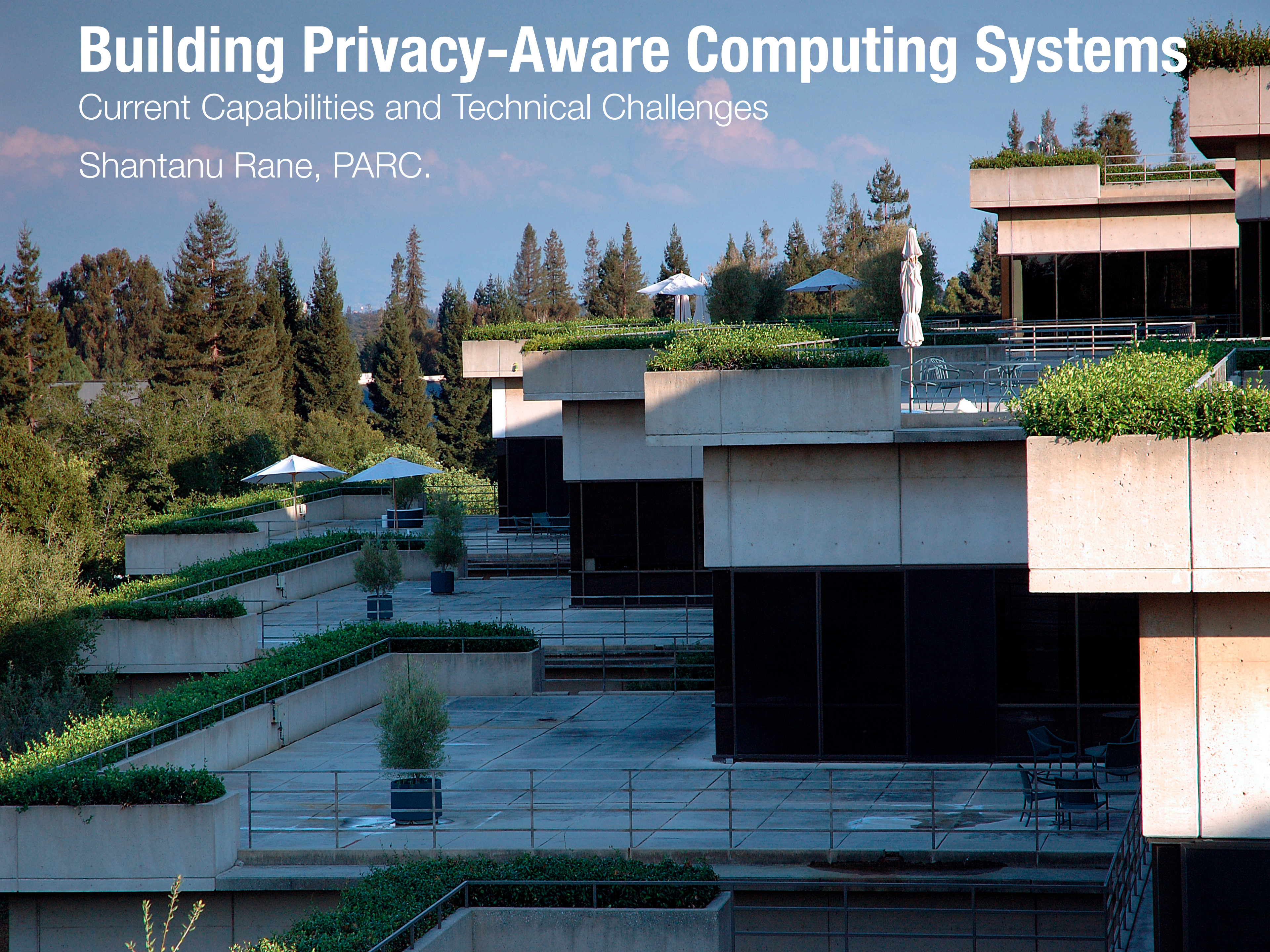
Building Privacy-Aware Computing Systems

Current Capabilities and Technical Challenges

Shantanu Rane, PARC.



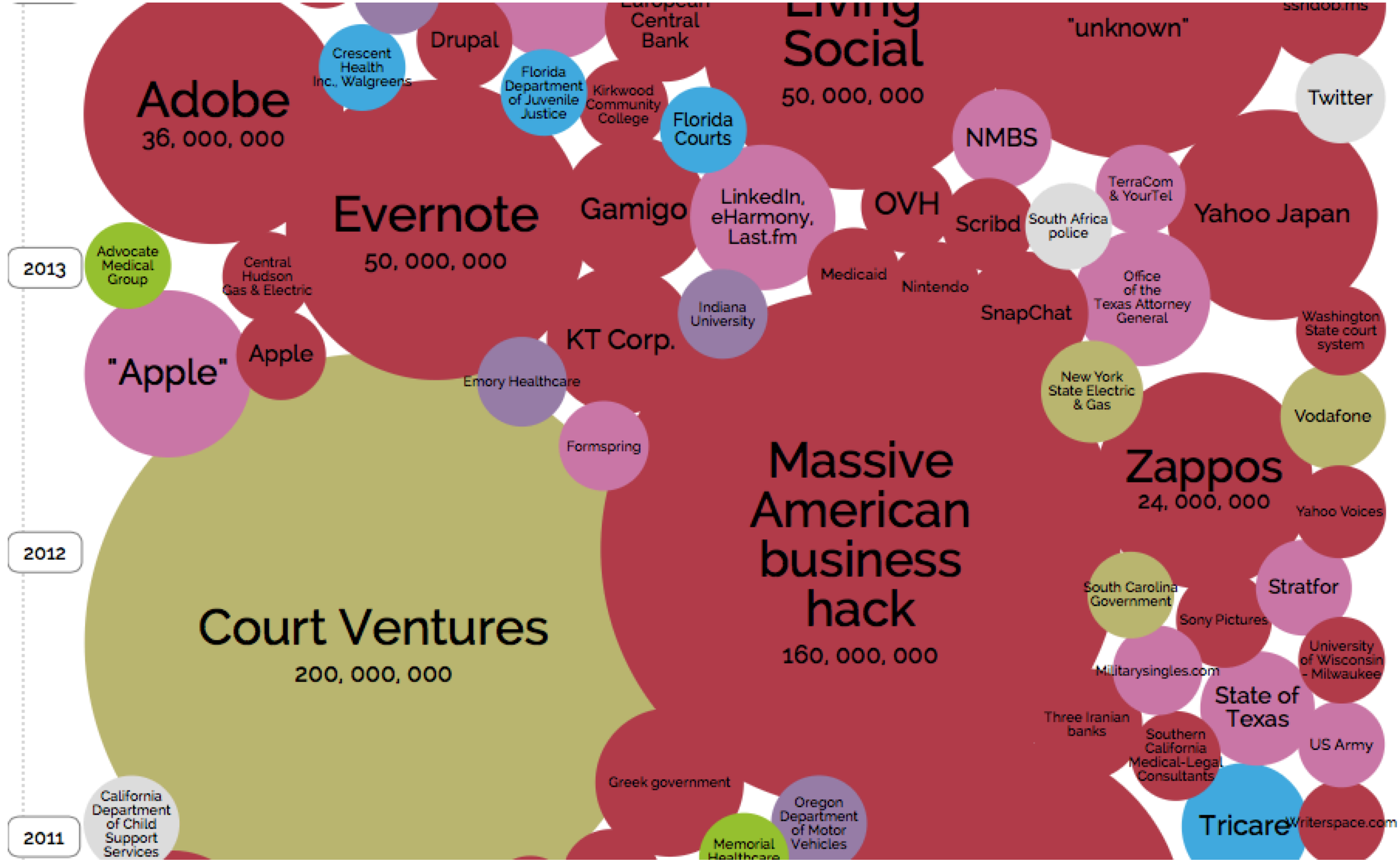
http://posttypography.com

- Social network data
- Smartphone app data
- Online shopping
- Car navigation data
- Biometrics
- Healthcare data
- Internet of things telemetry
- Smart grid pricing & usage
- Intellectual property
- Industrial diagnostics data
- Demographic data
- National security data

