GREAT LEAP

Best practices for data management

Tiziana Margaria

University of Limerick

tiziana.margaria@ul.ie

GREATLEAP

WORKING GROUP 3

Creation Analytical tools:
computational and **visualization tools** for the application to new research questions

Working group 3 will coordinate activities related to sharing, evaluating, and improving **analytical tools** for analyzing historical cause of death data.

It will focus on

(1) tool and methods adaptation by ensuring exchange between the fields of demography, epidemiology, and history and

(2) by establishing best practices across Europe that could be expanded worldwide.

Moreover, the latter will include **best practices** to move from the current research practice to open science.

Table of content

No good material for this audience

Data management, best practices

Many perspectives:

- database management, information systems
- statistics
- Python and Data science
- CRM (customer’s data – generate sales)
- The mathematics of relational databases
- …

verbs are **activities**

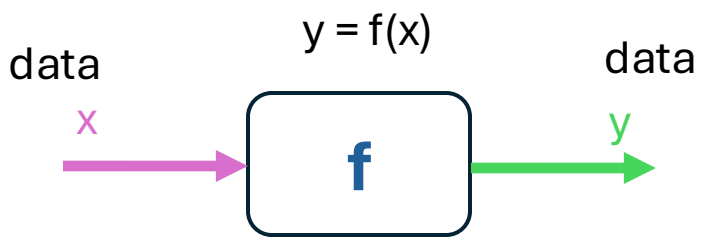
activities are **programs**

UML activity diagrams **are** programs

UML activity diagrams **are workflows**

Data vs. software

- Data per se is useless: “data cemetery”
- What “happens” with the data(sets)
  - Input/Output function
  - I/O relation



verbs
- send_a_mail_to (x)
- insert (record)
- check_non_empty (x)
- display_as_histogram (dataset)

verb

add 1

y = f(x) y = x + 1 (1,2) (7,8)

y = x^2 (1,1) (3,9)

compute_the_square_of

verb

y = convert_to_pdf (.doc)

verb

(slides.doc, slides.pdf)

(slides.doc, slides.svg)

(“x”.doc, “x”.pdf)

Data vs. software

- Data per se is useless: “data cemetery”
- What “happens” with the data(sets)
  - Input/Output function
  - I/O relation

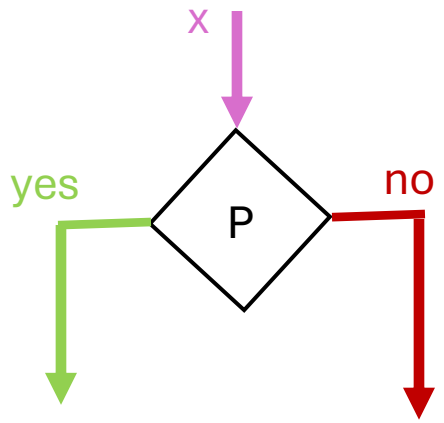
- Transformation: computation

  - **Format** → different format (R dataframes → .csv, TIFF→ PDF)

  - **Dataset** → aggregation/summary (table → histogram, map, graphics, or other representation)
    transcription problem: physical register or a .tiff image → record(s) in a database

Data vs. software

- Data per se is useless: “data cemetery”
- What “happens” with the data(sets)
  - Input/Output function
  - I/O relation



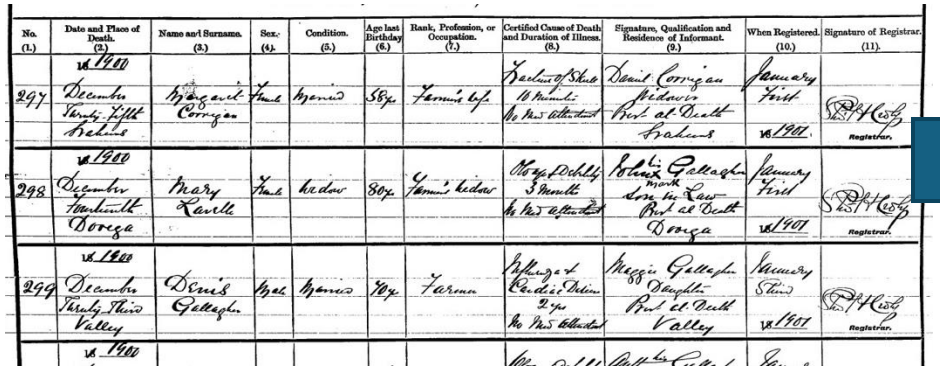
- Question: conceptually, mostly a filter (a “sieve”) based on a property P
  - Tests (3,5) is the pair (3,5) in the relation $y = x + 1$?
    in the relation $y = x^2$?
    (John, spinster)
  - Queries “give me all the death records where the deceased was younger than 20 and female”

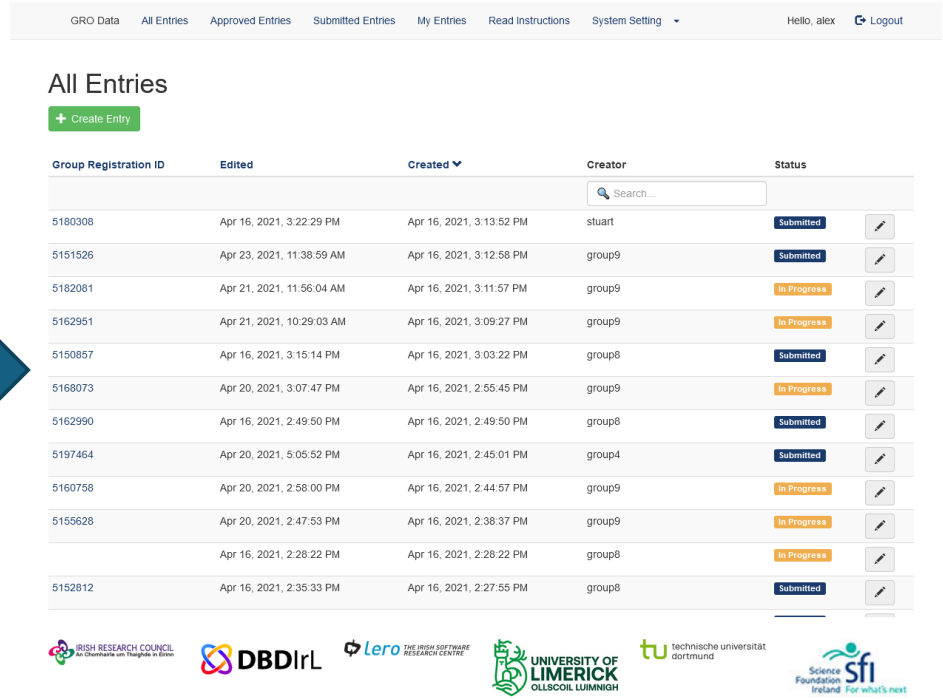
Case Studies around Historical Data Records

Tiziana Margaria, Ciara Breathnach,
Rachel Murphy, Alexander Schieweck, Enda O'Shea,
Daniel Sami Mitwalli, Marco Krumrey, Sebastian Teumert

Irish GRO - Death and Burial Records

- Death and Burial Data
- The data records are available as image scans, with some meta data
- We developed a Web Application to manually transcribe the records into a coherent format
- Afterwards the data is cleaned
- Used for many transcription campaigns (Transcribathons)
- Interfaculty team teaching award 2023





What are the Effects of the Old Age Pension Act?

Alexander Schieweck, Rachel Murphy & Ciara Breathnach

- From 1908 onwards, people who could prove that they are 70 years and older, were eligible for financial support.
- **Issues at the time:** People had no birth certificates, as they were born prior to the introduction of (standardized) birth certificates in Ireland.
- **Issues today:** Records are scanned, but not properly digitalized; Loss of record over time.

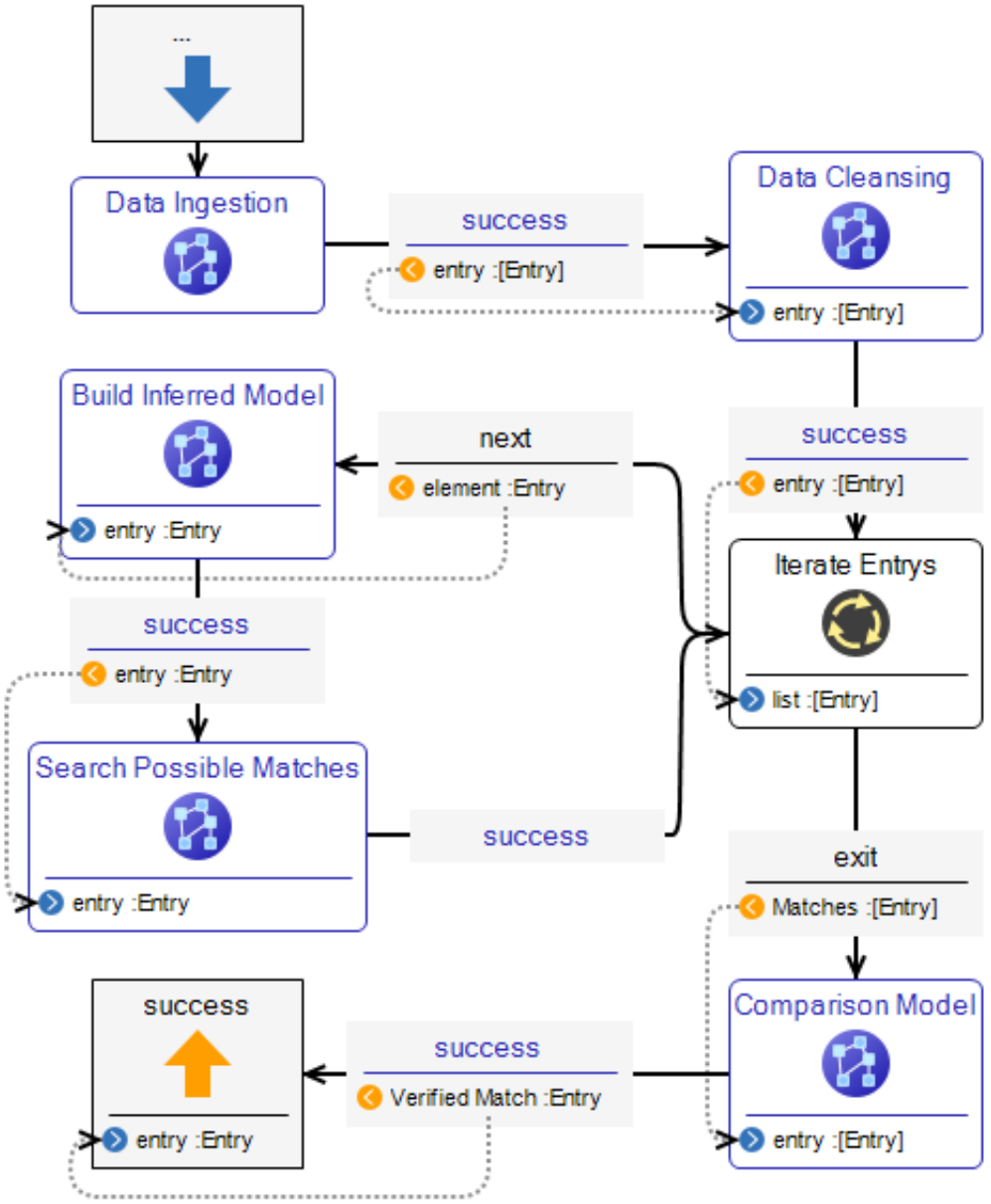
- **Case Study:** 132 people died age 70 and older in Dublin in 1911
  - 75 women, 57 men
  - 78 were widowed all of whom died at a clear address, 25 were returned as married, 15 spinsters, 11 bachelors & 2 unknown

