **as good as** random assignment because the assignment is unrelated to other factors that affect the outcome
    - Intuition:
- Sometimes the term "natural" experiments is used to introduce a quasi-experimental design
    - This means that the as good as random assignment was created by

Quasi-experiments and assumptions

- Every quasi-experimental design requires an assumption
- Leads to contentious debates over their validity

Quasi-experiments and assumptions

- Every quasi-experimental design requires an assumption
- Leads to contentious debates over their validity

Common Examples:

- **Difference-in-differences**: Compare outcomes in units that do and do not experience a treatment, before and after the treatment
  - Compare employment outcomes in states that change and do not change minimum wage (Card and Krueger (1993))
- **Regression Discontinuity Design**: Compare outcomes for units just above and just below some cutoff that determines a treatment
  - Compare economic outcomes for students just above and just below GPA cutoff to be admitted to college (Zimmerman (2014))
- **Instrumental variables**: Take a variable that moves the explanatory variable but is uncorrelated with the error term
  - Proximity of univerisities -> increased educational attainment -> higher earnings (Card (1995))

Causality 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

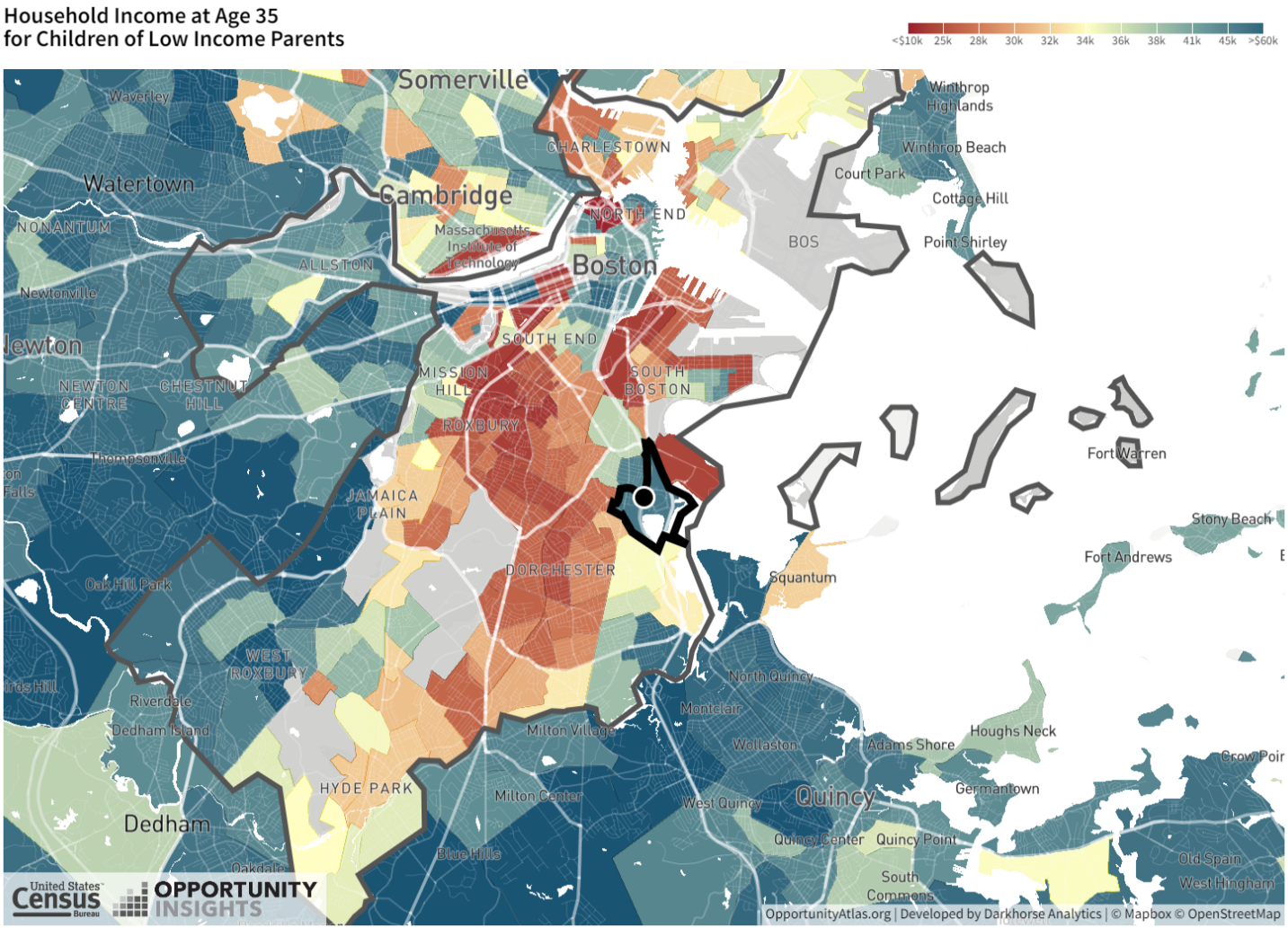
Causal Effects of Neighborhoods

- Ideal experiment: randomly assign children to neighborhoods and compare outcomes in adulthood
    - Any issues with this?
- How can we approximate this same thing?

