A Pipeline for Modeling Automated Scoring Using Python, R and Jupyter Notebooks

Nitin Madnani, Anastassia Loukina & Lei Chen

Machine Learning & Educational Assessment

A Pythonic Love Story

Nitin Madnani, Anastassia Loukina & Lei Chen

Educational Testing Service

- A non-profit educational organization founded in 1947, headquartered in Princeton, New Jersey (**N**≅3500).
- Designs and administers global as well as domestic educational assessments (GRE®, TOEFL®, PRAXIS® etc.)
- Conducts and publishes extensive research on psychometrics, statistics, cognitive science, and computer science.[1]
- **Mission**: To advance quality and equity in education by providing *fair* and *valid* assessments, research and related services.

[1] http://search.ets.org/researcher/

Two Parts

- **Part 1**: What makes educational assessment a challenging application for machine learning?

- **Part 2**: How does Python help us address some of these challenges at ETS?

Part 1

Educational Assessment

?

Machine Learning

Educational Assessments

Educational Assessments

Classroom Quiz

Educational Assessments

Classroom Quiz

Homework Assignment

Educational Assessments

Classroom Quiz

Homework Assignment

MOOC Assignments

Educational Assessments

Classroom Quiz

Homework Assignment

MOOC Assignments

K-12 Standardized Tests

Educational Assessments

Classroom Quiz

Teacher Certification

Homework Assignment

MOOC Assignments

K-12 Standardized Tests

Educational Assessments

Classroom Quiz

Practice Tests

Homework Assignment

Teacher Certification

MOOC Assignments

K-12 Standardized Tests

Educational Assessments

TOEFL/IELTS

Classroom Quiz

Practice Tests

Teacher Certification

Homework Assignment

K-12 Standardized Tests

MOOC Assignments

Educational Assessments

TOEFL/IELTS

Classroom Quiz

Practice Tests

GRE

Teacher Certification

Homework Assignment

K-12 Standardized Tests

MOOC Assignments

Educational Assessments

TOEFL/IELTS

Classroom Quiz

Practice Tests

GRE

Teacher Certification

Homework Assignment

GMAT

K-12 Standardized Tests

MOOC Assignments

Educational Assessments

TOEFL/IELTS

Classroom Quiz

Practice Tests

GRE

Teacher Certification

Homework Assignment

GMAT

K-12 Standardized Tests

GED

MOOC Assignments

Educational Assessments

Classroom Quiz

Homework Assignment

MOOC Assignments

Practice Tests

TOEFL/IELTS

GRE

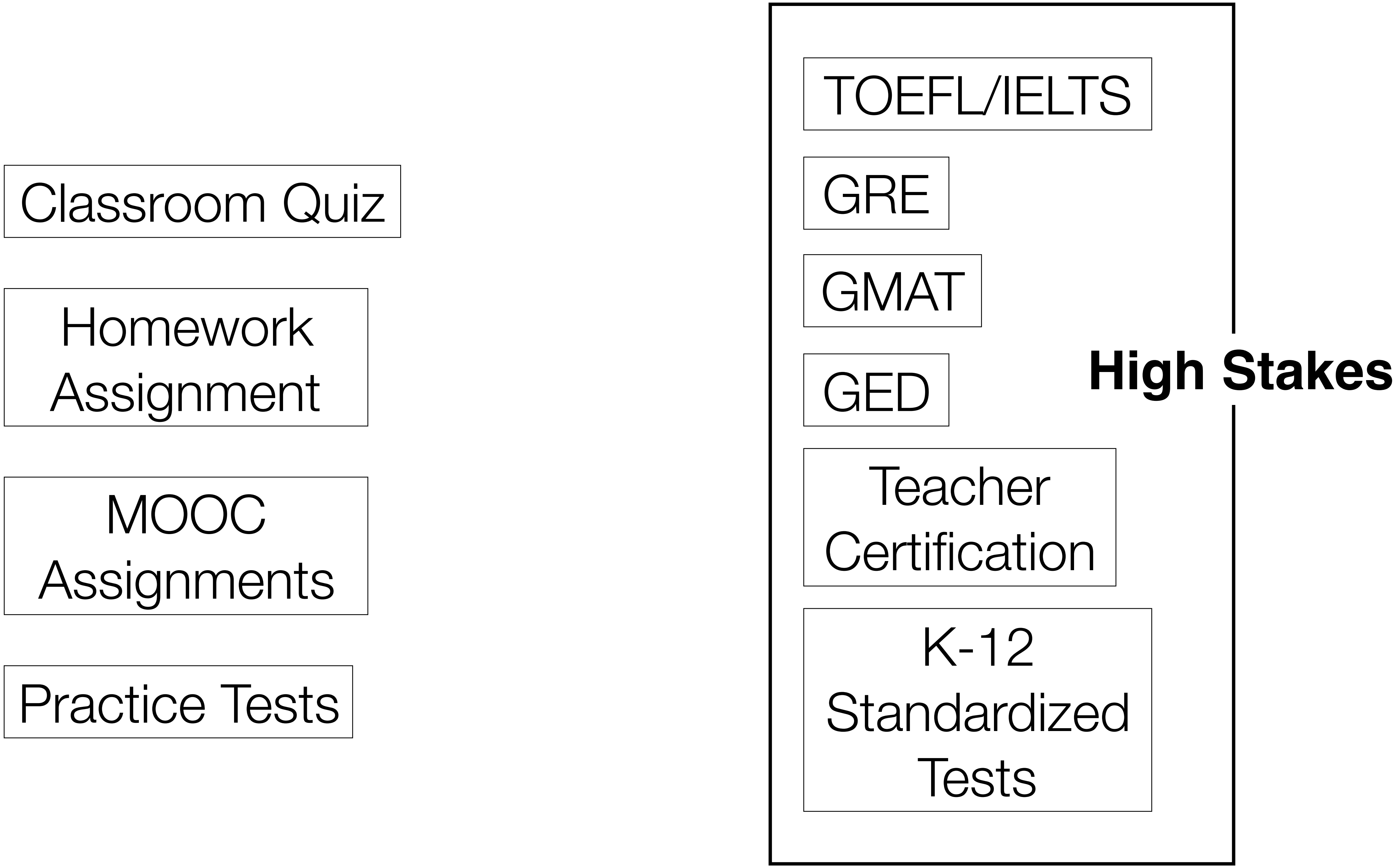
GMAT

GED

Teacher Certification

K-12 Standardized Tests

Educational Assessments



Classroom Quiz

Homework Assignment

MOOC Assignments

Practice Tests

TOEFL/IELTS

GRE

GMAT

GED

Teacher Certification

K-12 Standardized Tests

High Stakes

“High Stakes”

- A test with results that have **important**, **direct** consequences for the test-takers.

- A test-taker would want to **understand** what their score means and how it maps to what they did on the test.

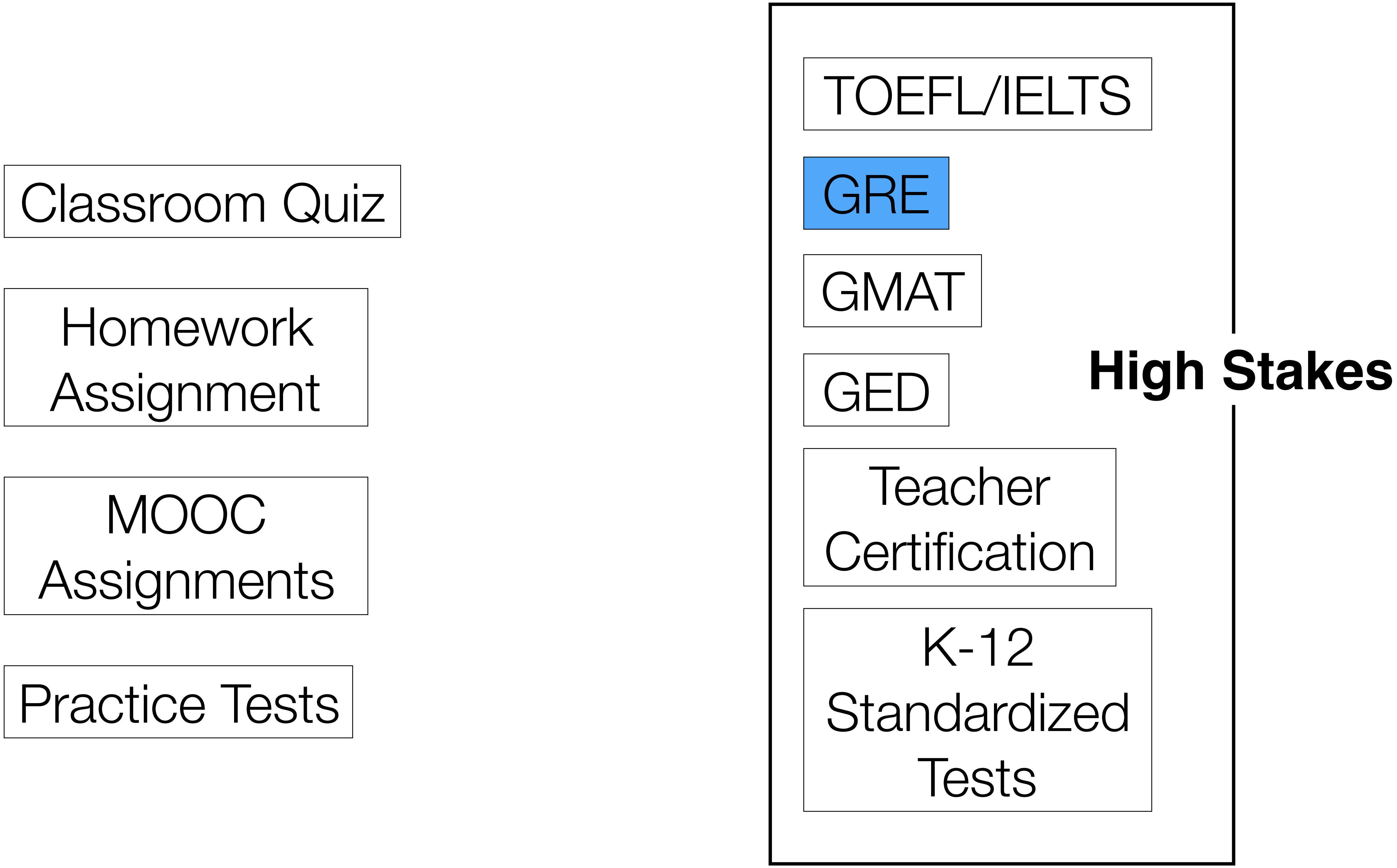
Educational Assessments

- Classroom Quiz
- Homework Assignment
- MOOC Assignments
- Practice Tests

High Stakes

- TOEFL/IELTS
- GRE
- GMAT
- GED
- Teacher Certification
- K-12 Standardized Tests

Educational Assessments



Classroom Quiz

Homework Assignment

MOOC Assignments

Practice Tests

TOEFL/IELTS

GRE

GMAT

GED

Teacher Certification

K-12 Standardized Tests

High Stakes

The GRE

- Graduate Record Examination, designed and administered by ETS.
- Used by at least 3000 colleges and universities across the world for graduate school applications to MS, MBA & PhD programs.[1]
- ~575,000 test-takers from ~200 countries between July 2013 and June 2014 (50% women, 45% men). [2]
- Three sections:
  - Verbal Reasoning
  - Quantitative Reasoning
  - Analytical Writing

[1] https://www.ets.org/s/gre/pdf/gre_aidi_fellowships.pdf

[2] http://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf

The GRE

- Graduate Record Examination, designed and administered by ETS.
- Used by at least 3000 colleges and universities across the world for graduate school applications to MS, MBA & PhD programs.[1]
- ~575,000 test-takers from ~200 countries between July 2013 and June 2014 (50% women, 45% men). [2]
- Three sections:
  - Verbal Reasoning
  - Quantitative Reasoning
  - Analytical Writing

[1] https://www.ets.org/s/gre/pdf/gre_aidi_fellowships.pdf

[2] http://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf

GRE Analytical Writing

“As people rely more and more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate.”

