Peter Cotton
Chief Data Scientist
Intech Investments

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Frequently Repeated Prediction

Supply and Demand

AI and the Future of Finance Conference
University of Waikato
Nov 30, 2020

microprediction™

Hello. I work for Intech — a leading equity quant manager

Quantitative Equity Specialist

- Founded by pioneering mathematician Dr. Robert Fernholz
- Based in West Palm Beach with Princeton and London offices
- Independently operated unit of Janus Henderson Investors

Distinctive Investment Approach

- Harness stock price volatility for alpha and risk management
- Rely only on advanced mathematics and portfolio rebalancing
- No dependence on forecasting stock returns

Important Investor Benefits

- Investment approach complements other equity managers'
- Volatility is an enduring alpha source
- Process is very customizable – 40% of AUM

33	Year Investing Heritage
$38	Billion in Assets Under Management
75	Employees Focused On Investing And Service
160	Institutional Clients in Five Continents
+10	Year Average Account Tenure
#6	Largest Global Equity Quantitative Manager
#7	Largest Defensive Equity Quantitative Manager

Thanks to key contributors

Eric Lou · 1st

CS + Math Student at Stanford University

- Wrote the front end
- Winning crawlers

Rusty Conover · 1st

Experienced and innovative software executive with a track record of building businesses, platforms and applications.

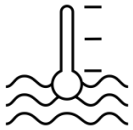
- MUIDs in Java, Julia, Rust
- ZK-MUID proofs
- Electricity

Outline: Demand and supply

1. Why realtime frequently repeated prediction is *almost* synonymous with Artificial Intelligence

2. A new way to supply it - a live, open source Automated Machine Learning network inhabited by algorithms written by anyone.

Example: Predicting water height for NOAA

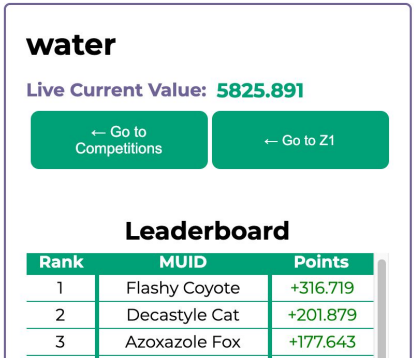


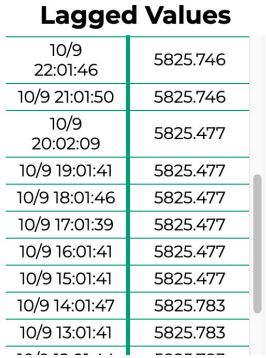
Instrumentation

```
def height():
    df = pd.read_csv('https://www.ndbc.noaa.gov/data/realtime2/21413.dart')
    return float(df.iloc[1,:].values[0].split(' ')[-1])
```

(Returns measured water height … somewhere … from NOAA)

Creates data stream





Data stream is predicted by dozens of competing time series algorithms, written by different authors using different tools, with access to different exogenous data.

Wanna predict it right now in a notebook?

https://github.com/microprediction/microprediction/blob/master/DefaultCrawler.ipynb

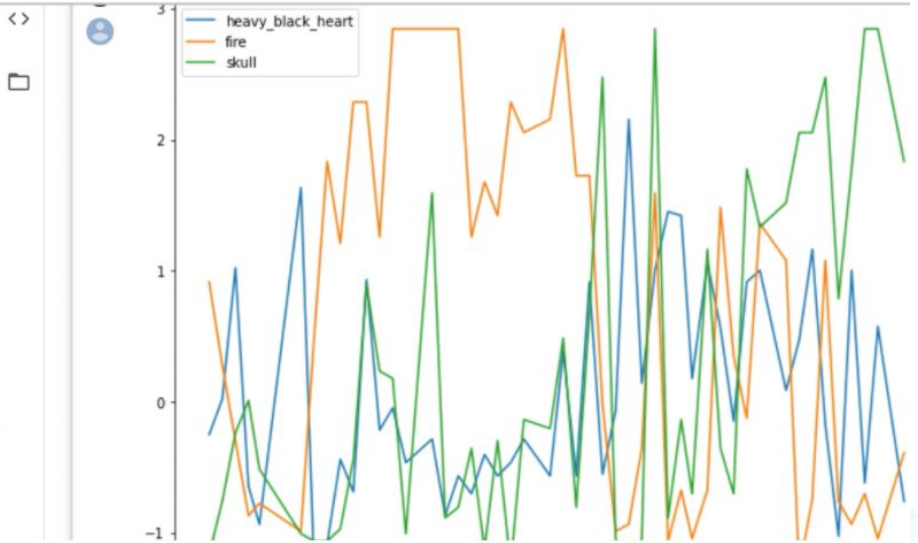
```
from microprediction import MicroCrawler

if __name__ == '__main__':
    crawler = MicroCrawler(difficulty=9)
    crawler.run()
```

1. Demand for Frequently Repeated Prediction

Reactions to the presidential debate

emojitracker: realtime emoji use on twitter

😂 3020230527	❤️ 1493008317	😍 1074792975	😭 1030120041	♻️ 967410648
😔 464765427	😩 461041589	☺️ 415426380	😁 412850245	👌 383838244
👀 295215459	🔥 278774933	💔 274223919	😌 272210739	💙 265066910



See https://www.microprediction.com/blog/tears_of_joy_standardizing_streaming_data

Use category #1: Auxiliary market predictions

Markets predict the mean of a stock well

Everything else (pretty much) is poorly predicted, because those prediction lack the discipline imposed by competition.