Causal Effects of Neighborhoods

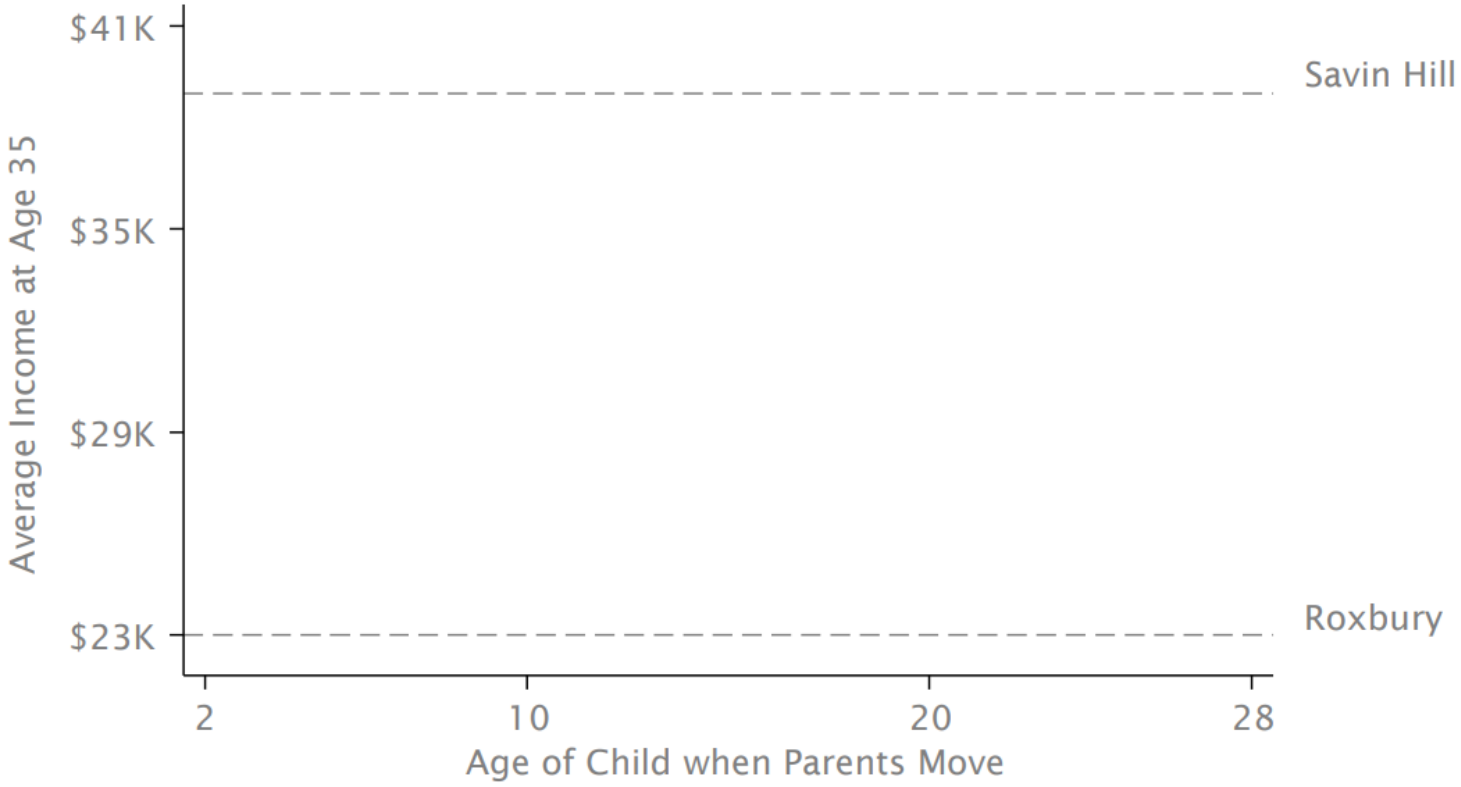
- Ideal experiment: randomly assign children to neighborhoods and compare outcomes in adulthood
    - Any issues with this?
- How can we approximate this same thing?
- Chetty and Hendren (2018) use a **quasi-experimental** design:
    - Sample of 3 million families that move across Census tracts
    - 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



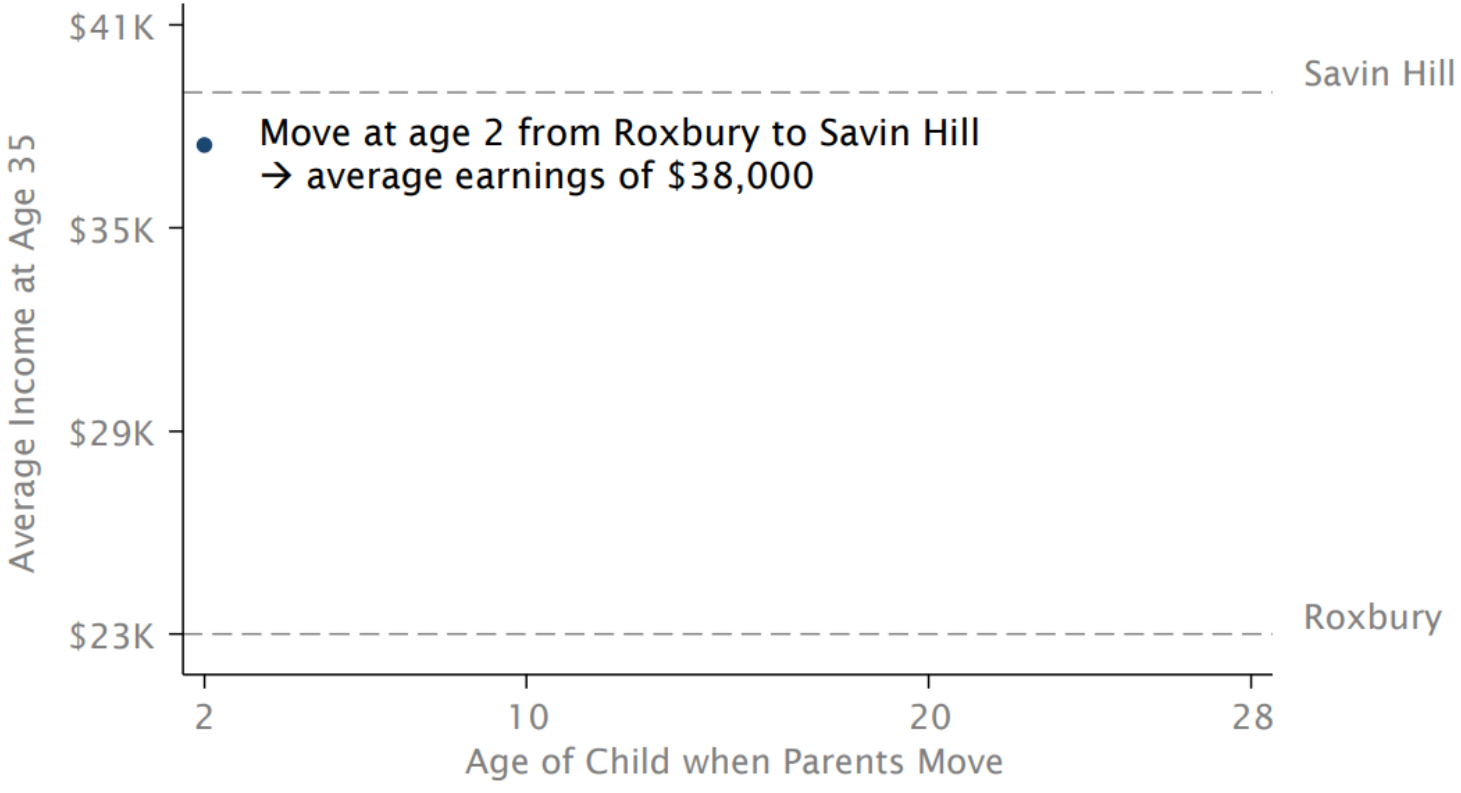
Opportunity Atlas of MA: Savin Hill outlined, Roxbury nextdoor.

