September 30, 2021

Generating Synthetic Data for Data Privacy in R

Claire McKay Bowen, Ph.D
Lead Data Scientist, Privacy and Data Security
cbowen@urban.org www.clairemckaybowen.com @ClaireMKBowen

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Overview

1. **Motivation:** What is data privacy? Why should we care?
2. **Background:** What has been done to address these issues?
3. **Methodology:** What is data synthesis?
4. **R Code:** How do you generate synthetic data?

Motivation

BRIEFING ROOM

Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government

JANUARY 20, 2021 • PRESIDENTIAL ACTIONS

"...advanc[e] equity for all, including people of color and others who have been historically underserved, marginalized, and adversely affected by persistent poverty and inequality."


ONE NATION, TRACKED
AN INVESTIGATION INTO THE SMARTPHONE TRACKING INDUSTRY FROM TIMES OPINION

Lack of Public Data Hampers COVID-19 Fight

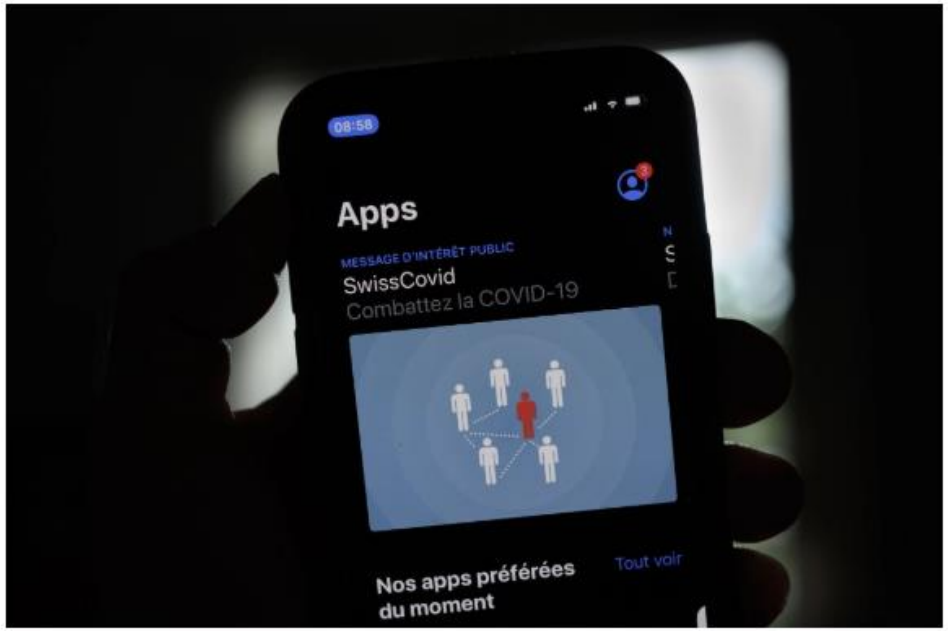
STATELINE ARTICLE August 3, 2020 By: Christine Vestal Topics: Health Read time: 7 min

A ventilator helps a COVID-19 patient breathe at a Houston hospital. Hospital data related to the coronavirus pandemic will now be collected by a private technology firm, rather than the Centers for Disease Control and Prevention. Epidemiologists say better COVID-19 data is needed to improve the nation's response.

David J. Phillip/The Associated Press

Google Promises Privacy With Virus App but Can Still Collect Location Data

Some government agencies that use the software said they were surprised that Google may pick up the locations of certain app users. Others said they had unsuccessfully pushed Google to make a change.



