CLASSIFYING LABORATORY TEST RESULTS USING MACHINE LEARNING

Joy (Sizhe) Chen, Kenny Chiu, William Lu, Nilgoon Zarei

AUGUST 31, 2018

TEAM

Joy (Sizhe) Chen

Kenny Chiu

William Lu

Nelly (Nilgoon) Zarei

AGENDA

- Background
- Project Scope
- Dataset
- Machine Learning Approach and Results
- Symbolic Approach and Results
- Pipeline Architecture
- Future Work

BACKGROUND

BACKGROUND

Lab Result
Specimen rejected | Test not performed. | No evidence of HCV infection.
No Bordetella pertussis DNA detected by PCR.
Result inconclusive. | Culture results to follow. | Varicella Zoster Virus | 'Isolated.'
'Organism identified as:' | Haemophilus influenzae | Biot… …rotypable (non encapsulated)

Project Goal: Automate the classification process!

Test …	…est Outcome	Organism Name
No	Negative	*Not Found
Yes	Negative	*Not Found
Yes	Indeterminate	*Not Found
Yes	Positive	Haemophilus influenzae

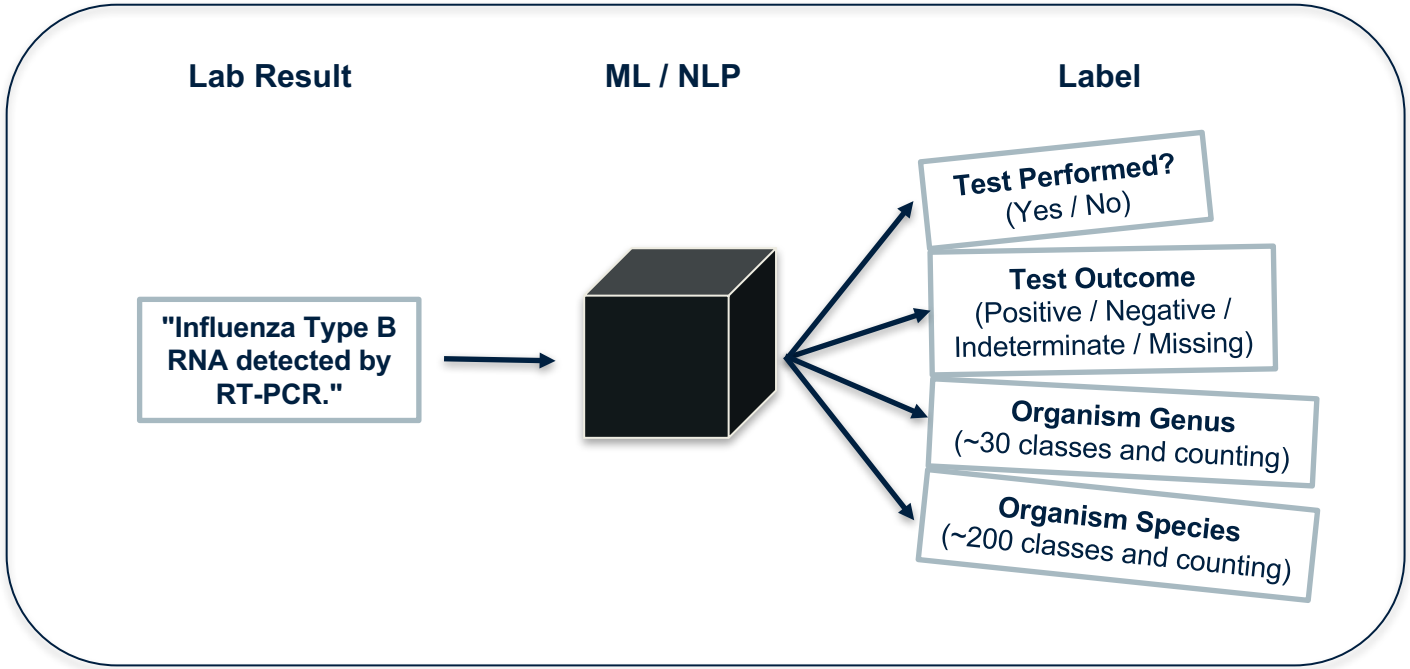
Semi-structured free form text data from lab reports containing raw test results

Manual classification process (expensive, slow)

Structured data used to analyze population-level disease trends

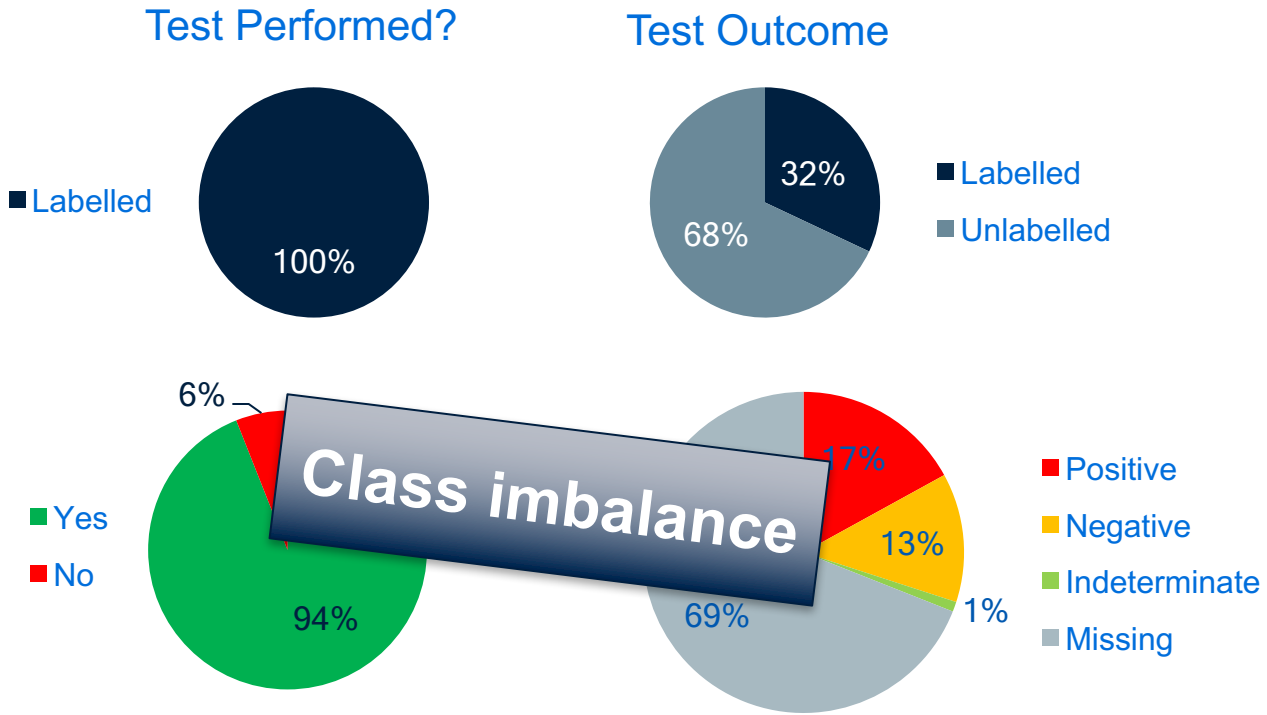
PROJECT SCOPE

Identify, implement, and test appropriate machine learning and natural language processing techniques for interpreting and labeling unstructured lab results



