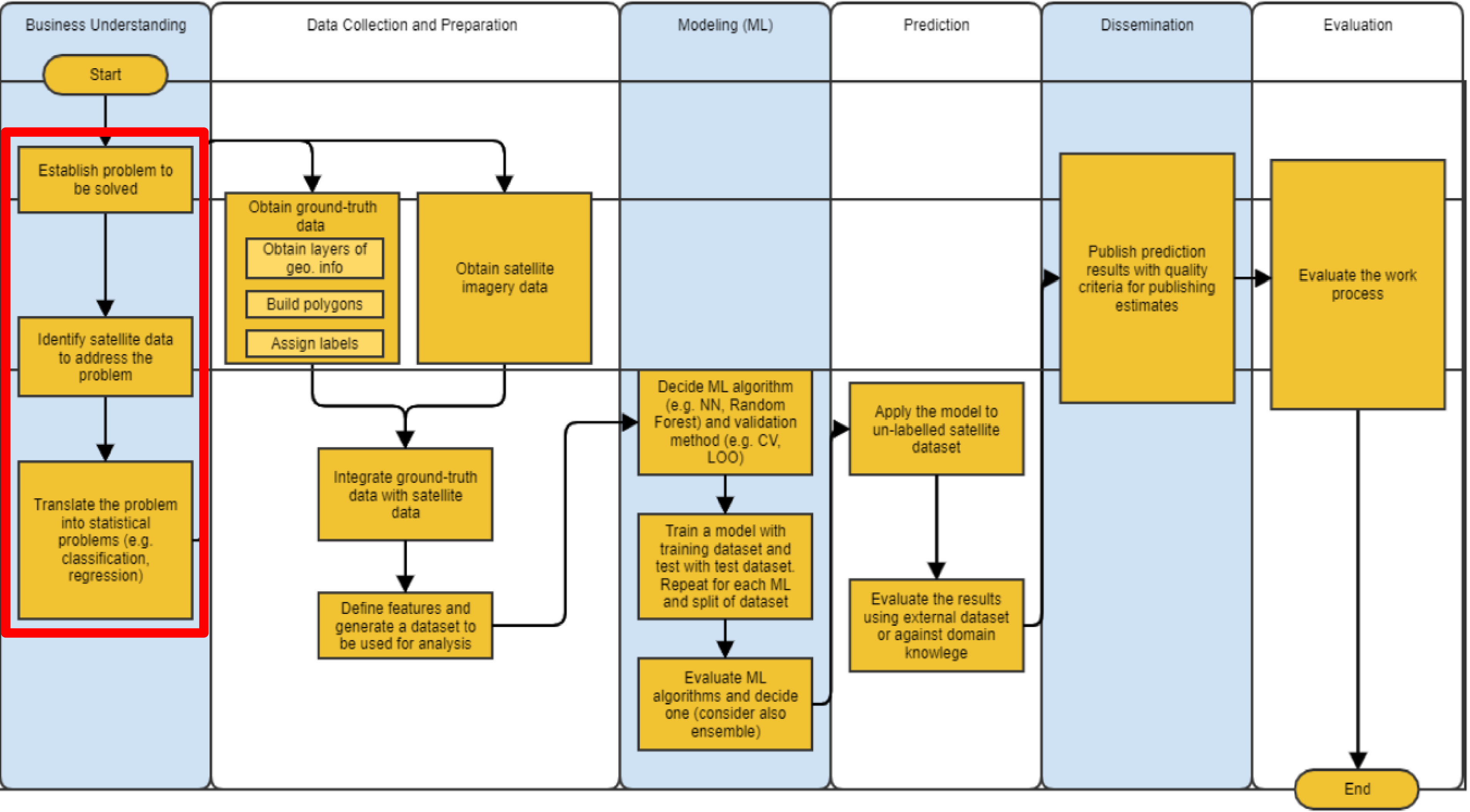
Aerial Image Address Use Classification using Machine Learning

Daniel Merkas
James Farnell
Australian Bureau of Statistics
Informing Australia's important decisions


Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Australian Bureau of Statistics

The Address Register (AR)

