Driving adaptive modeling with data science

Data Science Guest Lecture
Cardiff Metropolitan University
2022-11-21

Robert Caulk, *PhD*
Founder, Lead Software Developer

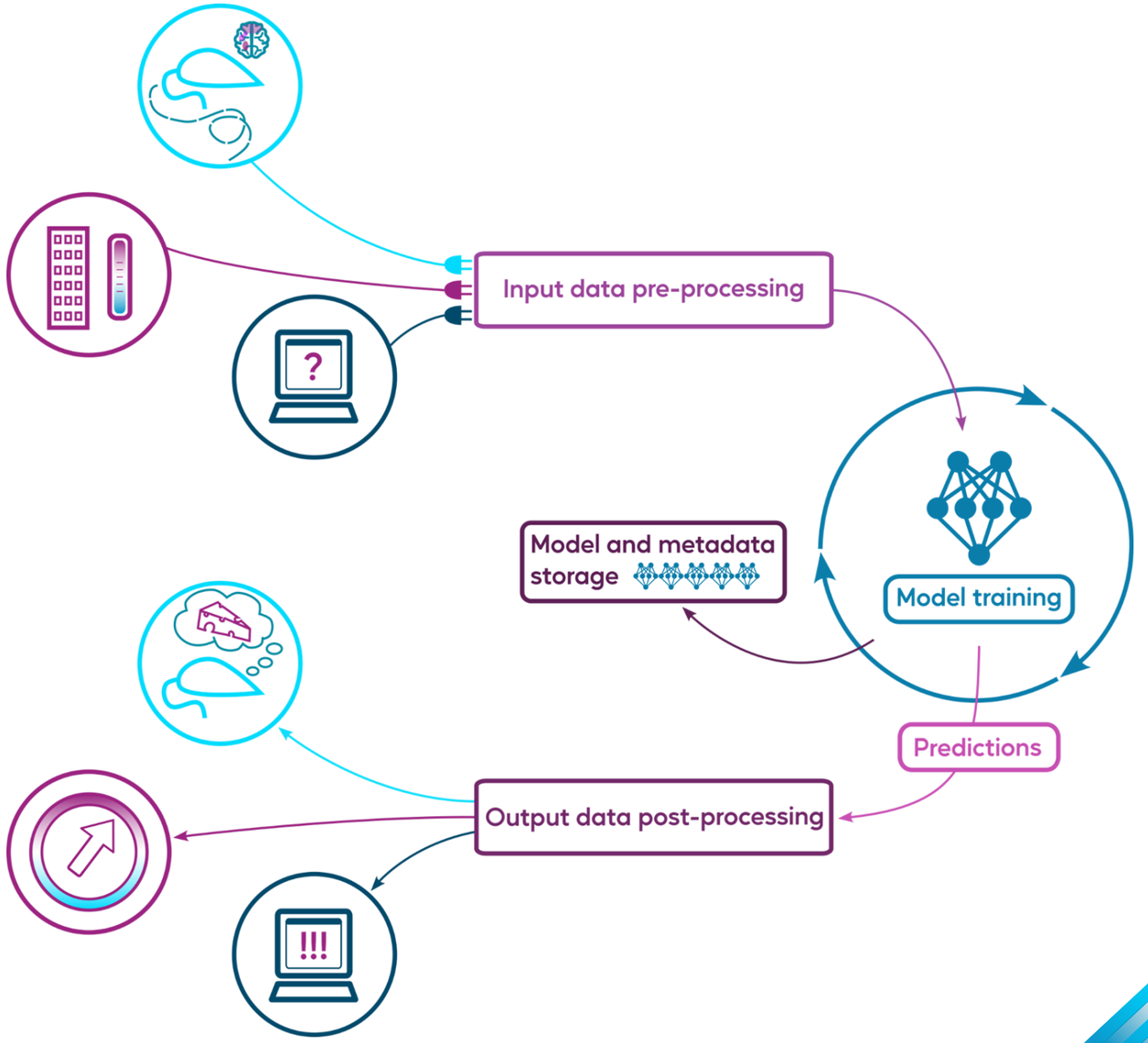
Elin Törnquist, *PhD*
Lead Research Scientist

Emergent Methods

What is Emergent Methods?

