Interpretability in Data-Centric ML

Alexandra (Ola) Zytek

Introduction to Data-Centric AI

IAP 2023

Massachusetts Institute of Technology

DATA TO AI

experimentation, we found out that security analysts highly value *easily interpretable features* when analyzing outputs of machine learning algorithms.

good accuracy as well as recall and that use human-understandable features. Our findings indicate that moderators would appreciate the ability to understand outputs based on such features. As

MGM has two core elements which perform interpretable feature extraction and selection. At the

of the process, even before an actual model is developed. For example, P11, referring to feature engineering, remarked: "*... this is the first step toward making interpretable models, even though we don't have any model yet.*". In particular, we found several data scientists complement feature

Roadmap

- Introduction to interpretable ML
- **Why** do we care about interpretable features?
- **What** are interpretable features ***really***?
- **How** do we get interpretable features?

- **Introduction to interpretable ML**
- **Why** do we care about interpretable features?
- **What** are interpretable features ***really***?
- **How** do we get interpretable features?

Location: Denfield

Population: 120

Median Income: $32,000

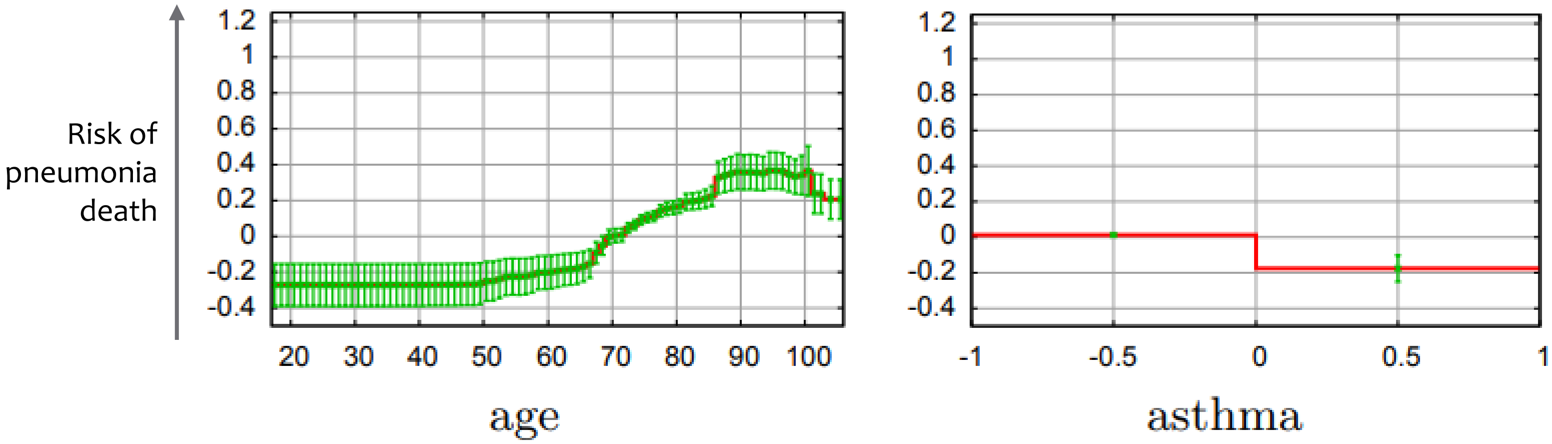
?

Median House Price: $198,000

Why do we need interpretable ML?

1. Debugging and validation
2. Reviewing decisions
3. Improving usability

Debugging and Validation



Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730

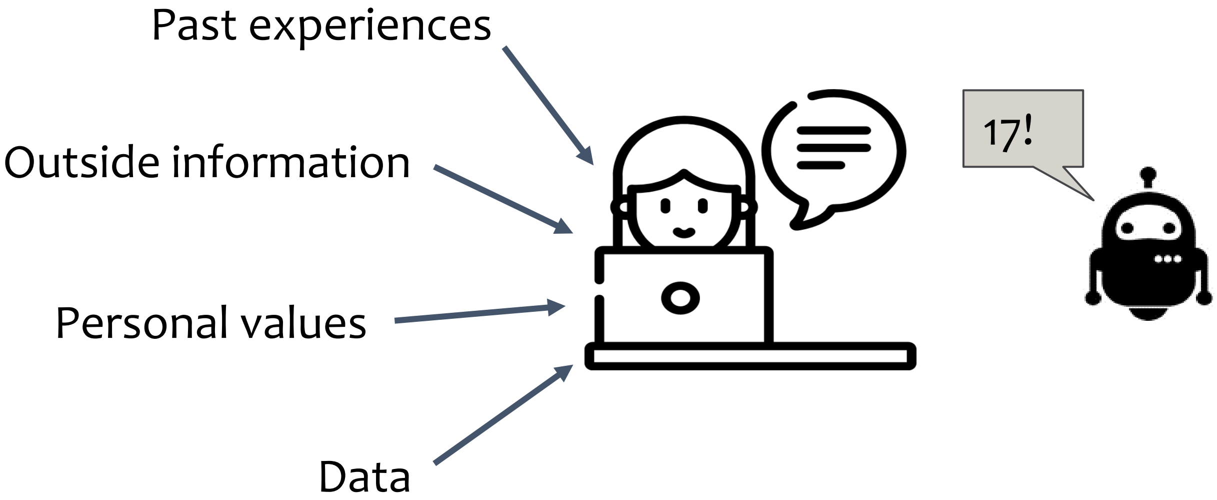
Reviewing Decisions

What self-driving cars can't recognize may be a matter of life and death

Engineers are racing to program artificial intelligence to recognize different scenarios that human drivers know inherently

Siddiqui, F. (2019, November 27). What self-driving cars can't recognize may be a matter of life and death. *Washington Post*.

