Model Quality

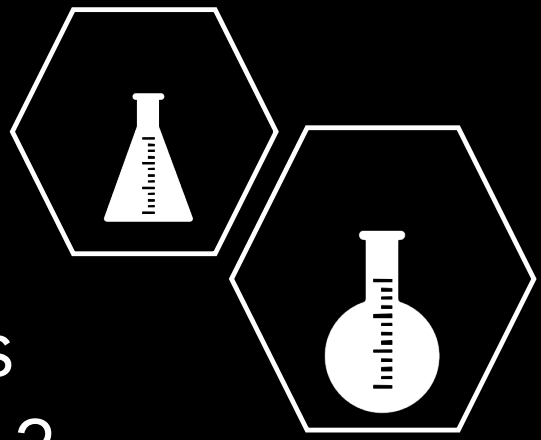
Model Selection, Evaluation, and Documentation

Jin Guo
SOCS McGill University

How do you know the model is doing what you intended to do?

Data Scientists/Model Developers

How do you know the model is doing what you intended to do?

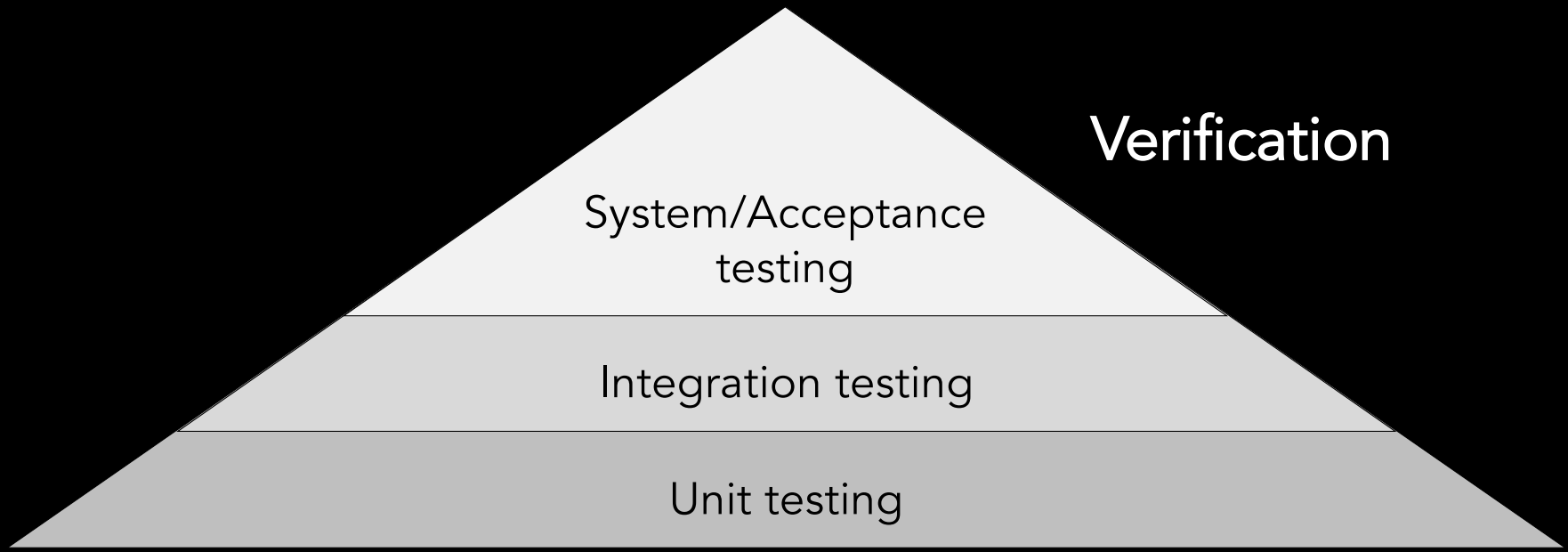


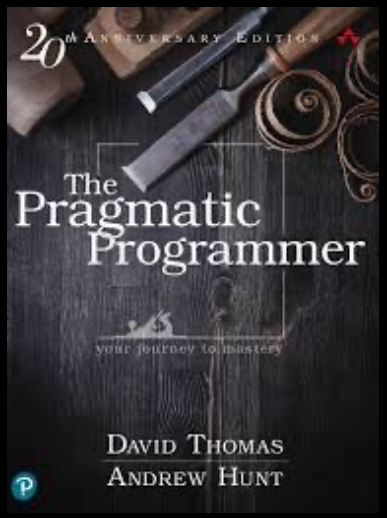
How do you know a program is doing what you intended to do?

Software Testing?

Validation

Verification





"All software you write *will* be tested—if not by you and your team, then by the eventual users—so you might as well plan on testing it thoroughly … "

Test Case Example during Unit Test

```
import org.junit.jupiter.api.Test;

import static org.junit.jupiter.api.Assertions.*;


class UndergradTest {

    @Test
    void getFirstName() {
        Student s = new Undergrad("001","Lily", "Joe");
        assertEquals("Lily", s.getFirstName());
    }
}
```

assertEquals method

Oracle

```
public static void assertEquals(Object expected,
                                Object actual)
```

Activity 1

- How do you, as a model developer, consider a model is performing reasonable? You can draw from your experience.