⋮

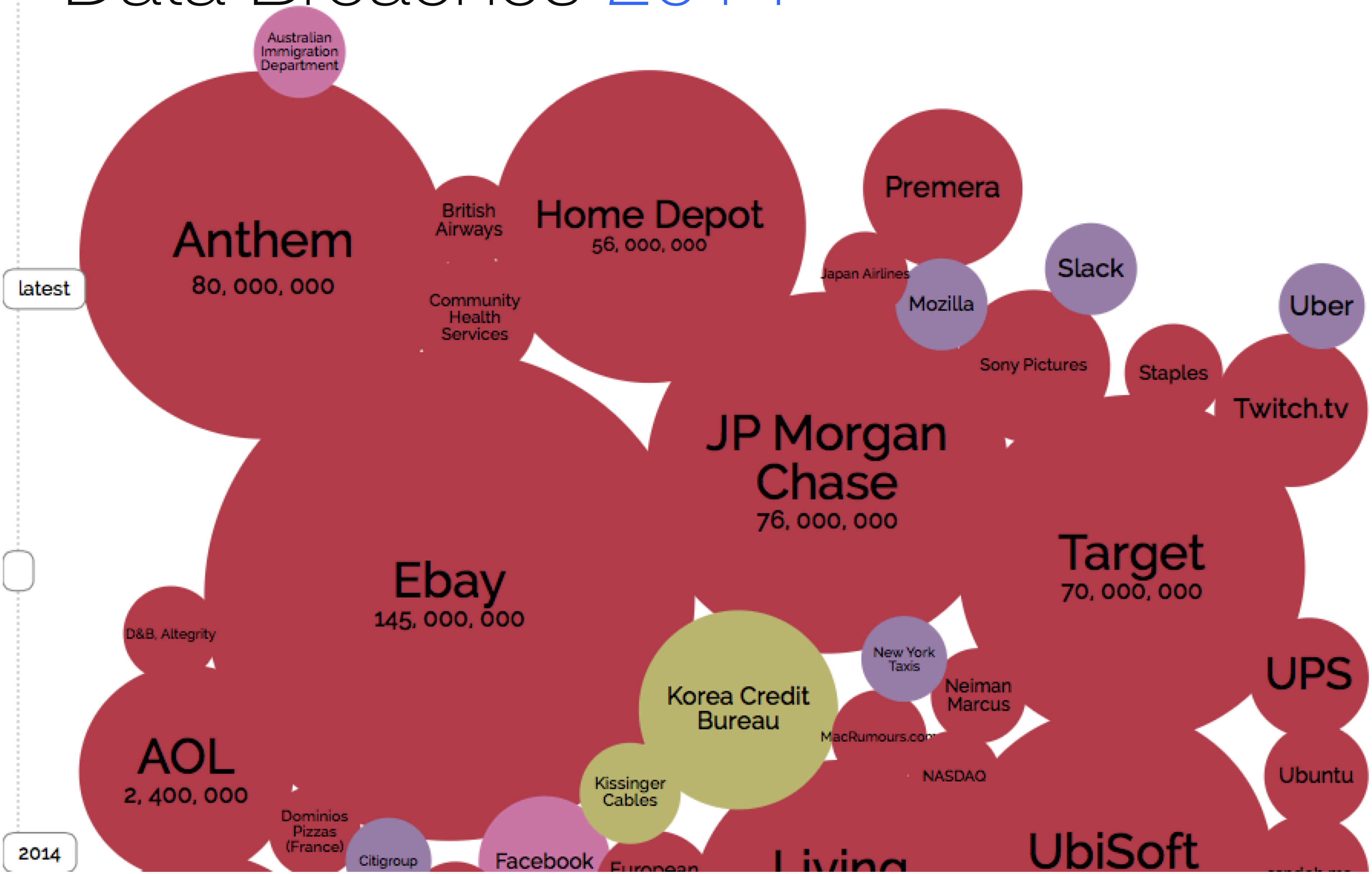
http://informationisbeautiful.net

Data Breaches 2011 2012 2013



http://informationisbeautiful.net

Data Breaches 2014

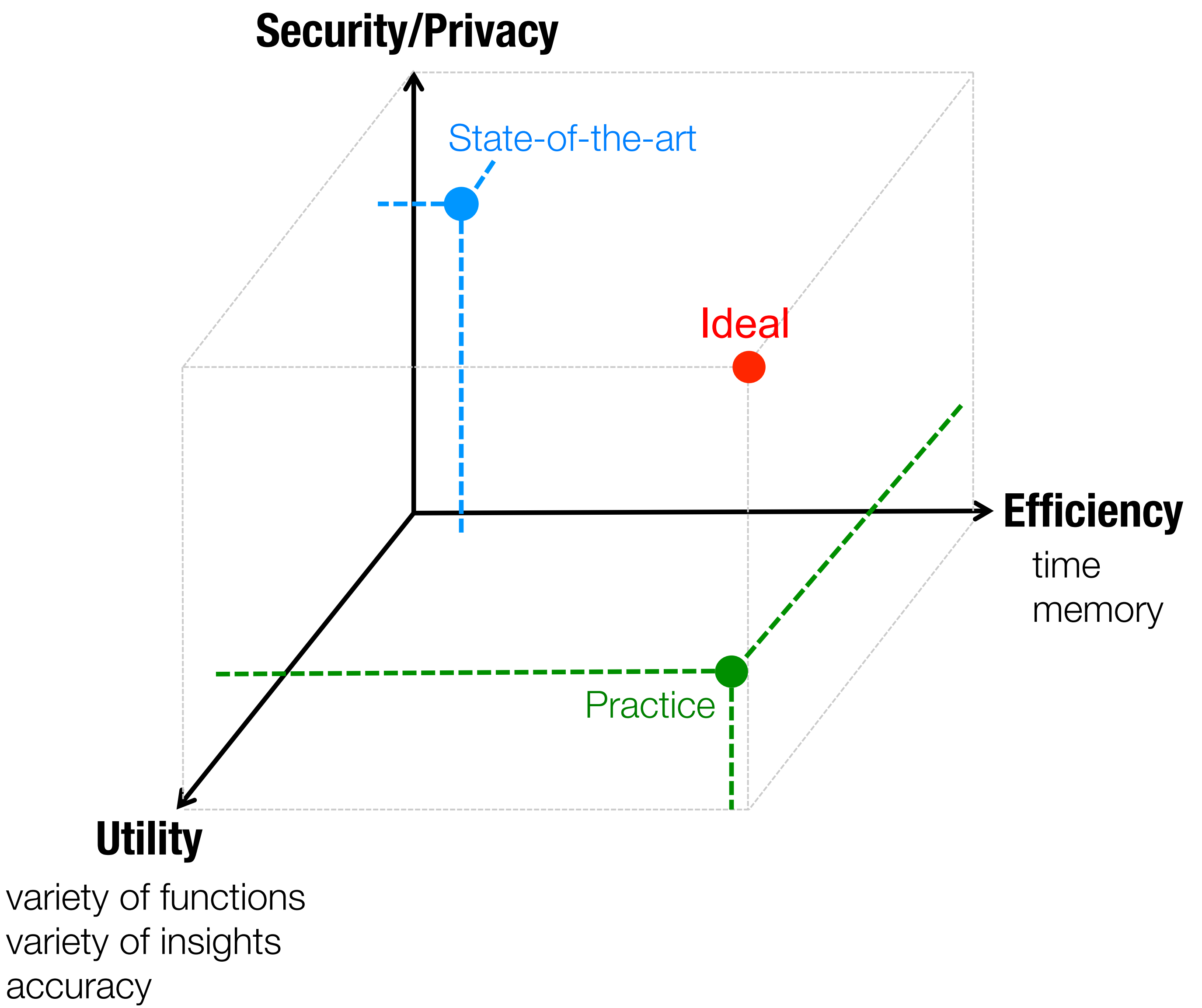


“**Recommendation 3**: … the NITRD agencies, should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.”

“…. create appropriate balance among economic opportunity, national priorities, and privacy protection.”

