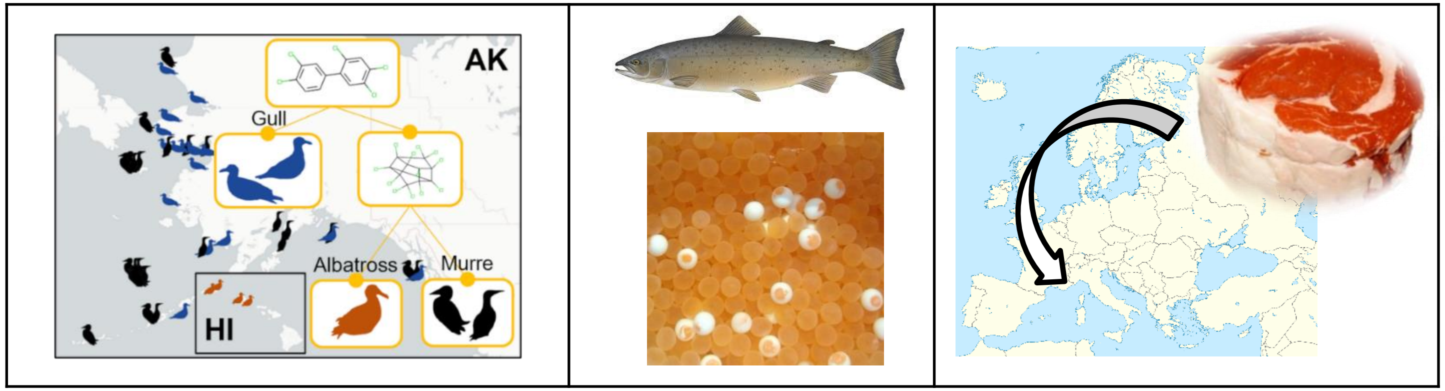
Science Ex Machina: Extracting Science from Data Using Statistical Models



NIST Isotope Metrology Webinar Series

Nathan A. Mahynski

Chemical Informatics Group

Chemical Sciences Division

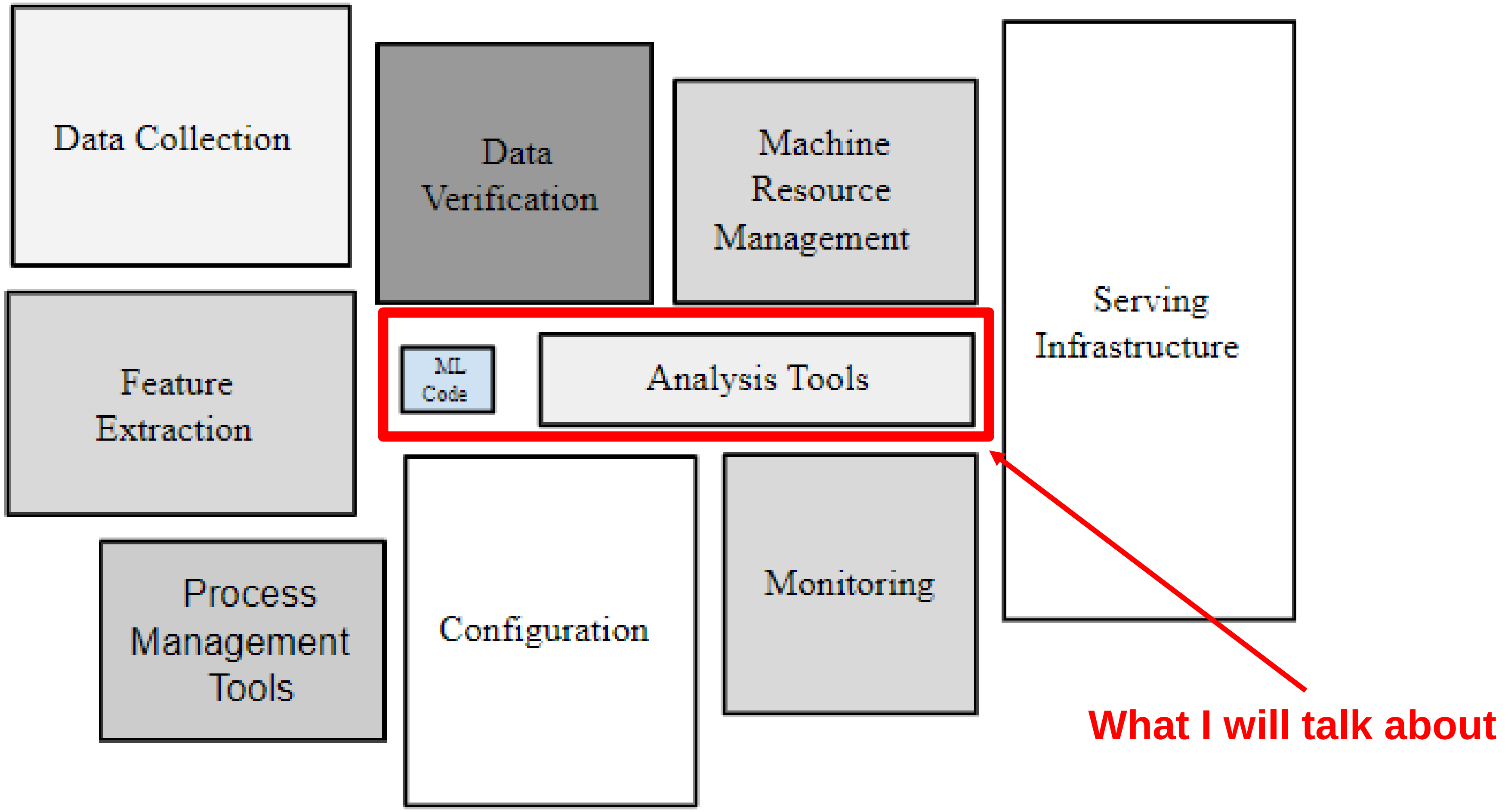
National Institute of Standards and Technology (NIST)

Gaithersburg, MD 20899

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

MATERIAL
MEASUREMENT
LABORATORY

Credit Where Credit Is Due

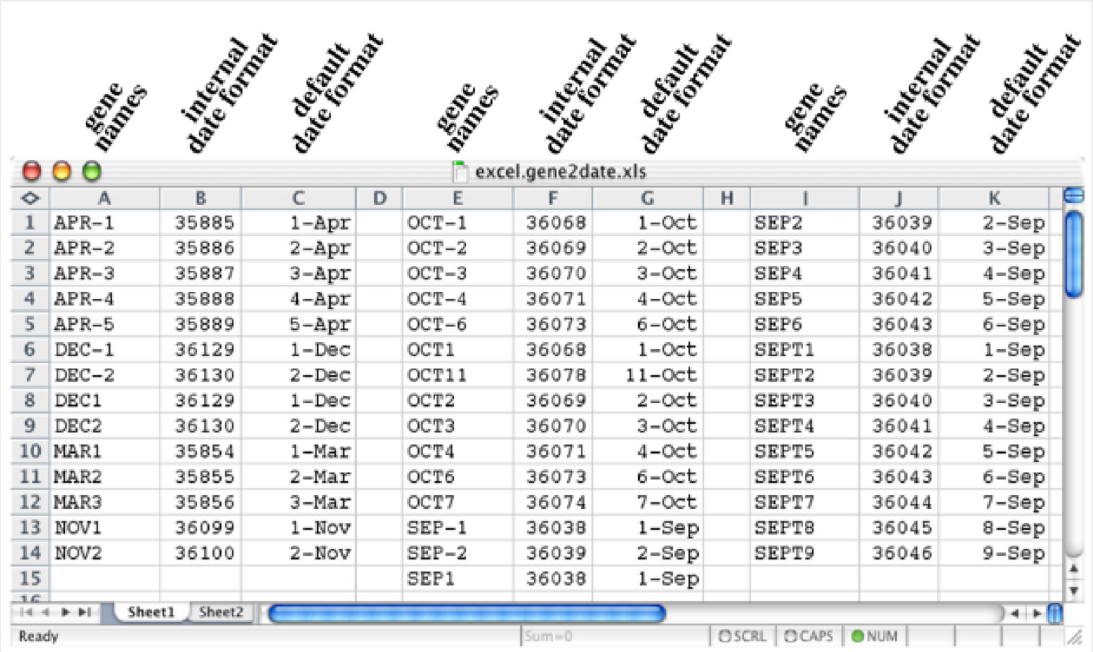


https://developers.google.com/machine-learning/crash-course/production-ml-systems

Why Can’t I Just Use Excel?

