



# Machine Learning: Opportunities and Pitfalls

Campbell R. Harvey  
Duke University and NBER

March 2019



# A Backtesting Protocol in the Era of Machine Learning

ROB ARNOTT, CAMPBELL R. HARVEY, AND HARRY MARKOWITZ

## Modeling Analysts' Recommendations via Bayesian Machine Learning



DAVID BEW, CAMPBELL R. HARVEY, ANTHONY LEDFORD,  
SAM RADNOR, AND ANDREW SINCLAIR

# The phone call

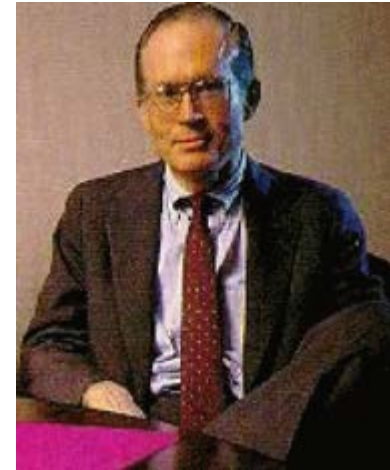
Early in my career, I get a telephone call my office line at about 9pm.

- The number is the familiar Goldman Sachs telephone number

# The phone call

Early in my career, I get a telephone call my office line at about 9pm.

- The number is the familiar Goldman Sachs telephone number
- To my shock, it is Fischer Black
- He has a problem with a table in my 1989 *Journal of Financial Economics* paper (which wasn't published until late 1990).



# The phone call

Black tells me he does not believe my Table 2

- He says that this is an example of “data mining”
- He argues that the predictability of stock returns I document is implausible and likely overfit



## Time-varying conditional covariances in tests of asset pricing models

Campbell R. Harvey \*

Table 2

Regressions of excess returns<sup>a</sup> for all NYSE common stocks (ranked by firm size) on the instrumental variables<sup>b</sup> based on monthly data from September 1941 to December 1987 (554 observations).

$$r_{j,t} = \delta_0 + \delta_1 xew_{t-1} + \delta_2 jan_t + \delta_3 xh3_{t-1} + \delta_4 junk_{t-1} + \delta_5 xdiv_{t-1} + \epsilon_{j,t}.$$

Portfolio	$\delta_0$	$\delta_1$	$\delta_2$	$\delta_3$	$\delta_4$	$\delta_5$	In-sample $\bar{R}^2$	Out-of-sample $\bar{R}^{2c}$
Decile 1	-0.024 (0.009)	0.142 (0.081)	0.091 (0.018)	6.746 (2.742)	31.035 (12.167)	5.316 (1.539)	0.179	0.120
Decile 2	-0.018 (0.007)	0.114 (0.055)	0.064 (0.013)	7.692 (2.456)	24.429 (8.688)	4.333 (1.124)	0.149	0.108
Decile 3	-0.017 (0.006)	0.104 (0.050)	0.052 (0.011)	7.561 (2.393)	23.307 (7.971)	4.237 (1.031)	0.131	0.100
Decile 4	-0.014 (0.006)	0.100 (0.046)	0.038 (0.010)	8.251 (2.365)	19.880 (7.344)	3.926 (0.945)	0.110	0.079
Decile 5	-0.014 (0.006)	0.072 (0.043)	0.034 (0.009)	9.008 (2.338)	18.600 (6.802)	4.018 (0.898)	0.107	0.082
Decile 6	-0.014 (0.005)	0.059 (0.042)	0.027 (0.009)	(8.310) (2.289)	(20.041) (6.437)	(4.057) (0.876)	0.094	0.068
Decile 7	-0.013 (0.005)	0.046 (0.041)	0.018 (0.008)	8.554 (2.370)	18.696 (6.536)	4.086 (0.851)	0.084	0.059
Decile 8	-0.011 (0.005)	0.019 (0.040)	0.016 (0.008)	8.818 (2.324)	15.731 (6.143)	3.750 (0.798)	0.075	0.047
Decile 9	-0.009 (0.005)	0.010 (0.039)	0.010 (0.007)	9.059 (2.482)	14.339 (6.060)	3.997 (0.787)	0.080	0.052
Decile 10	-0.007 (0.004)	0.005 (0.039)	0.000 (0.007)	7.749 (2.534)	11.001 (5.487)	3.562 (0.665)	0.067	0.032
Val. weight	-0.009 (0.004)	0.017 (0.038)	0.007 (0.007)	8.234 (2.436)	12.797 (5.608)	3.690 (0.703)	0.075	0.044

# The phone call

Table 2 tries to predict one-month ahead US stock returns using lagged information

- He argues the  $R^2$  is too high at 7.5%

Table 2

Regressions of excess returns<sup>a</sup> for all NYSE common stocks (ranked by firm size) on the instrumental variables<sup>b</sup> based on monthly data from September 1941 to December 1987 (554 observations).

$$r_{j,t} = \delta_0 + \delta_1 xew_{t-1} + \delta_2 jan_t + \delta_3 xh3_{t-1} + \delta_4 junk_{t-1} + \delta_5 xdiv_{t-1} + \epsilon_t.$$

Portfolio	$\delta_0$	$\delta_1$	$\delta_2$	$\delta_3$	$\delta_4$	$\delta_5$	In-sample $\bar{R}^2$	Out-of-sample $\bar{R}^{2c}$
Val. weight	-0.009 (0.004)	0.017 (0.038)	0.007 (0.007)	8.234 (2.436)	12.797 (5.608)	3.690 (0.703)	0.075	0.044

# The phone call

Looking back in time, this telephone call is ironic given my research agenda is to improve research practices in finance and to call out the data miners.

- I decide to replicate and test my model (in a true) out of sample

# Market timing replication

Table 5: Time Series Regression Coefficients, Original Data

	Cons	EW NYSE	Jan	Term Prem	Junk Spread	Div Spread	IS $R^2$	OOS $R^2$
Val. weight	-0.006	0.034	-0.010	6.477	13.400	3.250	0.076	0.033

Original sample to 1987 replicates well

- $R^2$  about the same



# Market timing validation

Table 5: Time Series Regression Coefficients, Original Data

	Cons	EW NYSE	Jan	Term Prem	Junk Spread	Div Spread	IS $R^2$	OOS $R^2$
Val. weight	-0.006	0.034	-0.010	6.477	13.400	3.250	0.076	0.033
1988-2018 Data								
Val. weight	0.012	0.067	-0.004	5.920	-7.510	1.344	0.012	-0.052

## Out of sample 1988-2018 fails

- $R^2$  now only 1.2% and not significant
- One-step ahead  $R^2$  now effectively zero

# Market timing validation

Table 5: Time Series Regression Coefficients, Original Data

	Cons	EW NYSE	Jan	Term Prem	Junk Spread	Div Spread	IS $R^2$	OOS $R^2$
Val. weight	-0.006	0.034	-0.010	6.477	13.400	3.250	0.076	0.033
1988-2018 Data								
Val. weight	0.012	0.067	-0.004	5.920	-7.510	1.344	0.012	-0.052

## Out of sample 1988-2018 fails

- $R^2$  now only 1.2% and not significant
- One-step ahead  $R^2$  now effectively zero
- Coefficients unstable

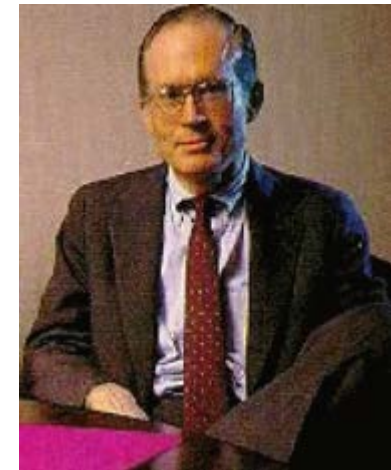
# Market timing validation

Table 5: Time Series Regression Coefficients, Original Data

	Cons	EW NYSE	Jan	Term Prem	Junk Spread	Div Spread	IS $R^2$	OOS $R^2$
Val. weight	-0.006	0.034	-0.010	6.477	13.400	3.250	0.076	0.033
1988-2018 Data								
Val. weight	0.012	0.067	-0.004	5.920	-7.510	1.344	0.012	-0.052