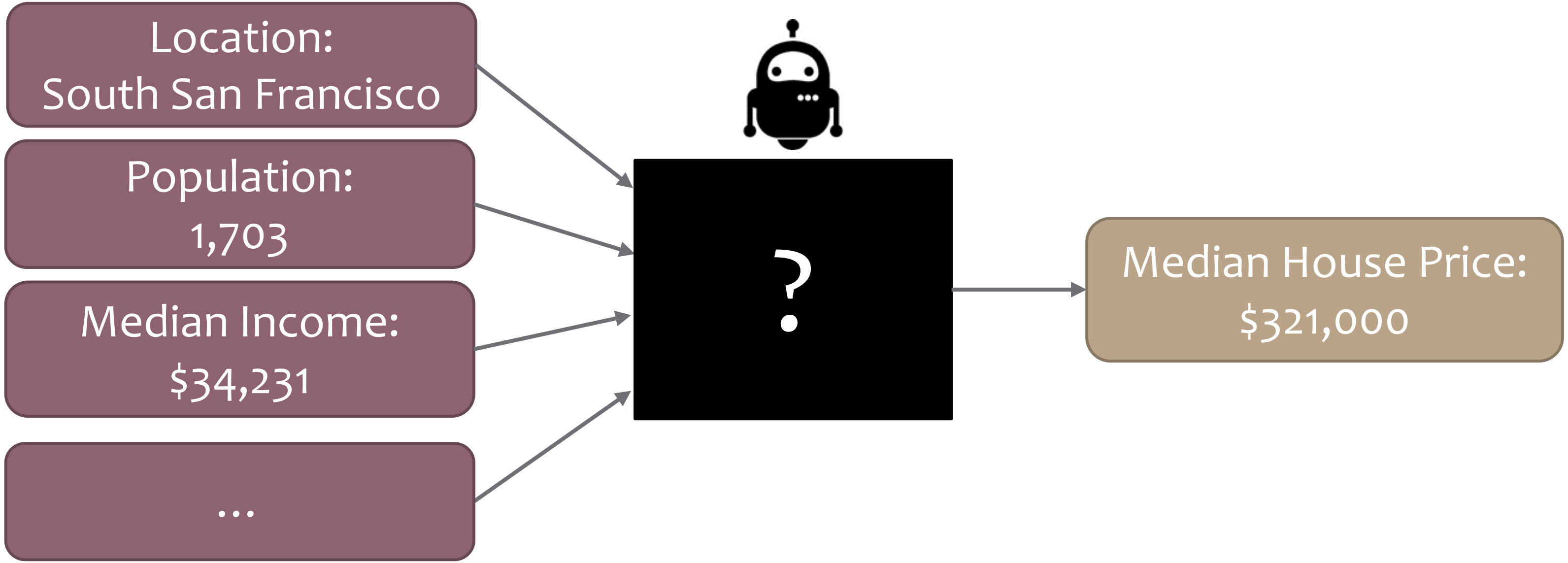
Improving Usability

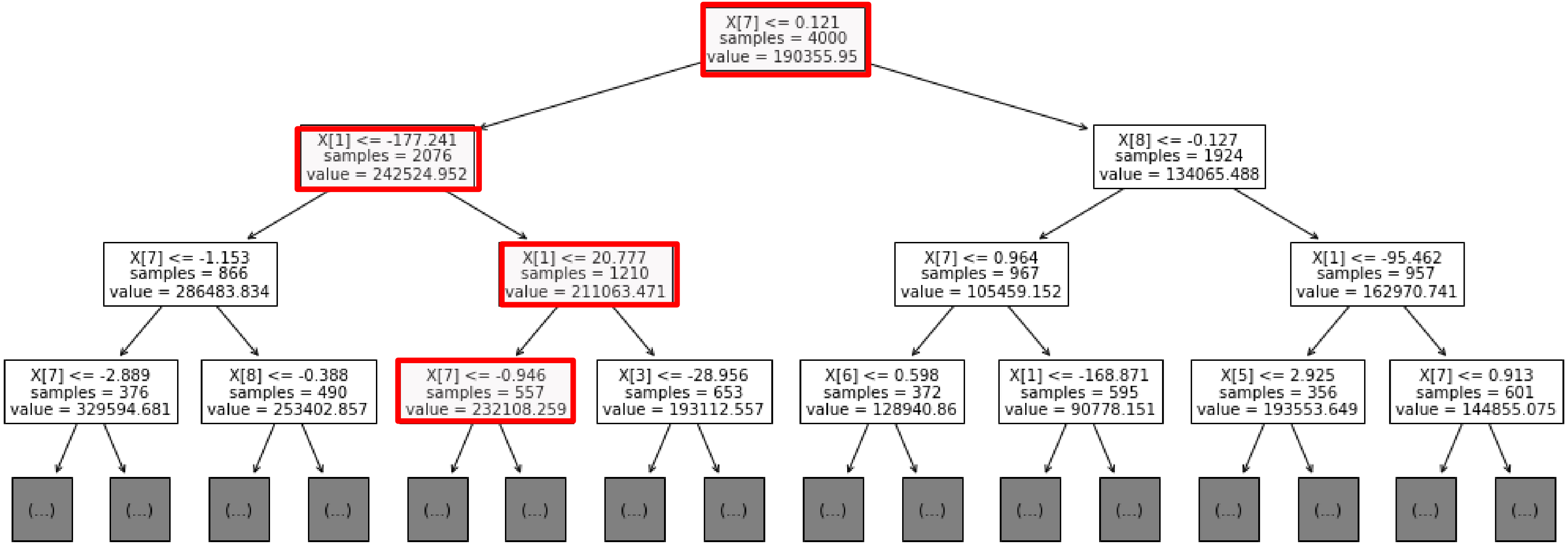


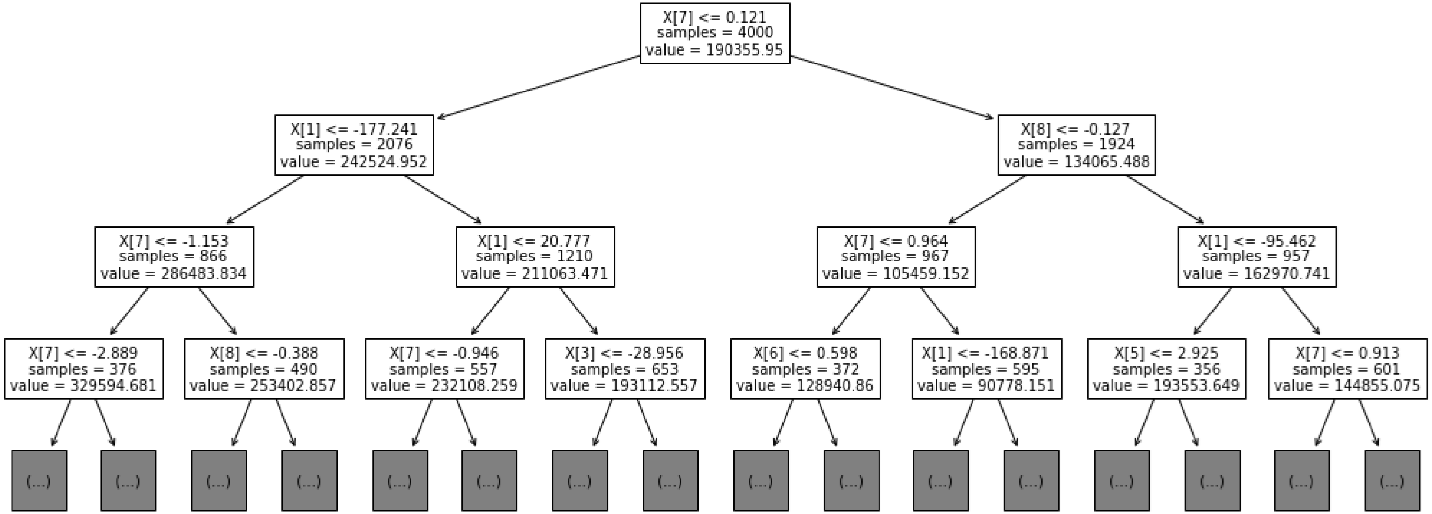
We need interpretable ML…

- When the problem formulation is **incomplete**
- When there is associated **risk**
- When **humans are involved** in decision-making