Directions: Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take.

Score 6. Outstanding

- articulates a clear and insightful position
- develops the position fully
- well-focused, well-organized analysis
- conveys ideas fluently and precisely
- demonstrates superior facility with English

...

Score 1. Fundamentally Deficient

- provides little/no evidence of understanding
- disorganized or extremely brief
- severe problems with sentence structure
- pervasive errors in grammar
- incoherent and meaning not clear

`https://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/scoring_guide`

Scoring essays

Given the stakes, our scoring methodology must maximize:

- **Accuracy**: how accurately does the assigned score measure the analytical skills of the test-taker?
- **Interpretability**: how easily can test-takers understand why they was assigned a particular score and what that score means?

Scoring essays

Given the stakes, our scoring methodology must maximize:

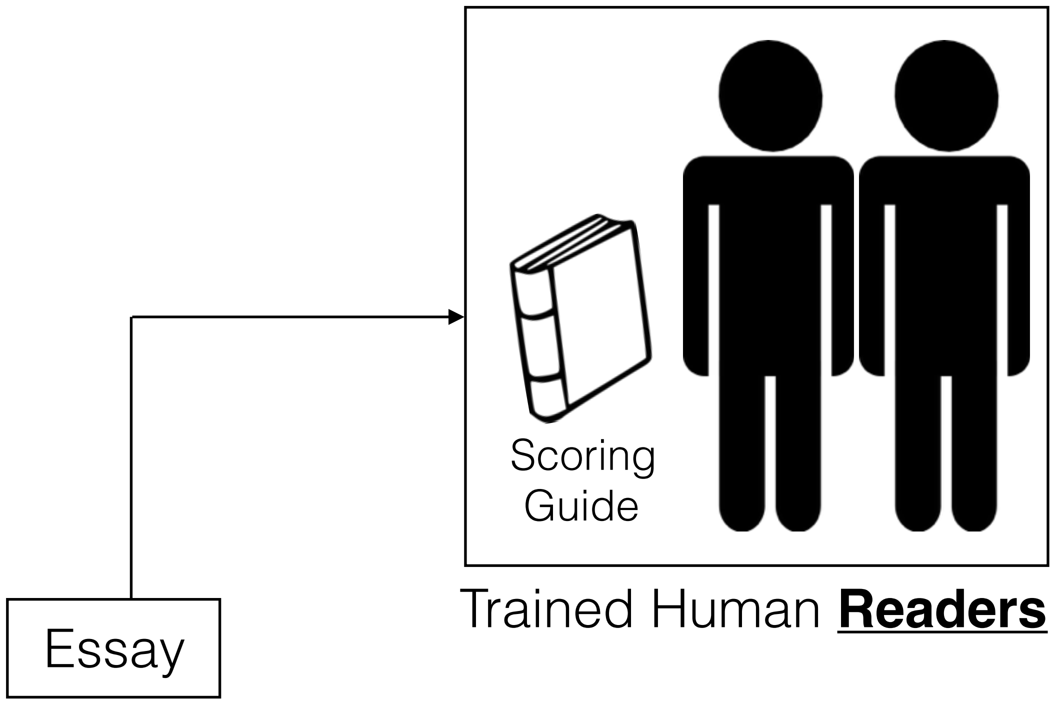
- **Accuracy**: how accurately does the assigned score measure the analytical skills of the test-taker?
- **Interpretability**: how easily can test-takers understand why they was assigned a particular score and what that score means?

It would also be nice to minimize:

- **Cost**: how efficiently can we score each test (how much money can we save the test-taker in fees)?

Scoring essays

Scoring essays

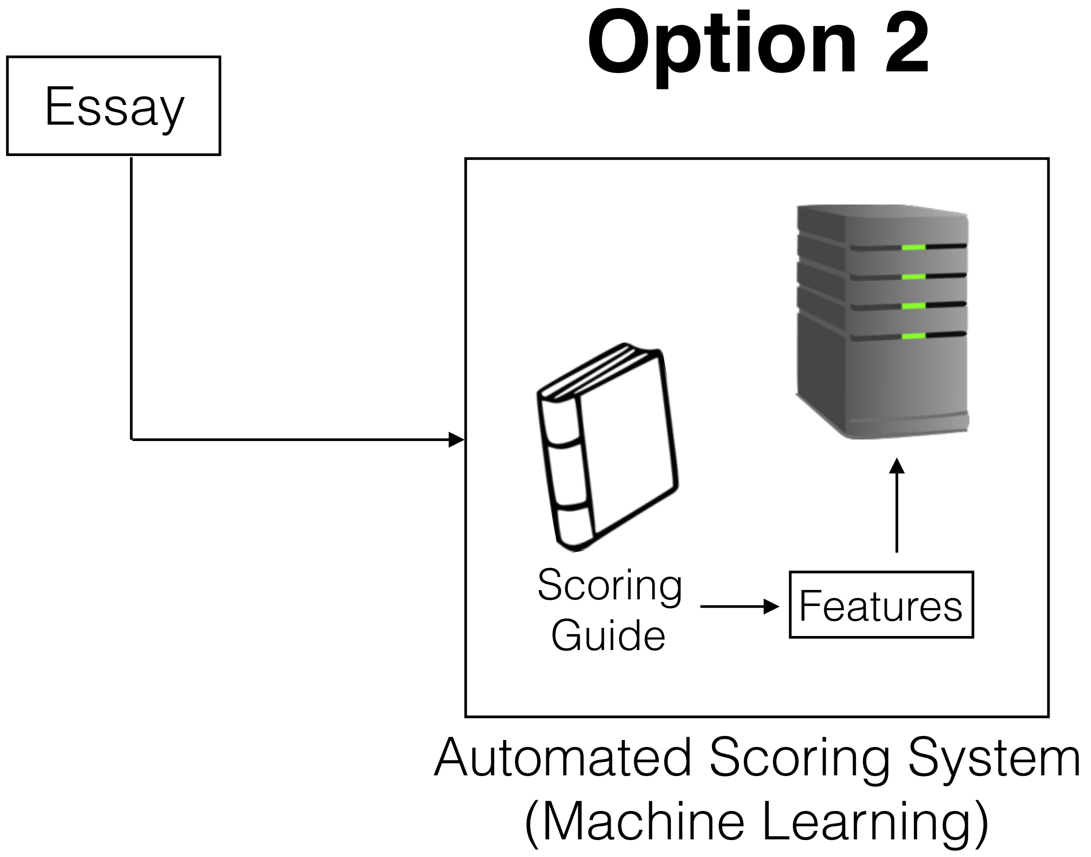


Trained Human **Readers**

Option 1

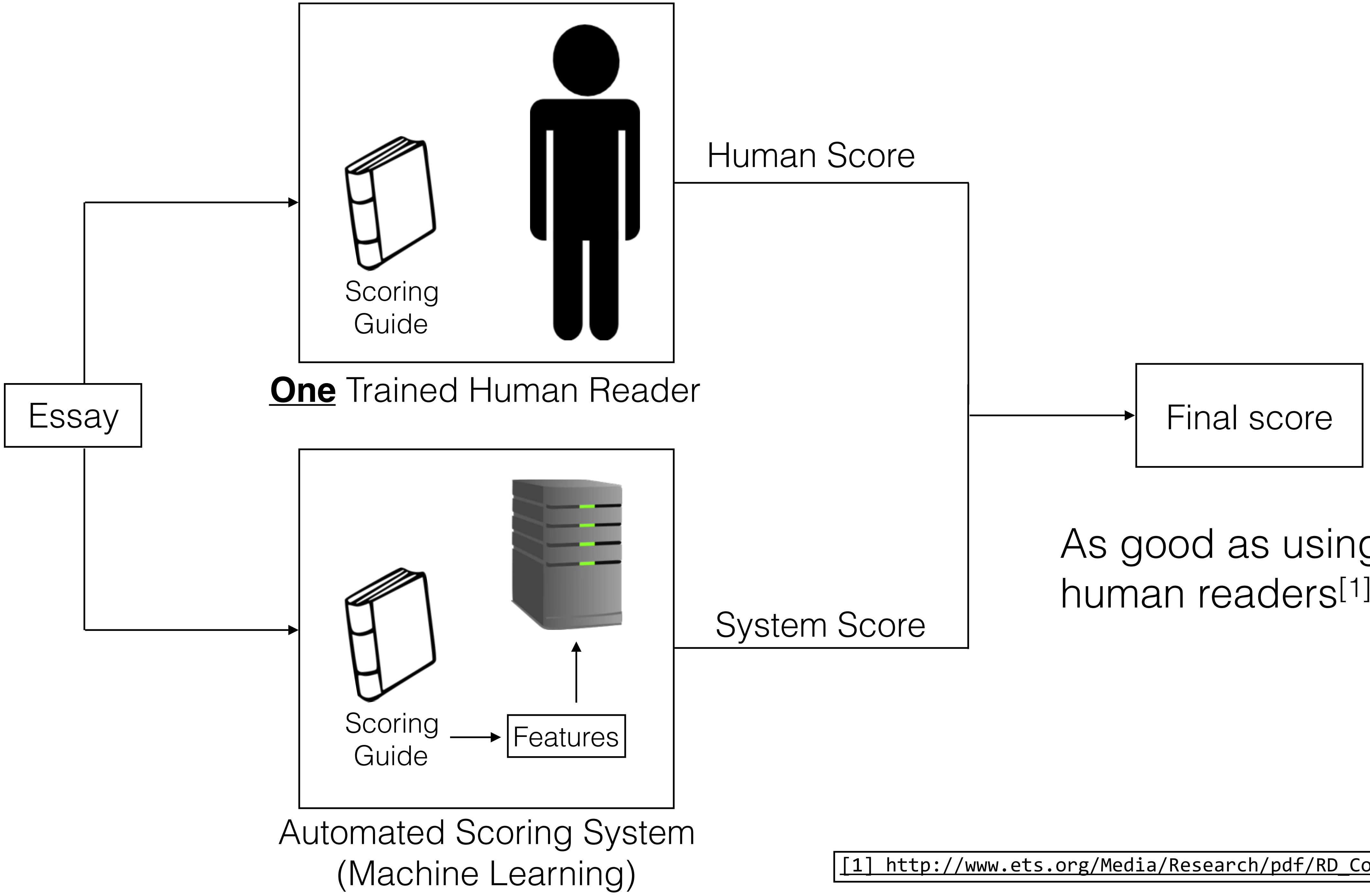
- High Accuracy
- Medium Interpretability
- High Cost

Scoring essays



- Medium Accuracy
- (Choice of) High Interpretability
- Low Cost

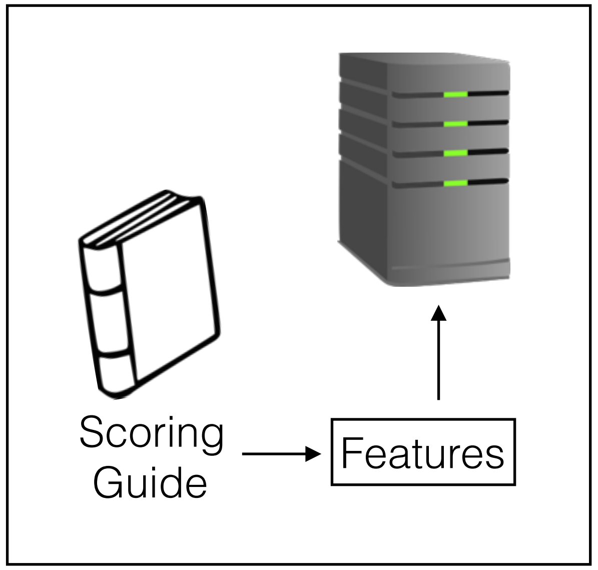
Scoring essays



As good as using two human readers[1].