- Volatilities,
- Correlations
- Bid-offer spreads
- Liquidity
- Trading costs
- Holding periods
- Client flow
- Response to inquiry
- Cover price

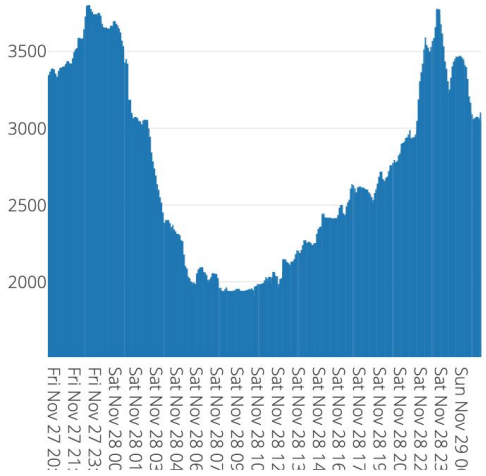
Use category #2: Prioritizing human work

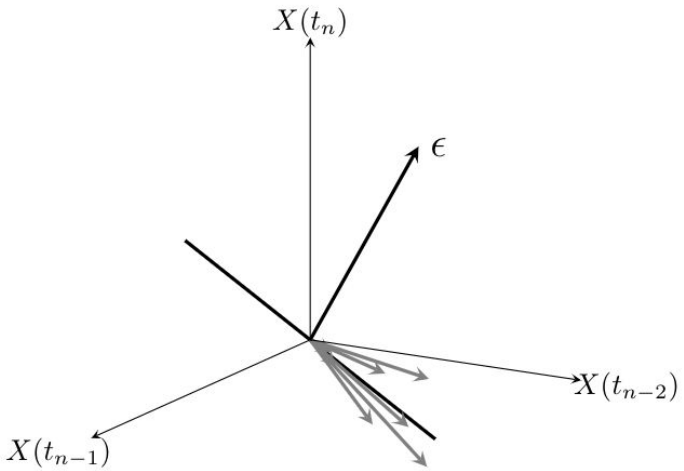
e.g. reference data cleaning

Probability that a record is changed?

Which records will be changed?

Use category #3: Enhancing live data feeds





Tagging.

Converting sporadic live data to continuous.

Discovering existing relationships

Predicting delayed data and partially filled data

Discovering good embeddings

Finding new exogenous data

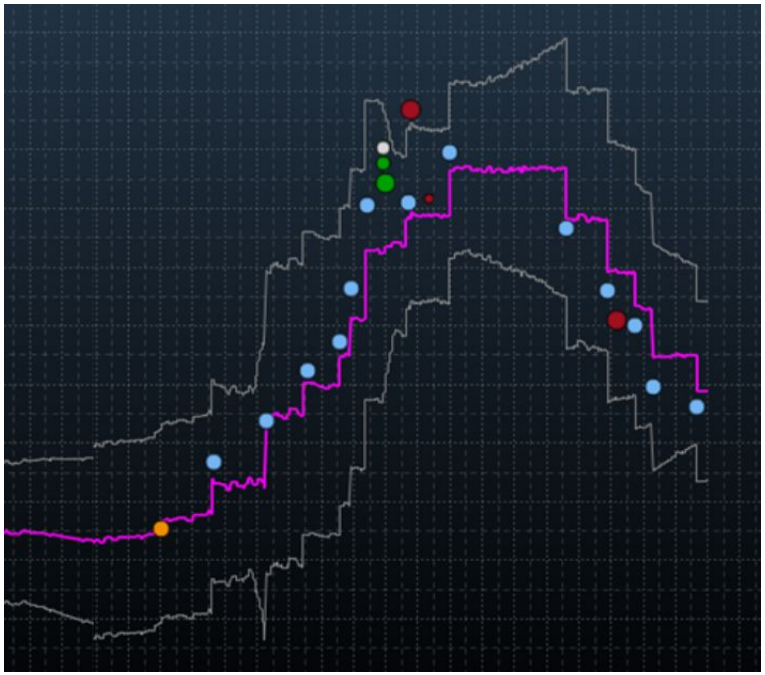
Discovering good proxies for truth

Use category #4: Live feature discovery

Chumming the water

Predicting quantities correlated with the quantity you truly care about

Determining which feature generation algorithms are suited to the task at hand



Use category #5: Enhancing business intelligence applications

Predicting numbers on dashboards

Highlighting unusual movements

Predicting human reaction to information, or not (false positives)

Enabling humans to track a larger amount of data in real time

Name	Price	10m	20m
USDJPY Curncy	110.96	-0.0324	+0.0458
EURUSD Curncy	1.1816	+0.0005	-0.0008
AUDJPY Curncy	84.32	+0.0389	+0.055

Use category #6: Ongoing performance analysis of models

Creating a feed and also a prediction of the same

Publishing model residuals

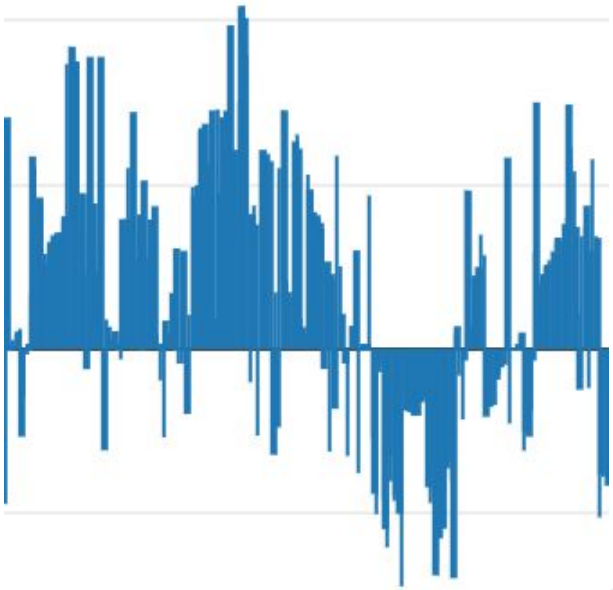
Using shifts in rankings of algorithms to detect regime changes

Using leaderboards to detect model drift

Identifying exogenous sources of data

Quality control of model inputs

Keeping quants honest !