Switzerland has asked Google to decouple the location setting requirement on Android phones from Bluetooth, which the country's virus alert app uses to detect nearby smartphones. Fabrice Coffrini/Agence France-Presse — Getty Images

By Natasha Singer

July 20, 2020

HEALTH

'It's not a pretty picture': Why the lack of racial data around COVID vaccines is 'massive barrier' to better distribution

Nada Hassanein USA TODAY

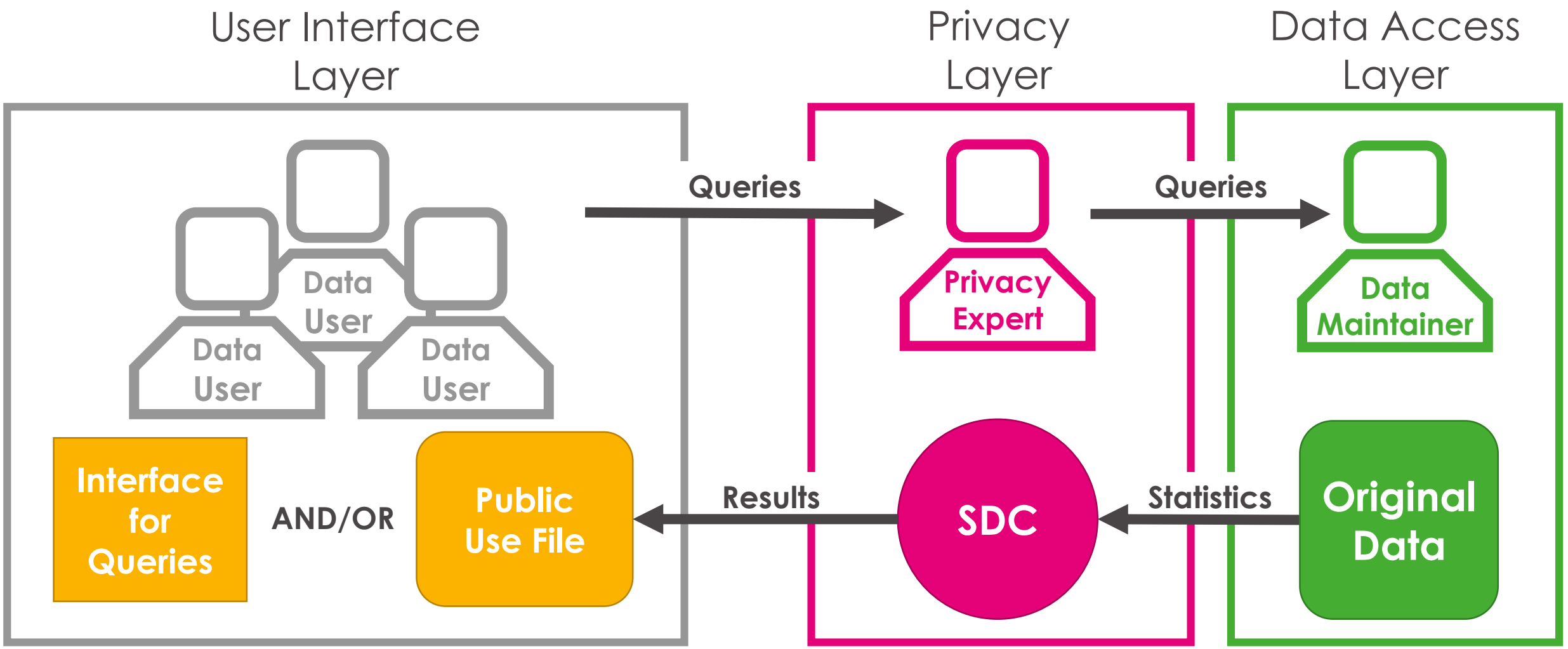
Published 5:30 a.m. ET Feb. 1, 2021 | **Updated 2:10 p.m. ET Feb. 1, 2021**

Abigail Echo-Hawk, chief research officer with Seattle Indian Health Board and a member of the Pawnee Tribe, gets a shot of the Moderna COVID-19 vaccine on Dec. 21. A colleague used a black pen to inscribe "For the (Heart) love of Native People" over the injection spot. *Karen Ducey, Getty Images*


