Unlocking the potential of the Integrated Data Infrastructure for research

RezBaz June 2026

Social Investment Agency
Toi Hau Tāngata

New Zealand Government
Te Kāwanatanga o Aotearoa

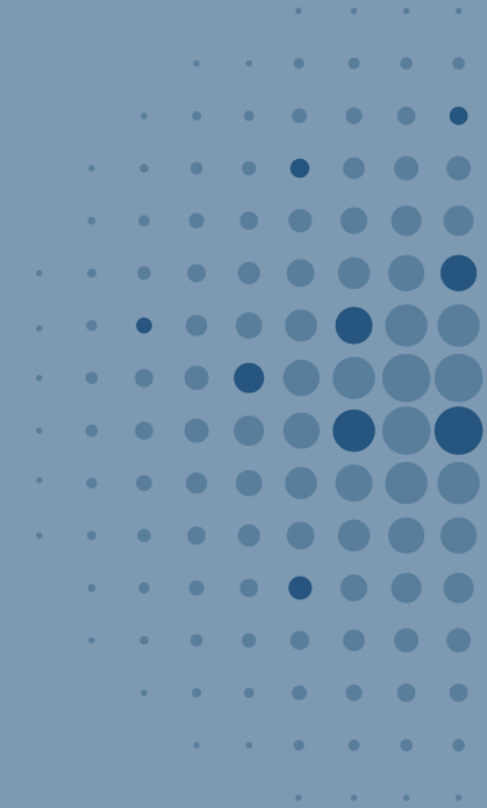
Insight arising from integration

Many insights arise from bringing together information that was once separate

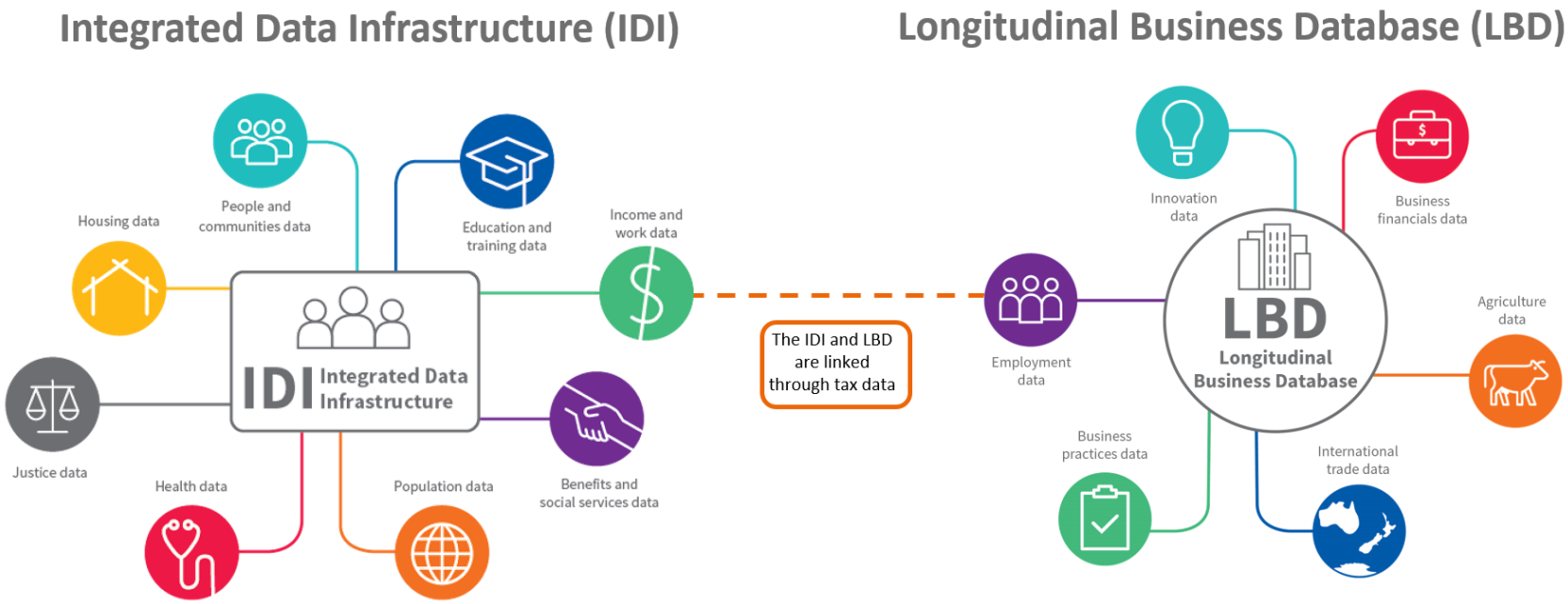
Integrated data is designed for this purpose

- What is the Integrated Data Infrastructure (IDI)
  - What is available and how it is integrated
  - Data protections in place
- What has been done with it
  - Examples of projects
  - How integration enables questions to be answered
- Some tips, tricks, and resources for using it
  - Significant data wrangling with steep learning curve
  - Range of resources to assist new researchers

What is the integrated data infrastructure?



World leading tool for research and analysis



https://www.stats.govt.nz/integrated-data/

The power of integrated data

Combine information

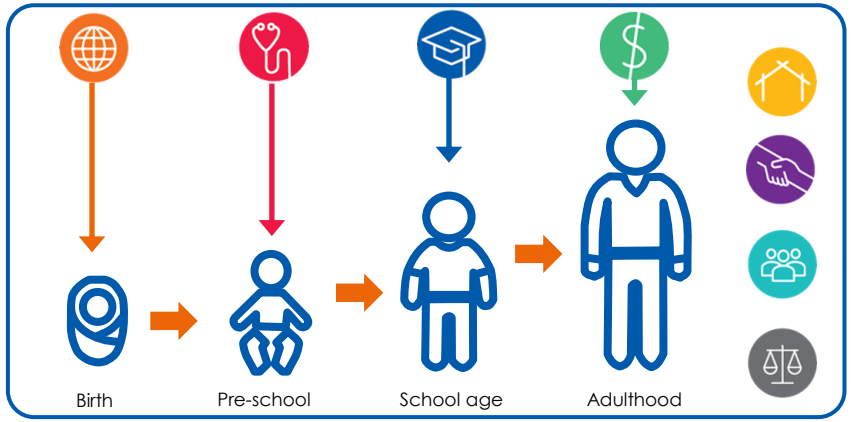
- from multiple domains
- from multiple life stages

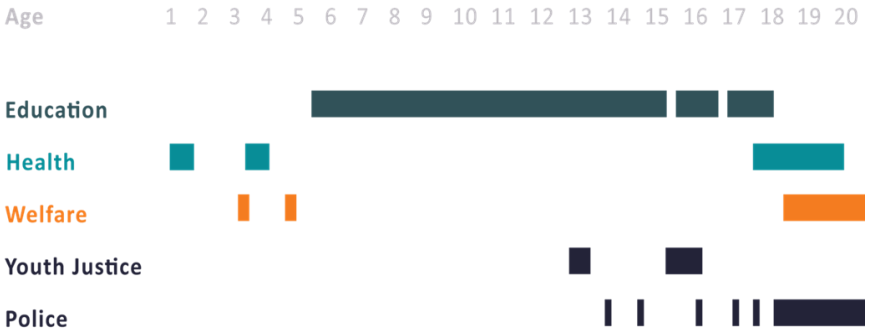
Construct detailed cross-sections

Follow experiences over time

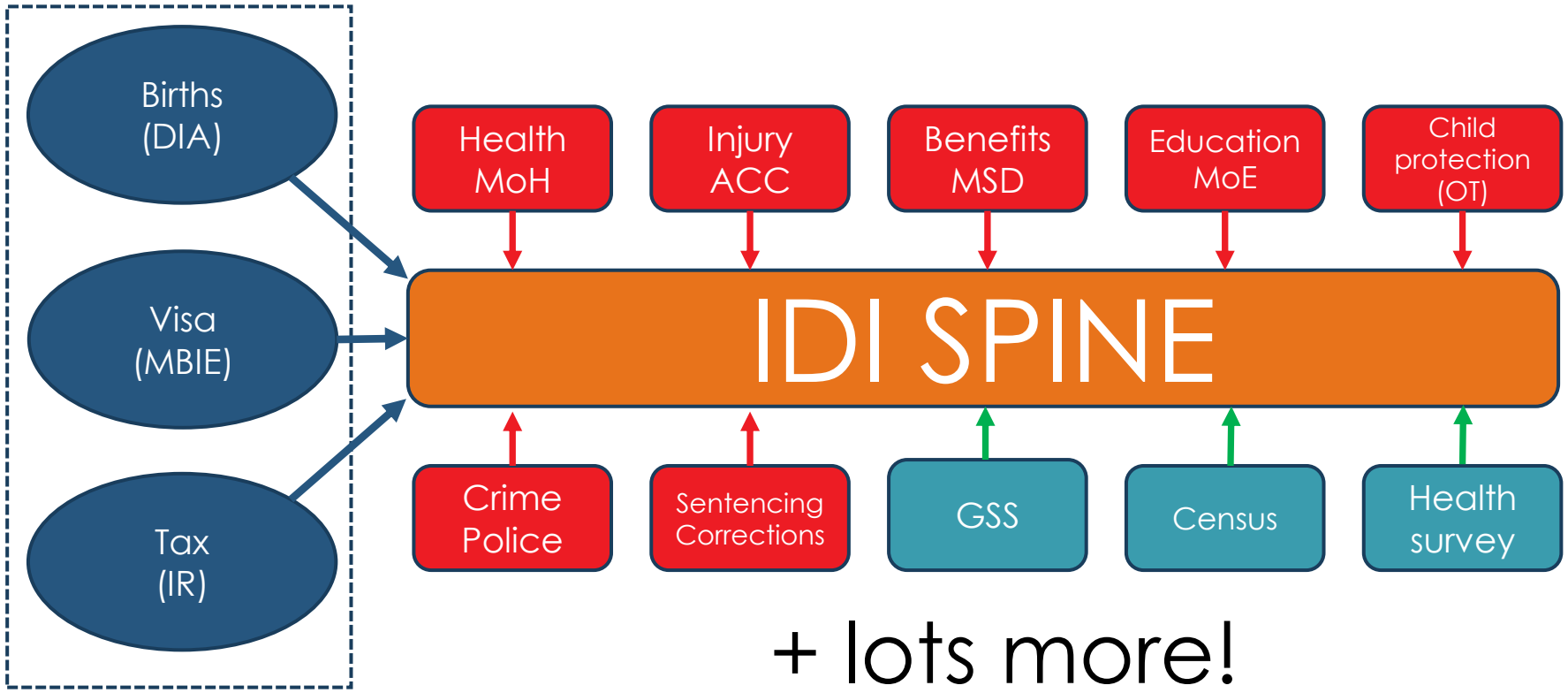
- Interactions between services
- Panel and longitudinal data

