Data Management & De-identification

RT2 Roadmap

Motivating Issues

Researchers degrees of freedom
Scientific misconduct
Publication bias
Failure to replicate

To achieve

Open materials, data, code, & access
Transparent reporting & disclosure
Reproducible & replicable results
Cumulative meta-analyses

Organized Workflow and File Management (OSF, Github)

Design	Conduct	Dissemination	Archiving
Pre-Registration	Data Management	Transparent Reporting & Disclosure	Data Repositories
Pre-Analysis Plans	Version Control	Preprints	Dynamic Documents
Power Planning	Open Notebooks (Jupyter/Docker)	Open Access	

Danae Roumis, Social Impact:

Impact evaluation of public health programs

SOCIAL IMPACT

ADVANCING DEVELOPMENT EFFECTIVENESS

- Monitoring, **evaluation**, strategic planning, capacity building
  - Impact evaluation, surveys & large-scale data collection
- 120+ countries in all regions & sectors
  - Agriculture & food security, democracy & governance, education, energy, environmental protection, health, humanitarian assistance, water & sanitation, youth development
- **USG** & non-USG
  - USAID, Millennium Challenge Corporation, State Dept., USDA, +
  - Foundations, NGOs, DFID, World Bank, +

Session

- Data management in context of open data efforts
- Open data and protecting research subjects
- De-identification process & tips

Motivation

- **Open data** heart of research transparency and reproducibility
- Funders, journals moving in this direction
- Some funders require data publication

Public *Permitted by law and subject to privacy, confidentiality, security*

Accessible *Convenient, usable formats that can be searched and downloaded*

Described *Documentation required for new users to understand how data*

Re-usable *Open license for new users to actually use **would add citable*

Complete *Finest level of granularity for maximize use (i.e. limited aggregation)*

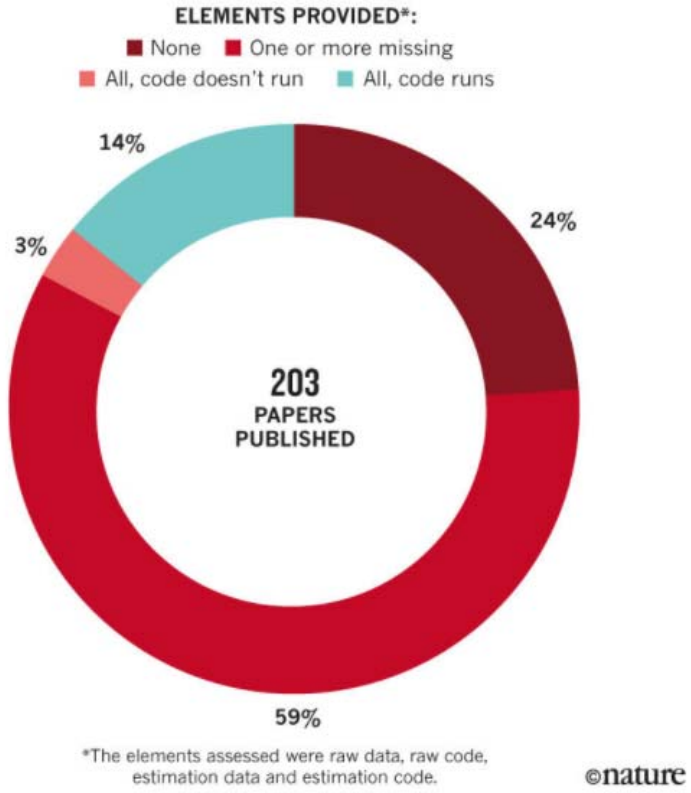
Timely *Preserve value with timely release*

Managed *Point of contact available to support new users, respond to questions*

Source: World Bank Open Data Toolkit

Motivation

- **Open data** heart of research transparency and reproducibility
- Funders, journals moving in this direction
- Some funders require data publication



Source: P. Gertler, S. Galiani & M. Romero (unpublished data)

See: https://www.nature.com/articles/d41586-018-02108-9

Motivation

- **Open data** heart of research transparency and reproducibility
- Funders, journals moving in this direction
- Some funders require data publication

- Benefits: transparency, accountability
- Risks: disclosure of personal information, sensitive data
  - Harm to respondents, distrust for future research

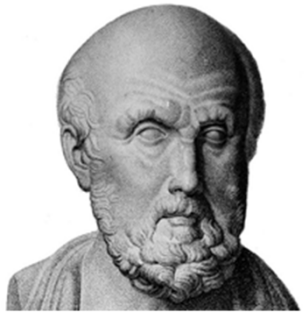
- Reinvigorated debate about **privacy**
  - Breaches, data privacy laws
  - Different fields: data privacy, responsible data, digital safeguarding

Balancing act

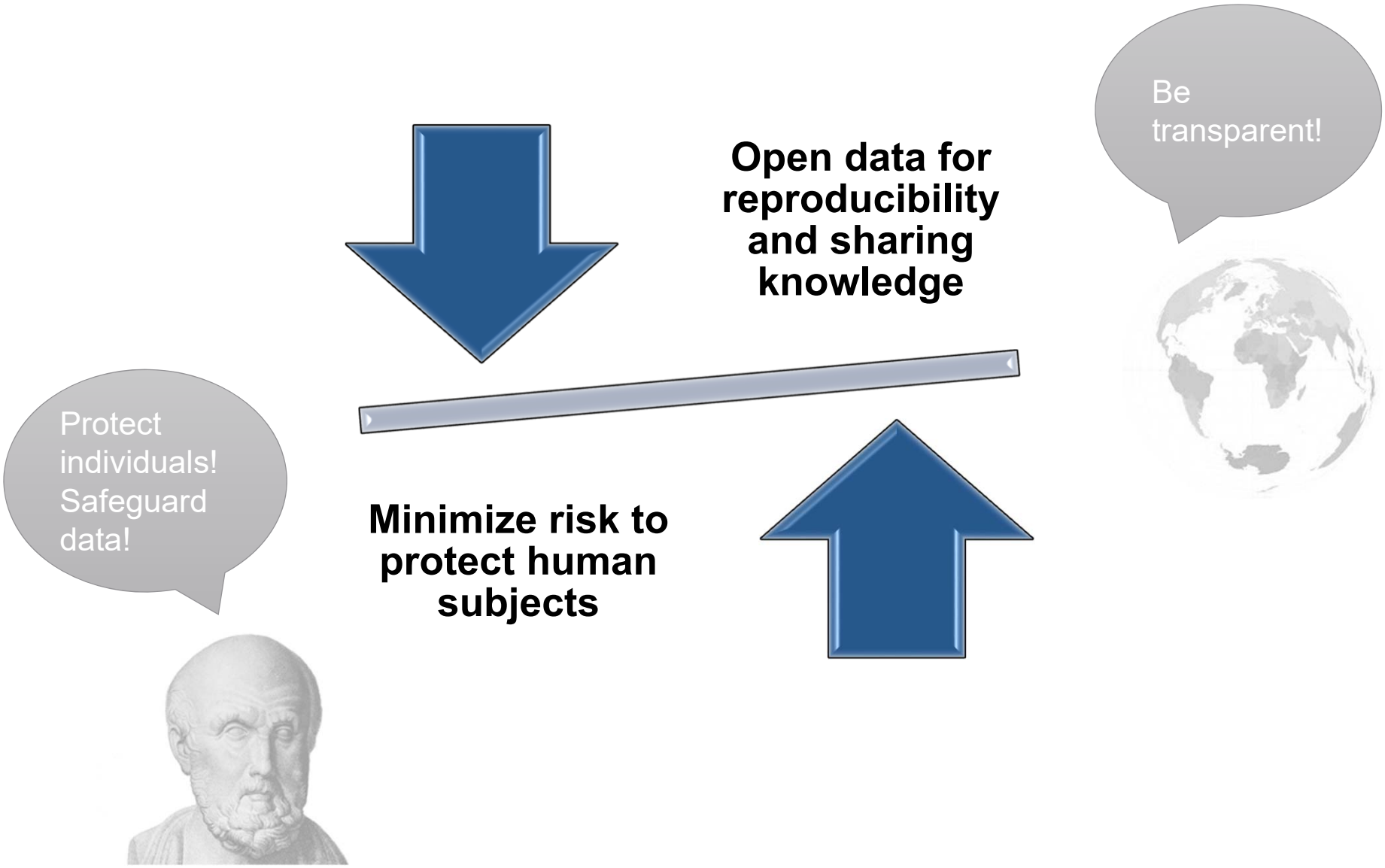
Balancing act

Be transparent!