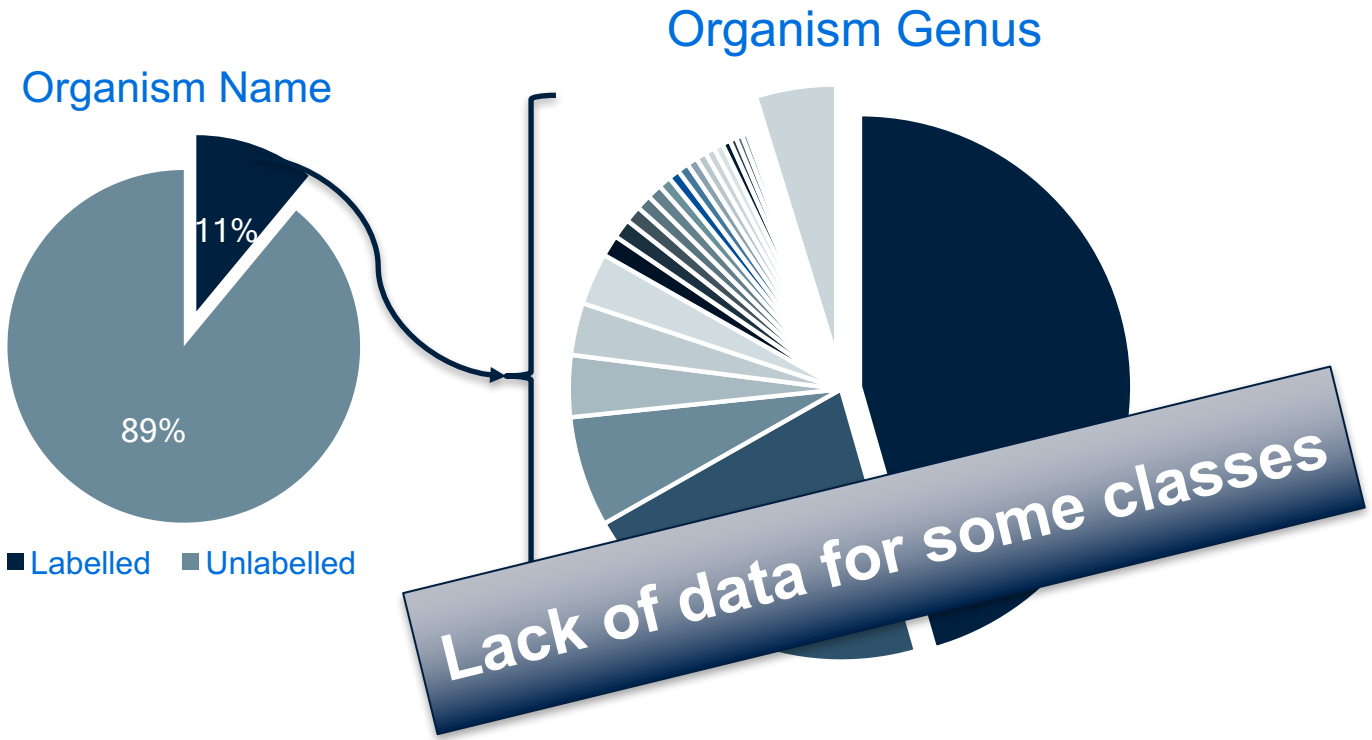
DATASET

~1 million rows; ~360K usable rows after filtering out proficiency tests and purely numeric results

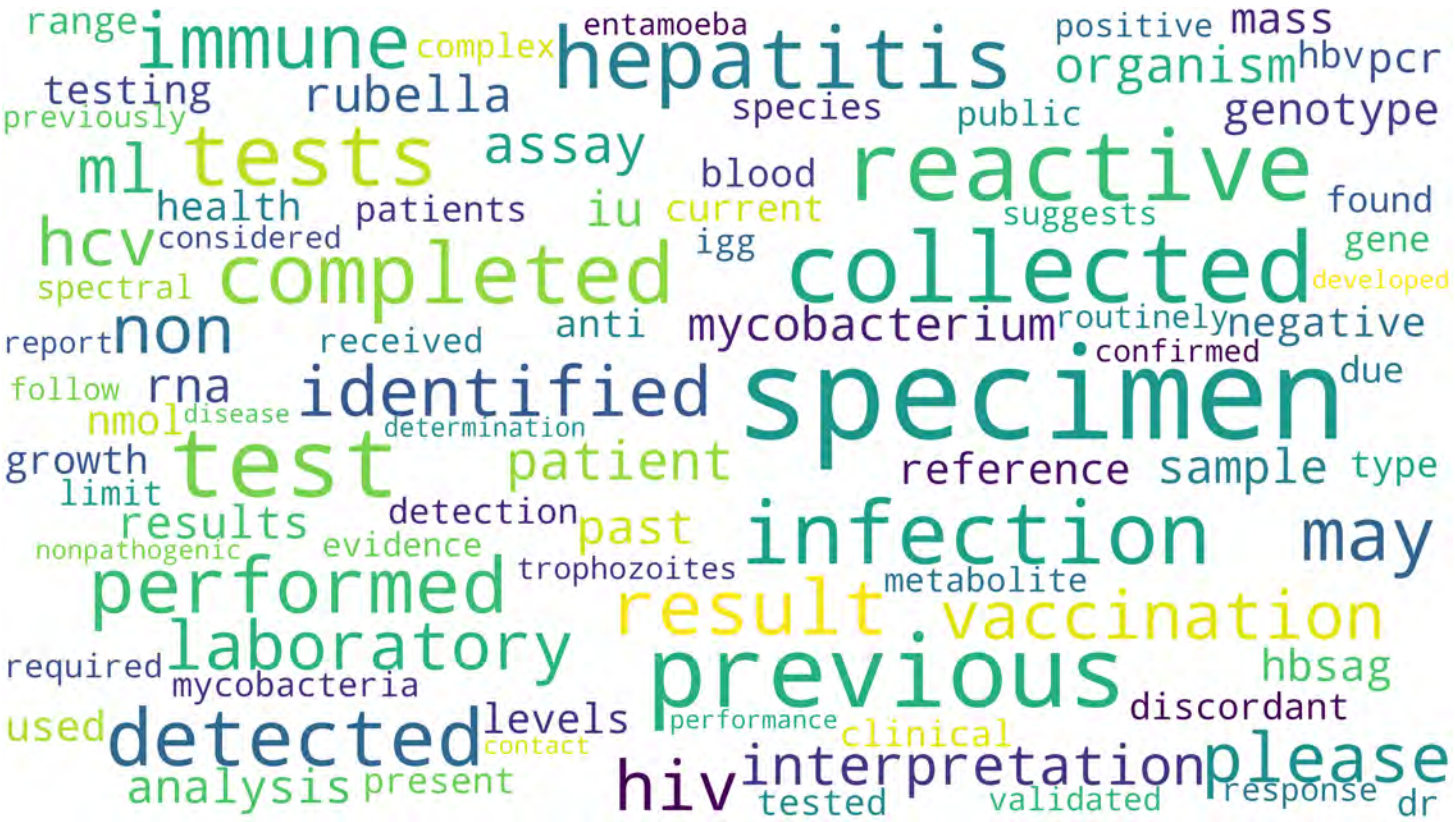


DATASET

~1 million rows; ~360K usable rows after filtering out proficiency tests and purely numeric results



