Statistically Valid Inferences from Privacy Protected Data

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Pew Research Center, 4/24/2025

[1] GaryKing.org/privacy. Based on APSR/AJPS/PA articles with subsets of {Georgie Evans, Meg Schwenzfeier, Adam D. Smith, Abhradeep Thakurta}

Progress in Social Science

Progress in Social Science

- Present

Progress in Social Science

- Present
- Future

Progress in Social Science

- Present
  - Social scientists have more data than ever
- Future

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)

- Future

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs

- Future

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future
  - We must liberate these datasets!

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future
  - We must liberate these datasets!
  - Academics, companies, governments, etc.: must get their privacy act together

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future
  - We must liberate these datasets!
  - Academics, companies, governments, etc.: must get their privacy act together
  - Goal today: data sharing without privacy violations

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future
  - We must liberate these datasets!
  - Academics, companies, governments, etc.: must get their privacy act together
  - Goal today: data sharing without privacy violations
  - How? Solving political problems technologically

Progress in Social Science

- Present
  - Social scientists have more data than ever
  - But a smaller % of data in the world than ever (about the people, groups, firms, countries we study)
  - Most is now locked up inside private companies and other orgs
  - The central unresolved issue: Privacy (of customers, citizens, firms, etc.)
- Future
  - We must liberate these datasets!
  - Academics, companies, governments, etc.: must get their privacy act together
  - Goal today: data sharing without privacy violations
  - How? Solving political problems technologically
  - I.e., everyone gets what they want, without balancing.

Solving Political Problems Technologically

Differential Privacy & Inferential Validity

A General Purpose, Statistically Valid DP Algorithm

The Algorithm in Practice

Convincing Facebook to Make Data Available

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via “constitutional design”)

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via “constitutional design”)

- Gary visits Facebook to persuade them to make data available

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?"

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica.

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto
  - Trusted 3rd party: Commission at Social Science One signs NDAs, agree not to publish from the data, chooses datasets, makes final decisions; can report publicly if Facebook reneges

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto
  - Trusted 3rd party: Commission at Social Science One signs NDAs, agree not to publish from the data, chooses datasets, makes final decisions; can report publicly if Facebook reneges
- Problem solved, without balancing $\rightsquigarrow$ agreements, announcements, funding, 30+ people assigned at Facebook

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto
  - Trusted 3rd party: Commission at Social Science One signs NDAs, agree not to publish from the data, chooses datasets, makes final decisions; can report publicly if Facebook reneges
- Problem solved, without balancing $\rightsquigarrow$ agreements, announcements, funding, 30+ people assigned at Facebook
- Just one issue:

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto
  - Trusted 3rd party: Commission at Social Science One signs NDAs, agree not to publish from the data, chooses datasets, makes final decisions; can report publicly if Facebook reneges
- Problem solved, without balancing $\rightsquigarrow$ agreements, announcements, funding, 30+ people assigned at Facebook
- Just one issue: Facebook's implementation plan was illegal!

Convincing Facebook to Make Data Available

Solving a Political Problem Technologically (via "constitutional design")

- Gary visits Facebook to persuade them to make data available
- In my hotel room packing, email arrives: "Hey what do we do about this?" This was Cambridge Analytica. (The worst timed lobby effort in history!)
- 3 days later: "Could you do a study of the 2016 election?"
- I'd love to, but I need 2 things & you'll only give me 1:
  - Complete access to data, people, etc. (like employees)
  - No pre-publication approval (like NO employees ever)
- We iterate, and I propose a 2-part solution
  - Outside academics: send proposals, no company veto
  - Trusted 3rd party: Commission at Social Science One signs NDAs, agree not to publish from the data, chooses datasets, makes final decisions; can report publicly if Facebook reneges
- Problem solved, without balancing $\rightsquigarrow$ agreements, announcements, funding, 30+ people assigned at Facebook
- Just one issue: Facebook's implementation plan was illegal!
- New Problem: Sharing data without it leaving Facebook

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models,

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)

- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data;

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer,

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal:

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy (seems to satisfy regulators et al.)

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy (seems to satisfy regulators et al.)
  - New Problem:

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy (seems to satisfy regulators et al.)
  - New Problem: Most DP algorithms are statistically invalid!

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy (seems to satisfy regulators et al.)
  - New Problem: Most DP algorithms are statistically invalid!
    - *unknown* statistical properties (usually *biased*)

Data Sharing Regime $\rightsquigarrow$ Data Access Regime

Solving Another Political Problem Technologically (via CS & Statistics)

- Data Sharing Regime: I give you data (maybe you sign DUA)
  - Venerable, but failing
  - Increasing public concern with privacy
  - Scholars discovered: de-identification doesn't work!
  - Nor does aggregation, query auditing, data clean rooms, legal agreements, restricted viewing, paired programmer models, etc.
  - Trusting researchers fails spectacularly at times (C.A.!)
  - Even trusting a researcher known to be trustworthy can fail
- Data Access Regime
  - Trusted server holds data; researchers as adversaries, can run any method $\rightsquigarrow$ noisy answer, a limited number of times
  - Goal: impossible to violate individual privacy; & possible to discover population level patterns
  - $\approx$ differential privacy (seems to satisfy regulators et al.)
  - New Problem: Most DP algorithms are statistically invalid!
    - *unknown* statistical properties (usually *biased*)
    - *no* uncertainty estimates

Solving Political Problems Technologically

Differential Privacy & Inferential Validity

A General Purpose, Statistically Valid DP Algorithm

The Algorithm in Practice

Theories of Inference: Statistics vs. CS

Theories of Inference: Statistics vs. CS

Population
⋮
Michael
Claudia
Michelle
Courtney
Neha
Monica
Alan
Mark Hugo
Katerina Eva
Kim

Mean income: $48

Quantity of Interest

Theories of Inference: Statistics vs. CS

Population	Sample
⋮	✗
Michael	✓
Claudia	✓
Michelle	✓
Courtney	✓
Neha	✓
Monica	✓
Alan	✓
Mark Hugo	✓
Katerina Eva	✓
Kim	✓
Mean income:	$48

Quantity of Interest

Theories of Inference: Statistics vs. CS

Population	Sample	$
⋮	✗	?
Michael	✓	122
Claudia	✓	76
Michelle	✓	145
Courtney	✓	96
Neha	✓	86
Monica	✓	127
Alan	✓	72
Mark Hugo	✓	132
Katerina Eva	✓	95
Kim	✓	134
Mean income:	$48	$108

Quantity of Interest

Usually no direct relevance

Theories of Inference: Statistics vs. CS

Population	Sample	$
⋮	✗	?
Michael	✓	122
Claudia	✓	76
Michelle	✓	145
Courtney	✓	96
Neha	✓	86
Monica	✓	127
Alan	✓	72
Mark Hugo	✓	132
Katerina Eva	✓	95
Kim	✓	134
Mean income:	$48	$108

Quantity of Interest

Usually no direct relevance

Theories of Inference: Statistics vs. CS

Population	Sample	$	+Privacy
⋮	✗	?	
Michael	✓	122	
Claudia	✓	76	
Michelle	✓	145	Noise & Censoring
Courtney	✓	96	
Neha	✓	86	
Monica	✓	127	
Alan	✓	72	
Mark Hugo	✓	132	
Katerina Eva	✓	95	
Kim	✓	134	
Mean income:	$48	$108	

Classical Inference

Quantity of Interest

Usually no direct relevance

Theories of Inference: Statistics vs. CS

Population	Sample	$	+Privacy	=dp$
⋮	✗	?		
Michael	✓	122		85
Claudia	✓	76		103
Michelle	✓	145		75
Courtney	✓	96		113
Neha	✓	86		125
Monica	✓	127		97
Alan	✓	72		101
Mark Hugo	✓	132		128
Katerina Eva	✓	95		83
Kim	✓	134		201
Mean income:	$48	$108		$111

Noise & Censoring

Quantity of Interest

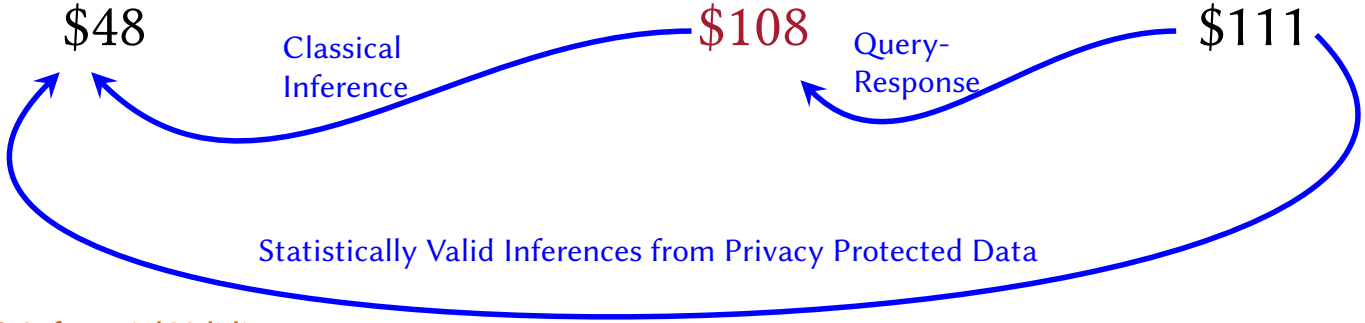
Usually no direct relevance

No direct relevance

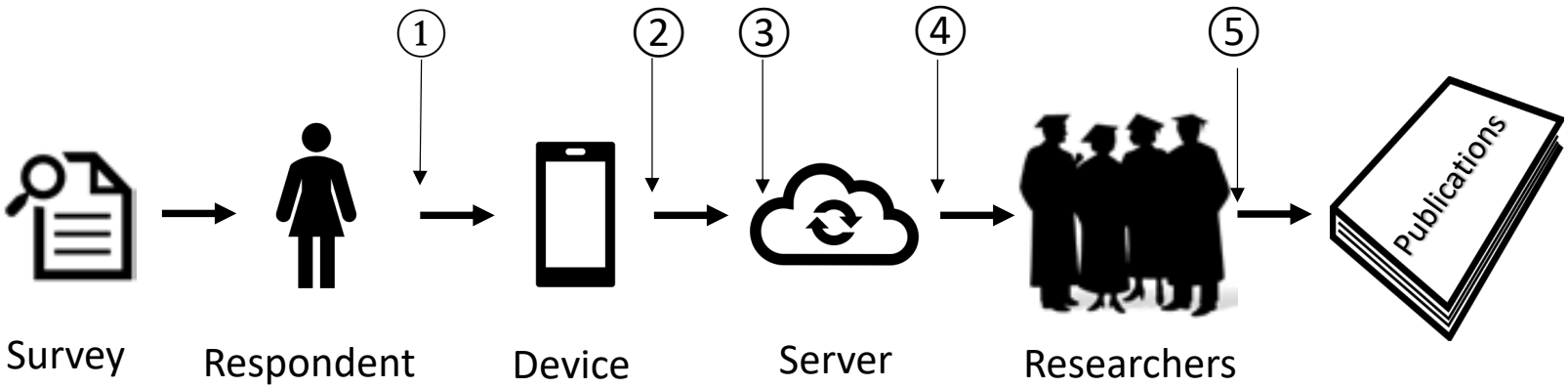
Theories of Inference: Statistics vs. CS

Population	Sample	$	+Privacy	=dp$
⋮	✗	?		
Michael	✓	122		85
Claudia	✓	76		103
Michelle	✓	145		75
Courtney	✓	96		113
Neha	✓	86		125
Monica	✓	127		97
Alan	✓	72		101
Mark Hugo	✓	132		128
Katerina Eva	✓	95		83
Kim	✓	134		201
Mean income:	$48	$108		$111

Noise & Censoring



Protecting Survey Data



Differential Privacy and its Inferential Challenges

Differential Privacy and its Inferential Challenges

- Estimators

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)
  - Including (D) or excluding (D') you doesn't change conclusions

$$\frac{\Pr[M(s, D) = m]}{\Pr[M(s, D') = m]} \in 1 \pm \epsilon$$

for all D, D', m

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)
  - Including (D) or excluding (D') you doesn't change conclusions

$$\frac{\Pr[M(s, D) = m]}{\Pr[M(s, D') = m]} \in 1 \pm \epsilon$$

  for all D, D', m
- Examples all proven to protect the biggest possible outlier

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)
  - Including (D) or excluding (D') you doesn't change conclusions

$$\frac{\Pr[M(s, D) = m]}{\Pr[M(s, D') = m]} \in 1 \pm \epsilon$$

  for all D, D', m
- Examples all proven to protect the biggest possible outlier
  - $M(\text{mean}, D) = \frac{1}{n} \sum_{i=1}^{n} c(y_i, \Lambda) + N\left(0, \frac{\Lambda}{n\epsilon}\right)$ (Λ, n, ϵ known)

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)
  - Including (D) or excluding (D') you doesn't change conclusions

$$\frac{\Pr[M(s, D) = m]}{\Pr[M(s, D') = m]} \in 1 \pm \epsilon$$

    for all D, D', m
- Examples all proven to protect the biggest possible outlier
  - $M(\text{mean}, D) = \frac{1}{n} \sum_{i=1}^{n} c(y_i, \Lambda) + N\left(0, \frac{\Lambda}{n\epsilon}\right)$ (Λ, n, ϵ known)
  - Or: mess with gradients, $X_i' X_i$, data, QOIs, etc.

