Curtin University

THICKSHAKE

HISTORICAL IMAGE CLASSIFICATION SYSTEM

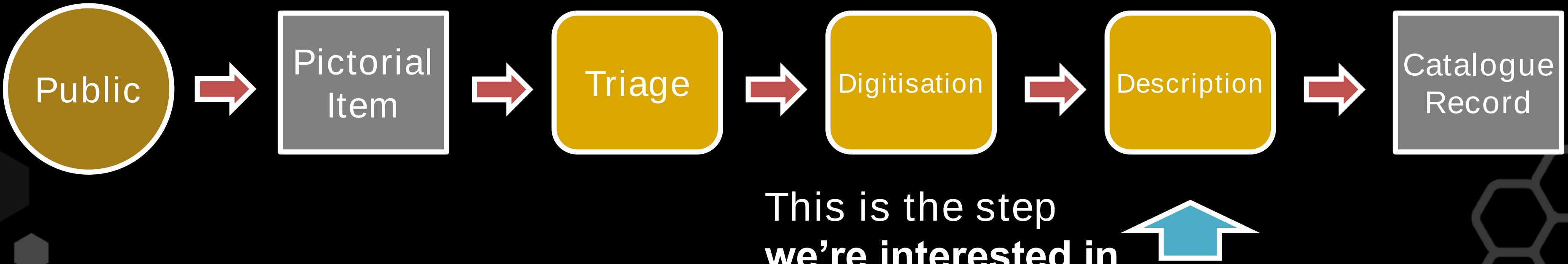
Mark Shelton | 16 February 2018

github.com/markshelton/thickshake

Project sponsored by the Pawsey Supercomputing Centre.

The State Library of Western Australia (SLWA) holds more than one million items in its pictorial collection.

In 2016-17, more than 30,000 items were added to the SLWA’s catalogue. This process is expensive and time-consuming.



Public → Pictorial Item → Triage → Digitisation → Description → Catalogue Record

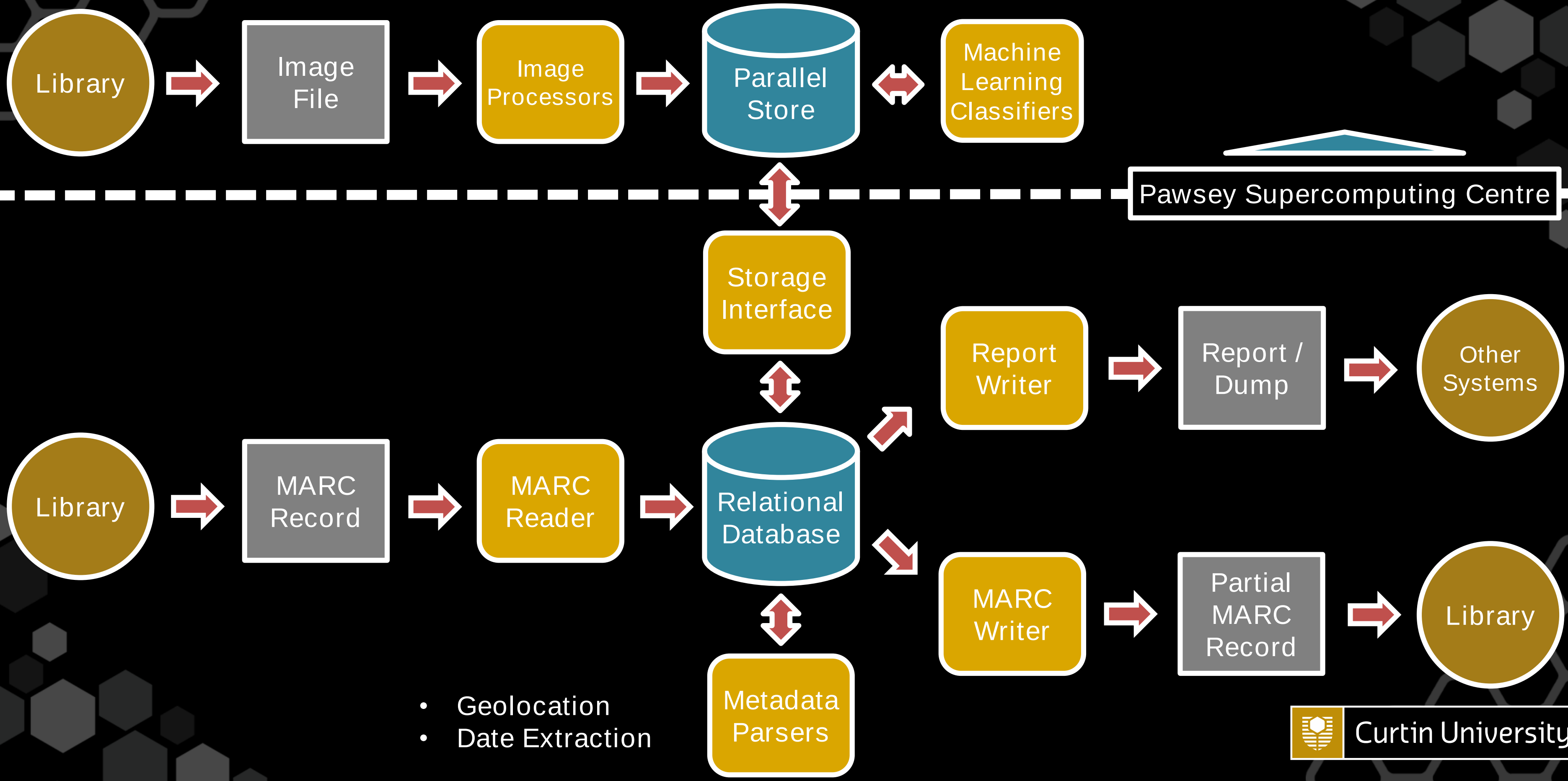
This is the step
we’re interested in

Contributions

- A flexible interface for manipulating library catalogue metadata
- A suite of functions that augment library catalogue metadata
- A back-end system that leverages high performance computing

