Best Practices

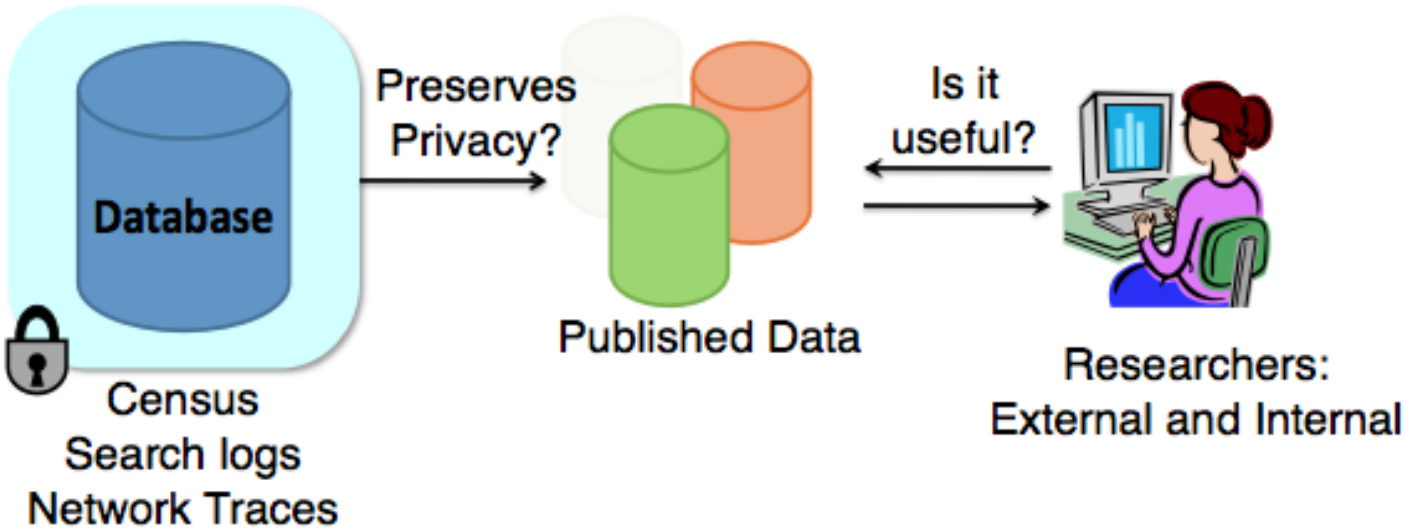
Daniel Kifer

Department of Computer Science & Engineering
Penn State University

Best Practices

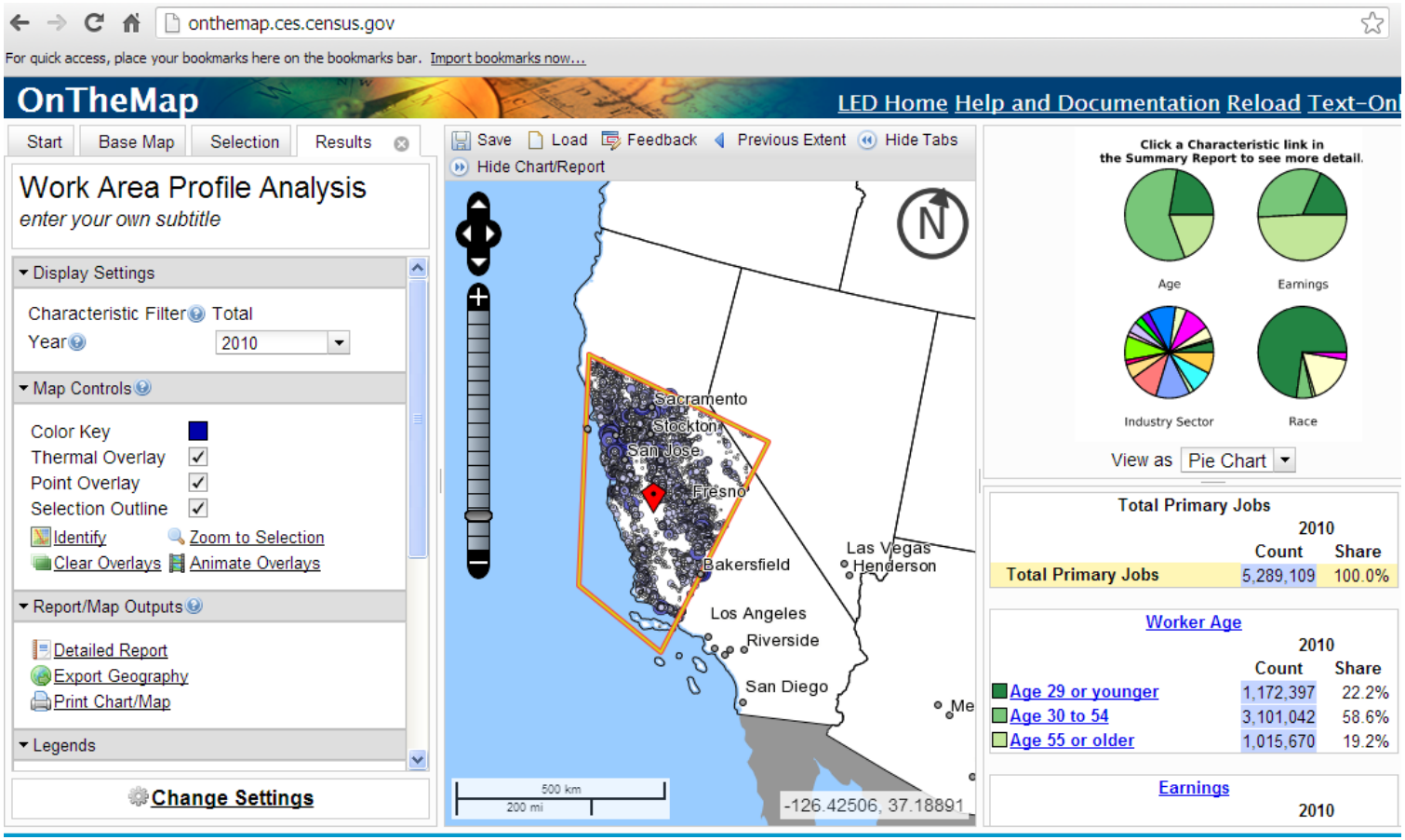
1. Details of the anonymizing algorithm must be public
2. Privacy definition is a contract
3. To randomize or not?
4. Start with utility
5. Expert implementation
6. Consider Data Enclave

