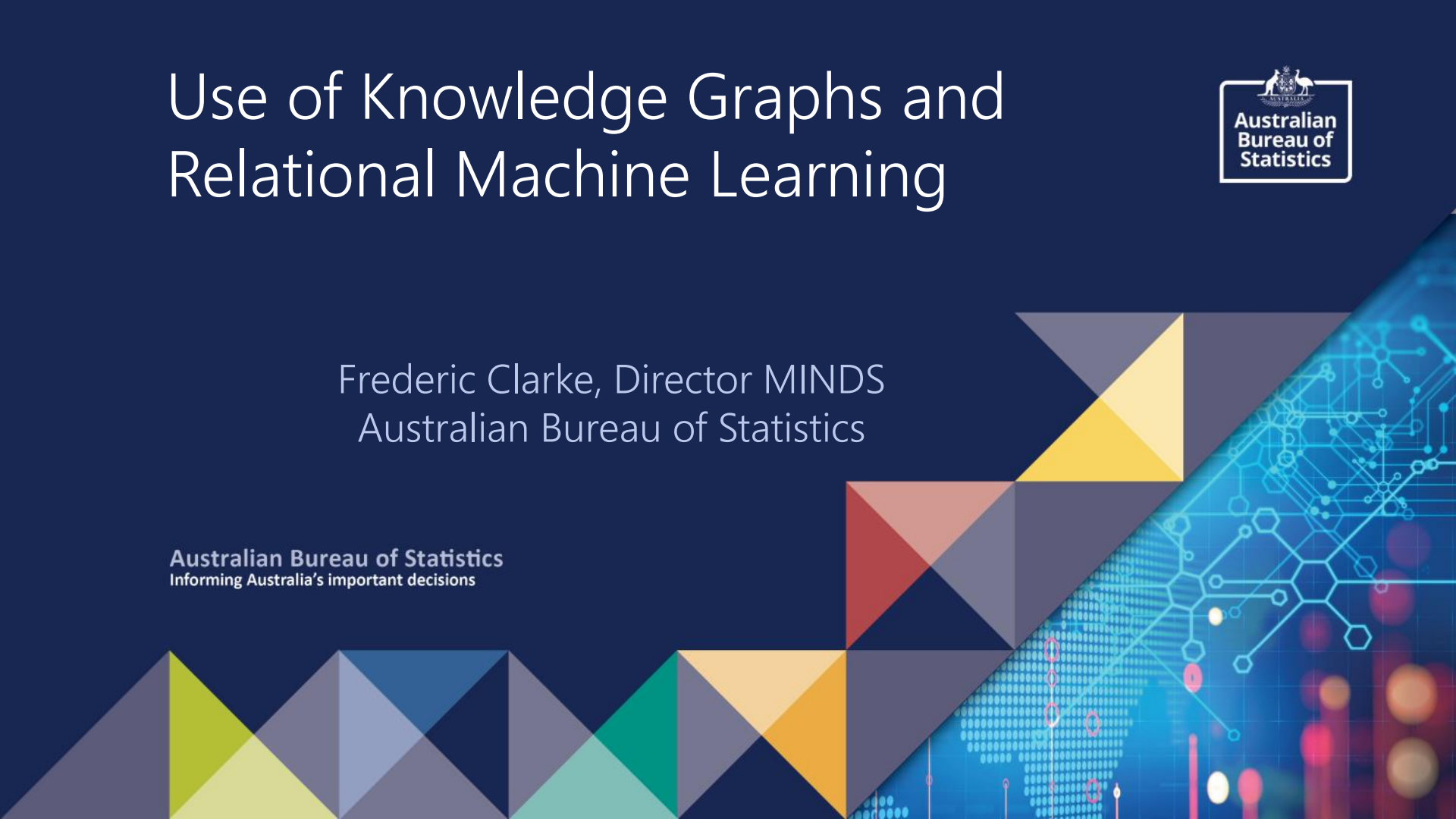
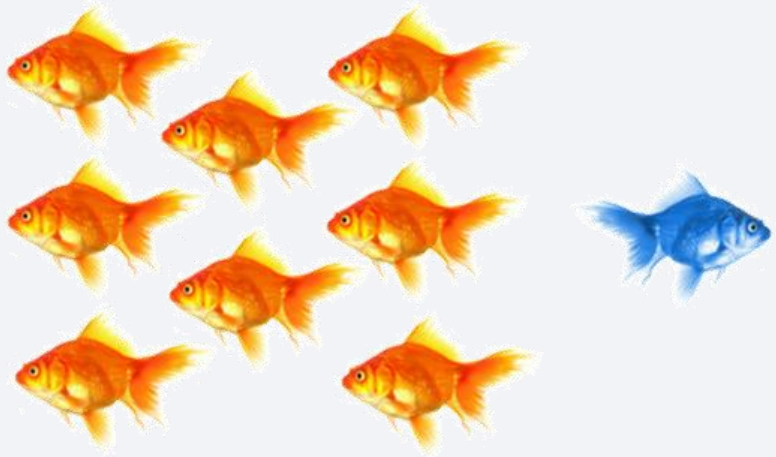
# Use of Knowledge Graphs and Relational Machine Learning

Frederic Clarke, Director MINDS
Australian Bureau of Statistics

Australian Bureau of Statistics
Informing Australia's important decisions



# Machine Intelligence and Novel Data Sources



## Multi-disciplinary team – maths, CS, KM

- Focus on hard problems, learn by doing
- Demonstrate solutions through prototypes

## Investigate, experiment, evaluate and inform

- Methods, technologies and models
- New data sources and applications
- ABS strategic capabilities and priorities
- Environmental trends, opportunities, threats