Differential Privacy and its Inferential Challenges

- Estimators
  - Classical Statistics: Apply statistic s to dataset D, $s(D)$
  - DP Mechanism: $M(s, D)$, with noise & censoring
    - Essential components of ensuring privacy
    - Fundamental problems for statistical inference
- The DP Standard (simplifying)
  - Including (D) or excluding (D') you doesn't change conclusions

$$\frac{\Pr[M(s, D) = m]}{\Pr[M(s, D') = m]} \in 1 \pm \epsilon$$

  for all D, D', m
- Examples all proven to protect the biggest possible outlier
  - $M(\text{mean}, D) = \frac{1}{n} \sum_{i=1}^{n} c(y_i, \Lambda) + N\left(0, \frac{\Lambda}{n\epsilon}\right)$ (Λ, n, ϵ known)
  - Or: mess with gradients, $X_i' X_i$, data, QOIs, etc.
- Statistical properties: usually biased, no uncertainty estimates

Solving Political Problems Technologically

Differential Privacy & Inferential Validity

A General Purpose, Statistically Valid DP Algorithm

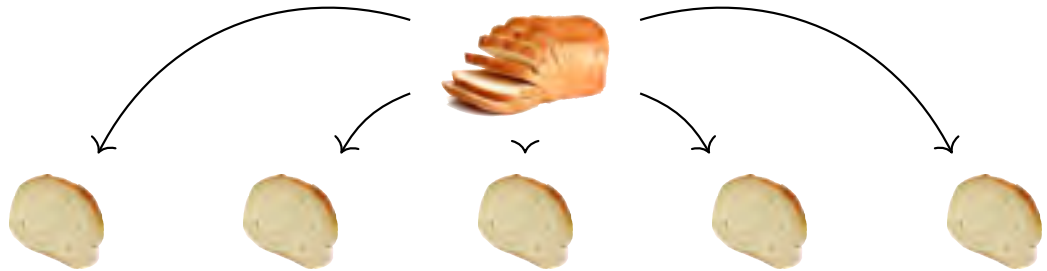
The Algorithm in Practice

A *Generic* Differentially Private Estimator

A *Generic* Differentially Private Estimator

Private data

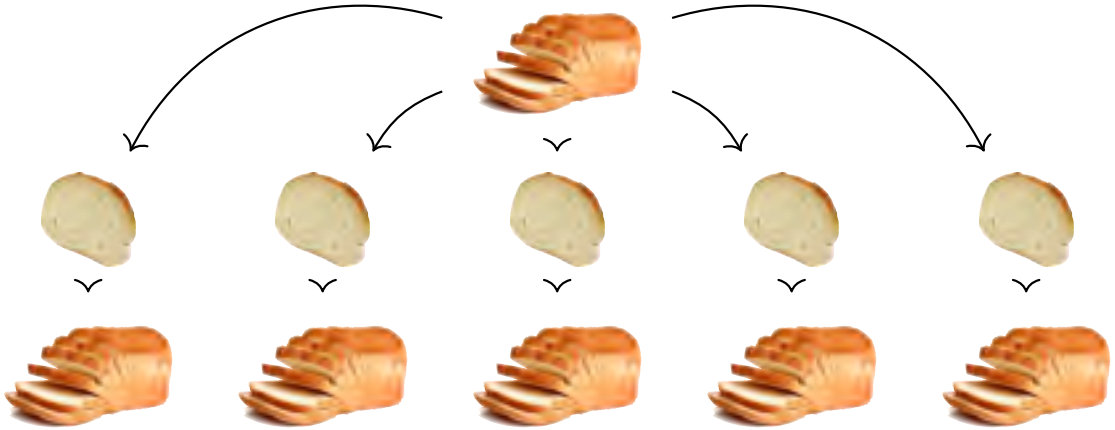
A *Generic* Differentially Private Estimator



