# Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Week 5

# Course outline

| Week | Description |
| --- | --- |
| 1 | Measurement, data types, data collection, data cleaning |
| 2 | Data manipulation, data exploration, visualization techniques |
| 3 | Probability, statistical distributions, descriptive statistics |
| 4 | Statistical hypothesis testing, quantifying confidence |
| 5 | Time series analysis, autoregression, moving averages |
| 6 | Linear regression, parameter estimation, model selection, evaluation |

Data & Inference

# WEEK 5A

# Today’s Lecture

| No. | Activity | Description | Time |
|---|---|---|---|
| 1 | Challenge | Detecting trends | 10 |
| 2 | Discussion | Trends and their importance | 10 |
| 3 | Case study | Trend following – commodity trading advisors | 10 |
| 4 | Analysis | Statistics for detecting trends | 20 |
| 5 | Demo | Techniques for identifying trends | 20 |
| 6 | Q&A | Questions and feedback | 10 |

# Definition of trend

- According to Wikipedia, trend may refer to:
- A fad, including fashion trends, or beauty trends.
- Trendsetter (or early adopter), person who starts (or follows early on) a fashion or technology trend before most other people
- Trendies, a teenage subculture common in Europe and the US from the 1990s-2010s
- Market trend, prolonged period of time when prices in a financial market are rising or falling faster than their historical average, also known as "bull" and "bear" markets, respectively.

# Class Poll 1

- What does the word “trend” mean to you?

**Slido.com**
**#99241**

# Trends

- Trends are useful for helping to understand what our community and peers are doing and thinking.
- We want to be able to identify trends based on data analysis.
- First we need data about the topic of interest and then we need to apply the correct statistical techniques.
- Statistical significance will also be important as we attempt to quantify the strength of the trend.

# The Bible and historical trends

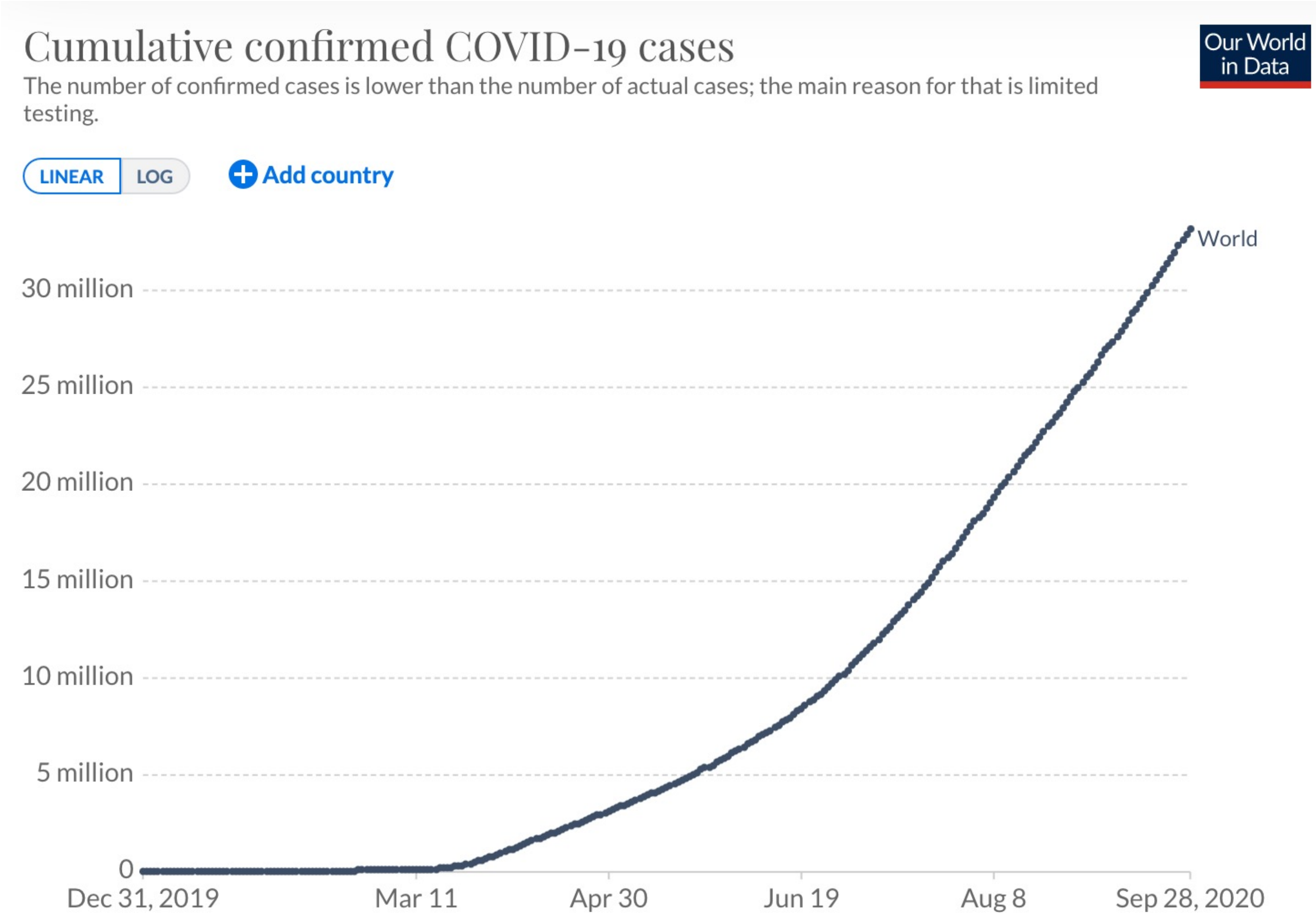
- Joseph interpreted the Pharaoh’s dream and warned Egypt that there would be 7 years of abundance followed by 7 years of famine.
- “Behold, seven years of great abundance are coming in all the land of Egypt; and after them seven years of famine will come, and all the abundance will be forgotten in the land of Egypt, and the famine will ravage the land.”
  - Genesis: 41:30

# Hurst and the Nile

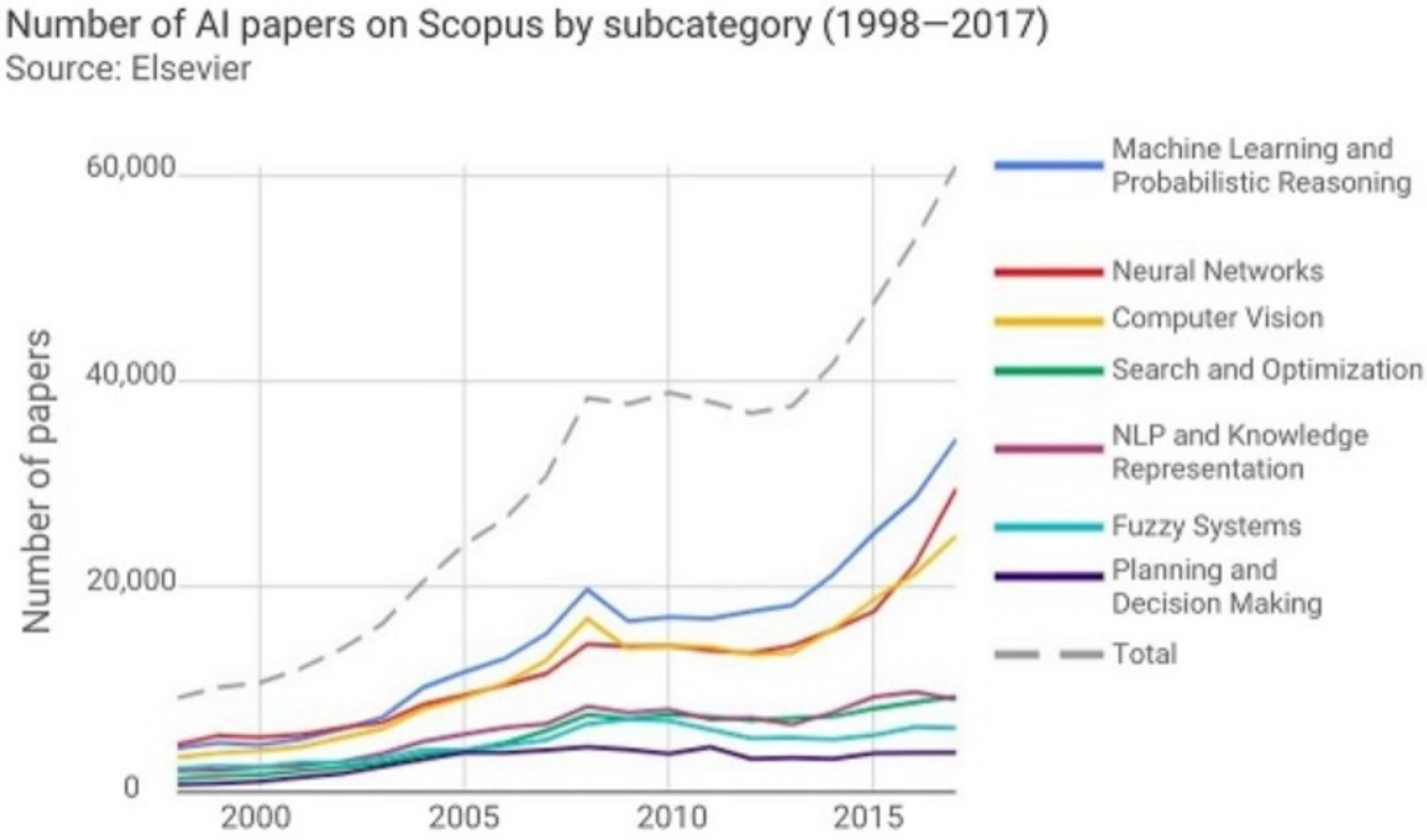
- Hurst (1900-1978) was an English hydrologist, who worked in the early 20th century on the Nile River Dam project.
- When designing a dam, the yearly changes in water level are of great importance in order to adapt the dam's storage capacity to the natural environment.
- Studying an Egyptian record of the Nile River's overflows over 847 years, Hurst observed that flood occurrences could be characterized as persistent.
- This means that heavier floods were accompanied by above average flood occurrences, while below average occurrences were followed by minor floods.