WE NEED RELIEF
NOW!
ED PAY

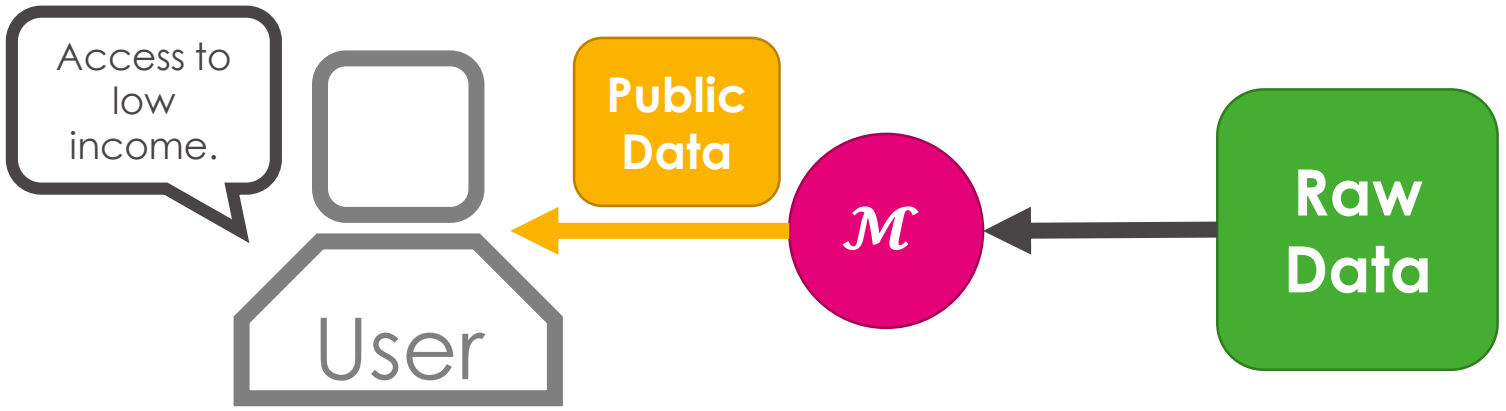
Background

How do we access confidential data?

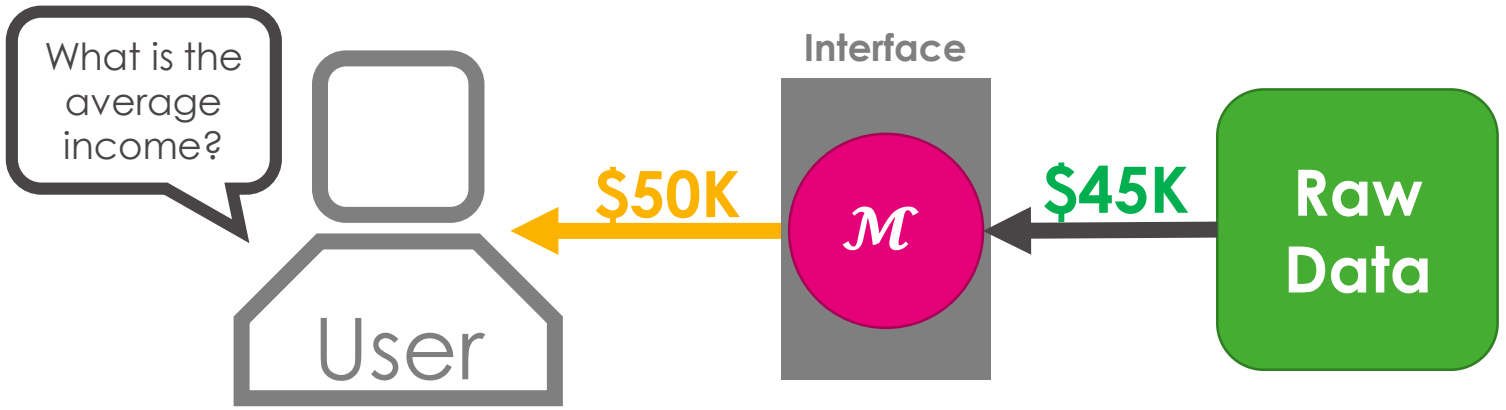


SDC / SDL = Statistical Disclosure Control or Limitation

Why is it hard to implementing this framework?



How do we **measure utility** (usefulness) and **disclosure risk** of the data?



How **much noise should be added** and how do you **limit the number of queries**?