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

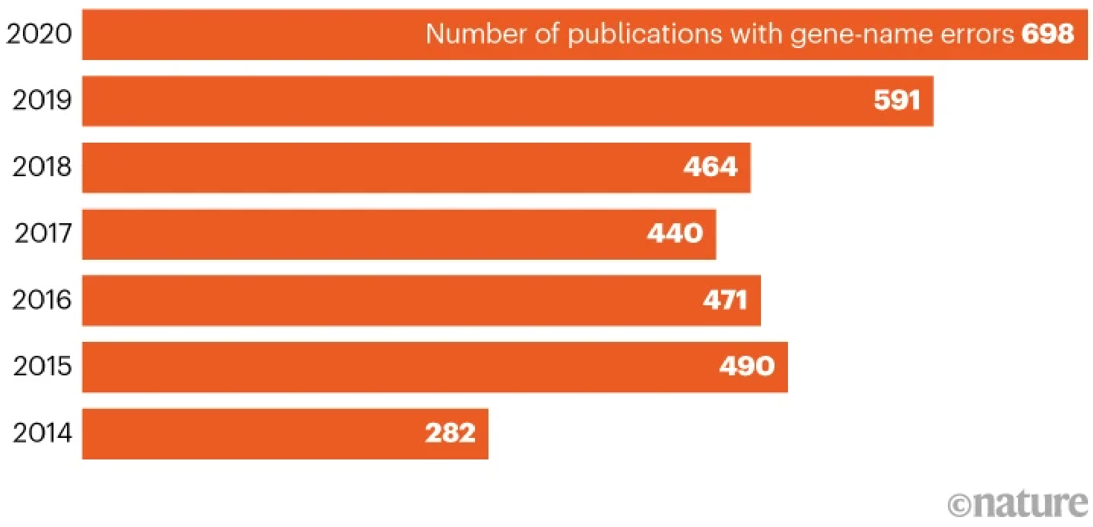


	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80

A GROWING PROBLEM

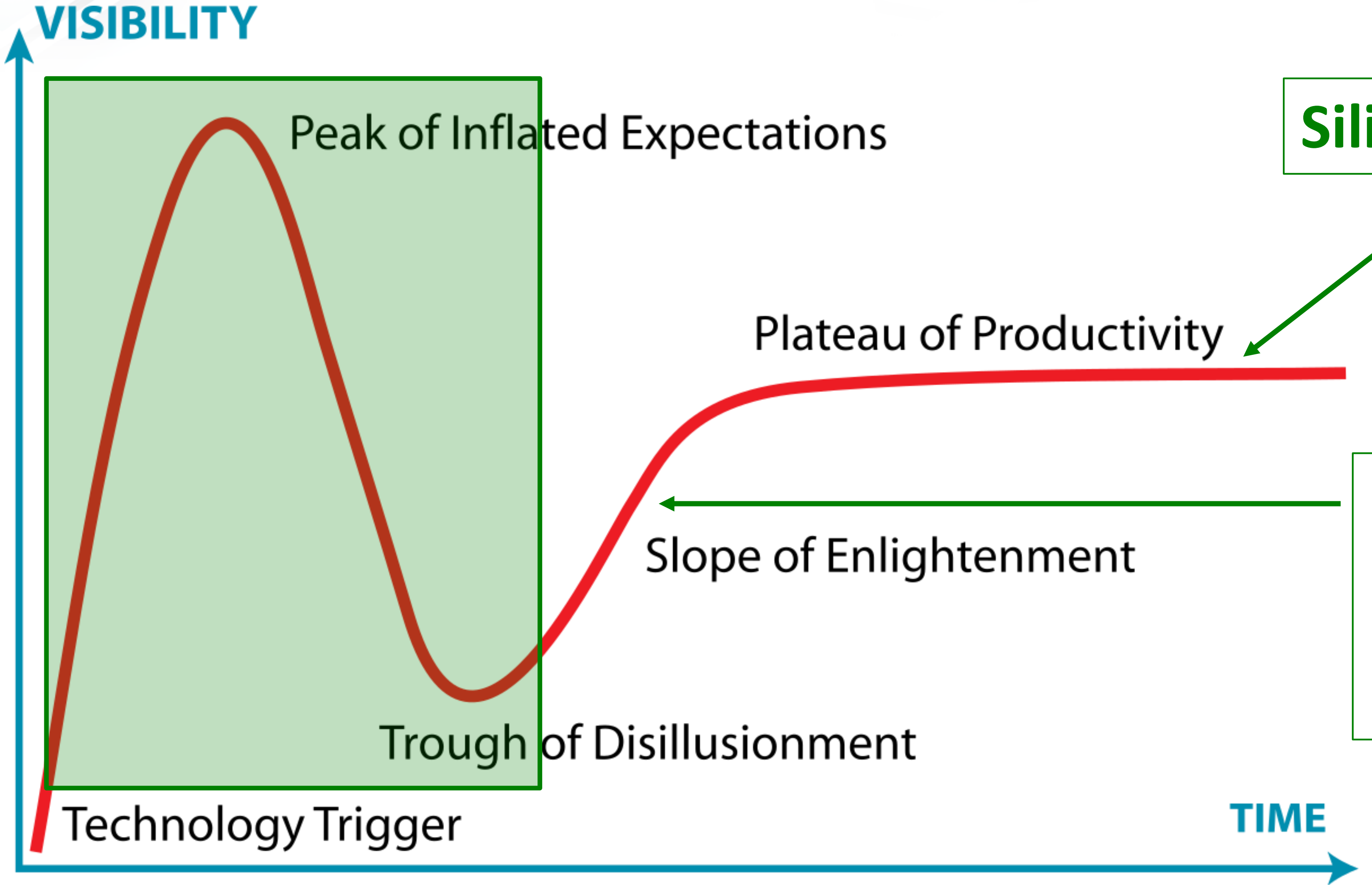
A 2016 analysis found that 20% of papers featuring gene names had errors created by spreadsheet autocorrect functions, but a bigger survey now finds the proportion is up to 30%. Since 2014, the number of papers with errors has increased significantly.



https://www.nature.com/articles/d41586-021-02211-4

How with AI/ML Affect Me and My Science?

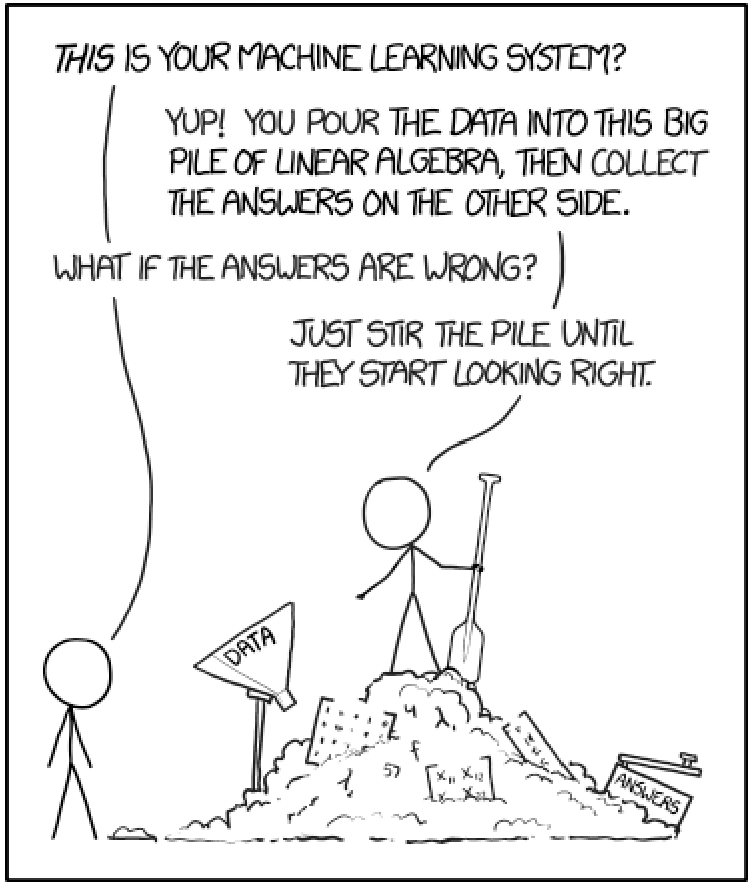
“People Who Use ML in Research”

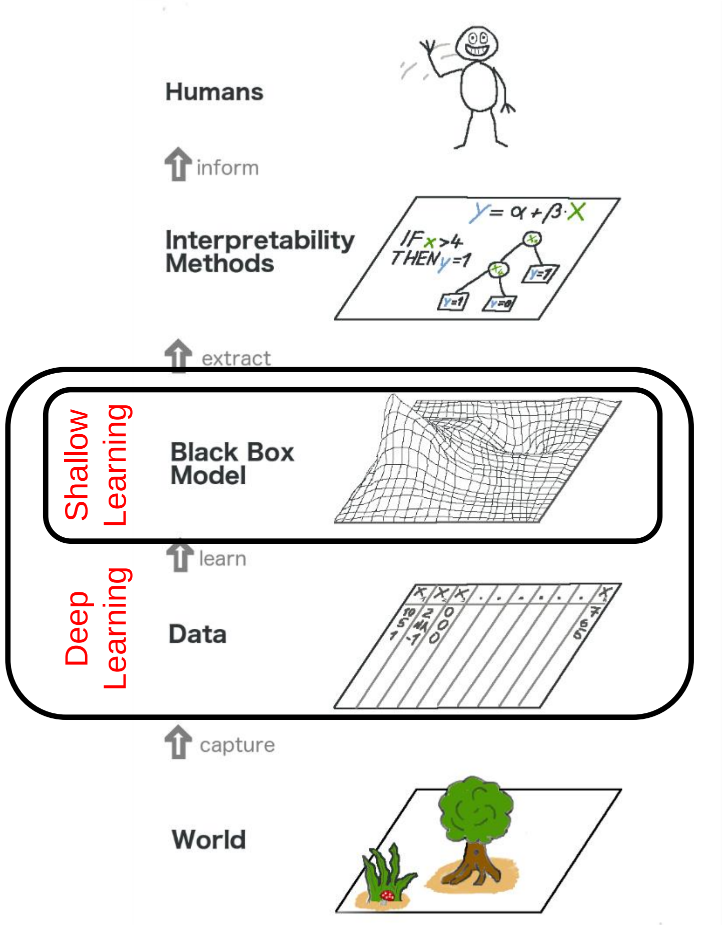


Silicon Valley

“People Who Do Research in ML”