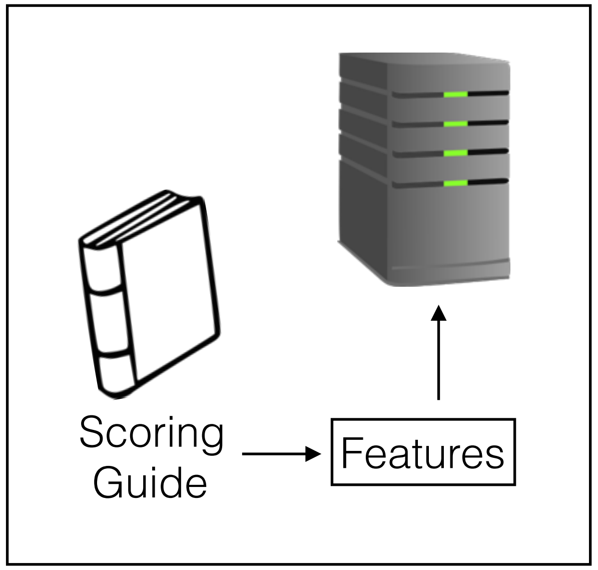
[1] http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf

E-rater



Automated Scoring System
(Machine Learning)

E-rater

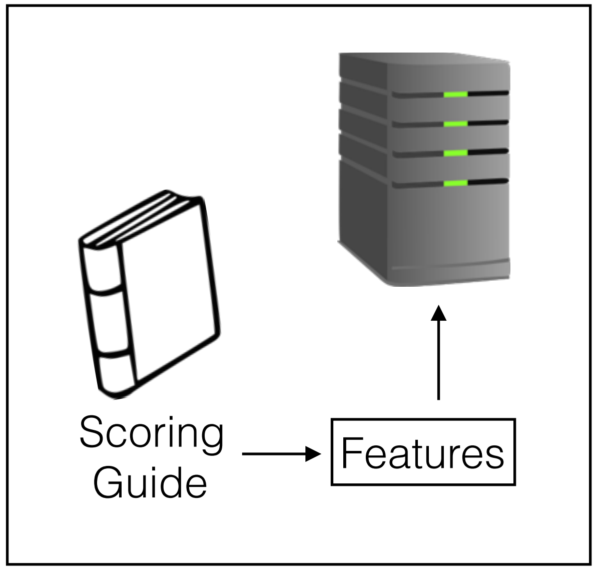


Automated Scoring System
(Machine Learning)

- “Essay Rater”
- Linear regression trained on older essays written to the same topic and scored by human readers.
- Features
  - **errors in grammar** (e.g., *subject-verb agreement*)
  - **usage errors** (*incorrect prepositions/articles*)
  - **mechanics errors** (*capitalization*, *spelling*)
  - **errors in style** (*repetitious word use*)
  - **discourse structure** (*presence of a thesis statement, main points*)
  - **vocabulary sophistication**
  - **essay organization**

Automated Essay Scoring With e-rater® V.2, The Journal of Technology, Learning, and Assessment, Volume 4(3), 2006

E-rater & Research



Automated Scoring System (Machine Learning)

- E-rater still an active area of research at ETS
  - Design new features; examine their effect on performance, and whether they overlap with existing features.
  - Try more sophisticated machine learning models (higher accuracy worth lower interpretability?)
- Last year, 10 new e-rater features proposed just for GRE!
- GRE one of a dozen assessments, e-rater one of many automated scoring engines
- Research untenable for a large group (>15 scientists) without a standardized pipeline.

Ideal research pipeline

Need an end-to-end machine learning pipeline that can:

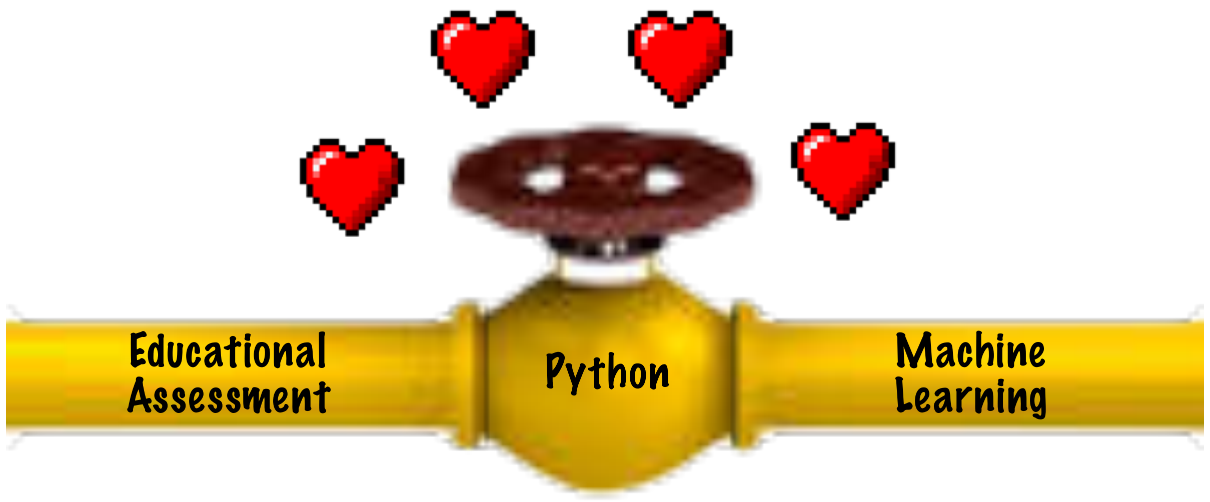
- Work on (almost) all platforms,
- Read features in any tabular format and clean it up,
- Efficiently apply filtering, scaling and transformations,
- Train any specified model with those features, and
- Generate a standardized, detailed report of performance on unseen essays.

Educational Assessment

?