Emergent Methods
www.emergentmethods.ai

- Open-source computational research company
- Specializing in generalizing adaptive modeling for time-series data
- Developers of creative ML softwares, FreqAI and JaiRevAI



What is FreqAI?

Real-time adaptive modeling toolkit for making actionable market forecasts

Emergent Methods
www.emergentmethods.ai

User-friendly machine learning sandbox

- 100% open-sourced code base
- Interactive knowledge base with over 10k posts (Discord)
- More than 15 unique developer contributors
- Hundreds of active users finding and reporting bugs

Generalized framework

- Foundational - connects a wide variety of open-source machine learning libraries
- Adaptive - Reinforcement Learning, Decision Trees, Neural Networks, SVM, DBSCAN
- Scientifically sound - industry standard outlier detection methods and statistically safe data handling

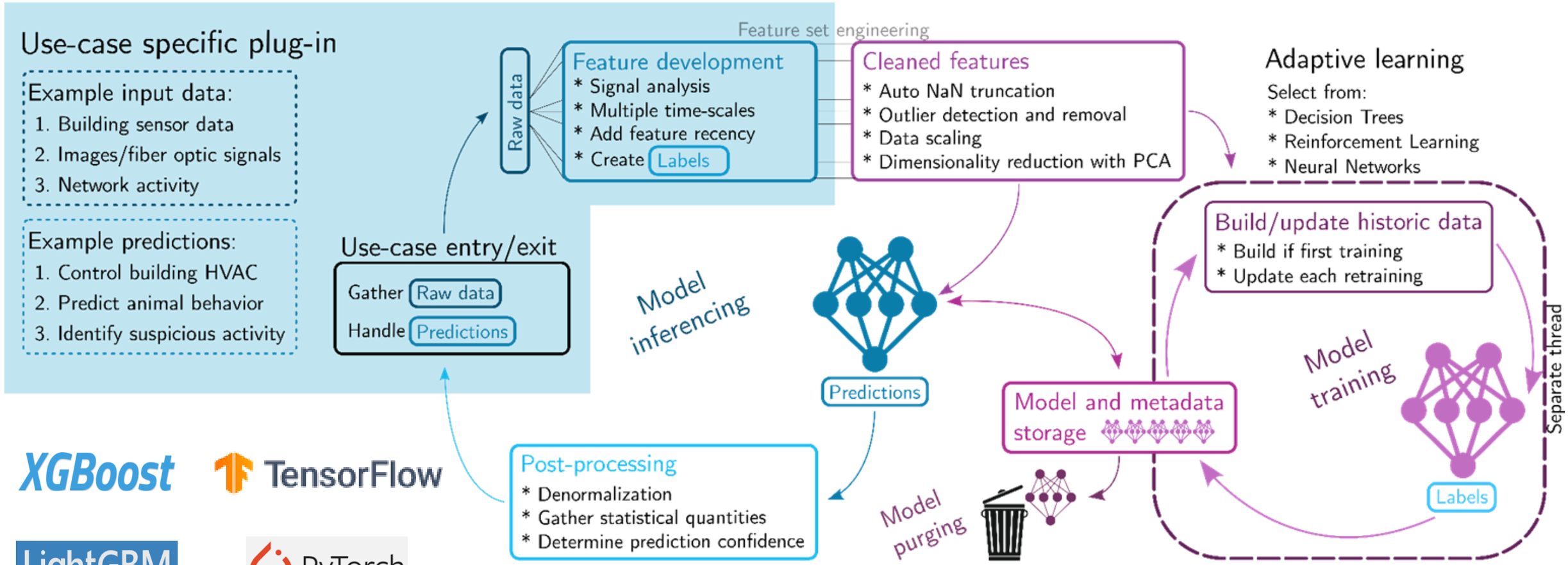
Journal of Open Source Software (under review)

