Disclaimer

The views expressed in this document are the author's and do not necessarily reflect those of the organizations he is affiliated with.

No investment decision or particular course of action is recommended by this presentation from the author

Machine Learning for Factor Investing: From theory to production

Tony Guida, Executive Director – Senior Quantitative PM/Research

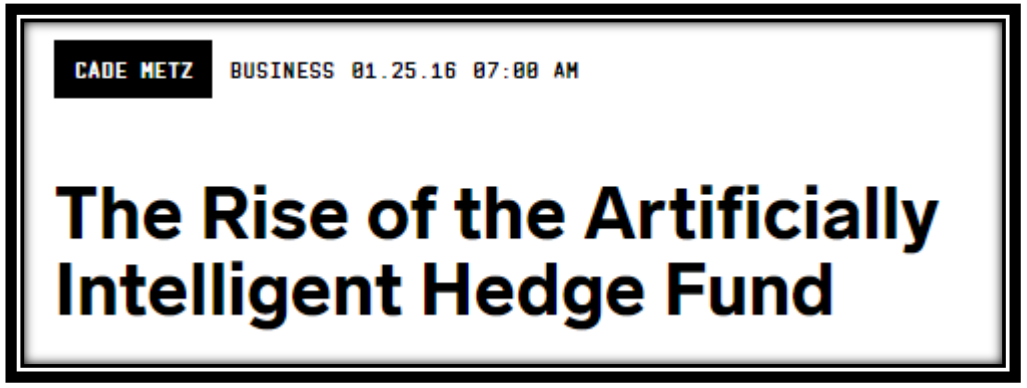
Editor in Chief : Journal of Machine Learning for Finance

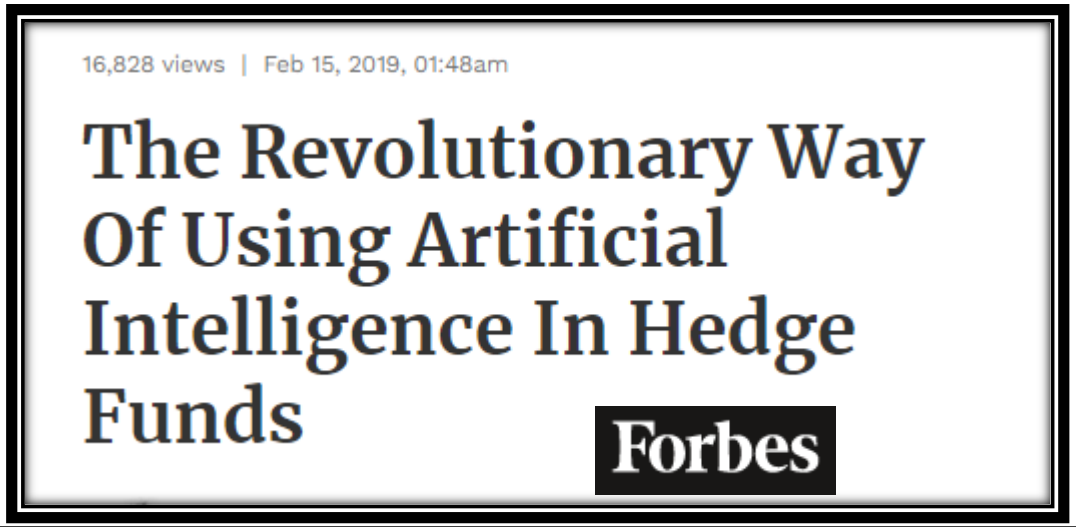
Why ML in production could be dangerous?

(a) Husky classified as wolf

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin Published in HLT-NAACL Demos 2016

Buzzword and misinterpretation





Buzzword and misinterpretation

Investing

AI Isn't Ready to Take Fund Manager Jobs Yet

Strategies based on artificial intelligence have underperformed as swings in investor sentiment have befuddled machines as well as humans this year.

By Ksenia Galouchko
October 2, 2019, 6:01 AM GMT+2

Data Science / AI / ML

Why machine learning hasn't made investors smarter

Why Most Companies Are Failing at Artificial Intelligence: Eye on A.I.

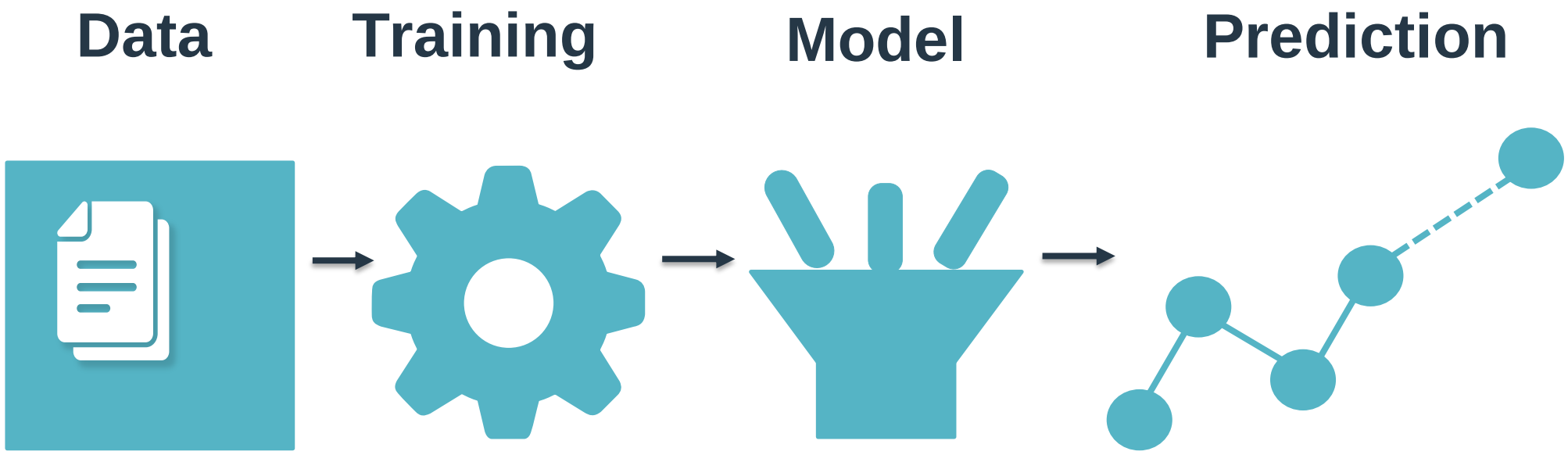
BY JONATHAN VANIAN
October 15, 2019 5:55 PM EST

Artificial Intelligence

AI's failure to live up to the hype is starting to put off investors

Investor enthusiasm for AI will wane with the first big failures – and it will be up to the industry to redefine the problems it is trying to solve

Generic definition of “Machine Learning”



Why this topic?

Given the **exponential increase in data availability**, the obvious temptation of any asset manager is to try to infer future returns from the abundance of attributes available at the firm level.

Current computational power allows to “test” almost all types of new characteristics/signals.

Sharing knowledge effect. Cross fertilization between Hard science and finance is increasing (implicitly and explicitly).

A need to innovate. Legacy approach for constructing Style/Factor equity portfolio has been delivering less return than 10 years ago.

What can we expect from ML in Factor Investing?

To test **more** characteristics/signals

To leverage on **non-linear** complex patterns, rule based

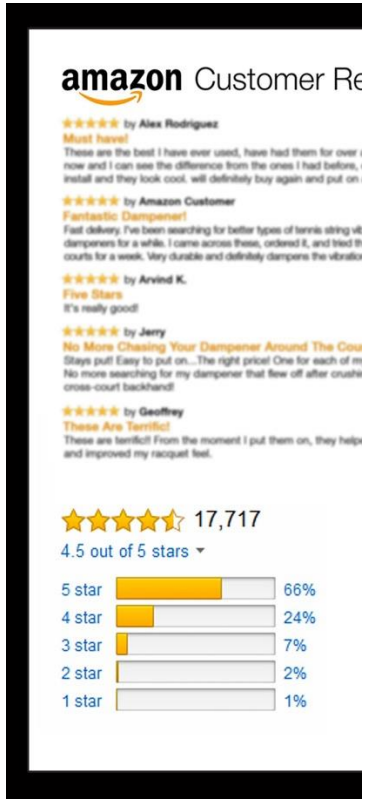
To **adapt and identify** to trends by re-running models

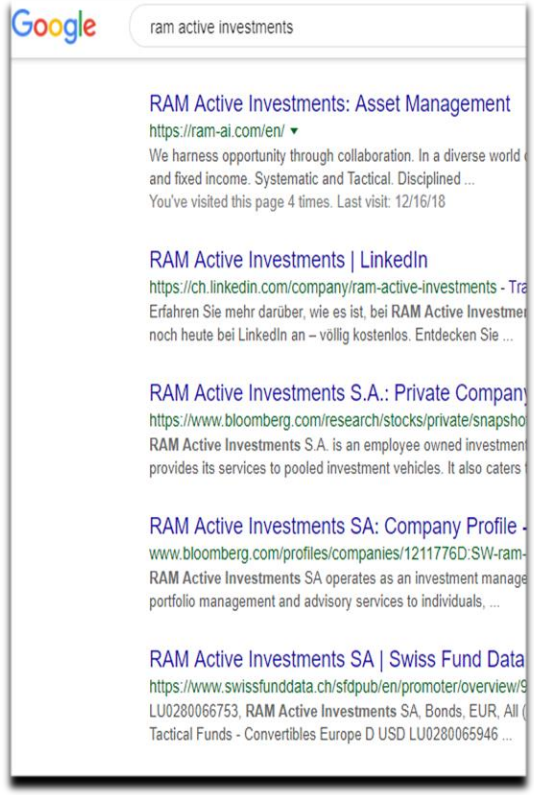
To **ensemble** more models, wisdom of the crowd

