Data Ledger

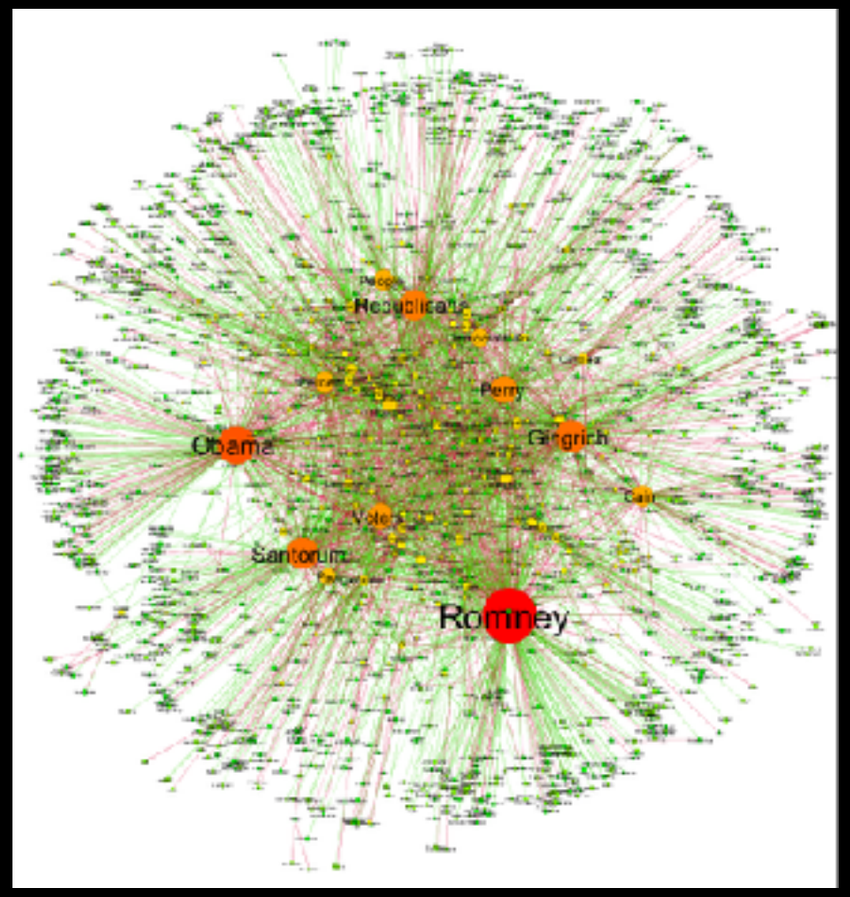
Data Privacy and Auditable Data Sharing

2019/10 Hackathon

A Common Big Data Issue

- Today, the scientific research in disciplines such as **social science, macroeconomy, medical science, generic engineering**, etc. are using the latest technologies such as big data and machine learning a lot.

- However, the issue of **data privacy and sensitivity** is inhibiting the research a lot.



