Introduction to Data Science and Analytics (DSC510)

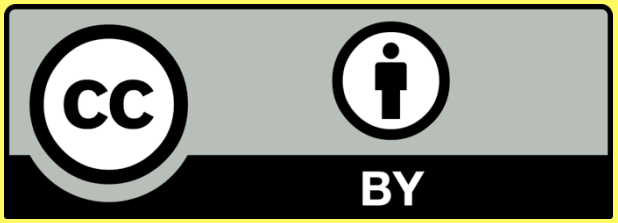
Privacy and Ethics in Data Science

George Pallis

Intended Audience

Designed for students with no programming background who want to have literacy in data and computing to better approach data science projects

- **Computational thinking**: a new way to approach problems through computing
  - Abstraction, decomposition, modularity,…
- **Data science**: a cross-disciplinary approach to solving data-rich problems
  - Machine learning, large-scale computing, semantic metadata, workflows,…



These materials are released under a CC-BY License

https://creativecommons.org/licenses/by/2.0/

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Artwork taken from other sources is acknowledged where it appears. Artwork that is not acknowledged is by the author.

Please credit as: Gil, Yolanda (Ed.) Introduction to Computational Thinking and Data Science. Available from http://www.datascience4all.org

If you use an individual slide, please place the following at the bottom: "Credit: http://www.datascience4all.org/"

As editors of these materials, we welcome your feedback and contributions.

Acknowledgments

ACI-1355475

- These course training materials were originally developed and edited by Yolanda Gil (USC) with support from the National Science Foundation with award ACI-1355475
- They are made available as part of http://www.datascience4all.org
- The course materials benefitted from feedback from many students at USC and student interns, particularly Taylor Alarcon (Brown University), Alyssa Deng (Carnegie Mellon University), and Kate Musen (Swarthmore College)
- We welcome new contributions and suggestions

Privacy

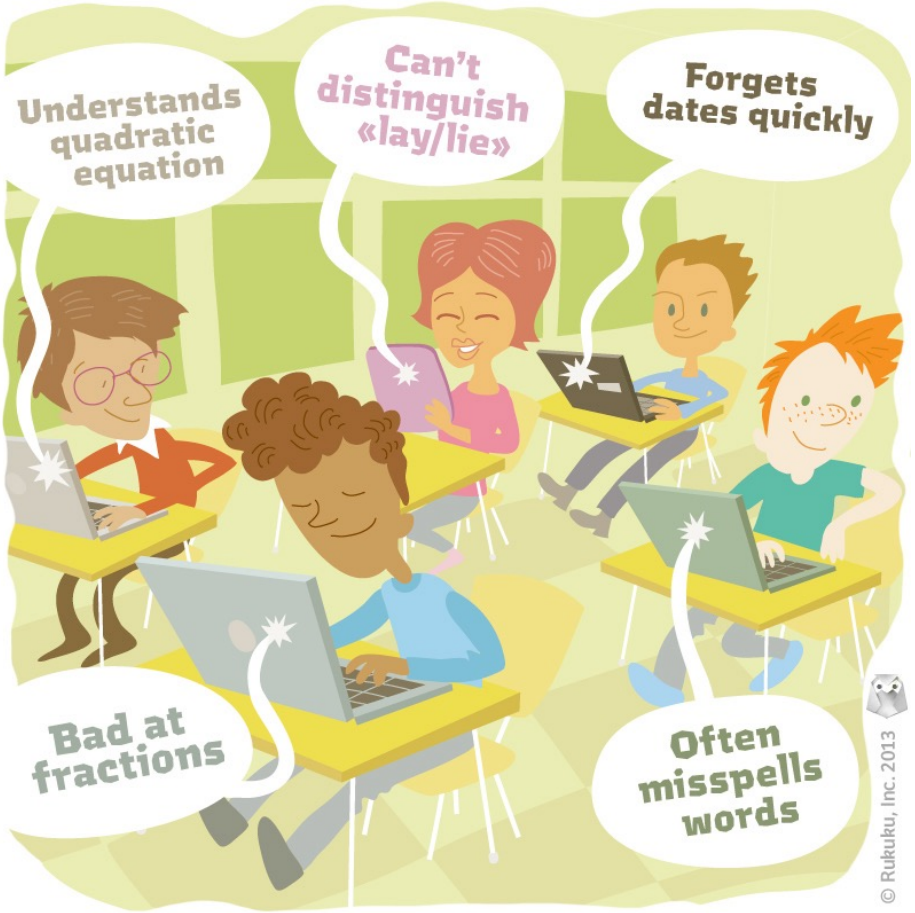
1. Privacy
   - Fair Information Practices
   - Managing sensitive data
   - Anonymizing sensitive data
   - Re-identifying datasets
2. Reproducibility
3. Societal value of data and data science

Privacy

The Rise of Privacy Concerns

- Science:
  - benefits of sharing clinical patient records
  - patients shall control access to their records
  - patients found to be altruistic:
    - willing to grant access for purpose of advancing science
- Government:
  - government and commercial use of data mining raises concerns about appropriate use of private citizen information,
  - e.g., data collected for the purpose of airline passenger screening should not be used for the enforcement of other criminal laws
- Open Web:
  - many users are happy to share private details on social webs
    - but would be rightfully upset was this data used for other purposes
  - content is shared between networks
    - not very transparent to the user
  - users need to be reassured about appropriate use of their data

Private and Sensitive Data



http://www.rukuku.com/blog/wp-content/uploads/student-data-and-personalization.jpg

http://www.crimeoff24.com/?p=4869

https://en.wikipedia.org/?title=Telephone_directory#/media/File:Telefonbog_ubt-1.JPG

Sensitive Data and Privacy

Sensitive Data

- Data about individuals and organizations that should not be freely disseminated and publicized
  - Health
  - Education
  - Finance
  - Demographic
  - Criminal
  - Location
  - Behavior

