Financial data science and machine learning

Ernie Chan, Ph.D.

QTS Capital Management, LLC.

About Us

- Ernie is a principal of QTS Capital Management, LLC., a CPO/CTA.
- Ernie was a ML researcher at IBM, and quant at Morgan Stanley, Credit Suisse, etc.
- Detailed biography at www.epchan.com

Why ML now?

- Single-factor, linear quant models have decaying alpha due to ease of replication.
- Advances in ML specifically address overfitting issues.
- New tools in ML introduce more transparency, less blackbox fitting.
- ML can be used for risk management and capital allocation, not primary signal generator.

Traditional Quant vs ML

Traditional Quant	ML
Few predictors	Numerous predictors
Traditional data (prices, fundamentals, etc.)	Alternative data (news, credit card transactions, etc.)
Linear	Nonlinear
Intuitive, easy to replicate	Unintuitive, same data -> very different models
Can't predict probability of success	Can predict probability of success
Arbitrary capital allocation	Logical capital allocation
Harder to overfit	Easy to overfit
Hard to generate multiple backtests for statistical assessment	Easy to generate multiple backtests for statistical assessment
Research+implementation: Easy!	Research+implementation: Very labor-intensive!

3 Steps

- Financial data science
  - Find problems with data, and scrub them.
  - Convert raw data into features.
- Machine learning
  - Use classification or regression techniques to make predictions.
- Trading strategy construction
  - Use predictions as input to a trading strategy.
  - Backtest various versions of strategy.

Data Is Nearly Everything

In the article

Data Challenges Are Halting AI Projects, IBM Executive Says

Arvind Krishna, IBM's senior vice president of cloud and cognitive software, said about 80% of the work with an AI project is collecting and preparing data.

(www.wsj.com/articles/data-challenges-are-halting-ai-projects-ibm-executive-says-11559035800)

Challenges of financial data science

- Ever-changing company names and tickers
- Dividend and split adjustment
- Survivorship bias
- Look-ahead bias of earnings data
- Structural breaks in pre-processed alternative data
- Averaging categorical features

Changing Symbols

- E.g. PCS -> TMUS, simple change.
- E.g. GOOG -> GOOG+GOOGL, split into 2 classes of stocks.
  - A common stock data provider has duplicated the price history of GOOG prior to the split and prepended it to GOOGL's history!
- Without a "Securities Master", very hard to find unique id of a stock, and to merge different databases.

Survivorship Bias

- Easier to show data has bias than to “prove” it doesn’t!
  - Check for Enron, Worldcom, Lehman Brothers, Pets.com, Toys-R-Us.
- Only stock database that I know free of survivorship bias is CRSP.com.