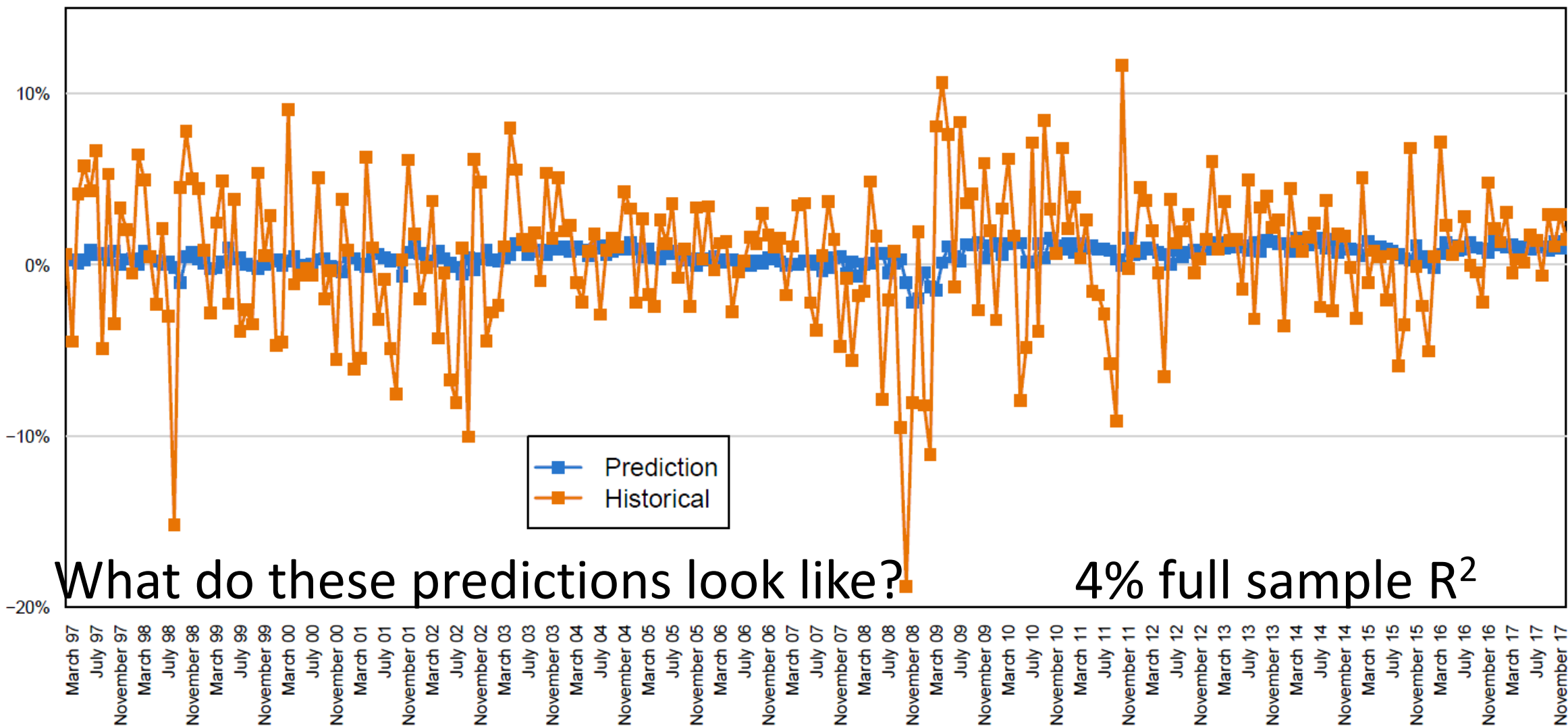
## Out of sample 1988-2018 fails

- $R^2$  now only 1.2% and not significant
- One-step ahead  $R^2$  now effectively zero
- Coefficients unstable



# Market timing validation

Variance of predictions (blue) is 4% of the variance of the actual returns (orange)



What do these predictions look like?

4% full sample  $R^2$

# What have we learned?

I am approached by a high-profile organization to assess their one-month ahead global equity forecasting model (call it GF)

# What have we learned?

## How the model works

- “...our model generates one-month-ahead forecasts on a strictly monthly basis. All forecasts are of MSCI indexes priced in USD.”
- “59 regional MSCI indexes. These cover all developed, emerging, and frontier markets, excepting only the following: Argentina, Bangladesh, Mauritius, Sri Lanka, Kuwait, Lebanon, Oman, Serbia, and UAE.”

# What have we learned?

## How the model was built

- The “model is the product of over two years of intensive data collection and statistical research.”
- “We have gathered and tested roughly 200 monthly variables for each market going back to the 1990s. These variables (some of them highly proprietary) cover market, economic, demographic, and political trends.”
- “All told, the model trains on over 3 million data points.”
- “Our statistical analysis is multi-stage and uses the most advanced machine learning algorithms.”

# What have we learned?

## How the model was built

- The “model is the product of over two years of intensive data collection and statistical research.”
- “We have gathered and tested roughly 200 monthly variables for each market going back to the 1990s. These variables (some of them highly proprietary) cover market, economic, demographic, and political trends.”
- “All told, the model trains on over 3 million data points.”
- “Our statistical analysis is multi-stage and uses the **most advanced machine learning algorithms.**”



# What have we learned?

## How accurate is the model?

- “Our accuracy is impressive”
- For “the ‘Big 25’ economies, our  $R^2$  is 0.96”
- For “the total world (all 59 economies), our  $R^2$  is 0.98”

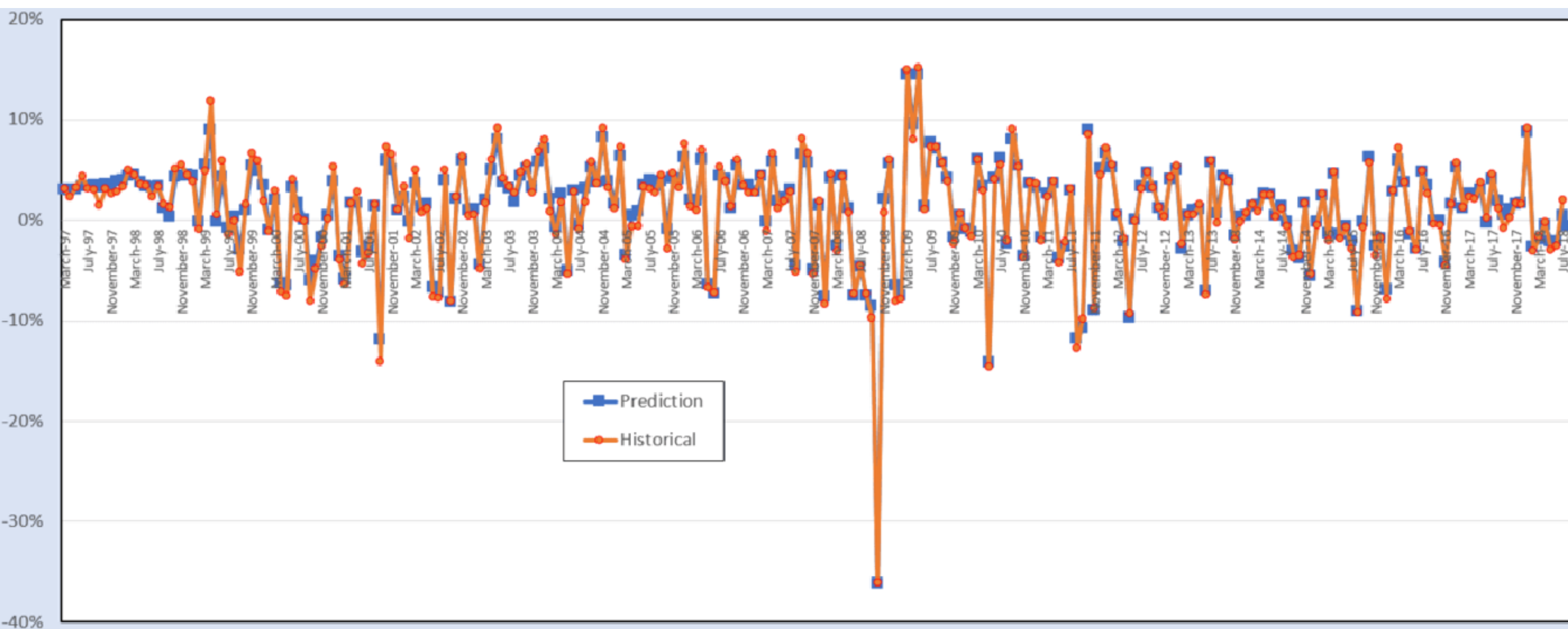
# What have we learned?

## How accurate is the model?

- “Our accuracy is impressive”
- For “the ‘Big 25’ economies, our  $R^2$  is 0.96”
- For “the total world (all 59 economies), our  $R^2$  is 0.98”

Maybe this is a typo and they mean 0.98% which would be consistent with my results of 1%  $R^2$

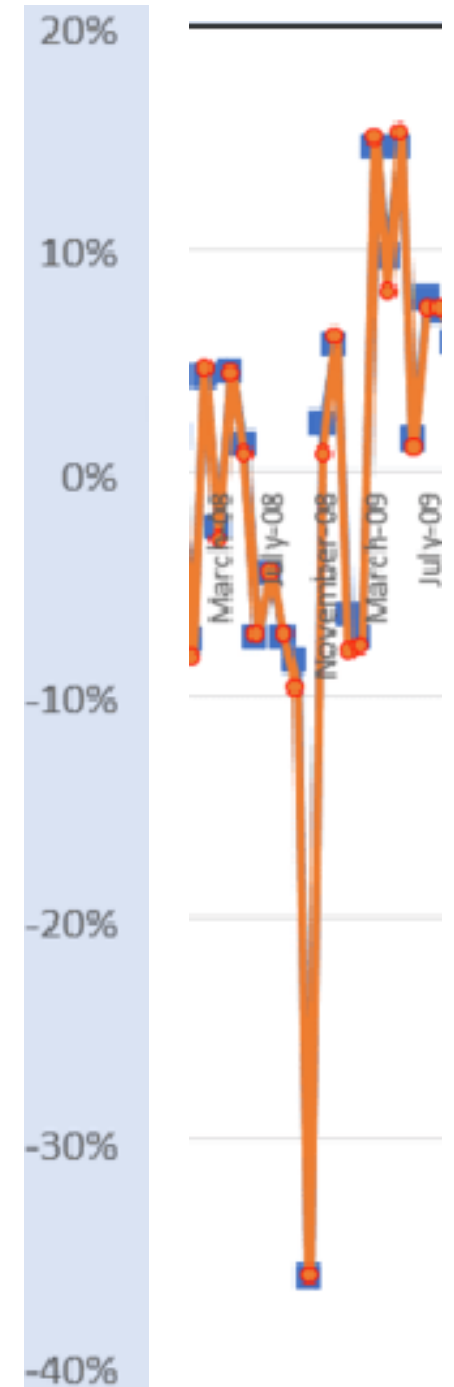
# What have we learned?



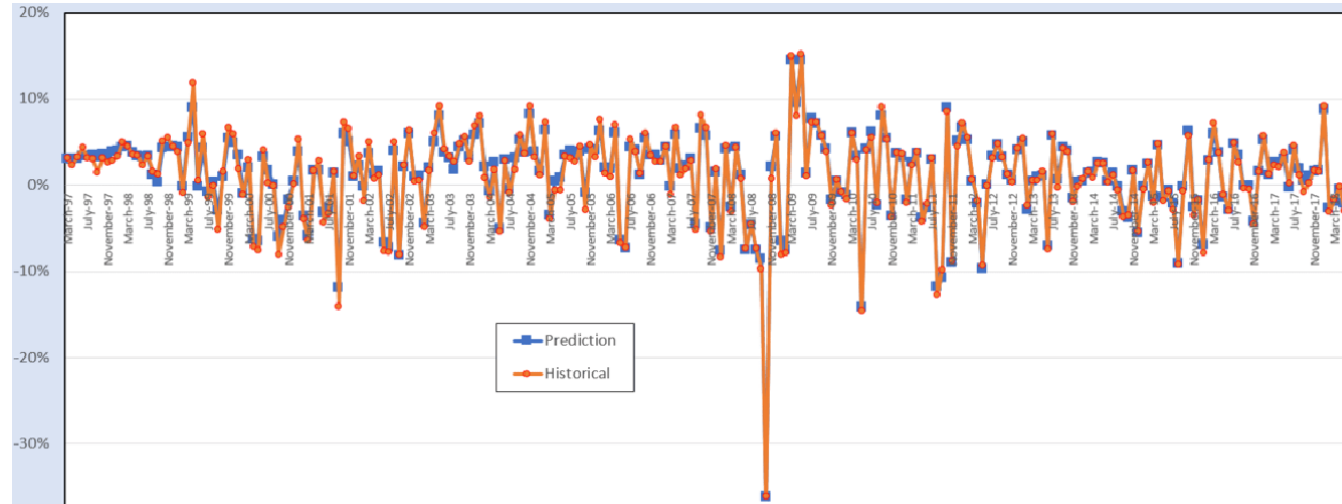
# What have we learned?

