MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

MIT STATISTICS + DATA SCIENCE CENTER

LABORATORY FOR INFORMATION & DECISION SYSTEMS
LIDS

An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?

Tamara Broderick

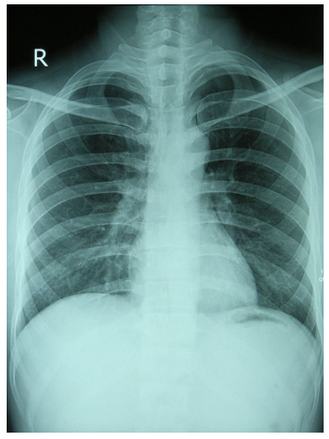
Associate Professor,
MIT

With R Giordano, J Huang, R Meager, Y Shen, D Wei

When can I trust decisions from data?

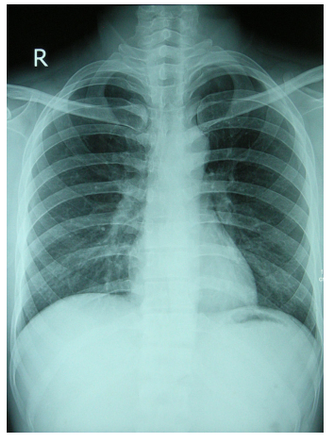
When can I trust decisions from data?

- More ubiquitous and black-box data analyses



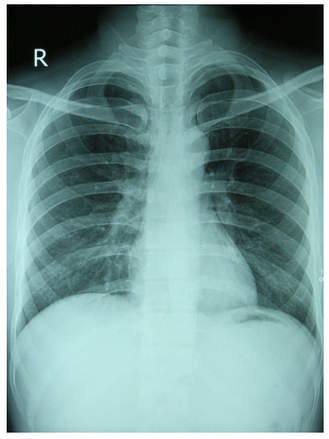
When can I trust decisions from data?

- More ubiquitous and black-box data analyses ➔ evaluation/checking increasingly critical



When can I trust decisions from data?

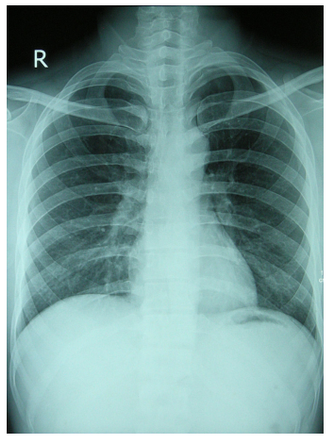
- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data

When can I trust decisions from data?

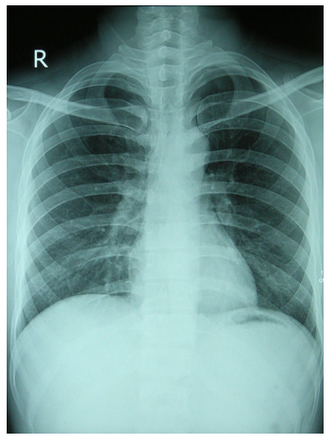
- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if a very small subset of data was instrumental to the original analysis

When can I trust decisions from data?

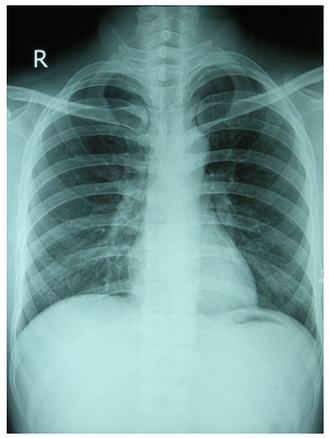
- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if a very small subset of data was instrumental to the original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we can drop 1 data point to flip the sign of the effect

When can I trust decisions from data?

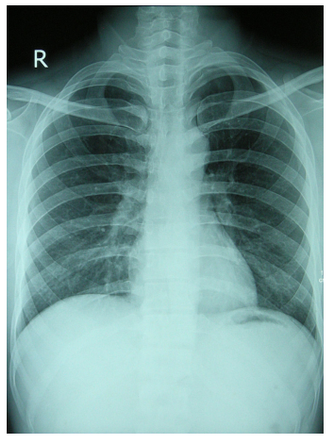
- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if a very small subset of data was instrumental to the original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we can drop 1 data point to flip the sign of the effect
  - E.g. out of 57,477 preferences (matchups) in Chatbot Arena, we can drop 2 to change the top-ranked LLM

When can I trust decisions from data?

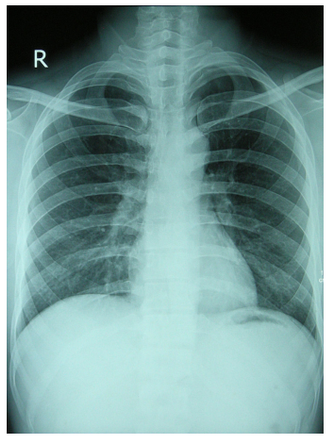
- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if a very small subset of data was instrumental to the original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we can drop 1 data point to flip the sign of the effect
  - E.g. out of 57,477 preferences (matchups) in Chatbot Arena, we can drop 2 to change the top-ranked LLM
- **Challenge**: Impossibly costly to check every data subset

When can I trust decisions from data?

- More ubiquitous and black-box data analyses → evaluation/checking increasingly critical



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
- Might worry about generalizing if a very small subset of data was instrumental to the original analysis
  - E.g. in a study of microcredit with ~16,500 data points, we can drop 1 data point to flip the sign of the effect
  - E.g. out of 57,477 preferences (matchups) in Chatbot Arena, we can drop 2 to change the top-ranked LLM
- **Challenge**: Impossibly costly to check every data subset
- **We show**: an *approximation* is fast, automatable, accurate + empirical & theoretical analysis of our method

Roadmap

Roadmap

- Why might existing evaluation methods not fully handle generalization already?

Roadmap

- Why might existing evaluation methods not fully handle generalization already?
- Our proposed flag: can I drop a small fraction of data points to change conclusions?

Roadmap

- Why might existing evaluation methods not fully handle generalization already?
- Our proposed flag: can I drop a small fraction of data points to change conclusions?
- Checking directly is too expensive, so we provide an approximation (with a guarantee on quality)

Roadmap

- Why might existing evaluation methods not fully handle generalization already?
- Our proposed flag: can I drop a small fraction of data points to change conclusions?
- Checking directly is too expensive, so we provide an approximation (with a guarantee on quality)
- Our method reveals many data analyses are robust but some aren't

Roadmap

- Why might existing evaluation methods not fully handle generalization already?
- Our proposed flag: can I drop a small fraction of data points to change conclusions?
- Checking directly is too expensive, so we provide an approximation (with a guarantee on quality)
- Our method reveals many data analyses are robust but some aren't
- And this non-robustness isn't just reflecting other well-known issues (non-significance, heavy tails, misspecification, etc.); it's reflecting signal-to-noise

Why care about dropping data subsets?

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling
  - More realistic: future population different from analyzed population

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct
  - More realistic: future population different from analyzed population

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct
  - More realistic: future population different from analyzed population, model is misspecified

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want
  - More realistic: future population different from analyzed population, model is misspecified

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy, small fractions of data missing not-at-random

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy, small fractions of data missing not-at-random
- In all these cases, we'd be concerned if dropping a very small fraction of data changed our conclusions

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy, small fractions of data missing not-at-random
- In all these cases, we'd be concerned if dropping a very small fraction of data changed our conclusions
- Concerns not specific to any one field (not a gotcha)

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy, small fractions of data missing not-at-random
- In all these cases, we'd be concerned if dropping a very small fraction of data changed our conclusions
- Concerns not specific to any one field (not a gotcha)
- No current check is a cure-all, and ours isn't either

Why care about dropping data subsets?

- A first (non-exhaustive, illustrative) motivating example: an intervention that affects only a few people
- Standard evaluation tools (e.g. p-values, ideas from computer science) aim to address generalization, but make assumptions that aren't precisely true in practice
  - Assumptions include: i.i.d. sampling, model is correct, output is what we want, data missing at random
  - More realistic: future population different from analyzed population, model is misspecified, reporting a convenient proxy, small fractions of data missing not-at-random
- In all these cases, we'd be concerned if dropping a very small fraction of data changed our conclusions
- Concerns not specific to any one field (not a gotcha)
- No current check is a cure-all, and ours isn't either
  - Even when doesn't bother you, should be up front about it

We need & provide an approximation

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation:

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation (uses influence scores):

[*Many* many use infl scores to drop data, but not worst case fraction: e.g. Jaeckel 1972; Hampel 1974; Cook 1977; Cook, Weisberg 1980; Belsley+ 1980 ch 2.1; Huh, Park 1990; Koh, Liang 2017; Beirami+ 2017, Giordano+ 2019a]

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation (uses influence scores):

[*Many* many use infl scores to drop data, but not worst case fraction: e.g. Jaeckel 1972; Hampel 1974; Cook 1977; Cook, Weisberg 1980; Belsley+ 1980 ch 2.1; Huh, Park 1990; Koh, Liang 2017; Beirami+ 2017, Giordano+ 2019a]

  - Can run when decision is based on minimizers of smooth empirical loss plus optional penalty (and beyond).