Private data

Partition

A *Generic* Differentially Private Estimator

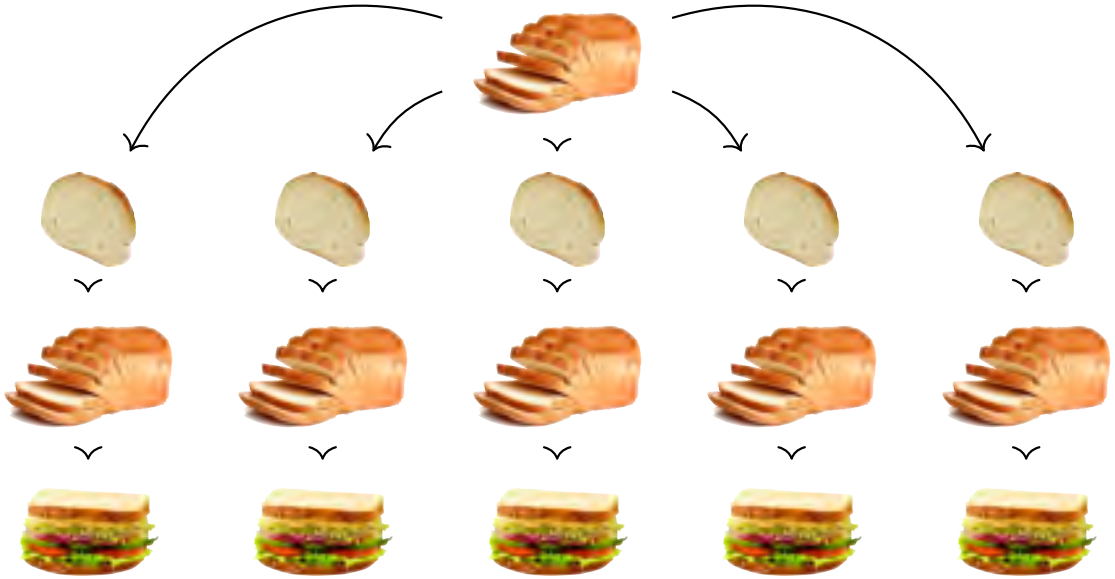


Private data

Partition

Bag of little bootstraps

A *Generic* Differentially Private Estimator



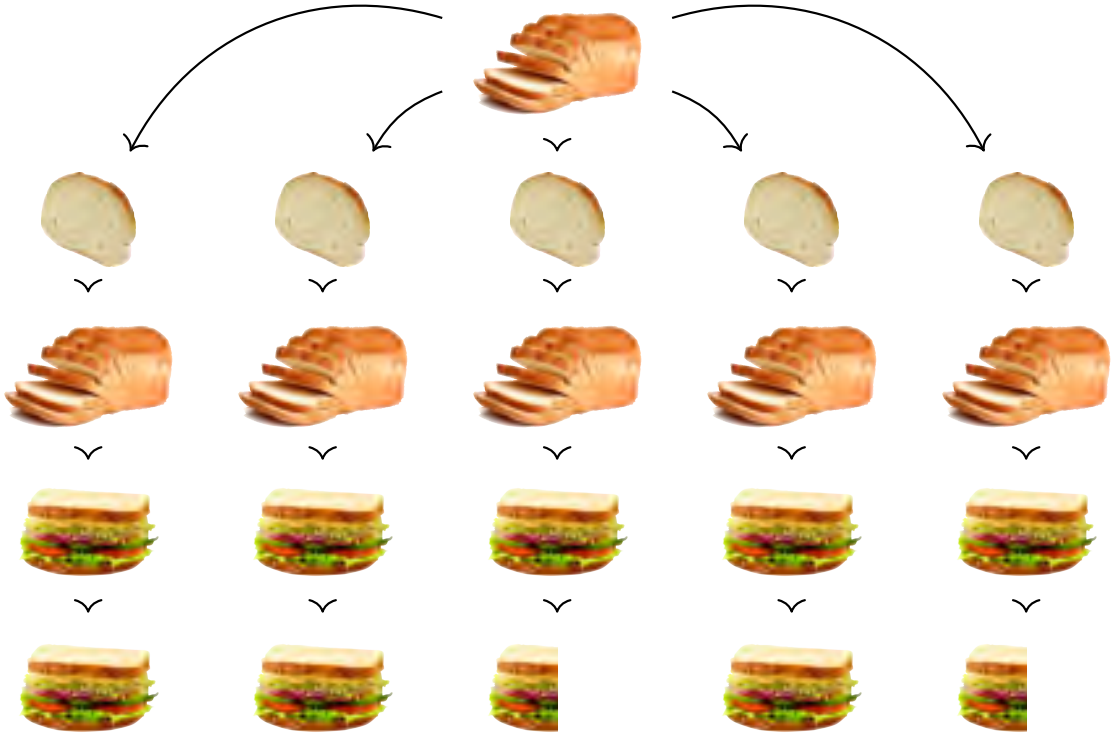
Private data

Partition

Bag of little bootstraps

Estimator

A *Generic* Differentially Private Estimator



Private data

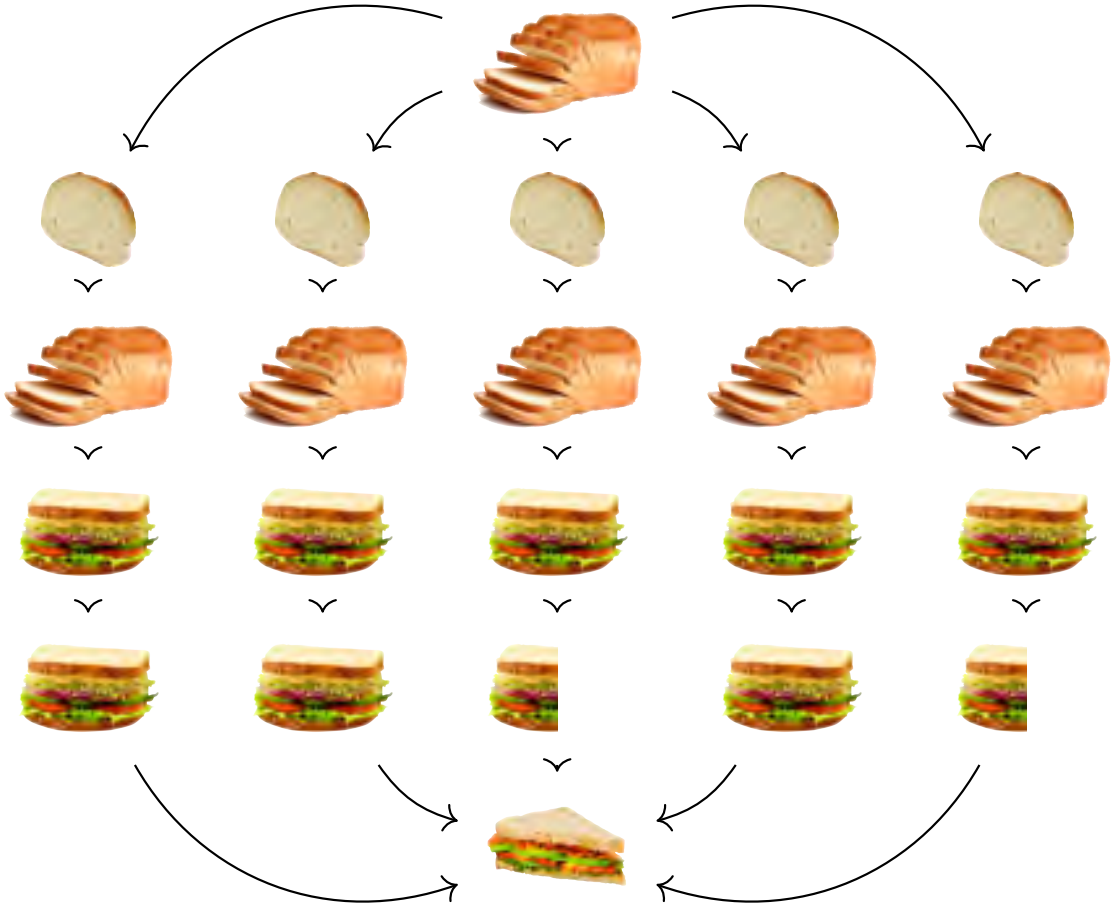
Partition

Bag of little bootstraps

Estimator

Censor

A *Generic* Differentially Private Estimator



Private data

Partition

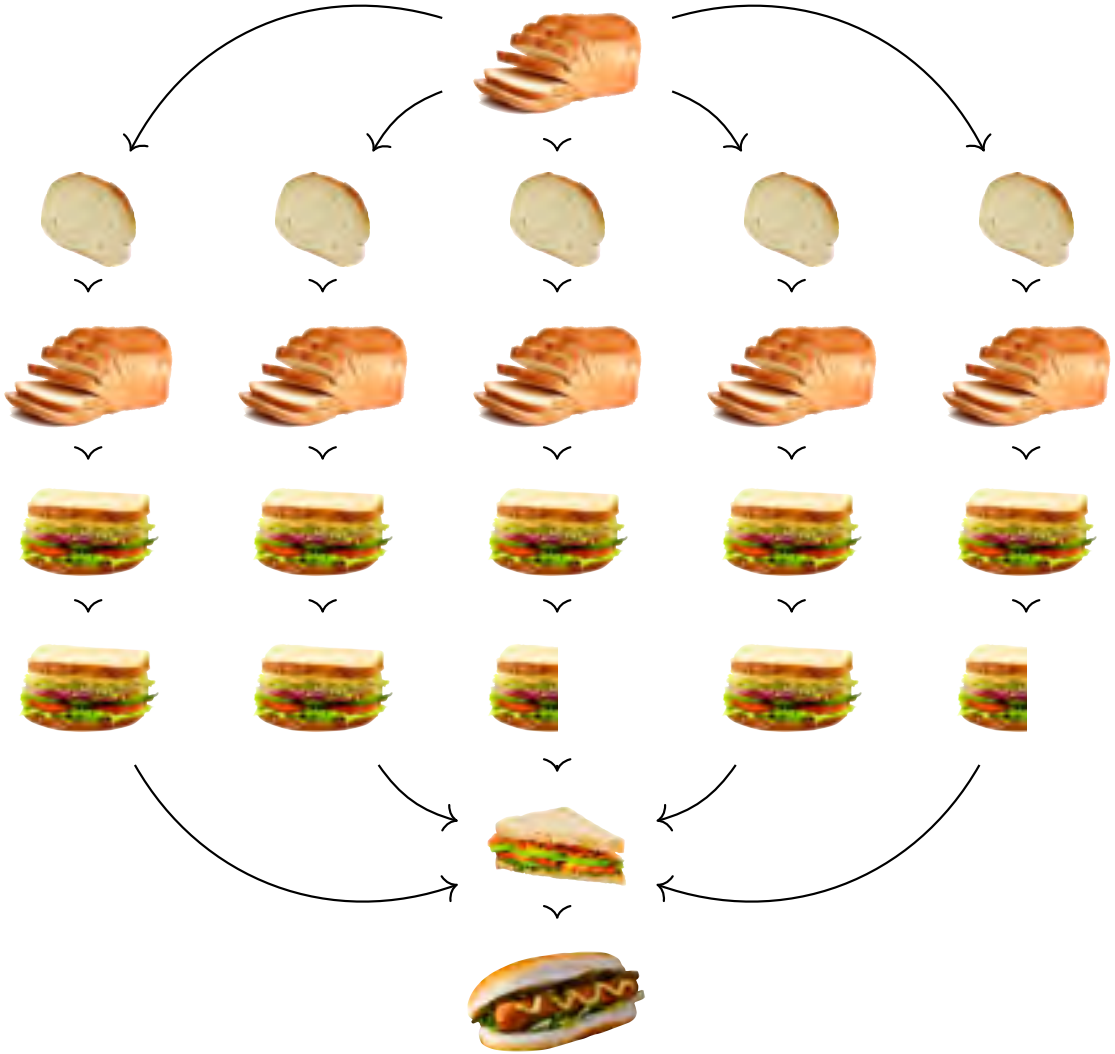
Bag of little bootstraps

Estimator

Censor

Average

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

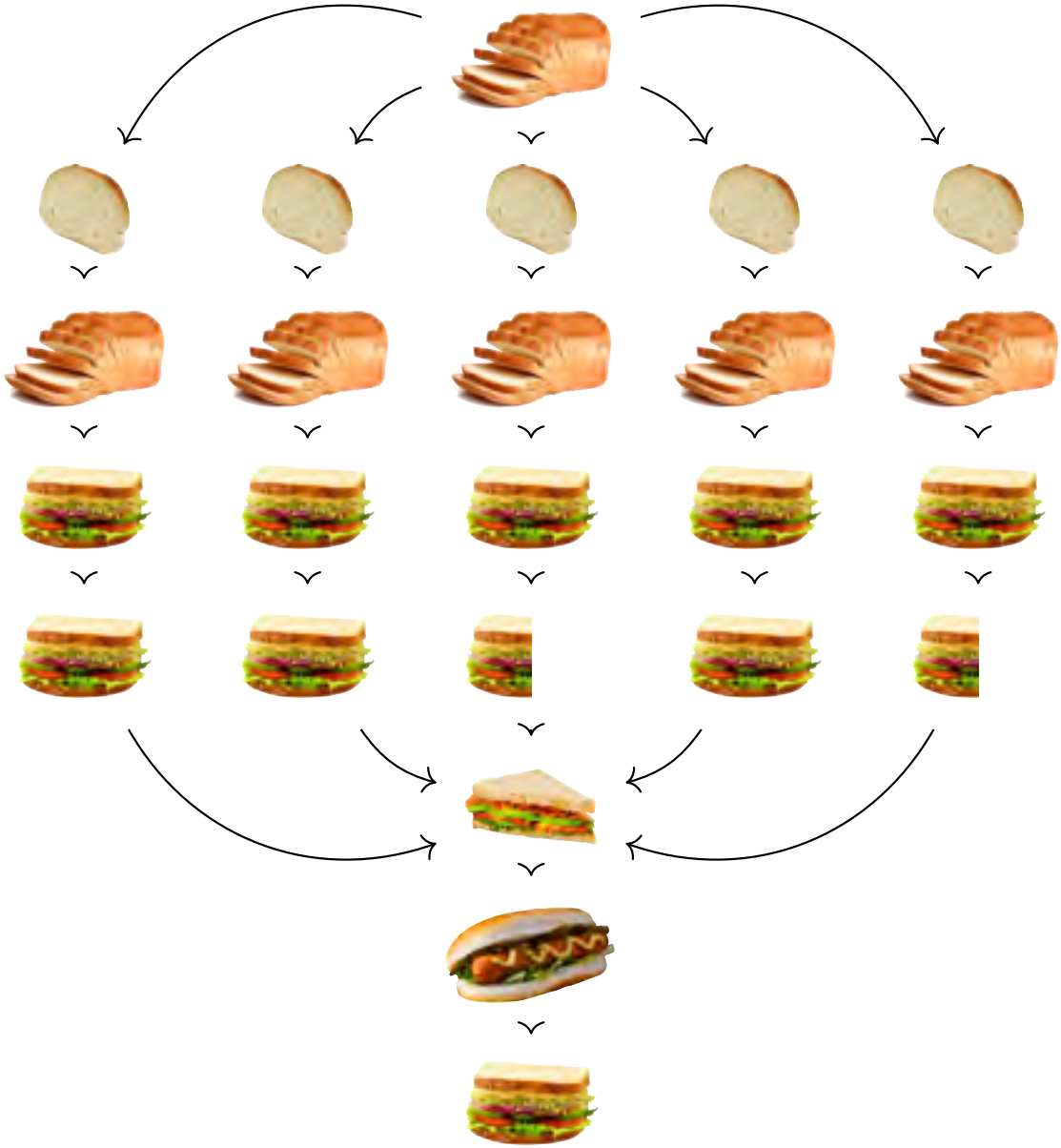
Estimator

Censor

Average

Noise

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

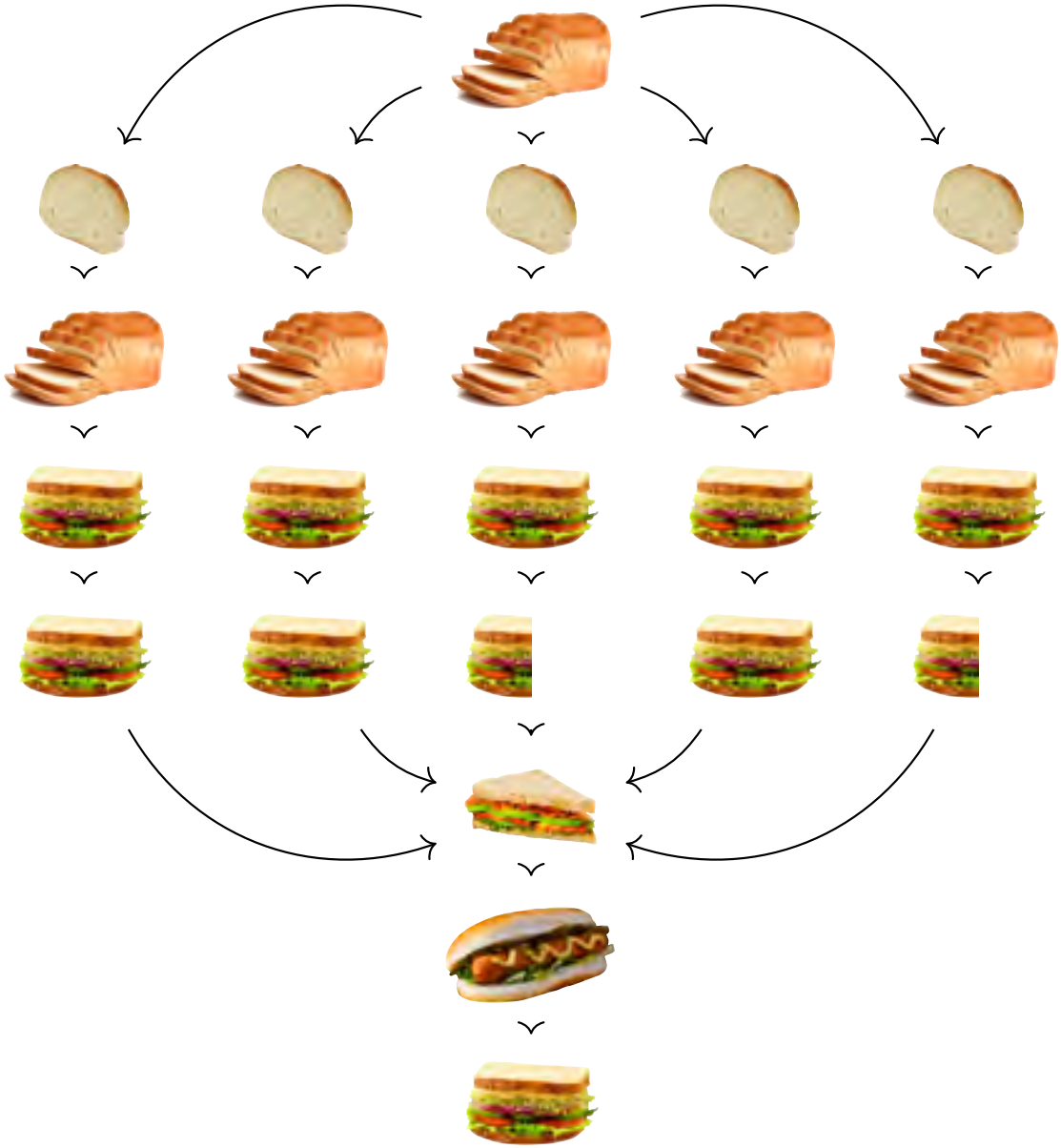
Censor

Average

Noise

Bias Correction

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

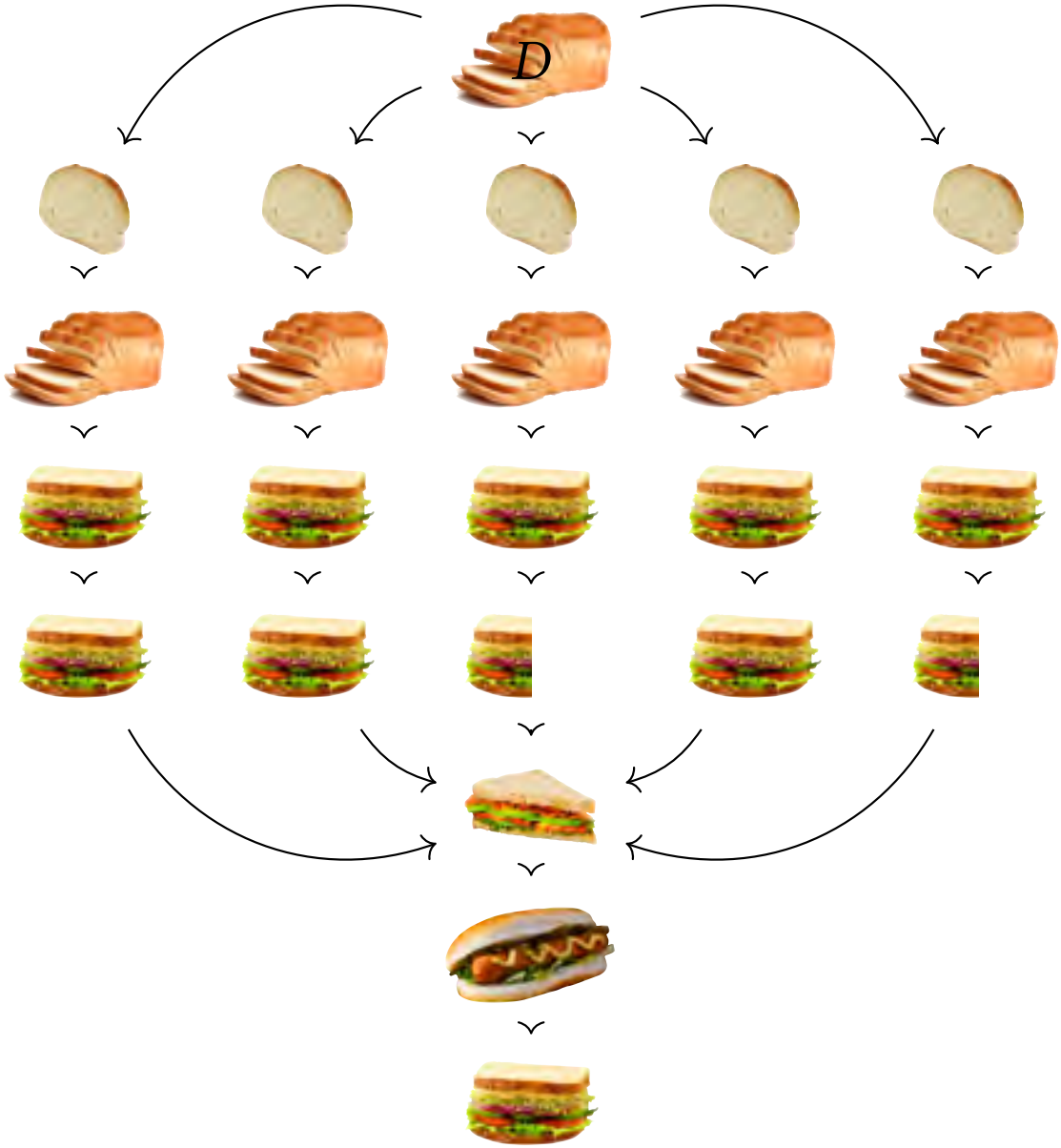
Censor

Average

Noise

Bias Correction
(& variance estimation)

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

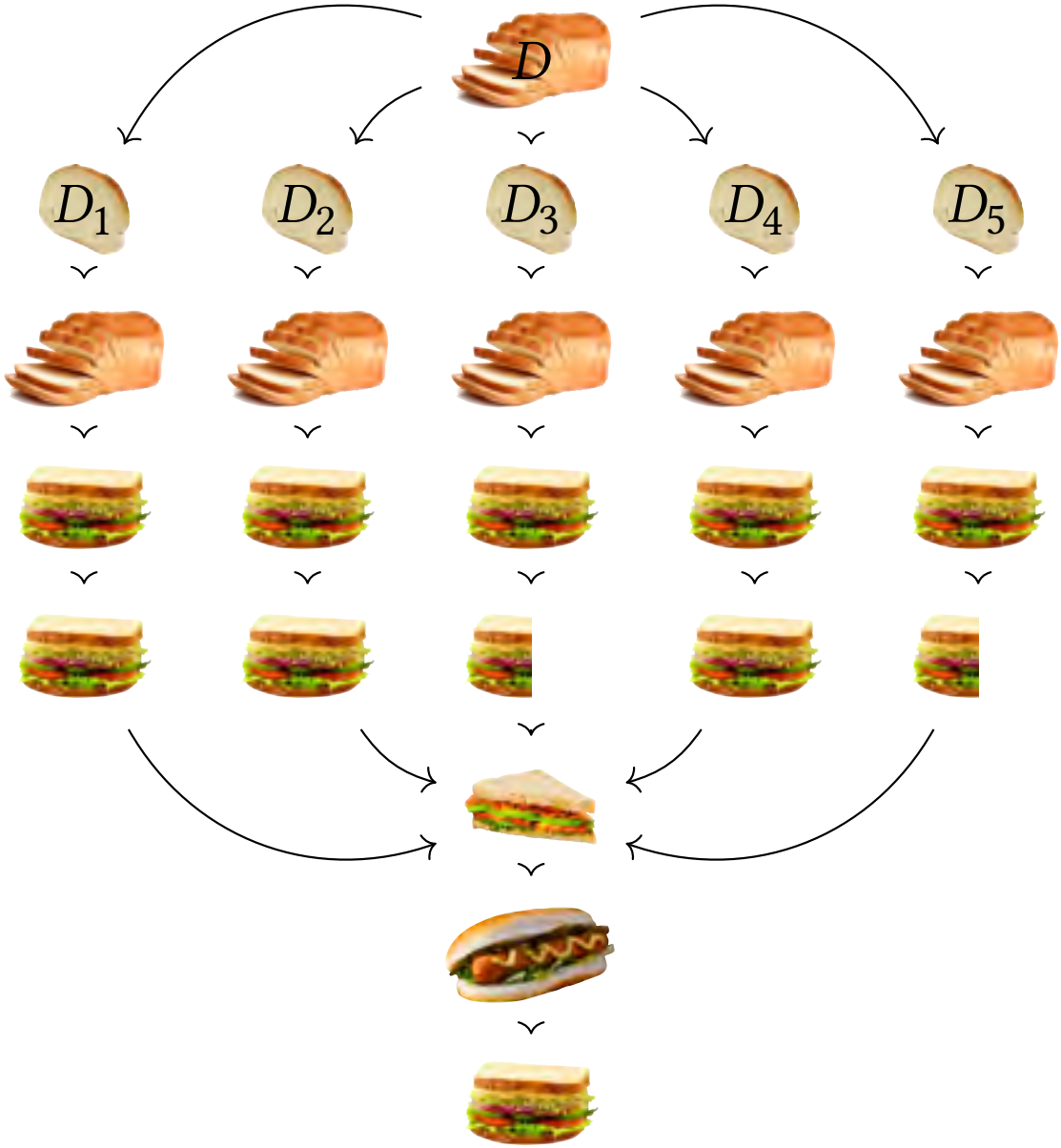
Censor

Average

Noise

Bias Correction
(& variance estimation)