# Agenda

The strategic context

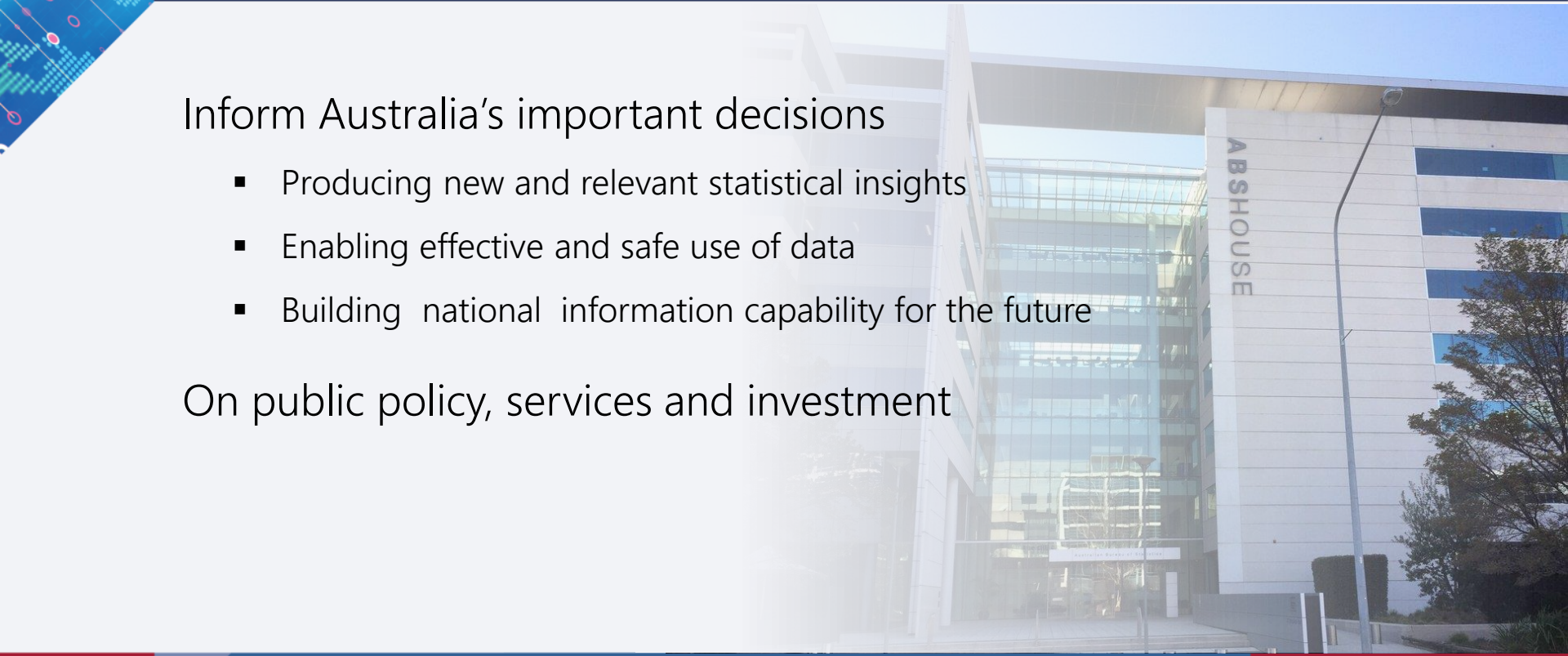
A motivating example

ABS use of KL and RL

GLIDE and current work

# Strategic context

# ABS Mission




## Inform Australia's important decisions

- Producing new and relevant statistical insights
- Enabling effective and safe use of data
- Building national information capability for the future

## On public policy, services and investment

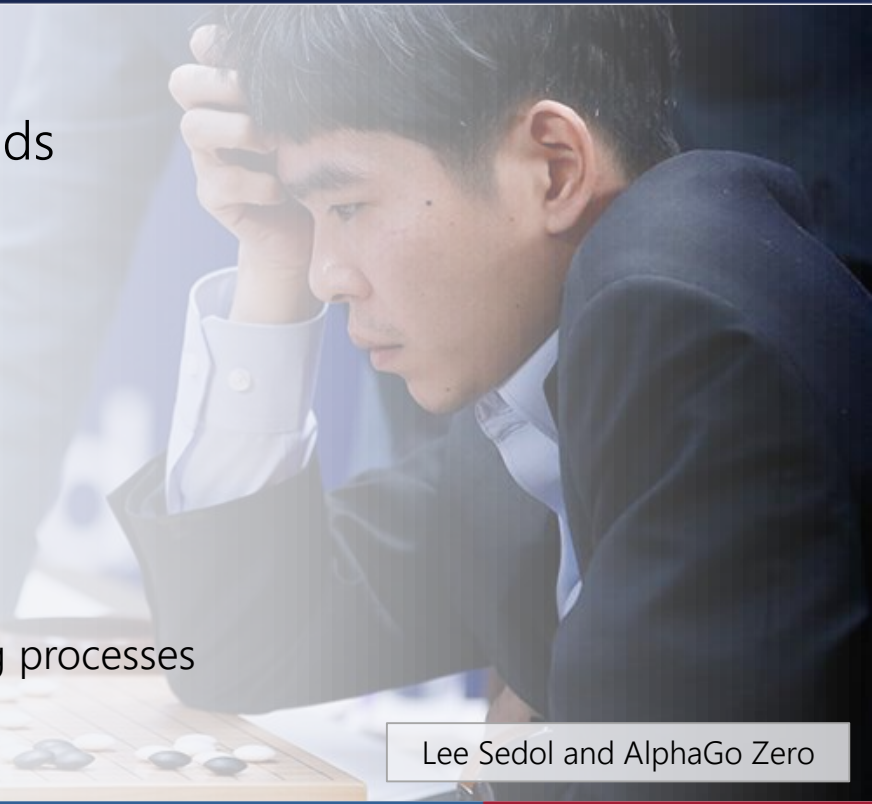
# Disruptive change

## Driven by powerful global megatrends

- Intelligent Machines
- Digital Connectedness
- Data-driven World

## Impact on government is profound

- Overturns extant business models
- Challenges traditional decision-making processes



Lee Sedol and AlphaGo Zero

# Complex systems

Most economic, social and environmental systems are complex

- Many interacting entities of different types
- Dynamic and non-linear relationships
- Emergent system structure and behaviour

Connectedness is an essential condition for complexity

Complex systems underlie most ‘wicked’ problems

# Big data

Personal Devices

Imagery Systems

Smart Meters

Product Scanners

Environmental Monitors

Telematics Devices

Diverse new sources of human-generated and machine-generated data

Can be used for statistical purposes

- Improve quality, time-to-market and cost-effectiveness of mainstream statistical products
- Create new statistical products that fill information gaps