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation (uses influence scores):

[*Many* many use infl scores to drop data, but not worst case fraction: e.g. Jaeckel 1972; Hampel 1974; Cook 1977; Cook, Weisberg 1980; Belsley+ 1980 ch 2.1; Huh, Park 1990; Koh, Liang 2017; Beirami+ 2017, Giordano+ 2019a]

  - Can run when decision is based on minimizers of smooth empirical loss plus optional penalty (and beyond).
  - We provide theory to support its accuracy.

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation (uses influence scores):

[*Many* many use infl scores to drop data, but not worst case fraction: e.g. Jaeckel 1972; Hampel 1974; Cook 1977; Cook, Weisberg 1980; Belsley+ 1980 ch 2.1; Huh, Park 1990; Koh, Liang 2017; Beirami+ 2017, Giordano+ 2019a]

  - Can run when decision is based on minimizers of smooth empirical loss plus optional penalty (and beyond).
  - We provide theory to support its accuracy.
  - Any non-robustness is conclusive since we can re-run the analysis (just once) to confirm.

We need & provide an approximation

- Consider Angelucci et al (2015), the randomized controlled trial of microcredit in Mexico. *Fantastic reproducibility!*
  - The study included over 16,000 households.
  - There are ~10^{53} subsets of size 16 (0.1% of 16,000).
  - If data analysis takes 1 second, check takes >10^{46} years.
  - GPUs / parallel computing / etc won't save you!
- We provide an approximation (uses influence scores):

[*Many* many use infl scores to drop data, but not worst case fraction: e.g. Jaeckel 1972; Hampel 1974; Cook 1977; Cook, Weisberg 1980; Belsley+ 1980 ch 2.1; Huh, Park 1990; Koh, Liang 2017; Beirami+ 2017, Giordano+ 2019a]

  - Can run when decision is based on minimizers of smooth empirical loss plus optional penalty (and beyond).
  - We provide theory to support its accuracy.
  - Any non-robustness is conclusive since we can re-run the analysis (just once) to confirm.
  - We provide theory & experiments to show non-robustness reflects a low signal-to-noise ratio in the data analysis.

What makes an analysis non-robust?

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)

- Lottery in Oregon, USA

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)

• Lottery in Oregon, USA

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
    - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas

• Lottery in Oregon, USA

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas
- **Using Bayes or more-tailored models isn't a panacea**

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas
- **Using Bayes or more-tailored models isn't a panacea**
  - Meager 2022: multilevel (Bayesian) model for simultaneous analysis of all 7 RCTs of microcredit

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
    - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas
- **Using Bayes or more-tailored models isn't a panacea**
  - Meager 2022: multilevel (Bayesian) model for simultaneous analysis of all 7 RCTs of microcredit
  - Thoughtful modeling design choices

• Lottery in Oregon, USA

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas
- **Using Bayes or more-tailored models isn't a panacea**
  - Meager 2022: multilevel (Bayesian) model for simultaneous analysis of all 7 RCTs of microcredit
  - Thoughtful modeling design choices
  - Can drop <0.03% of data to change if 0 is in the 95% credible interval, for microcredit effect across countries

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
    - Finkelstein et al 2012, >21,000 data points (surveys)
    - $p < 0.01$ for a positive effect of lottery win on a particular measure of health (# days good health in past 30 days)
      - We can drop 10 points (0.05% of data) to change signif
- **Using bootstrap confidence intervals isn't a panacea**
  - Can still change top-ranked LLM in various arenas
- **Using Bayes or more-tailored models isn't a panacea**
  - Meager 2022: multilevel (Bayesian) model for simultaneous analysis of all 7 RCTs of microcredit
  - Thoughtful modeling design choices
  - Can drop <0.03% of data to change if 0 is in the 95% credible interval, for microcredit effect across countries
  - Can drop <0.1% of data to change the effect sign

What makes an analysis non-robust?

What makes an analysis non-robust?

- **It’s not just that everything is non-robust**

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
  - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data
- **Removing outliers isn't a panacea**

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village
  - Original analysis removes the largest responses

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village
  - Original analysis removes the largest responses
  - We find: can drop 3 points of >4,000 & change signif.

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  - Effect of cash transfers on consumption in poor households, Angelucci & De Giorgi 2009, >10,000 points
    - We find: must drop >4% data to change sign and/or significance (in the approximation)
  - MT-Bench LLM rankings: (in the approximation) we must drop >1% of preferences (matchups) to change top-1
    - Carefully designed prompts (vs. user submitted), expert annotators (vs. crowd-sourced preferences)
  - Djokovic robustly top-1 tennis player from recent data
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village
  - Original analysis removes the largest responses
  - We find: can drop 3 points of >4,000 & change signif.
- **It's not just misspecification**: see above and next slide

What makes an analysis non-robust?

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2)$$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

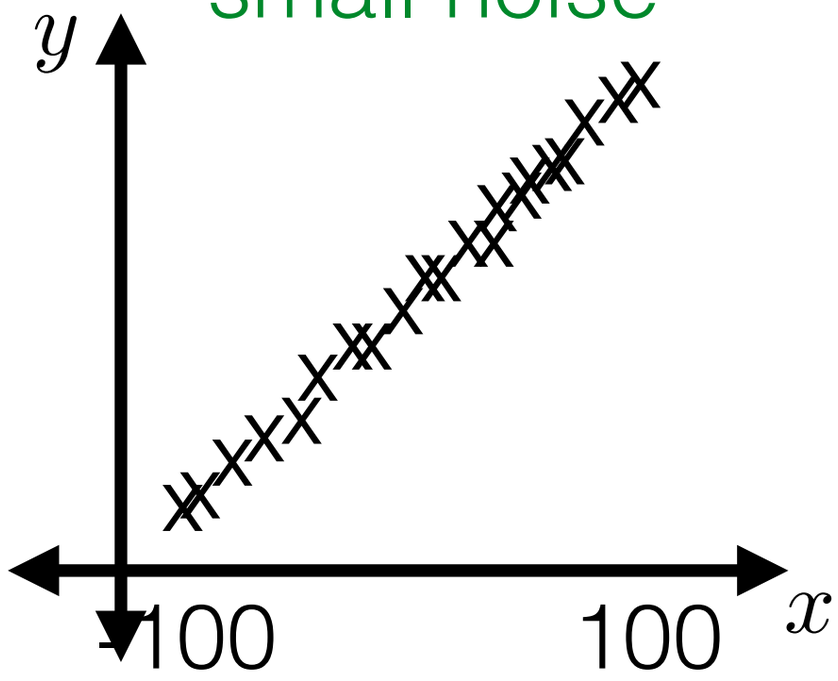
What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \dfrac{\sigma_\epsilon}{\sigma_X}$

small noise



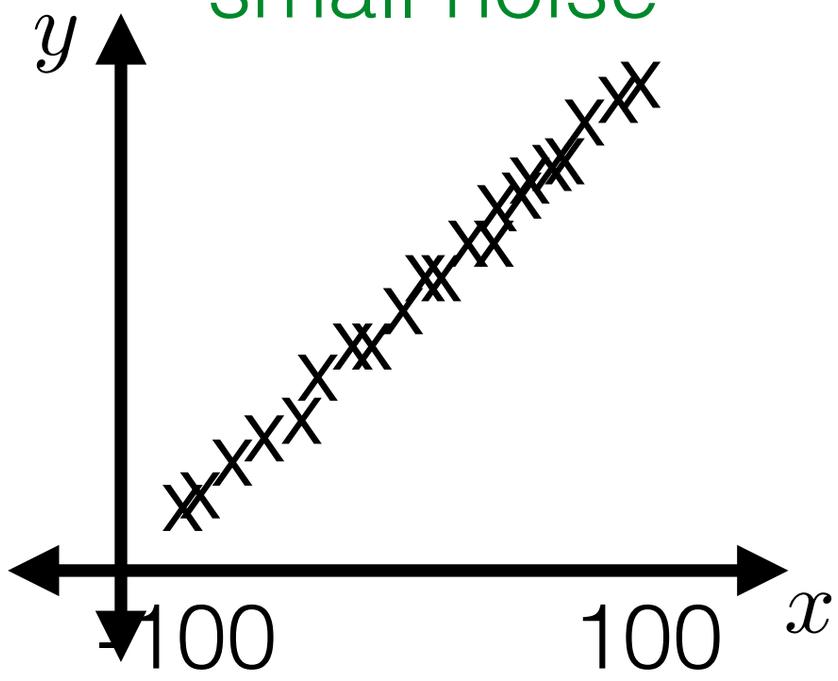
What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