[PCAST Report, May 2014]

Privacy Research vs Deployment



Outline

1. Data analytics setting
2. Privacy preserving tools
   - Computational
   - Statistical
3. Reflections on future directions

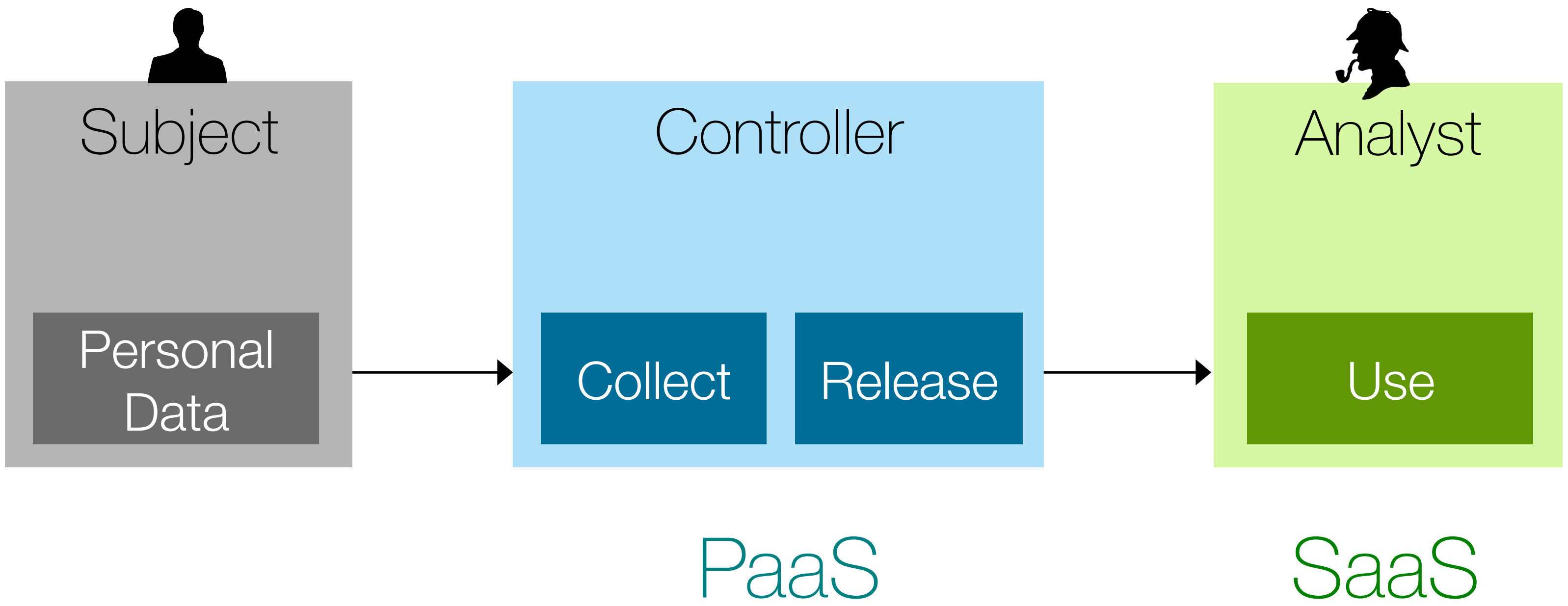
Data | Computation | Insights

Owner | | Analyst

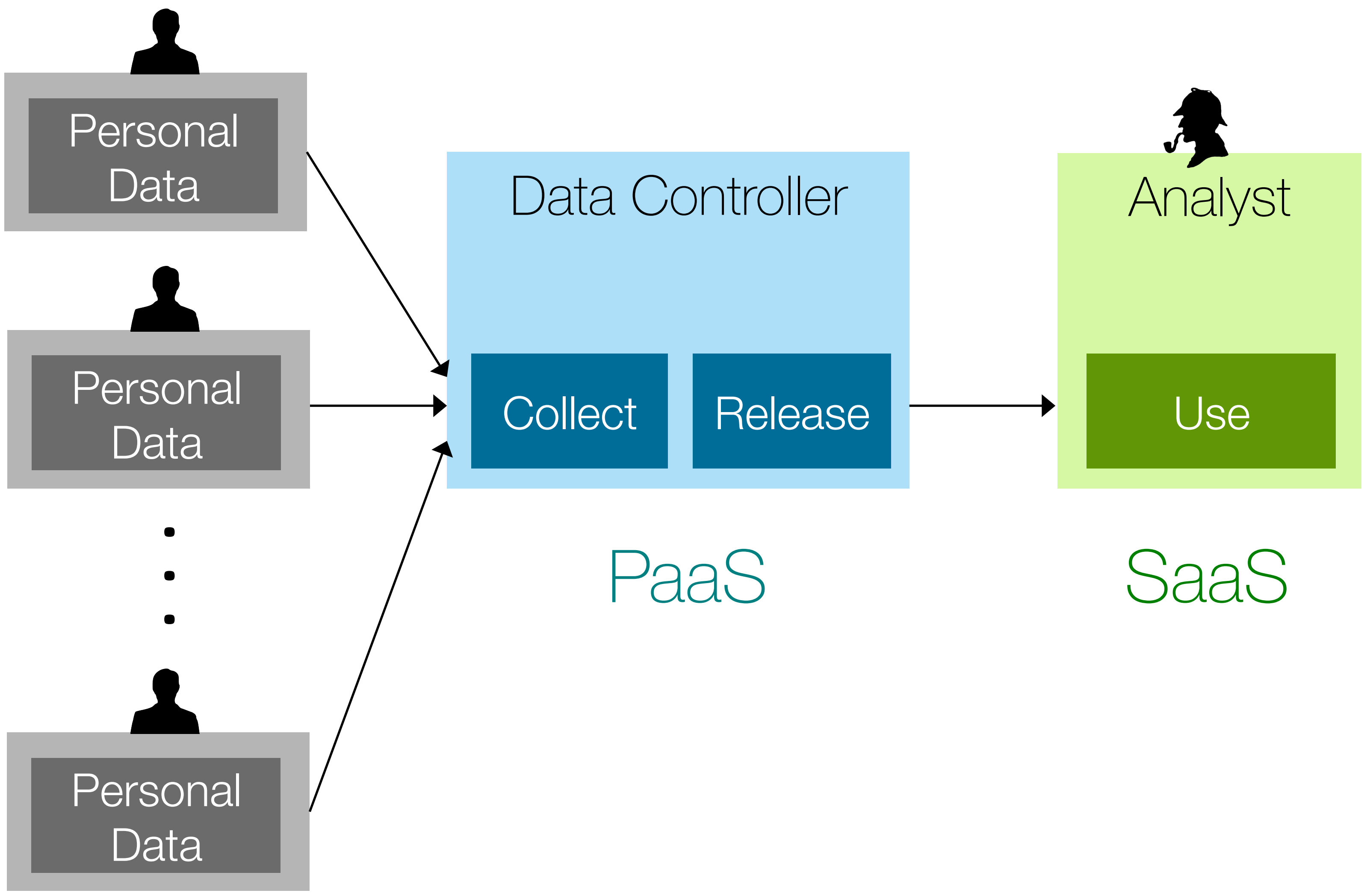
1

The Data Analytics Setting

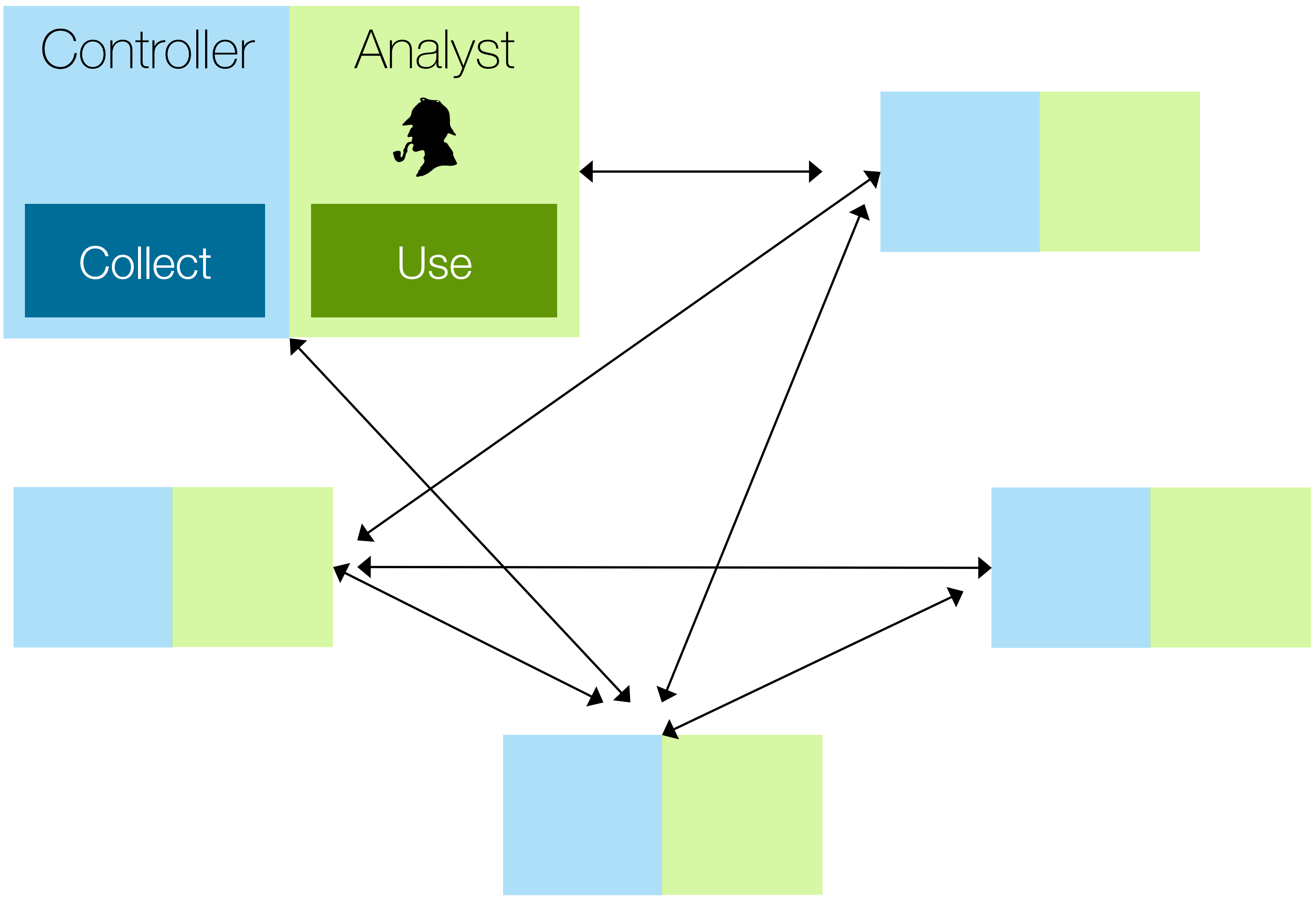
Data Analytics Setting



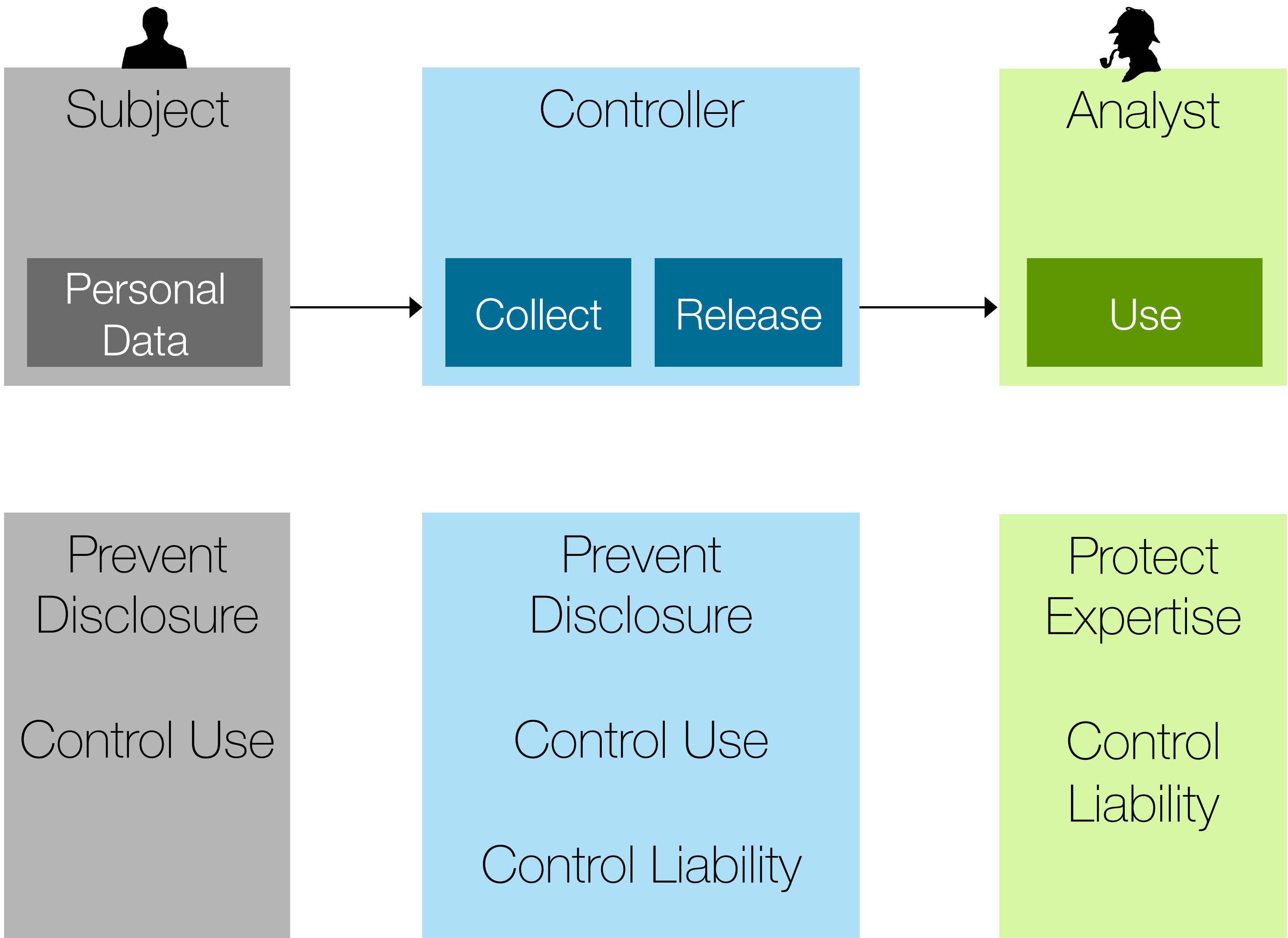
Personal Privacy Setting



Enterprise Privacy Setting

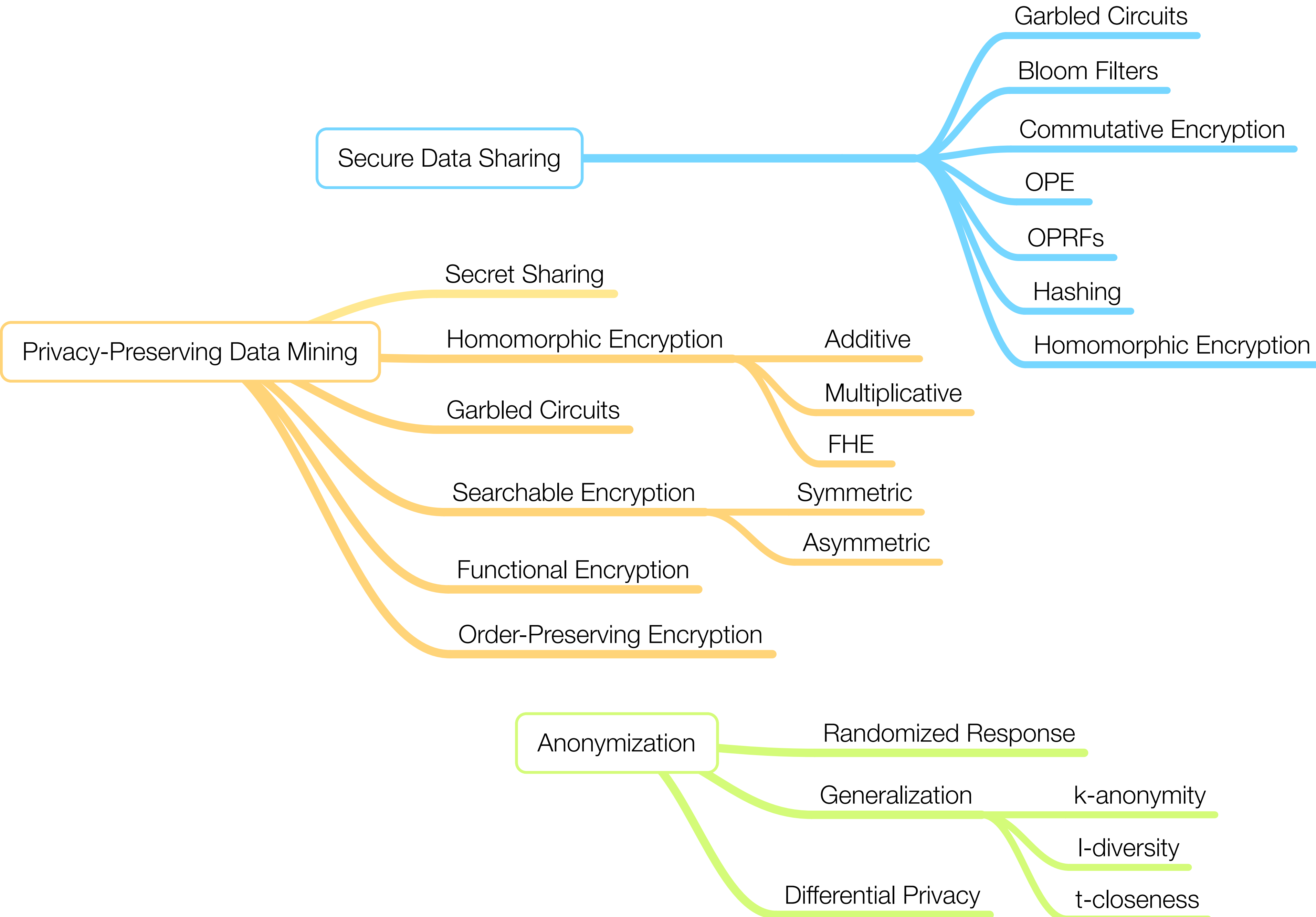


Privacy & Security Requirements



2

Tools, their capabilities & limitations


Secure Data Sharing
Garbled Circuits
Bloom Filters
Commutative Encryption
OPE
OPRFs
Hashing
Homomorphic Encryption
Privacy-Preserving Data Mining
Secret Sharing
Homomorphic Encryption
Additive
Multiplicative
FHE
Garbled Circuits
Searchable Encryption
Symmetric
Asymmetric
Functional Encryption
Order-Preserving Encryption
Anonymization
Randomized Response
Generalization
k-anonymity
l-diversity
t-closeness
Differential Privacy

Privacy-Preserving Data Sharing

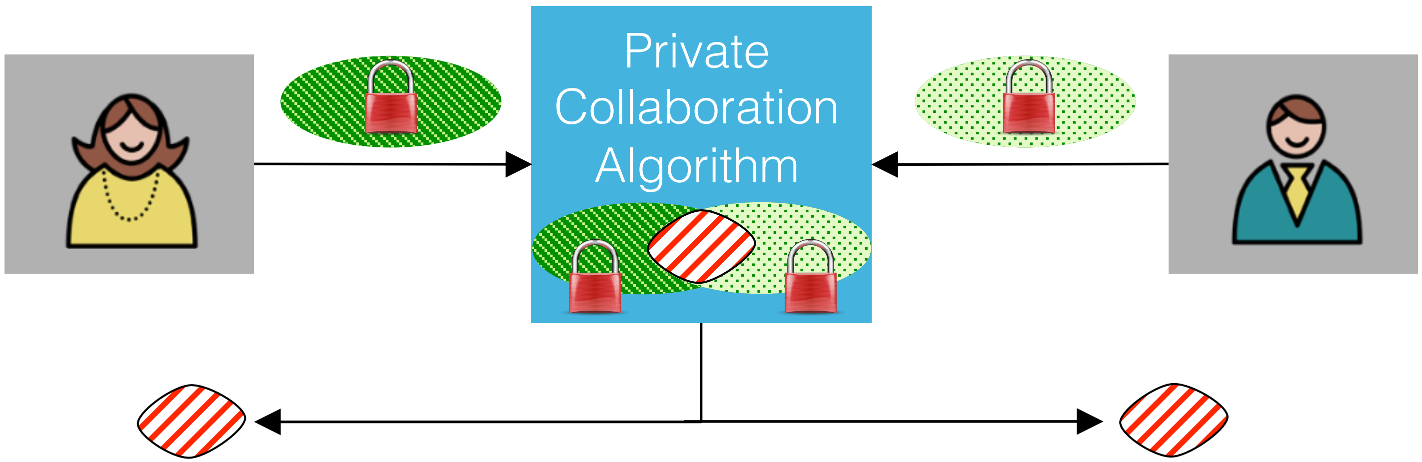
Privacy questions