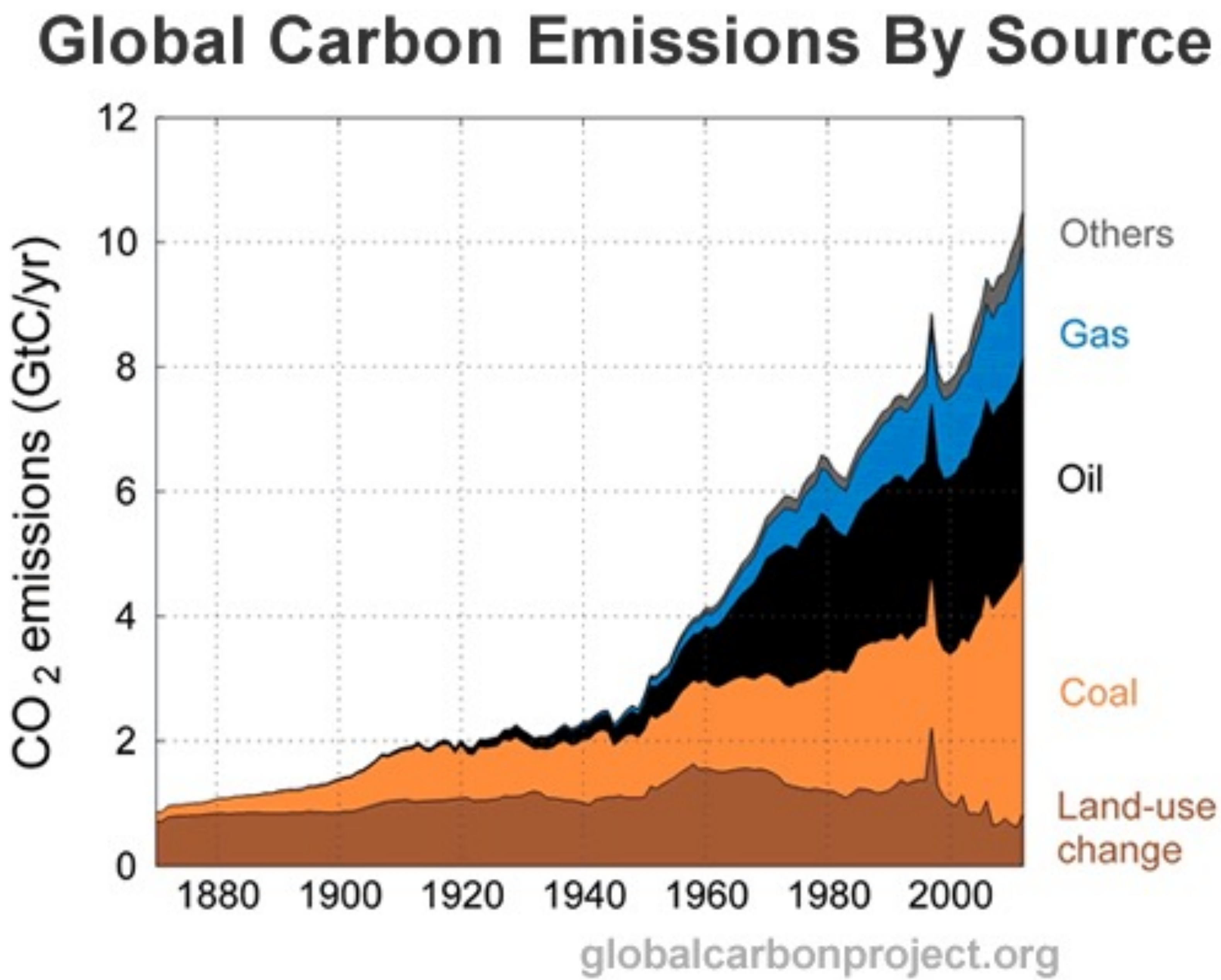
# Trend following

# COVID-19 cases

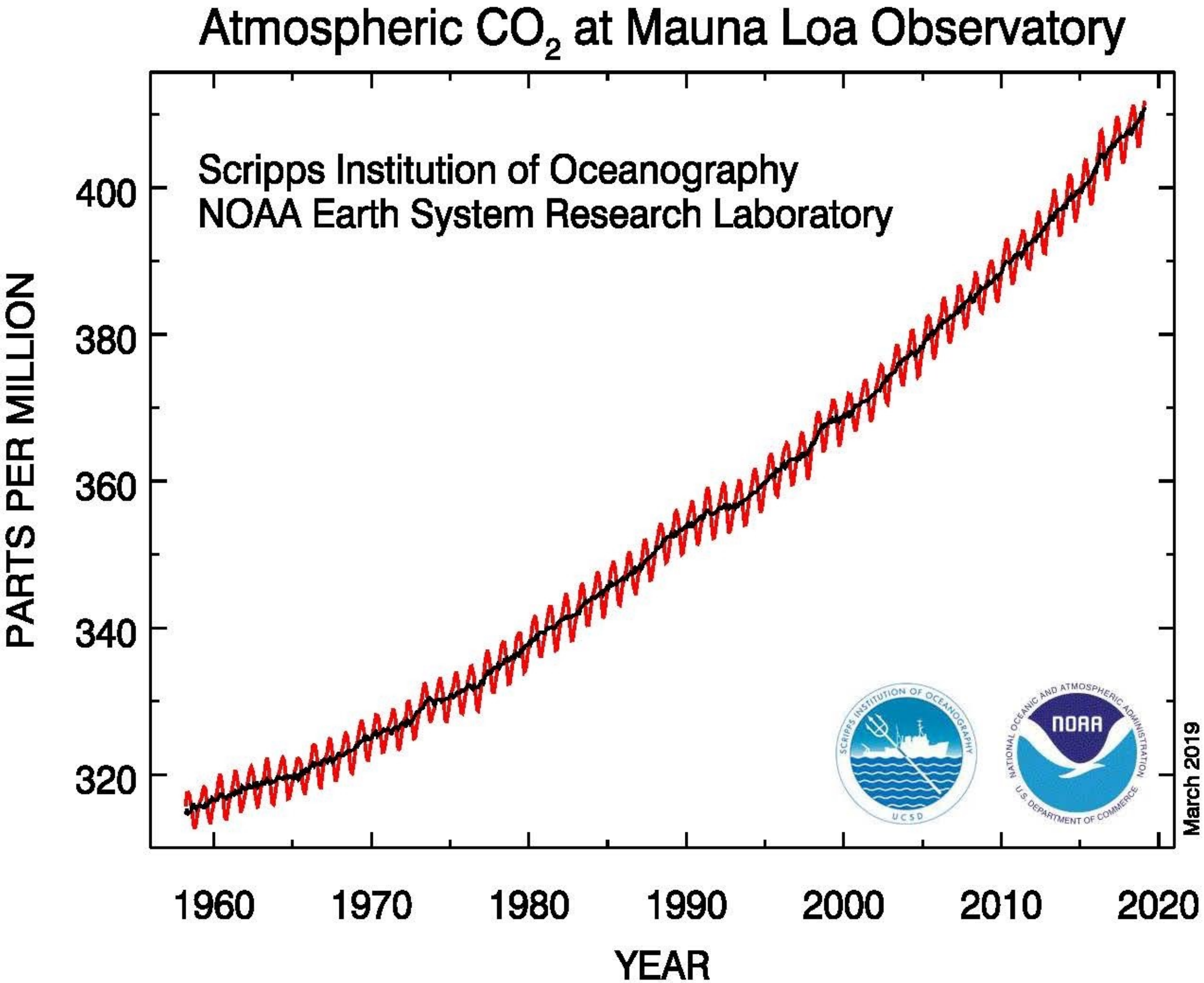


# AI Trends

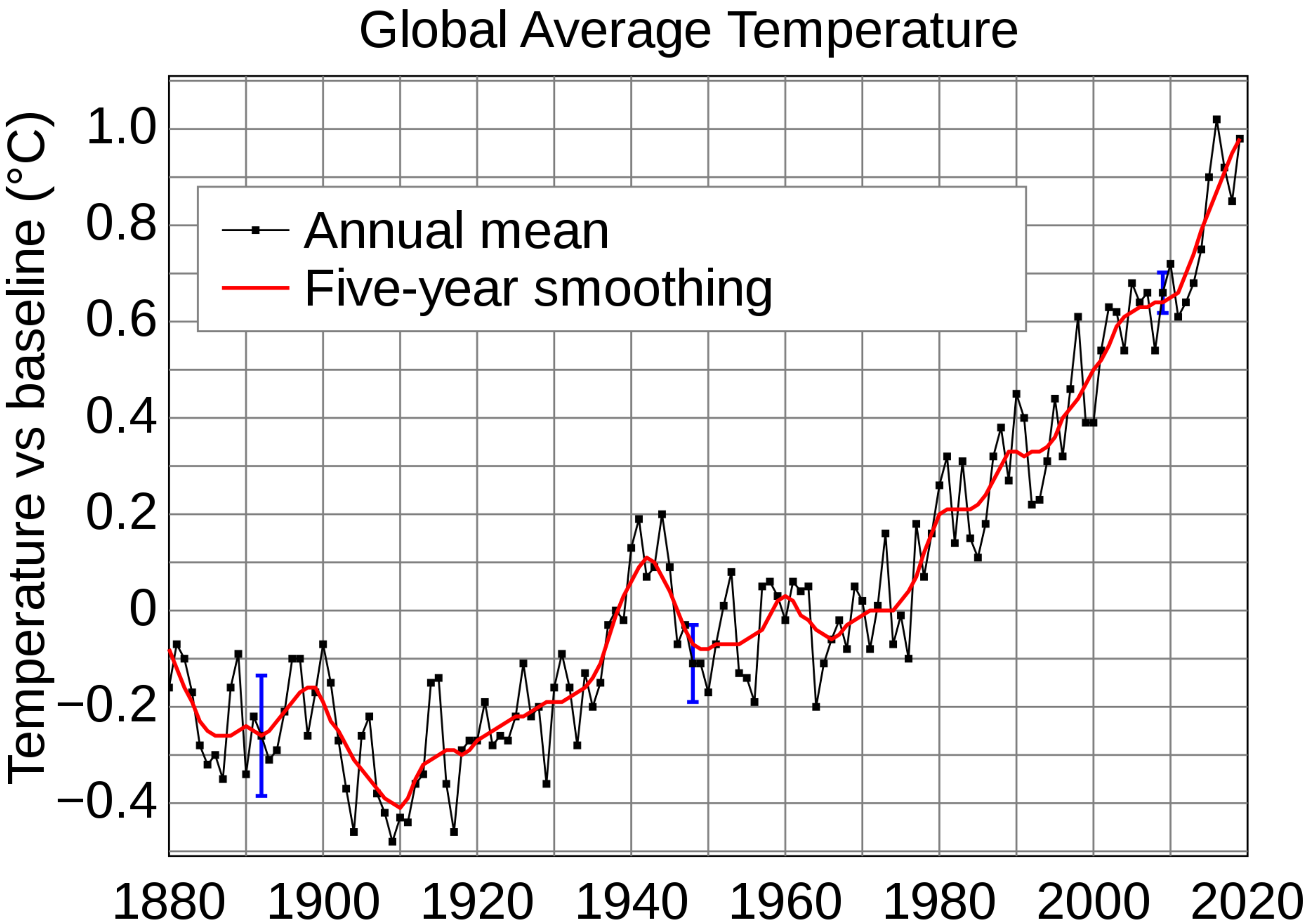

Number of AI papers on Scopus by subcategory (1998—2017)
Source: Elsevier
Number of papers
60,000
40,000
20,000
0
2000
2005
2010
2015
Machine Learning and Probabilistic Reasoning
Neural Networks
Computer Vision
Search and Optimization
NLP and Knowledge Representation
Fuzzy Systems
Planning and Decision Making
Total


Global Carbon Emissions By Source
CO 2 emissions (GtC/yr)
0
2
4
6
8
10
12
1880
1900
1920
1940
1960
1980
2000
Others
Gas
Oil
Coal
Land-use change
globalcarbonproject.org

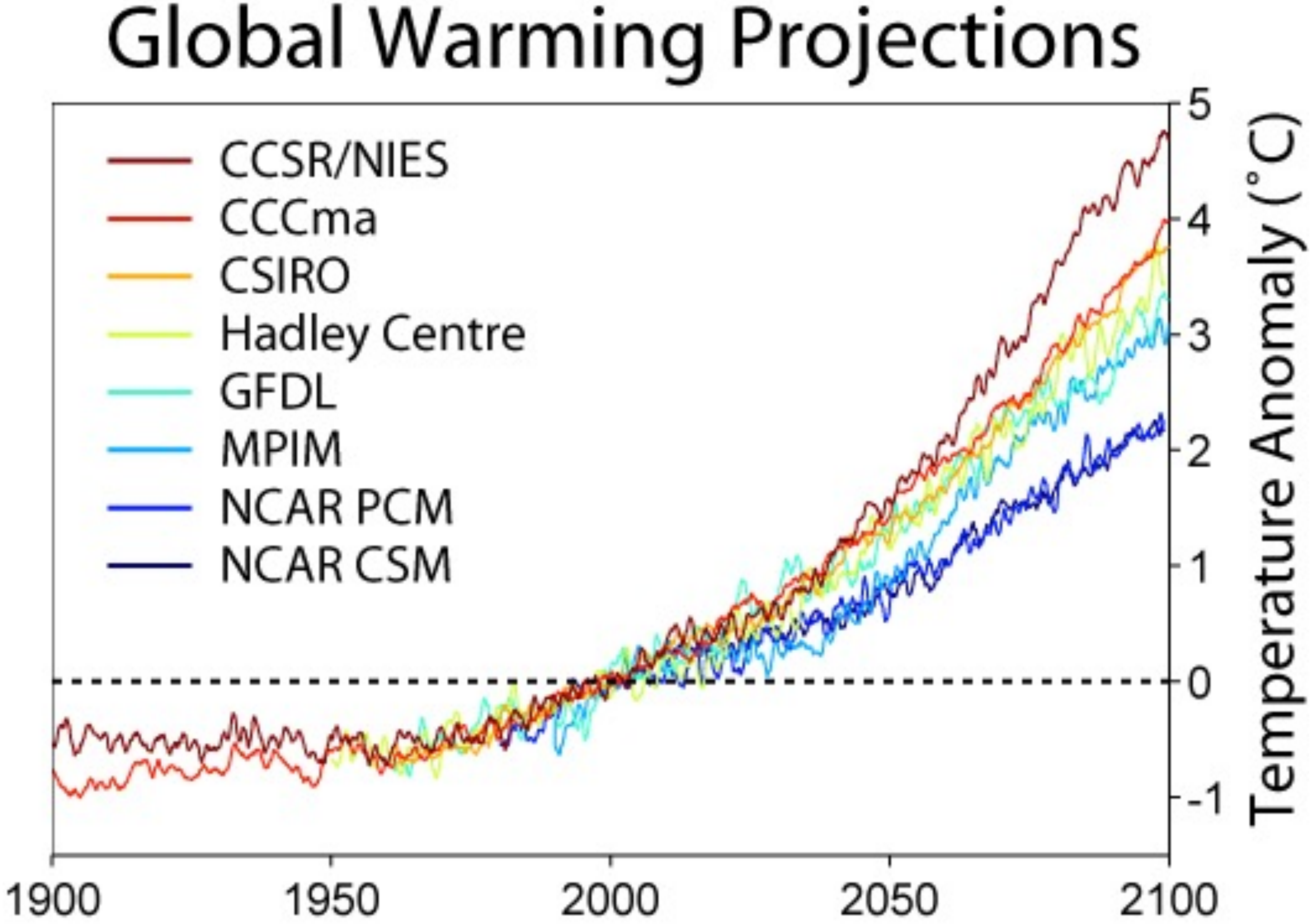
# Carbon Concentration



# Global Warming



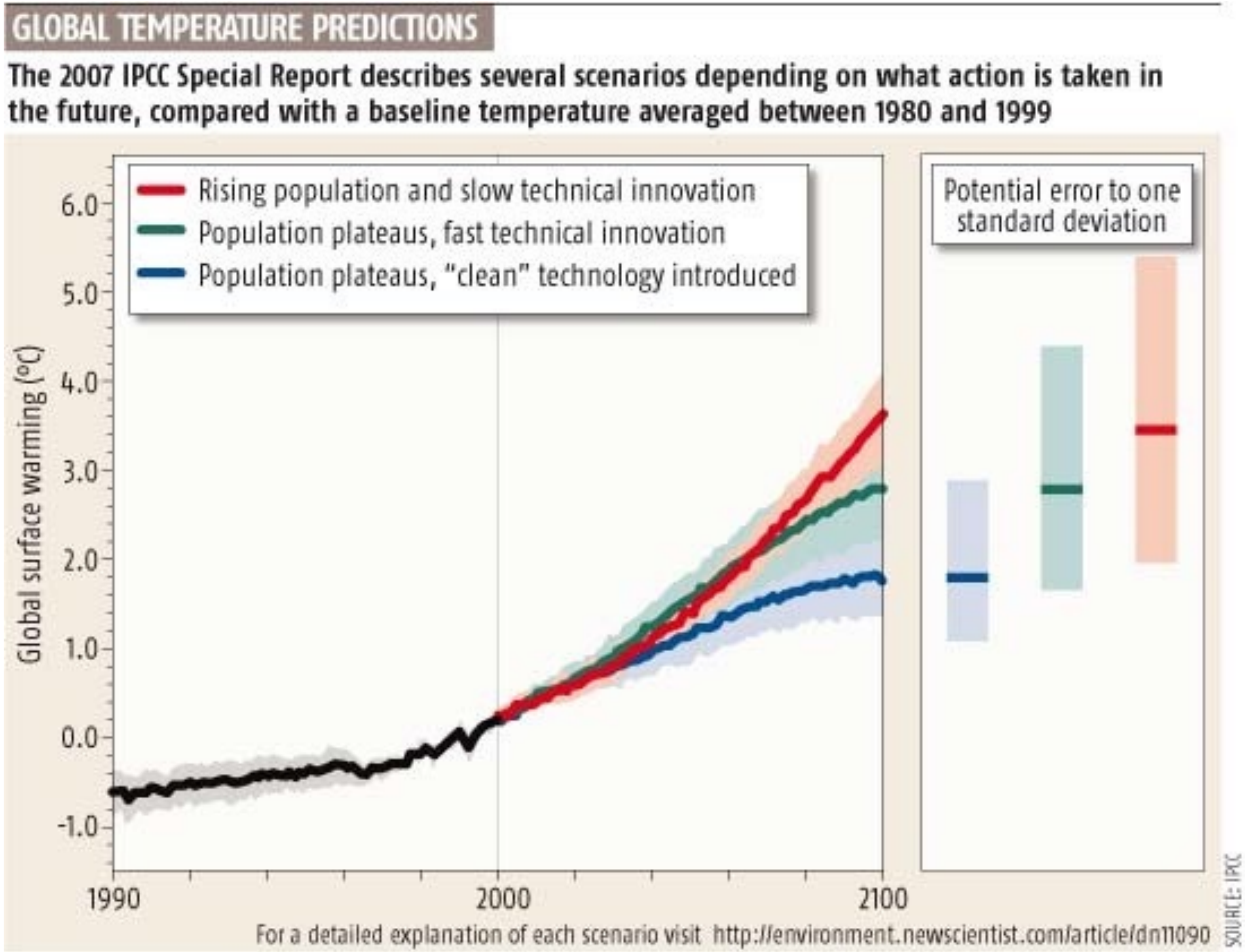

Global mean surface temperature change since 1880. Baseline temperature is about 14 °C.
Source: [NASA GISS]()



Global Warming Projections
CCSR/NIES
CCCma
CSIRO
Hadley Centre
GFDL
MPIM
NCAR PCM
NCAR CSM
Temperature Anomaly (°C)
5
4
3
2
1
0
-1
1900
1950
2000
2050
2100


Source: IPCC

# IPCC Scenarios

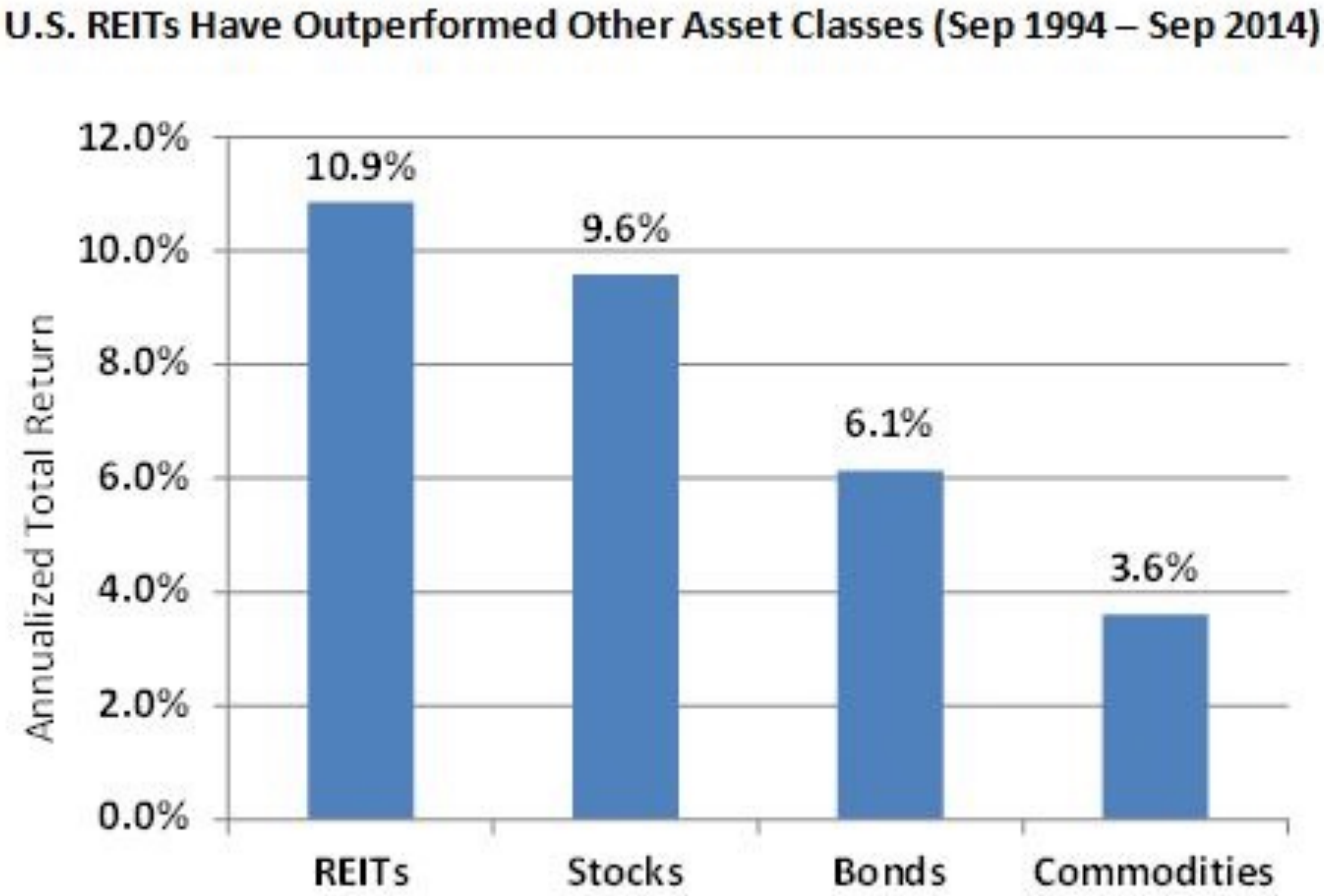


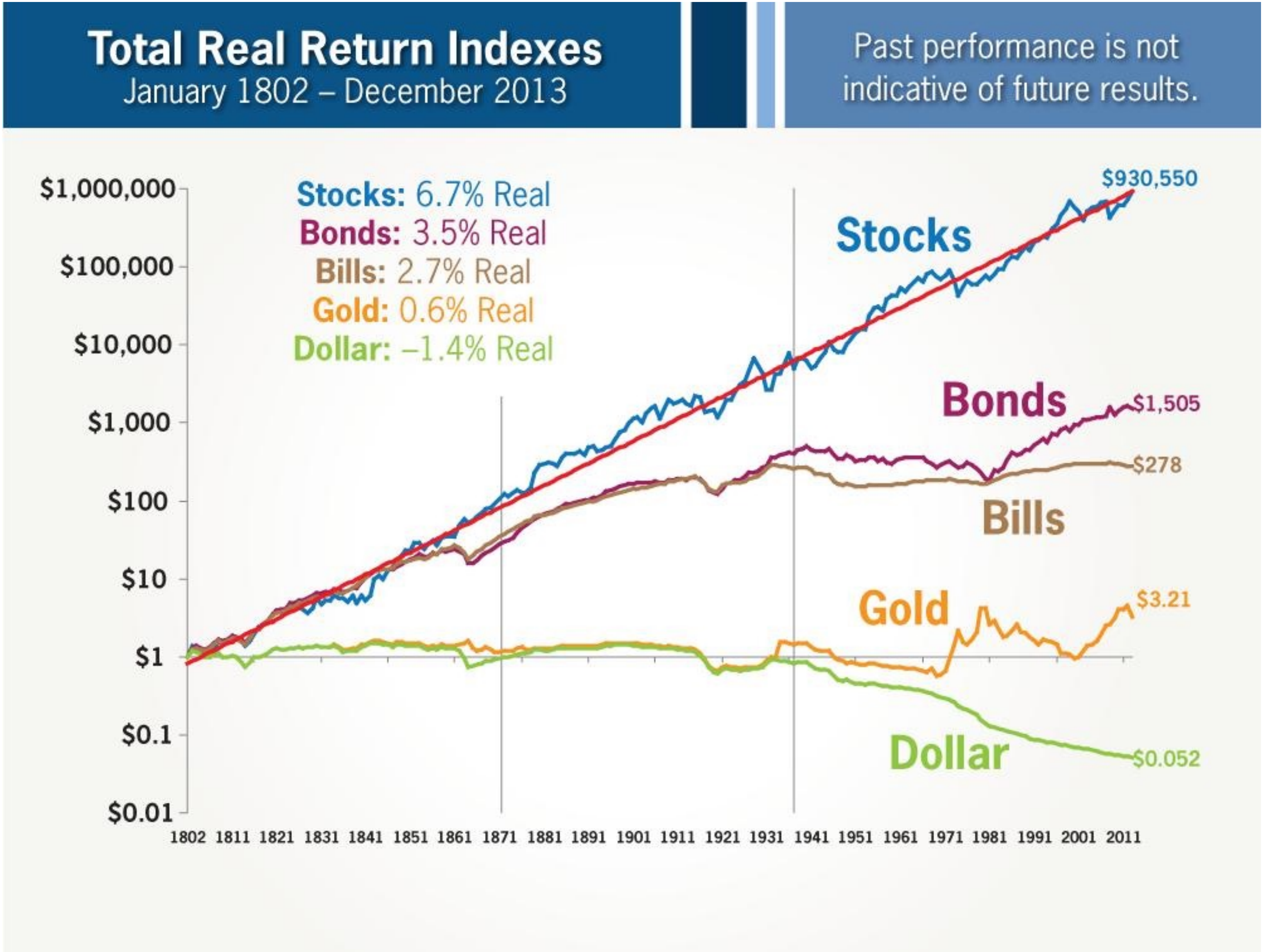
# Class Poll 2

- Which scenario do you think is most likely?
  - A: Rising population and slow technical innovation
  - B: Population plateaus, fast technical innovation
  - C: Population plateaus, clean technology introduced