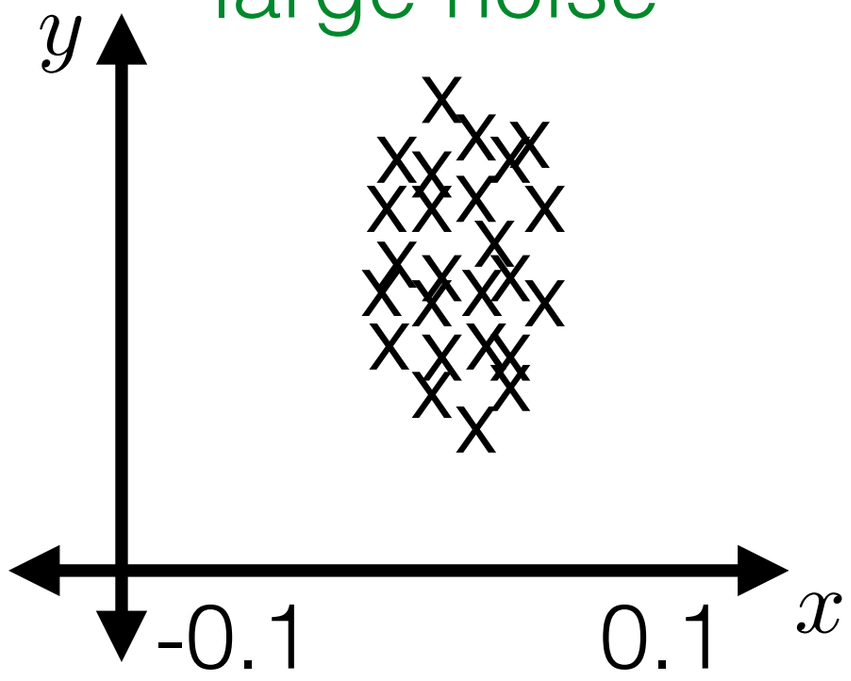
$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

small noise



large noise



What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

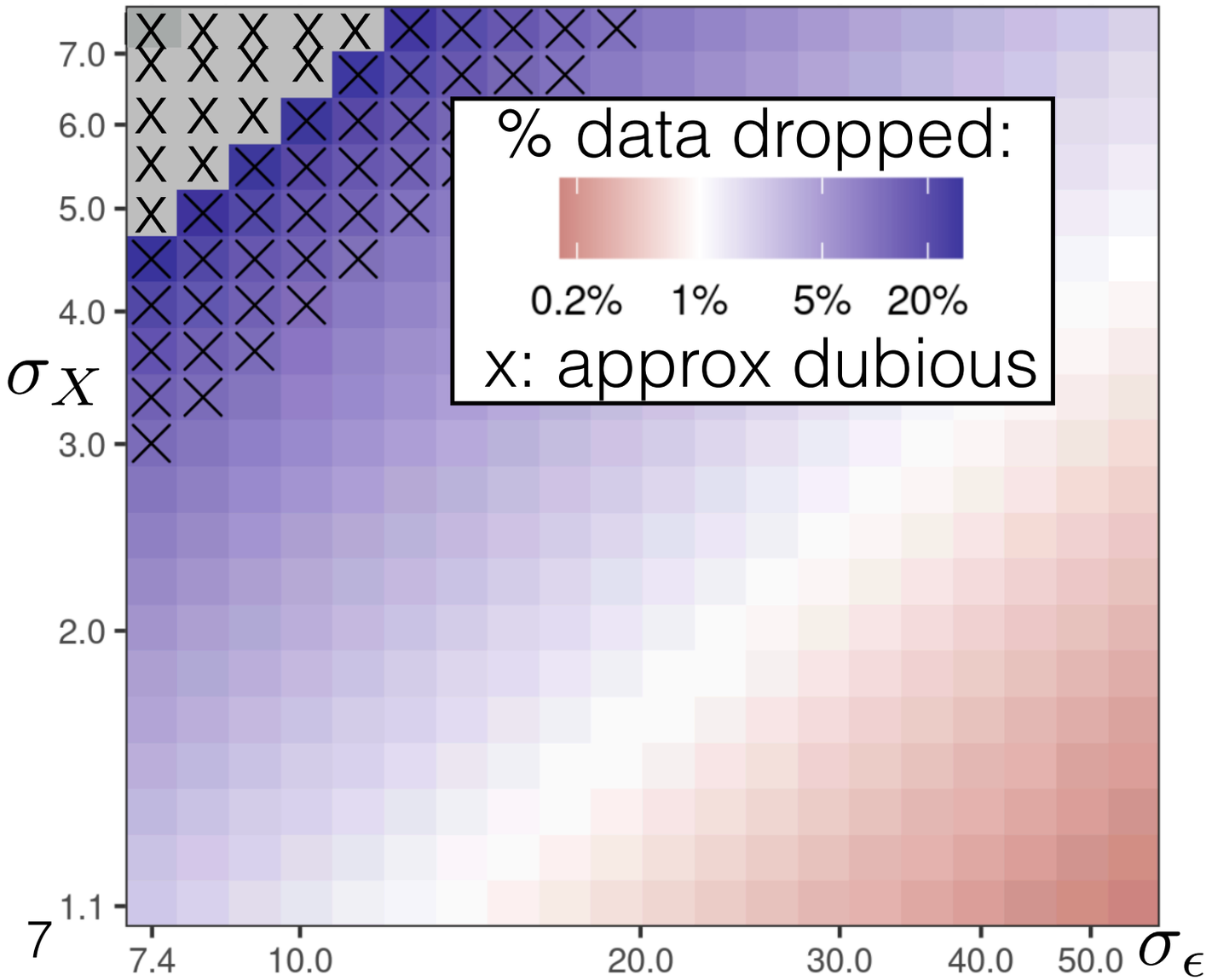
- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