## How accurate is the model?

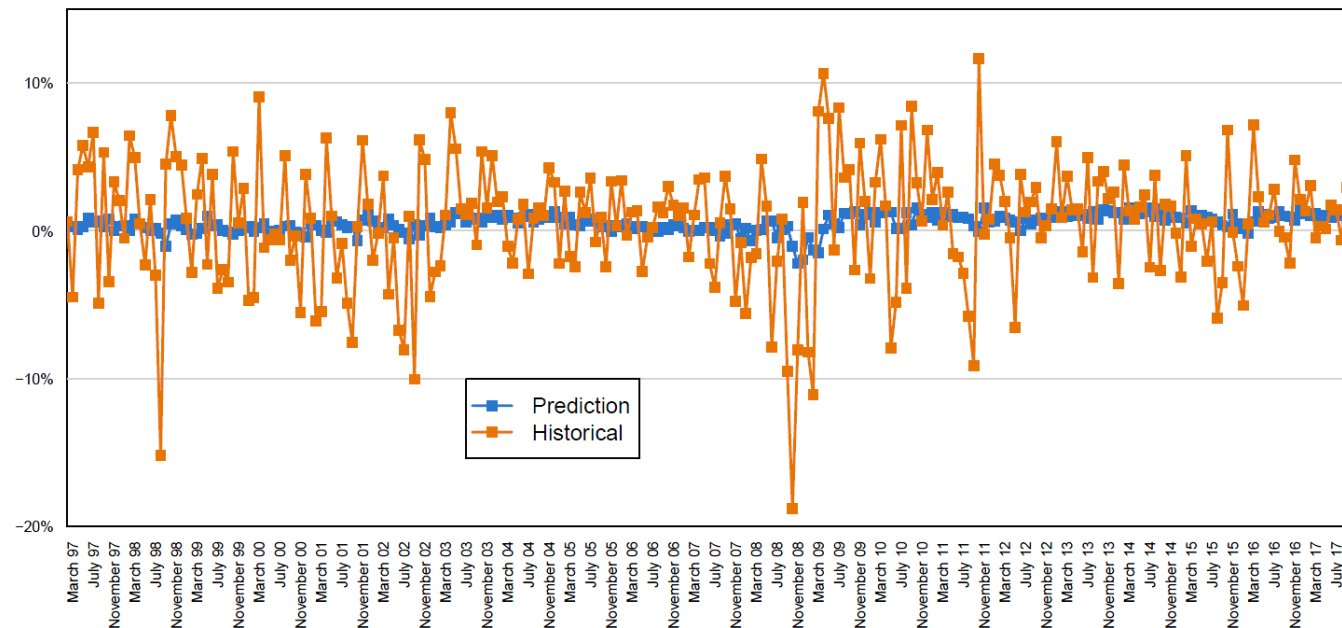
- Worst month of the global financial crisis had a -35% return (October 2008)
- Blue square is the prediction; Red dot the realized return
- Forecast (made in September 2008) was -35%!



# What have we learned?



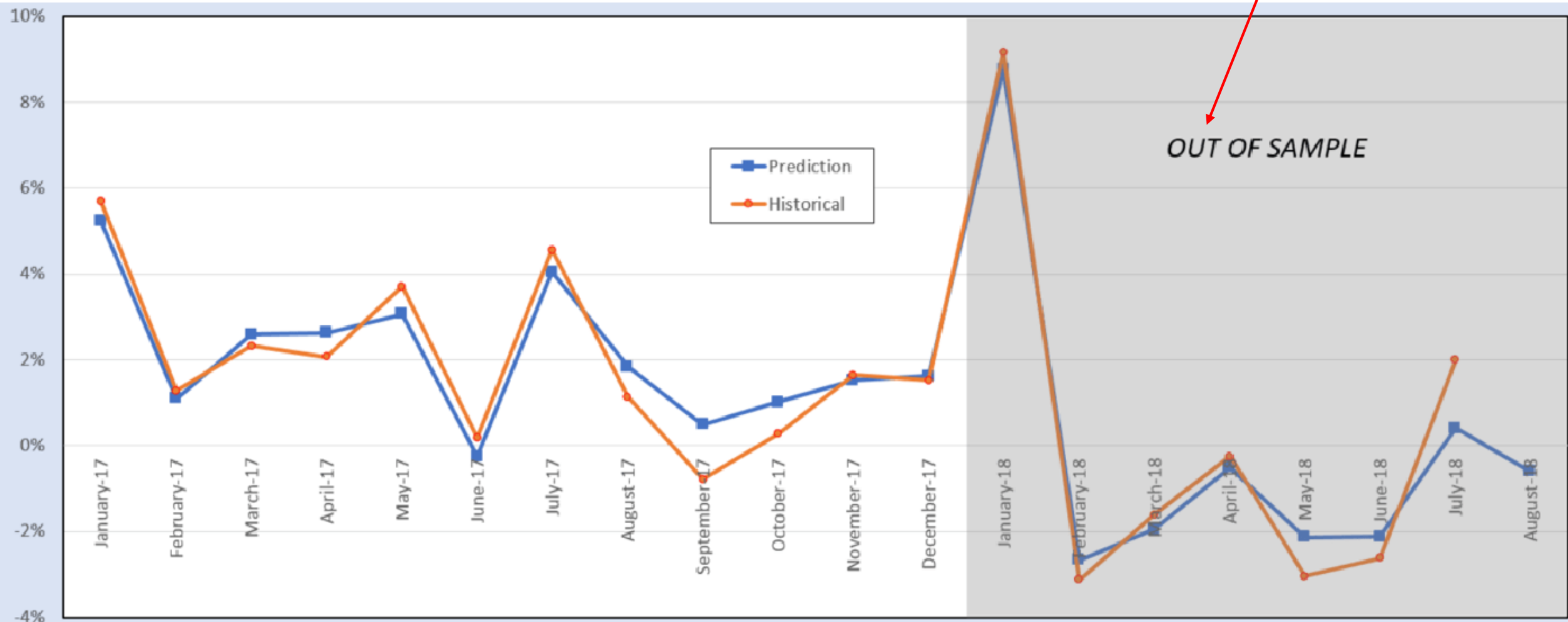
GF model  $R^2=0.98$



Harvey model  $R^2=0.04$

# What have we learned?

True out of sample??



# What have we learned?

## Further development of the model

- ... “run alternative independent variables”
- ... “generate ‘now-cast’ forecasts at any time of the month”
- “These and other improvements are in store”

# The era of machine learning

Applications of these advanced methods requires knowledge of the limitations of these methods.

[T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.



# The era of machine learning

Applications of these advanced methods requires knowledge of the limitations of these methods.

[T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.




25 years in global  
forecast model

# The era of machine learning

Applications of these advanced methods requires knowledge of the limitations of these methods.

[T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.

Important, I will  
return to this point

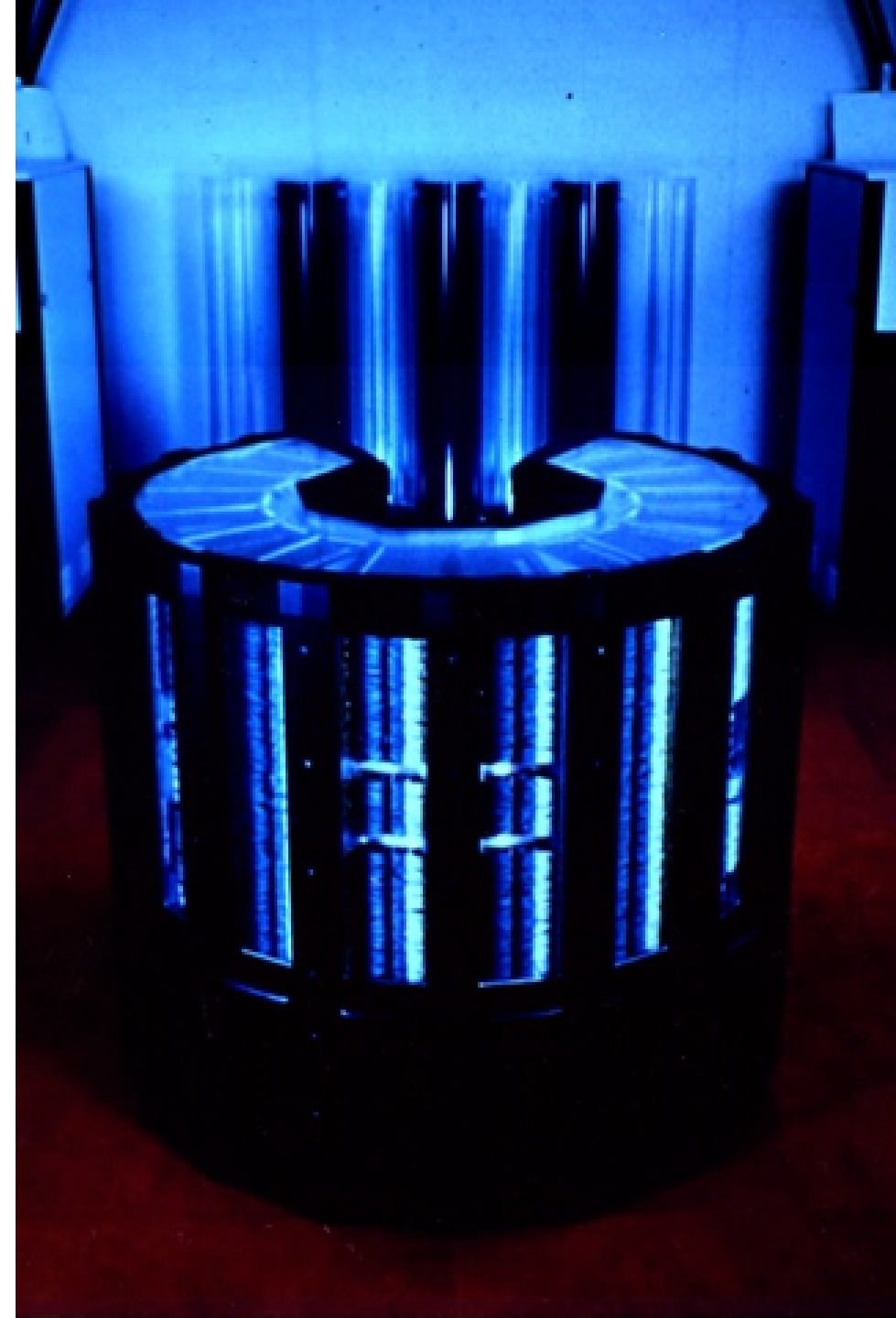


# Rise of the machines

Cray 2 is the world's fastest supercomputer: 1985-1990

- CPU: 1.9 GFLOPs\*
- Weight: 5,500 pounds
- Cost: \$32 million (current \$)