Look-ahead bias

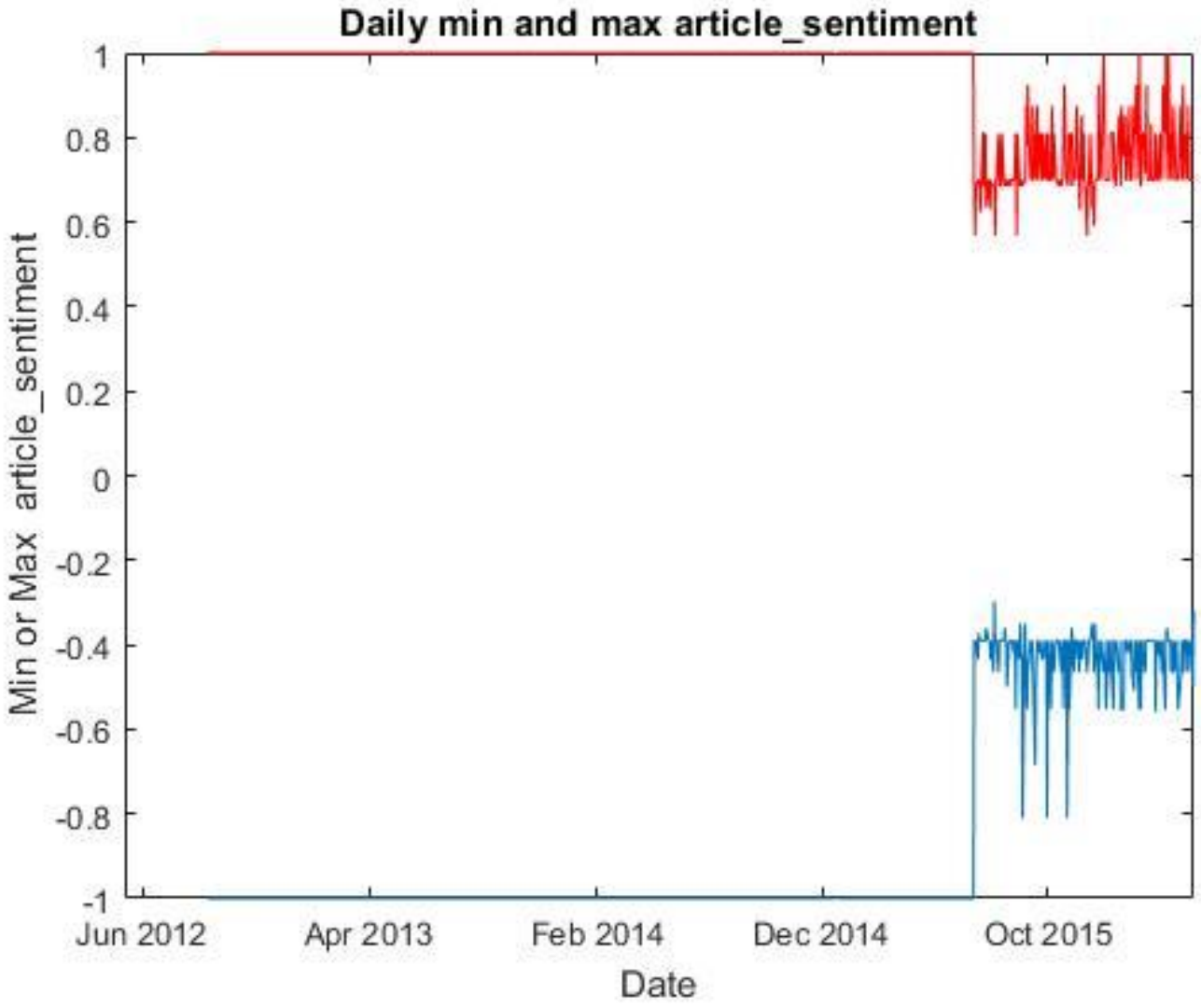
- Many earnings announcement databases have look-ahead bias.
  - Actual announcement date is only known after-the-fact.
  - Backtesting a strategy should only use *expected*, not actual, announcement date.
  - Wall Street Horizon is the only earnings announcement database without this bias.

