Decoding De-Identification for Public Agencies

Office of Privacy and Data Protection

Aug. 27, 2020

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Overview

Why de-identify?

De-identification and risk

Where to start

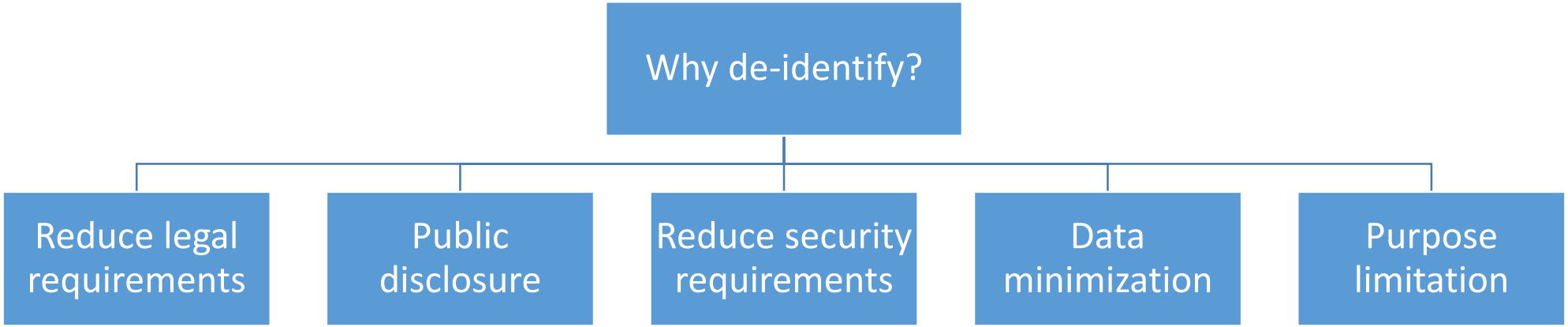
What next

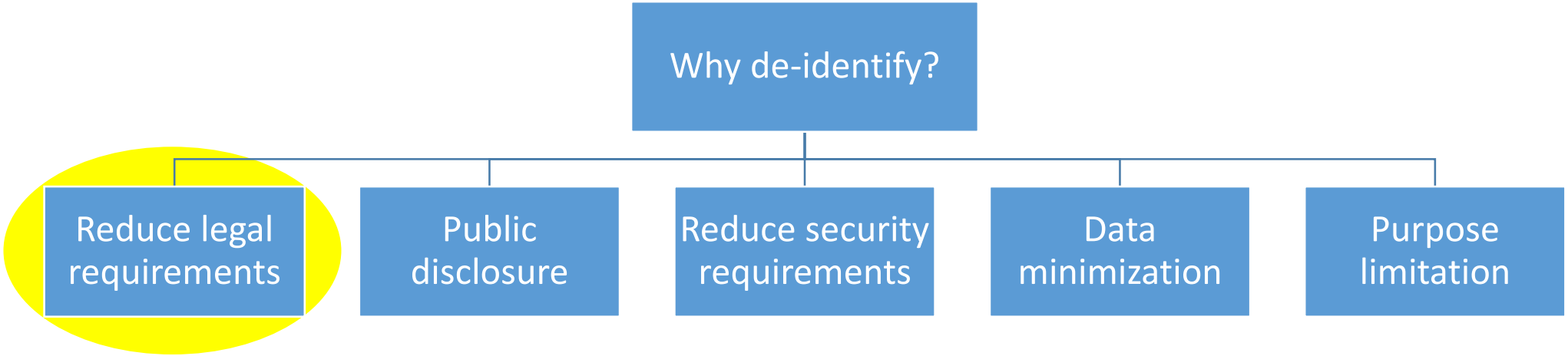
Aggregate reporting

Why de-identify?