Moving to a Higher Mobility Area and Income



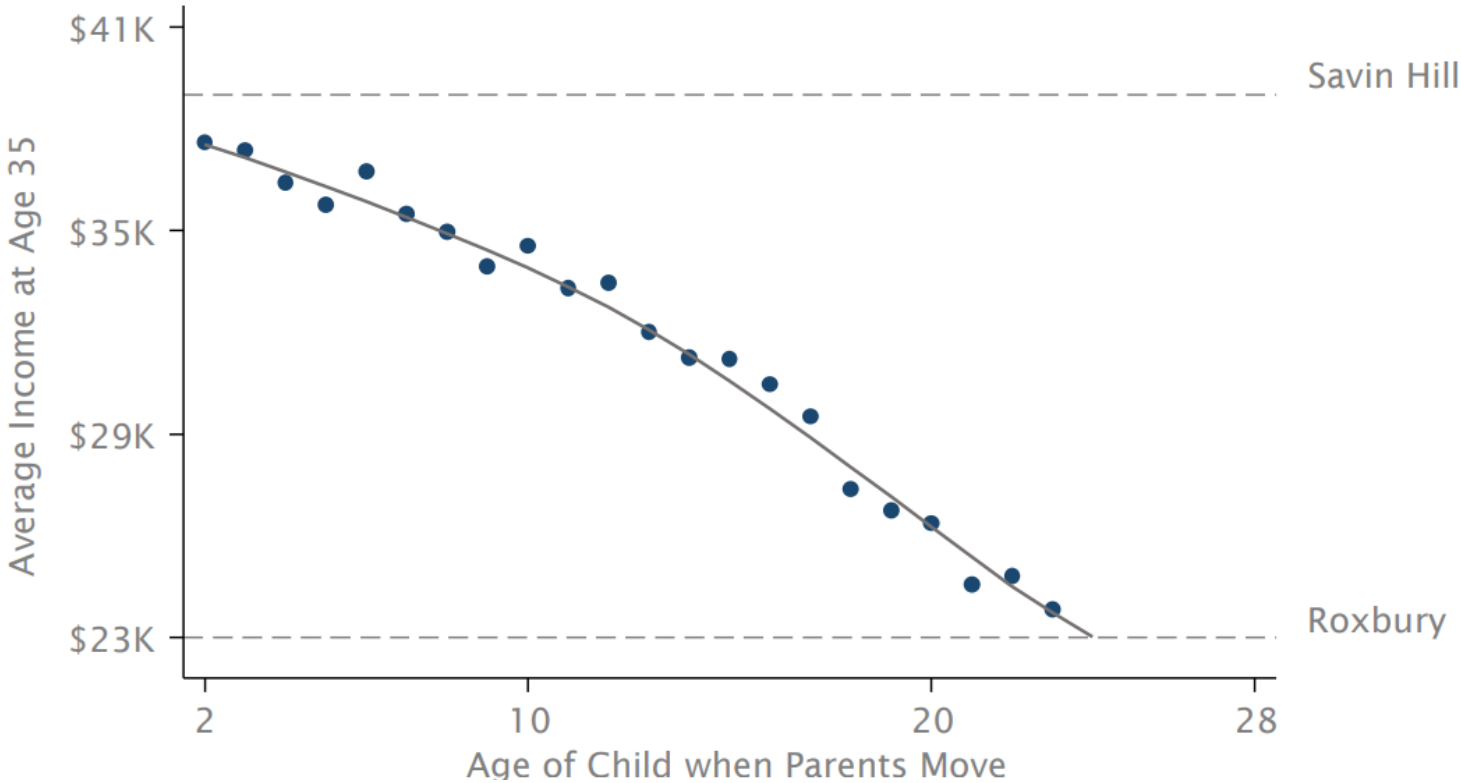
Chetty and Hendren (2018).

