Too many factors!

Equity Quant in a Post-Truth World

Quantitative Strategy
Global Markets

N Firoozye

October 2016

The subtleties of anomaly prediction

In-Sample, many strategies can look very appealing

Non-standard strategies' excess returns

Significant excess returns can be devised from following astrological strategies (e.g., Lunar Cycles)

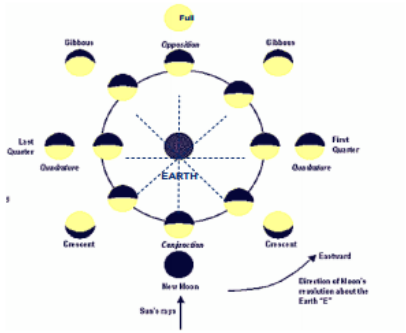
- Shorting the index on the new moon and going long on the full moon in order to make profits in the falling markets. „
- The proposed trading strategy can be summarised as follows: **buy an index on the new moon** (if this is a non-trading day, buy on the next trading day), **hold till the full moon** (usually 14-15 days), **sell the index on the full moon** (similarly, if non-trading day, sell on the next trading day).
- DAX buying would from 1959 to 2010 resulted in a return of 173%.
- DAX with *Buy on New Moon, Selling on Full Moon* would result in 756% return over the same period

Relative performance of the various stock indicies and using moon trading dates

	Period	Amount invested	Index Performance	Moon trading (buy new/sell full)	Compare with index %	Moon trading (buy full/sell new)	Compare with index %
FTSE 100	1984 - 2010	£1,000	£5,130	£12,116	269	£2,036	25
S&P 500	1928 - 2010	£1,000	£63,894	£1,502,689	2388	£2,571	2
DAX	1959 - 2010	£1,000	£17,361	£75,689	457	£3,971	18
EUROXX 50	1986 - 2010	£1,000	£3,084	£4,918	188	£1,770	37
Hang Seng	1964 - 2010	£1,000	£202,867	£778,722	385	£52,801	26
CAC 40	1987 - 2010	£1,000	£2,477	£3,698	183	£1,554	38
			Average	645%		24%	
				>100%		<100%	
			Conclusion	Over perform the index		Underperform compared to the index	

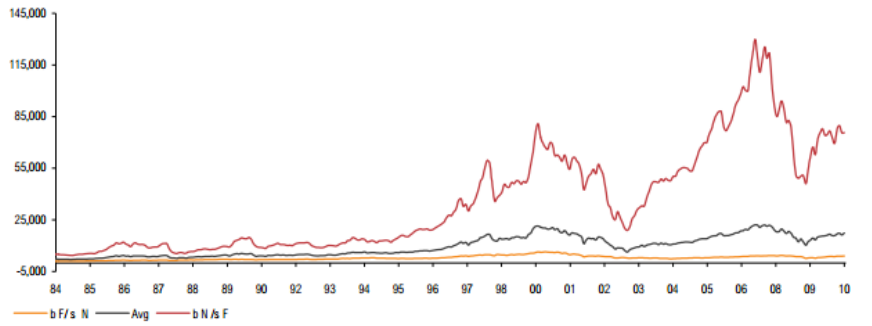
Source: RBS

The Lunar Cycle – The Moon's phases. Phases seen from Earth (E)



Source: Encyclopedia Americana International

Long Only DAX vs Long New, Short Full Moon Strategy



Source: RBS

Source: Sheer Lunacy Staring at the Heavens, RBS 7 July 2010 http://www.markettiming.nl/img/image/file/RBS_MoonTrading13Jun10.pdf

Weather-based strategies are statistically significant

Predictive performance – using indicators to predict Quant Factor returns (Novy-Marx, 2015)

- **NYC Temperature**
  - **Market, Size, Value, Long-Run Reversals, Asset Growth, Asset Turnover,** all have strongly significant negative betas to NYC Temperature
  - **Return-on-assets, Earnings-to-price, Gross-Margins, Earnings Momentum,** all have strongly significant positive betas to NYC Temperature
  - **Cold temperatures are good for market, small-cap and value,** while **hot temperatures are good for earnings-related anomalies**
- **Global Warming Anomaly:**
  - **Value**, **Long-Run Reversals, Earnings-to-Price and Investment (Change in PPE and Inventories) factors** have a **strongly significant** negative beta to global warming
  - **Gross-Margins** has a strongly significant positive beta to global warming
  - **Global warming is bad for value** and **long-run reversal strategies, good for market power**
- **El Niño** (Pacific Temperature Anomaly)
  - **Accruals, Beta Arbitrage (long low-beta, short high-beta), Gross-Profitability, Gross-Margins, Idiosyncratic Volatility factors (among others)** all have a strongly significant positive beta to El Niño
  - **Earnings quality based strategies do well when the El Niño is in full force**