Protect individuals! Safeguard data!



Balancing act



Secret aid worker: we don't take data protection of vulnerable people seriously

Personal information leaked in sensitive contexts can spark violence, discrimination, exclusionary policies. Yet my NGO shares confidential data freely

The horror stories abound. Highly sensitive data is routinely emailed openly among staffers, without encryption. Personally-identifiable data is stored in the organisation's cloud storage without protocols for who can and cannot access it, and how this data can be used or not used. There are no guidelines as to what data should be collected in the first place, and how to collect it in a secure manner. There is no data anonymisation that would remove personally identifiable information from what's collected. Informed consent protocols, if they exist within specific programmes, are inconsistent across the whole organisation and are not routinely enforced. Much of what should be "confidential" is accessible to all staff and even outside consultants.

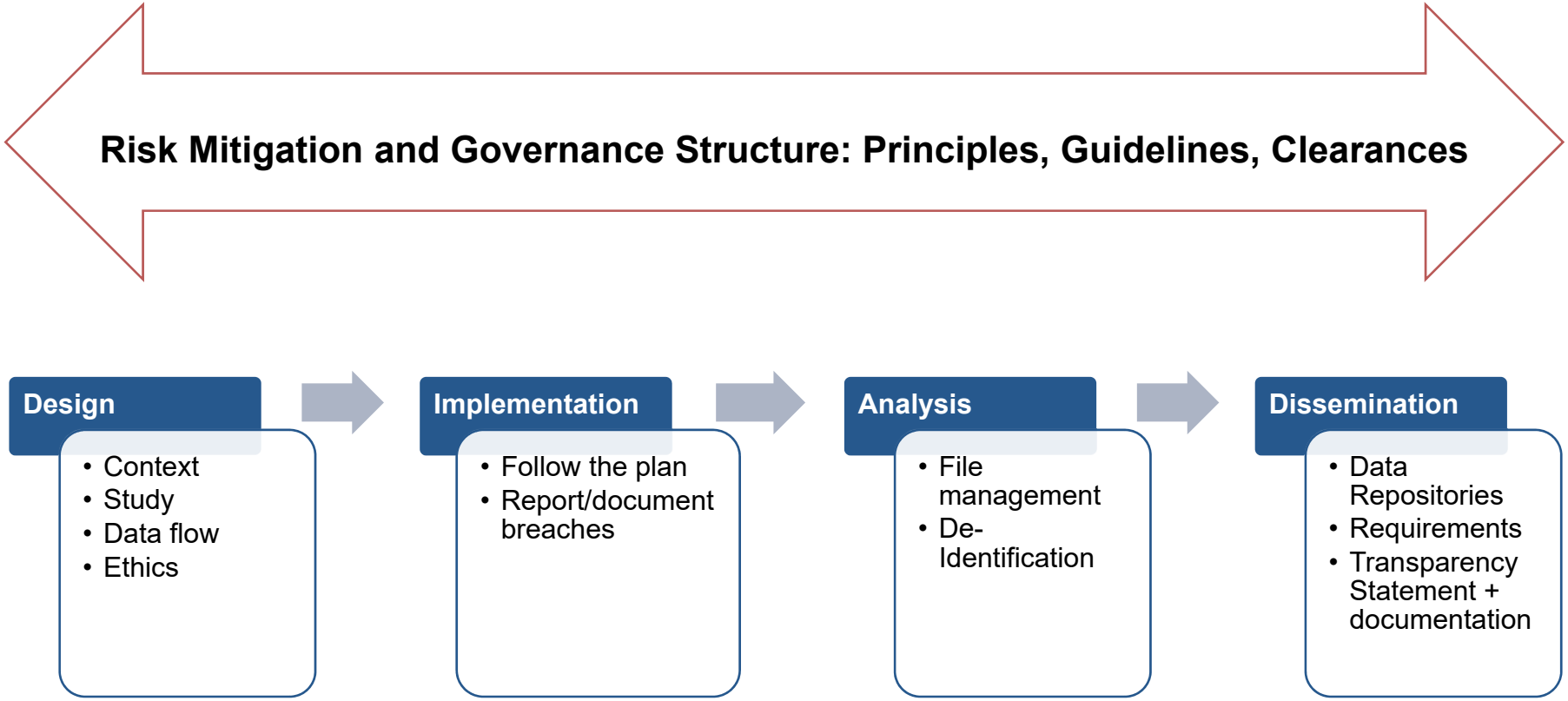
So what, you might say, what's the worst that could happen? Consider, for instance this scenario: You provide direct cash transfers to individuals. The recipients of the programme are selected by their level of vulnerability. The ruling party in the state is generally suspicious of foreign aid organisations, and believes that you are using these cash transfers to assist their political enemies. They then get hold of a list of addresses of your beneficiaries and all names in a household as well as detailed information about their financial status. The ruling party uses the data to harass and intimidate what they perceive are western-supported enemies of the party.