De-identification

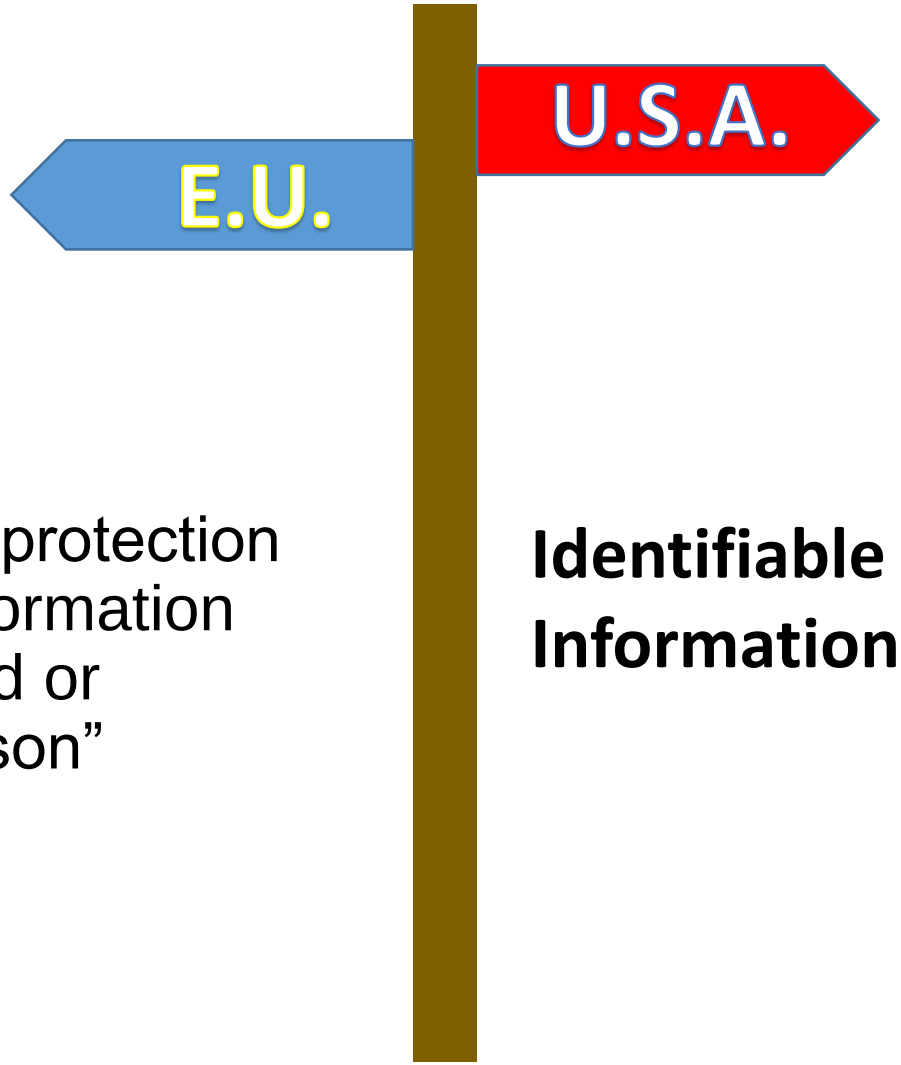
A process to reduce the likelihood of a person's identity being revealed by removing or hiding identifiable information


Why de-identify?
Reduce legal requirements
Public disclosure
Reduce security requirements
Data minimization
Purpose limitation



Most privacy laws only apply to *identifiable* information

Scope of privacy laws

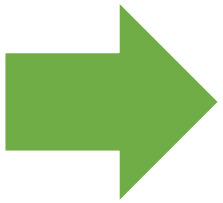


GDPR Recital 26

"The principles of data protection should apply to any information concerning an identified or identifiable natural person"