Methodology

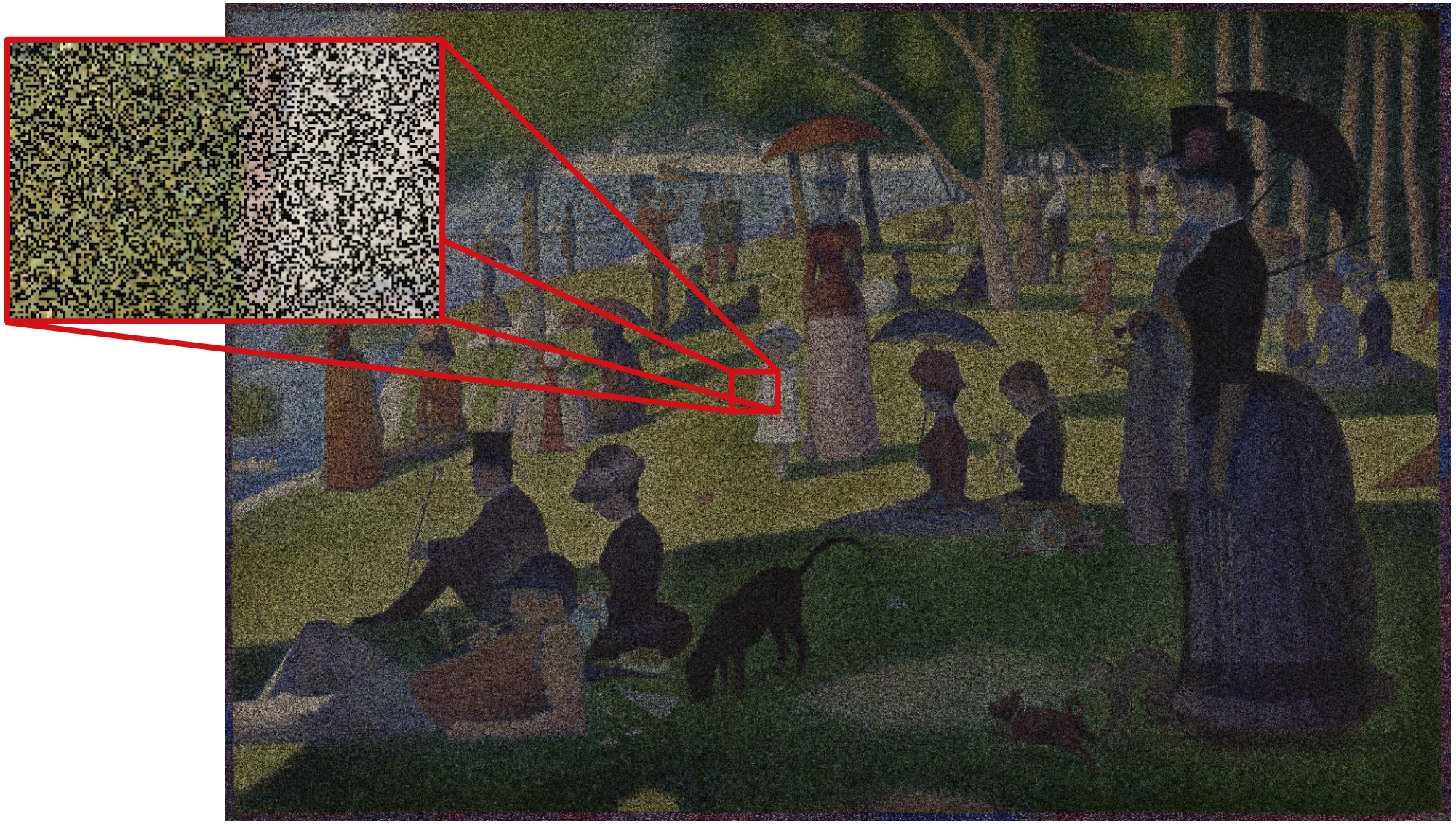
What are ways to protect data?