Source: The Guardian, Secret Aid Worker

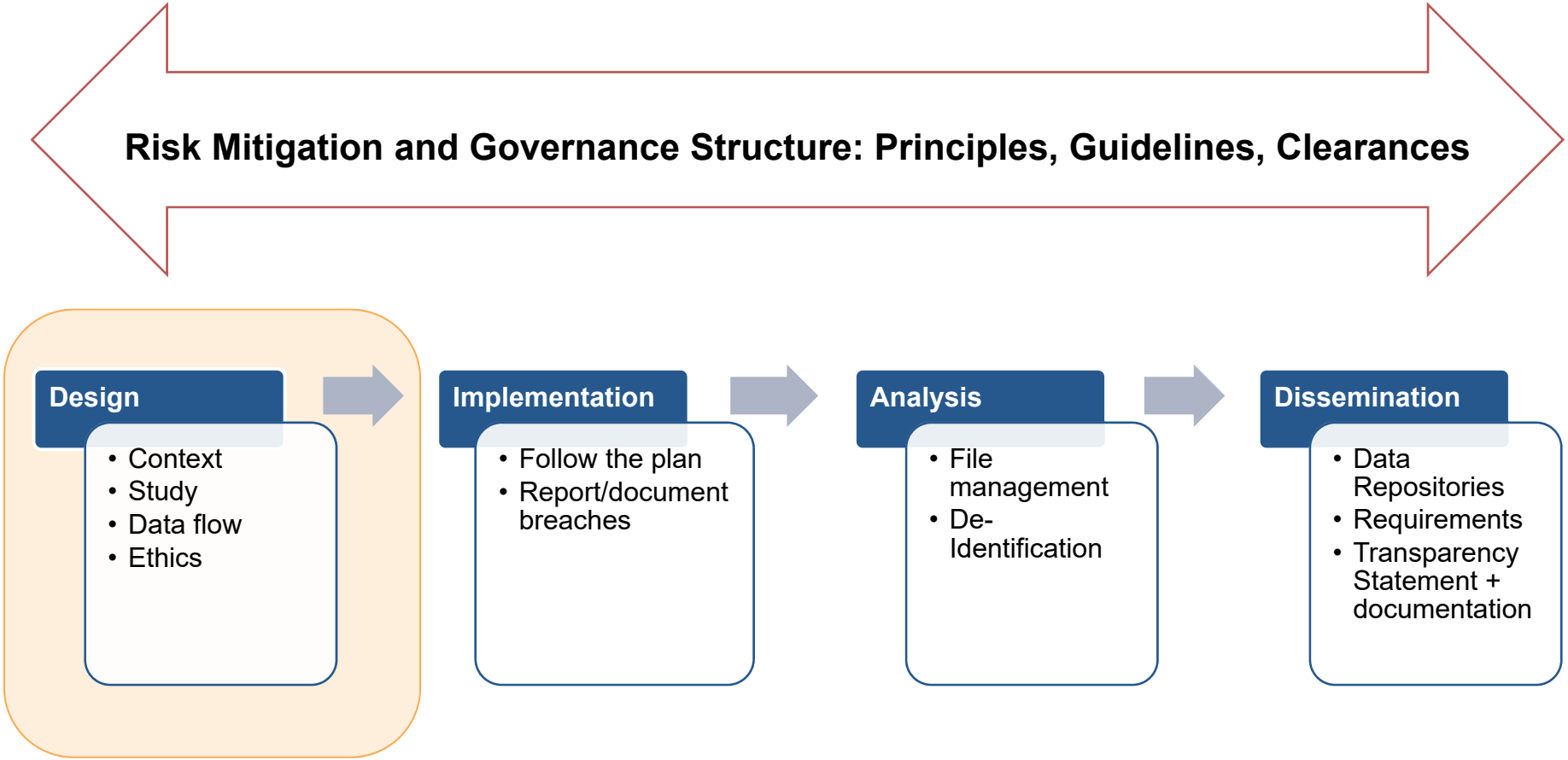
Challenge

- No one-size-fits-all approach to open data
  - Not so simple as “just post the data”
- Typically not considered until end of the research, at publication
  - Many touch-points throughout research process

Understand research process



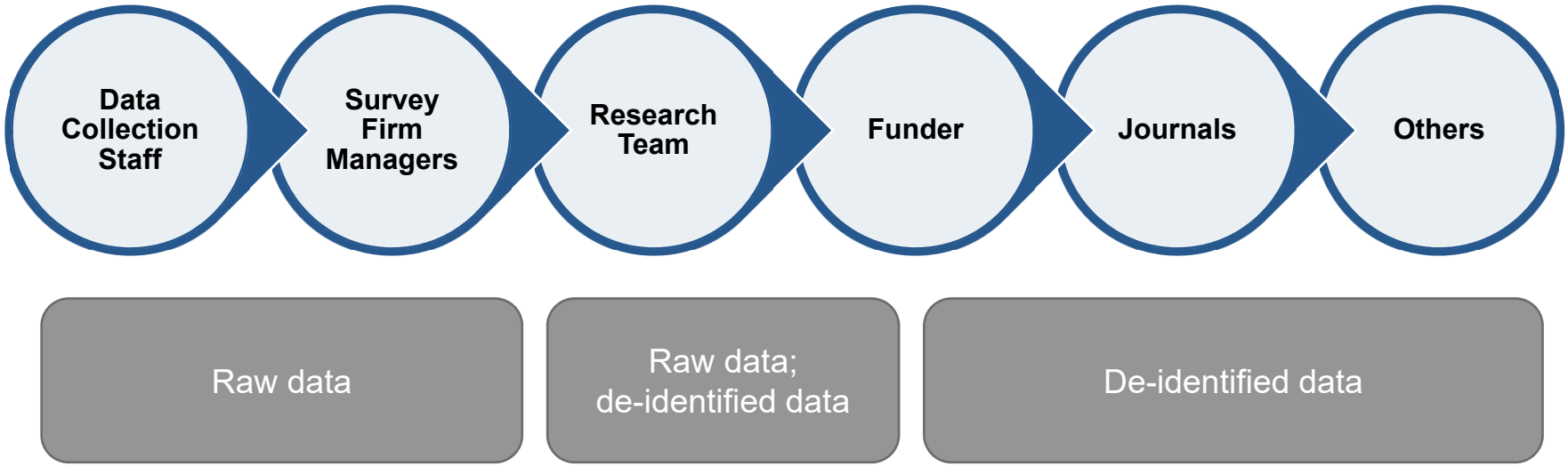
Understand research process



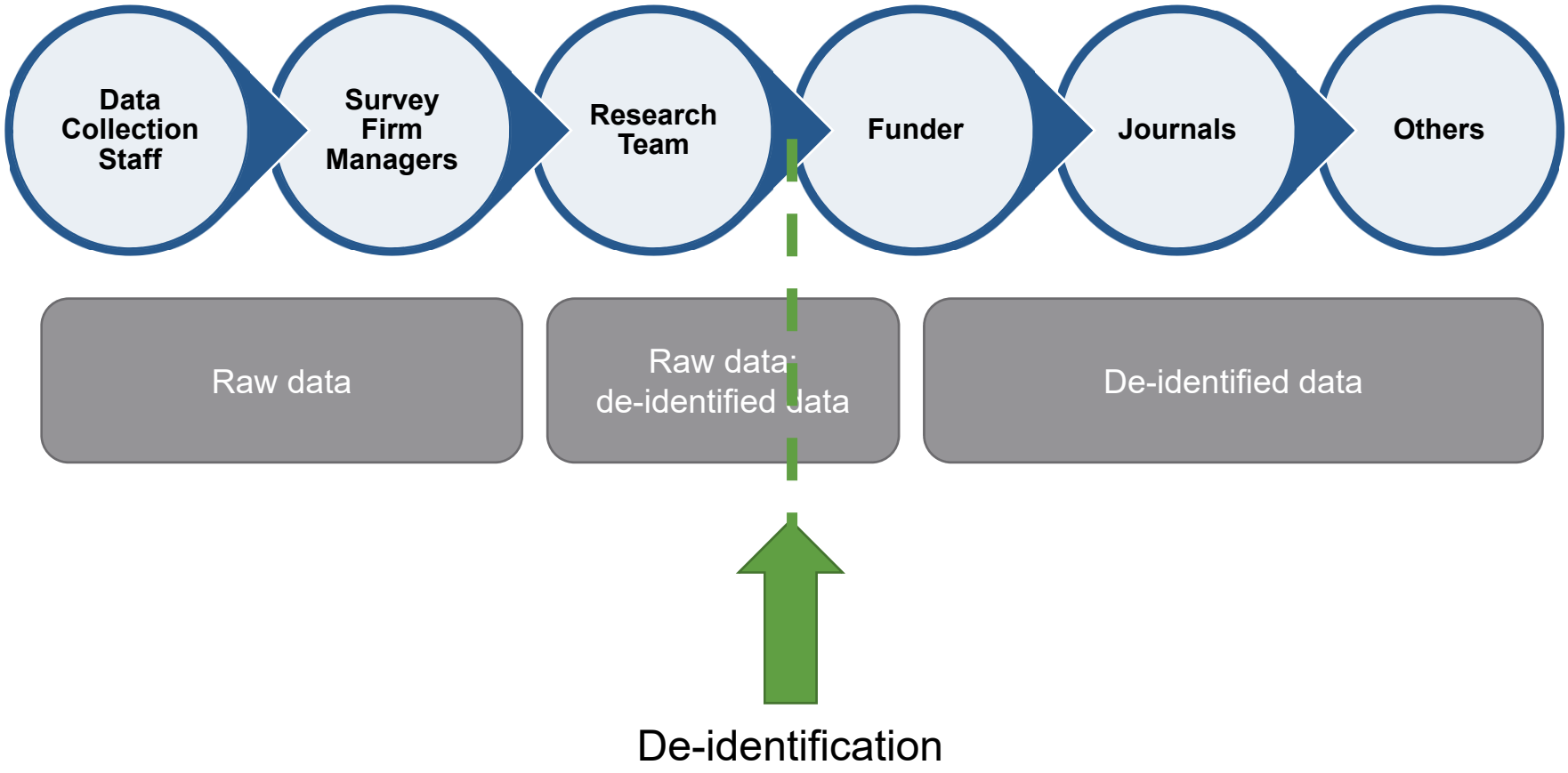
Consider context & study

- Context:
  - Sensitivity and risk may be context-dependent
  - Privacy and data protection laws
    - Where funder and research located may matter
    - US: Common Rule
    - EU: General Data Protection Regulation (GDPR)
    - Data Protection Laws of the World Map
- Study details:
  - Identification strategy, data sources and ownership, linkage with other datasets, sample and population, rounds of data collection
- Informs study's data management plan