To be **less biased** than trad. dogmatic quant. approach

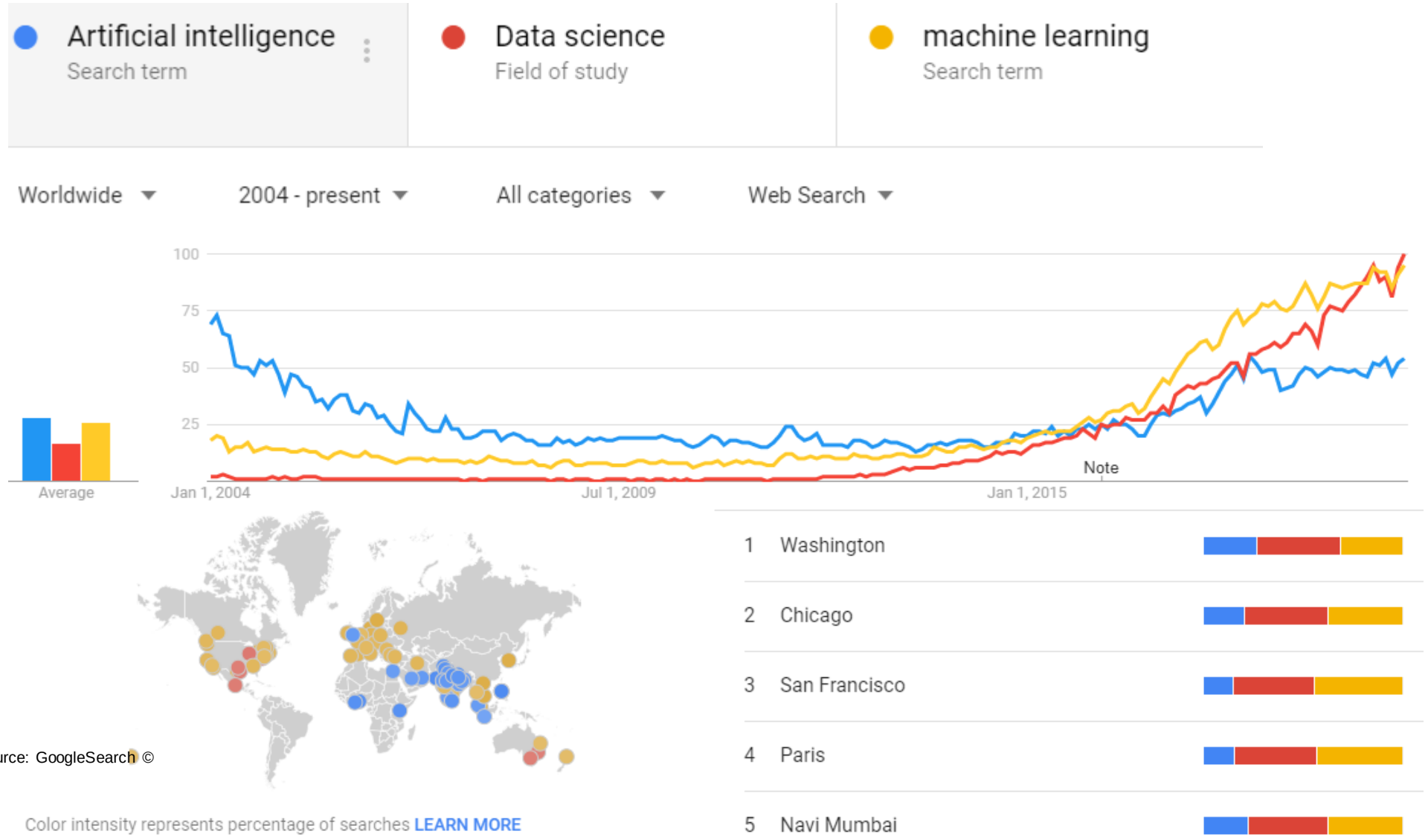
Table of content

A society much more digital due to Technology



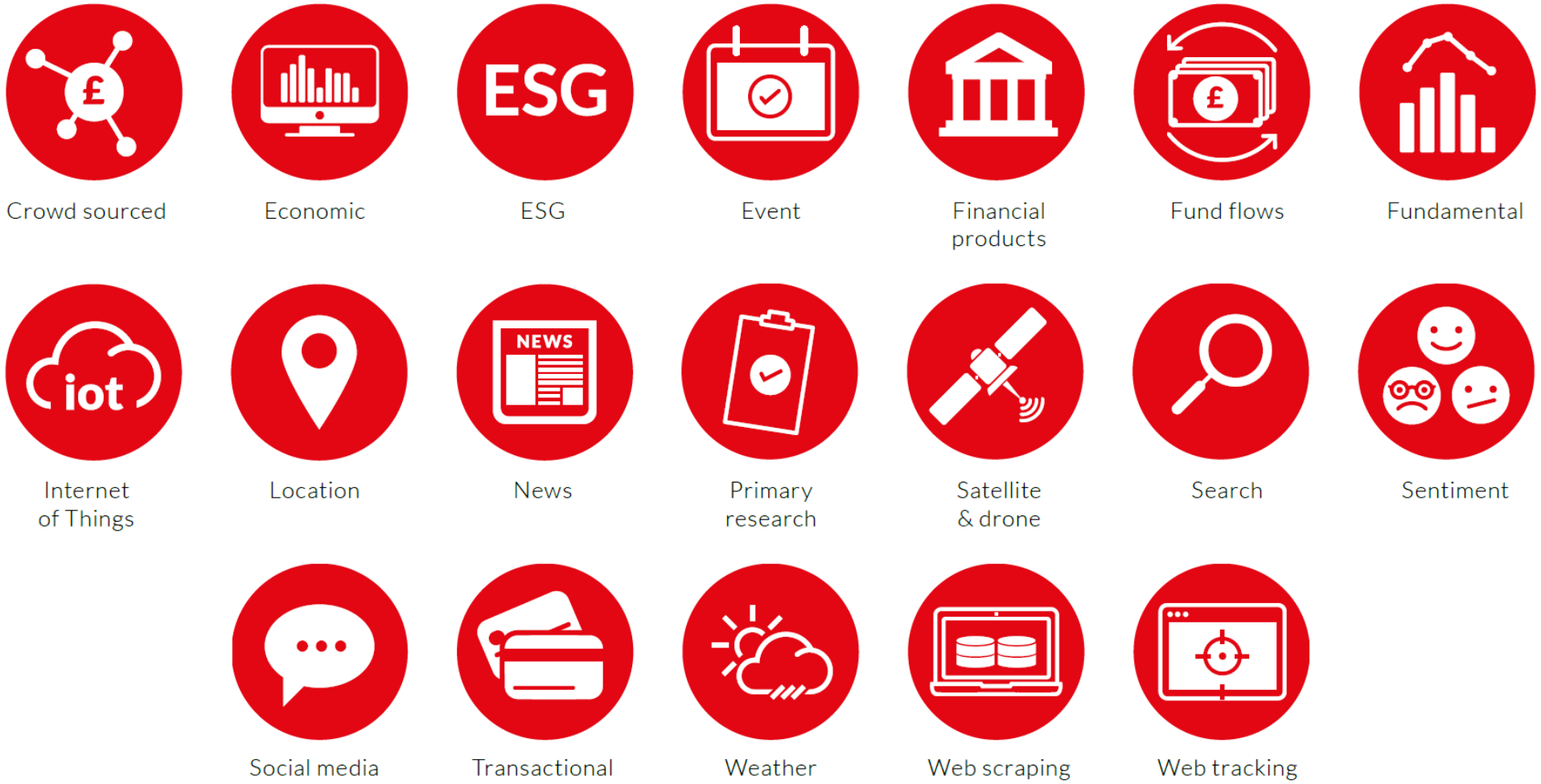


Current global interest in ML



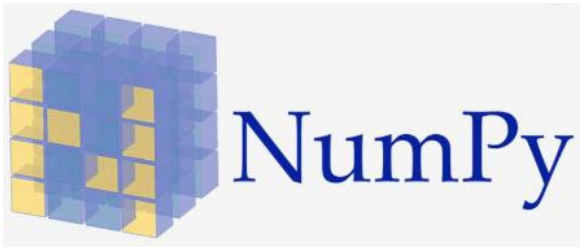
Source: GoogleSearch ©

Alternative/Big data



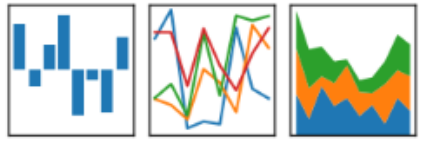
Source: Neudata ©

Open sourced tools



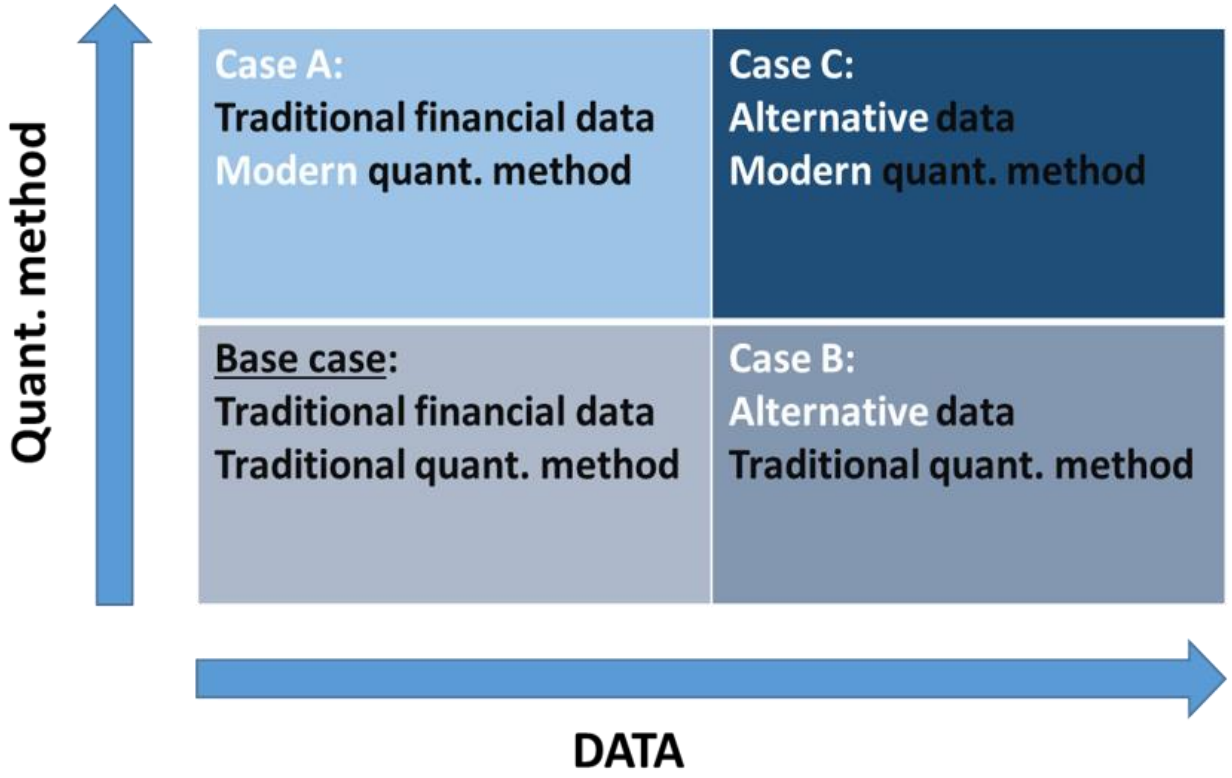
pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



seaborn

Where are we ? Law of diffusion of innovation



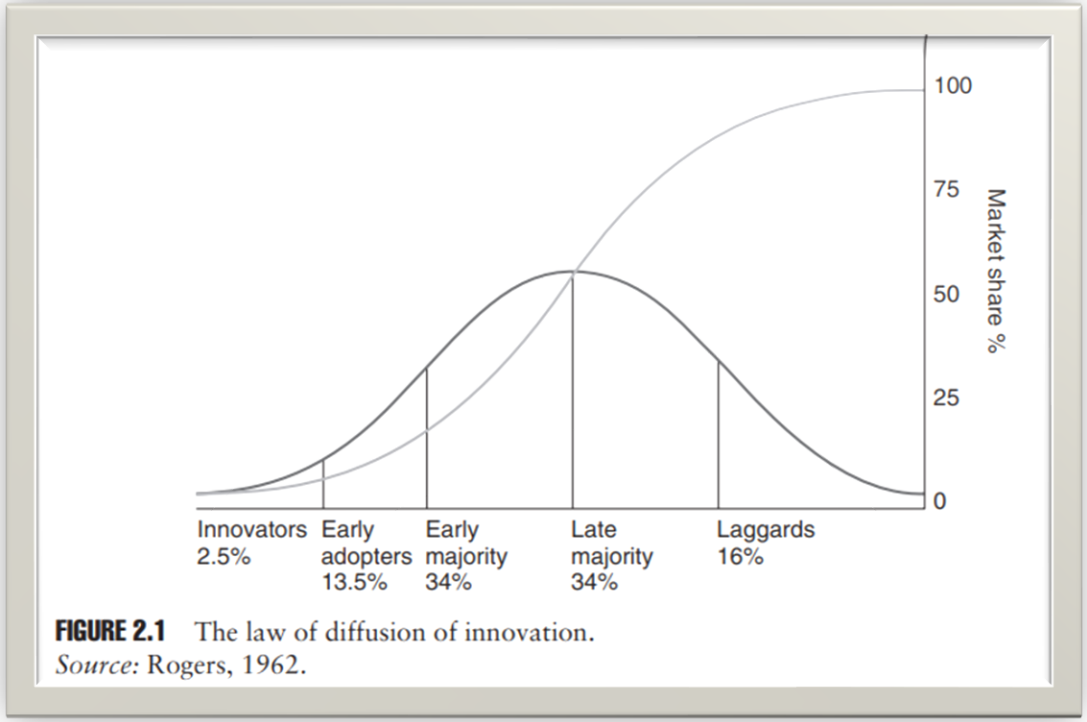


FIGURE 2.1 The law of diffusion of innovation.
Source: Rogers, 1962.