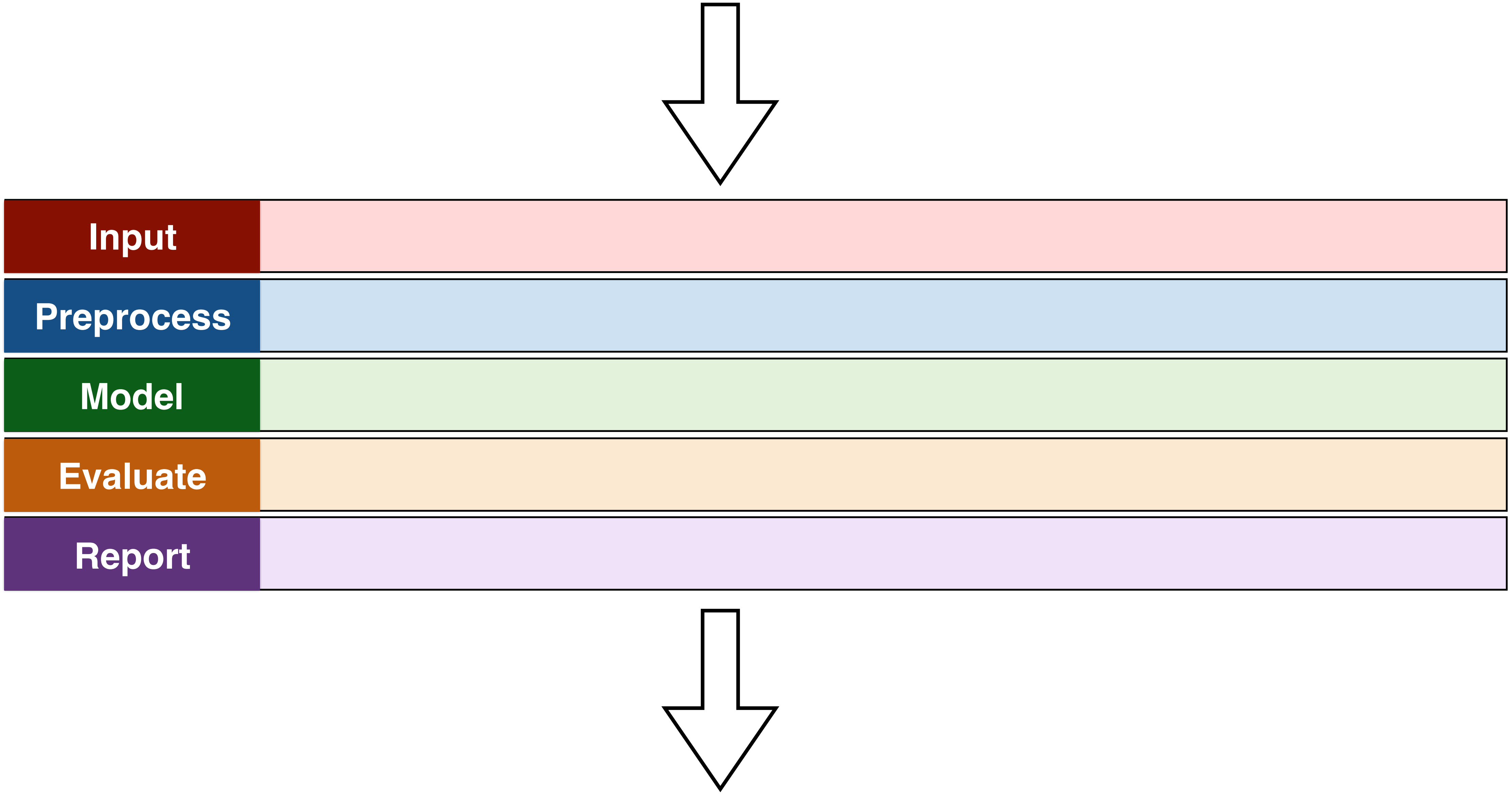
Machine Learning

Part 2



Python Pipeline

Input



Input

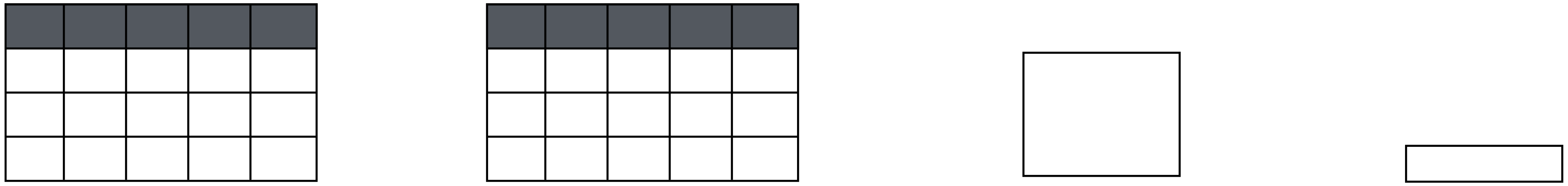
Preprocess

Model

Evaluate

Report

final self-contained report

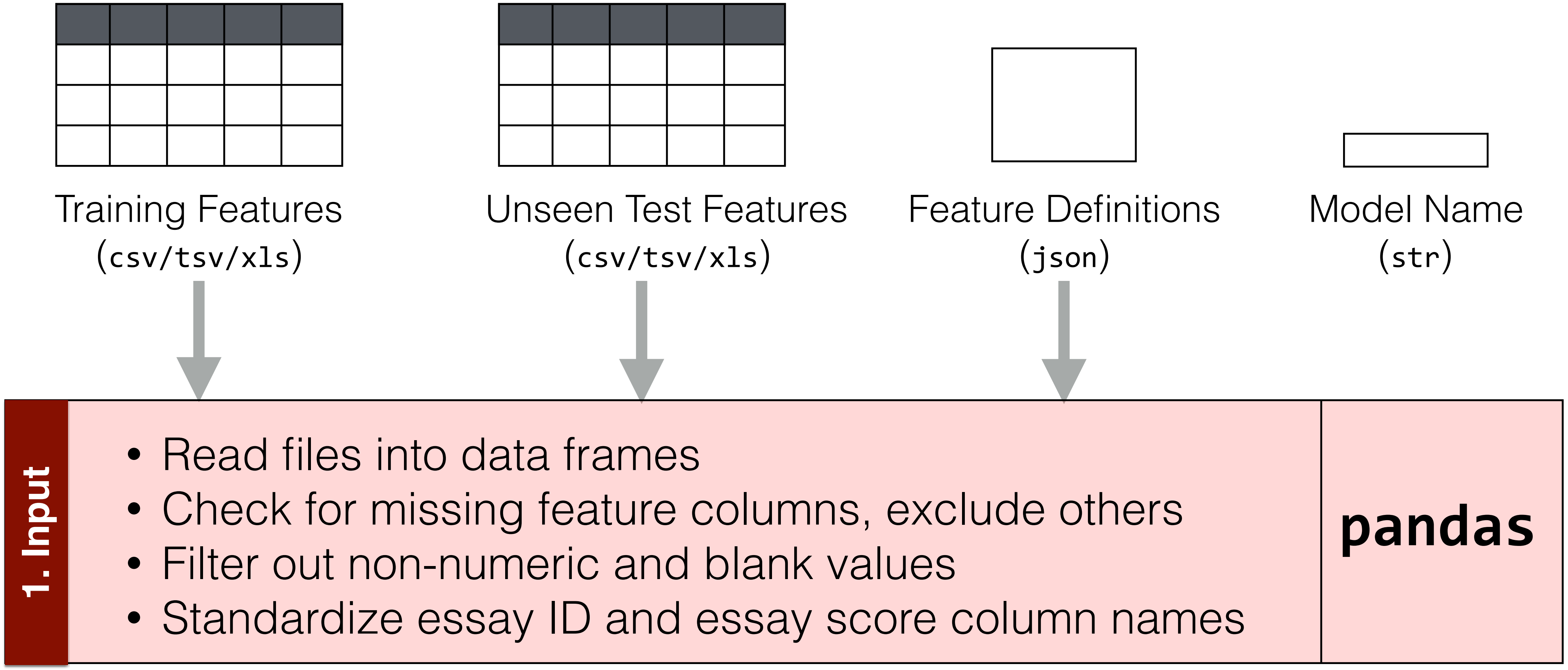


Training Features
(`csv/tsv/xls`)

Unseen Test Features
(`csv/tsv/xls`)

Feature Definitions
(`json`)

Model Name
(`str`)



Training Features
(csv/tsv/xls)

Unseen Test Features
(csv/tsv/xls)

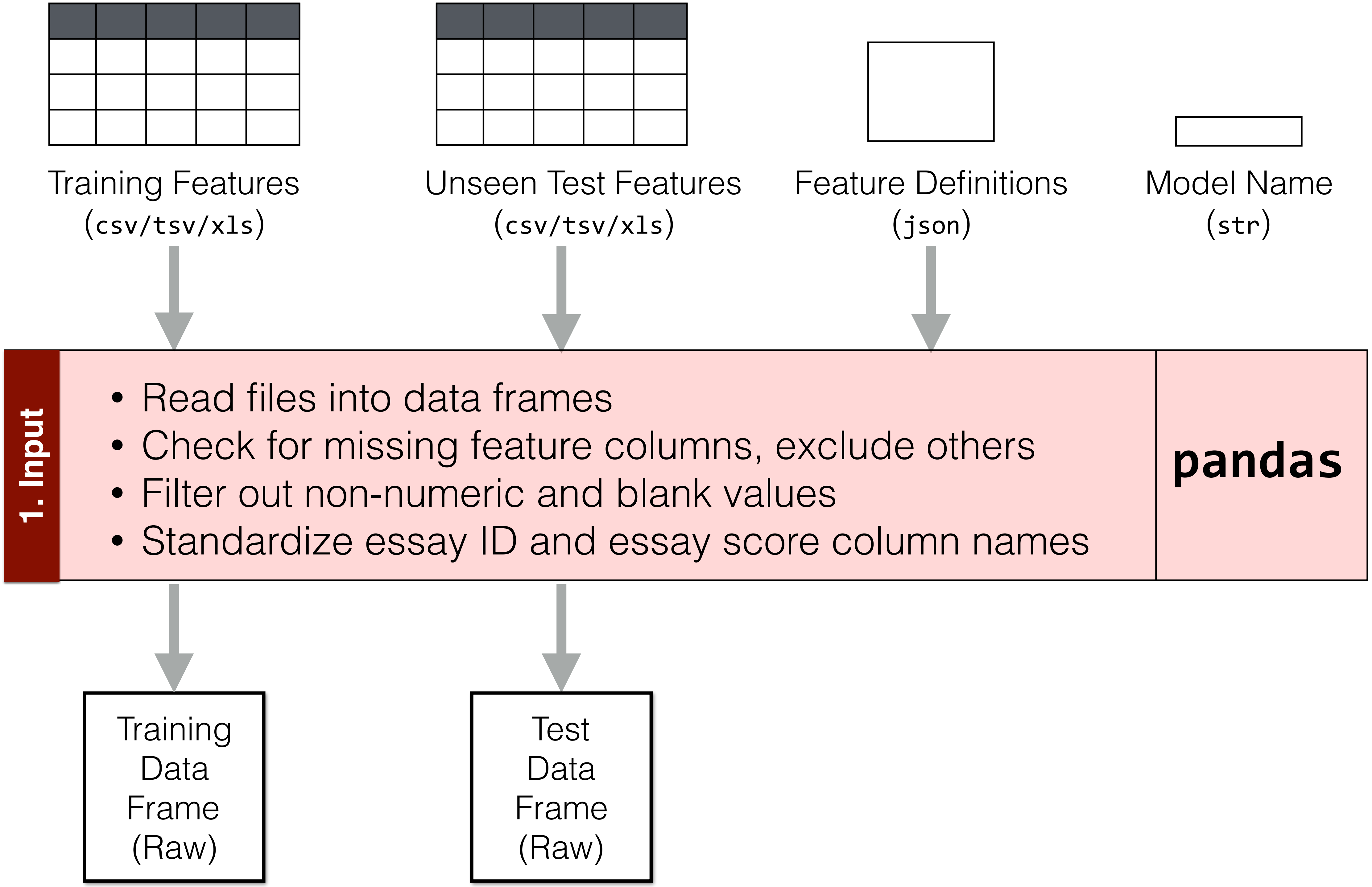
Feature Definitions
(json)

Model Name
(str)

1. Input

- Read files into data frames
- Check for missing feature columns, exclude others
- Filter out non-numeric and blank values
- Standardize essay ID and essay score column names

pandas



Training Features (csv/tsv/xls)

Unseen Test Features (csv/tsv/xls)

Feature Definitions (json)

Model Name (str)

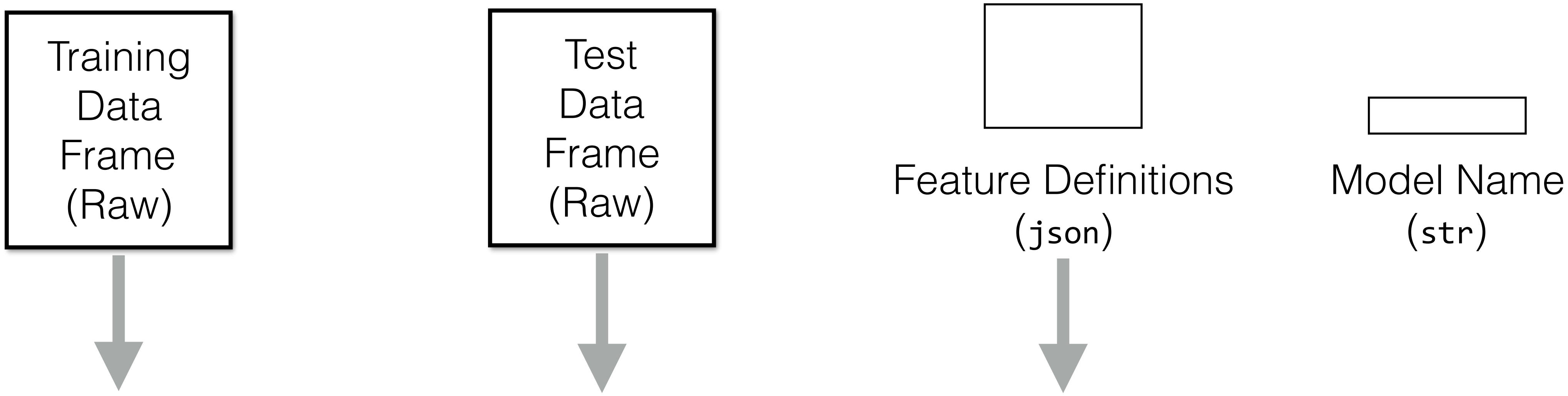
1. Input

- Read files into data frames
- Check for missing feature columns, exclude others
- Filter out non-numeric and blank values
- Standardize essay ID and essay score column names

pandas

Training Data Frame (Raw)

Test Data Frame (Raw)