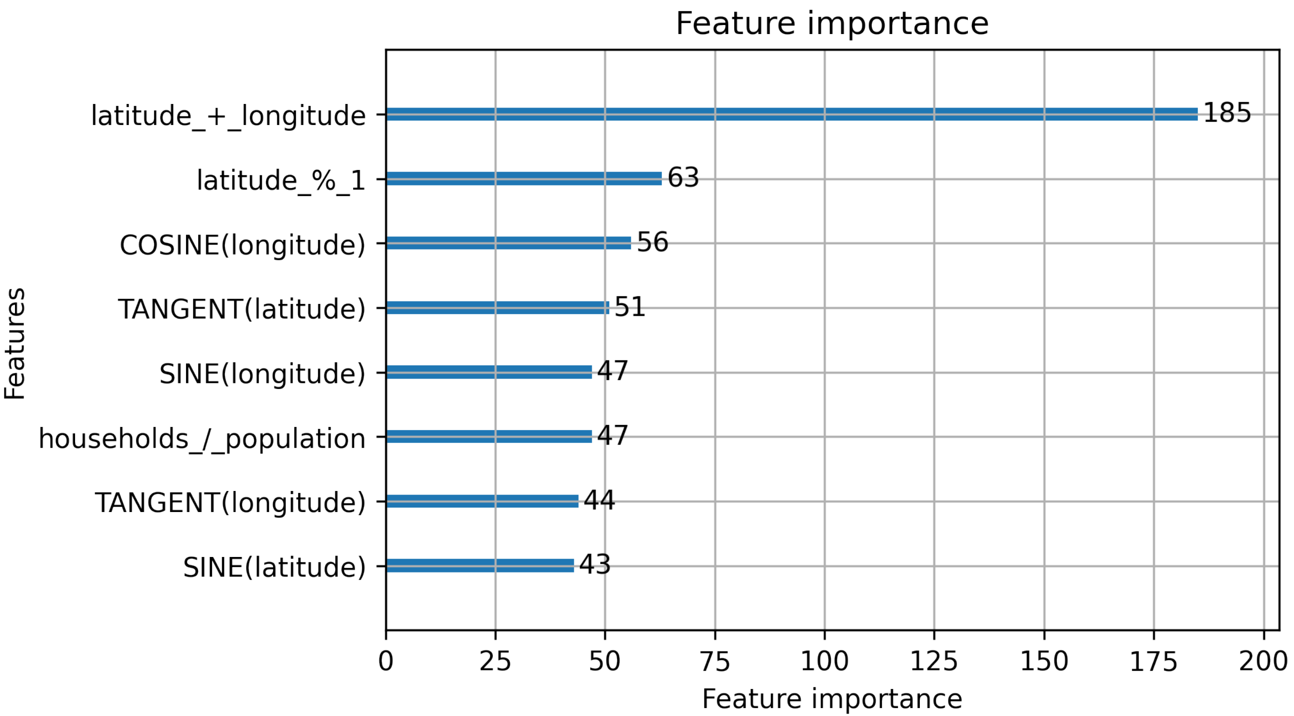
- Introduction to interpretable ML
- **Why do we care about interpretable features?**
- **What** are interpretable features ***really***?
- **How** do we get interpretable features?

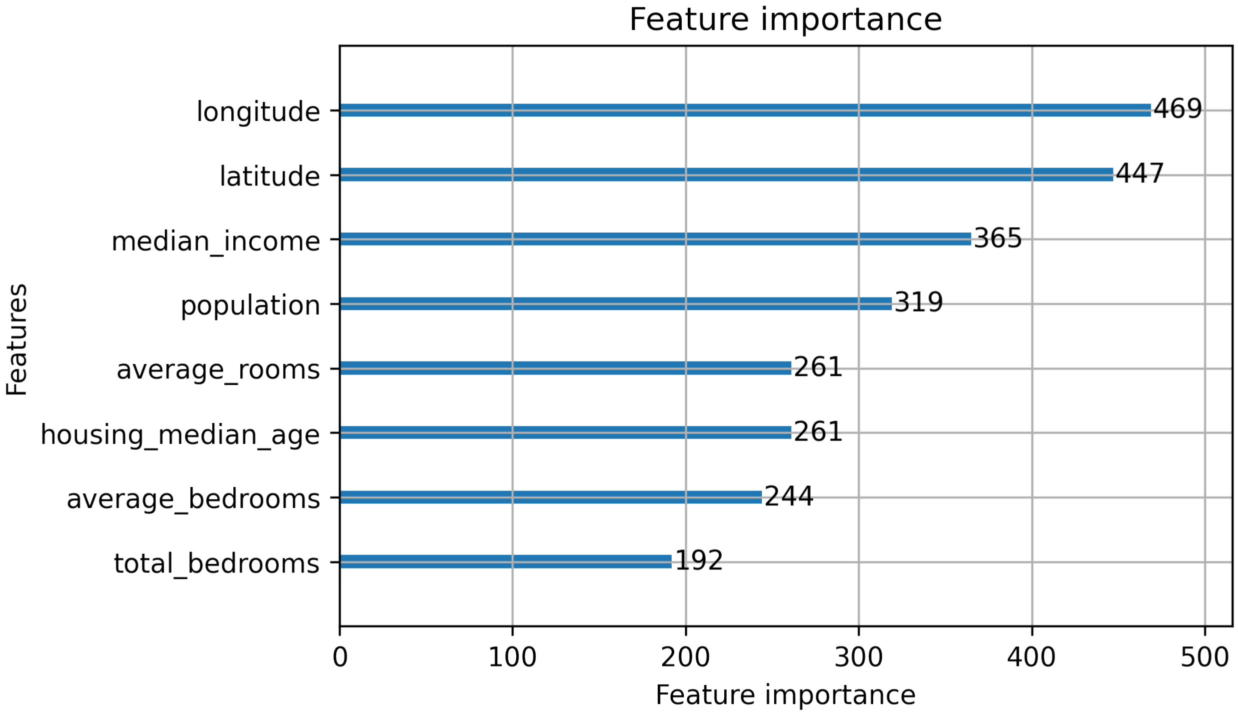

Location:
South San Francisco
Population:
1,703
Median Income:
$34,231
...
?
Median House Price:
$321,000


X[7] <= 0.121
samples = 4000
value = 190355.95
X[1] <= -177.241
samples = 2076
value = 242524.952
X[8] <= -0.127
samples = 1924
value = 134065.488
X[7] <= -1.153
samples = 866
value = 286483.834
X[1] <= 20.777
samples = 1210
value = 211063.471
X[7] <= 0.964
samples = 967
value = 105459.152
X[1] <= -95.462
samples = 957
value = 162970.741
X[7] <= -2.889
samples = 376
value = 329594.681
X[8] <= -0.388
samples = 490
value = 253402.857
X[7] <= -0.946
samples = 557
value = 232108.259
X[3] <= -28.956
samples = 653
value = 193112.557
X[6] <= 0.598
samples = 372
value = 128940.86
X[1] <= -168.871
samples = 595
value = 90778.151
X[5] <= 2.925
samples = 356
value = 193553.649
X[7] <= 0.913
samples = 601
value = 144855.075
(...)