\*1.9 billion floating point operations per second



# Rise of the machines

iPhone Xs:\*

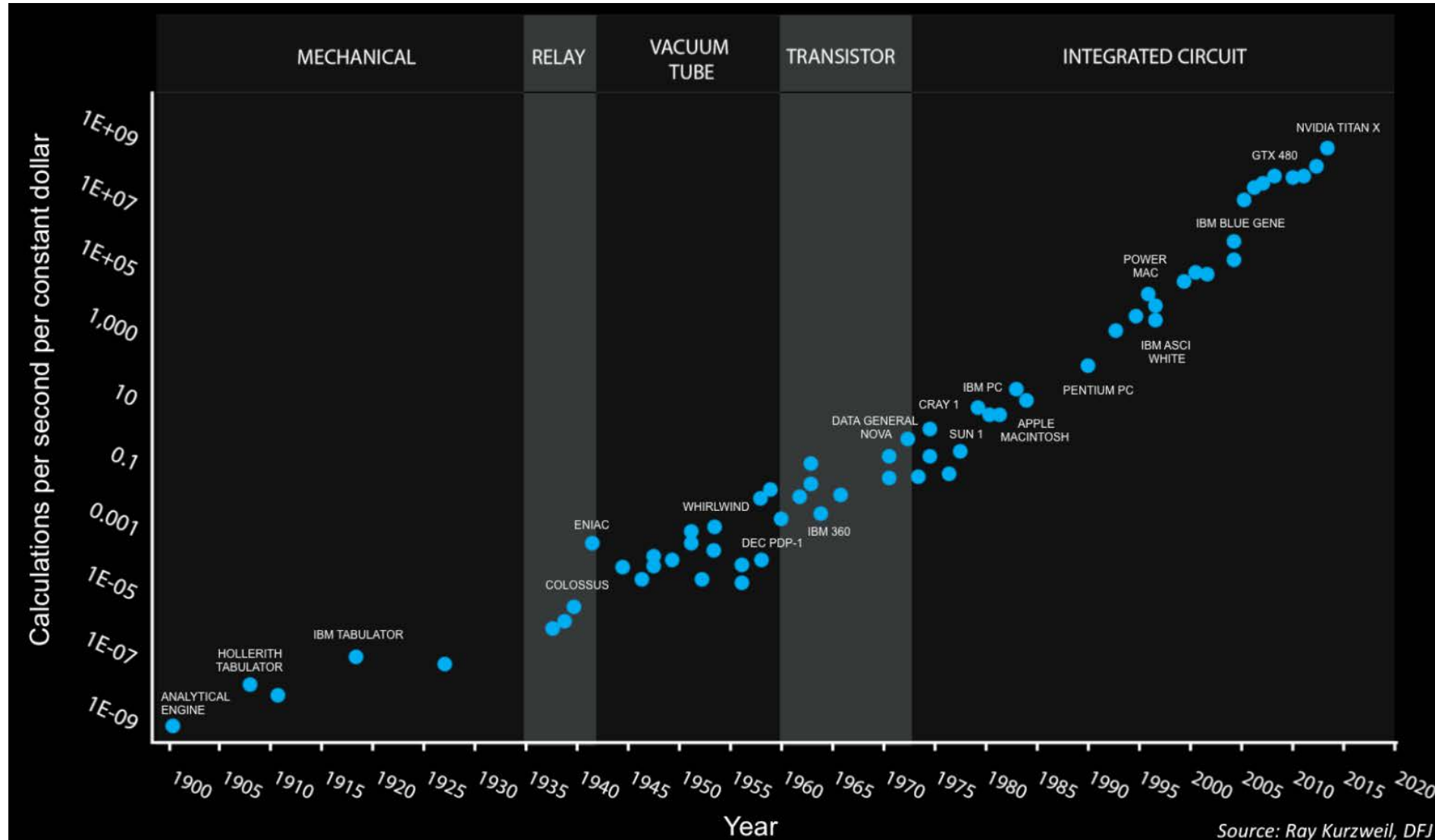
- 5,000 GFLOPs (9x faster than 2017)
- 512 GB storage
- 6.24 oz
- \$1,000

\*A12 Bionic. <https://www.apple.com/iphone-xs/a12-bionic/>  
Also note the Apollo guidance system had only 4K of RAM.



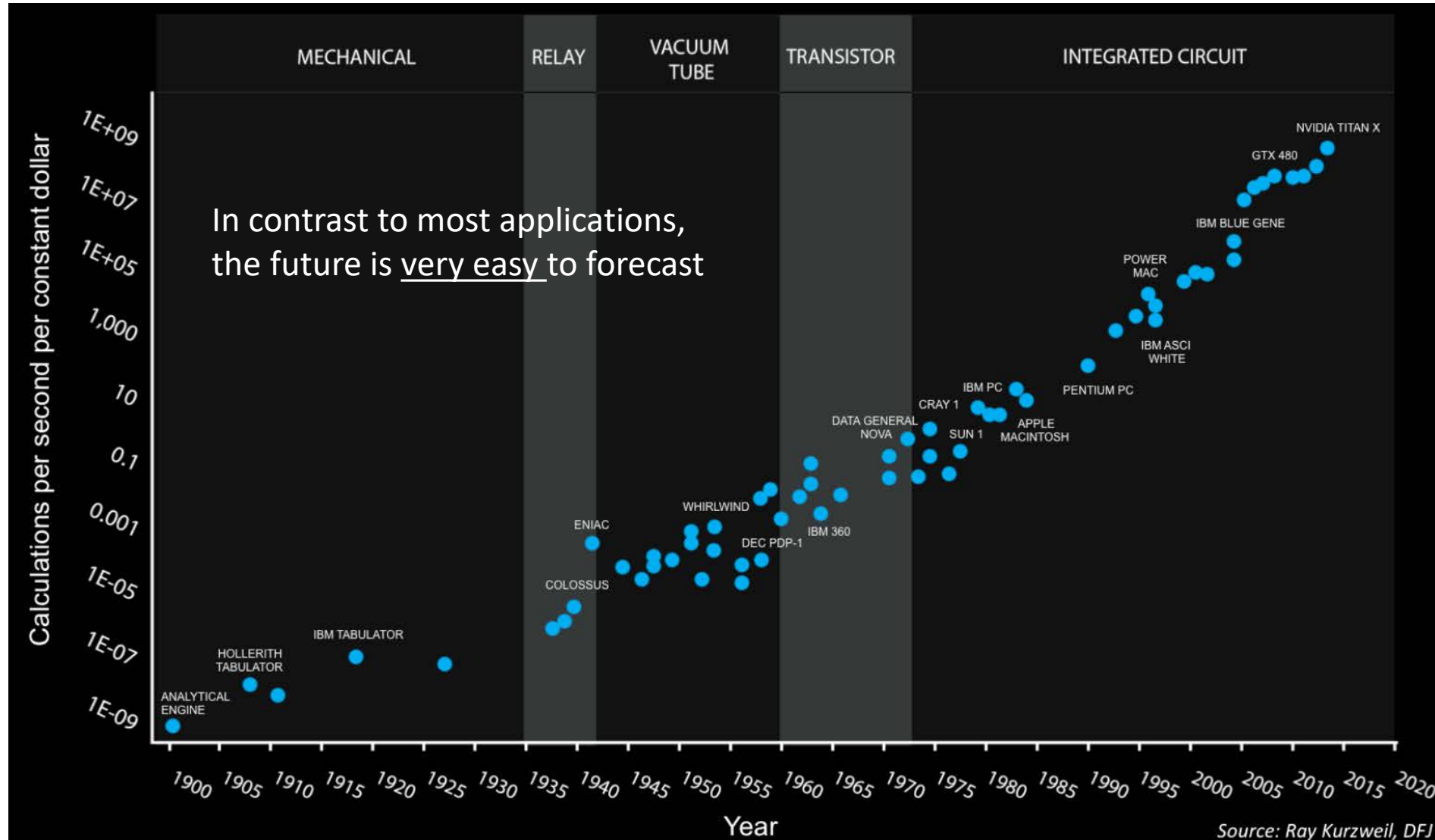
# Rise of the machines

## Revolution #1: Computing Power



# Rise of the machines

## Revolution #1: Computing Power



# Rise of the machines

## Revolution #2: Data generation and storage

### Cost per GB

- 1981 - \$300,000
- 1987 - \$50,000
- 1990 - \$10,000
- 1994 - \$1,000
- 1997 - \$100
- 2000 - \$10
- 2004 - \$1
- 2010 - \$0.1
- 2019 - \$0.01