Privacy

- Desire to limit the dissemination of sensitive data
- Lots of technology, but:
  - Unclear requirements
  - Unclear behaviors

Sensitive Data

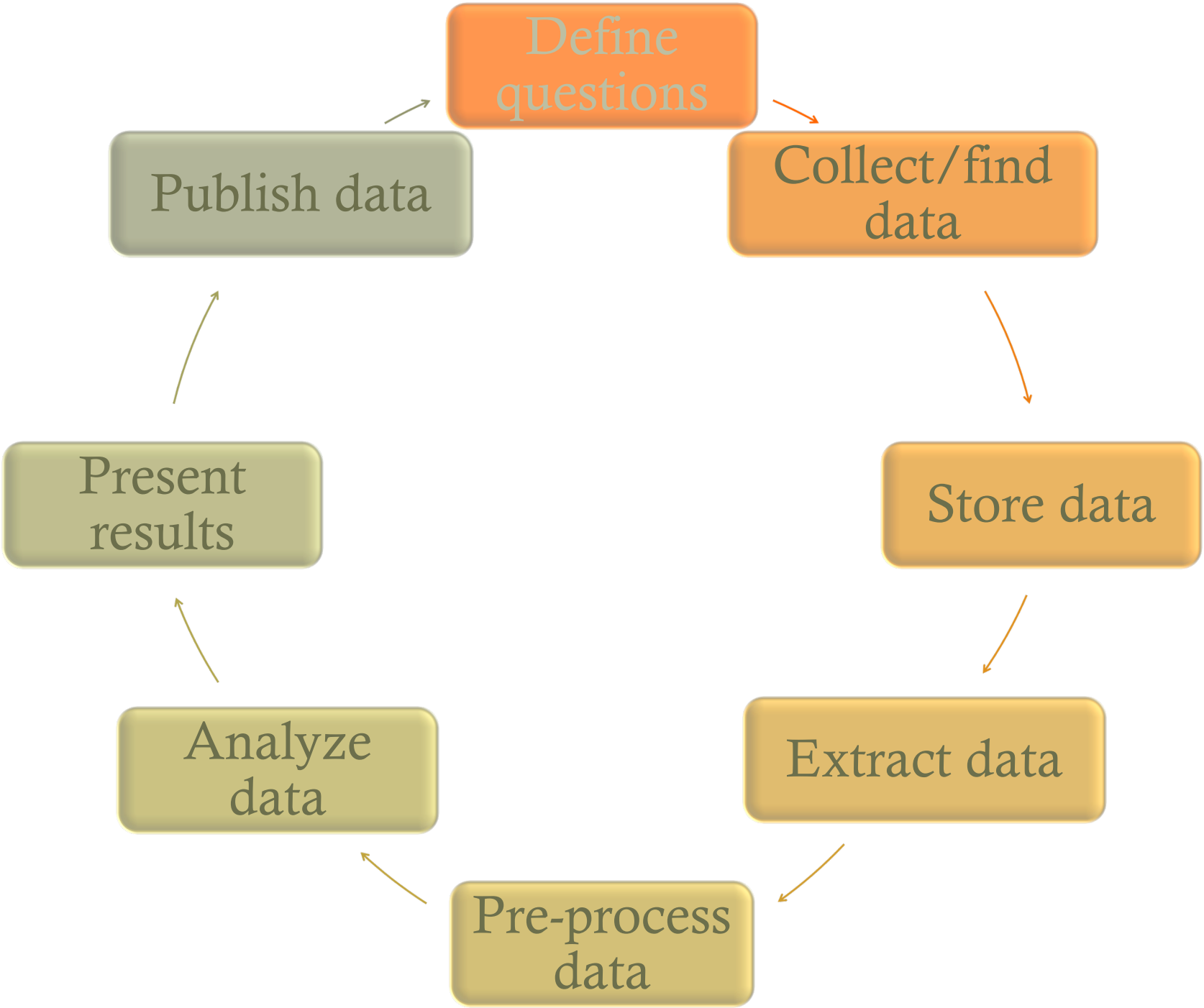
identifying values **sensitive attribute**