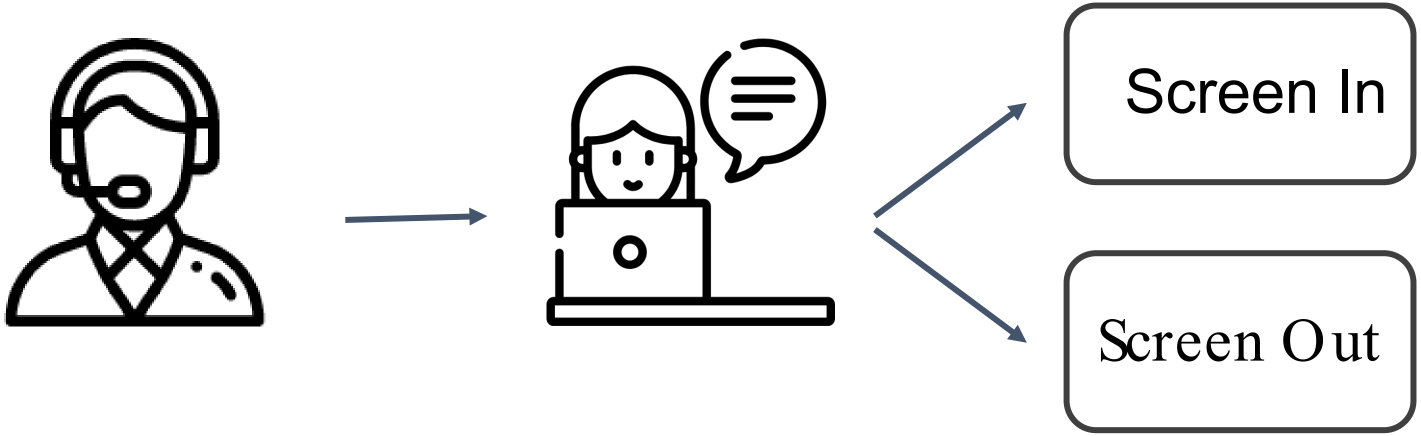


X[7] <= 0.121
samples = 4000
value = 190355.95
X[1] <= -177.241
samples = 2076
value = 242524.952
X[8] <= -0.127
samples = 1924
value = 134065.488
X[7] <= -1.153
samples = 866
value = 286483.834
X[1] <= 20.777
samples = 1210
value = 211063.471
X[7] <= 0.964
samples = 967
value = 105459.152
X[1] <= -95.462
samples = 957
value = 162970.741
X[7] <= -2.889
samples = 376
value = 329594.681
X[8] <= -0.388
samples = 490
value = 253402.857
X[7] <= -0.946
samples = 557
value = 232108.259
X[3] <= -28.956
samples = 653
value = 193112.557
X[6] <= 0.598
samples = 372
value = 128940.86
X[1] <= -168.871
samples = 595
value = 90778.151
X[5] <= 2.925
samples = 356
value = 193553.649
X[7] <= 0.913
samples = 601
value = 144855.075
(...)


Feature importance
latitude_+_longitude
latitude_%_1
COSINE(longitude)
TANGENT(latitude)
SINE(longitude)
households_/_population
TANGENT(longitude)
SINE(latitude)
185
63
56
51
47
47
44
43
Features
0
25
50
75
100
125
150
175
200
Feature importance

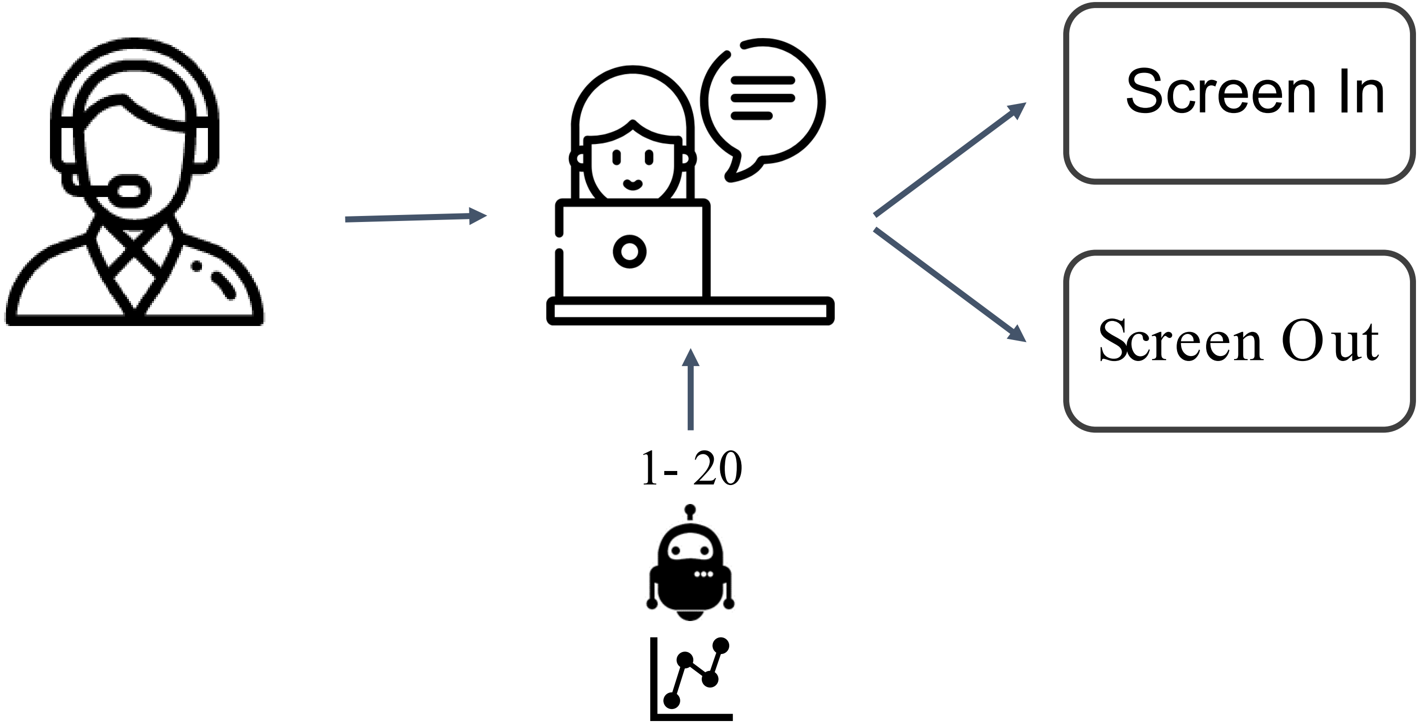

Feature importance
longitude
latitude
median_income
population
average_rooms
housing_median_age
average_bedrooms
total_bedrooms
469
447
365
319
261
261
244
192
0
100
200
300
400
500
Feature importance
Features

Case Study: Child Welfare



Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. https://doi.org/10.1109/TVCG.2021.3114864

Case Study: Child Welfare



Search feature | Category | All Contributions | Click “Show All Factors“ to enable Search and Filter

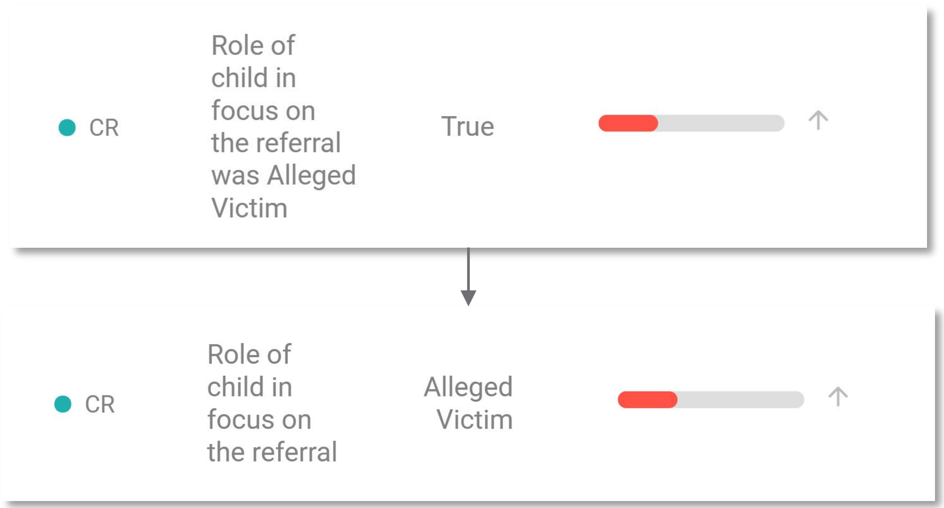
Category	Factor	Value	Contribution
DG	Age range of child in focus	<1 year	
DG	Age of the child in focus at time of referral	0	
RO	Number of other children (non victims) on the referral	0	

Problem: Confusing Features

●	PH	Count of days the child in focus was in a child welfare placement in the last 365 days	1
●	PH	Count of days the child in focus was in a child welfare placement in the last 730 days	1