Predict or track outcomes





A list of all identities – the IDI Spine



Deterministic vs Probabilistic linking

Name: Joseph Blogs
DOB: 31 May 1971
Passport Number: ABCDEF
IRD Number: 123456

Health Records

Name: Joe Blogs
DOB: 31 May 1971
IRD Number: 123456

Name: Jo Bloggs
DOB: 31 May 1971

Name: Jo Blogs
DOB: 31 May 1971

IDI Data sets vary in history and currency

Pre-1985 2000 2010 2026

Health

Education and Training

Benefits and Social Services

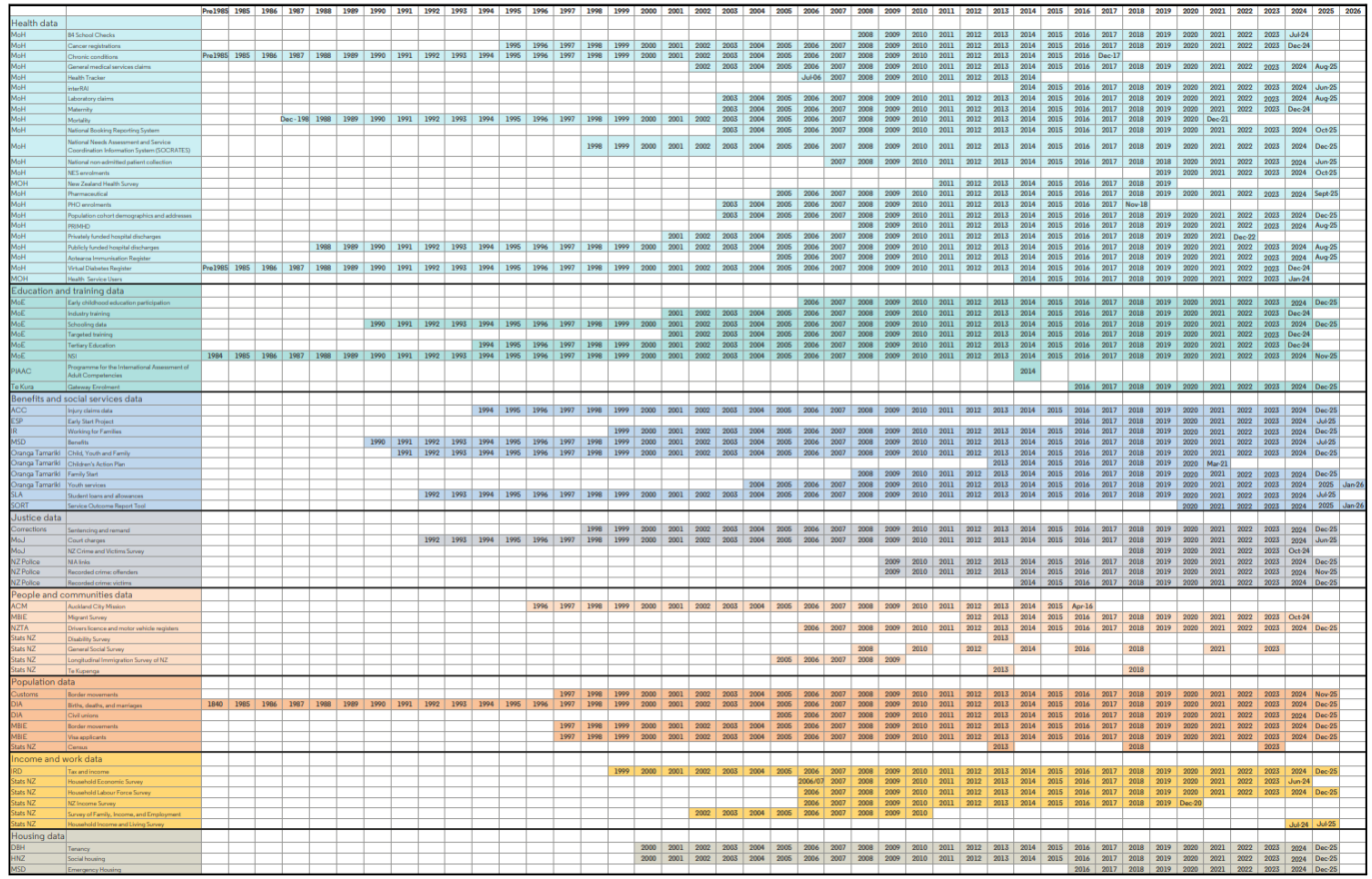
Justice

People and communities

Population Data

Income and Work

Housing Data



https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/data-in-the-idi/

‘Five Safes’ keep integrated data safe

Safe People

Only approved researchers can access or view microdata.

Safe Projects

Data can only be used for research projects in the public interest.

Safe Settings

Research takes place in secure data labs which Stats controls.

Safe Data

Access is granted only to the data that is needed for the research.

Safe Output

Confidentiality rules protect against privacy breaches.
All output is checked by Stats NZ to confirm it is safe.

Data is de-identified

Information supplied to Stats NZ

- **Name:** Star Thinker
- **Date of birth:** 29 February 1933
- **IRD:** 123-123-123
- **NHI:** 0123456789
- **Address:** 123 Enlightenment Terrace, Researchville

Information visible to researchers

- **snz_uid:** 4545454545
- **Birth month:** May
- **Birth year:** 1981
- **snz_ird_uid:** 111111
- **snz_moh_uid:** 22222
- **address_uid:** 99999
- **Meshblock:** 4507

Confidentiality rules limit data release

Microdata output guide specifies output rules

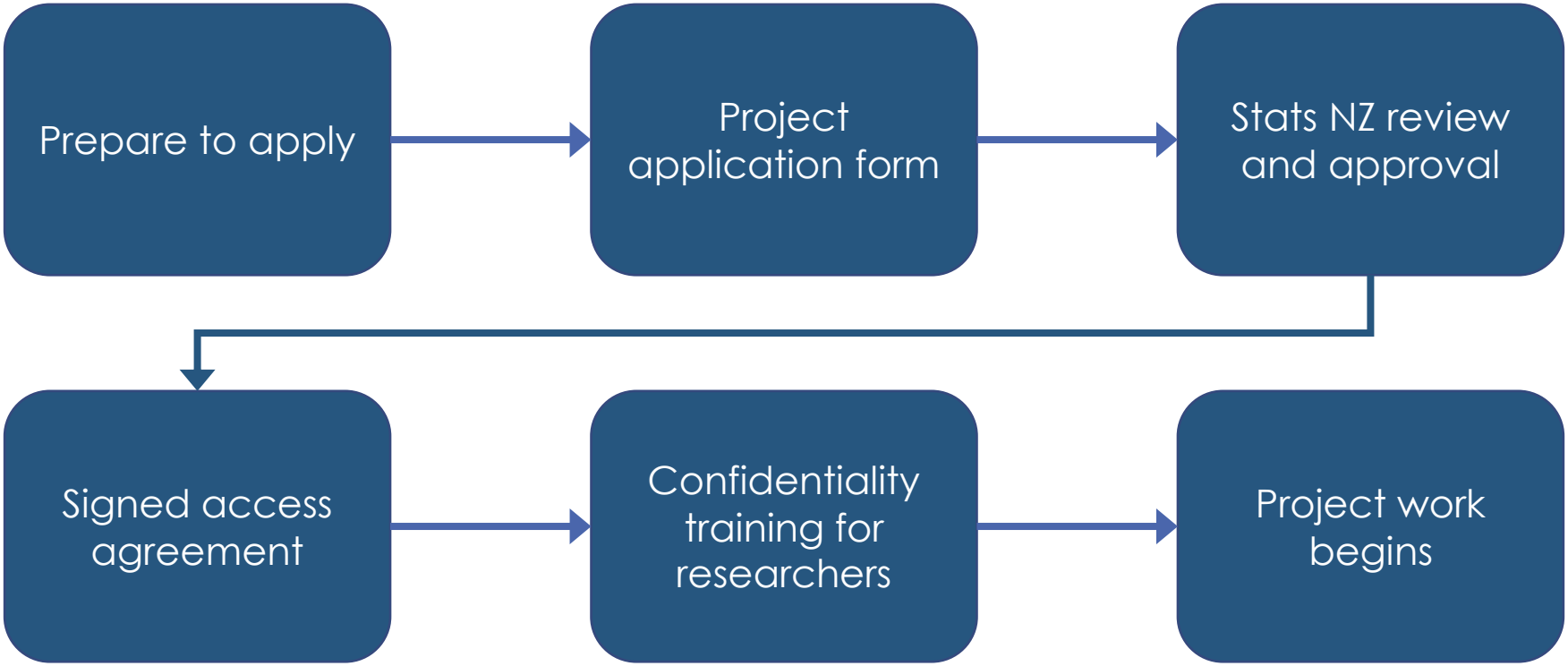
Four most common rules:

- Random rounding of counts to base 3
- Counts of fewer than 6 people are suppressed
- Totals for fewer than 20 people are suppressed
- Values that reflect a single organization are suppressed

Example process

Output	Raw	Released
People in Researchville	62	60
Total income for people in Researchville	$999999	$999999
Academics in Researchville	17	18
Total income for academics in Researchville	$777777	Suppress
Award winners in Researchville	5	Suppress
Total income for award winners in Researchville	$555555	Suppress
Employees of Top Research Inc	8	Suppress

Applying for access



Apply to use microdata for research
www.stats.govt.nz/integrated-data/apply-to-use-microdata-for-research/

Not perfect – has strengths and limitations

Rich range of data

Skilled and in-demand roles

Secure access

Enforced privacy protections

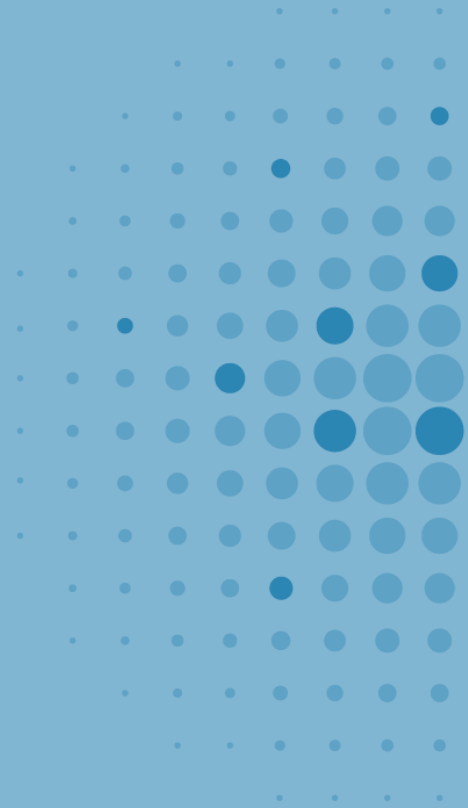
Range of data quality and documentation shortcomings

Technical capability and knowledge requirements

Administrative process
Barriers to collaboration

Small number studies limited

What has been done with integrated data?

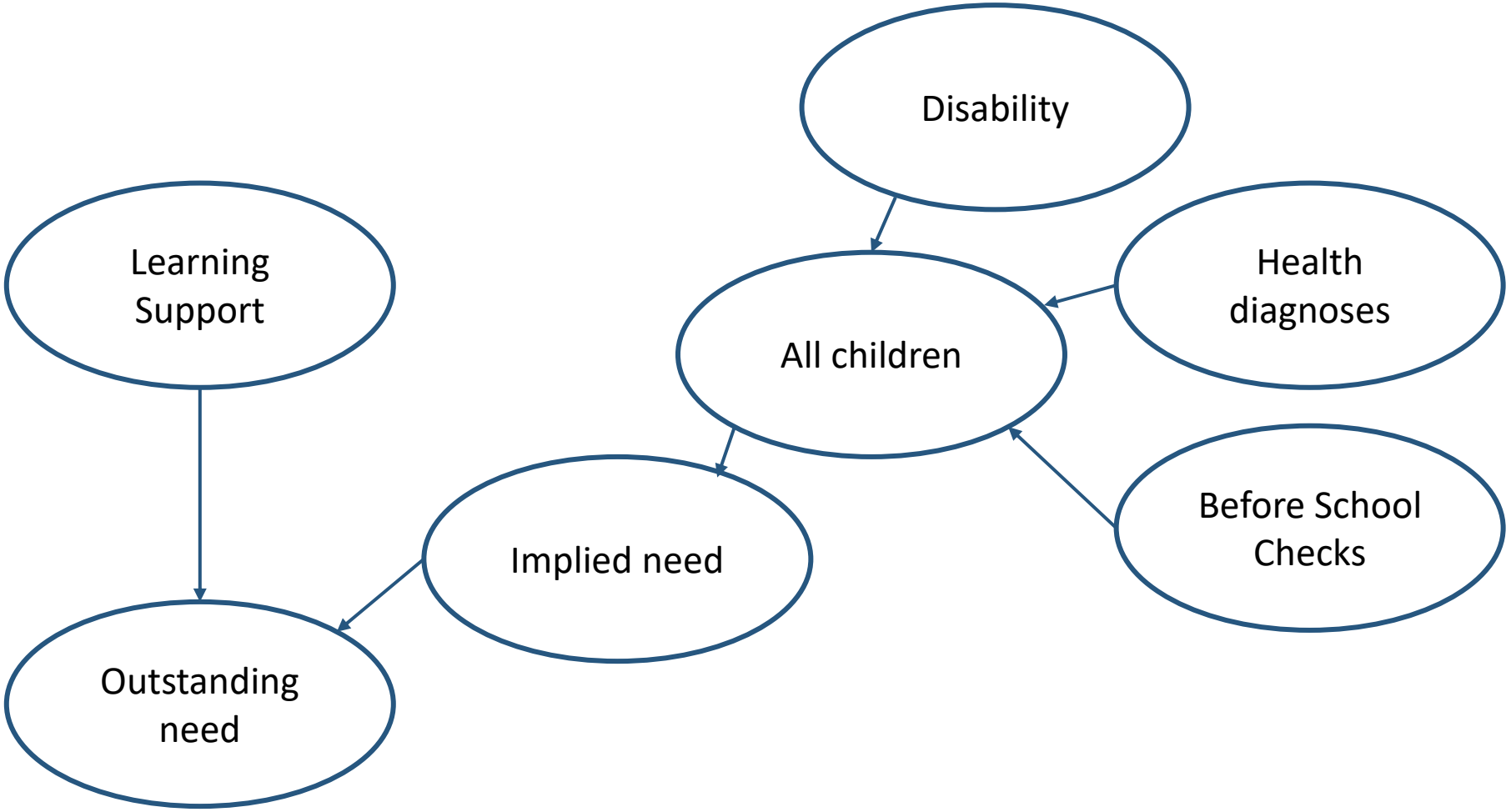


Types of questions the IDI is good at

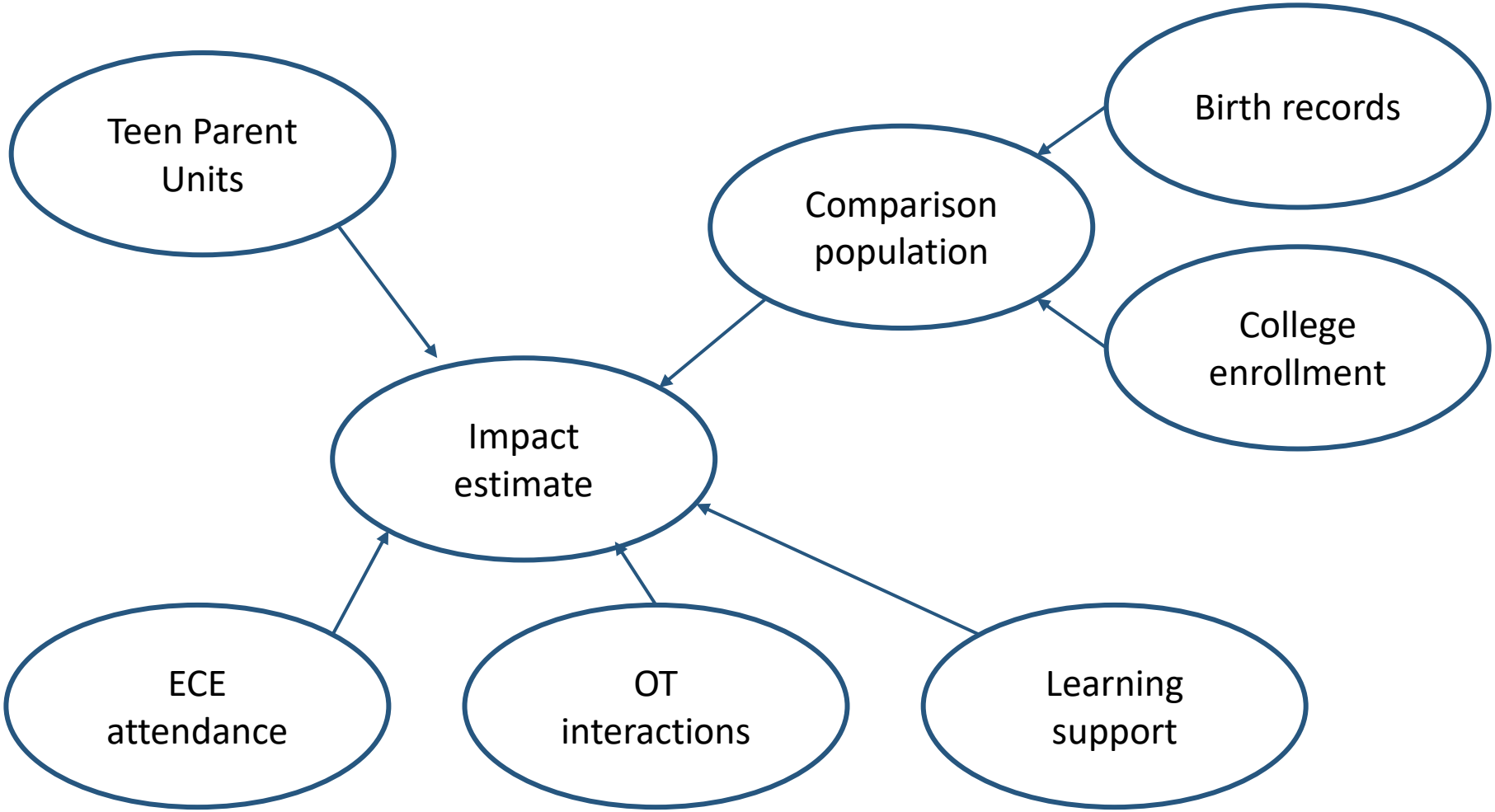
Descriptive | Inferential | Predictive

Overlaps	Unmet needs	Impact	Lifecourse	Simulation
What don't I know about my clients? Finding people who have an interaction with agency A (or other characteristic) and looking at what other agencies know about them.	*'Who else should we provide services to?'* The inverse of overlaps – who are the people who we don't see in agency A, but they look like agency A clients?	*'Are we making a difference?'* The richness of data in the IDI can help in identifying comparison groups; the longitudinal nature allows for follow-up.	'What leads to the outcomes we want (or don't want)?' Longitudinal analysis that could be defined by an experience/ outcome at the start, middle or end of period.	*'Where should we take action? What would be the likely impact?'* Many models exist in the IDI that can be used for forecasting or 'what if' analysis.