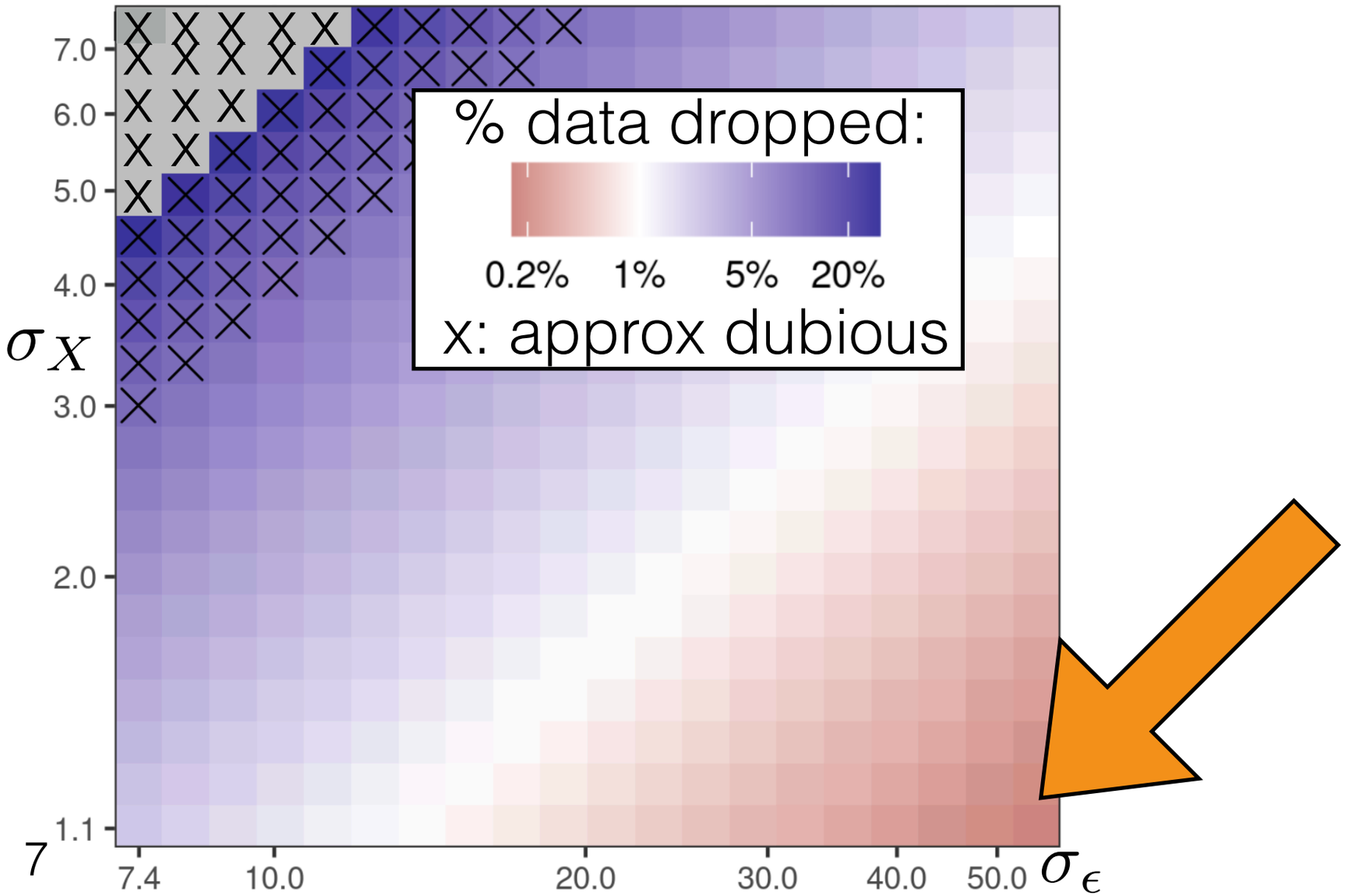


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

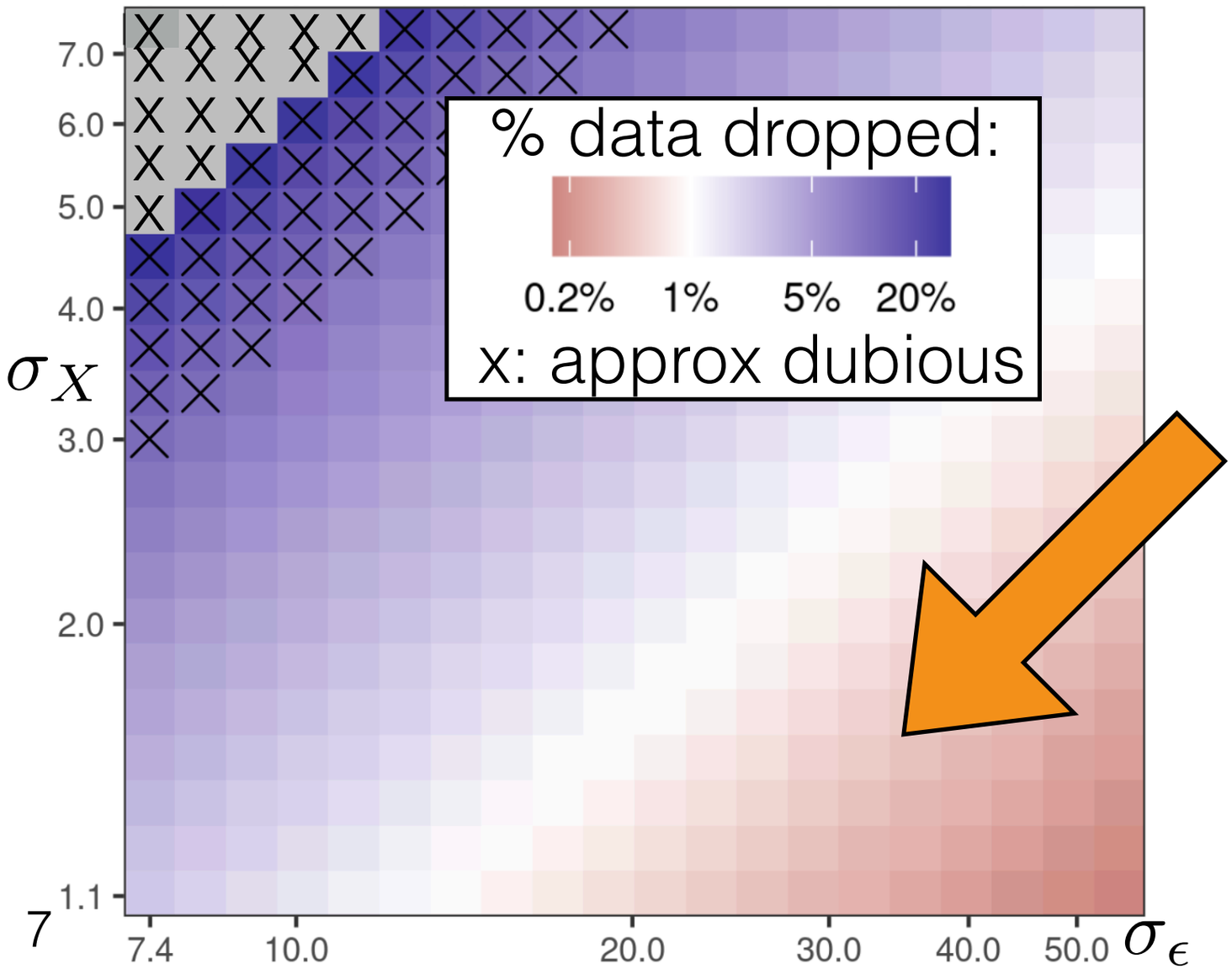


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

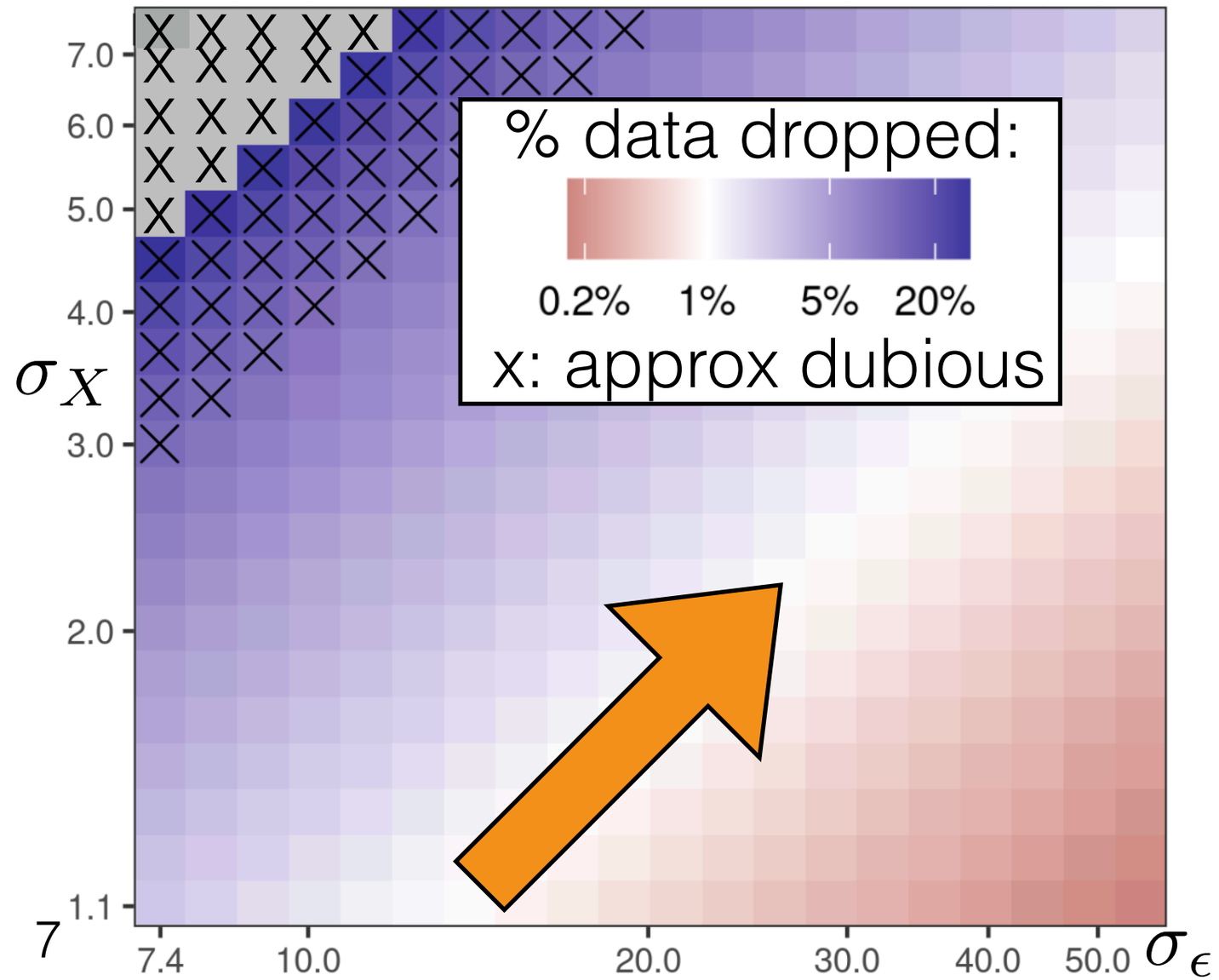


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

