On Computational Thinking, Inferential Thinking and "Big Data"

Michael I. Jordan
University of California, Berkeley

November 16, 2015

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables
- Science in exploratory mode (e.g., astronomy, genomics)
  - *inferential issue*: massive number of hypotheses

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables
- Science in exploratory mode (e.g., astronomy, genomics)
  - *inferential issue*: massive number of hypotheses
- Measurement of human activity, particularly online activity, is generating massive datasets that can be used (e.g.) for personalization and for creating markets

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables
- Science in exploratory mode (e.g., astronomy, genomics)
  - *inferential issue*: massive number of hypotheses
- Measurement of human activity, particularly online activity, is generating massive datasets that can be used (e.g.) for personalization and for creating markets
  - *inferential issues*: many, including heterogeneity, unknown sampling frames, compound loss function

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables
- Science in exploratory mode (e.g., astronomy, genomics)
  - *inferential issue*: massive number of hypotheses
- Measurement of human activity, particularly online activity, is generating massive datasets that can be used (e.g.) for personalization and for creating markets
  - *inferential issues*: many, including heterogeneity, unknown sampling frames, compound loss function

- And then there are the computational issues

What Is the Big Data Phenomenon?

- Science in confirmatory mode (e.g., particle physics)
  - *inferential issue*: massive number of nuisance variables
- Science in exploratory mode (e.g., astronomy, genomics)
  - *inferential issue*: massive number of hypotheses
- Measurement of human activity, particularly online activity, is generating massive datasets that can be used (e.g.) for personalization and for creating markets
  - *inferential issues*: many, including heterogeneity, unknown sampling frames, compound loss function

- And then there are the computational issues
  - and, most notably, the interactions of computational and inferential issues

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"

A Job Description, circa 2015

- Your Boss: “*I need a Big Data system that will replace our classic service with a personalized service*”
- “*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*”

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"
- "*It should only improve as we collect more data; in particular it shouldn't slow down*"

A Job Description, circa 2015

- Your Boss: "*I need a Big Data system that will replace our classic service with a personalized service*"
- "*It should work reasonably well for anyone and everyone; I can tolerate a few errors but not too many dumb ones that will embarrass us*"
- "*It should run just as fast as our classic service*"
- "*It should only improve as we collect more data; in particular it shouldn't slow down*"
- "*There are serious privacy concerns of course, and they vary across the clients*"

Some Challenges Driven by Big Data

- Big Data analysis requires a thorough blending of computational thinking and inferential thinking

Some Challenges Driven by Big Data

- Big Data analysis requires a thorough blending of computational thinking and inferential thinking
- What I mean by computational thinking
  - abstraction, modularity, scalability, robustness, etc.

Some Challenges Driven by Big Data

- Big Data analysis requires a thorough blending of computational thinking and inferential thinking
- What I mean by computational thinking
  - abstraction, modularity, scalability, robustness, etc.
- Inferential thinking means (1) considering the real-world phenomenon behind the data, (2) considering the sampling pattern that gave rise to the data, and (3) developing procedures that will go “backwards” from the data to the underlying phenomenon

Some Challenges Driven by Big Data

- Big Data analysis requires a thorough blending of computational thinking and inferential thinking
- What I mean by computational thinking
  - abstraction, modularity, scalability, robustness, etc.
- Inferential thinking means (1) considering the real-world phenomenon behind the data, (2) considering the sampling pattern that gave rise to the data, and (3) developing procedures that will go “backwards” from the data to the underlying phenomenon
  - merely computing “statistics” or running machine-learning algorithms generally isn’t inferential thinking
  - a focus on confidence intervals---not just “outputs”

The Challenges are Daunting

- The core theories in computer science and statistics developed separately and there is an oil and water problem
- Core statistical theory doesn't have a place for runtime and other computational resources
- Core computational theory doesn't have a place for statistical risk

Outline

- Inference under privacy constraints
- Inference under communication constraints
- Inference (confidence intervals) and parallel, distributed computing

Part I: Inference and Privacy

with John Duchi and Martin Wainwright

Privacy and Data Analysis

- Individuals are not generally willing to allow their personal data to be used without control on how it will be used and now much privacy loss they will incur
- "Privacy loss" can be quantified via differential privacy
- We want to trade privacy loss against the value we obtain from "data analysis"
- The question becomes that of quantifying such value and juxtaposing it with privacy loss