**Slido.com**
**#99241**

# Trends in different asset classes

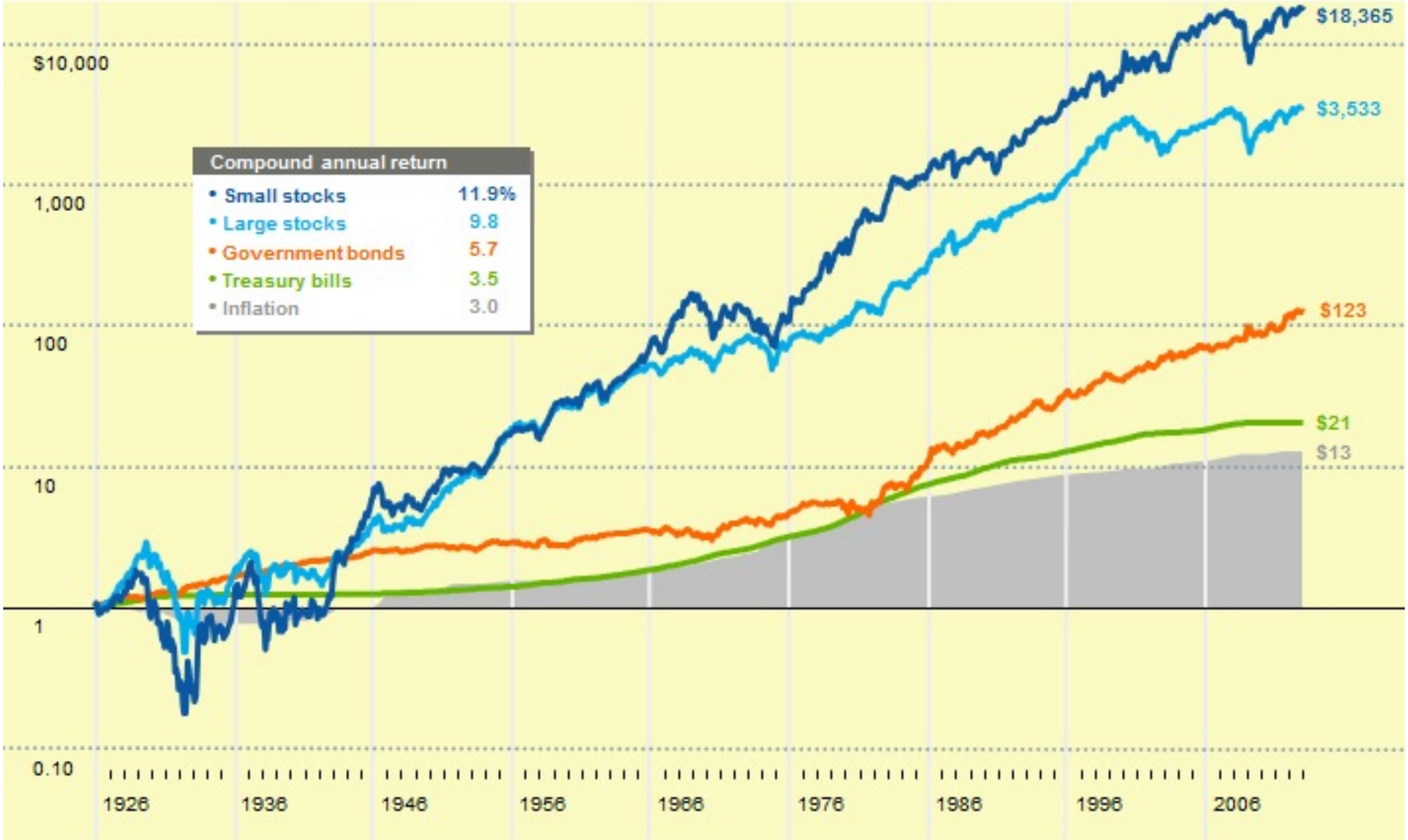





Source: www.aaii.com

**Ibbotson® SBBI®**

Stocks, Bonds, Bills, and Inflation 1926–2012



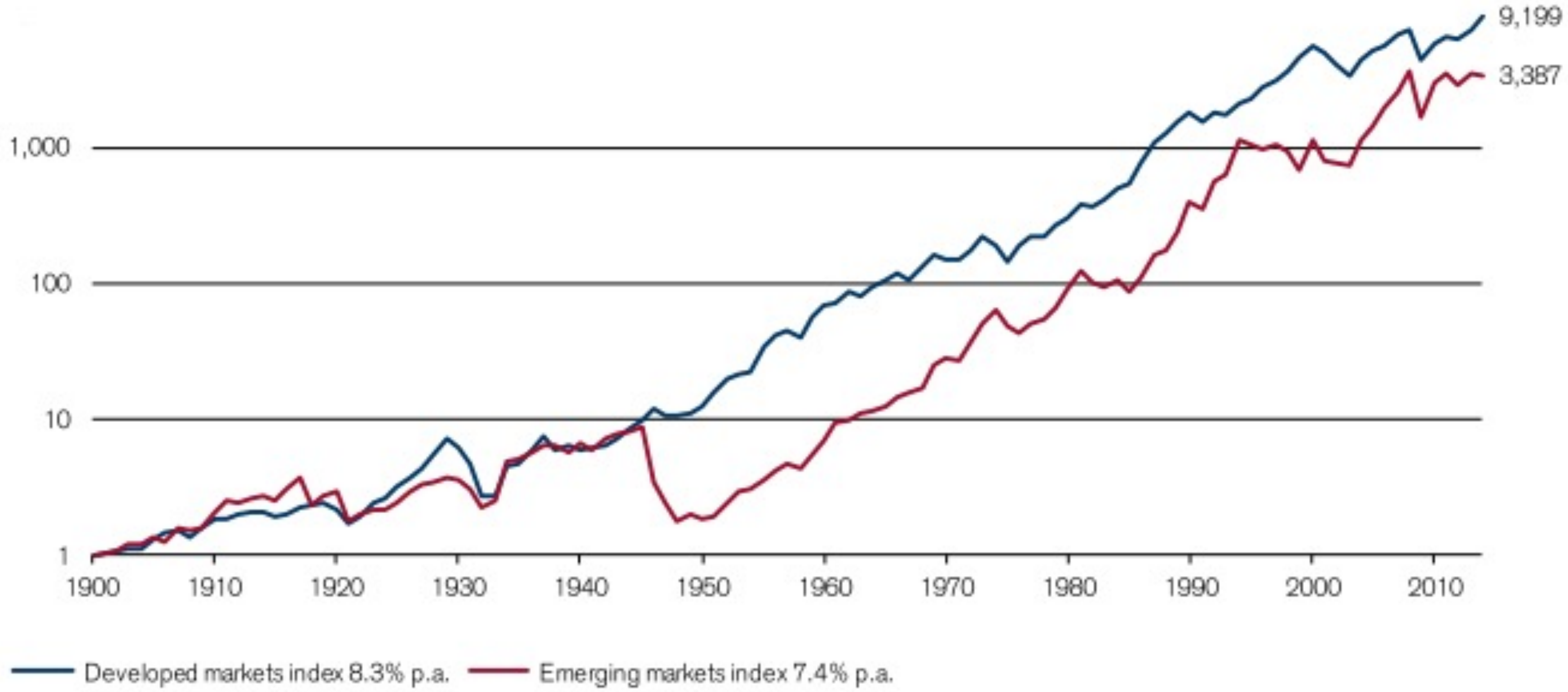

**Past performance is no guarantee of future results.** Hypothetical value of $1 invested at the beginning of 1926. Assumes reinvestment of income and no transaction costs or taxes. This is for illustrative purposes only and not indicative of any investment. An investment cannot be made directly in an index. 

# Developed and Emerging

Figure 2

**Long-run emerging and developed market returns, 1900–2013**

Source: Elroy Dimson, Paul Marsh and Mike Staunton using data from DMS database, MSCI Barra, and S&P/IFCG



# Class Poll 3

- Harry Brown's permanent portfolio fixes weights for stocks, bonds, gold, and cash (treasury bills) with the following percentages:

**Slido.com**
**#99241**

# Is there a trend?

- Being able to use a statistical methodology to detect trends is important for a variety of applications
- Some forecasting models depend on identification of trends
- The optimal model to use depends highly on knowledge about the existence or lack of trends
- Early warning systems aim to detect the formation of trends before some adverse event occurs:
  - Natural disasters
  - Terrorism events
  - Trending topics in twitter

# Moving averages

- A moving average may be viewed as a convolution or a low-pass filter.
- A simple moving average of the last n observations is given by:

$$s_t = \frac{1}{n}\sum_{i=1}^{n} x_{t-i+1}$$

- By smoothing the time series, it can remove the effects of seasonality.
- Note that it gives equal weight to both new and old observations

# COVID-19 Monitoring Quiz

- For monitoring COVID-19 daily cases, it is recommended to use an n-day simple moving average where n is:
- 3, 7, 10, 15

**Slido.com**
**#99241**

# Exponentially weighted moving averages

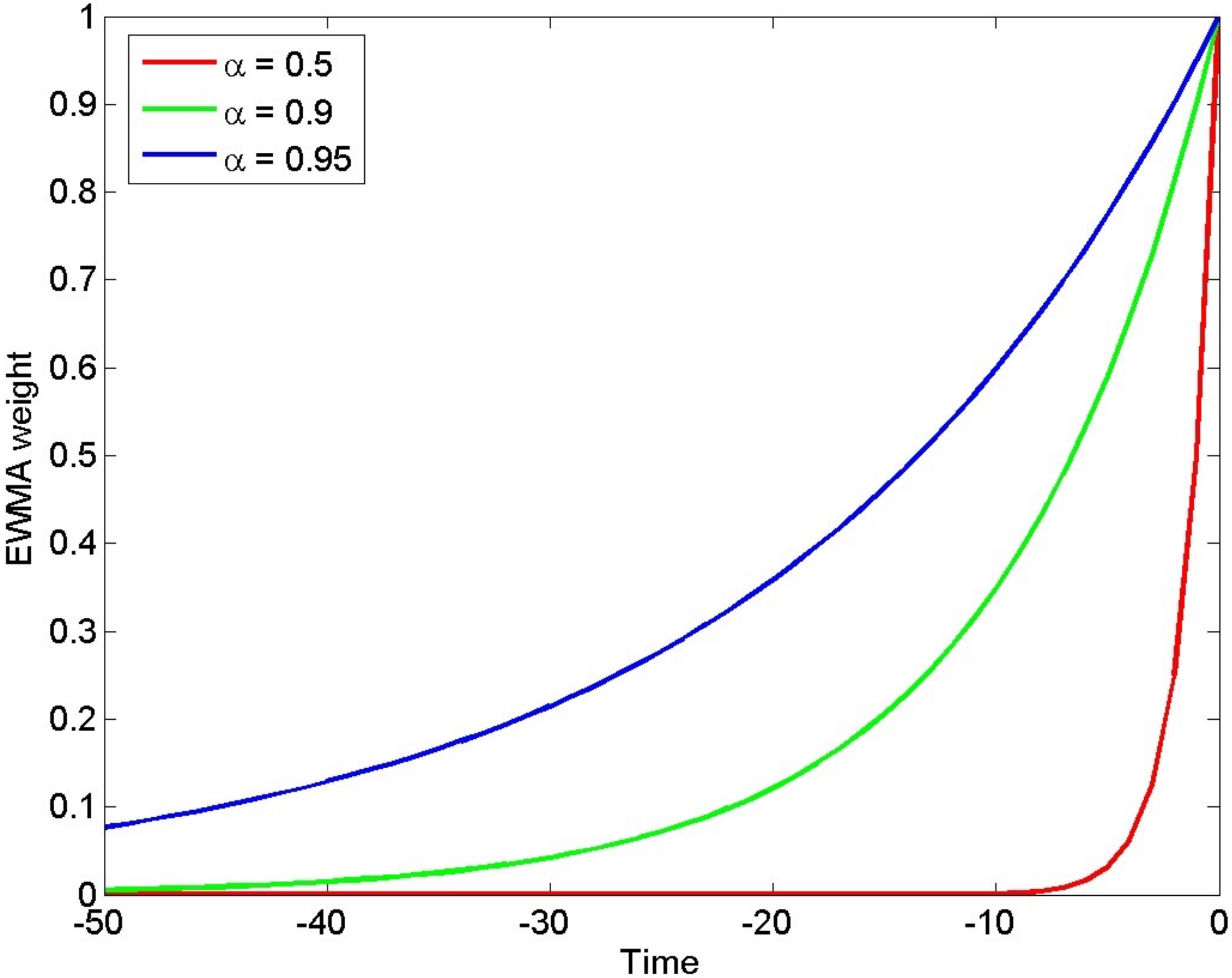
- EWMA employs exponentially decreasing weights to discount the influence of old observations:

$$s_t = \alpha x_t + (1-\alpha)s_{t-1}$$

- The smoothing factor, $0<\alpha<1$,determines the rate of decay of old information
- The EWMA may also be expressed via the number of time periods *n* using the smoothing factor, *α=1/n, giving:*

$$s_t = \left(\frac{1}{n}\right)x_t + \left(\frac{n-1}{n}\right)s_{t-1}$$

# EWMA weights



# Moving average crossover

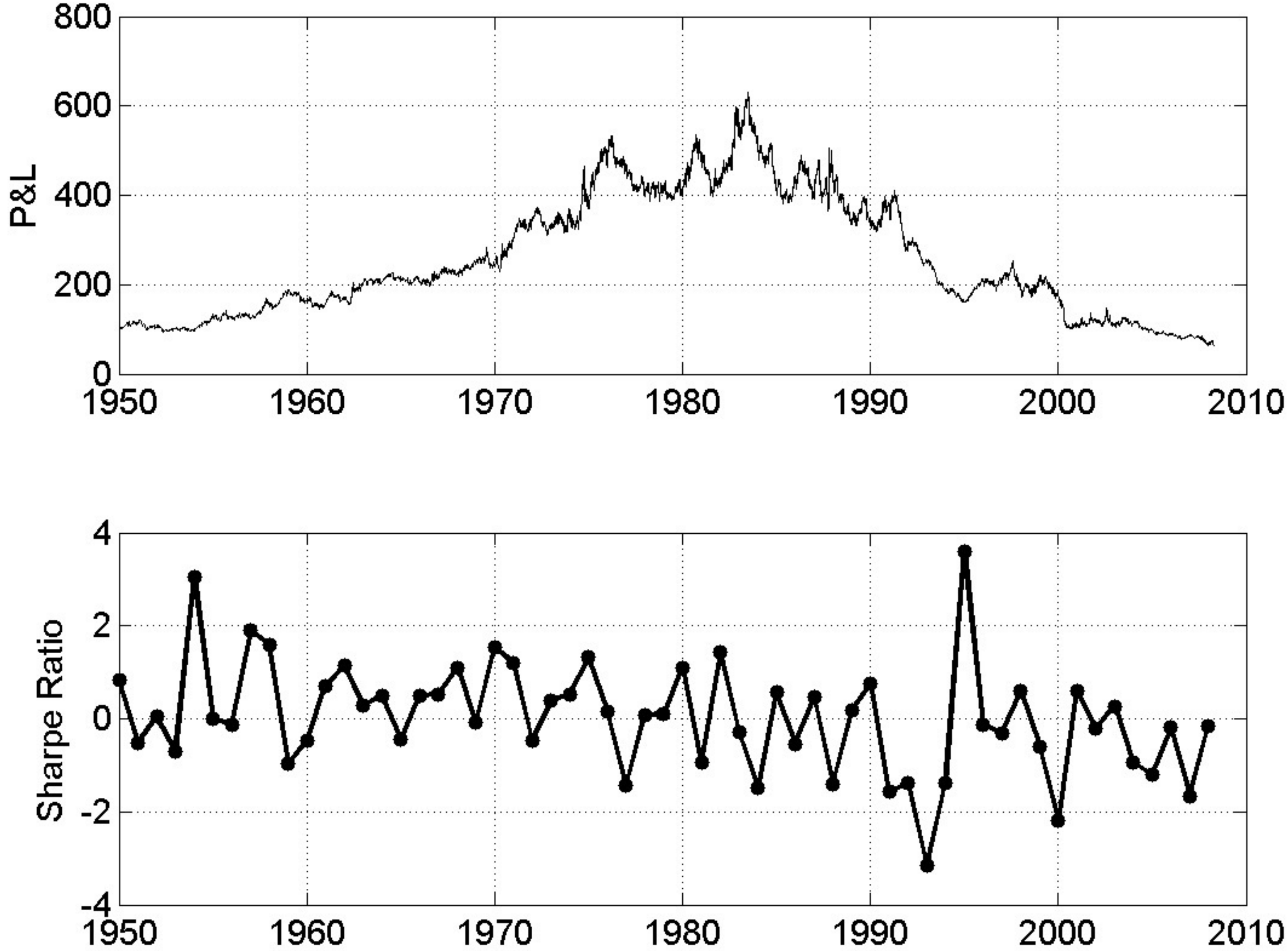
- In finance, trends are often identified using the crossover of two moving averages.
- For example, one could use a 50-day and 200-day moving average.
- SMA(p,50) will oscillate around the SMA(p,200) providing buy and sell signals.
- Buy when $SMA(p_t,50) > SMA(p_t,200)$
- Sell when $SMA(p_t,50) < SMA(p_t,200)$