Chapter 2 -Taming Big Data – "Big Data and Machine Learning in Quantitative Investment"

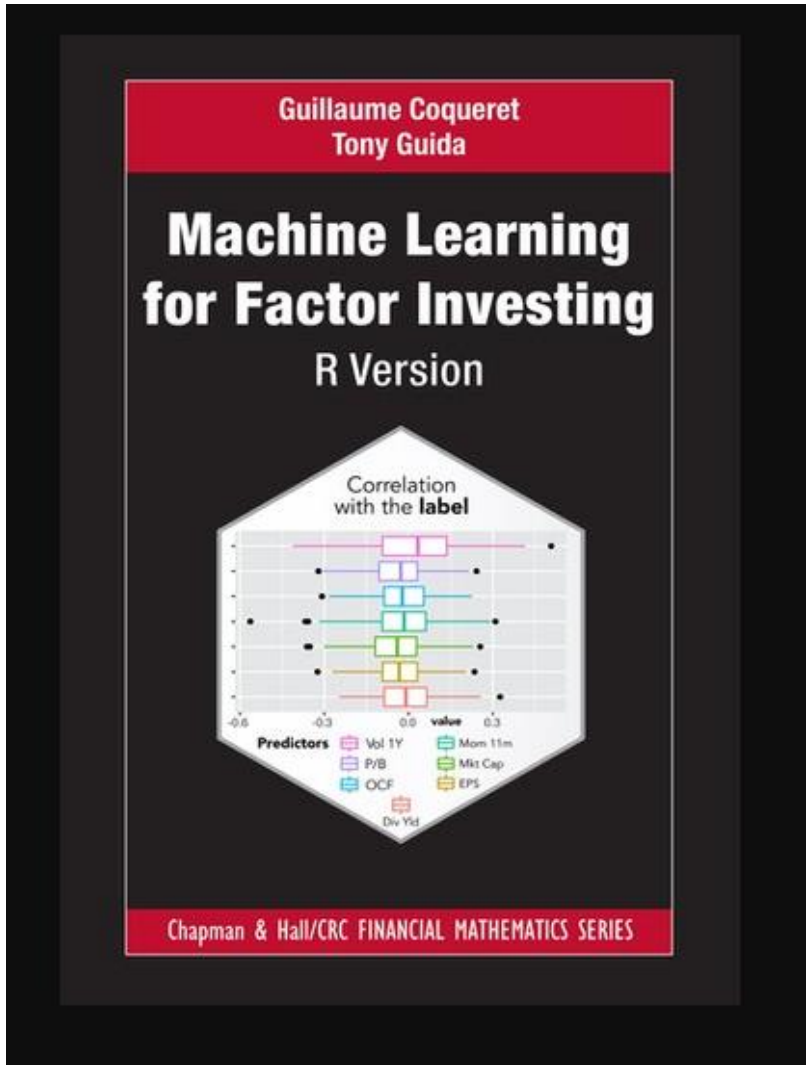
Beyond Machine Learning buzzword

Intelligence is NOT consciousness

Technology is NEVER deterministic

- You have to : have a ML "**Why**"
- You have to : **frame** the question, put financial and economic structure in dataset. **Your** domain knowledge.
- You have to : **understand** and know your **Data**

Machine Learning for Factor Investing book



Chapter 6 Tree-based methods

Classification and regression trees are simple yet powerful clustering algorithms popularized by the monograph of Breiman et al. (1984). Decision trees and their extensions are known to be quite efficient forecasting tools when working on tabular data. A large proportion of winning solutions in ML contests (especially on the Kaggle website[12]) resort to improvements of simple trees. For instance, the meta-study in bioinformatics by Olson et al. (2018) finds that boosted trees and random forests are the top 2 algorithms from a group of 13, excluding neural networks.

Recently, the surge in Machine Learning applications in Finance has led to multiple publications that use trees in portfolio allocation problems. A long, though not exhaustive, list includes: Ballings et al. (2015), Patel, Shah, Thakkar, and Kotecha (2015a), Patel, Shah, Thakkar, and Kotecha (2015b), Moritz and Zimmermann (2016), Krauss, Do, and Huck (2017), Gu, Kelly, and Xiu (2020b), Guida and Coqueret (2018a), Coqueret and Guida (2020) and Simonian et al. (2019). One notable contribution is Bryzgalova, Pelger, and Zhu (2019) in which the authors create factors from trees by sorting portfolios via simple trees, which they call *Asset Pricing Trees*.

In this chapter, we review the methodologies associated to trees and their applications in portfolio choice.

Workflow for ML Based portfolio construction

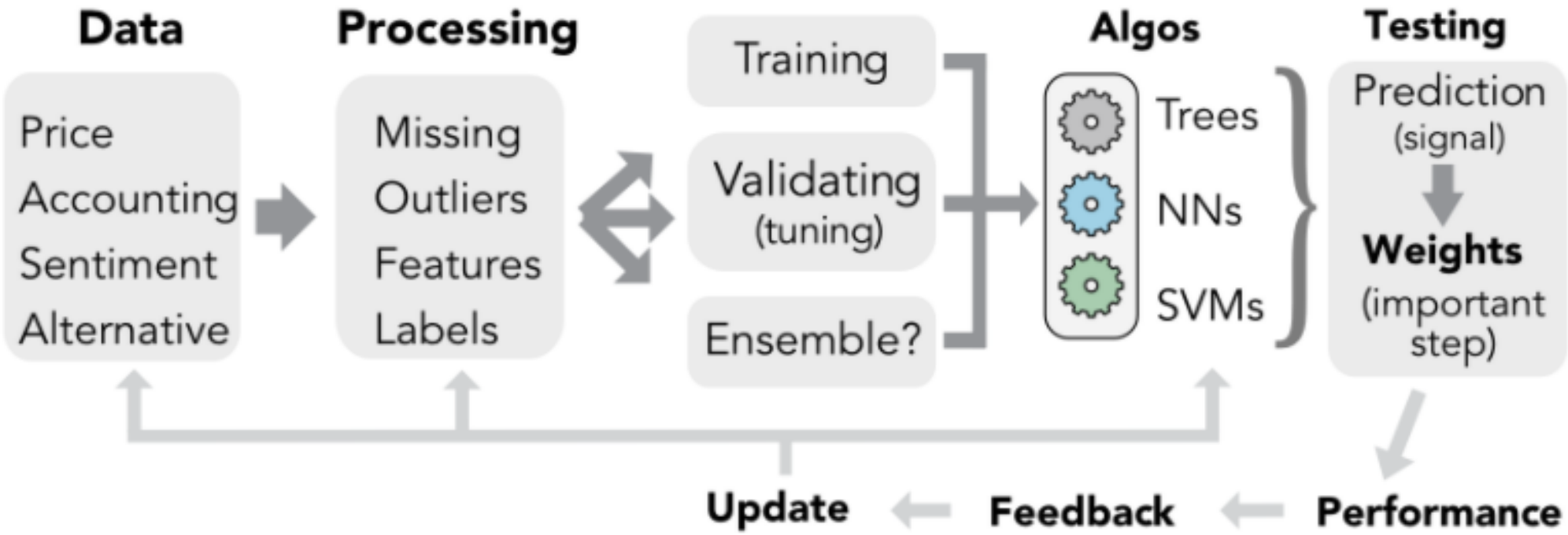
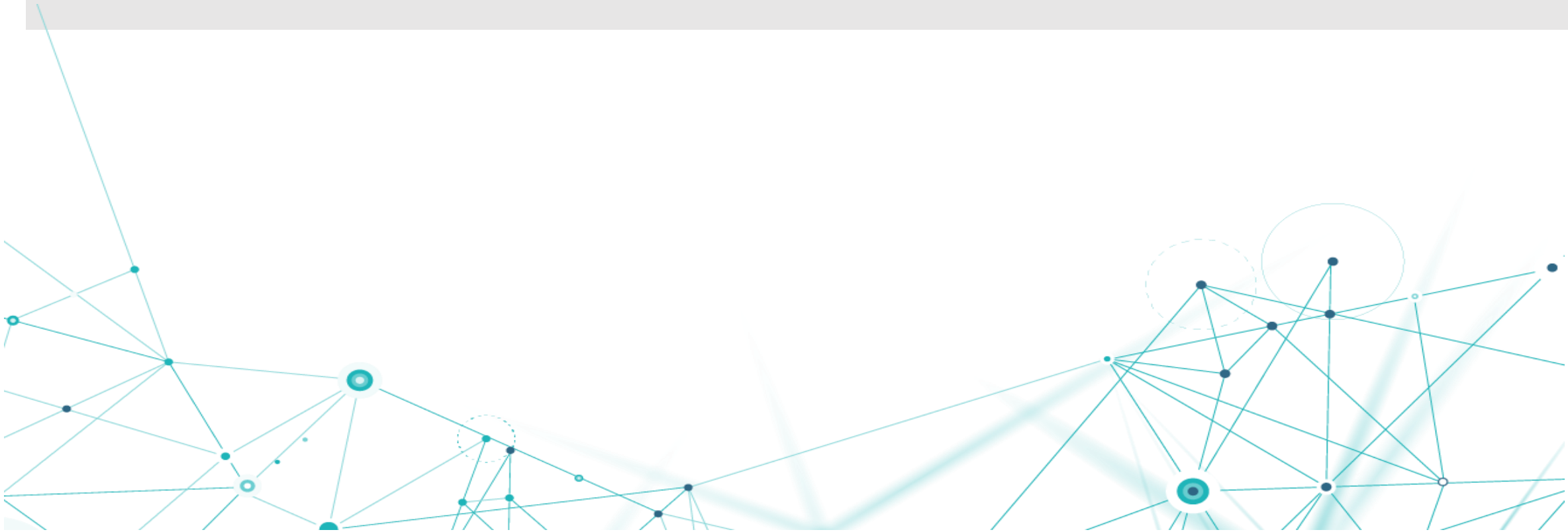


FIGURE 2.1: Simplified workflow in ML-based portfolio construction.

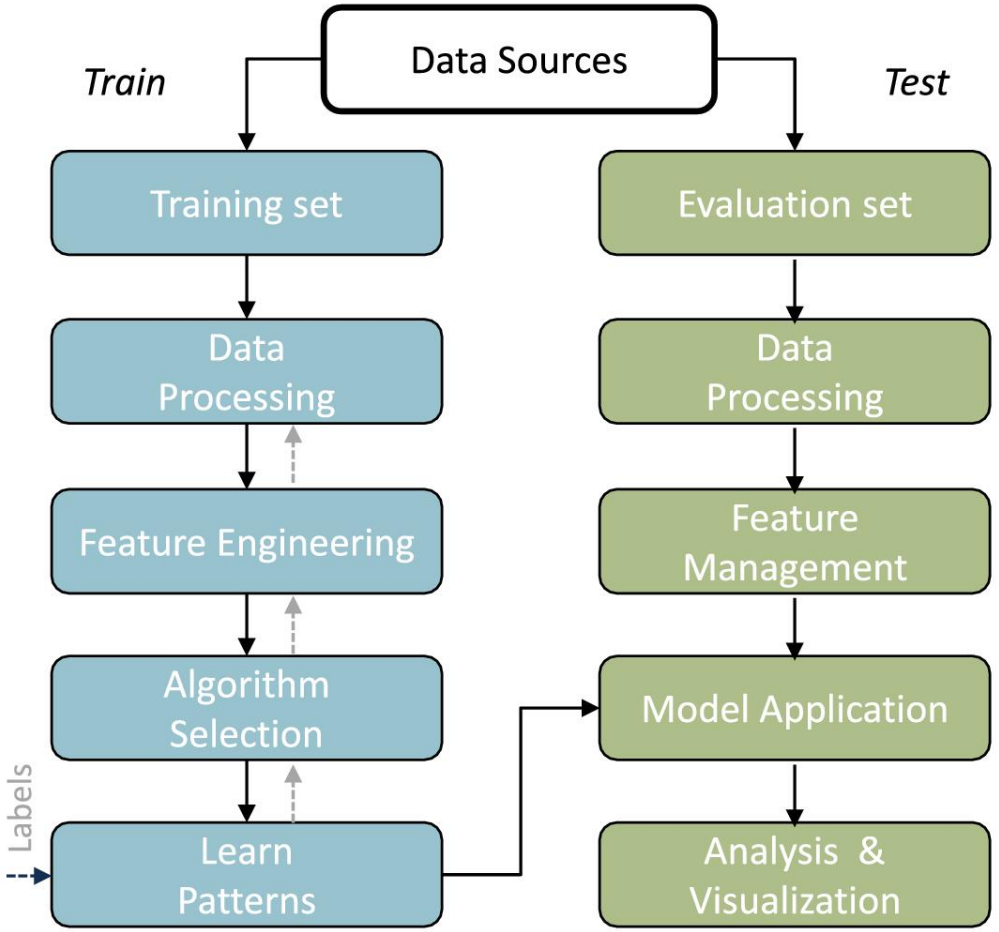
1: E.D.A



Exploratory Data Analysis

- Data curation, outliers check, missing data
- Data distribution
- Data correlation, causality
- Data visualisation
- Etc…