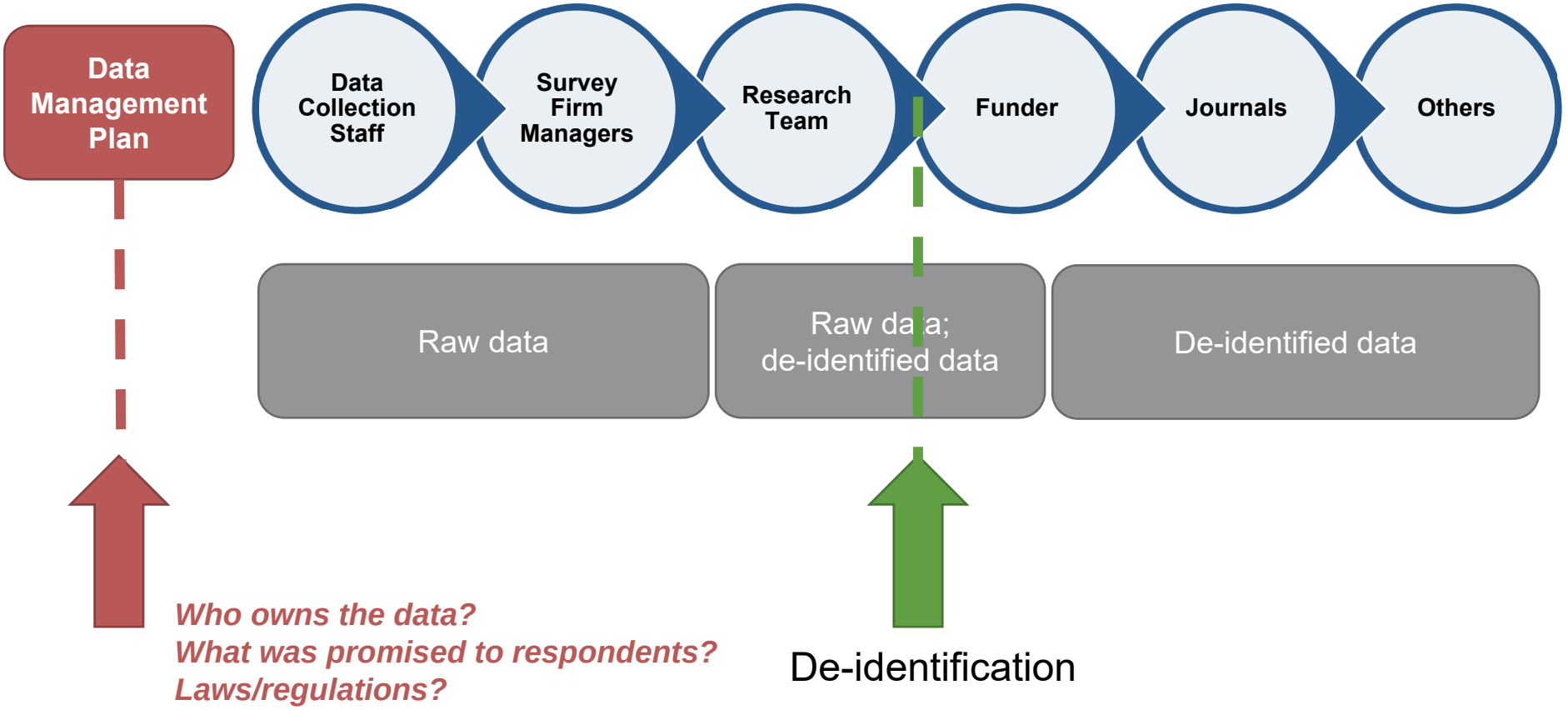
Understand your data flow



Understand your data flow



Understand your data flow



Data ownership

What kind of data do you use?

Publicly accessible?

Administrative: Institutions, individual

Primary: Institutions

Primary: Individuals

Who owns it? Under what conditions obtained?

What was promised at informed consent?

Can it be de-identified?

Ethics

- 3 basic principles
  - **Justice** – Fair distribution of benefits and burdens; consideration of vulnerable populations
  - **Respect for persons** – Informed consent, assent and parental consent
  - **Beneficence** – Understand benefits and risks of participation; protection of confidentiality

Ethics

- Institutional review board (IRB)
- Management of approved process throughout study
- Monitoring and reporting breaches
- All data handlers: common understanding
- Ideally: all data handlers with PII – IRB training
  - NIH course
  - CITI courses

- Not all IRBs are the same – check qualifications (HHS-certified?), expertise…IRBs don't guide open data efforts and don't oversee de-identification

Ethics & open data

- Collect only what you need
- Informed consent must address open data efforts
  - Confidentiality – right-size the promise
  - Transparency – data access and use
- Data publication based on conditions to which respondents consented

Consent: Examples

“Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."

"If you decide to be in this study, the study researchers will get information that identifies you and your personal health information. This may include information that might directly identify you, such as your name and address. This information will be kept for the length of the study and a fixed period afterwards (x- years). After that time it will be destroyed or de-identified, meaning we will replace your identifying information with a code that does not directly identify you.”

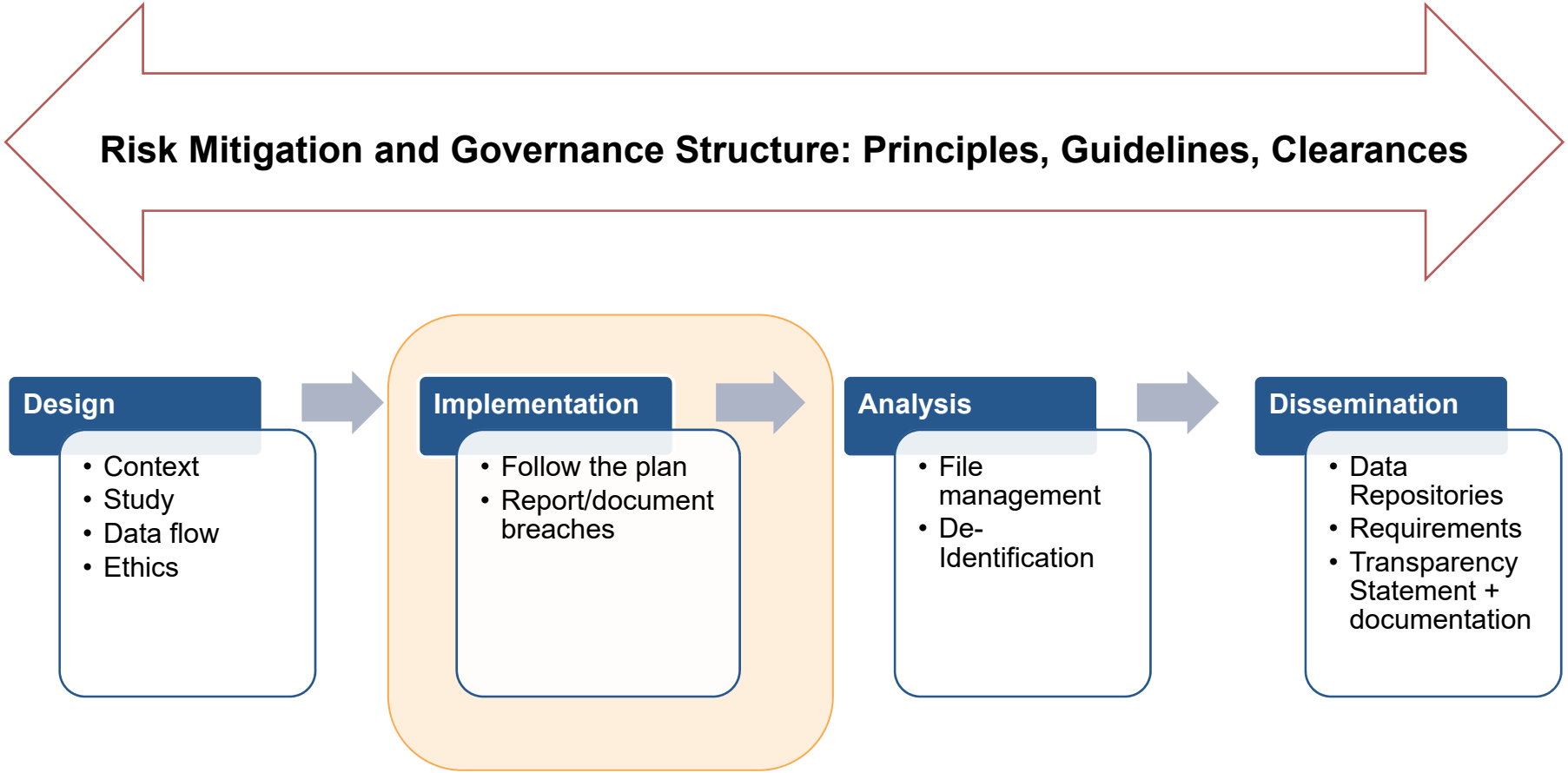
"All personally identifying information collected about you will be destroyed once it is no longer needed for the study.”