The Problems with Blind Modeling

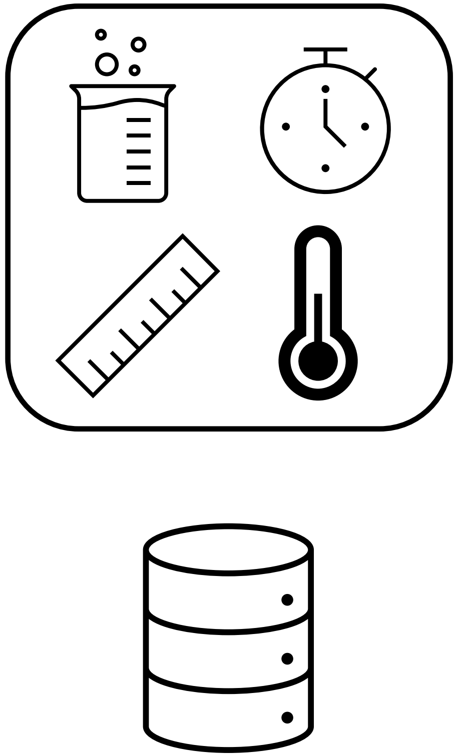

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS



https://christophm.github.io/interpretable-ml-book/agnostic.html

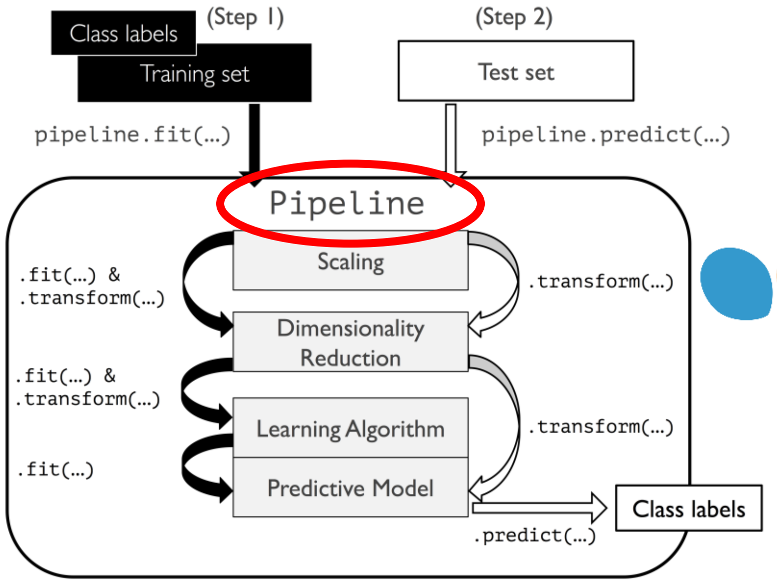
Our Approach

1. Collection



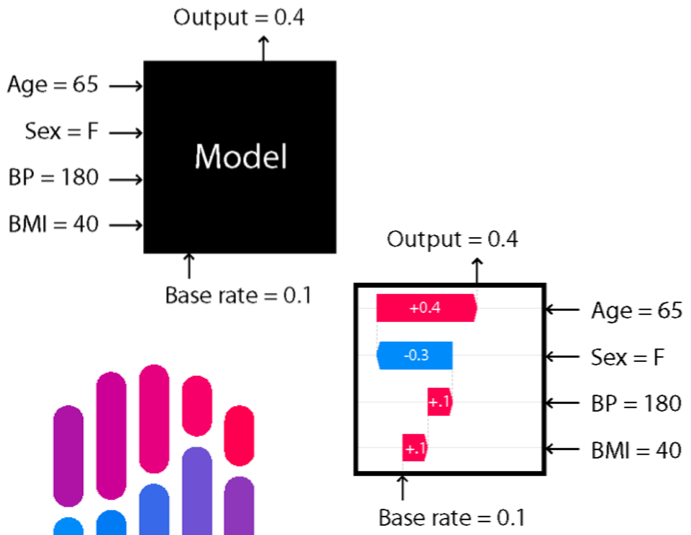
2. Analysis

AutoML to optimize architecture, (nested) CV for hyperparameters



Python Machine Learning 2nd Ed., Raschka & Mirjalili (2017).

3. (Feature-based) Explanations

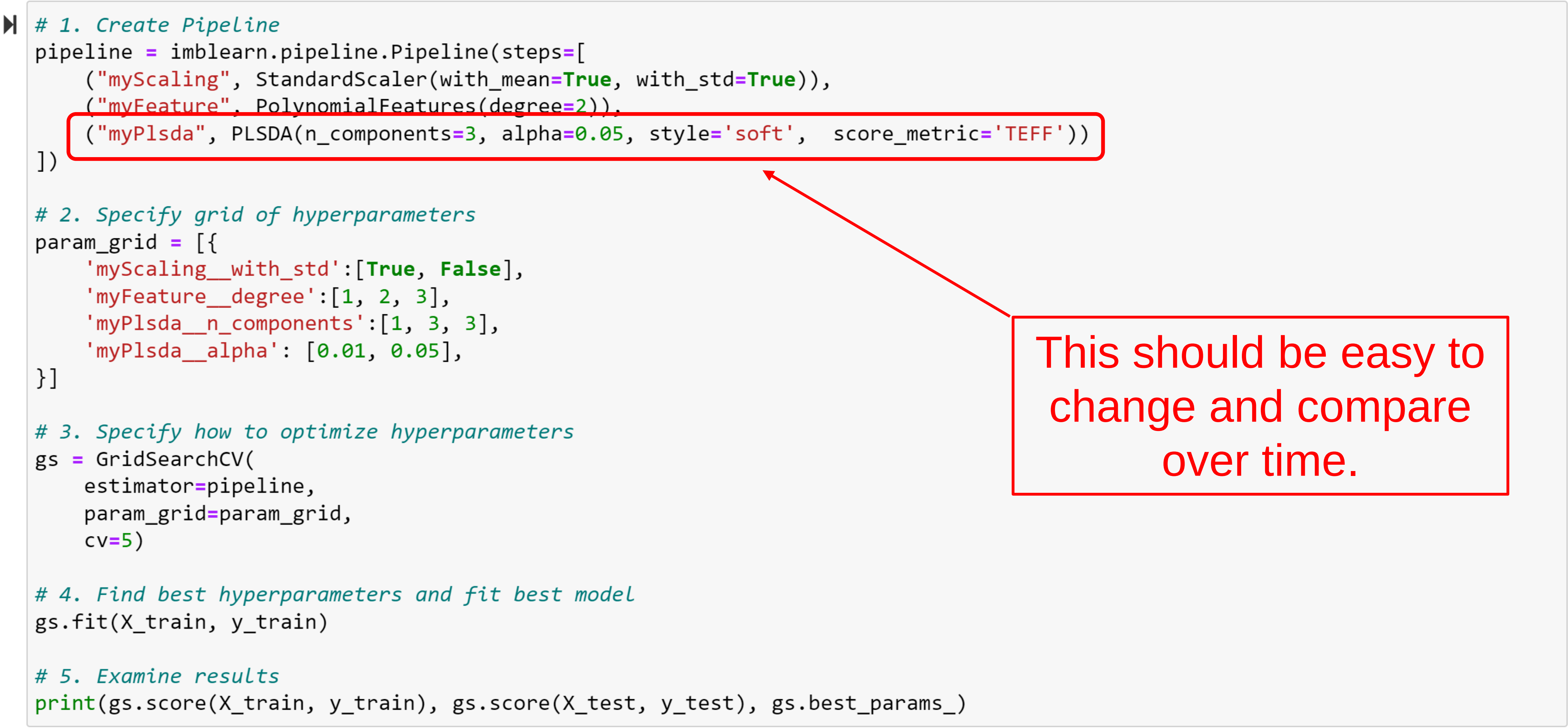


SHAP

Model Agnostic

https://shap.readthedocs.io

Tools Should Be Simple to Use



```
# 1. Create Pipeline
pipeline = imblearn.pipeline.Pipeline(steps=[
    ("myScaling", StandardScaler(with_mean=True, with_std=True)),
    ("myFeature", PolynomialFeatures(degree=2)),
    ("myPlsda", PLSDA(n_components=3, alpha=0.05, style='soft',  score_metric='TEFF'))
])

# 2. Specify grid of hyperparameters
param_grid = [{
    'myScaling__with_std':[True, False],
    'myFeature__degree':[1, 2, 3],
    'myPlsda__n_components':[1, 3, 3],
    'myPlsda__alpha': [0.01, 0.05],
}]

# 3. Specify how to optimize hyperparameters
gs = GridSearchCV(
    estimator=pipeline,
    param_grid=param_grid,
    cv=5)

# 4. Find best hyperparameters and fit best model
gs.fit(X_train, y_train)

# 5. Examine results
print(gs.score(X_train, y_train), gs.score(X_test, y_test), gs.best_params_)
```

This should be easy to change and compare over time.