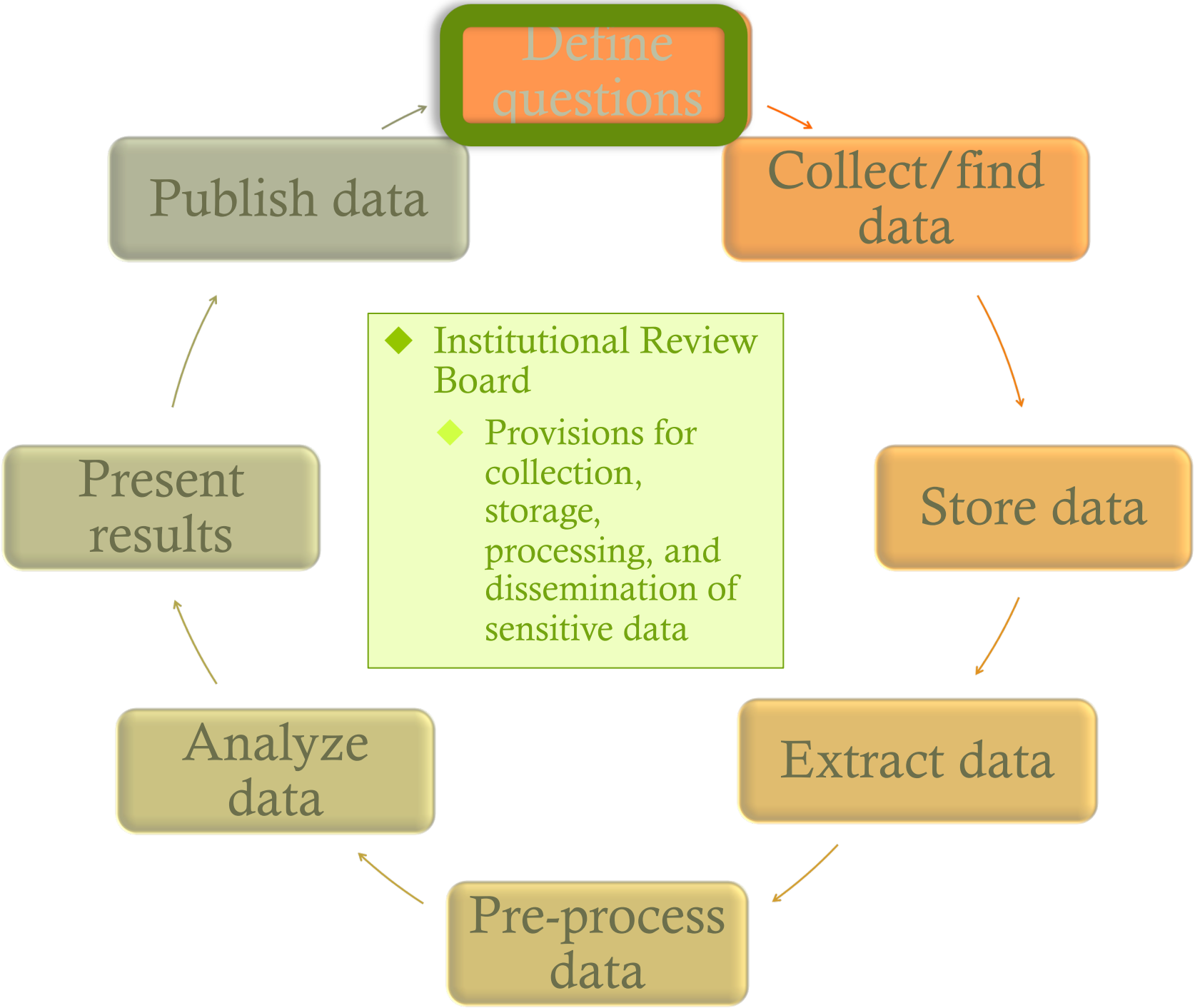
	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

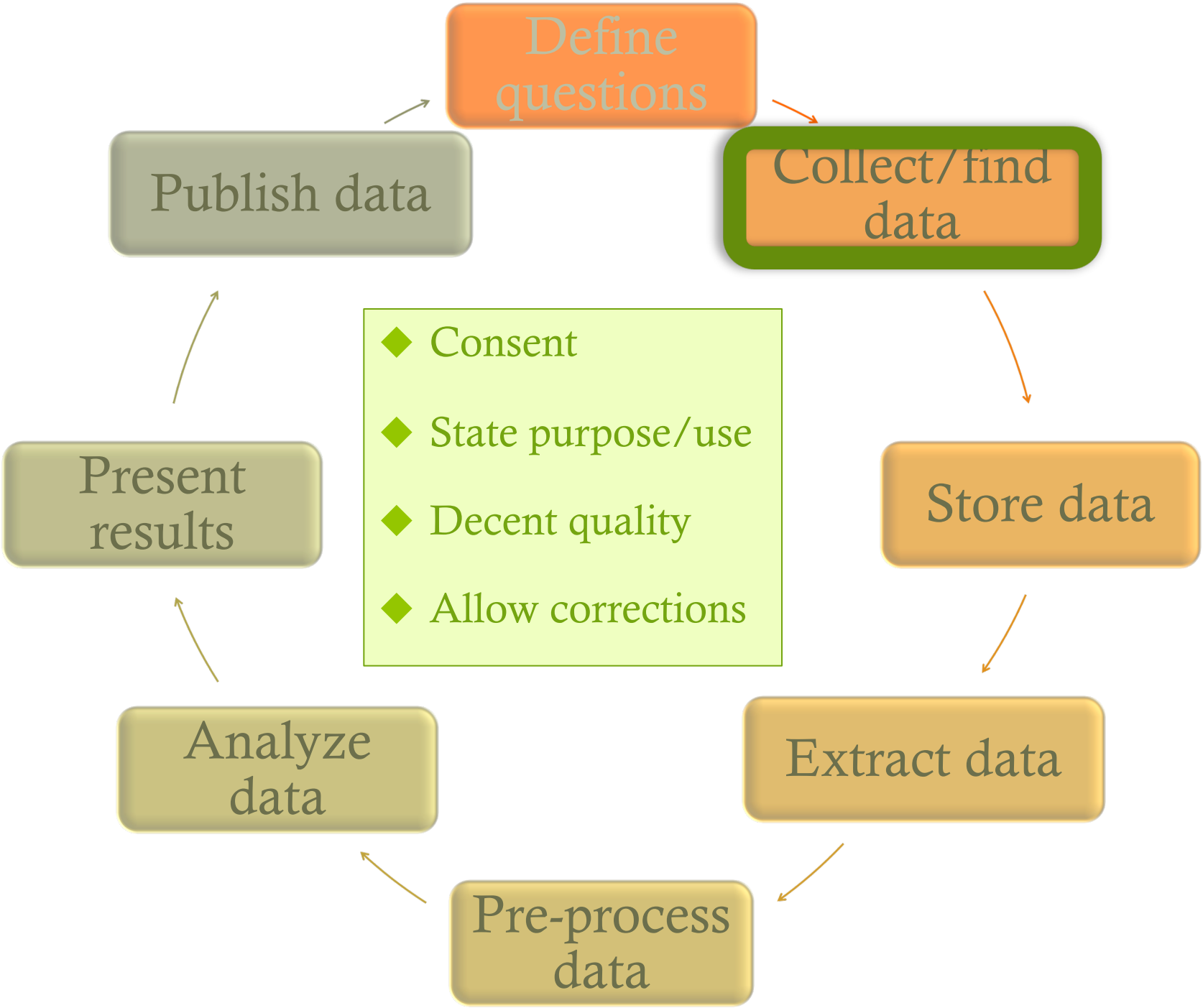
OECD's Eight Principles of Fair Information Practices