Docs: https://www.freqtrade.io/en/latest/freqai/

Adaptive modeling core engine

Designed for real-time modeling of time-series systems

Emergent Methods
www.emergentmethods.ai

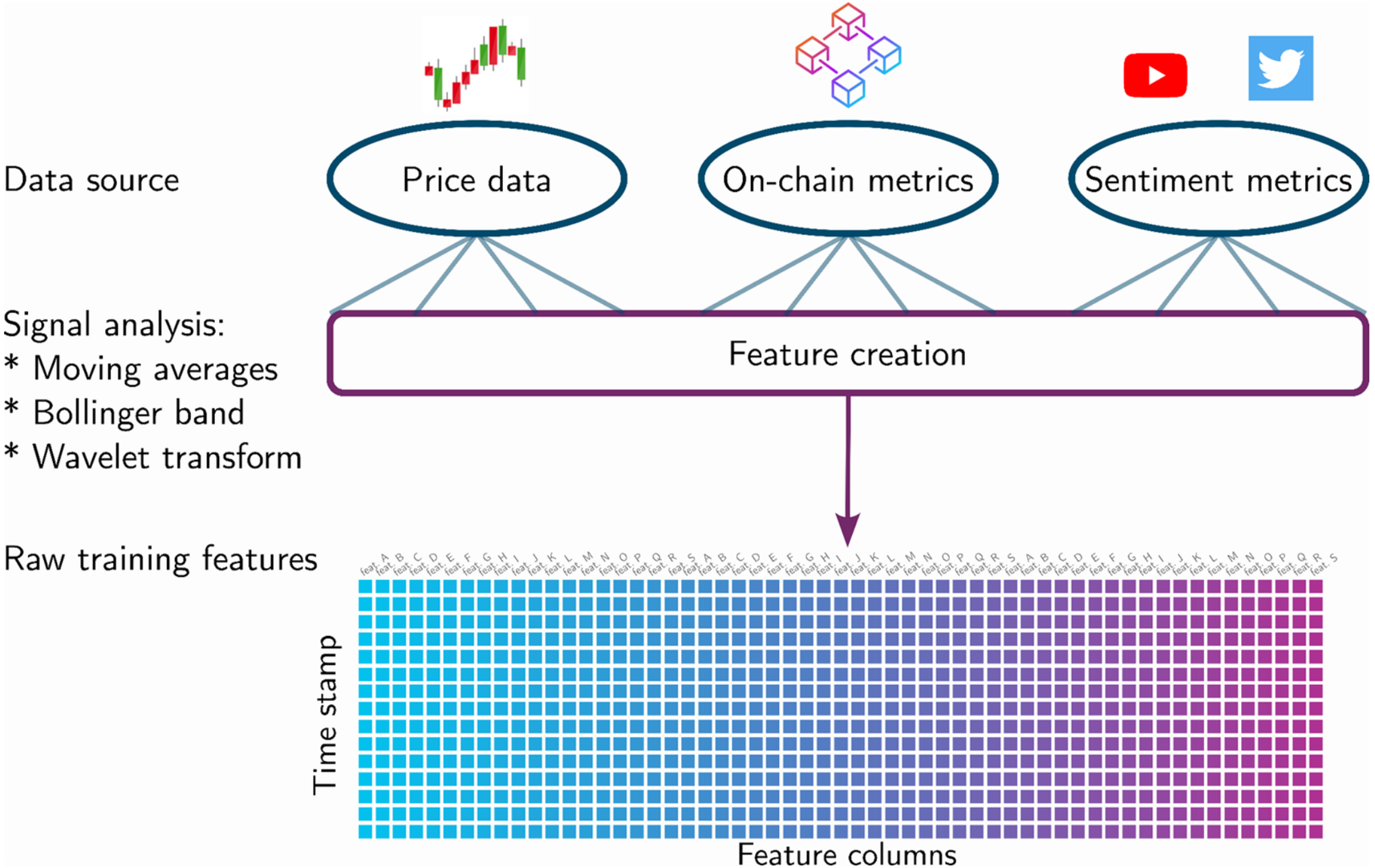


XGBoost TensorFlow LightGBM PyTorch CatBoost

Adaptive modeling core engine

Feature engineering

Emergent Methods
www.emergentmethods.ai



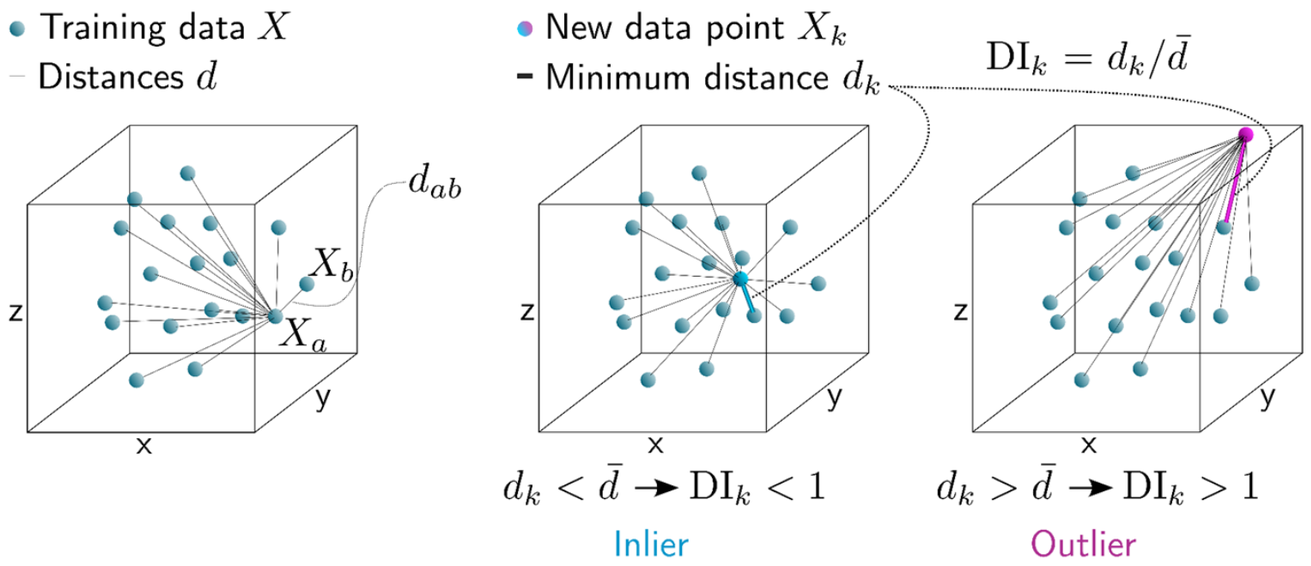