1. How to share common data w/o revealing unique data?
2. How to privately ascertain whether data is worth sharing or purchasing?

Applications

Cyber threat mitigation, recommendation engines, data monetization

Private Set Intersection



- ✓ Can be implemented in many ways with classical cryptographic tools, e.g., Bloom filters, hashing, RSA-style encryption, etc.
- ✓ Can be made secure against malicious participants.
- ✗ Supports a very specific operation, e.g., efficient for PSI, but very inefficient for count queries.
- ✗ Hard to use with noisy data.

Privacy-preserving Data Mining

Privacy Questions

1. Which queries are possible given available privacy primitives?
2. How to preserve database privacy and query privacy?

Applications

Federated search, Healthcare analytics, Data quality assessment, Education analytics, Call graph analysis, Transportation analytics, too many to list.

Functions

sum	
product	set intersection
mean	set union
variance	set cardinality
distances	histogram
polynomials	max/min
correlation	selection
filtering	classification
graph processing	edit distances

Homomorphic Cryptosystems

Additive

[Paillier 99, Damgard-Jurik 01]

$$E(x)E(y) \equiv E(x+y)$$

Multiplicative

[El Gamal 85]

$$E(x)E(y) \equiv E(xy)$$

2-DNF homomorphic

[Boneh, Goh, Nissim 05]

$$e(E(x), E(y)) \equiv F(xy)$$
$$F(xy+uv) \equiv F(xy)F(uv)$$

Fully homomorphic

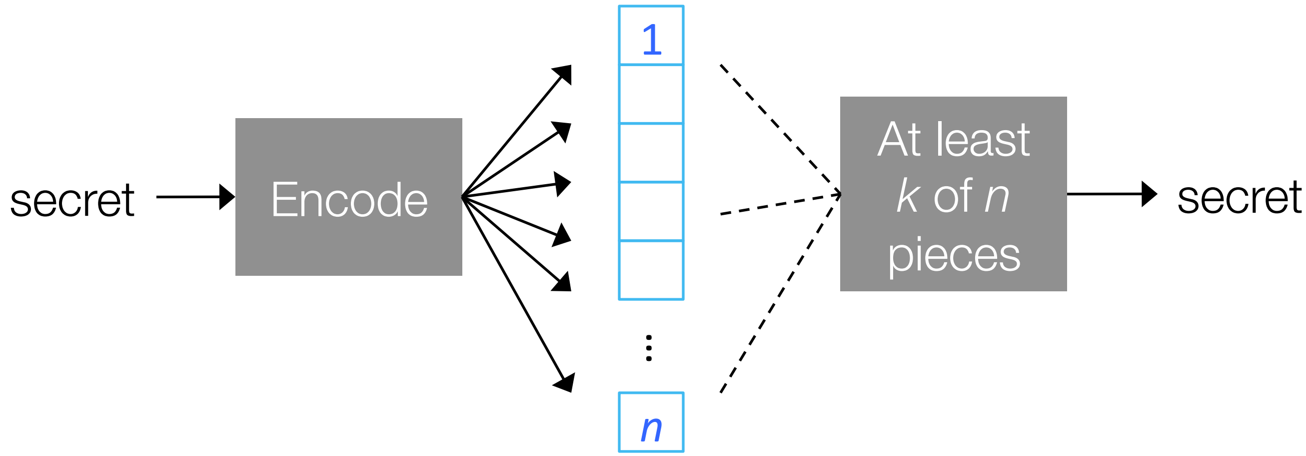
[Gentry, 09]
[Gentry, Halevi, Vaikunthanathan 10]
[Brakerski, Vaikunthanathan 10]

$$E(x+y) \equiv E(x) + E(y)$$
$$E(x)E(y) \equiv E(xy)$$

Homomorphic Cryptosystems

- Enables outsourced cloud computing for rich variety of functions.
- Some formulations, e.g., Ring Learning With Errors, are resistant to quantum computing attacks.
- Memory access patterns reveal information about data elements. (cf. ORAM)
- Most schemes were developed for semi-honest parties. For malicious parties, use ZKP, but this increases complexity.
- Data is growing faster than computational power. Moore’s law won’t save us from the complexity of FHE.

Secret Sharing



Can be achieved using error correcting codes. [Shamir, 1979]

✓ At the heart of information-theoretically secure multiparty computation. [BGW,1988][CCD,1988]. Each party computes functions of shares, which are combined to obtain a function of the secret.

✓ Computationally efficient. Tolerates < $n/3$ cheaters for arbitrary functions.