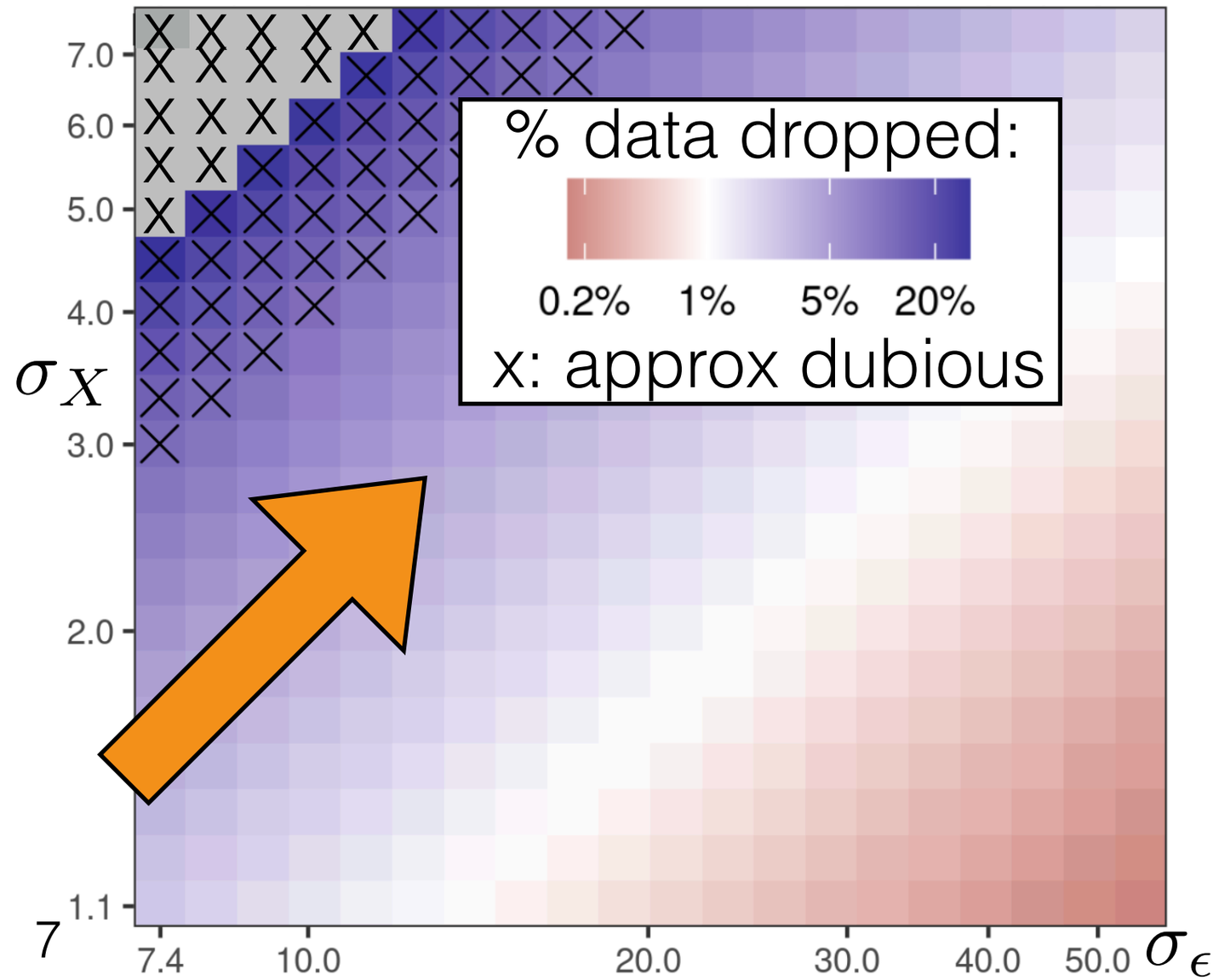


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

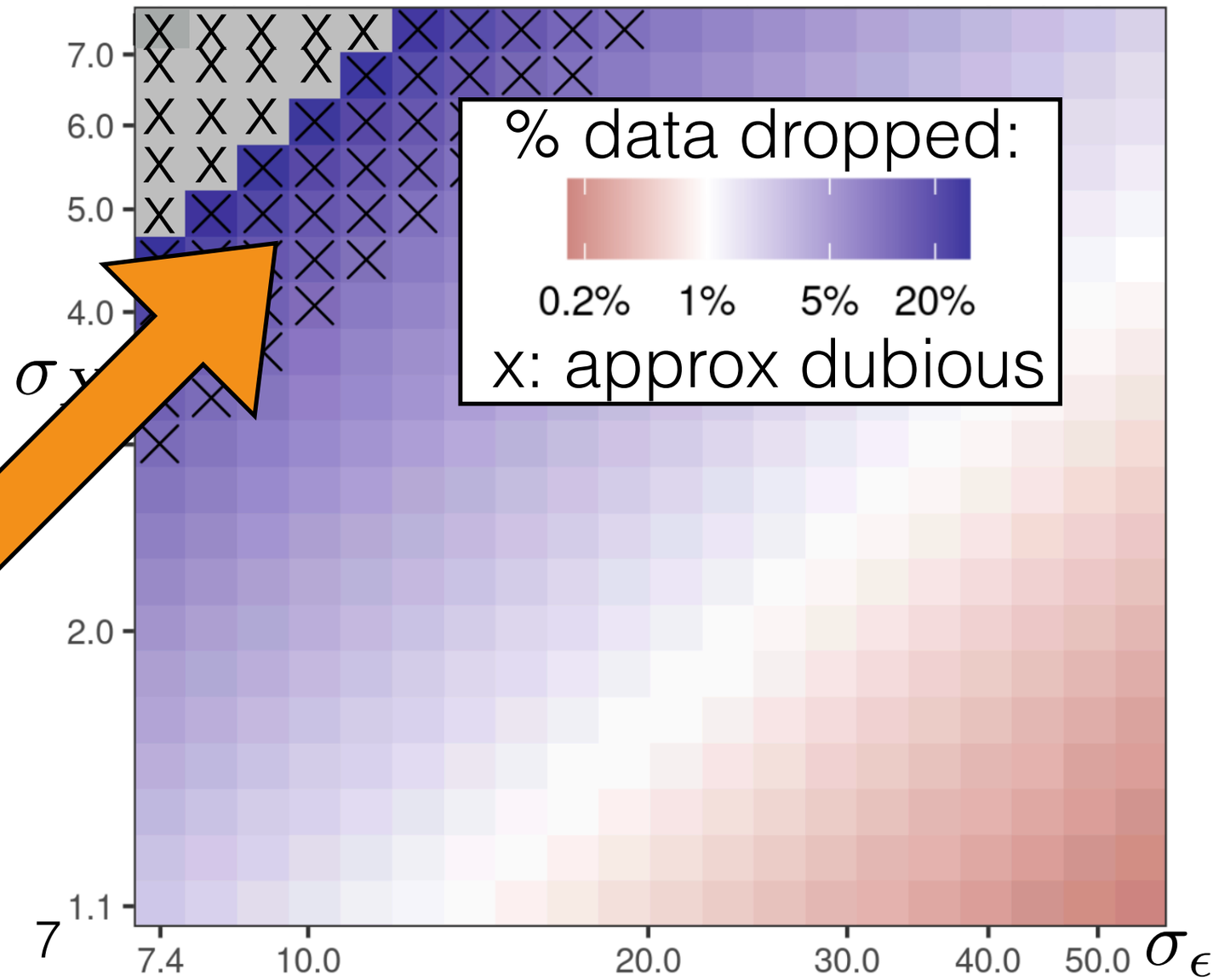


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

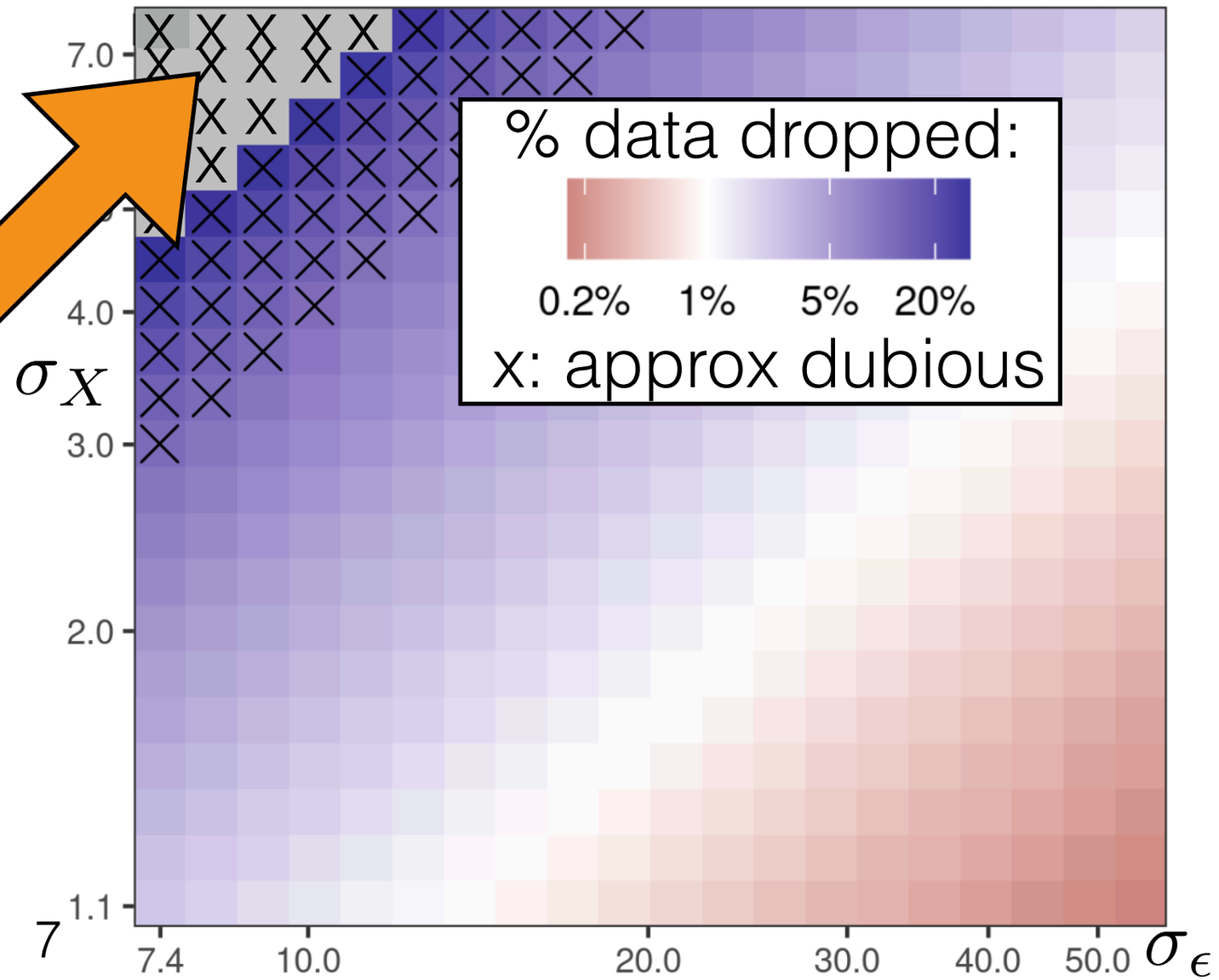


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