DATASET


range
immune
complex
entamoeba
hepatitis
positive
mass
organism
hbvpcr
testing
rubella
species
public
genotype
previously
ml
tests
assay
blood
reactive
health
patients
iu
current
suggests
found
hcv
considered
igg
gene
completed
collected
spectral
developed
non
anti
mycobacterium
routinely
negative
report
received
confirmed
follow
rna
identified
specimen
due
nmol
disease
determination
growth
test
patient
reference
sample
type
limit
results
detection
past
infection
may
nonpathogenic
evidence
trophozoites
metabolite
performed
result
vaccination
laboratory
previous
hbsag
required
mycobacteria
discordant
used
detected
levels
performance
clinical
contact
please
analysis
present
hiv
interpretation
tested
validated
response
dr

DATASET

- Lab results may be incomplete sentences and may contain typographical errors

  BCCDC seretype: non froup 5 | Final | 12/Jun/2009 | Sputum | Streptococcus pneumoniae | STUDY

  Isolate not | Salmonella species

- Lab results may contain contradictory information

  TEST NOT PERFORMED | Galactomannan testing is valid only for Haematology and lung transplant patients with no recent antifungal exposure | Test performed at Provincial Laboratory of Public Health, Edmonton

  Organism identified as: | Neisseria meningitidis nongroupable | Upon further investigation | Organism identified as: | **Moraxella osloensis** | by 16S rRNA gene sequence analysis.

DATASET

- One organism may be positive, while another may be negative

> NEGATIVE for Shiga toxin stx1 and stx2 genes by PCR. | Isolate serotyped as: | Escherichia coli | not | O157:H7

- Lots of negative organisms may be mentioned in the result full description

> **Rhinovirus** or **Enterovirus** detected by multiplex NAT. | | **Adenovirus** detected by multiplex NAT. | | Multiplex NAT is capable of detecting Influenza A and B, Respiratory Syncytial Virus, Parainfluenza 1, 2, 3, and 4, Rhinovirus, Enterovirus, Adenovirus, Coronaviruses HKU1, NL63, OC43, and 229E, hMetapneumovirus, Bocavirus, C. pneumoniae, L. pneumophila, and M. pneumoniae. | | MULTIPLE INFECTION DETECTED

MACHINE LEARNING APPROACH

- Automatically learn patterns from existing categorized data to categorize new data

- Data is represented in terms of **features**

- Machine learning model has a number of **parameters**
  - During training, old data is used to optimize the parameter values
  - During classification on new data, a computation is performed on the new data's features and the optimized parameter values in order to determine the classifications

- Parameters are **fitted to the training data**, thus allowing the model to learn.

RESULTS – BINARY TEST OUTCOME

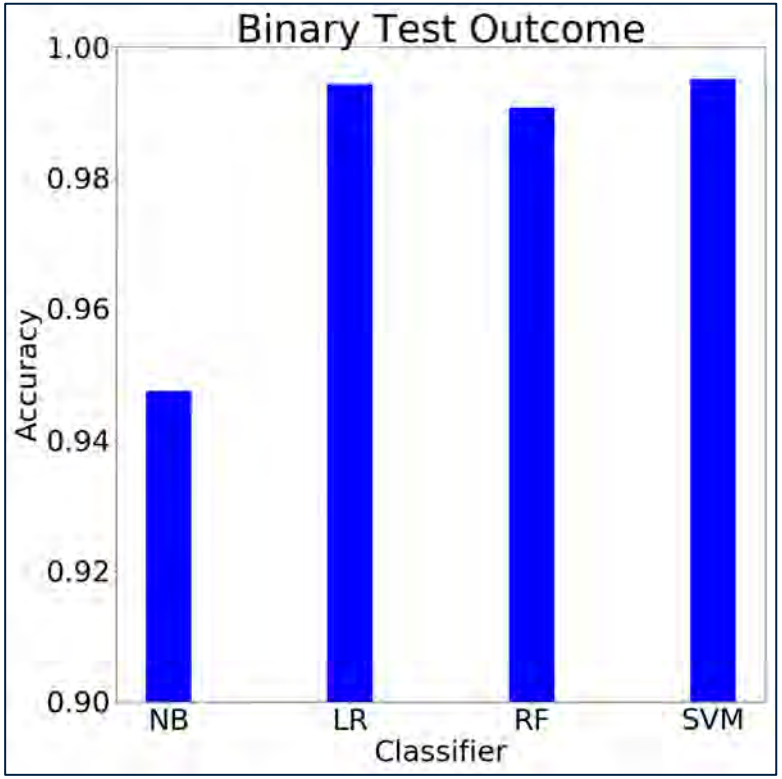
- Started with trivial case: **binary Test Outcome (Positive / Negative)**

- **Bag-of-words**: represent document by vector of integers that denote number of times each unigram (single word) appears
    - simple and convenient but loses word ordering information

“Unable to differentiate between Streptococcus mitis and Streptococcus pneumoniae.”

Unigram	Count
differentiate	1
identified	0
…	…
streptococcus	2
unable	1

RESULTS – BINARY TEST OUTCOME

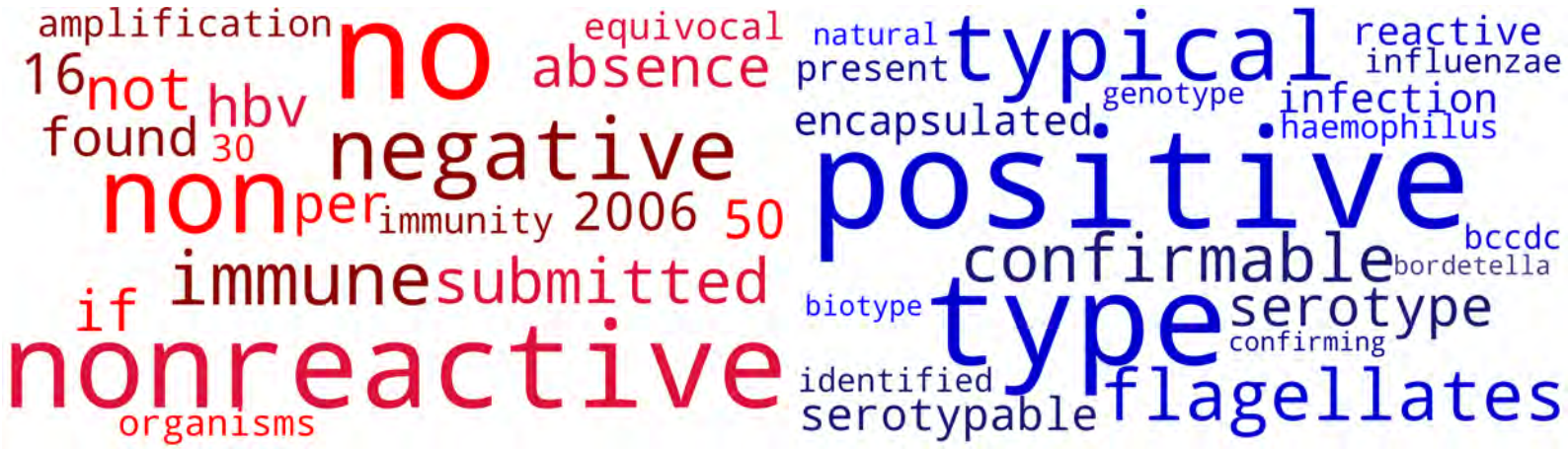


RF (100 trees)	Predicted *Positive*	Predicted *Negative*	Recall
True *Positive*	3860	41	99%
True *Negative*	16	2987	99%
Precision	99%	99%	

SVM (Linear)	Predicted *Positive*	Predicted *Negative*	Recall
True *Positive*	3885	16	99%
True *Negative*	9	2994	99%
Precision	99%	99%	