✗ Must keep track of inter-participant communications. Not much is known for computation with n=3 parties! [Wang, Ishwar, Rane, 2014]

Garbled Circuits & Oblivious Transfer

	0	1
0	0	1
1	1	1

OR Gate

Garble

	m_0	m_1
k_0	$E(k_0 \| m_0 , 0)$	$E(k_0 \| m_1 , 1)$
k_1	$E(k_1 \| m_0 , 1)$	$E(k_1 \| m_1 , 1)$

[Ex from Prabhakaran's Crypto Notes, 14]

Alice produces garbled circuit for function *f*

Alice provides her keys corresponding to her input to Bob

Bob obtains his keys from Alice via 1-of-2 OT

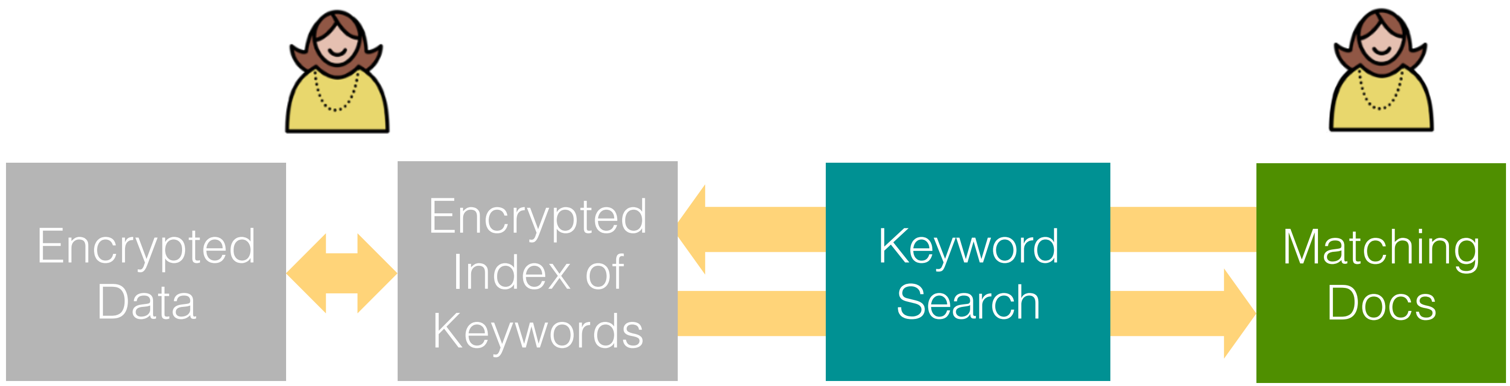
Bob evaluates circuit by decryption using his and Alice's keys

Implementations: Fairplay [Malkhi, Nisan, Pinkas, Sella, 04]

GCs: Advantages and Limitations

- ✓ General primitive for secure computation. [Yao, 86]
- ✓ Speed-up: Free XORs, row reduction [Pinkas, Schneider, Smart, Williams 09] [Kolesnikov, Schneider 08].
- ✓ Very impressive recent results on Levenshtein distance, Hamming distance, AES. [Huang, Evans, Katz, Malka, 11].
- ✗ Circuits can be extremely complex for data-mining tasks such as classification, clustering, etc., especially with > 2 parties.
- ✗ Circuit design and garbling requires in-house expertise.

Searchable Encryption

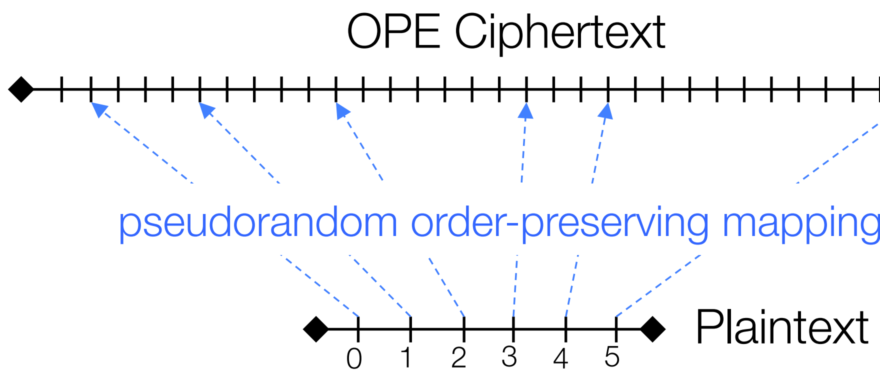


Symmetric constructions based on ORAMs [Song, Wagner, Perrig, 00]. [Curtmola, Garay, Kamara, Ostrovsky, 06]

Public-key construction based on bilinear maps on elliptic curves. [Boneh, Di Crescenzo, Ostrovsky, Persiano, 04]

- Compatible with conjunctive, subset, range queries [Boneh, Waters, 07].
- Can be vulnerable to repeated queries.
- Public-key methods leak document identifiers.

Order-Preserving Encryption



Weaker cryptographic technique where ciphertexts preserve order

- Need knowledge about data values [Agarwal, Kiernan, Srikant, Xu, 04]
- One-shot method with hyper-geometric sampling [Boldyreva, Chenette, Lee, O'Neill, 09, 11]

✓ Supports range queries, median finding, and is deployed within cryptDB. [Ala Popa, Redfield, Zeldovich, Balakrishnan, 11, 12, 13]

✗ Ciphertext expansion can be prohibitive.

Anonymization

Privacy Questions

1. Which attributes are sensitive?
2. How to anonymize sensitive attributes?
3. What is the privacy-utility tradeoff for analytics on output data?
4. What is the risk of re-identification via external linkage?

Applications

Disclosure control methods for advertising, healthcare, smart grid, education analytics, etc.

Masking

John Smith	32	92043	American	Heart Disease
Kei Takamura	34	92043	Japanese	Cancer
Sarah Jones	38	92043	American	Cancer
Cesar Vincent	37	92306	French	Viral Infection

askdhsf	32	92043	American	Heart Disease
lkjljhflgl	34	92043	Japanese	Cancer
rwithgd	38	92043	American	Cancer
vmbnvc	37	92306	French	Viral Infection

Replaces PII with pseudonymous identifiers

✓ Easy and fast. Identify sensitive attributes and hash them.

✓ High utility, as long as only a few attributes are masked.

✓ HIPAA compliant.

✘ Masking does *not* preserve privacy

askdhsf	32	92043	American	Heart Disease
lkjljhflgl	34	92043	Japanese	Cancer
rwithgd	38	92043	American	Cancer
vmbnvc	37	92306	French	Viral Infection

+

Kei Takamura	92043	Japanese Instructor

→

askdhsf	32	92043	American	Heart Disease
Kei Takamura	**34**	**92043**	**Japanese**	**Cancer**
rwithgd	38	92043	American	Cancer
vmbnvc	37	92306	French	Viral Infection

MA Governor medical records [Sweeney 02]

NYT re-identification of AOL Search Data [Barbaro, Zeller, 06]

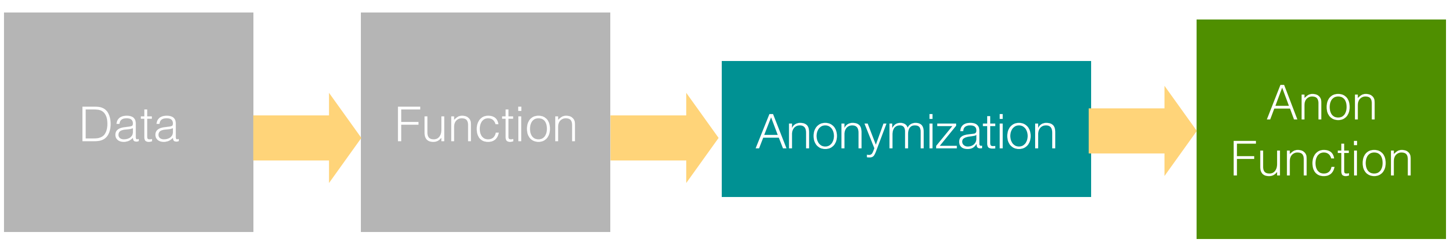
“Innocuous” DNA Statistics [Homer et al. 08]

De-anonymization of Netflix database [Narayanan, Shmatikov 08, 11]

Anonymization Methods

Input perturbation / generalization (e.g., k-anonymity)

Output perturbation (e.g., differentially private mechanisms)



k-anonymity and variants

32	American	92043	Heart Disease
34	Japanese	92043	Cancer
38	American	92043	Cancer
37	French	92306	Viral Infection

$k = 4$

[30, 40]	*	92***	Heart Disease
[30, 40]	*	92***	Cancer
[30, 40]	*	92***	Cancer
[30, 40]	*	92***	Viral Infection

A record is indistinguishable from k-1 other records w.r.t. anonymized attributes. [Sweeney, 02]

Multidimensional methods available [LeFevre, DeWitt, Ramakrishnan 06]

k-anonymity and variants

✓ Stronger protection than simple masking.