- Face Detection
- Text Recognition

- Subject Recognition
- Landmark Recognition



Library

Image File

Image Processors

Parallel Store

Machine Learning Classifiers

Pawsey Supercomputing Centre

Storage Interface

Library

MARC Record

MARC Reader

Relational Database

Report Writer

Report / Dump

Other Systems

MARC Writer

Partial MARC Record

Library

Metadata Parsers

- Geolocation
- Date Extraction

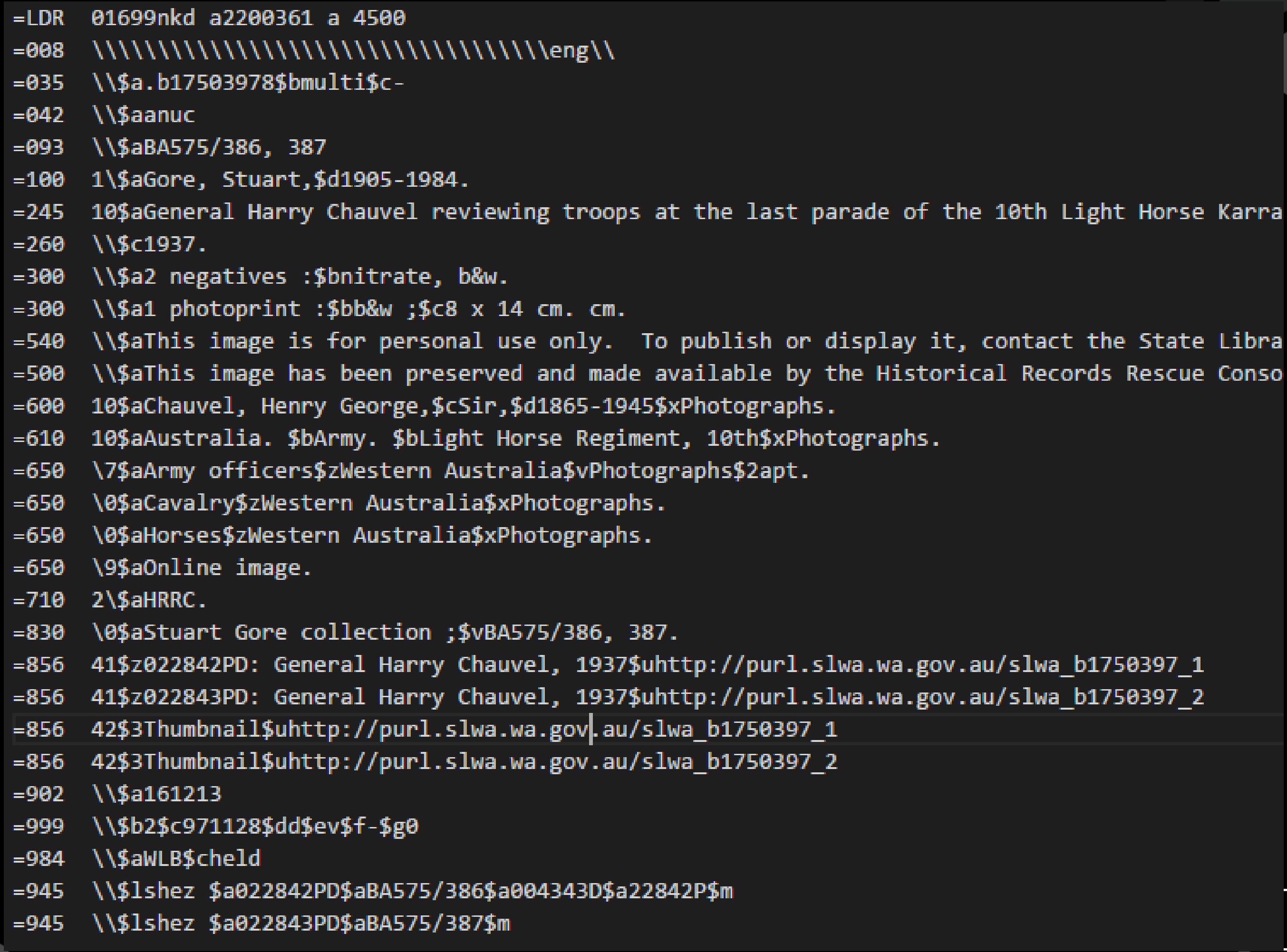
Curtin University

Structure

- Library Interface
- Metadata Parsing
- Image Processing
- Machine Learning

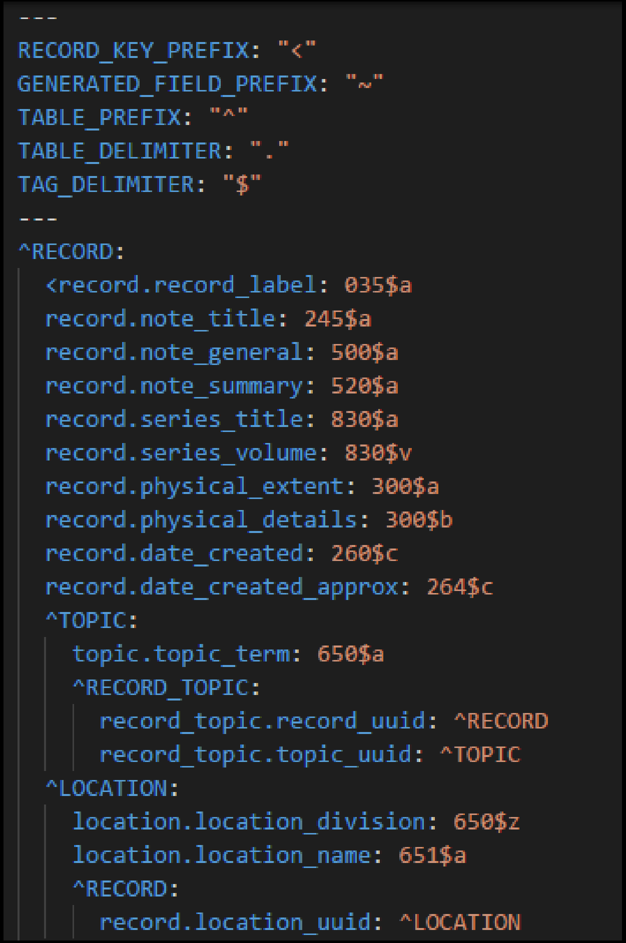
LIBRARY INTERFACE

Curtin University



```
=LDR  01699nkd a2200361 a 4500
=008  \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\eng\\
=035  \\$a.b17503978$bmulti$c-
=042  \\$aanuc
=093  \\$aBA575/386, 387
=100  1\$aGore, Stuart,$d1905-1984.
=245  10$aGeneral Harry Chauvel reviewing troops at the last parade of the 10th Light Horse Karra
=260  \\$c1937.
=300  \\$a2 negatives :$bnitrate, b&w.
=300  \\$a1 photoprint :$bb&w ;$c8 x 14 cm. cm.
=540  \\$aThis image is for personal use only.  To publish or display it, contact the State Libra
=500  \\$aThis image has been preserved and made available by the Historical Records Rescue Conso
=600  10$aChauvel, Henry George,$cSir,$d1865-1945$xPhotographs.
=610  10$aAustralia. $bArmy. $bLight Horse Regiment, 10th$xPhotographs.
=650  \7$aArmy officers$zWestern Australia$vPhotographs$2apt.
=650  \0$aCavalry$zWestern Australia$xPhotographs.
=650  \0$aHorses$zWestern Australia$xPhotographs.
=650  \9$aOnline image.
=710  2\$aHRRC.
=830  \0$aStuart Gore collection ;$vBA575/386, 387.
=856  41$z022842PD: General Harry Chauvel, 1937$uhttp://purl.slwa.wa.gov.au/slwa_b1750397_1