Use category #7: Enhancing business intelligence applications

Human surveillance is popular (dashboards etc)

Datable Llama	+0.433
Boost Mole	+0.431
Hamal Bass	+0.235
Carryover	+0.2
Mesole Mammal	-0.014
Eyas Stoat	-0.1

Directional change prediction

Anomaly detection … from changes in leaderboards?

Regime changes … from changes in leaderboards?

Predicting human reaction to information, or not (false positives)

Enabling humans to track a larger amount of data in real time

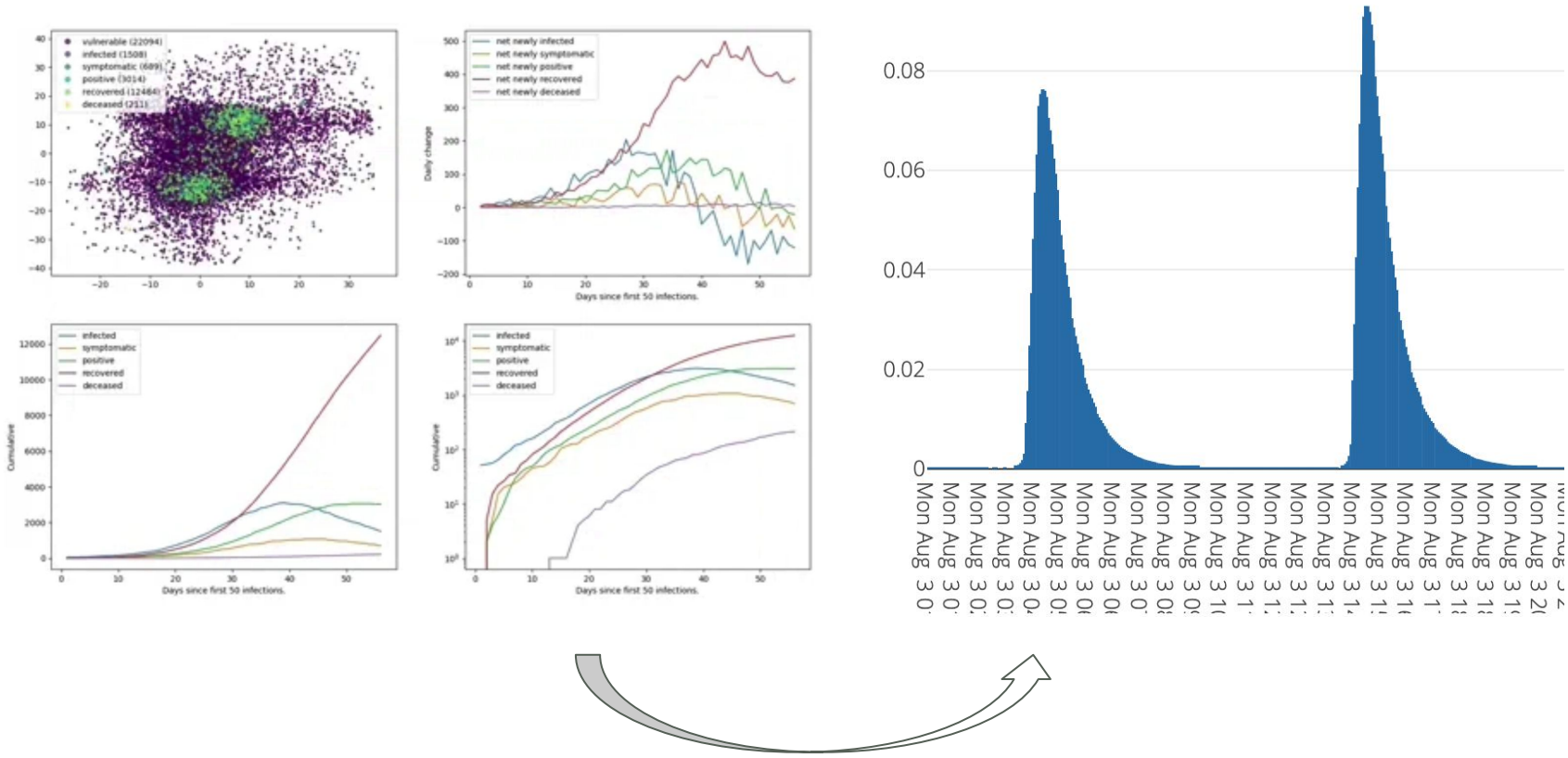
Use category #8: Fairness and explanation

Discovering hidden bias (proxies for race, redlining)