Uses

Census

Survey Frames

Data Linkage

AR

100k New Quarterly

15 Million Addresses

Maintenance Processes

Admin Model and Business Rules

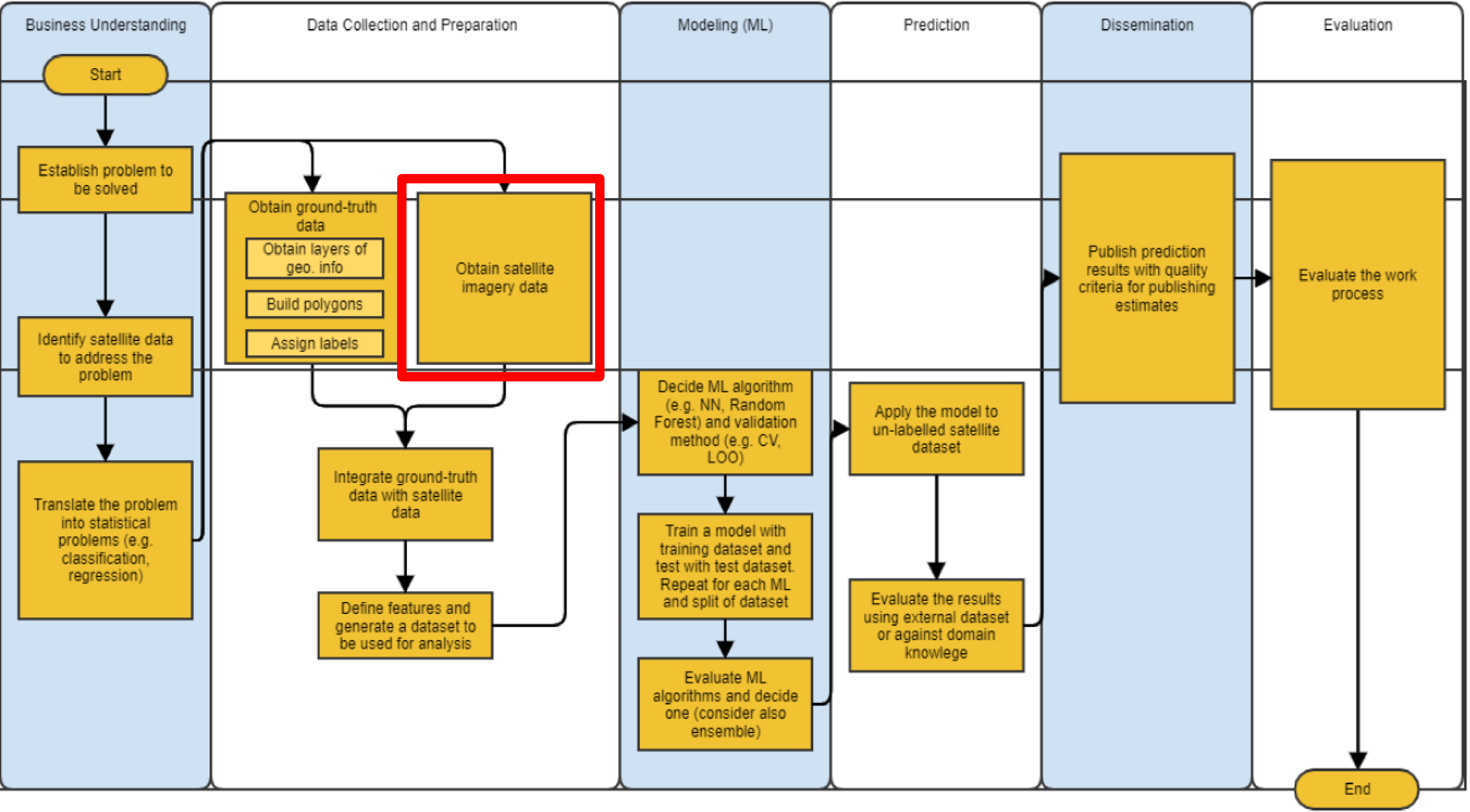
68%*

Desktop Canvassing

32%

AIR

(Automated Image Recognition)



Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Image Pipeline

Address
GANSW1047032

link

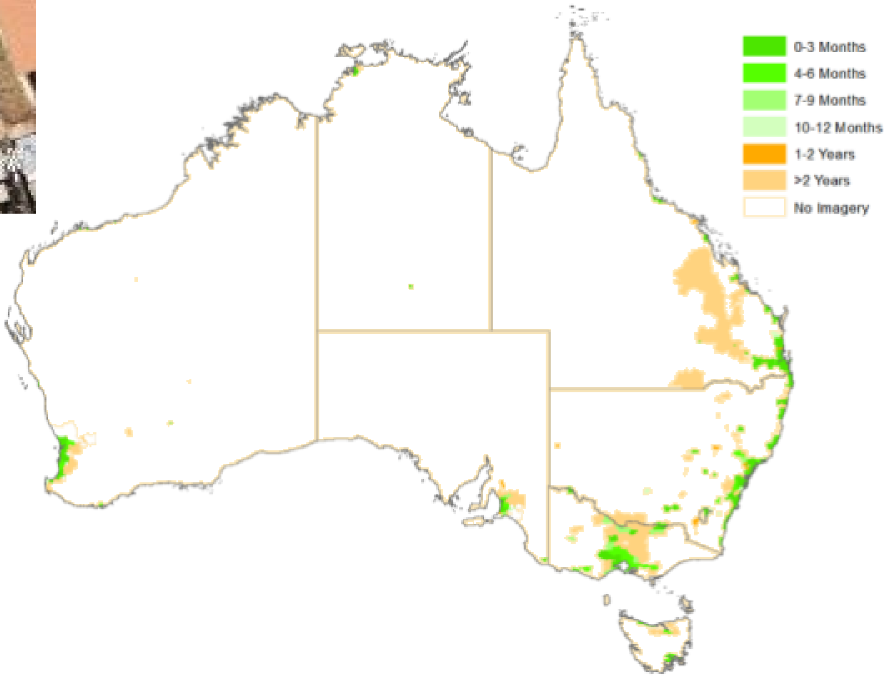
Lat/Long Point
-37.8136, 144.9631

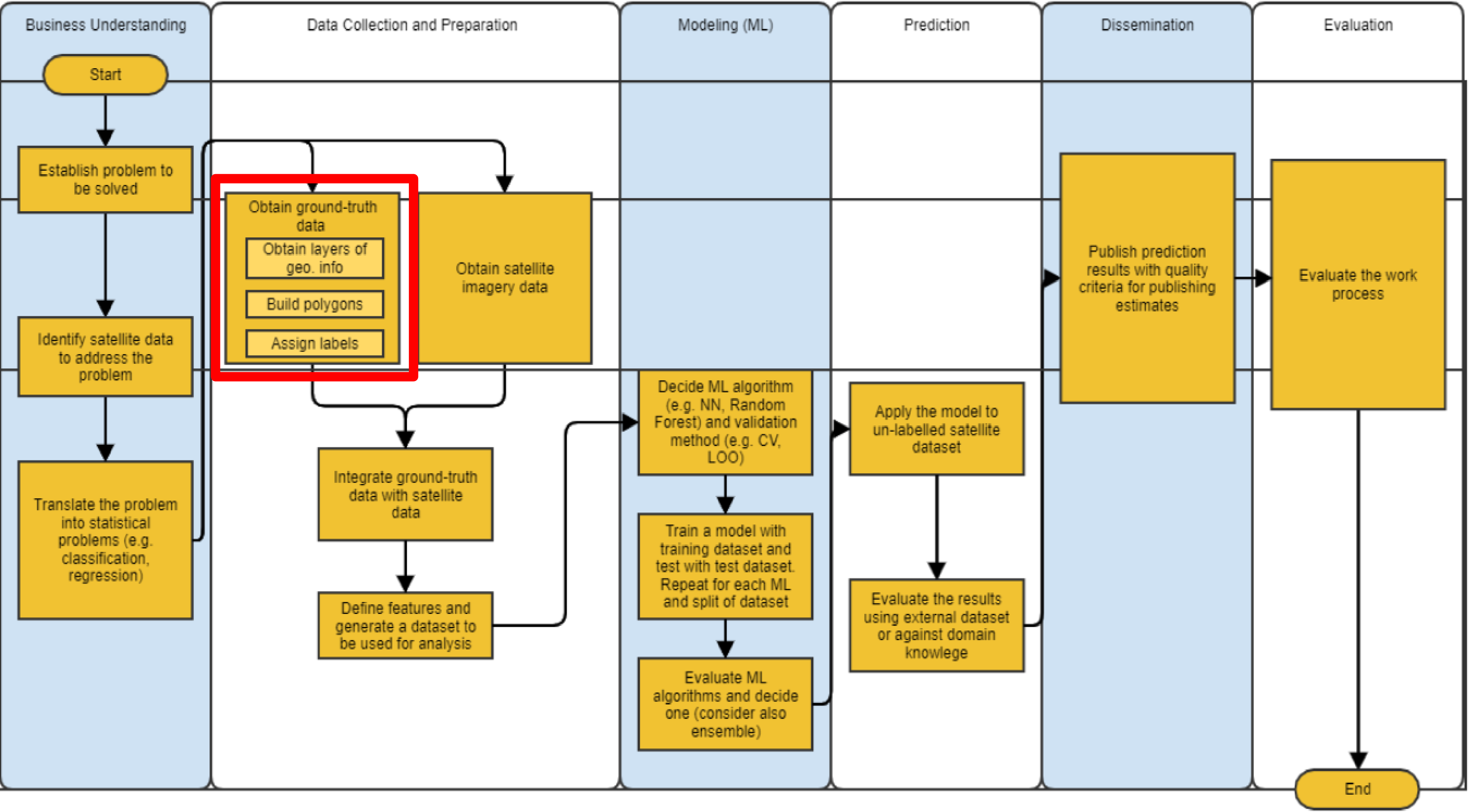
API call

Commercial Provider
-37.8136, 144.9631

- Coverage ✓
  89% of Australian population
- Recency ✓
  within 12 months
- Resolution ✓
  <23cm/pixel

download
150 x 150 pixel,
~30 x 30 meter,
lat/long center





Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Parcel Overlay

- 75% of addresses link to cadasters
- Cadaster ≈ Address boundaries
- Overlay identifies small, complex addresses
- Improves speed / accuracy of hand-classification