=856  41$z022843PD: General Harry Chauvel, 1937$uhttp://purl.slwa.wa.gov.au/slwa_b1750397_2
=856  42$3Thumbnail$uhttp://purl.slwa.wa.gov.au/slwa_b1750397_1
=856  42$3Thumbnail$uhttp://purl.slwa.wa.gov.au/slwa_b1750397_2
=902  \\$a161213
=999  \\$b2$c971128$dd$ev$f-$g0
=984  \\$aWLB$cheld
=945  \\$lshez $a022842PD$aBA575/386$a004343D$a22842P$m
=945  \\$lshez $a022843PD$aBA575/387$m
```

We have developed a system that maps MARC records onto a relational database. The interface works on a user-defined map.



```
---
RECORD_KEY_PREFIX: "<"
GENERATED_FIELD_PREFIX: "~"
TABLE_PREFIX: "^"
TABLE_DELIMITER: "."
TAG_DELIMITER: "$"
---
^RECORD:
  <record.record_label: 035$a
  record.note_title: 245$a
  record.note_general: 500$a
  record.note_summary: 520$a
  record.series_title: 830$a
  record.series_volume: 830$v
  record.physical_extent: 300$a
  record.physical_details: 300$b
  record.date_created: 260$c
  record.date_created_approx: 264$c
  ^TOPIC:
    topic.topic_term: 650$a
    ^RECORD_TOPIC:
      record_topic.record_uuid: ^RECORD
      record_topic.topic_uuid: ^TOPIC
  ^LOCATION:
    location.location_division: 650$z
    location.location_name: 651$a
    ^RECORD:
      record.location_uuid: ^LOCATION
```

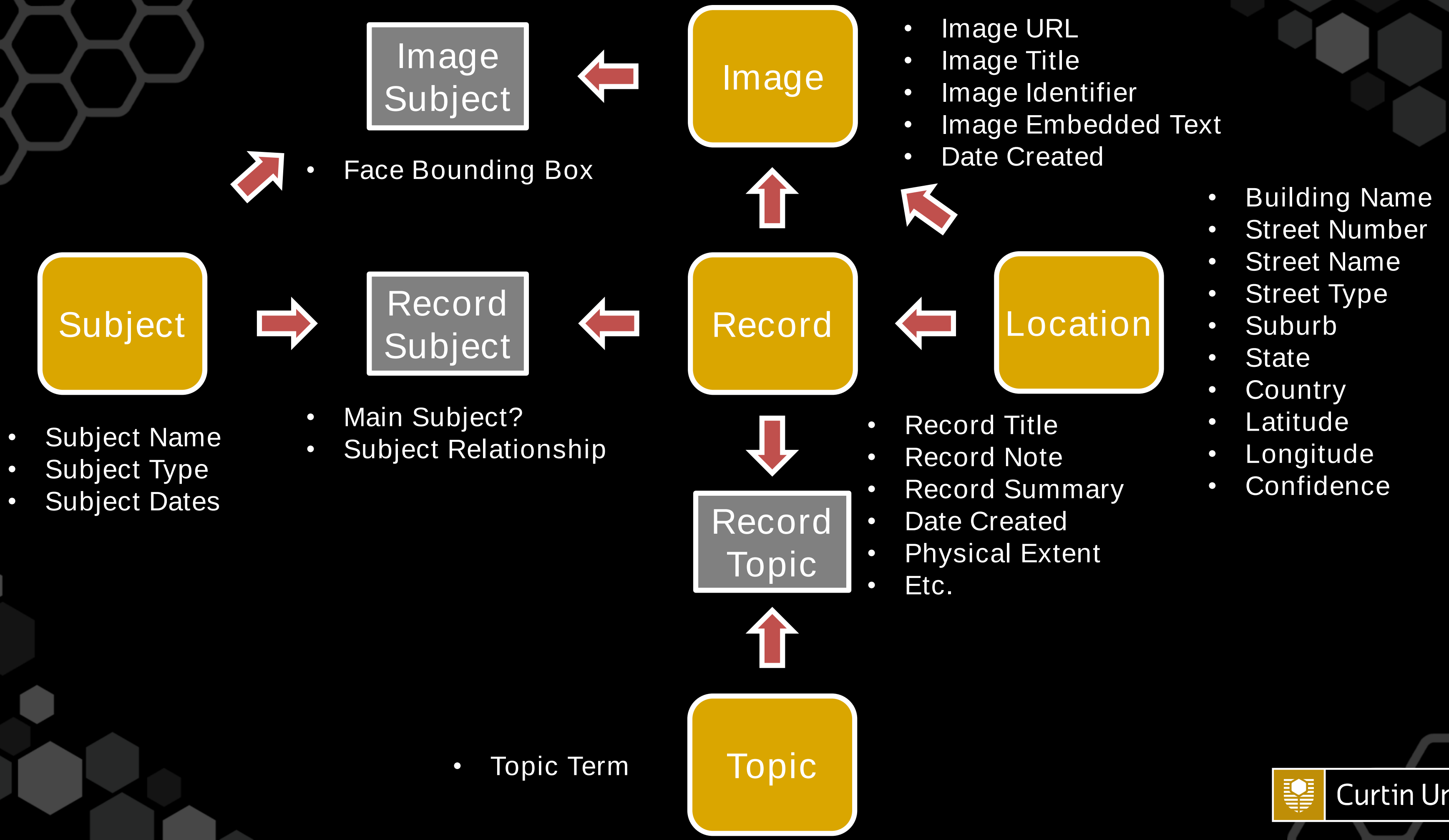


Image Subject

Image

- Image URL
- Image Title
- Image Identifier
- Image Embedded Text
- Date Created

- Face Bounding Box

- Building Name
- Street Number
- Street Name
- Street Type
- Suburb
- State
- Country
- Latitude
- Longitude
- Confidence

Subject

Record Subject

Record

Location

- Subject Name
- Subject Type
- Subject Dates

- Main Subject?
- Subject Relationship

- Record Title
- Record Note
- Record Summary
- Date Created
- Physical Extent
- Etc.