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

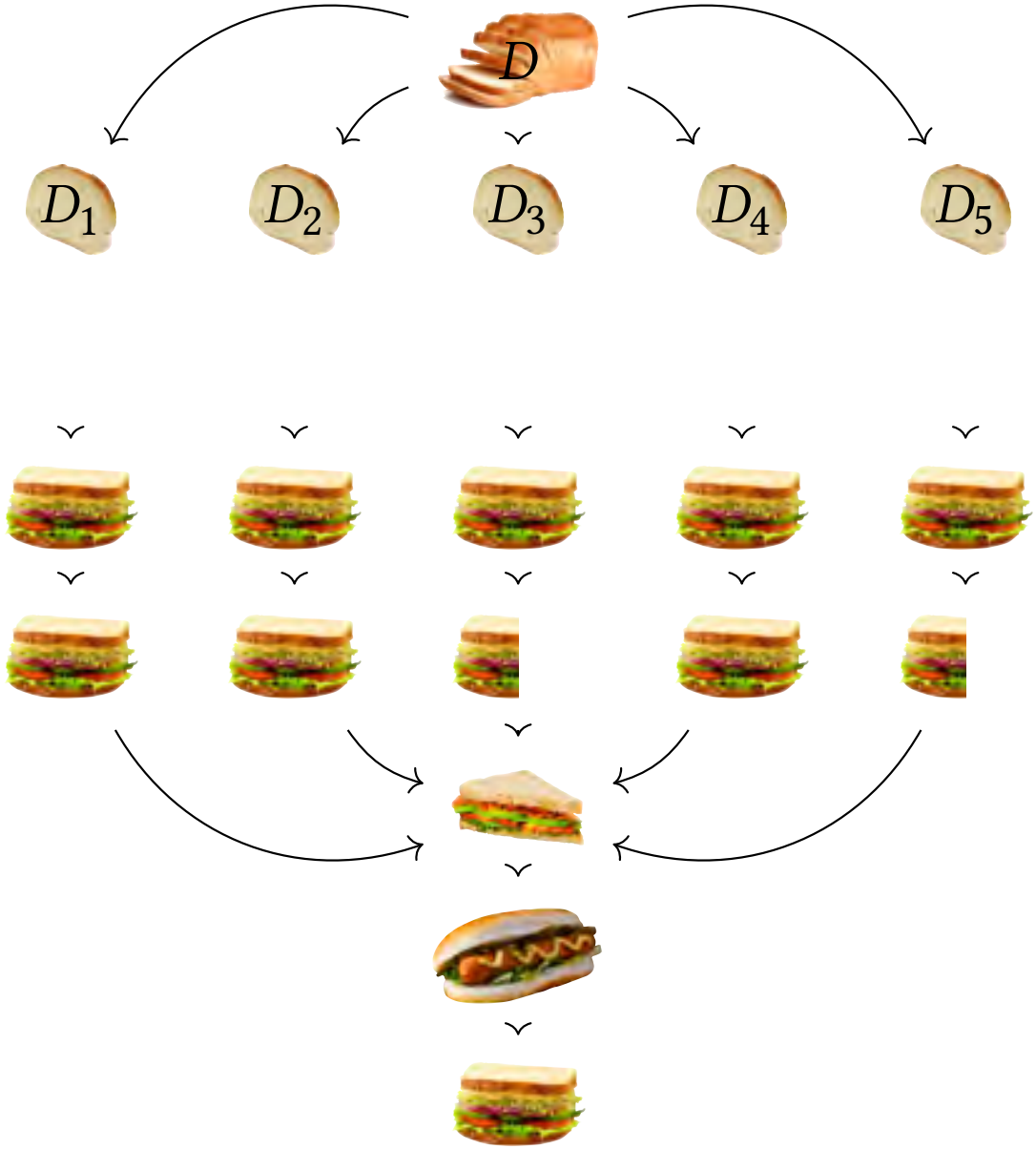
Censor

Average

Noise

Bias Correction
(& variance estimation)

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

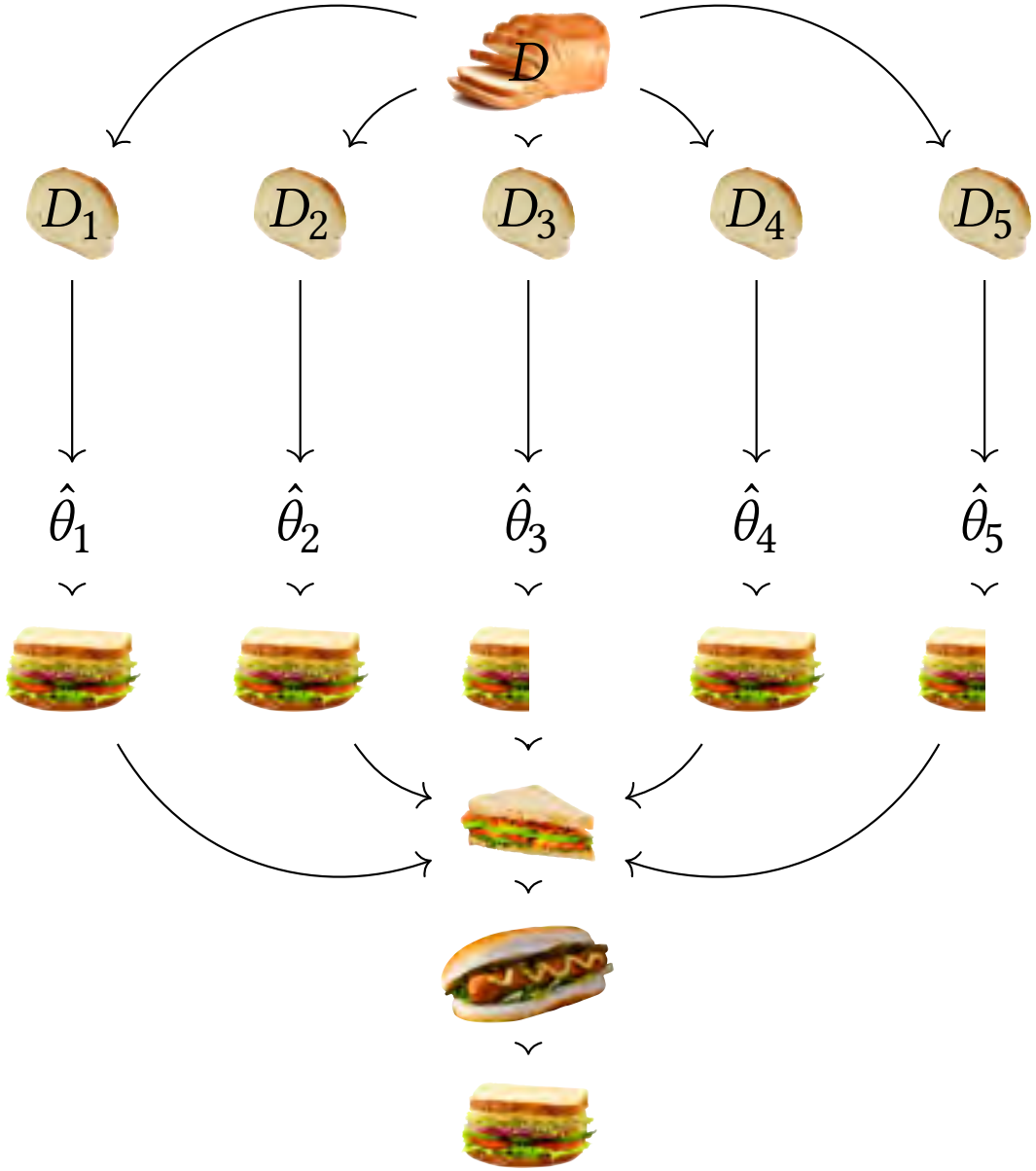
Censor

Average

Noise

Bias Correction
(& variance estimation)

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

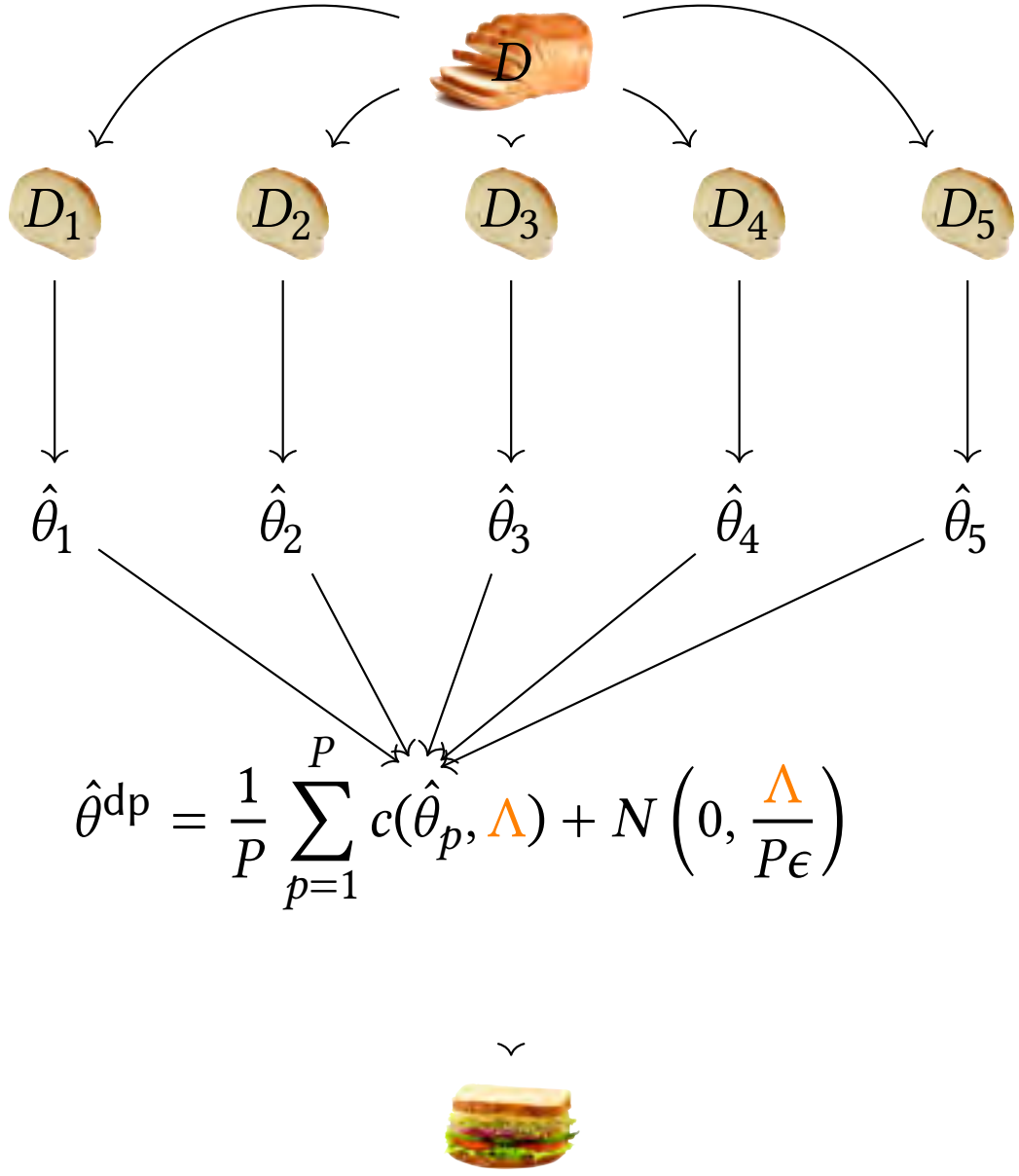
Censor

Average

Noise

Bias Correction
(& variance estimation)