- How do that compare model evaluation with the unit test practice for traditional software source code? What are the transferable consideration, and what are not?
- Summarize your comparison on Miro.

Model Evaluation VS Software Unit Testing

- Evaluation Means
- Evaluation Objective
- What do to in the case of unsatisfied evaluation outcome
- Quality of the evaluation itself

…

Metrics for Machine Learning Model

Activity 2

1.Identify the metrics you have used before. Describe its context and why you chose those metrics.
2.Have you used metrics not in this table before? If so, describe it and its context.
3.Any observations form this table.

	Accuracy	Root mean Square error	True positive Rate	False Positive Rate	Precision	Recall	F-measure		Kononenko and Bratko's information score
Algorithm	Acc	RMSE	TPR	FPR	Prec	Rec	*F*	AUC	K & B
NB	71.7	0.4534	0.44	0.16	0.53	0.44	0.48	0.7	48.1118
C45	75.5	0.4324	0.27	0.04	0.74	0.27	0.4	0.59	34.2789
3NN	72.4	0.5101	0.32	0.1	0.56	0.32	0.41	0.63	43.3682
RIP	71	0.4494	0.37	0.14	0.52	0.37	0.43	0.6	22.3397
SVM	69.6	0.5515	0.33	0.15	0.48	0.33	0.39	0.59	54.8934
Bagging	67.8	0.4518	0.17	0.1	0.4	0.17	0.23	0.63	11.3004
Boosting	70.3	0.4329	0.42	0.18	0.5	0.42	0.46	0.7	34.4795
RF	69.23	0.47	0.33	0.15	0.48	0.33	0.39	0.63	20.7763