Predicting coefficients ex-post

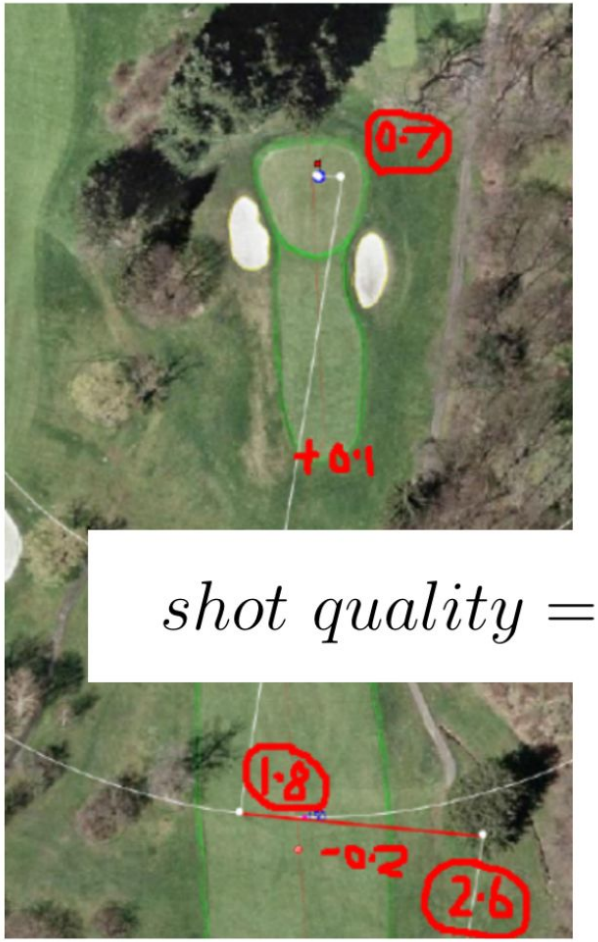
Usage category #9: Surrogate models

Competing and combining surrogate models for agent based epidemic modeling



https://www.microprediction.org/stream_dashboard.html?stream=pandemic_infected

Use mega-category #10, #11, #12... Control systems



V = avg(# shots to finish hole) - 1

$$shot\ quality = V(before) - V(after) - 1$$

Endless possibilities

Category	Example Sub-cat.	Example Sub-sub cat.
Recognition	Image	Facial
Search	Travel	Personalization
Recommendation	Ad-tech	Click-throughs
Government	Open cities	Flight status
Sales and CRM	Repeat shopping	Visitation
Internet of things	Homes	Usage
Environment	Air	Pollutants
Transport	Driving	Distracted driver
Manufacturing	Industrial control	Predictive maintenance
Agriculture	Juice	Orange juice
Finance	Investment banking	Commercial loans
Energy	Power	Wind
Medicine	Inventory	Hospital stays

2. Supplying Frequently Repeated Predictions

Algorithms Play Continuous Lottery Games

All day, every day

traffic_absolute_speed

Live Current Value: 49.7

← Go to Competitions | ← Go to Z1

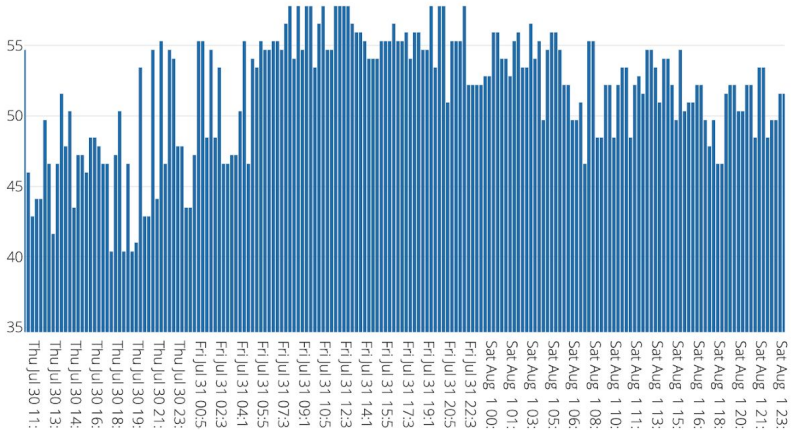
Leaderboard

Rank	MUID	Points
1	Staboy Bat	+231.03
2	Doodle Bee	+161.968
3	Chatty Fly	+101.565
4	Behest Mammal	+97.258
5	Decastyle Cat	-8.737
6	Lettable Clam	-8.816
7	Bedabble Toad	-9
8	Statesboy Cat	-10.129
9	Flammable Cod	-10.5

Lagged Values

Timestamp	Data
8/4 22:39:32	49.7
8/4 22:19:33	49.08
8/4 21:59:33	49.7
8/4 21:39:32	49.08
8/4 21:19:32	49.7
8/4 20:59:34	49.08
8/4 20:39:32	49.08
8/4 20:13:40	54.05
8/4 19:53:40	36.03
8/4 19:33:41	50.95
8/4 19:13:39	49.7
8/4 18:53:42	45.98
8/4 18:34:07	49.7
8/4 18:13:44	45.98
8/4 17:53:41	45.98
8/4 17:33:48	47.22
8/4 17:13:41	50.33