Step back simple example of ML pipeline



Danish Rafique, Luis Velasco, Michael Frosz, "Machine Learning for Network Automation: Overview, Architecture, and Applications [Invited Tutorial]," J. Opt. Commun. Netw. **10,** D126-D143 (2018); https://www.osapublishing.org/jocn/abstract.cfm?uri=jocn-10-10-d126

Creating the dataset

Instances >>> ~ 600 000

Results | Messages

	DataDate	dateRetumPerf	SecID	Fact2	Fact3	Fact4	Fact7	Fact8	Fact9	Fact10	Fact11	Fact12	Fact13	Fact14	Fact16	Fact17	Fact18	Fact20	Fact21	Fact22	Fact23	Fact25	Fact26	Fact27
1	2010-12-31	2011-12-31	1195515867	19	70	94	94	13	99	15	34	55	65	42	32	28	21	27	31	49	38	53	53	76
2	2014-08-31	2015-08-31	570191681	44	NULL	83	48	11	7	19	72	72	28	33	32	31	21	30	22	31	51	13	14	49
3	2012-01-31	2013-01-31	290849864	55	66	98	15	3	100	7	2	2	58	43	14	38	25	19	22	11	11	10	11	65
4	2012-02-29	2013-02-28	324622763	69	69	99	27	4	100	7	2	1	65	42	14	39	16	18	21	6	6	13	15	62
5	2012-08-31	2013-08-31	1528063821	73	70	99	61	4	99	14	2	2	47	41	19	38	16	25	26	43	45	83	84	50
6	2012-03-31	2013-03-31	1850474528	61	67	99	27	4	100	11	2	1	60											
7	2012-11-30	2013-11-30	2021474168	75	74	99	60	4	NULL	13	2	10	47											
8	2017-01-31	2018-01-31	408403206	42	31	61	68	10	2	26	94	92	14											
9	2013-01-31	2014-01-31	1593812596	65	73	99	53	4	NULL	18	3	10	47											
10	2016-05-31	2017-05-31	352206094	NULL	NULL	NULL	NULL	NULL	74	NULL	77	77	33											
11	2009-03-31	2010-03-31	1251799406	55	65	49	30	10	100	11	3	1	83											
12	2007-02-28	2008-02-29	1092089186	36	69	21	97	8	90	10	74	63	33											
13	2016-09-30	2017-09-30	1171513832	30	32	60	67	10	100	27	89	83	13											
14	2003-07-31	2004-07-31	1668572152	48	30	70	35	8	91	30	3	1	41											
15	2010-08-31	2011-08-31	1317881264	20	4	87	96	15	99	10	29	56	62											
16	2008-02-29	2009-02-28	31085621	46	81	19	24	7	63	5	77	62	51											
17	2009-04-30	2010-04-30	512957258	52	62	50	32	10	100	9	3	1	84											
18	2012-04-30	2013-04-30	2143460900	67	68	99	27	4	100	10	2	1	60											
19	2016-10-31	2017-10-31	1415589206	20	NULL	NULL	64	11	58	35	79	73	15											
20	2011-01-31	2012-01-31	486144046	21	70	95	95	13	99	12	35	55	65											
21	2013-03-31	2014-03-31	156902714	54	64	99	67	1	100	19	5	35	36											
22	2009-08-31	2010-08-31	290508352	29	3	53	94	18	100	19	1	1	71	55	43	28	33	45	37	1	1	47	58	74
23	2010-09-30	2011-09-30	1527687499	20	4	87	96	15	99	12	29	56	62	42	30	35	29	32	34	31	26	40	43	75
24	2016-06-30	2017-06-30	2006847672	NULL	NULL	NULL	NULL	NULL	80	NULL	75	74	23	NULL	41	56	37	30	26	74	64	87	82	50
25	2007-12-31	2008-12-31	156190601	47	84	12	83	7	100	8	90	77	53	72	57	60	62	64	72	31	30	18	15	29
26	2014-07-31	2015-07-31	1573508350	48	NULL	89	84	12	6	19	71	71	31	32	31	32	25	31	23	35	61	51	61	53
27	2007-06-30	2008-06-30	1049253878	38	70	26	97	6	96	10	81	70	32	18	67	23	79	75	75	34	42	71	71	30
28	2007-03-31	2008-03-31	400280762	36	70	21	97	8	90	11	74	62	32	17	68	53	81	81	74	34	51	71	77	30
29	2016-08-31	2017-08-31	1510213818	32	30	60	70	10	100	24	89	83	14	23	44	58	29	36	28	83	56	94	88	51
30	2011-02-28	2012-02-29	231369713	19	65	95	58	15	100	11	23	45	63	48	31	34	26	24	31	57	59	60	62	78
31	2008-04-30	2009-04-30	474472098	45	81	20	24	6	65	6	78	62	51	77	54	34	53	59	67	14	10	38	26	26

Some features examples

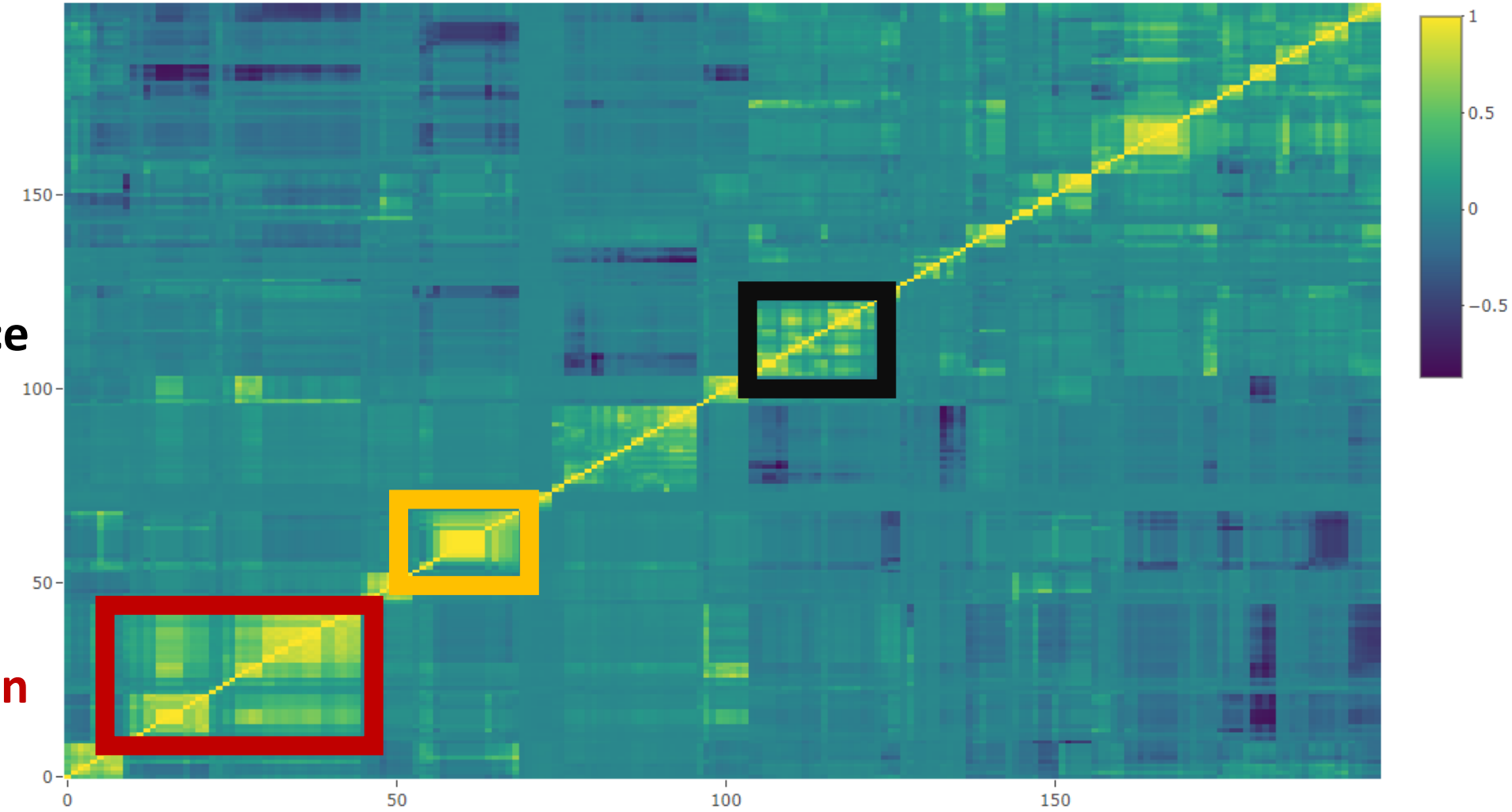
- *Fundamental trailing*
- *Price based*
- *Volume based*
- *Risk based*
- *Composites*