Identifiable Information

If only identifiable information is protected

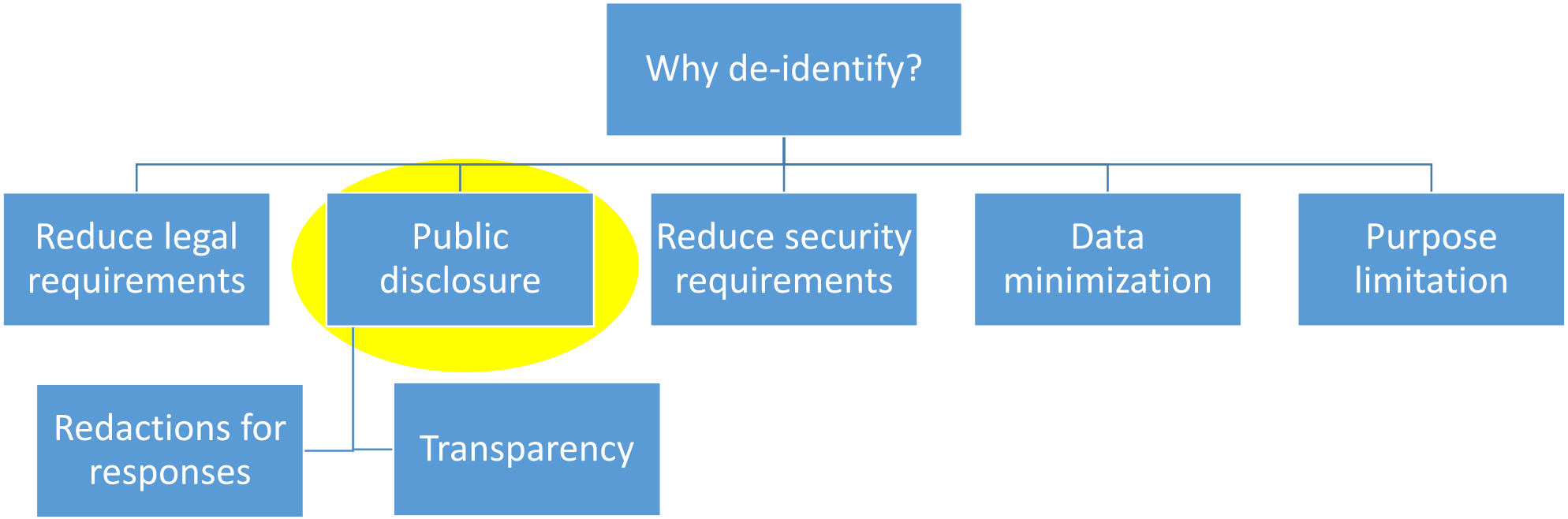


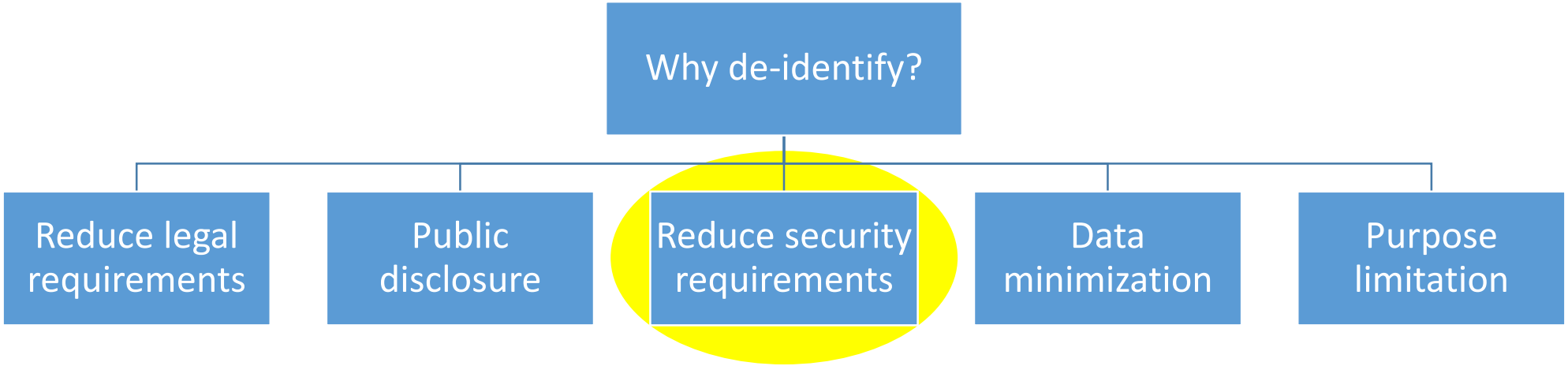
then de-identified information is not

GDPR – "The principles of data protection should therefore not apply to . . . information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous"

HIPAA – "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information"

GLBA – "Personally identifiable financial information does not include . . . information that does not identify a consumer, such as aggregate information or blind data that does not contain personal identifiers"


Why de-identify?
Reduce legal requirements
Public disclosure
Reduce security requirements
Data minimization
Purpose limitation
Redactions for responses
Transparency


Why de-identify?
Reduce legal requirements
Public disclosure
Reduce security requirements
Data minimization
Purpose limitation

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Subject to public disclosure

Not subject to public disclosure

Category 1 – Public Information

. . . information that can be or currently is released to the public. It does not need protection from unauthorized disclosure, but does need integrity and availability protection controls.

Category 2 – Sensitive Information

. . . may not be specifically protected from disclosure by law and is for official use only. Sensitive information is generally not released to the public unless specifically requested.

Category 3 – Confidential Information

. . . information that is specifically protected from either release or disclosure by law . . .

Category 4 – Confidential Information Requiring Special Handling

. . . information that is specifically protected . . . and for which [there are especially strict requirements and serious consequences could come from improper disclosure]

Subject to public disclosure

Not subject to public disclosure

Category 1 – Public Information

. . . information that can be or currently is released to the public. It does not need protection from unauthorized disclosure, but does need integrity and availability protection controls.

Category 2 – Sensitive Information

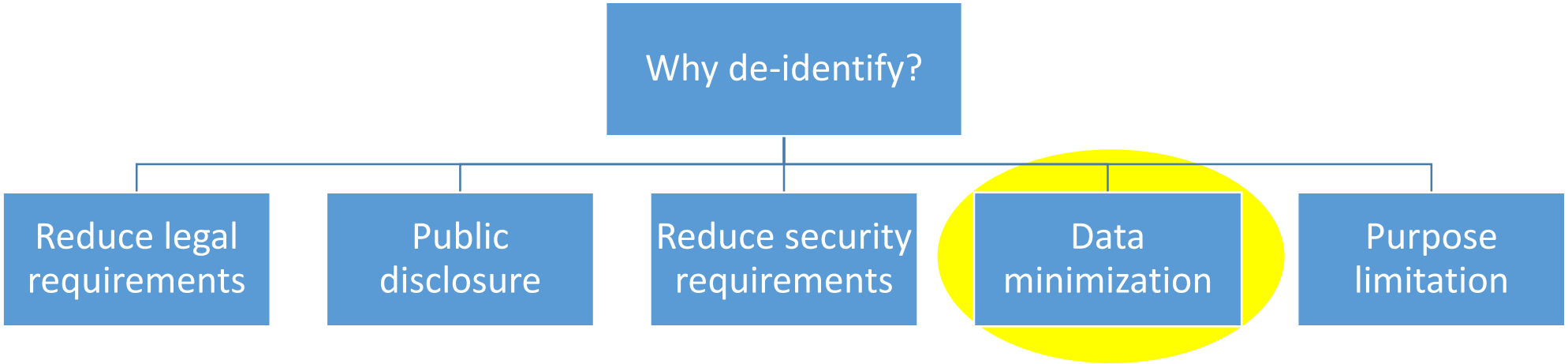
. . . may not be specifically protected from disclosure by law and is for official use only. Sensitive information is generally not released to the public unless specifically requested.