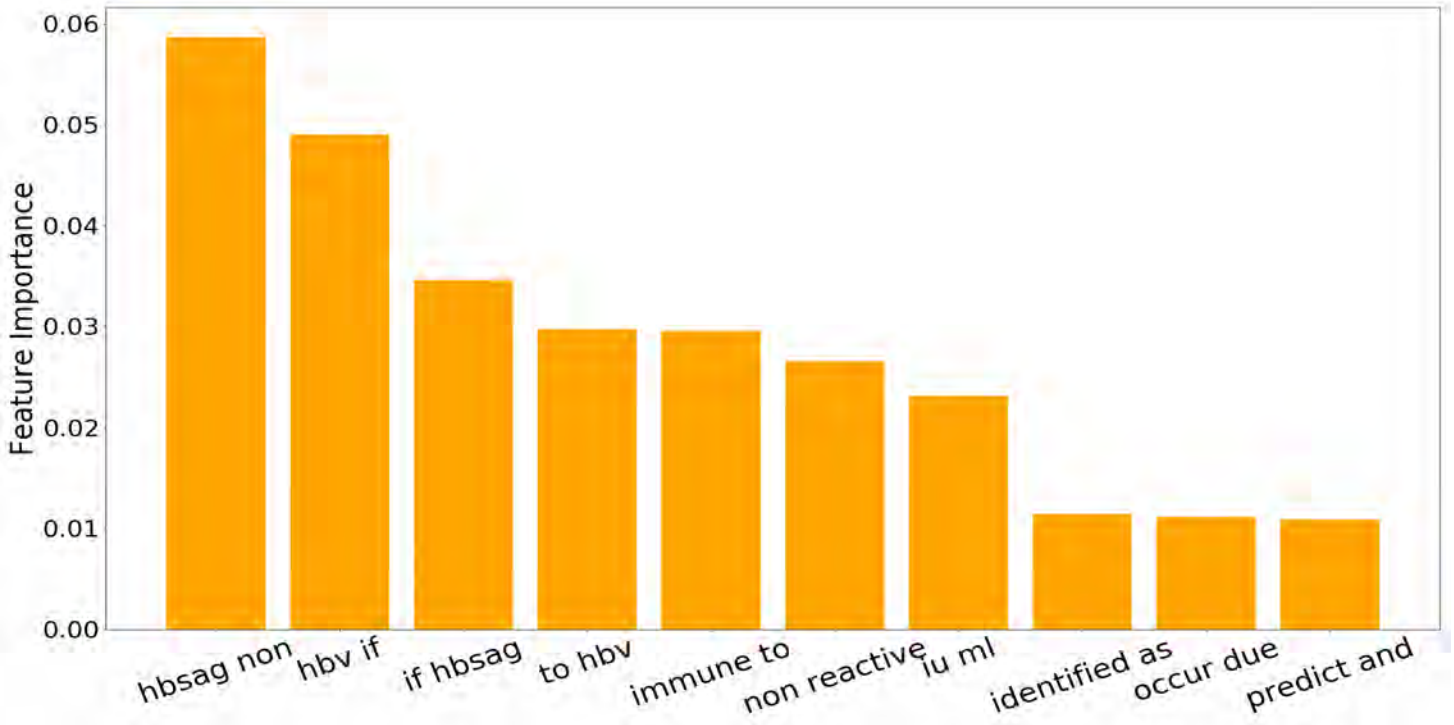
RESULTS – BINARY TEST OUTCOME

Important unigrams for Negative and Positive based on Logistic Regression weights