Consent: Examples

Funders may also have suggested language, e.g.:

"Our study is funded by the Millennium Challenge Corporation, an agency that provides assistance to other countries' development projects, and is being carried out by Evaluation Firm and Survey Firm. The interview is expected to take X minutes. Any information you provide that can identify you will be kept strictly confidential by the parties conducting this study, including MCC employees, employees of the survey firm, and other researchers, to the maximum extent permitted by the laws of the United States of America and the laws of Country. These users will use data for statistical purposes only."

Understand research process



Data storage and handling

- Define formats (collection, storage) and how affect risk and accessibility
- Data handling
  - Encrypted, between collaborators and authorized personnel
  - Signed agreements for data collection teams
- Digital storage
  - Permissions & password protected folders
  - Check for funder-specific guidelines
  - For example, USG federal agencies need to ensure end point encryption software AES-256 encryption or above
  - Additional resources available here

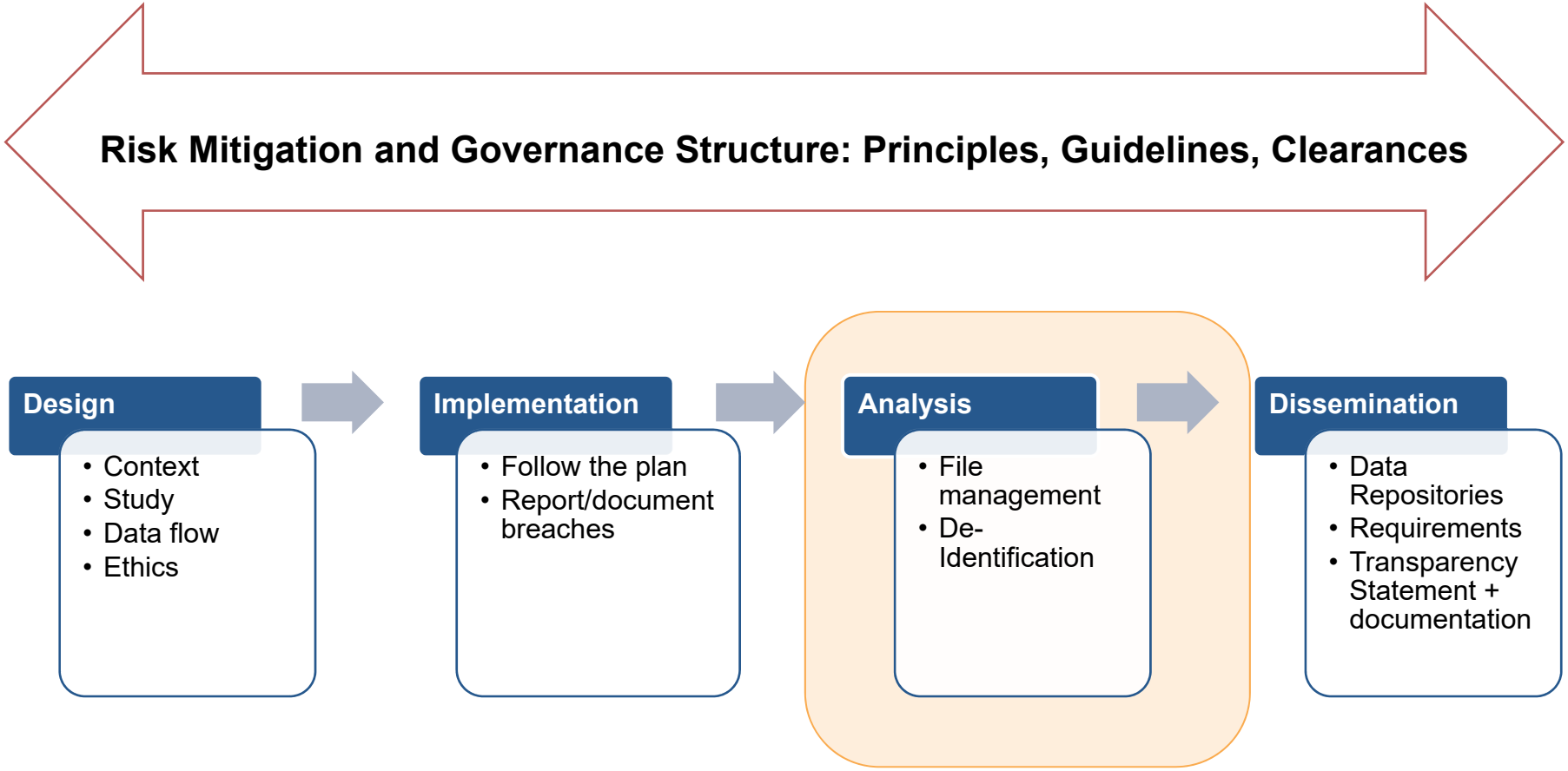
- Difficult to align across data handlers except through contractual obligations and oversight

Data breaches

- How to respond?
- How to report?
- How mitigate/change Data Management Plan?

- How does this affect your ability to de-identify for open data in future?

Understand research process



De-identification

- "Removes identifying information from a dataset so that individual data cannot be linked with specific individuals."
  - – NIST.IR.8053, here

- **Goal: reduce risk of disclosure, maintain usability of data**

De-identification

“De-identified data refers to data through which a link to a particular individual cannot be established. This often involves “scrubbing” the identifiable elements of personal data, making it “safe” in privacy terms while attempting to retain its commercial and scientific value.”

- Future of Privacy forum, here

Anonymisation is the process of turning data into a form which does not identify individuals and where identification is not likely to take place. This allows for a much wider use of the information. The Data Protection Act controls how organisations use ‘personal data’ – that is, information which allows individuals to be identified.”

“Anonymisation is of particular relevance now, given the increased amount of information being made publicly available through Open Data initiatives and through individuals posting their own personal data online.”