Problem: Confusing Language

"The 'true' and 'false' is hard to interpret… Would rather have a positive statement (e.g., no perpetrator named)" –Child Welfare Screener

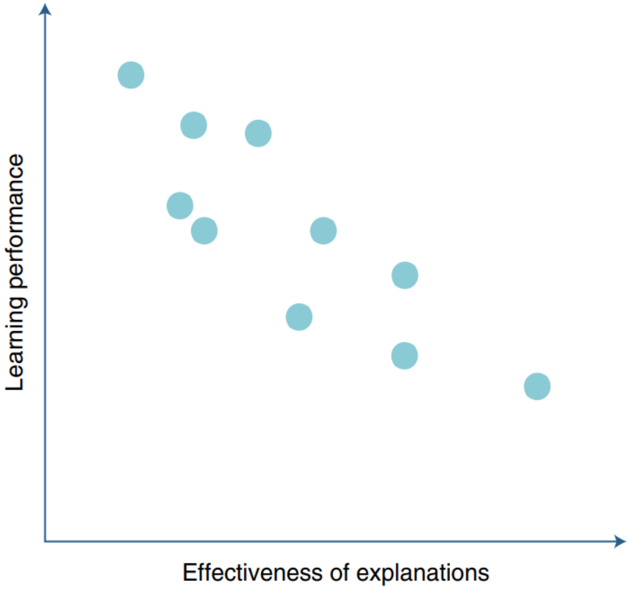


Problem: Irrelevant Features

“2 parents have missing date-of-birth is shown as a significant blue bar which I can’t imagine is protective.” – Child Welfare Screener

Performance and Interpretability

- Interpretability leads to…
  - ✕ More efficient training
  - ✕ Better generalization
  - ✕ Fewer adversarial examples
- The interpretability-performance tradeoff is (mostly) a myth

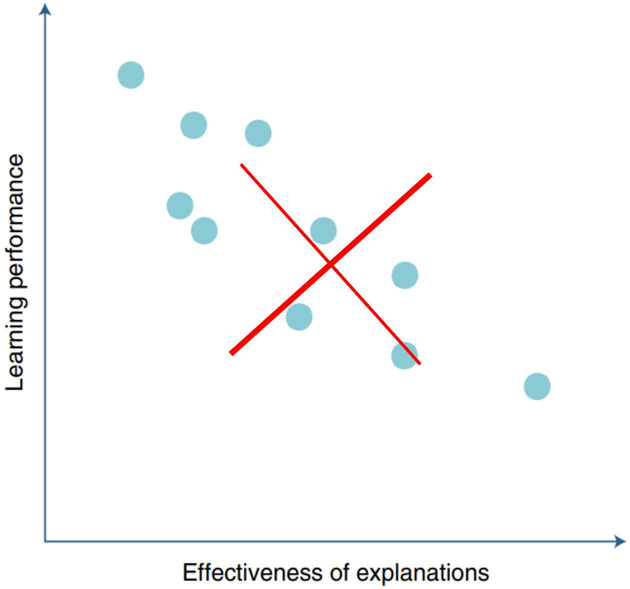


Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). *Adversarial Examples Are Not Bugs, They Are Features* (arXiv:1905.02175).

Performance and Interpretability

- Interpretability leads to…
  - × More efficient training
  - × Better generalization
  - × Fewer adversarial examples
- The interpretability-performance tradeoff is (mostly) a myth



Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). *Adversarial Examples Are Not Bugs, They Are Features* (arXiv:1905.02175).

- Introduction to interpretable ML
- **Why** do we care about interpretable features?
- **What are interpretable features *really*?**
- **How** do we get interpretable features?

What are interpretable features *really*?

The features that are most useful and meaningful to the user

Example: Housing Price Prediction

Area Quality (numeric)	Average House Size (numeric)	Most Common House Color (categorical)	Normalized Median Income (numeric)	x12 (numeric)

Zytek, A., Arnaldo, I., & Liu, D. (2022). The Need for Interpretable Features: Motivation and Taxonomy. *SIGKDD Explorations*, 24(1).

	Area Quality (numeric)	Average House Size (numeric)	Common House Color (categorical)	Normalized Median Income (numeric)	x12 (numeric)
Readable	✓	✓	✓	✓	

	Area Quality (numeric)	Average House Size (numeric)	Common House Color (categorical)	Normalized Median Income (numeric)	x12 (numeric)
Readable	✓	✓	✓	✓	
Understandable	✓	✓	✓		

	Area Quality (numeric)	Average House Size (numeric)	Common House Color (categorical)	Normalized Median Income (numeric)	x12 (numeric)
Readable	✓	✓	✓	✓	
Understandable	✓	✓	✓		
Relevant	✓	✓			

	Area Quality (numeric)	Average House Size (numeric)	Common House Color (categorical)	Normalized Median Income (numeric)	x12 (numeric)
Readable	✓	✓	✓	✓	
Understandable	✓	✓	✓		
Relevant	✓	✓			
Abstract Concept	✓				

- Introduction to interpretable ML
- **Why** do we care about interpretable features?
- **What are interpretable features *really*?**
- **How do we get interpretable features?**

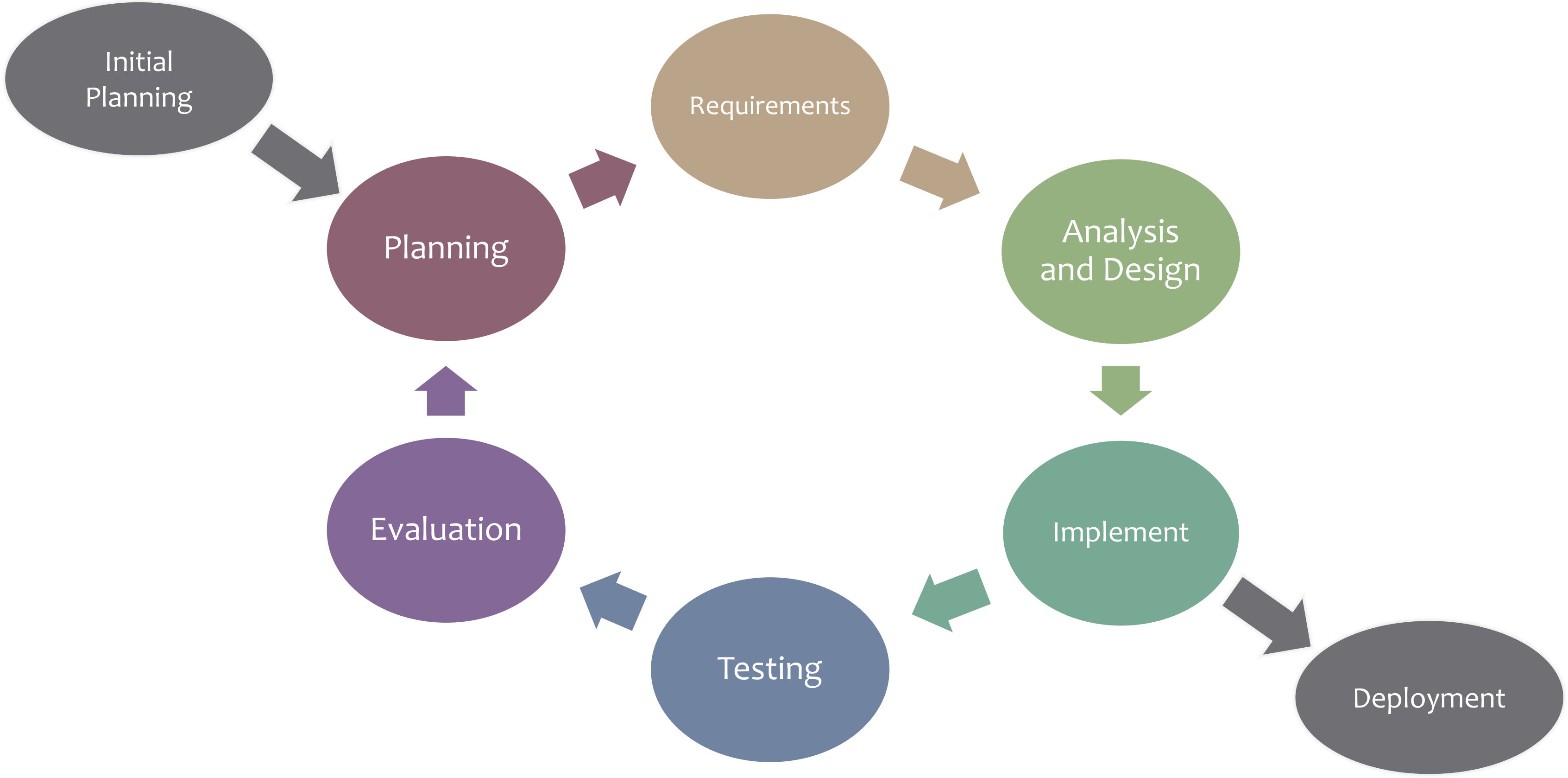
“[Feature engineering] is the first step to making an interpretable model, even if we don’t have a model yet” – Data Scientist