Characterizing the parameter space

Dissimilarity Index

$$d_{ab} = \sqrt{\sum_{j=1}^{p}(X_{a,j} - X_{b,j})^2}$$

$$\bar{d} = \sum_{a=1}^{n}(\sum_{b=1}^{n}(d_{ab}/n)/n)$$

$$DI_k = d_k/\bar{d}$$

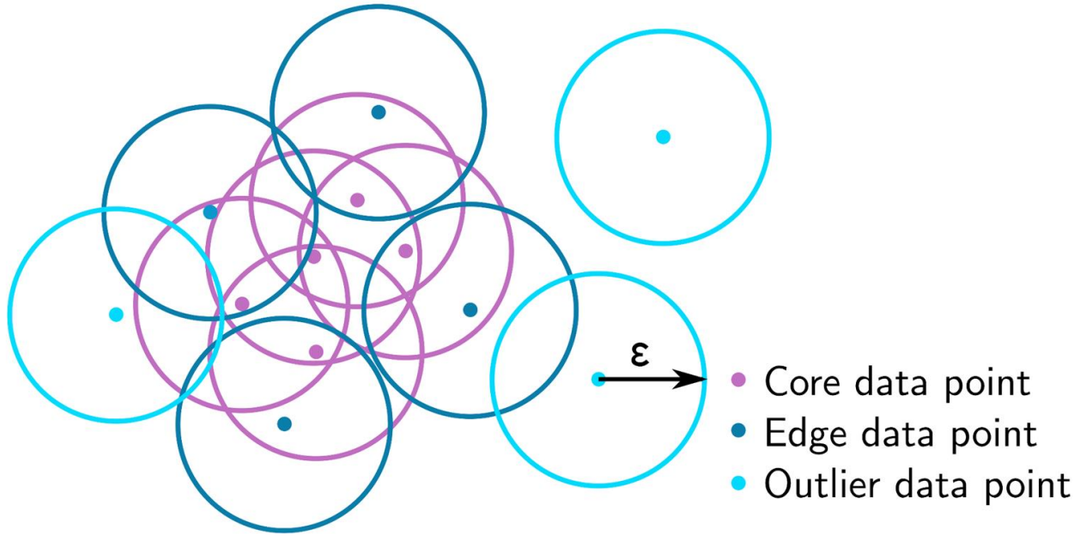


Characterizing the parameter space

Outlier detection

Emergent Methods
www.emergentmethods.ai

- Support Vector Machine - plane fitting
  - Linear SVM - fast, but likely too coarse
- DBSCAN - clustering
