Placing administrative data at the heart of the UK population statistics system; how the total survey error framework is helping to inform the use and design of integrated data solutions

Nicola Rogers, Centre for International Migration, Office for National Statistics, United Kingdom

Dr Louisa Blackwell, Methods Data and Research, Office for National Statistics, United Kingdom

Presentation Outline

- Office for National Statistics, UK Population & Migration Statistics Transformation Programme
- International migration and the use of Home Office administrative data
- An error framework for longitudinally linked administrative data
- Further work in progress

ONS Population Statistics Transformation

- We are the UK's largest independent producer of official statistics and its recognised national statistical institute
- We are responsible for collecting and publishing statistics related to the economy, population and society at national, regional and local levels
- We also conduct the census in England and Wales every 10 years
- Working in partnership across the Government Statistical Service (GSS), ONS are progressing a programme of work to **put administrative data at the core of our evidence on international migration by 2020**
- Our work to transform migration statistics is joined up with the Admin Data Census programme to transform population statistics – with admin data also at the core of population statistics by 2020

How we currently measure international migration

- International Passenger Survey (IPS) at UK ports- intentions based – how long migrants plan to stay or leave

- Outputs include long-term and short-term migration, migrant population, using UN definitions

- **High demand from our users – including calls from parliamentary committees - for better evidence on migration**

- Changing policy context means decision-makers and the public need more evidence on migration, including the impact and contribution of migration on the economy and society. Existing survey data doesn't provide sufficient depth and there are gaps in the evidence base

- All available sources being considered

Our latest research

- January research engagement report updating users on our population and migration statistics transformation journey
- This shows progress towards using administrative data, bringing sources together to fill gaps in coverage.
- We have researched linked immigration, education, health and annualised income records.
- Underlines the importance of linkage and triangulation across a range of sources – no single source has all the answers
- Home Office administrative data is a critical and valuable source in this research

Outline of Exit checks data collection

Biometric Enrolment

Visa Application

Visa Issued

OVERSEAS VISA POST

VALID VISA PERIOD

T I M E L I N E

ISA's ELECTRONIC RECORDS

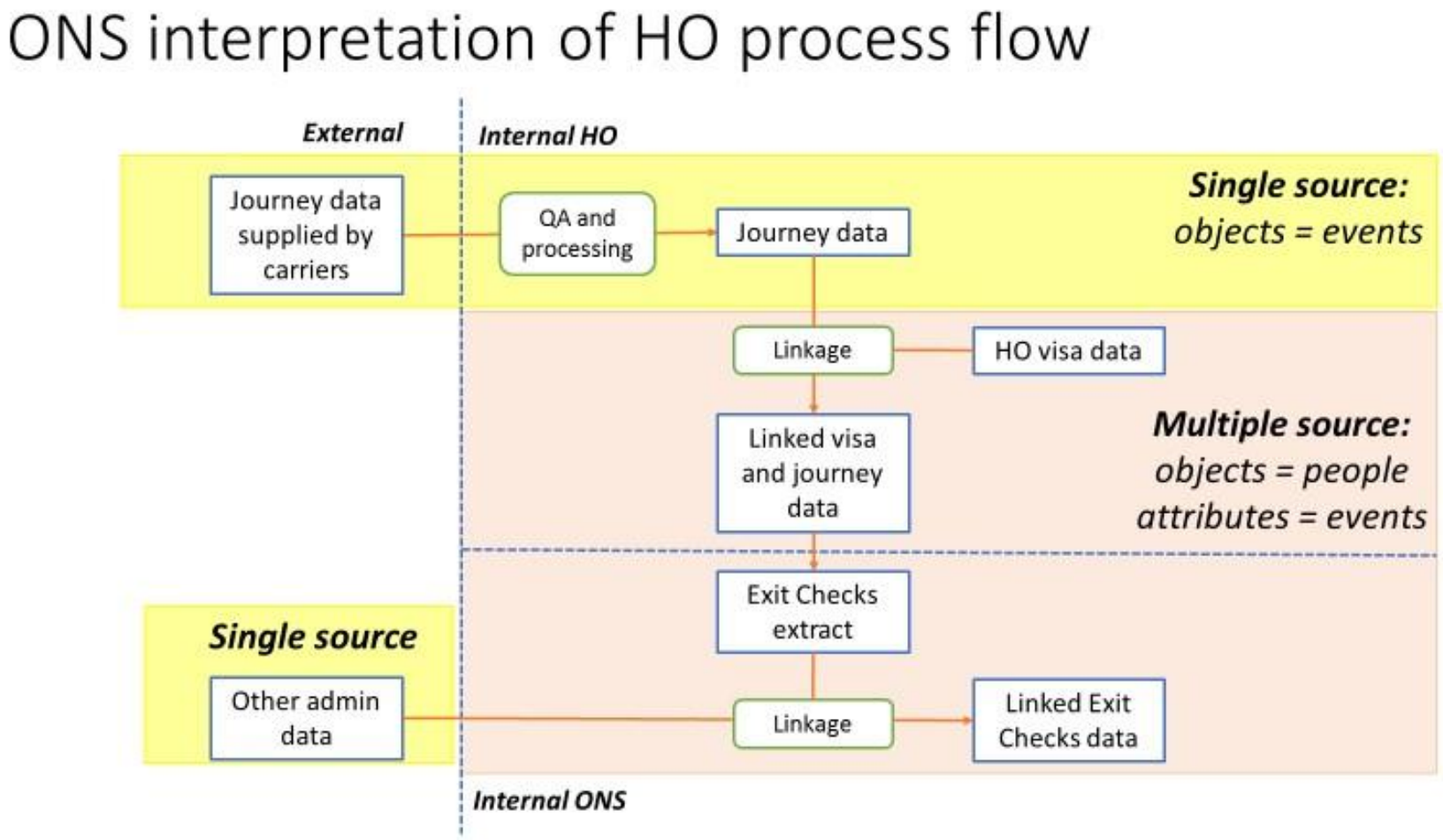
08/04/2015	IABS	Biometric Enrolment
09/04/2015	Proviso/CRS	Visa Application
10/04/2015	Proviso/CRS	Visa Issued
01/05/2015	API	Inbound Passenger Record
01/05/2015	PCP	Passport Scanned
01/05/2015	SBVP	Secure ID Biometric Check against IABS biometrics
14/05/2015	API	Outbound Passenger Record
10/10/2015	Proviso/CRS	Visa Expires