Record Topic

- Topic Term

Topic

Curtin University

Test Dataset Overview

- Records – 3,048
- Images – 10,106
- Subjects – 4,030
  - People (2,036)
  - Top: EL Mitchell, Betty Smith, AH Stone
- Topics – 1,722
  - Top: Interiors, Hotels, Streets

METADATA PARSING

Mappify API
2,500 request limit per day (free)
Based on PSMA G-NAF dataset

Image Title

311688PD: Durham House building premises of Wrightson Dance Studios, The Inn Trim hairdressers, Galore House (no. 842) and Marjorie Young Antiques (no. 836) Hay Street, Perth, December 1982

→

Parsed Address

building_name: None
street_number: ‘836-842’
street_name: 'Hay’
street_type: 'Street'
suburb: 'Perth’
state: 'WA'
location_type: ‘parsed’

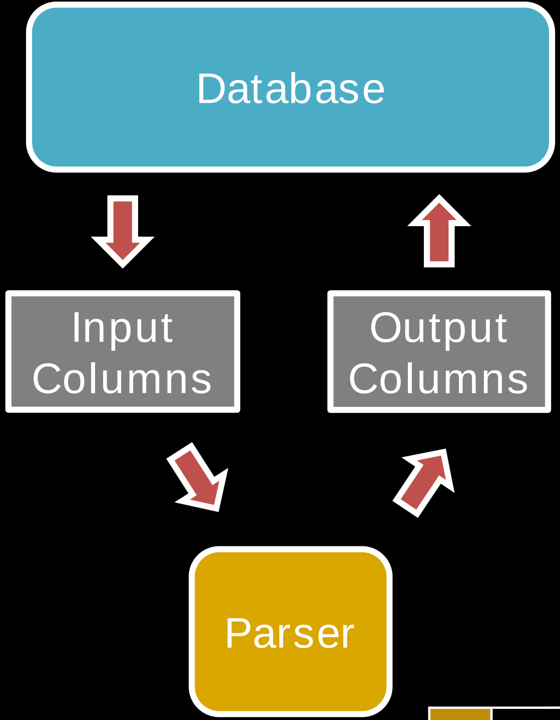
→

Geocoded Address

building_name: None
street_number: '838-842’
street_name: 'Hay’
street_type: 'Street'
suburb: 'Perth’
post_code: '6000’
state: 'WA'
longitude: 115.85448
latitude: -31.95236,
confidence: 0.05,
location_type: 'geocoded’