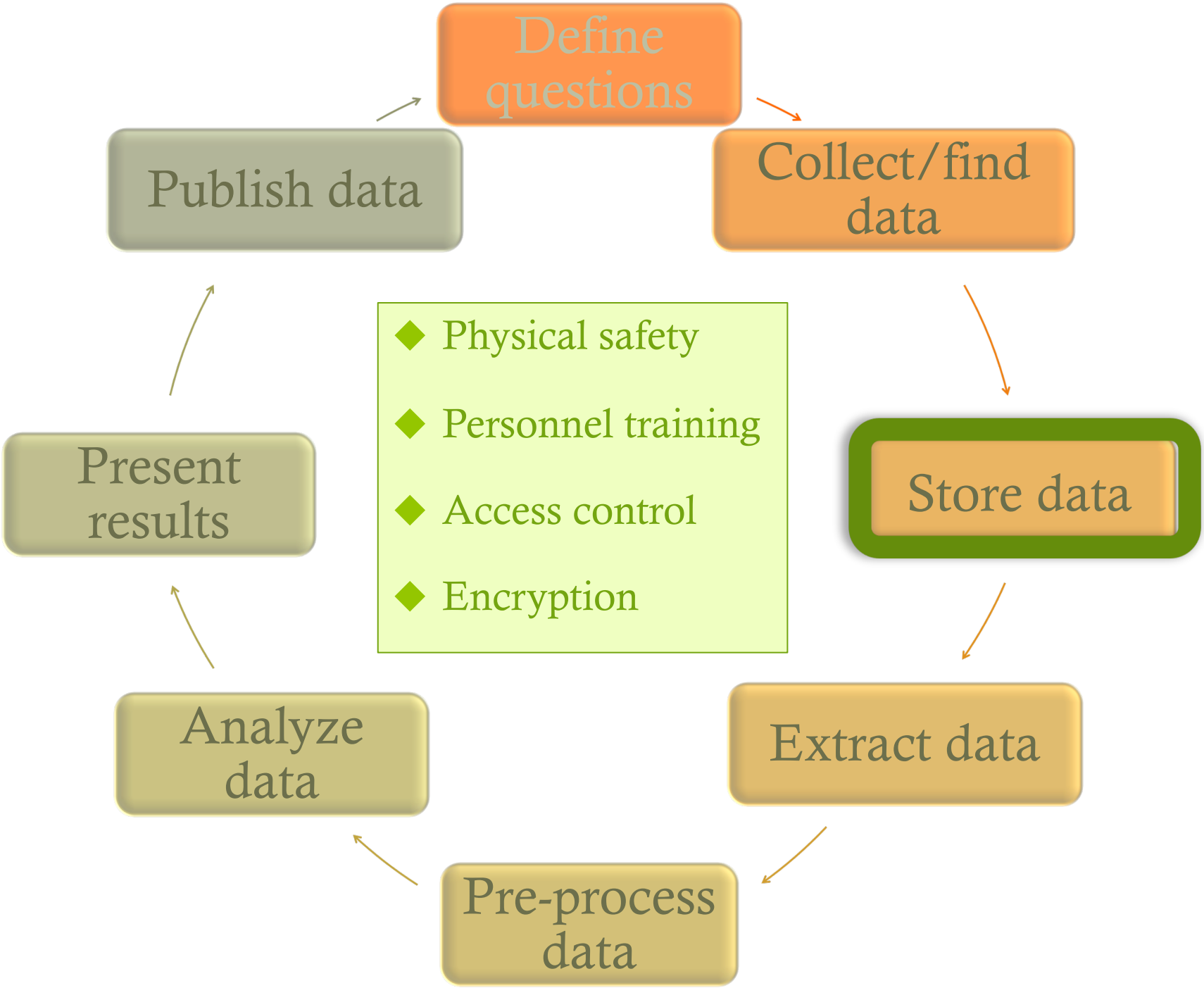
- A framework for privacy protection
- Protect use
  - Collection for a purpose
  - Use only for authorized purpose
  - Accountability throughout these principles