The CensusIRL Application

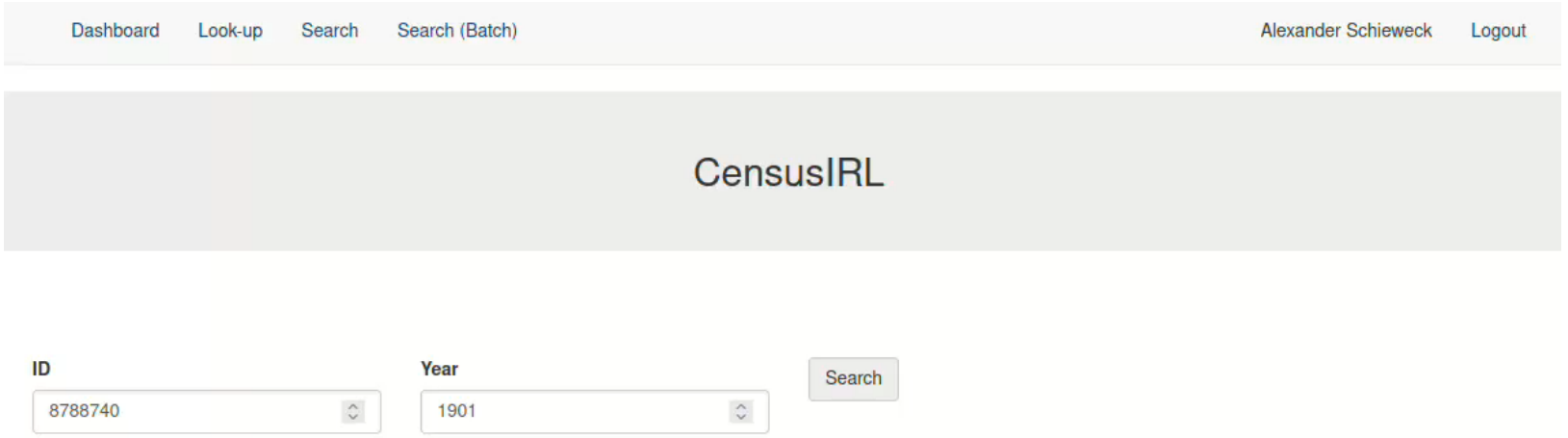
CensusIRL allow to trace people through time based on the Irish Census Records.

- Model (= design) a process to match entries from the 1901 census to the 1911 census
- Generalize: the process should work with the N^{th} Census

Anonymised entries are not included in the scope



Demo: Explore Census Data



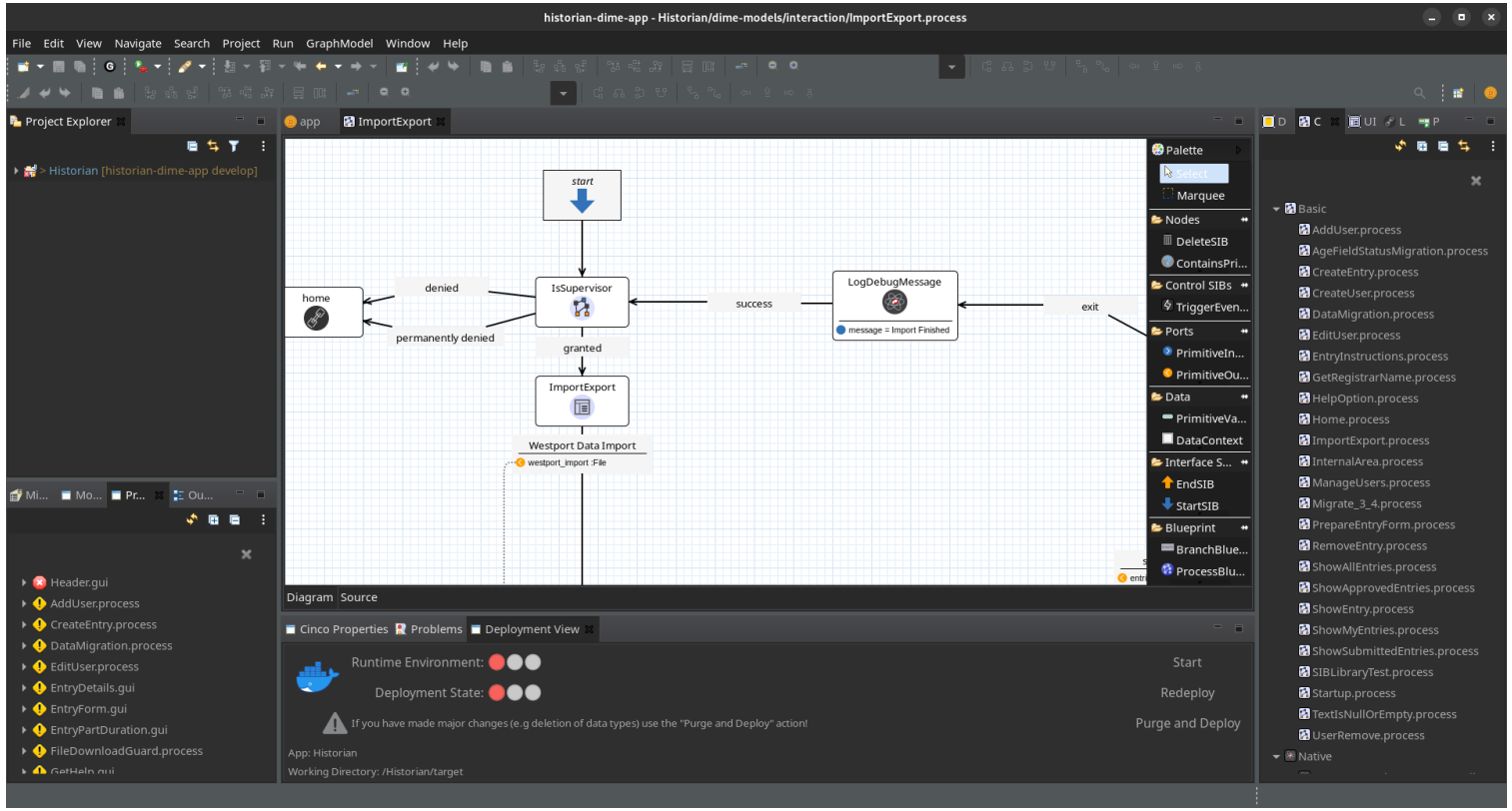
Low-Code Development via DIME

The tool used to develop the Historian DIME App (HDA) & CensusIRL

- TU Dortmund & UL
- Based on the popular Eclipse IDE

Utilizes different graphical model types to model and generate web applications:

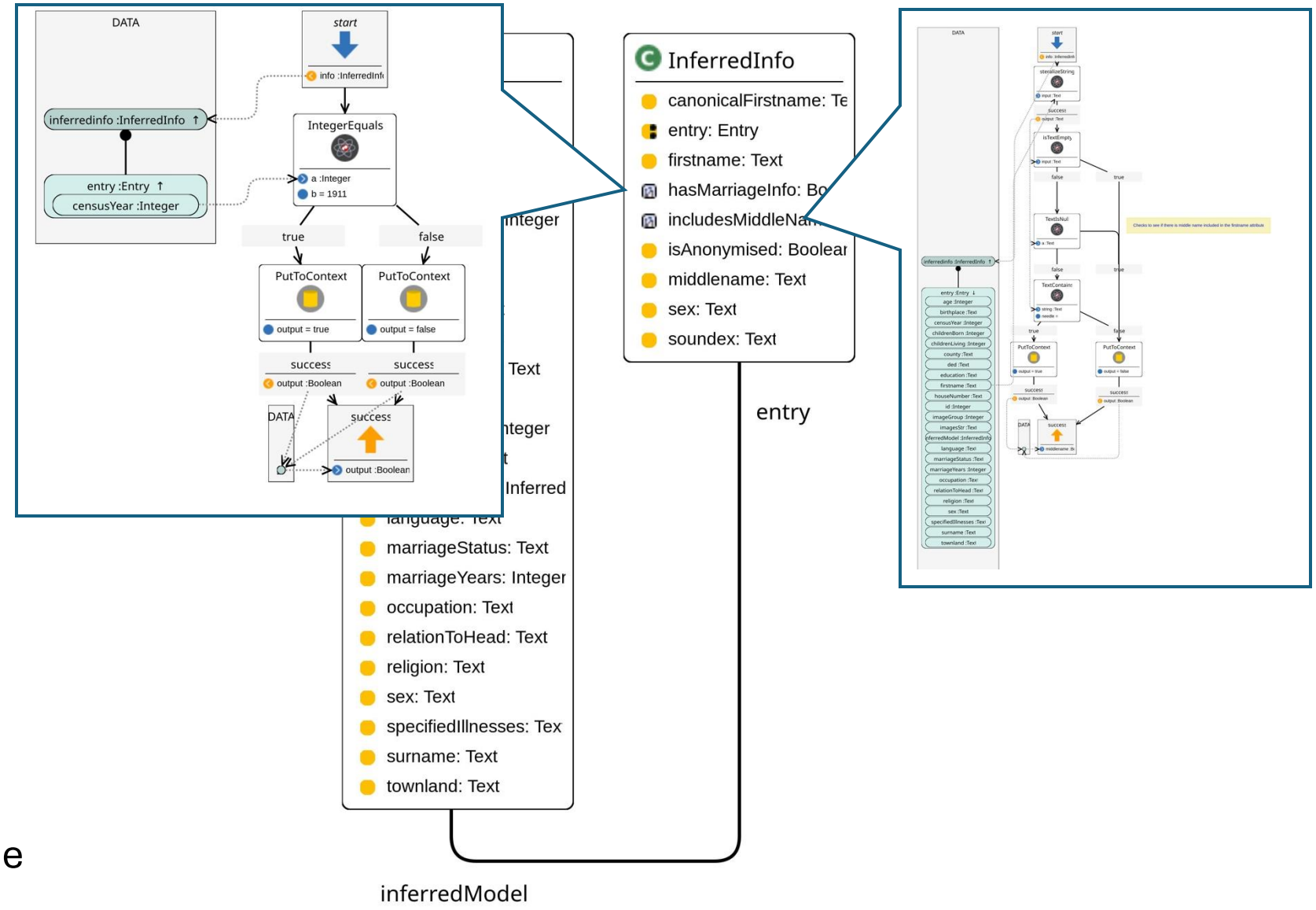
1. Data Model
2. Process Model
3. UI Model



Online: https://gitlab.com/scce/dime

Data Model

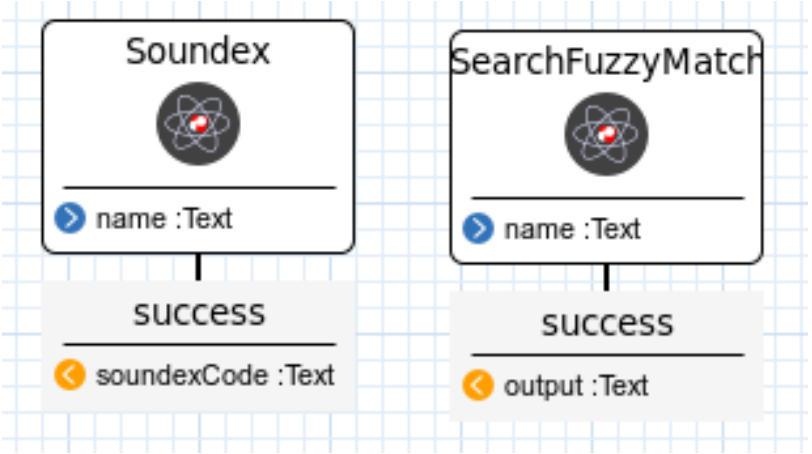
- Census data contains 23 data points per record
  - 1901 & 1911 census contain the same columns
- Pre-process **InferredInfo** to store additional information, which can be created out of the available census data
- **SearchingModel** describes the database query which will be used to filter the census records
- Stored in PostgreSQL database



What can we match?