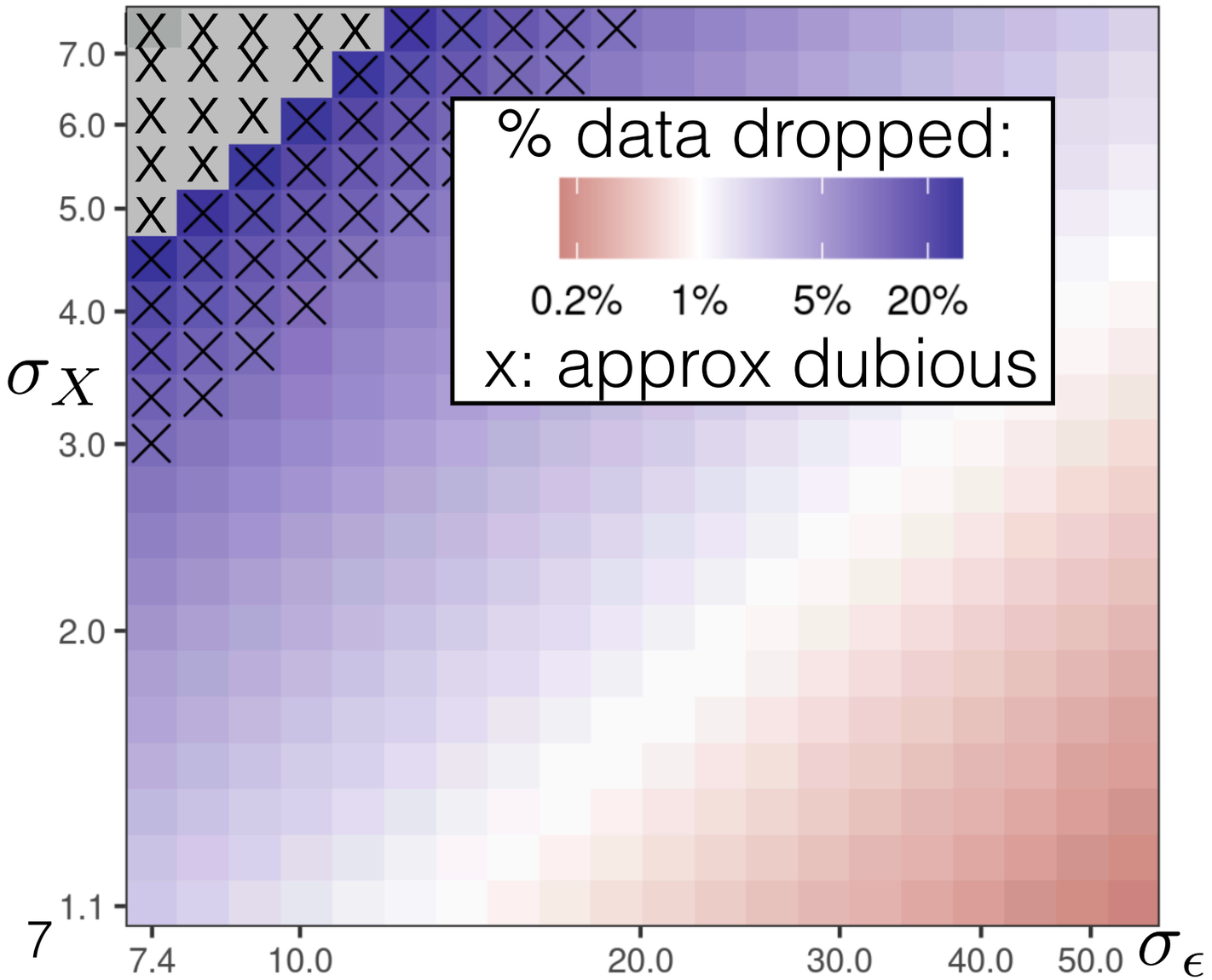


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$

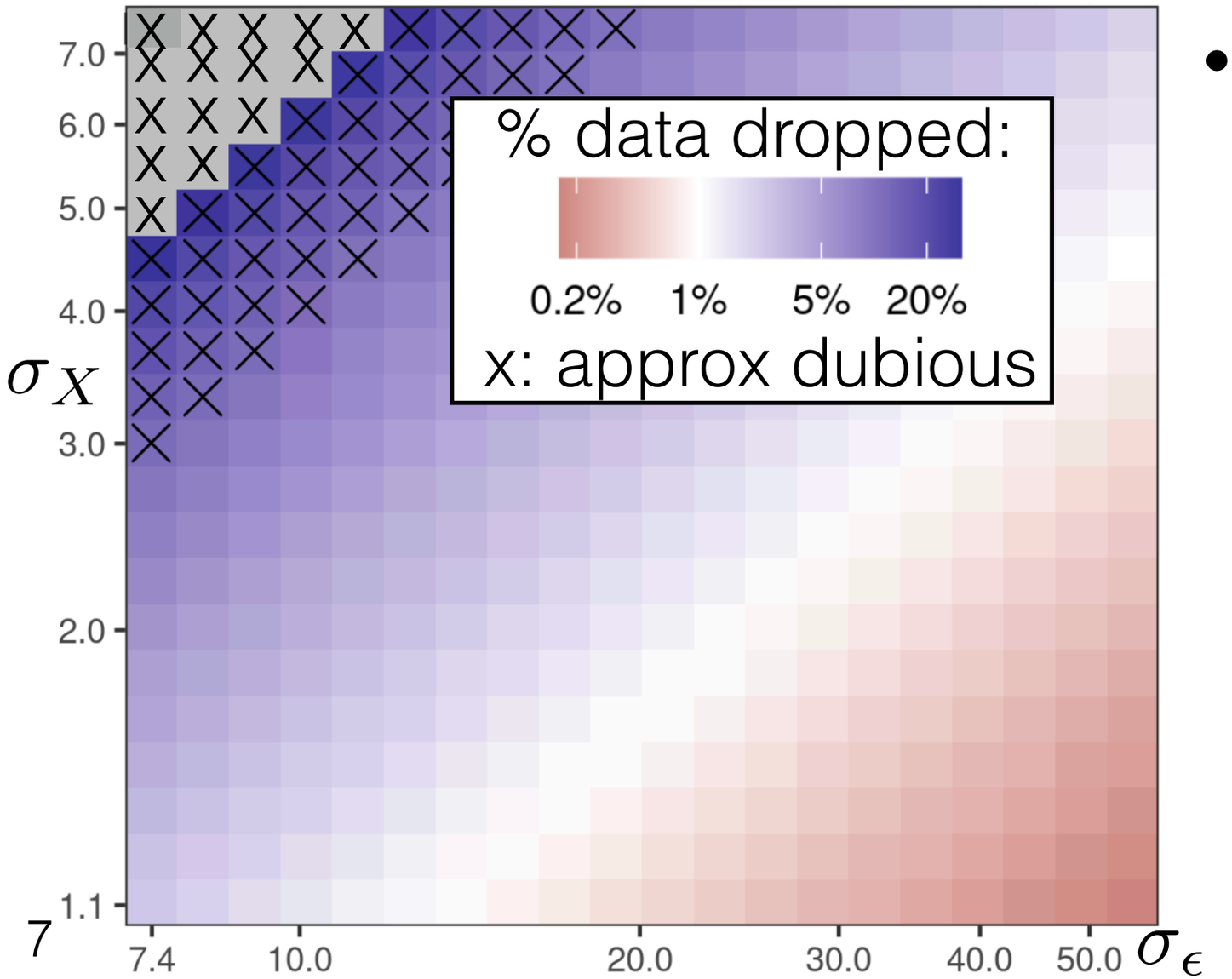


What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$



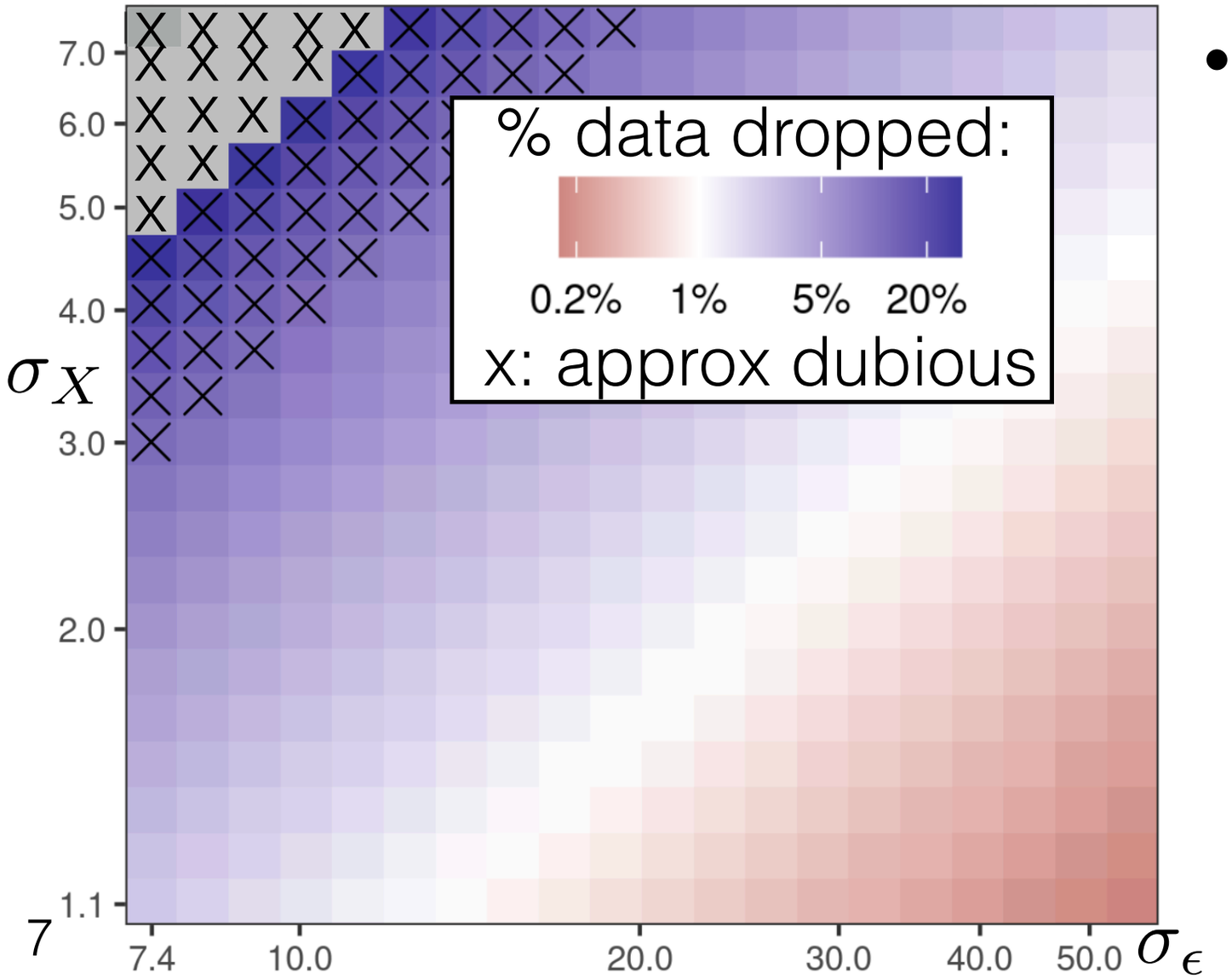