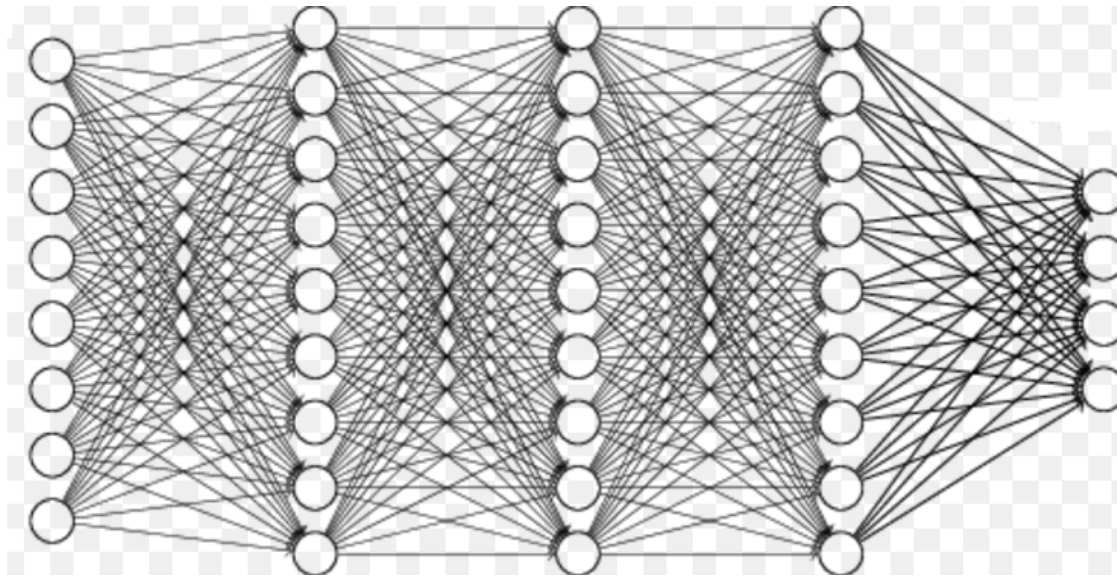
Campbell R. Harvey 2019



# Rise of the machines

## Revolution #3: Methods

- Maturing of combined methods from statistics, computer science, engineering and mathematics





# Rise of the machines

## Revolution #4: Open-source code

- Coding is a shared activity. No longer necessary to ‘reinvent the wheel’.



# Build better together

Host and review code, manage projects and build  
your best software alongside 31 million developers.

# Prerequisites

## Basic questions: Data

- How much data are available?
- What is the signal to noise ratio?
  - (For example, tick data is useless for lower frequency signals, such as value)
- What is the quality of the data?
- How independent are the data?

# Prerequisites

## GF model

- Approximately 22 years of monthly data
  - 59 markets
  - 200 predictors
  - “approximately 3 million data points”
- However, these markets are correlated with the most important being the US equity market. Here, 264 monthly returns are hit with 200 predictors!

# Prerequisites

## Basic questions: Method

- Is the machine learning method appropriate for the particular application?
- How much data does the method require?
- Can we learn from applications in other sciences?
- Now for a positive example!

# Star gazing

## Galaxy Zoo: Supernovae project

- Thousands of amateur astronomers are asked to help classify Supernovae (SN) in massive citizen science project
- Three way classification:
  - Very likely SN
  - Possible SN
  - Not likely SN

# Star gazing

## Galaxy Zoo: Supernovae project

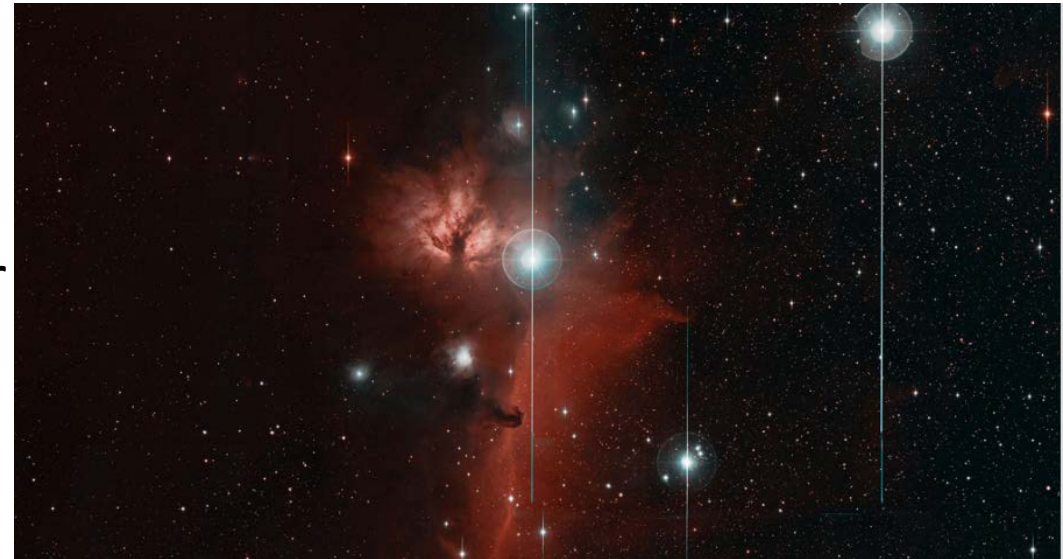
- Thousands of amateur astronomers are asked to help classify Supernovae (SN) in massive citizen science project
- Three way classification:
  - Very likely SN
  - Possible SN
  - Not likely SN
- “True” classification from Caltech’s Palomar Transient Factory



# Star gazing

## How should the classifications be combined?

- Thousands of astronomers looking at about 10,000s possible SN
- They might provide input on 5-10 possible SN
  - Is it best to look at the average classification?
  - What about a simple majority vote (yes or no)?
  - Should we take the track record of the astronomer into account?
  - Can we allow for both improvement in the track record through time as well as fatigue?



# Star gazing

Amateur astronomers are serious!



Francis Longstaff

Allstate Chair in Insurance and Finance

THE ASTROPHYSICAL JOURNAL, 791:24 (9pp), 2014 August 10  
© 2014. The American Astronomical Society. All rights reserved. Printed in the U.S.A.

## A FORMAL METHOD FOR IDENTIFYING DISTINCT STATES OF VARIABILITY IN TIME-VARYING SOURCES: SGR A\* AS AN EXAMPLE

L. MEYER<sup>1</sup>, G. WITZEL<sup>1</sup>, F. A. LONGSTAFF<sup>2</sup>, AND A. M. GHEZ<sup>1</sup>

<sup>1</sup> Department of Physics and Astronomy, University of California, Los Angeles, CA 90095-1547, USA

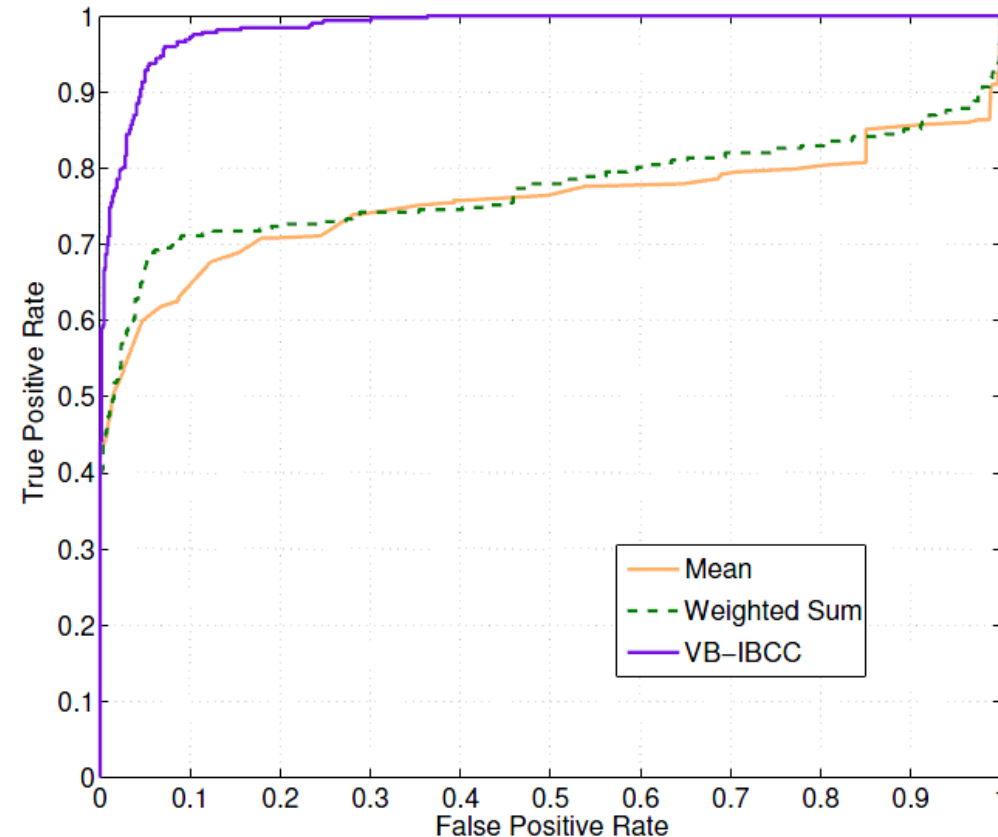
<sup>2</sup> UCLA Anderson School of Management, University of California, Los Angeles, CA 90095-1481, USA



# Star gazing

## Ideal application of independent Bayesian classifier combination (IBCC)

- Analysis by Simpson, Roberts, Psorakis and Smith (2013) showed IBCC had dramatic improvements over using a simple rule like the average
- Allowing for a 10% error rate, the correct classification went from 65% (simple average) to 97% (IBCC)



# Star gazing and stock picking

## Striking similarities to investment manager evaluating analysts' recommendations

- Thousands of objects (companies) and thousands of astronomers (analysts)
- Subjects do not cover all objects (companies) – only a subset (sparsity)
- Classification:

very likely	↔	buy
possible	↔	hold
not likely	↔	sell

# Star gazing and stock picking

Striking similarities to investment managers evaluating analysts' recommendations

- In addition, it is reasonable to assume a differential degree of skill among the analysts