Category 3 – Confidential Information

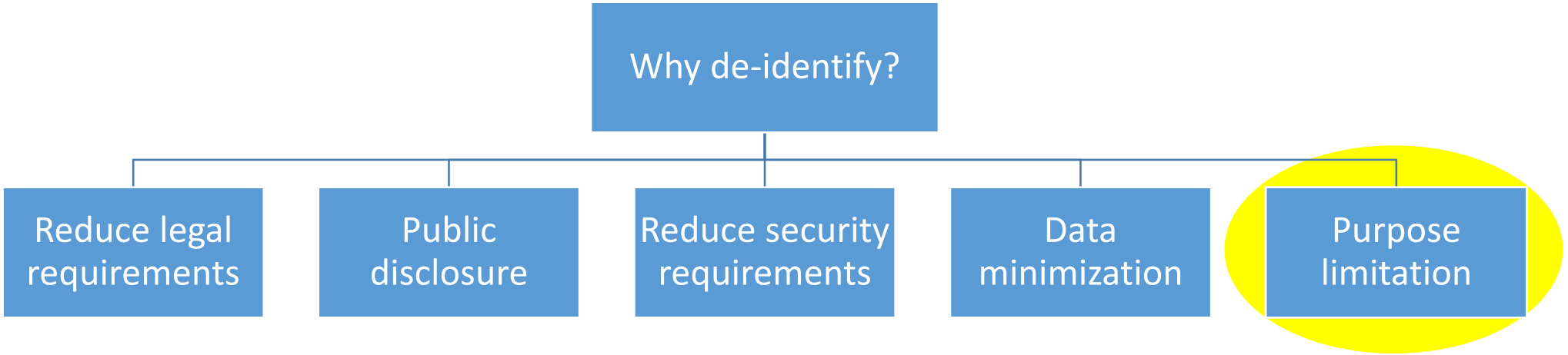
. . . information that is specifically protected from either release or disclosure by law . . .

Category 4 – Confidential Information Requiring Special Handling

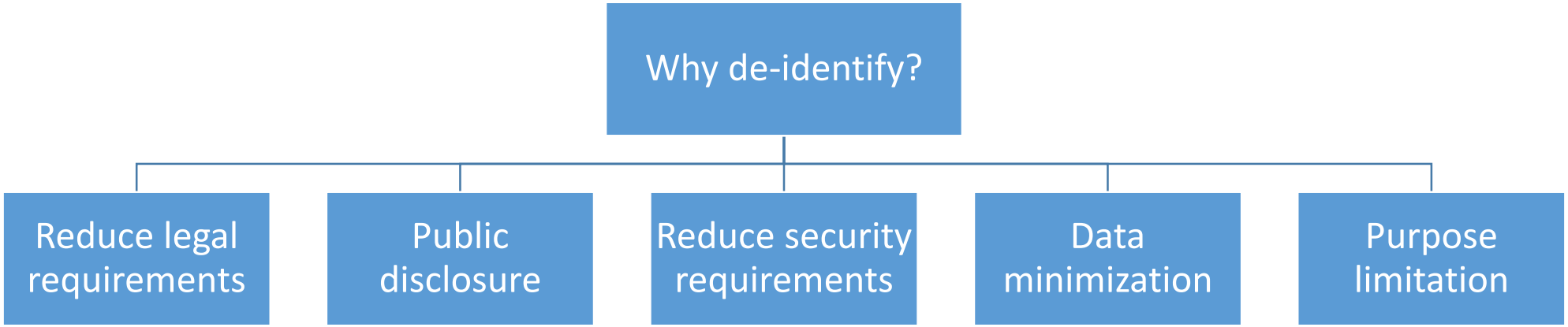
. . . information that is specifically protected . . . and for which [there are especially strict requirements and serious consequences could come from improper disclosure]



The data minimization principle applies across the information lifecycle, including collection, *use*, and *disclosure*



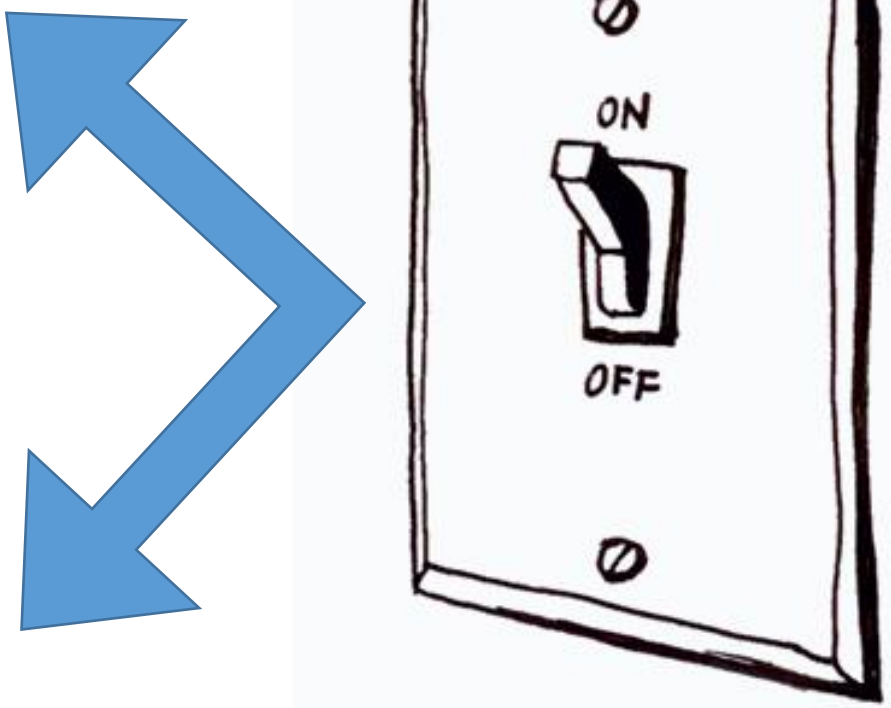
De-identification *may* allow information to be used for purposes that would otherwise be incompatible with the original reason the information was collected