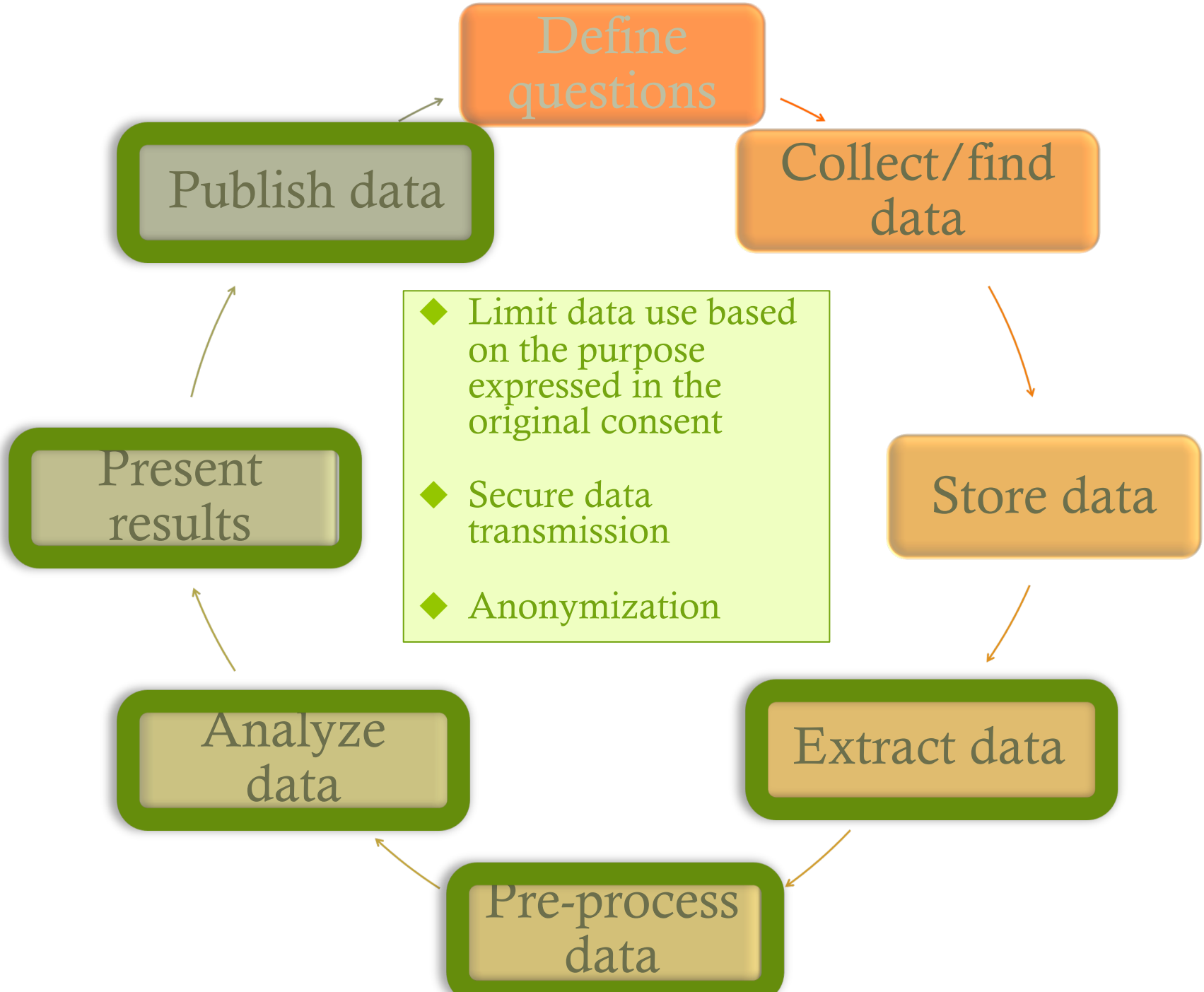
Principle	Description
Collection limitation	The collection of personal information should be limited, should be obtained by lawful and fair means, and, where appropriate, with the knowledge or consent of the individual.
Data quality	Personal information should be relevant to the purpose for which it is collected, and should be accurate, complete, and current as needed for that purpose.
Purpose specification	The purposes for the collection of personal information should be disclosed before collection and upon any change to that purpose, and its use should be limited to the purposes and compatible purposes.
Use limitation	Personal information should not be disclosed or otherwise used for other than a specified purpose without consent of the individual or legal authority.
Security safeguards	Personal information should be protected with reasonable security safeguards against risks such as loss or unauthorized access, destruction, use, modification, or disclosure.
Openness	The public should be informed about privacy policies and practices, and individuals should have ready means of learning about the use of personal information.
Individual participation	Individuals should have the following rights: to know about the collection of personal information, to access that information, to request correction, and to challenge the denial of those rights.
Accountability	Individuals controlling the collection and use of personal information should be accountable for taking steps to ensure the implementation of these principles.


Define questions
Collect/find data
Store data
Extract data
Pre-process data
Analyze data
Present results
Publish data


Define questions
Collect/find data
Store data
Extract data
Pre-process data
Analyze data
Present results
Publish data
Institutional Review Board
Provisions for collection, storage, processing, and dissemination of sensitive data


Define questions
Collect/find data
Store data
Extract data
Pre-process data
Analyze data
Present results
Publish data
Consent
State purpose/use
Decent quality
Allow corrections


Define questions
Collect/find data
Store data
Extract data
Pre-process data
Analyze data
Present results
Publish data
Physical safety
Personnel training
Access control
Encryption


Define questions
Collect/find data
Store data
Extract data
Pre-process data
Analyze data
Present results
Publish data
Limit data use based on the purpose expressed in the original consent
Secure data transmission
Anonymization

Anonymization Techniques

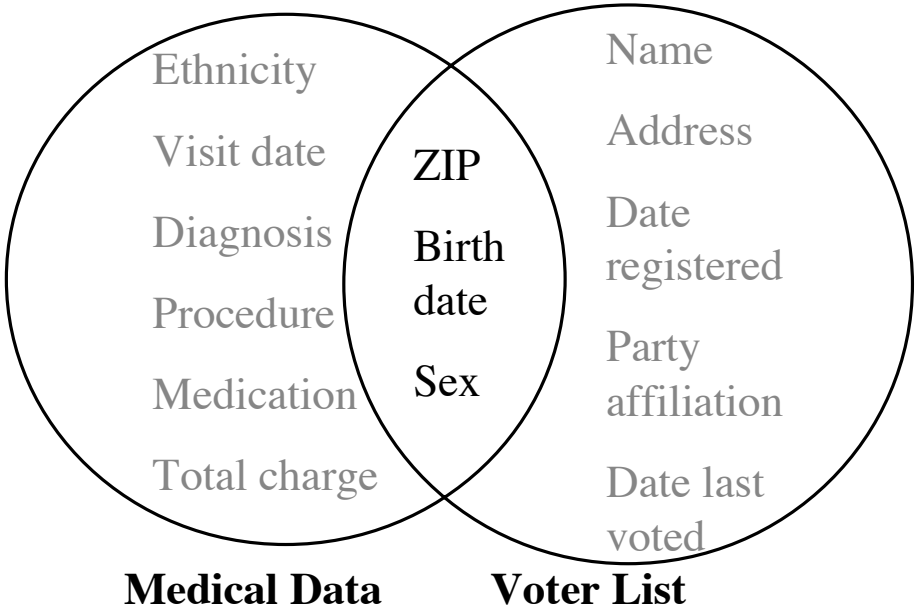
- Replace identifiers with randomly-generated identifiers
  - Eg: “Jane Krakowski” -> “Patient6479”
- Abstraction: Replace values by ranges
  - Eg: Check-in date: 3/1/16 -> Check-in date: Spring 2016
  - Eg: Replace zip code by state
- Cluster data points and replace individuals by their cluster centroid
  - Eg Ages: 21, 25, 28, 27, 18 -> 5 individuals with nominal age of 24
- Remove values
  - Eg: Omit birth date