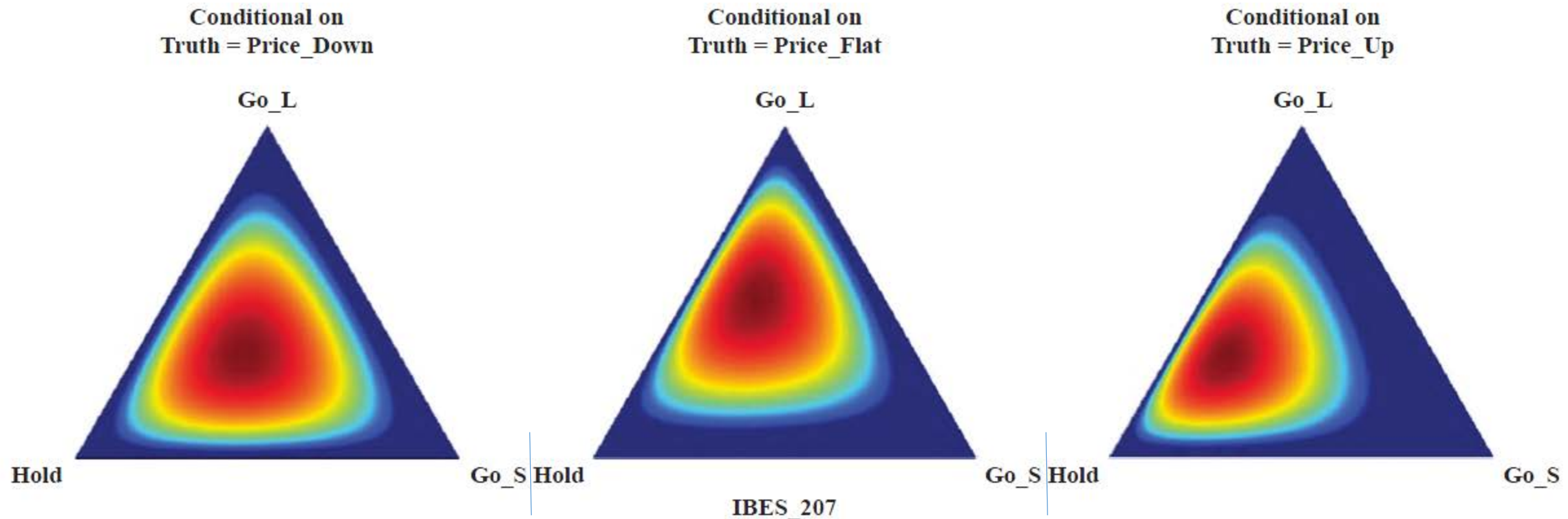
# Star gazing and stock picking

## IBCC improves performance

Side	Model	Mean	Vol	Alpha	Alpha <i>t</i> -Stat
Long–Short	Brok_Flw_LS	4.54	13.92	4.65	2.09
	IBCC_Rol_LS	5.07	11.01	5.30	2.69
	IBCC_Exp_LS	6.50	12.66	6.64	3.35
	Both_Rol_LS	7.99	15.43	8.15	3.18
	Both_Exp_LS	7.88	16.00	8.09	3.12

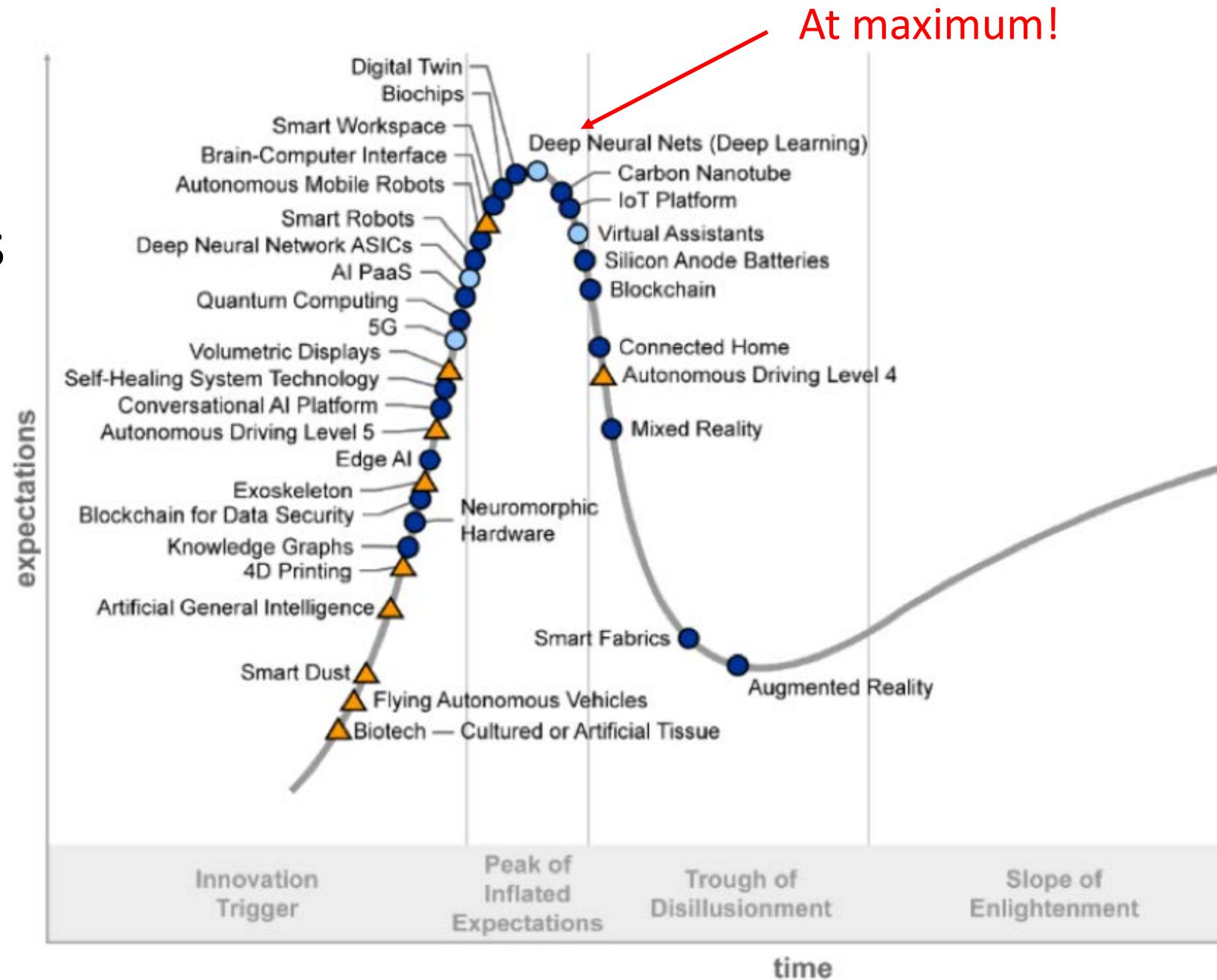
# Star gazing and stock picking

IBCC allows us to evaluate individual analysts through time



Time to step back

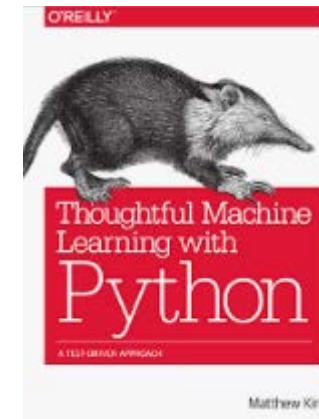
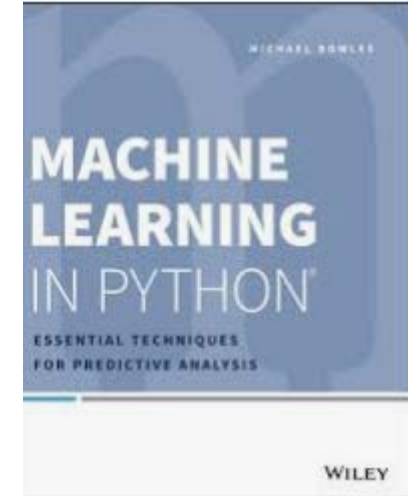
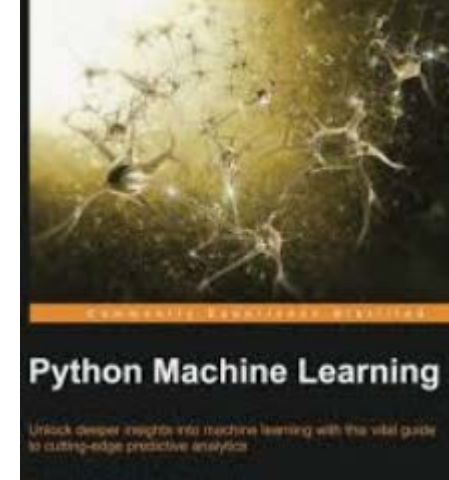
Unfortunately, there is a lot hype about machine learning



# Time to step back

Tempting to hit any problem with machine learning

- Tools are freely available and off the shelf
- It is important to match the problem with the method – and to make sure there is enough high quality data
- Star gazing application of IBCC is an exemplar example
- There are many pitfalls

