Applying Modern Crypto to Advance Social Science

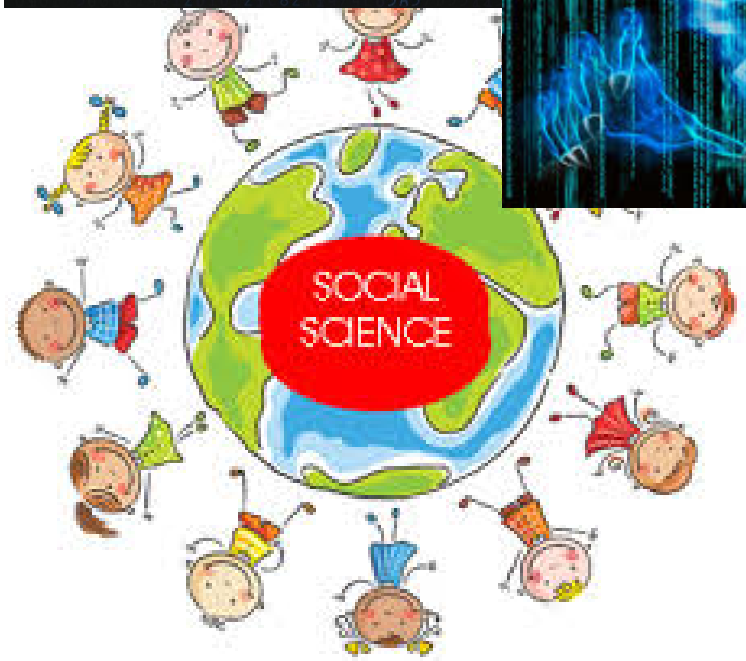
Ran Cohen (MIT & Northeastern)
Yarkin Doroz (NJIT)
Shafi Goldwasser (MIT)
Jason Owen-Smith (IRIS/University of Michigan)
Kurt Rohloff (NJIT)

Social Science

Relationships among individuals within a societv

Social Science + PRIVACY

Relationships among individuals within a societv



Massachusetts
Institute of
Technology

Alfred P. Sloan
FOUNDATION

INSTITUTE FOR
RESEARCH ON
INNOVATION & SCIENCE

IRIS

- Institute for Research on Innovation and Science
- Consortium of major research universities
- Founded in 2015
- Located at University of Michigan

Understand, explain, and improve
the public value of academic research

Funded Academic Research

- 2015 in US: **$214** were invested in academic research for every man, woman, and child
- **Goal:** develop human knowledge and improve quality of life and well being
- How well do we understand, explain, and improve those effects?
- IRIS provides **high quality, trusted, secure data** to answer this questions

IRIS Data

- IRIS collects from the member universities:
  - Financial data
  - Personal administrative data
  - Scientific investments
  - Vendors
  - Federal awards and subawards transactions
- Process the data
  - Fill missing records, ensure consistent terms, add crosswalks
- Produce secondary de-identified data
  - Remove Personally Identifiable Information (PII)
  - Personal name, date of birth, university name…
- Partners with US Census
  - Statistics on people
  - Employment records
  - Vendor characteristics

IRIS First Data Release

- 19 universities
  - $11B in 2014 federal R&D (16% of total)
- Transaction level data
  - 162,694 federal and non-federal sponsored projects
  - 333,565 individuals
    - 28,641 Post-Docs
    - 76,295 Grad Students
    - 87,195 Undergrads
  - $18.1B in vendor spending to ~81,000 establishments
  - $6B in subcontracts to other performers
- Links to abstracts etc for federal awards (NIH, NSF, USDA)
- Individual level links to dissertation information
- Title 13 crosswalks to LEHD, LBD, ACS, Decennial Census (available only through the FSRDC system)

Research on IRIS Data

Measuring the Economic Value of Research: The Case of Food Safety

Kaye Husbands Fealing, Julia I. Lane, John L. King, and Stanley R. Johnson editors