A *Generic* Differentially Private Estimator



Private data

Partition

Bag of little bootstraps

Estimator

Censor

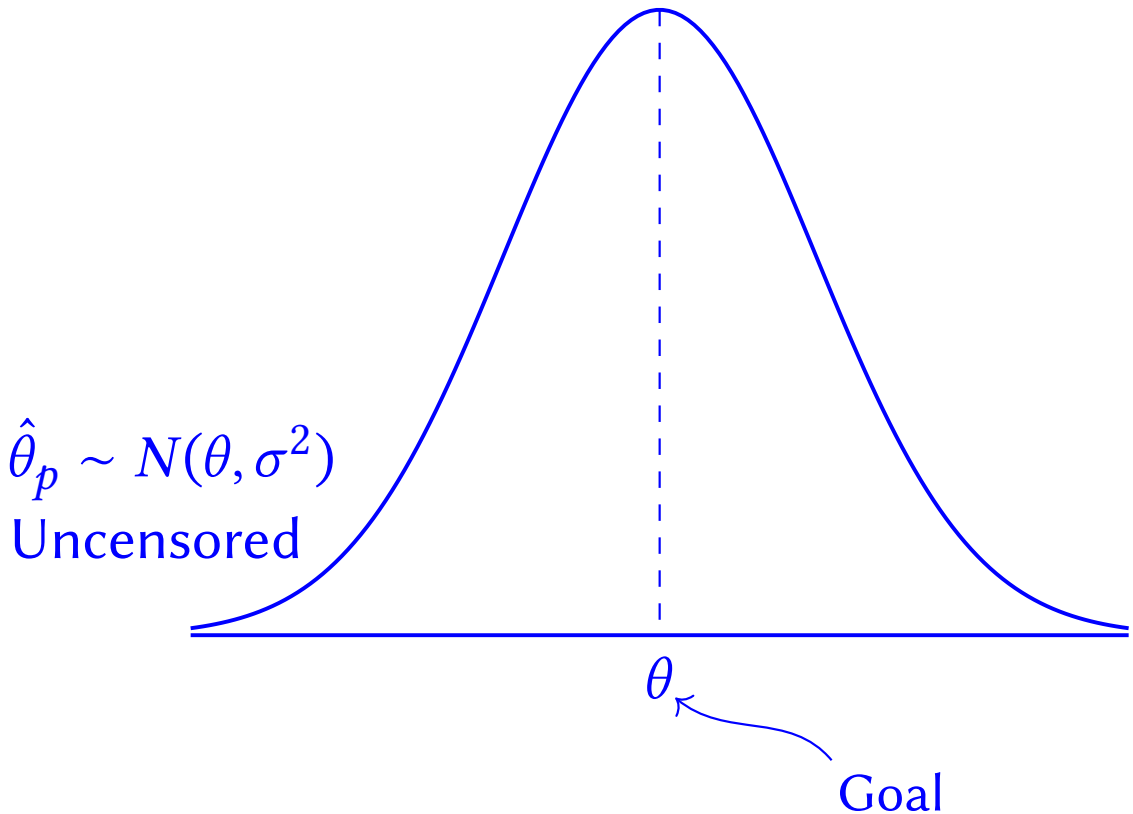
Average

Noise

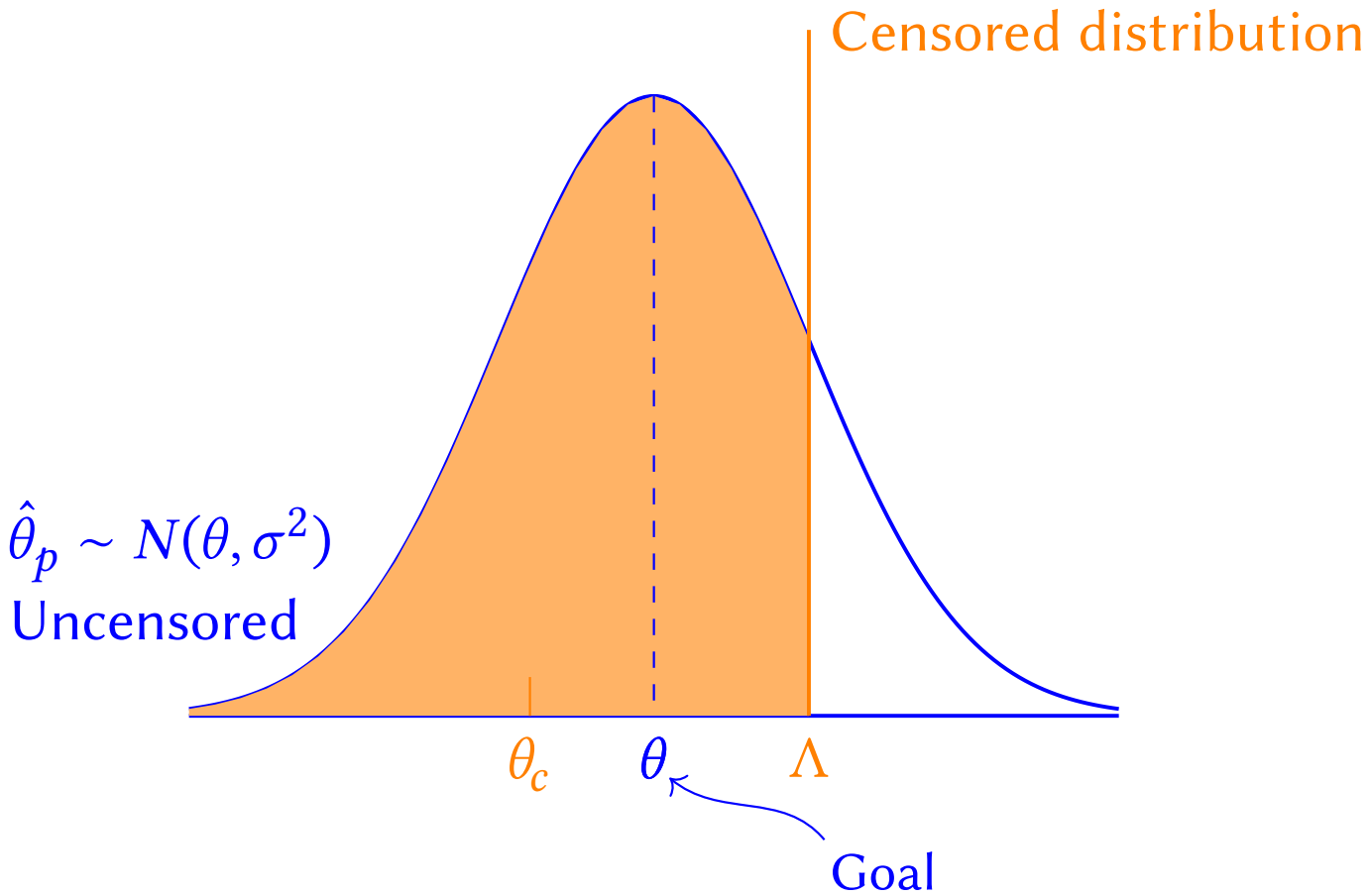
Bias Correction
(& variance estimation)

Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)

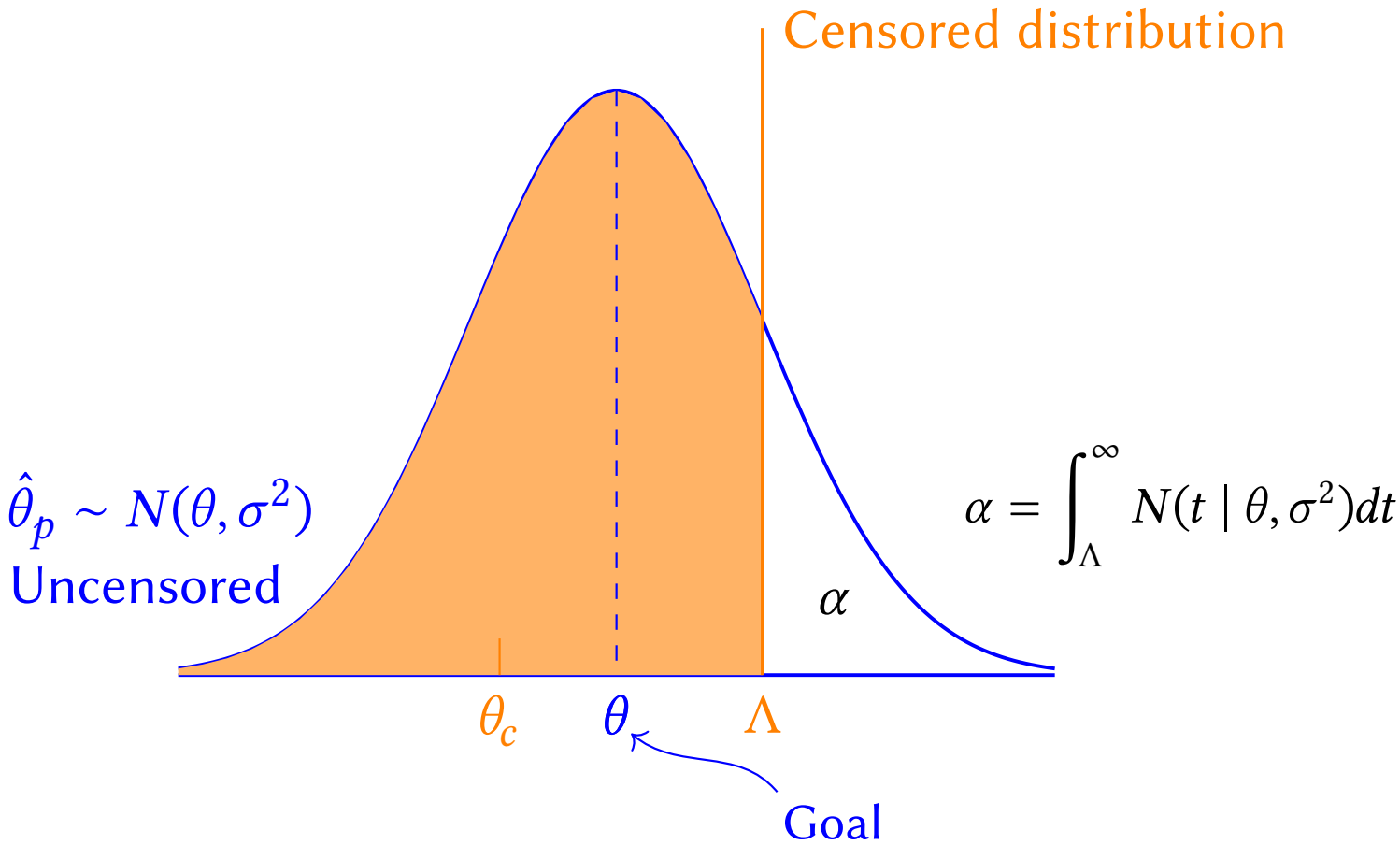
Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



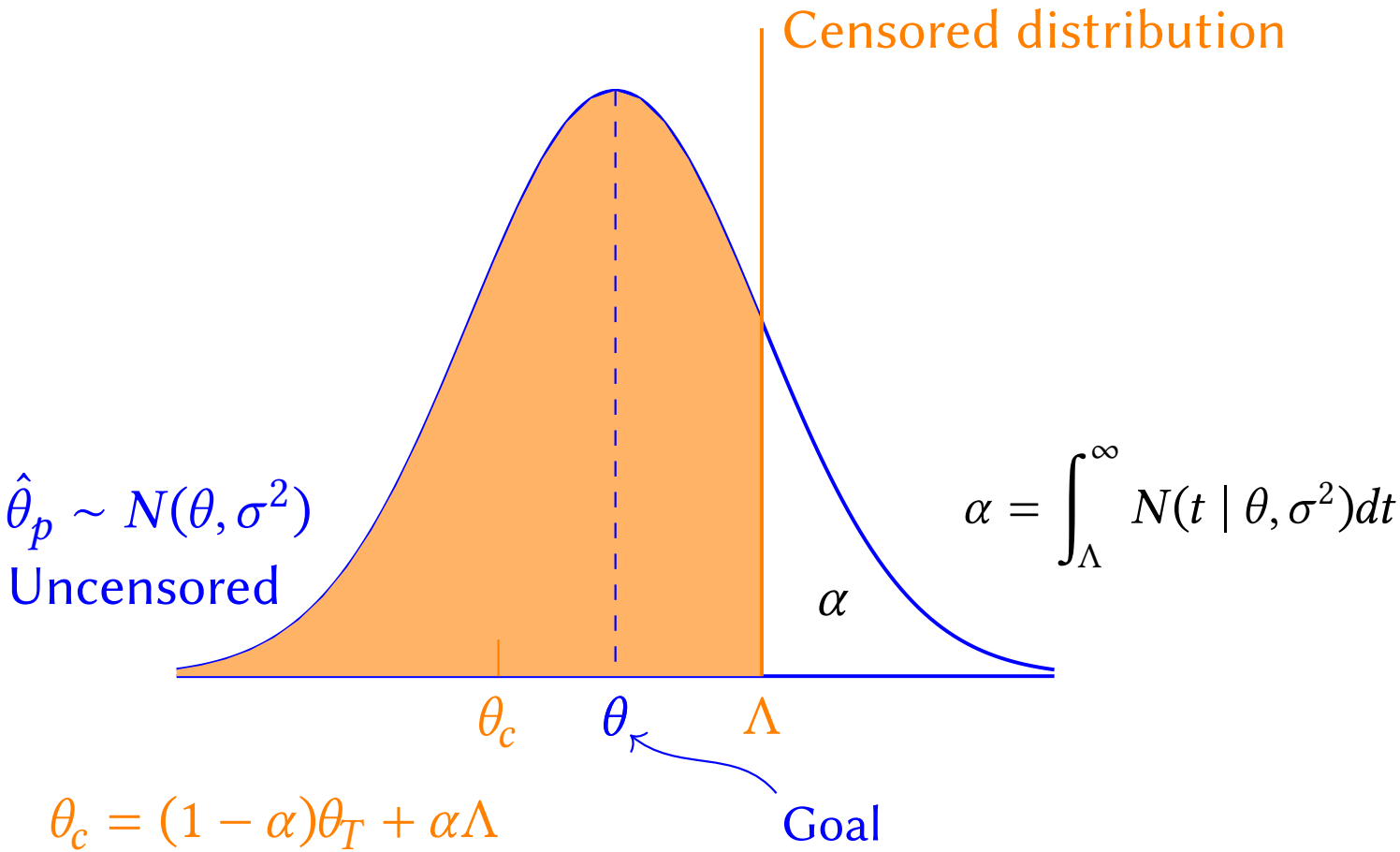
Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



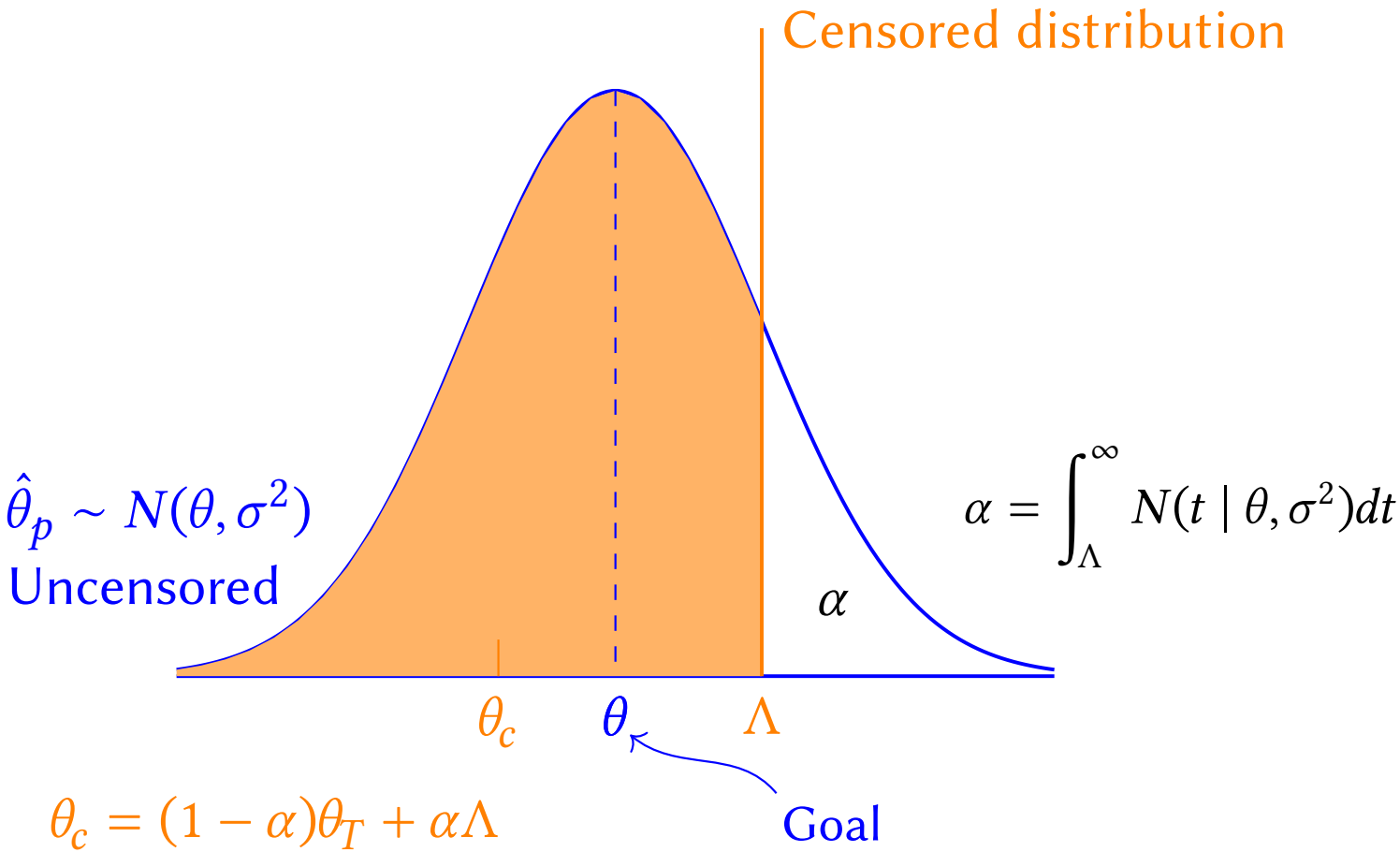
Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)