Other statistically-significant non-standard strategies

Endless combinations of indicators work surprisingly well (Novy-Marx)

- **Astrological Occurrences**
  - **Conjunctions between Mars and Saturn**
    - **Market does** significantly worse when Mars and Saturn are in conjunction.
    - **High to Low Quality** does significantly better when Mars and Saturn are in conjunction
  - **Angle between Jupiter and Saturn**
  - **Number of Sunspots**
    - Well-documented coincidence between Maunder minimum and Dutch Tulip Mania (1637) and South Sea Bubble (1720), Black Monday (1987), Start of Great Recession (2007)
- **Political Events**
  - Democrat in the White House
    - **Market and Size** do significantly better
    - **Earnings to Price, Idiosyncratic volatility** do significantly worse

These results are striking, and quite surprising. In fact, some readers may be inclined to reject some of this paper's conclusions solely on the grounds of plausibility. I urge readers to consider this option carefully, however, as doing do so entails rejecting the standard methodology on which the return predictability literature is built.

Source: R Novy-Marx, Predicting anomaly performance with politics, the weather, global warming, sunspots and the stars http://rnm.simon.rochester.edu/research/PPiCToAPA.pdf

Take-aways : Don’t put all your faith in the back-test

In-Sample Performance can be very appealing

- How do we ensure performance is repeated?

- In Sample (**IS**) Performance (the “Back-test”)
  - Can be very deceptively appealing
  - Can be verified using standard statistical methods
    - Many predictors can be statistically significant in sample (IS).
    - Almost nobody adjusts for multiple tests (Do you know what trials didn’t work?)
- Out of sample (**OOS**) Performance (e.g., the “Live” period)
  - May be completely unrelated to IS performance
  - Tend to be when the strategy is proven
  - On the other hand, can be just plain lucky

The “factor zoo”

Where do all the factors come from?

Now we have a zoo of new factors.

- John H Cochrane, Presidential Address: Discount Rates, AFA, 2011

How many factors are there?

Fama and French started empirical factor research, post

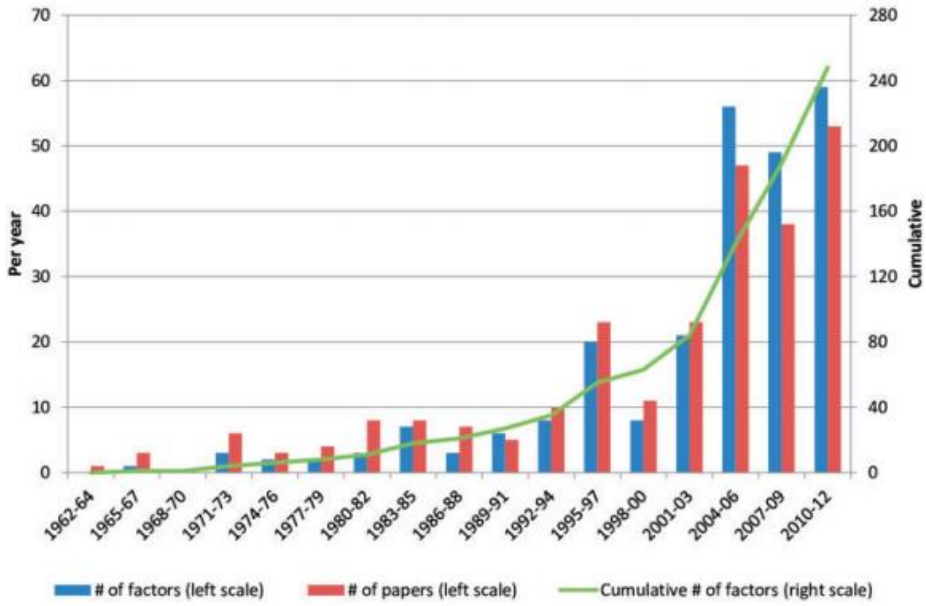
- **Fama French** –
  - 3 factor,
  - 5 factor in 1993,
  - 5 (different) factors 2003? recently. Possibly 7?
- Recent paper 15 – why some stocks underperform, some outperform
- **MSCI Barra** – 82 factors
  - Foundations of Factor Investing https://www.msci.com/documents/10199/71b6daf5-9e76-45ff-9f62-dc2fcd8f2721 differentiates between those earning a risk premium and those which just explain returns. – Value, Size, Momentum, Quality, Volatility, HighDiv
  - Some are country specific, but 26 are styles
- **S&P Capital IQ** - 400 factors or 450 factors (from website) or over 500 signals (http://marketintelligence.spglobal.com/documents/products/Alpha_Factor_Library_v2.pdf)
  - *Access over 500 stock selection and industry-specific signals spanning seminal academic literature and the latest practitioner expertise, coupled with S&P Capital IQ's Quantamental Research articles*

Even more factors have been published

Reputable journals

- Bob Arnott
- Total Academic Factors ‘discovered’
  - 315
  - Many hundreds more tested

Factors and Publications*



** … and the Cross-Section of Expected-Returns, Campbell Harvey, Yan Liu, Heqing Zhu, Review of Fin Stud, Dec, 2015, 10.1093/rfs/hhv059*

Data snooping, p-hacking and the crisis in science

Does it have anything to do with underperforming quant funds?