Utilities box

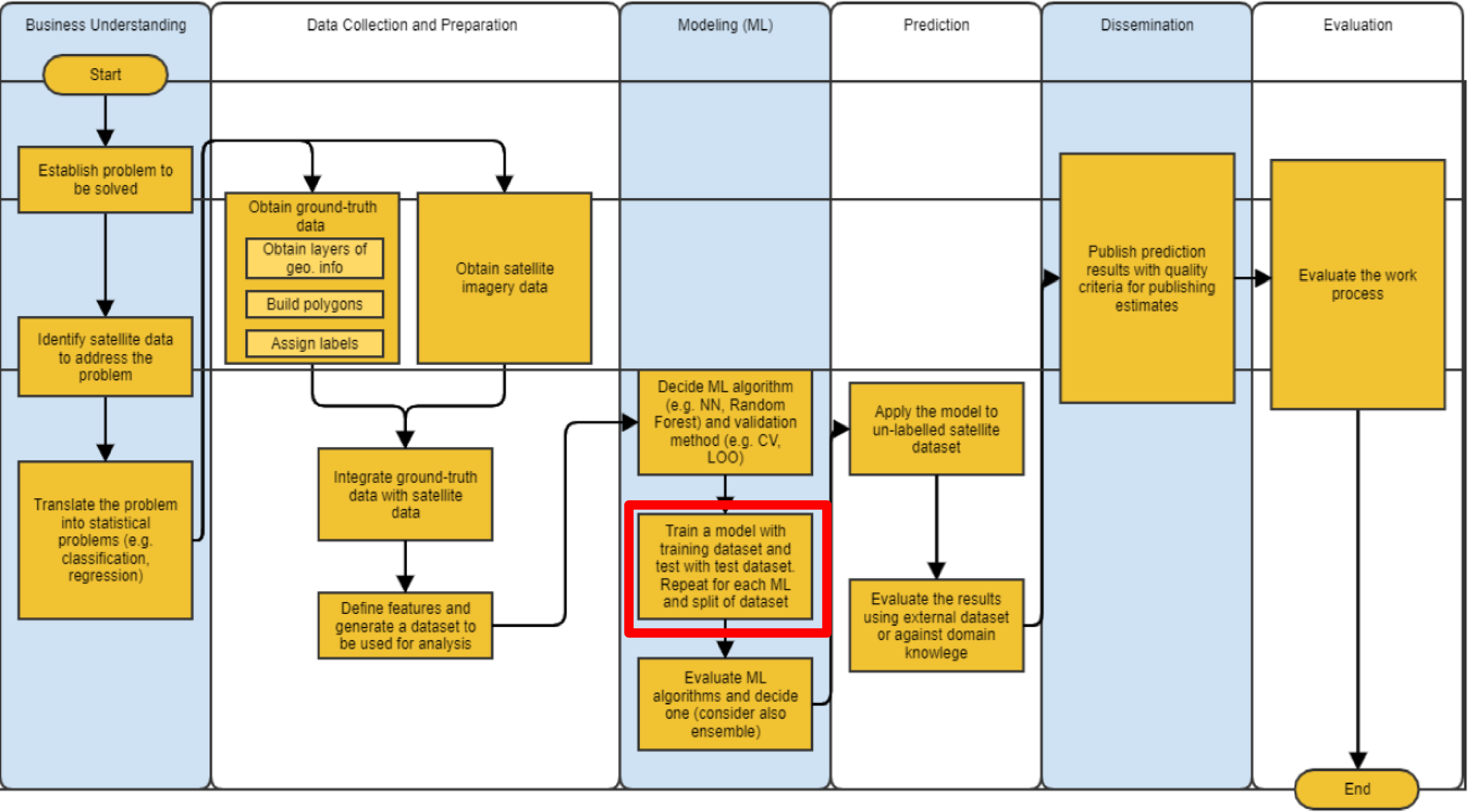
Parcel offcut

Common property

Easement

Image Categories

Private Dwelling	Vacant Property	Under Construction	High Density	Other/Commercial



Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Training Data

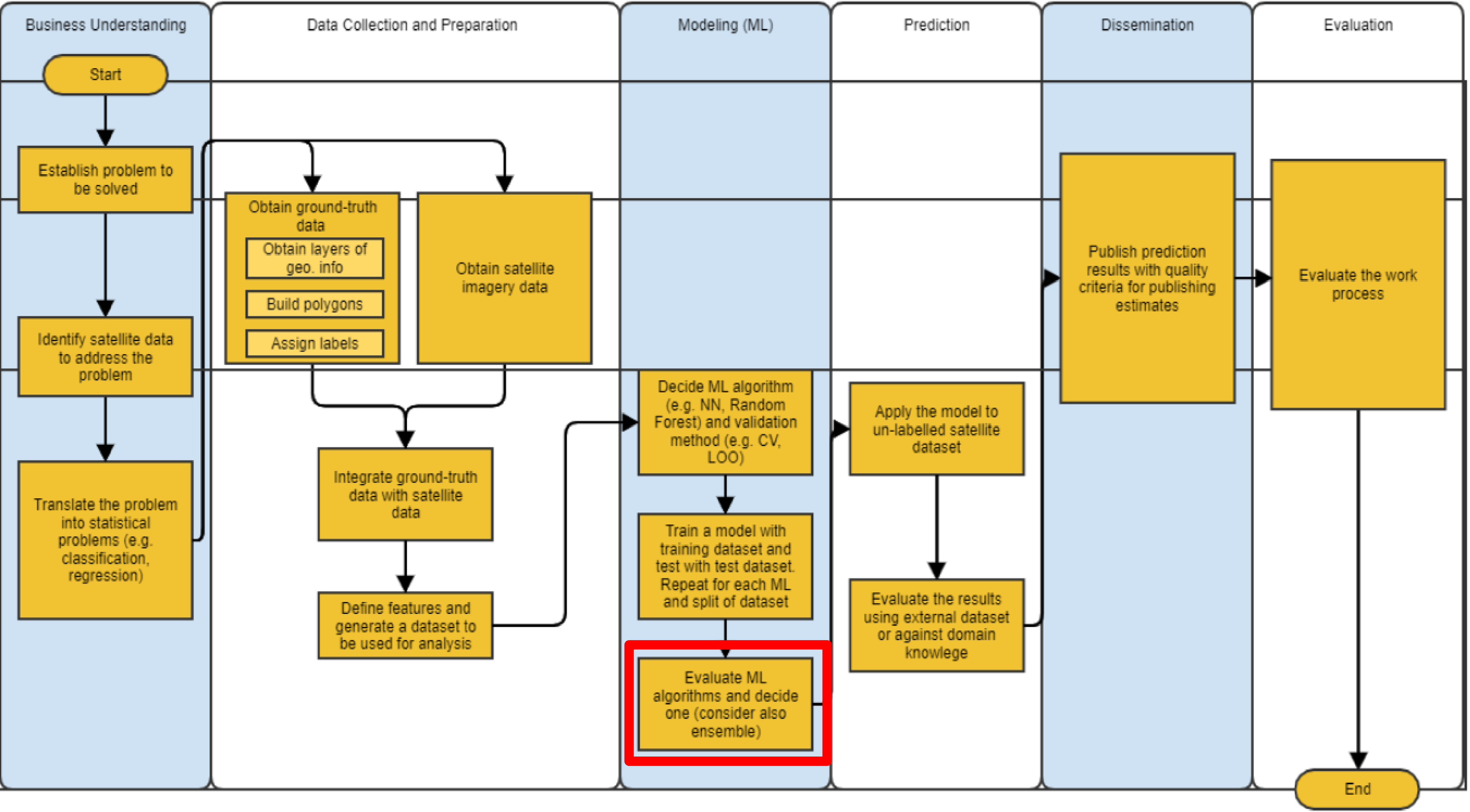
- Hand classified images split 80/20 training/validation
- Class balance ≈ real world frequency
  - Higher accuracy on larger classes
- Purposive, maximum variation sample
  - More generalizable with fewer images
- Iterative decision boundary improvement
  - Classify un-labelled images
  - Visual scan for mistakes
  - Correct classification
  - Feed into training set
  - Re-train