The matching happens on the following properties:

- **First Name**, which is generalized
- **Last Name**, using Soundex to account for similar names
- **Sex**
- Approximate **birthyear**, +/- 5 years
- **Birthplace**
- **County**
- **Marriage Information**
  - Once married, a person will be seen as married or widow/widower, not single anymore



List of common First Name variations

STANDARD	ABBREVIATION or PET name or LATIN NAME or IRISH NAME
Abigail	Abby, Abigail, Abigeal, Nabby, Abagail, Abel, Abbey, Abigal
Abraham	Abrahame, Ab, Abr, Abe, Abby
Agnes	Agnus, Aggie, Assie
Aidan	Aodh, Mogue, Moses, Edanus
Aloysius	Aloyisius, Aloysuis
Albert	Bert, Albie, Bertie, Albertus, Alberti
Alexander	Ales, Alexr, Alec, Alex

https://www.rootsireland.ie/help/first-names/

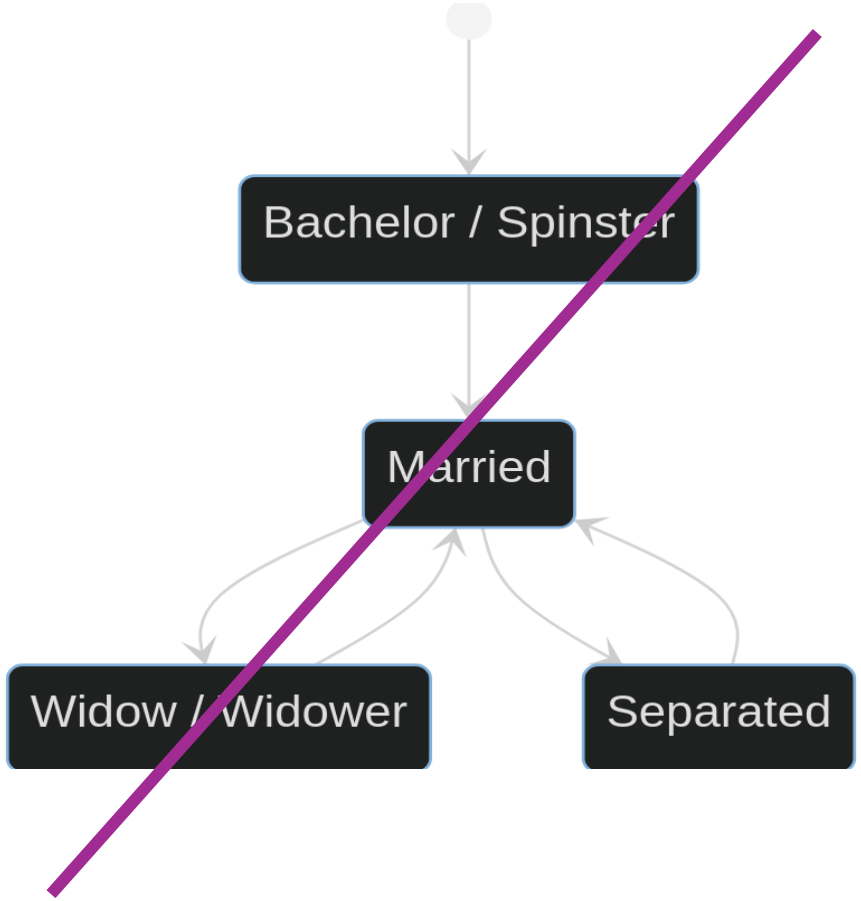
Challenges with Records Matching

	1901 Census	1911 Census	Civil Marriage
Forename	Willm Fitz Henry	Fitz-Henry	William
Surname	Plummer	Plummer	Plummer

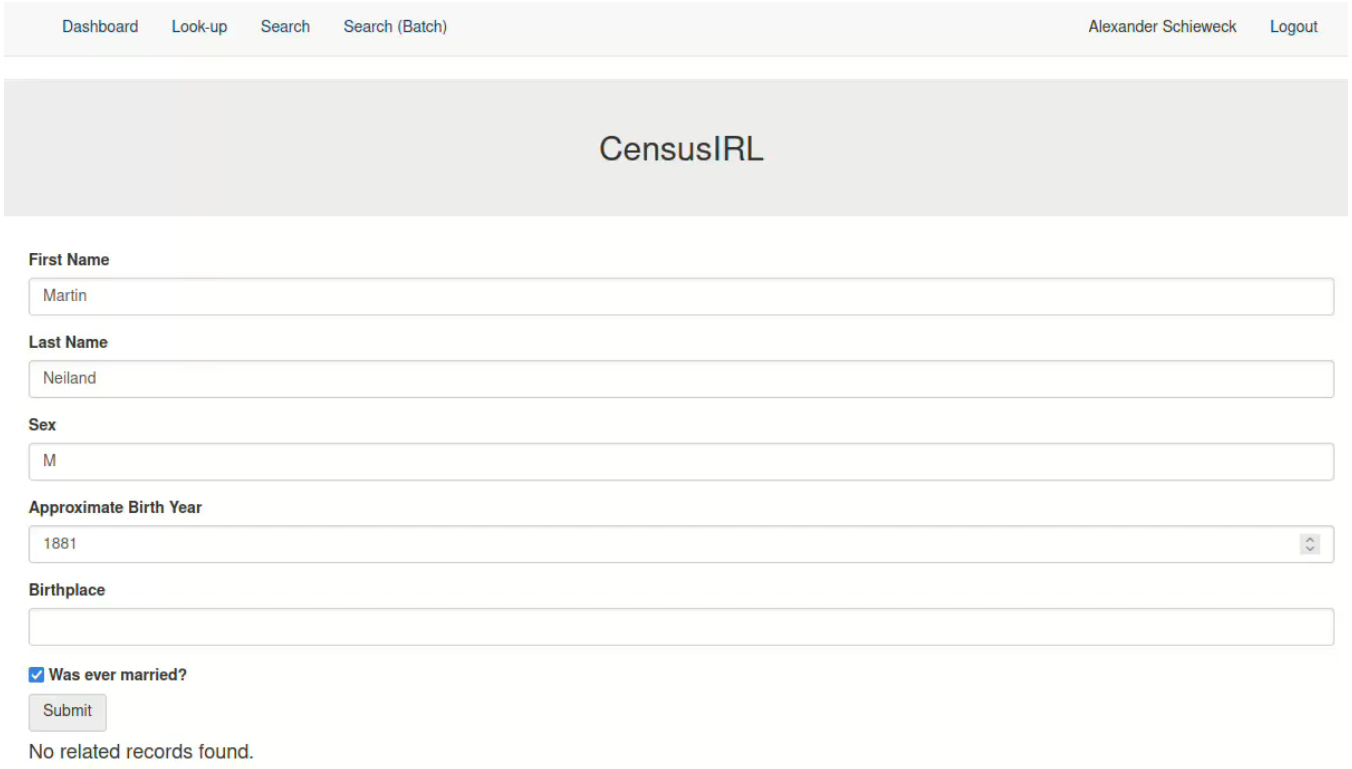
	Maiden name	1901 Census	1911 Census
Forename	Bridget	Bridget	Bridget
Surname	Redmond	D'Arcy	Meagher

Challenges with Records Matching

- Marriage information not intuitive
- A lot of misspellings
- 227 distinct status descriptions, incl.
  - ‘No’
  - ‘Illegitimate’
  - ‘Still Single’
  - ‘unmarried (unfortunately)’
  - ‘Very Munch’
  - ‘On the Look Out’
  - ‘Already stated’
  - ‘Not yet’
  - ‘Not yet either’
  - ‘Not of Course’.



