ETH zürich

Social Data Science

Lecture 6: Trends

Dr. David Garcia

So Far

- Foundations of Social Data Science
  - Social Data Science is question-driven: theory, methods, and data
- Social Science theories
  - The friendship paradox and other sampling biases
  - Culture dynamics in fashion and growth processes of collective aggregation
  - Social Impact Theory and its relationship to social media
- Methods in Social Data Science
  - Correlation and bootstrapping
  - Distribution fitting and introduction to regression
  - Data management and manipulation in R
- Digital trace datasets
  - Google trends and World Bank data
  - TwitteR: user profiles, timelines, and search

Notes:

- Course homepage: https://www.sg.ethz.ch/teaching/sds/
- Course Moodle: https://moodle-app2.let.ethz.ch/course/view.php?id=2985
- Learning Goals of this lecture:
  - To identify patterns of social trends in collective aggregates
  - To understand introductory principles of time series modeling and analysis
  - To be able to access bitcoin data and other JSON APIs

Notes:

Outline

Social Trends

Social Trend

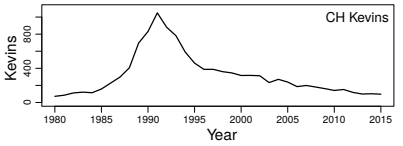
The temporal tendency or direction of social behavior.
The change in size over time of a collective aggregate.

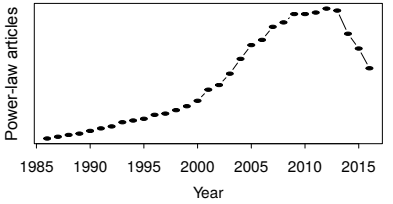
Examples of social trends:

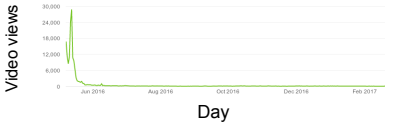
- Daily views for a Youtube Video
- Yearly amount of babies with a name
- Monthly unemployment or crime rate in a city

Social trends can reveal:

1. How a group reacts to external events
2. The strength of social influence between people
3. Which behavior anticipates another







Notes:

Notes:

- In Lecture 04 we focused on the **final size** of collective aggregates
- In this lecture, we study how the aggregates **change over time**
- Examples of social trends from previous lectures:
  1. Yearly amount of babies named Kevin (Lecture 03)
  2. Yearly amount of papers with power-laws (Lecture 04)
  3. Daily amount of views of a Youtube video tweeted by Justin Biever (Lecture 05)

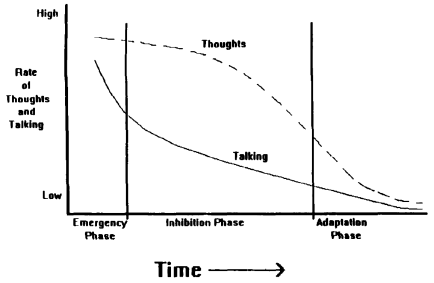
Collective reactions to external events

Social stage model of collective coping (Pennebaker & Harber)

- A group of people reacts to an external event (first applied to emergencies)
- Trend of talking and thinking about the event
- Applications to emotion research (Lectures 07 and 09)

1. Emergency phase:
   High levels of thinking and talking
2. Inhibition/satiation phase:
   Fast decrease in talking, slower in thinking
3. Adaptation phase:
   Both talking and thinking go back to baseline

- **Applications beyond emergencies:**
  Fast vs slow relaxation of collective reactions



Exogenous and endogenous and bursts

- Types of social trends towards the peak of a collective aggregate of a community
- Examples with Google trends volume
- **Exogenous burst:** Spike created by an **external event**
  - Unexpected terrorist attack creates a very fast increase, the event is *external to the community*
- **Endogenous burst:** Peak driven by **social influence in the community**
  - Anticipation for a movie and word of mouth *within a community* creates a slower increase

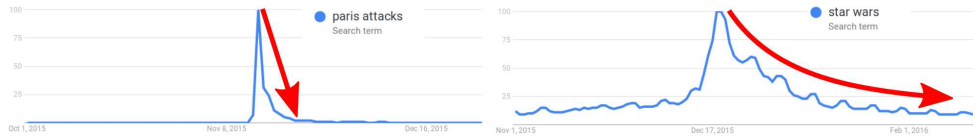
Notes:

- A Social Stage Model of Collective Coping: The Loma Prieta Earthquake and The Persian Gulf War. J. Pennebaker and K. Harber. Journal of Social Issues (1993)
  http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4560.1993.tb01184.x/pdf

Notes:

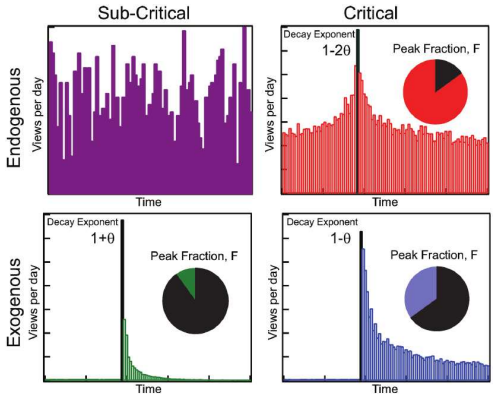
- Left: Google search volume in the US for 'Paris attacks'
- Right: Google search volume in the US for 'Star Wars'

Subcritical and critical dynamics



- Types of social trends after the peak of a collective aggregate of a community
- Examples with Google trends volume
- **Subcritical dynamics:** Fast relaxation due to **social influence weaker than novelty decay**
  - Interest in the US about the attacks was limited, the peak relaxed fast
- **Critical dynamics:** Slow relaxation due to **social influence stronger than novelty decay**
  - Hype about Star Wars movie kept people talking and searching about it