Issues: heterogeneity, bias and dimensionality

ABS operates in an increasingly congested and contested environment

Web Applications


Social Media Services

Logistics Systems

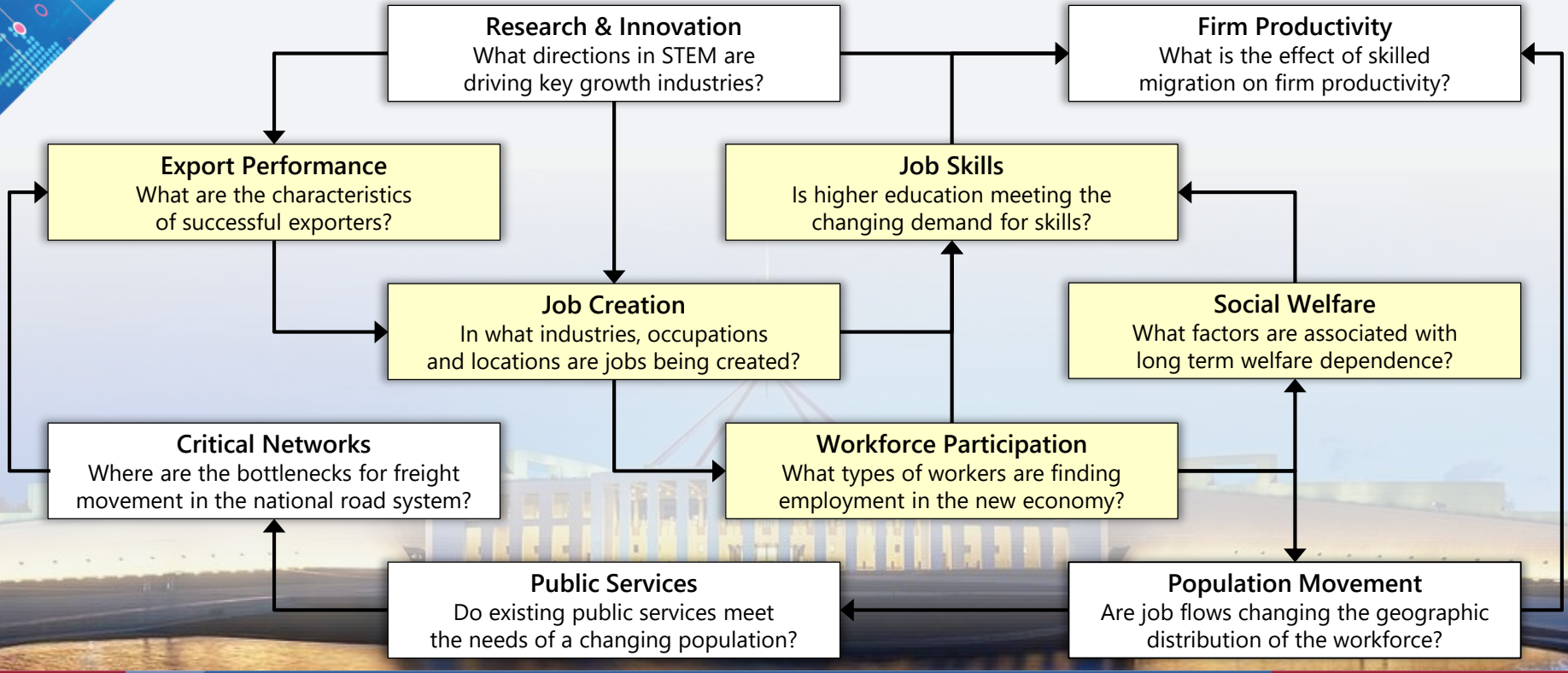
Accounting Systems

Administrative Collections

Surveys and Censuses

# Motivating example

# Connectedness of policy concerns



# Next generation analytical capability

## Built on system-centric information models

- Composable, interpretable and semantically precise
- Join up interrelated concept and data spaces
- Connect individuals and groups in multiple ways

## Dynamically integrates heterogeneous multisource data

- Structured and unstructured
- Cross-sectional and longitudinal temporal linkages

# Next generation analytical capability

Enables multiple analytical perspectives and objectives

- Exploration (pattern sensing) – finding statistical features and correlations
- Explanation (model building) – testing hypotheses about the observed data
- Extrapolation (system simulation) – projecting beyond known cases

# ABS use of KG and RL

# Knowledge graphs

Australian Bureau of Statistics

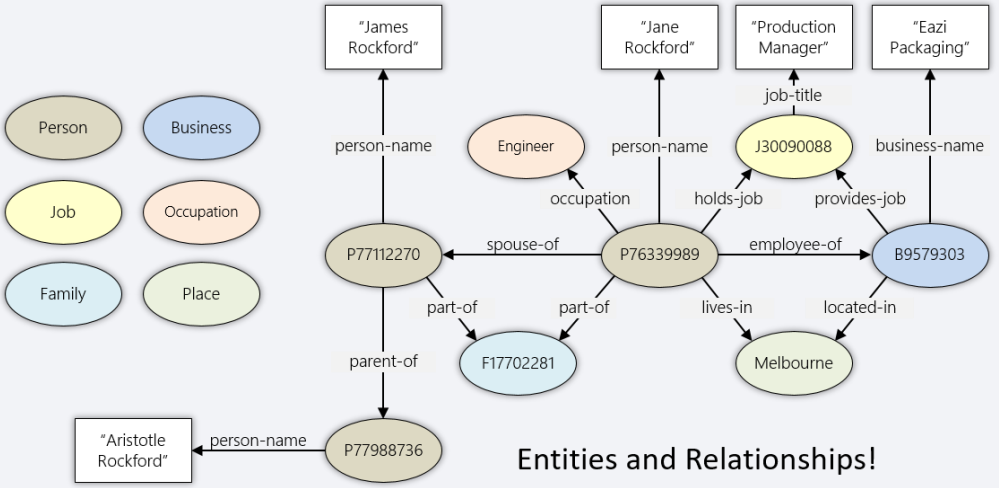
System is depicted as a graph of entities and relationships

- Entity – individual thing or group of things
- Relationship – association between entities
- Entities interact through relationships of analytical interest

Use W3C Semantic Web formalism for knowledge graphs

- Graph composition (standard: RDF)
- Semantic modelling (standard: OWL)
- Knowledge discovery (standard: SPARQL)