  - Challenging to define hyperparameters

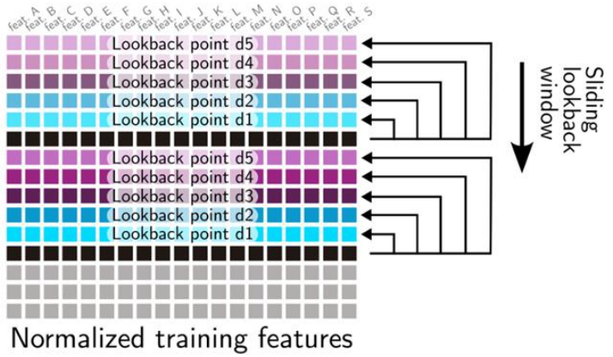


Adding creative metrics

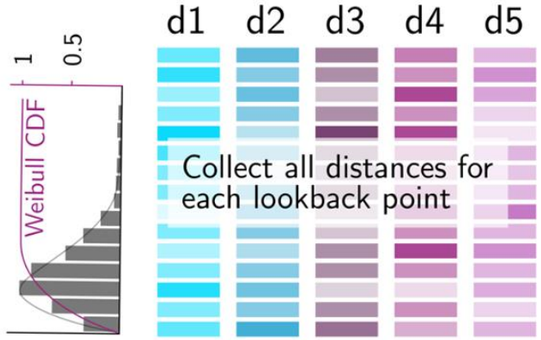
Inlier metric

Emergent Methods
www.emergentmethods.ai

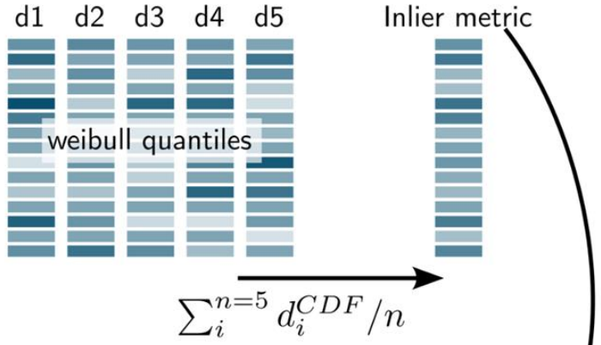
1. Compute distances between each point and the lookback points



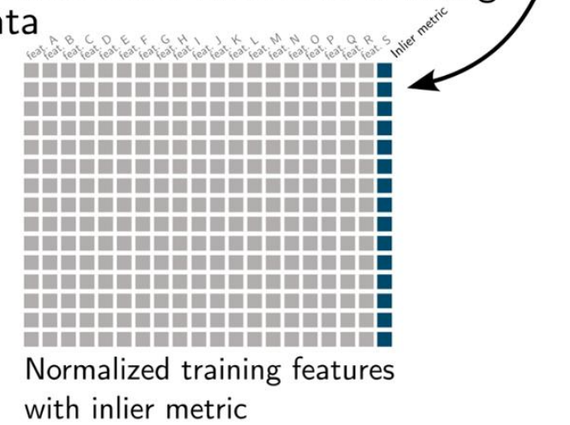
2. Fit weibull distributions to each lookback point