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk

Attributed to von Neumann by Enrico Fermi, as quoted by Freeman Dyson in "A meeting with Enrico Fermi" in Nature427 (22 January 2004) p. 297

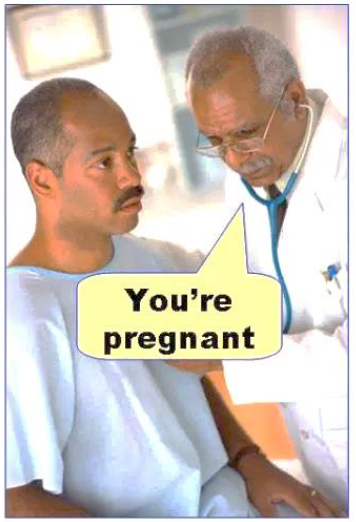
False Positives, False Negatives

Significance — Control some of the Type I (*incorrectly significant*) errors. Don't bet on duds!

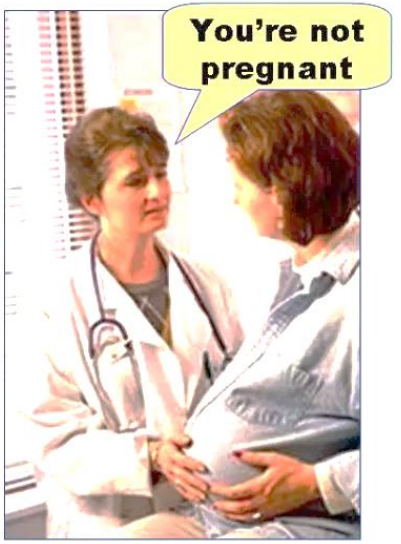
- **Exploratory Analysis is usually done on a data-set**
- **Hypotheses are tested**

- **Type I errors**
  - Strategies which are **incorrectly significant.**
  - Look good but lose money.
- **Type II errors**
  - Result in conservativeness
  - Can be damaging if far too conservative
- Type I errors are generally worse for trading strategies than Type II errors

Type I error
(false positive)



Type II error
(false negative)



Munchausen's Statistical Grid

If it's not significant, refine the analysis and try again!

The more analysis you do, the more thorough you seem

MUNCHAUSEN'S STATISTICAL GRID,WHICH MAKES ALL TRIALS SIGNIFICANT

Graham Martin, Neurosurgical Unit. Wellington Hospital, Wellington, New Zealand

Published: 29 December 1984, Volume 324, No. 8417-8418, p1457, 29 December 1984

he meets the deep but unformulated expectations of the sick for a sense of fraternity". Norman Bethune is a culture hero of the Chinese people, who died in the service of the Communist revolution. Harry Timbers died of typhus while working in the Ukraine during the famine of the 1920s. Very recently, a young American, Charles Clements, described his long march from Air Force pilot in Vietnam through medical school to a hard and dangerous life as doctor among the guerrillas in El Salvador, in *Witness to War*.

But being a selfless martyr and missionary is not what all the turmoil in the community is about, nor what the committees investigating the "failures" of medical education address themselves to. It is meeting patient expectations and at the same time fulfilling professional principles and range of knowledge. How find the people? How determine their promise? How then, proceed to instill the knowledge and the emotional response?

The division along these lines is too sharp. To be a socially minded and humanitarian doctor it ought not be necessary to go off and fight plague in Manchuria, or die in anticolonialist crusades. What the AAMC committee seeks, I believe, is what has been sought since the scientific revolution turned medicine around—science and humanity in medical practice. Francis Peabody, in the 1920s, commented that the "secret of the care of the patient was caring for the patient". Whether that can be attained by curriculum juggling, demanding poetry as a prerequisite course for admission to medical school, or assigning students to group instruction rather than attendance at lectures, remains to be seen.

Institution for Social and Policy Studies,
Yale University,
New Haven,
Connecticut, USA

GEORGE SILVER

Occasional Grid

MUNCHAUSEN'S STATISTICAL GRID, WHICH MAKES ALL TRIALS SIGNIFICANT

THE best-known story of the travels of the famous Baron von Munchausen[1] is his description of the thaw at the beginning of the Russian spring. Stopping to sleep while crossing the trackless snows of the steppe, he hitched his horse to a small post sticking out of the snow. The next morning he found himself in the centre of a village and his horse hanging by its bridle from the cross on the top of the church steeple. To get the horse down he had to shoot the bridle through with his pistol.