Cambridge University Press, December 2017

Why the US science and engineering workforce is aging rapidly

David Blau and Bruce A. Weinberg

Proceedings of the National Academy of Sciences

Early Edition: approved February 14, 2017

STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census

Catherine Buffington, Benjamin Cerf, Christina Jones, and Bruce A. Weinberg

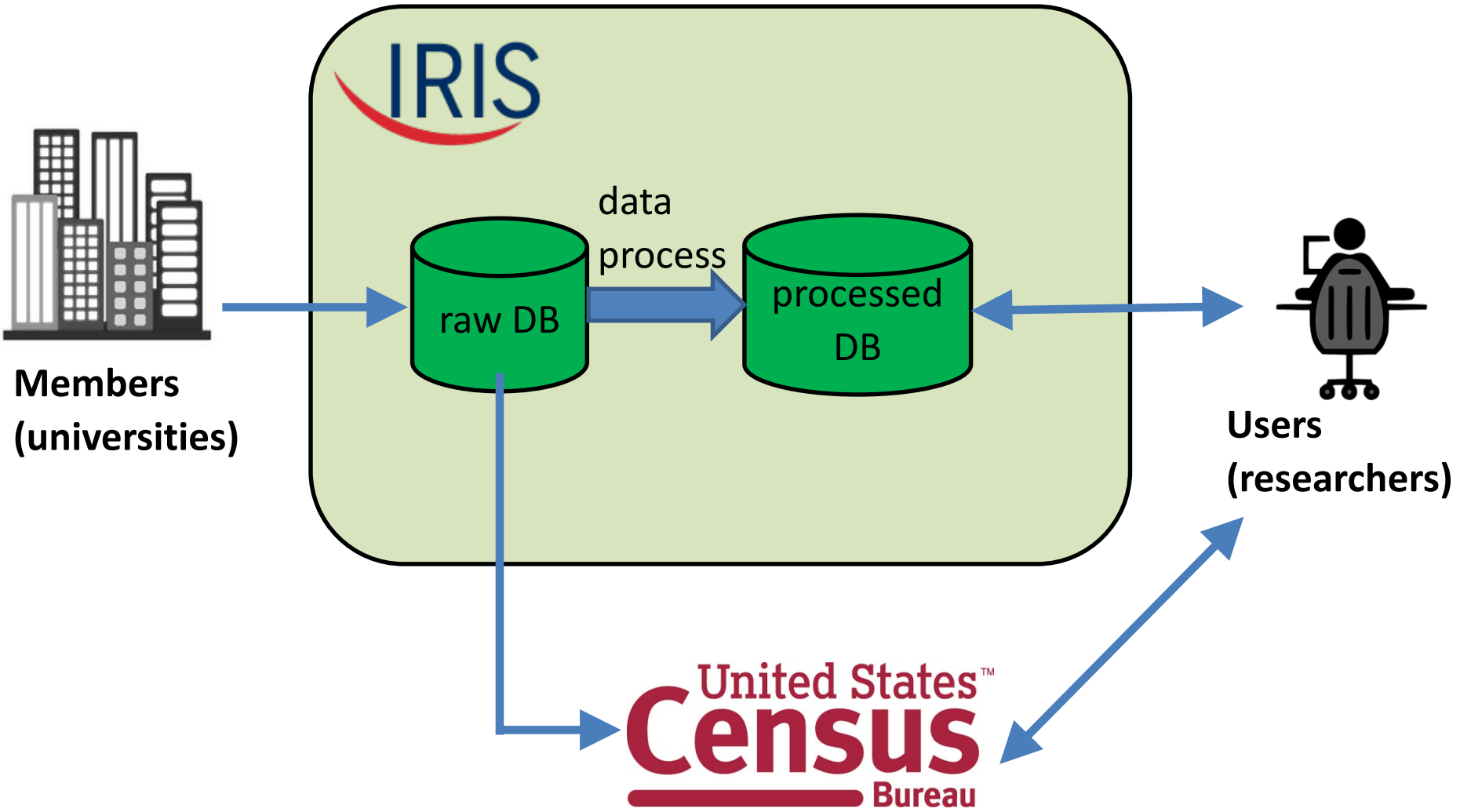
American Economic Review May 2016

Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients

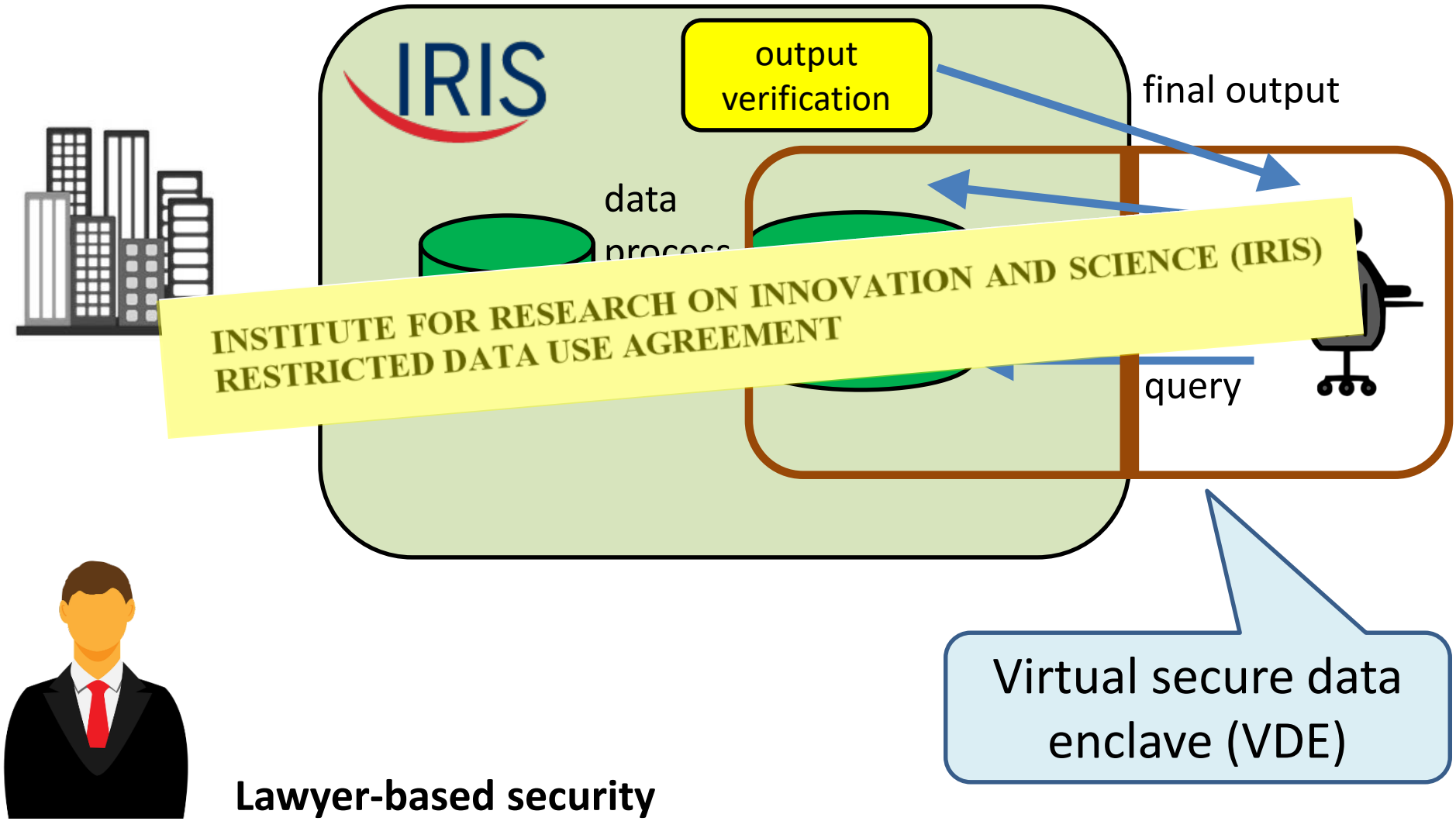
Nikolas Zolas, Nathan Goldschlag, Ron Jarmin, Paula Stephan, Jason Owen- Smith, Rebecca F. Rosen, Barbara McFadden Allen, Bruce A. Weinberg, Julia I. Lane

***Science* 11 December 2015**

The System Today



Accessing IRIS Data



IRIS Data Use Agreement

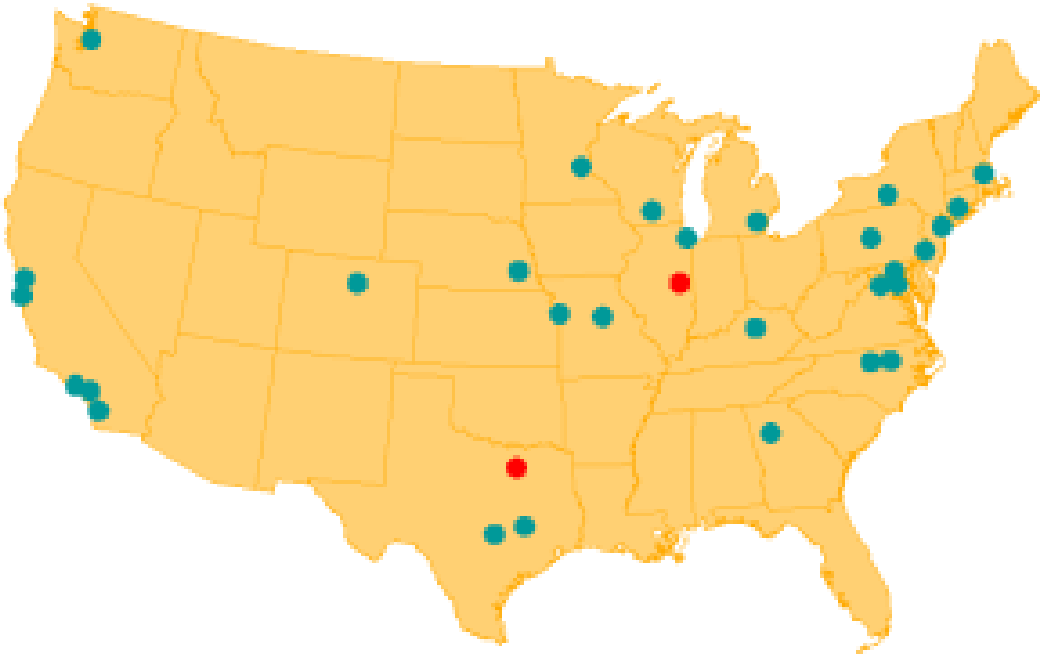
4. User agrees to limit their work with the Data as follows:
   a. Not to use or further disclose the Data or any information contained therein other than as permitted by this Agreement or required by applicable law.
   b. Not to attempt to extract, copy, or otherwise remove the Data or any part of it from the Enclave.
   c. To report to Michigan any use or disclosure of the Data or any part of it not authorized by this Agreement of which User or any Authorized Party becomes aware.