Moving to a Higher Mobility Area and Income



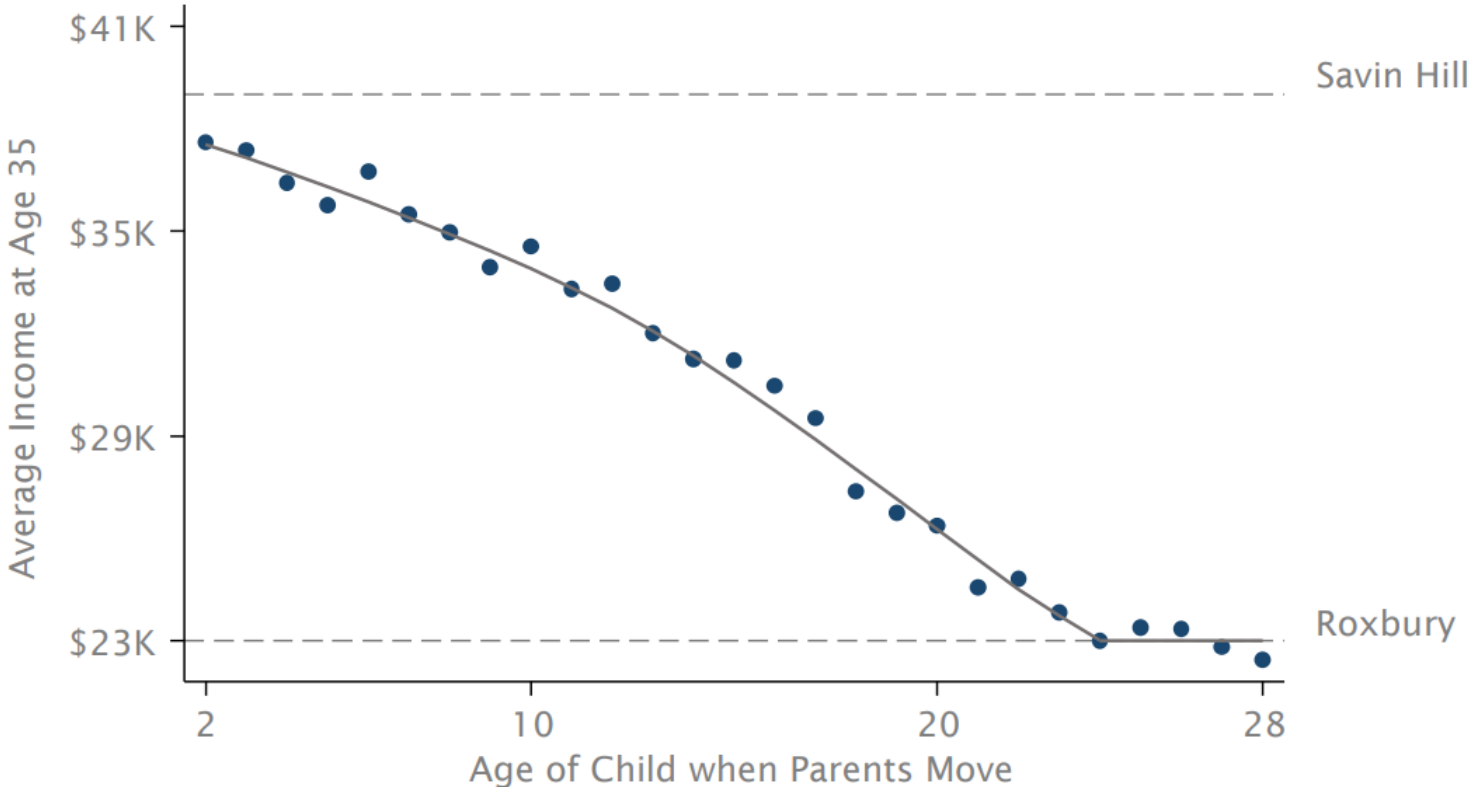
Chetty and Hendren (2018).

Moving to a Higher Mobility Area and Income



Chetty and Hendren (2018).

Moving to a Higher Mobility Area and Income



Chetty and Hendren (2018).

One issue: families differ a ton

- Each family is different

1. Some families are rich, some are poor
2. Some families are more educated, some are less educated
3. Some families are religious, some are not

- Each of these differences could affect:

1. If they move
2. When they move
3. Where they move
4. Children's income mobility
5. Much more...

- We can't possible control for all of this, let alone measures some of it

Fixed effects!

- Fixed effects are a way to control for unobserved variables that are constant along some dimension
    - This dimension could be time, space, individual, etc.
- Fixed effects remove the variation between units, leaving only the variation within units
- Chetty and Hendren (2018) employ fixed effects to isolate within-family variation
    - They do many other things too, but this is the simplest to folow
- Every family has its own unique characteristics that affect children's outcomes
- Fixed effects control for these characteristics by effectively removing the within-family mean from each observation