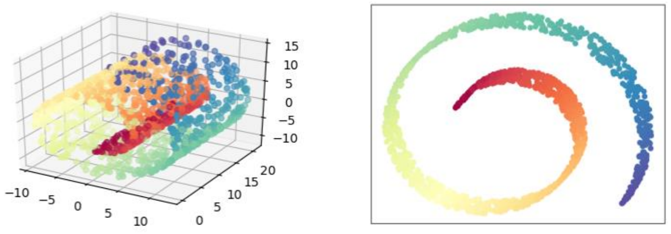
Coding to a Standard API

N > 10

Conventional Chemometrics

Often linear dimensionality reduction

"Scores" ⟶ $X = TP^T + E$



e.g., PCA, PCR, PLS(-DA), SIMCA

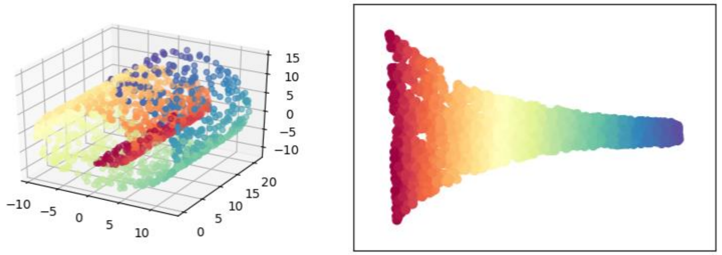
Only **global** properties considered

N > 100

Topological Methods

"Non-linear dimensionality reduction"

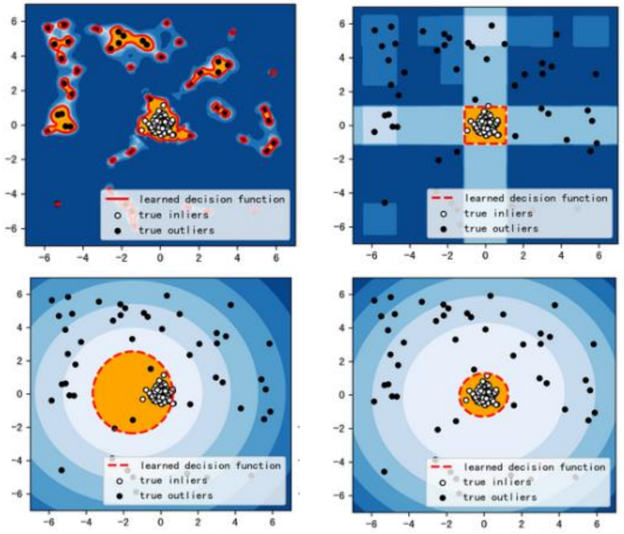
"Embedding" ⟶ $T = f(X)$



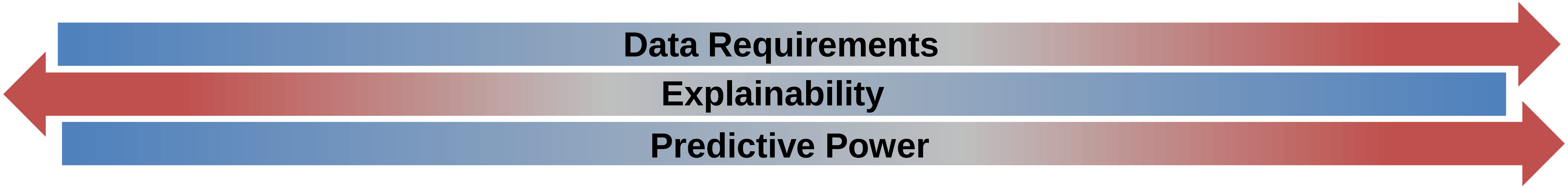
e.g., Isomap, LLE, t-SNE, UMAP, PaCMAP

Local properties now considered

"Machine Learning"/AI



e.g., VAE, Deep NN, pyOD



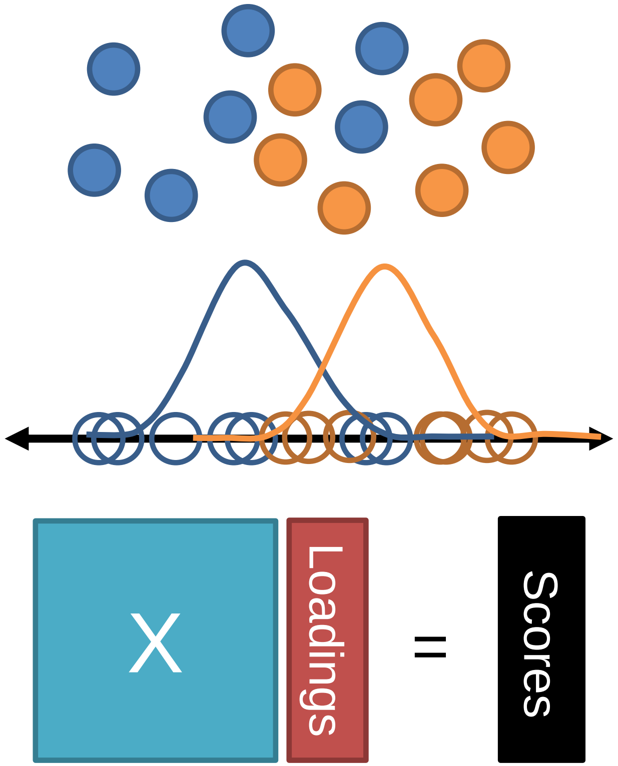
Data Requirements

Explainability

Predictive Power

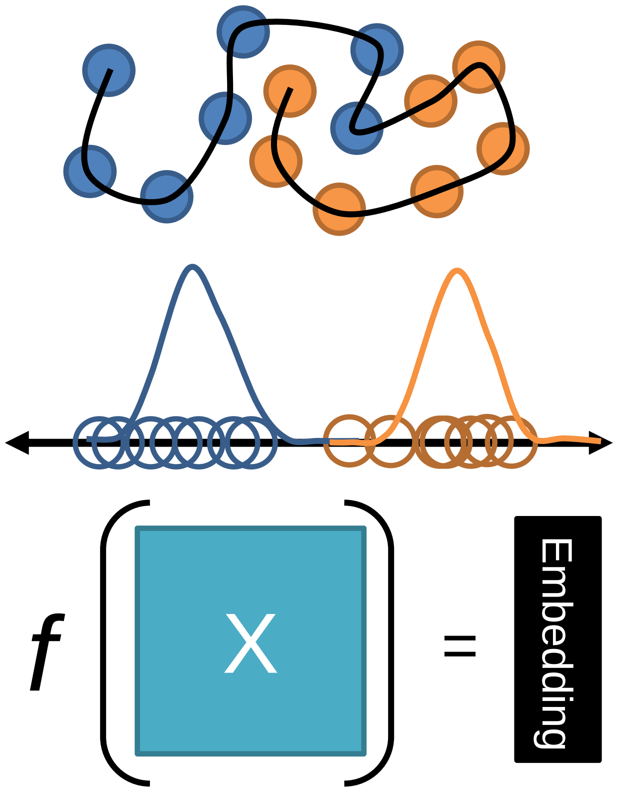
Meaningful Representations and Explanations

(Linear) Projection Methods



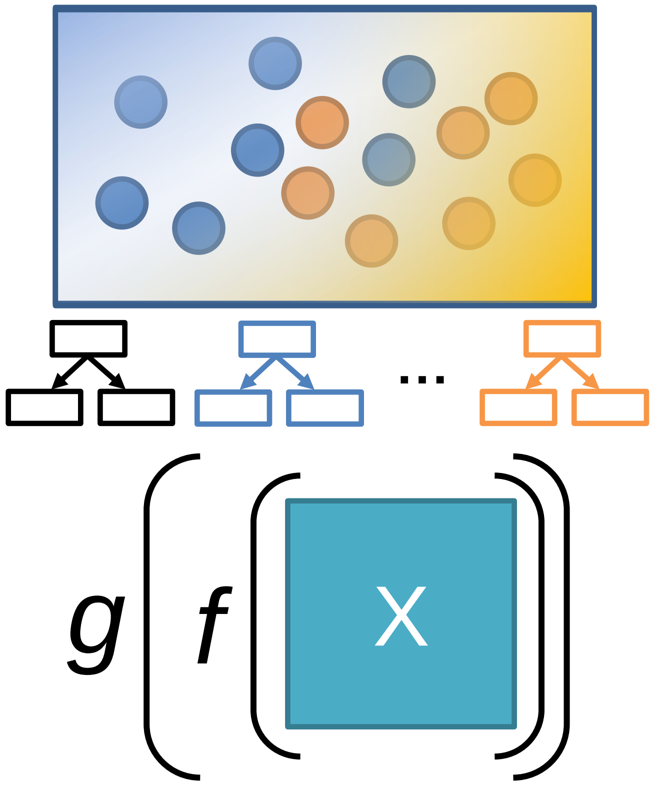
Statistically meaningful latent space
Easy to explain with loadings

Manifold Learning (Non-linear)



Statistically meaningful latent space
Use, e.g., SHAP to explain

Generic Machine Learning



Calibrate to get meaningful latent space
Use, e.g., SHAP to explain

A Multitude of Models and Explanations

A **“Rashomon set”** is an ensemble of almost equally high performing models.

- Can be **very different** black boxes with a different perspective or explanation of the same event or observation.