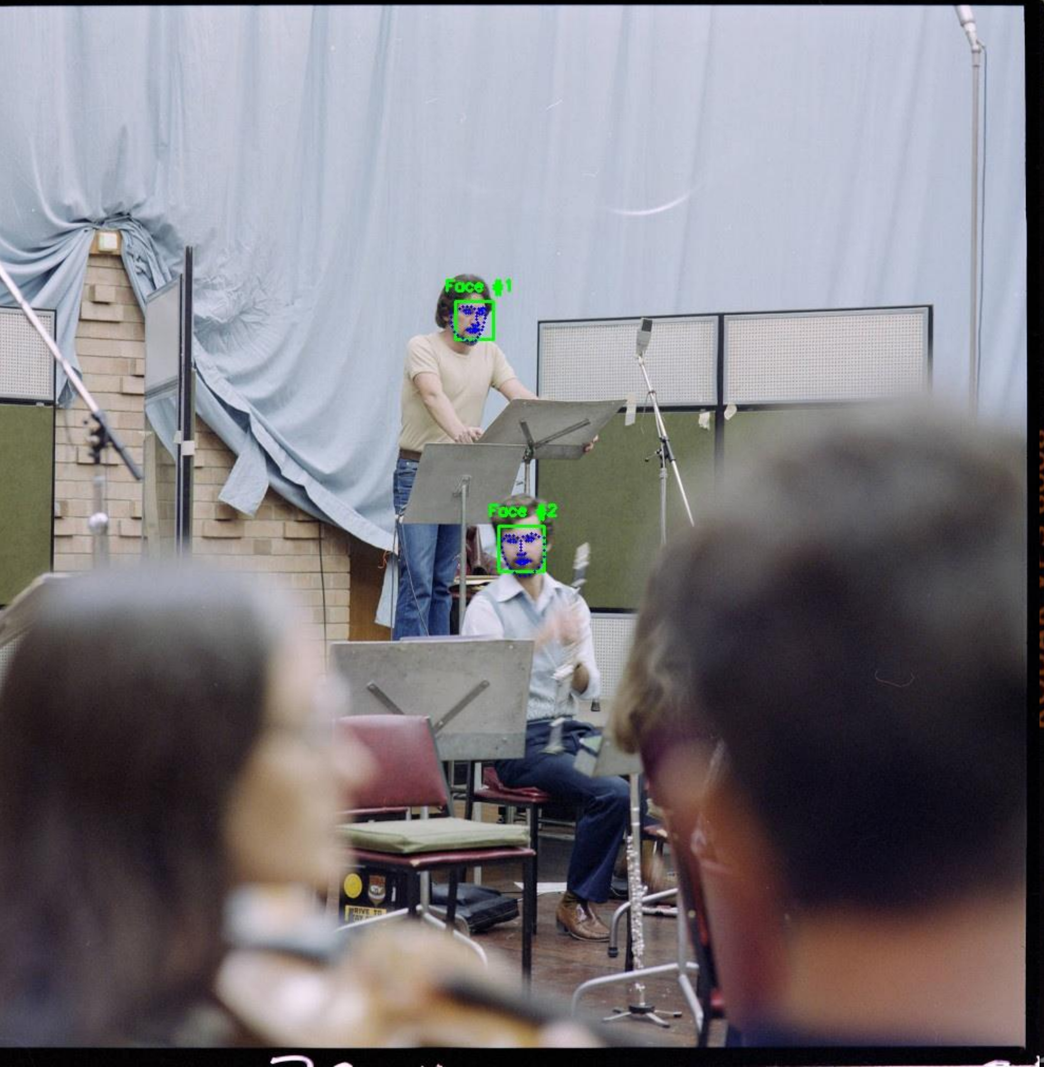
Metadata Parsing Wrapper

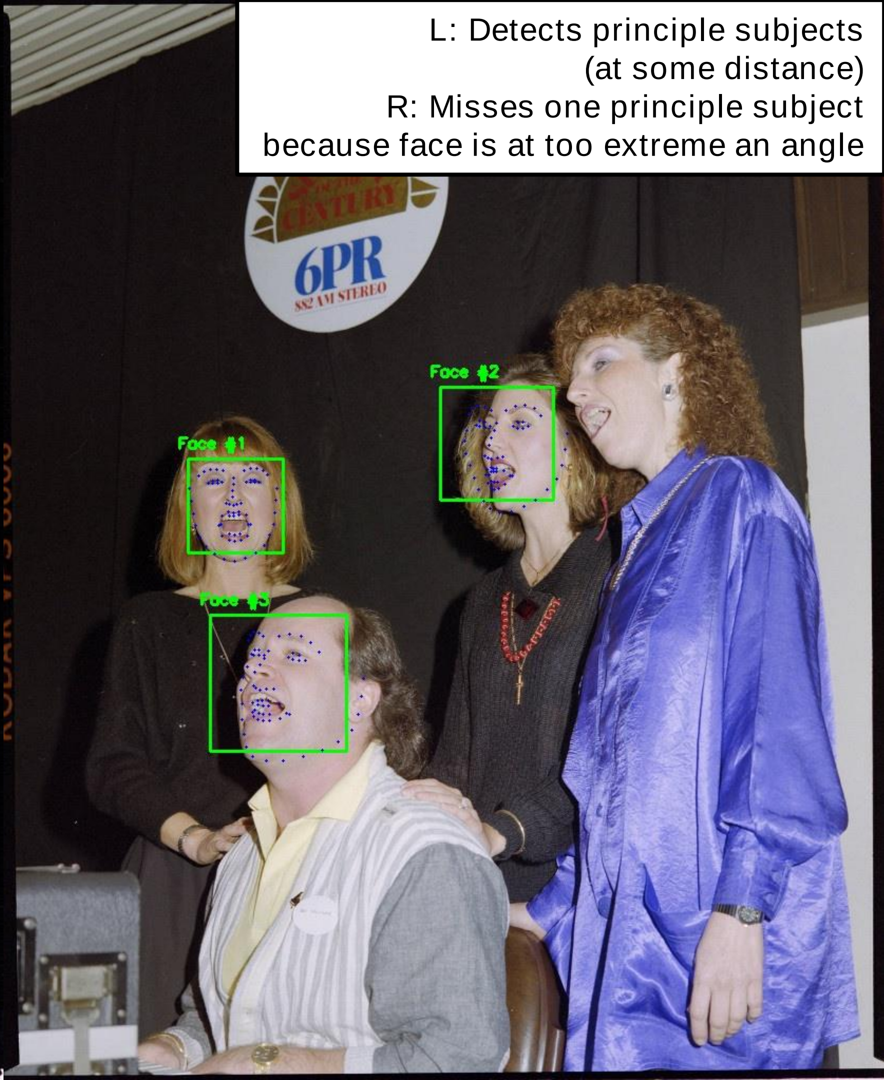
Input Table: image
Input Columns: [“image note”]
Parser Function: extract location
Parser Arguments: None
Output Table: location
Output Map: {
    "index": "image_uuid",
    "building_name": "building_name",
    "street_number": "street_number
    etc.
}



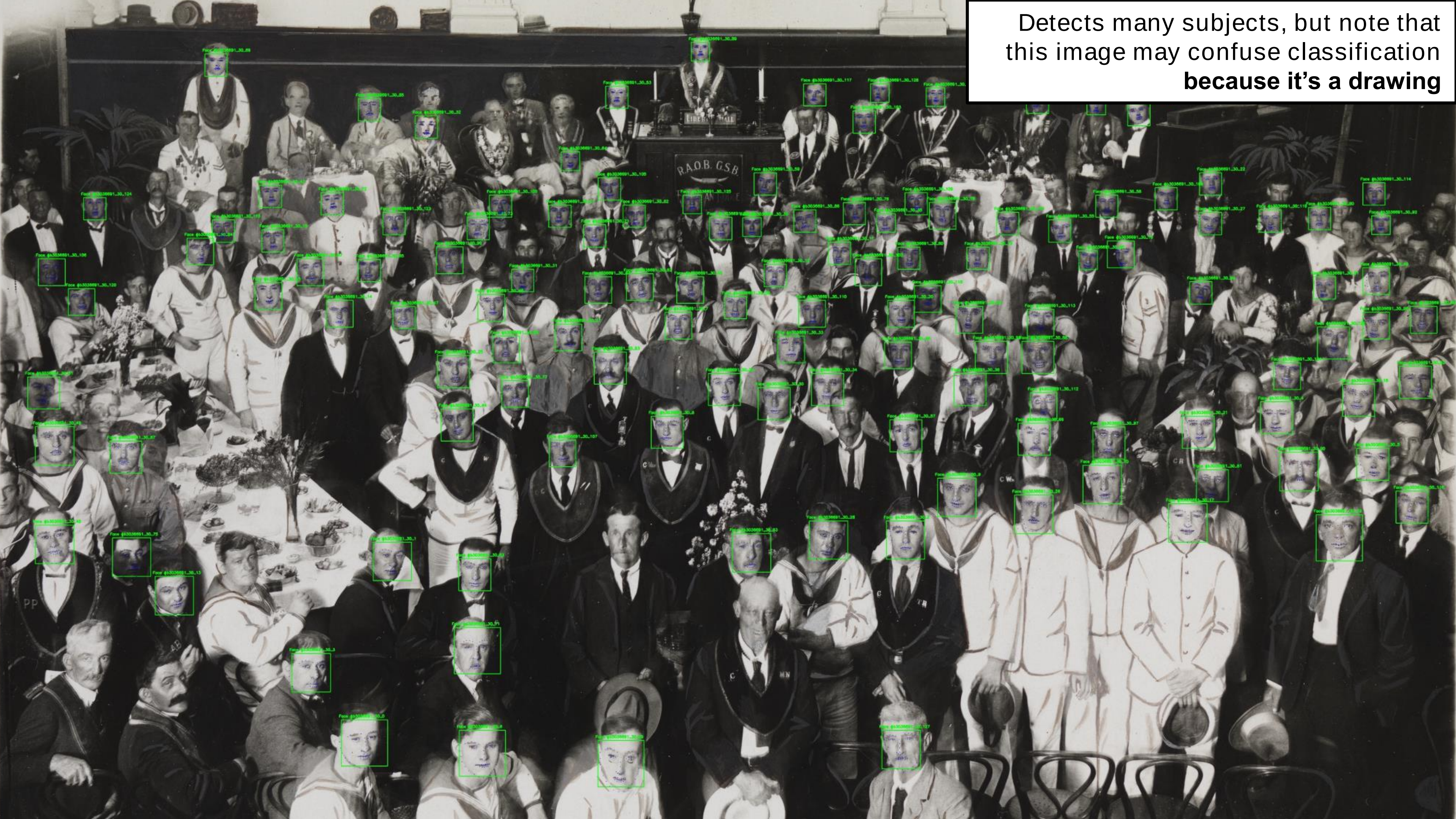
Curtin University

IMAGE PROCESSING





L: Detects principle subjects
(at some distance)
R: Misses one principle subject
because face is at too extreme an angle