Modeling collective responses



Model of social trends combining two ingredients:

1. **Burst exogeneity**
   - **Endogenous:** peak generated by internal dynamics
   - **Exogenous:** peak triggered by an external event
2. **Criticality of social influence**
   - **Subcritical:** social influence absent or weaker than novelty decay
   - **Critical:** social influence stronger than novelty decay

Source: Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS (2008)

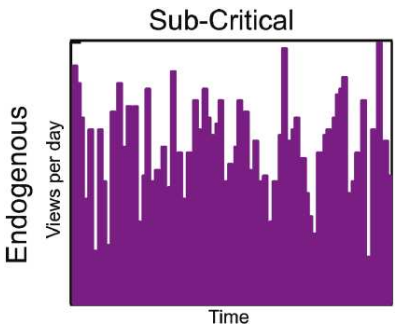
Notes:

- Left: Google search volume in the US for 'Paris attacks'
- Right: Google search volume in the US for 'Star Wars'

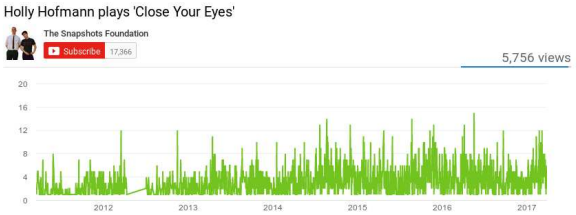
Notes:

- Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS 2008.
  http://www.pnas.org/content/105/41/15649.abstract
- The Decay exponents ($1-\Theta$, $1+\Theta$, $1-2\Theta$) are explained in detail in the above reference

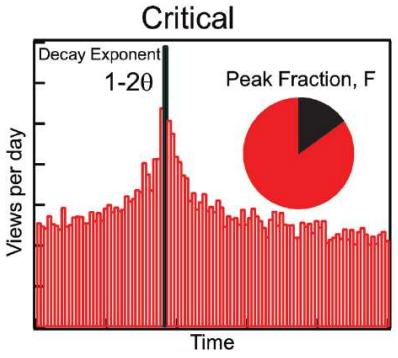
Endogenous subcritical



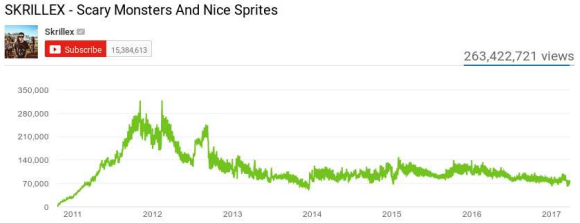
- Endogenous: absence of strong spike
- Subcritical: social influence absent or weaker than novelty decay
- Trend: noisy signal of incoherent activity
  - No identifiable peak
- Example: most of content in social media



Endogenous critical



- Endogenous: peak present without spikes
- Critical: social influence stronger than novelty decay
- Trend: large wave with slow increase and decrease
  - Less than 20% total volume at the peak
- Example: content that becomes popular in social media



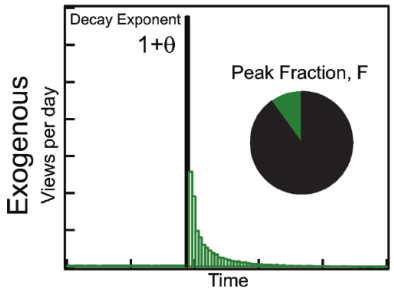
Notes:

- Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS 2008.
  http://www.pnas.org/content/105/41/15649.abstract

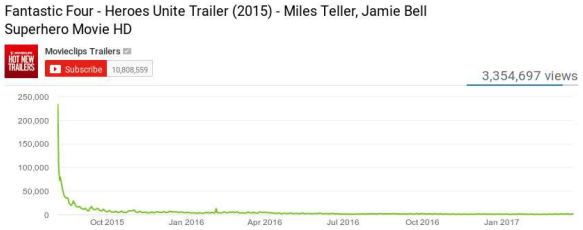
Notes:

- Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS 2008.
  http://www.pnas.org/content/105/41/15649.abstract

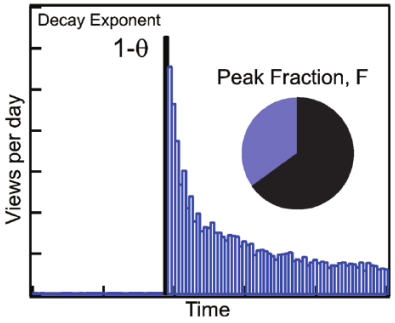
Exogenous subcritical



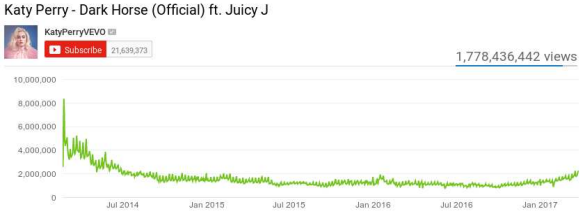
- Exogenous: peak as spike produced by an external event
- Subcritical: Social influence weaker than novelty decay
- Trend: sudden growth and decrease
  - More than 80% total volume at the peak
- Example: junk content featured in mass media



Exogenous critical



- Exogenous: Spike generated by external factor
- Critical: Social influence stronger than novelty decay
- Trend: Spike growth with slow decay
  - Between 20% and 80% total volume at the peak
- Example: Central channels creating popular content



Notes:

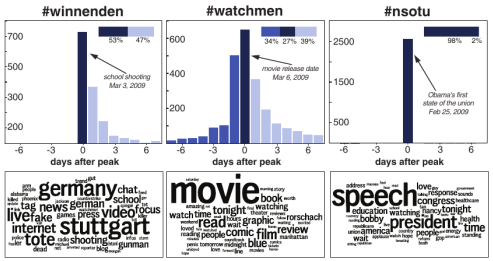
- Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS 2008.
  http://www.pnas.org/content/105/41/15649.abstract