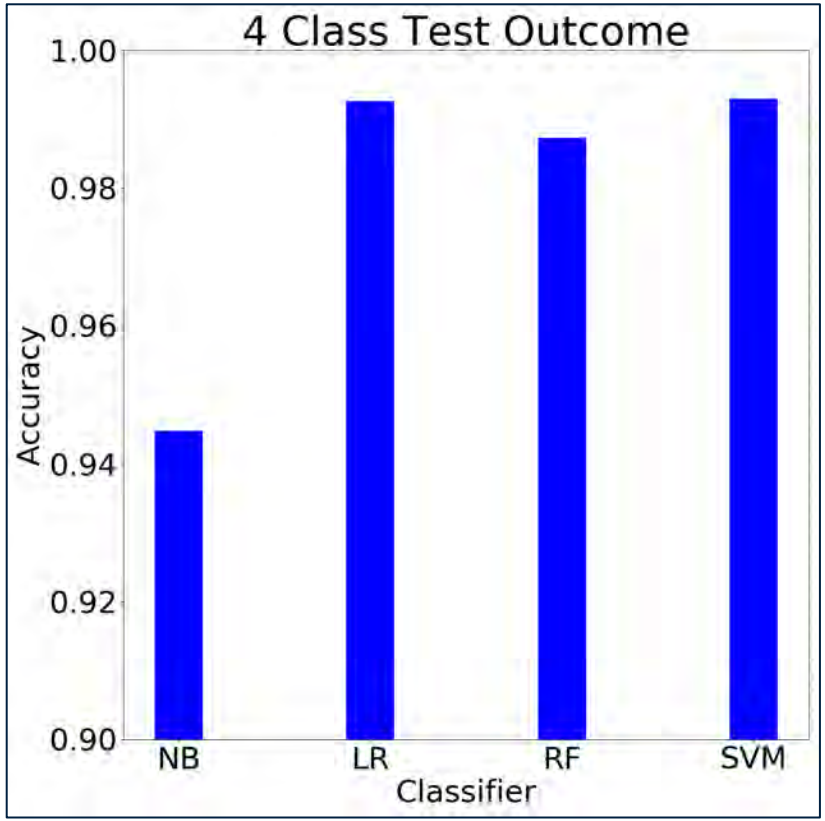


RESULTS – BINARY TEST OUTCOME

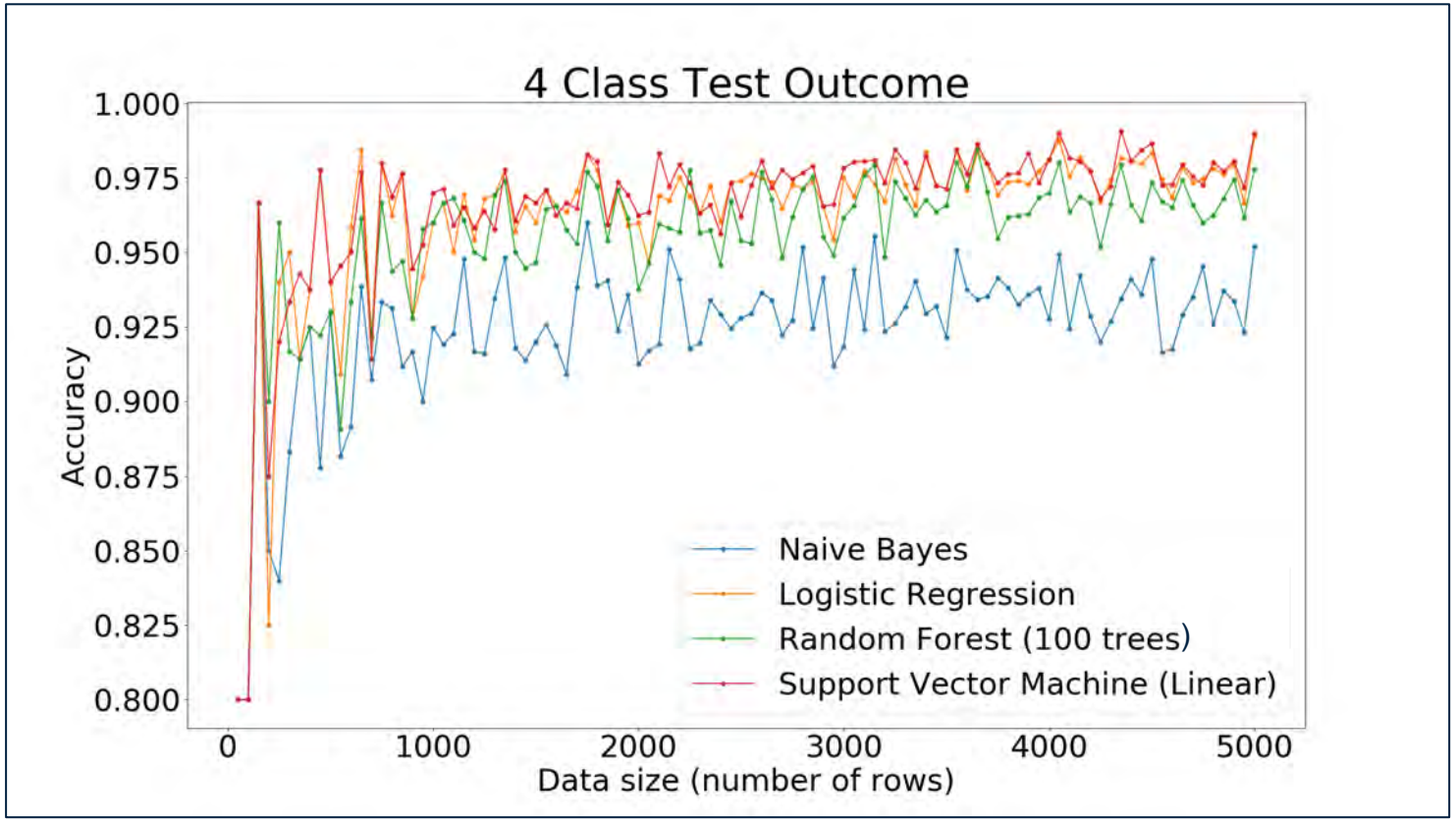
Important bigrams for Test Outcome as ranked by Random Forest



RESULTS – 4 CLASS TEST OUTCOME

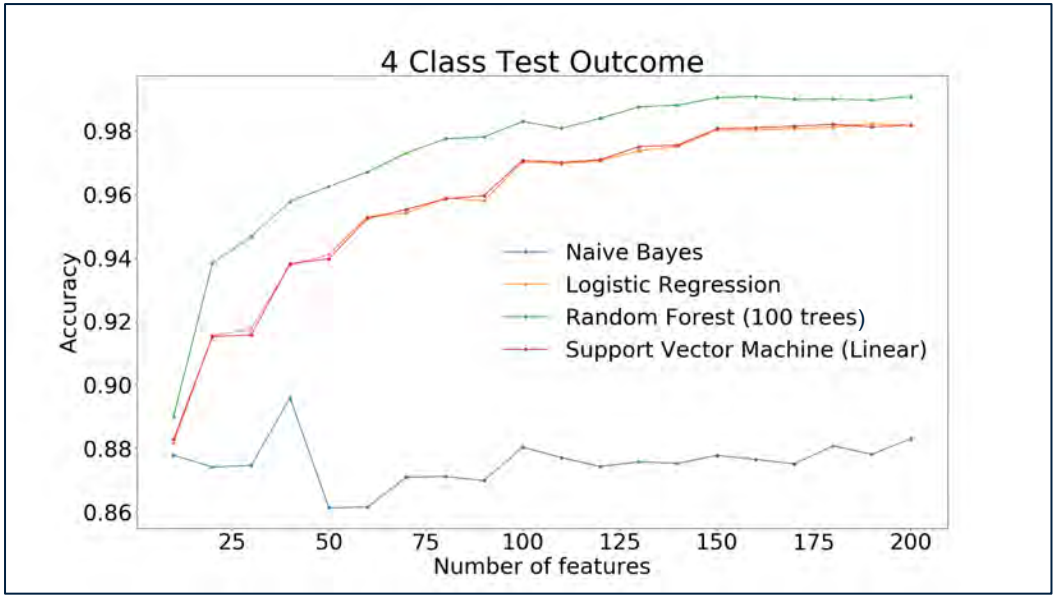


RESULTS – 4 CLASS TEST OUTCOME



RESULTS – FEATURE SELECTION

- Remove unhelpful features to **prevent overfitting** and **speed up training**.



- For example, *Test Outcome* classifiers still do well with only 200 unigram features!

RESULTS – TEST PERFORMED

Support Vector Machine (Linear): **98% accuracy**

SVM (Linear)	Predicted *Yes*	Predicted *No*	**Recall**
True *Yes*	67696	411	99%
True *No*	947	3475	79%
Precision	99%	89%	

- **Class imbalance** caused the classifier to over-predict the majority class.

RESULTS – TEST PERFORMED

- Strategies to fix this:
    - **Down-sampling** – in the training set, randomly throw out rows from the majority class until classes are balanced.
        - Disadvantage: throws out too much training data.
    - **Up-sampling** – in the training set, randomly duplicate rows from the minority class until classes are balanced.
        - Disadvantage: takes too long to train.

RESULTS – TEST PERFORMED

- **Class reweighting** – during training, penalize the classifier more for misclassifying minority rows.

Support Vector Machine (Linear): **98% accuracy**

SVM (Linear)	Predicted *Yes*	Predicted *No*	**Recall**
True *Yes*	66355	1800	97%
True *No*	429	3945	90%
Precision	99%	69%	

- Disadvantage: Reduces false positives at the expense of false negatives.

RESULTS – TEST PERFORMED

Add **bigrams** (pairs of consecutive words) and **trigrams** (triples of consecutive words) to the feature space to **boost interpretability** but at the cost of **introducing duplicates**.

Most important *Test Performed* features (ranked by Random Forest)	
Unigrams only	**Unigrams, bigrams, and trigrams**
performed	missing
not	test not
test	**test not performed**
missing	not performed
routinely	performed
patient	not

SYMBOLIC APPROACH FOR ORGANISM NAME

- **Problems** with the machine learning approach:
    - **Data-hungry** – there are not enough labelled rows for some organisms
    - **Can’t find new organisms** – there is no complete dictionary of organism names, so an approach is needed
- We must consider an **alternative approach** for classifying organism name.

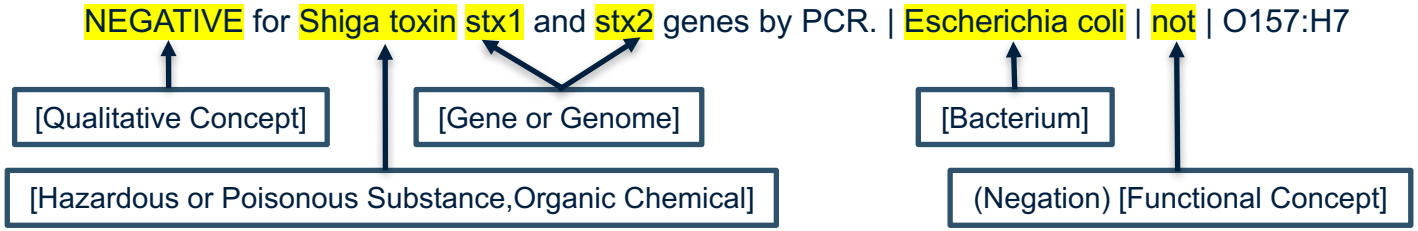
MACHINE LEARNING VS. SYMBOLIC

	Machine Learning	Symbolic
Description	Automatically learn patterns from existing categorized data (“training set”) to categorize new data (“test set”)	Tag each word by referring to a knowledge base, then apply domain rules to categorize data
Pros	• Adapts to new coding styles • More robust to typos and grammatical errors	• More interpretable • Can find labels that do not already exist in the database
Cons	• Data hungry • Long training time • Requires domain knowledge	• Long tagging time • Requires significant domain knowledge

METAMAP

***MetaMap* application**: annotates text with UMLS Metathesaurus concepts

- e.g. *Bacterium*, *Functional Concept*, *Finding*



Usages:

1. Extract all recognized *Bacterium* and *Viruses* as microorganisms
2. Generalize classifiers by including UMLS concepts as classifier inputs

RESULTS – ORGANISM GENUS

- Training stage: **construct dictionary** of all existing organisms in the database.