Problems with Anonymization Techniques

- Limited use for research
  - Too coarse-grained
- Re-identification
  - Re-identification is often trivial
    - E.g., anomymized list of students admitted showing undergraduate university and average GPA
  - Re-identification is possible with high certainty in many cases
    - By linking the anonymized dataset with other public data that is not anonymized

Examples of Re-Identification through Linking Data: (I) Medical Records

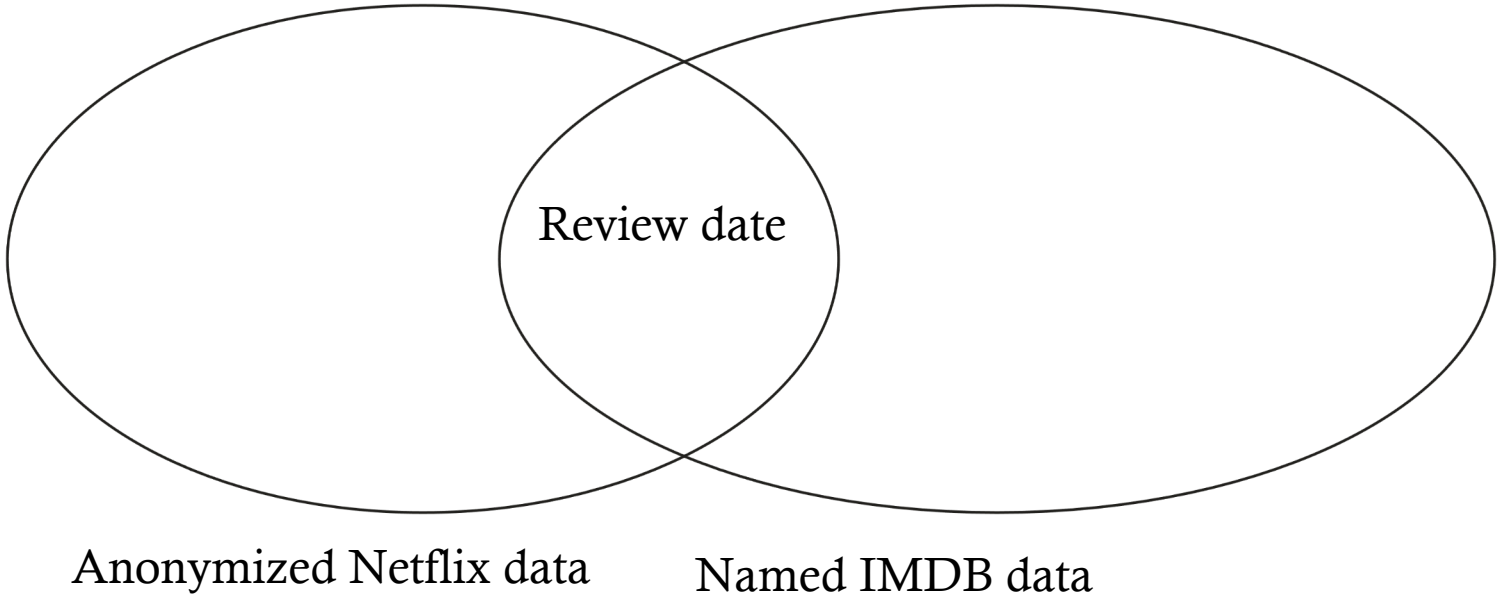
- 87% of the population can be uniquely identified based solely on birthdate, sex, and zip code
  - Most datasets even if anonymized contain this information



http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html

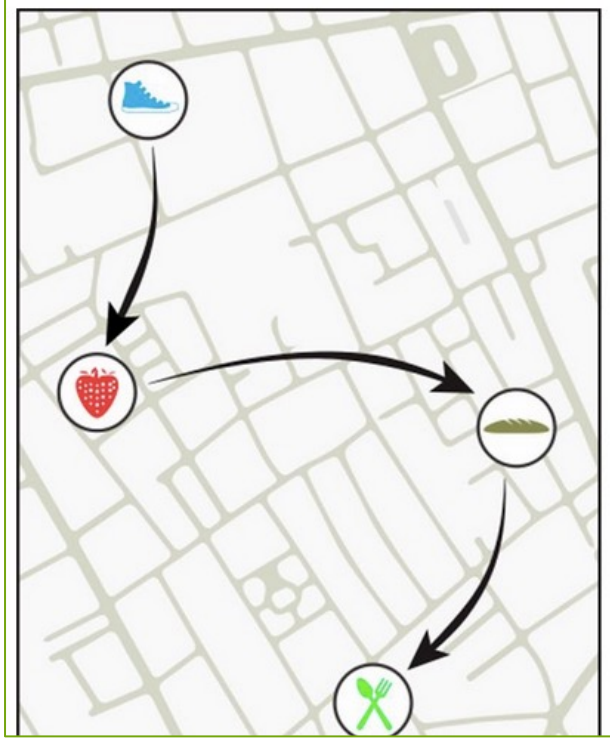
Examples of Re-Identification through Linking Data: (II) Opinions

- Published anonymized data about reviews
  - Public dataset contained reviews that were not anonymized and could be mapped based on the date