1

   d. To ensure that any Authorized Parties understand and agree to the same restrictions and conditions that apply to the User under this Agreement.
   e. Not to use the information contained in the Data to identify the individuals whose information is contained in the Data, nor to contact them under any circumstances.
   f. Not to take screen shots or other video or image grabs of any displayed data.
   g. That if the identity of any person or institution should be discovered inadvertently; (i) no use will be made of this information, (ii) Michigan will be advised of the incident within one (1) business day of User's discovery of the incident, (iii) the information

Lawyer-based security

Accessing Census Data



Census data is protected by

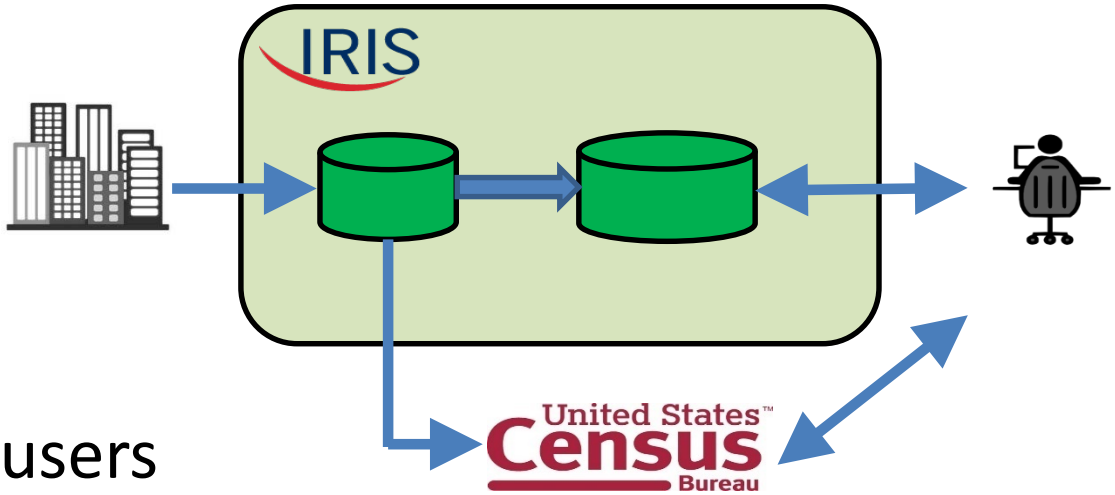
Title 13

29 FSRDC (Federal Statistical Research Data Centers)

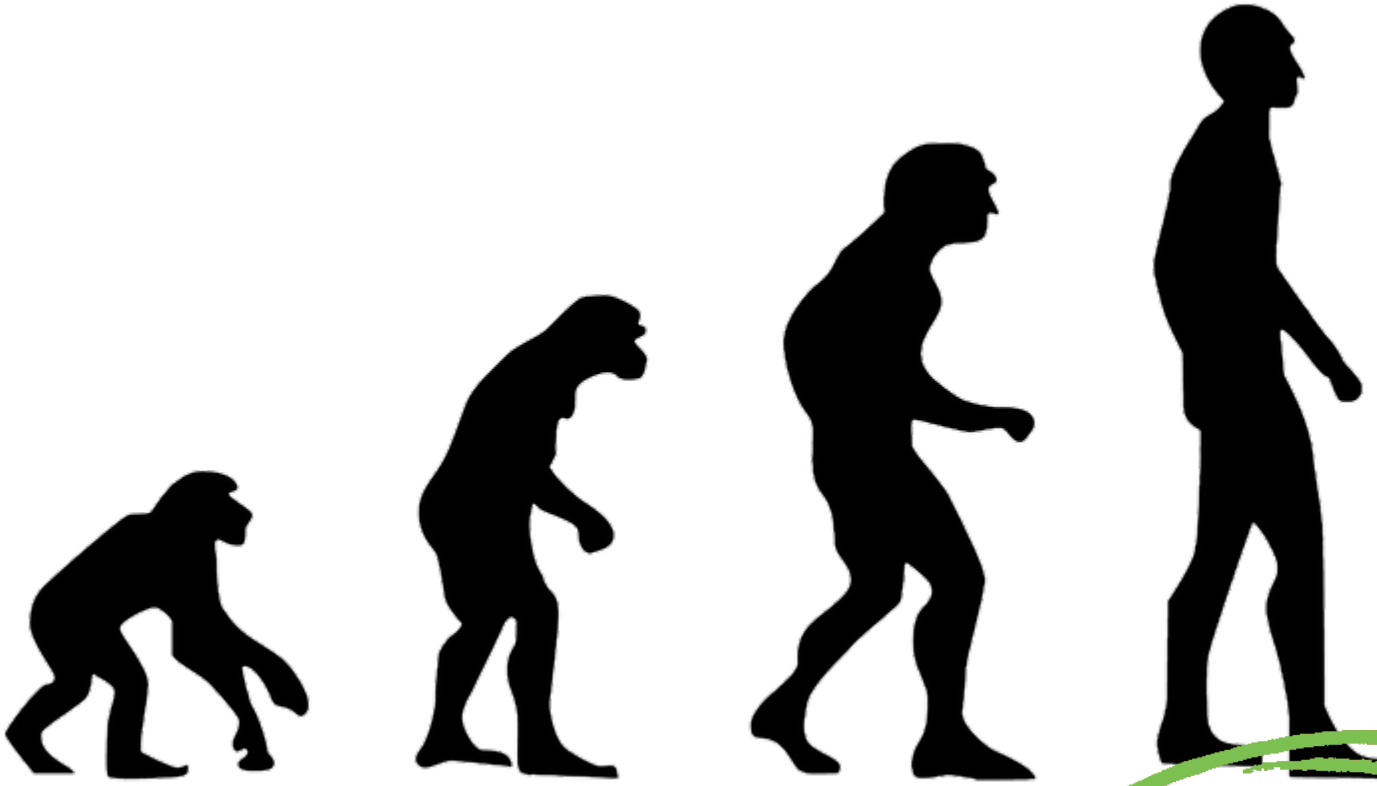
United States™ Census Bureau

The System Today

- **Pros:**
  - It works!
- **Cons:**
  - Data is visible to users
  - A lot of bureaucracy
  - Manual verification of the output
  - Data visible to IRIS