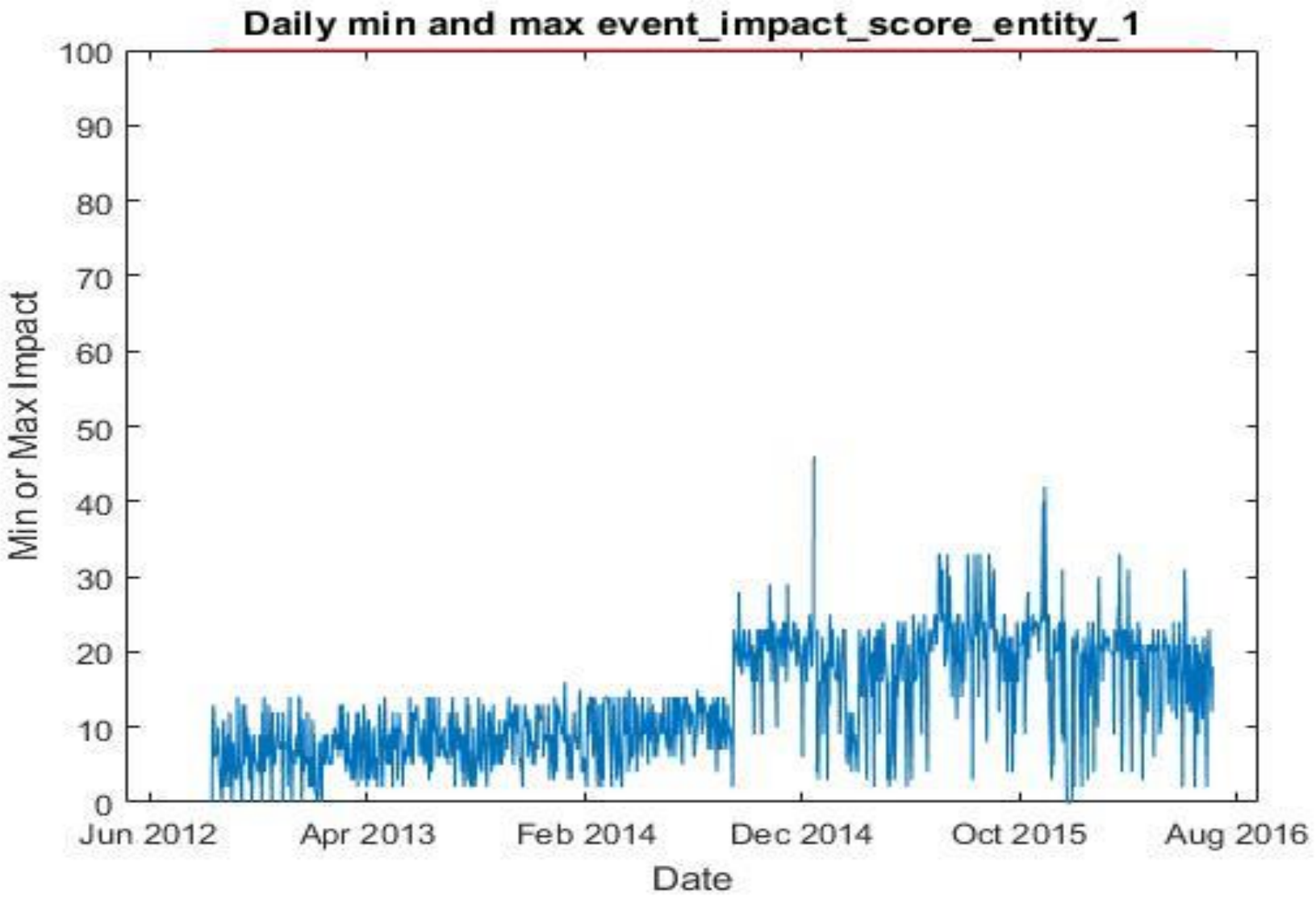
Look-ahead bias

- Many fundamental stock database uses "revised" data.
  - E.g. Regular Compustat vs Point-in-time Compustat.
  - E.g. Sharadar SF1 database distinguishes between "As Reported" (AR) and "Most Recent" (MR) data.
- Revised data cannot be used for backtest.

Structural breaks

- How does one check if alternative (e.g. news sentiment) data is sound?
  - Check for structural breaks in statistics of data features.
  - E.g. daily min and max of sentiment score over time.
  - E.g. daily min and max of impact score over time.
  - (Data from a startup news sentiment data vendor.)


Daily min and max article_sentiment
Min or Max article_sentiment
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Jun 2012
Apr 2013
Feb 2014
Dec 2014
Oct 2015
Date


Daily min and max event_impact_score_entity_1
Min or Max Impact
100
90
80
70
60
50
40
30
20
10
0
Jun 2012
Apr 2013
Feb 2014
Dec 2014
Oct 2015
Aug 2016
Date

What ML models to use?