# Knowledge graphs – simple example



# Analytical domains

Systems are partitioned the into context-specific analytical domains

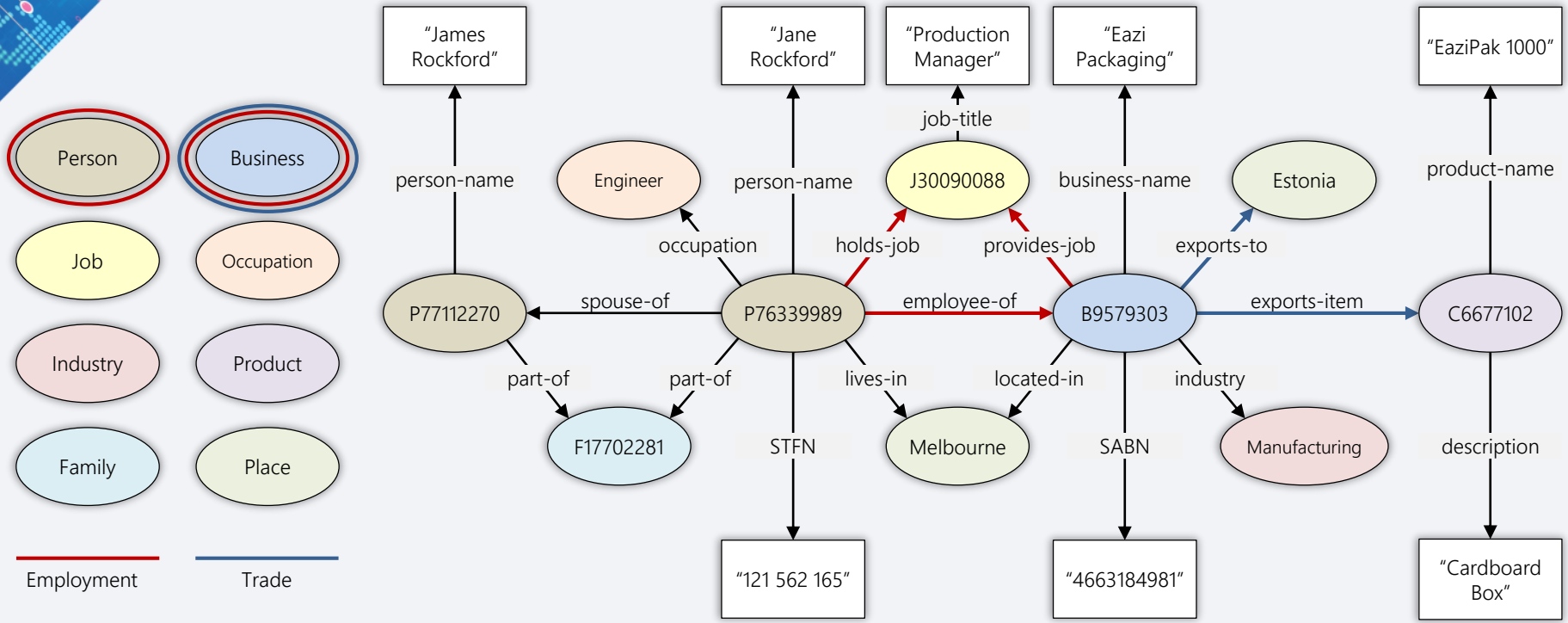
- Example: Trade, Employment, Production, Education, Welfare, etc

Each domain has set of associated entity types and relationships

- Basic entity types can exist in multiple domains
- Relationships are usually context-specific and so bound to one domain

Domains are connected through common entity types

# Analytical domains – simple example



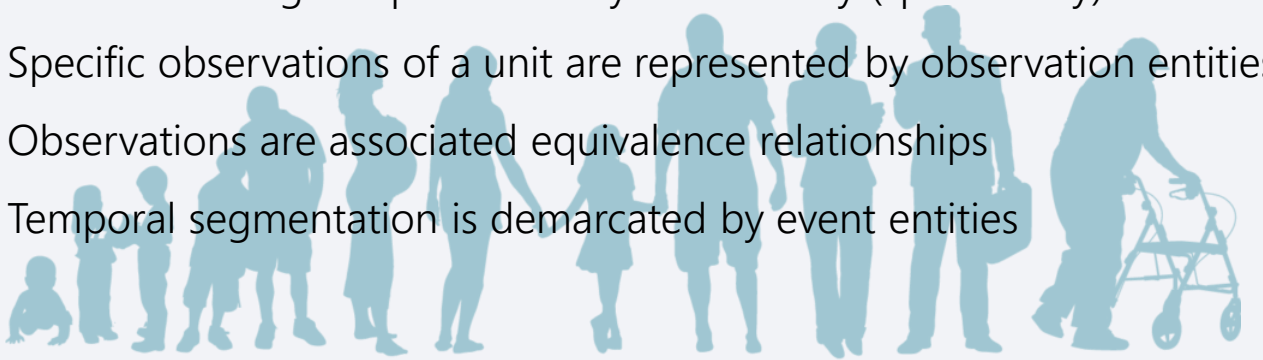
# Units and observations

Observable characteristics of an entity can change over time

- Example: person name, address, marital status, and employment details

Entities are associated with their respective observations in data

- Real world thing is represented by a unit entity (spine entity)
- Specific observations of a unit are represented by observation entities
- Observations are associated equivalence relationships
- Temporal segmentation is demarcated by event entities



# Associating observations – example

## Example: observations of the same person in different data sets

- Record-3 is much later that Record-1 and Record-2
- Significant events: Change of Residence, Marriage, Graduation

**Record-1**

| Family Name | Given Name | Address | DOB | Country of Birth | Sex | Marital Status | Occupation | STFN | Observation date |
|---|---|---|---|---|---|---|---|---|---|
| Smith | Jane | 1 Long Street Broadford VIC | 05-08-1985 | Australia | F | Single | Student | | 10-02-2005 |

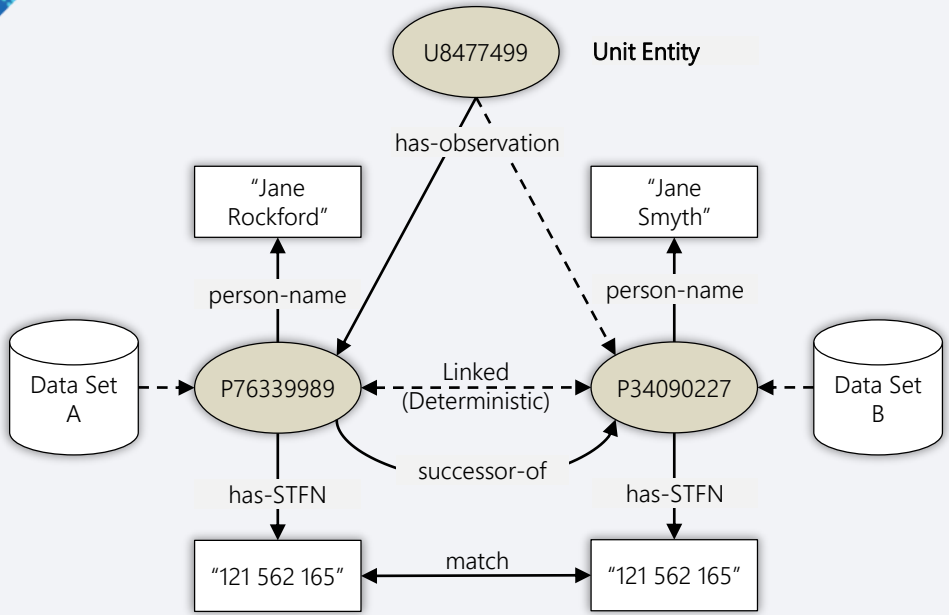
**Record-2**

| Family Name | Given Name | Address | DOB | Country of Birth | Sex | Marital Status | Occupation | STFN | |
|---|---|---|---|---|---|---|---|---|---|
| Smyth | Jane | 1 Long Street Broadford VIC | 05-08-1985 | | F | Single | Student | 121 562 165 | 17-05-2005 |