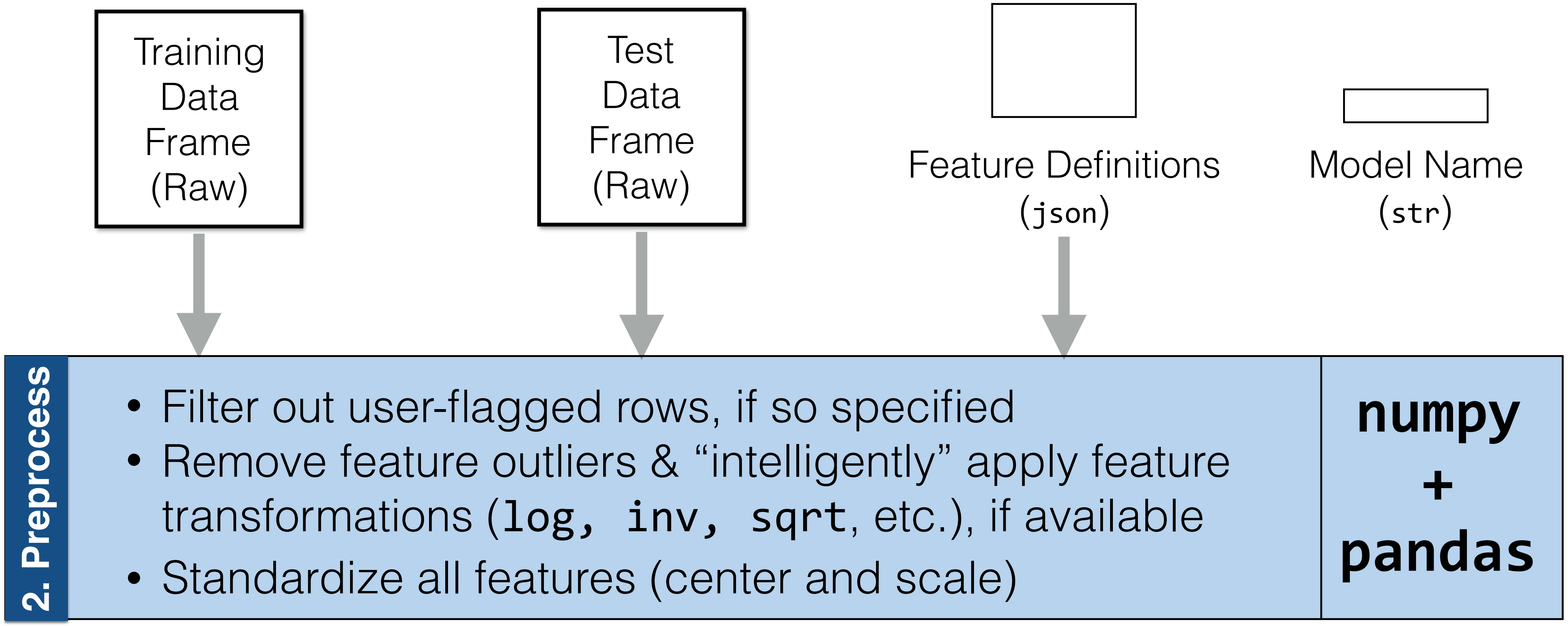


Training Data Frame (Raw)

Test Data Frame (Raw)

Feature Definitions (`json`)

Model Name (`str`)



Training Data Frame (Raw)

Test Data Frame (Raw)

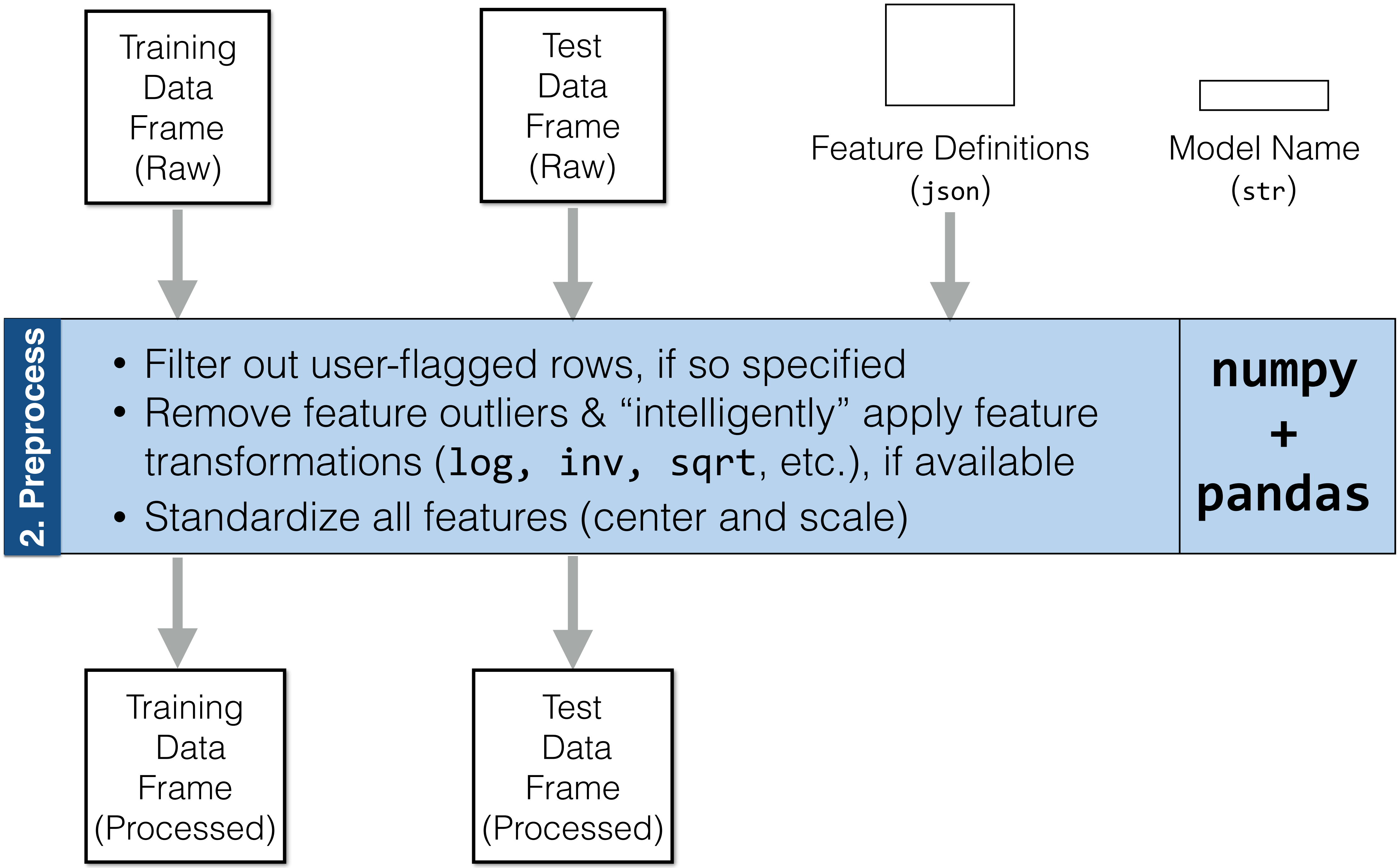
Feature Definitions (`json`)

Model Name (`str`)

2. Preprocess

- Filter out user-flagged rows, if so specified
- Remove feature outliers & “intelligently” apply feature transformations (`log`, `inv`, `sqrt`, etc.), if available
- Standardize all features (center and scale)

numpy + pandas



Training Data Frame (Raw)

Test Data Frame (Raw)

Feature Definitions (`json`)

Model Name (`str`)

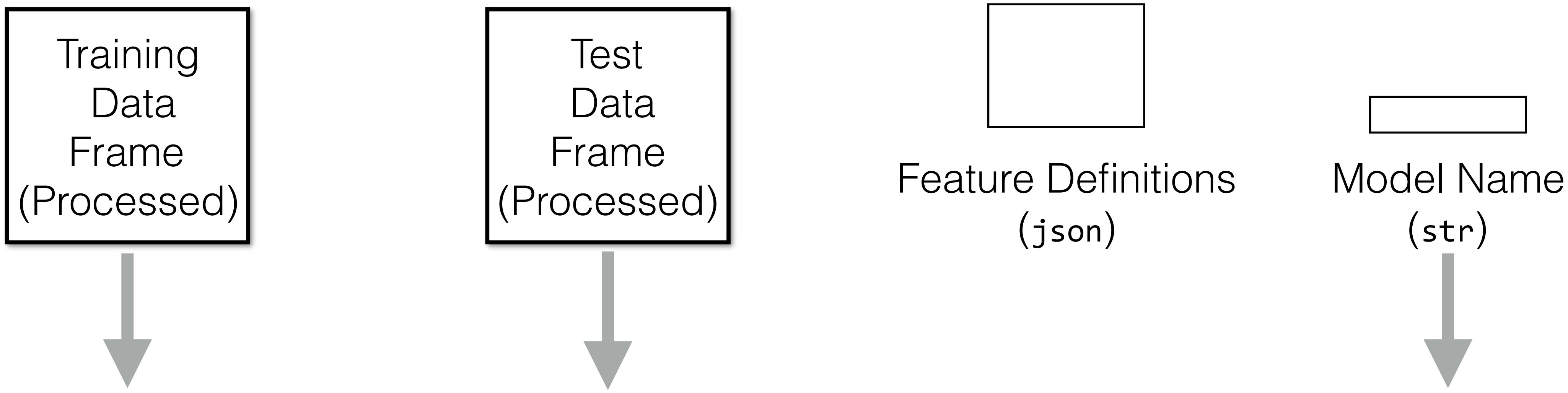
2. Preprocess

- Filter out user-flagged rows, if so specified
- Remove feature outliers & “intelligently” apply feature transformations (`log, inv, sqrt`, etc.), if available
- Standardize all features (center and scale)

numpy + pandas

Training Data Frame (Processed)

Test Data Frame (Processed)