- UK ICO, here

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM
FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY
Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA
Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

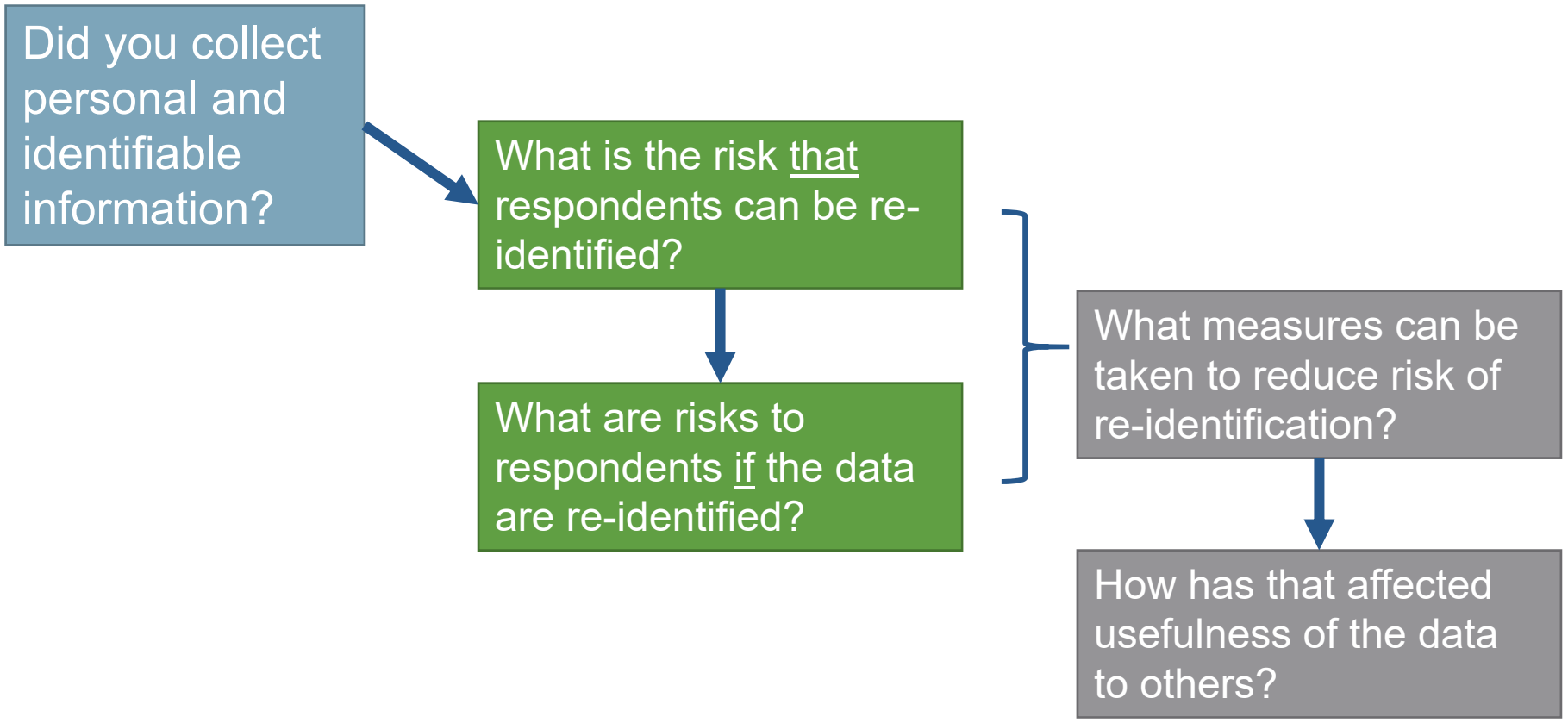
DE-IDENTIFIED DATA
Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA
Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

Source: Future of Privacy Forum, link here

De-identification



Did you promise confidentiality in informed consent?

Who might want access? Why? What means do they have to do so? What kinds of risks? Financial, legal, medical, reputational

No one size fits all but best practices; Researcher responsibility to assess risk & take action

De-identification

- **Understand context & study data**
- **Assess risk**
- **Conduct de-identification**
- **Test de-identification**
- **Publish data & documentation**

Assessing risk

- **Using dataset on its own**
  - Forward & backward re-identification
  - Indirect identifiers, small populations, inferential disclosure
  - Consider group harm
- **By linking with other datasets**
  - Linkage, “jigsaw”
  - Not necessarily limited to (free or) publicly available data
  - HIPAA suggests 3 conditions:
    - de-identified data are **unique or “distinguishing”**
    - **data source** with which it could be linked
    - **mechanism to relate** de-identified & identified data sources
- **Intruder scenarios**
  - Consider motivations, existing info, and means to re-identify
  - Spontaneous recognition, nosy neighbor, “prosecutor” scenario

Assessing risk

- **UK Office for National Statistics - Intruder scenarios:**
  - attempt to match a number of individuals or other statistical units in the data with other sources; these will share a characteristic of interest such as ethnic group, household composition or occupation
  - attempt to find a specific individual, household or business as they have knowledge of information unique to this record (including, perhaps, response knowledge)
  - attempt to find an arbitrary record in order to demonstrate that the published data are not secure
  - attempt to find a record in the survey and use this information to find out more about this record externally; this is the reverse of the previous types of attack

Source: UK ONS Policy for social survey microdata

Assessing risk

"What is an acceptable level of identification risk for an expert determination?

There is no explicit numerical level of identification risk that is deemed to universally meet the "very small" level indicated by the method. The ability of a recipient of information to identify an individual (i.e., subject of the information) is dependent on many factors, which an expert will need to take into account while assessing the risk from a data set. This is because the risk of identification that has been determined for one particular data set in the context of a specific environment may not be appropriate for the same data set in a different environment or a different data set in the same environment. As a result, an expert will define an acceptable "very small" risk based on the ability of an anticipated recipient to identify an individual."

Source: HHS HIPAA de-identification guidelines

De-identification

Direct Identifiers	**Removal** As early as possible; create key; access by approved personnel	• Individual names • Phone numbers • Addresses • GPS coordinates
Indirect Identifiers	**Permutation** In some cases, may consider suppressing	• Area names • Linkable geographic codes • Account numbers • Check text fields in survey
Sensitive Data	Highly context-dependent; judgment of researcher	• Income/expenditures, durable assets • Political views & activity • Trauma, abuse, medical information • Witness or participation in illegal activity

De-identification

Well-commented, replicable code

Direct Identifiers	**Removal** As early as possible; create key; access by approved personnel	• Individual names • Phone numbers • Addresses • GPS coordinates
Indirect Identifiers	**Permutation** In some cases, may consider suppressing	• Area names • Linkable geographic codes • Account numbers • Check text fields in survey
Sensitive Data	Highly context-dependent; judgment of researcher	• Income/expenditures, durable assets • Political views & activity • Trauma, abuse, medical information • Witness or participation in illegal activity

Consider…

- **Names of persons**
- **Names of institutions/affiliations**
- **Geographic information (names, codes)**
- **Elements of dates (except year) directly related to a respondent, ages**
- **Identifying numbers (e.g. accounts, IDs)**
- **Identifying web information**
- **Biometric/anthropometric information**
- **Rare occupations, conditions, religious sects, languages, ethnicities, political affiliations, etc.**
- **Photos, video, or audio**
- **Free text or open entry data**
- **Variable combinations, including within small geographic areas**
- **Linkage with other existing datasets**

Consider…

- **Names of persons**
- **Names of institutions/affiliations**
- **Geographic information (names, codes)**
- **Elements of dates (except year) directly related to a respondent, ages**
- **Identifying numbers (e.g. accounts, IDs)**
- **Identifying web information**
- **Biometric/anthropometric information**
- **Rare occupations, conditions, religious sects, languages, ethnicities, political affiliations, etc.**
- **Photos, video, or audio**
- **Free text or open entry data**
- **Variable combinations, including within small geographic areas**
- **Linkage with other existing datasets**

USG/health, HIPAA Safe Harbor:
https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

Consider…

- **Names of persons**
- **Names of institutions/affiliations**
- **Geographic information (names, codes)**
- **Elements of dates (except year) directly related to a respondent, ages**
- **Identifying numbers (e.g. accounts, IDs)**
- **Identifying web information**
- **Biometric/anthropometric information**
- **Rare occupations, conditions, religious sects, languages, ethnicities, political affiliations, etc.**
- **Photos, video, or audio**
- **Free text or open entry data**
- **Variable combinations, including within small geographic areas**
- **Linkage with other existing datasets**