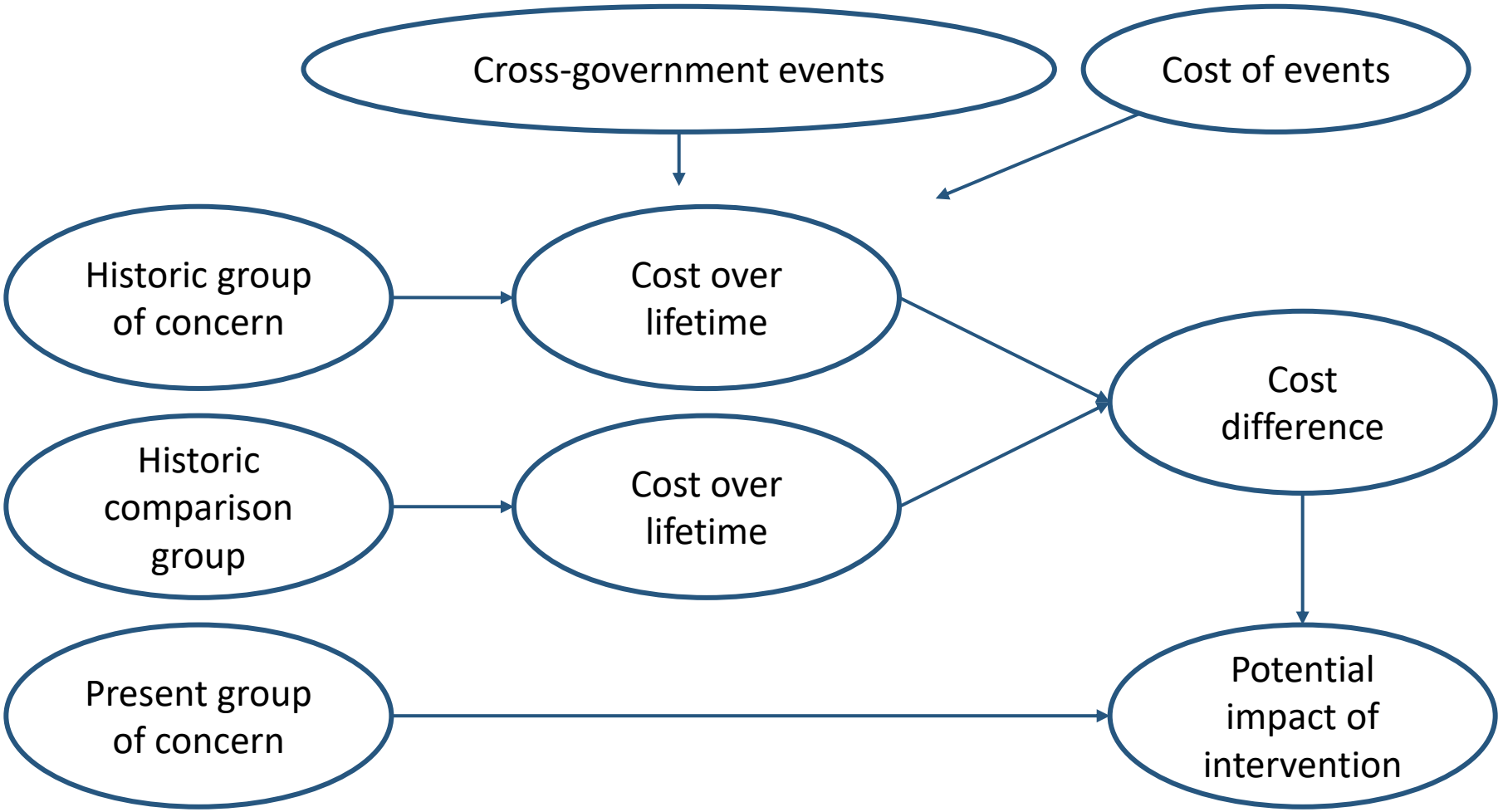
Highest learning needs review



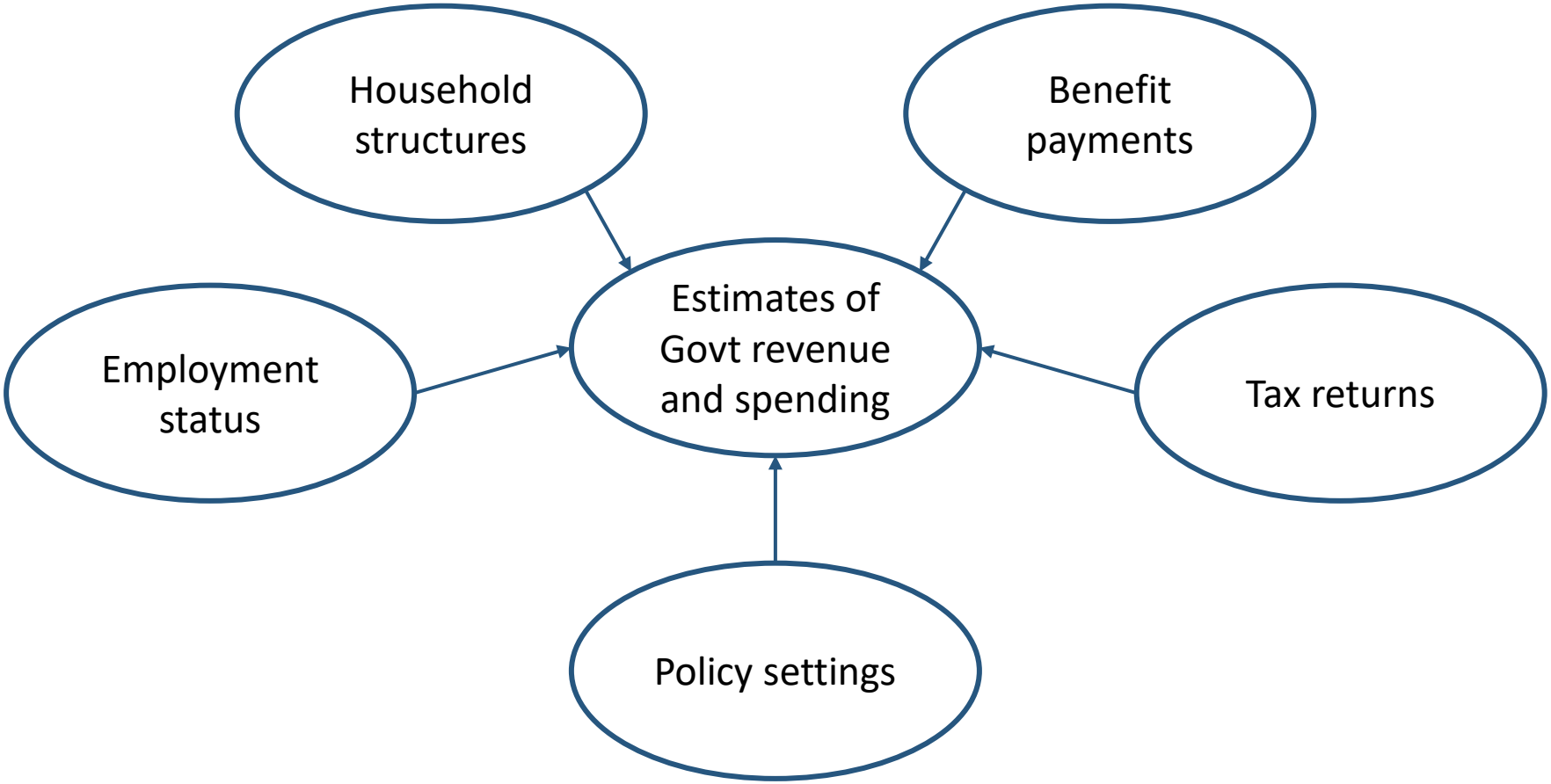
Long-term impact of teen parent units



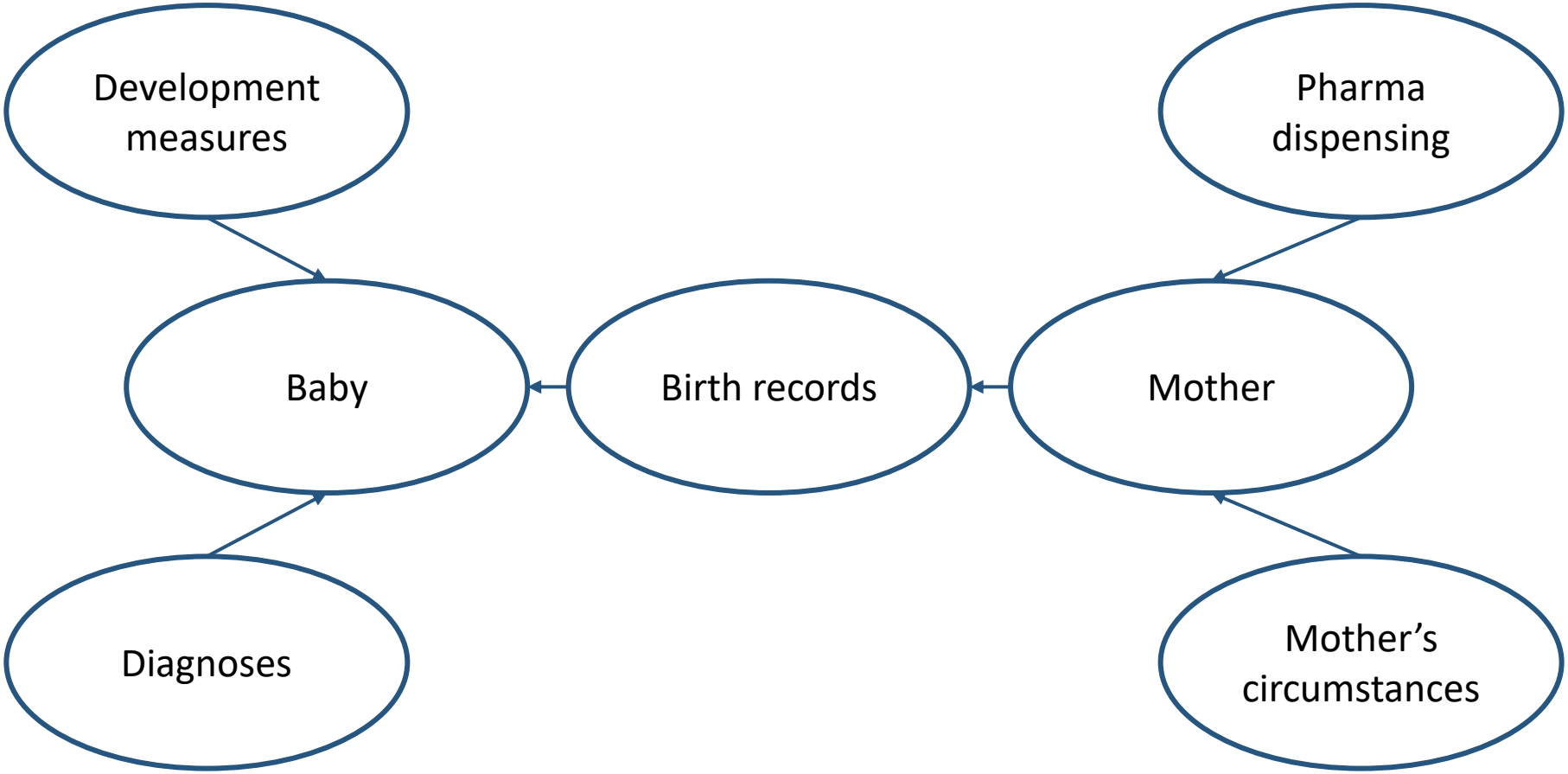
Changes in trajectory



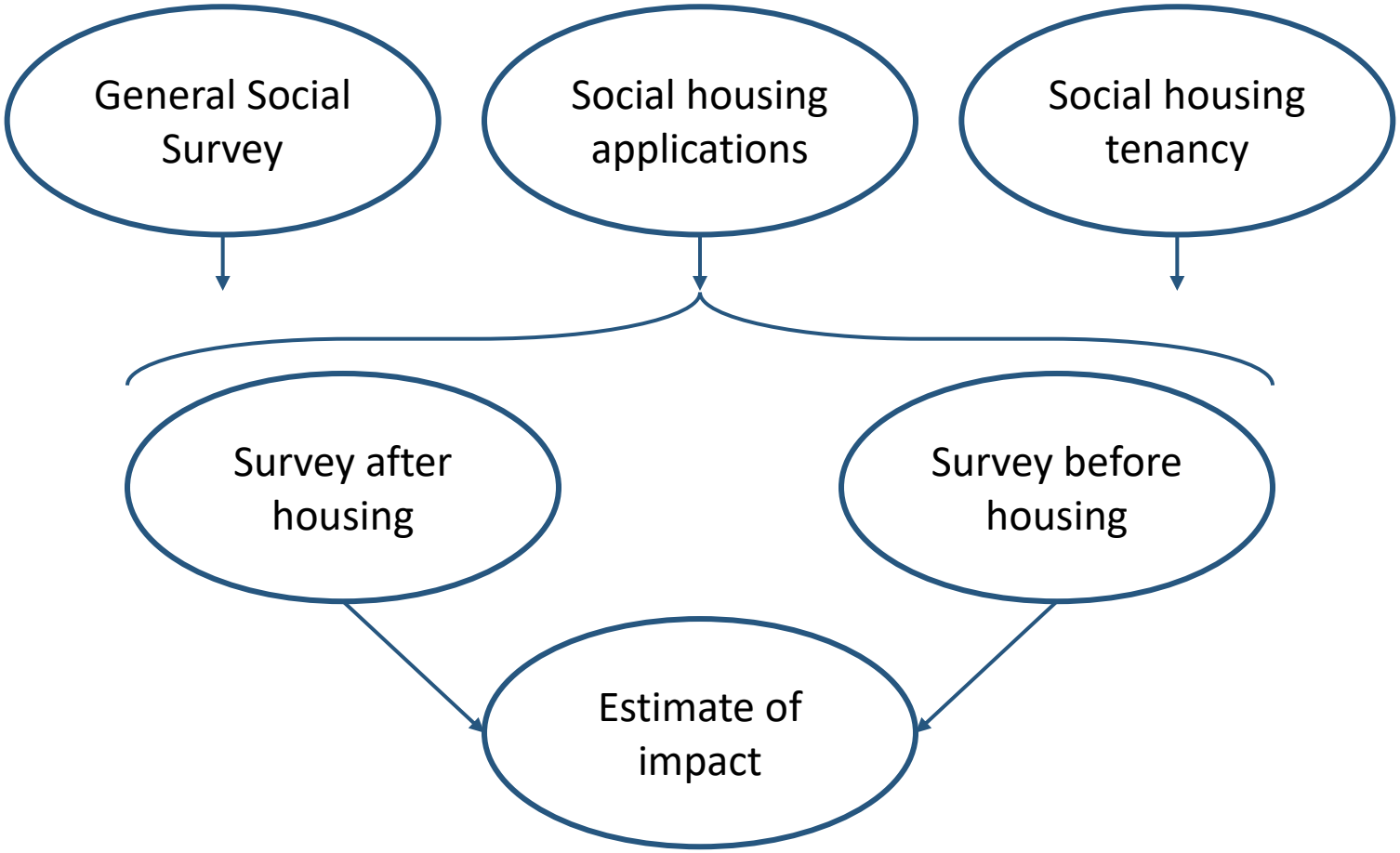
Tax policy microsimulation modelling



Effect of maternal antibiotic use of babies



Wellbeing impact of social housing



What advice makes it easier to use integrated data?

Lifecycle of an IDI project (our experience)