3. Compute point quantiles from weibull distributions

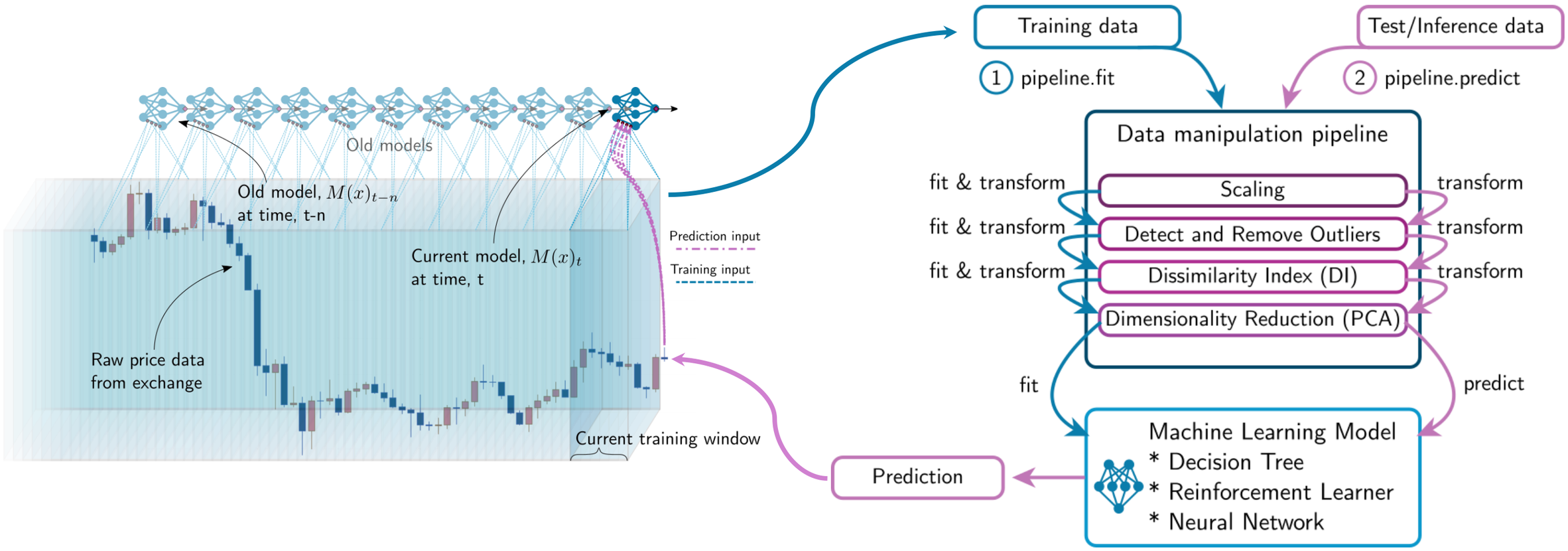


4. Include inlier-metric in training data



Adaptive modeling core engine

Data handling

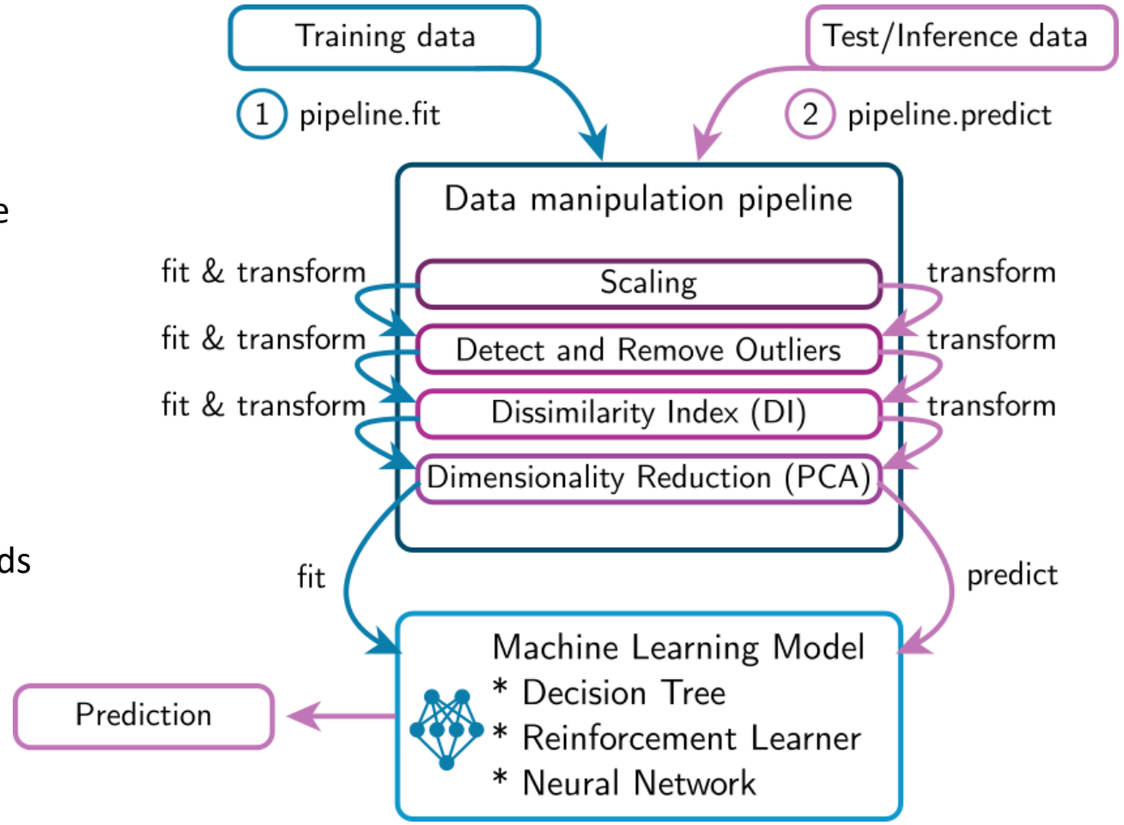


Adaptive modeling core engine

Common pitfalls

Emergent Methods
www.emergentmethods.ai

- Improper NaN handling
  - Don’t blindly fill/replace NaNs
- Improper normalization
  - Normalize test/prediction features to training parameter space
- Tossing the kitchen sink in the dishwasher
  - Preference features from other data sources over redundant signal analysis on the same data source
  - Don’t underestimate the value of computational performance
- Naive stacking of outlier detection/dimensionality reduction methods