**Record-3**

| Family Name | Given Name | Address | DOB | Country of Birth | Sex | Marital Status | Occupation | STFN | |
|---|---|---|---|---|---|---|---|---|---|
| Rockford | Jane | 32 King Street Lalor VIC | 05-08-1985 | Australia | F | Married | Engineer | 121 562 165 | 15-07-2015 |

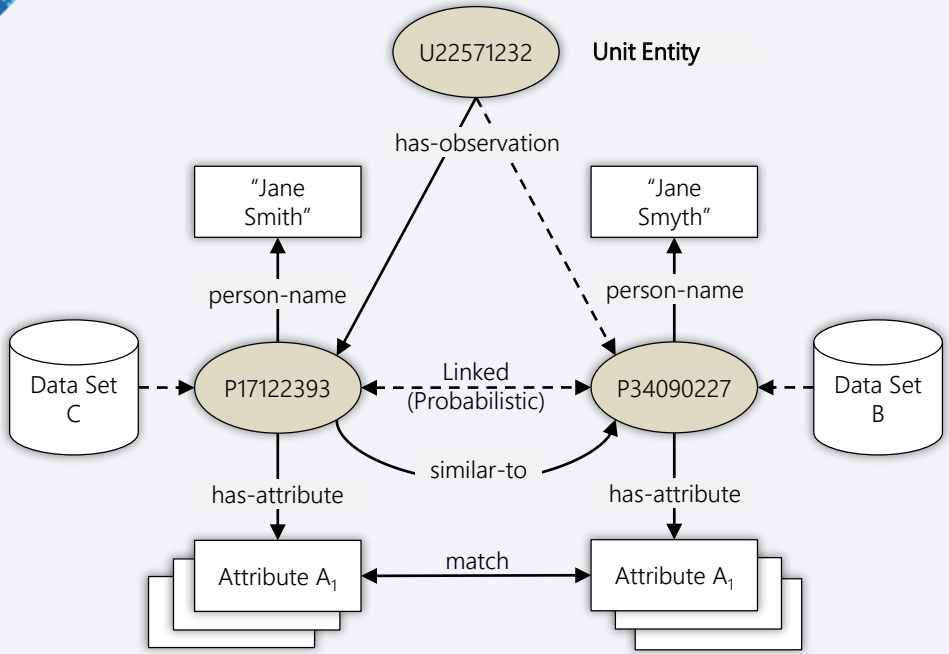
# Associating observations – example




## Deterministic association

- Current: identifier match using common unique key
- Future: fact match using deductive rules (FOL)
- Multiple class inheritance

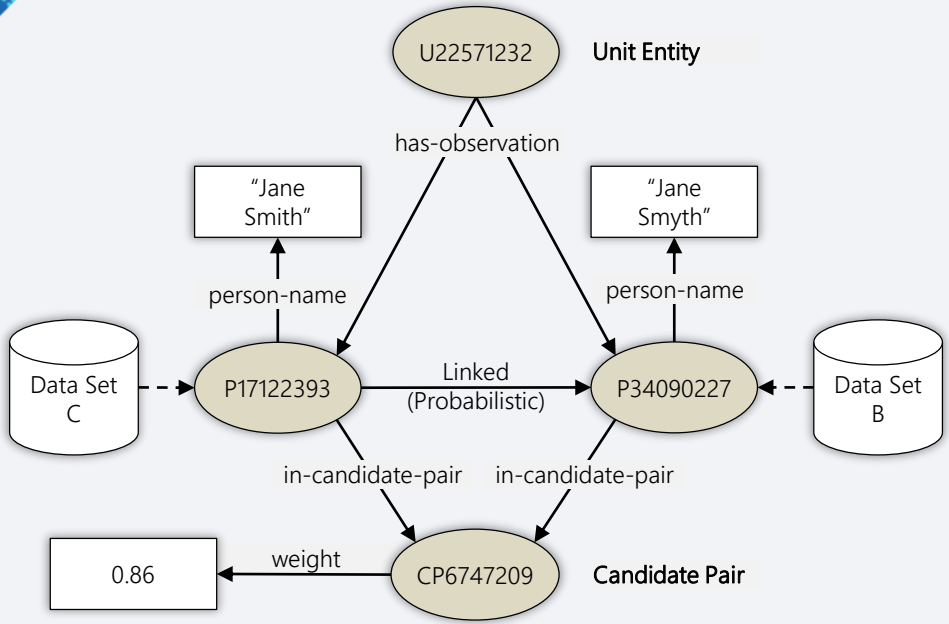
# Associating observations – example




## Probabilistic association

- Current: similarity match on entity characteristics using Sunter-Fellegi model
- future: similarity match on entity characteristics evaluation or relational characteristics using machine learning
- Multiple class inheritance