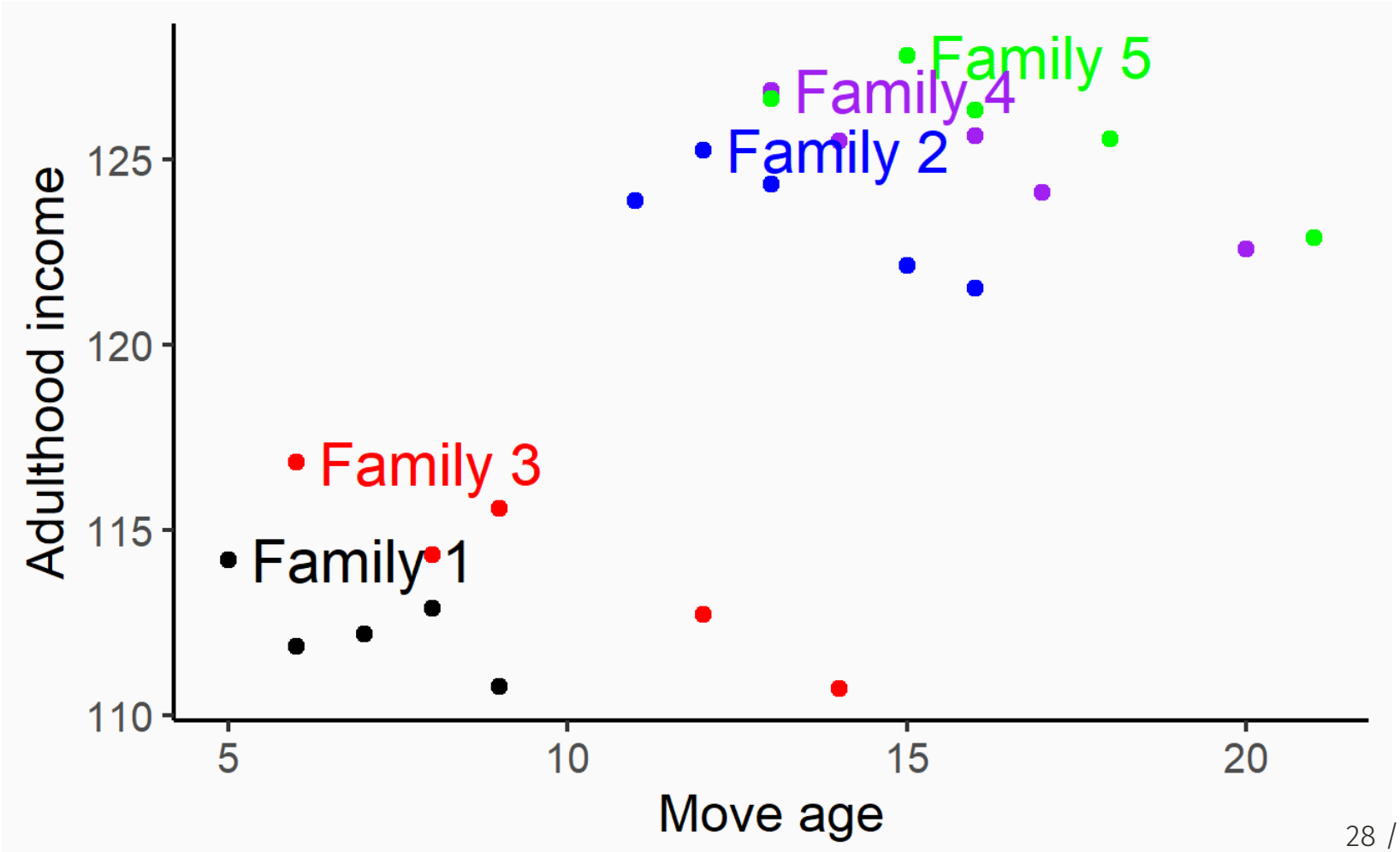
Simplified dataset of mobility

- Let's look at a hypothetical dataframe for 5 families with 5 children each that move from the same low mobility neighborhood to a high mobility neighborhood

```
## # A tibble: 25 × 4
## # Groups:   family_id [5]
##    family_id child_id age_moved income
##        <int>    <int>     <dbl>  <dbl>
##  1         1        1         5   114.
##  2         2        1        16   122.
##  3         3        1         8   114.
##  4         4        1        20   123.
##  5         5        1        21   123.
##  6         1        2         9   111.
##  7         2        2        11   124.
##  8         3        2        14   111.
##  9         4        2        16   126.
## 10         5        2        18   126.
## # ℹ 15 more rows
```