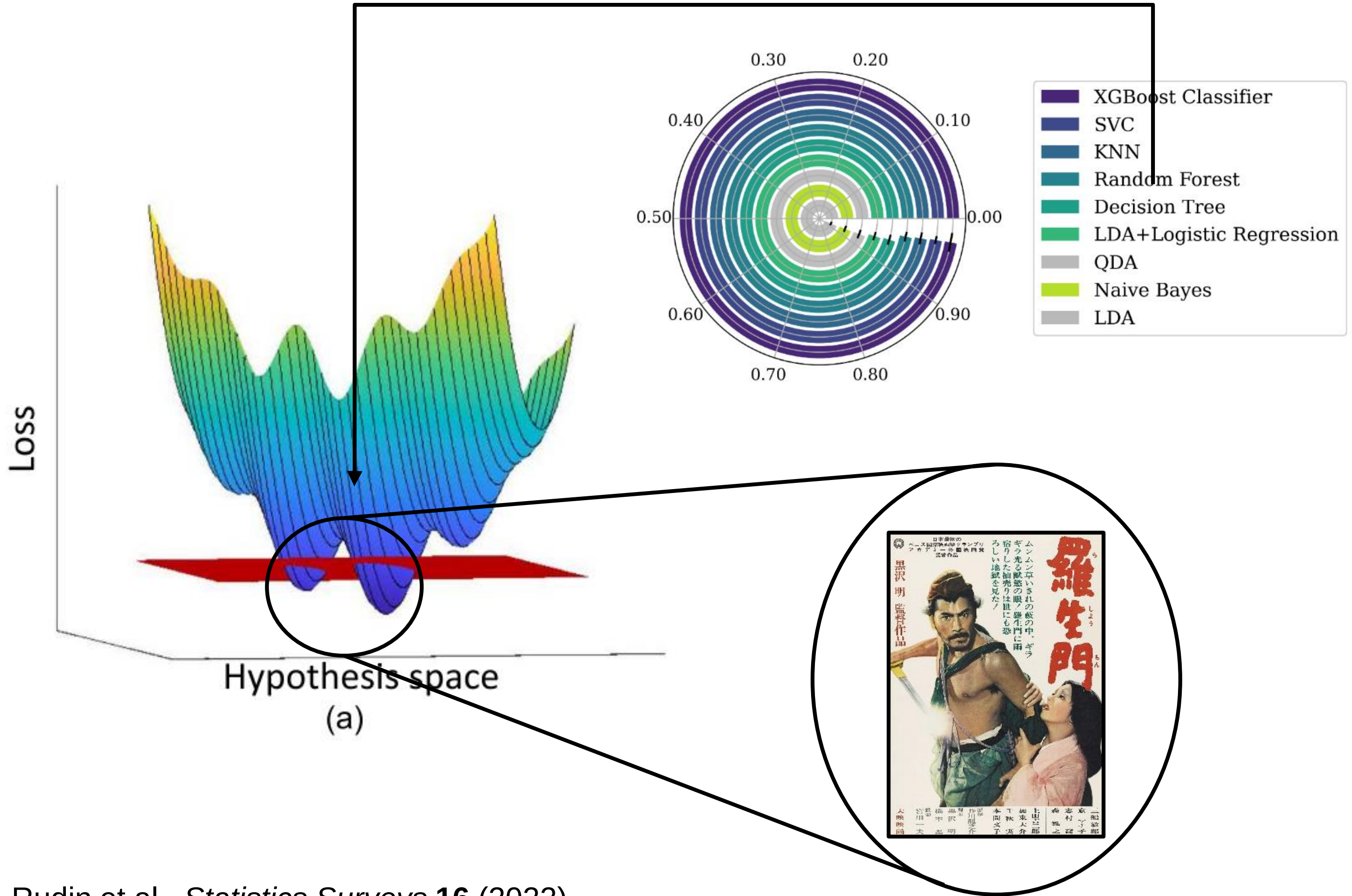
Which one(s), if any, is “correct”?

Large RS often appear when you have more information/measurements than you need.

- Large databases
- Correlated measurements

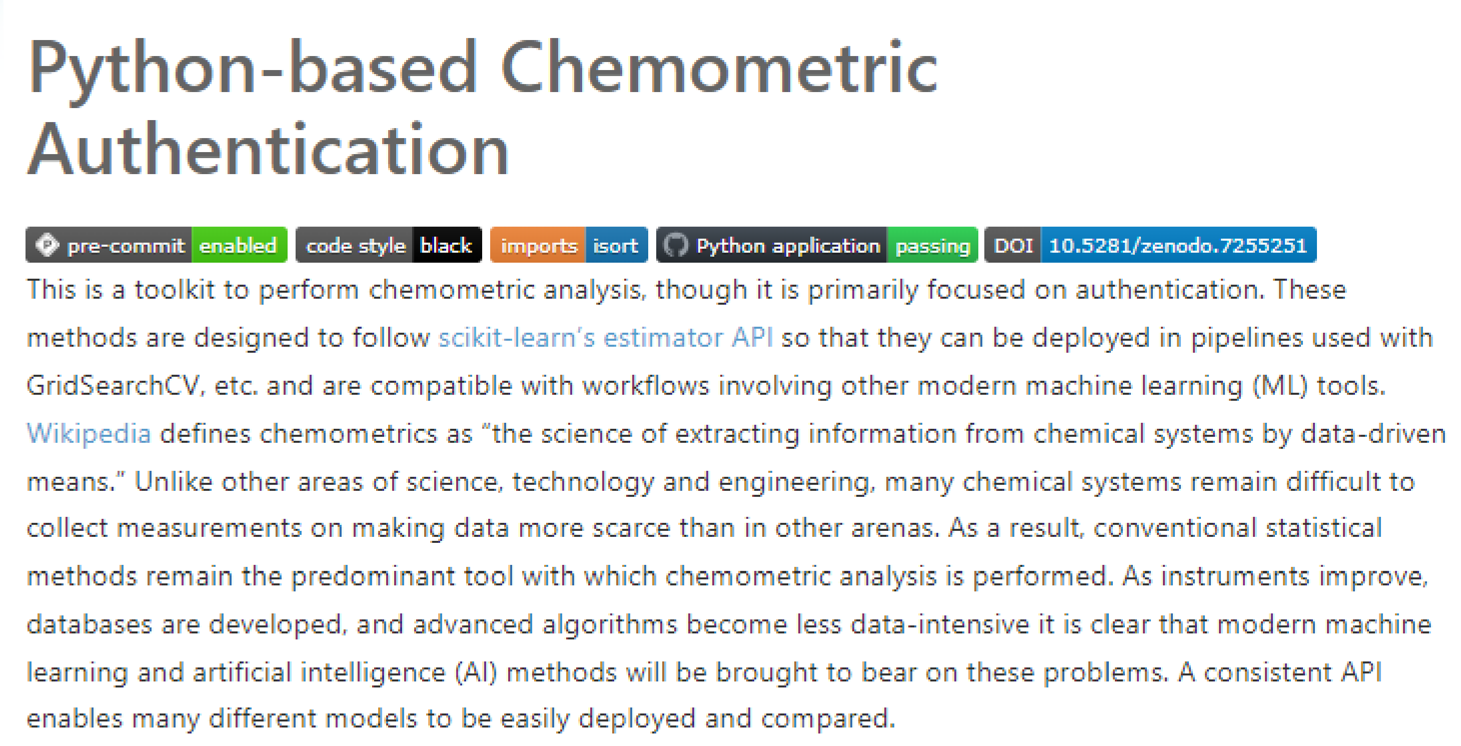
Under weak assumptions, a large RS must contain a simple (interpretable?) model.

Best models are ~97% accurate



Rudin et al., *Statistics Surveys* **16** (2022).

Our Community Resource in Development



Python-based Chemometric Authentication

pre-commit enabled | code style black | imports isort | Python application passing | DOI 10.5281/zenodo.7255251

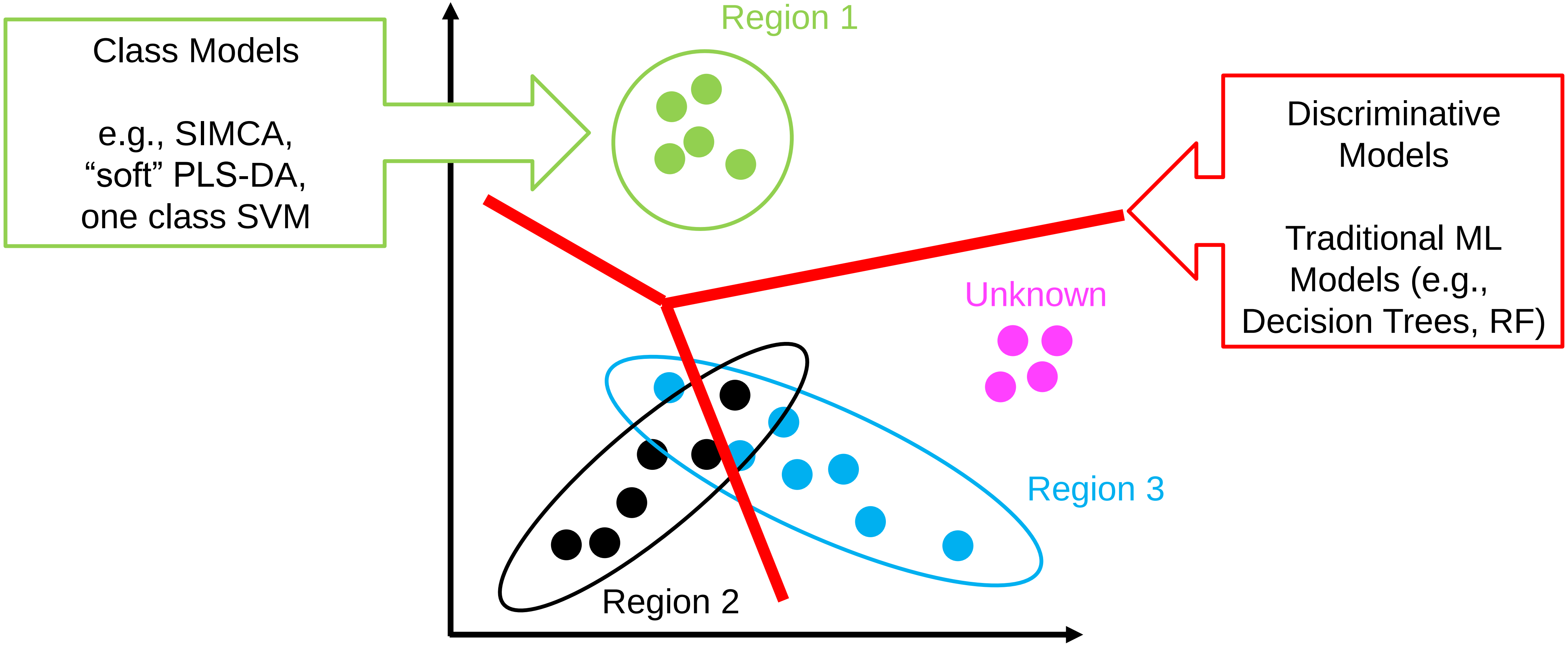
This is a toolkit to perform chemometric analysis, though it is primarily focused on authentication. These methods are designed to follow scikit-learn's estimator API so that they can be deployed in pipelines used with GridSearchCV, etc. and are compatible with workflows involving other modern machine learning (ML) tools. Wikipedia defines chemometrics as "the science of extracting information from chemical systems by data-driven means." Unlike other areas of science, technology and engineering, many chemical systems remain difficult to collect measurements on making data more scarce than in other arenas. As a result, conventional statistical methods remain the predominant tool with which chemometric analysis is performed. As instruments improve, databases are developed, and advanced algorithms become less data-intensive it is clear that modern machine learning and artificial intelligence (AI) methods will be brought to bear on these problems. A consistent API enables many different models to be easily deployed and compared.