Class	Images
Private Dwellings	11,100
Under Construction	2,200
Vacant Property	4,400
High Density	1,500
Other/Commercial	1,700

Augmented Data

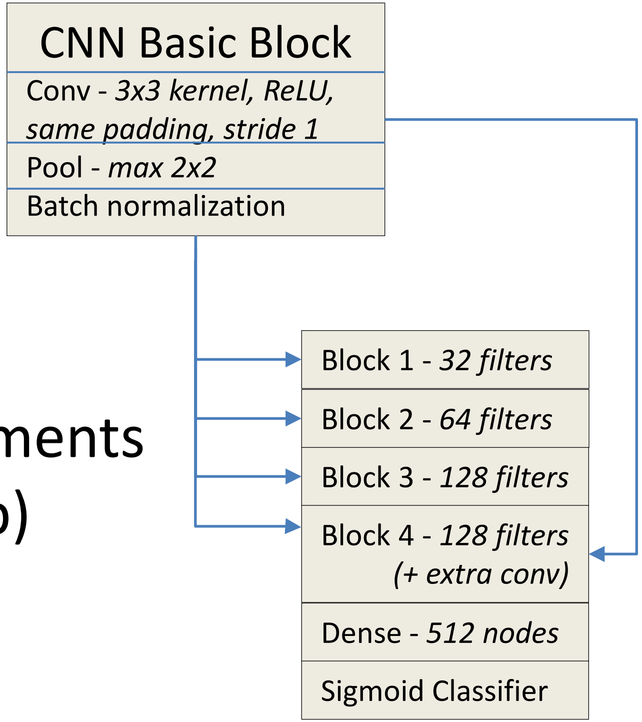
- Random transforms for each training image, each epoch
- Prevents model from memorizing training images
- Transformation limited to realistic class examples

+ Zoom / Aspect ratio shift
+ Brightness shift
+ Skew


Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

AIR Architecture

- 5 convolutional layers
  + 1 dense hidden layer
- No residual links
- R / Keras / TensorFlow
- Moderate computational requirements
  (150 epochs in 20 hours on laptop)
- Strong augmentation
  + L2/ Ridge regularization