Detects many subjects, but note that this image may confuse classification **because it's a drawing**

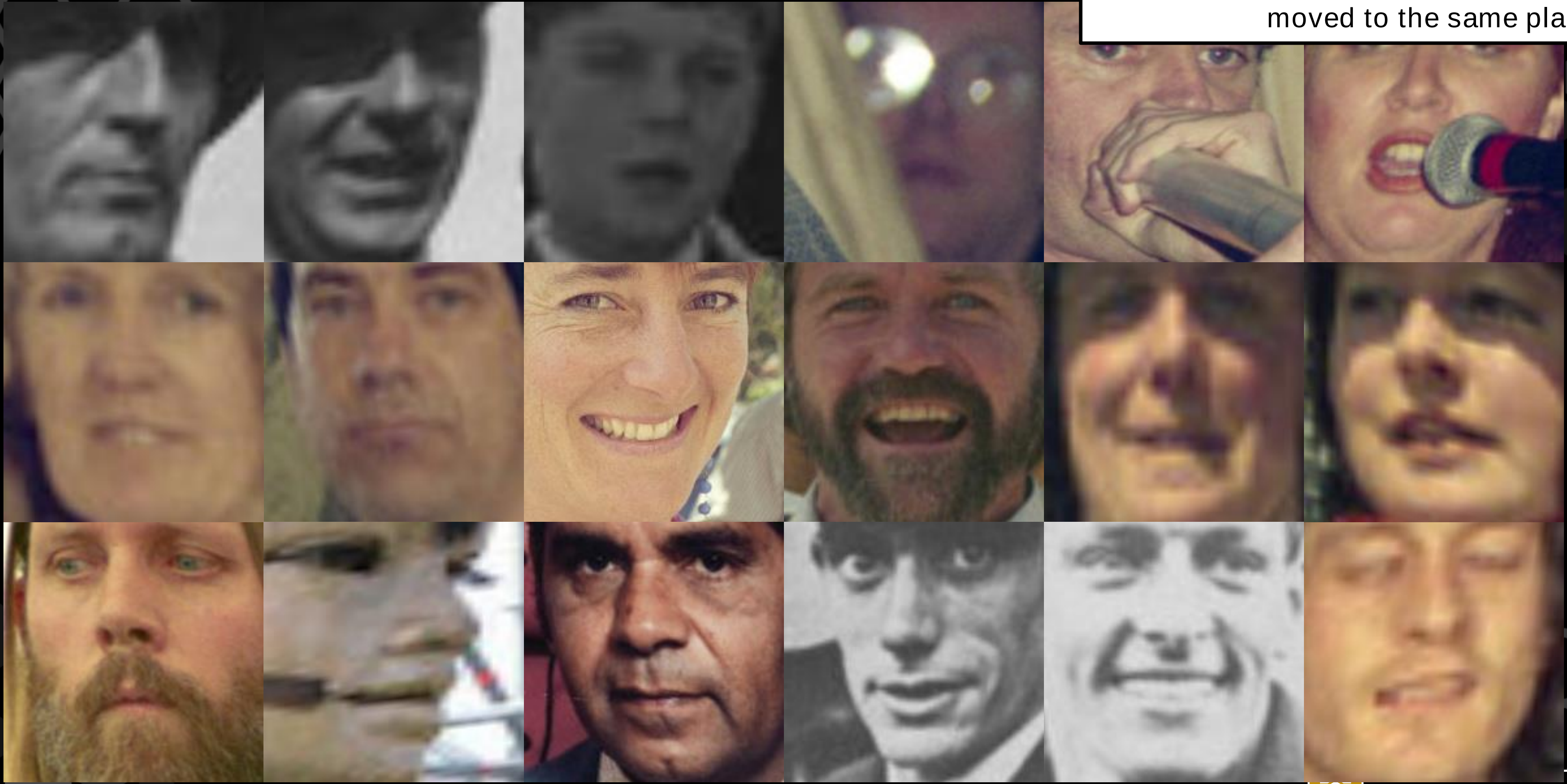
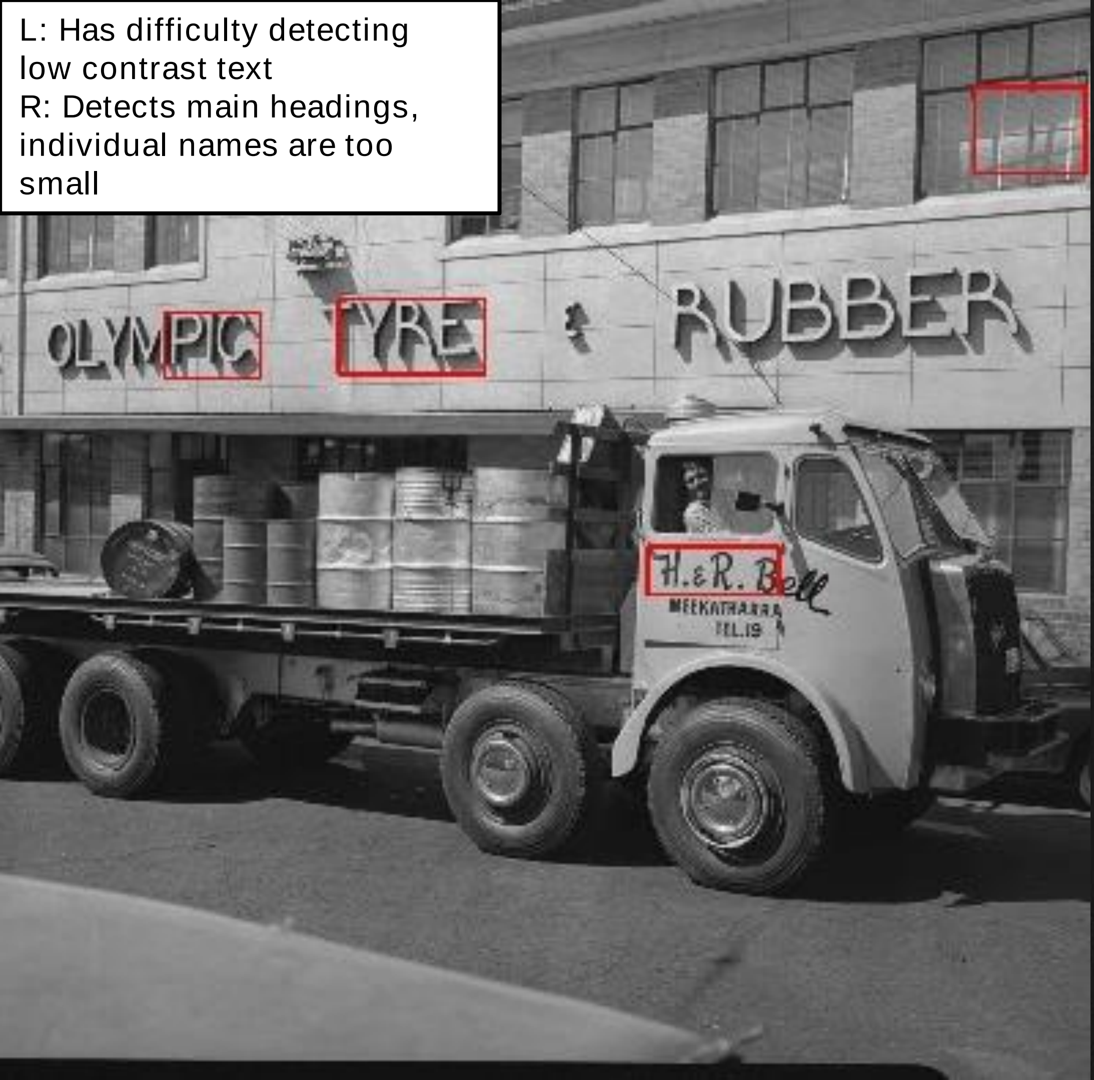
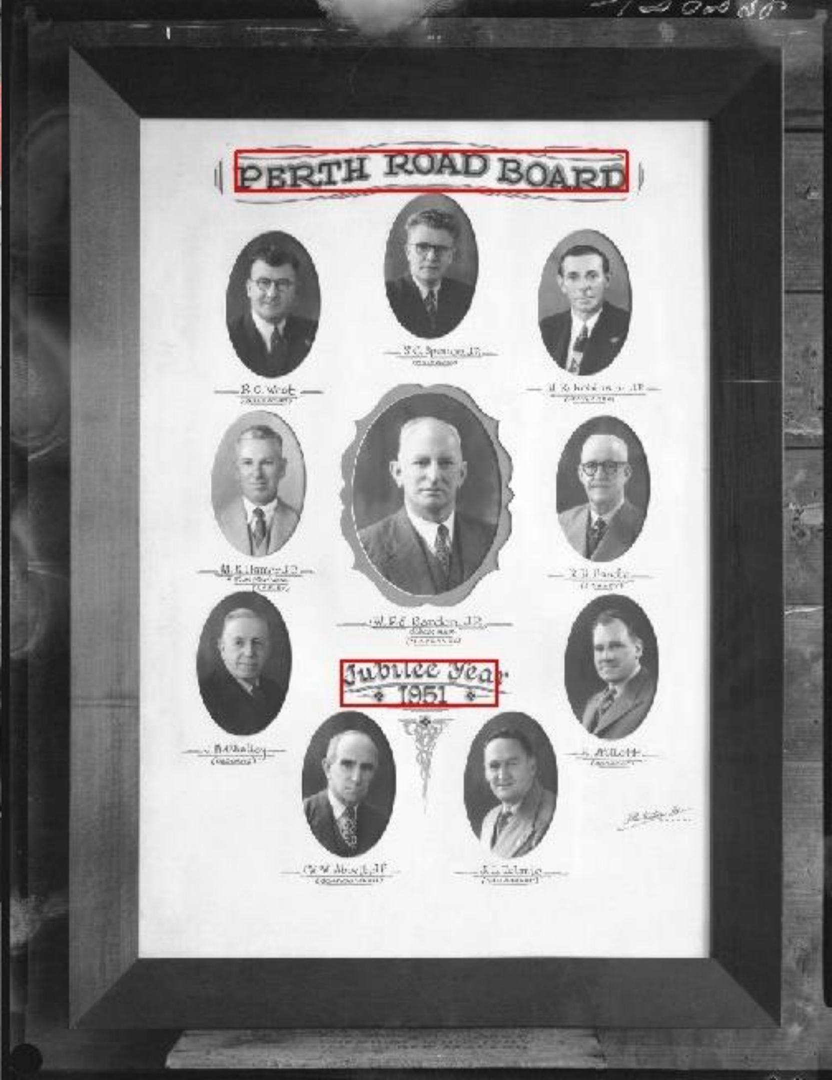