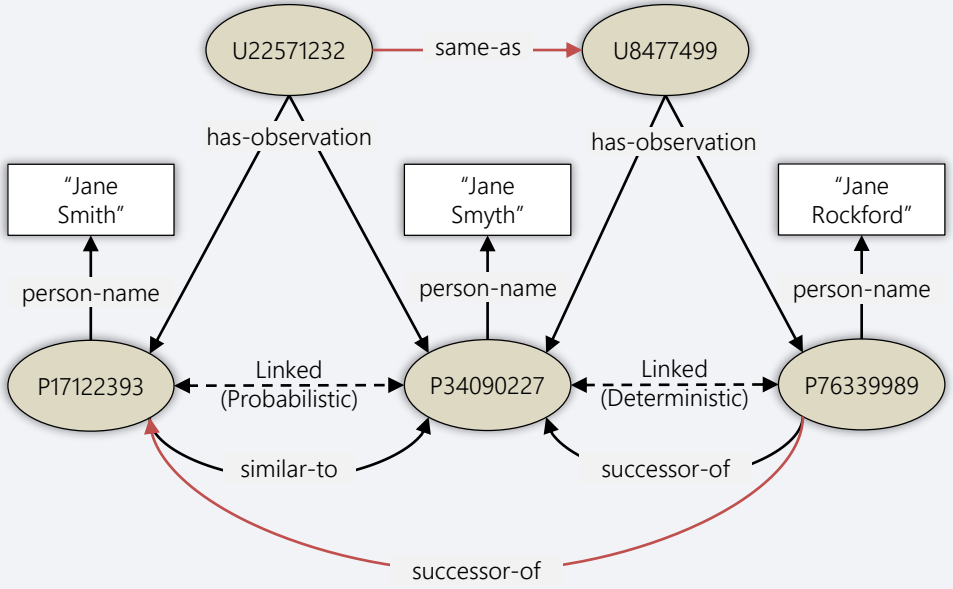
# Representing weights/probabilities



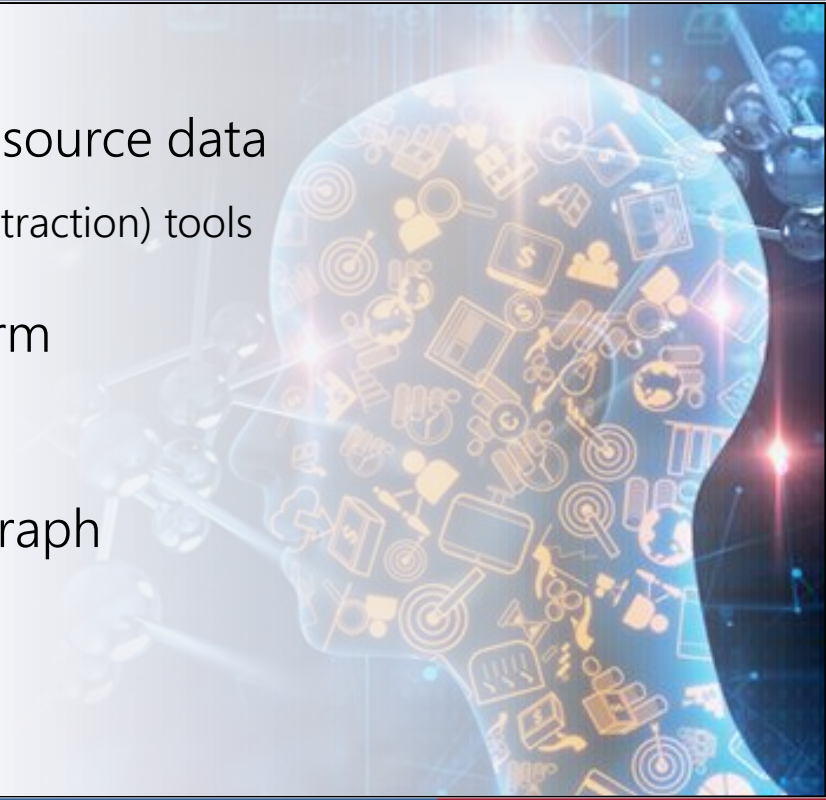

## Accommodates SF pairwise linking

- Candidate pairs with associated weights
- An observation can be liked to more than one CP
- Much better represented in a property graph construct
- Can we combine the two paradigms?

# Associating observations – example



# Combining data



Extract entities and relationships from source data

- Automatically by content analysis (fact extraction) tools

Create the knowledge graph in SW form

- Asserted facts from source data

Associate observation entities in the graph

- Inferred facts from reasoning processes

# Relational learning

Based on pattern of connections among entities

- Extend scope of probabilistic linking beyond IID assumption

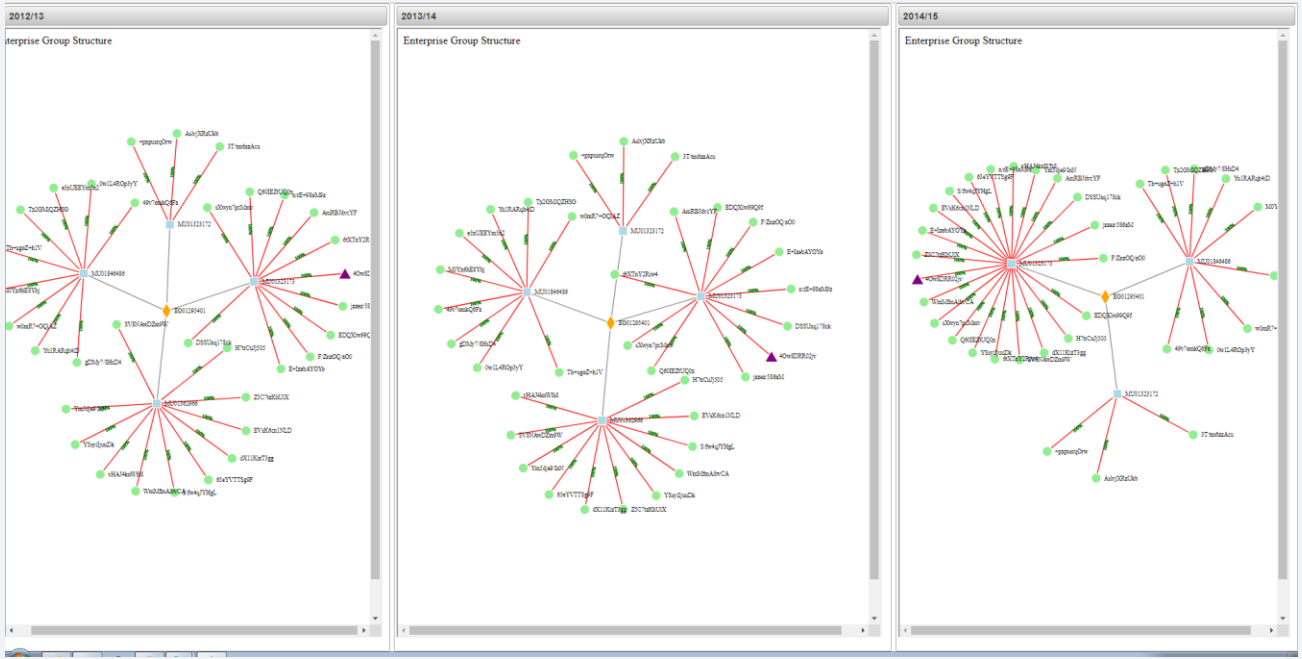
Associate entities across time and in disparate data sources

- Needed when there are no reliable common identifiers
- Example: persons by family and household relationships