- We use a **two-part algorithm** for classifying *Organism Genus* label.
  - First, look at *Test Outcome* classification.
    - If *Test Outcome* is negative, *Organism Genus* is "*Not Found" by definition.

Rhinovirus or **Enterovirus** detected by multiplex NAT. | | **Adenovirus** detected by multiplex NAT. | | Multiplex NAT is capable of detecting Influenza A and B, Respiratory Syncytial Virus, Parainfluenza 1, 2, 3, and 4, Rhinovirus, Enterovirus, Adenovirus, Coronaviruses HKU1, NL63, OC43, and 229E, hMetapneumovirus, Bocavirus, C. pneumoniae, L. pneumophila, and M. pneumoniae. | | MULTIPLE INFECTION DETECTED

- This approach achieves ~85% accuracy.
  - Fails mostly on rows with **lots of negative organisms**.

RESULTS – ORGANISM SPECIES

- Training stage: **construct dictionary** mapping genus to possible species.
  - Uses existing genus and species labels in the database.

- We use a **two-part algorithm** again:
  - First, look at *Organism Genus* classification.
    - If *Organism Genus* is “*Not Found”, *Organism Species* is “*Not Found”.
  - Then, look at the list of organisms recognized by MetaMap:
    - Filter the list, keeping all species corresponding to the *Organism Genus* classification.
    - Arbitrarily pick an organism.

- This approach achieves ~50% accuracy.

RESULTS – TEST OUTCOME GENERALIZABILITY

- Original test outcome classifier **did not generalize to unlabelled dataset**:

Result Full Description	Test Outcome Prediction
Growth of mycobacteria to be identified. | | 16A306 | Mycobacterium gordonae	*Missing
Mycobacterium tuberculosis complex | Identification of species to follow. | 11S458 | Mycobacterium tuberculosis	*Missing
…	…

- Classifier **overfitted** to organism names in training set.
  - Classifier did not recognize “Mycobacterium” as an organism name because **it did not appear in the training set**.

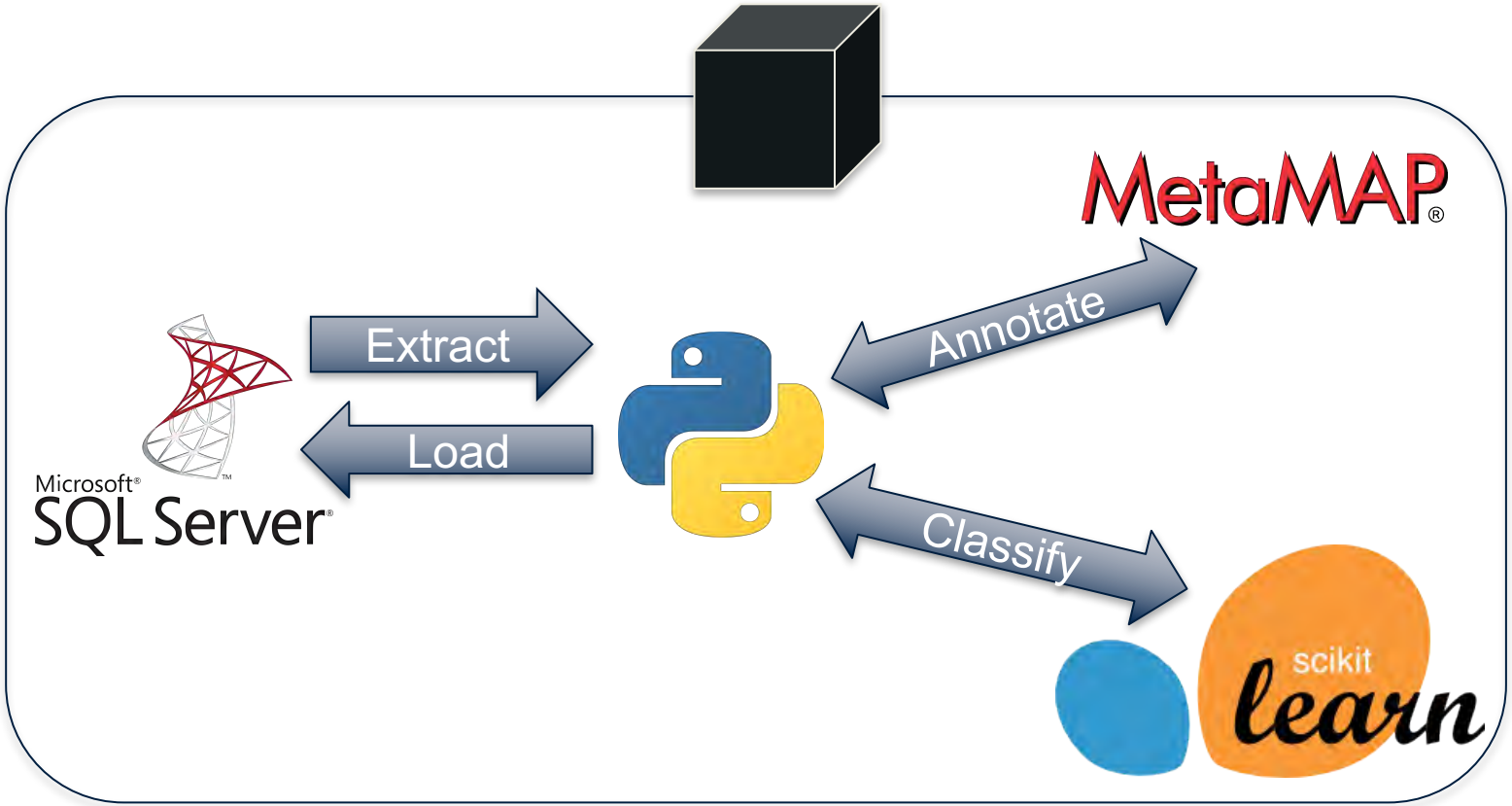
RESULTS – TEST OUTCOME GENERALIZABILITY

- Solution: **replace organism names** with special "_ORGANISM_" token.
  - Use MetaMap to identify organism names in the input text.