  - PCA is great but it may magnify/buffer other methods
- Training a single model to do it all
  - Passing 10 rows x 100 columns vs 100 columns x 10 rows
  - Don’t assume globality



Ongoing experiment (3 weeks)

Configuration

Emergent Methods
www.emergentmethods.ai

- Compare performance on three popular gradient boosted decision tree algorithms
- Aggregate/select best prediction confidence (***Consumer***)
- Track resource usage of each algorithm
- Server details: four identical 12 core Xeon X5660 2.8 GHz 64gb DDR3

Consumer XGBoost LightGBM CatBoost

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

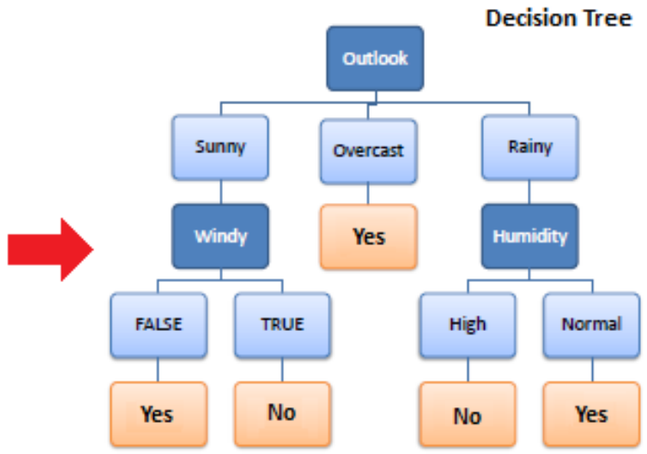


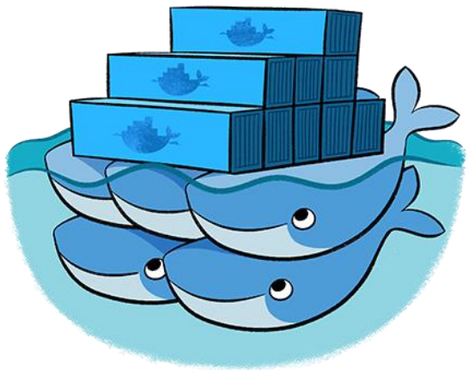
image source: https://www.saedsayad.com/decision_tree.htm