Determine policy and research question

Define population and period of interest

Create indicators and build dataset

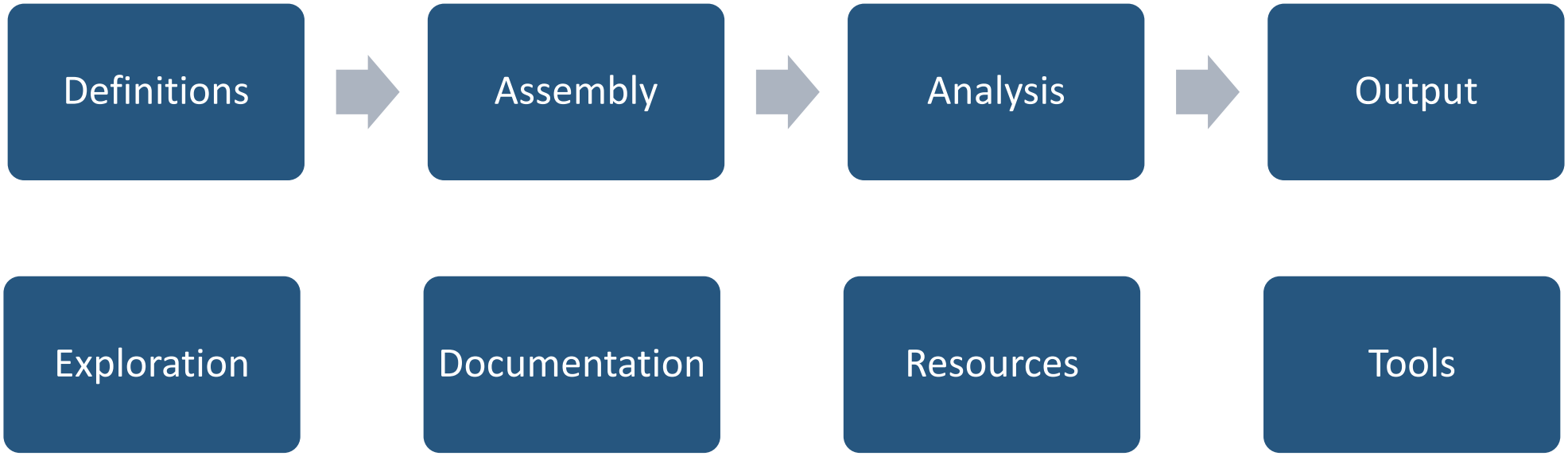
Actual analysis

Sense-making and implications for policy

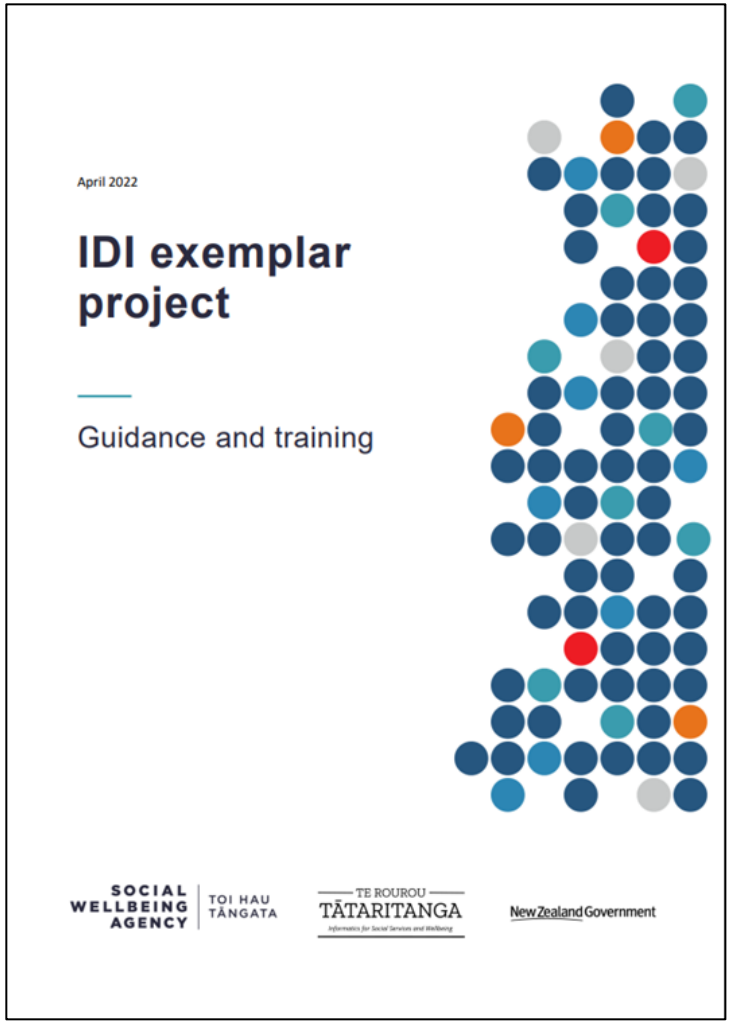
Communication of findings

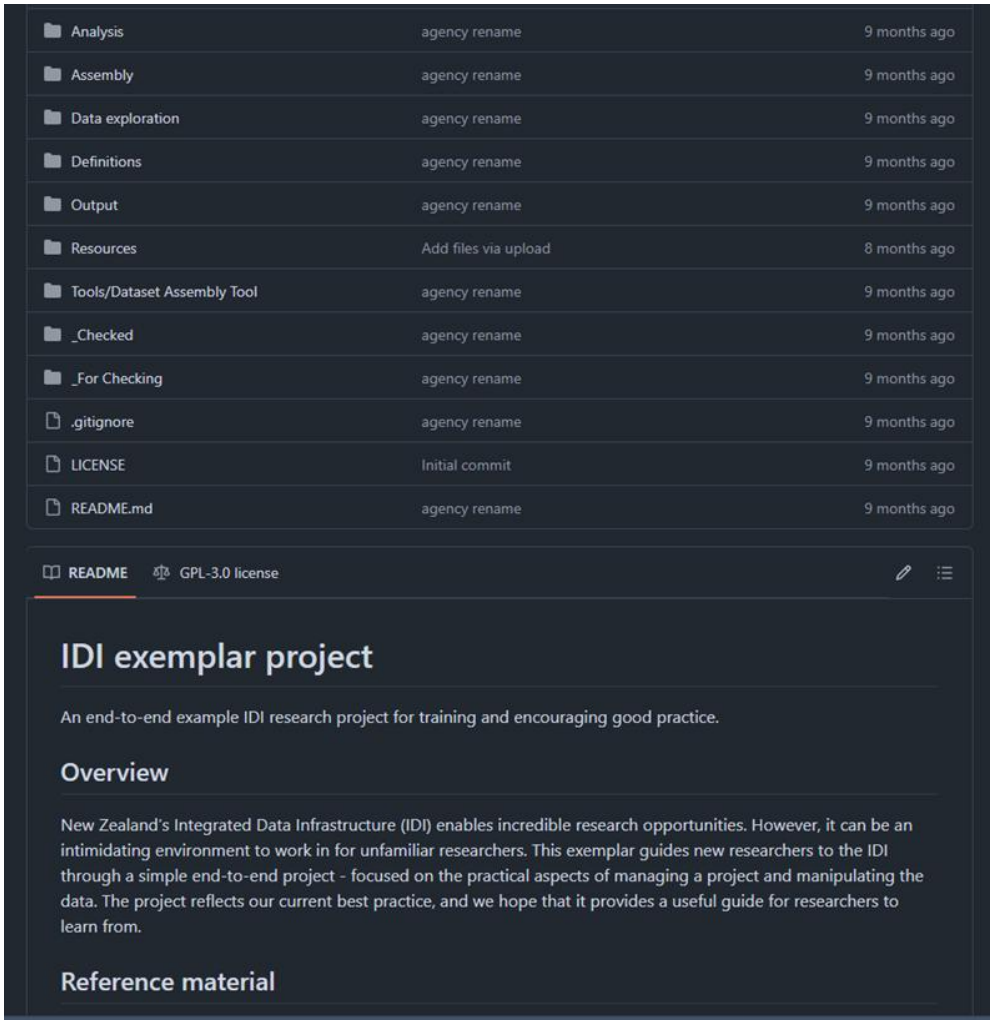
In IDI

Structure projects using consistent patterns



Start with a small project





Recalibrate expectations

Example task 1:

Compare self-reported life satisfaction against every other measure in the General Social Survey (GSS).

Non-technical perspective:

Concern that large number of crosstabs will be time consuming to create.

Analytic approach:

Quick and straightforward.

Only one input table, already arranged for analysis. Repetitive processing done by computer not researcher.

Example task 2:

Count the number of benefit recipients with children who have diabetes.

Non-technical perspective:

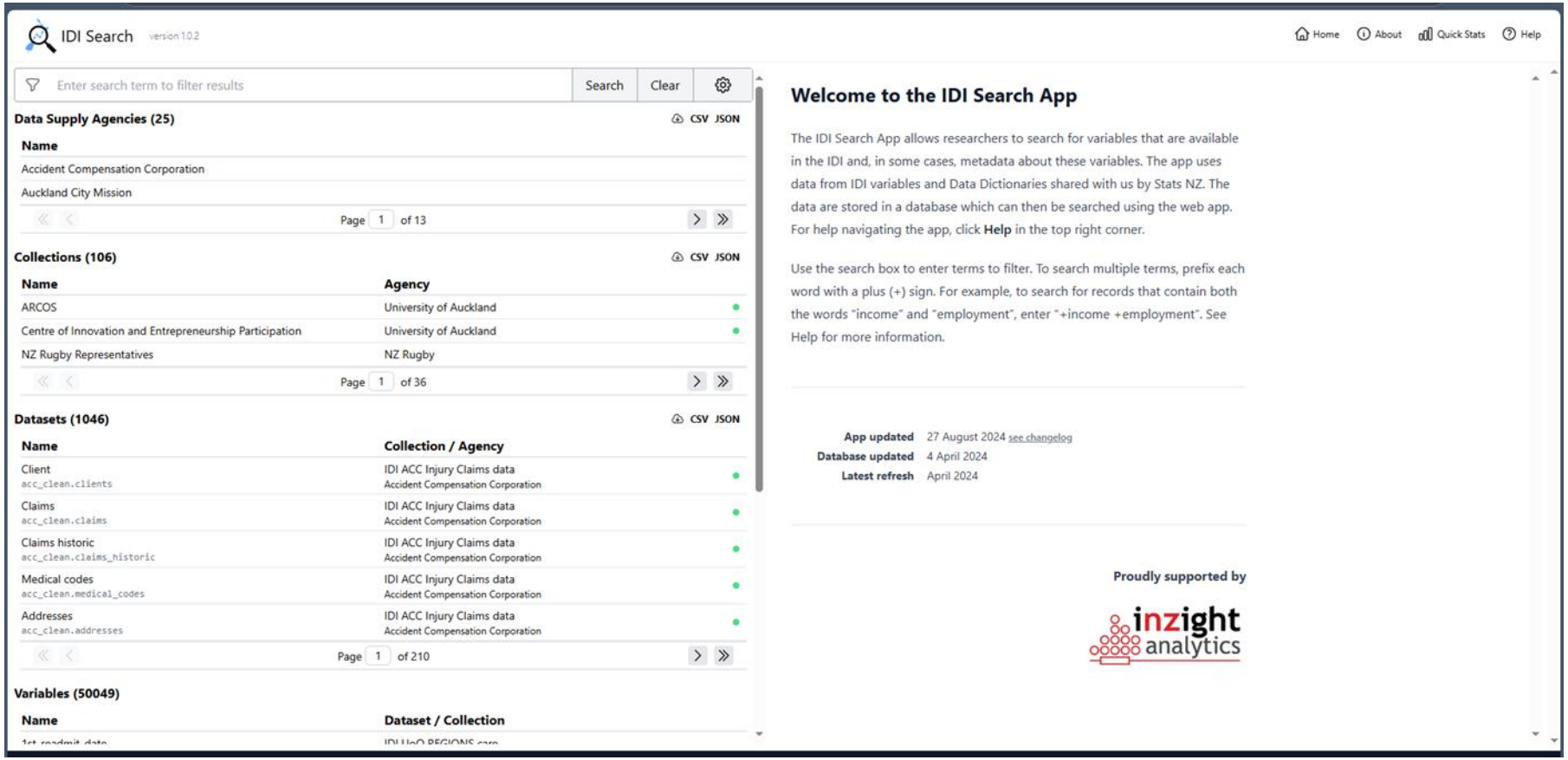
Straightforward as only a single number, benefit receipt, children, and diabetes are all unambiguous concepts.