Area under the ROC curve

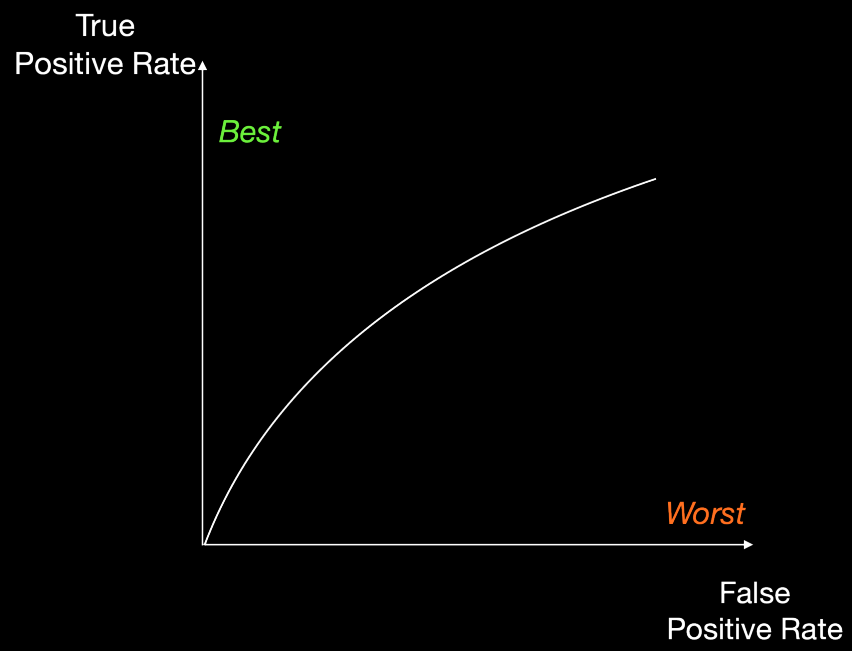
Common Metrics

Accuracy	*(TP + TN) / (P + N)*
Precision	*TP / (TP + FP)*
Recall	*TP / (TP + FN)*
F-measure	$\frac{(1+\alpha)Precision * Recall}{\alpha * Precision + Recall}$

	Model Pred Positive	Model Pred Negative
Actual Positive	*True Positive(TP)*	*False Negative(FN)*
Actual Negative	*False Positive(FP)*	*True Negative(TN)*

Input	Actual Output	Model Output
1	*No*	*Yes*
2	*No*	*No*
3	*Yes*	*No*
.....	*.....*	*.....*

ROC Analysis

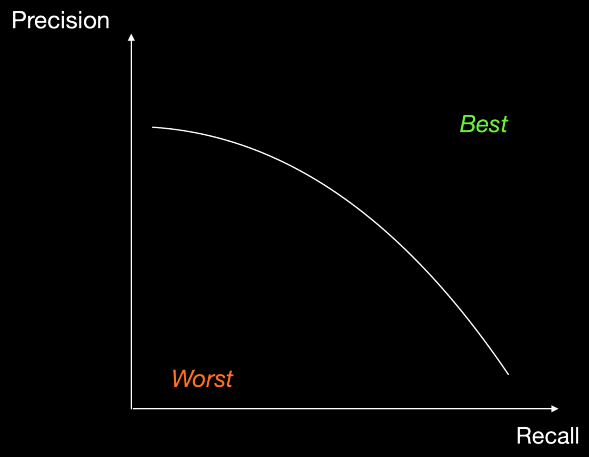


	Model Pred Positive	**Model Pred Negative**
Actual Positive	*True Positive(TP)*	*False Negative(FN)*
Actual Negative	*False Positive(FP)*	*True Negative(TN)*

True Positive Rate (Hit Rate) $TP / (TP + FN)$

False Positive Rate (Fallout) $FP / (FP + TN)$

Precision-Recall (PR) Curves



	Model Pred Positive	Model Pred Negative
Actual Positive	*True Positive(TP)*	*False Negative(FN)*
Actual Negative	*False Positive(FP)*	*True Negative(TN)*

Input	Actual Output	Model Output
1	*No*	*0.8 -> Yes*
2	*No*	*0.7 -> No*
3	*Yes*	*0.75 -> No*
.....	*.....*	*.....*
14	*Yes*	*0.4*

Domain Specific Metrics

- Mean Average precision
  - Information Retrieval

- ROUGE, BLEU score
  - Text generation

…

Other properties of the model

- Fairness : group and individual level
  - Equalized odds
  - Demographic parity
  - …
- Explainability: model and individual level
  - Complexity
  - Representativeness
  - …