Privacy

query

database

Privacy

query

$\downarrow$

database

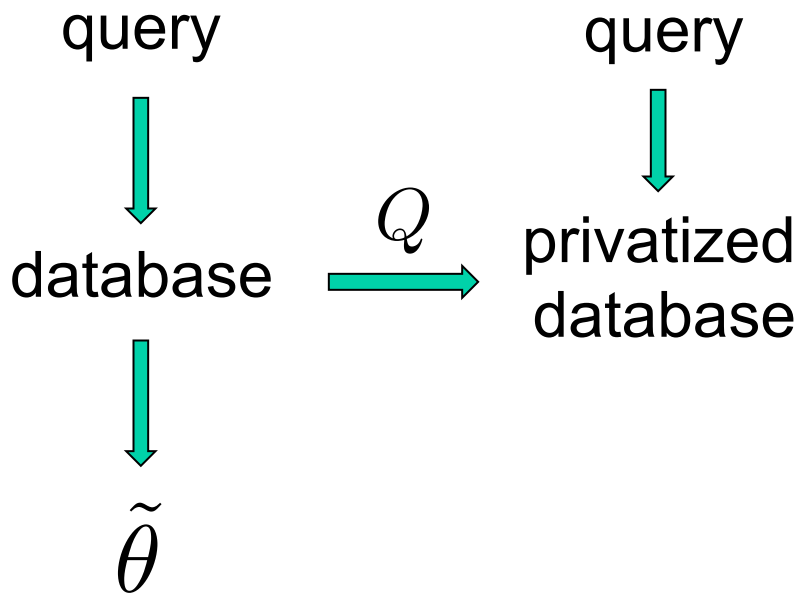
$\downarrow$

$$\tilde{\theta}$$

Privacy

$$\begin{tikzcd} \text{query} \arrow[d] & \\ \text{database} \arrow[r, "Q"] \arrow[d] & \text{privatized database} \\ \tilde{\theta} & \end{tikzcd}$$

Privacy



Privacy

$$
\begin{tikzcd}
\text{query} \arrow[d] & & \text{query} \arrow[d] \\
\text{database} \arrow[d] \arrow[rr, "Q"] & & \text{privatized database} \arrow[d] \\
\tilde{\theta} & & \hat{\theta}
\end{tikzcd}
$$

Privacy

$$
\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] \\
\text{database} \arrow[r, "Q"] \arrow[d] & \text{privatized database} \arrow[d] \\
\tilde{\theta} & \hat{\theta}
\end{tikzcd}
$$

Classical problem in differential privacy: show that $\hat{\theta}$ and $\tilde{\theta}$ are close under constraints on Q

Inference

query

$\downarrow$

database

$\downarrow$

$$\tilde{\theta}$$

Inference

$$
\begin{tikzcd}
 & \text{query} \arrow[d] \\
P \arrow[r, "S"] & \text{database} \arrow[d] \\
 & \tilde{\theta}
\end{tikzcd}
$$

Inference

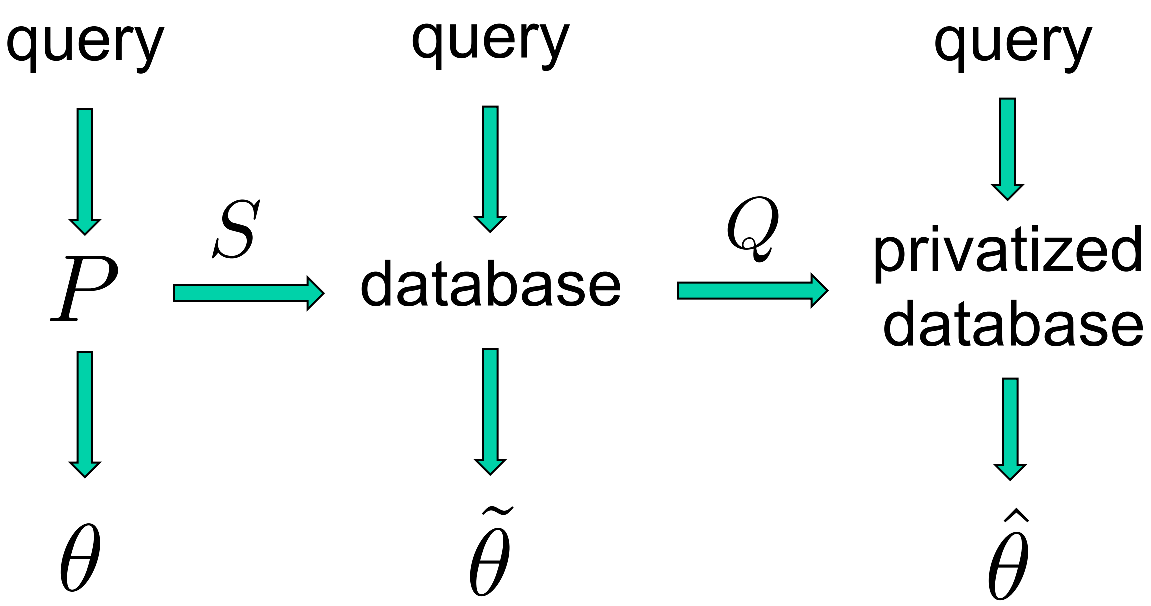
$$
\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] \\
P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[d] \\
\theta & \tilde{\theta}
\end{tikzcd}
$$

Inference

$$
\begin{tikzcd}
\text{query} \arrow[d] & \text{query} \arrow[d] \\
P \arrow[r, "S"] \arrow[d] & \text{database} \arrow[d] \\
\theta & \tilde{\theta}
\end{tikzcd}
$$

Classical problem in statistical theory: show that $\tilde{\theta}$ and θ are close under constraints on S

Privacy and Inference



The privacy-meets-inference problem: show that θ and $\hat{\theta}$ are close under constraints on Q and on S