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–26.

Methods for Interpretable Features

1. Including the user
2. Using interpretable feature transformations
3. Using interpretable feature generation

Iterative Design Process (for Features)



Collaborative Feature Engineering

- Improve the process of crafting ***human-generated*** features
- Crowd source feature generation
- Allow domain experts to directly participate

Flock: Choosing features through comparisons

1. Machine-generate features for a prediction task
2. Crowd-generate features
3. Cluster crowd-generated features
4. Iterate on inaccurate model nodes

Cheng, J., & Bernstein, M. S. (2015). Flock: Hybrid Crowd-Machine Learning Classifiers. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 600–611.







The first painting is probably a Monet because it has lilies in it, and looks like Monet’s style. The second probably isn’t Monet because Monet doesn’t normally put people in his paintings.

Split by conjunctions…

The first painting is probably a Monet because it has lilies in it, and looks like Monet's style. The second probably isn't Monet because Monet doesn't normally put people in his paintings.

Cluster using any clustering algorithm…

The first painting is probably a Monet because it has lilies in it
It has flowers
The painting including lilies
There are flowers and lilies in the painting

Crowd-source an aggregated feature label

Does the painting have flowers/lilies?

Results

- Flock outperforms:
  - Original features/data used directly
  - Machine-engineered features only
  - Crowd classifications
- And generates interpretable features, ie.
  - Contains flowers
  - Is abstract
  - Does not contain people

Ballet: Feature Engineering with Feedback

- Abstract away model building/training/evaluating
- Write features with only simple Python

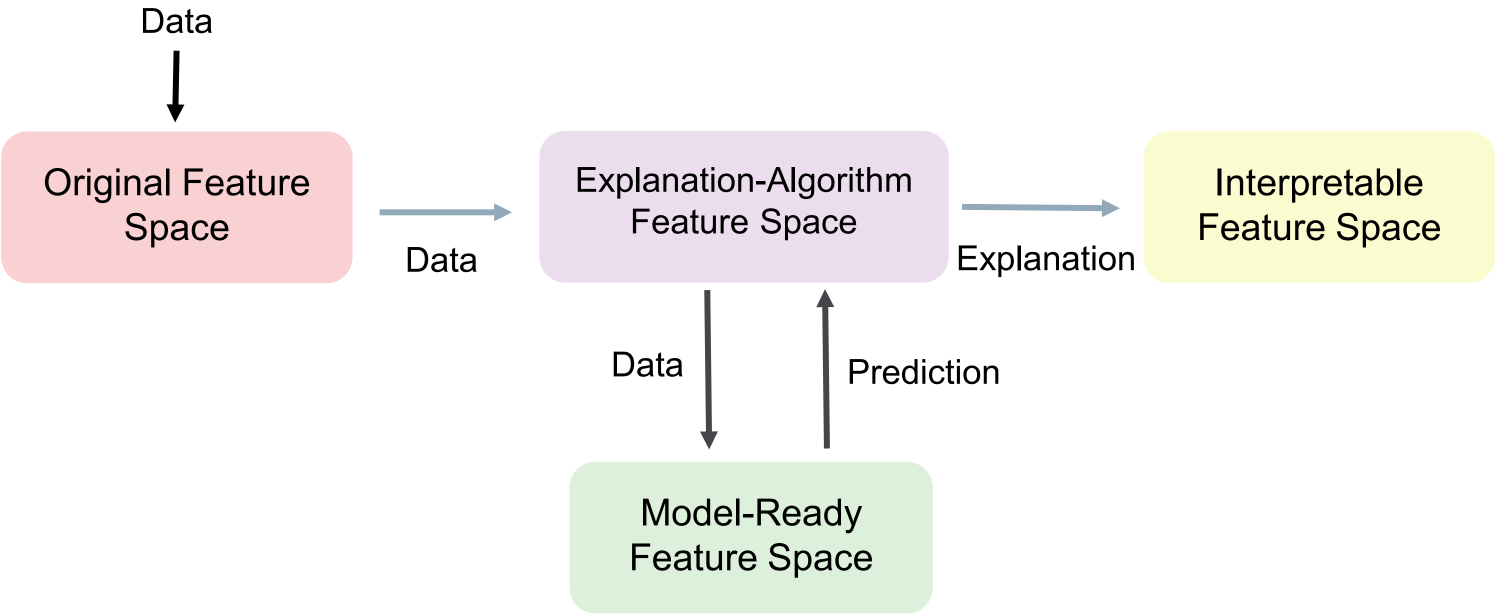
```python
def hi_lo_age(dataset):
    """Whether users are older than 30 years"""
    from sklearn.preprocessing import binarize
    threshold = 30
    return binarize(dataset["users"]["age"]
        .values.reshape(-1,1), threshold)
```

Smith, M. J., Cito, J., Lu, K., & Veeramachaneni, K. (2021). Enabling collaborative data science development with the Ballet framework. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-39.