Evolution of Modern Crypto



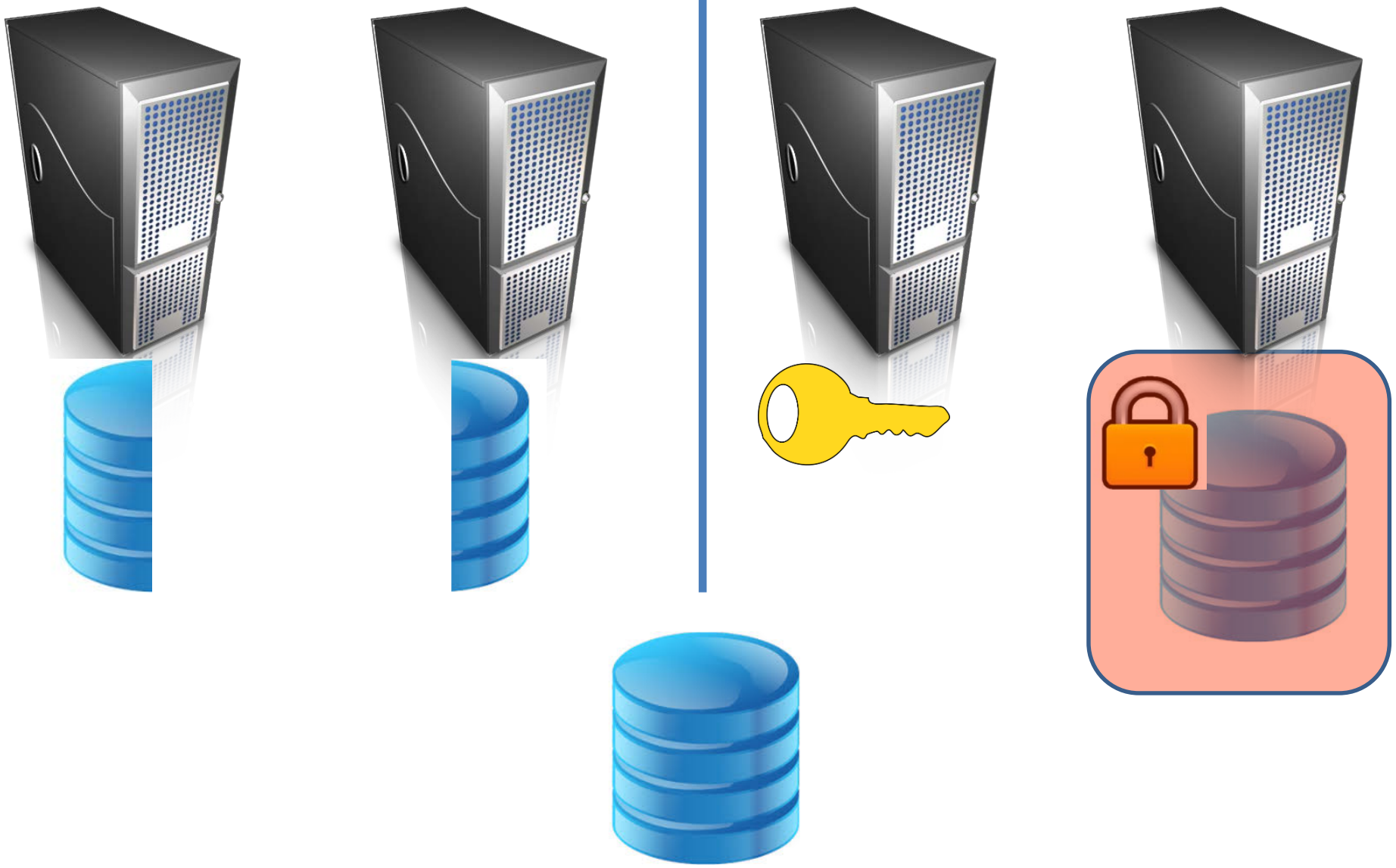
Feasibility

Asymptotic efficiency

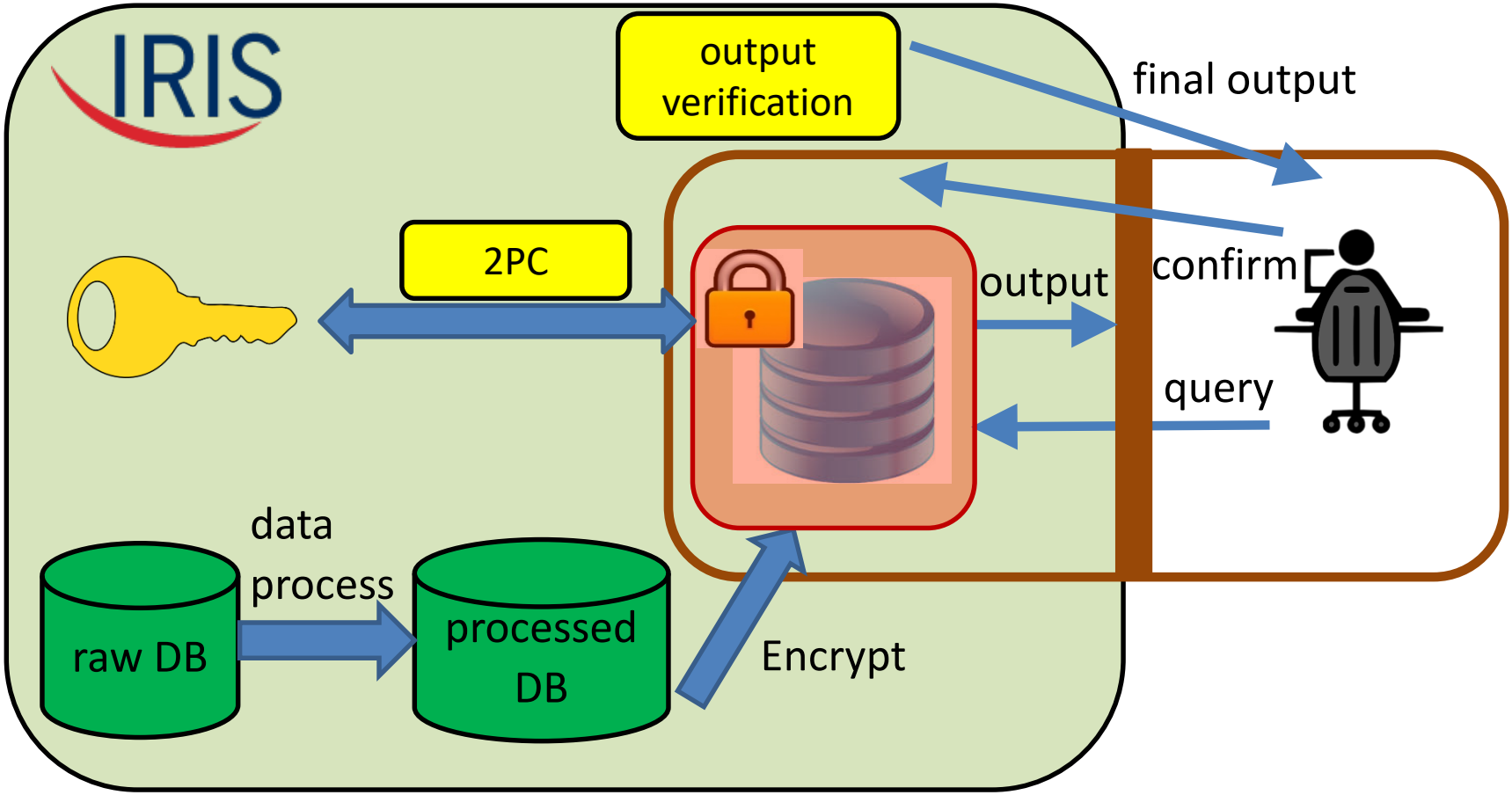
Concrete efficiency

Proof of concept

MPC/FHE

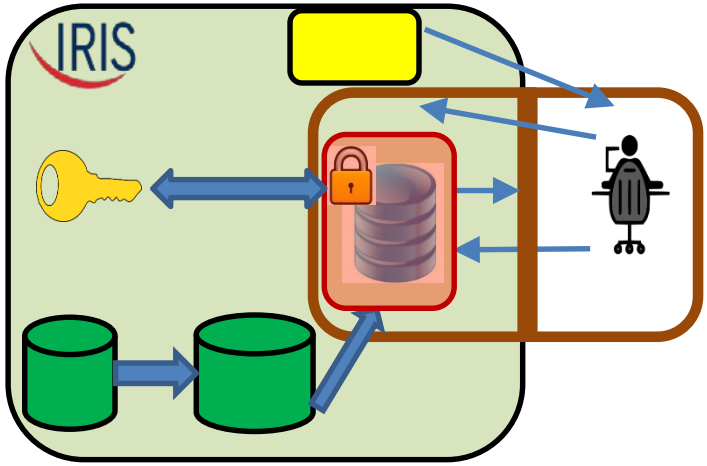


Phase 1 – Encrypted DB



Phase 1 – Encrypted DB

- **Pros:**
  - Data is hidden from users
  - Less bureaucracy
  - Control over the leakage

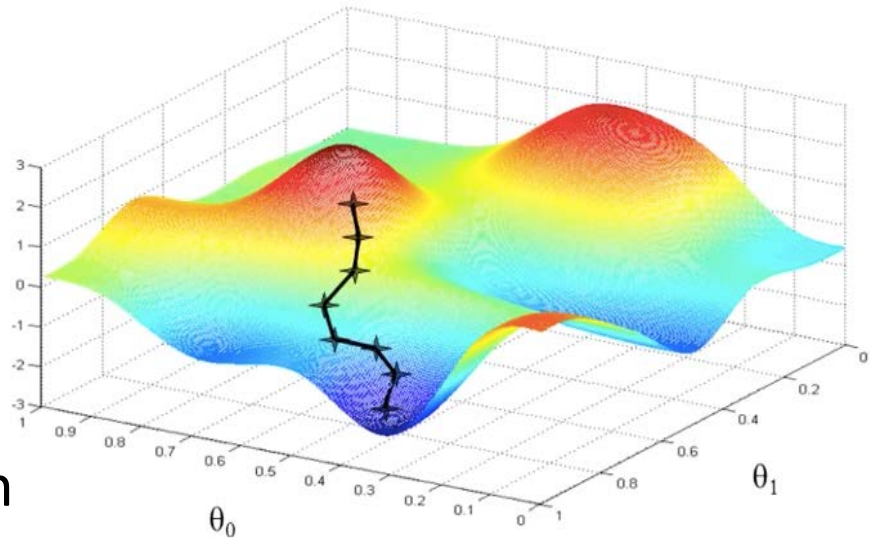


- **Cons:**
  - Increased runtime
  - Manual verification of the output
  - Data visible to IRIS

Proof of Concept

Research question: explain the size, composition, and level of student participation of grant funded research teams

- Regression analysis
  - OLS regression
  - Logistic regression
  - Poisson regression
  - Probit regression
  - Negative binomial regression



- Gradient descent
  - Iterative optimization algorithm to find local miminum
- Implemented using the **PALISADE FHE library**

Concrete Queries

1) Percentage of students funded on grant
   Computed using **OLS regression**

2) Student participation (at least one student paid on the grant)
   Computed using **Logistic regression**

3) Project size (number of total people employed by a grant)
   Computed using **Poisson regression**