https://pychemauth.readthedocs.io/en/latest/

In 4th beta release

https://colab.research.google.com/

Fundamentally Different Types of Models



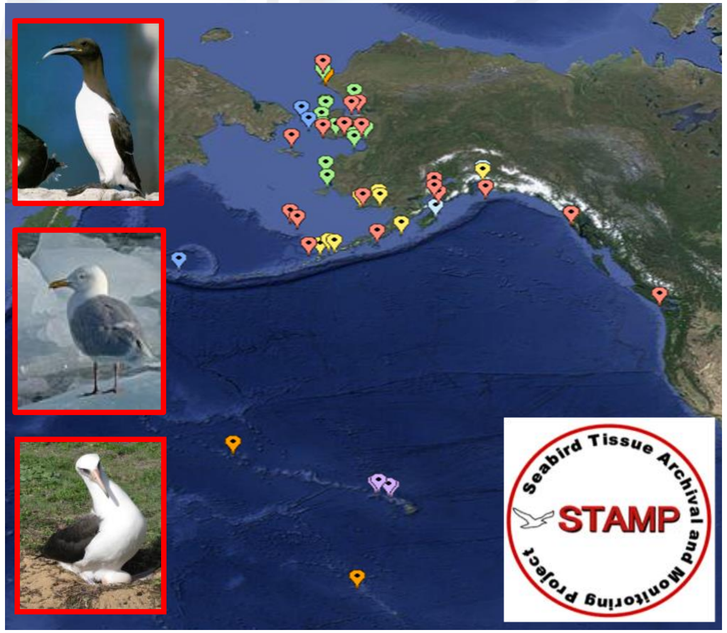
Class Models

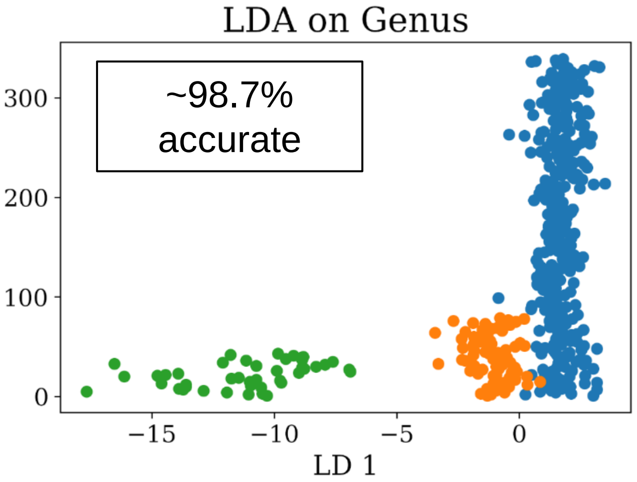
e.g., SIMCA, “soft” PLS-DA, one class SVM

Discriminative Models

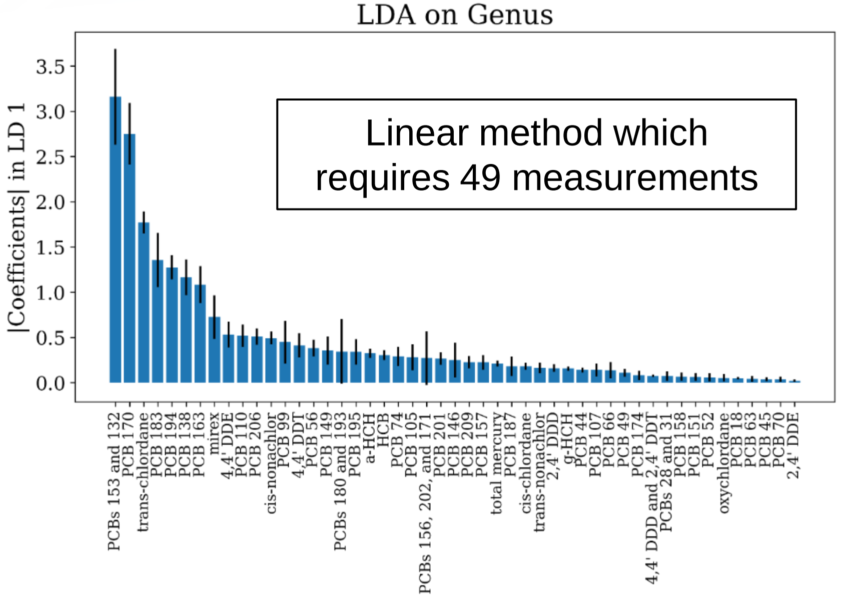
Traditional ML Models (e.g., Decision Trees, RF)

Interpretable Models of Pacific Seabirds

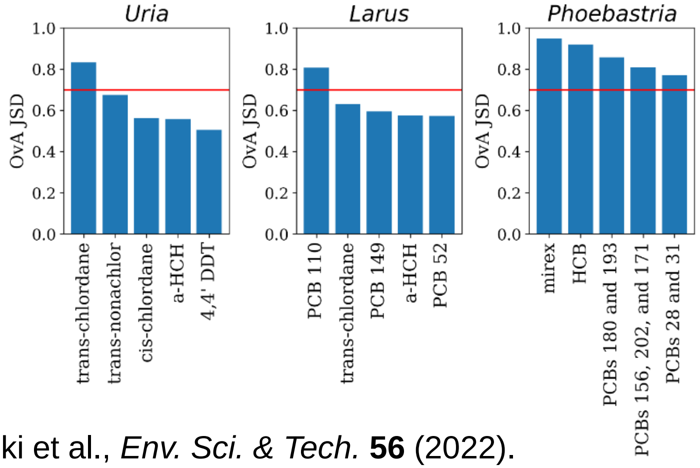




~98.7% accurate

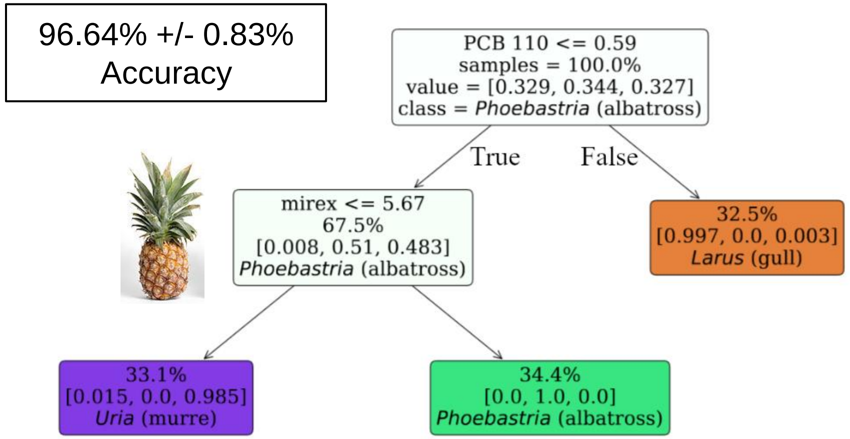


Linear method which requires 49 measurements



Mahynski et al., *Env. Sci. & Tech.* **56** (2022).