Analytic approach:

Very challenging.

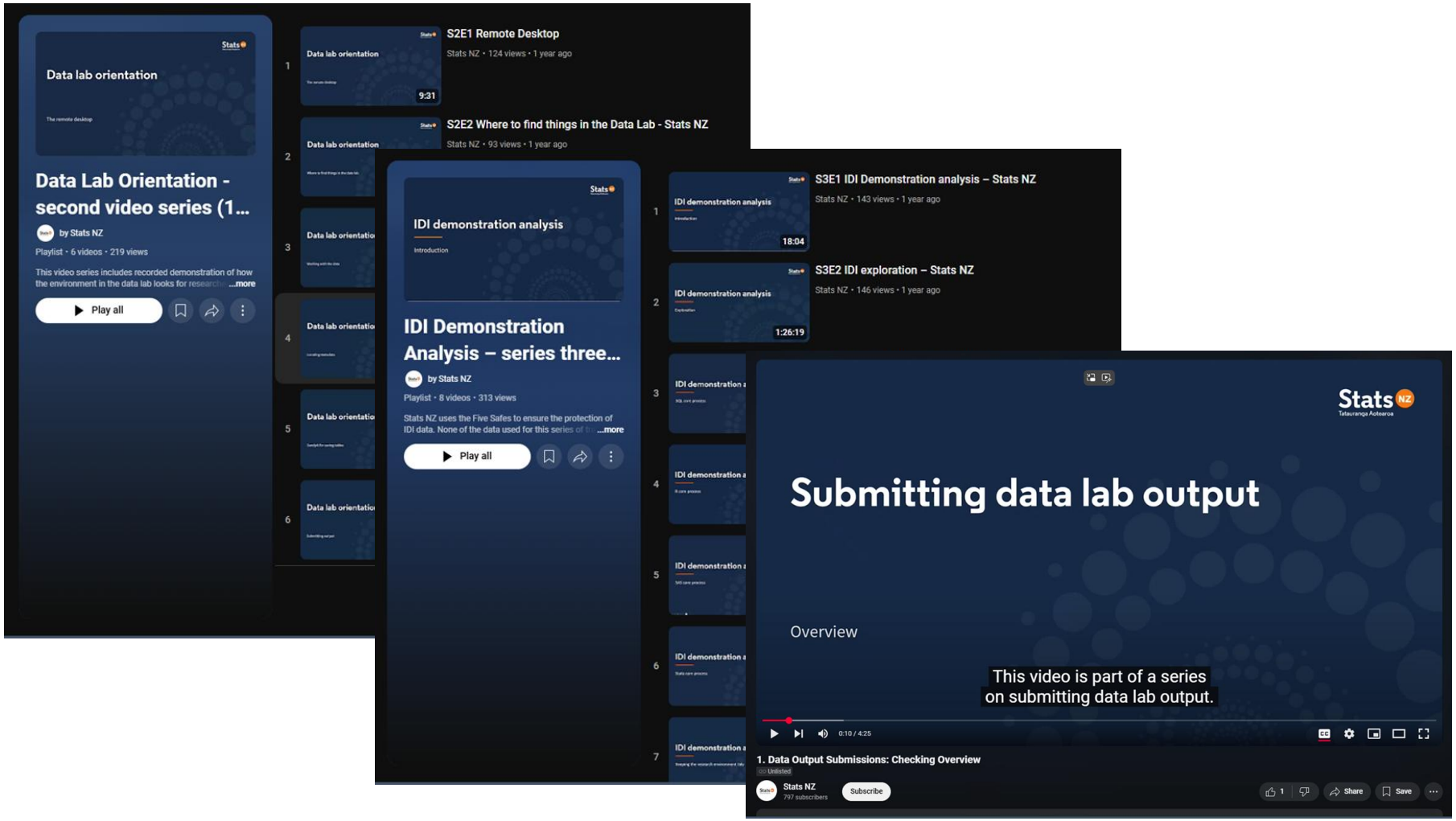
Diabetes must be constructed from a range of source tables. Multiple ways to define parenting status – may need to test and compare approaches.