Examples of Re-Identification through Linking Data: (III) Behavior Patterns

- Four spatiotemporal points are enough to uniquely re-identify 90% of individuals
- Even data sets that provide coarse information for all dimensions provide little anonymity



shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5

http://science.sciencemag.org/content/347/6221/536.full

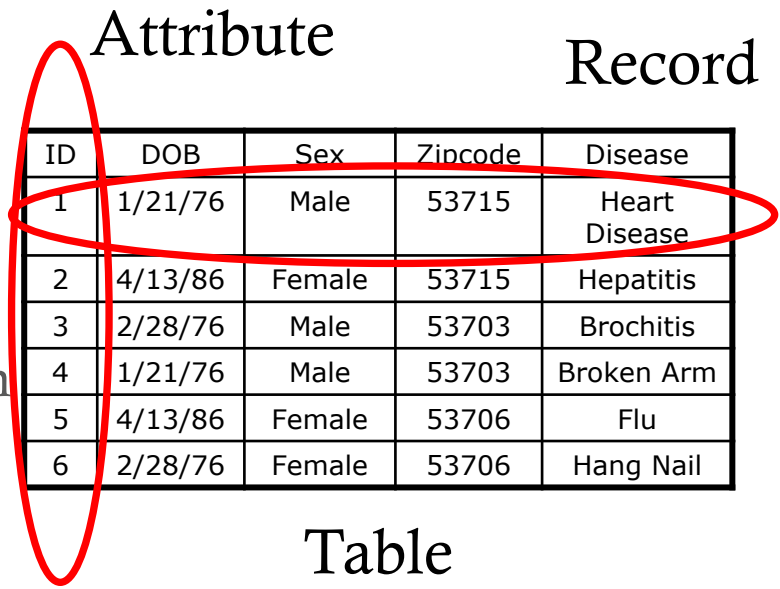
Addressing the Problems of Simple Anonymization Techniques

- Provide guarantees that re-identification will not be possible within some bounds
  - Eg: can only map a given individual to a set of 50 individuals

1. k-anonymization
2. l-diversity
3. t-closeness
4. Differential privacy

Terminologies

- **Key Attribute:** uniquely identifies an individual directly
  - Name, Address, Cell Phone
  - Always removed before release!
- **Quasi-Identifier:** A set of attributes that can be potentially linked with external information to re-identify entities
  - ZIP code, Birth date, gender
  - Can be removed, but utility will degrade
- **Sensitive Attribute:** A set of attributes that need to be released to researchers but raises privacy concerns
  - Medical record, wage, etc.



ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	53715	Heart Disease
2	4/13/86	Female	53715	Hepatitis
3	2/28/76	Male	53703	Brochitis
4	1/21/76	Male	53703	Broken Arm
5	4/13/86	Female	53706	Flu
6	2/28/76	Female	53706	Hang Nail

Privacy Requirements in Data Sharing

ID	Job	Sex	Age	Disease
1	Engineer	Male	35	Hepatitis
2	Engineer	Male	38	HIV
3	Lawyer	Male	38	Flu
4	Writer	Female	30	Flu
5	Writer	Female	31	Hepatitis
6	Actor	Female	31	Hepatitis
7	Actor	Female	32	HIV

- Objective is to preventing the following disclosures

1. **Membership disclosure:** link a particular individual to a table (E.g. Bob is an engineer -> he is sick)

2. **Identification disclosure:** link an individual to a particular record (E.g. Alice is 30 year old -> writer)

3. **Attribute disclosure:** undiscover a new (sensitive) attribute about an individual
(E.g. Writer Alice is 30 year old -> flu)

Addressing Anomymization Problems:

k-Anonymity

- A dataset has k-anonymity if at least k individuals share the same identifying values

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

k=2

K-Anonymity

- A table is "k-anonymized" if each record cannot be identified from other k-1 records when only "quasi-identifiers" are considered
- Alice who is
  - A writer
  - 30 years old
  - How many records can we link her to?

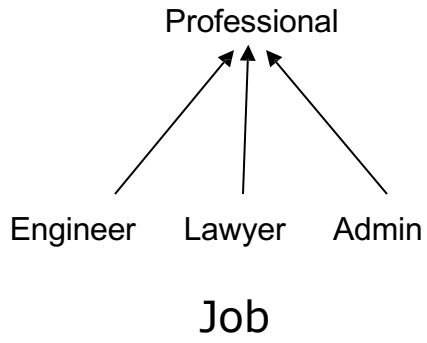
	ID	Job	Sex	Age	Disease
EC1	1	Professional	Male	[35-40)	Hepatitis
	2	Professional	Male	[35-40)	HIV
	3	Professional	Male	[35-40)	Flu
EC2	4	Artist	*	[30-35)	Flu
	5	Artist	*	[30-35)	Hepatitis
	6	Artist	*	[30-35)	Hepatitis
	7	Artist	*	[30-35)	HIV

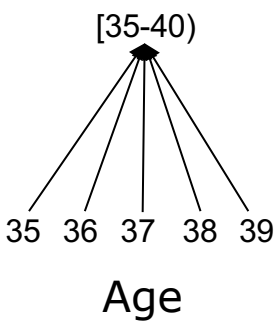
- k-Anonymity ensures that an individual cannot be linked to a record with probability > 1/k

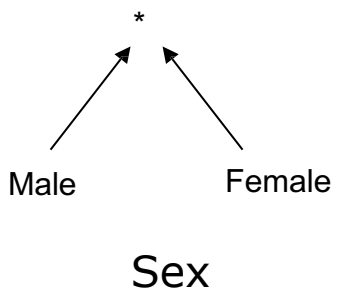
Generalization and Suppression

- Many ways to achieve k-anonymity

1. Generalization
   - Replace with less-specific but semantically-consistent values
2. Suppression
   - Data from certain attribute is removed and replaced with *







Other Anonymization Methods

- Perturbation
  - Add noise/errors
  - For example, change age by ± 2 years
- Data generation
  - Generate similar "fake" data to construct ECs

Example of *k*-Anonymity & Generalization

The Microdata

QID			SA
Zipcode	Age	Gen	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

The Generalized Table

QID			SA
Zipcode	Age	Gen	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- 3-Anonymous table
  - The adversary knows Alice's QI values (47677, 29, F)
  - The adversary does not know which one of the first 3 records corresponds to Alice's record.