The Baron's medical writings were extensive but not much appreciated till this century. In 1951 Asher[2] published his description of the tall-tale-telling, peripatetic, poly-hospital, surgery-seeking-patient—the syndrome now universally known by the Baron's name.

Recently the Baron's statistical writings have come to light, and with them a statistical device, Munchausen's statistical grid. The great virtue of this device is that it can resurrect a significant result from any foundering therapeutic trial. First the patients in the trial are allocated to five classes (eg, age groups). Then the five classes are divided by sex, to make ten subgroups. The experimenter declares a probability of $p = 0 \cdot 05$,—ie, one in twenty—to be significant, and now there is a fifty-fifty chance that one of the subgroups will produce the desired result. Should the experimenter still have bad luck, subdividing the classes along some other line—eg, Scandinavian and non-Scandinavian patients—provides a second chance of winning; it should give twenty subclasses and considerably improve the chances of a significant result.

The Baron's system has substantial commercial applications. Once a drug has been found to have a probably significant effect in one of the cells, say in Scandinavian women aged 51–60, it may be marketed among the Scandinavians in Minnesota. Since it works for those aged 51–60, it probably works for others, and to confine it to Scandinavian women would be both sexist and racist. Thus, Munchausen's grid can rescue any trial from otherwise certain obscurity.

The Baron foresaw the criticisms that future generations might level at his device; according to him the arguments would run like this (see Carnap[3]). Suppose you have an urn containing red and blue balls, evenly mixed but in unknown proportions—possibly even no red balls at all. If in nineteen random tries you pulled out nineteen blue balls, you might suspect that there were few or no red balls in the urn, and that your chance of getting a red ball on the twentieth try was rather small. The Baron refuted this argument; everyone knew that if your wife had nineteen sons in succession the chance of having a daughter the next time was still fifty-fifty. On this basis he asserted the principle that all events were statistically totally independent of all others, and that therefore all the cells in his grid must be treated as though they were totally independent events like tossing a coin or the determination of the sex of a baby. Thus, there was no similarity between the events in any of the cells in this grid. For instance, in drug trials there was no similarity between Scandinavian men aged 51–60 and those aged 41–50. Therefore it was quite unreasonable to expect some evidence that a drug which worked for one group worked for other similar groups. It was mere cynicism to say that if a drug was ineffective in nine of ten cells in his grid, this said something about its potency.

Curiously, despite his profound intellect, the Baron did not carry this principle of the total independence of all events into his personal life. He shot an acquaintance when the coin they were playing with turned up heads nine times in a row, and discarded a mistress when the butler said nineteen times out of twenty that she was "not at home".

Munchausen's grid has another more subtle application when its use is concealed. If you wish to study, for instance, the relation between backache and changes on the lumbar spine X-ray, you examine a large but unspecified number of features (say forty) of the X-ray, and relate them to the presence or absence of pain. Statistical tests are applied, using a probability level of one in twenty, and you find that there is more back pain with square osteophytes than with round ones (a discrimination that only you and your colleagues can make). The main message of your paper then becomes that square osteophytes cause back pain. You ignore the second statistically significant result among the forty or so tested, that there is a relation between the amount of bowel gas on the film and back pain, because you are not a gastroenterologist. This method has been described as "casting the net widely" a technique known for centuries to improve the chances of catching a fish. Moreover, submitting a larger number of factors to statistical examination not only improves your chances of a positive result but also enhances your reputation for diligence.