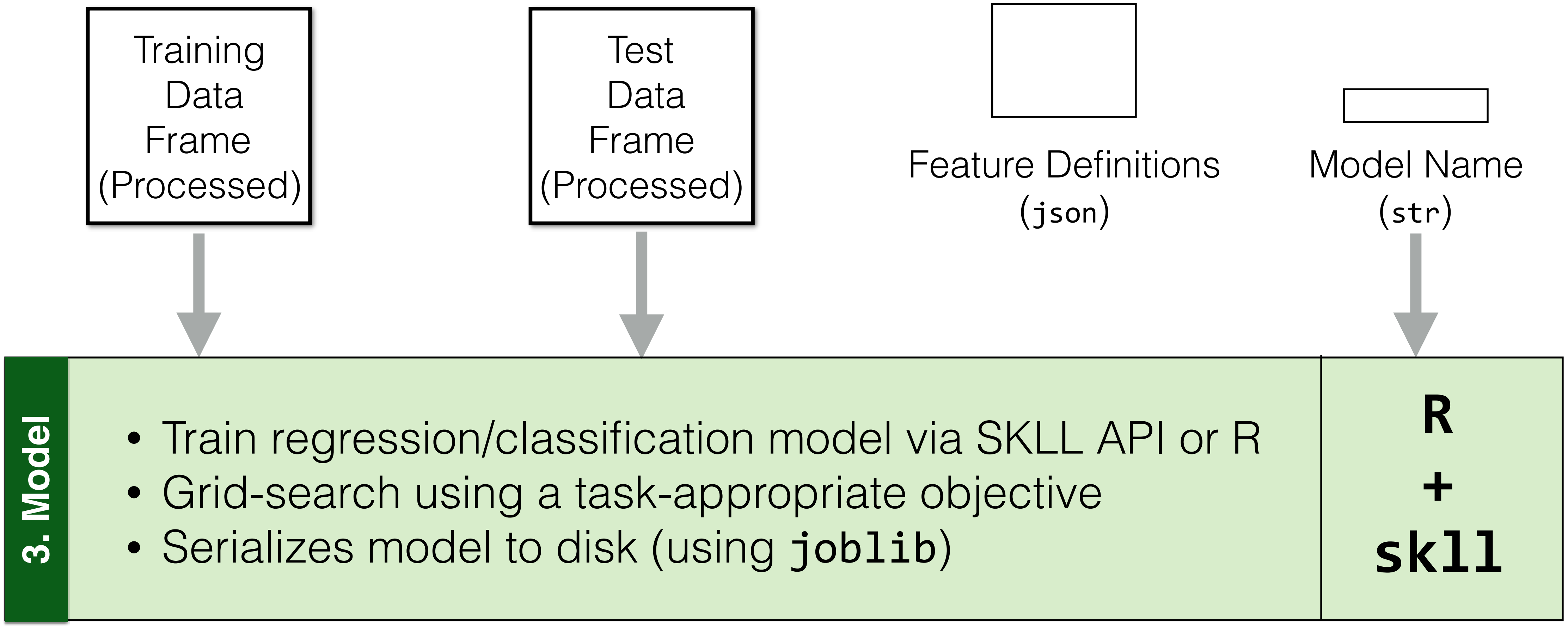


Training Data Frame (Processed)

Test Data Frame (Processed)

Feature Definitions (`json`)

Model Name (`str`)



Training Data Frame (Processed)

Test Data Frame (Processed)

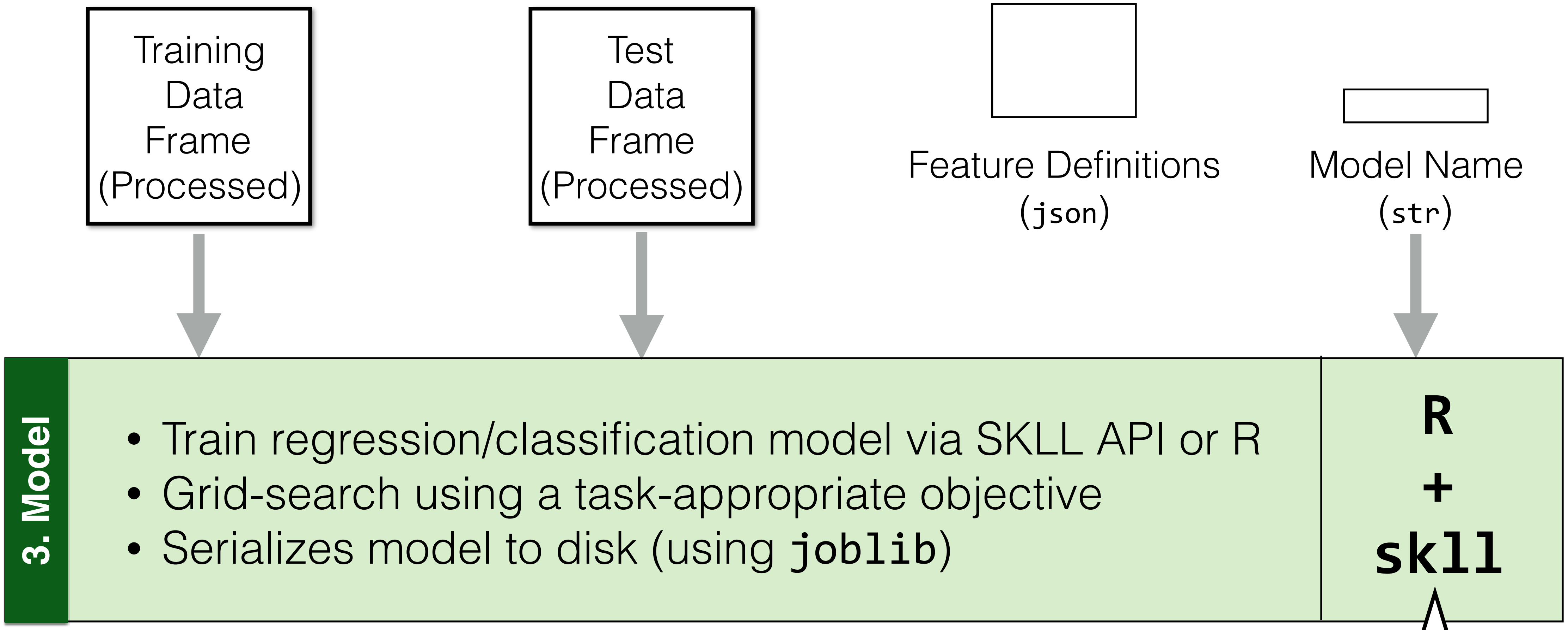
Feature Definitions (`json`)

Model Name (`str`)

3. Model

- Train regression/classification model via SKLL API or R
- Grid-search using a task-appropriate objective
- Serializes model to disk (using **`joblib`**)

R
+
`skll`



Training Data Frame (Processed)

Test Data Frame (Processed)

Feature Definitions (`json`)

Model Name (`str`)

3. Model

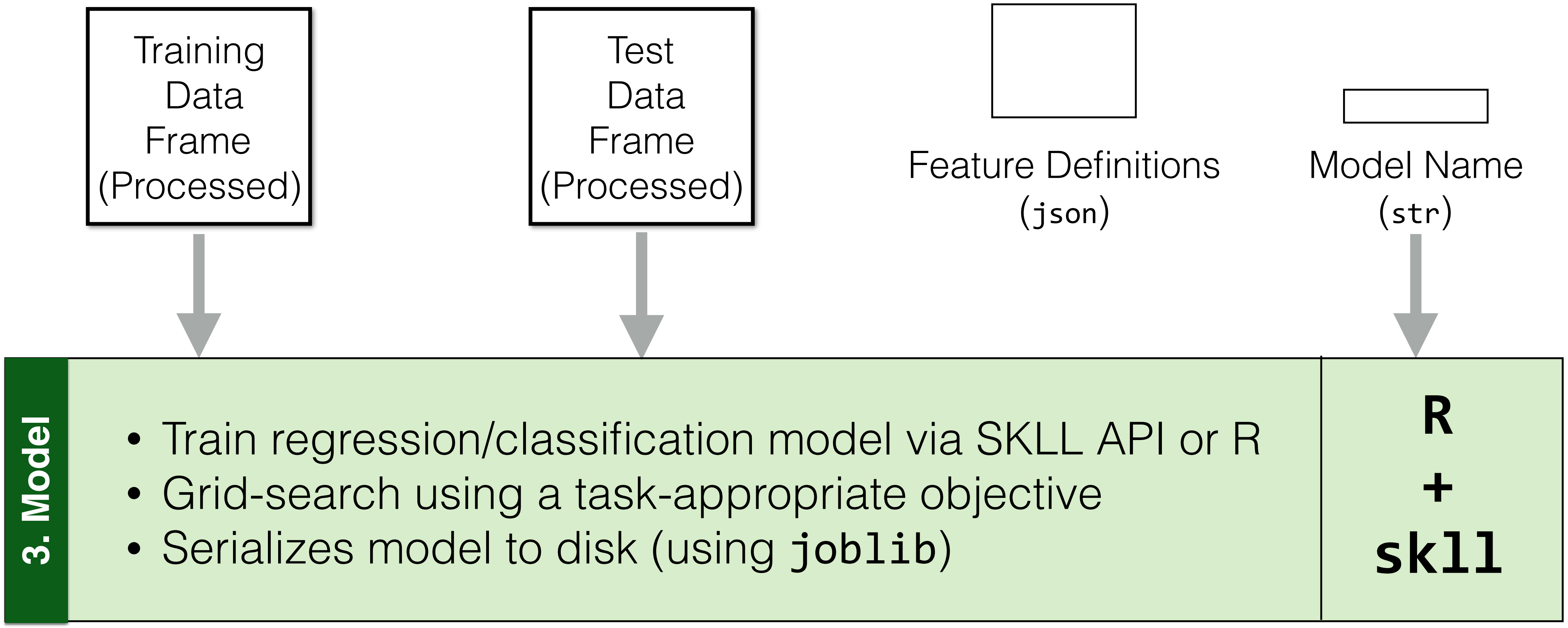
- Train regression/classification model via SKLL API or R
- Grid-search using a task-appropriate objective
- Serializes model to disk (using `joblib`)

R + `skll`

SKLL (pronounced “skull”) provides an API and command-line utilities to make it much simpler to run common `scikit-learn` experiments with pre-generated features.

(Presented by **@dsblanch** at PyData 2013 & 2014)

`https://github.com/EducationalTestingService/skll`



Training Data Frame (Processed)

Test Data Frame (Processed)

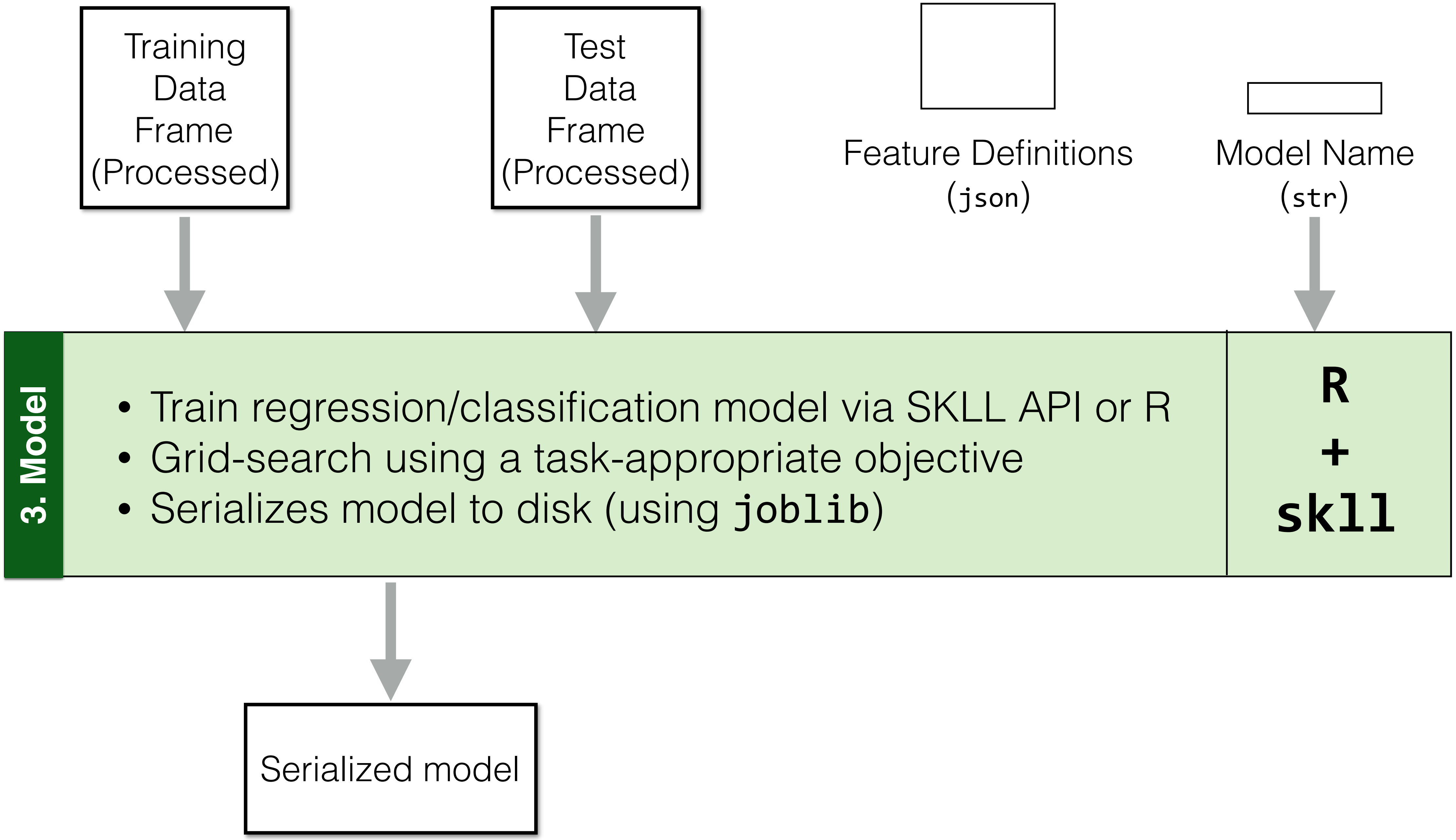
Feature Definitions (`json`)