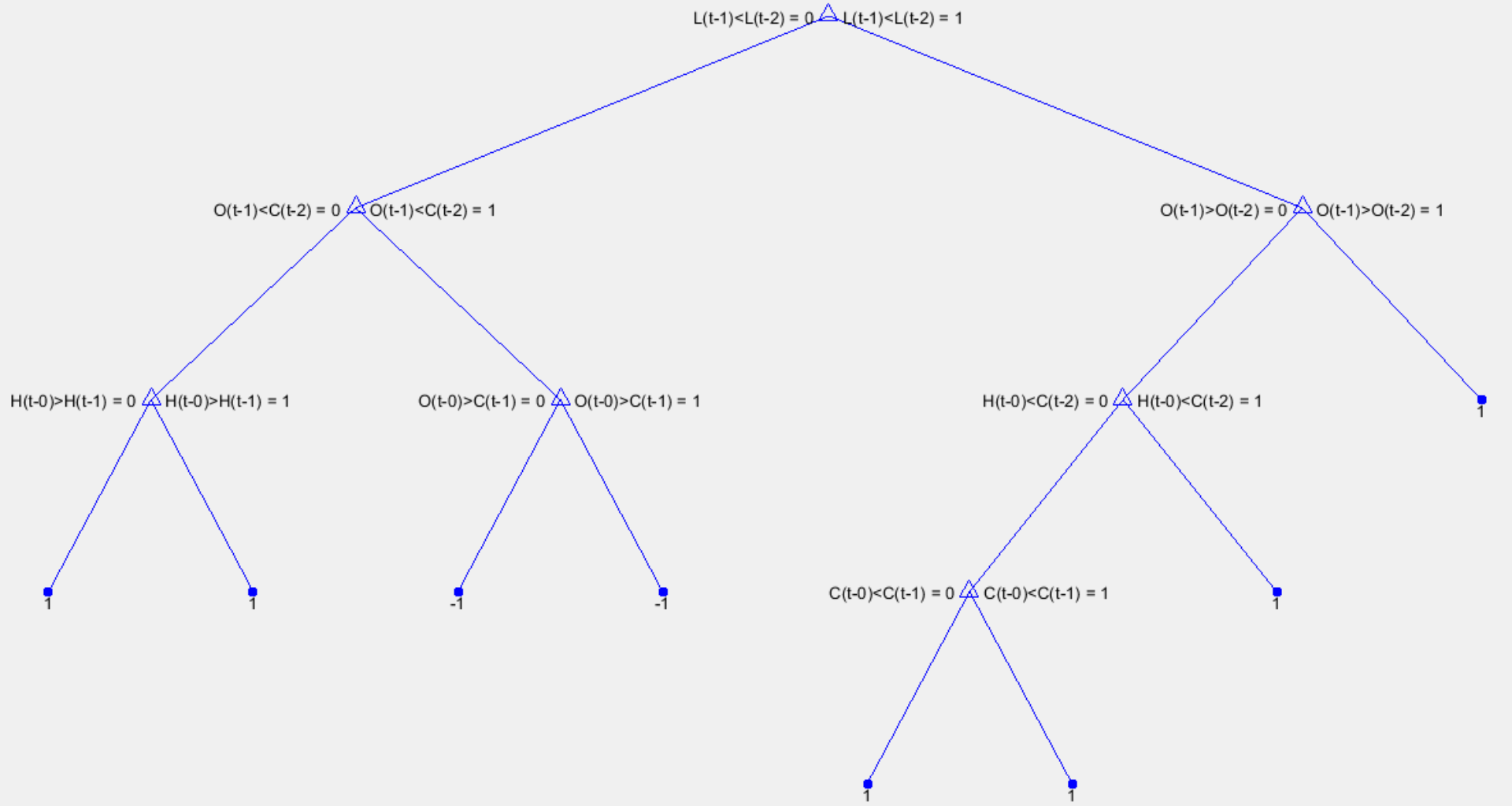
- Shallow vs deep ML models
- **Linear/Logistic regression** is a “shallow” model.
  - Simple (non-hierarchical) structure, little nonlinearity.
- Simple models often work best with few features.
- Deeper models work better when features are numerous and their (nonlinear) interactions important.
- **Random forest** is most commonly used deeper model.

Random Forest (RF)

- Random forest is an ensemble of classification (or regression) trees.
- Tree is a method of select features one-by-one.
- After picking one predictor, we segment data (parent node) into 2 subsets (child nodes) based on an inequality on that predictor.
- The predictor is picked based on minimizing the Gini's Diversity Index (GDI), in the responses within each child node.
  - GDI=$1 - p_+^2 - p_-^2$
  - $p_{+/-}$ is the fraction of observations with positive/negative returns.
  - Perfect classifier gets GDI=0.

Random Forest

- Repeat process on each child, until
  - Child size becomes too small; or
  - GDI in responses in the children are no smaller than the parent's.


L(t-1)<L(t-2) = 0
L(t-1)<L(t-2) = 1
O(t-1)<C(t-2) = 0
O(t-1)<C(t-2) = 1
O(t-1)>O(t-2) = 0
O(t-1)>O(t-2) = 1
H(t-0)>H(t-1) = 0
H(t-0)>H(t-1) = 1
O(t-0)>C(t-1) = 0
O(t-0)>C(t-1) = 1
H(t-0)<C(t-2) = 0
H(t-0)<C(t-2) = 1
C(t-0)<C(t-1) = 0
C(t-0)<C(t-1) = 1
1
1
-1
-1
1
1
1
1
1

Random Forest Regression

- Is classification or regression more important for predicting returns?
  - (See https://twitter.com/lopezdeprado/status/1142815691151745025)

Metalabelling

- Metalabelling is a general ML method to improve on a base model. (See Lopez de Prado.)
- Good candidates for base model:
  - Shallow ML model.
  - Time series model like ARIMA or GARCH.
  - Simple factor model.
  - Simple technical analysis/indicators.
  - Fundamental and discretionary strategies.
  - **Your current trading strategy.**

What is the meta label?