# Demonstration on S&P500

# Technical trading rules

- Brock et al. (1992)
- Oscillator: $s_t = \text{EWMA}(p_t,n_s) - \text{EWMA}(p_t,n_l)$
- Buy (sell) when short period moving average moves above (below) the long period moving average
- Buy when $s_t > 0$ and sell when $s_t < 0$
- Could reduce trading costs using a band
- We consider the rule $(n_s,n_l) = (1,50)$ assuming trading costs of 20 basis points

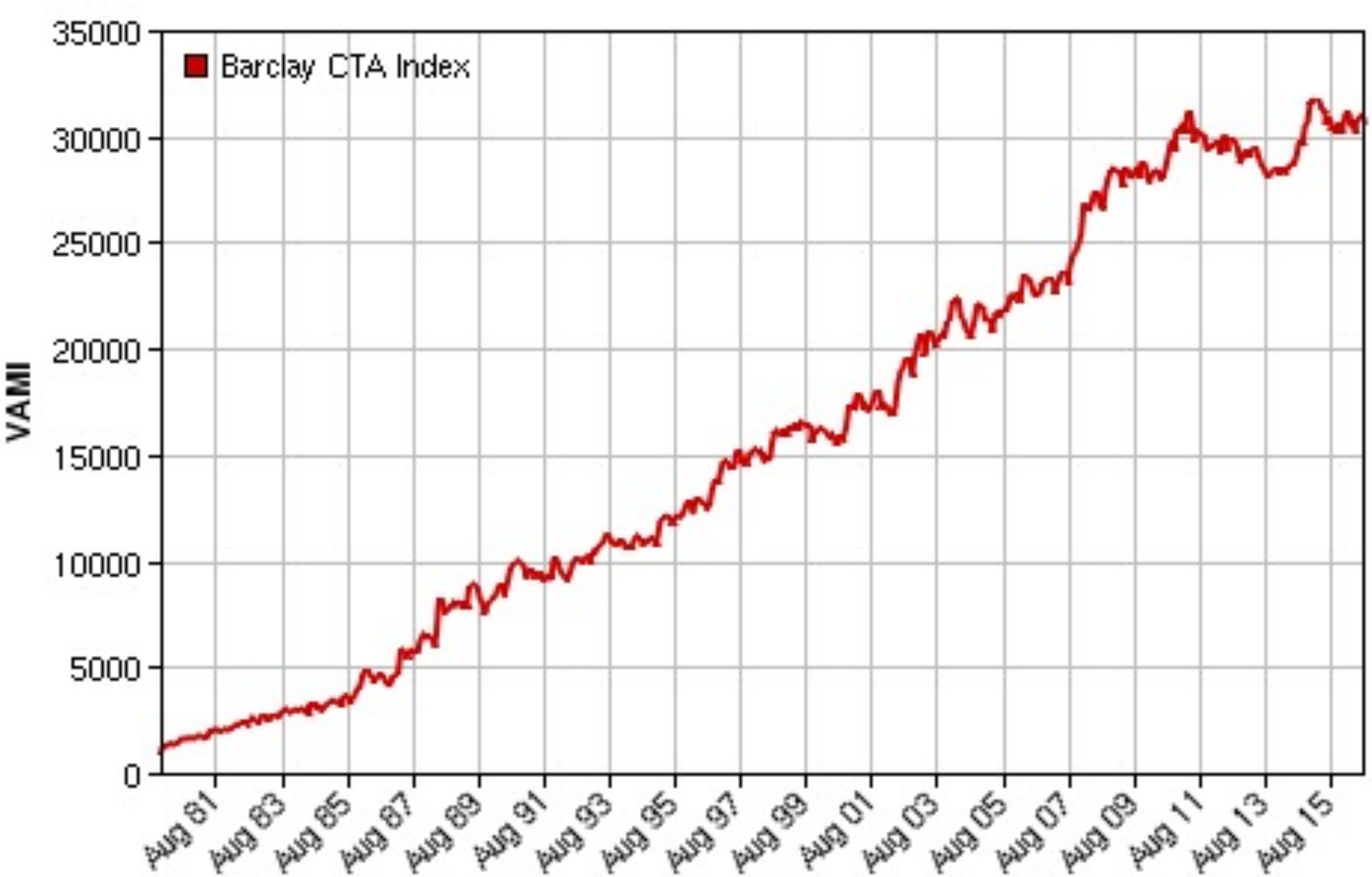
# TTR on S&P500 during 1950-2008



# CTAs versus stockmarket


www.rcgmanagedfutures.com

# CTA performance



# Sharpe ratio

- The Sharpe ratio is a measure of the excess return (or risk premium) per unit of risk in an investment asset or a trading strategy:

$$SR = \frac{\mu}{\sigma}$$

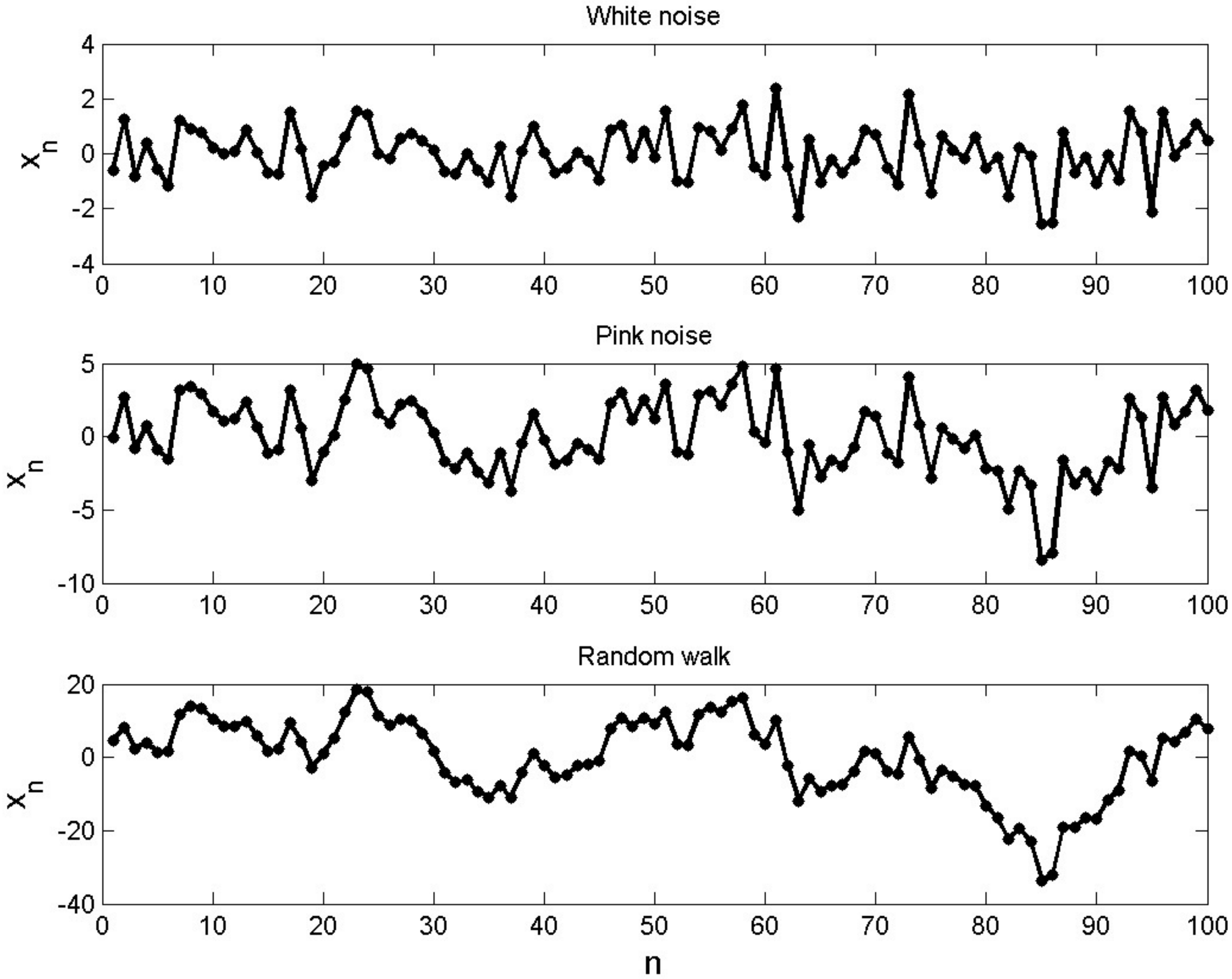
  where μ is the mean and σ is the standard deviation of the return time series
- In some cases the mean and standard deviation relate to the excess return above a benchmark or risk free asset

# CTA performance (1980-2016)

| Metric | Value |
|---|---|
| Compound Annual Return | 9.78% |
| Sharpe Ratio | 0.38 |
| Worst Drawdown | 15.66% |
| Correlation with S&P500 | 0.01 |
| Correlation with US Bonds | 0.14 |
| Correlation with World Bonds | 0.00 |

http://www.barclayhedge.com

# Fractal time series



# V-ratio test

- The variance ratio test assesses the null hypothesis that a univariate time series y is a random walk.
- The null model is

$$y(t) = c + y(t-1) + e(t),$$

- where c is a drift constant and e(t) are uncorrelated innovations with zero mean.

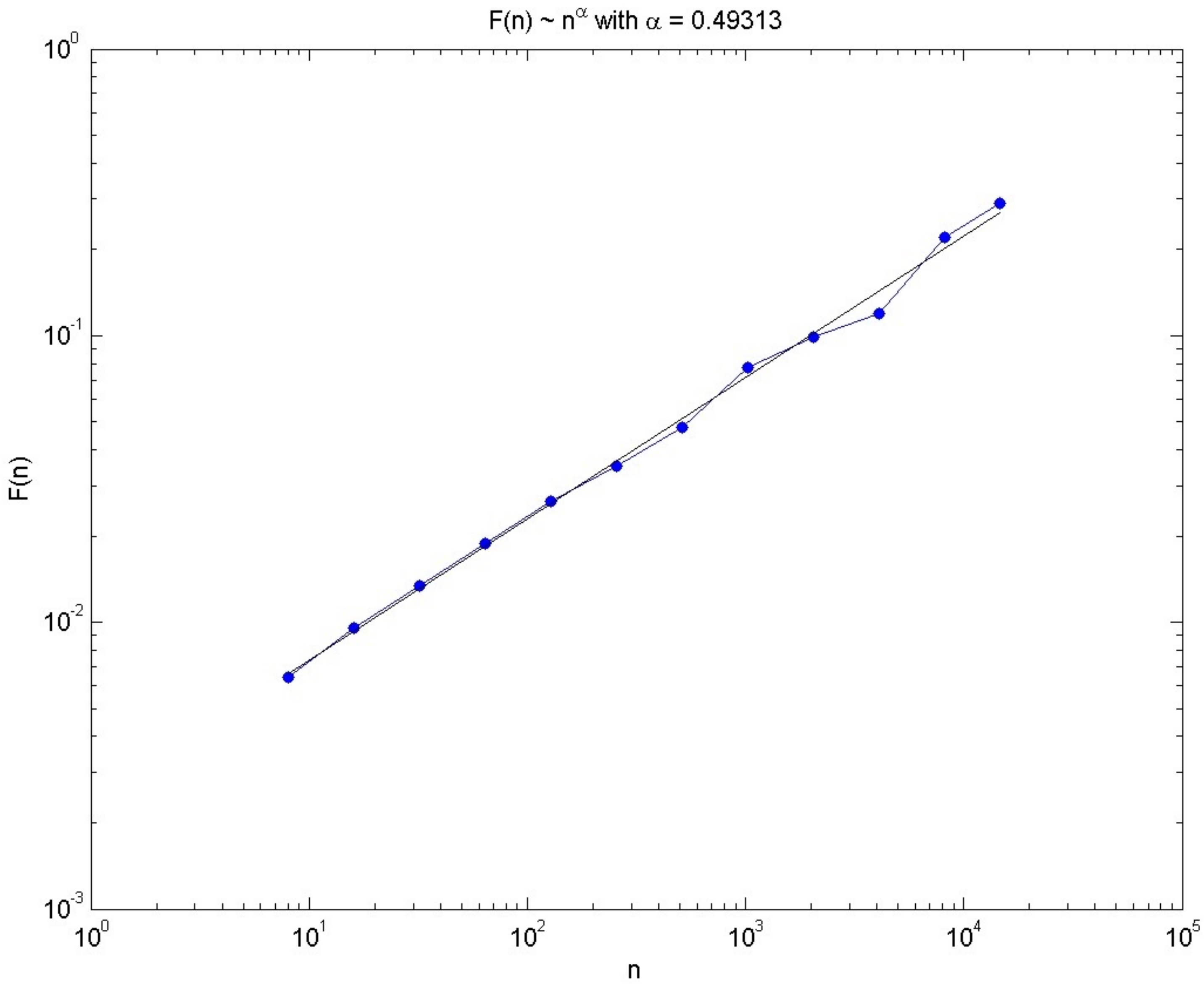
# Detrended Fluctuation Analysis

- The raw time series *x(i)* is first integrated to give *y(i); i=1,...,N*
- For each length scale, n, *y(i)* is divided into segments of equal length, n
- In each segment, the data is detrended by subtracting the local linear least squares fit, $y_n(k)$. The root-mean-square fluctuation of this integrated and detrended time series is given by

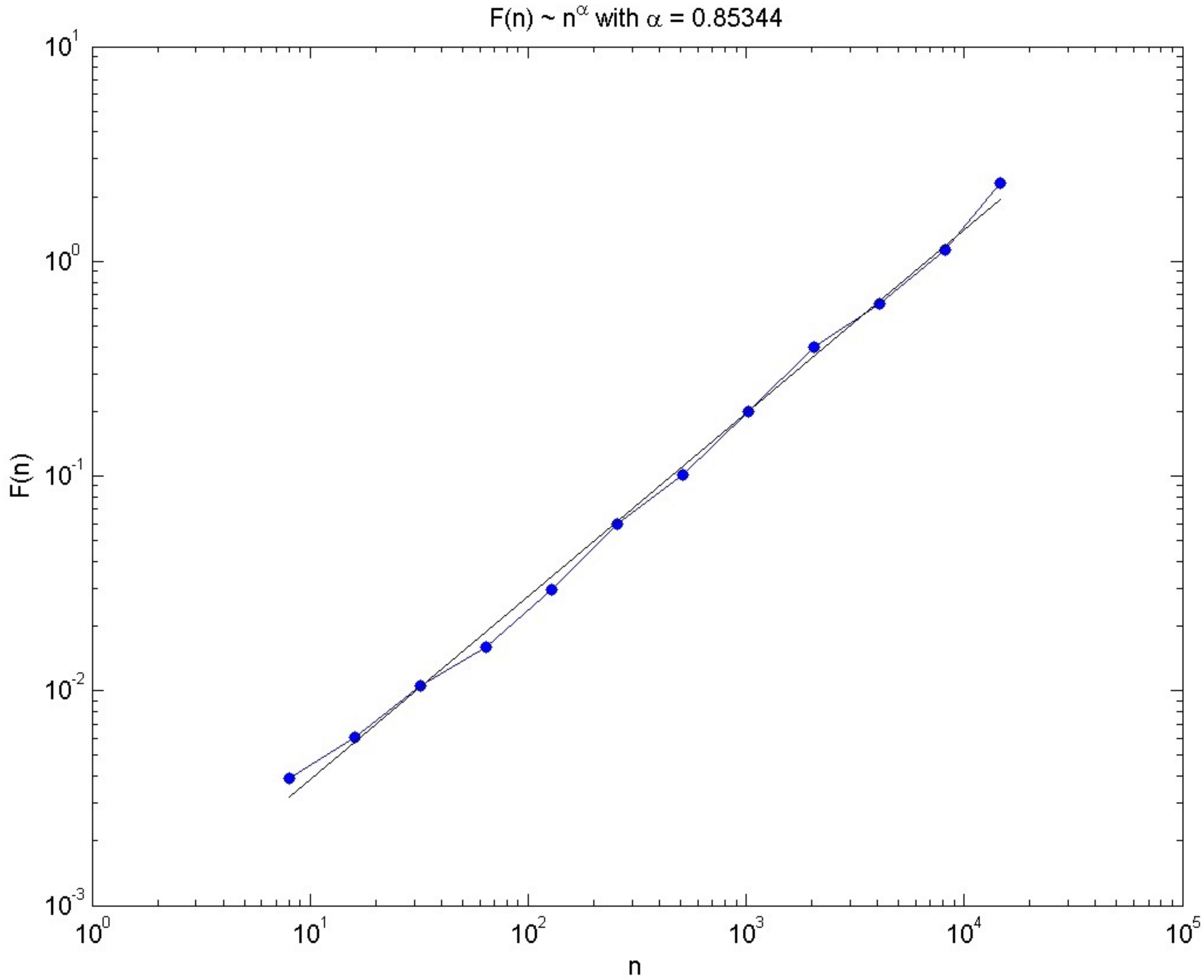
$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y(k) - y_n(k)\right)^2}$$

- We then calculate the average fluctuation F(n) for each segment n.
- If the scaling is approximately given by $F(n) = c\, n^{\alpha}$, we can estimate α by calculating the slope of log F(n) versus log n.
- A linear relationship on a log-log plot indicates the presence of power law (fractal) scaling and β=2α-1where β is the slope of the power spectrum (β =0,1,2 for white noise, pink noise and a random walk).