Model Name (`str`)

3. Model

- Train regression/classification model via SKLL API or R
- Grid-search using a task-appropriate objective
- Serializes model to disk (using **`joblib`**)

R
+
`skll`



Training Data Frame (Processed)

Test Data Frame (Processed)

Feature Definitions (`json`)

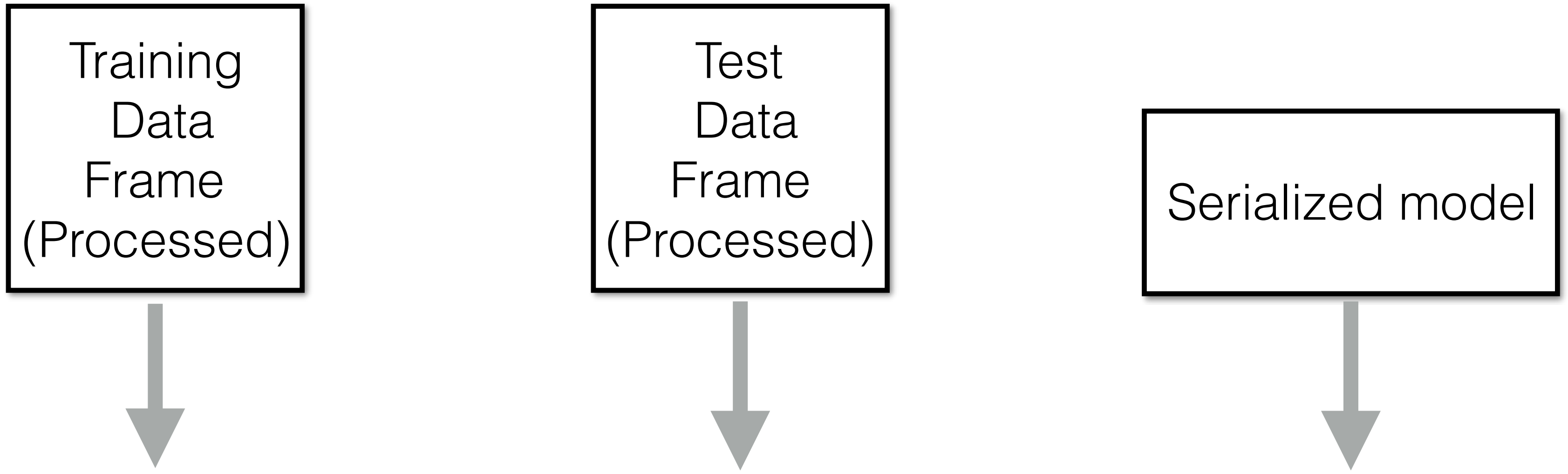
Model Name (`str`)

3. Model

- Train regression/classification model via SKLL API or R
- Grid-search using a task-appropriate objective
- Serializes model to disk (using **`joblib`**)

`R`
+
`skll`

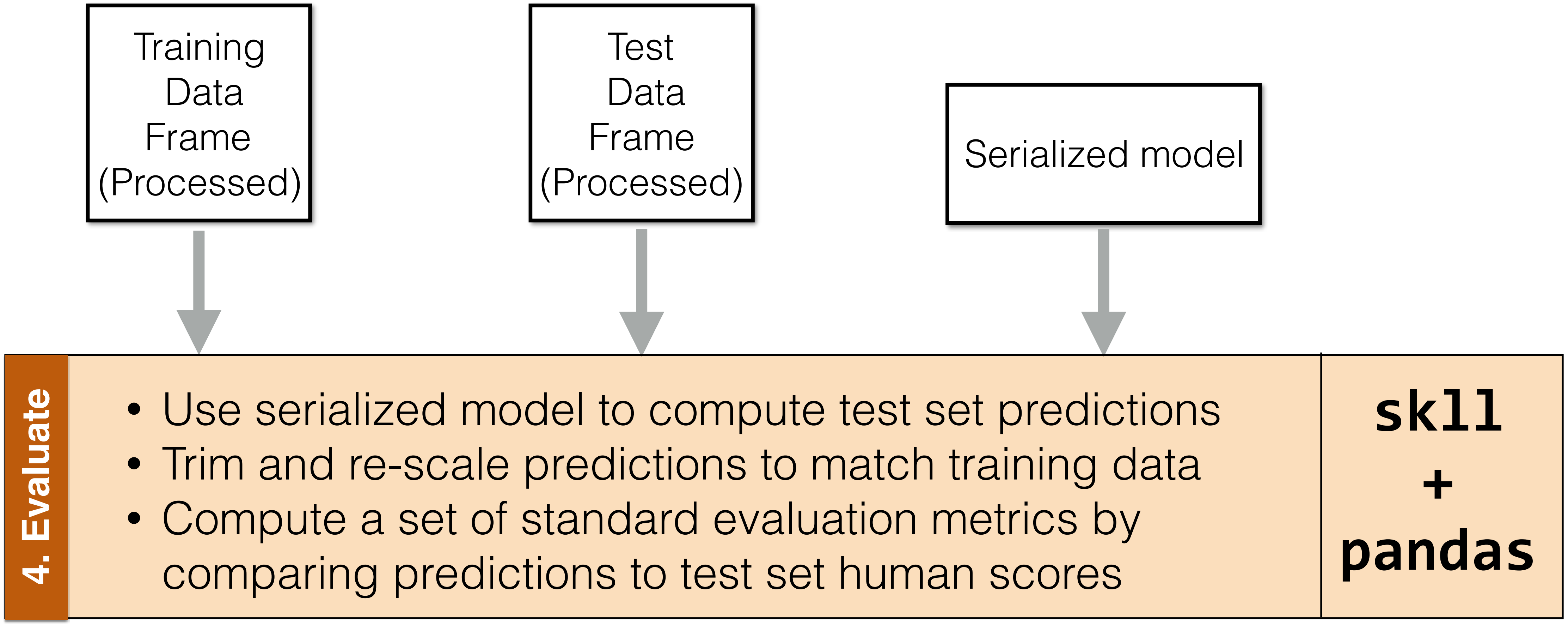
Serialized model



Training Data Frame (Processed)

Test Data Frame (Processed)

Serialized model



Training Data Frame (Processed)

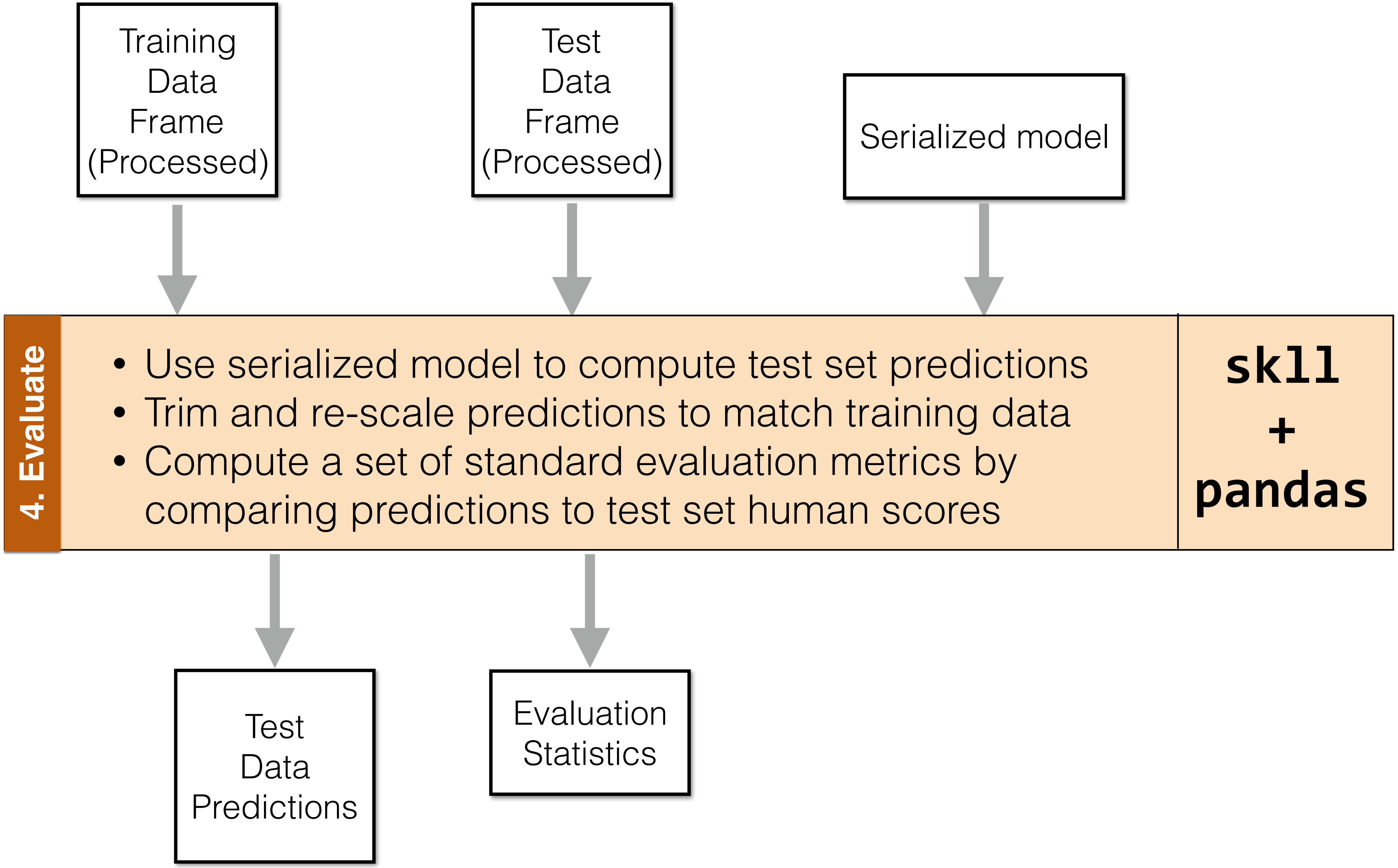
Test Data Frame (Processed)