Ongoing experiment (3 weeks)

Distributed deployment

Emergent Methods
www.emergentmethods.ai

- Multiple instance message communication with websockets (***Consumer***)
- GitLab continuous integration + Docker swarm experimental prototyping and deployment



websockets

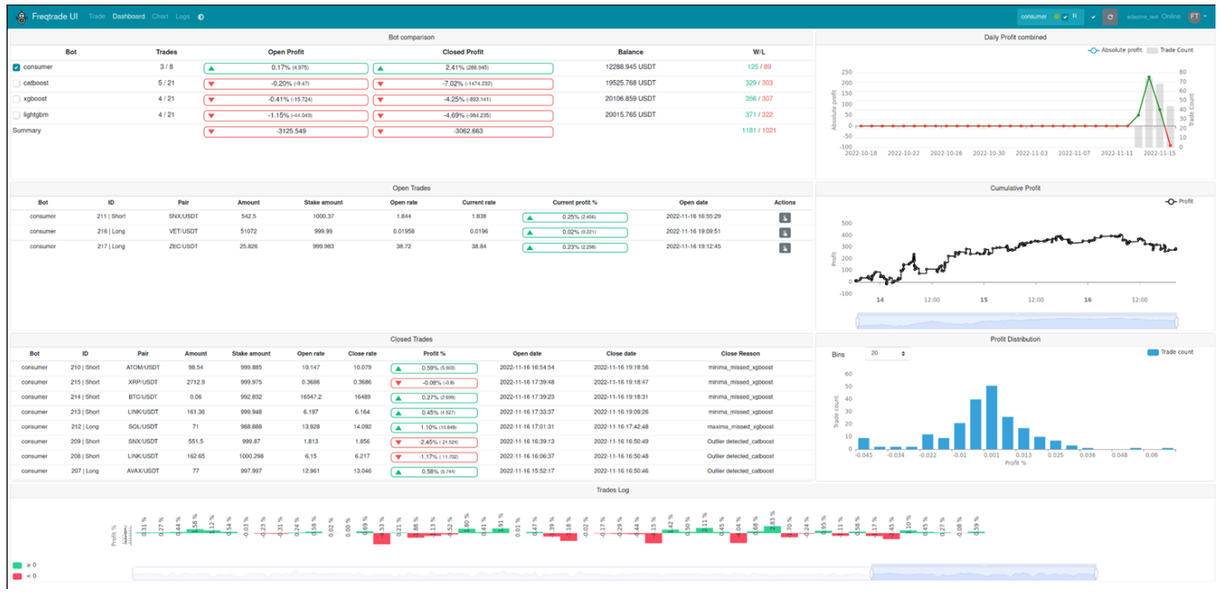
GitLab

Ongoing experiment (3 weeks)

Preliminary results

Emergent Methods
www.emergentmethods.ai

	Training time [s]	Inference time [s]	Profit [%]	RAM usage [Gb]	CPU load [%]
Consumer	N/A	N/A	2.4	2.5	5
XGBoost	91.7	0.17	2.2	5.4	35
LightGBM	169.8	0.05	1.8	5.8	37
CatBoost	256.4	0.52	0.6	6.0	40



Real-time data handling

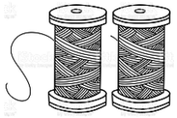
Challenges

Crash resilience

Data collection/storage

Time to prediction

Prediction handling



Threading

Who are Emergent Methods?

Emergent Methods
www.emergentmethods.ai

TEAM

Robert Caulk, PhD
Founder
/ Lead Developer
France

Elin Törnquist, PhD
Lead Research Scientist
Sweden

Tim Pogue
Lead Large-Scale
Systems Engineer
USA

Wagner Costa Santos
Software Developer
/ Data Analyst
Brazil

Andy Lawless
Software Developer
/ Quality Assurance
UK

Steven Caulk, MBA
Head of Business
Logistics
USA

contact@emergentmethods.ai