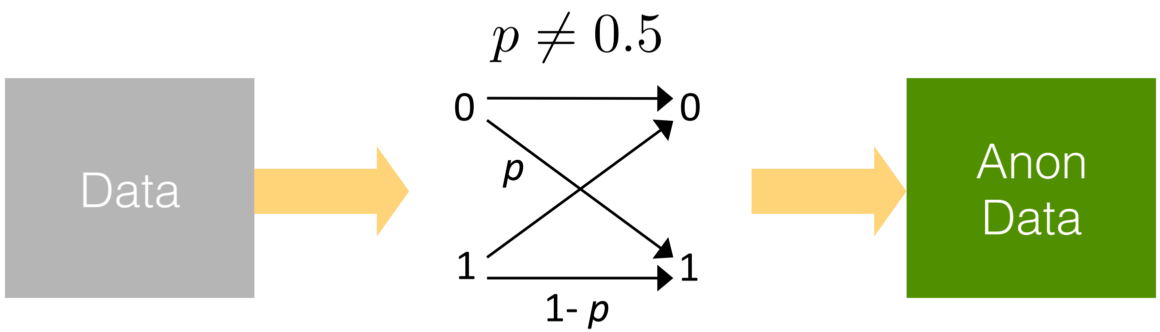
✗ Leaks information if sensitive attribute has low diversity, e.g., all patients have cancer.

✗ ℓ-diversity addresses diversity issue, but susceptible to skewness attacks on attribute values in an equivalence class. [Machanavajjhala et al. 07]

✗ t-closeness address skewness, but destroys useful correlations in the process. [Li, Li, Venkitasubramanian, 07] [Domingo-Ferrer and Torra, 2008]

Randomized Response

Binary case: Given *p*, estimate % of 0/1 [Warner 65]



Post-Randomization [Kooiman, Willenborg, Gouweleeuw 98]

$$\begin{bmatrix} a_{1,1} & \cdots & a_{\ell,\ell} \\ \vdots & \ddots & \vdots \\ a_{\ell,1} & \cdots & a_{1,\ell} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_\ell \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_\ell \end{bmatrix}$$

Original PDF　　Perturbed PDF

Randomized Response

✓ Simple: usually add noise to the data.

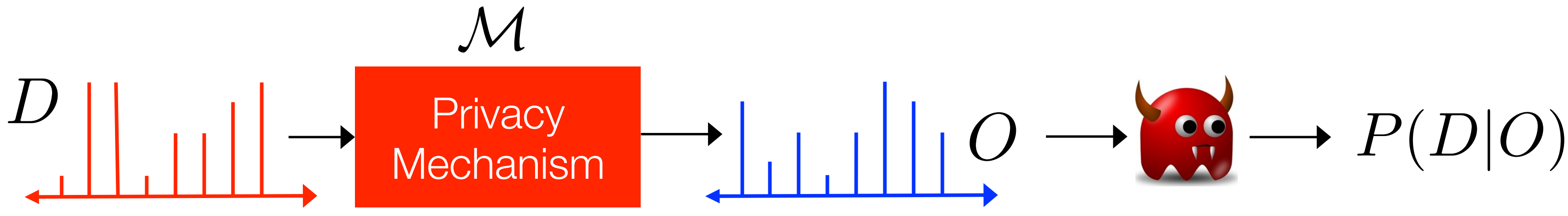
✓ Good for aggregate statistics e.g., PMFs, means, etc.

✗ Not suitable for many common tasks, e.g., max / min.

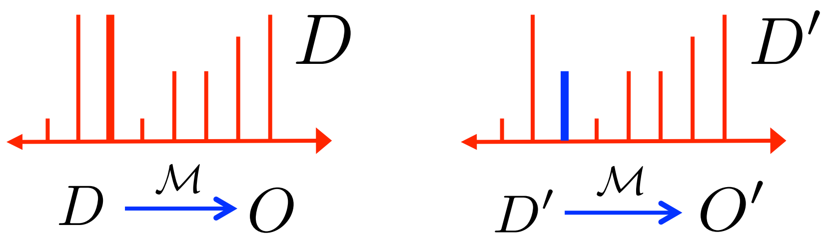
$$\begin{bmatrix} a_{1,1} & & \\ & \ddots & \\ & & a_{1,\ell} \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_\ell \end{bmatrix} = \begin{bmatrix} q_1 \\ \vdots \\ q_\ell \end{bmatrix}$$

✗ Privacy-utility tradeoff degrades very rapidly upon composition, as PRAM matrices can become poorly conditioned. [Lin, Wang, Rane, 12]

Differential Privacy



Perfect privacy $\Rightarrow P(D|O) = P(D)$ useless in practice.



Need $\dfrac{P(O \in \mathcal{S}|D)}{P(O' \in \mathcal{S}|D')} \leq e^{\epsilon}$

Differential Privacy: Output is **insensitive** to any single element in D. Thus D and D' appear statistically indistinguishable to an adversary.

[Dwork, 06, 08, 09]

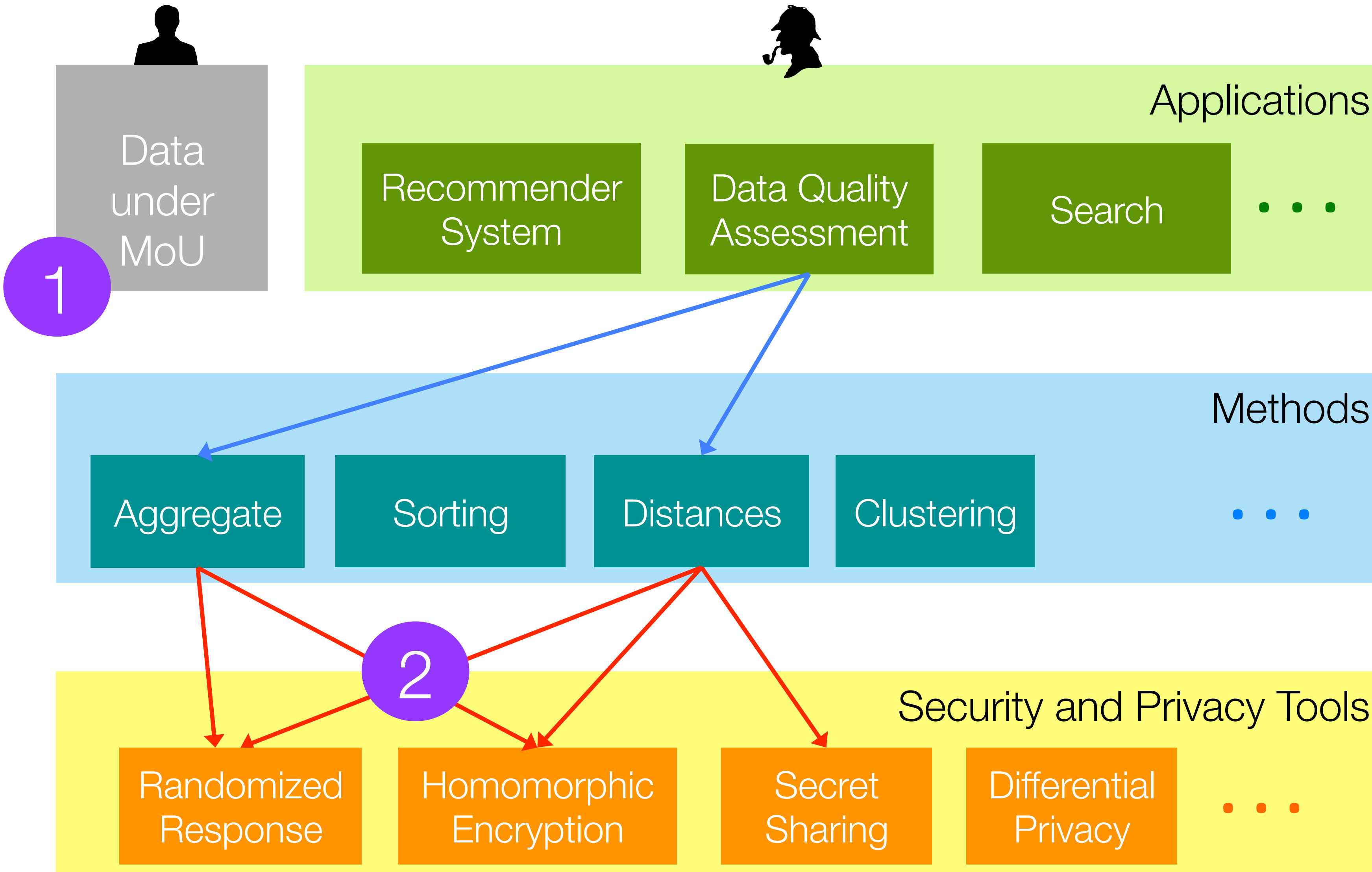
Differential Privacy

- ✓ Provides strong protection against adversaries with background information, unlike *k*-anonymity. [Kasiviswanathan, Smith, 08]
- ✓ Additively composable, i.e., if two mechanisms provide DP, then their cascade provides DP (albeit lower privacy than before).
- ✗ Treats all records as equally private, heavily obfuscates rare values.
- ✗ Noise variance is proportional to sensitivity of the function being published. Hard to determine. [Nissim, Raskhodnikova, Smith 07]
- ✗ Privacy deteriorates with the number of queries. [Dwork 10]