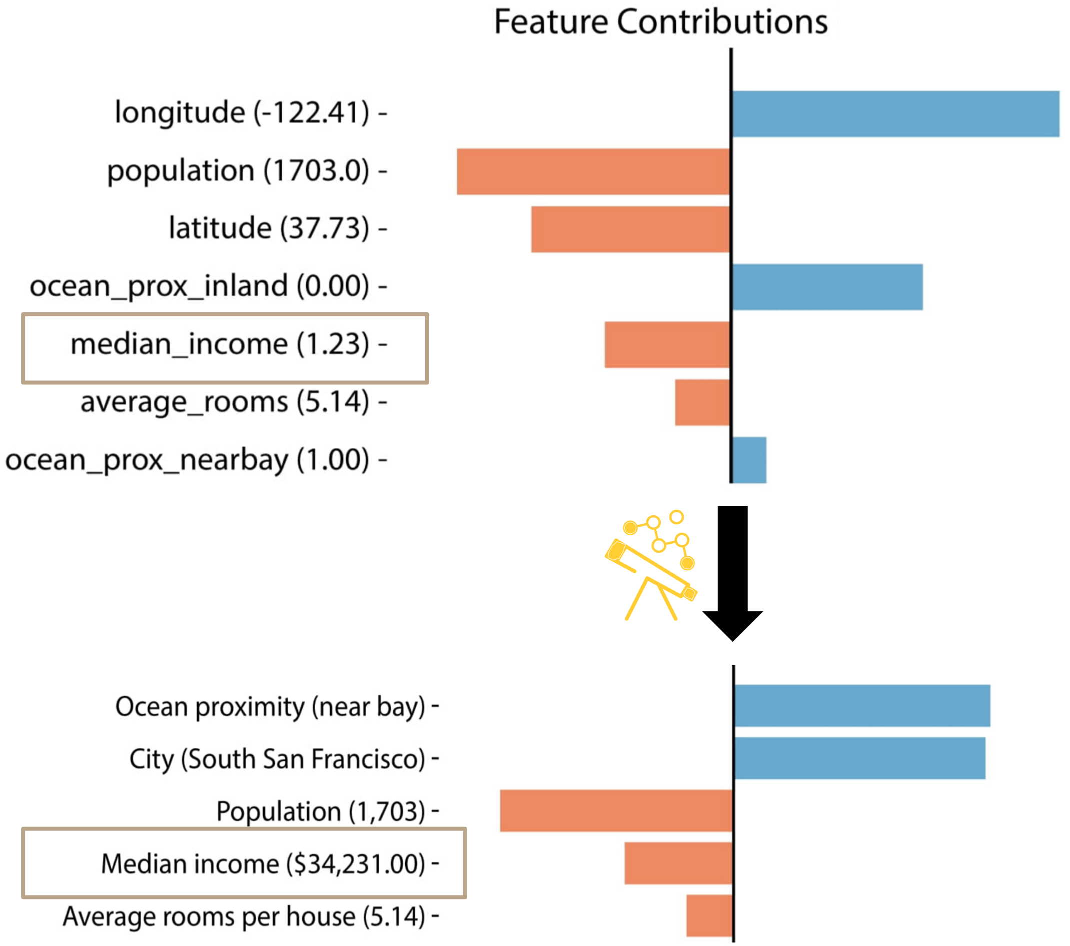
Methods for Interpretable Features

1. Including the user
2. Using interpretable feature transformations
3. Using interpretable feature generation

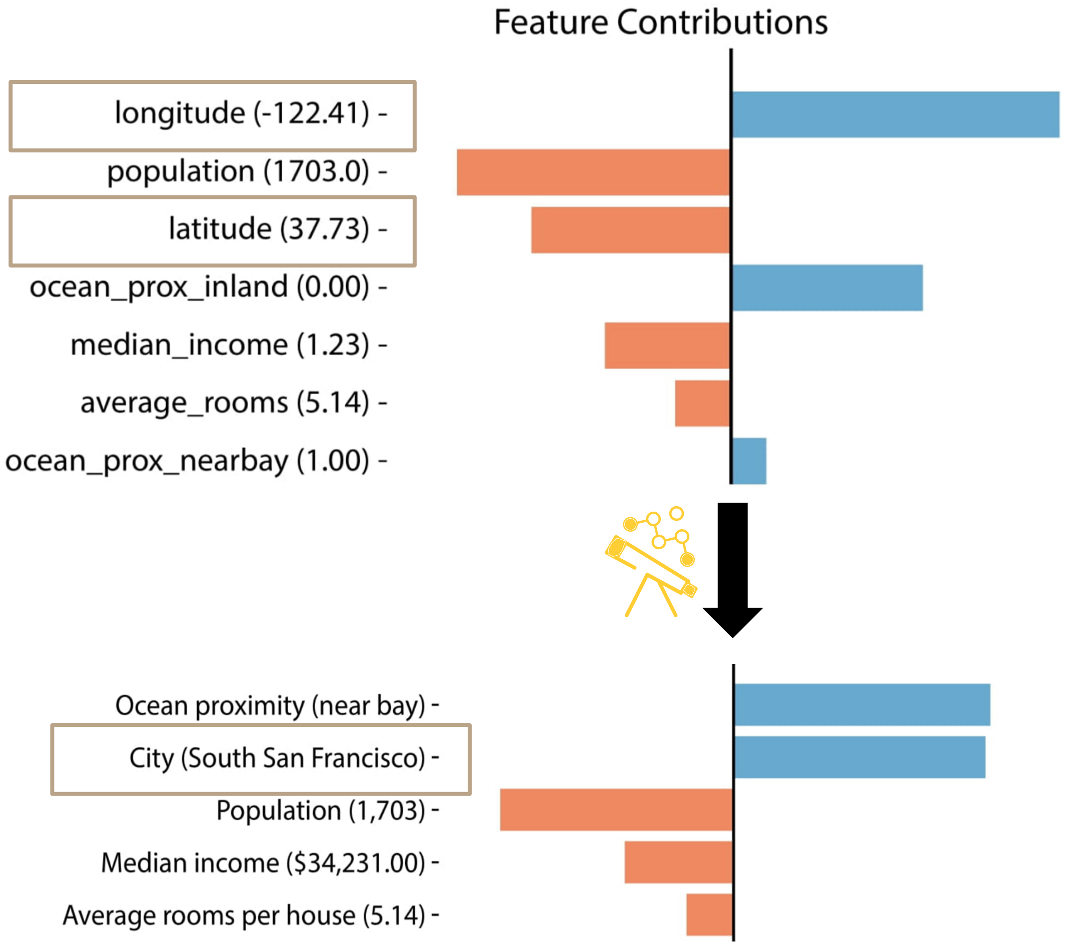
Pyreal: System for Interpretable Transforms