KEY EVENTS

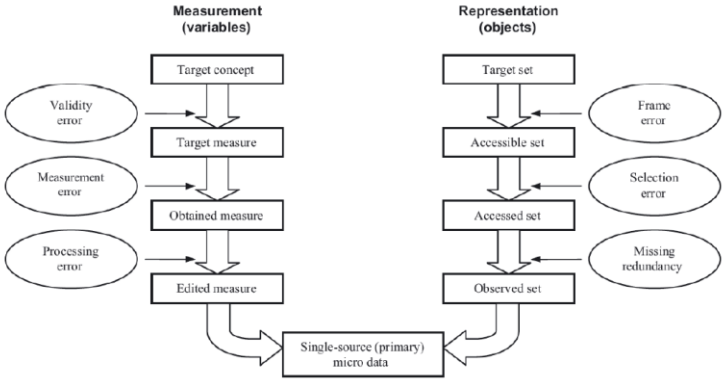
ISA Identity S

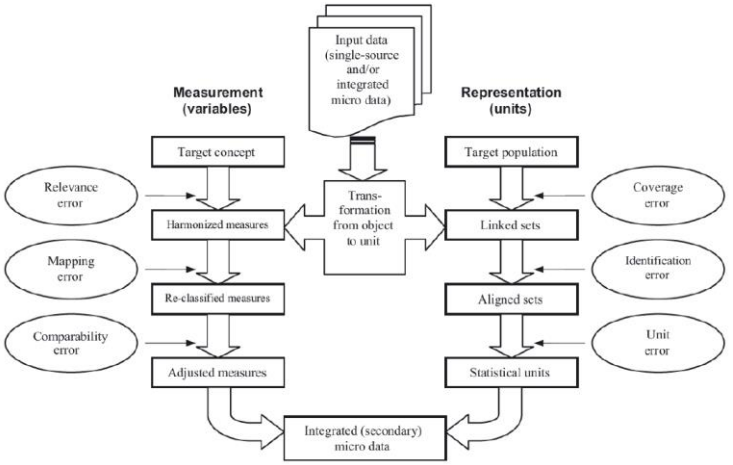
Out of Country Leave Expired Compliant

Complex linkage of Exit Check data

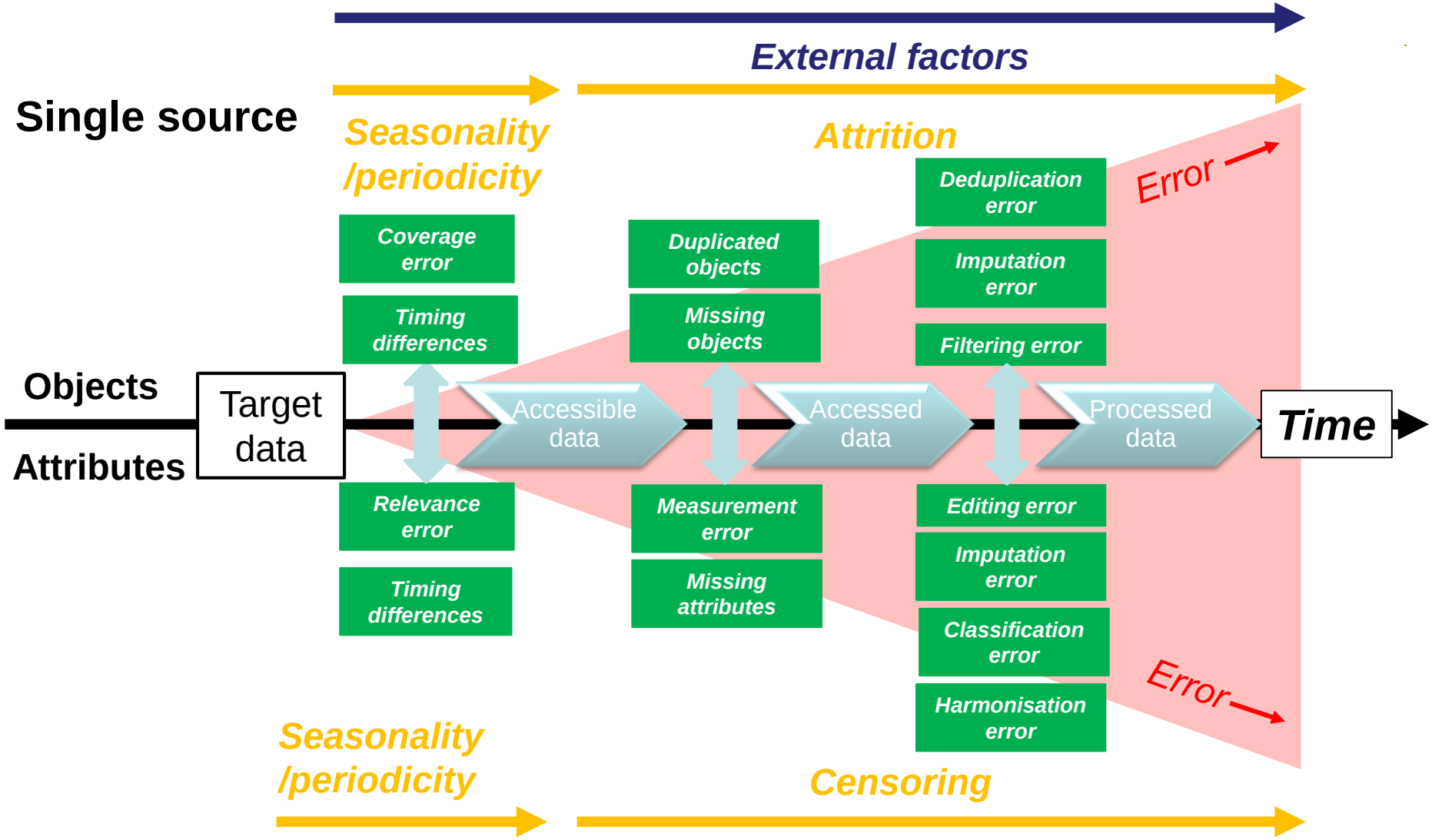


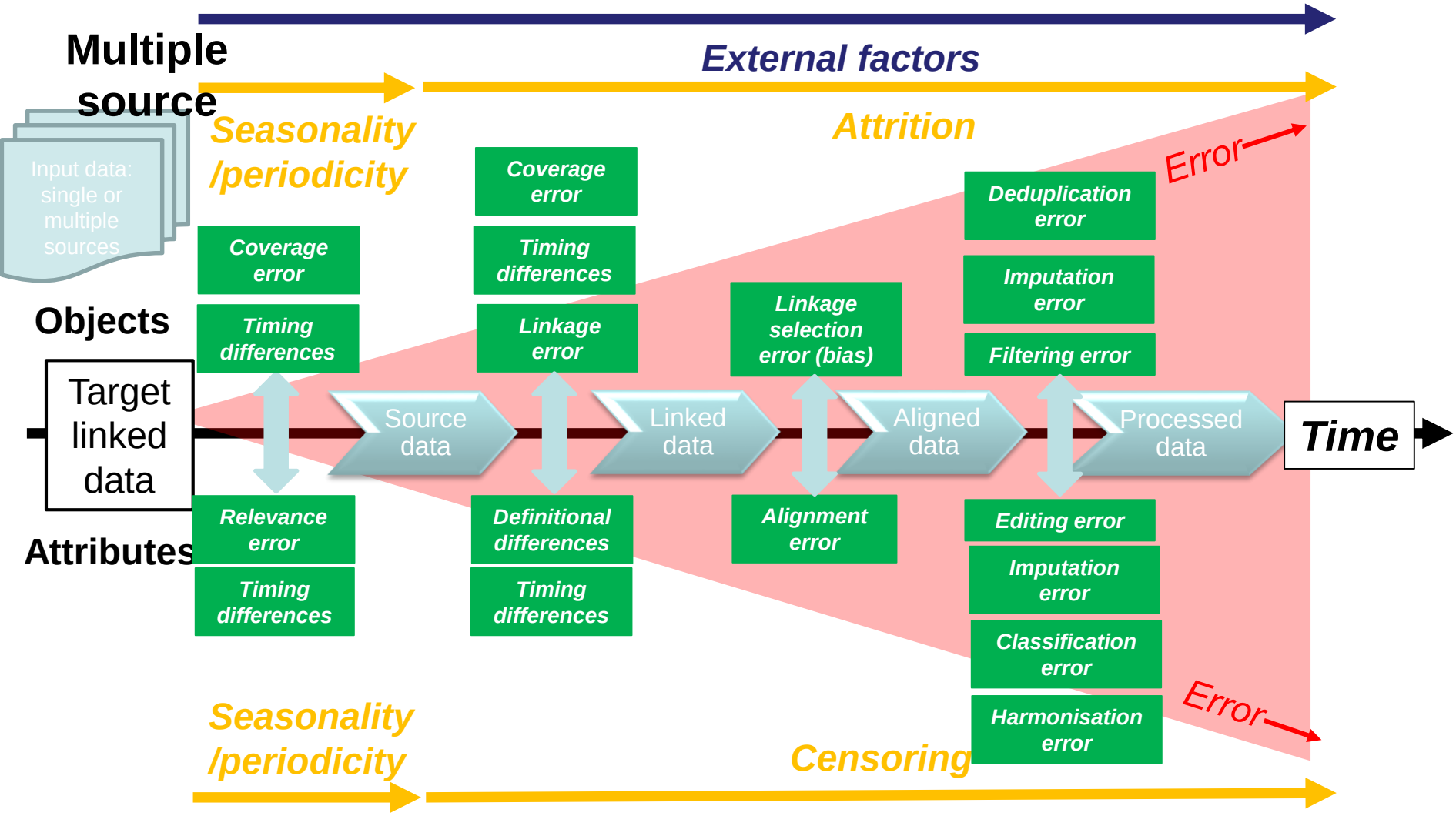
SNZ quality framework: Phases 1 and 2

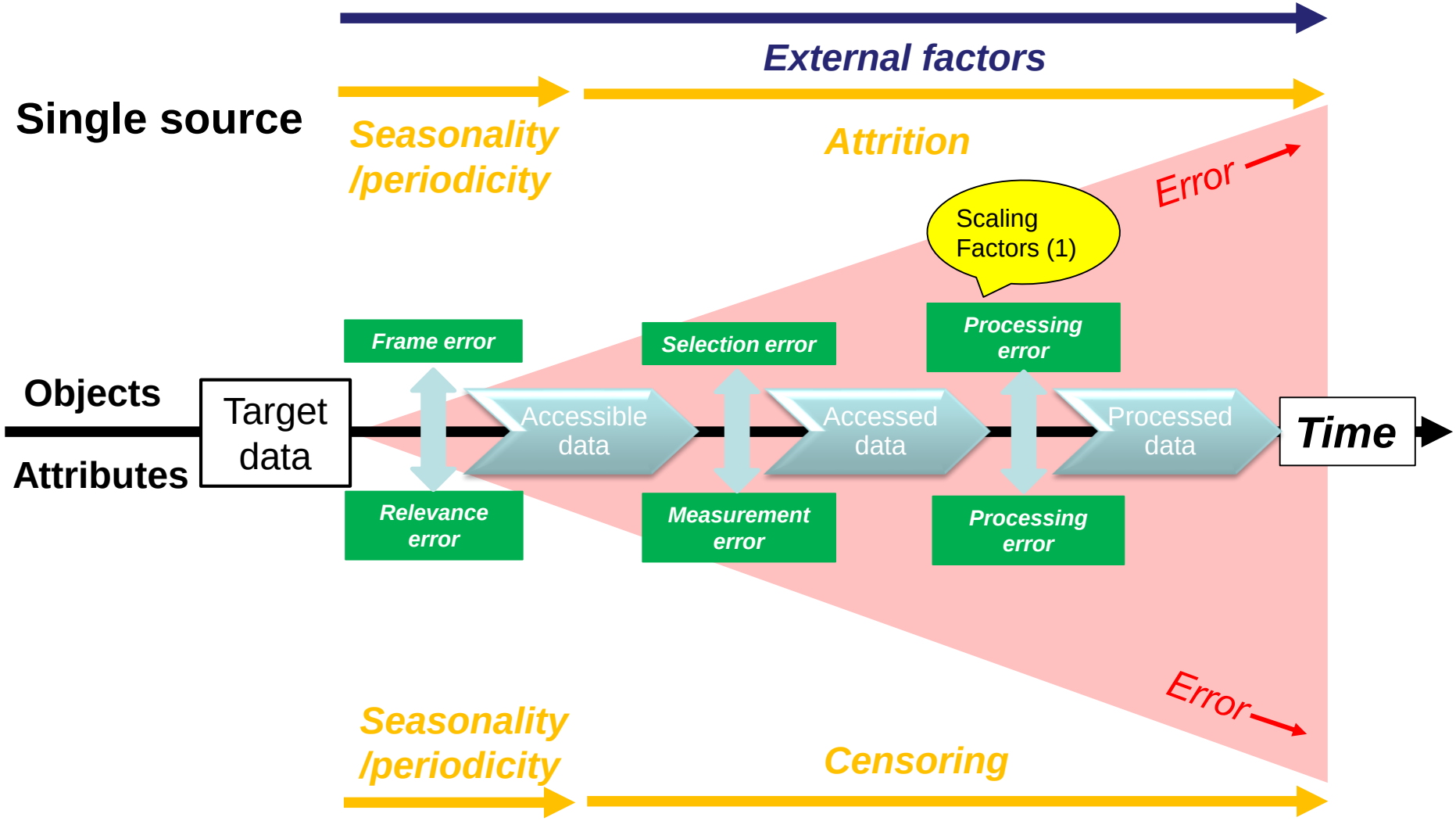


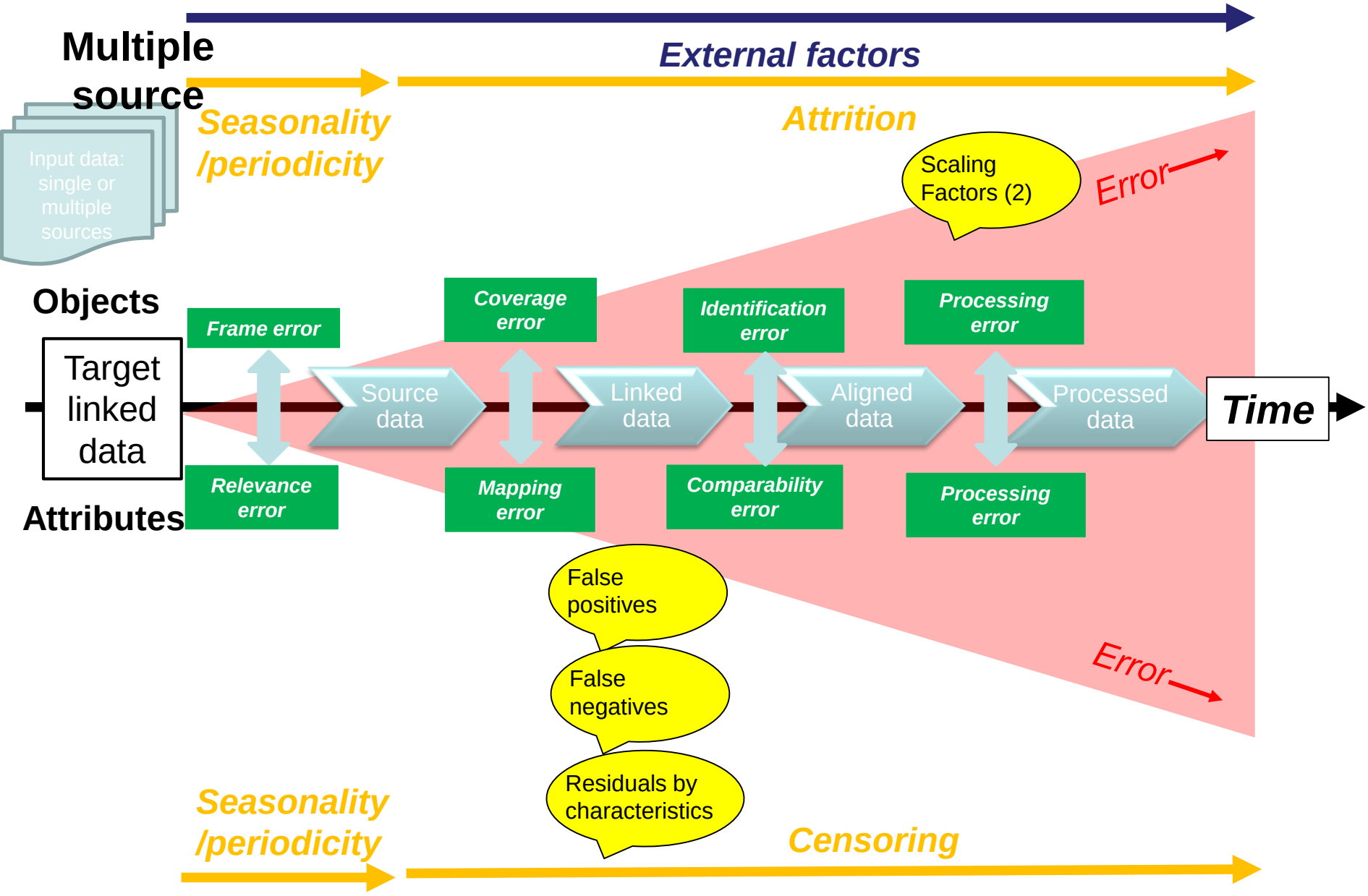


- Extremely valuable taxonomy of error to help us to understand quality issues