Runtime

Two test databases

- 10,000 records, 11 properties
  - ~23 seconds per iteration

	10,000	
	# iterations	Total time (sec)
OLS	2	46
Logistic	8	185
Poisson	11	250

Runtime

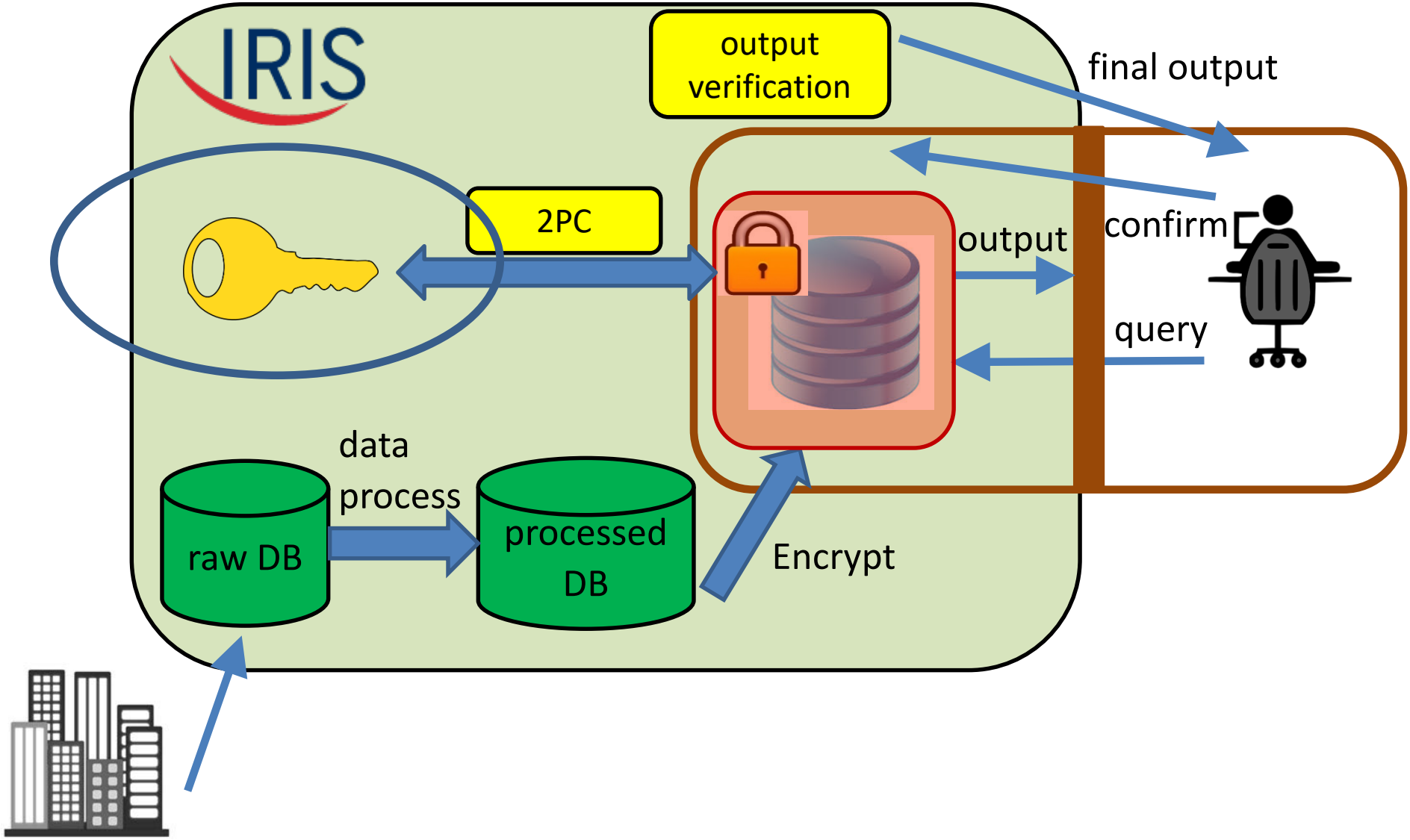
Two test databases

- 10,000 records, 11 properties
  - ~23 seconds per iteration
- 50,000 records, 14 properties
  - ~80 seconds per iteration
- 7 digits of precision

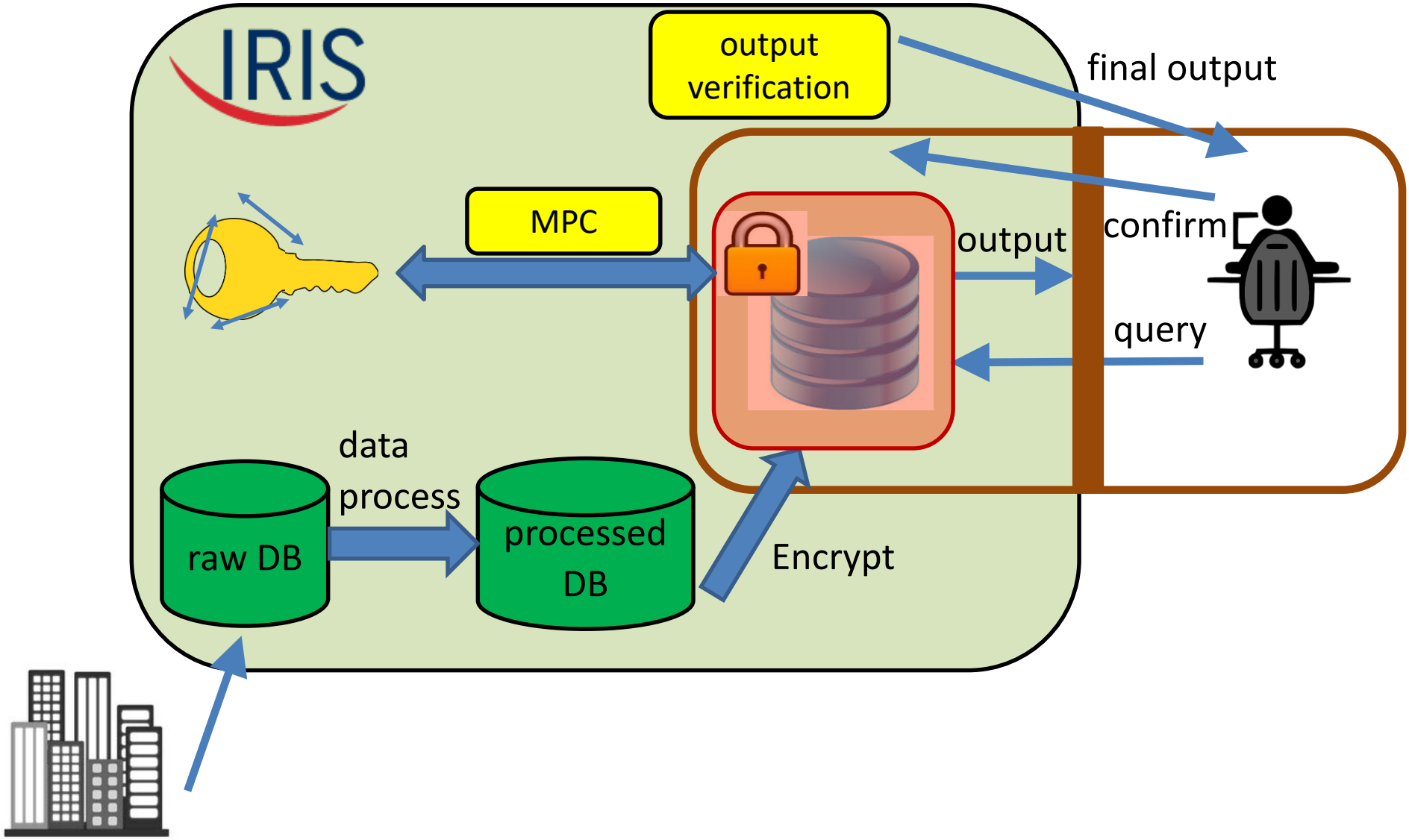
	10,000		50,000	
	# iterations	Total time (sec)	# iterations	Total time (sec)
OLS	2	46	2	160
Logistic	8	185	9	720
Poisson	11	250	13	1040