Between and Within variation

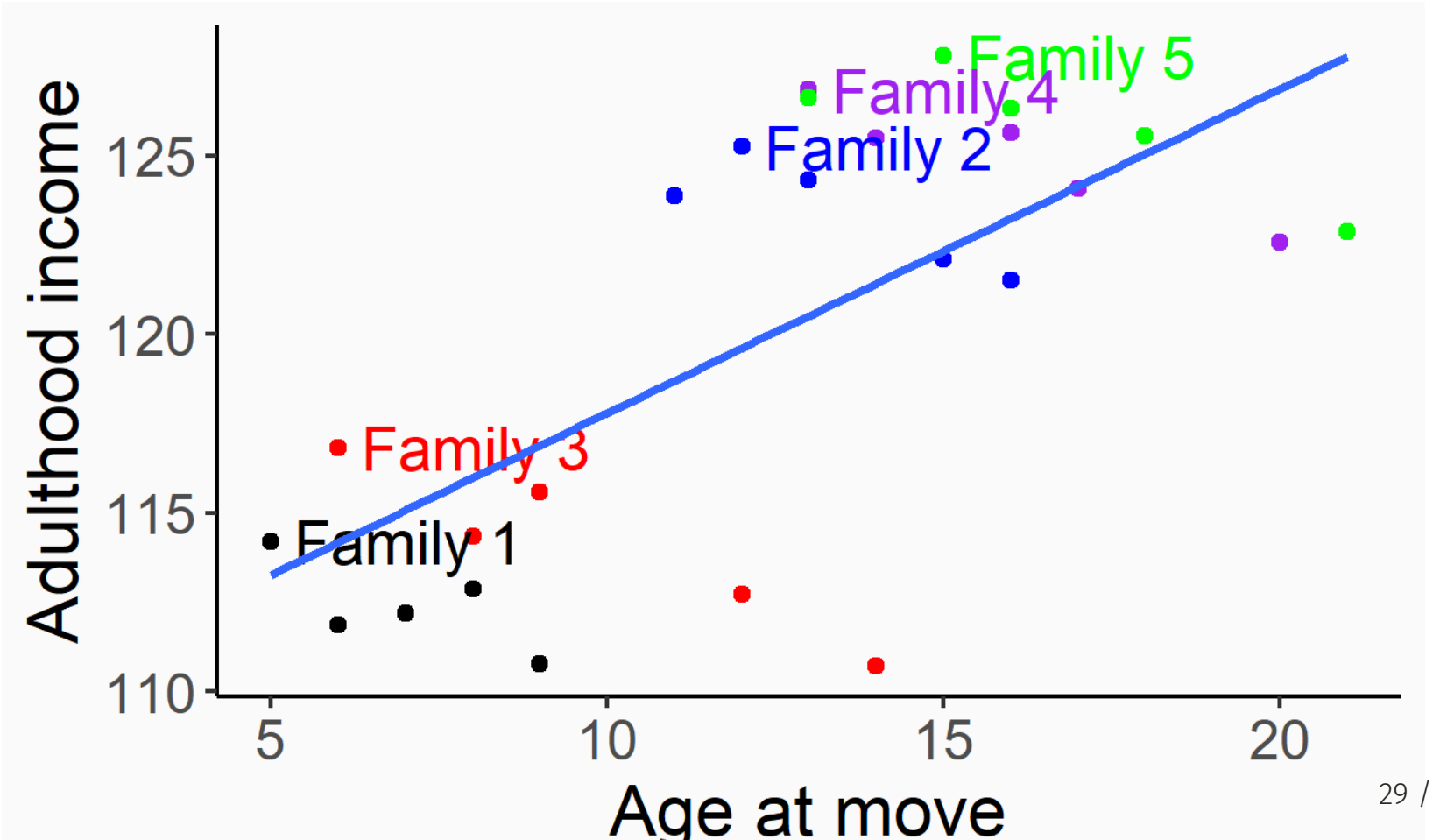
- Below I plot the fake data



ChatGPT and GitHub copilot helped me create this

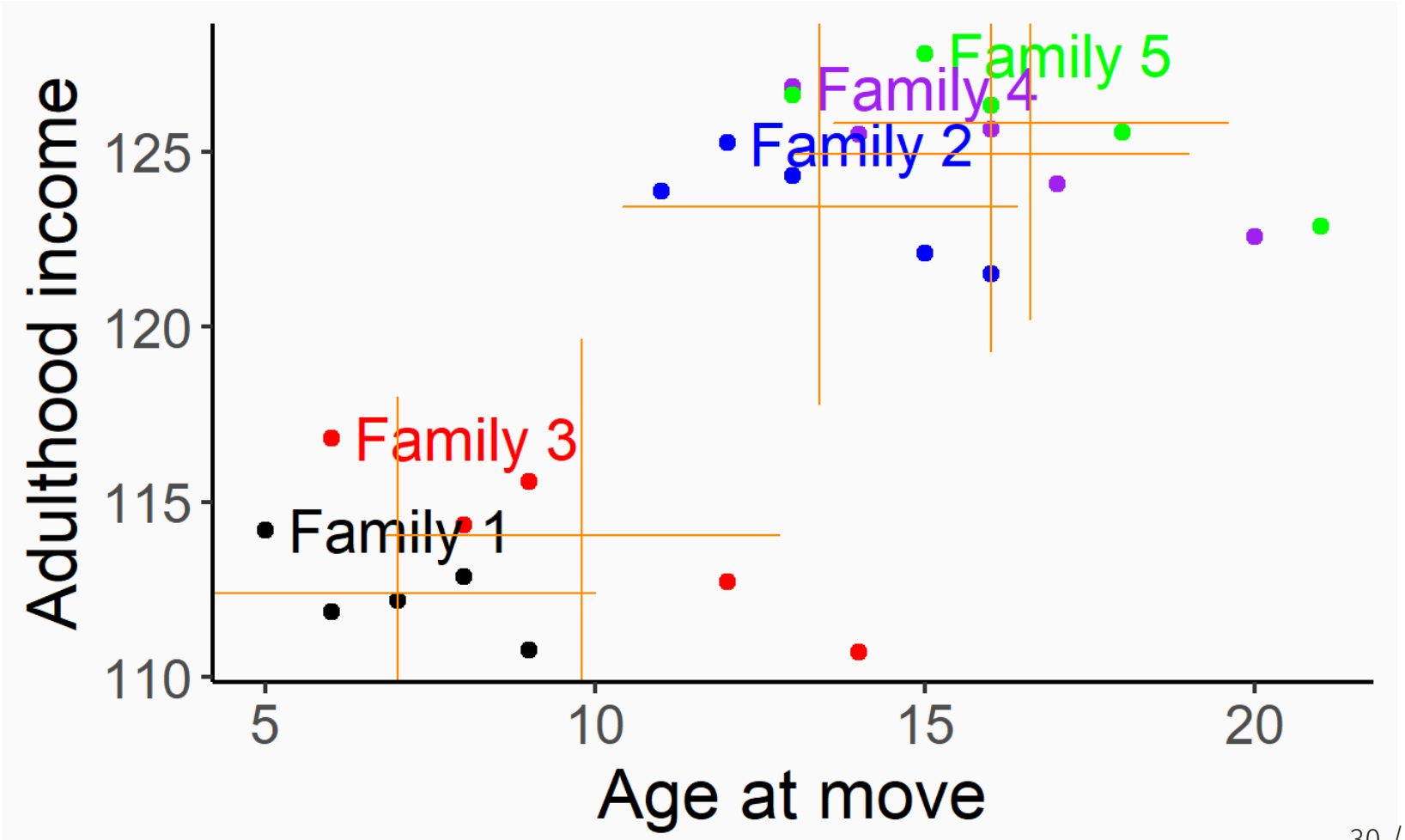
Between and Within

- If I just regress (pooled OLS), I get an increasing relationship with age moved!