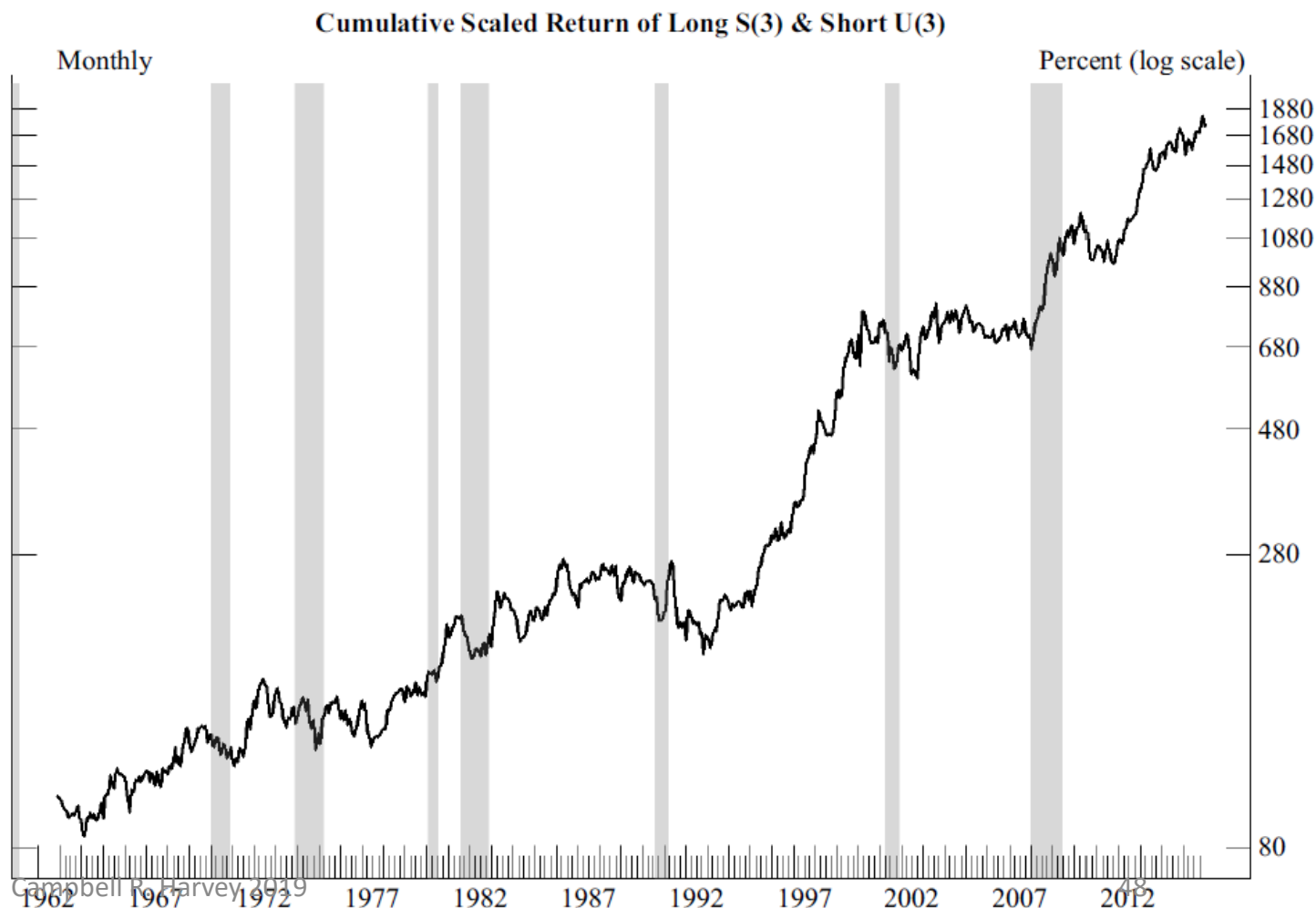


# Exhibit A

## EXHIBIT 1

Long-Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015

- Real data
- Impressive Sharpe (highly “significant”)
- Consistent method
- Good performance in GFC
- No significant correlation with known factors
- Turnover less than 10% per year





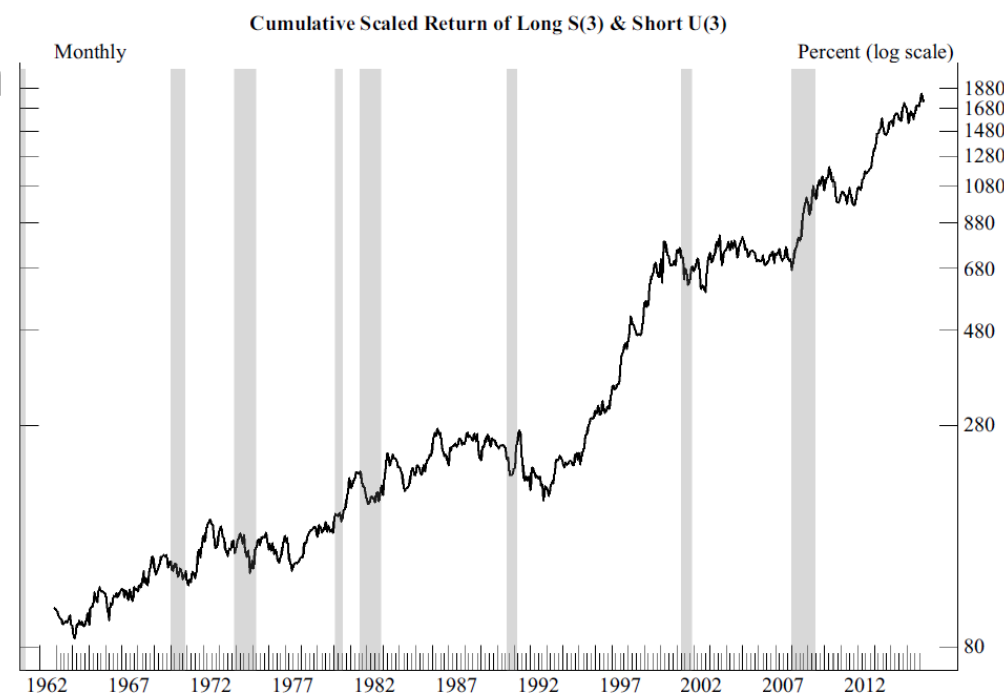
# Exhibit A

## Long S(3) and short U(3)

- Long equally weighted portfolio of stocks with S as the third letter of their ticker symbol
- Short equally weighted portfolio of stocks with U as the third letter of their ticker symbols
- I tried thousands of combinations to get the best one
- No economic foundation!

### EXHIBIT 1

Long-Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015



# Exhibit B

**2.1 million factors!**

*p*-hacking:

Evidence from two million trading strategies

Tarun Chordia

Amit Goyal

Alessio Saretto\*

August 2017

## Abstract

We implement a data mining approach to generate about 2.1 million trading strategies. This large set of strategies serves as a laboratory to evaluate the seriousness of

# Exhibit B

## The best factor

- Here is an example of a top five factor in the 2-million factor paper!

$(CSHO-CSHPRI)/MRC4$

# Exhibit B

- In words:

(Common Shares Outstanding – Common Shares Used to Calculate EPS)

# Exhibit B

- In words:

(Common Shares Outstanding – Common Shares Used to Calculate EPS)  
Rental Commitments – 4<sup>th</sup> year

# Best practices for research in quantitative finance

The time is right to develop a research protocol

- Goals of the protocol:
  - Provide asset owners a set of due diligence questions
  - Improve the research culture

# #1. Research motivation

- a) Does the model have a solid economic foundation?
- b) Did the economic foundation or hypothesis exist *before* the research was conducted?

# #1. Research motivation

- Data-mined factors should be treated with much more skepticism than a factor from economic theory (such as illiquidity, skewness, etc.). Some factors, such as the one scaled by 4<sup>th</sup> year rental payments should be thrown out no matter how good they look in sample
- Beware of ex post rationalization of a data mined result (it is called HARKing – Hypothesizing After the Results are Known)

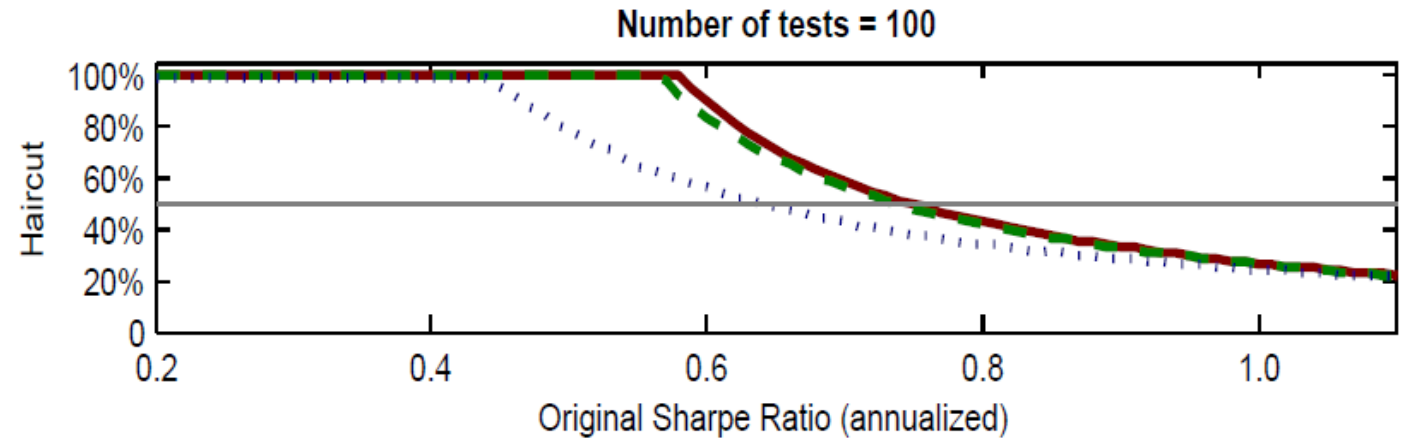


## #2. Multiple testing and statistical methods

- a) Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful) and are the researchers aware of the multiple-testing issue?
- b) Is there a full accounting of all possible interaction variables if interaction variables are used?
- c) Did the researchers investigate all variables set out in the research agenda or did they cut the research as soon as they found a good model?