What are ways to protect data? Suppression

What are ways to protect data? Generalization

What are ways to protect data? Sampling

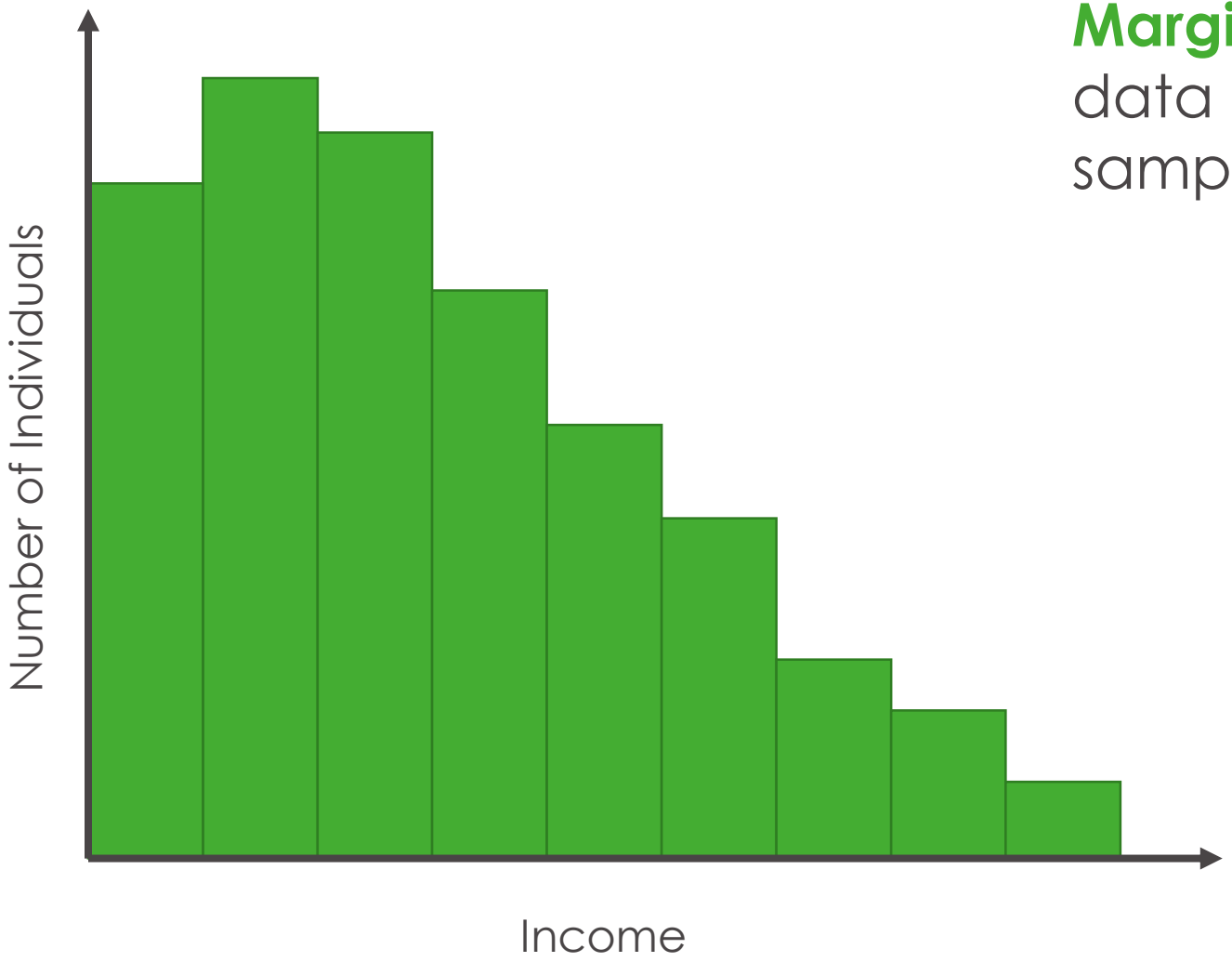
What are ways to protect data? Sampling