- We can detect if dropping a small fraction of data suffices to change conclusions

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$



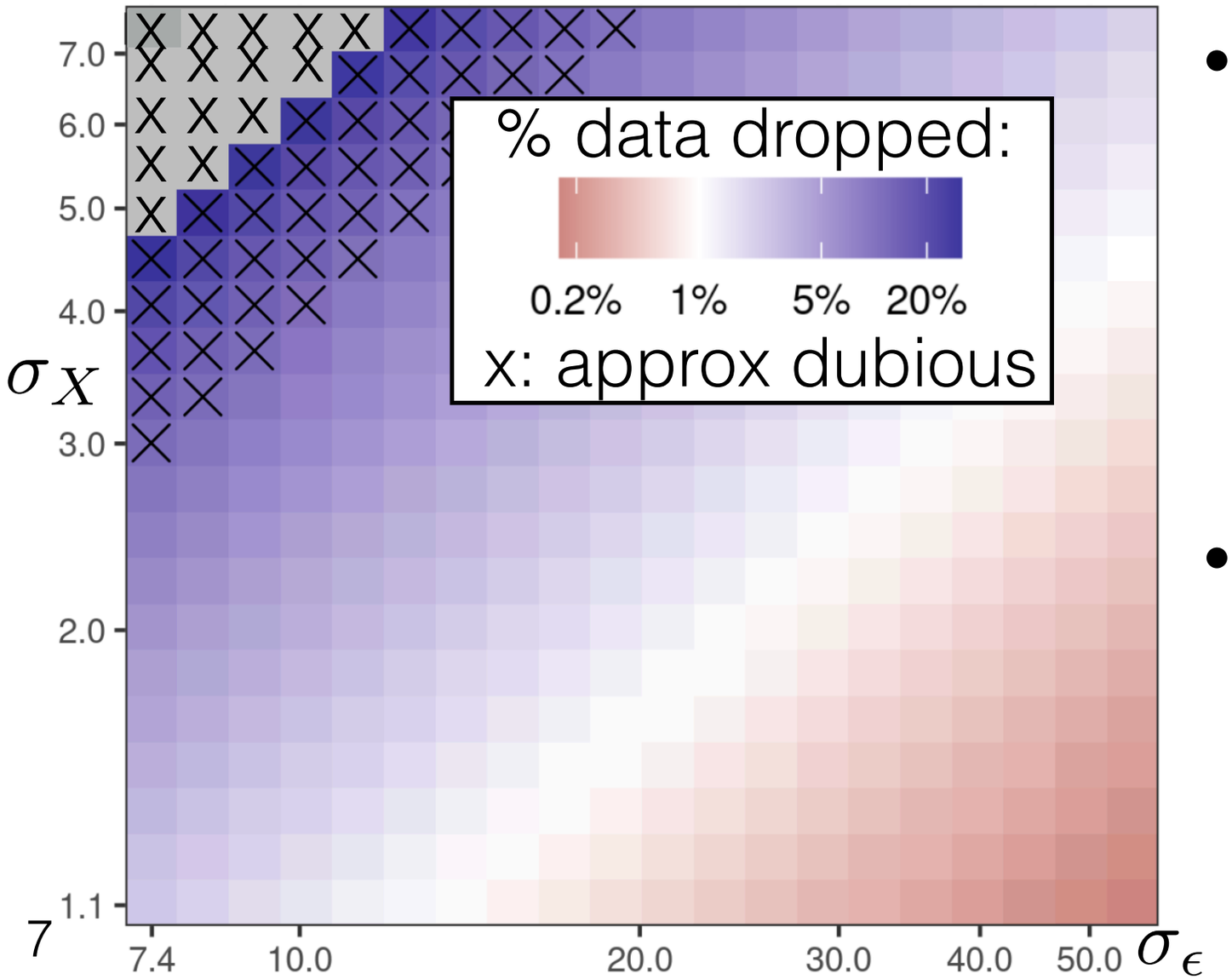
- We can detect if dropping a small fraction of data suffices to change conclusions
- If it's small, we can say how small