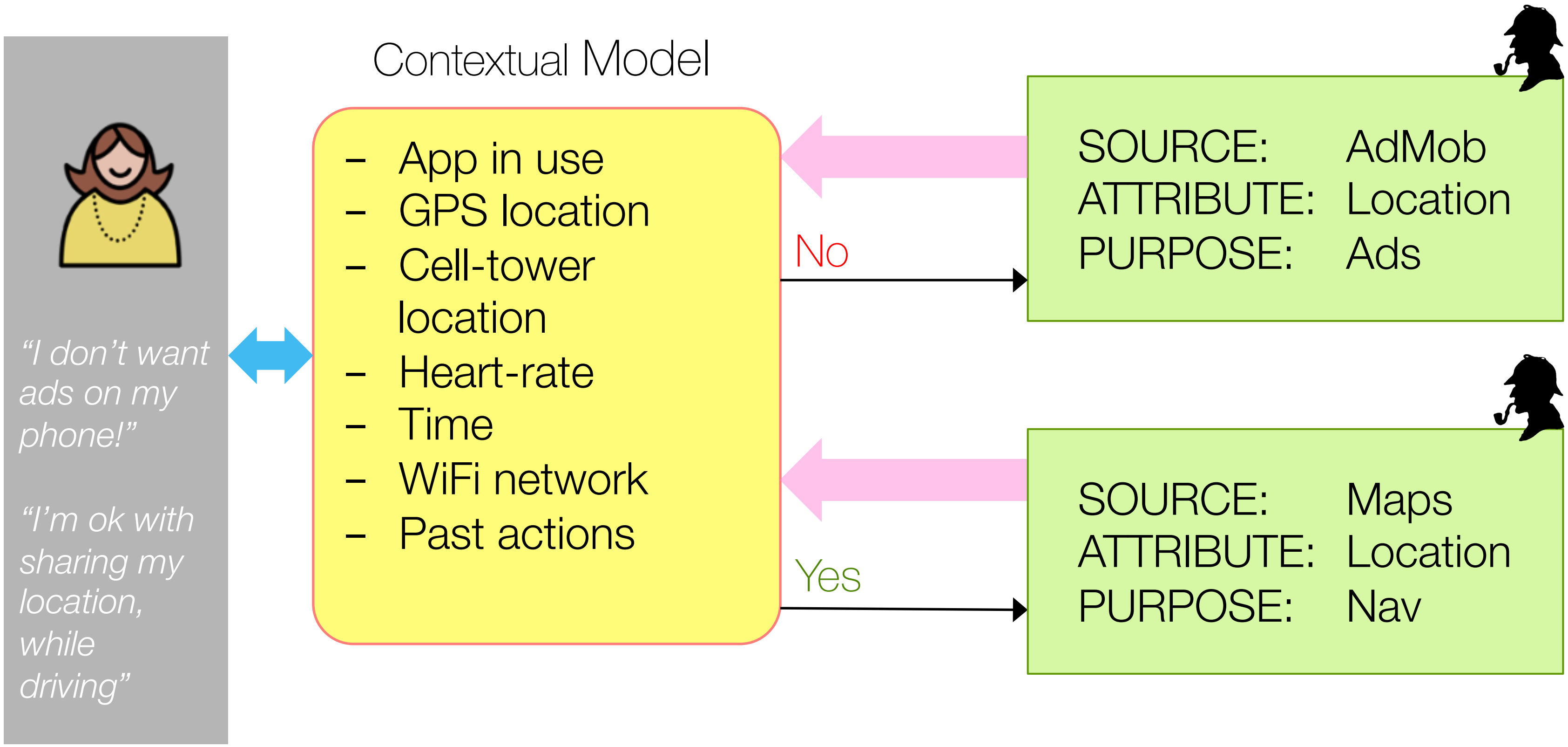
3

Reflections on future directions

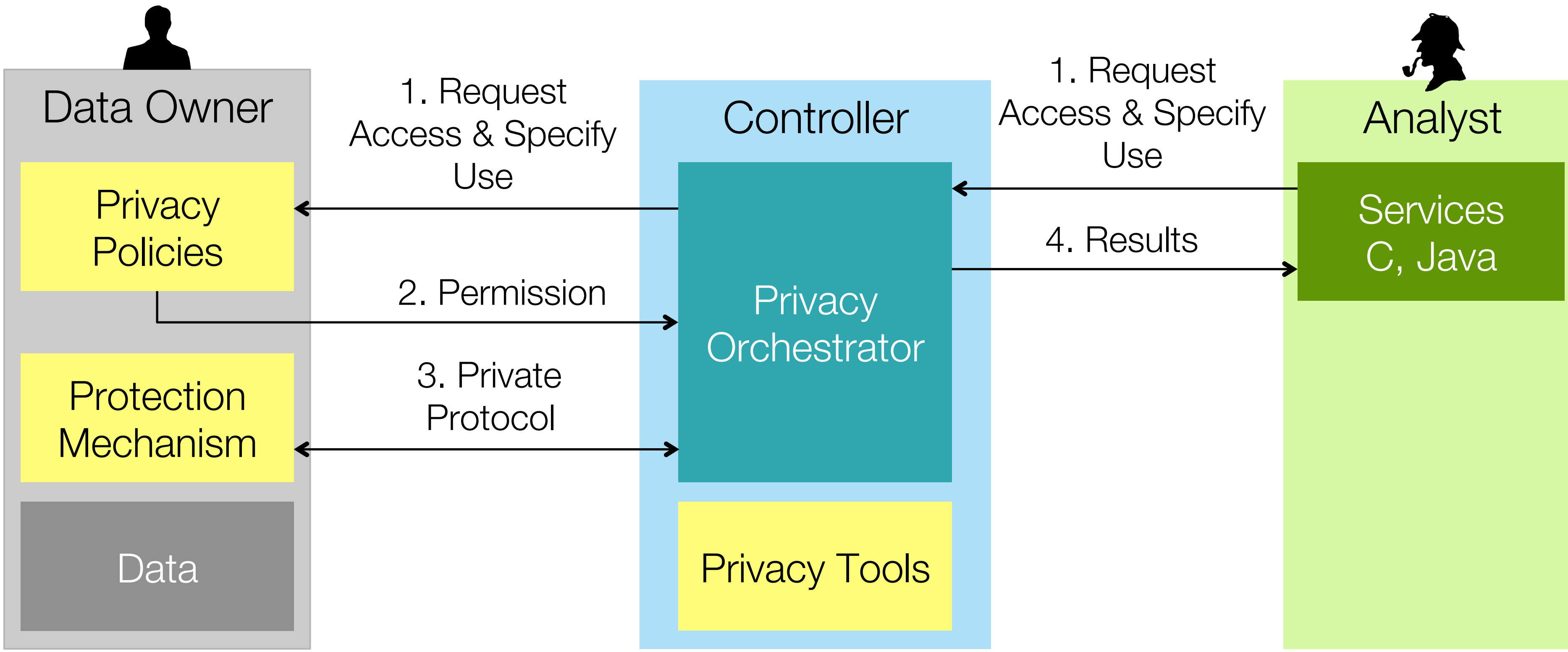
How We Achieve Privacy Today



Owner-controlled Privacy Policies



Orchestrating a Data Transaction



Match users' requests for data against owners' privacy policies.

Rewrite analytics programs using one or more privacy tools.

Update policies using feedback from previous computations.

Conclusions

Multiple computational and statistical primitives can be leveraged for privacy in computation.

Need a way to assess and select methods according to their privacy-utility-efficiency tradeoffs.

Need interdisciplinary outlook (beyond crypto)

- Statistics: New paradigms, e.g., Differential privacy
- Machine learning: Support for legacy analytics.
- Domain-specific languages: Policy & Querying languages
- Signal processing: Dimensionality reduction