96.64% +/- 0.83% Accuracy



PCB 110 <= 0.59
samples = 100.0%
value = [0.329, 0.344, 0.327]
class = *Phoebastria* (albatross)

True

False

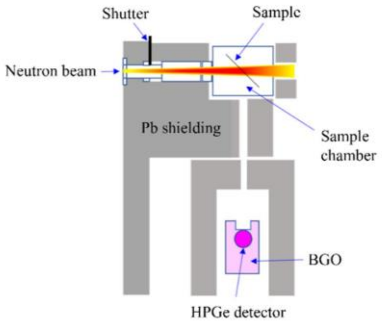
mirex <= 5.67
67.5%
[0.008, 0.51, 0.483]
Phoebastria (albatross)

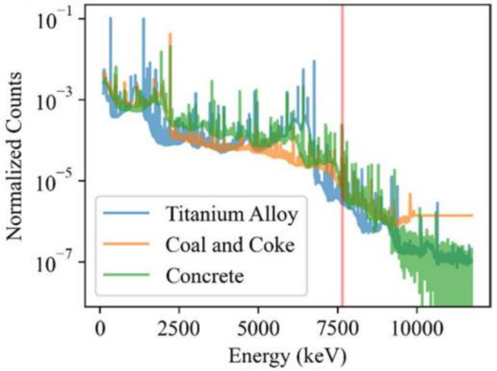
32.5%
[0.997, 0.0, 0.003]
Larus (gull)

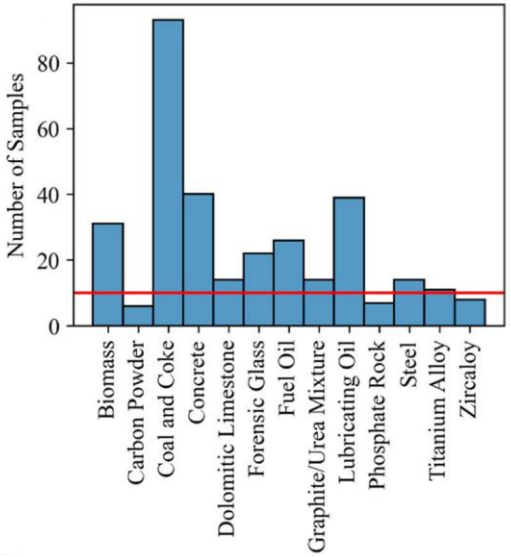
33.1%
[0.015, 0.0, 0.985]
Uria (murre)

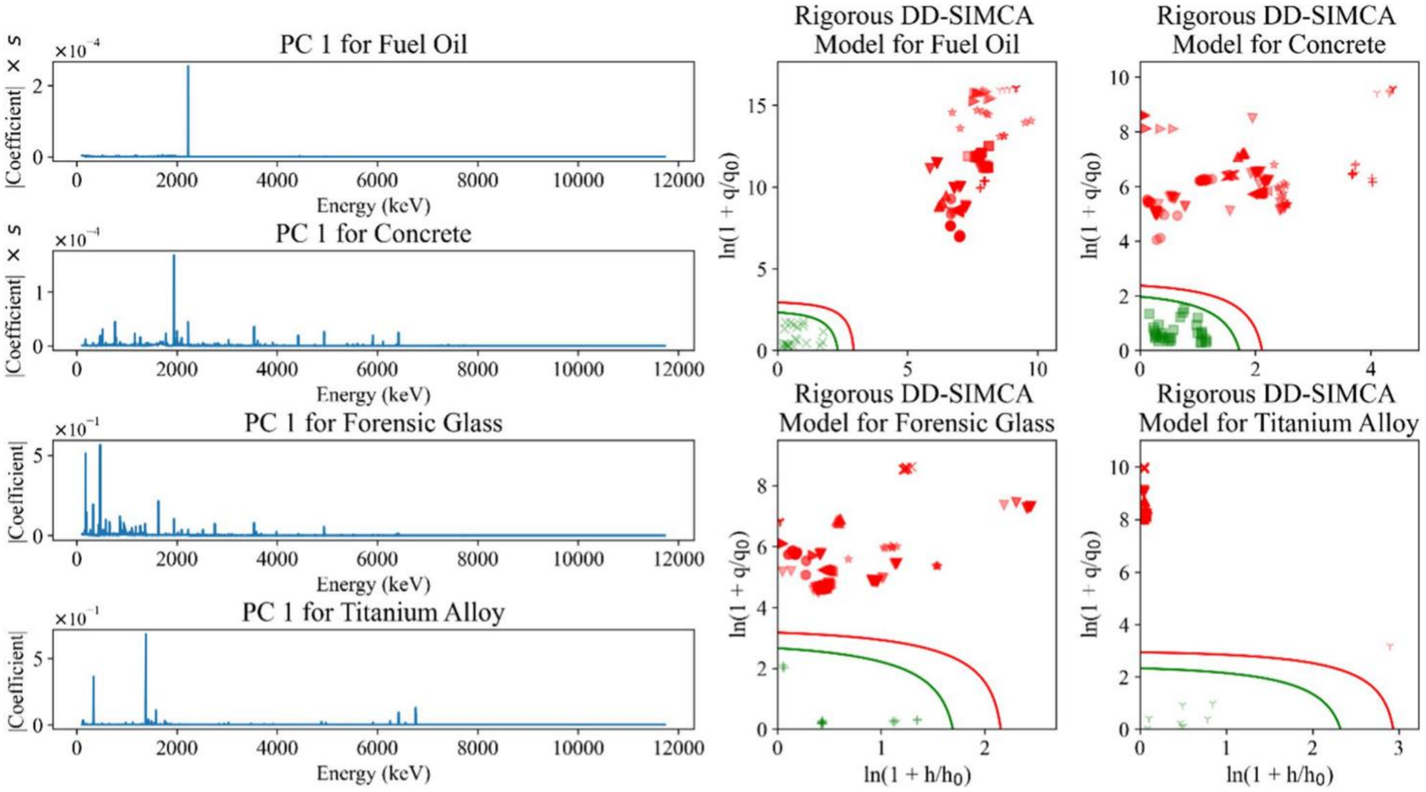
34.4%
[0.0, 1.0, 0.0]
Phoebastria (albatross)

Material Authentication using PGAA





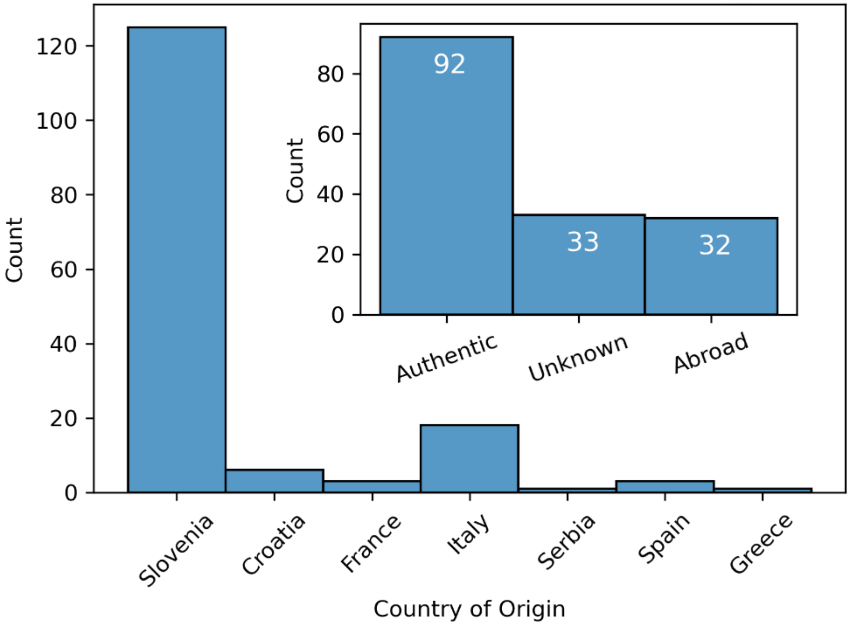


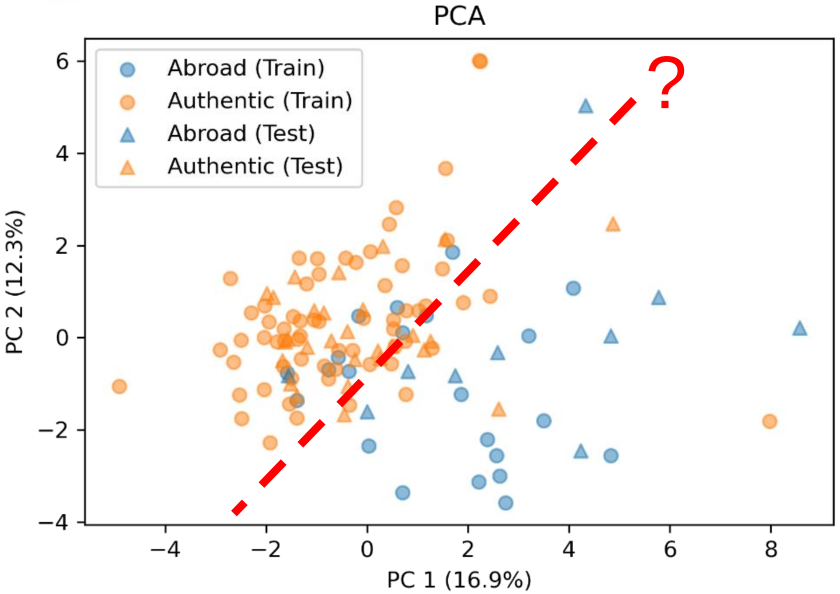


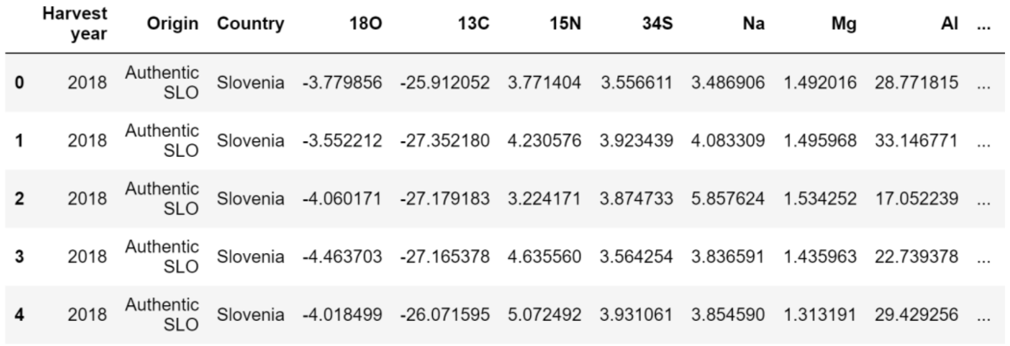
Mahynski et al., *J. Radioanl and Nucl. Chem.* ***332*** (2023).