Between and Within

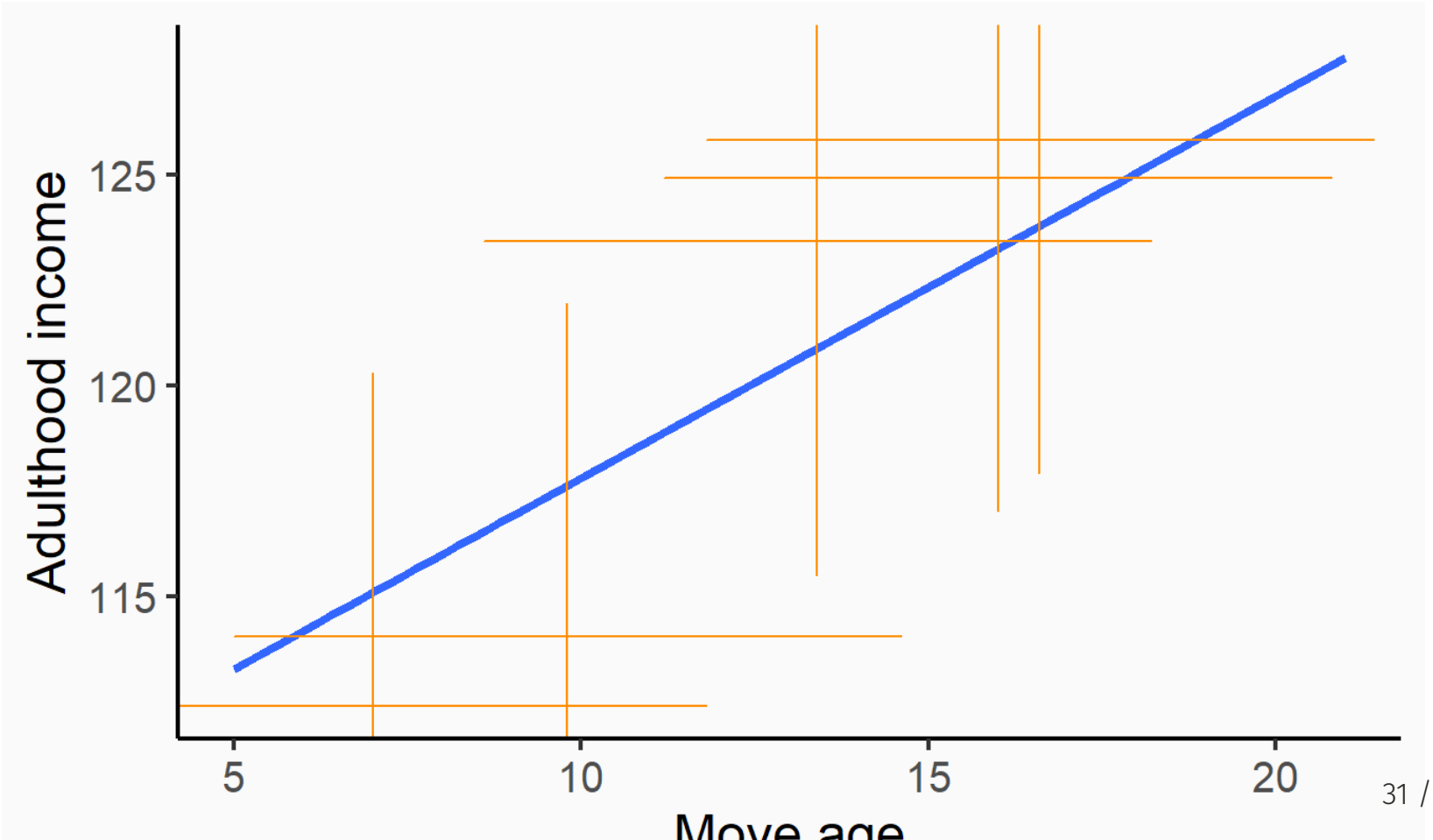
- BETWEEN variation is the variation between means of each family