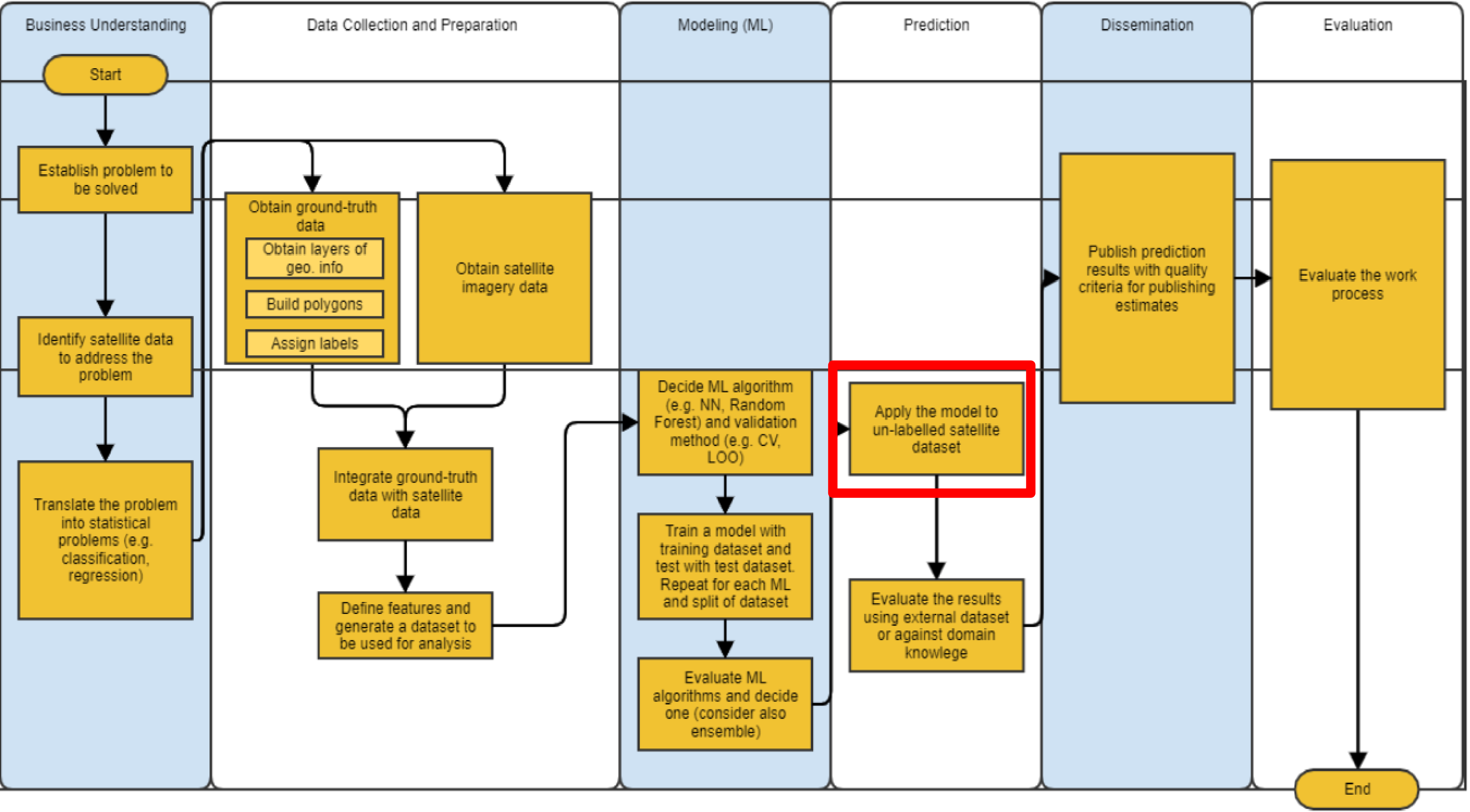


Australian Bureau of Statistics

Evaluation Set

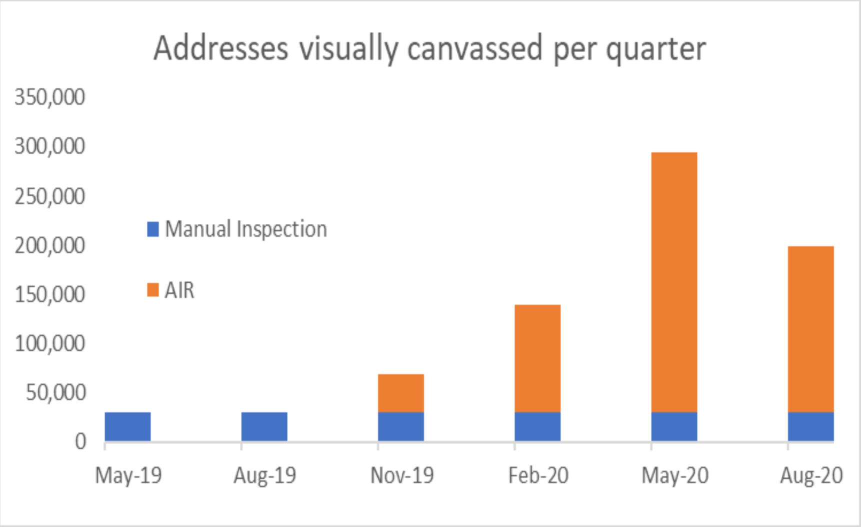
- Representative, randomly sampled, hand-classified set
- Green
  - Clear category definitions
  - Trusted without/above admin
- Red
  - Complex categories
  - Requires confirming admin /desktop canvassing