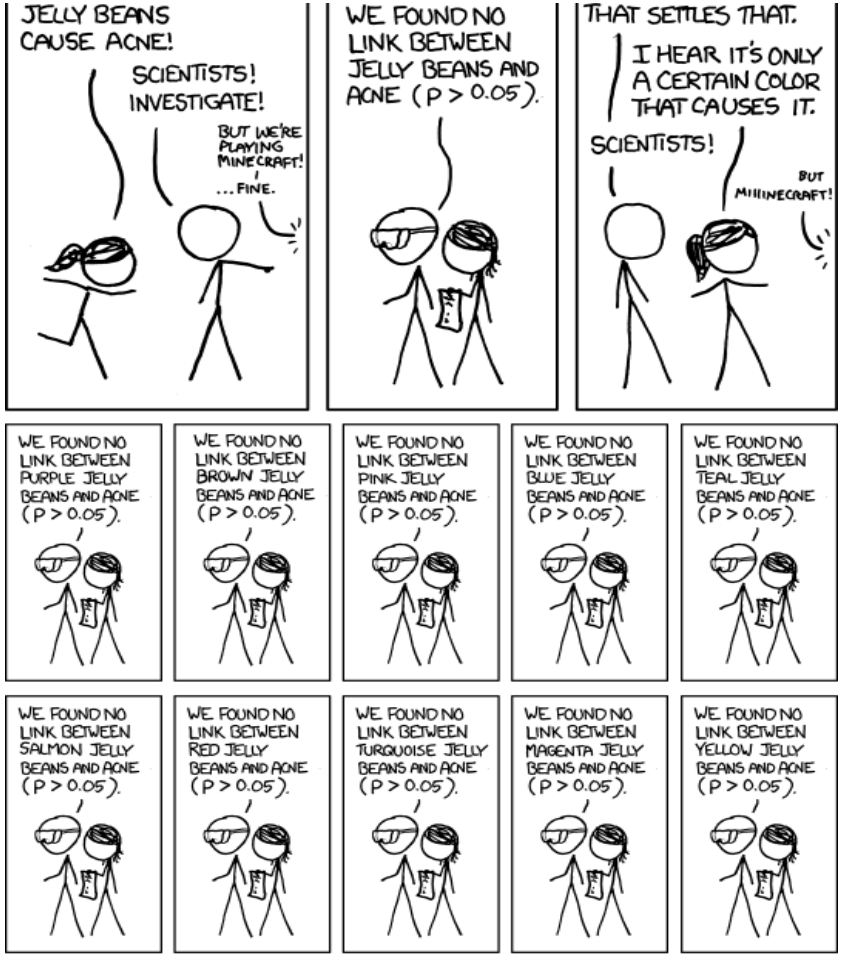
Neurosurgical Unit,
Wellington Hospital,
Wellington, New Zealand

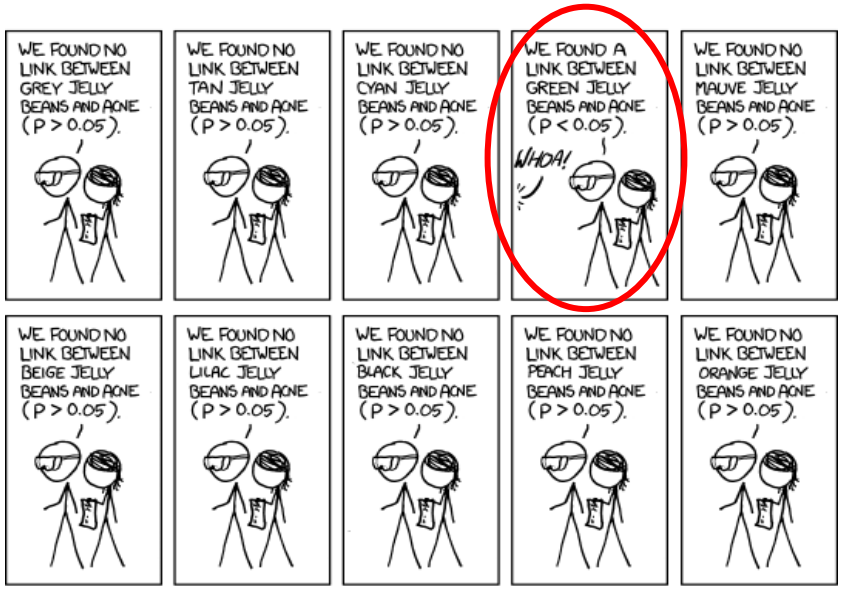
GRAHAM MARTIN

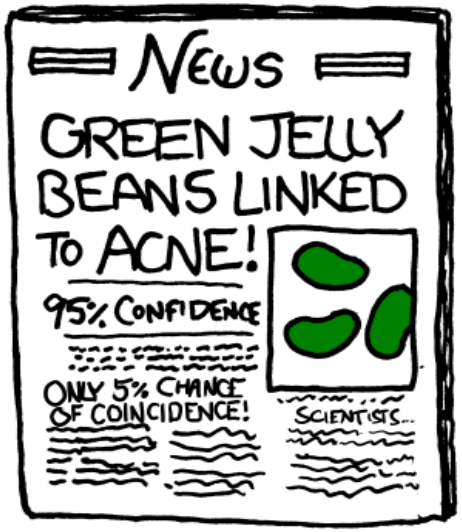
1. Raspe RE, et al (1785). The singular travels, campaigns and adventures of the Baron Munchausen. London: Cresser Press, 1948
2. Asher R. Munchausen's syndrome. *Lancet* 1951; i: 339–41.
3. Carnap R. Mathematics, an introduction to its spirit and use. San Francisco: Freeman WH, 1979: 153

Munchausen's Statistical Grid (in action)

If it's not significant, refine the analysis and try again!