Narrative network of US Elections 2012, Computational Sociology, Wikipedia

Privacy

Data Privacy in Research

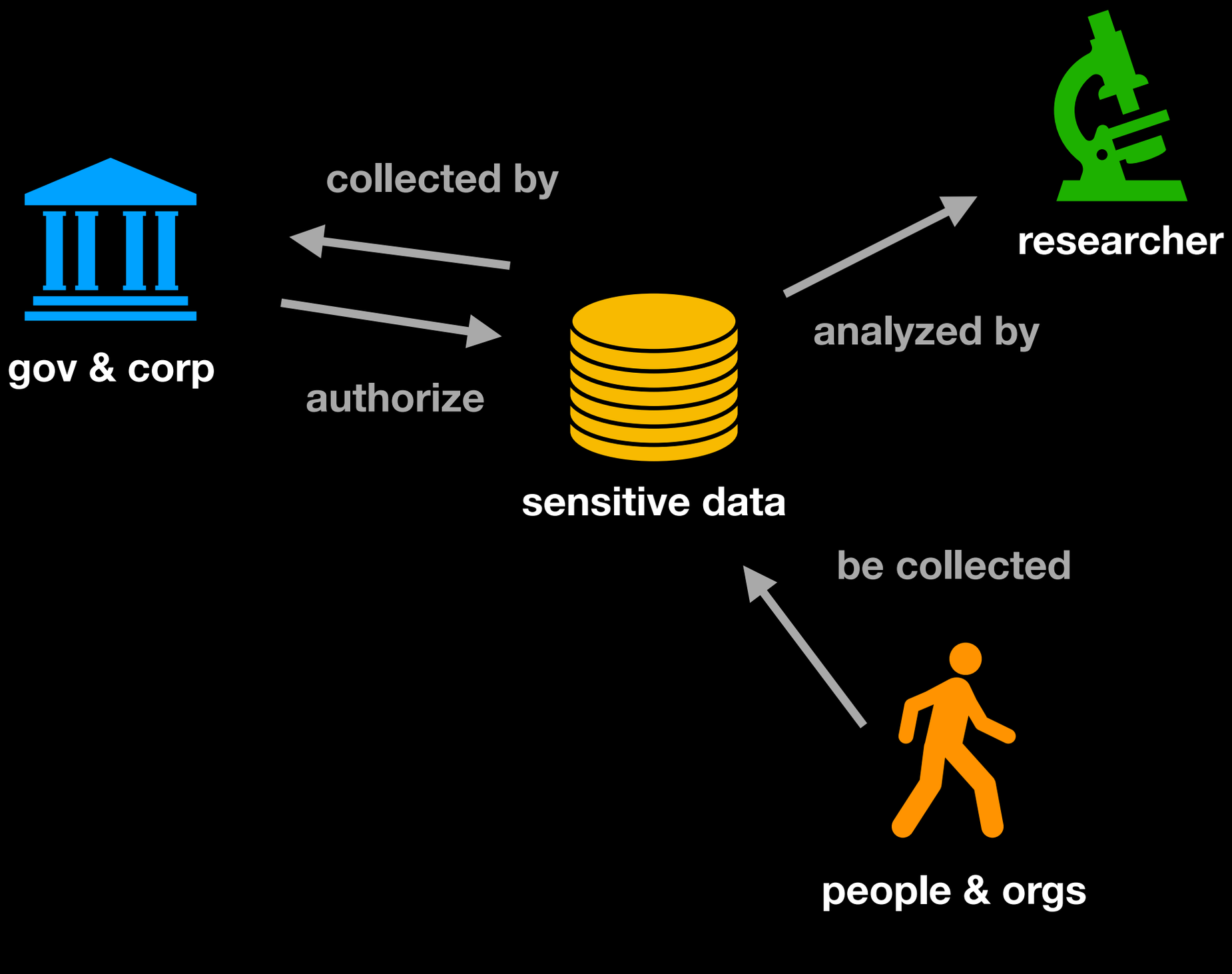
- The current **big data owners** such as governments and corporations **DON'T want to share the data** to researcher, unless in some extremely restricted computer environment.

- The researchers may unintentionally **LEAK the sensitive and confidential datasets** to unauthorized people (or got hacked), and hurt the society as a result. This is extremely risky.

- **Data sharing across institutes** is difficult among researchers due to the privacy concern.

- The true data owners such as people and enterprises are **NOT aware** that their data are used in research, and usually have **NO rights and opportunities to say NO**.

Data Privacy in Research



Investigation Process

- We surveyed the data work issues of more than 100 scientists from dozens of universities and institutes (e.g. UC Berkeley, Fudan University, etc.) mainly in East Asia and US from the year of 2017, by face-to-face and phone interview.

- Privacy issue is a challenging issue for many researchers in **social science** especially, and common in fields that involves human subjects.

- A good amount of valuable datasets are only available to researchers after quite time-consuming application process, and can only be accessed in a specified offline computer, to avoid privacy and security risks.

Example

- Researcher: Professor Zhang, Fudan University
- Topic: Government Finance Policy
- Dataset: Enterprise Dataset from Ministry of Industry and Commerce
- Privacy: High
- Difficulty to Apply Dataset: High

What do Stakeholders Want?

- **Governments and Corporations**: secure data; track how data is used by who

- **Researchers**: analyze data securely with low cost and no leakage

- **People and Enterprises**: know who is using my data, how and why; the right to say NO when privacy is a concern

Reach Balance

- **Sensitive Data Security** v.s**. Data Accessibility**

- **Create Value in Research** v.s. **Data Ownership by Everyone**

How Blockchain Helps?