Web application – Find Matching Data

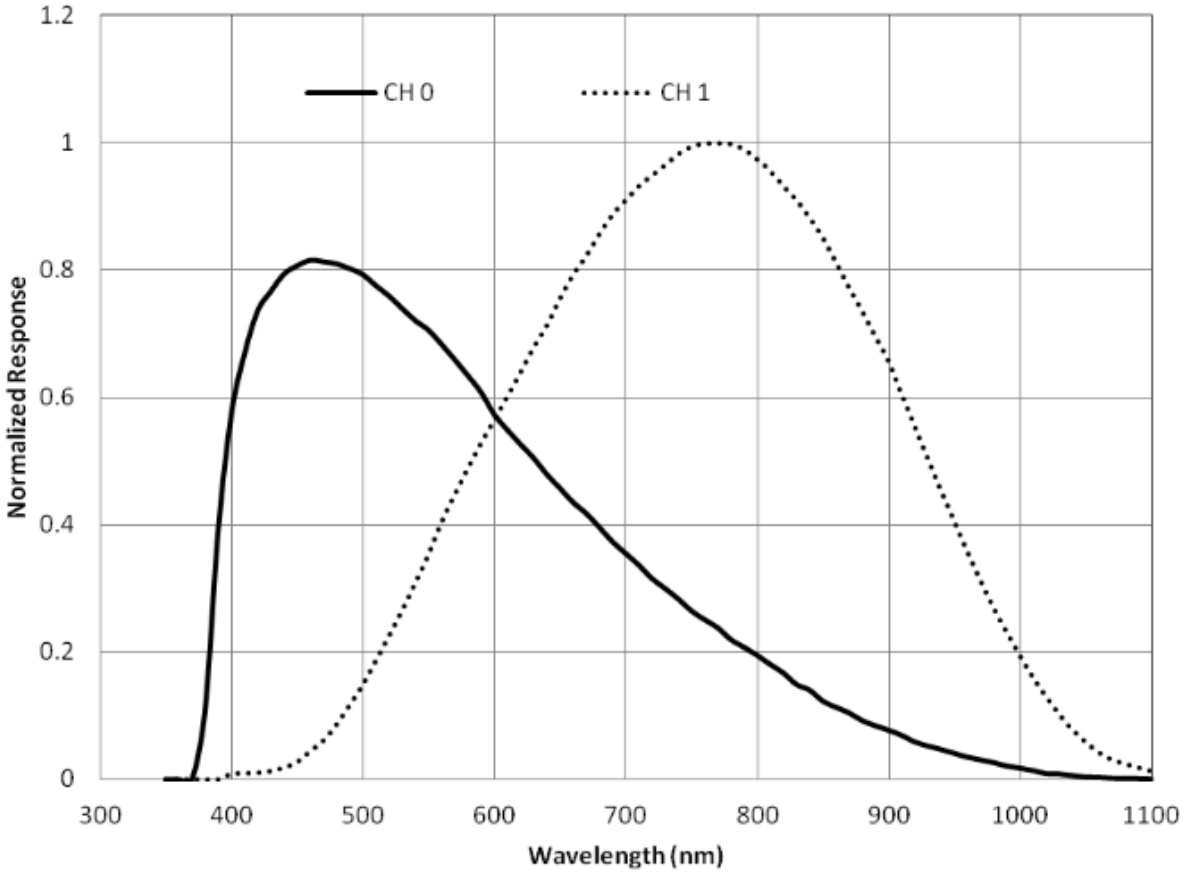


CRT·AI
SFI Centre for Research Training in Artificial Intelligence

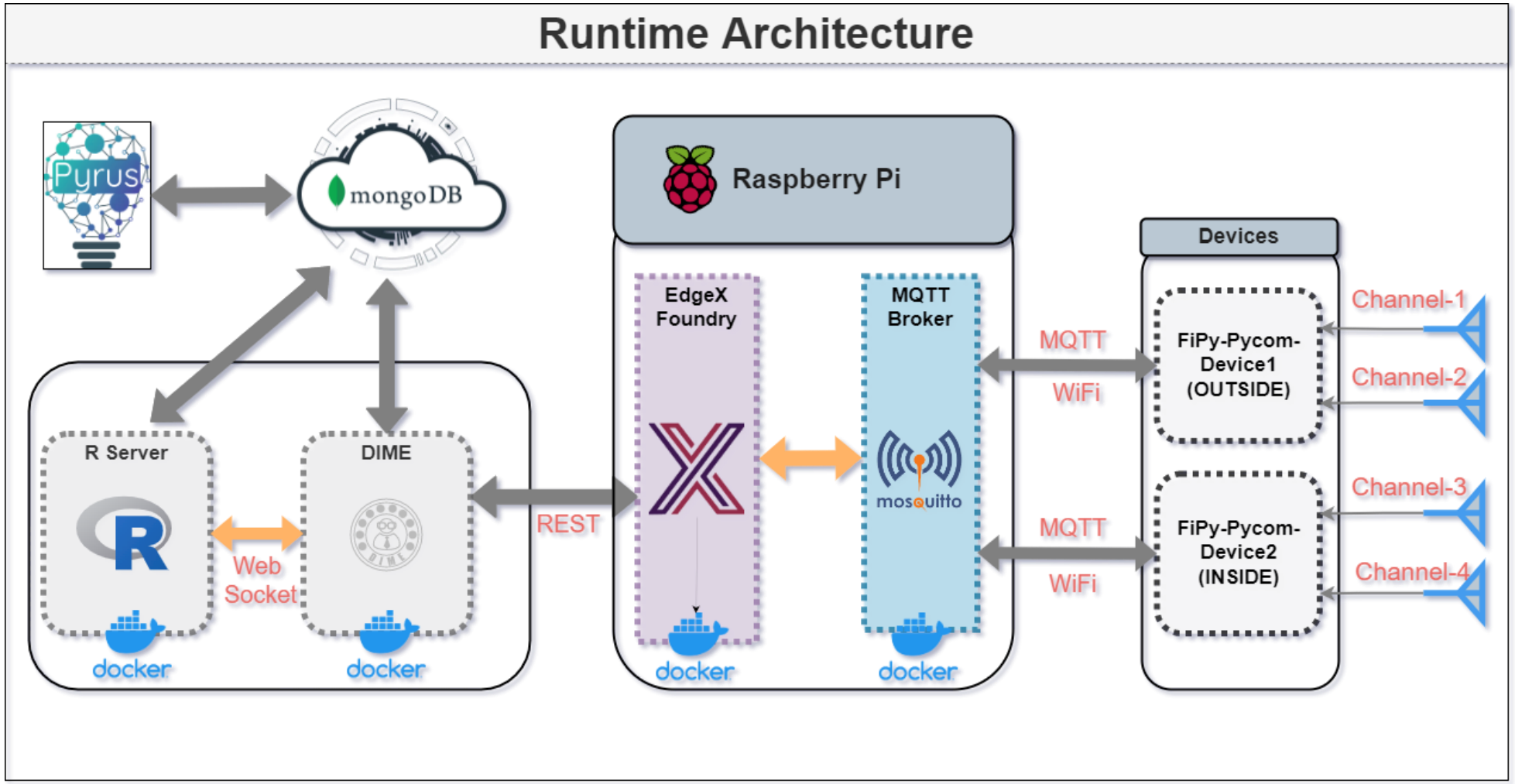
Case Study: *Low-code re-engineering for IoT Applications*

Chaudhary, H. A. A., Guevara, I., Singh, A., Schieweck, A., John, J., Margaria, T., & Pesch, D. (2023). Efficient Model-Driven Prototyping for Edge Analytics. *Electronics*, *12*(18), 3881. https://doi.org/10.3390/electronics12183881