Serialized model

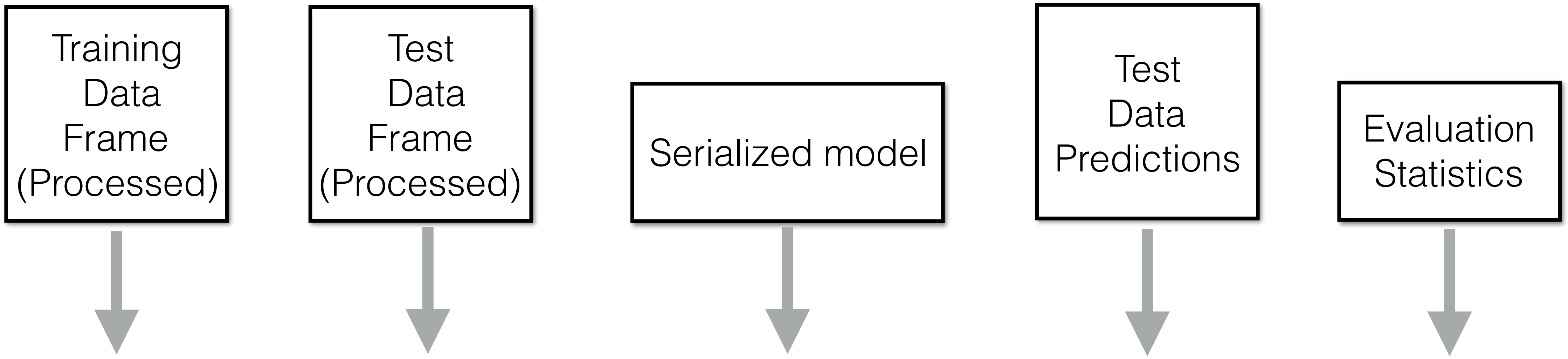
4. Evaluate

- Use serialized model to compute test set predictions
- Trim and re-scale predictions to match training data
- Compute a set of standard evaluation metrics by comparing predictions to test set human scores

`skll` + `pandas`



- Use serialized model to compute test set predictions
- Trim and re-scale predictions to match training data
- Compute a set of standard evaluation metrics by comparing predictions to test set human scores



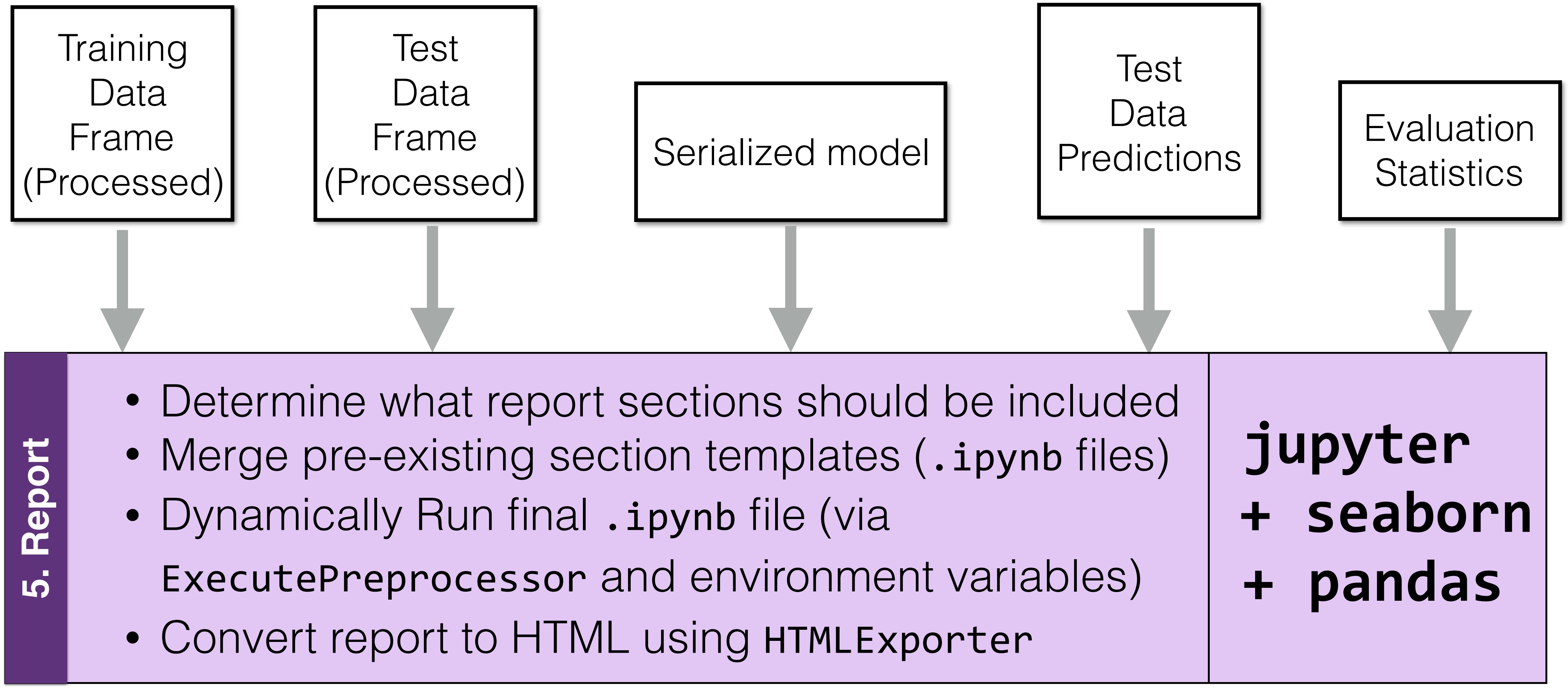
Training Data Frame (Processed)

Test Data Frame (Processed)

Serialized model

Test Data Predictions

Evaluation Statistics



Training Data Frame (Processed)

Test Data Frame (Processed)

Serialized model

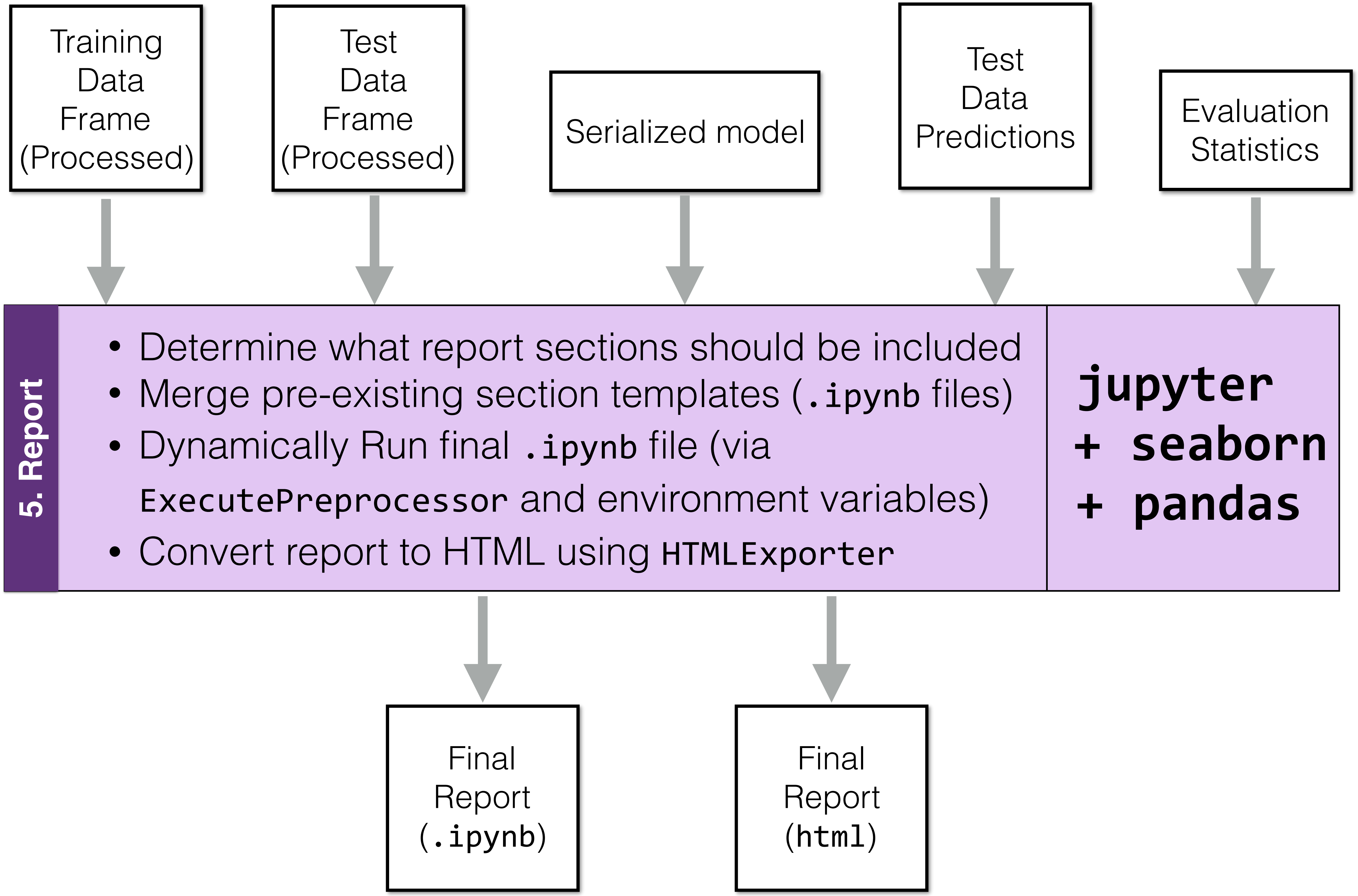
Test Data Predictions

Evaluation Statistics

5. Report

- Determine what report sections should be included
- Merge pre-existing section templates (`.ipynb` files)
- Dynamically Run final `.ipynb` file (via `ExecutePreprocessor` and environment variables)
- Convert report to HTML using `HTMLExporter`

`jupyter`
`+ seaborn`
`+ pandas`



- Determine what report sections should be included
- Merge pre-existing section templates (`.ipynb` files)
- Dynamically Run final `.ipynb` file (via `ExecutePreprocessor` and environment variables)
- Convert report to HTML using `HTMLExporter`

`jupyter`
`+ seaborn`
`+ pandas`

Demo

Summary

- Machine learning in high-stakes educational assessment requires additional number crunching to verify accuracy and interpretability.
- Need a pipeline to compare a large number of research experiments using a standardized, easy-to-read report.
- The scientific Python stack makes it super easy to implement all stages of the pipeline!
- In progress
    - Release under open-source license (2016 release)
    - A CherryPy/JS web-app to allow wider reach

Questions?

https://github.com/EducationalTestingService

https://github.com/desilinguist

@haikuman