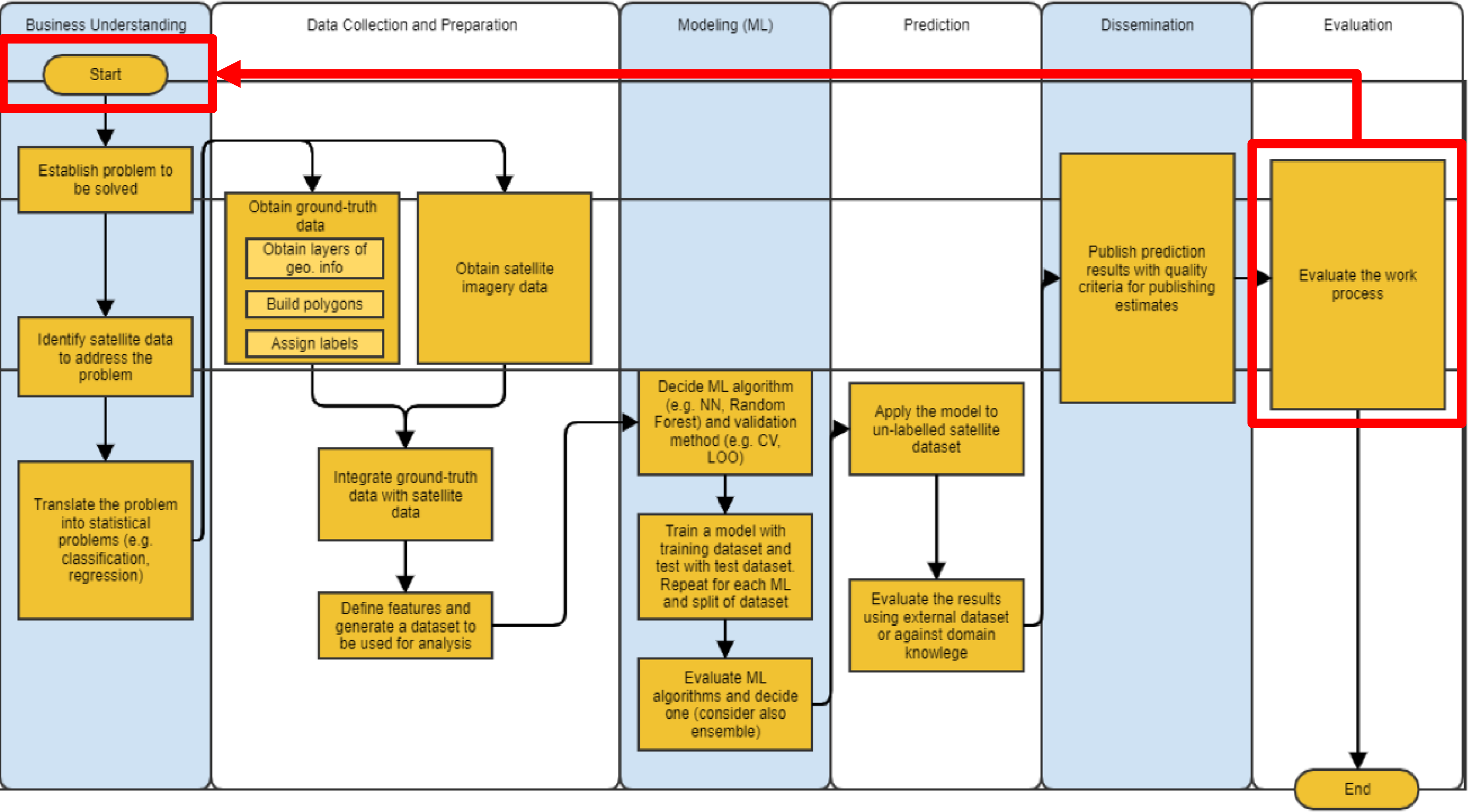
Class	Rate	F1 Score	$\frac{TPR + TNR}{2}$
Overall		**82.2%**	**81.2%**
Private Dwellings	63%	90%	83%
Under Construction	4%	63%	82%
Vacant Property	19%	77%	83%
High Density	6%	56%	74%
Other/Commercial	8%	48%	67%


Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Timeliness

- Desktop Canvassers
  - High accuracy
  - 1000 addresses per week
- AIR
  - Good accuracy for major address types
  - Deployed to:
    - Simple new addresses
    - Monitoring construction
    - Monitoring demolitions
    - Re-inspecting addresses





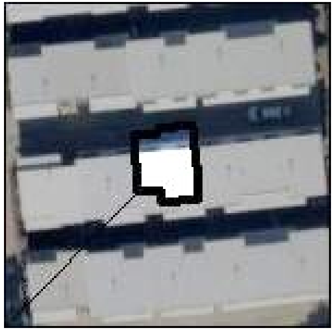
Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Model Refresh

- Parcel cutouts
- More training images
- Unbalanced classes
- Deeper ML network
- Representative evaluation set

Performance Evaluation – Parallel Run

- Parallel run on 25,000 images
- Major Changes
  - Less reliance on desktop canvassing ~15%
    eg: *Flat roofed residential buildings classified directly to Private Dwelling*
  - Coverage improvement ~2.8%
    eg: *Under Construction/Vacant to Private Dwelling*



Questions?