- The label (or metaLabel) is whether the prediction of the base model is correct.
  - MetaLabel=+1(0) if base prediction correct(incorrect).
  - MetaLabel=+1(0) if RT trade generated by base strategy is profitable(unprofitable).
- Meta-model will also output *probability* that base model prediction is correct.
- If metaPrediction=0, or if metaPrediction=+1 but with probability ~ 0.5, can refrain from taking base trade.
- Probability can be used for capital allocation.

Example of metalabeling

- Create your favorite technical indicators to use as features
  - See https://github.com/bukosabino/ta
  - What about supplementary data such as implied volatilities? Commodity, bond, FX indices?
- Create metaLabels, with base model = LR model previously created.
- Train a RF model using these.
- A successful metalabel model: http://www.quantsportal.com/does-meta-labeling-add-to-signal-efficacy/

From ML to Trading Strategy

- Output from classification model=Prob(Up, Dn).
- Simple Strategy:
  - Buy and hold for one bar if Prob(Up) $\geq$ Threshold
  - Sell and hold for one bar if Prob(Dn) $\geq$ Threshold
  - Otherwise flat
- Performance metrics: CAGR, Sharpe Ratio, maxDD, maxDDD, Calmar Ratio, Equity Curve

Capital weighting

- Prob(Up, Dn) can also be used to decide trade size.
- E.g. Weight (in $)=Prob(Up)-0.5
  - Negative weight means short position.
  - If we have multiple assets *s*, can normalize weights s.t. $\sum_s |weights(s)| = 1$.

Robust Backtesting

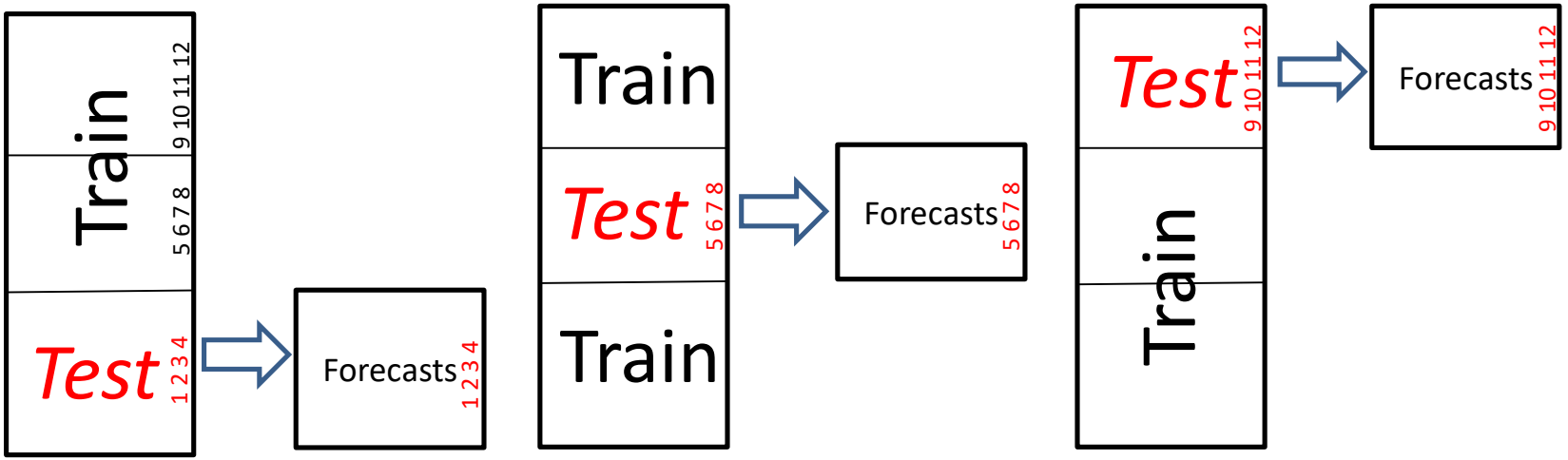
- Even if backtest performance is computed on test set, it is still one sample out of random distribution of price paths.
  - Good/bad performance due to chance?
- If strategy created by ML signals: try different random seeds.
  - Plot a histogram of resulting Sharpe ratios.