Features correlation example

Risk Price based

Volume

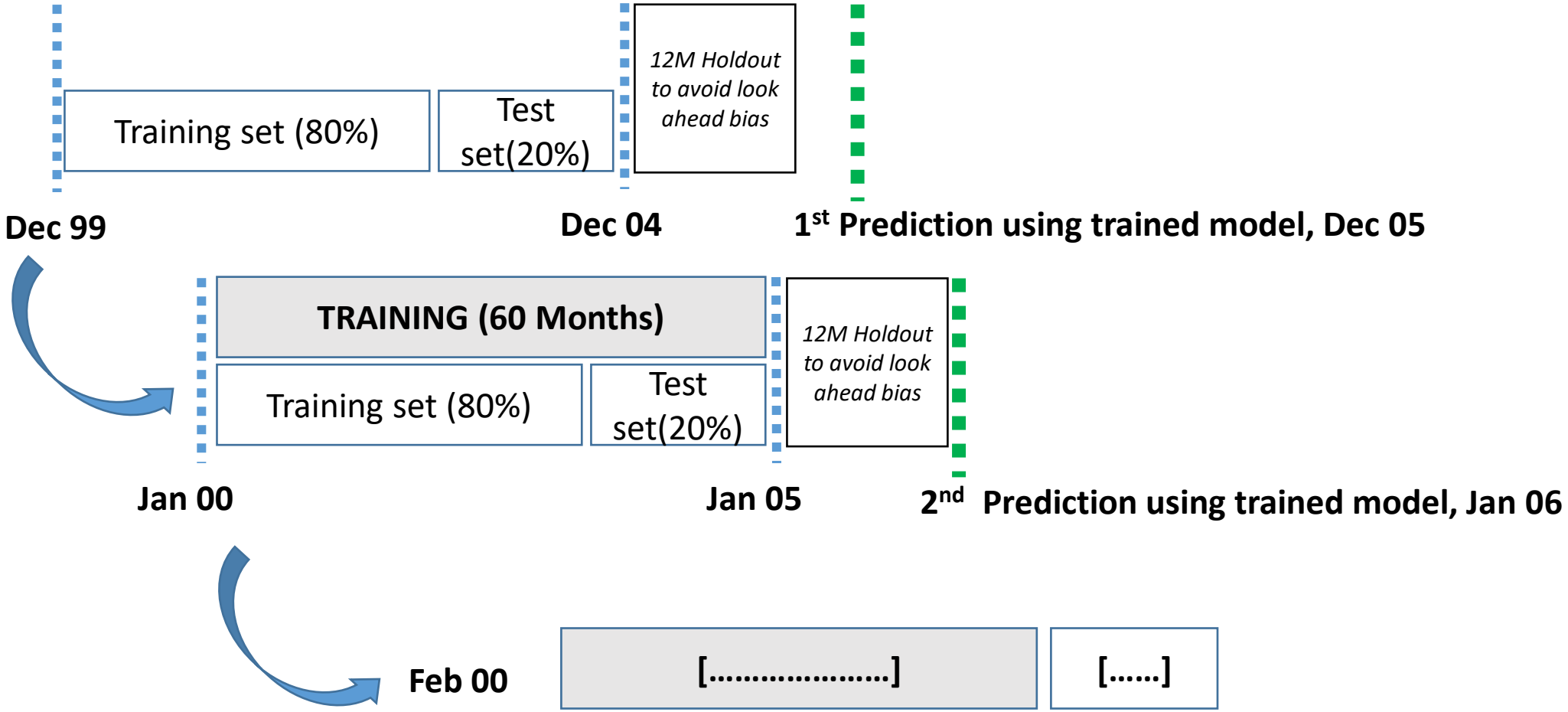
Valuation



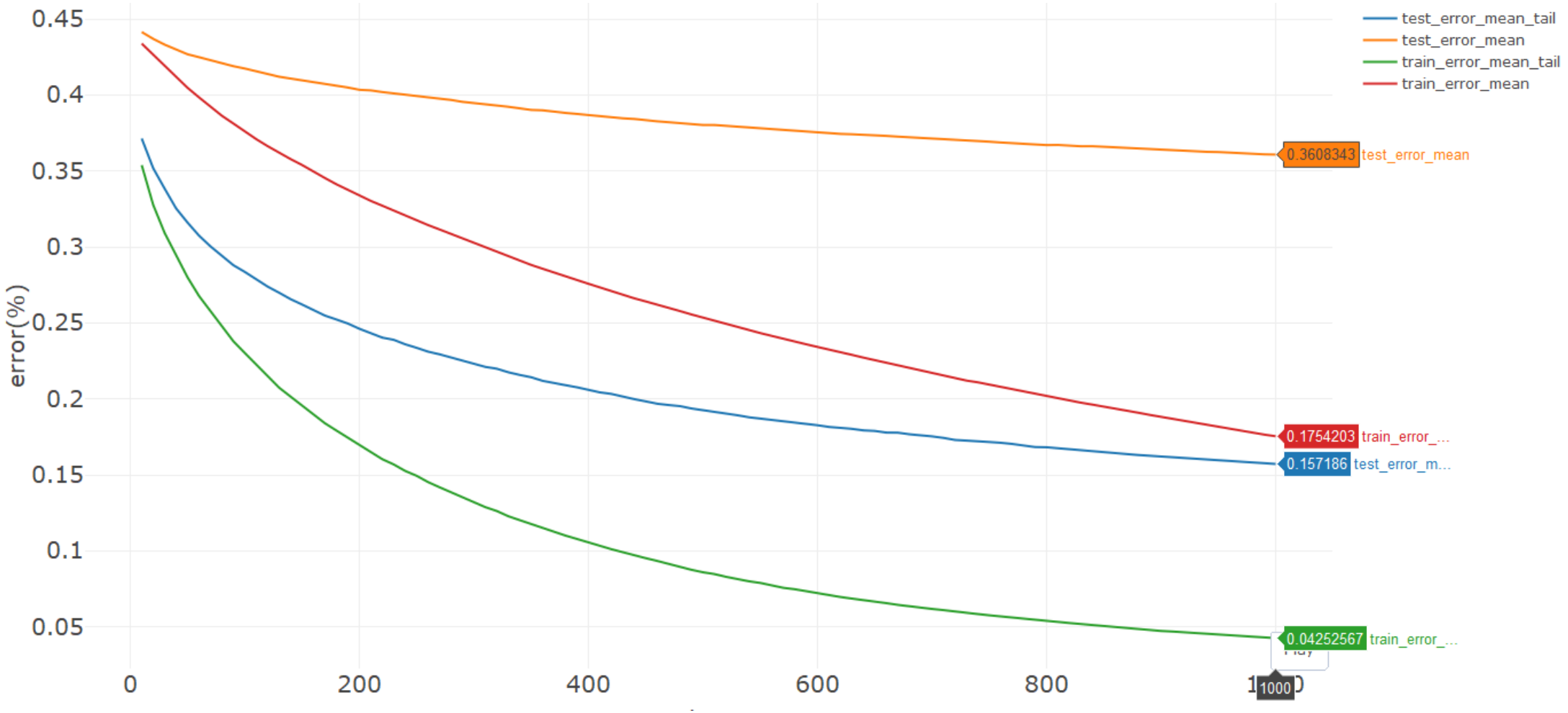
Source: Guida, Coqueret. Chapter 7, Ensemble Learning Applied to Quant Equity – Big Data and Machine Learning in Quantitative Investment

Rolling Windows for training (case for 12M forward)

In this example we use a rolling window of **60 months** to predict the **12M forward performance** of a stock.



What tails training does on accuracy



Source: Hypothetical exercice based on a different yet similar datasets (World including EM)

2: Empirical ML Asset Pricing

First and deadliest enemy: EGO

- Forward Looking Data
- Backtest overfitting
- Stale training
- Model complexity
- Features selection bias
- Etc….

Example of Backtesting protocol

A Backtesting Protocol in the Era of Machine Learning*

Rob Arnott

Research Affiliates, Newport Beach, CA 92660, USA

Campbell R. Harvey

Fuqua School of Business, Duke University, Durham, NC 27708, USA
National Bureau of Economic Research, Cambridge, MA 02912, USA

Harry Markowitz

Harry Markowitz Company, San Diego, CA 92109, USA

ABSTRACT

Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning's early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, which may be influenced by others' actions and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. While investors need to be cautious—indeed, more cautious than in past applications of quantitative methods—these new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.

JEL: G11, G14, G17, C11, C58

Keywords: *Machine Learning, Data Science, Data Mining, Backtesting, Overfitting, Interpretable Classification, Interpretable Policy Design, Trading, Strategies, Anomalies, Selection Bias, Research Protocol.*

Hyperparameters:

- **The learning rate**, η: it is the step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features and η actually shrinks the feature weights to make the boosting process more conservative.
- **The maximum depth**: it is the longest path (in terms of node) from the root to a leaf of the tree. Increasing this value will make the model more complex and more likely to be overfitting.
- **Regression λ**: it is the L^2 regularization term on weights (mentioned in the technical section) and increasing this value will make model more conservative.
- **gamma:** minimum loss reduction required to make a further partition on a leaf node of the tree. The larger, the more conservative the algorithm will be.

model	max_depth	eta	round	eval_metric	subsample	col_by_sample
XGB	5	1%	150	error	0.8	0.8

Source: Hypothetical exercice based on a different yet similar datasets (World including EM)

3: Added Value of Data engineering

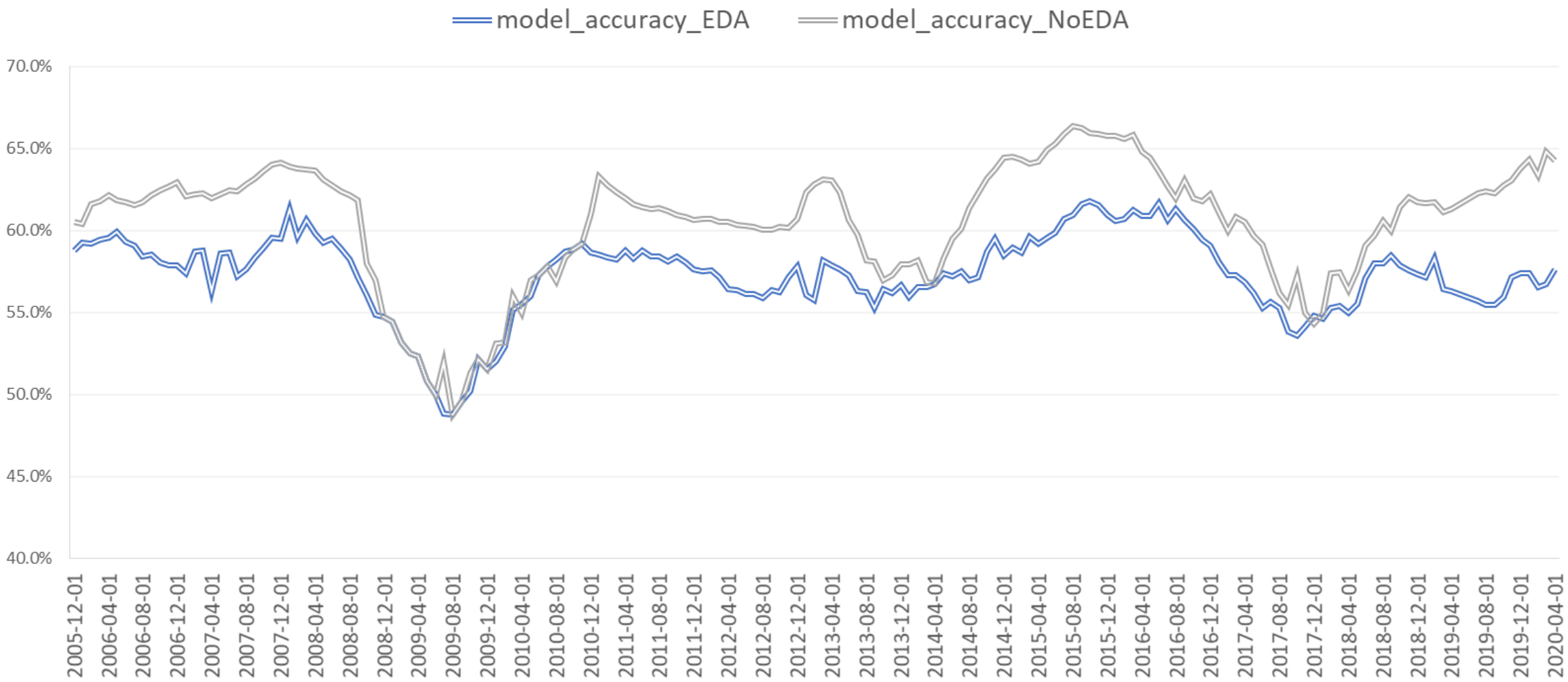
Objective, Data and Protocol

We will only predict **12M future** performance
We will use a boosted tree classification ML model
Our Investment universe is composed of **global stocks (~2000)**
Full dataset from Dec-1999 until April-2020.
Dataset contains alt and traditional data.
Data engineering from **E**xploratory **D**ata **A**nalysis (**EDA**) is based on:

- *Training on tails (extreme quantile from Label/fit cross section) training*
- *Outliers removals*
- *Low-coverage instance (row) removal*
- *Low-coverage feature (column) removal*