Result Full Description	Test Outcome Prediction
Growth of _ORGANISM_ to be identified. | | 16A306 | _ORGANISM_	Positive
ORGANISM complex | Identification of species to follow. | 11S458 | _ORGANISM_	Positive
…	…

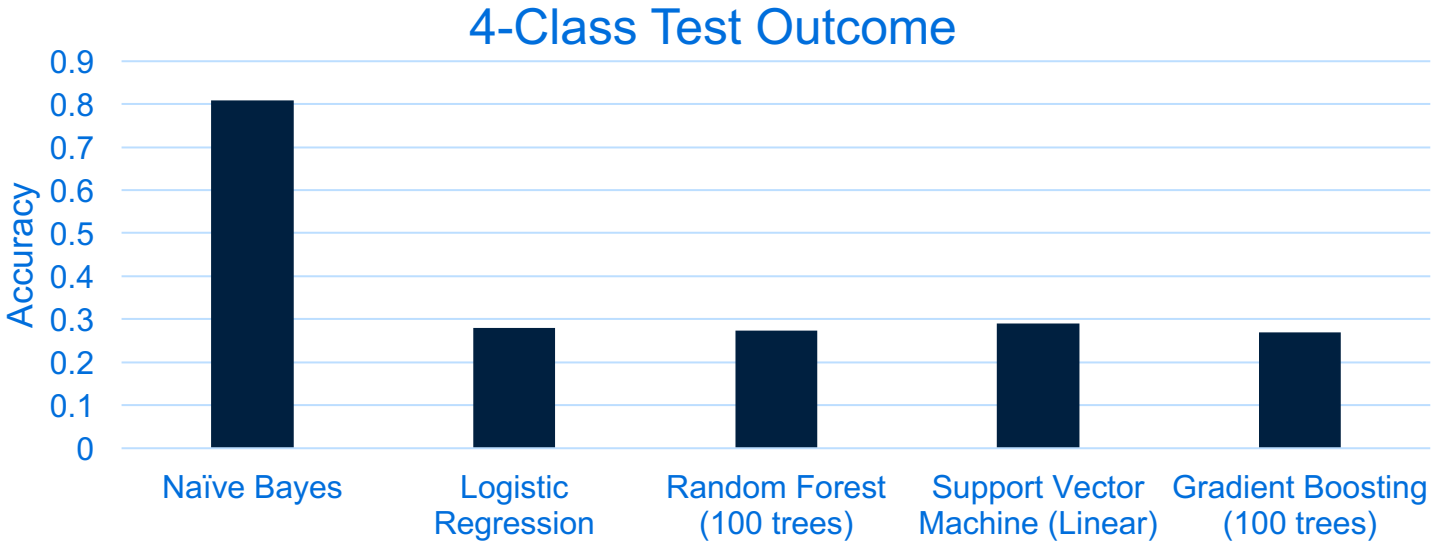
Feature engineering

PIPELINE ARCHITECTURE



FUTURE WORK

- Use all labelled **antibody tests** as training set, use all labelled **NAT/PCR tests** as testing set.



- Naïve Bayes likely worked well **by chance**.
- This hints that we should train **separate classifiers for different test types**.

FUTURE WORK

- Classify data at the **observation level** to detect which organisms were positive.

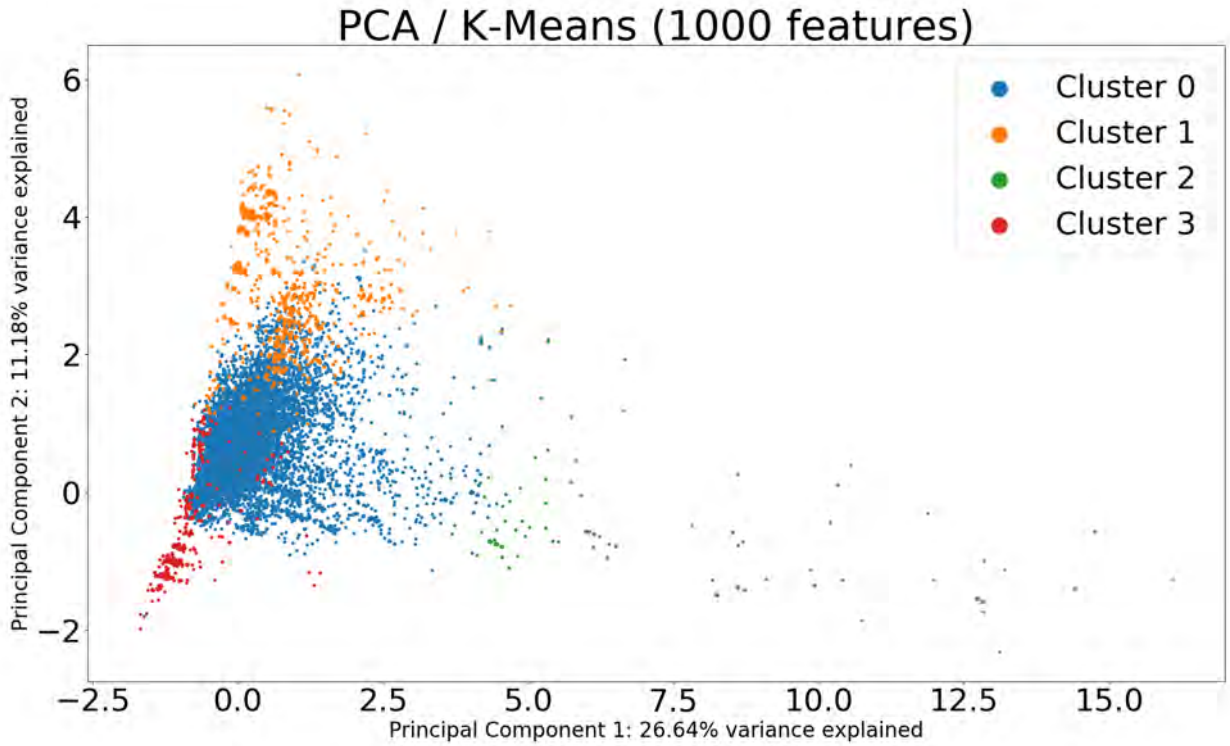
NEGATIVE for Shiga toxin stx1 and stx2 genes by PCR. | Isolate serotyped as: | Escherichia coli | not | O157:H7

Result Description	Test Outcome	Organism Name
NEGATIVE for Shiga toxin stx1 and stx2 genes by PCR.	Negative	Shiga toxin stx1 / stx2
Isolate serotyped as: | Escherichia coli | not | O157:H7	Positive	Escherichia coli non-o157 h7

- **Challenge:** no labelled data given at the observation level.
  - Workaround: Train at the test level, classify at the observation level.
  - Either relabel data **manually** or use **clustering**.

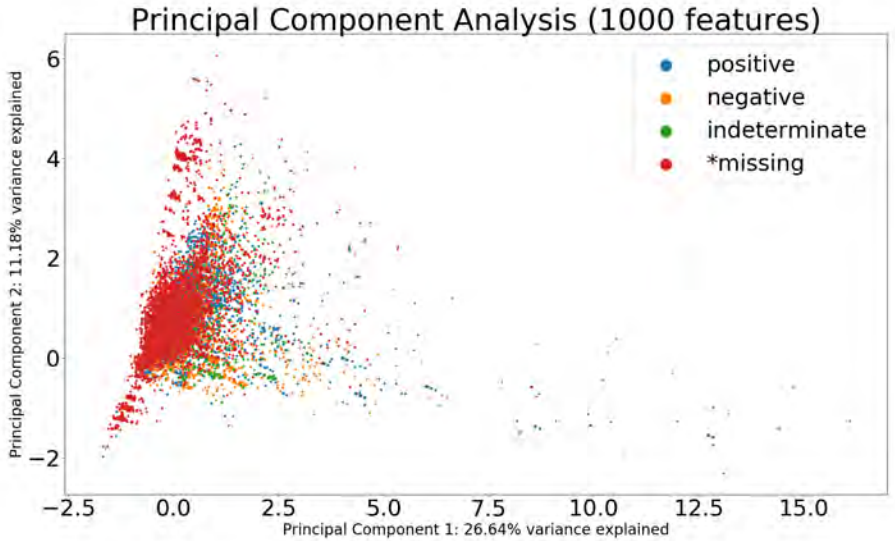
FUTURE WORK

- Use **clustering** to find patterns in **unlabelled data**.