Photo distortion is caused by normalisation – all eyes and lips are moved to the same place.

L: Has difficulty detecting low contrast text
R: Detects main headings, individual names are too small





MACHINE LEARNING

Face Predictions

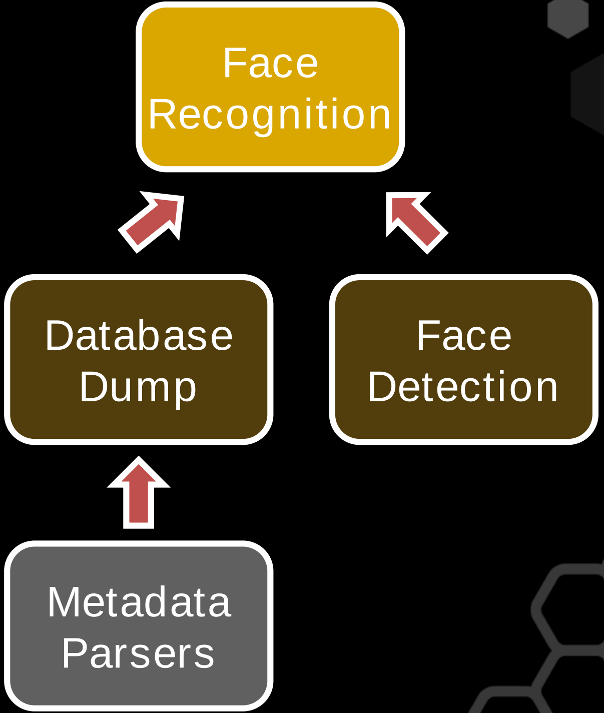
$$\text{label}_{\text{pred}} = \text{label}_{\text{known}} + \text{features}_{\text{metadata}} + \text{features}_{\text{face}}$$