Robust Backtesting

- If features are based on prices only, can simulate multiple price paths for backtesting.
- Simulated prices (returns) can be based on
  - Sampling with replacement, or
  - Time series models
    - Linear: ARIMA+GARCH
    - Nonlinear: LSTM, GAN (Generative Adversarial Network)

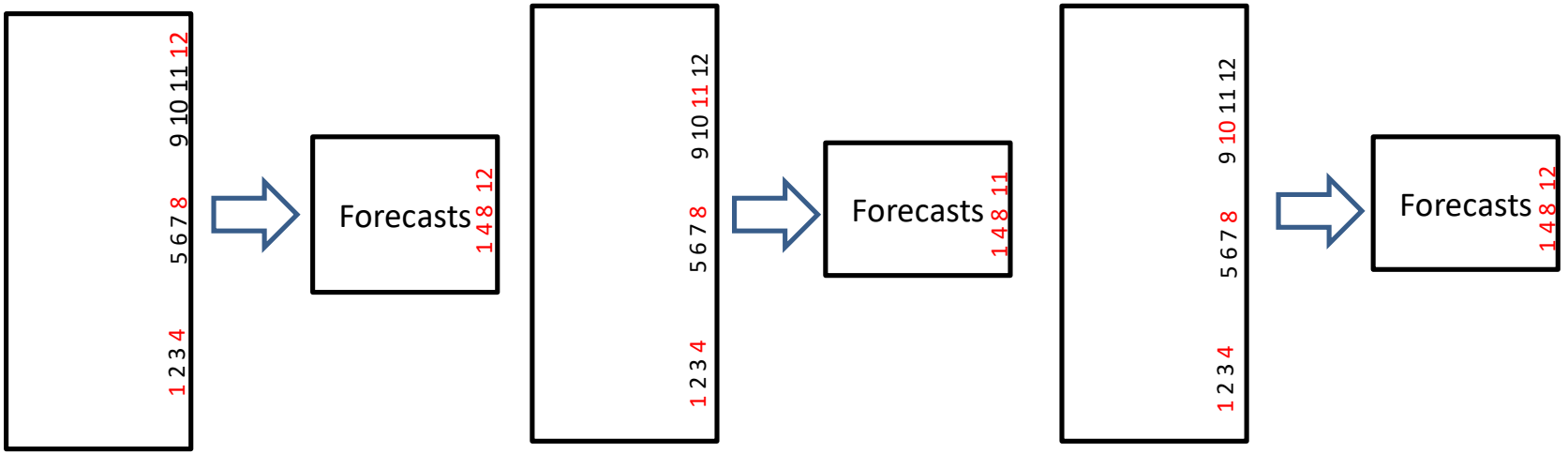
CV Backtest

- Ordinary CV generates only 1 backtest path.
  - E.g. 3 *non-overlapping test folds*, each with 4 rows
  - E.g. Forecasts on t=1, 2, …, 12 are unique.
  - Single OOS forecasted time series: t=1, 2, …, 12



CPCV* Backtest

- CPCV generates multiple backtest paths.
  - Allow test folds to overlap.
  - E.g. test sets below overlap on t=1,4,8.
  - Hence 3 different forecasts on 1,4,8.



*Combinatorial Purged CV, see AFML

Where to start

- Books:
  - Chan: “Machine Trading: Deploying Computer Algorithms to Conquer the Markets”
  - Murphy: “Machine Learning: A Probabilistic Perspective”
  - López de Prado: “Advances in Financial Machine Learning”
- Websites:
  - www.quantresearch.org/
  - github.com/hudson-and-thames

Where to start?

- Python
  - Scikit-learn (all inclusive ML package)
  - LightGBM (random forests with boosting)
  - Keras (deep learning)
- Matlab
  - Statistics and Machine Learning Toolbox
  - Deep Learning Toolbox

Where to start

- Course:
  - "Lifecycle of Trading Strategy Development"
    - Online: epchan.com/workshops
    - In-person: London, Dec 9-11, www.globalmarkets-training.co.uk/datascience.html

Download my talks at

www.epchan.com