- These reasons are not exhaustive
- When working on de-identification issues, consider why you're de-identifying to help guide decisions and standards

De-identification and risk

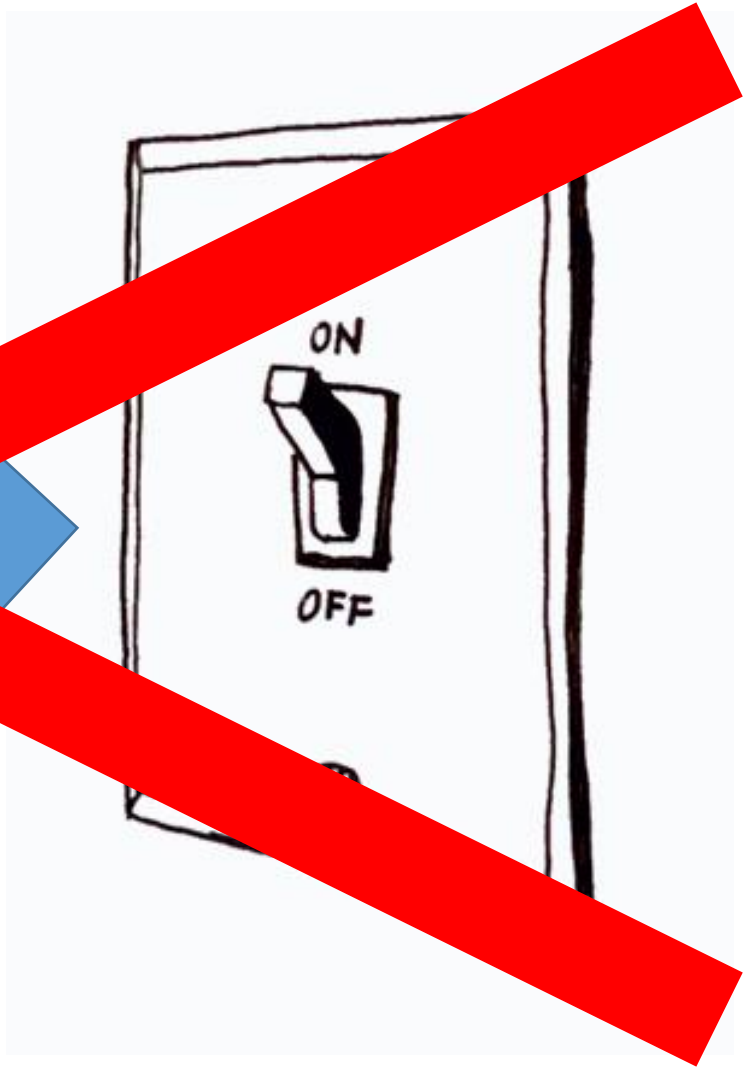
Identifiable information



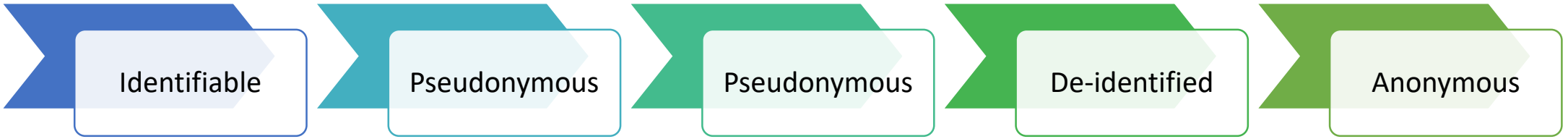
De-identified information

le information

tified information



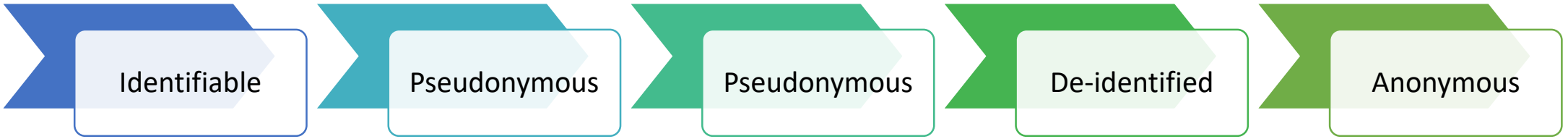
Spectrum of identifiability



Spectrum of identifiability

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Glossary Break



Different laws use different terms for protected information.
For example:

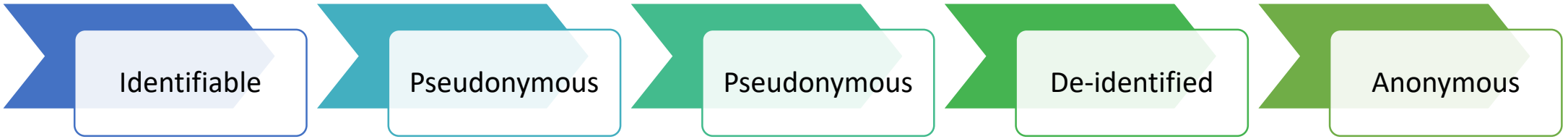
- Personal information
- Personally identifiable information
- Individually identifiable information
- Information that can be readily associated with an individual

For today, assume identifiable information is a broader term that includes these variations

Spectrum of identifiability

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Glossary Break



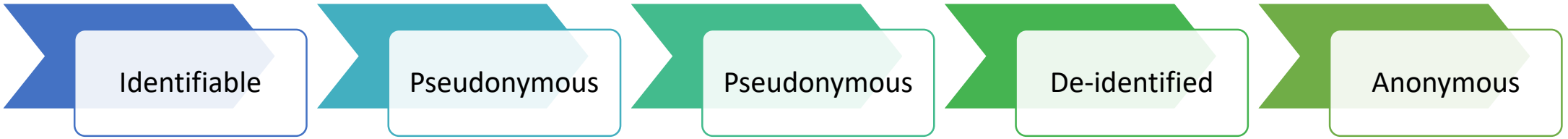
Information that allows re-identification when combined with other information. Could include:

- Data with a unique system or research ID assigned to individuals
- Data with indirect identifiers that allow re-identification

Spectrum of identifiability

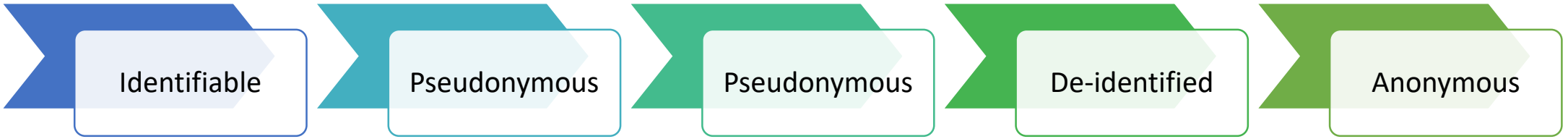
WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Glossary Break



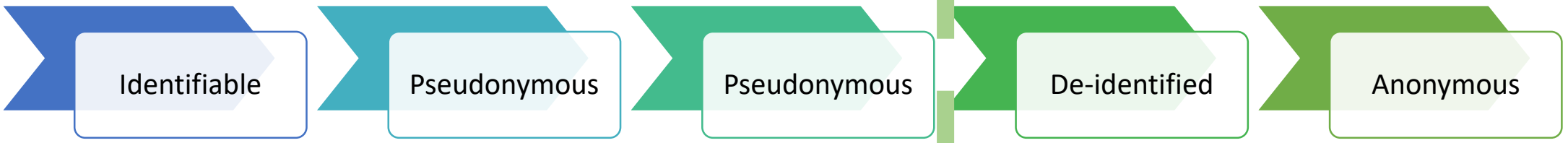
There's room for debate; definitions and interpretations vary.
For today, assume "de-identified," "anonymous," and "nonpersonal" can be used interchangeably.

Spectrum of identifiability



Spectrum of identifiability

Risk of re-identification ≠ 0



Risk-based determination

Where to start

Where to start – identify applicable laws

- Understand where the protections for your information come from, even if the laws or regulations provide little guidance about what is identifiable
- Don’t be afraid to look to non-applicable laws for guidance (just don’t rely on them)

Where to start – identify applicable laws

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Gramm-Leach-Bliley Act (GLBA)

Personally identifiable financial information does not include:

Information that does not identify a consumer, such as aggregate information or blind data that does not contain personal identifiers such as account numbers, names, or addresses.

Where to start – identify applicable laws

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Family Educational Rights and Privacy Act (FERPA)

[Personally identifiable information] includes, but is not limited to –

(a) The student's name;

(b) The name of the student's parent or other family members;

(c) The address of the student or student's family

(d) A personal identifier, such as the student's social security number, student number, or biometric record;

(e) Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;

(f) Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; or

(g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.

Where to start – identify applicable laws

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Family Educational Rights and Privacy Act (FERPA)

[Personally identifiable information] includes, but is not limited to –

(a) The student's name;

(b) The name of the student's parent or other family members;

(c) The address of the student or student's family

(d) A personal identifier, such as the student's social security number, student number, or biometric record;

(e) Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;

(f) Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; or

(g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.