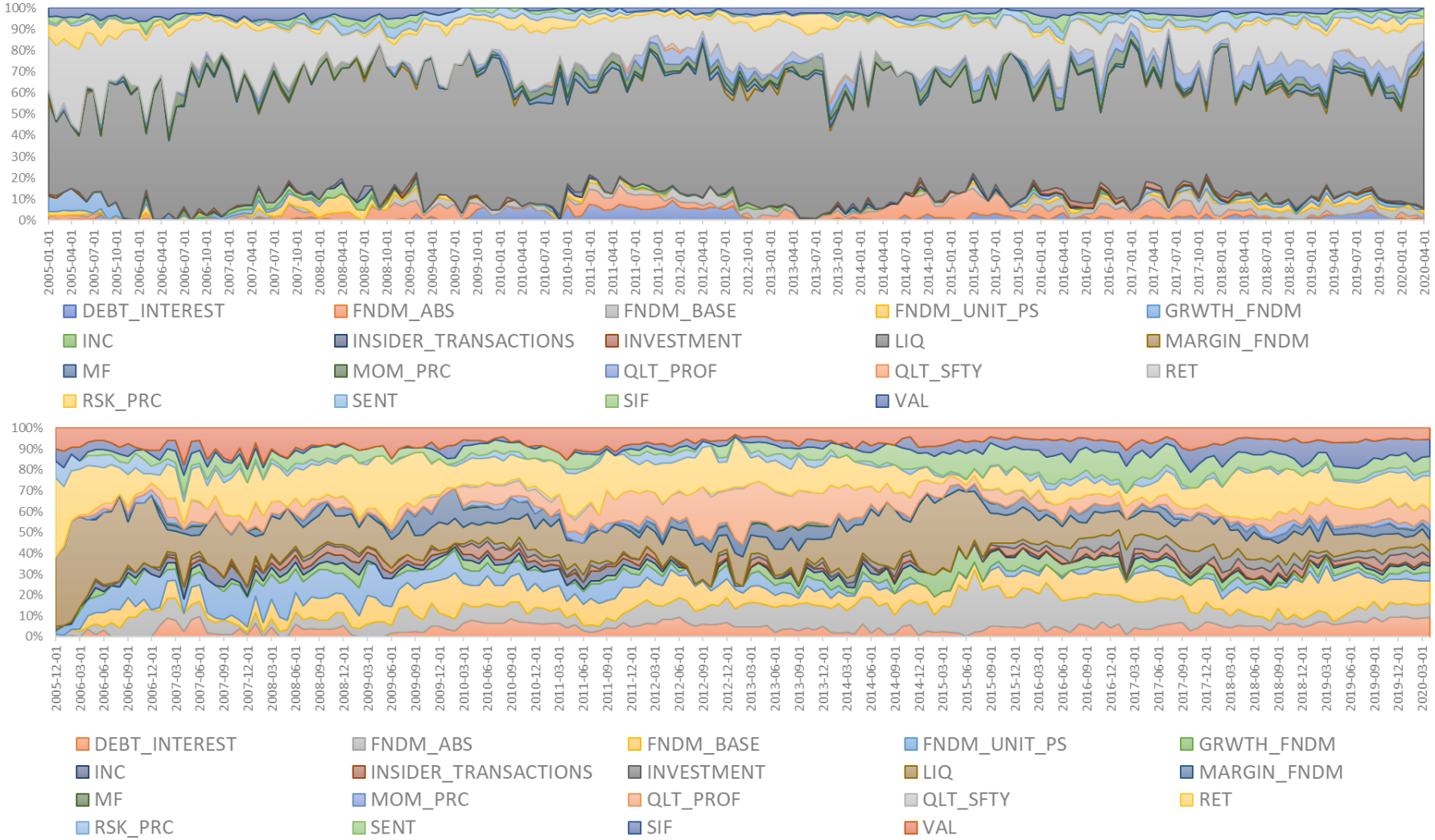
- **(~ 400) features**, monthly normalised in percentile
- We use a rolling window of 5 years- **80% Training 20% Testing**
- Compare the results according to 3 angles : accuracy, interpretability and out of sample performance analytics where we will dig into the data engineering.

Data engineering: What's the impact of NOT doing it



Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

Normalised Feature's importance breakdown



Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

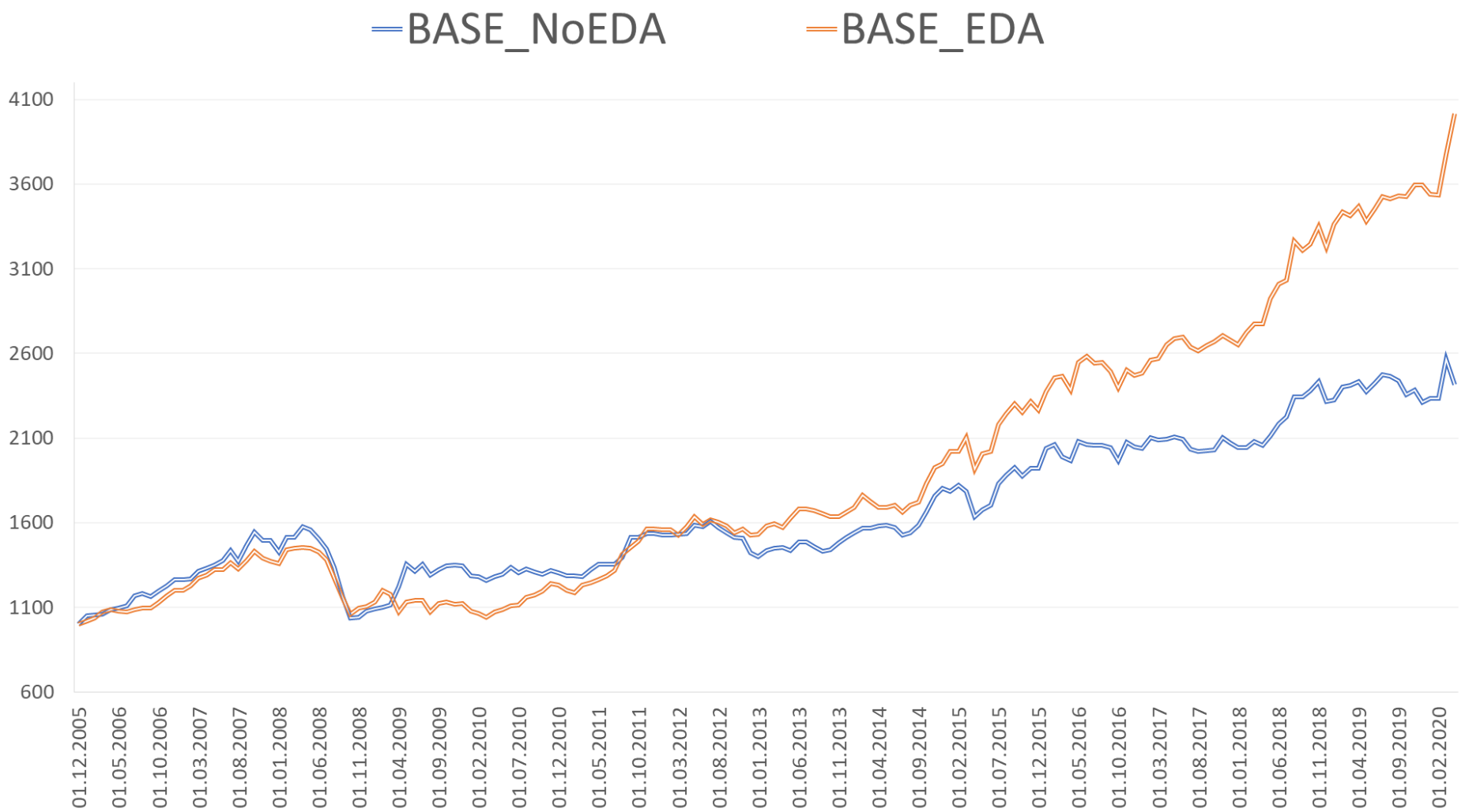
Turnover analysis for both models

BASE_EDA	Average of D1	Average of D2	Average of D3	Average of D5	Average of D4	Average of D6	Average of D7	Average of D8	Average of D9	Average of D10
2006	234%	476%	584%	611%	603%	603%	613%	546%	464%	279%
2007	293%	618%	712%	791%	772%	782%	752%	718%	635%	232%
2008	266%	581%	714%	760%	746%	750%	765%	750%	673%	364%
2009	383%	669%	800%	843%	845%	847%	835%	792%	713%	398%
2010	353%	650%	744%	741%	782%	760%	746%	684%	650%	290%
2011	255%	637%	781%	883%	812%	845%	793%	710%	633%	360%
2012	168%	535%	745%	856%	823%	852%	785%	679%	563%	380%
2013	229%	586%	763%	857%	856%	852%	814%	721%	636%	380%
2014	392%	653%	700%	752%	721%	747%	747%	688%	530%	364%
2015	314%	554%	662%	674%	694%	644%	565%	577%	445%	341%
2016	309%	583%	700%	731%	697%	677%	678%	655%	549%	322%
2017	234%	465%	620%	556%	619%	564%	629%	656%	568%	355%
2018	353%	640%	756%	745%	783%	723%	726%	705%	640%	320%
2019	429%	734%	772%	821%	807%	794%	779%	731%	685%	330%
2020 (as of end of April)	379%	704%	779%	806%	811%	819%	804%	782%	682%	335%

BASE_NoEDA	Average of D1	Average of D2	Average of D3	Average of D5	Average of D4	Average of D6	Average of D7	Average of D8	Average of D9	Average of D10
2006	688%	513%	622%	690%	654%	729%	715%	648%	586%	356%
2007	344%	653%	775%	791%	746%	804%	772%	708%	607%	282%
2008	238%	527%	790%	754%	748%	843%	804%	723%	614%	339%
2009	270%	591%	889%	846%	905%	866%	826%	776%	716%	407%
2010	586%	757%	783%	897%	870%	939%	903%	854%	731%	366%
2011	755%	739%	859%	877%	906%	898%	838%	795%	681%	394%
2012	425%	728%	618%	803%	743%	787%	821%	778%	631%	471%
2013	535%	791%	739%	789%	816%	794%	791%	661%	577%	319%
2014	730%	597%	663%	719%	698%	739%	753%	674%	600%	307%
2015	760%	677%	685%	662%	670%	621%	608%	544%	503%	279%
2016	659%	559%	609%	516%	606%	476%	530%	524%	464%	229%
2017	628%	610%	661%	605%	650%	580%	531%	569%	552%	324%
2018	588%	554%	592%	670%	626%	712%	655%	683%	650%	368%
2019	779%	781%	776%	793%	766%	844%	804%	740%	623%	329%
2020 (as of end of April)	863%	689%	751%	887%	818%	828%	821%	783%	605%	287%

Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

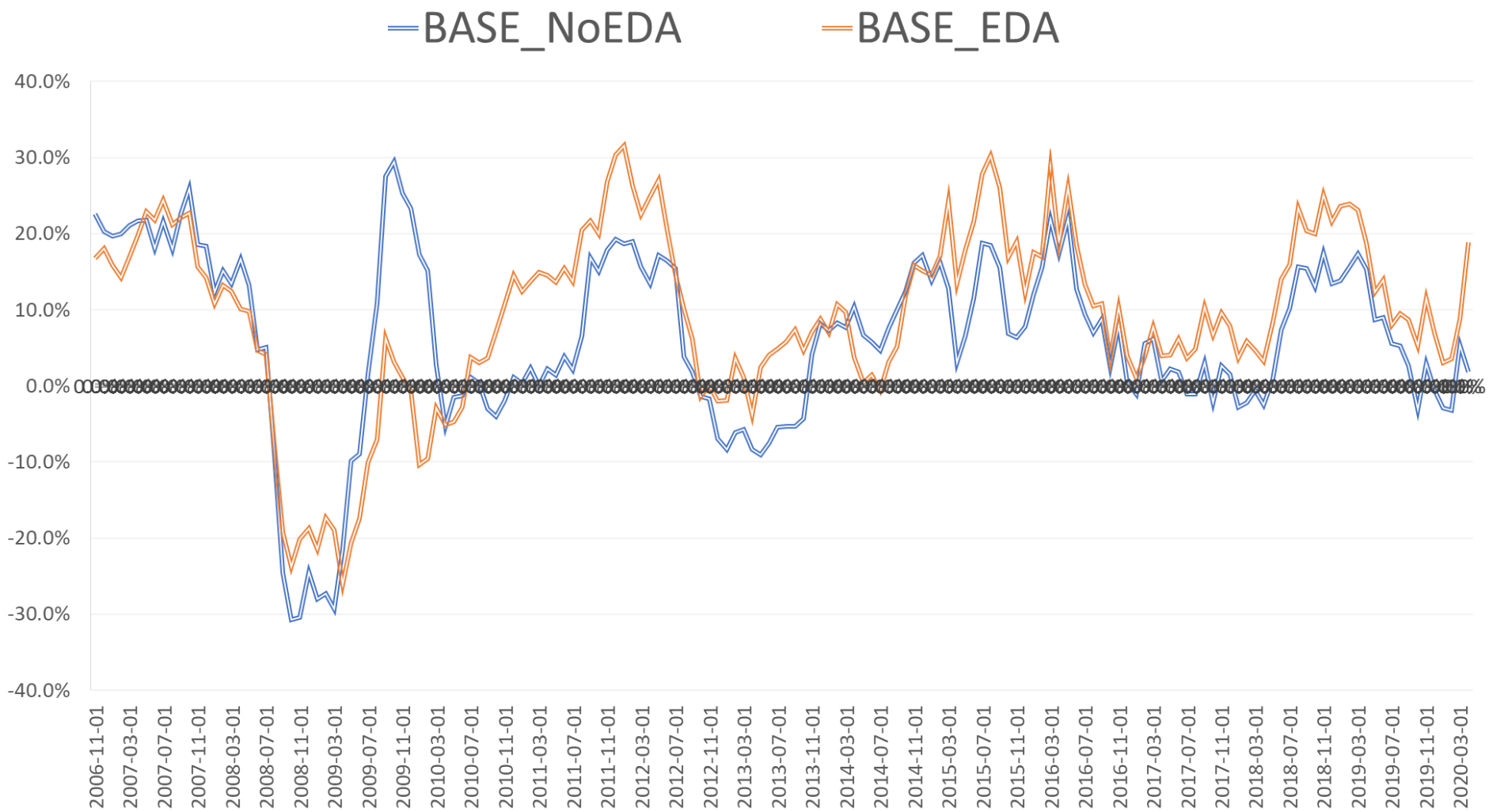
First things first: L/S performance from ML models



Performance is Gross of TC and gross of management fees. Backtested performance in USD using London fixing. It is neither an offer nor an invitation to buy or sell investment products and may not be interpreted as investment advice.

Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

Rolling 12 months performance



Performance is Gross of TC and gross of management fees. Backtested performance in USD using London fixing. It is neither an offer nor an invitation to buy or sell investment products and may not be interpreted as investment advice.

Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

Data engineering: What's the impact of NOT doing it

Portfolio Long D10/Short D1. Legs equal weighted. Cash Neutral.	Performance p.a. (USD gross of TC and mgmt. fees)	Data engineering	Volatility	Sharpe (risk free assumed zero)	Run time (seconds)	Turnover (Yearly average 2-ways)
BASE_EDA	10.4%	ALL	10.6%	0.99	720	362%
BASE_NoEDA	4.2%	NONE	10.8%	0.38	4200	450%
BASE_NoEDA1	6.7%	ONLY TAILS	12.1%	0.55	1120	469%
BASE_NoEDA2	4.8%	ONLY COVERAGE	10.7%	0.45	3800	497%
BASE_NoEDA3	5.0%	ONLY OUTLIERS	9.72%	0.51	3750	518%

4: ML Factor Investing vs. traditional Factor Investing

Protocol for ML

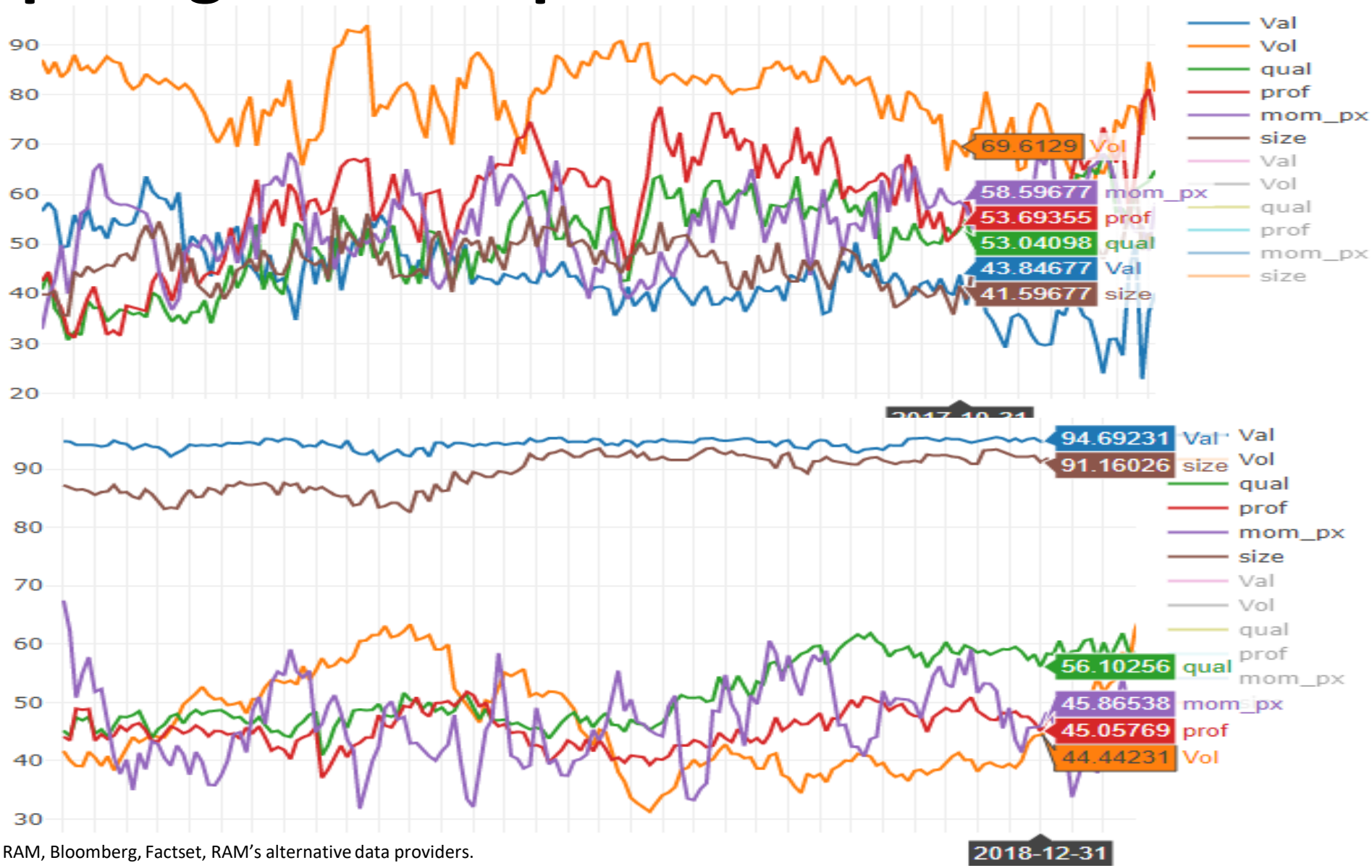
We will predict **1M future** performance
We will use a boosted tree classification ML model
Our Investment universe is composed of **global stocks including EM (~1700)**
Full dataset from Dec-1999 until May-2020. **Style datasets** are subpart of the global.
Stocks are filtered according absolute and relative metrics for **MCAP** and **ADV**.
Data engineering for training is based on:

- *Training on tails (extreme quantile from Label/fit cross section) training*
- *Outliers removals (from label and features)*
- *Low-coverage instance (row) removal*
- *Low-coverage feature (column) removal*

- **(~ 200) features**, monthly normalised in percentile
- We use a rolling window of 5 years- **80% Training 20% Testing**

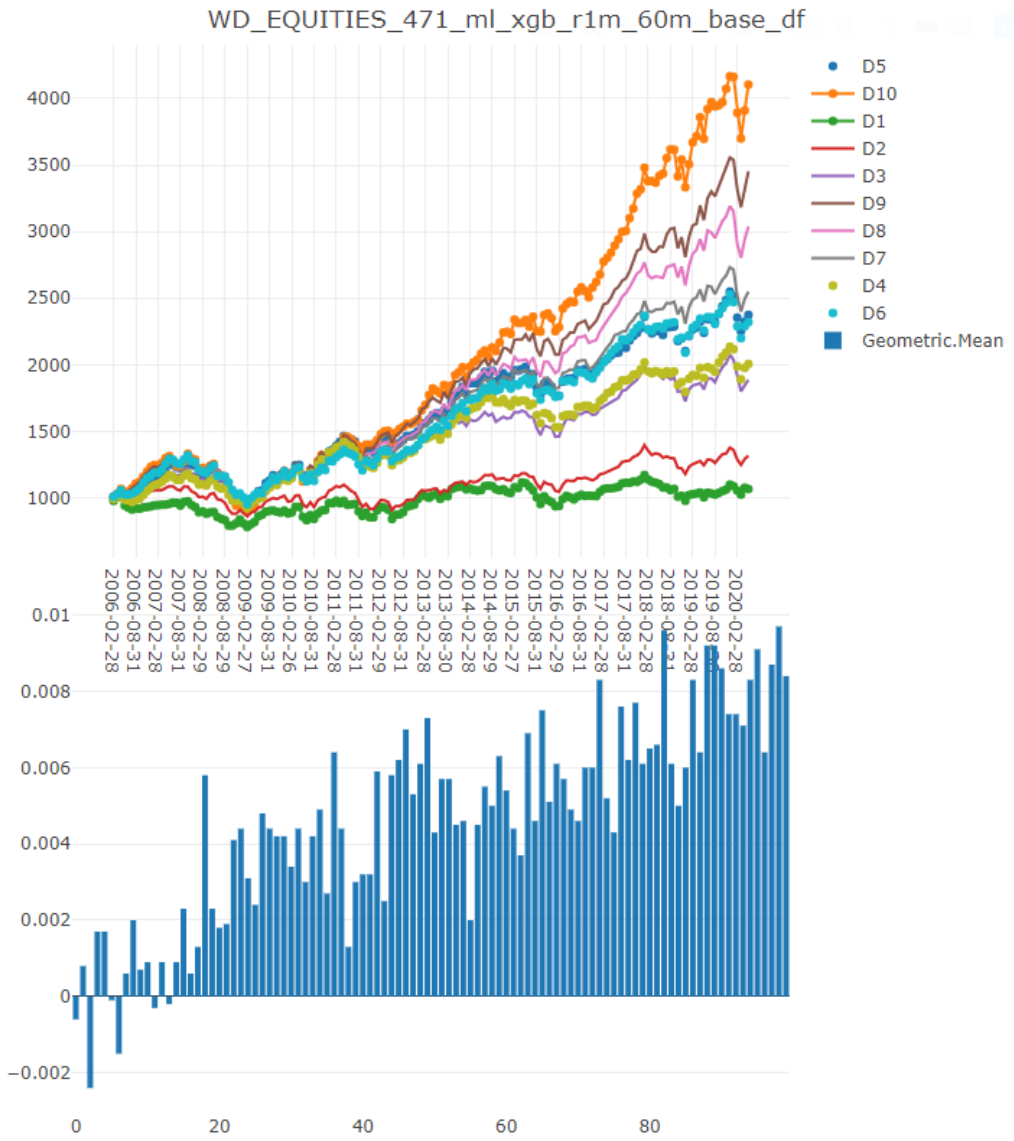
From the ML output (probability of outperforming) we create a signal and we construct portfolio from top/bottom decile (around 150 stocks in each portfolio).