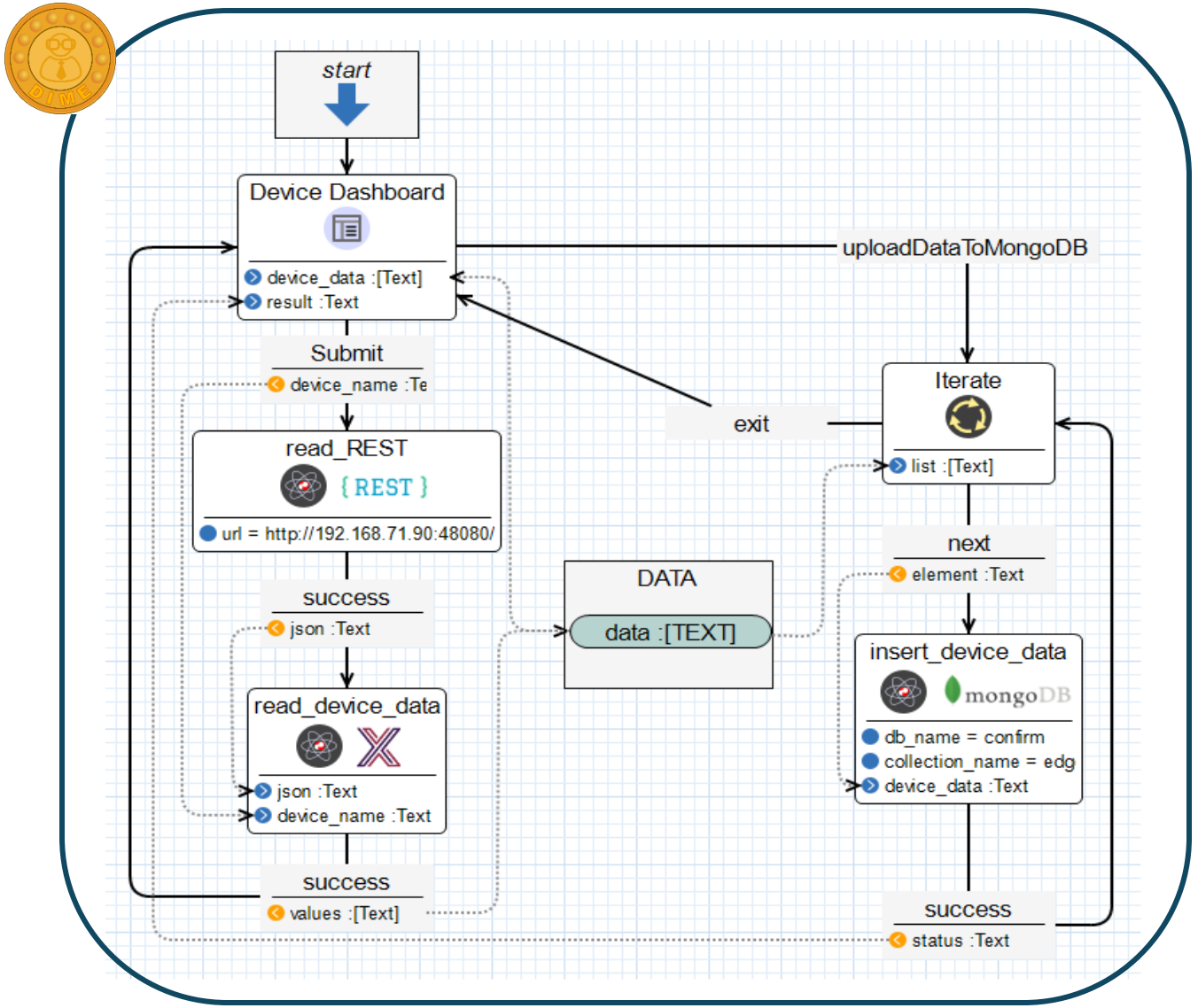
Experimental Setup



Edge Analytics Use Case: infrastructure and communications

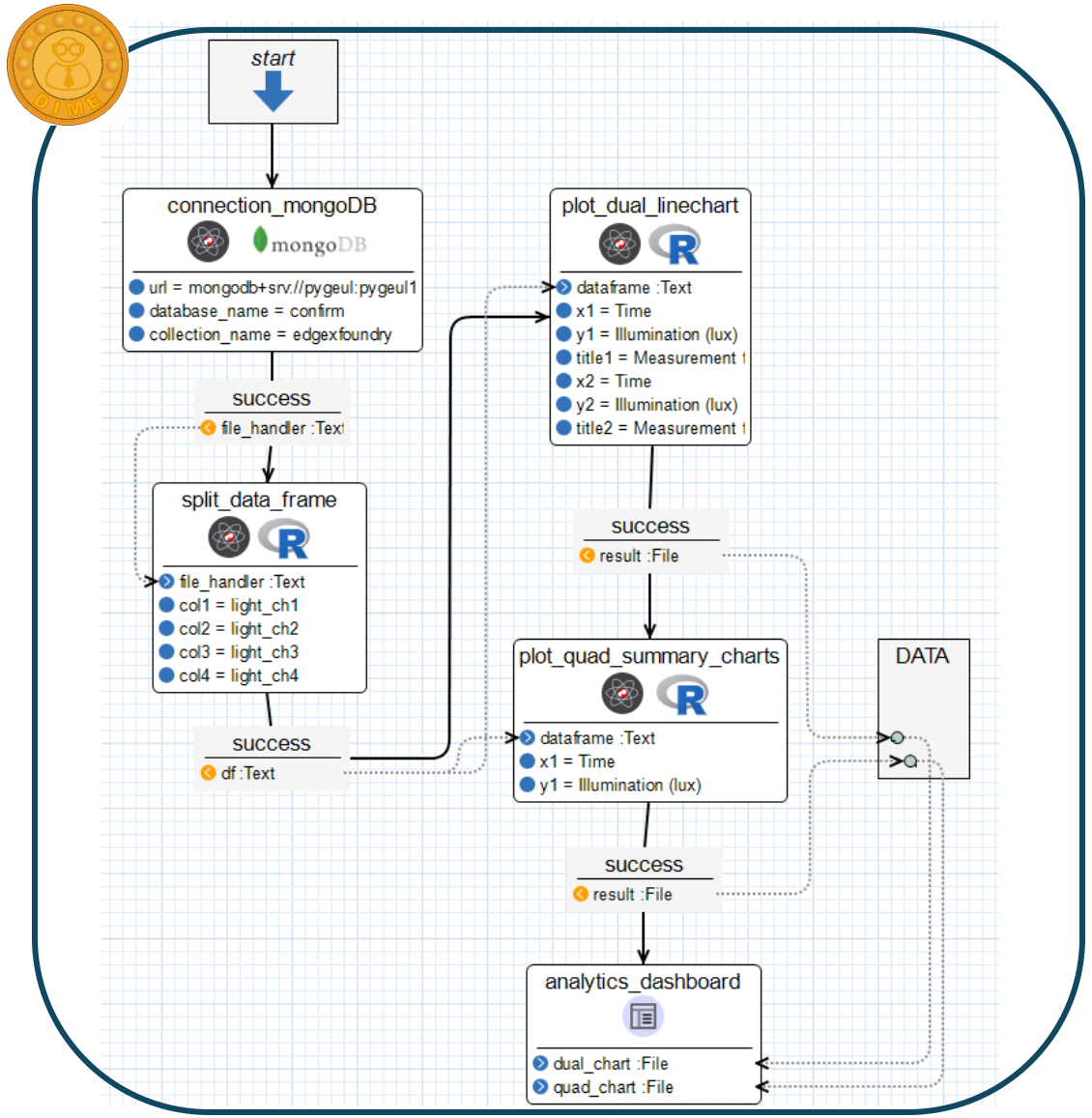


Data Acquisition from EdgeX Foundry and Data Ingestion into MongoDB



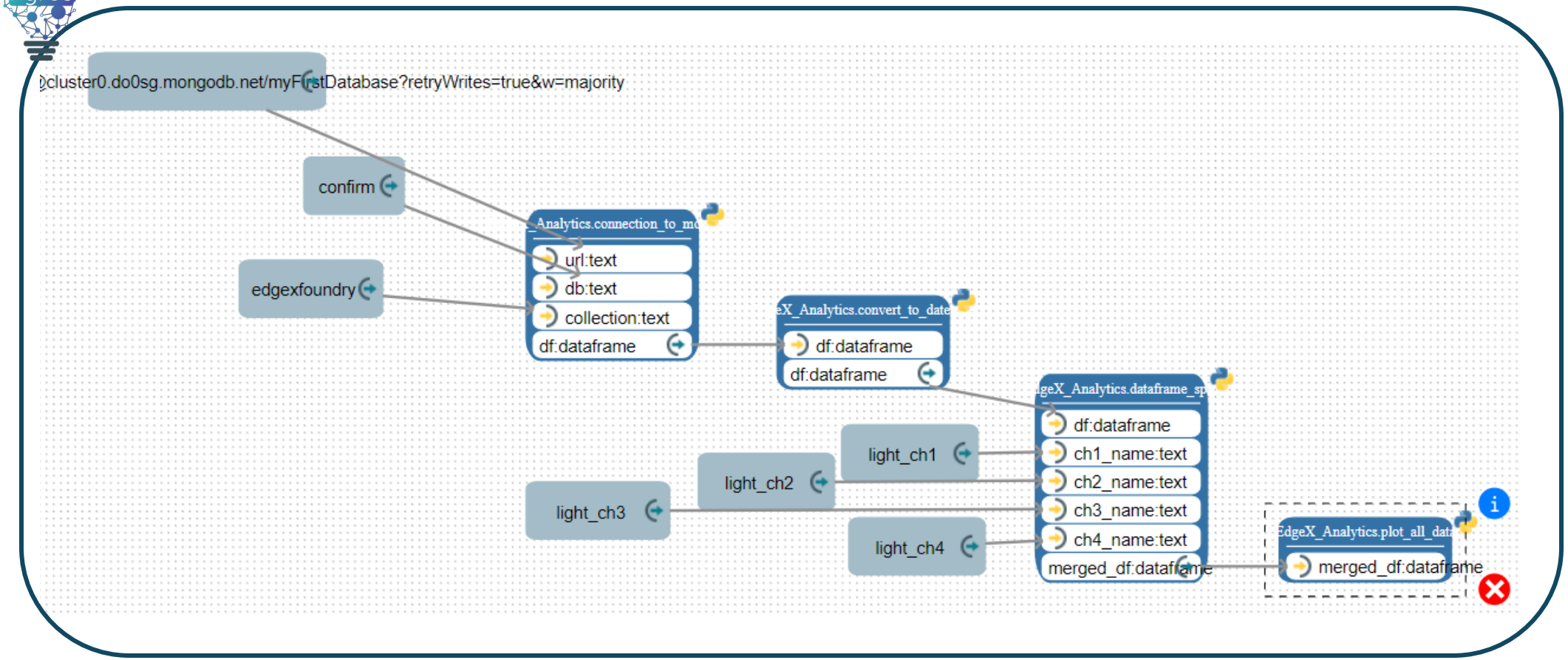
Process in DIME

Analytics Dashboard in R



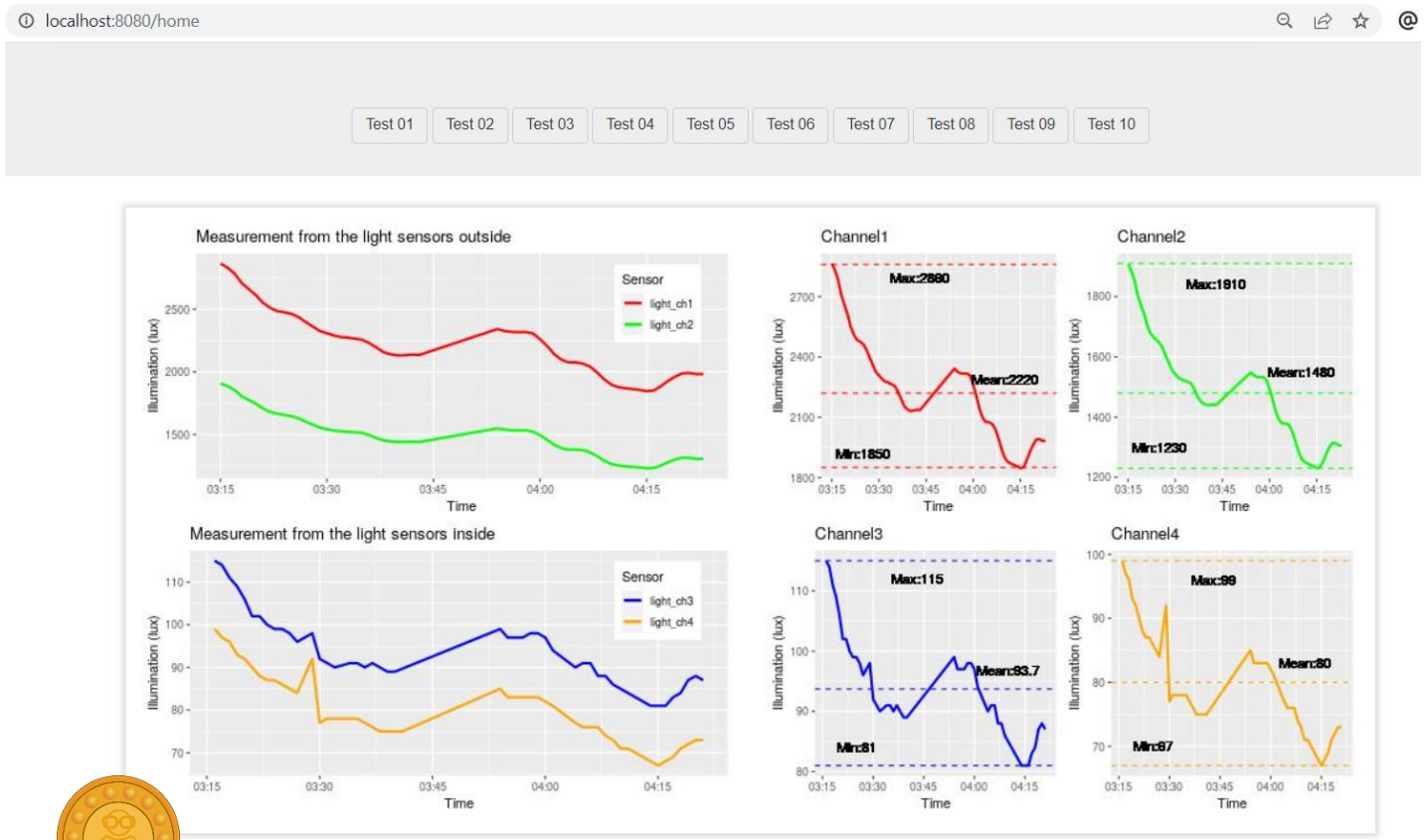
Process in DIME

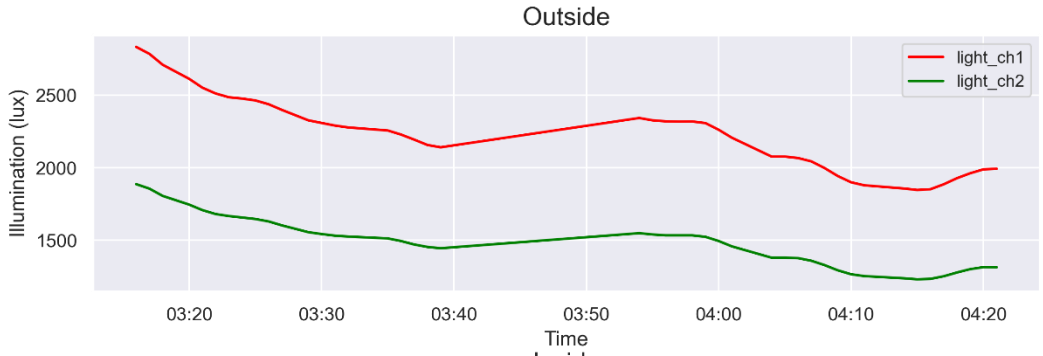
Analytics Dashboard in Python

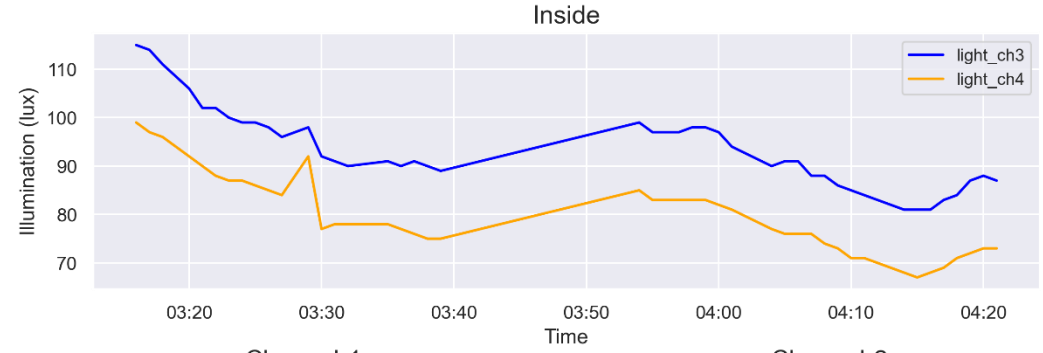


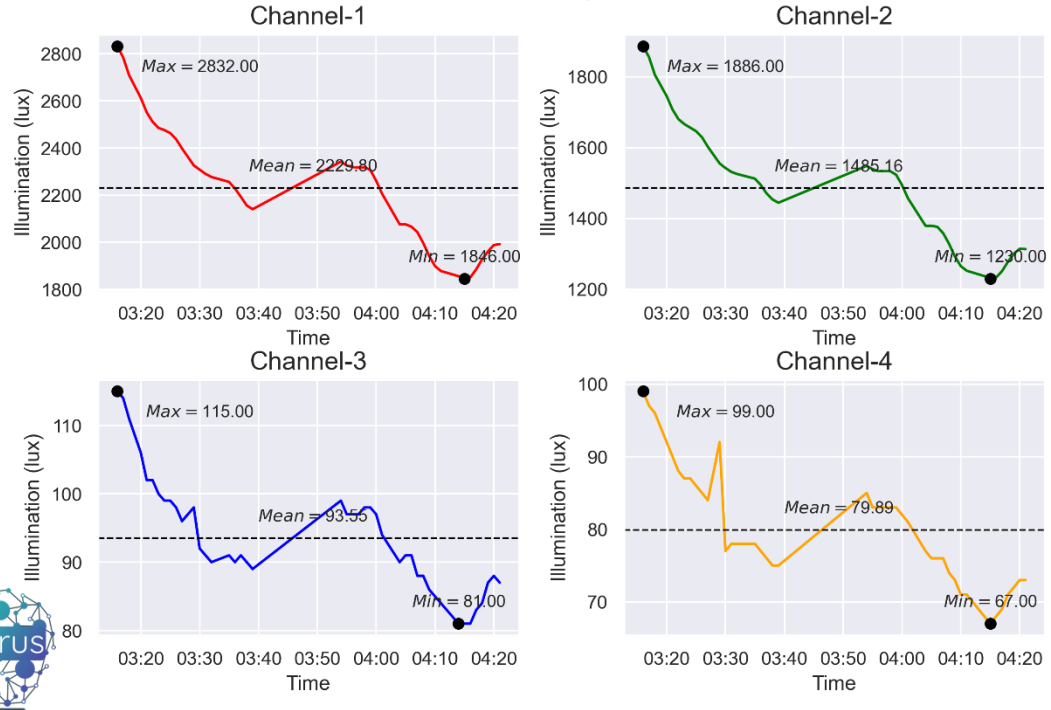
Pyrus pipeline

Results: The Data Visualization Dashboards: DIME & Pyrus