Activity 3

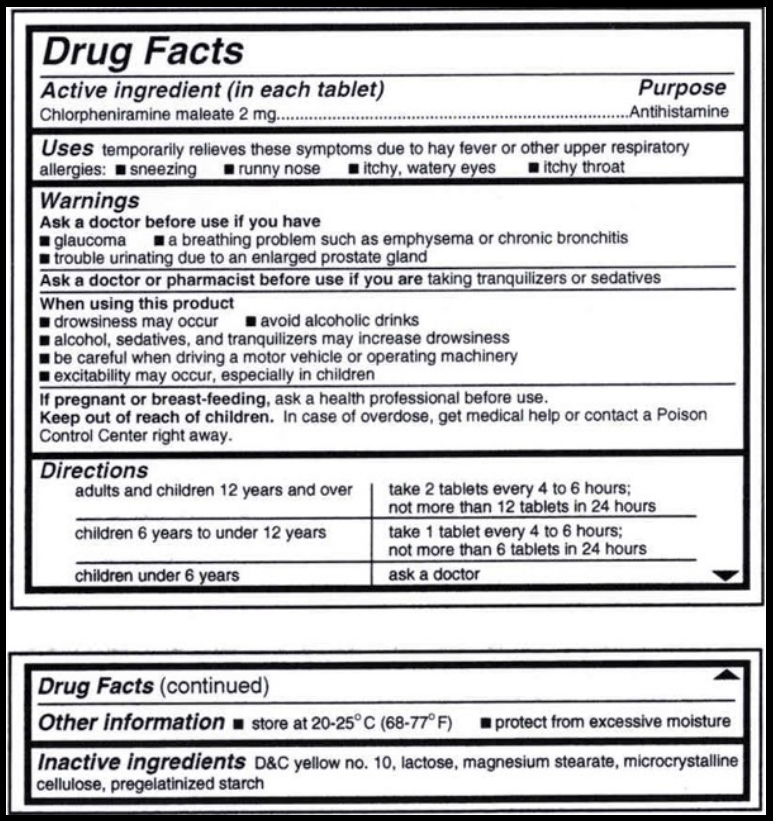
- What content do/should you include in the model documentation?

- What content do you need as a model user?

Model Documentation

- Examples:
  - https://keras.io/api/applications/vgg/
  - https://huggingface.co/bert-base-multilingual-cased
  - https://modelcards.withgoogle.com/object-detection

Model Documentation



Drug Facts

Active ingredient (in each tablet) — **Purpose**
Chlorpheniramine maleate 2 mg ..... Antihistamine

Uses temporarily relieves these symptoms due to hay fever or other upper respiratory allergies: ■ sneezing ■ runny nose ■ itchy, watery eyes ■ itchy throat

Warnings

Ask a doctor before use if you have
■ glaucoma ■ a breathing problem such as emphysema or chronic bronchitis
■ trouble urinating due to an enlarged prostate gland

Ask a doctor or pharmacist before use if you are taking tranquilizers or sedatives

When using this product
■ drowsiness may occur ■ avoid alcoholic drinks
■ alcohol, sedatives, and tranquilizers may increase drowsiness
■ be careful when driving a motor vehicle or operating machinery
■ excitability may occur, especially in children

If pregnant or breast-feeding, ask a health professional before use.
Keep out of reach of children. In case of overdose, get medical help or contact a Poison Control Center right away.

Directions

adults and children 12 years and over	take 2 tablets every 4 to 6 hours; not more than 12 tablets in 24 hours
children 6 years to under 12 years	take 1 tablet every 4 to 6 hours; not more than 6 tablets in 24 hours
children under 6 years	ask a doctor

Drug Facts (continued)

Other information ■ store at 20-25°C (68-77°F) ■ protect from excessive moisture

Inactive ingredients D&C yellow no. 10, lactose, magnesium stearate, microcrystalline cellulose, pregelatinized starch

Image from: https://www.fda.gov/drugs/resources-you-drugs/over-counter-medicine-label-take-look

Model Card

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. 2019.

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

On Thursday:

More on Model Quality (Saskia)

Model -> System