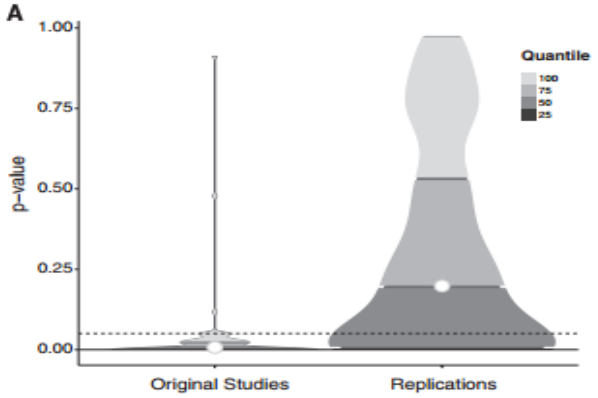
Source: XKCD (Significant): https://xkcd.com/882/

Data-snooping, Data-Dredging, P-Hacking – What's the big deal?

Science is facing a crisis of lack of reproducibility

- Hundreds of major studies in psychology and medicine have failed to be reproducible*
- Open-sourced data initiatives to publish both failed and successful trials are underway for future reproducibility
- Wasted research funds are estimated to be over $28bn/year***

Density plots of original and replication P values (100 replications by 270 Authors)**



PLOS MEDICINE

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

nature International weekly journal of science

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 | Corrected: 28 July 2016

The Atlantic

Psychology's Replication Crisis Can't Be Wished Away

It has a real and heartbreaking cost.

ED YONG | MAR 4, 2016 | SCIENCE

Slate

FUTURE TENSE THE CITIZEN'S GUIDE TO THE FUTURE. APRIL 19 2016 9:21 AM

FROM SLATE, NEW AMERICA, AND ASU

Cancer Research Is Broken

There's a replication crisis in biomedicine—and no one even knows how deep it runs.

2.1k 669 83

By Daniel Engber

*Source: *Why Most Published R* ://dx.doi.org/10.1371/journal.pmed.002012*

**Estimating the reproducibility of psychological science, Open Science Collaboration, Science Magazine,27 Aug 2015, Vol 349, 6251 - http://science.sciencemag.org/content/sci/349/6251/aac4716.full.pdf*

*** The Economics of Reproducibility in Preclinical Resesarch, Freedman-Cockburn-Simcoe, PLOS Biology June 9 2015, http://journals.plos.org/plosbiology/article?id=10.1371%2Fjournal.pbio.1002165*

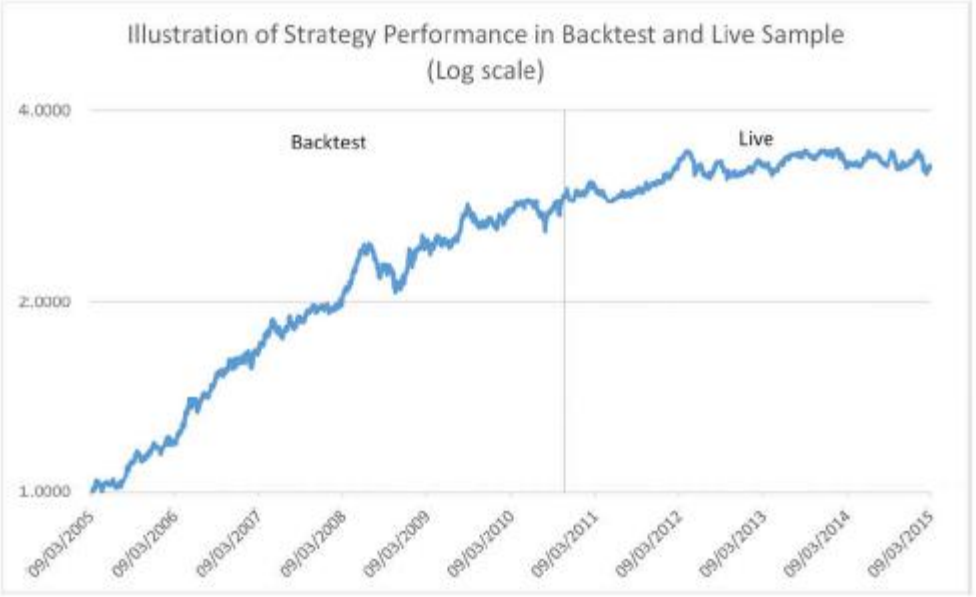
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106The *Extent and Consequences of –p-hacking in Science*

This could never happen in finance, right?

It happens far too often!

Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance

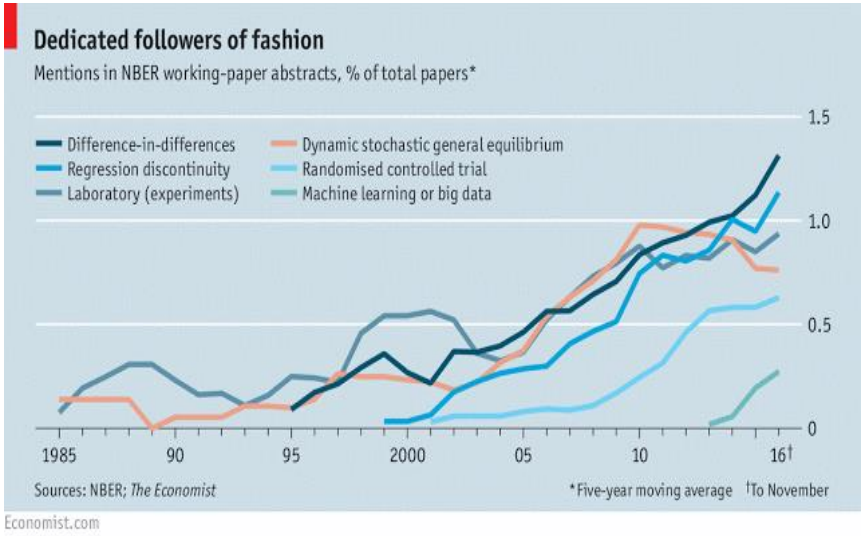
David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu



It is likely to get worse

Easily fit ML models (with hundreds of parameters) mean that we are likely to see much more overfitting

Economists are also using machine learning and big data more frequently



Standard Methods for preventing overfitting

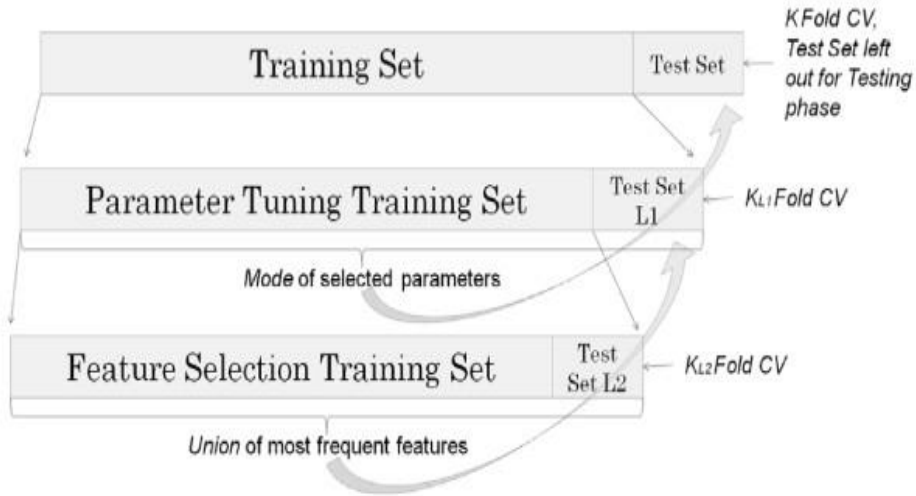
Train-Test-Holdout, or
Cross-Validation then Verification

Original data is split into
- Training Set
- Test Set

- Training set may be iteratively split into
  - **Test sets**
  - **Train sets**
  - Reused and iterated (e.g., Kx(Train/Test) via **Cross-Validation)** until model features and parameters are finalised

- Test or **Holdout Set**, hidden until after model is chosen, to estimate live performance

- Careful use of method should avoid overfitting
- **In practice, overfitting is still pernicious**



Now that we're careful, could anything go wrong?

Hacking data science competitions

Ignore the data, just fit to the holdout!
Holdout overfitting!

Dashboard — Public Leaderboard - Heritage Health Prize

This leaderboard is calculated on approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.

#	Δ1w	Team Name * in the money	Score	Entries
1	–	EXL Analytics *	0.443793	555
2	–	POWERDOT	0.447651	671
3	–	Dolphin	0.450403	555
4	↑1	jack3	0.451425	455
5	↓1	Hopkins Biostat	0.451569	444
6	–	Xing Zhao	0.453081	161
7	–	Old Dogs With New Tricks	0.454096	370
8	–	Areté Associates	0.454424	112
9	–	Alice Sasandr	0.454670	376
10	↑9	J.A. Guerrero	0.454728	173