Notes:

- Robust dynamic classes revealed by measuring the response function of a social system. R. Crane, D. Sornette. PNAS 2008.
  http://www.pnas.org/content/105/41/15649.abstract

Classification of Twitter hashtag trends

- The daily amount of tweets with a hashtag follow social trends
- The text of tweets gives us an idea of which contents follow which trends



- #winnenden: School shooting (Exogenous critical)
- #watchmen: Movie release date (Endogenous critical)
- #nsotu: Political speech (Exogenous subcritical)
- Ignored hashtags: Endogenous subcritical

Computational Finance

Notes:

- Dynamical classes of collective attention in twitter. J. Lehmann, B. Gonçalves, J. J. Ramasco, C. Cattuto. WWW '12 2012.
  http://dl.acm.org/citation.cfm?id=2187871

Notes:

Computational Finance

review articles

Computational Finance

Application of Computer Science techniques to practical problems in finance. Also known as algorithmic trading

Fundamentals of computational finance:

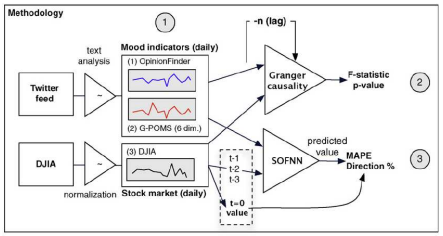
- **Identification of which behavior anticipates another**
- Learning from historical data to make financial predictions
- Computerized traders that can take risk into account

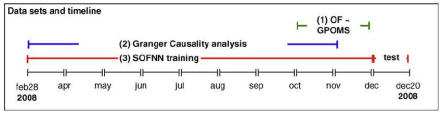
Social media in computational finance:

- Focus on social media signals as predictors of prices
- Automatic traders using large-scale data from social media

Twitter mood and the stock market

- Predicting the Dow Jones Industrial Average (DJIA) using mood on daily tweets
- Popular application of social media in computational finance
- *"We find an accuracy of 86.7% in predicting the daily up and down changes"*
- Issues:
  1. Missing stationarity test in time series analysis
  2. Black-box predictions hard to interpret
  3. Short test period, missing profit measurement





- We will learn more about research on mood in Lectures 07-09
- In this lecture, we will learn a formal method to test if a signal **leads** price changes

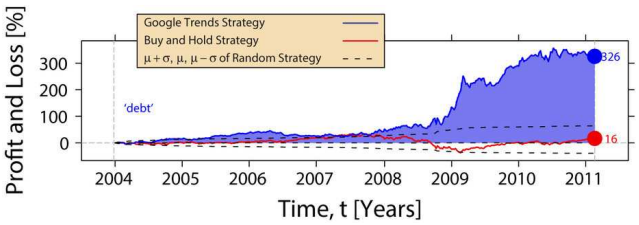
Notes:

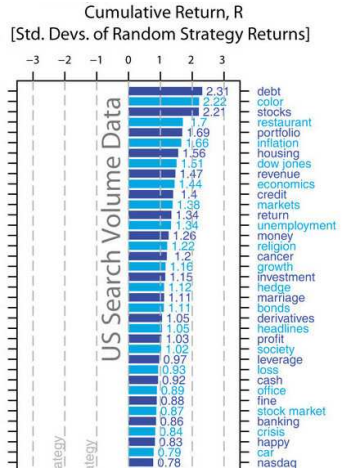
- Algorithmic Trading Review. P. Treleaven, M. Galas, V. Lalchand. Communications of the ACM, Vol. 56 No. 11, Pages 76-85. 2013
  http://cacm.acm.org/magazines/2013/11/169035-algorithmic-trading-review/fulltext

Notes:

- Twitter mood predicts the stock market. J. Bollen, H. Mao, X. Zeng. Journal of Computational Science 2011.
  http://www.sciencedirect.com/science/article/pii/S187775031100007X

Social Data Science Stories: Trading with Google Trends





- Google Trends of many terms to predict the DJIA
- Above: high profit of trading based on the trend of "debt"
- Right: top trend terms by generated profit
- Terms were chosen based on "relatedness to the concept of stock markets"

Predicting the DJIA with any Google trend

keyword	t-stat	keyword	t-stat	keyword	t-stat	keyword	t-tstat
multiple sclerosis	-2.1	Chevrolet Impala	-1.9	Moon Buggy	-2.1	labor	-1.5
muscle cramps	-1.9	Triumph 2000	-1.9	Bubbles	-2.0	housing	-1.2
premenstrual syndrome	-1.8	Jaguar E-type	-1.7	Rampage	-1.7	success	-1.2
alopecia	2.2	Iso Grifo	1.7	Street Fighter	2.3	bonds	1.9
gout	2.2	Alfa Romeo Spider	1.7	Crystal Castles	2.4	Nasdaq	2.0
bone cancer	2.4	Shelby GT 500	2.4	Moon Patrol	2.7	investment	2.0

- Independent validation with the same method but random keywords
- Table: Return of trading strategies based on trends for each word
- t-stat: Indicates the average profit of a trading strategy
- Terms like "Moon Patrol" or "Shelby GT 500" are better than "debt"

Notes:

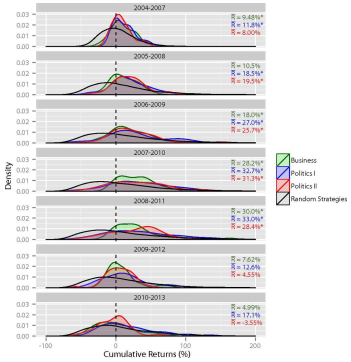
- Quantifying Trading Behavior in Financial Markets Using Google Trends. T. Preis, H. S. Moat, H. E. Stanley. Nature 2013. http://www.nature.com/articles/srep01684