Where to start – identify applicable laws

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Washington State Breach Notification Law (RCW 42.56.590)

First name or first initial and last name in combination with any one or more of the following elements:

1. SSN;
2. Driver's license # or WA ID #;
3. Account number, credit or debit card number, or any other security code, access code, or password that would permit access to an account;
4. Full date of birth;
5. Private key … that is used to authenticate or sign an electronic record;
6. Student, military, or passport ID #
7. Health insurance policy # or health insurance ID #;
8. Health information;
9. Biometric data (see Chapter 19.375 RCW and Chapter 40.26 RCW);
10. User name or email address in combination with a password or security answers that permits account access*****
11. Any combination of the above elements that would enable identity theft*

*** Does not require first name/initial or last name combination**

Where to start – identify applicable laws

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Health Insurance Portability and Accountability Act (HIPAA)

The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names
(B) All geographic subdivisions smaller than a state [except the first three digits of ZIP]
(C) All elements of dates (except year) for dates that are directly related to an individual
(D) Telephone numbers
(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers
(M) Device identifiers and serial numbers
(F) Email addresses
(N) Web Universal Resource Locators (URLs)
(G) Social security numbers
(O) Internet Protocol (IP) addresses
(H) Medical record numbers
(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers
(Q) Full-face photographs and any comparable images
(J) Account numbers
(R) Any other unique identifying number, characteristic, or code [except a unique code not shared with others]; and
(K) Certificate/license numbers

AND

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Where to start – recognize types of identifiers

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

"Direct identifiers" and "indirect identifiers" are commonly used

- Imprecise terms with no universal definitions

Direct identifier = information uniquely associated with an individual (1:1)

Indirect identifiers = information not uniquely associated with an individual, but that could be used in combination with other information to identify a person (1:many)

Where to start – recognize types of identifiers

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Direct identifier examples

Name
Street name or street address or post office box
Telephone and fax numbers
Email address
Social security number
Certificate/license numbers
Vehicle identifiers and serial numbers
URLs and IP addresses
Full-face photos and other comparable images
Medical record numbers, health plan beneficiary numbers, and other account numbers
Device identifiers and serial numbers
Biometric identifiers, including finger and voice prints

Where to start – recognize types of identifiers

WaTech
Washington Technology Solutions
Washington's Consolidated Technology Services Agency

Indirect identifier examples

Demographic information (*e.g.*, age, gender, race, ethnicity, languages)
Geographic information (*e.g.*, census tract, zip code)
Facility name
Employment information
Time/date of event (*e.g.*, birth, death, hospital admission or discharge)
Diagnosis or procedure information

Where to start – recognize types of identifiers

Be aware of the possibility for other unique information

- Information that describes physically observable characteristics is more likely to be identifiable
- Be wary of free text in records or data

e.g., A therapist's note that says, "patient was concerned about decreasing volume of gold coins in his money pool"

Where to start – more data, more problems

- The more variables or data points included, the higher the likelihood of re-identification
- Consider location data
  - One data point may reveal nothing about a person
  - But twenty data points a day could reveal who they are, where they live, where they work, where they shop, where their kids go to school, and that they're seeing a psychiatrist

Where to start – consider context/audience

- Who will have access to the information/who is it being shared with?
- Will they sign an agreement that they will not attempt to re-identify people?
- Do they have other information that would allow them to identify people?
- What is the nature of the information? What are the potential consequences if a person is re-identified?

What next

Redaction tools

- De-identification is pointless if the appropriate information is removed using the wrong tools
  - Placing a box over text in word does not remove the content
  - Placing redactions in Adobe without finalizing the redactions does not remove the content
  - Hiding a tab in Excel does not remove the content
  - Metadata may include identifiable information even if a record itself is appropriately redacted
- Ensure staff who release or publish data have access to appropriate tools and know how to use them

Statistical analysis

- It is possible to minimize information loss by using statistical analysis instead of making across the board determinations of which data elements to remove
- HIPAA specifically contemplates that a person with
  - “appropriate knowledge”
  - of “statistical and scientific principles”
  - can determine that “the risk is very small that the information could be used, alone or in combination with other reasonably available information,
  - by an anticipated recipient
  - to identify an individual who is a subject of the information”
- Although not specifically contemplated in other laws, reasonable to assume an expert can help determine whether there is a risk of re-identification

Other de-identification methods

- Completely removing information is *not* the only method to de-identify
- Data masking
  - Original data is obscured to hide sensitive information
  - Particularly useful for setting internal user controls where different staff need access to different levels of information
  - Different from encryption – masking is not reversible, the user should not be able to retrieve the original data
  - *e.g.*, SSN visible for some users, shows up as xxx-xx-xxxx for other users
  - *e.g.*, all variables replaced with fake data to create test data
- Perturbation – adding random noise to original values
- Data swapping – exchange values between different records
- Generalization – replacing precise values with more general values
- Microaggregation – statistical method to report average values for groups of people with similar indirect identifiers, instead of individual values

Business needs

- Different methods of removing data have different impacts on the resulting product
  - Completely removing all identifiers may make data useless for intended purpose
  - Generalizing data allows more data to remain, but makes it difficult to make inferences about small changes or make comparisons (over time or across populations)
  - Perturbation may help maintain statistical accuracy, but transparency is reduced because you have introduced fictional data

Interaction with Public Records Act

- Records maintained by public agencies concerning government functions are public records subject to disclosure under Washington's Public Records Act unless a specific exemption applies
- The Act "shall be liberally construed and its exemptions narrowly construed…"
- An agency follows de-identification best practices when publishing data; but what happens if someone makes a request for more details?
- Potential conflict exists when there are no direct identifiers involved, but an agency believes a requester knows the identity of a person in responsive records
  - Media request for protected incident reports when the name of a person involved has already been published
  - Requests for inmate records that would mean little to the general public but would very likely be identifiable to other inmates

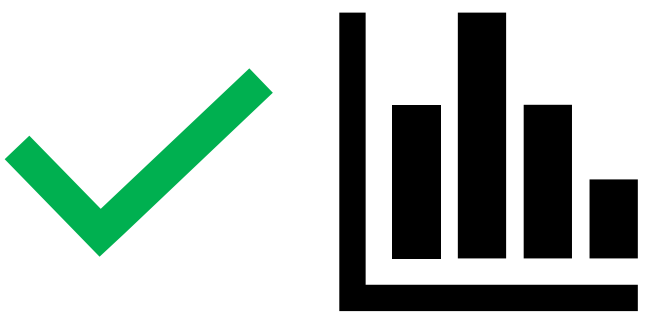
Document your decisions

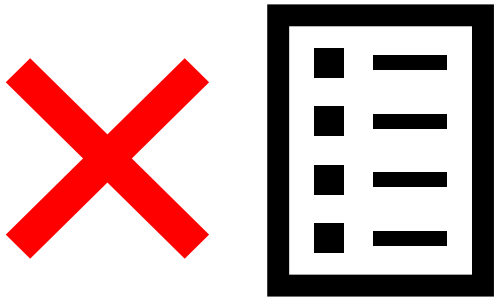
- When you have made agency decisions, document and explain them
- The appropriate level of formality will vary
  - Rules provide the most authority and receive public comment
  - Formal policies help demonstrate the official agency position for both internal and external audiences
  - Informal standards are the most flexible
- The standards do not need to (and probably shouldn't) be absolute
  - Even with overarching guidelines, agencies may need to make project-by-project determinations
- Having documented standards helps:
  - Ensure internal compliance and consistency
  - Explain decisions to external stakeholders such as constituents, legislators, and agency partners
  - Defend actions with courts or regulators

Aggregate Reporting

Aggregate reporting – Releasing summaries or counts instead of individual level information

- When dealing with very sensitive information, aggregate reporting is almost always preferable to individual/record level reporting
- Can be static or dynamic
- Different than data aggregation





Aggregate reporting *does not* completely remove risk

- Risk of re-identification should be one factor when considering granularity of reporting standards (e.g., date ranges, age ranges, geographic areas)
- Risk of re-identification rises when the number of people in an overall population (the denominator) or the number of people with a specific characteristic (the numerator) decrease
  - More variables/cross-tabulations = higher risk of re-identification
- Risk of re-identification rises if the identity of the people in the population is known, or the identity of respondents to a survey are known
- Practical risk of re-identification rises with dynamic reporting

Gender	Took test	Passed test	Percentage
Male	15	1	6.7
Female	20	10	50
Total	35	11	31.4

Gender	Took test	Passed test	Percentage
Male	15	1	6.7
Female	20	10	50
Total	35	11	31.4

Suppressing small numbers helps reduce risk

- *e.g.*, DOH standard typically requires suppressing cells with counts from 1 through 9 (0<n<10), HCA standard typically requires suppressing cells with counts from 1 through 10 (0<n<11)

Gender	Took test	Passed test	Percentage
Male	15	1	6.7
Female	20	10	50
Total	35	11	31.4

Suppressing the small number alone is not sufficient

Took test	Passed Test	Percentage
15	0	0

Suppression standards typically allow a count of 0 to be published

- But on a case-by-case basis counts of all or none may not be appropriate

Took test	Passed Test	Percentage
15	*	<5%

- Assume publisher has taken steps to suppress 0, and implemented top and bottom coding on the percentage.
- Top and bottom coding means establishing an upper and lower boundary for published values. In this case, assume:
  - Any percentage greater than 95% is published as >95%
  - Any percentage lower than 5% is published as <5%
- Because of the small sample size, publishing <5% reveals that the true count is 0.

Aggregate reporting wrap-up

Aggregate reporting is an effective way to provide information to the public while reducing the risk of re-identification

Suppression reduces the risk of re-identification, but has significant drawbacks

- Difficult to report on small population sizes without significant information loss
- Susceptible to being undone using other available data

Consider ways to increase sample size

- *e.g.*, by county instead of city, by school instead of class

Consider ways to blur data to prevent small numbers

- *e.g.*, use age range, roll-up into categories

These methods have different tradeoffs; determine how information will be used before developing reporting standards

Review final product for unexpected results

Takeaways

- Understand what you are trying to accomplish by de-identifying
- Understand how de-identification decisions will impact business needs
- Identify and evaluate applicable laws
  - Laws that don't apply to you may still be instructive
- Accept that de-identification is a risk exercise
- Be creative, consider all available options including whether information can be released in aggregate
- Document decisions and standards
- Continually review (standards and individual products), consider context, adjust as necessary

Questions?