What are ways to protect data? Noise Infusion

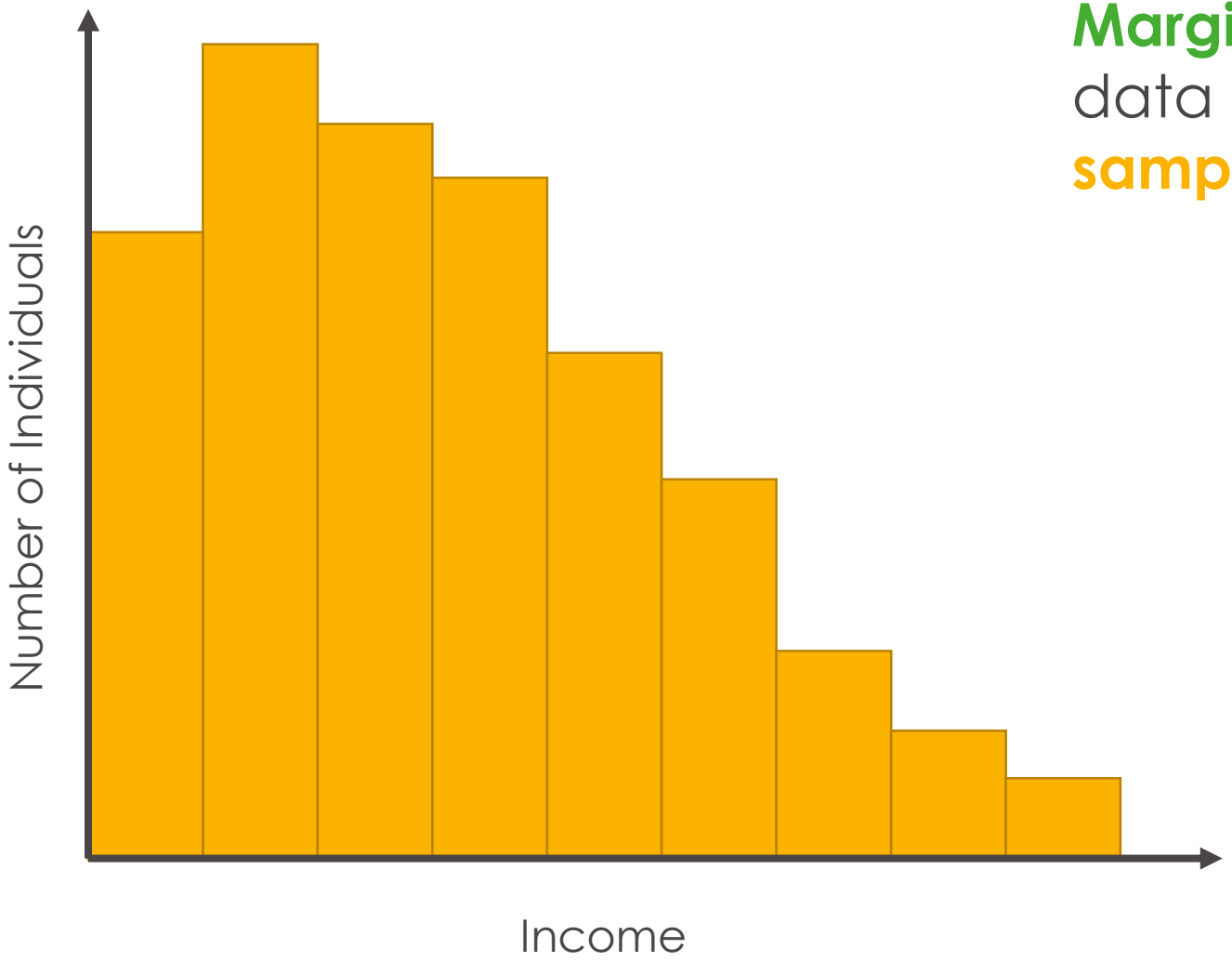
What are ways to protect data? Synthetic Data

What is non-parametric data synthesis?