☑ TIP

Consider naming variables to indicate identifiers, making them less likely to miss later on

☑ TIP

If you are a Stata user, and know what you're looking for in variable names or text fields, consider using:

-**lookfor**-

-**strpos()**-

USG/health, HIPAA Safe Harbor:
https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

De-identification

- **Order matters**
  - **Remove direct identifiers and address geographic identifiers first**, to limit amount of other data manipulation needed
  - Drawback - loss of transparency on study locations
  - Consider restricted-access dataset

- **Researcher judgment to determine how data treated**
  - **Continuous:** Grouping, top-/bottom-coding, winsorizing, rounding
  - **Categorical:** Coarsening, collapsing categories
  - **String:** Redacting or pseudonymizing

- **Decisions justified against risk assessment**
  - For sensitive data, consider observability & stability of the attribute

Other types of data

- **Principles are the same!**
  - Protection of subjects vs. usability
  - Commensurate with assessed risk

- Qualitative
  - Redacting personal, sensitive information
  - Pseudonymization for people, places
  - Consider generalizability, research purpose
- Geospatial
  - Aggregating, masking, jittering
- Media (audio, photo, video)
  - Blurring, bleeping
  - Includes data collected for quality control

☑ TIP

Flag potentially identifying information while you are coding or reading qualitative transcripts.

&

Keep a de-identification log.

Qualitative de-identification log

Original text

Interview recorded: 3pm, 10 October 2011
Interviewee: Julius Smith
DoB: 9 September 2005
School: Green Lanes Primary School

I live on Clementine Lane so I walk to school every day. I live
a flat with my parents and my Uncle Jermaine. When I get h
from school I watch TV. I don't like reading but I like watchin
Harry Potter films. My favourite subject at school is art. My
teacher is Mr Haines and he is very nice. I used to get bullied
Neil and Chris but I told Mr Haines and they stopped.

I play football for Junior Champs, and we are good. I play
midfield.

Anonymised text

Interview recorded: October 2011
Interviewee ref: 2011/67
School year: Key Stage 1
School local authority area: Lynenham District Council

I live in [LM51 postcode] so I walk to school every day. I live with [family members]. When I get home from school I watch TV. I don't like reading but I like watching Harry Potter films. My favourite subject at school is art. My teacher is Mr [teacher's name] and he is very nice. I used to get bullied by [other pupils] but I told [the teacher] and they stopped.

I play football for [a local team], and we are good. I play midfield.

Source - UK ICO [Anonymization code of practice]

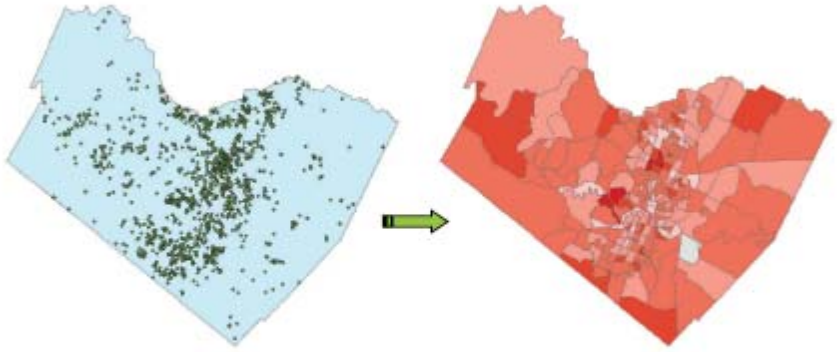
Log

example anonymisation log:

Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Spain	European country
p3	Manchester	Northern metropolitan city or English provincial city
p2	20th June	June
p2	Amy (real name)	Moira (pseudonym)
Int2		
p1	Francis	my friend
p8	Station Road primary school	a primary school
p10	Head Buyer, Produce, Sainsburys	Senior Executive with leading supermarket chain

Source: [UK Data archive]

Spatial de-identification



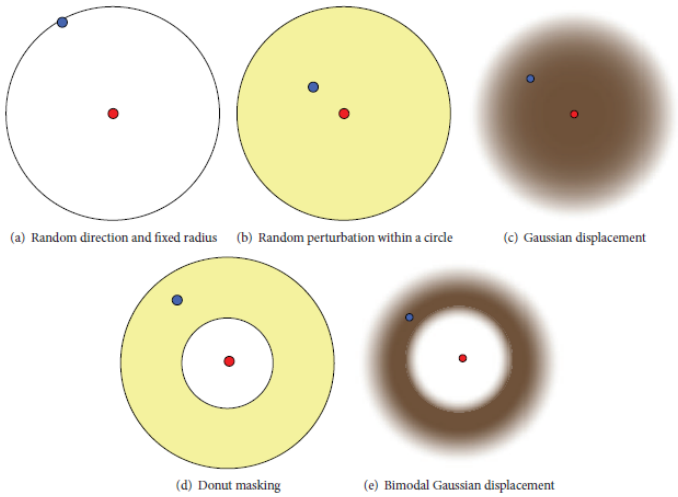


FIGURE 5: Graphical representation of common geographic masking techniques. The red dot indicates the original location and the blue dot one of the many possible masked locations.

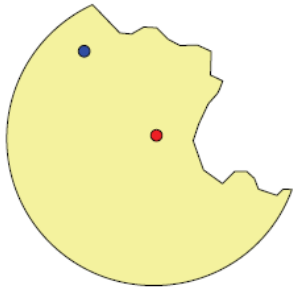
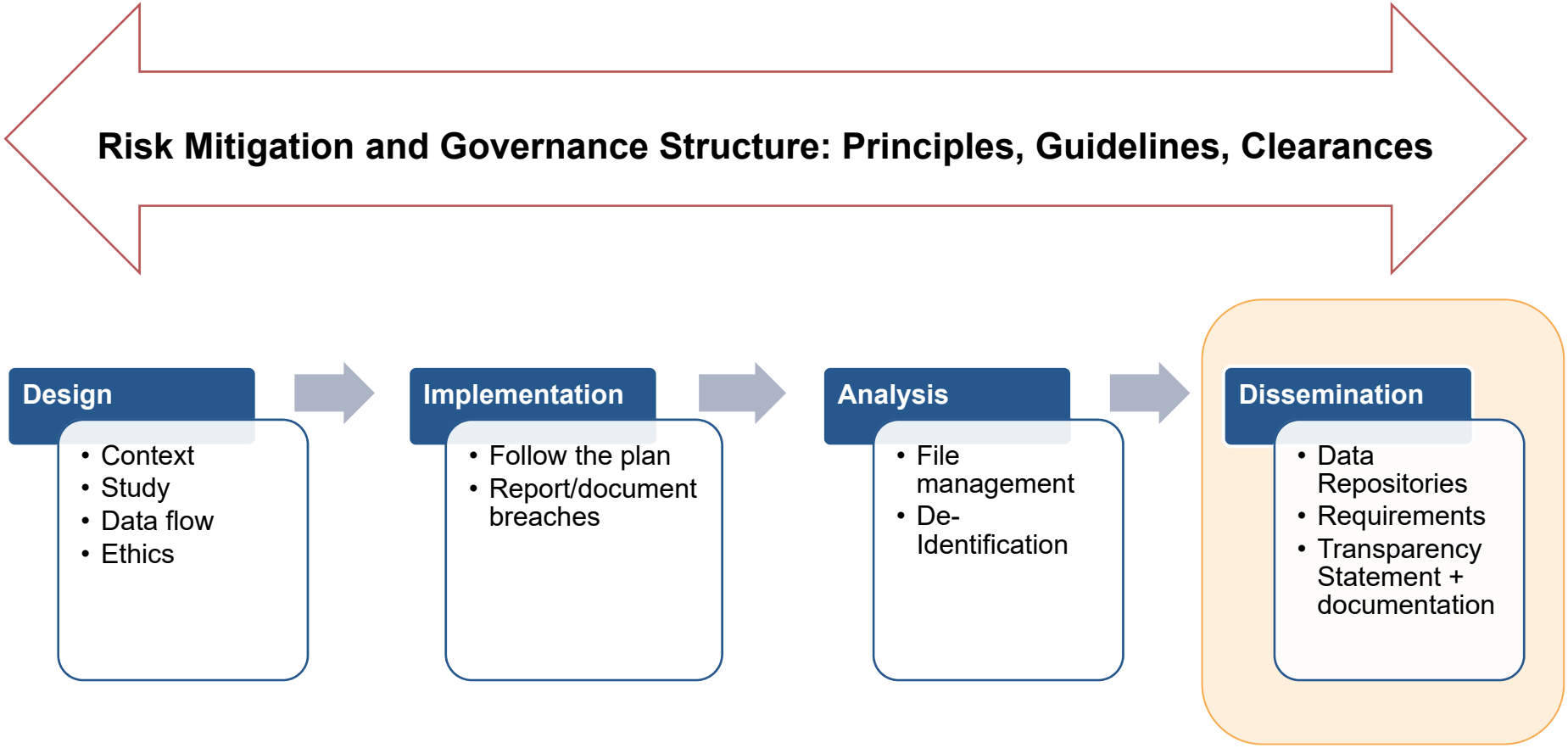