- Taken it a step further to explicitly acknowledge error arising from the longitudinal dimension


External factors
Single source
Seasonality /periodicity
Attrition
Error
Coverage error
Timing differences
Duplicated objects
Missing objects
Deduplication error
Imputation error
Filtering error
Objects
Attributes
Target data
Accessible data
Accessed data
Processed data
Time
Relevance error
Timing differences
Measurement error
Missing attributes
Editing error
Imputation error
Classification error
Harmonisation error
Error
Seasonality /periodicity
Censoring


Multiple source
Input data: single or multiple sources
Objects
Target linked data
Attributes
External factors
Seasonality /periodicity
Attrition
Error
Coverage error
Timing differences
Relevance error
Timing differences
Coverage error
Timing differences
Linkage error
Definitional differences
Timing differences
Linkage selection error (bias)
Alignment error
Deduplication error
Imputation error
Filtering error
Editing error
Imputation error
Classification error
Harmonisation error
Source data
Linked data
Aligned data
Processed data
Time
Seasonality /periodicity
Censoring
Error


External factors
Single source
Seasonality /periodicity
Attrition
Error
Scaling Factors (1)
Frame error
Selection error
Processing error
Objects
Target data
Accessible data
Accessed data
Processed data
Time
Attributes
Relevance error
Measurement error
Processing error
Error
Seasonality /periodicity
Censoring