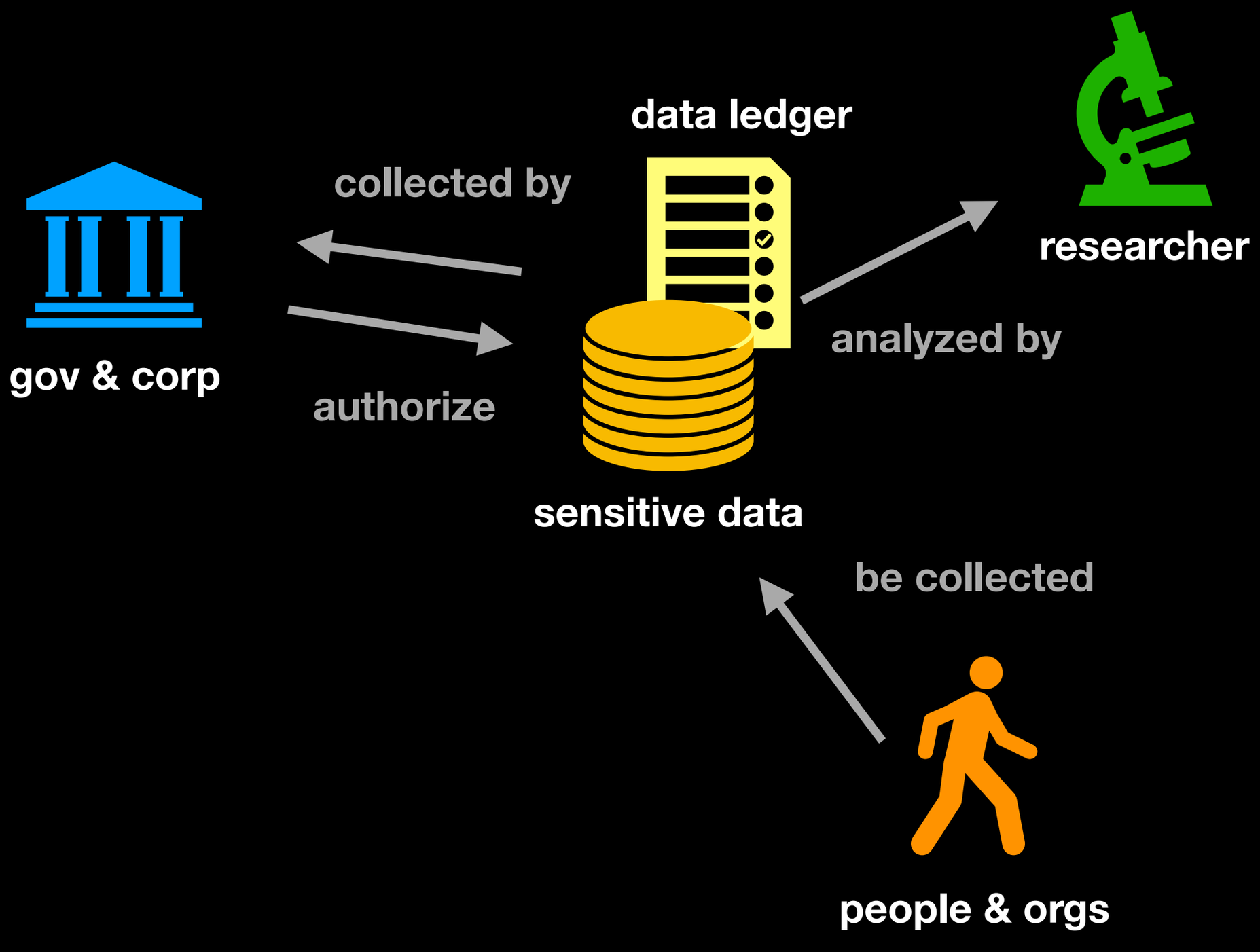
- **Transparency:** Similar to transaction ledger, we need a data usage ledger for auditing

- **Automation:** The data ledger should work automatically with consensus via smart contract

- **Security:** The researchers should be identified and data needs to be encrypted; reputation of researchers taken into account

- **Payment**: Researcher got rewards or punishment via smart contract; set up data market by tracking ledger

Trustable Data Usage in Research



Benefits to All

- **Governments and Corporations**: less concern about security; transparency about know how the data is used; may make more dataset available to researchers

- **Researchers**: keep track of data usage; get access to data sooner; make data analysis reproducible

- **People and Enterprises**: transparency about big data usage in research; say NO when issues are found.

- **Data Democracy**: In long term, the most of **data ownership will be back into the hands** of people and enterprises themselves, but still make some data usable by other groups including governments, corporations, researchers, etc.

How to Build with Blockchain

- **Identity:** researchers need to be verified to use the data;
- **Data Repositories**: a robust and secure layer of dataset infrastructure should be built to share data peer-to-peer or via a public registry
- **Data Usage Ledger:** a dataset toolkit and SDK that records all queries in ledger; built on top of an efficient public chain or side chain; support payment of data market potentially

Challenges

- **Privacy Issue of Data Ledger:** the private info of people and orgs in ledge should be recorded in a safe approach; the data ledger should not be the new source of data leakage;

- **Capability of Data Authorization by People**: need to add one module to enable people to start/stop authorizing the data to researchers when needed;

- **Cultural Shift**: Educate people to support the valuable research when needed, instead of stopping all data authorization because of fear.