Notes:

- Predicting financial markets with Google Trends and not so random keywords. D. Challet, A. B. H. Ayed. arXiv 2013.
  https://arxiv.org/abs/1307.4643

The stock market as a complex adaptive system

- Figure: Distribution of profits when trading with Google Trends for three year periods, between 2004 and 2011
- Cumulative profits reached an average of more than 30% between 2006 and 2011
- For more recent years, profit disappears



The stock market is a complex adaptive system

- Profit-seeking traders improve their strategies
- The behavior of the system reacts to new information
- Successful trading can be a *self-defeating prophecy*

If a social trend makes profit in trading, soon it won't

Market efficiency

Eugene Fama

Efficient market hypothesis (Eugene Fama)

Prices fully adapt to all available information

- **Weak form**: Prices adapt to all past publicly available information
- **Semi-strong form**: Prices adapt to all past and newly produced publicly avaliable information
- **Strong form**: Prices adapt to all information, even hidden "insider" information

- Different markets can have different levels of efficiency, or even be inefficient
- Market inefficiency can create price fluctuations due to scarce of unreliable information

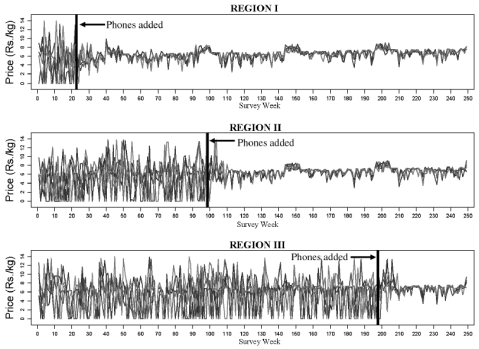
Notes:

- Quantifying the semantics of search behavior before stock market moves. Curme, C., Preis, T., Stanley, H. E. and Moat, H. S. PNAS (2014) http://www.pnas.org/content/111/32/11600

Notes:

- Eugene Fama. Wikipedia 2017. https://en.wikipedia.org/wiki/Eugene_Fama
- Efficient Market Hypothesis. Wikipedia 2017. https://en.wikipedia.org/wiki/Efficient-market_hypothesis

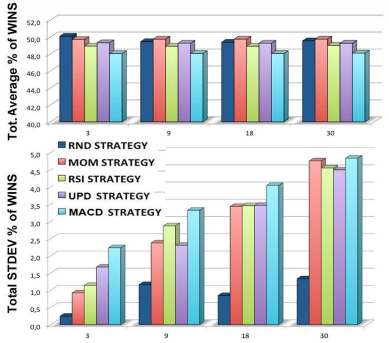
Market efficiency and price fluctuations



- Weekly sardine prices in three fish markets in Kerala, India
- Large fluctuations due to difficulty to price fish when guessing with limited information
- Mobile phones allows fast, cheap, and widespread access to price information
- The market becomes more efficient (semi-strong form)

Market efficiency: Implications for trading

- In efficient markets, it is impossible to trade for profit:
  - using publicly available information (weak form)
  - or using private information too (strong form)
  - then investors achieve high returns by chance or by adopting risk in trading
- Study comparing standard strategies to random traders:
  - Accuracy measured as % of wins over time windows
  - Accuracies are similar for random and standard traders
  - Risk of trading: std of accuracy in predictions
  - Standard traders are much riskier than random ones



Notes:

- The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector. R. Jensen. Q J Econ (2007) https://doi.org/10.1162/qjec.122.3.879

Notes:

- Are Random Trading Strategies More Successful than Technical Ones? A. Biondo, A. Pluchino, A. Rapisarda, D. Helbing. PLoS One, 2013. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068344

Exercise 06: Bitcoin time series analysis

Step	Exercise 6
Plan	Test if use of Bitcoin leads price increases
Data Retrieval	Use JSON APIs to gather network use and pricing data
Data Processing	Time series construction and transformation
Analysis	Stationarity and Granger causality tests
Conclusion	Does transaction volume predict changes in price?
Critique	How robust are the results? What could be the explanations?

Bitcoin JSON APIs

Notes:

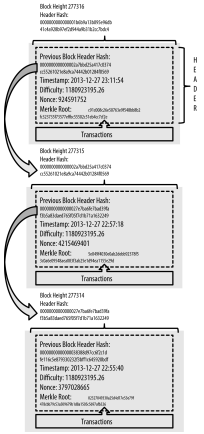
Notes:

The Bitcoin cryptocurrency

The Bitcoin Criptocurrency (BTC):
Online payment P2P protocol that works as a currency without a central authority

- Designed by "Satoshi Nakamoto" in 2008, open source system since 2009
- Initial use for illegal activities, now accepted by many legal businesses
- Famous for its wild price fluctuations and bubbles
  - In 2010, two pizzas worth were purchased for 10,000 BTC
  - Currently worth approximately 10 Million USD
- Importance of cybersecurity: Mt. Gox, the first Bitcoin trading platform, filed bankruptcy in 2014 after reporting that 744.000 BTC were stolen
- Currently hundreds of "altcoins" with similar technology

The Blockchain technology



- Bitcoin is mined solving a computationally expensive problem
- The supply of BTC saturates at a maximum of 21 Million BTC
- Every 10 minutes on average, a block is generated with the miner rewards and transactions that record the transfer of Bitcoin
- **The Blockckain**
  - distributed ledger that contains all the generated blocks
  - works as a public database with the history of all transactions
  - once recorded, the data in a block cannot be retroactively altered
- Just like monetary transactions, *smart contracts* can be implemented on blockchain technologies
  - Simple scripts that define *conditional transactions* without third parties
  - Gambling, auctions, credit enforcement, performance-based payouts...