Marginal Tables – a basic synthetic data approach by using random sampling

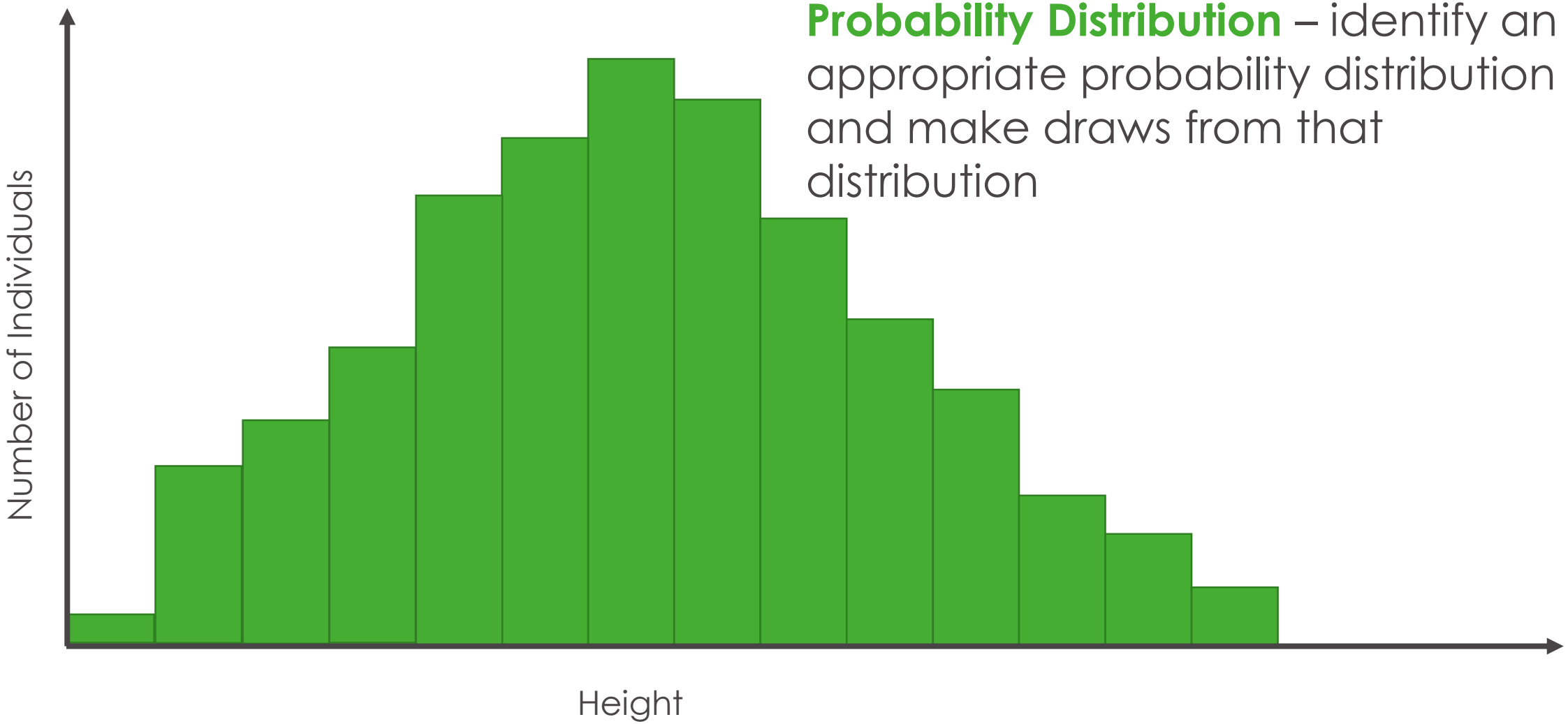
What is non-parametric data synthesis?