ChatGPT and GitHub copilot helped me create this

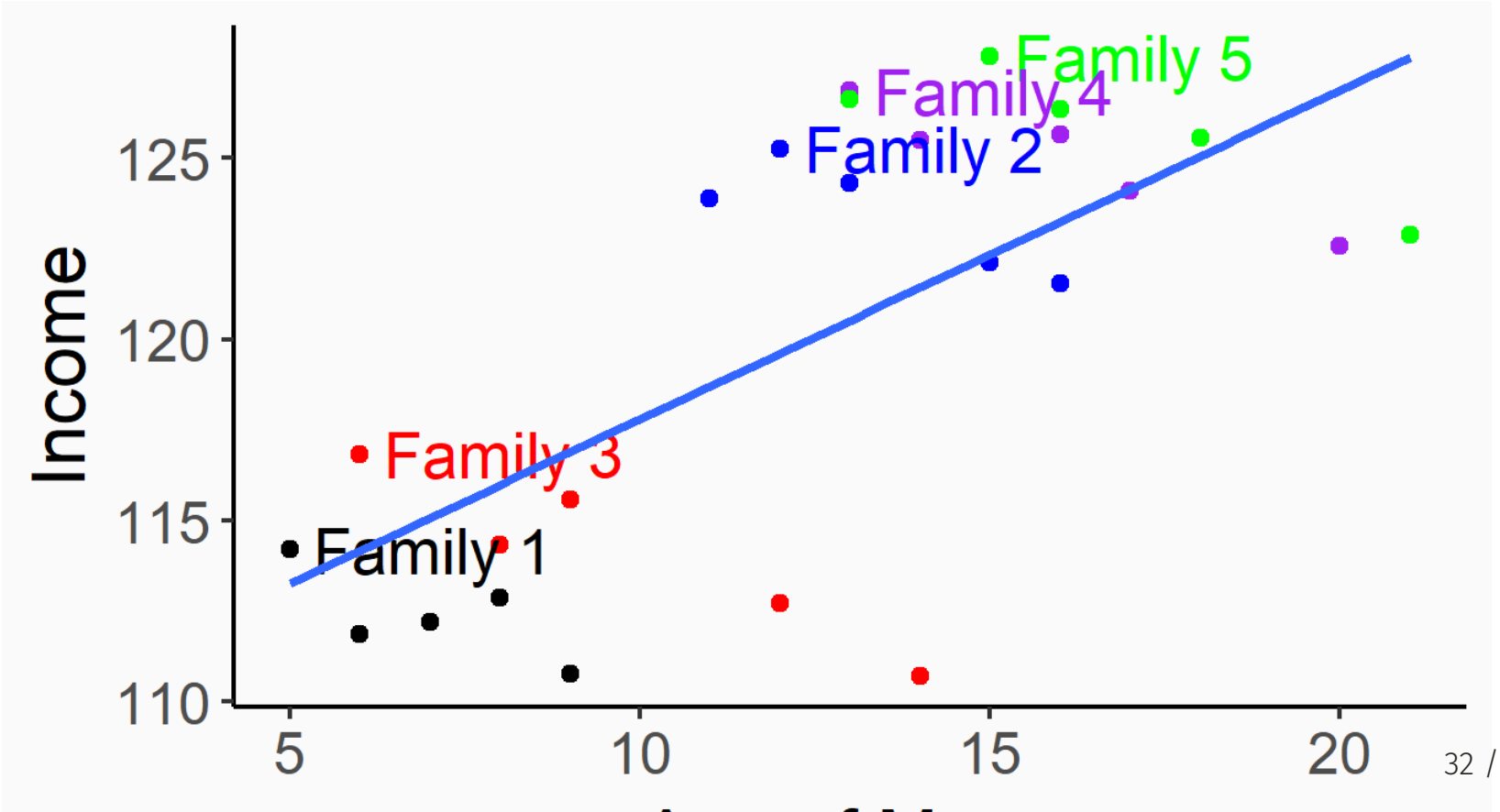
Between and Within

-Seriously, Only look at those means! The individual child variation within families does not matter



Between and Within

- Within variation treats the orange crosses as their own axes and looks at variation within family
- We basically slide the axes on top of each other and analyze *that* data



Removing between variation

- Chetty and Hendren (2018) use fixed effects to remove between-family variation
- But what does that mean?
- How do we actually do this?
  - Let's look at a stylized model

Stylized model of Chetty and Hendren (2018)

$$\text{Income}_i = \sum_{m=0}^{m=30} \beta_m I(\text{Age at move}_i = m) + \epsilon_i$$

but ϵ_i includes all that family variation, which is an omitted variable! That will create bias.

- We really have something like this:

$$\text{Income}_i = \sum_{m=0}^{m=30} \beta_m I(\text{Age at move}_i = m) + (\alpha_f + \nu_i)$$

$I()$ means "indicator function", which equals one when the age at move was m

De-meaning

- Just like we de-meaned our plots, we can de-mean our data
- What happens if we subtract the mean of each variable for each family from each observation?
  - Well $\bar{\alpha}_f = \alpha_f$, so it is just gone!

$$\text{Income}_i - \bar{\text{Income}}_i = \sum_{m=0}^{m=30} \beta_m I(\text{Age at move}_i = m) - I(\text{Age at move}_i = m) + \dots$$

- This is called a **fixed effect** model
- By construction, ν_i is no longer correlated with family characteristics
- Yes, I know that the average of $I(\text{Age at move}_i = m)$ is a strange concept
  - Think of it as the probability of moving at age m for each family

Fixed effects don't make Causality

- Fixed effects don't just make a regression causal
    - There may be other omitted variables that are correlated with the explanatory variable, not "absorbed" by your fixed effects
- 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?

Fixed effects don't make Causality

- Fixed effects don't just make a regression causal
    - There may be other omitted variables that are correlated with the explanatory variable, not "absorbed" by your fixed effects
- 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

Fixed effects elsewhere

- Fixed effects are extremely popular in applied economics
- Any time you have panel data, you can bet a fixed effects model is attempted
    - Panel data is (usually) a dataset where you track a unit (individual, county, etc.) over time
- Plus, they form the backbone of difference-in-difference analysis

Too many fixed effects?

- One challenge with fixed effects:
  - Each fixed effect is like adding a variable for each level of the fixed effect (each individual, each year, etc.)
  - This reduces the degrees of freedom in your model
  - As you add more fixed effects, you need more data to keep statistical power

The actual model in Chetty and Hendren (2018):

$$\text{Income}_i = \alpha_{qosm} + \sum_{m=1}^{m=30} b_m I(m_i = m)\Delta_{odps} + \sum_{s=1980}^{1987} \kappa_s I(s_i = s)\Delta_{odps} + \epsilon_{2i}$$

has over 200,000 fixed effects!

Next lecture: Fixed effects and difference-in-differences