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$



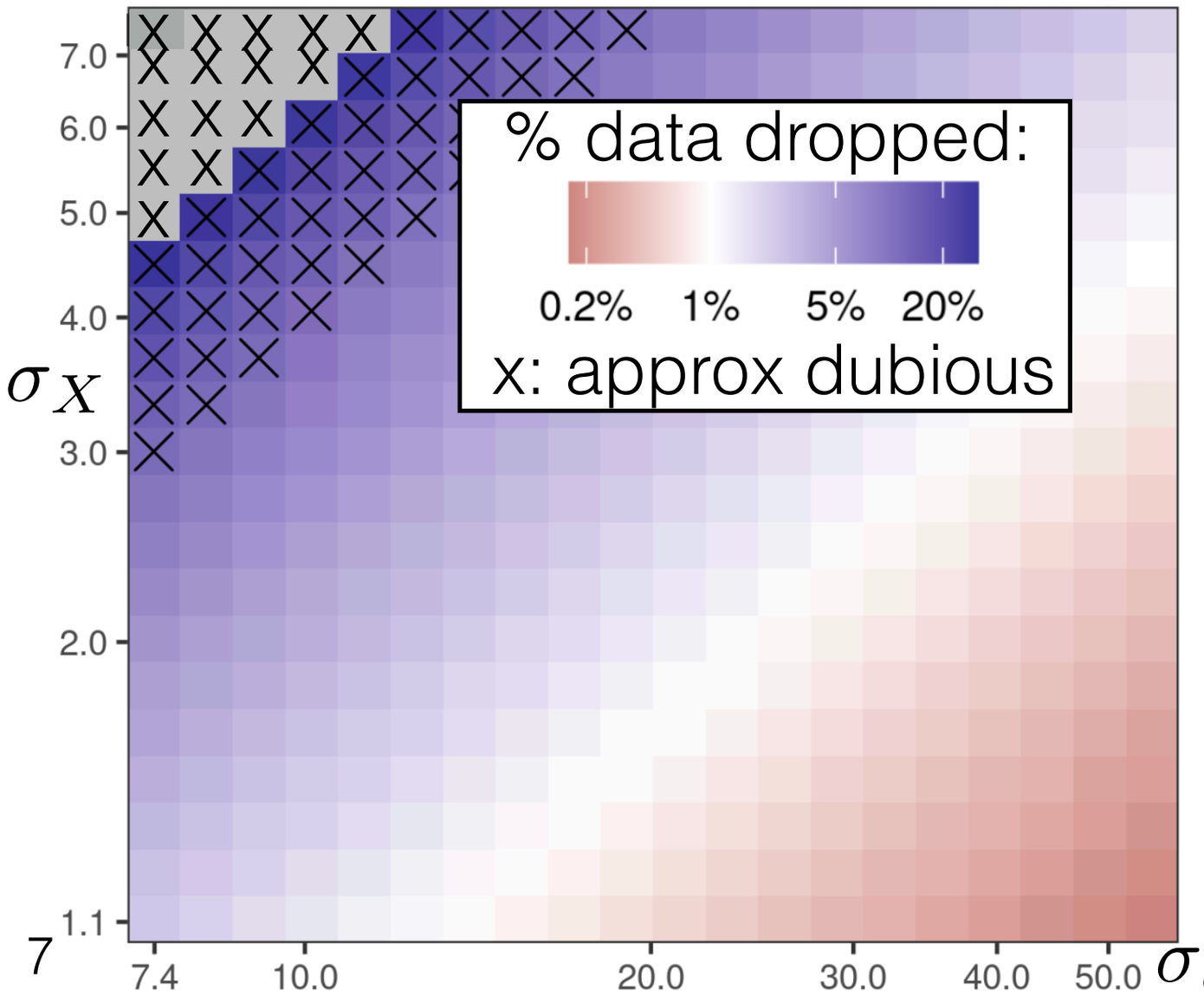
- We can detect if dropping a small fraction of data suffices to change conclusions
  - If it's small, we can say how small
- Sensitivity tracks the signal-to-noise ratio

What makes an analysis non-robust?

- Simulations from linear model with Gaussian noise

$$y_n = \theta x_n + \epsilon_n, \quad \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \quad x_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2), \quad \theta = 0.5$$

- Can we flip sign of $\hat{\theta}$ by dropping some of 5,000 points?
- Signal = size of change of interest: $\Delta = |\hat{\theta}|$
- Noise = estimate of the (scaled) asymptotic std dev: $\approx \frac{\sigma_\epsilon}{\sigma_X}$



- We can detect if dropping a small fraction of data suffices to change conclusions
  - If it's small, we can say how small
- Sensitivity tracks the signal-to-noise ratio
- Not decisive: misspecification, means, heavy tails, gross outliers

Wrapping up

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check):
  Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024
- **Better approximations?** Kuschnig et al 21; Moitra, Rohatgi 23; Freund, Hopkins 23; Hu et al 24; Huang, Burt, Nguyen, Shen, Broderick 25; etc.

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures

- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024
- **Better approximations?** Kuschnig et al 21; Moitra, Rohatgi 23; Freund, Hopkins 23; Hu et al 24; Huang, Burt, Nguyen, Shen, Broderick 25; etc.
- **Reproducibility as prereq:** Haibe-Kains+ *Nature Matters Arising* 2020

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures

- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check):
  Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024
- **Better approximations?** Kuschnig et al 21; Moitra, Rohatgi 23; Freund, Hopkins 23; Hu et al 24; Huang, Burt, Nguyen, Shen, Broderick 25; etc.
- **Reproducibility as prereq:** Haibe-Kains+ *Nature Matters Arising* 2020
- **Suite of checks**: Broderick+ *Science Advances* 2023

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024
- **Better approximations?** Kuschnig et al 21; Moitra, Rohatgi 23; Freund, Hopkins 23; Hu et al 24; Huang, Burt, Nguyen, Shen, Broderick 25; etc.
- **Reproducibility as prereq:** Haibe-Kains+ *Nature Matters Arising* 2020
- **Suite of checks**: Broderick+ *Science Advances* 2023
- Existing validation/uncertainty methods can mislead in some **spatial** settings, and we offer better options in certain cases

Wrapping up

- Evaluation methods give *proxies* of performance, not direct measures
- We present a way to **check** if there is a very small fraction of data you can drop to change conclusions (+analysis of our check): Broderick, Giordano, Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? 2020 (alphabetical authors)
- **p-hacking isn't robust to small-data dropping:** michaelwiebe.com/blog/2021/01/amip, rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html
- **LLMs:** Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings 2026, **Bio:** Shiffman+ 2024, **MCMC:** Nguyen+ 2024
- **Better approximations?** Kuschnig et al 21; Moitra, Rohatgi 23; Freund, Hopkins 23; Hu et al 24; Huang, Burt, Nguyen, Shen, Broderick 25; etc.
- **Reproducibility as prereq:** Haibe-Kains+ *Nature Matters Arising* 2020
- **Suite of checks**: Broderick+ *Science Advances* 2023
- Existing validation/uncertainty methods can mislead in some **spatial** settings, and we offer better options in certain cases:
  - **Validation**: Burt, Shen, Broderick. Consistent Validation for Predictive Methods in Spatial Settings. 2025. **Uncertainty**: Burt*, Berlinghieri*, Bates, Broderick. Smooth Sailing: Lipschitz-Driven Uncertainty Quantification for Spatial Association 2025. & Burt, Berlinghieri, Broderick. Wrong Model, Right Uncertainty: Spatial Associations for Discrete Data with Misspecification. 2025.