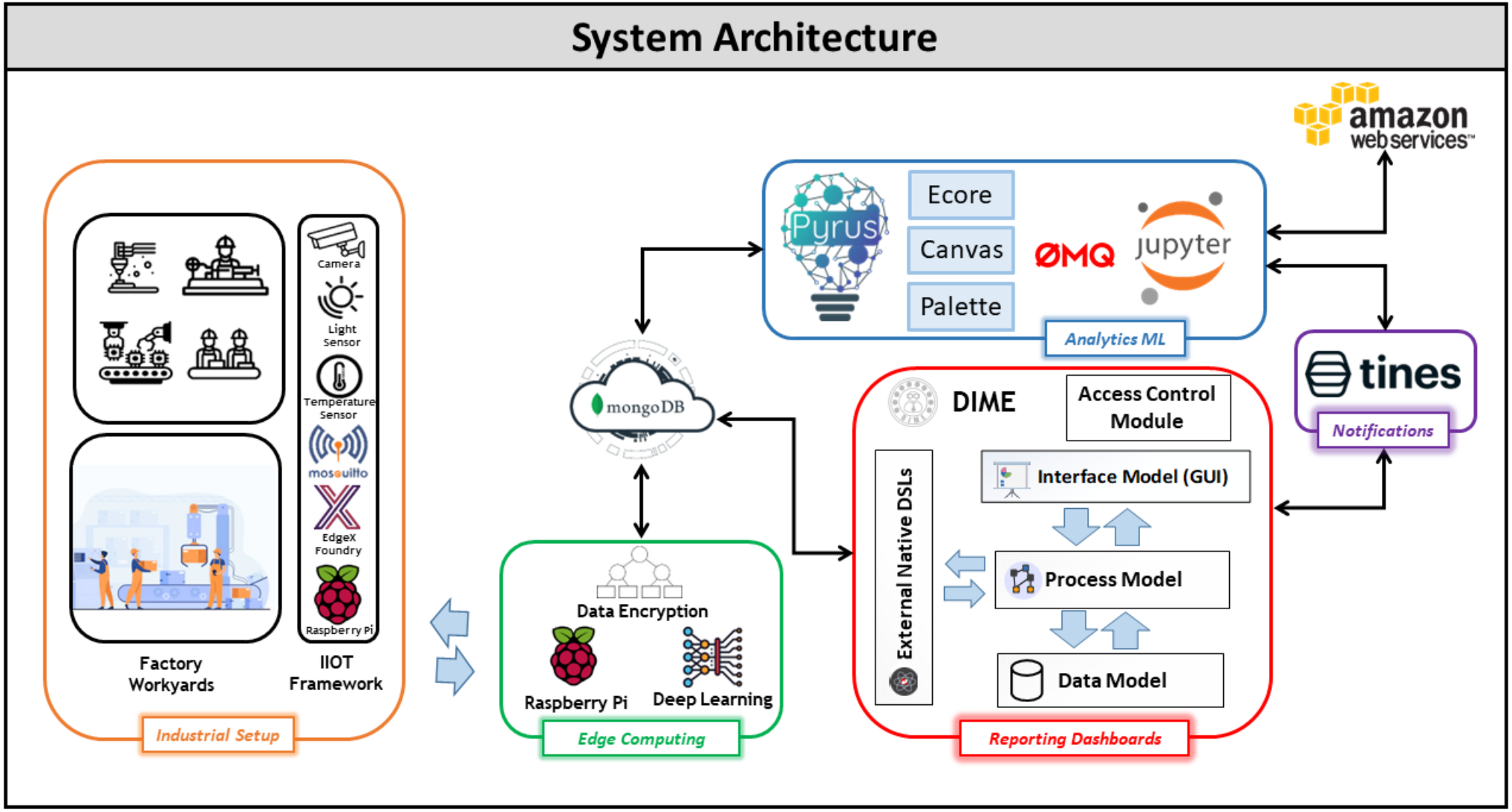




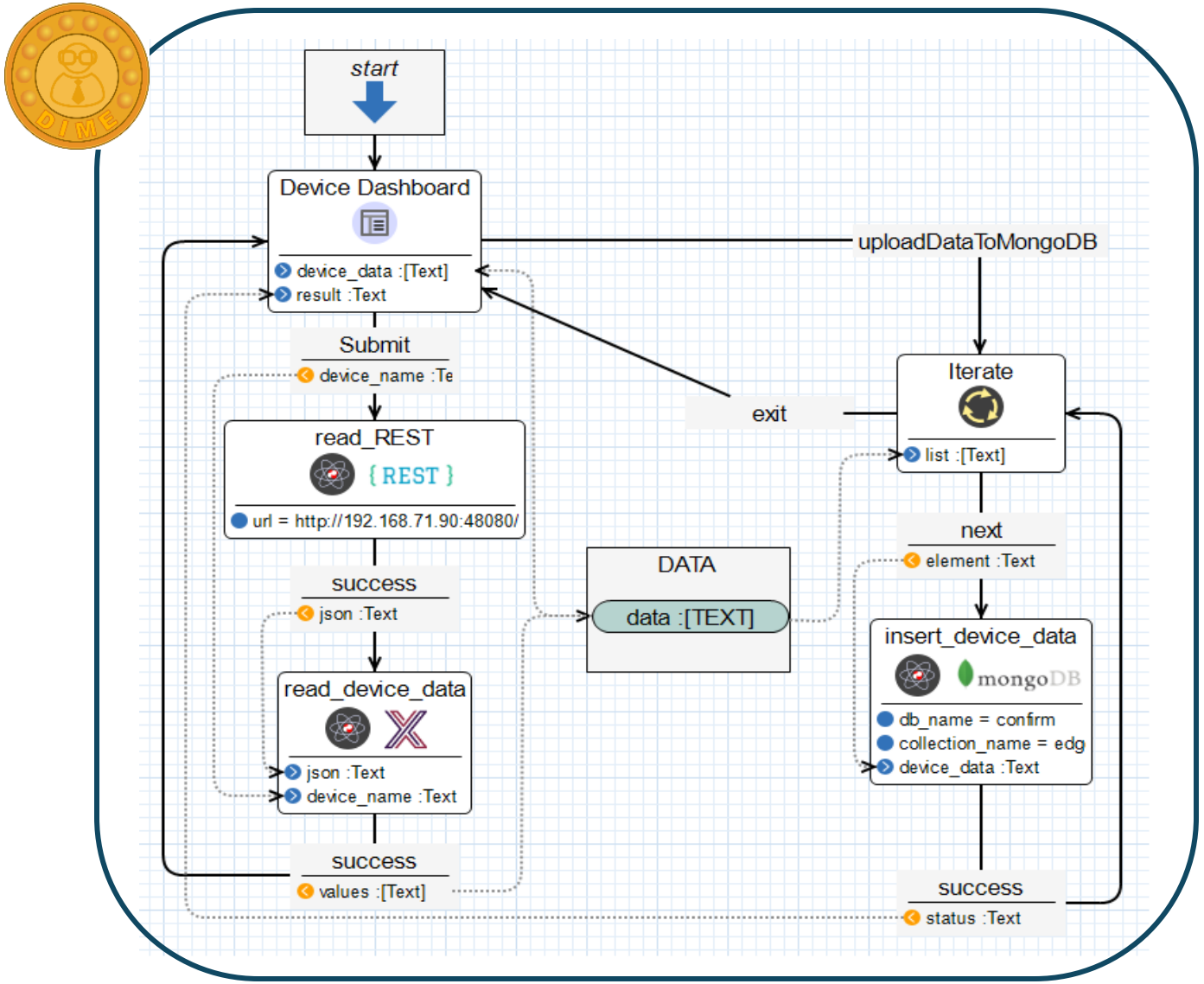
Case Study:
Safety Measures Enforcement in the Workyard

Chaudhary, H.A.A and Guevara, Ivan and John, Jobish and Singh, Amandeep and Ghosal, Amrita and Pesch, Dirk and Margaria, Tiziana. (2022).
Model-Driven Engineering in Digital Thread Platforms: A Practical Use Case and Future Challenges.
In: Leveraging Applications of Formal Methods, Verification and Validation. Practice. ISoLA 2022.
Lecture Notes in Computer Science, vol 13704. Springer, Cham.
https://doi.org/10.1007/978-3-031-19762-8_14