# DFA of S&P500 returns



# DFA of S&P500 absolute returns



# Time series

- Sampling frequency;
- Trends;
- Mean reversion;
- Simple moving averages
- Exponential moving averages
- Persistence and DFA

# Matlab functions

- vratiotest
- dfa,
- tsmovavg
- movavg

# Q&A

Data & Inference

# WEEK 5B

# Today's Lecture

| No. | Activity | Description | Time |
|---|---|---|---|
| 1 | Challenge | Real-time decision-making | 10 |
| 2 | Discussion | Difficulties of testing and going live | 10 |
| 3 | Case study | Quantopian, Collective2 | 10 |
| 4 | Analysis | Autoregression and moving averages | 20 |
| 5 | Demo | Applications for human activity | 20 |
| 6 | Q&A | Questions and feedback | 10 |

# Assignment 3

| Number | Description | |
|---|---|---|
| 1 | Daily energy intake | Statistical significance: evidence of deviation from recommendation? |
| 2 | Guinness enjoyment | Statistical significance: comparison of Ireland and other countries |
| 3 | World Bank indicators for fertility and GDP per capita | Scatter plot: several years and many countries.  Color-coded for countries |
| 4 | Monthly house price index | Calculate monthly returns, annualized return and study autocorrelation |
| 5 | FTSE100 index daily prices | Daily returns, annualized return. Compare houses and stockmarket performance. |

# Quiz

- Which of the following games would you play?
  - ➢ Game A gives a 95% probability to win \$10 but a 5% probability of losing \$1,000.
  - ➢ Game B gives a 5% probability to win \$1,000 but 95% probability of losing \$10.
- **slido.com #21740**

# Expectancy

- Expected profit holds the key to making the correct decision:
- $E = Prob_{win} * Amount_{win} + Prob_{loss} * Amount_{loss}$
- Let's insert the numbers:
  - Game A: E = 0.95*$10 – 0.05*$1000 = -$40.5
  - Game B: E = 0.05*$1000 – 0.95*$10 = +$40.5
- Suppose we play 20 games, then 19 of them equate with the 0.95 probability and 1 of them with the 0.05 probability.
  - Game A: Profit = 19*$10 – 1*$1000 = -$810
  - Game B: Profit = 1*$1000 – 19*$10 = +$810
- Definitely makes sense to play game A!

# Automated decision-making

- Decision-making takes up time and valuable resources.
- Automated decision-making is an attractive idea as it can potentially avoid mundane tasks and save time.
- But what are the steps required before handing over so much power to a computer?
- What evidence would you need to see before trusting a machine?

# The Machine

- The machine relies on an algorithm to process data and suggest actions.
- This algorithm is based on a model.
- A model is simply a caricature of reality.
- The model attempts to describe the evolution of a real-world system.
- Such a model will inevitably make mistakes due to uncertainty in the observations, parameters and structure.
- When and how can we justify using a model for decision-making?

# Human experts versus machines

- Traditionally important decisions have been made by human experts.
- Predictive analytics due to availability of data, computational resources and machine learning/statistical techniques.
- Accounting for uncertainty is key for quantifying the confidence underlying the decision-making process.

# Intelligent decision-support

- Predictive analytics help organisations understand trends and patterns in human activity
- Facilitating the identification of relationships between explanatory variables and key performance indicators (sales, profit, risk reduction, efficiency)
- Having impacts on finance, insurance, energy, healthcare, government,…

# More data or better models?

- In many disciplines, the focus is often on obtaining new predictive variables.
- In medicine, traditional model structure is a logistic regression.
- A powerful nonlinear classifier is consistently superior to logistic regression, offering a relative improvement in performance of 36% in predicting the outcome across six different medical datasets.

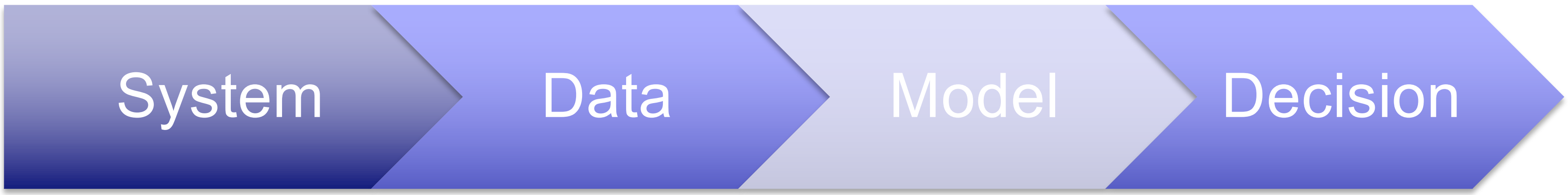
Tsanas, Little, McSharry (2013). In “Handbook of systems and complexity in health”, Springer. New York.

# Human experts


Decision-making by human experts is often referred to as discretionary or judgmental.

Different humans are affected by psychology, emotions and a host of other factors that make such decision-making hard to replicate, explain or justify after the fact.

# Automated Decision-making

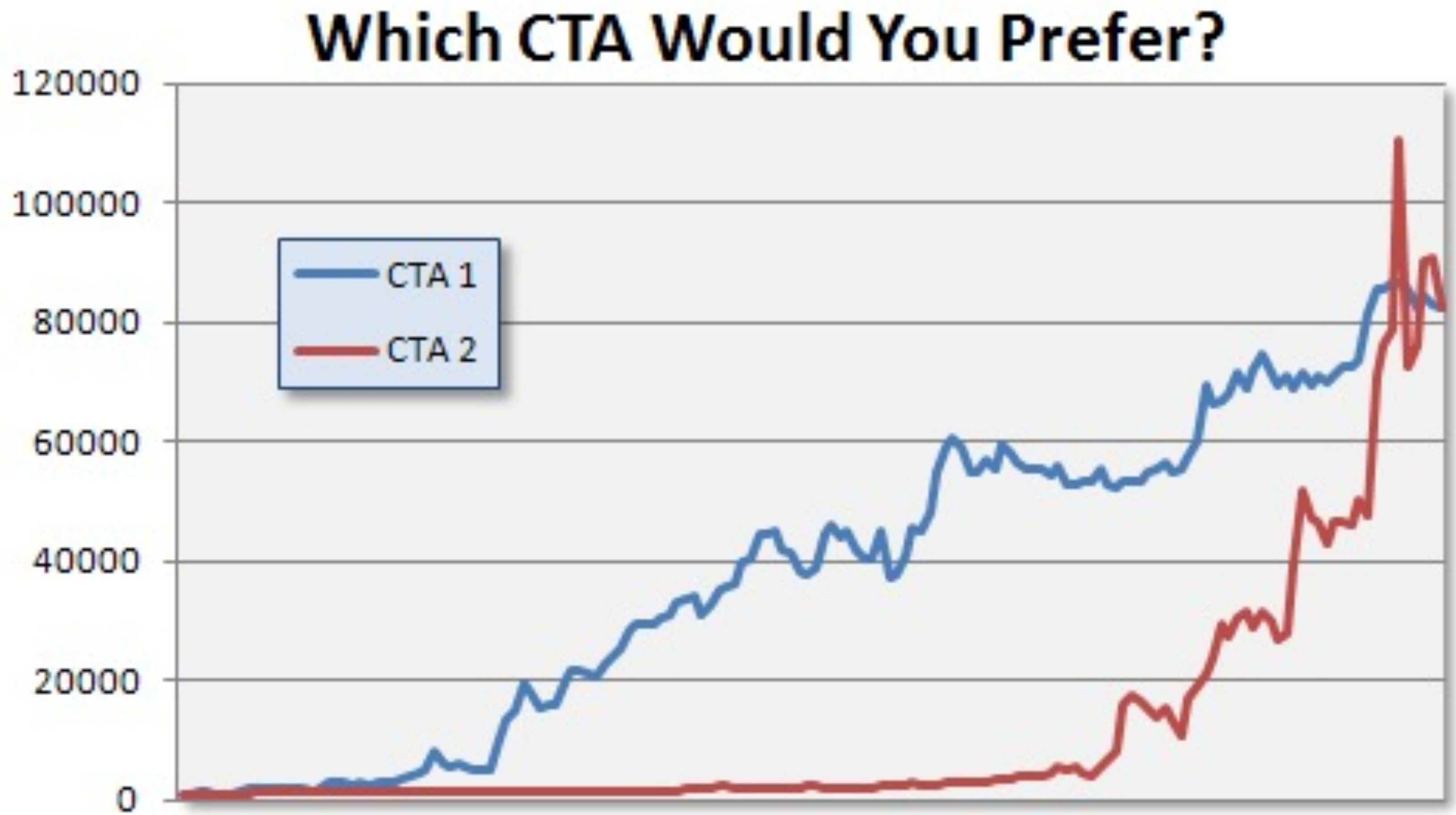



Decision-making by machines is often referred to as automated, systematic or algorithmic.

Machines are affected by psychology and emotions and the transparency of the process makes it easy to replicate, explain or justify after the fact.

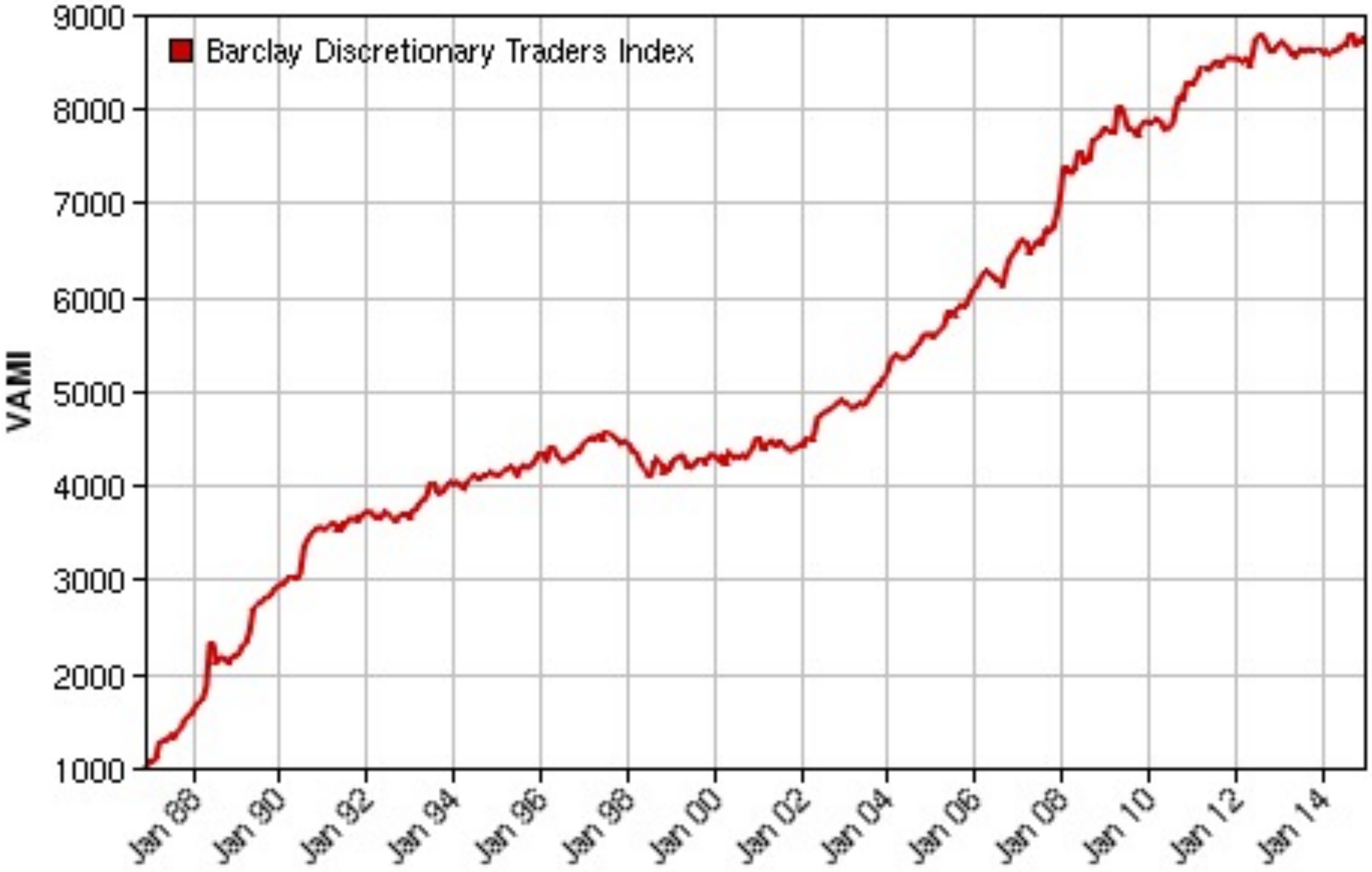
However the process between the system and decision is subject to much uncertainty!

# Metrics can help us decide




Class Poll 1
slido.com
**#21740**

# Discretionary Performance



# Discretionary performance

| Metric | Value |
|---|---|
| Number of programs | 137 |
| Compound Annual Return | 8.05% |
| Sharpe Ratio | 0.60 |
| Worst Drawdown | 10.67% |
| Correlation with S&P500 | 0.01 |
| Correlation with US Bonds | 0.17 |
| Correlation with World Bonds | 0.00 |

http://www.barclayhedge.com

# Systematic Performance

# Systematic performance

| Metric | Value |
|---|---|
| Number of programs | 457 |
| Compound Annual Return | 8.23% |
| Sharpe Ratio | 0.37 |
| Worst Drawdown | 22.07% |
| Correlation with S&P500 | -0.04 |
| Correlation with US Bonds | 0.09 |
| Correlation with World Bonds | -0.04 |

http://www.barclayhedge.com

# Modeling Challenges

- 1. Obtaining access to sufficient amounts of high quality historical data to construct a model.
- 2. Data is required for both the explanatory variables and target (e.g. key performance indicators).
- 3. Evaluating the performance of a model using statistical procedures.
- 4. Ensuring the the performance metrics are likely to be applicable in the future.

# Implementation Challenges

- Real-time decision-making requires access to live data.
- The time stamps on the data must be carefully validated.
- Model must be able to run fast enough.
- A delay between model output and decision could reduce the benefits.

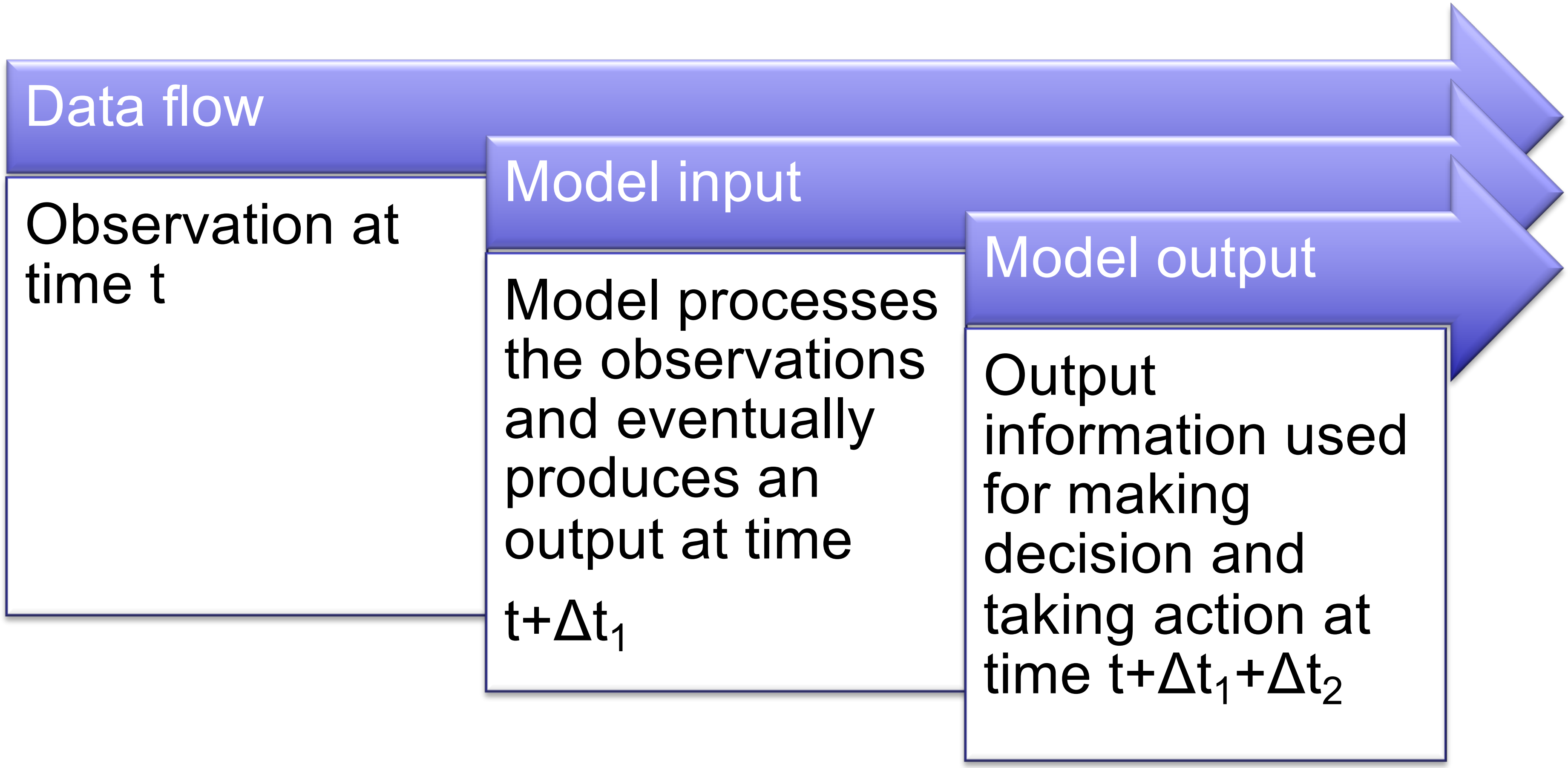
# Evaluation – avoid over-fitting

- The first step is to ensure that the historical performance of the model can actually be achieved in practice.
- This is best done by using a walk-forward approach and making sure that the model is constructed using data in the time window [t-Ttrain, t] and evaluated on data in the time window [t, t+Ttest].
- This procedure is known as cross-validation and avoids over-fitting issues.