WHAT'S NEXT

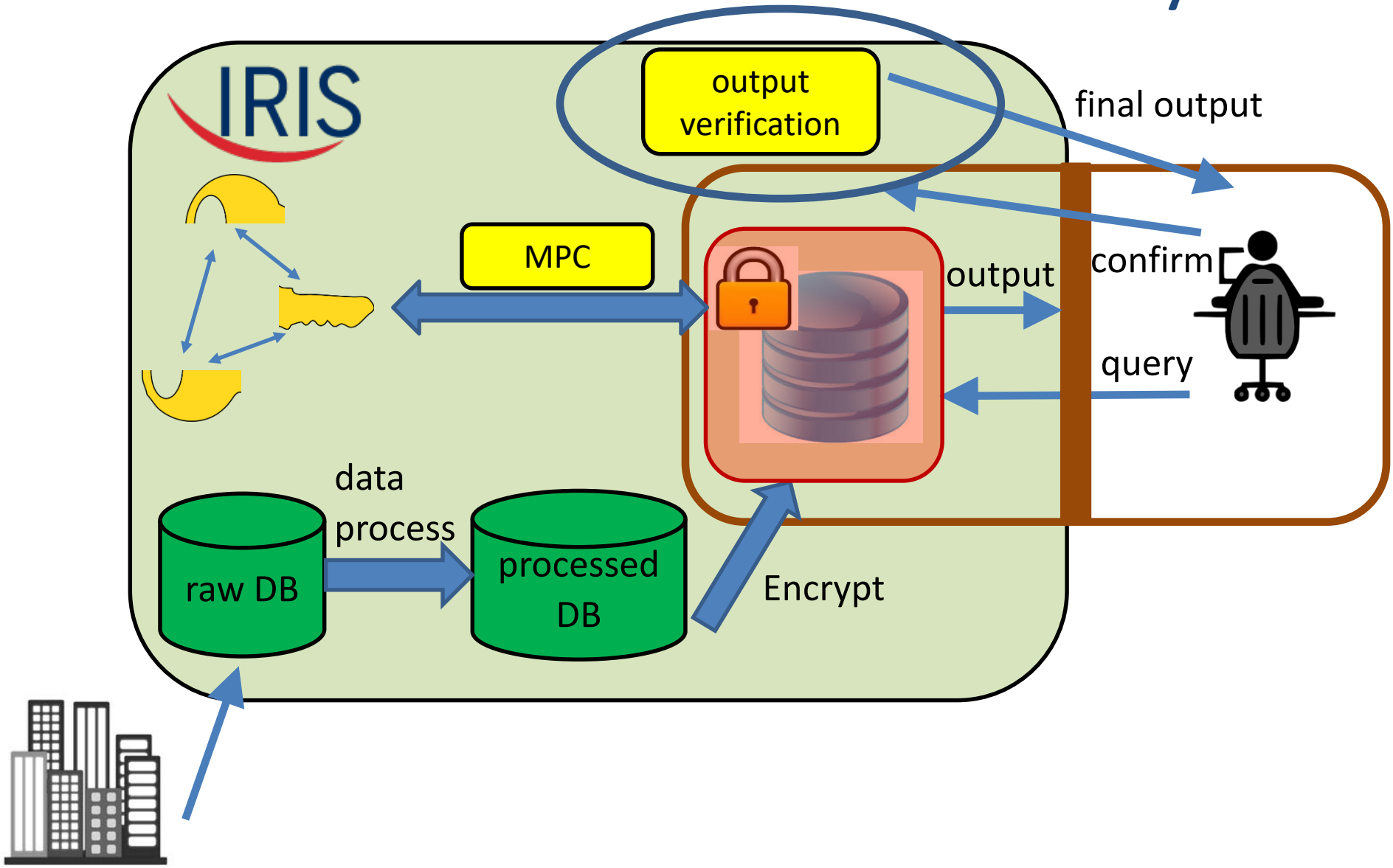
Phase 2 – Distributed Key



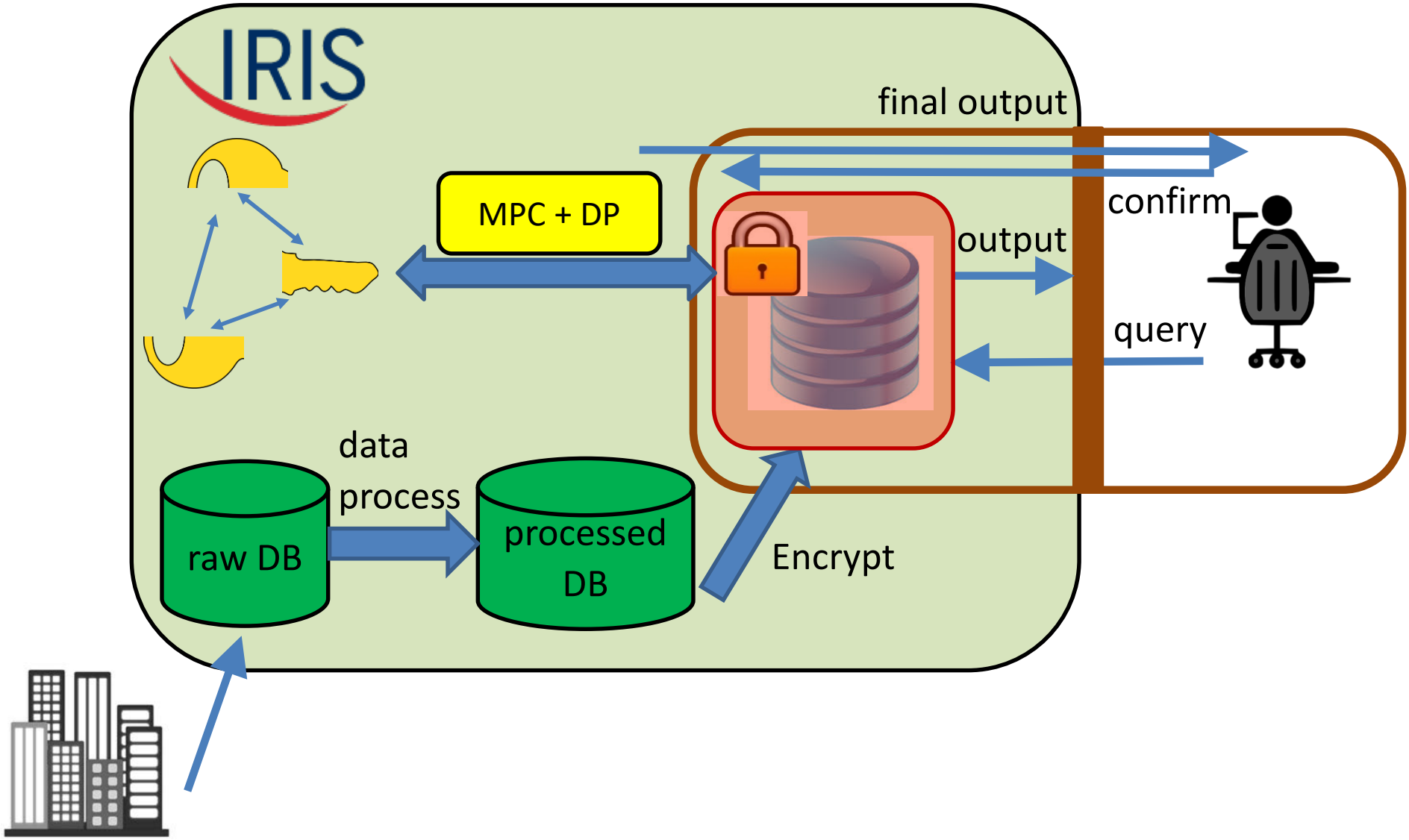
Phase 2 – Distributed Key



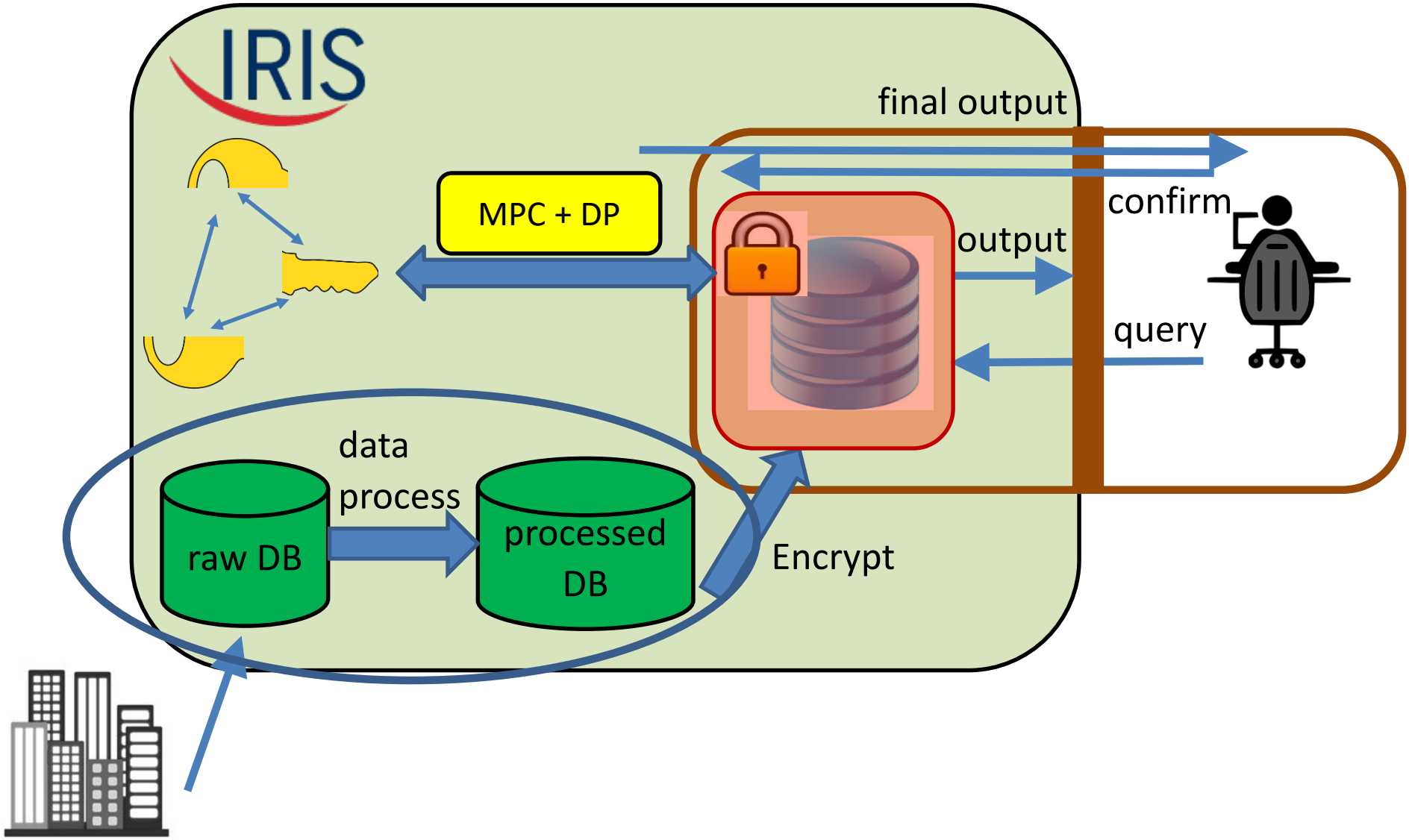
Phase 3 – Differential Privacy



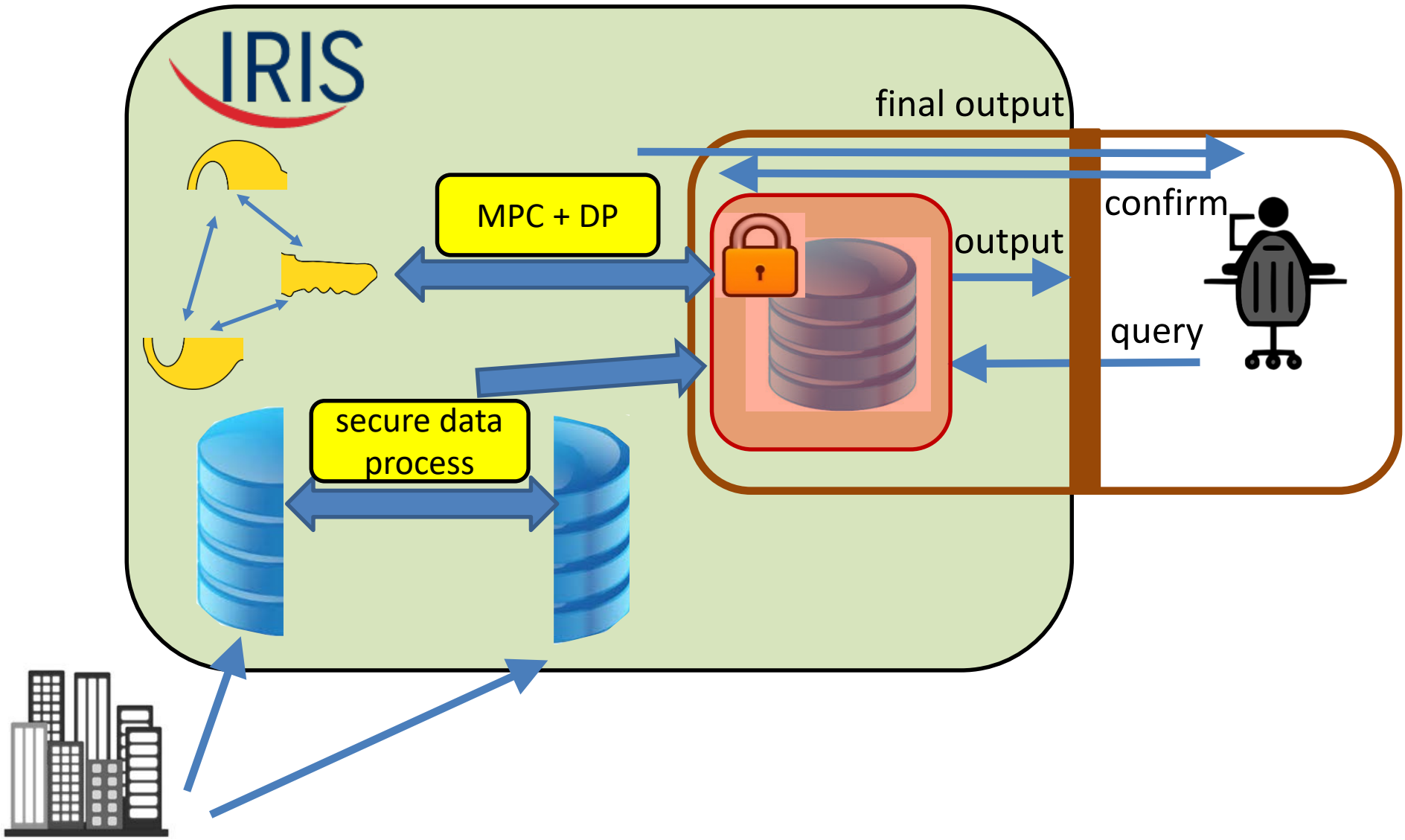
Phase 3 – Differential Privacy



Phase 4 – Secure Data Process



Phase 4 – Secure Data Process



Summary

- Protect IRIS data using FHE & MPC
- Better privacy – more data contributors
- Less bureaucracy – more researchers
- Conduct real social science research

Thank You