FUTURE WORK

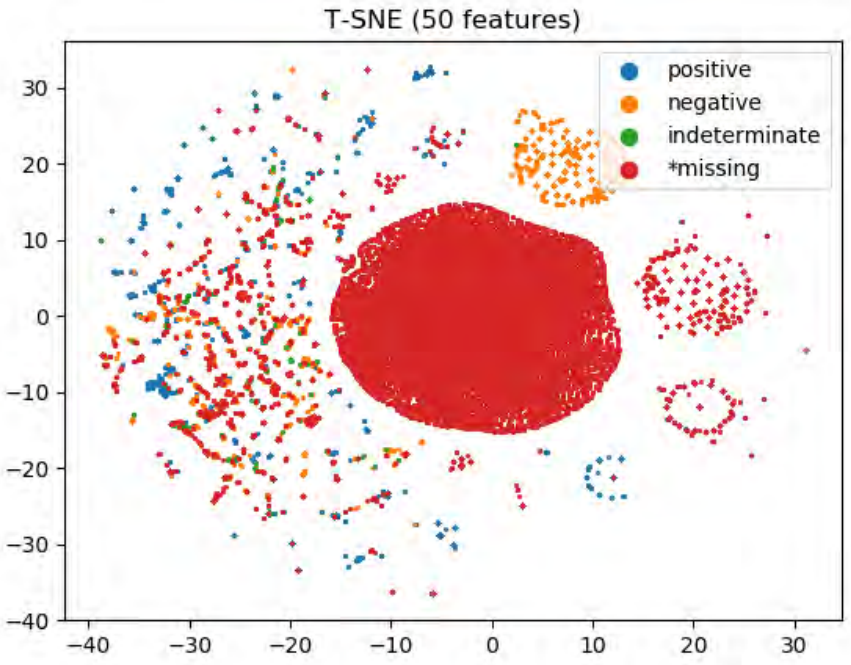
- **Principal Component Analysis** – project the data into a 2D space that explains the most variance between the data points



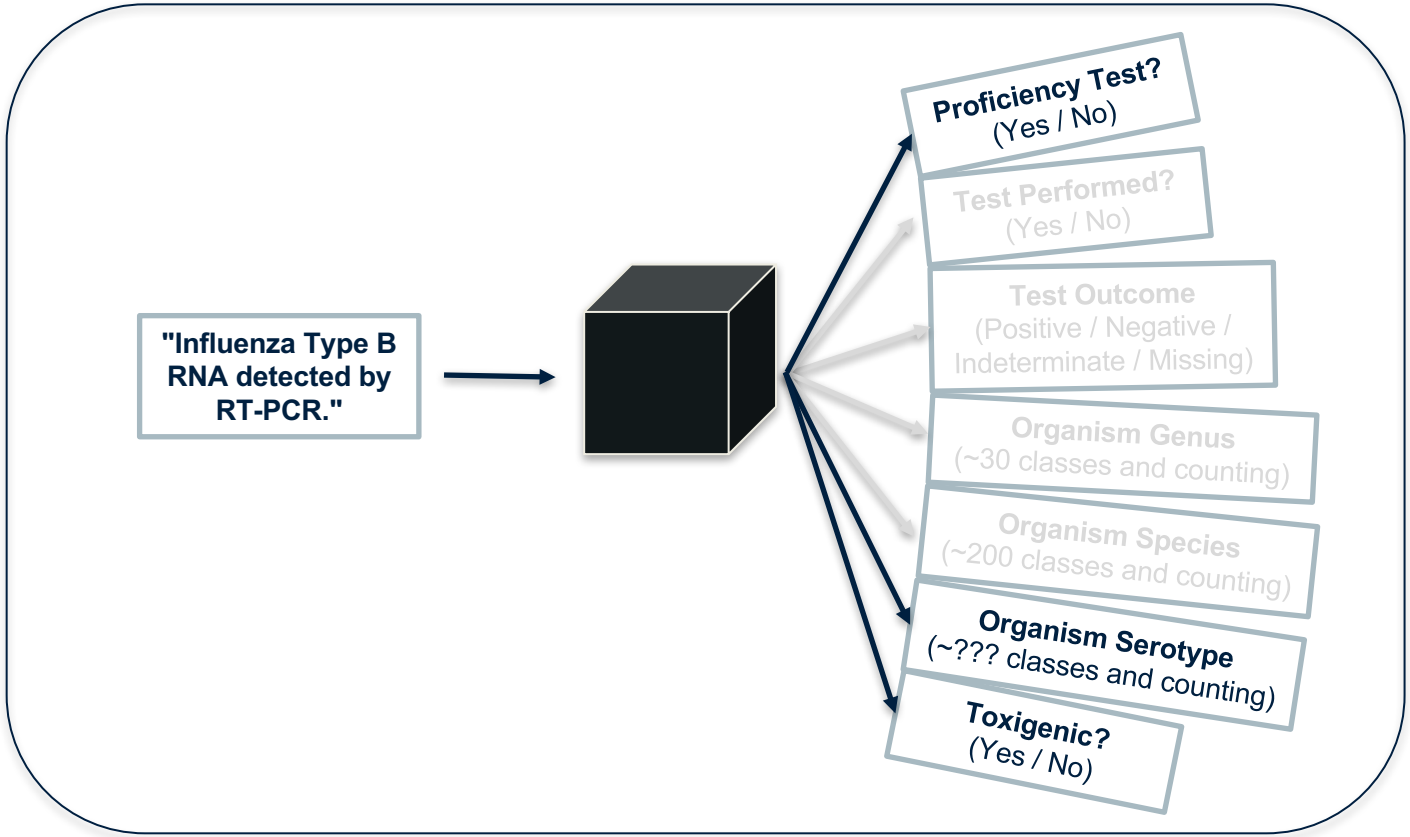
- This hints that we should try other clustering methods (hierarchical, etc.)
- However, **sum of variance explained is below 50%**, so interpretation is dangerous.

FUTURE WORK

- **t-distributed stochastic neighbour embedding** – identify a 2D “surface” that the data points reside on, and create a visualization by flattening that surface



FUTURE WORK


"Influenza Type B RNA detected by RT-PCR."
Proficiency Test? (Yes / No)
Test Performed? (Yes / No)
Test Outcome (Positive / Negative / Indeterminate / Missing)
Organism Genus (~30 classes and counting)
Organism Species (~200 classes and counting)
Organism Serotype (~??? classes and counting)
Toxigenic? (Yes / No)

FUTURE WORK

- **Ensemble methods (stacking):** Flag a row for human processing if enough classifiers disagree.

Individual Classifier Prediction				Final Prediction
NB	LR	RF	SVM	
Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	No	Flag
No	No	No	Yes	No
No	Yes	Yes	No	Flag

- Look into **classifier confidence measures** to flag rows as well.

FUTURE WORK

- Training separate classifiers for separate test types, replacing organism names, and classifying at the observation level are **not well tested**.

- Classifier still exhibits **generalizability issues**.

- Future work should aim to improve generalizability.
    - Feature engineering (removing dates, etc.)
    - Meet with a domain expert to obtain a list of domain-specific **stop words**.

QUESTIONS?