Challenges:

- Labels are per record (not image) -> uncertainty
  - Investigating semi-supervised learning
- Most subjects have few observations

Processing Tree

- Check if store already contains the output of this function
- If not:
  - Check if store contains output of each function dependency
  - If not: Run dependency function and store output (incrementally)
  - Run main function and store output (incrementally)
- Check if database already contains output of this function
- If not: Map each row in store to record in database



CONCLUSIONS

Automated Build System

App Container

* Python 3.5
* Tensorflow 1.4.0
* OpenCV 3.3.1
* dlib 19.7

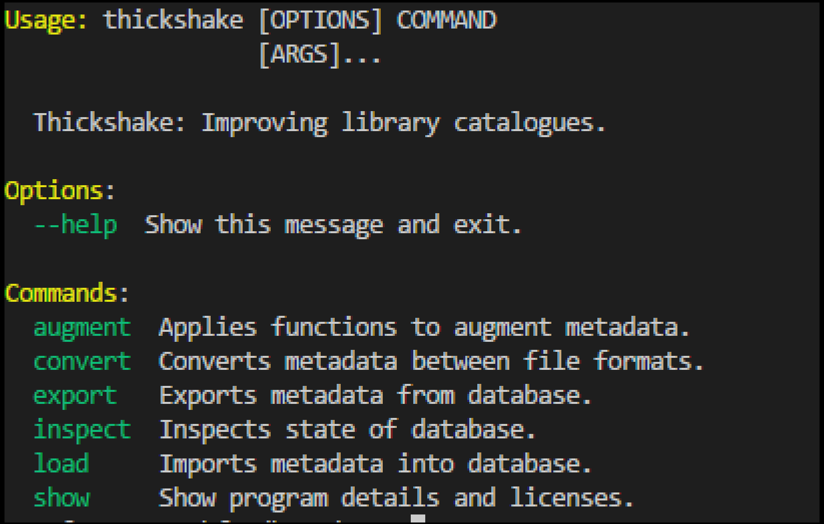
Database Container

* PostgreSQL 10.1

Make commands:

```
make start # loads and builds images, creates volume, creates virtual network
make stop # saves environment, stops containers, removes virtual network
make restart # stops and restarts containers and virtual network
make jupyter # opens jupyter service in default internet browser
make shell # opens interactive session with app container
make push # tags app image and pushes image to DockerHub
```

Command Line Interface

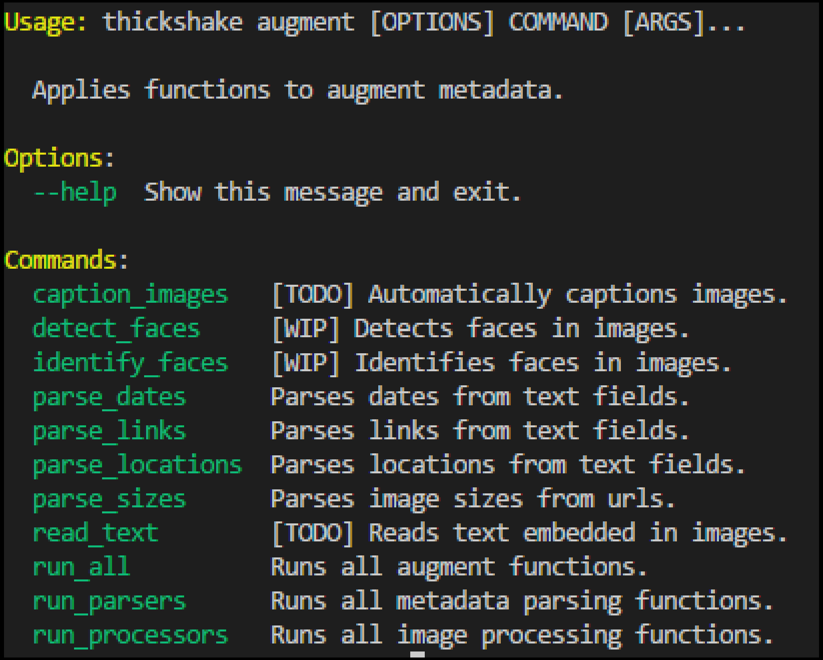


```
Usage: thickshake [OPTIONS] COMMAND
                  [ARGS]...

  Thickshake: Improving library catalogues.

Options:
  --help  Show this message and exit.

Commands:
  augment   Applies functions to augment metadata.
  convert   Converts metadata between file formats.
  export    Exports metadata from database.
  inspect   Inspects state of database.
  load      Imports metadata into database.
  show      Show program details and licenses.
```



```
Usage: thickshake augment [OPTIONS] COMMAND [ARGS]...

  Applies functions to augment metadata.

Options:
  --help  Show this message and exit.

Commands:
  caption_images    [TODO] Automatically captions images.
  detect_faces      [WIP] Detects faces in images.
  identify_faces    [WIP] Identifies faces in images.
  parse_dates       Parses dates from text fields.
  parse_links       Parses links from text fields.
  parse_locations   Parses locations from text fields.
  parse_sizes       Parses image sizes from urls.
  read_text         [TODO] Reads text embedded in images.
  run_all           Runs all augment functions.
  run_parsers       Runs all metadata parsing functions.
  run_processors    Runs all image processing functions.
```

Contributions

- A flexible interface for manipulating library catalogue metadata
- A suite of functions that augment library catalogue metadata
- A back-end system that leverages high performance computing

Next Steps

- Continue to build machine learning functions (e.g. landmark recognition, image captioning)
- Integrate Thickshake with OldPerth to get as much of the catalogue on the map as possible

Thank you …

- Joshua Hollick (HIVE, Lead Supervisor)
- Andrew Woods (HIVE, Supervisor)
- Debra Jones (SLWA, Supervisor)
- Sussanah Soon (HIVE, Collaborator)
- William Olman (HIVE, Collaborator)
- Barbara Patison (SLWA, Advisor)
- Adrian Bowen (SLWA, Advisor)
- Catherine Kelso (SLWA, Advisor)
- David Ong (SLWA, Advisor)
- Maciej Cytowski (Pawsey, Advisor)

Curtin University

THICKSHAKE

HISTORICAL IMAGE CLASSIFICATION SYSTEM

Mark Shelton | 16 February 2018

github.com/markshelton/thickshake

Project sponsored by the Pawsey Supercomputing Centre.