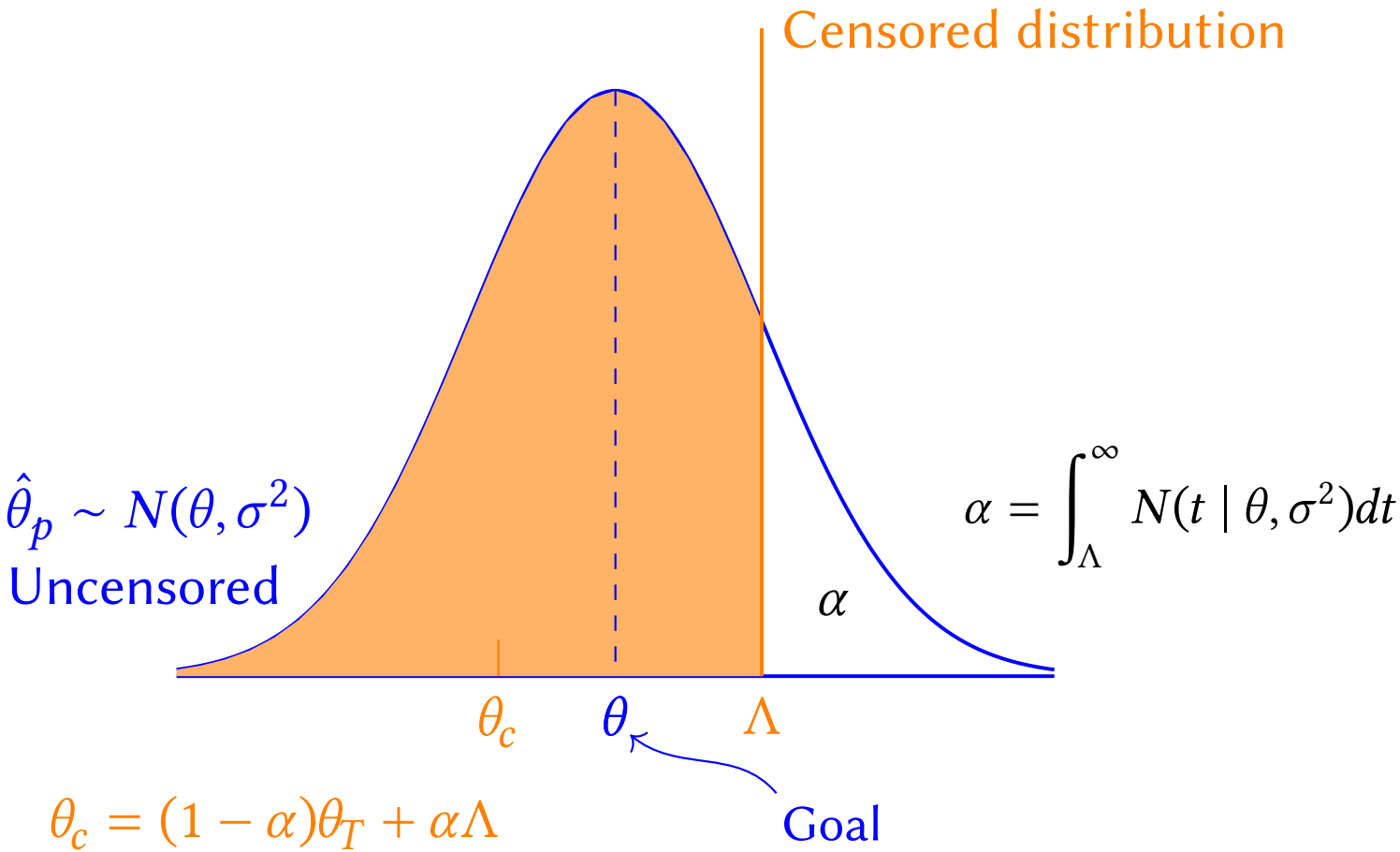


Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



Equations: 2

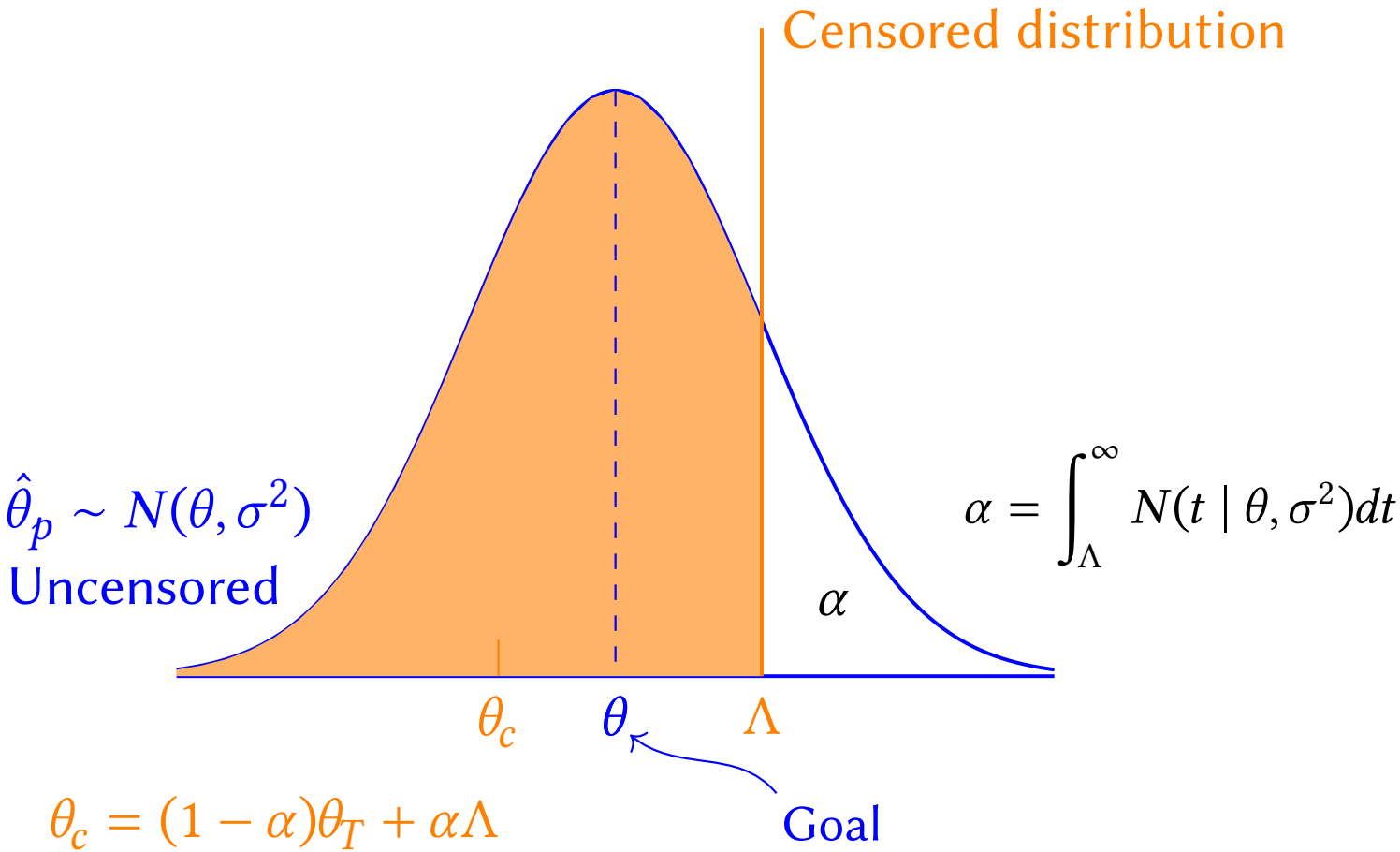
Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



Equations: 2

Unknowns: θ, σ^2, α, θ_c

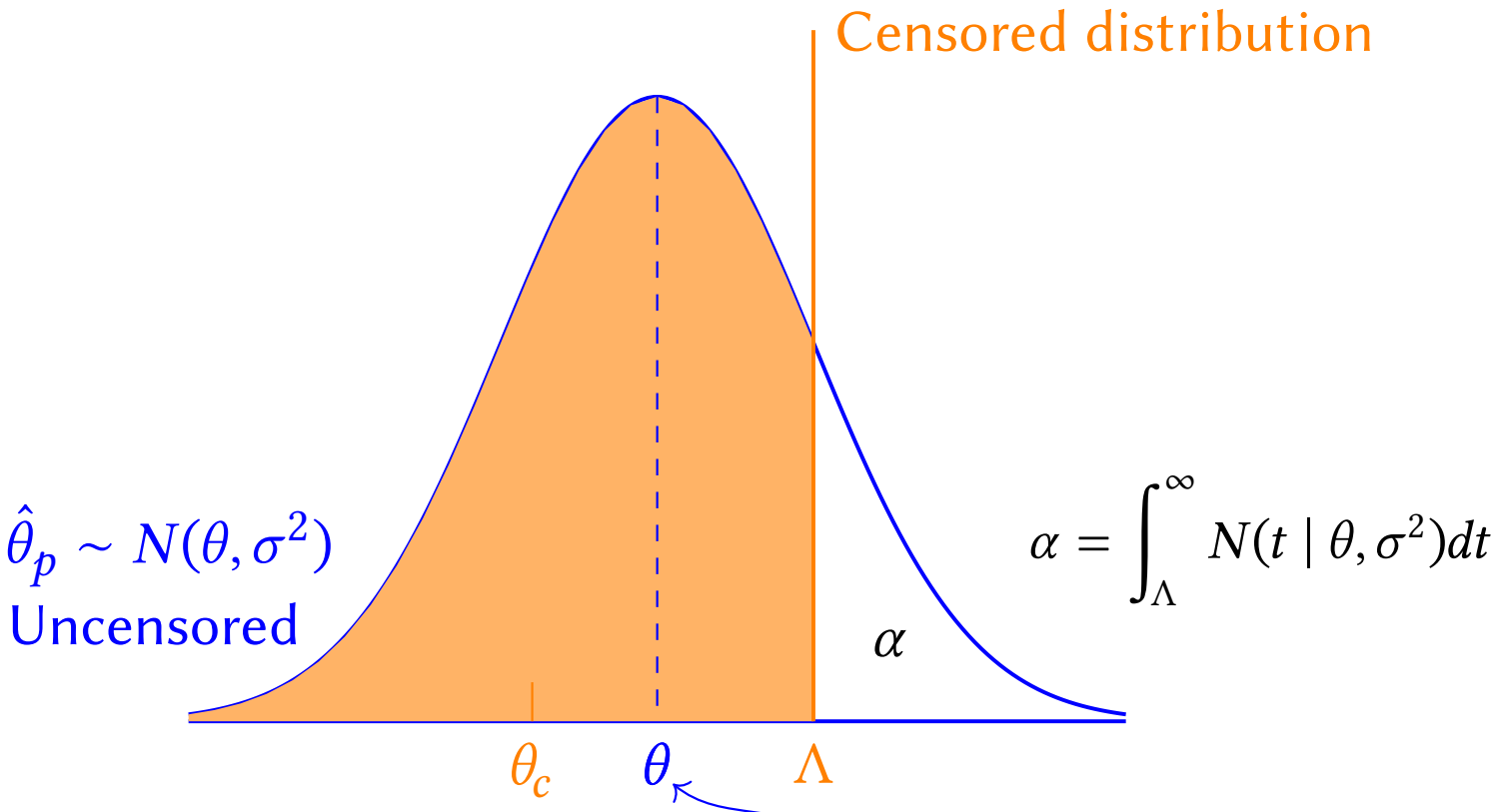
Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



Equations: 2

Unknowns: θ, σ^2, α, ~~θ_c~~ ✗

Bias Correction of: $\hat{\theta}^{\text{dp}} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}_p, \Lambda) + N\left(0, \frac{\Lambda}{P\epsilon}\right)$ (Λ, P, ϵ known)



$$\alpha = \int_{\Lambda}^{\infty} N(t \mid \theta, \sigma^2)dt$$

$\theta_c = (1-\alpha)\theta_T + \alpha\Lambda$

Disclose: $\hat{\theta}^{\text{dp}}$, $\hat{\alpha}^{\text{dp}}$

Equations: 2

Unknowns: θ, σ^2, ~~α~~, ~~θ_c~~ ✗ ✗

Variance Estimation

Variance Estimation

- Simulate estimates via standard (Clarify) procedures:

$$\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}} \sim N\left(\begin{bmatrix} \hat{\theta}^{\text{dp}} \\ \hat{\alpha}^{\text{dp}} \end{bmatrix}, \begin{bmatrix} \hat{V}(\hat{\theta}^{\text{dp}}) & \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) \\ \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) & \hat{V}(\hat{\alpha}^{\text{dp}}) \end{bmatrix}\right)$$

Variance Estimation

- Simulate estimates via standard (Clarify) procedures:

$$\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}} \sim N\left(\begin{bmatrix} \hat{\theta}^{\text{dp}} \\ \hat{\alpha}^{\text{dp}} \end{bmatrix}, \begin{bmatrix} \hat{V}(\hat{\theta}^{\text{dp}}) & \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) \\ \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) & \hat{V}(\hat{\alpha}^{\text{dp}}) \end{bmatrix}\right)$$

Functions of disclosed params

Variance Estimation

- Simulate estimates via standard (Clarify) procedures:

$$\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}} \sim N\left(\begin{bmatrix} \hat{\theta}^{\text{dp}} \\ \hat{\alpha}^{\text{dp}} \end{bmatrix}, \begin{bmatrix} \hat{V}(\hat{\theta}^{\text{dp}}) & \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) \\ \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) & \hat{V}(\hat{\alpha}^{\text{dp}}) \end{bmatrix}\right)$$

Functions of disclosed params

- Bias correct simulated params:

$$\{\tilde{\theta}^{\text{dp}}, \hat{\sigma}^2_{\text{dp}}\} = \text{BiasCorrect}\left[\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}}\right]$$

Variance Estimation

- Simulate estimates via standard (Clarify) procedures:

$$\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}} \sim N\left(\begin{bmatrix} \hat{\theta}^{\text{dp}} \\ \hat{\alpha}^{\text{dp}} \end{bmatrix}, \begin{bmatrix} \hat{V}(\hat{\theta}^{\text{dp}}) & \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) \\ \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) & \hat{V}(\hat{\alpha}^{\text{dp}}) \end{bmatrix}\right)$$

Functions of disclosed params

- Bias correct simulated params:

$$\{\tilde{\theta}^{\text{dp}}, \hat{\sigma}^2_{\text{dp}}\} = \text{BiasCorrect}\left[\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}}\right]$$

- Standard error: Standard deviation of $\tilde{\theta}^{\text{dp}}$ over simulations

Variance Estimation

- Simulate estimates via standard (Clarify) procedures:

$$\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}} \sim N\left(\begin{bmatrix} \hat{\theta}^{\text{dp}} \\ \hat{\alpha}^{\text{dp}} \end{bmatrix}, \begin{bmatrix} \hat{V}(\hat{\theta}^{\text{dp}}) & \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) \\ \widehat{\text{Cov}}(\hat{\alpha}^{\text{dp}}, \hat{\theta}^{\text{dp}}) & \hat{V}(\hat{\alpha}^{\text{dp}}) \end{bmatrix}\right)$$

Functions of disclosed params

- Bias correct simulated params:

$$\{\tilde{\theta}^{\text{dp}}, \hat{\sigma}^2_{\text{dp}}\} = \text{BiasCorrect}\left[\hat{\theta}^{\text{dp}}, \hat{\alpha}^{\text{dp}}\right]$$

- Standard error: Standard deviation of $\tilde{\theta}^{\text{dp}}$ over simulations
- Bias correction: reduces bias *and* variance