## #2. Multiple testing and statistical methods

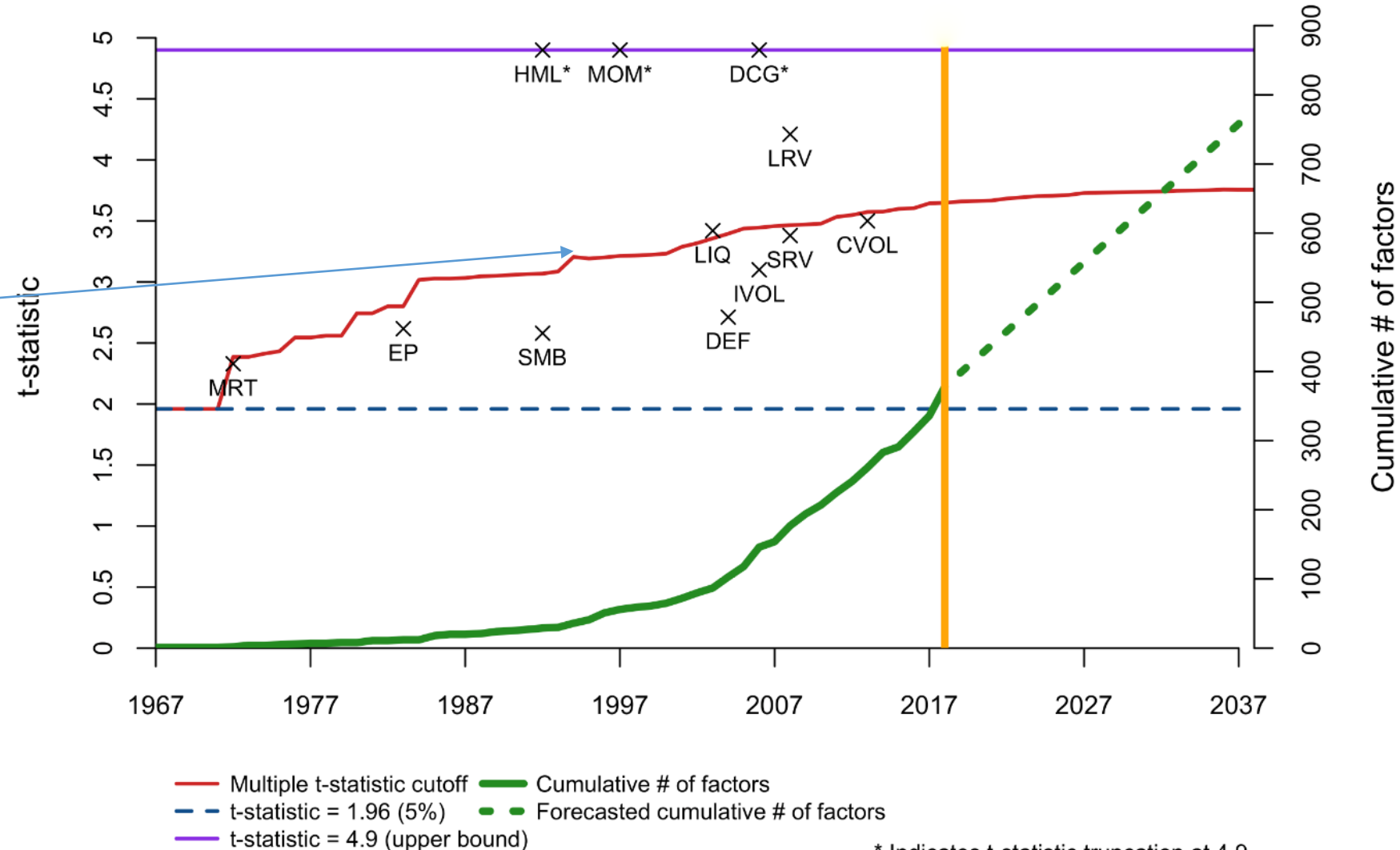
- The more variables tried – the higher the Sharpe ratio needs to be in sample.
- Harvey and Liu (2015) show that the haircut to Sharpe ratio is non-linear



## #2. Multiple testing and statistical methods

The hurdle for a new discovery must clear a higher hurdle.

Realize that low-hanging fruit has already been picked.



## #2. Multiple testing and statistical methods

### Beware of the parallel universe problem

- A researcher compiles a list of 20 variable to test for prediction. The first one “works”. The researcher stops and claims a single test (so no penalty)
- However, in a parallel universe, the researcher might try #20 first and only hit #1 after 20 tests
- Both researchers have multiple testing problems.

## #2. Multiple testing and statistical methods

### Beware of the parallel universe problem

- A researcher compiles a list of 20 variable to test for prediction. The first one “works”. The researcher stops and claims a single test (so no penalty)
- However, in a parallel universe, the researcher might try #20 first and only hit #1 after 20 tests
- Both researchers have multiple testing problems.



# #3. Data and sample choice

- a) Do the data chosen for examination make sense? And, if other data are available, does it make sense to exclude these data?
- b) Did the researchers take steps to ensure the integrity of the data?
- c) Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
- d) If outliers are excluded, are the exclusion rules reasonable?
- e) If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

# #3. Data and sample choice

Report to major national pension:

- “Abstracting from the financial crisis, we conclude that active management of both equity and fixed income has significantly contributed to the returns of the fund.”
- The global financial crisis should not be excluded from statistical analysis

## #4. Cross-validation

- a) Are the researchers aware that true out-of-sample tests are only possible in live trading?
- b) Are steps in place to eliminate the risk of out-of-sample “iterations” (i.e., an in-sample model that is later modified to fit out-of-sample data)?
- c) Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?



## #4. Cross-validation

Almost all academic research assumes zero transactions costs

- Backtested results are inflated not just by data mining but also because reasonable transaction costs are excluded.
- Product is often motivated by academic results.
- HML, for example, involves shorting small cap stocks – that is expensive
- Also, this basic mistake biases performance evaluation

## #5. Model dynamics

- a) Is the model resilient to structural change and have the researchers taken steps to minimize the overfitting of the model dynamics?
- b) Does the analysis take into account the risk/likelihood of overcrowding in live trading?
- c) Do researchers take steps to minimize the tweaking of a live model?

## #6. Complexity

- a) Does the model avoid the curse of dimensionality?
- b) Have the researchers taken steps to produce the simplest practicable model specification?
- c) Is an attempt made to interpret the predictions of the machine learning model rather than using it as a black box?

## #6. Complexity

### *Regularization:*

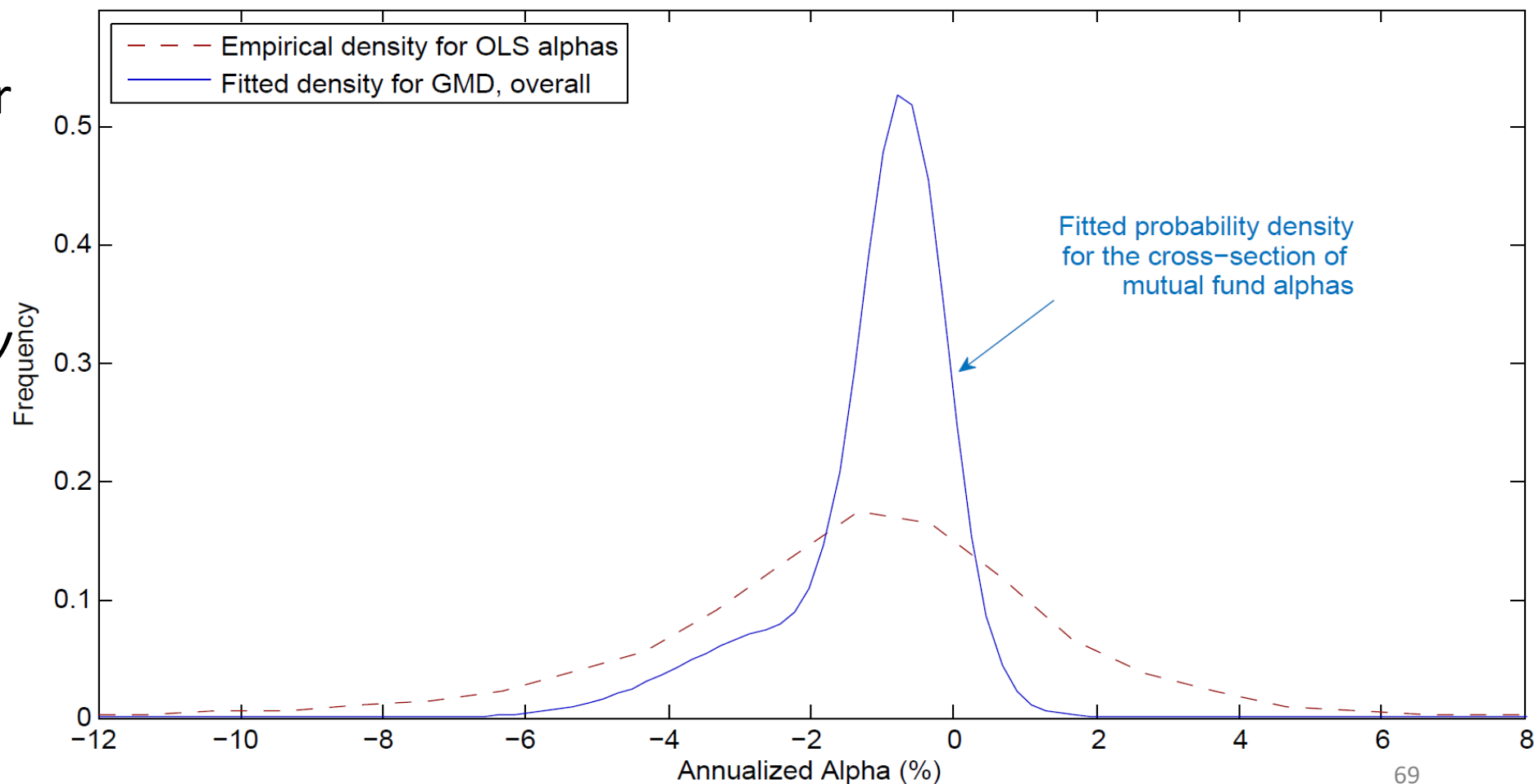
- Introduce additional constraints to achieve model simplification that helps prevent model overfitting.

[T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.

# #6. Complexity

## Regularization:

- INQUIRE 2017 paper “Detecting Repeatable Performance” (published in *Review of Financial Studies*, 2018), we impose a parametric distribution on the cross-section of alphas.



## #7. Research culture

- a) Does the research culture reward quality of the science rather than finding the winning strategy?
- b) Do the researchers and management understand that most tests will fail?
- c) Are expectations clear (that researchers should seek the truth not just something that works) when research is delegated?

# Conclusions

- Protocols are scientifically proven to improve outcomes (e.g., pilot or hospital checklists)
- Goals are to arm asset owners with the right questions and eliminate some obvious false positives
- The goal is not to eliminate all false positives – that is easy, just reject every strategy/manager. In doing this, we will miss many discoveries.
- My newest research focuses on the trade off of false positives and missed discoveries – an area that is unexplored in finance. **Maybe next time!**

# False (and Missed) Discoveries in Financial Economics

**Campbell R. Harvey**

*Duke University, Durham, NC 27708 USA*

*National Bureau of Economic Research, Cambridge, MA 02138 USA*

**Yan Liu\***

*Texas A&M University, College Station, TX 77843 USA*

<https://ssrn.com/abstract=3073799>



# Contact

W: <http://www.duke.edu/~charvey>

M: cam.harvey@duke.edu

T: @camharvey

SSRN: <http://ssrn.com/author=16198>

PGP: E004 4F24 1FBC 6A4A CF31 D520 0F43 AE4D D2B8 4EF4