Detect events that involve changes in the structure of groups

- Example: business reorganisation, closure, takeover

# Detecting change over time



Weisfeiler-Lehman (WL) kernel (de Vries, 2013) (Weisfeiler & Lehman, 1968) and the Intersection Tree Path (ITP) kernel (Losch, et al., 2012)

# RL approaches

## Graph kernel learning

- Represent the structural form of a graph for use in kernel learning algorithms
- Kernels: Weisfeiler-Lehman (WL), Intersection Tree Path (ITP)

## Tensor factorisation

- Manipulates 3-order adjacency tensor of the knowledge graph
- Estimate probability distribution over possible states of graph
- Algorithms: RESCAL, Complex Embedding, HOLE, TransE

# Overview of GLIDE

# GLIDE

Capability vision for enabling informed decision making

- About policy, services, investment and (possibly) regulation

Based on insights derived from different types of analysis

- Exploratory analysis, hypothesis testing, system simulation

Using a dynamic evidence base from diverse data sources

- Surveys, admin collections, sensors, transactions, web content, …

**Graph-Linked Information Discovery Environment**

# What GLIDE will provide

Australian Bureau of Statistics

One platform for data analysis and linking

- Browser interface – rich, interactive, navigable, context-sensitive visualisation
- Program interface (R, Python) – statistical and econometric modelling
- Spatial, temporal and compositional perspectives of problem
- Deterministic and probabilistic linkage methods – common key, Sunter-Fellegi
- Extraction of entities and facts from heterogeneous data
- Reusable, plug-and-play models and components

**Graph-Linked Information Discovery Environment**

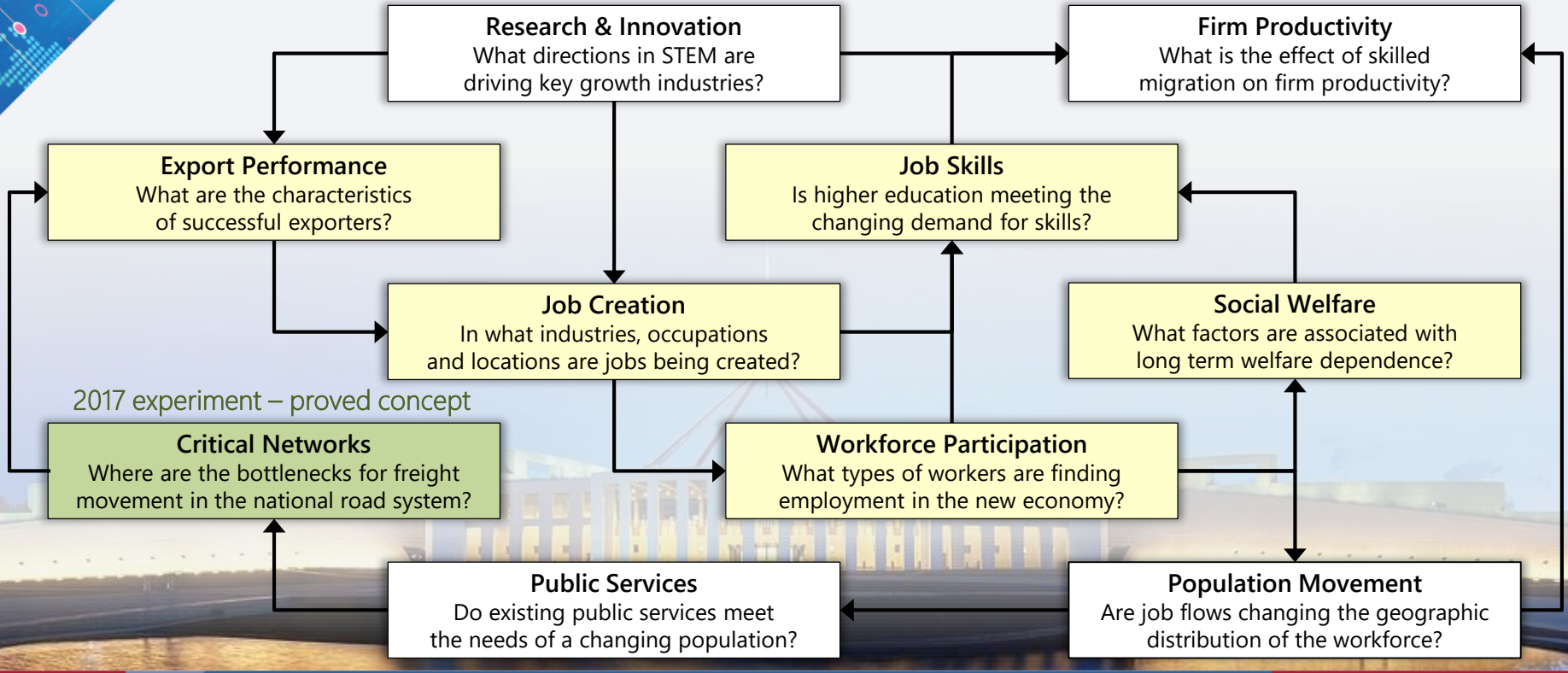
# What GLIDE will provide

A set of extensible, interoperable 'data pools'

- Multiple structured or unstructured data sets
- Longitudinally linked data (given social license)
- Different entity and relationship types in the form of a knowledge graph

**Graph-Linked Information Discovery Environment**

# Pathway initiatives



# Pathway initiatives – freight movement

Rese
What
driving

Export Performance
What are the characteristics of successful exporters?