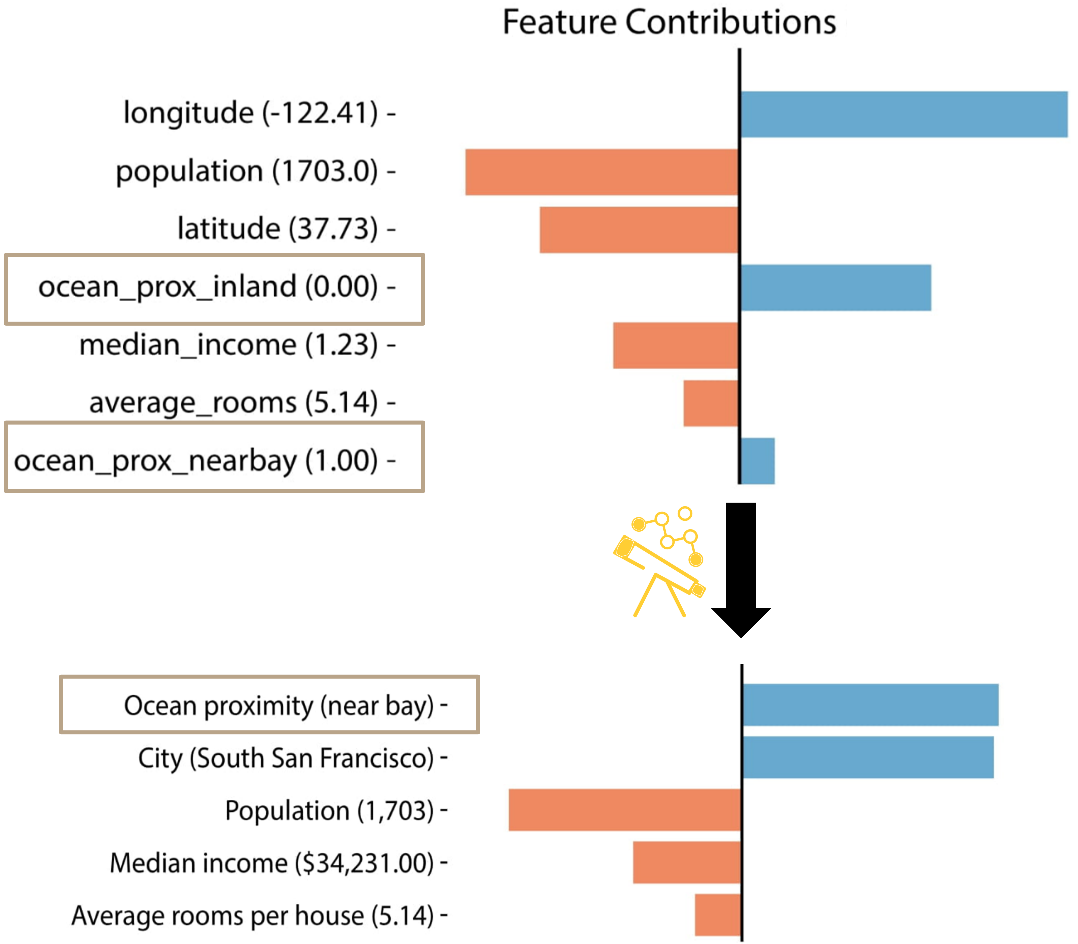


https://dtail.gitbook.io/pyreal/



Feature Contributions
longitude (-122.41) -
population (1703.0) -
latitude (37.73) -
ocean_prox_inland (0.00) -
median_income (1.23) -
average_rooms (5.14) -
ocean_prox_nearbay (1.00) -
Ocean proximity (near bay) -
City (South San Francisco) -
Population (1,703) -
Median income ($34,231.00) -
Average rooms per house (5.14) -


Feature Contributions
longitude (-122.41) -
population (1703.0) -
latitude (37.73) -
ocean_prox_inland (0.00) -
median_income (1.23) -
average_rooms (5.14) -
ocean_prox_nearbay (1.00) -
Ocean proximity (near bay) -
City (South San Francisco) -
Population (1,703) -
Median income ($34,231.00) -
Average rooms per house (5.14) -


Feature Contributions
longitude (-122.41) -
population (1703.0) -
latitude (37.73) -
ocean_prox_inland (0.00) -
median_income (1.23) -
average_rooms (5.14) -
ocean_prox_nearbay (1.00) -
Ocean proximity (near bay) -
City (South San Francisco) -
Population (1,703) -
Median income ($34,231.00) -
Average rooms per house (5.14) -

Methods for Interpretable Features

1. Including the user
2. Using interpretable feature transformations
3. Using interpretable feature generation

Mind the Gap Model (MGM)

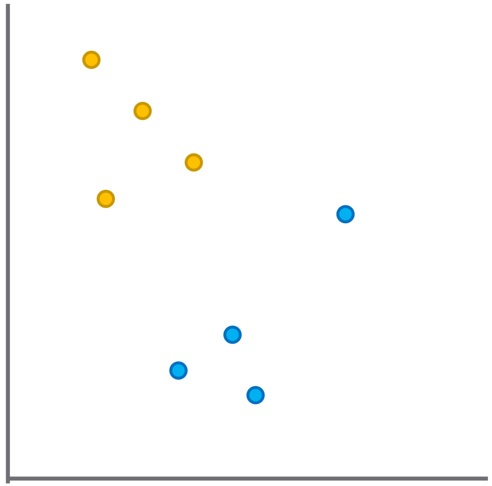
1. Assign features to groups with AND or OR

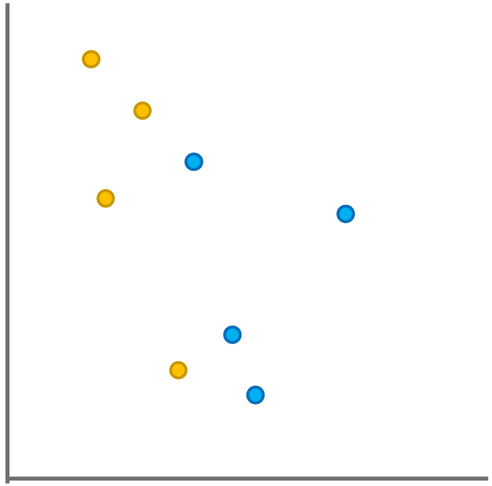
lays eggs	backbone AND tail	toothed OR predator

Kim, B., Shah, J. A., & Doshi-Velez, F. (2015). Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. *Advances in Neural Information Processing Systems*, *28*.

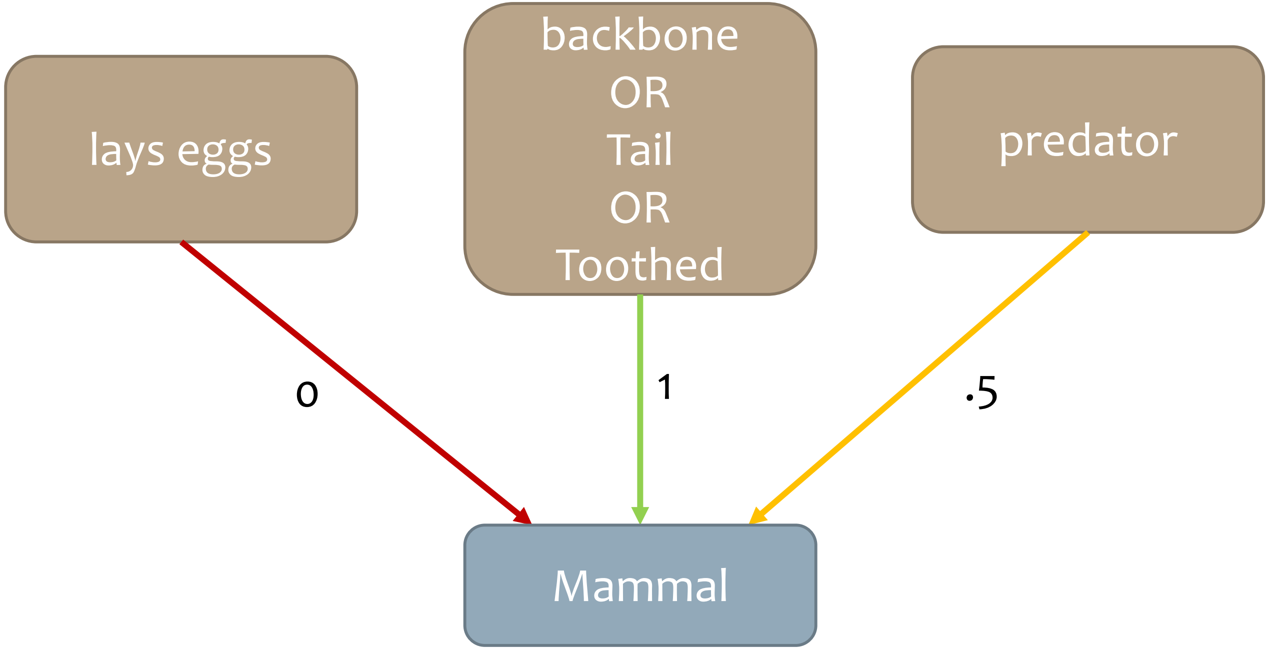
Mind the Gap Model (MGM)

2. Identify groups maximize separation and iterate





Mind the Gap Model (MGM)



Conclusion

1. ML models are only as interpretable as their features
2. Interpretable features are those that are meaningful to the user
3. Interpretable features are generated by including users, focusing on interpretable transforms, and using feature generation algorithms that consider interpretability

Lab

- Use explanation algorithms to identify flawed data