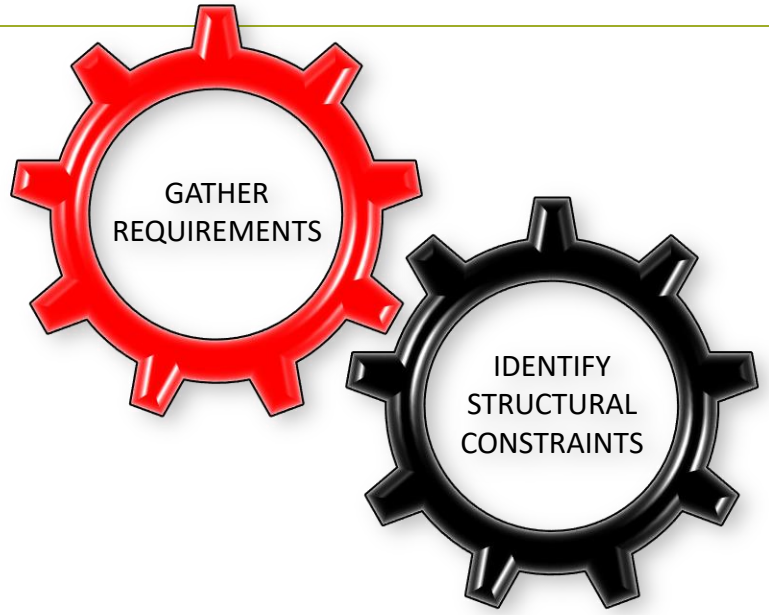

Multiple source
External factors
Input data: single or multiple sources
Seasonality /periodicity
Attrition
Scaling Factors (2)
Error
Objects
Target linked data
Attributes
Frame error
Coverage error
Identification error
Processing error
Source data
Linked data
Aligned data
Processed data
Time
Relevance error
Mapping error
Comparability error
Processing error
False positives
False negatives
Residuals by characteristics
Error
Seasonality /periodicity
Censoring

Statistical design: gathering requirements

Statistical design: gathering requirements

What are the topics of interest?
What are the key concepts being investigated?
What key variables are required?
Who are the stakeholders – what are their interests (e.g. policy)
What is the frequency of outputs/outcomes to be measured?
How timely do the outputs need to be?
What form do the outputs take and how can/ will they be disseminated (Datasets? Reports?)
What is the population of interest? Is a control population needed for comparison?

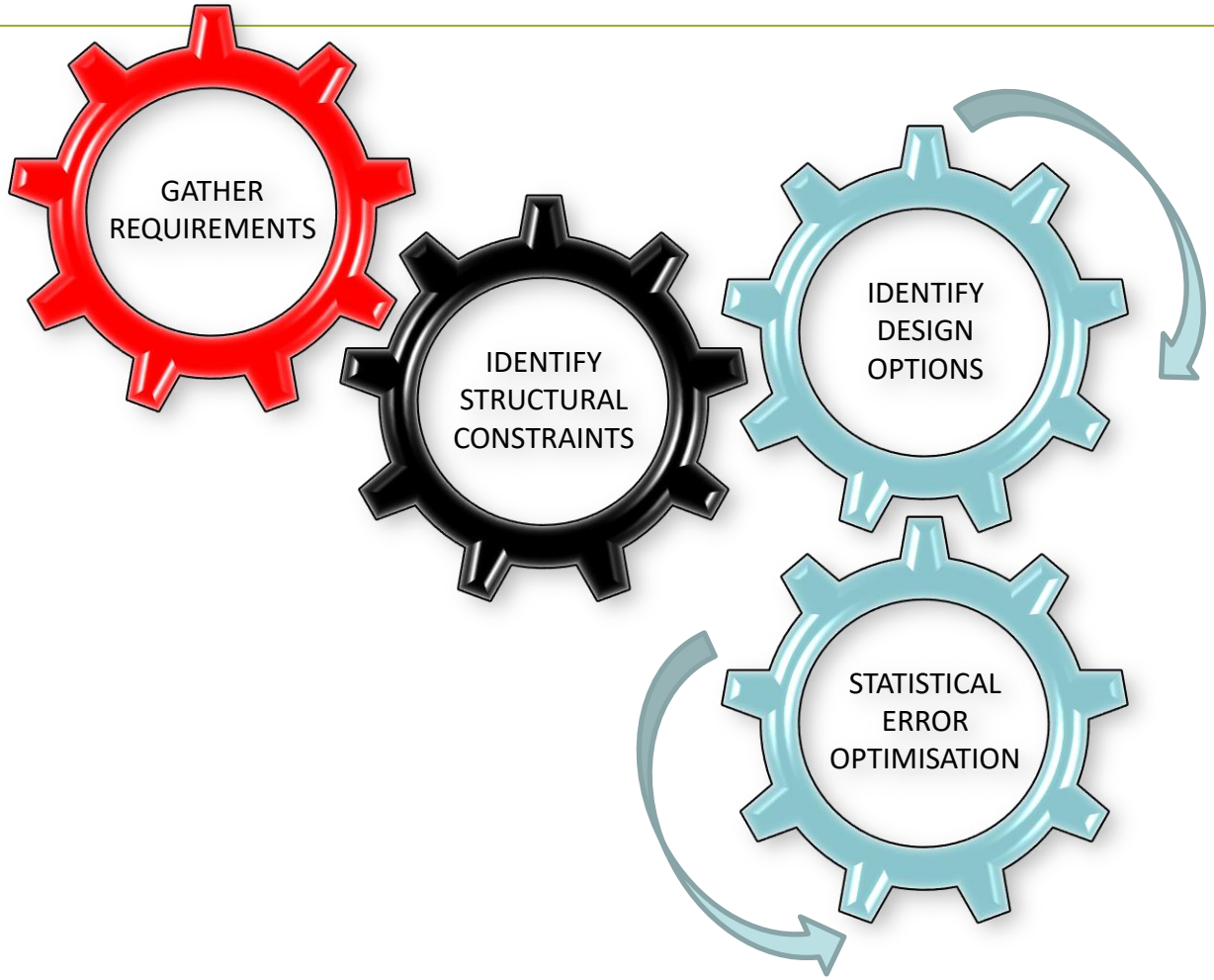
Statistical design: identifying the structural constraints