In what
and locatio

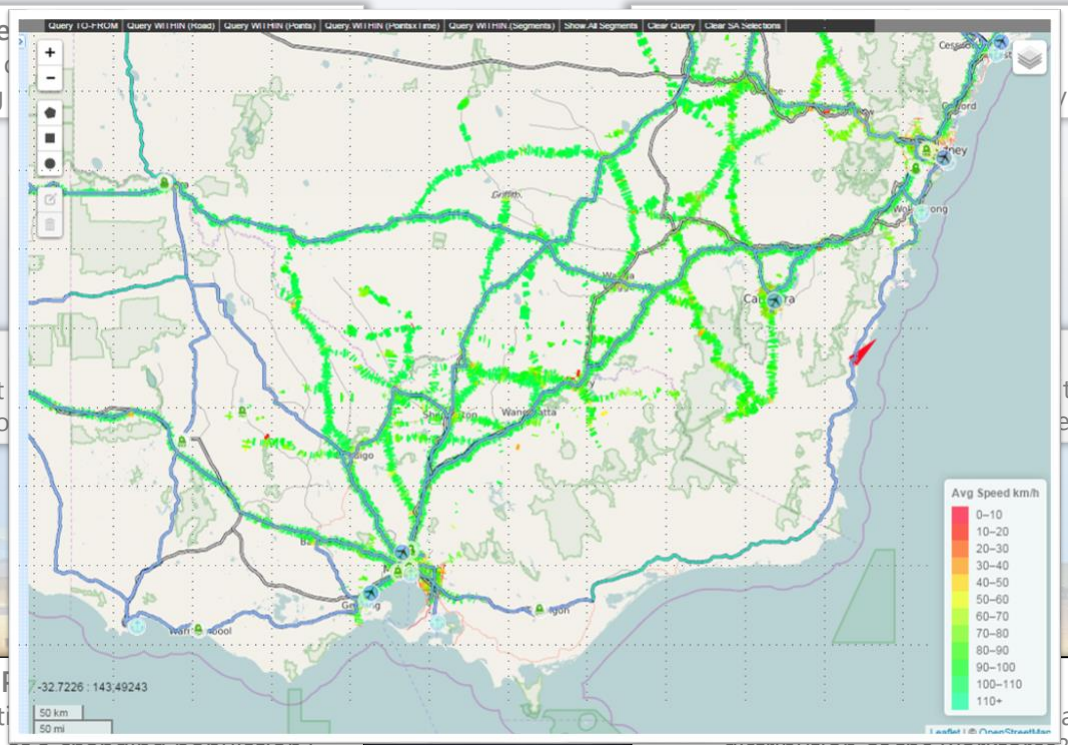
**Critical Networks**
Where are the bottlenecks for freight movement in the national road system?

Do existi
the needs of a changing population?

?

th
e?

aphic
distribution of the workforce?



# Pathway initiatives – freight movement

**Research & Innovation**
What di
driving k

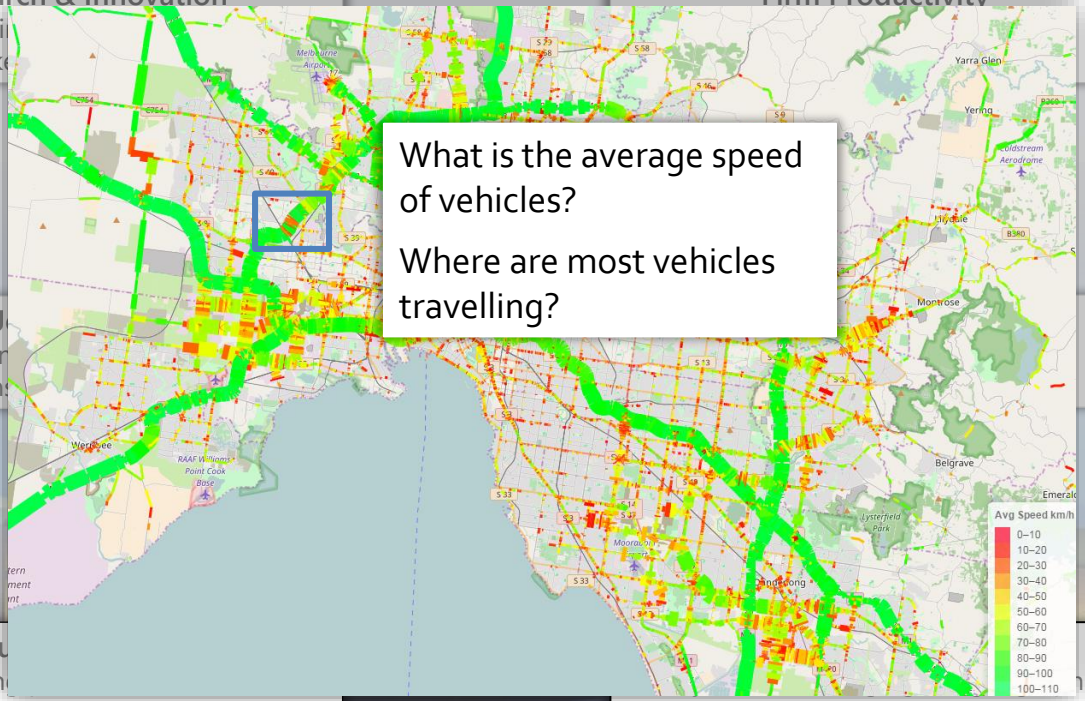
**Export Performance**
What are the characteristics of successful exporters?

**J**
In what i
and location

**Critical Networks**
Where are the bottlenecks for freight movement in the national road system?

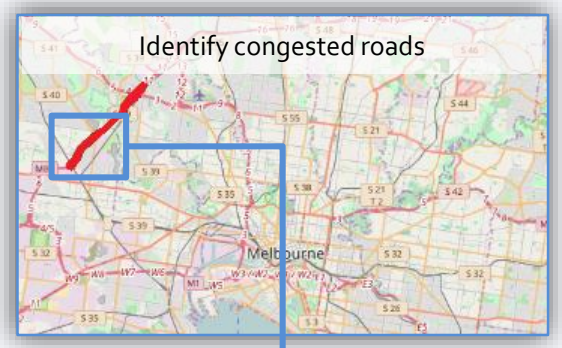
**Pu**
Do existin
the needs of a changing population?

**Firm Productivity**

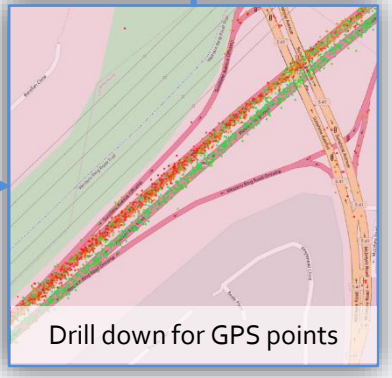



distribution of the workforce?

# Freight movement – drilling down





# Pathway initiatives – successful exporters

Research & Innovation
What direc
driving key

Firm Productivity

**Export Performance**
What are the characteristics of successful exporters?

Full extract BAS and BCS data

- 7 million businesses per year
- Complex relationships

Monthly export transactions data

- 800,000+ records per month

BLADE data

- Business Characteristics Survey (BCS)
- Business Activity Statements (BAS)

Merchandise export records

- Held by Department of Home Affairs (DHA)
- Lodged by exporters and agents via (CCF)

ABS business unit model (LE, EG, TAU)

Classifications (AHECC, SITC, ANZSIC, SISCA, AGSO, …)

Job
indu
ons a

Pub
Do existing
the needs of a changing population?

distribution of the workforce?

Australian Bureau of Statistics

# Questions?