Determining the Authenticity of Slovenian Strawberries

Data courtesy of Prof. Nives Ogrinc and Ms. Lidija Strojnik



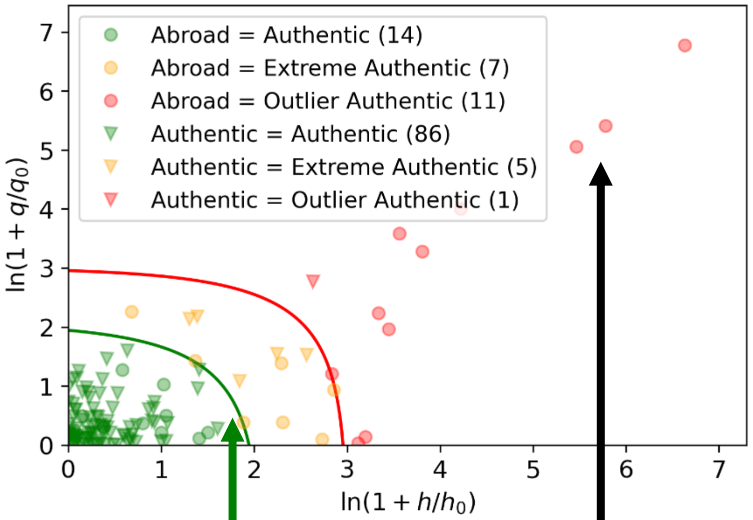




	Harvest year	Origin	Country	18O	13C	15N	34S	Na	Mg	Al	...
0	2018	Authentic SLO	Slovenia	-3.779856	-25.912052	3.771404	3.556611	3.486906	1.492016	28.771815	...
1	2018	Authentic SLO	Slovenia	-3.552212	-27.352180	4.230576	3.923439	4.083309	1.495968	33.146771	...
2	2018	Authentic SLO	Slovenia	-4.060171	-27.179183	3.224171	3.874733	5.857624	1.534252	17.052239	...
3	2018	Authentic SLO	Slovenia	-4.463703	-27.165378	4.635560	3.564254	3.836591	1.435963	22.739378	...
4	2018	Authentic SLO	Slovenia	-4.018499	-26.071595	5.072492	3.931061	3.854590	1.313191	29.429256	...

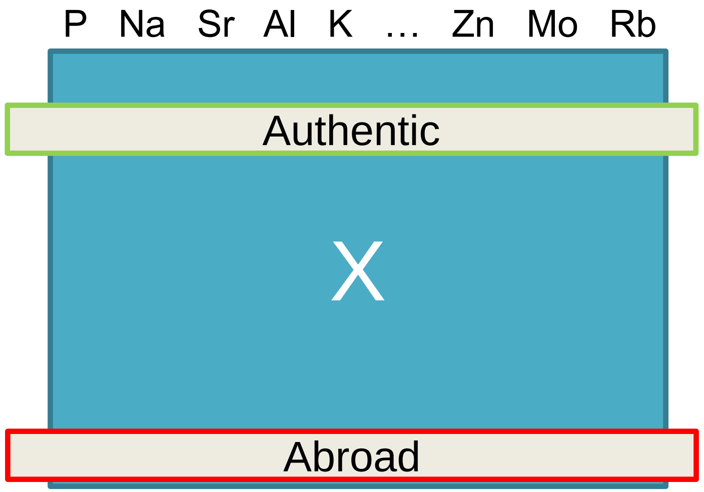
4 Stable Isotope Ratios (C, N, O, S)
19 Trace Elements (> LOD 80% of samples)
70/30 train/test split of the data

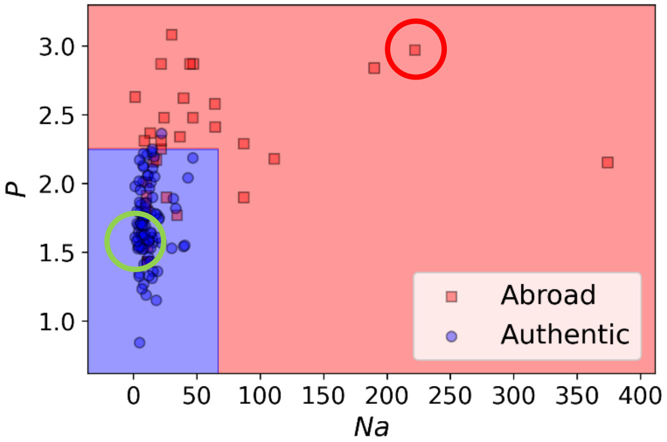
Which Model is More Useful?



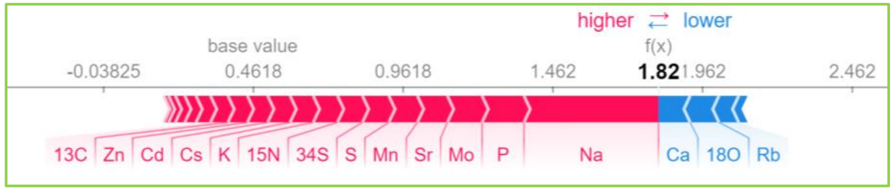
$$d^2 = N_h \frac{h}{h_0} + N_q \frac{q}{q_0}$$

$$d_{crit}^2 = \chi^{-2}(1 - \alpha, N_h + N_q)$$

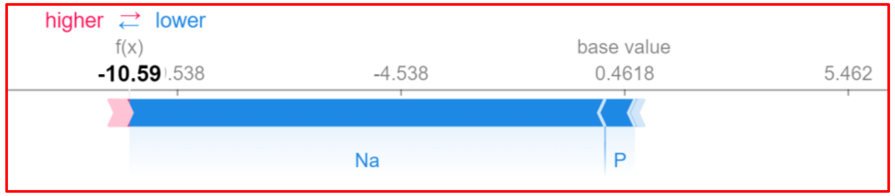




$$f = d_{crit} - d$$



1.54 3.83



222 2.97

```
{'simca__alpha': 0.05,
 'simca__n_components': 3,
 'simca__scale_x': True,
 'simca__style': 'dd-simca'}
```

Pomerantsev & Rodionova, *J. Chemom.* **28** (2014).
Lundberg & Lee, *NIPS* (2017).

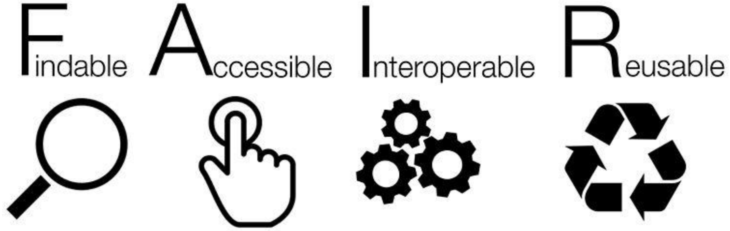
Looking into the Future

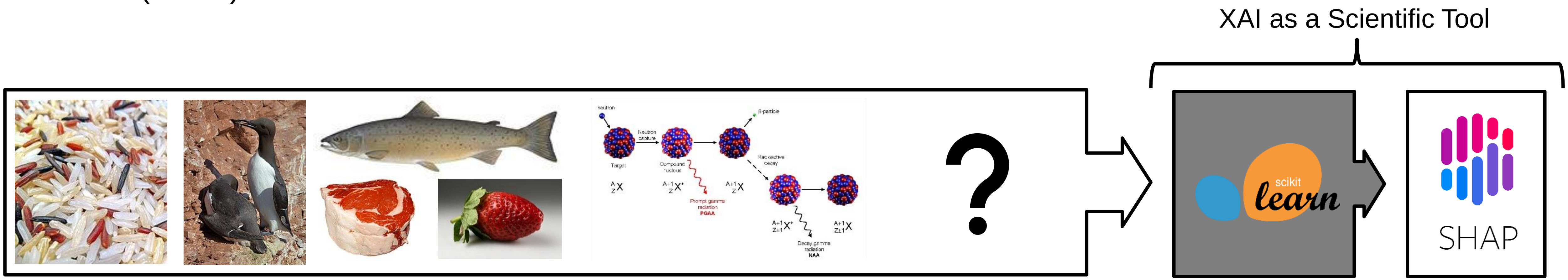
Standardized APIs enable many pipelines or models (“black boxes”) to be easily compared.

- Enables continuous improvement of models and pipelines
- Ensures long-term interoperability as new models and techniques are developed
- Enables best-practices to be routinely evaluated
- Relies on continuous development of FAIR data(bases)

Chemical Informatics Group @NIST

https://www.nist.gov/mml/csd/chemical-informatics-group




XAI as a Scientific Tool