Details of the anonymizing algorithm



- Details and parameters of anonymization algorithm must be made public.
- Statistical analyses are only valid if we know what was done to the data.

On The Map



- Census data product
- Redesigned 2008 by Machanavajjhala, Kifer, Abowd, Gehrke, Vilhuber
- Purpose:
  - provide rigorous privacy guarantees
  - allow details of anonymizing algorithm to be released

Privacy definitions are contracts

- If input table is
  - Then output
- Privacy definitions
  - Contracts algorithms must follow
  - Ensure algorithms are safe
  - Can contract be abused?
    - yes for k anonymity

Zip Code	Age	Nationality	Disease
13053	25	Indian	Cold
13068	39	Russian	Stroke
13053	27	American	Flu
14850	43	American	Cancer
14850	57	Russian	Cancer
14853	40	Indian	Cancer

Zip Code	Age	Nationality	Disease
130**	< 40	*	Cold
130**	< 40	*	Stroke
130**	< 40	*	Flu
1485*	≥ 40	*	Cancer
1485*	≥ 40	*	Cancer
1485*	≥ 40	*	Cancer

Privacy definitions are contracts

- Type of language to look for:
  - Whether or not Bob has cancer, the output data will probably be the same
  - Whether or not the Doe family participates, energy data will probably be the same.
  - Example: differential privacy (may not lead to useful energy data).

To randomize or not?

- Pesky neighbor repeatedly asks you if today is your birthday.
  - You initially answer "no"
  - Until one day you decline to answer (it is probably your birthday).
- Private policy: always decline to answer (I can neither confirm nor deny).
- Randomized policy: lie with some probability
  - Preserves privacy
  - Increases utility

Start with utility

1. Identify analyses end users want.
2. Identify aggregate information they need.
3. Identify privacy risks contained in the information.
4. Decide on privacy/utility tradeoff.
5. Then design algorithm.
6. Redesign as requirements change.

Expert Implementation

- Someone has to implement anonymizing algorithm.
- Experienced programmer is not enough.
- Need experienced programmer with privacy expertise.
  - Floating point computations (how computers store numbers) can breach privacy.
  - Use of a computer's random number generator can breach privacy.
  - Ok for normal code. Not ok when security is important.

Data enclaves

- Benefits of anonymized data:
  - Low maintenance/distribution cost.
- Drawback:
  - Limitations on data quality
  - Statistical analysis more complex.
- Alternative: data enclave
  - Secure computing facility.
  - User travels on site. Cannot bring electronics in or out.
- Benefits:
  - Higher data quality
  - Statistical analysis is easier
- Drawback:
  - High cost.
  - Low availability.