Data Values

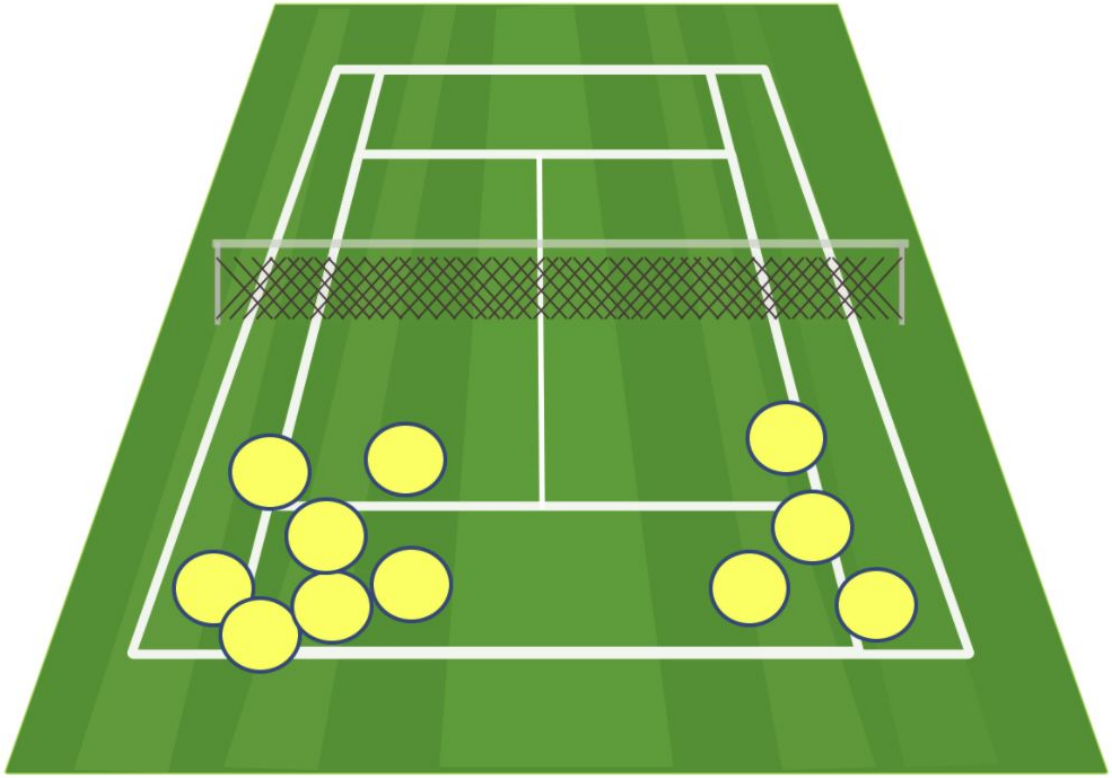


Algorithms authored by anyone

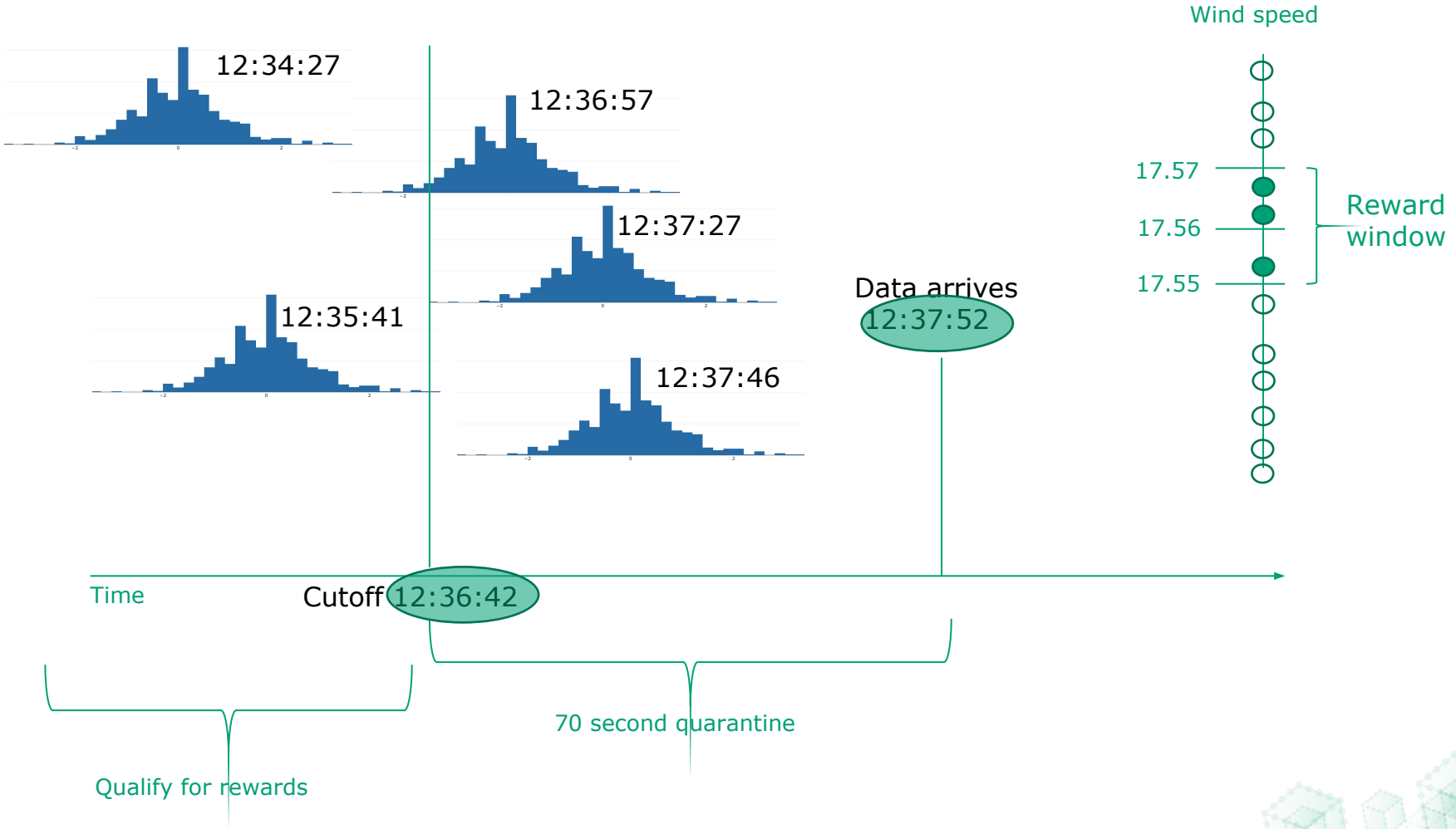
Live data published by anyone

Algorithms submit 225 scenarios

Why not point estimates?