Solving Political Problems Technologically

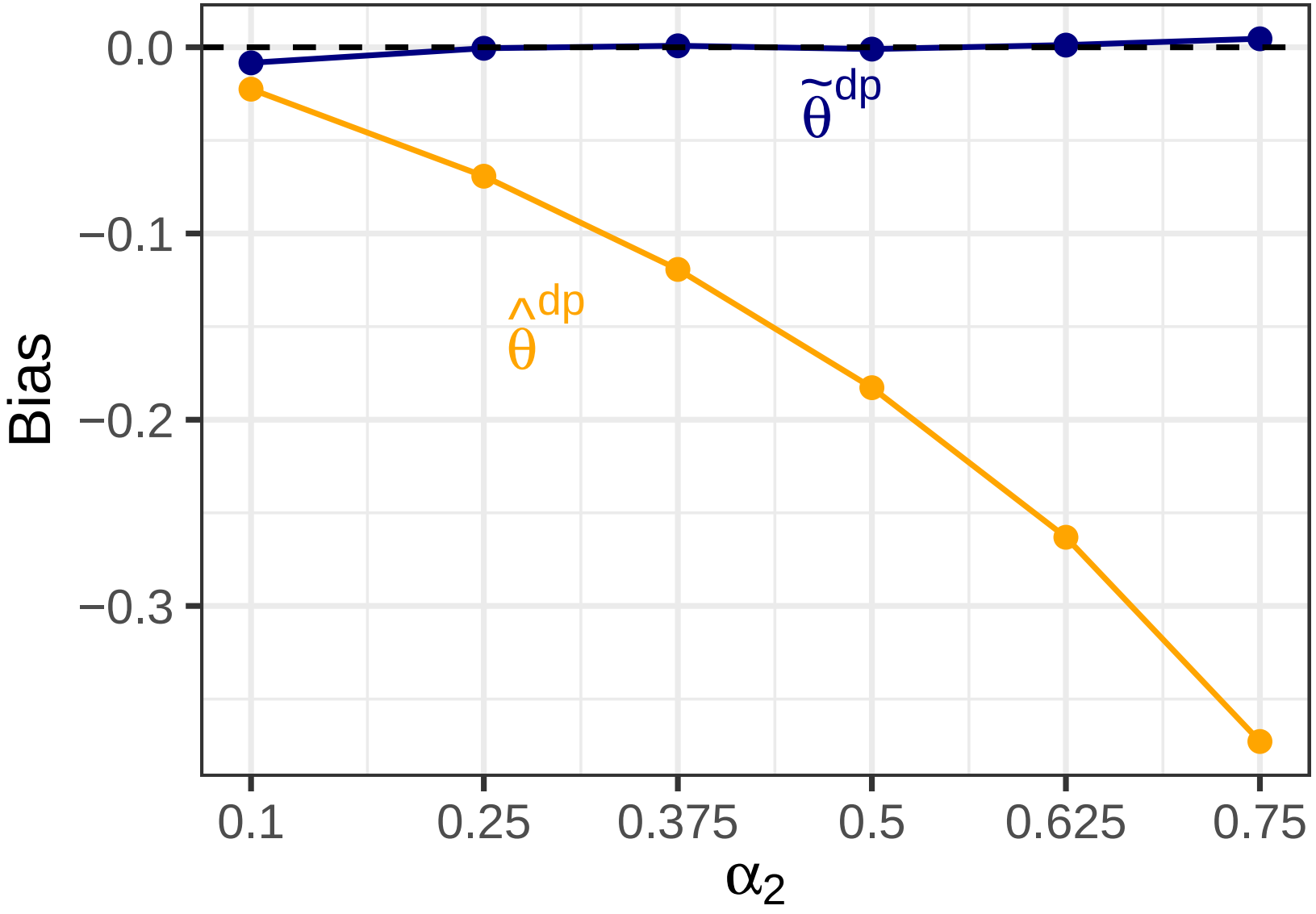
Differential Privacy & Inferential Validity