Review metadata resources

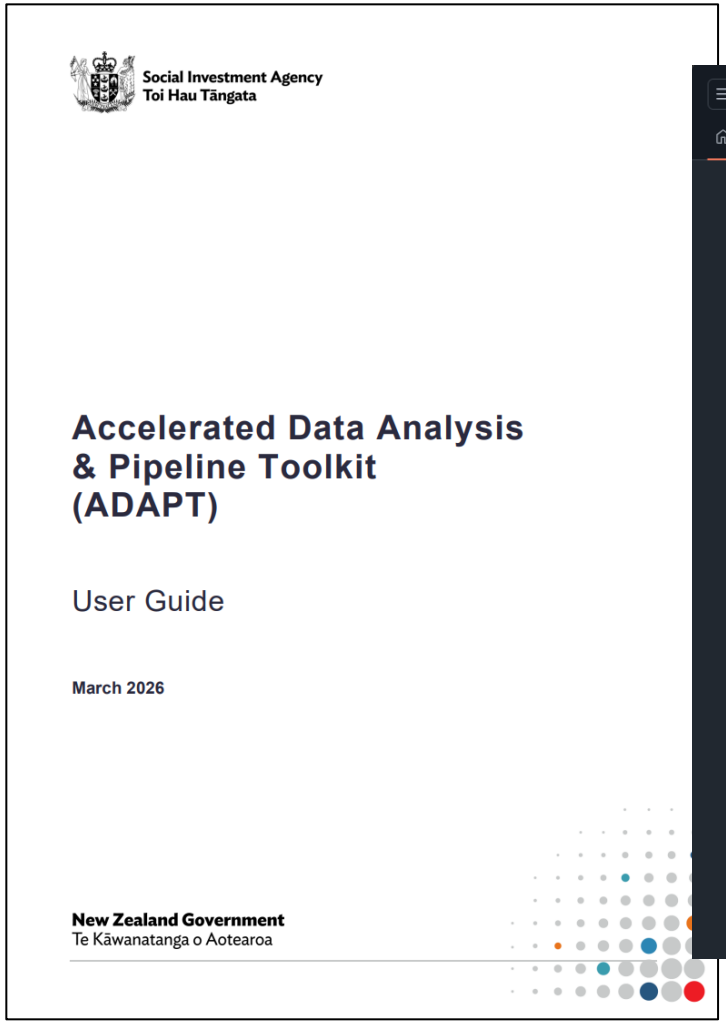


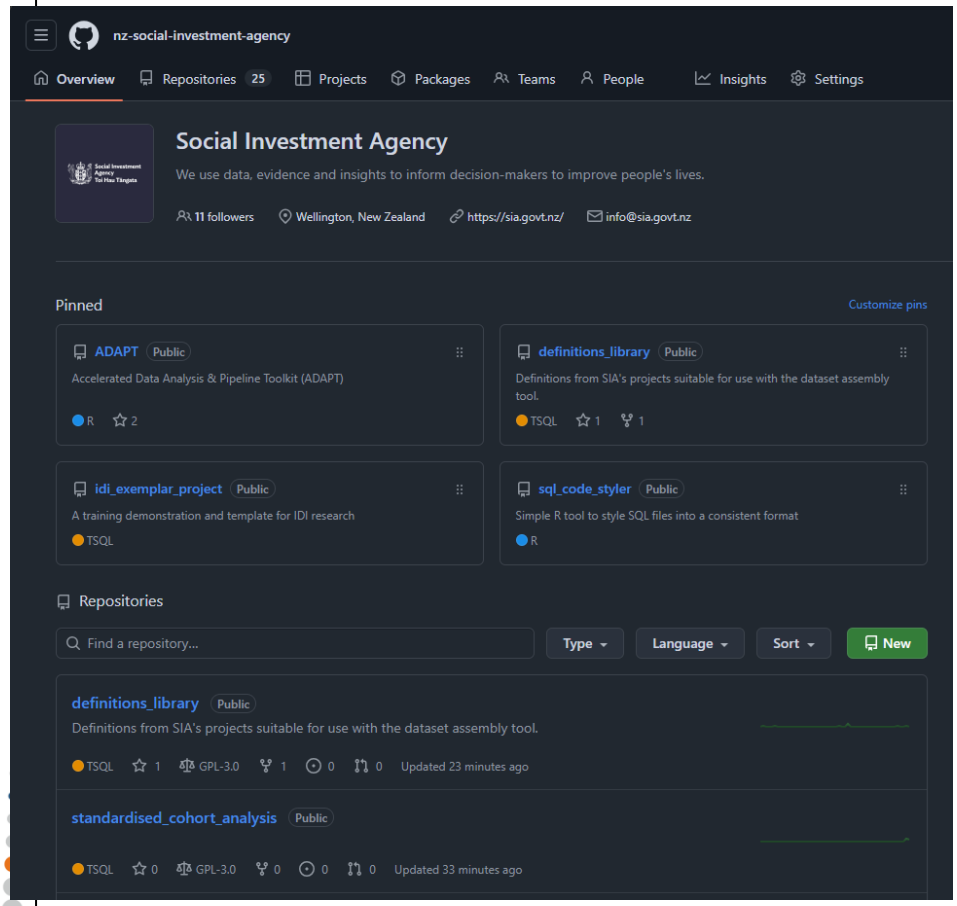
https://idisearch.terourou.org/

Video series are available



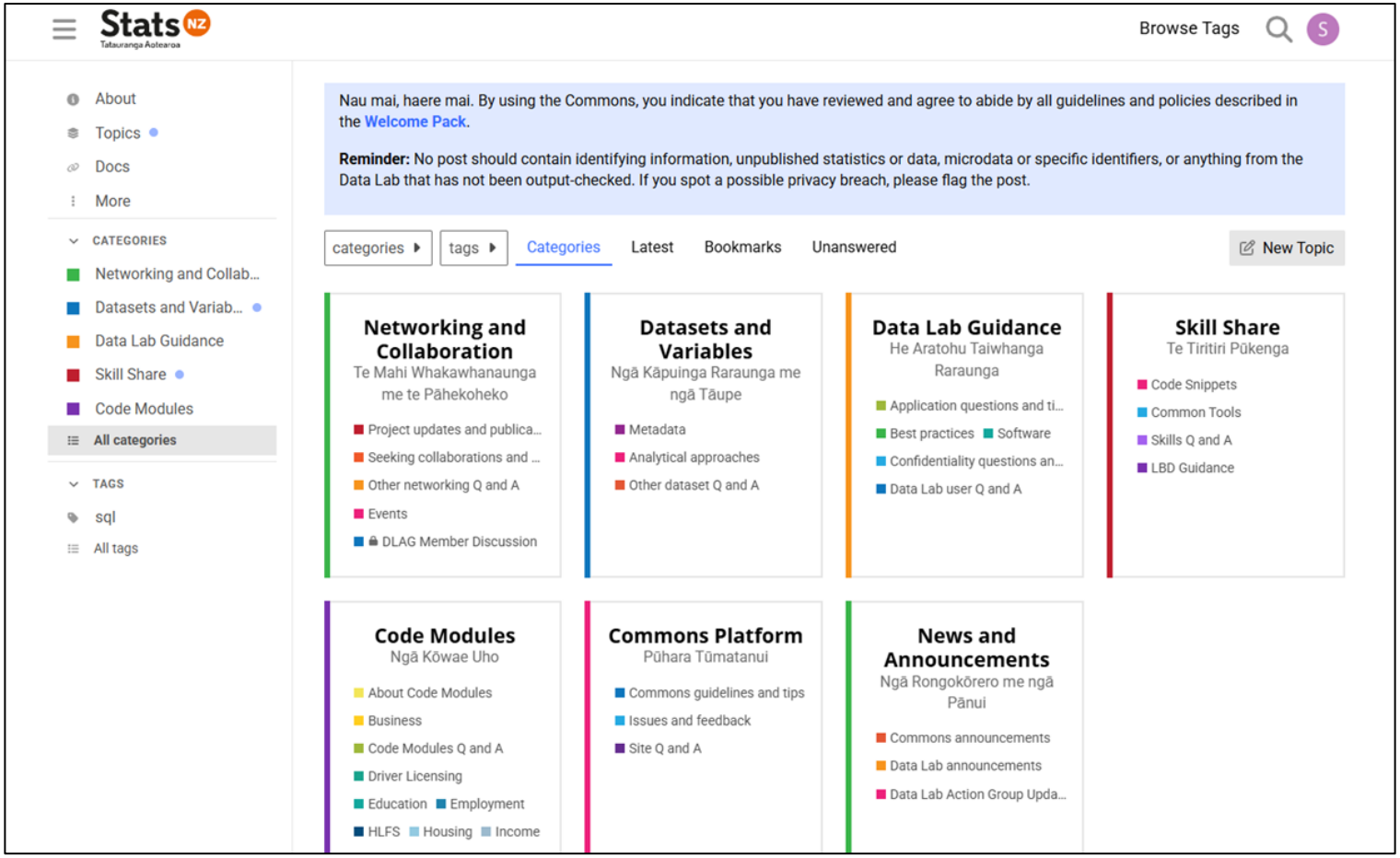
Build on existing tool and resources





https://github.com/nz-social-investment-agency

Connect with the research community



https://idcommons.discourse.group/

Plan for confidentiality rules

Design research within rules

- Estimate population size
- List subgroups you want to analyse
- Will every subgroup be large enough?

- Random rounding adds noise
- Will your results be robust with this noise?

- Track entity counts through analysis, difficult to add retrospectively

Make output process easy

- Stats NZ check 80+ submissions every week
- A small amount of extra effort on each output submission adds up to days of extra effort

- Spend a little more time ensuring submission is correct and clear
- Save the checker time and save yourself delays

- Watch the video series on good output practice before your first submission

Distinguish between different table layouts

Tidy rectangular source

ID	region	age	income
1	north	younger	200
2	north	older	400
3	north	younger	100
4	south	older	200
5	south	younger	100
6	south	older	0
7	south	older	400
8	south	younger	300
9	south	younger	400
10	south	younger	400
11	north	older	100
12	north	older	300
13	north	older	0
14	north	younger	400
15	north	younger	200
16	north	older	300

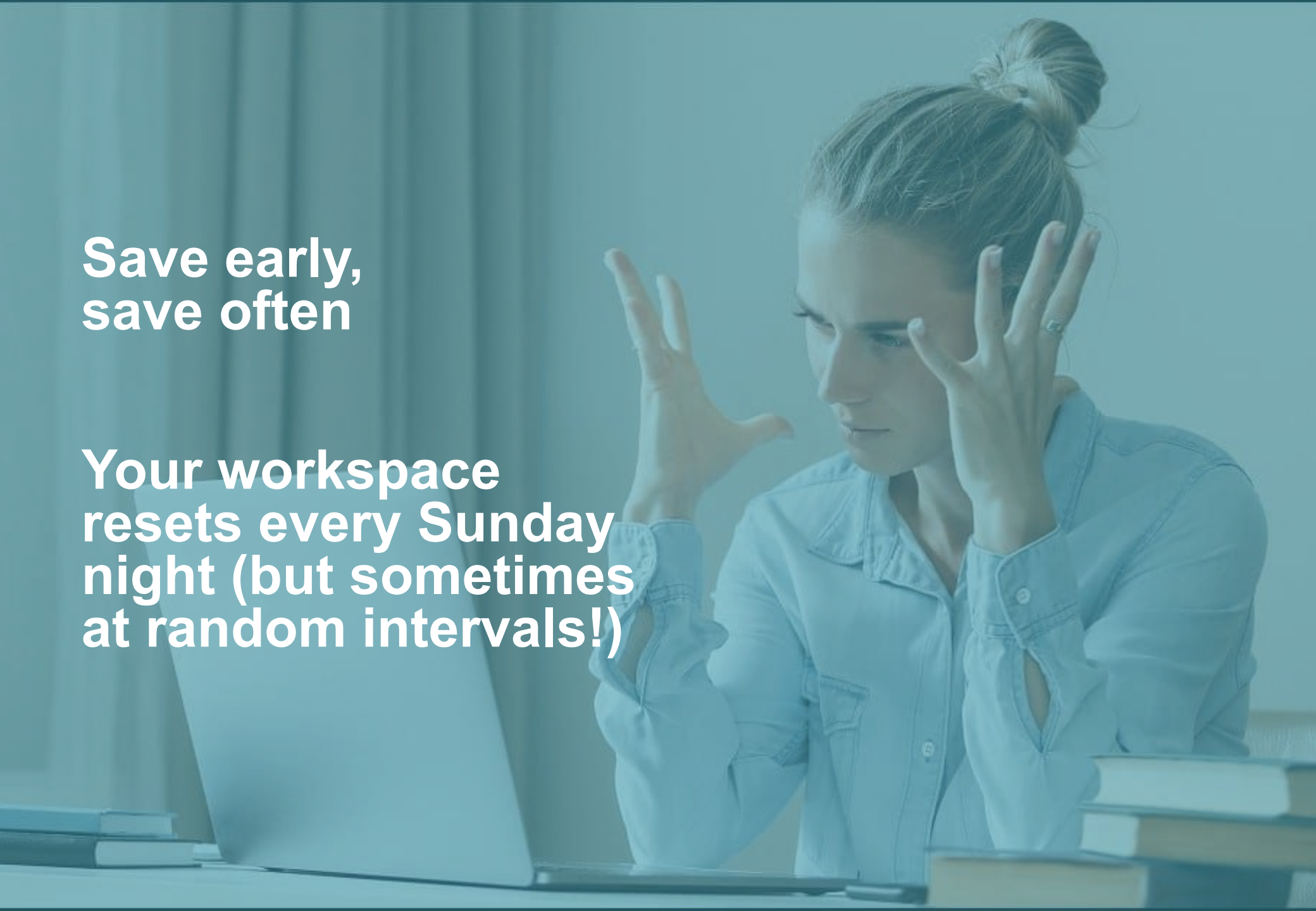
Long-thin results

region	age	count	total income
north	older	5	1100
north	younger	4	900
south	older	3	600
south	younger	4	1200
-	older	8	1700
-	younger	8	2100
north	-	9	2000
south	-	7	1800
-	-	16	3800

Presentation results

count	younger	older
north	4	5
south	4	3

total income	younger	older
north	900	1100
south	1200	600


Save early,
save often
Your workspace
resets every Sunday
night (but sometimes
at random intervals!)

Simon Anastasiadis

info@sia.govt.nz

sia.govt.nz

Social Investment Agency
Toi Hau Tāngata

New Zealand Government
Te Kāwanatanga o Aotearoa