Quarantine

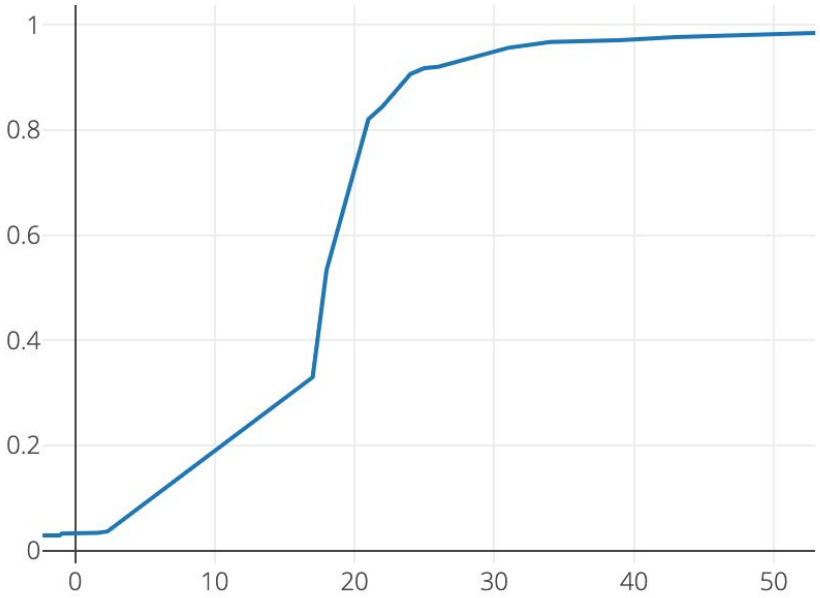

12:34:27
12:36:57
12:37:27
12:35:41
12:37:46
Wind speed
17.57
17.56
17.55
Reward window
Data arrives
12:37:52
Time
Cutoff 12:36:42
70 second quarantine
Qualify for rewards

Implied Percentiles

Every incoming data point implies a new data point …

$$z = F(x)$$

where F is the "community" distribution function



Cumulative distribution for NY Electricity Production (Wind) 1 hr ahead

Stacking Lotteries

Those "market implied" percentiles are themselves the subject of lottery games (via normal quantile function)

Approximately N(0,1)

z1~fcx~70

Go to Stream →

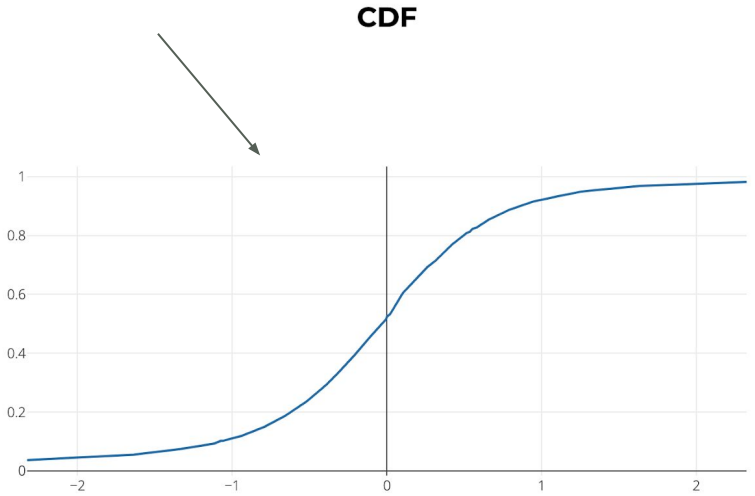
Horizon: 70 sec | 310 sec | **910 sec** | 3555 sec

Leaderboard

Rank	MUID	Points
1	Comal Cheetah	+7.101
2	Decastyle Cat	+5.595
3	Exhalable Cat	+5.338
4	null	+0.8
5	Flammable Cod	-3.793
6	Cellose Bobcat	-15.04

Lagged Values

Timestamp	Z-Score
8/4 16:01:06	-0.51996
8/4 16:00:11	0.55508
8/4 15:59:18	-0.00448
8/4 15:58:05	-0.04721
8/4 15:57:04	1.48253
8/4 15:56:03	0.55323
8/4 15:55:05	-0.00446
8/4 15:54:03	-0.00445
8/4 15:53:04	0.58113
8/4 15:52:11	-1.29729
8/4 15:51:05	1.24723
8/4 15:50:06	-0.5483
8/4 15:49:04	-0.55082
8/4 15:48:04	-0.53286
8/4 15:47:04	-1.07452
8/4 15:46:06	-0.39177
8/4 15:45:13	0.26186

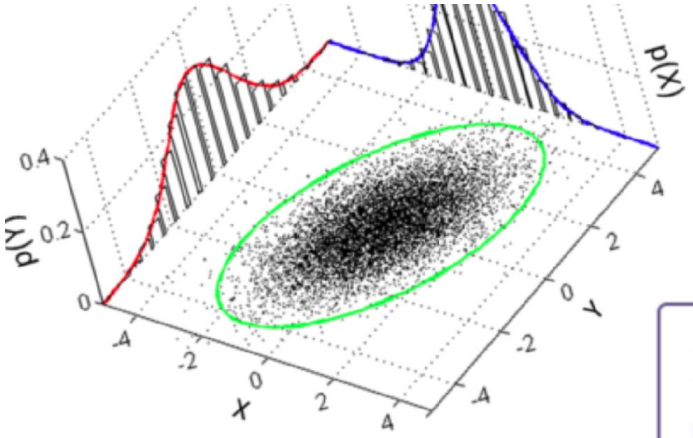


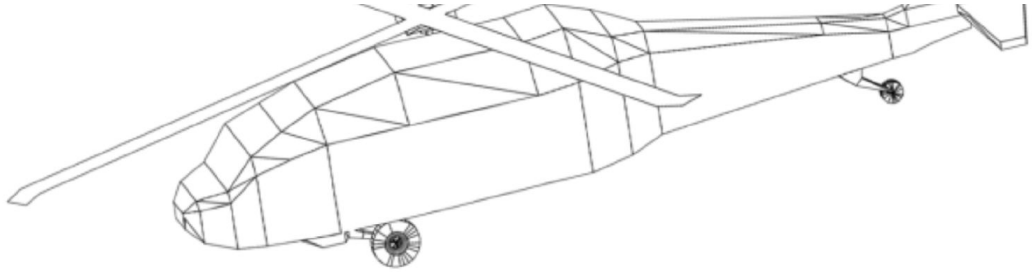
Algorithms predicting small deviations from standard normal