Statistical design: identifying the structural constraints

Do the data exist?
What are the known quality issues?
Timeliness and availability of time series
Geographical coverage
Population coverage and completeness
Topic coverage and completeness
Observability (hard to measure topics/populations)
Funding
Legislative and access
Ethical
Technical capacity
Scope and scalability
Coherence across sources, definitions and classifications
External harmonisation requirements
Availability and consistency of identifiers
Do we have measured quality indicators for the sources?

Statistical design: balancing design and error management



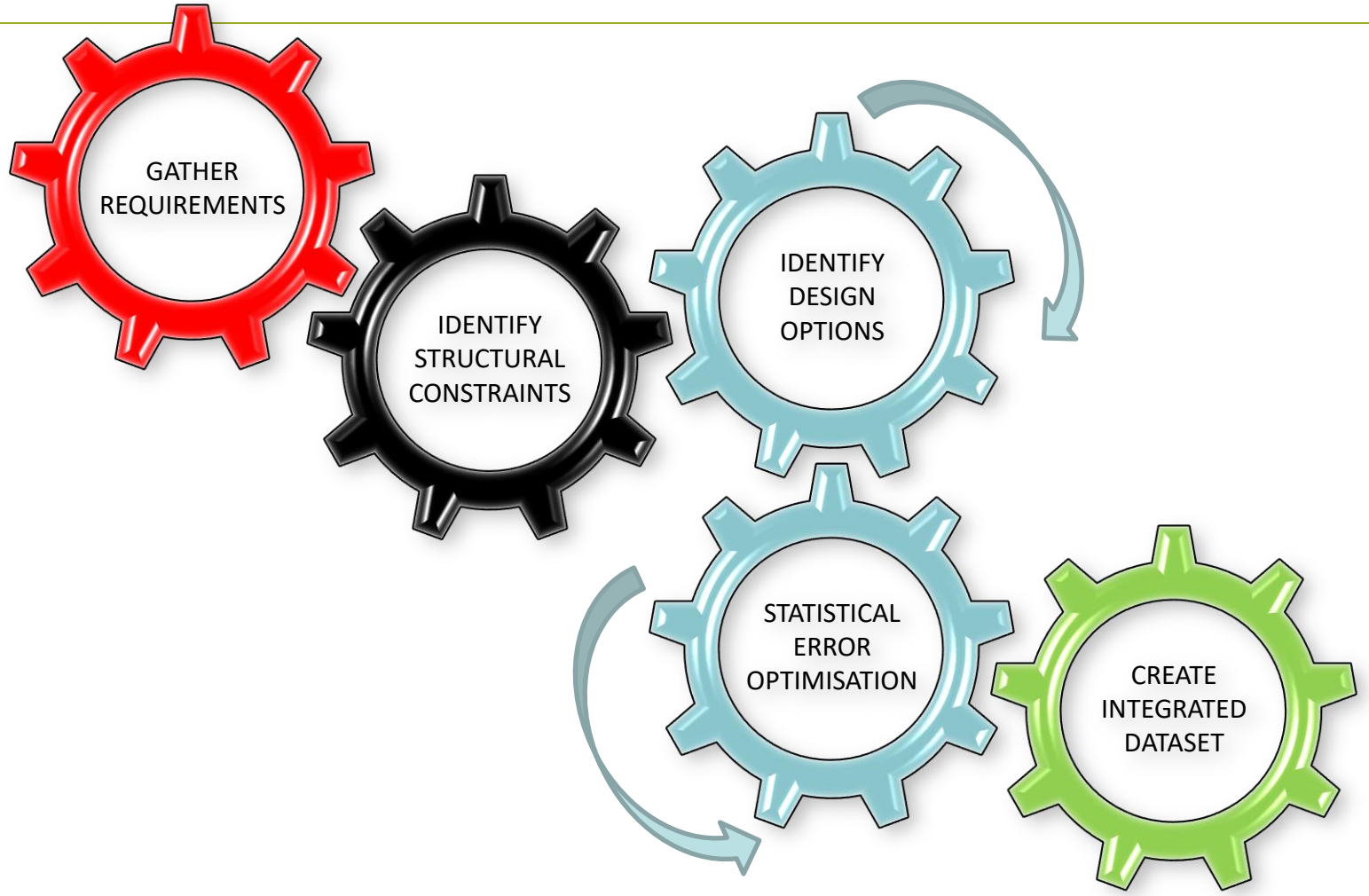
Statistical design: balancing design and error management

Design decisions:

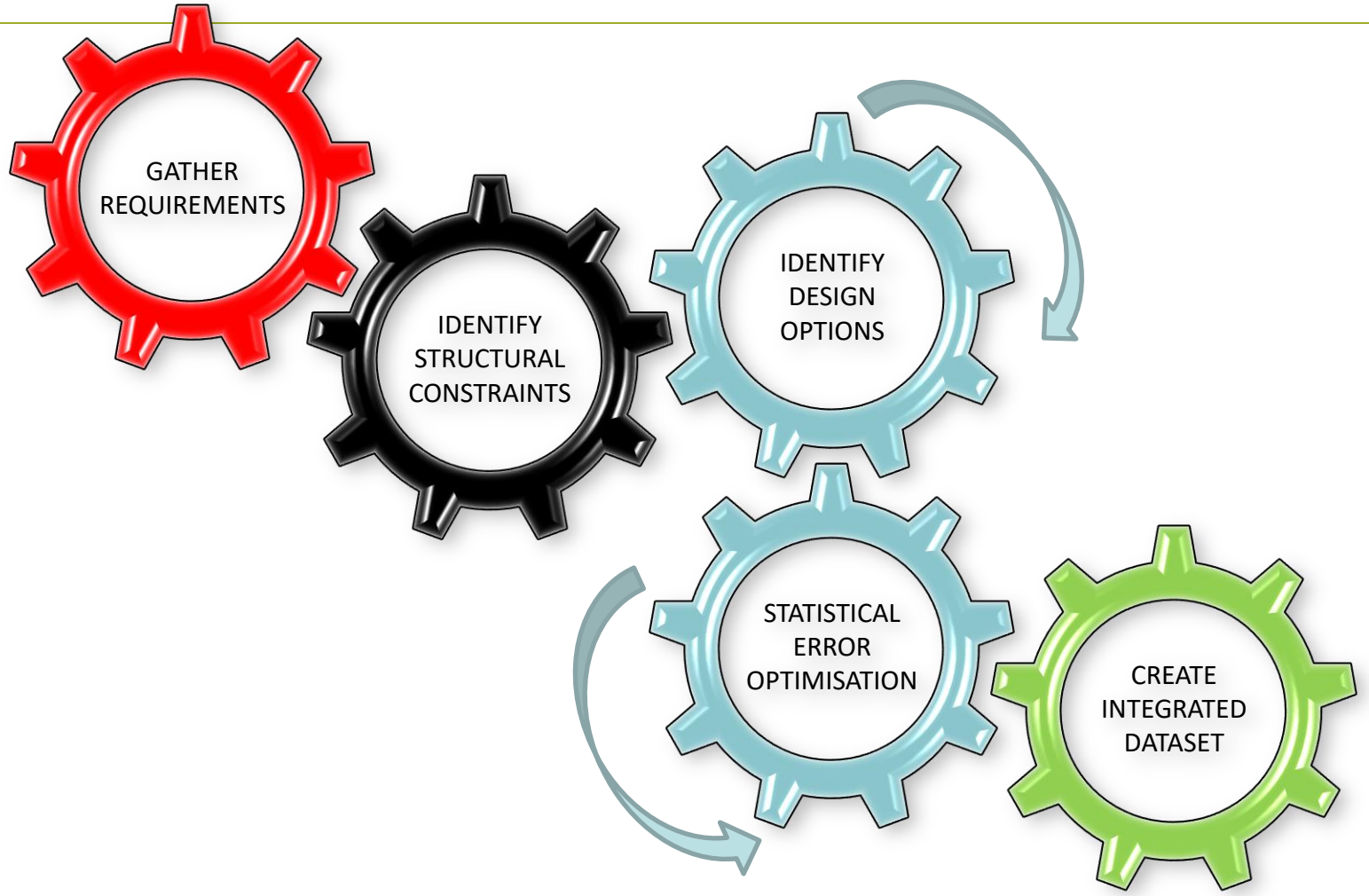
- Total population?
- Admin vs survey vs big data vs qualitative
- Sampling design and strategy
- Mode of collection (e.g. paper or online)
- Alignment of time referencing
- Periodicity – continuous vs repeated measures
- Granularity/disaggregation
- Retrospective vs prospective data integration
- Appetite for linkage error/quality management
- Sequence of matching
- ……..

User needs	Design options	Error
What trying to measure: Research questions Coverage and granularity Statistical precision	**Assess against criteria:** Fit for purpose Timeliness Replicability Relevance Coherence …. etc	**Assess against:** TSE for admin data + longitudinal dimension

Statistical design: implementation



Statistical design: implementation



Further work in progress

- Event history analysis to understand migration patterns and characteristics and explore alternative definitions of a short-term/ long-term circular migrant
- Exploring the validity of alternative approaches to imputation of unobserved migration events
- ‘Signs of life’ research using additional data linkage
- A population estimation framework for linked administrative data

Acknowledgements

With special thanks to contributing authors:

Sarah Cummins and Anvitha Patalay, Office for National Statistics

and Professor Peter W Smith, Southampton University

Also special thanks to contributing teams at the UK Home Office including statisticians, operational researchers, immigration system extracts and data suppliers

Discussion

- We have struggled to conceptualise how error propagates through linked administrative data through time. Are there existing models or frameworks that we can draw on?

- To what extent can we / should we seek to manage the errors that are introduced before we receive the data? Are the issues for single sources and complex linked sources different?

- Under what conditions should we impute missing values in longitudinal data? Are there guidelines?

Any questions?