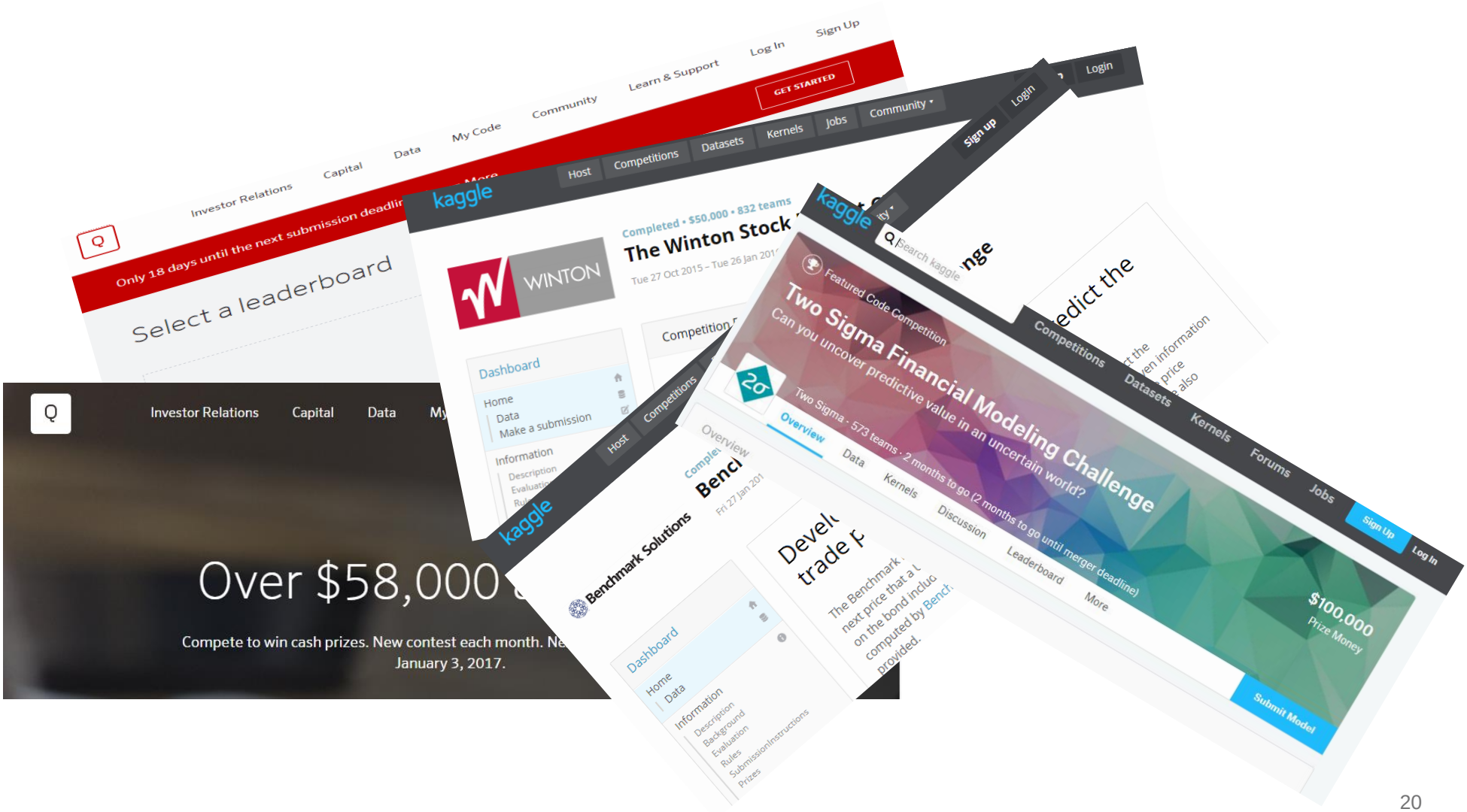
Public leaderboard of the Heritage Health Prize (Source)

http://blog.mrtz.org/2015/03/09/competition.html

Heritage Health Prize http://www.heritagehealthprize.com/c/hhp/leaderboard/public

Even Quant funds can fall prey to holdout overfitting!

Some strategies are only about in-sample performance
Just like another case of survivorship bias!



Take-aways

It doesn’t work unless you (pretty much) know what you’re looking for

- Multiple Testing allows us to find almost anything
- Strong Priors can prevent testing completely spurious relationships
  - Don’t “touch” your data very often!

How do you avoid data mining?

OOS performance is good but may just be luck

- **Panel Data**
  - **Different asset classes, currencies, exchanges, tenors, expiries, futures**
  - Koijen et al, Carry
  - Asness et al, Value & Momentum Everywhere
  - Morris et al, The substance in styles: why alternative risk sources should become standard
  - Morris, What if equity indices are like soybeans? How applying fixed income tools (and agnosticism) to equity index futures can add value
- **Strong priors / Domain Specific Knowledge / Intuition**
- **Explain Causality, not just Predictiveness**

- **Promising methods but not always applicable: Data snooping detection**
  - Multiple Testing adjustments
    - Holm/Bonferroni/BHY
  - Bootstrap detection:
    - White, Romano-Wolf, Model-Confidence-Spread
    - Harvey-Liu-Zhu