Combine Percentiles

Some seemingly univariate series of games are actually copulas

Pitch and Yaw implied compulas - from MIT SciML helicopula challenge





helicopter_theta

Live Current Value: -0.67467

← Go to Competitions | ← Go to Z1

Leaderboard

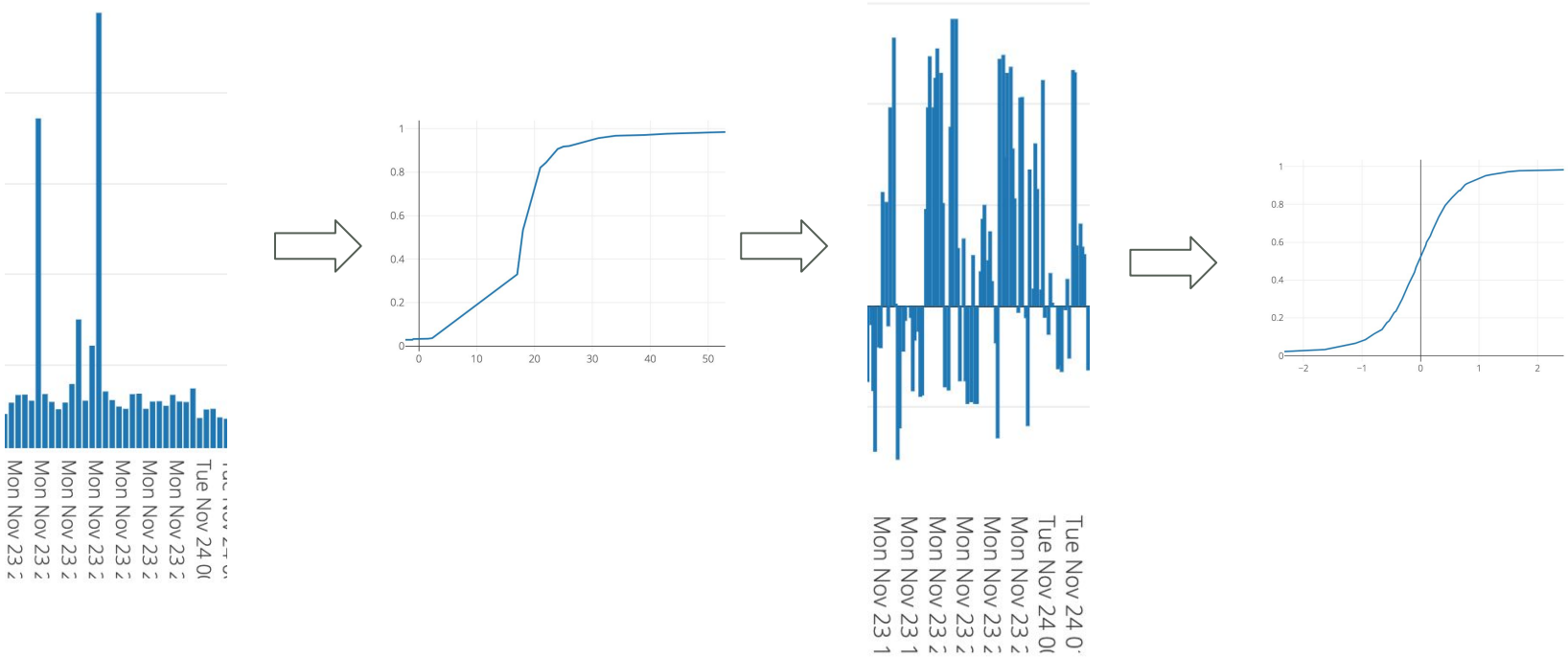
Rank	MUID	Points

Lagged Values

Timestamp	Data
7/8 08:54:09	-0.67467
7/8 08:47:09	-0.67589
7/8 08:40:09	-0.67386
7/8 08:33:08	-0.68115
7/8 08:26:08	-0.68034
7/8 08:22:12	-0.68155