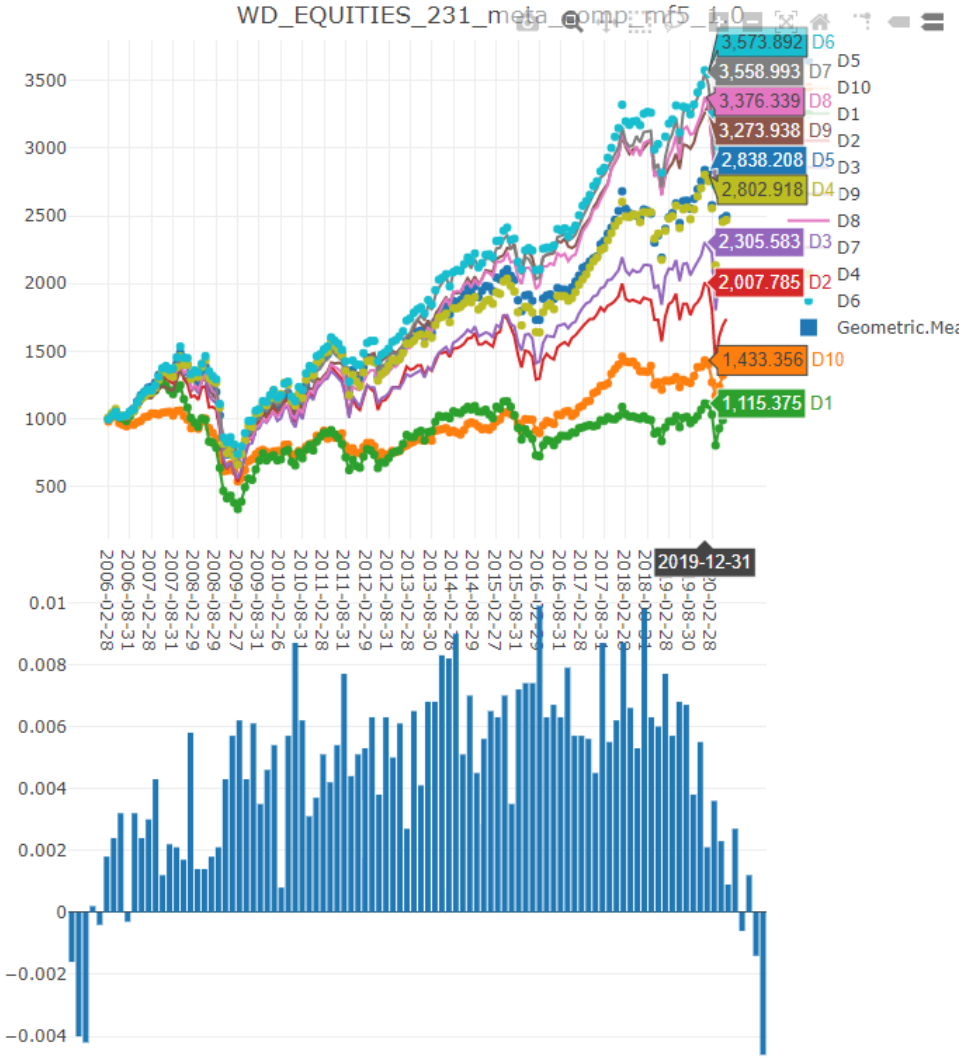
Comparing Factor Exposure



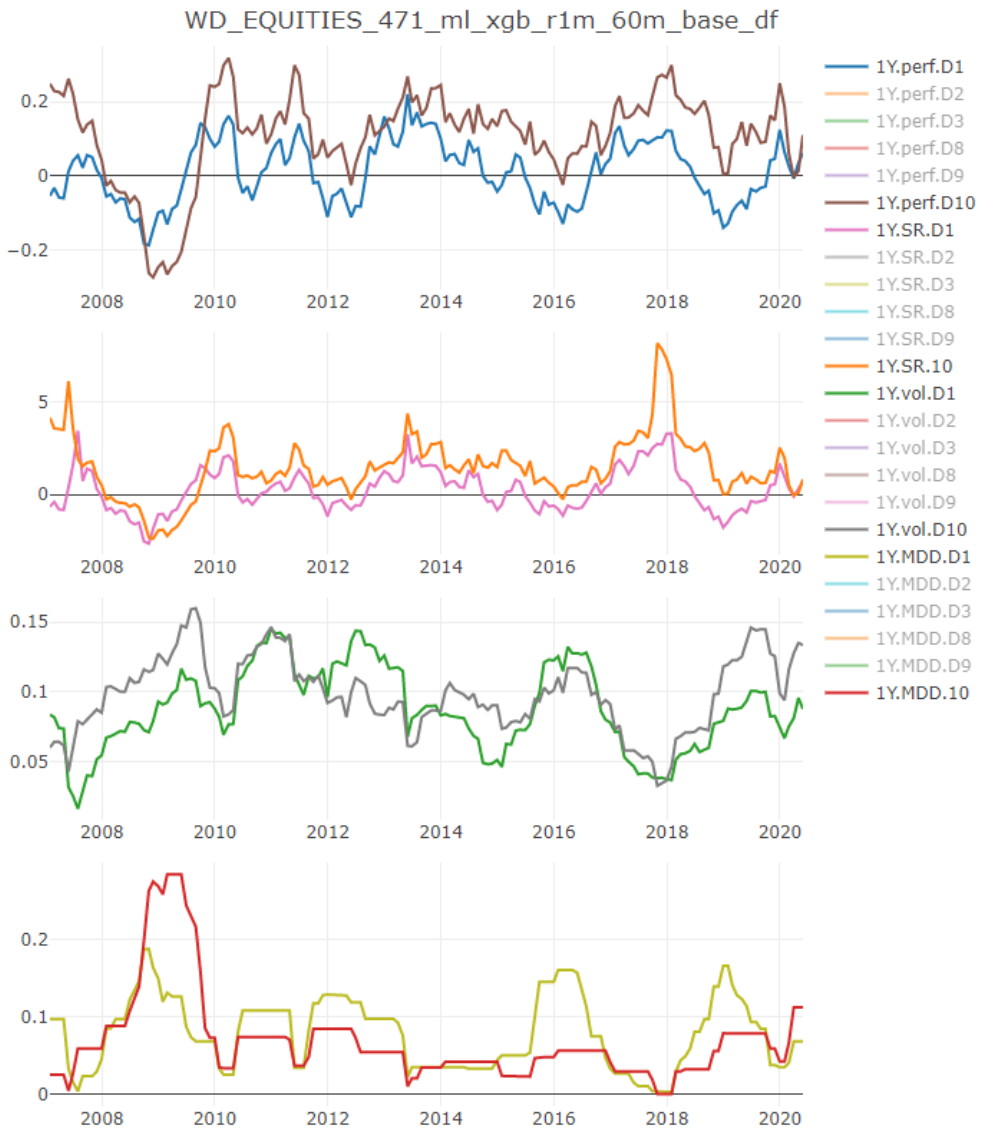
Source: RAM, Bloomberg, Factset, RAM's alternative data providers.

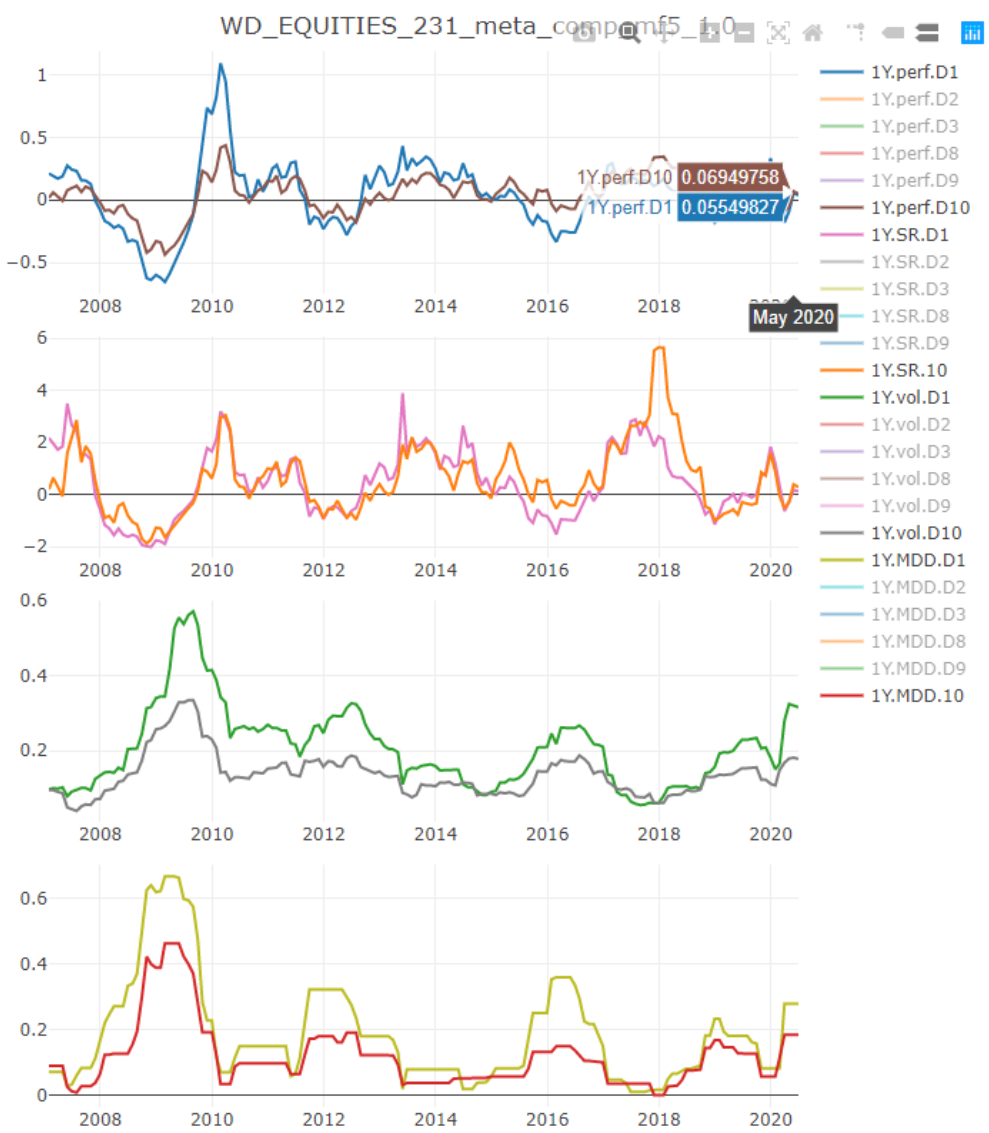
Comparing Monotonicity



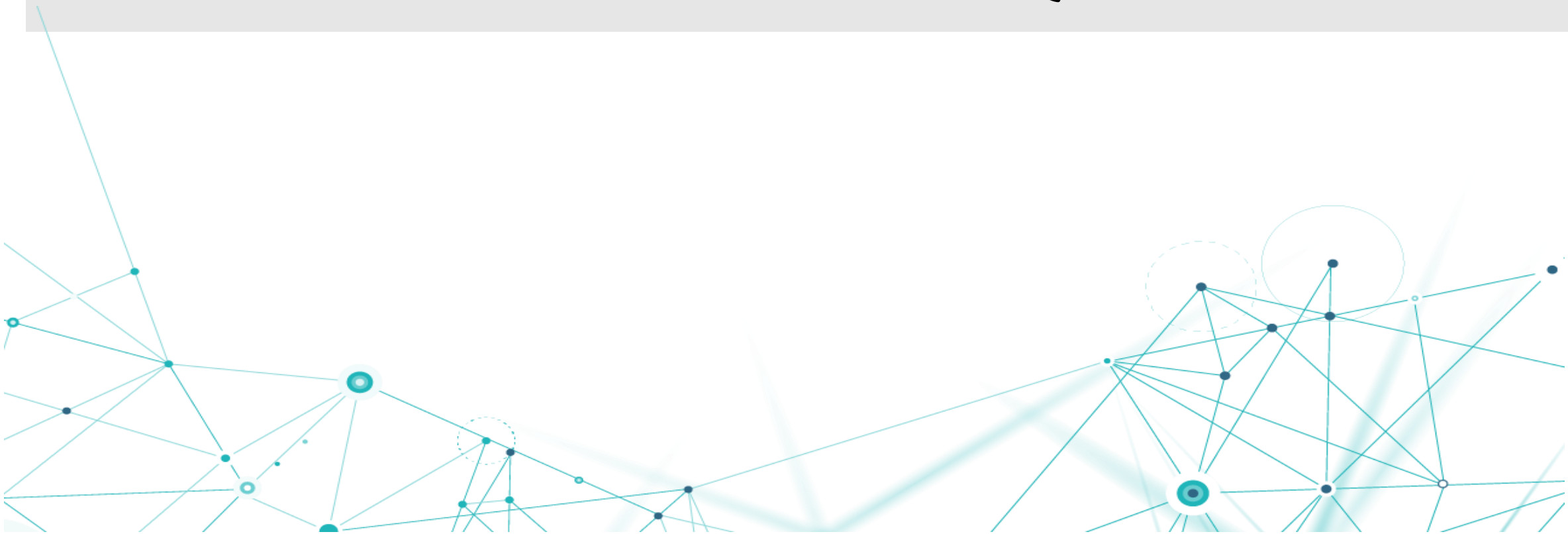


Comparing Performance





Conclusion and Q&A



Conclusion

- Machine learning is not new but a “**new**” way for doing **research** today.
- ML used with traditional data proved to add a **non-linear adaptative** component to alpha prediction
- Matter of survival to be capable on **onboarding**, **analysing** and **implementing**
- Big/alt data in the investment toolbox.