FIGURE 6: Example of geographic masking technique (i.e., random placement within a circle) using an additional spatial filter to constrain displacement. The red dot represents the original location; the yellow area represents all possible locations for the masked location; and the blue dot represents one possible masked location selected randomly. This filter can be used to avoid placement in areas where logically no population resides (such as water bodies or parks) or to limit displacement to a particular enumeration unit (such as the same census tract or postal code).

Source: Paul A. Zandbergen 2014. Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Department of Geography, University of New Mexico, Albuquerque, NM 87131, USA*

Understand research process



Dissemination

- **Research documentation linked to the data**
  - Design, pre-analysis plan, sampling
  - Questionnaires, informed consent
  - Reports, Data, Code
  - Transparency Statement

- **Restricted access (requests), if de-identification limits usability**
  - Must still be within terms of informed consent
  - Example – MCC requires new users to sign and submit a Restricted-Access Certification Form when requesting restricted-access data

- Data use agreements, MOUs
- Cold room, data centers
- Timing depends. Some rules of thumb…
  - 6-12 months following publication of report
  - With journal publication

De-identification

- **Understand context & study data**
- **Assess risk**
- **Conduct de-identification**
- **Test de-identification**
- **Publish data & documentation**
- **Be kind to replicators?**

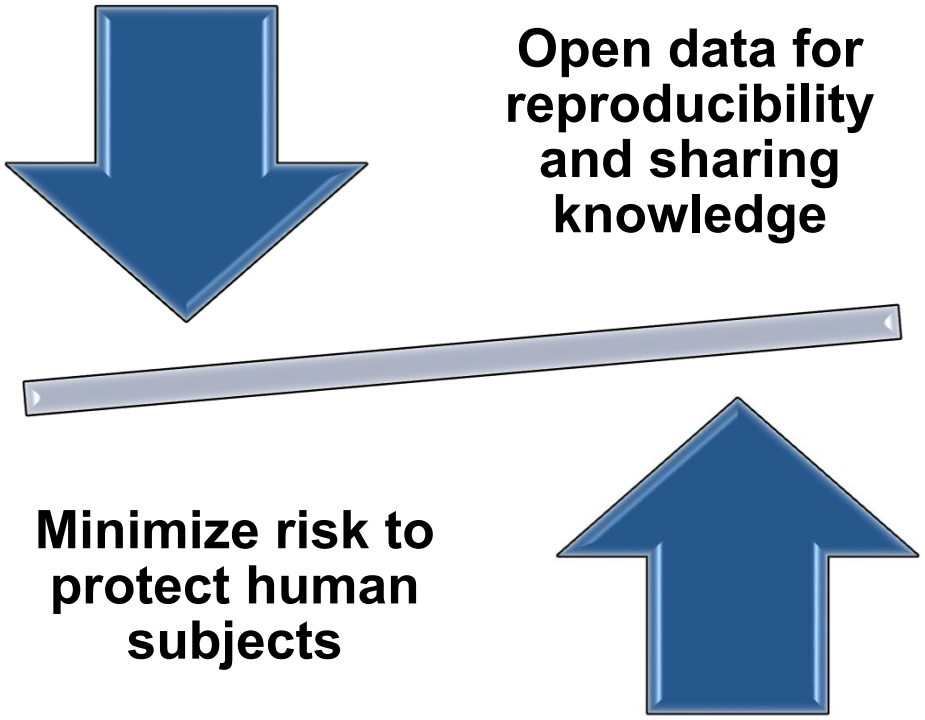
De-identification & replication

- Before starting, consider – WHEN can/should you consider de-identification?
- If run all analysis on identifiable data – who has access to that identifiable data? Will others be able to replicate your analysis in the future?
- Consider if de-identification can largely be done before analysis, so that underlying data is shareable

Key takeaways

- **Open data heart of research transparency and reproducibility**
- **Think early & often to facilitate open data**
- **Plan early**
  - Data Management Plan
  - Ethics & informed consent
  - Data handling, management
  - Risk assessment, de-identification, data publication
  - De-identification vs. analysis/replicability
  - Data publication + supplementary materials
  - Share everything else if data can’t be shared?
  - Set up of restricted access?

Thank you


Open data for reproducibility and sharing knowledge
Minimize risk to protect human subjects

Resources

Privacy & Data Protection by Country (here and here)

USG

De-Identification of Personal Information (NIST.IR 8053)
HIPAA Privacy Rule
Belmont Report
USAID ADS Chapter 508
GovLab assessment of data privacy policies
Open Data Compliance guide for USAID Open Data Policy
MCC Microdata Guidelines
NIH human subjects research training

Resources

EU General Data Protection Regulation

- GDPR Full text; Article 29 Working Party Guidelines & documents
- EU ICO
  - Guide to the GDPR
  - GDPR Key Definitions
  - ICO guide to “what is personal data?”
  - ICO guide on determining what is personal data
  - ICO guide to data controllers versus processors
  - 12 Steps to Prepare for the GDPR
  - Getting ready for the GDPR
  - GDPR compliance checklists
  - Guide to Data Protection
- Other:
  - Community call on GDPR – Etherpad notes
  - Primer on the GDPR
  - Top 10 Operational Impacts of the GDPR
  - Paper on Personal Data and Encryption in the GDPR
  - How GDPR overlaps with US regulations

Resources

Other

Responsible Data Hackpad
Responsible Data community
Responsible Data Handbook
Digital Safeguarding Training Tools
Girl Effect Digital Safeguarding Policy
Oxfam Responsible Data Policy
MEASURE Evaluation Data Security, Privacy, and Confidentiality
UNHCR Data protection policy
CGIAR Open data & data management policy
GODAN Responsible data in agriculture & nutrition
UK ICO Privacy impact assessment code of practice
UK ICO Data sharing code of practice
BITSS Ethics in Social Science Presentation
Transparency and Openness Promotion (TOP) Guidelines
Utrecht University Data Management Resources
IPA Data Publication Guidelines
Data Sharing Practices (De-Identified Datasets)
More consent language examples

Resources

Admin Data

Administrative data and reproducibility

JPAL Administrative data resources

Administrative Data Guidelines

MIT Sample data use agreements

Data Management

ICPSR

Data Management Planning Tool

Dataverse Data Management Plan guidance

Data Archiving and Networked Services (DANS)

Utrecht University Data Management materials

Handbook of the Modern Development Specialist

Resources

Data Repositories

World Bank's Microdata Catalog

Harvard's Dataverse

UMichigan: ICPSR

Journals → Mendeley Data

US – Data.gov

EU - Data Archiving & Networked Services

World Bank Open Data

Questions?

"I'm sure there are better ways to disguise sensitive information, but we don't have a big budget."