Edge Analytics Use Case: Infrastructure and Communications

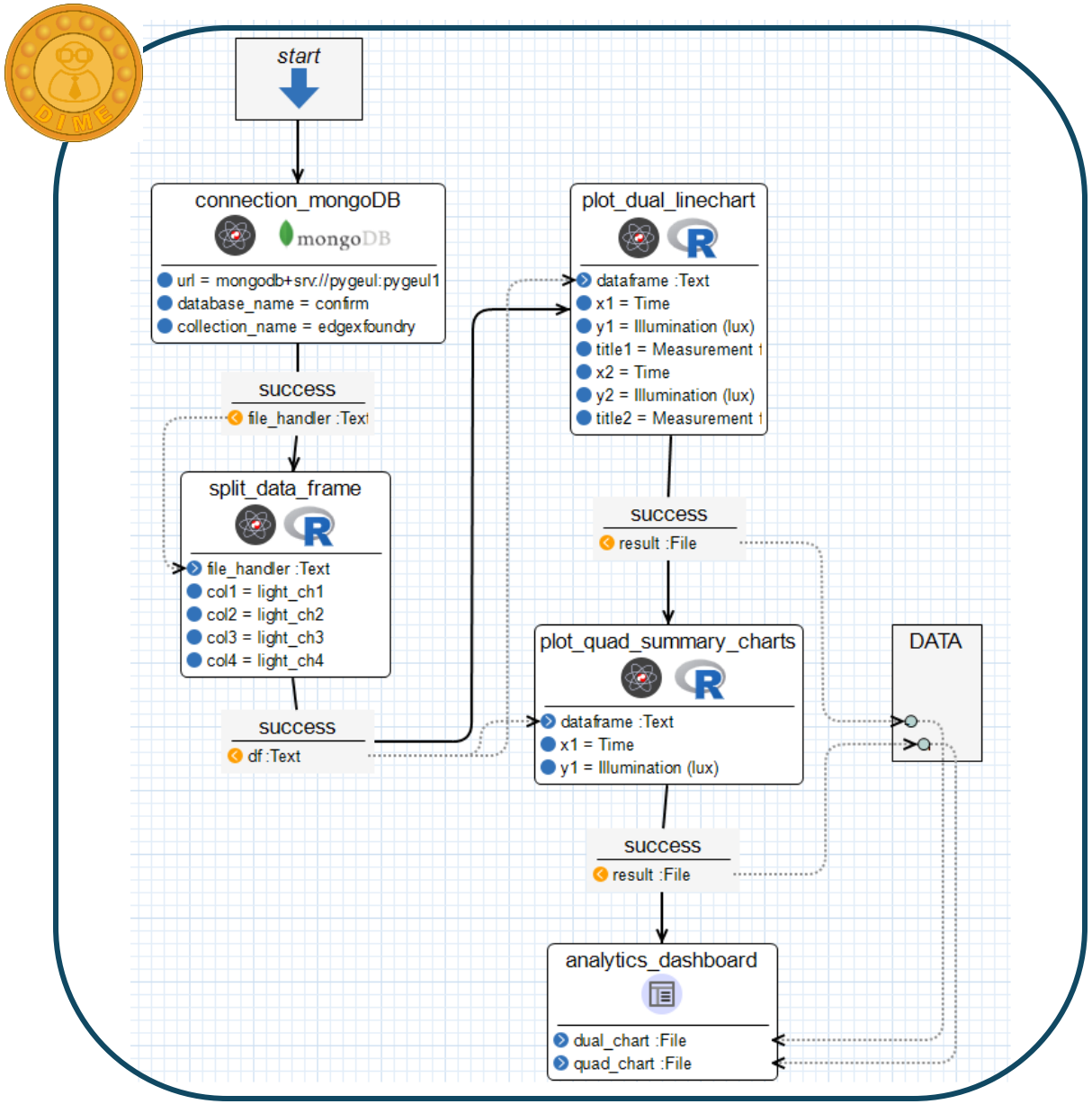


Data Acquisition from EdgeX Foundry and Data Ingestion into MongoDB



Workflow in DIME

Analytics Dashboard in R

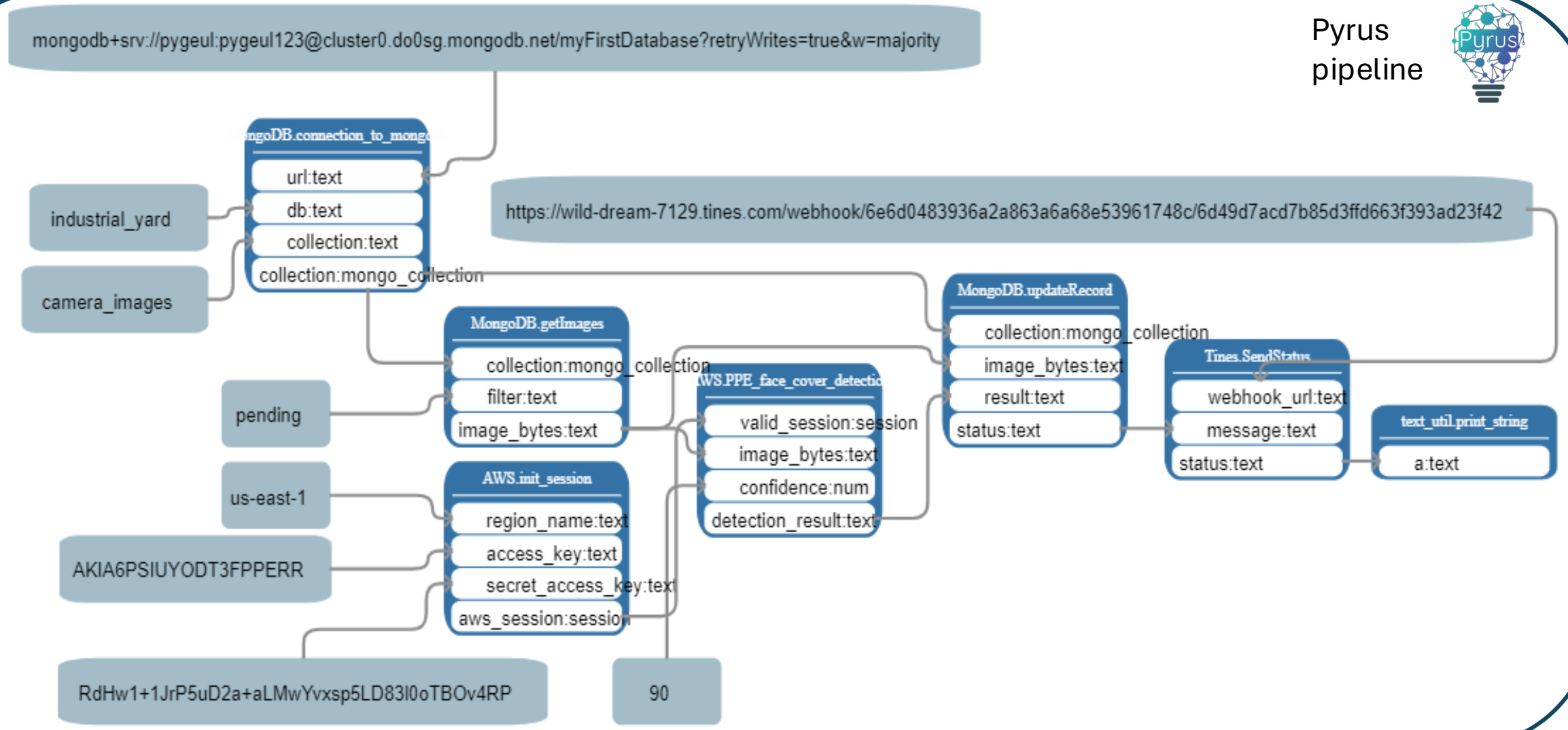


Process in DIME

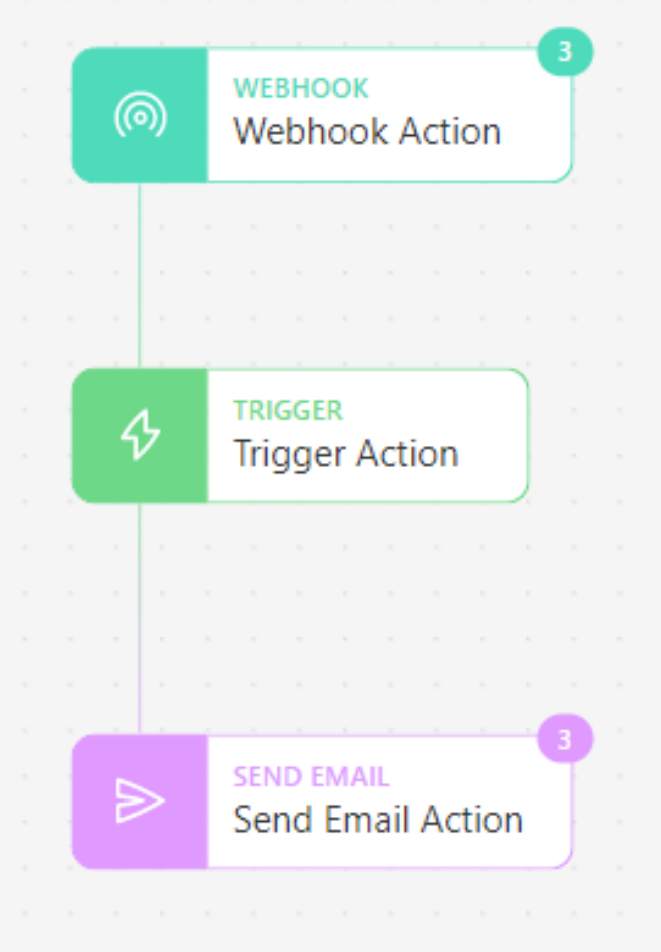
Python Analytics Dashboard in Pyrus

CRT·AI
SFI Centre for Research Training in Artificial Intelligence

Pyrus pipeline



Email Notification Pipeline in Tines



Best Practice of Data Management

FAIR and XAIR and where

FAIR: https://www.go-fair.org/fair-principles/

XAIR: https://www.kg-alliance.org/kga-wg-xai-24-4/

Where: where is your data?

(this morning, on my news)

WIRED

MARCO SCHIAFFINO SECURITY 06.03.2025

Il contrattacco dell'Europa agli Usa parte dal cloud

Un nuovo standard per gestire i servizi nel rispetto delle norme europee. Obiettivo: contrastare lo strapotere delle Big Tech statunitensi

Europe's counterattack to the US starts from the cloud

A new standard for managing services in compliance with European regulations.
Objective: to counter the excessive power of US Big Tech

https://www.wired.it/article/cloud-europa-normative-provider/

Best Practice of Data Management: FAIR

The FAIR Guiding Principles for scientific data management and stewardship
Mark Wilkinson et al. (2016) – Nature
https://www.nature.com/articles/sdata201618

Guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise ***machine-actionability*** (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Covers data and metadata