Optics Analogy

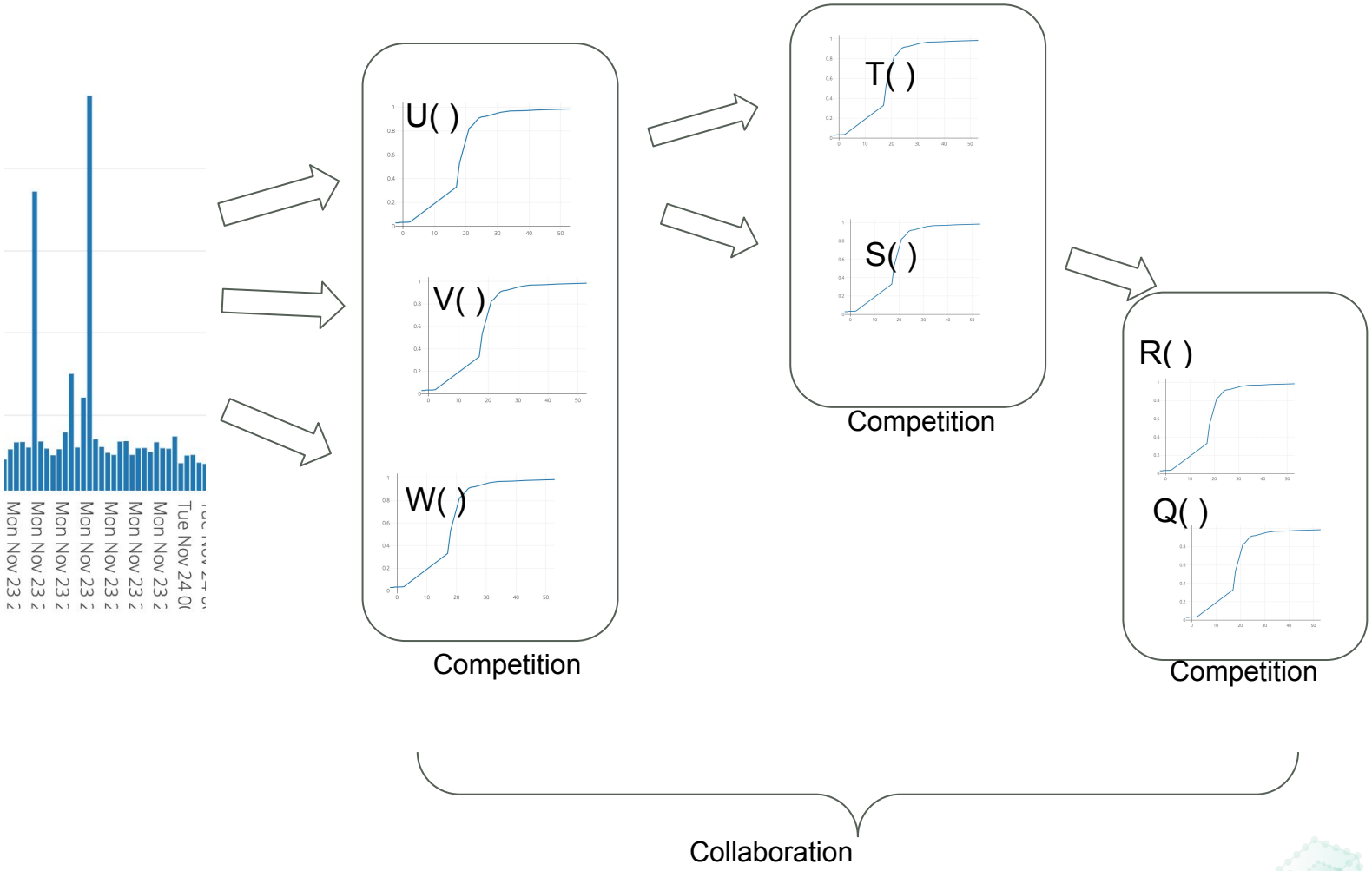
Keep “lensing” until you get N(0,1)



Composition of monotone functions, each contributed by one or more algorithms

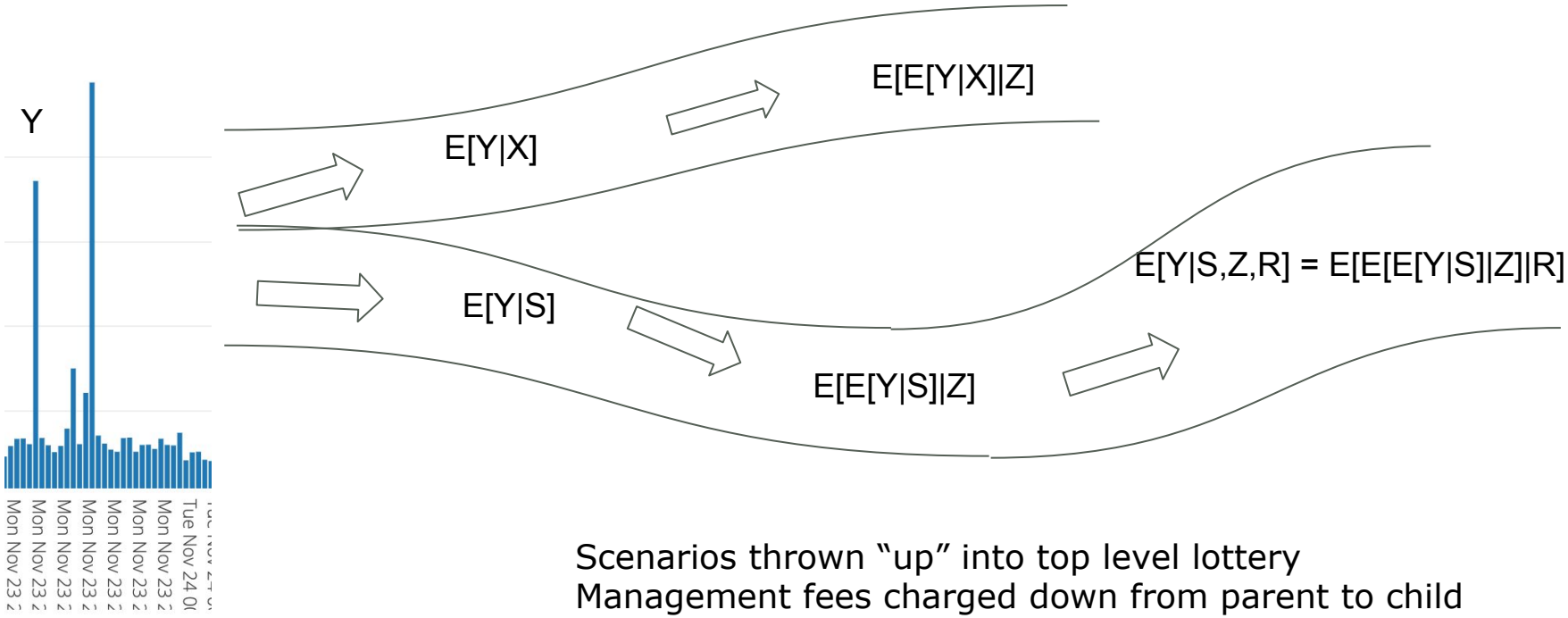
Pathways in the Collective Probability Brain

Scenarios “thrown” up to top level lottery



Law of Iterated Expectations

Pathways grow and shrink based on the economics



Scenarios thrown "up" into top level lottery
Management fees charged down from parent to child

Point estimates are a special case - shift
Exogenous data is a special case - shift arbitrarily

Wanna Play?

```
from microprediction import MicroCrawler

if __name__ == '__main__':
    crawler = MicroCrawler(difficulty=9)
    crawler.run()
```

(Modify the crawler to use whatever analytics you like)