# Paper testing - avoid data snooping

- However it is still possible to make mistakes and cheat by accident.
- For example by peaking at data at time t+1 and using it to build a model at time t.
- In order to avoid any doubt, it is best to test the model in real-time and collect statistics day by day.
- This process of ensuring that the model can only access previous information is sometimes called paper testing.

# Time sequence schedule



# Independent Evaluation

- Because of the many difficulties just highlighted around evaluation, it is often best to use an independent source to demonstrate credibility.
- This approach deals with the issues around data quality and time sequencing.
- By processing in real-time, the model can only make decisions and suggest actions in the correct time sequence.

# Trading Algorithms

- What percentage of trades in US equities is performed by algorithms?

Class Poll 2
slido.com
**#21740**

# Social networking for investors

- Covestor.com
  – We make it easy to select money managers to help you meet your investment goals.
- Marketocracy.com
  – Invest with skill, not luck.
- TenStocks.com
  – Capital at work.

# Social networking for developers

- Collective2.com
  - Automate your trading. Trade stocks, options, futures, or forex. No human required.
- Quantopian.com
  - Your ideas, uour code, your trades, your algorithmic investing platform.
- QuantConnect.com
  - Design and trade algorithmic trading.

# Case Study Collective2

All performance results are hypothetical. Trading is risky and you can lose money. Learn more

## Leaderboard

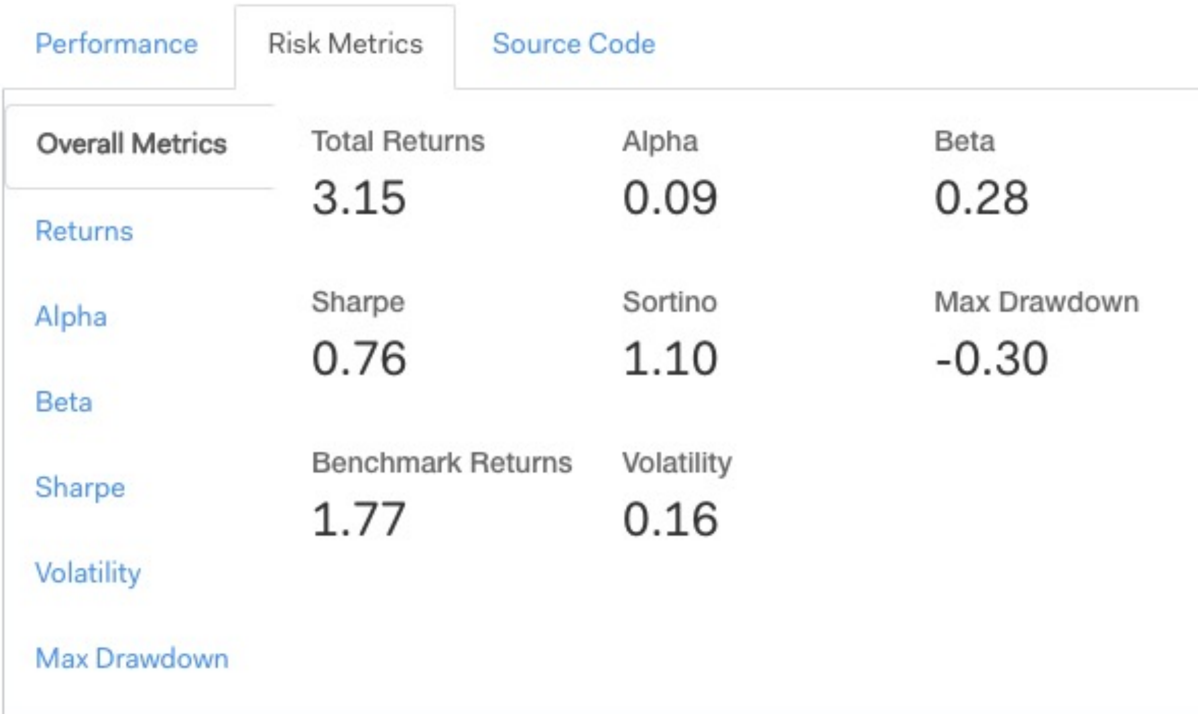
Currently showing Popular (All instruments) trading All | Stocks | Options | Futures | Forex

| STRATEGY NAME | TRADES | SUB FEE | All time | SUG MIN CAP | % Profitable | MAX DD | RETURN | ACTION |
|---|---|---|---|---|---|---|---|---|
| FEATURED STRATEGY<br>EF Futures<br>Manager: EmanueleFrisa | FUTURES | $99 /month | +50.3% +$13,500 | $25,000 | 51.65% | (20.9%) | +50.3% Annual Return since Nov 21, 2016 | + WATCH SIM LIVE |
| US Bonds<br>Manager: CalvRoss | FUTURES | $147 /month | +162.1% +$82,000 | $100,000 | 76.74% | (3.4%) | +162.1% Cumul. Return since Mar 22, 2018 | + WATCH SIM LIVE |
| Spy Strategy<br>Manager: CFOSolutions | STOCKS | $100 /month | +8.8% +$27,500 | $15,000 | 55.00% | (4.9%) | +8.8% Annual Return since Feb 26, 2016 | + WATCH SIM LIVE |
| Dutch Volatrader<br>Manager: DutchVolatrader | STOCKS | $100 /month | +27.3% +$9,250 | $15,000 | 81.54% | (5.3%) | +27.3% Annual Return since Aug 11, 2017 | + WATCH SIM LIVE |
| Volatility Invest<br>Manager: AlexanderG | STOCKS | $50 /month | +11.2% +$9,500 | $15,000 | 56.00% | (6.2%) | +11.2% Annual Return since Apr 01, 2017 | + WATCH SIM LIVE |
| SpxLearning<br>Manager: PlusHellZone | STOCKS | $99 /month | +15.1% +$8,350 | $15,000 | 100.00% | (7.2%) | +15.1% Cumul. Return since Feb 06, 2018 | + WATCH SIM LIVE |
| Longevity<br>Manager: SustainablePace | FOREX | $15 /month | +16.8% +$970 | $10,000 | 64.47% | (7.8%) | +16.8% Cumul. Return since Jan 17, 2018 | + WATCH SIM LIVE |
| Low Vol AllCap Equity<br>Manager: NTLLC-GDSingh | STOCKS | $29 /month | +15.3% +$11,700 | $15,000 | 64.50% | (8.0%) | +15.3% Annual Return since Mar 20, 2016 | + WATCH SIM LIVE |
| WXYZ<br>Manager: Teffub | FUTURES | $125 /month | +24.4% +$19,400 | $90,000 | 72.73% | (8.2%) | +24.4% Cumul. Return since Jan 10, 2018 | + WATCH SIM LIVE |
| Trend Factor<br>Manager: Flaneur | FUTURES | $199 /month | +17.6% +$15,200 | $60,000 | 66.29% | (8.6%) | +17.6% Annual Return since Jun 02, 2017 | + WATCH SIM LIVE |

https://collective2.com

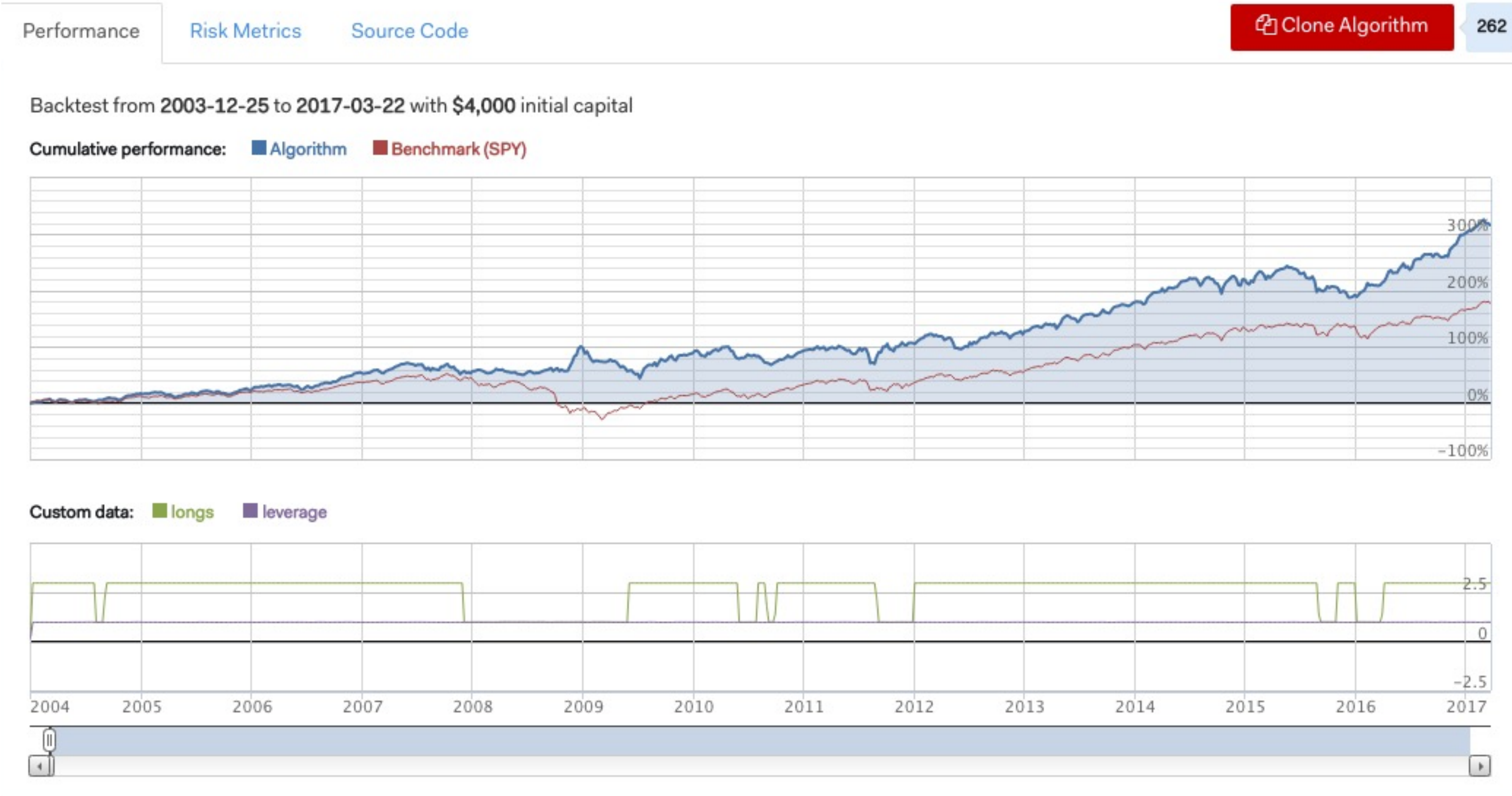
# Case Study - Quantopian

- Social finance websites
- https://www.quantopian.com
- Risk Metrics for EFT-Rotation-Strategy:




https://www.quantopian.com/posts/etf-rotation-strategy

# Quantopian - Performance



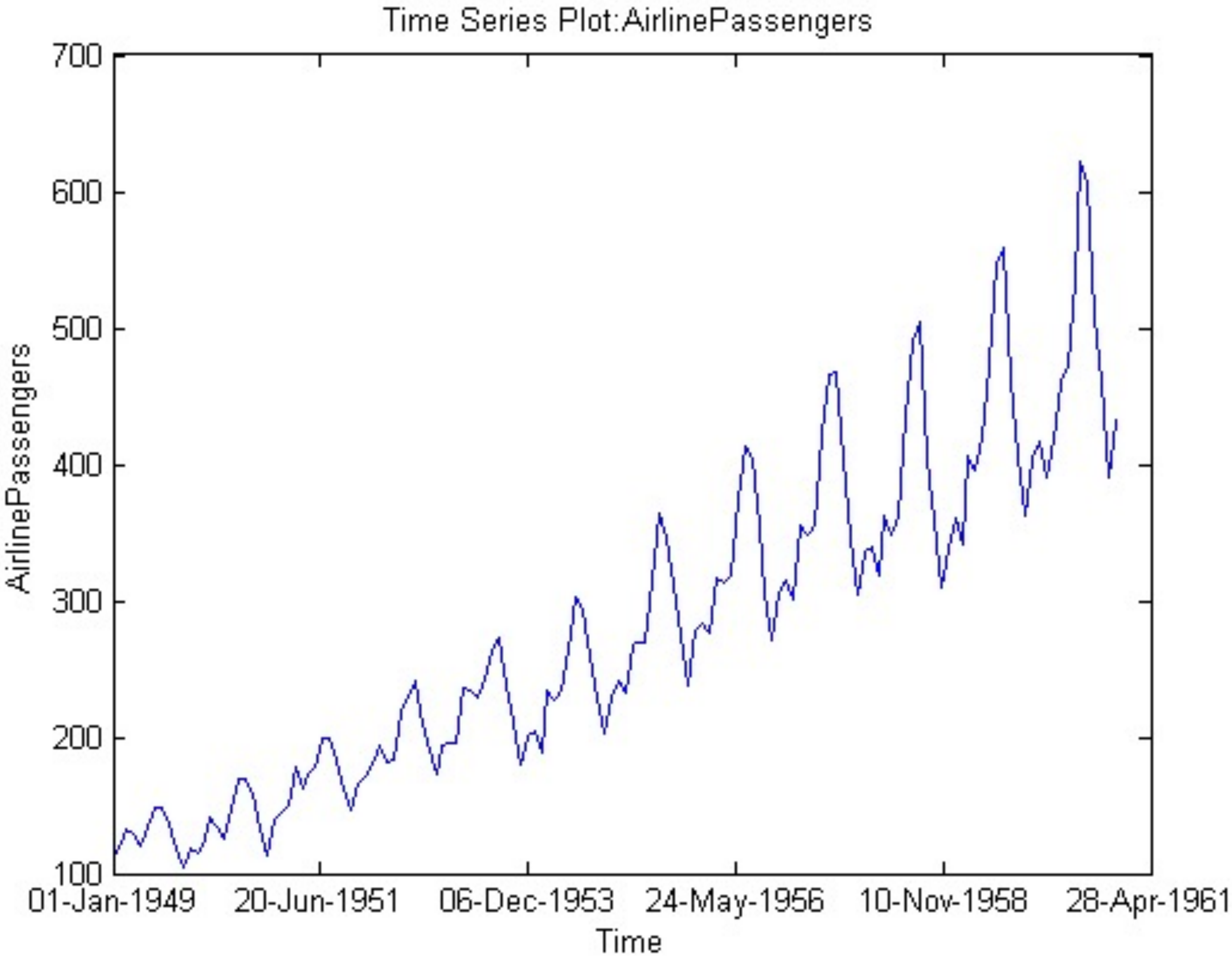

https://www.quantopian.com/posts/etf-rotation-strategy

# Hege Fund Rewards

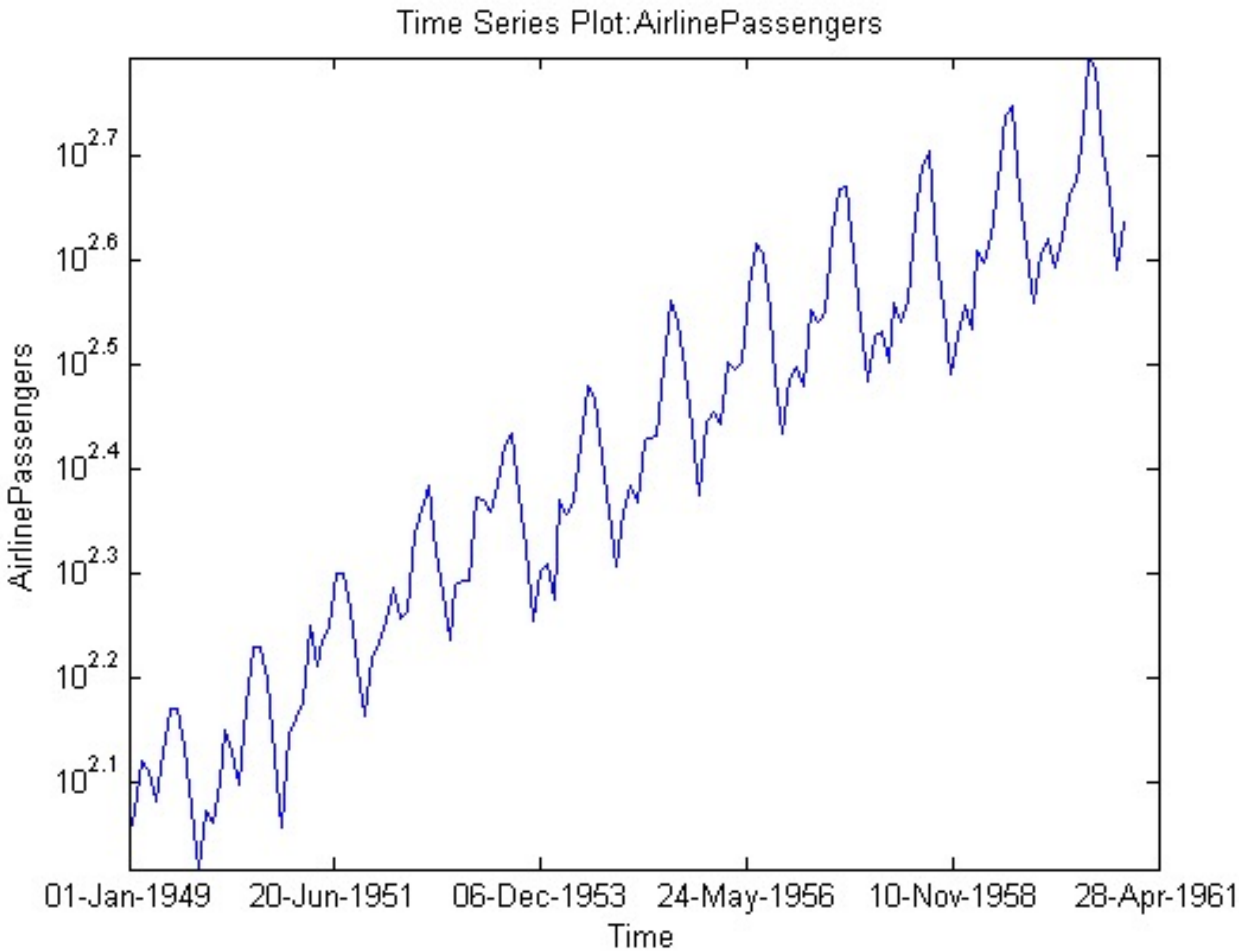
- Annual income in 2018 for quantitative hedge fund owner (James Simons, Renaissance Technologies) in US$:
- 200 million
- 400 million
- 800 million
- 1,600 million