A General Purpose, Statistically Valid DP Algorithm

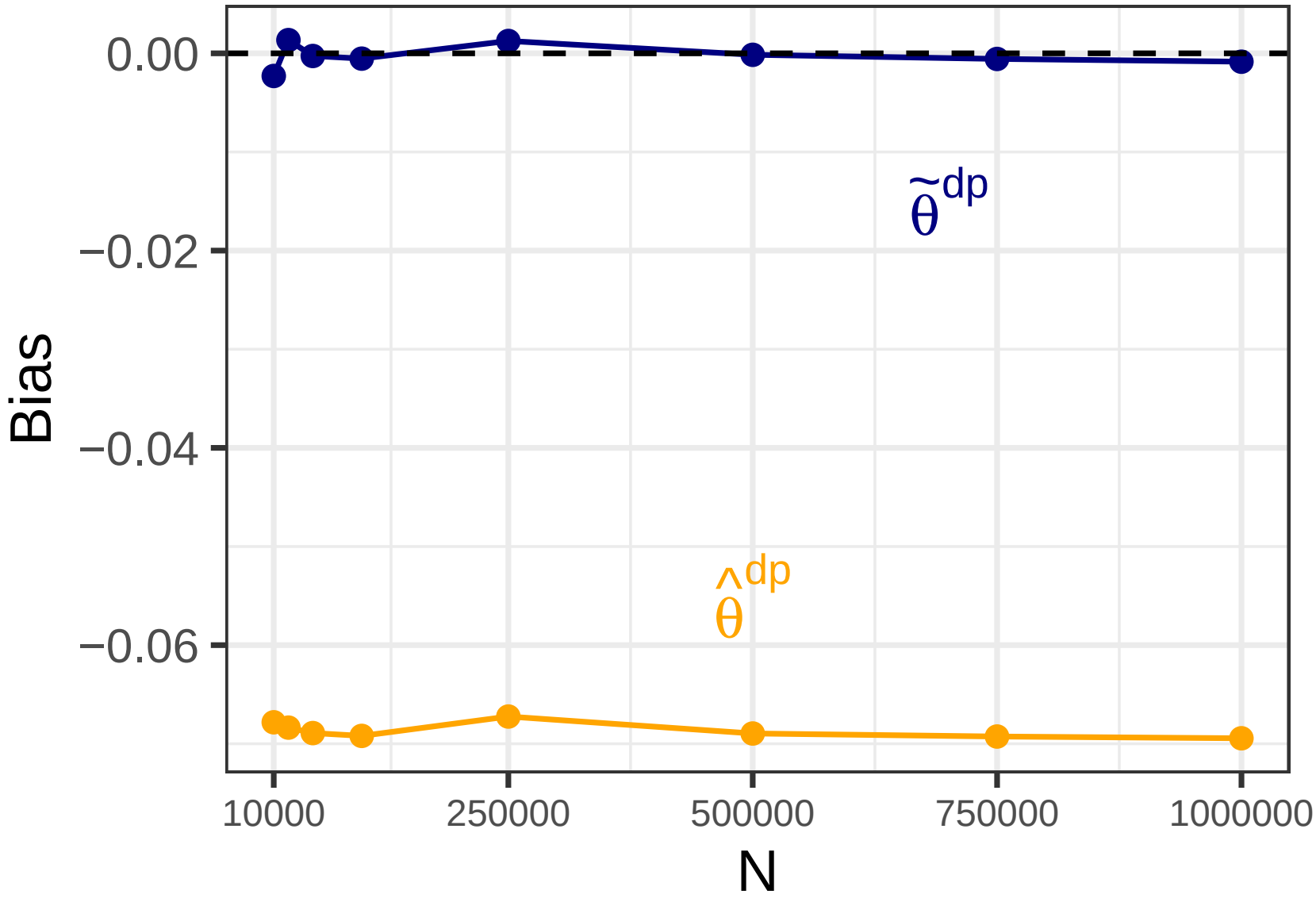
The Algorithm in Practice

Simulations: Finite Sample Evaluation

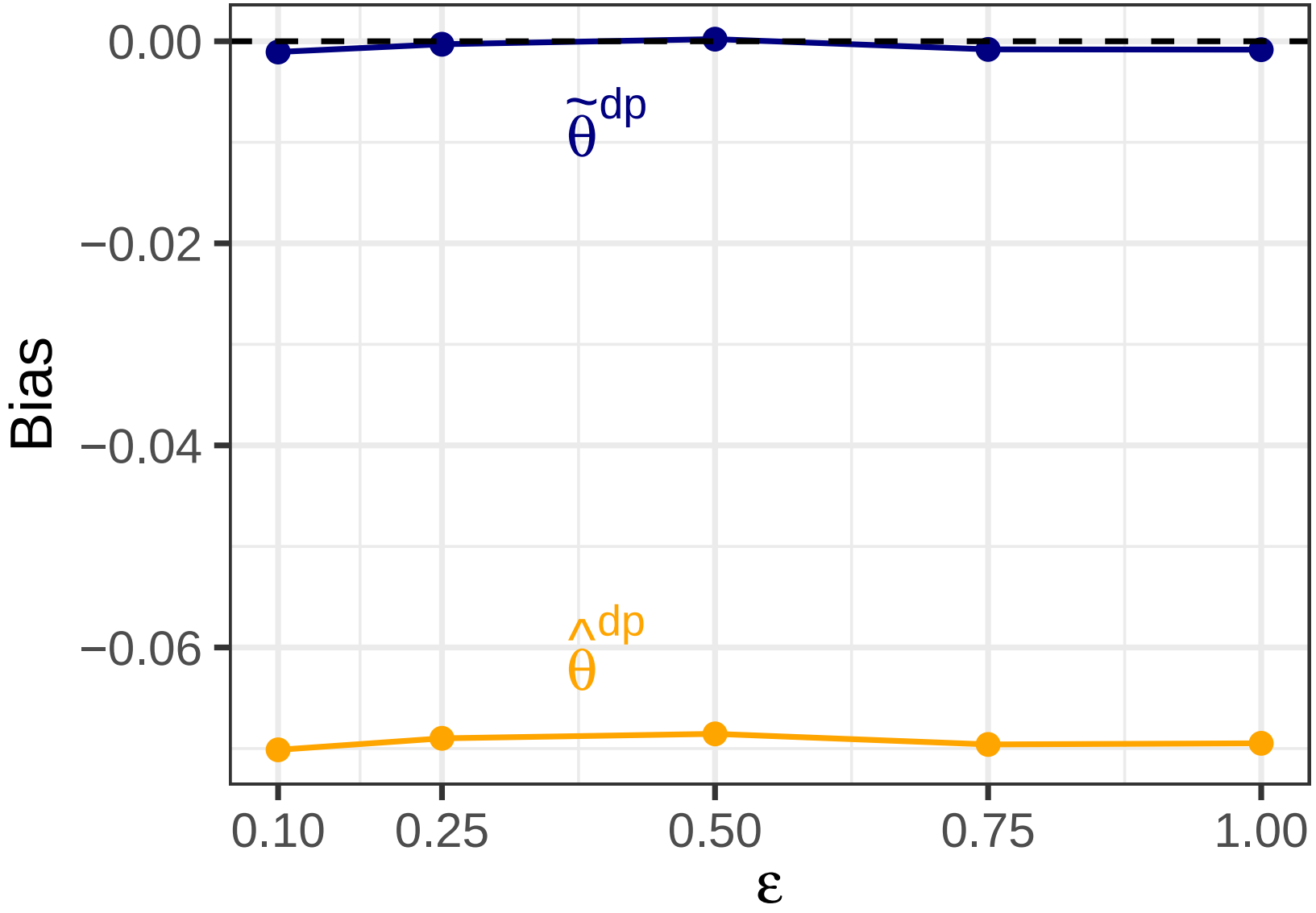
Simulations: Finite Sample Evaluation



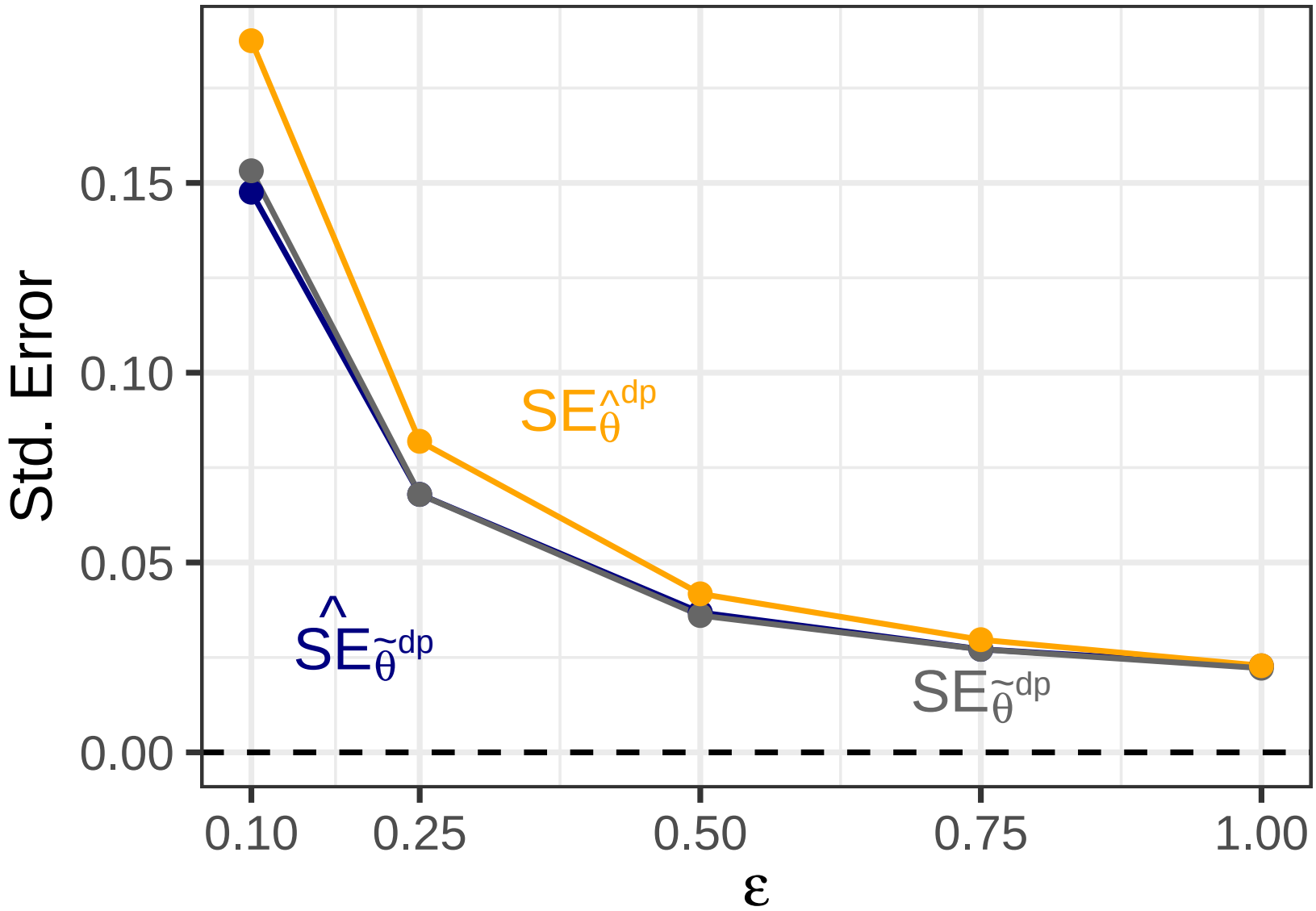
Simulations: Finite Sample Evaluation



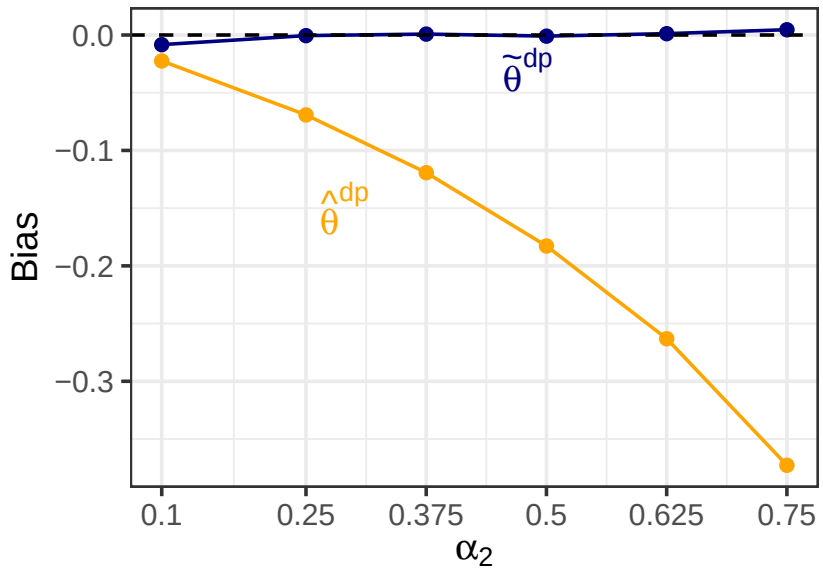
Simulations: Finite Sample Evaluation

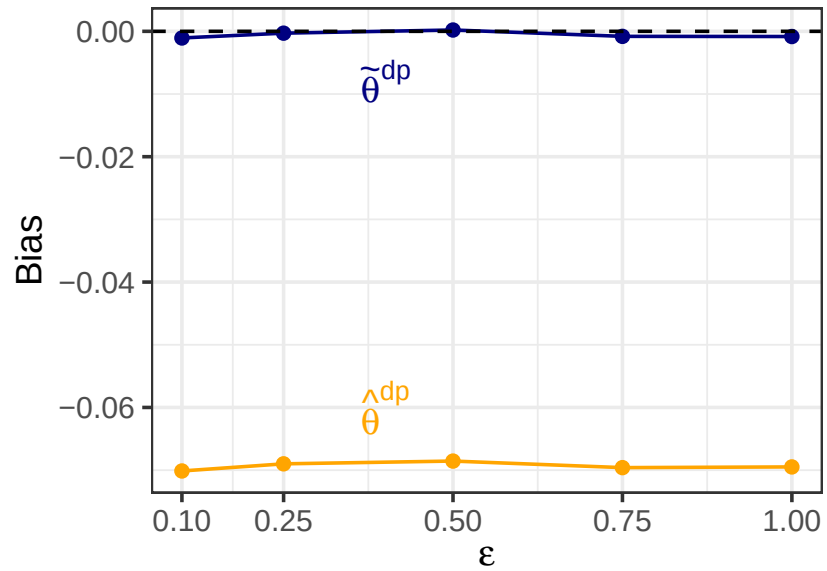


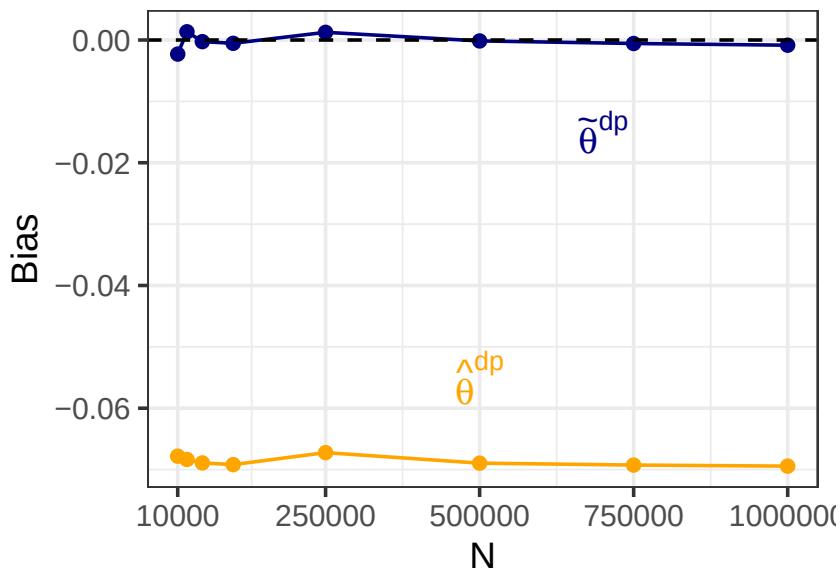
Simulations: Finite Sample Evaluation

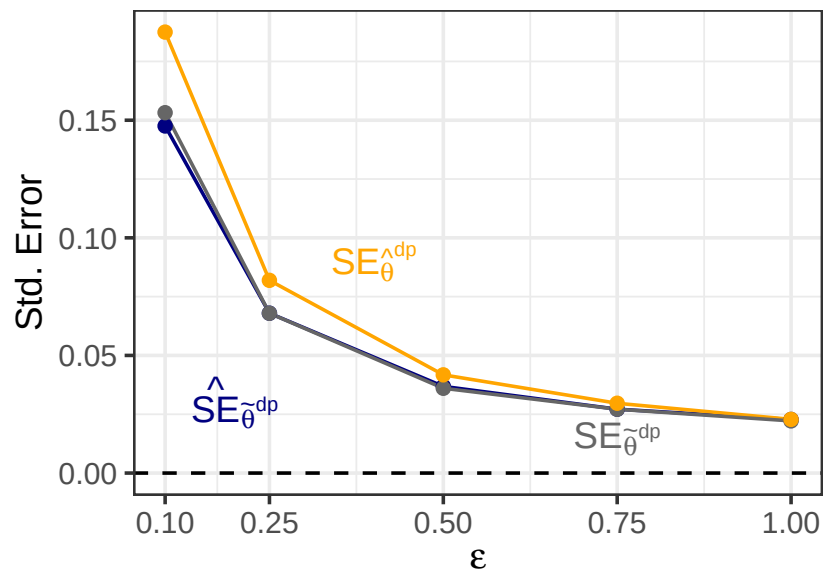


Simulations: Finite Sample Evaluation

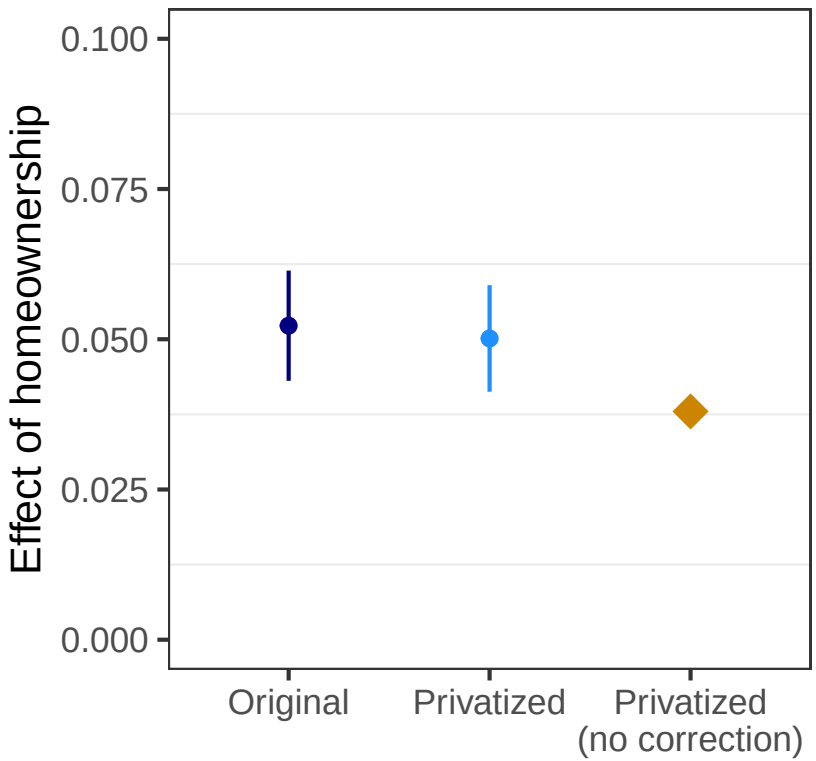




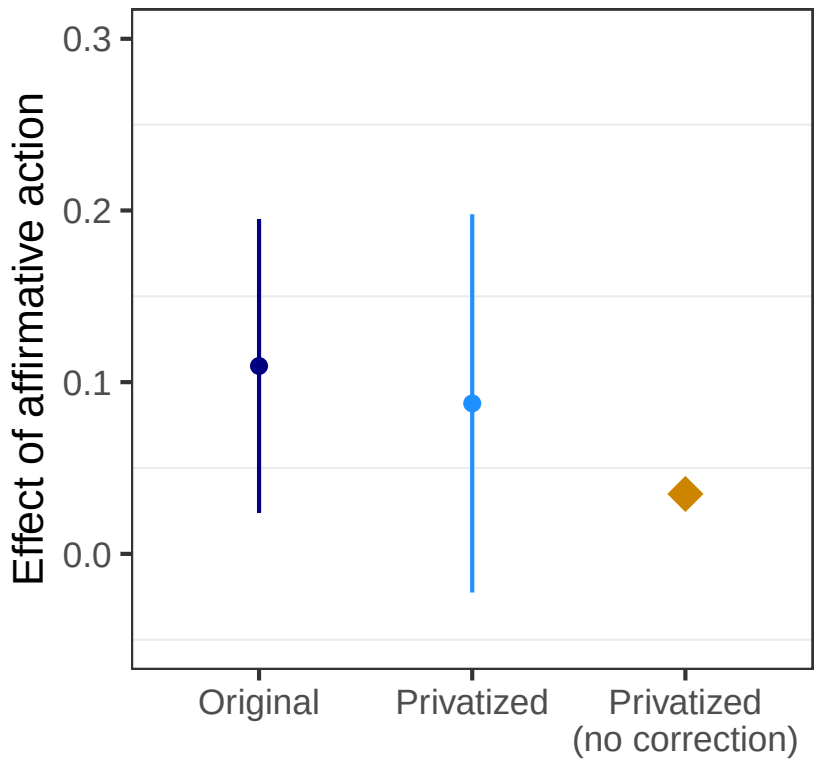




Similar Empirical Results, Larger CIs



(a) Yoder (APSR, 2020)



(b) Bhavnani and Lee (AJPS, 2019)

Concluding Remarks

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance
  - Valid uncertainty estimates,

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance
  - Valid uncertainty estimates,Computationally efficient

Concluding Remarks

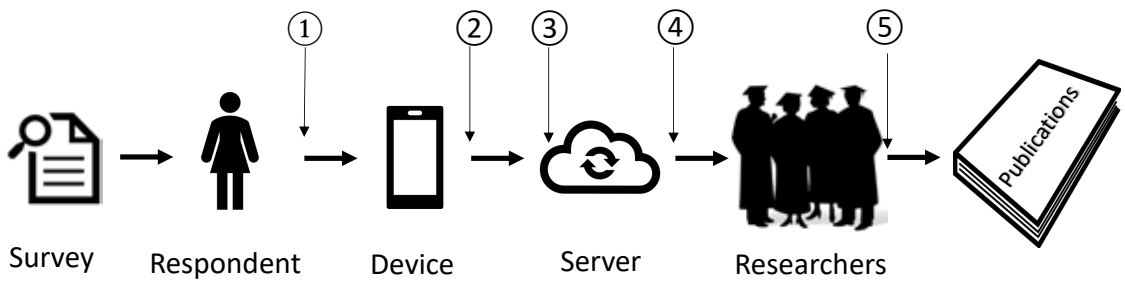
- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance
  - Valid uncertainty estimates,Computationally efficient
  - Solves political problems technologically

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance
  - Valid uncertainty estimates,Computationally efficient
  - Solves political problems technologically
- Community based, Open Source Software: OpenDP.org

Concluding Remarks

- Data sharing $\rightsquigarrow$ data access
  - DP protects individual privacy
  - Enables inference to private database, not population
  - Usually biased, no uncertainty estimates
  - Fails to protect society from fallacious scientific conclusions
- Inferential validity
  - A scientific statement: not necessarily correct, but must have:
  - known statistical properties & valid uncertainty estimates
- Proposed algorithm
  - Generic: almost any statistical method or quantity of interest
  - Statistically unbiased, lower variance
  - Valid uncertainty estimates,Computationally efficient
  - Solves political problems technologically
- Community based, Open Source Software: OpenDP.org
- Lots of options, research in progress:



Articles, software, slides, videos: GaryKing.org/privacy

Articles, software, slides, videos: GaryKing.org/privacy

- Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. "Statistically Valid Inferences from Privacy Protected Data" *American Political Science Review*

Articles, software, slides, videos: GaryKing.org/privacy

- Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. "Statistically Valid Inferences from Privacy Protected Data" *American Political Science Review*
- Georgina Evans, Gary King, Adam D. Smith, Abhradeep Thakurta. "Differentially Private Survey Research" *American Journal of Political Science*

Articles, software, slides, videos: GaryKing.org/privacy

- Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. "Statistically Valid Inferences from Privacy Protected Data" *American Political Science Review*
- Georgina Evans, Gary King, Adam D. Smith, Abhradeep Thakurta. "Differentially Private Survey Research" *American Journal of Political Science*
- Georgina Evans, Gary King. "Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset" *Political Analysis*

Appendix

Properties of Differential Privacy

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss
  - OK for worst case scenerio; unhelpful in practice

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss
  - OK for worst case scenerio; unhelpful in practice
- Privacy Budget

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss
  - OK for worst case scenerio; unhelpful in practice
- Privacy Budget
  - Composition: ϵ_1-DP and ϵ_2-DP is $(\epsilon_1 + \epsilon_2)$-DP

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss
  - OK for worst case scenerio; unhelpful in practice
- Privacy Budget
  - Composition: ϵ_1-DP and ϵ_2-DP is $(\epsilon_1 + \epsilon_2)$-DP
  - Can limit maximum risks across analyses & researchers

Properties of Differential Privacy

- Post-processing: if $M(s, D)$ is DP, so is $f[M(s, D)]$
  - Useful for bias corrections
- Privacy risk quantified (ϵ), instead of 0/1 for re-ID
  - Helpful mathematically; insufficient in applications
- Real privacy loss $\ll$ maximum privacy loss
  - OK for worst case scenerio; unhelpful in practice
- Privacy Budget
  - Composition: ϵ_1-DP and ϵ_2-DP is $(\epsilon_1 + \epsilon_2)$-DP
  - Can limit maximum risks across analyses & researchers
  - When the budget is used, no new analyses can ever be run