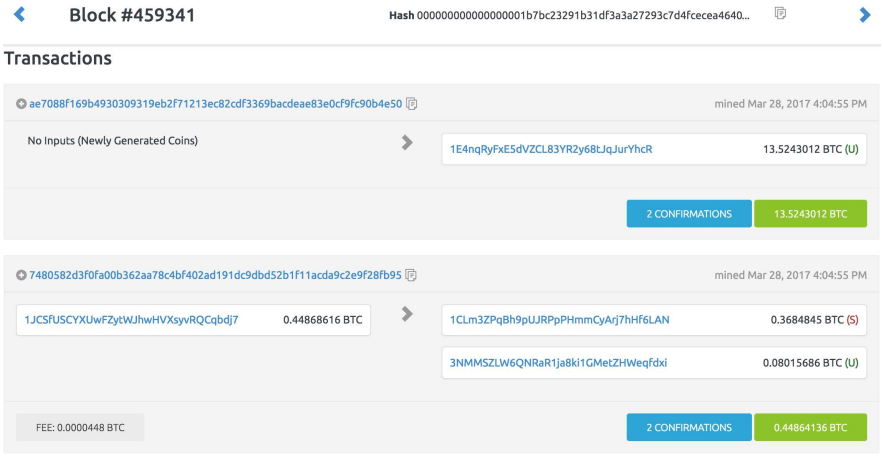
Notes:

- A brief introduction to Bitcoin, Blockchain and cryptocurrency by PwC FinTech `http://www.pwc.com/us/en/financial-services/fintech/bitcoin-blockchain-cryptocurrency.html`
- You can find a complete timeline of the growth of bitcoin via `http://historyofbitcoin.org/` or more about bitcoin history `https://en.wikipedia.org/wiki/History_of_bitcoin`

Notes:

- Figure source `https://napkinfinance.com/napkin/bitcoin-blockchain/`
- More information about blockchain can be found in such an in-depth guide like `https://blockgeeks.com/guides/what-is-blockchain-technology/`.

Bitcoin Transactions



Bitcoin JSON APIs

CoinDesk BLOCKCHAIN

- Public JSON APIs provide data on Bitcoin markets and Blockchain transactions
- Sending HTTP requests to URLS with parameters, we get data in JSON format
- JSON is short for "JavaScript Object Notation"
  - easy for humans to read and write
  - easy for machines to parse and generate
- An example of JSON Data from CoinDesk Bitcoin Price Index

sample.json

```
{"bpi":{"2017-03-01":1230.016,"2017-03-02":
1260.924,"2017-03-03":1290.786,"2017-03-04":
1267.68,"2017-03-05":1277.685},"disclaimer":"This data was
produced from the CoinDesk Bitcoin Price Index. BPI value data
returned as USD.","time":{"updated":"Mar 6, 2017 00:03:00
UTC","updatedISO":"2017-03-06T00:03:00+00:00"}}
```


Notes:

- Screenshot of block content captured from https://blockexplorer.com
- Example of two unrelated transactions in the same block
- More about bitcoin transactions can be found via https://en.bitcoin.it/wiki/Transaction and http://www.coindesk.com/information/how-do-bitcoin-transactions-work/

Notes:

JSON data request to the CoinDesk API

The CoinDesk Bitcoin Price Index (BPI or XBP)

Measures a weighted average of Bitcoin prices across exchange platforms that meet certain quality criteria

The CoinDesk API has two endpoints (URLs) for data access:

- BPI real-time data http://api.coindesk.com/v1/bpi/currentprice.json
- BPI historical data http://api.coindesk.com/v1/bpi/historical/close.json

Example of how to set up an API call:

```
> BPRequestStartDay <- "2017-03-01"
> BPRequestEndDay <- "2017-03-05"
> url = paste0("http://api.coindesk.com/v1/bpi/historical/close.json?start=",
  BPRequestStartDay,"&end=", BPRequestEndDay)
> url
[1] "http://api.coindesk.com/v1/bpi/historical/close.json?start=2017-03-01&end=2017-03-05"
```


CoinDesk Historical Data API

Parameter meanings and request-response example:

Historical BPI data

We offer historical data from our Bitcoin Price Index through the following endpoint:

http(s)://api.coindesk.com/v1/bpi/historical/close.json

By default, this will return the previous 31 days' worth of data. This endpoint accepts the following optional parameters:

- **?index=[USD/CNY]**The index to return data for. Defaults to USD.
- **?currency=<VALUE>**The currency to return the data in, specified in ISO 4217 format. Defaults to USD.
- **?start=<VALUE>&end=<VALUE>** Allows data to be returned for a specific date range. Must be listed as a pair of start and end parameters, with dates supplied in the YYYY-MM-DD format, e.g. 2013-09-01 for September 1st, 2013.
- **?for=yesterday**Specifying this will return a single value for the previous day. Overrides the start/end parameter.