Class Poll 3
slido.com
**#21740**

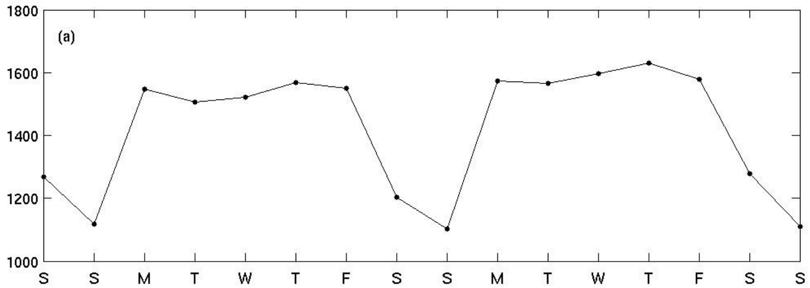
# Airline Passengers

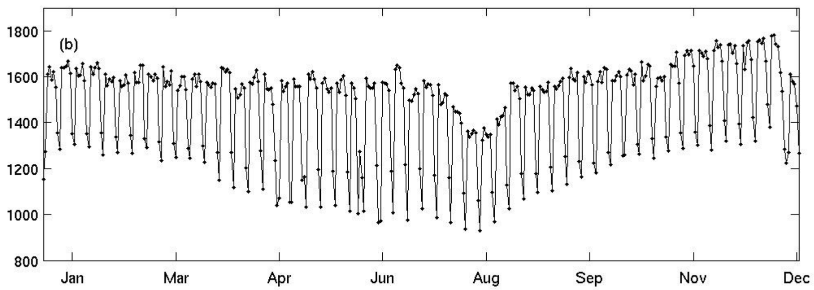


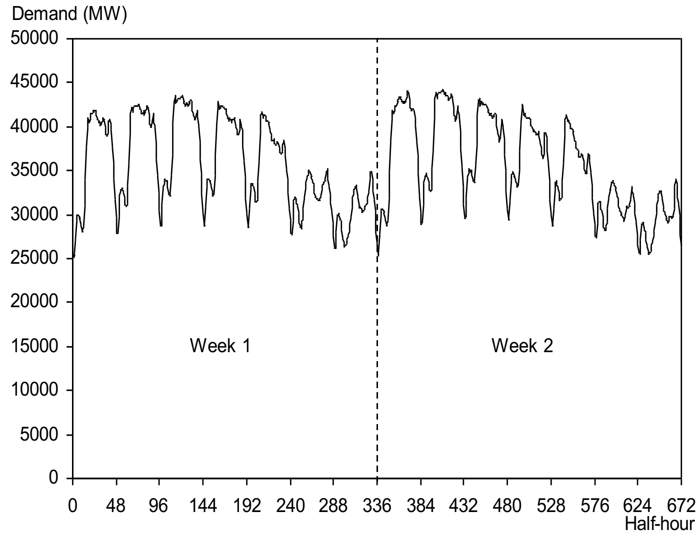
# Airline Passengers – log scale

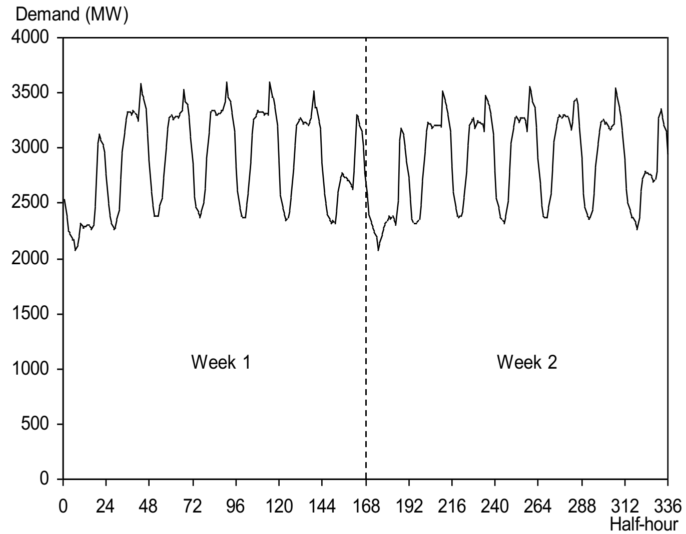


# Collective human behaviour: Electricity demand



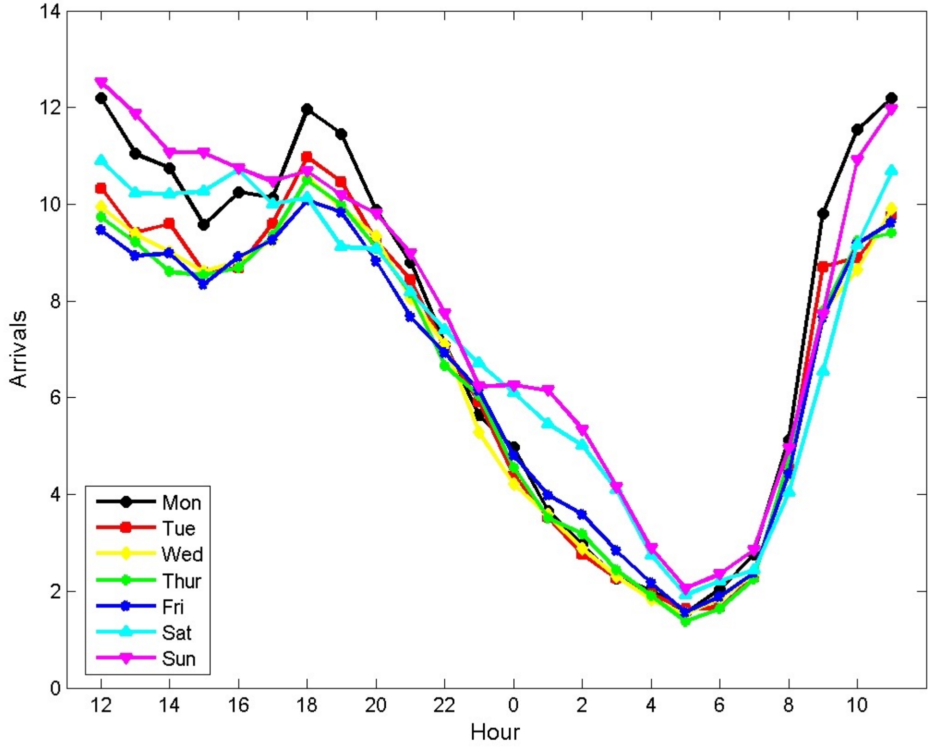


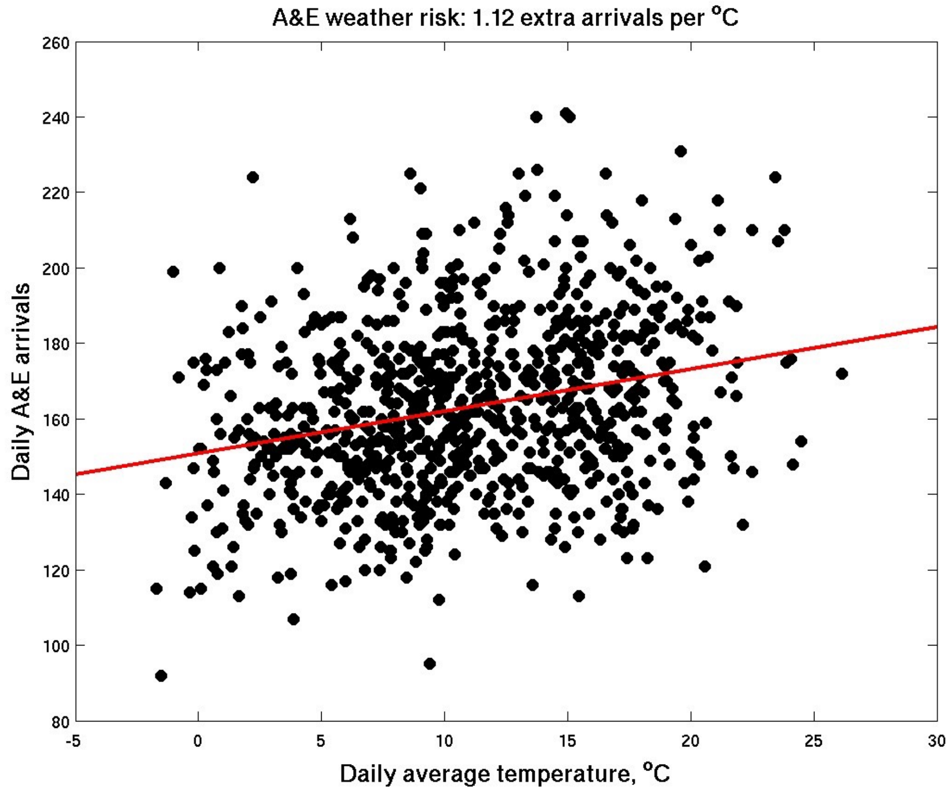
McSharry, PE, Bouwman, S and Bloemhof, G. (2005). Probabilistic forecasts of the timing and magnitude of peak electricity demand. IEEE Transactions on Power Systems, 20(2): 1166-1172

Taylor, JT, de Menezes, LM and McSharry, PE (2005). A comparison of methods for forecasting electricity demand up to a day ahead. International Journal of Forecasting, 22(1): 1-16

Taylor, JT and McSharry, PE (2007). Short-Term Load Forecasting Methods: An Evaluation Based on European Data. IEEE Transactions on Power Systems, 22(4): 2213 - 2219

# A&E arrivals




A&E arrival rate at JR hospital depends on the day of the week, time of day, social events and the weather.

# Autocorrelation

- The autocorrelation of a random variable X is

$$\rho(t,s) = \frac{E\left[(X_t - \mu)(X_s - \mu)\right]}{\sigma^2}$$

- If X is second-order stationary, then with $k=|t-s|$, we have

$$\rho(k) = \frac{E\left[(X_t - \mu)(X_{t+k} - \mu)\right]}{\sigma^2}$$

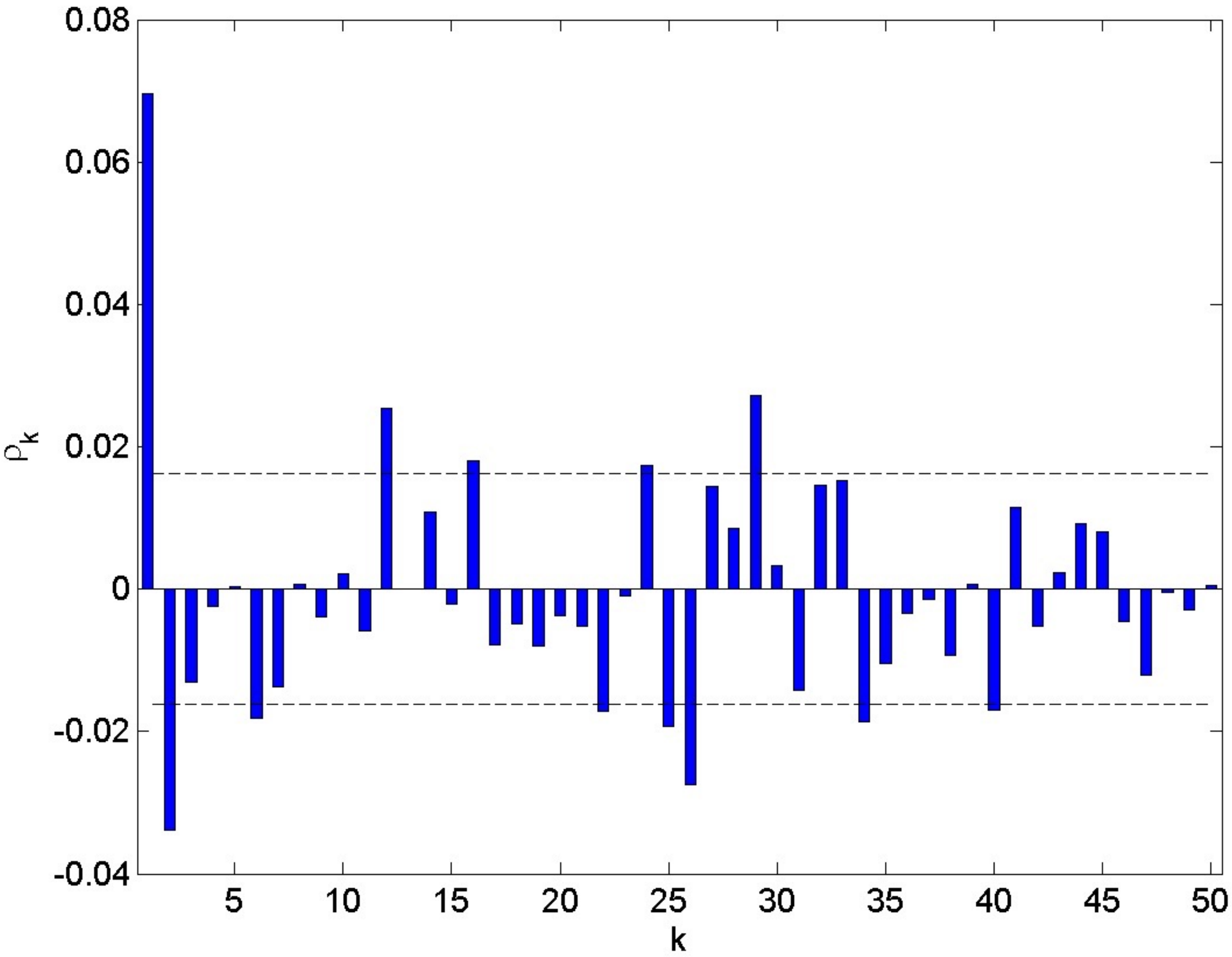
- For time series samples,

$$\rho_k = \frac{\frac{1}{n-k}\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sigma^2}$$

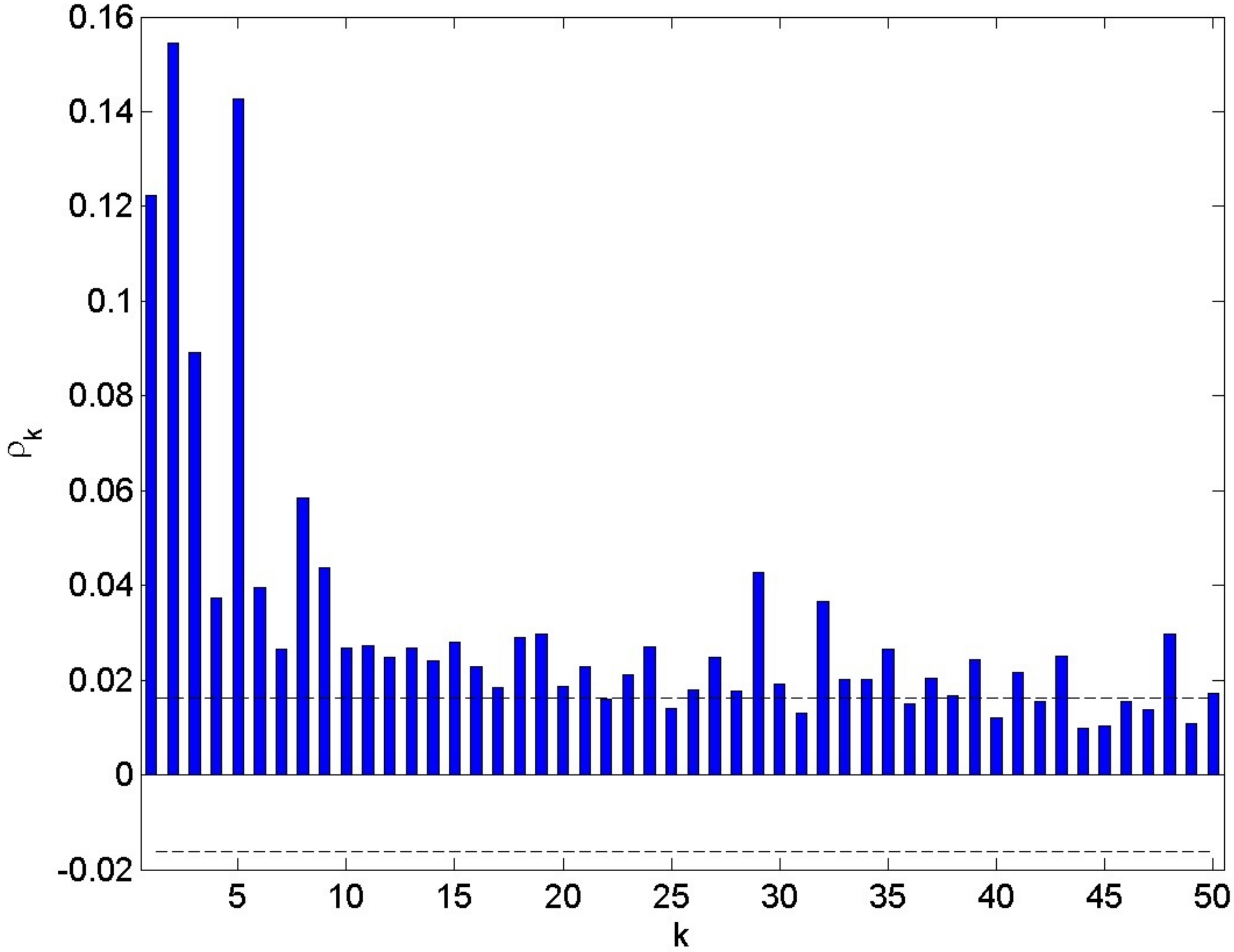
# Autocorrelation

- As the autocorrelation represents the cross-correlation of a variable with itself, it quantifies the nature of the relationship between observations in a time series as a function of the time lag between them
- The autocorrelation can be used to identify periodic signals (such as weekday effects) which may be masked by noise
- Significant values can be identified using the 95% confidence intervals (corresponding to a normal distribution) at $\pm 1.96/\sqrt{n}$ where *n* samples are employed
- Note that the correlation coefficient only detects linear relationships

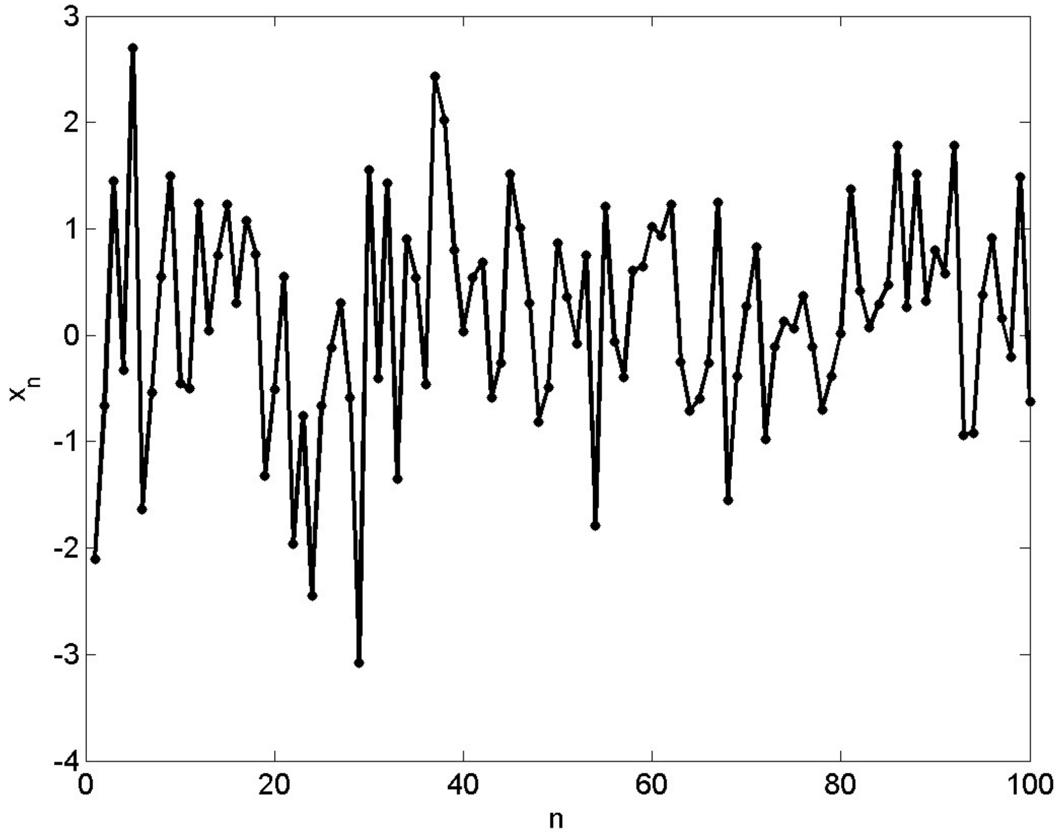
# ACF of returns



# ACF of squared returns

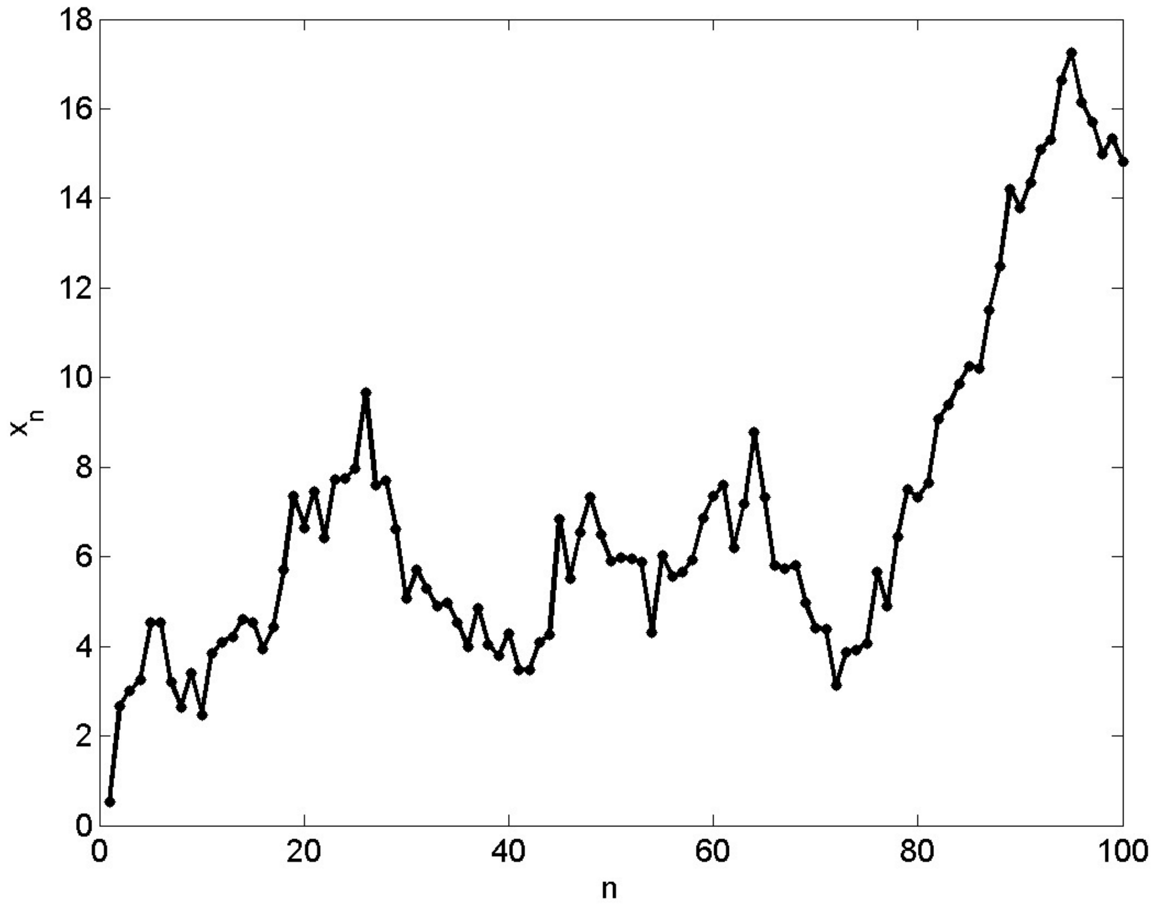


# White noise




- $x_n = \varepsilon_n$, $\varepsilon_n \sim N(0,\sigma^2)$
- Power spectral density is flat; components at all frequencies are equally likely

# Random walk




- $x_n = x_{n-1}+\varepsilon_n$, $\varepsilon_n \sim N(0,\sigma^2)$

# Moving average MA(q)

- MA(q) is given by

$$x_n = \varepsilon_n + \sum_{i=1}^{q} \theta_i \varepsilon_{n-i}, \quad \varepsilon_n \sim N(0,\sigma^2)$$

$$= \Theta(\mathrm{L})\varepsilon_{\mathrm{n}}, \qquad \Theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$$

- This models the current observation as a function of current and lagged unobservable shocks
- The memory is increased by larger q

# Autoregressive moving average ARMA(p,q)

- ARMA(p,q) is given by

$$x_n = \sum_{i=1}^{p} \phi_i x_{n-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{n-i} + \varepsilon_n, \quad \varepsilon_n \sim N(0,\sigma^2)$$

$$= \frac{\Theta(\mathrm{L})}{\Phi(\mathrm{L})} \varepsilon_{\mathrm{n}},$$

- This model has the increased flexibility of both the AR and MA specifications

# AR(1)

- AR(1):

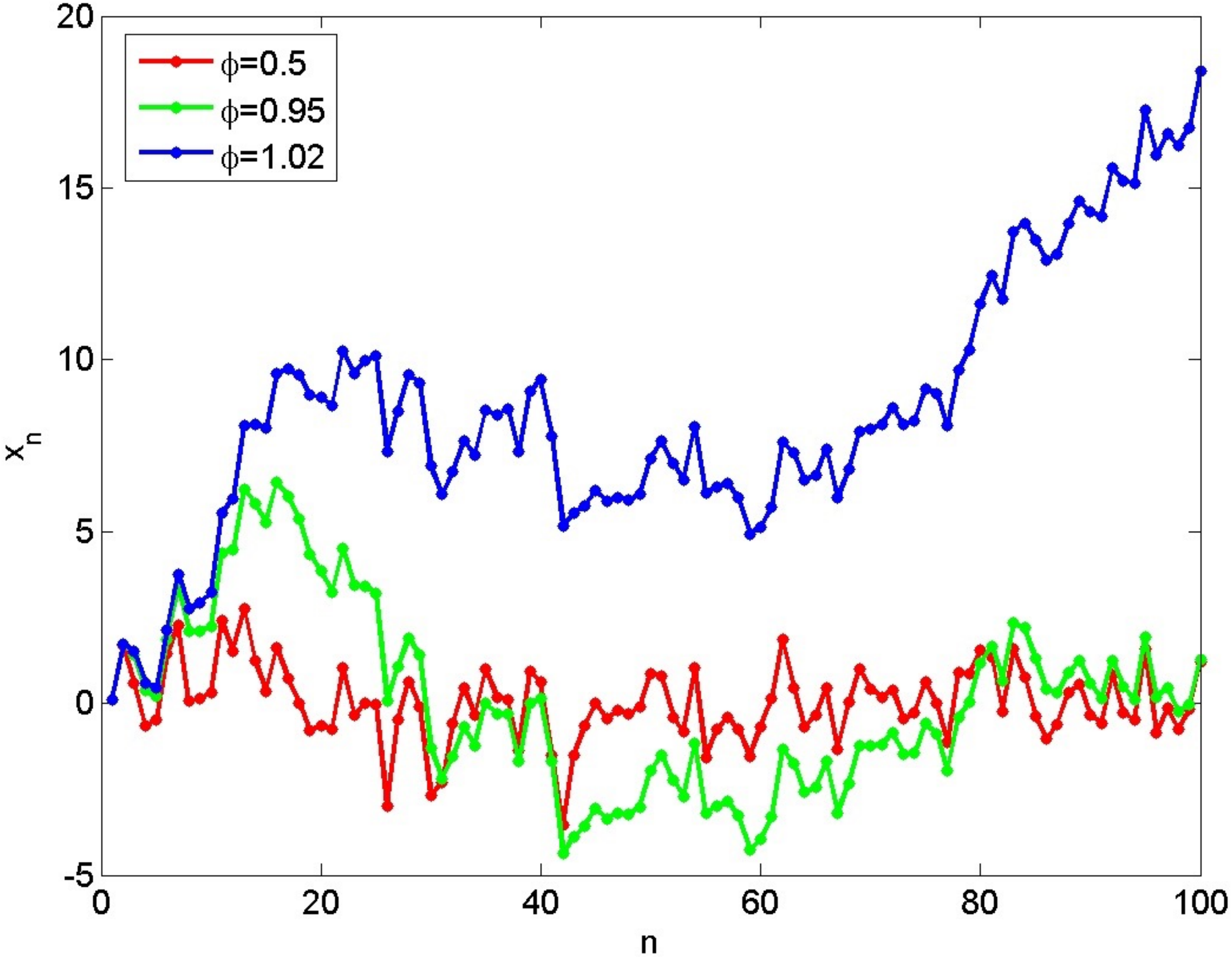
$$x_n = \phi x_{n-1} + \varepsilon_n$$

- Autocorrelation:

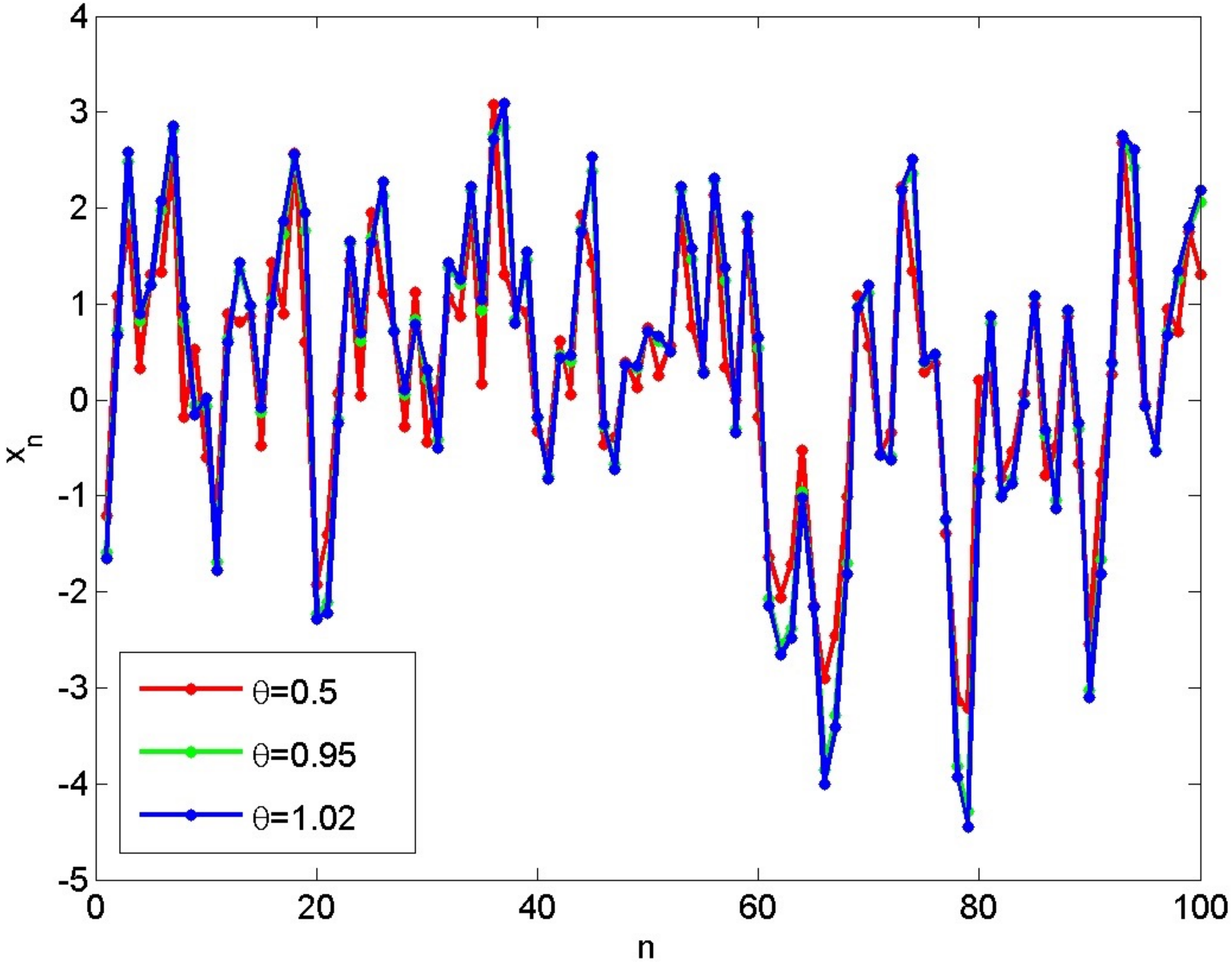
$$\rho(\tau) = \phi^{\tau}$$

- As ρ=Φ, we can understand the relevance of the link between an exploratory visualisation analysis and time series modelling
- AR(1) is a discrete-time version of an Orstein-Uhlenbeck stochastic process

# AR(1) examples



# MA(1) examples



# Time series analysis

- Autocorrelation;
- Human activity; Seasonality;
- Time of day, month, year
- Visualizing seasonality
- Representing seasonality
- Testing for seasonality

# Matlab functions

- acf
- arima
- month, weekday,
- toy

# Q&A