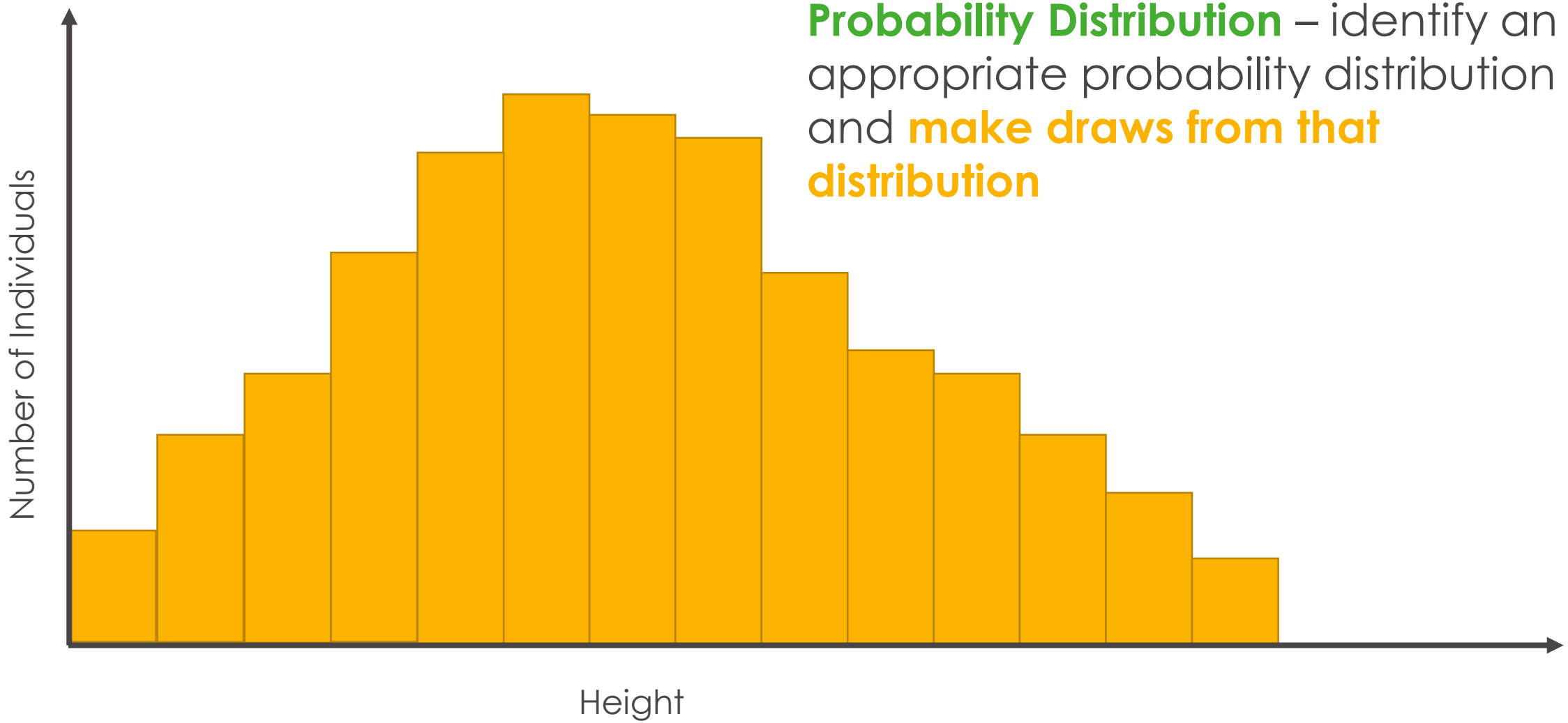
Marginal Tables – a basic synthetic data approach by **using random sampling**

What is parametric data synthesis?

Probability Distribution – identify an appropriate probability distribution and make draws from that distribution



What is parametric data synthesis?



Probability Distribution – identify an appropriate probability distribution and **make draws from that distribution**

Summary

- Motivation
- Background
- Synthetic Data
  - Non-parametric
  - Parametric
- R Coding

Contact Me


cbowen@urban.org


www.clairemckaybowen.com


/in/bowenclaire


@ClaireMKBowen

WE ARE HIRING A DATA SCIENTIST!

WE ARE HIRING A DATA SCIENTIST!

The successful candidate will have:

- A Masters or a Bachelors in Computer Science, Mathematics, Statistics, Economics, Public Policy, or a related field.
- At least 2 years of programming experience in a professional or academic environment with either Python or R.
- Ability to apply and implement research principals and goals into code.
- Willingness to learn and adapt to changes in work assignments, deadlines, and team environment.
- Strong organizational skills and the ability to manage competing deadlines.
- Ability to communicate clearly with both technical and non-technical audiences.
- Interest in domestic economic and social policy issues.