Sample Request: `http(s)://api.coindesk.com/v1/bpi/historical/close.json?start=2013-09-01&end=2013-09-05` ***Sample JSON Response:***


```
{"bpi":{"2013-09-01":128.2597,"2013-09-02":127.3648,"2013-09-03":127.5915,"2013-09-04":120.5738,"2013-09-05":120.5333},"disclaimer":"This data was produced from the CoinDesk Bitcoin Price Index. BPI value data returned as USD.","time":{"updated":"Sep 6, 2013 00:03:00 UTC","updatedISO":"2013-09-06T00:03:00+00:00"}}
```


Notes:

- You can find how to specify the request url format in http://www.coindesk.com/api/
- About the CoinDesk Bitcoin Price Index: http://www.coindesk.com/price/bitcoin-price-index/

Notes:

CoinDesk API Documentation http://www.coindesk.com/api/

CoinDesk API request example in R

- Using a URL constructed as in previous slides:

```
# read text lines from a connection
> raw.data <- readLines(url, warn = "F")
> raw.data
[1] "{\"bpi\":{\"2017-03-01\":1230.016,\"2017-03-02\":1260.924,\"2017-03-03\":1290.786,\"
  2017-03-04\":1267.68,\"2017-03-05\":1277.685},\"disclaimer\":\"This data was produced from
   the CoinDesk Bitcoin Price Index. BPI value data returned as USD.\",\"time\":{\"updated\"
  :\"Mar 6, 2017 00:03:00 UTC\",\"updatedISO\":\"2017-03-06T00:03:00+00:00\"}}"
```

- The result is a string that needs to be interpreted as JSON

Read JSON data with rjson or jsonlite

- libraries "rjson" and "jsonlite" convert R strings into structured objects and vice-versa
- With the raw data we got from the API, we can interpret it as JSON with *fromJSON()*
- The result is an object with attributes that contain the data, like a group of values

```
library(rjson)
# alternatively
# library(jsonlite)

raw <- fromJSON(raw.data)
```

```
raw                 List of 3
  bpi :List of 5
  ..$ 2017-03-01: num 1230
  ..$ 2017-03-02: num 1261
  ..$ 2017-03-03: num 1291
  ..$ 2017-03-04: num 1268
  ..$ 2017-03-05: num 1278
  disclaimer: chr "This data was produced from the…
  time :List of 2
  ..$ updated : chr "Mar 6, 2017 00:03:00 UTC"
  ..$ updatedISO: chr "2017-03-06T00:03:00+00:00"
```


Notes:

Notes:

The Blockchain.info API

The Blockchain.info API provides historical data from the **Bitcoin blockchain**

- Making HTTP requests we get JSON data on individual transactions, trasaction volume, etc
- Example of parameters for the "transactions per second" historical data

Charts API Get the data behind Blockchain's charts

This method can be used to get and manipulate data behind all Blockchain.info's charts.

URL: https://api.blockchain.info/charts/$chartName?timespan=$timespan&rollingAverage=$rollingAverage&start=$start&format=$format&sampled=$sampled
Method: GET
Example: https://api.blockchain.info/charts/transactions-per-second?timespan=5weeks&rollingAverage=8hours&format=json

- **$timespan** Duration of the chart, default is 1 year for most charts, 1 week for mempool charts. (Optional)
- **$rollingAverage** Duration over which the data should be averaged. (Optional)
- **$start** Datetime at which to start the chart. (Optional)
- **$format** Either JSON or CSV, defaults to JSON. (Optional)
- **$sampled** Boolean set to 'true' or 'false' (default 'true'). If true, limits the number of datapoints returned to ~1.5k for performance reasons. (Optional)

Notes:

More details on: https://blockchain.info/api/charts_api

Introduction to Time Series Analysis in R

Notes:

Time series in R

- Time series are sets of values X_t measured in consecutive moments in time t
- We use the "ChickEgg" data example with chicken population and egg production in the US
- Data is already set up as time series objects, with time data and values

```
# data("ChickEgg") from library(lmtest)
> chicken = ChickEgg[,"chicken"]
> chicken
Time Series:
Start = 1930
End = 1983
Frequency = 1
[1] 468491 449743 436815 444523 433937
  389958 403446 423921 389624 418591
  438288 422841 476935 542047 582197 ...
```

```
# data("ChickEgg") from library(lmtest)
> egg = ChickEgg[,"egg"]
> egg
Time Series:
Start = 1930
End = 1983
Frequency = 1
[1] 3581 3532 3327 3255 3156 3081 3166 3443
   3424 3561 3640 3840 4456 5000 5366 5154
   5130 5077 5032 5148 5404 5322 ...
```


Time series visualization

- We visualize time series with time points t on the x axis and values (X_t) on the y axis

```
# plot "chicken population"
plot(index(chicken), chicken, type="l")
```


```
# plot "egg production"
plot(index(egg), egg, type="l")
```

- You can also use the plot() method of the time series object: *plot(chicken)*

Notes:

- Chickens, Eggs, and Causality, or Which Came First? W. Thurman and M. Fisher. American Journal of Agricultural Economics. 1988. http://www.sungpark.net/ChickensEggs.pdf

Notes:

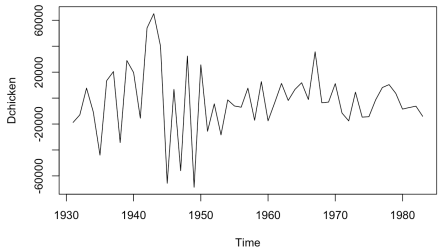
Window and difference

- You can select a period of time with the *window()* function

```
chicken_window <- window(chicken, 1960, 1980)
plot(chicken_window)
```

- The *diff()* function calculates differences, e.g $\Delta X_t = X_t - X_{t-1}$

```
Dchicken <- diff(chicken, differences=1)
plot(Dchicken)
```



Autoregression modelling

- The value of X_t might depend on its previous values
- One approximation is the AR(1) model: X_t linearly depends on X_{t-1}

$$X_t = c + \varphi X_{t-1} + \varepsilon_t$$

- c is a time-independent intercept of the value of X_t
- φ is the autocorrelation coefficient, measures how much of X_{t-1} is retained in X_t
- ε_t are the residuals (errors) of the model. They satisfy some assumptions:
  - are normally distributed with zero mean
  - do not depend on X_{t-1}
  - are not correlated over time: $\rho(\varepsilon_t, \varepsilon_{t-1}) \sim 0$

Notes:

Notes:

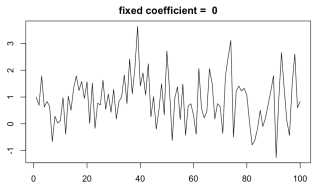
Examples of AR(1) time series

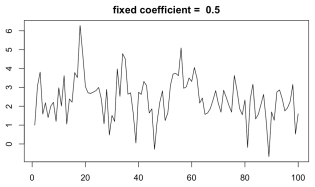
- AR(1) model:

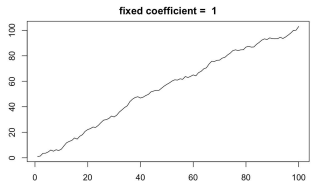
$$X_t = c + \varphi X_{t-1} + \varepsilon_t$$

- Eexamples with φ set to 0, 0.5, and 1
- Right: code to simulate the time series below

```
phi <- 0 # or 0.5, 1
c <- 1
n <- 100
X <- numeric(length=n)
X[1] <- 1
for(i in 2:n){
  X[i] <- c + phi*X[i-1] + rnorm(1)
}
```



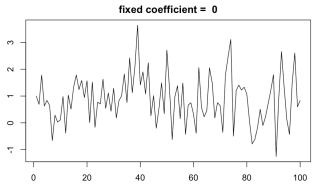




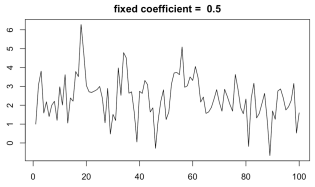
Stationary time series

Stationary time series have constant mean and variance

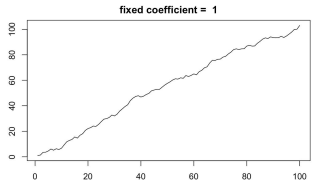
- If the mean or std changes a lot over different periods, the time series is non-stationary
- Non-stationary time series have $\varphi \geq 1$
- For a non-stationary X_t, ΔX_t *might* be stationary



Stationary



Stationary



Non-stationary

Notes:

In the plots, "fixed coefficient" is ϕ

Notes:

Stationarity test of *Eggs*

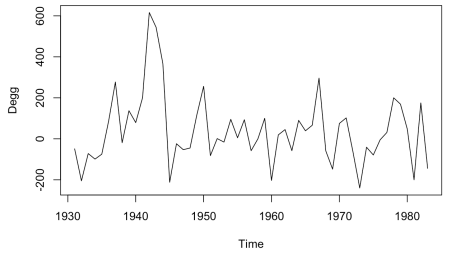
- We test if egg production, *Eggs*, is stationary
- check the confidence interval of model parameters to decide if $\varphi < 1$
- e.g. see the values for the term 'egg_lag1'


```
> n <- length(egg)
> egg_lag0 <- egg[2:n]
> egg_lag1 <- egg[1:(n-1)]
> m <- lm(egg_lag0~egg_lag1)
> confint(m)
                  2.5 %     97.5 %
(Intercept) -42.8621562 495.375262
egg_lag1      0.9091027   1.015661
```

- 1 is between 0.9091027 and 1.015661: **The time series is not stationary**

Stationarity test of Δ*Eggs*

- We test if changes in egg production, Δ*Eggs*, is stationary
- check the confidence interval of model parameters to decide if $\varphi < 1$
- e.g. see the values for the term 'Degg_lag1'




```
> n <- length(Degg)
> Degg_lag0 <- Degg[2:n]
> Degg_lag1 <- Degg[1:(n-1)]
> m <- lm(Degg_lag0~Degg_lag1)
> confint(m)
                  2.5 %     97.5 %
(Intercept) -21.5846230 71.7582909
Degg_lag1     0.1026647  0.6359425
```

- both 0.1026647 and 0.6359425 are far below 1: **The time series is stationary**

Notes:

Notes:

Linear regression with multiple independent variables

- Our first regression models had only one independent variable, but can have more
- The output variable can be a linear combination of multiple independent variables and a stochastic term

$$Y = \beta_0 + \sum_{i=1}^{m} \beta_i Z_i + \varepsilon$$

- β_0 is the intercept
- $\beta_0, \ldots, \beta_m$ are the coefficients of each of the independent variables
- ε is the error term or residual

Granger non-causality test

- A Granger (non-)causality test checks if one time series anticipates (or leads) another
  - We test if knowing one time series allows us to improve the prediction of another
  - It does not prove causation, but the lack of a result evidences absence of causation
  - It does not account for confounding factors
  - It is better called a **non-causality test**
- To test if X "Granger causes" Y, fit two models and test which one is better

  Model 1: $Y_t = c + \varphi Y_{t-1} + \varepsilon_t$
  Model 2: $Y_t = b + \alpha Y_{t-1} + \beta X_{t-1} + \xi_t$

- Condition: Y_t must be stationary (i.e. $\varphi < 1$)
- Granger test steps:
  1. Fit AR models: Model 1 and Model 2
  2. Evaluate if Model 2 is better than Model 1. If so, then X_{t-1} is informative for predicting Y_t

Notes:

Notes:

Evaluating models: The Akaike Information Criterion (AIC)

- The AIC measures the **relative** quality of statistical models for a given data set

$$AIC = 2k - 2ln(\hat{L})$$

- $\hat{L}$ = likelihood of the model. Often maximized for the parameters of the model
- k = the number of parameters of the model
- The AIC measures information loss by the model: **The higher, the worse**
- The term $2k$ penalizes overfitting with too many parameters
- The AIC works as a *rule of thumb* to select the best model
- For linear regression k is the number of coefficients plus one (the intercept)