Attacks on *k*-Anonymity

- k-anonymity does not provide privacy if:
  - Sensitive values lack diversity
  - The attacker has background knowledge

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl does not have heart disease

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Addressing Anomymization Problems: l-Diversity

- A dataset has l-diversity if the individuals that share the same identifying values have at least l distinct values for the sensitive attribute

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

l=1

The Similarity Attack on *l*-Diversity

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-diversity does not consider semantic meanings of sensitive values

Addressing Anomymization Problems: t-Closeness

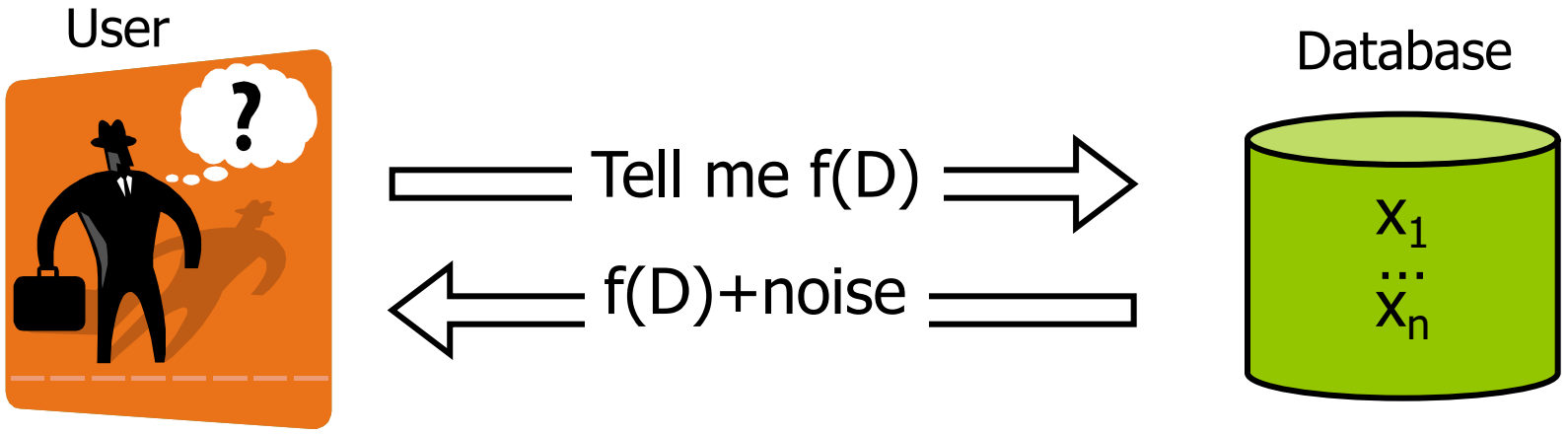
- A dataset has t-closeness if the individuals that share the same identifying values have values for the sensitive attribute that are within a threshold t of diversity

  - Threshold is mathematically defined for the data

Differential Privacy

- **Only method that provides mathematical guarantees of anonymity**
- Main problem addressed: Taking an individual i off a dataset reveals their sensitive attribute information
  - Eg: retrieving aggregate data before removal, then retrieving aggregate data after removal, and then comparing the difference will give us the sensitive attribute of i
- Main idea: Differential privacy adds "noise" to the retrieval process so that such comparisons do not give us the actual sensitive attribute information
  - "noise" is mathematically defined for the data

One Primitive to Satisfy Differential Privacy: Add Noise to Output



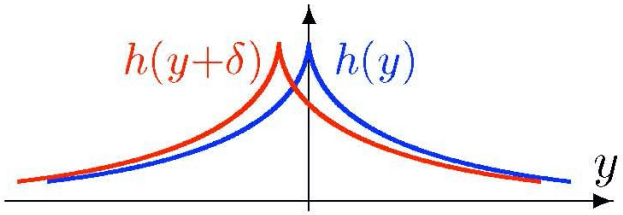
- Intuition: f(D) can be released accurately when f is insensitive to individual entries $x_1, \ldots x_n$
- Global sensitivity $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$
  - Example: $GS_{average} = 1/n$ for sets of numbers between 0 and 1
- Theorem: $f(x) + Lap(GS_f / \varepsilon)$ is ε-indistinguishable
  - Noise generated from Laplace distribution

Sensitivity with Laplace Noise

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$



Sliding property of $\mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

$\delta = f(x) - f(x') \leq \mathsf{GS}_f$

Summary: Threats to Privacy

- Privacy requirements are not well articulated
  - People want benefits in exchange for data
- Unclear that we are able to limit collection and publication
  - Unique behavior of people (we don't read legal contracts)
  - Human error, not without consequences
- Mounds of sensitive data about individuals is readily available in the open web
  - Open web already contains sensitive information that should not be available and violates privacy acts
  - Lots of commercial data with personal information is for sale
- Limited understanding of anonymization and other privacy technologies
  - Linking to public datasets leads to re-identify individuals

Societal Value of Data and Data Science

Granting Access to Private Records: Health Information

- Anonymized information is often not useful for research
  - Too coarse grained
- Private information has great value
  - Tradeoff with quality of treatment
  - Incentivized through first access to new treatments
  - Altruism
- Giving up privacy for pre-specified ***uses***
  - Eg: for specific medical study, not for insurance purposes, not for employers, not for social studies

There is zero privacy anyway, get over it

Although you can upload your data using a pseudonym, there is no way to anonymously submit data. Statistically speaking it is really unlikely that your medical and genetic information matches that of someone else. By uploading you do not only disclose information about yourself, but also about your next kinship (parents and siblings), that shares half of a genome with you. Before uploading any genetic data you should make sure that those people approve of you doing so. This is especially important if you have monozygotic twin, who shares all of your genome!

Privacy

1. Privacy
   - Fair Information Practices
   - Managing sensitive data
   - Anonymizing sensitive data
   - Re-identifying datasets
2. Societal value of data and data science