Best Practice of Data Management: FAIR

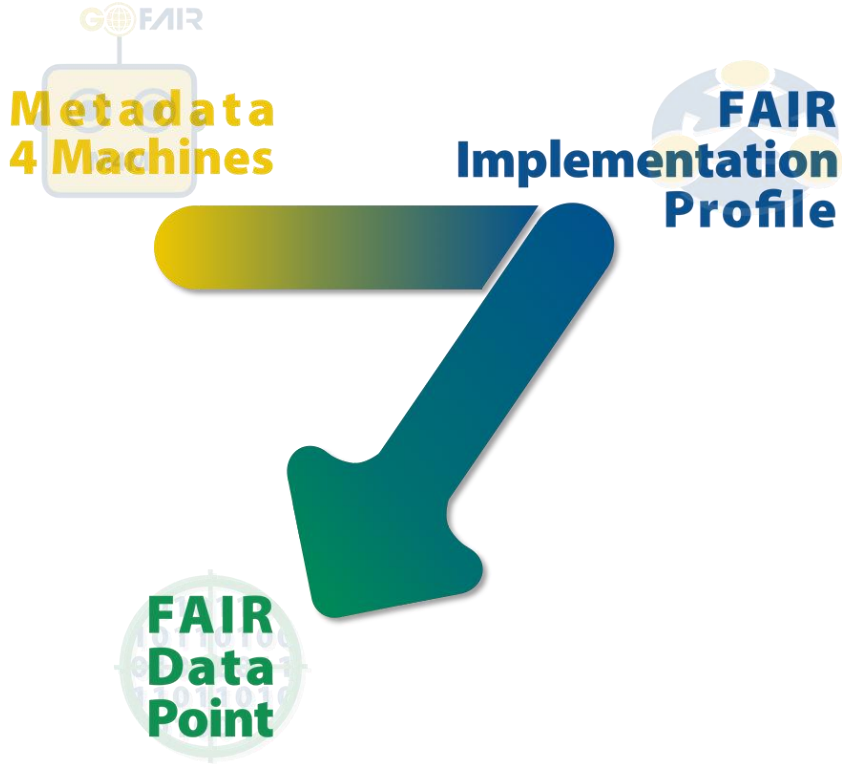
Operationalization:

GO FAIR organization:

https://www.go-fair.org/fair-principles/

Description of the principles:
https://www.go-fair.org/wp-content/uploads/2022/01/FAIRPrinciples_overview.pdf

A practical “how to” guidance to go FAIR:
Three-point FAIRification Framework
https://www.go-fair.org/how-to-go-fair/



FAIR Data Points (FDP)
or
FAIR Digital Objects (FDO)

Best Practice of Data Management: FAIR

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- **F1. (Meta)data are assigned a globally unique and persistent identifier.**
- **F2. Data are described with rich metadata (defined by R1 below).**
- **F3. Metadata clearly and explicitly include the identifier of the data they describe.**
- **F4. (Meta)data are registered or indexed in a searchable resource.**

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

Best Practice of Data Management: FAIR

Accessible

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

- **A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.**
  - **A1.1 The protocol is open, free, and universally implementable.**
  - **A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.**
- **A2. Metadata are accessible, even when the data are no longer available**.

Best Practice of Data Management: FAIR

Interoperable

The data usually need to be integrated with other data.
In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- **I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**
- **I2. (Meta)data use vocabularies that follow FAIR principles.**
- **I3. (Meta)data include qualified references to other (meta)data.**

Best Practice of Data Management: FAIR

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- **R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.**
- **R1.1. (Meta)data are released with a clear and accessible data usage license. R1.2. (Meta)data are associated with detailed provenance.**
- **R1.3. (Meta)data meet domain-relevant community standards**.

Best Practice of Data Management: XAIR

Led by the **Knowledge Graph Alliance (KGA)**
https://www.kg-alliance.org

Working group on **Explainable-AI-ready** data and metadata principles:
https://www.kg-alliance.org/kga-wg-xai-24-4/

The « Explainable-AI-ready data and metadata principles » (XAIR) WG embarks on a mission to define principles that ensure data and models are inherently explainable, aligning with the urgent need for transparency and interpretability in AI applications. This group emphasizes that « Data are XAIR to the degree that they are semantically enriched so that best use can be made of interpretable learning techniques, » including both inductive learning and logical deduction. With the advent of the European Commission's Battery Regulation mandating digital product passports (DPPs) for an expanding range of products, the necessity for data to be explainable-AI-ready (XAIR) is not just a preference but a burgeoning legal requirement. The initiative encapsulates its ethos in the slogan "FAIR and XAIR data," advocating for a future where data's reliability is underpinned by its explainability, ensuring it meets both ethical standards and regulatory compliance.

Best Practice of Data Management: **considerations**

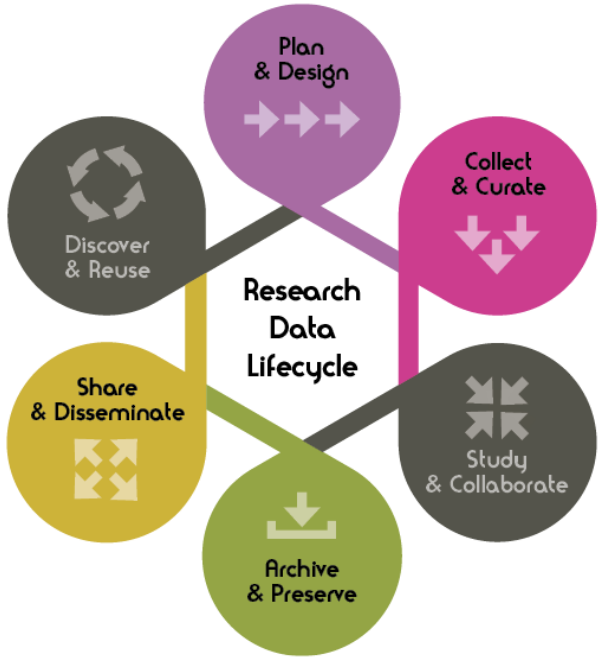
- **Reproducibility:** Document the workflows (= the "how") usually insufficient, usually not delivered together with the data.
  See **Nature Protocols**: https://www.nature.com/nprot/
- **Loopholes:** "Legal" data - Subscribing to a repository, them scraping of DBs and reselling it as a new product
- **Languages:** of the data, of the tools (IT languages, human languages)
- **Logging**: who has done what (changes, updates, but also simple access), for accountability and forensics
- **GDPR:** Data Protection Officers, beyond DPOs. E.g. Irish Computer Society with the ADPOs – very successful
- **Courses** on research integrity, GDPR, research ethics (forms for ethics: too much, too little, too case by case)
- What we do not have is courses on **sustainability** in and for IT
- Example of the Irish GRO recently changing URLs of their published resources: published links don't work anymore

How to Practise Good Research Data Management

Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

How to RDM:
https://dri.ie/how-to-rdm/

How to Practise Good Research Data Management



Data Management Planning

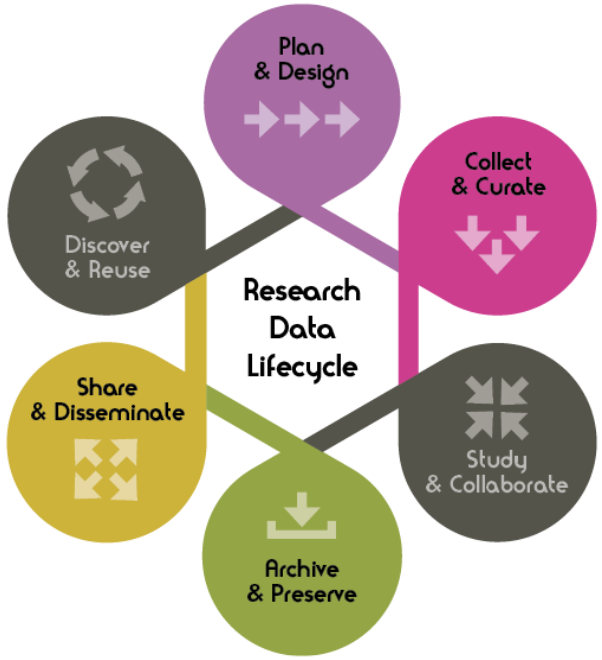
A Data Management Plan (DMP) is a document that can help you to articulate the role that data will play in your research inquiry.
It should outline how you will control the data you collect and/or create over the course of the project, as well as the roles and responsibilities of any collaborators or partners.

It is a living document that you can return to at any time to guide decisions around what data ultimately needs to be shared and how you will do that.

DMPs further help demonstrate transparency, openness and return on public investment by describing how the data can be made discoverable, accessible, and reusable.

Still not sure whether creating a DMP is right for you?
The Australian Research Data Commons (ARDC) answers the question
Why do I need a data management plan?
(https://ardc.edu.au/resources/aboutdata/data-management-plans/)

How to Practise Good Research Data Management



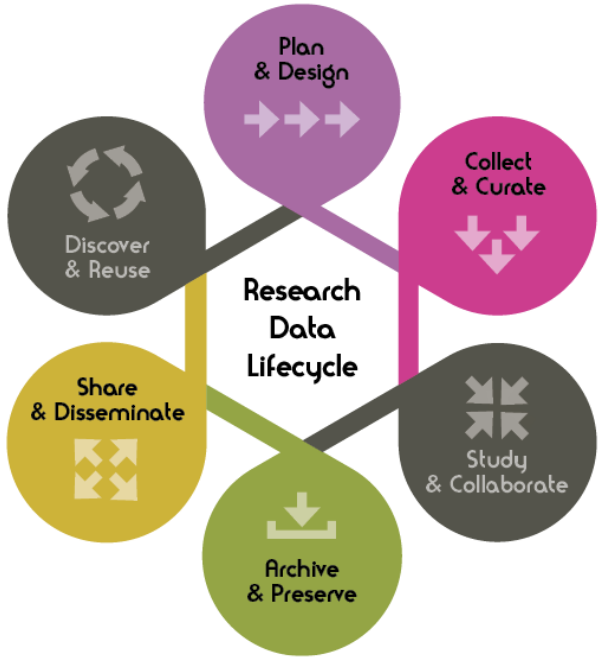
Tools for writing DMPs:

- DMPonline is a UK-based online tool for writing DMPs. Create an account to gain access to a DMP template or browse examples of DMPs submitted by other researchers.

- The DMPTool is the original free, open source application for creating DMPs. The Funder Requirements registry provides templates tailored to the particular information requested by certain funding bodies.

- LIBER (Association of European Research Libraries) maintains a Data Management Plan Catalogue "to inspire researchers and others in the process of writing a Data Management Plan."

How to Practise Good Research Data Management



DMP tips and advice:

Irish Research Council's Tips on Data Management Plans

Research Ireland's Guidance on Data Management Plans

DRI: For more useful tips and links to resources that can help you to fill out the sections of your DMP, download the DRI's RDM Resource Pack.

Table of content

What does Data Management encompass?

- Data types
  basic, complex, …
- Data description
  meaning, provenance, units of measure, properties, …
- Metadata
  types, format, semantic types, …
- Operations on data
  insert, delete, modify, test, CRUD, set_to_null, impute, …
- Datasets:
  - properties (size, format, location (URI?), open access, raw, curated, standards, …)
  - governance (policies, actors, FAIR, XAIR, live/static, …)

Software as Data

- Manage the collection of programs, models, as if they were data to curate properly
- Open source: opportunity and risk
- How to manage the "community":
  - sustainable? – if there are no resources…
  - governance? - centralized is a risk, and it costs money and time

→ DAOs (Distributed Autonomous Organizations),

using Distributed Ledger Technology

e.g. in the community we are building around the LCNC platform

SCCE:
Sustainable Computing for Continuous Engineering

- Components are quite stable
- Systems (must) evolve continuously
- Better choose technologies that are change-friendly
- Models are more understandable than code (→ user-friendly, accessible)
- Models are more portable than code (obsolete technologies → migration)
- Models + code generation help produce less code, in particular less handwritten code, which is faulty code
- "Left shift" of verification and checks:

  correct and improve together the models (at design time)

  instead of painstakingly chasing bugs in the source code (over and over again) at runtime

  "needle In haystack"

The “One Thing” Approach

prevention,
instead of repair

GREAT LEAP

Best practices for data management

Tiziana Margaria

University of Limerick

tiziana.margaria@ul.ie