Fitting AR models

- We fit models with various independent variables using a "+" on the formula of *lm()*
- For example, fit a model depending on lag1 data from both Degg and Dchicken

```
> n <- length(Degg)
> Degg_lag0 <- Degg[2:n]
> Degg_lag1 <- Degg[1:(n-1)]
> Dchicken_lag0 <- Dchicken[2:n]
> Dchicken_lag1 <- Dchicken[1:(n-1)]
> m <- lm(Degg_lag0 ~ Degg_lag1 + Dchicken_lag1)
> m
Call:
lm(formula = Degg_lag0 ~ Degg_lag1 + Dchicken_lag1)

Coefficients:
(Intercept)      Degg_lag1  Dchicken_lag1
 20.2337621       0.4489163     -0.0008422
```


Notes:

- Wiki page about AIC https://en.wikipedia.org/wiki/Akaike_information_criterion
- Remember the explanation and examples about Likelihood on Lecture 04
- Examples of parameters in the previous slide:
  - Model 1: The parameters are c and φ, thus $k = 2$
  - Model 2: The parameters are b, α, and β, thus $k = 3$
- The error term ε is not a parameter!

Notes:

A Granger causality example with $\Delta Chicken$ and $\Delta Eggs$

- What comes first, the chicken or the egg?
- $\Delta Chicken$ and $\Delta Eggs$ are stationary
  1. Does $\Delta Chicken$ Granger-cause $\Delta Eggs$?

     Using lm notation, we test if the AIC of
     $\Delta Eggs_t \sim \Delta Eggs_{t-1} + \Delta Chicken_{t-1}$
     is lower than the AIC of
     $\Delta Eggs_t \sim \Delta Eggs_{t-1}$
  2. Does $\Delta Eggs$ Granger-cause $\Delta Chicken$?

     Using lm notation, we test if the AIC of
     $\Delta Chicken_t \sim \Delta Chicken_{t-1} + \Delta Eggs_{t-1}$
     is lower than the AIC of
     $\Delta Chicken_t \sim \Delta Chicken_{t-1}$

Does $\Delta Chicken$ Granger-cause $\Delta Eggs$?

$\Delta Eggs_t \sim \Delta Eggs_{t-1}$

- Fit model 1 with only ΔEgg data and measure AIC

```
> n <- length(Degg)
> Degg_lag0 <- Degg[2:n]
> Degg_lag1 <- Degg[1:(n-1)]
> m1 <- lm(Degg_lag0~Degg_lag1)
> AIC(m1)
[1] 680.9628
```

$\Delta Eggs_t \sim \Delta Eggs_{t-1} + \Delta Chicken_{t-1}$

- Fit model 2 including $\Delta Chicken$ data and measure AIC

```
> Dchicken_lag0 <- Dchicken[2:n]
> Dchicken_lag1 <- Dchicken[1:(n-1)]
> m2 <- lm(Degg_lag0 ~
           Degg_lag1 + Dchicken_lag1)
> AIC(m2)
[1] 682.3894
```

* The AIC of model 2 is **higher** than the AIC of model 1

\- $\Delta Chicken$ does not Granger-cause $\Delta Eggs$

Notes:

Notes:

Does $\Delta Eggs$ Granger-cause $\Delta Chicken$?

$$\Delta Chicken_t \sim \Delta Chicken_{t-1}$$

- Fit model 1 with only $\Delta Chicken$ data and measure AIC

```
> n <- length(Dchicken)
> Dchicken_lag0 <- Dchicken[2:n]
> Dchicken_lag1 <- Dchicken[1:(n-1)]
> m1 <- lm(Dchicken_lag0~Dchicken_lag1)
> AIC(m1)
[1] 1206.906
```

$$\Delta Chicken_t \sim \Delta Chicken_{t-1} + \Delta Eggs_{t-1}$$

- Fit model 2 including $\Delta Eggs$ data and measure AIC

```
> Degg_lag0 <- Degg[2:n]
> Degg_lag1 <- Degg[1:(n-1)]
> m2 <- lm(Dchicken_lag0 ~
           Dchicken_lag1 + Degg_lag1)
> AIC(m2)
[1] 1198.925
```

* The AIC of model 2 is **lower** than the AIC of model 1

\- $\Delta Eggs$ Granger-causes $\Delta Chicken$

The meaning of Granger causality

- **We have not proven causality!**
- A model in which eggs cause chickens can explain our results
  - It is faster and cheaper to turn an egg into a chicken than to get a chicken to produce eggs
  - Excess production in eggs can be used to produce more chicken

Explanation involving third factors or confounds can always be possible:

- Seasonal effects can create lagged increases in production
- Higher temperature can first accelerate the production of eggs and prolong the life of chicken
- Economic growth: Egg production takes less time to use investments than chicken farming

What we did: We found evidence that a model of changes in chicken population improves by including changes in egg production

Notes:

Notes:

Summary

1. Social trends as responses of communities
   - Endogenous vs exogenous peaks: external factors producing spikes
   - Critical vs subcritical behavior: social influence creating slow increases and decreases
2. Computational finance and social media
   - Google trends terms are not profitable
   - The efficient market hypothesis: Trading is hard
3. JSON APIs
   - A brief introduction to Bitcoin and its related APIs
   - How to gather and process JSON data about Bitcoin
4. Time Series analysis
   - Stationarity tests
   - Granger non-causality testing with AIC

Quiz time!

1. Only my friends watch my YouTube videos. If the BBC talks about one, will it have endogenous or exogenous dynamics?
2. Will it be critical or subcritical?
3. What is the percentage of volume at the peak in an Endogenous Subcritical collective reaction?
4. Which is the most profitable term when trading based on Google trends?
5. If we manage to predict Bitcoin prices, does it mean that it is an inefficient market?
6. How often are blocks added to the Bitcoin blockchain?
7. If I get a φ estimate of 3, is the time series stationary?
8. If people becoming vegan convice others to become vegetarian, does $\Delta Eggs$ Granger-cause $\Delta Chicken$?

Notes:

Notes: