

Dynamic Topic Modeling for Bitcoin Message Fora

Elisabeth Bommers

Cathy Yi-Hsuan Chen

Ernie Gin Swee Teo

Wolfgang K. Härdle

Marco Linton

Ladislaus von Bortkiewicz Chair of Statistics

Sim Kee Boon Institute for Financial Economics

International Research Training Group

Humboldt-Universität zu Berlin

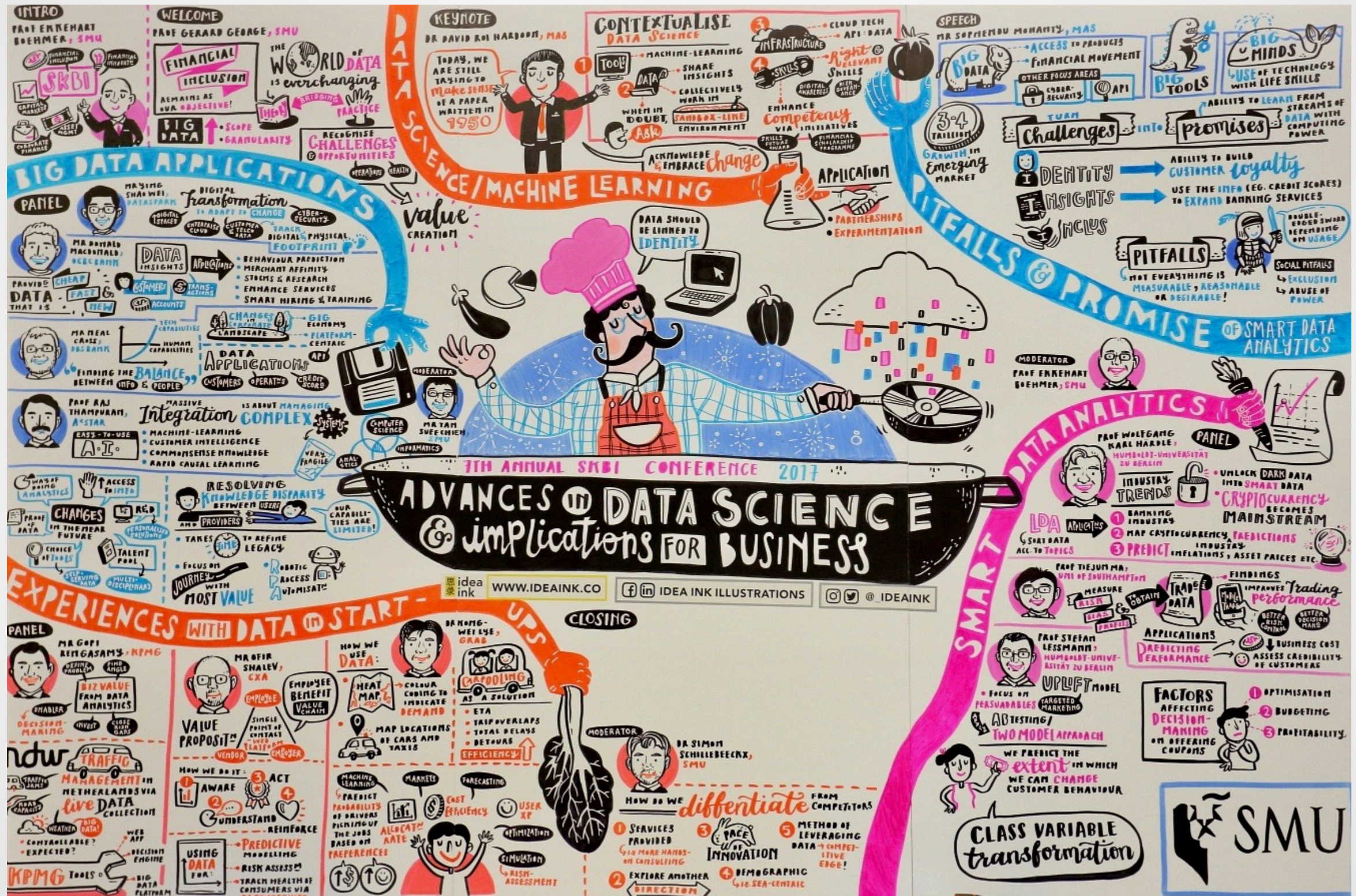
lvb.wiwi.hu-berlin.de

www.case.hu-berlin.de

irtg1792.hu-berlin.de



Smart Data Analytics



Smart Data Analytics


Machine Learning not only provides new tools: it solves a different problem!

- § ML produces predictions of y from x
- § ML manages to uncover generalizable patterns
- § ML discovers flexible data relations without overfitting
- § Traditional E'tcs likes to produce good estimates of parameters β that underlie the relationship between y and x
- § Estimates are to be consistent
- § Rely on assumptions on DGP

Old Ansatz: the past tells us.

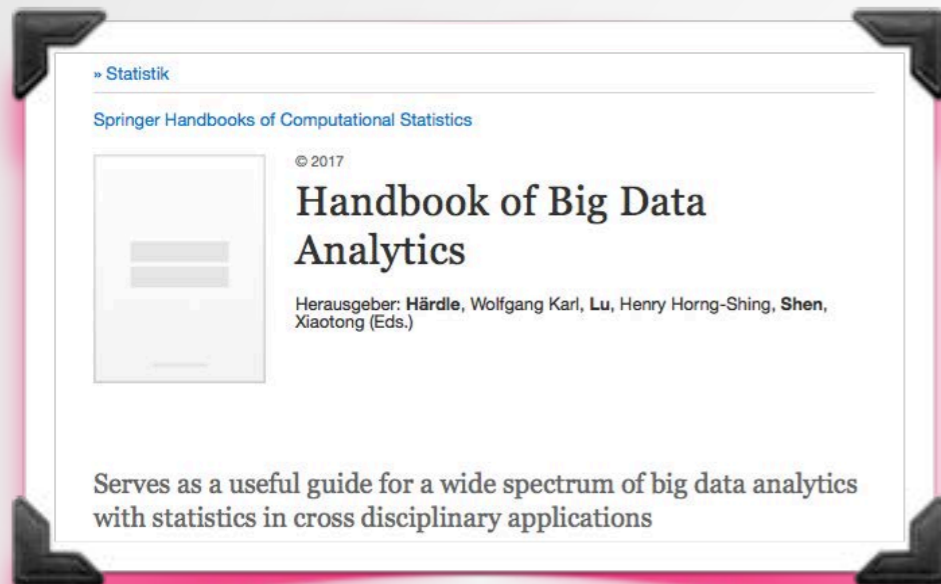
New Ansatz: how rich is this?


$$\mathbf{E}(X_{n+1} \mid X_1, \dots, X_n) = X_n$$

$$P(t, T) = E_{Q_*} \left[\frac{B(t)}{B(T)} \mid \mathcal{F}(t) \right]$$


Quantlets

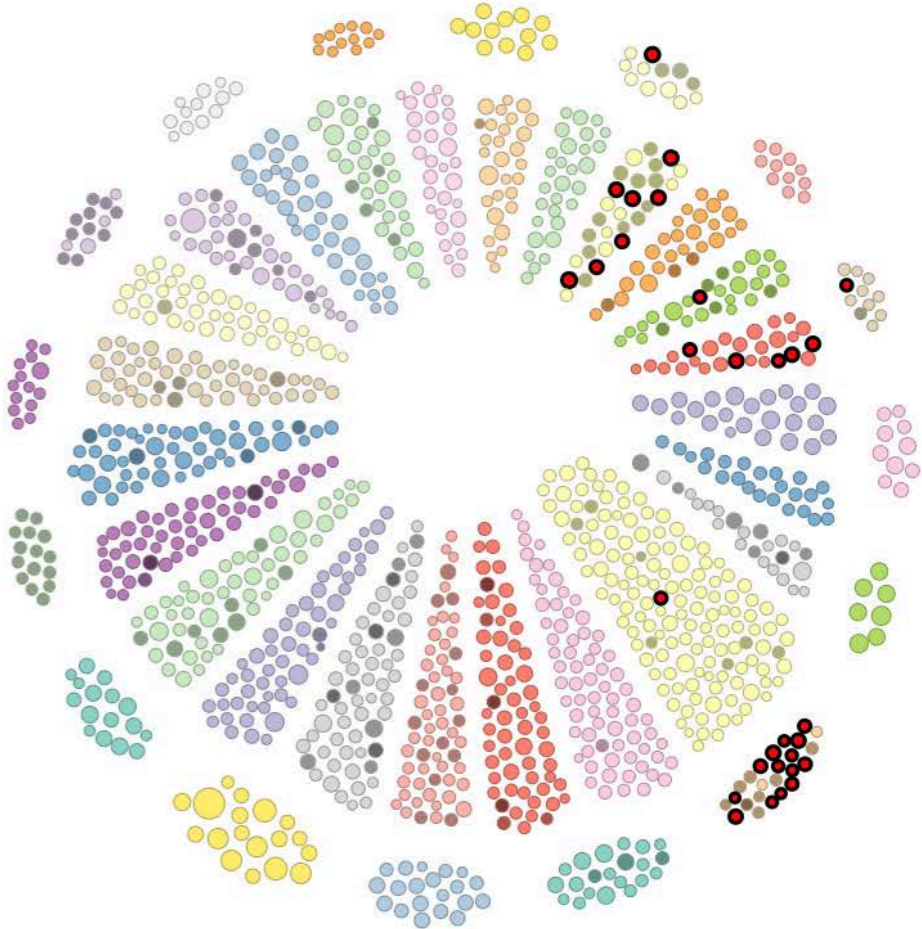
§ Unlock the value of dark data to transform it into smart data




QuantNetXploRer
Full version [About](#)

Found items: 30

- MVAclusbank
- MVAcontbank2
- MVAdenbank3
- MVAncabank
- MVAncabanki
- MVApcabanki
- MVAboxbank6
- MVAhisbank1
- MVAhisbank2
- MVAdraftbank4
- MVAdenbank
- MVAashbank
- MVAscabank45
- MVAscabank56
- MVAscabank456
- MVAandcur2
- MVAparcoo1
- MVAregbank
- MVAcontbank3
- MVAfacebank10
- MVAfacebank50
- MVAboxbank1
- MVAandcur
- MVAaper
- MVAdisfbank
- MVAdenbank2
- MVApcabank
- MVApcabankr
- BCS_NPCAbiplot
- BCS_PCAbiplot



Cryptocurrencies

§ 1032 cryptos

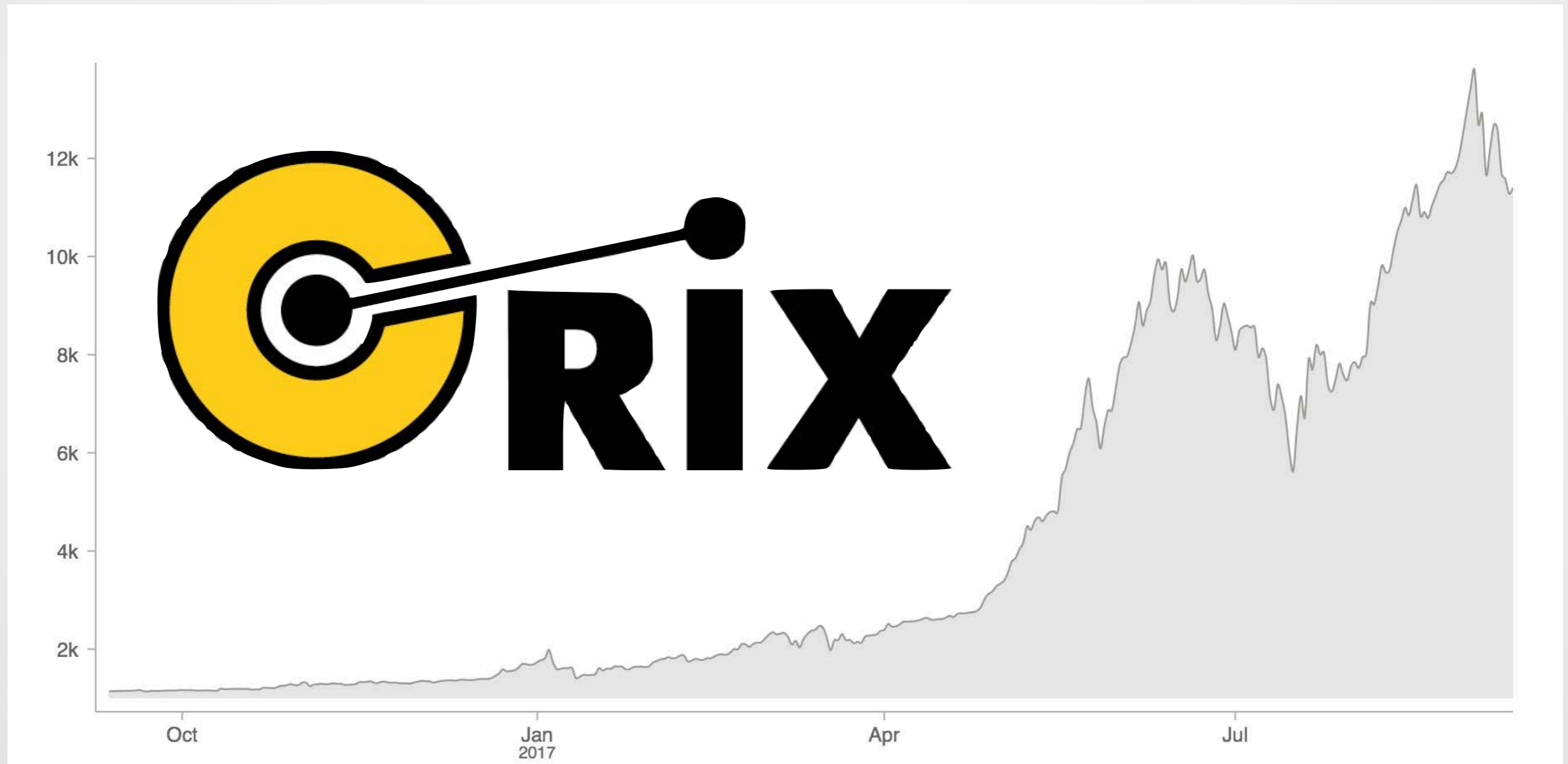
§ coinmarketcap.com

§ Social media mining

All ▾		Currencies ▾	Assets ▾	USD ▾	Next 100 →		View All
#	Name	Market Cap	Price	Circulating Supply	Volume (24h)	% Change (24h)	Price Graph (7d)
1	Bitcoin	\$71,647,732,142	\$4326.58	16,559,900 BTC	\$1,524,440,000	4.46%	
2	Ethereum	\$29,184,784,983	\$308.60	94,571,259 ETH	\$554,242,000	6.35%	
3	Bitcoin Cash	\$9,239,401,086	\$557.42	16,575,325 BCH	\$222,691,000	4.50%	
4	Ripple	\$8,413,405,786	\$0.219420	38,343,841,883 XRP *	\$111,308,000	2.96%	
5	Litecoin	\$3,681,991,715	\$69.61	52,892,832 LTC	\$340,665,000	7.43%	
6	Dash	\$2,472,328,466	\$327.35	7,552,485 DASH	\$17,477,800	2.42%	
7	NEM	\$2,354,859,000	\$0.261651	8,999,999,999 XEM *	\$3,713,270	1.42%	
8	Monero	\$1,733,819,935	\$115.05	15,070,536 XMR	\$26,394,800	2.91%	
9	IOTA	\$1,728,361,962	\$0.621818	2,779,530,283 MIOTA *	\$41,866,000	25.37%	
10	Ethereum Classic	\$1,468,462,694	\$15.39	95,433,422 ETC	\$121,315,000	6.25%	
11	OmiseGO	\$1,270,289,662	\$12.92	98,312,024 OMG *	\$88,641,100	14.11%	
12	NEO	\$1,171,955,000	\$23.44	50,000,000 NEO *	\$31,934,600	5.44%	
13	Qtum	\$871,872,500	\$14.78	59,000,000 QTUM *	\$159,643,000	13.71%	
14	BitConnect	\$869,213,320	\$130.25	6,673,372 BCC	\$7,305,680	5.16%	
15	Lisk	\$791,123,846	\$7.06	112,131,215 LSK *	\$14,824,500	6.40%	

CRIX

§ Unlock the value of dark data to transform it into smart data



Current cryptos in CRIX

Coin	Name	Price (in \$)	Market Cap (in \$K)	Volume (in \$K)
1	btc	4164.98	68,968,724	771,493
2	eth	296.48	28,036,288	363,008
3	xrp	0.21	8,240,628	137,781
4	ltc	67.48	3,569,122	216,855
5	dash	317.75	2,399,639	20,634
6	xem	0.25	2,291,980	2,729
7	xmr	111.65	1,682,479	15,143
8	iot	0.56	1,545,608	23,500
9	etc	14.58	1,391,278	142,793

Cryptos

- § Cryptocurrencies have become mainstream
- § Domain of technophiles and radicals ?
- § traded on many exchanges (blockchain)
- § bottomless source of information
- § Crowd wisdom a powerful indicator of major events



What can be done?

- § Dynamic topic modelling (DTM), text mining and machine learning
- § detect events: new trends in currencies, fraudulent schemes
- § unsupervised DTM opinion evolution of topics
- § test hypothesis of self-fulfilling prophecies and herding behaviour
- § Smailovic et al. (2013) : improve prediction from Twitter sentiments

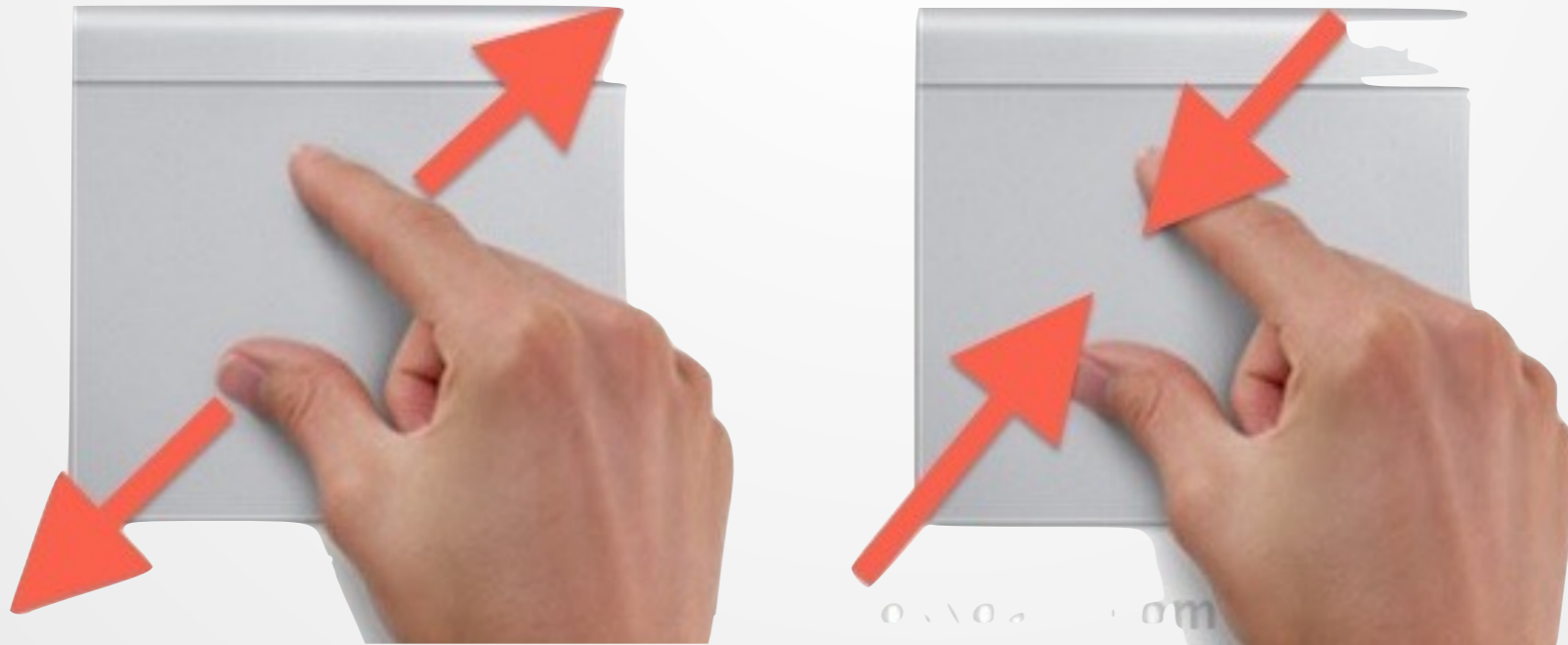
Introduction

- § Amount of unstructured dark data is growing
- § Conventional ways: read, search and links?
- § Theme based exploration offers more possibilities



Desiderata

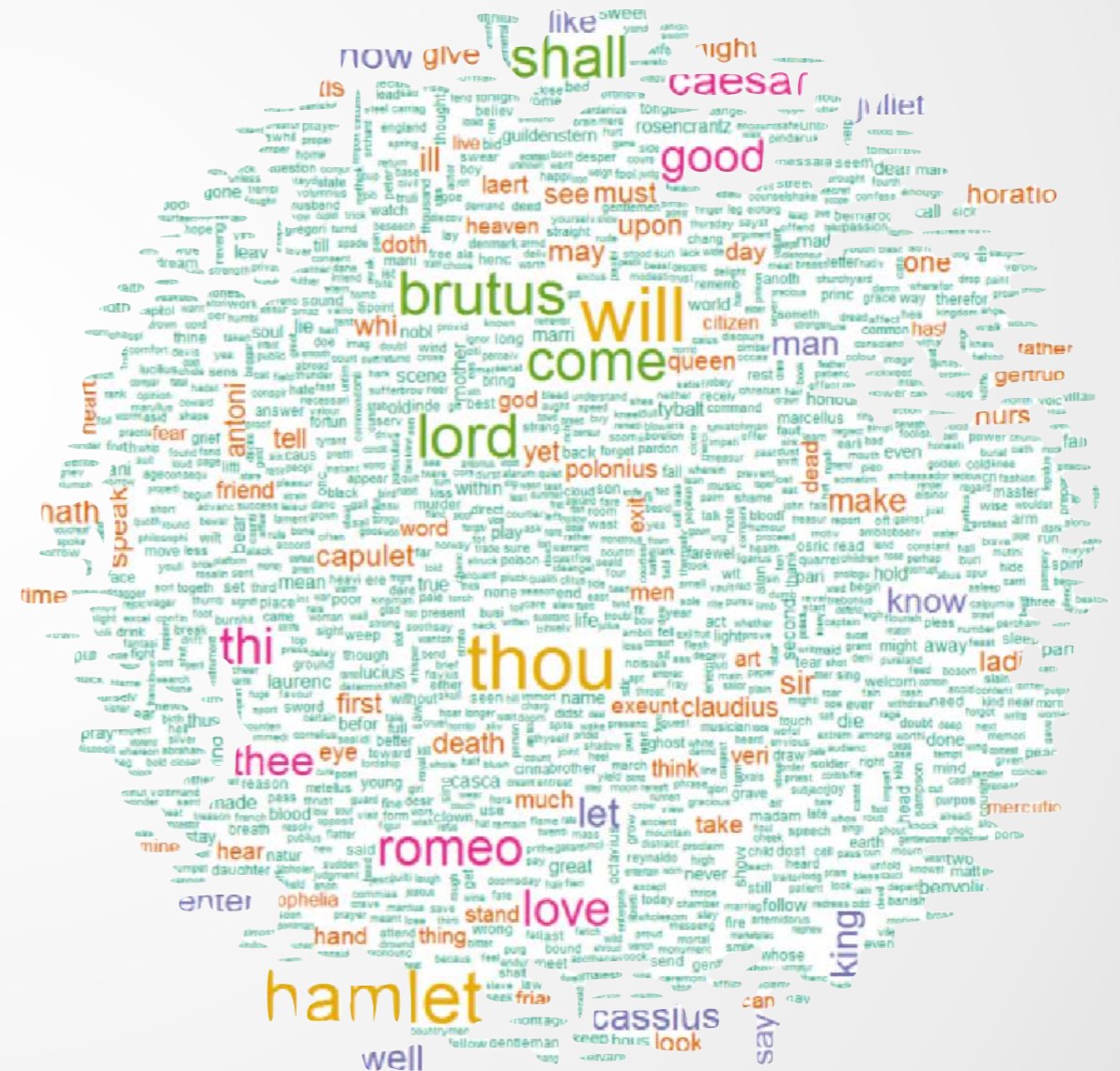
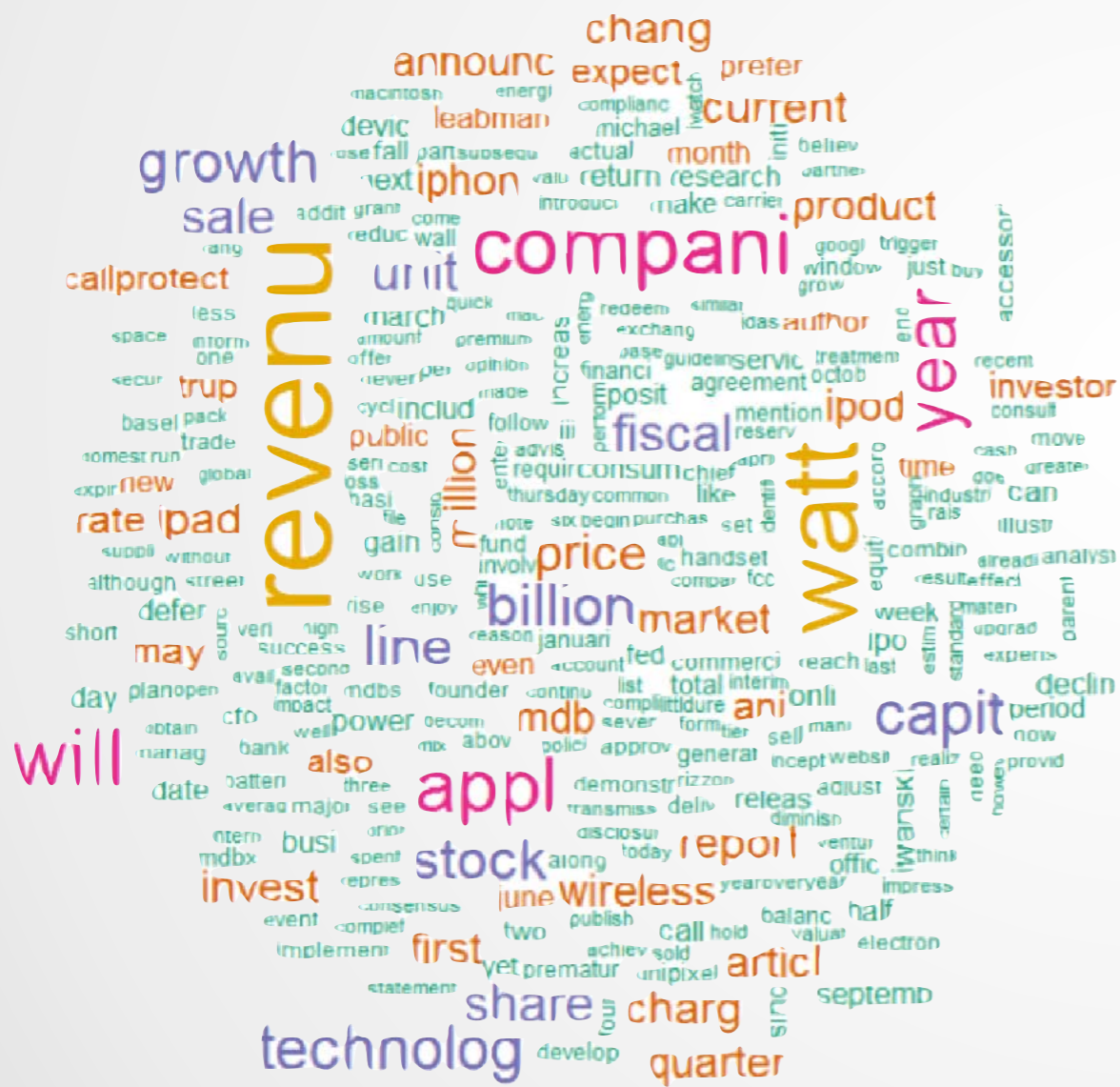
- § label/ tag all documents with corresponding topics
- § „Zoom in“ and „zoom out“ to find specific or broader themes
- § Find new connections and changes over time
- § But to read all the available documents is not in human power
- § Solution: probabilistic topic modeling



Word distribution

Topic: NASDAQ

Topic: Shakespeare



Latent Dirichlet allocation (LDA)

- § Simple topic model
- § Document may exhibit several topics
- § Each topic is a distribution over a fixed vocabulary
- § Topics are generated first, before the documents
- § Data is arising from a generative process

LDA: Generative Process

Words are generated in a two-stage process:

1. Randomly choose a distribution over topics.
2. For each word in the document
 - a) Randomly choose a topic from the distribution over topics
 - b) Randomly choose a word from the corresponding distribution over the vocabulary.

Example 1

- Observed data
 - 2 documents
 - 155 distinct words
- §? number of topics

15. **Silent Night! Holy Night!**
 From the Third (unpublished) Part of "HYMNS AND MUSIC FOR THE YOUNG," By permission of the Author.

1. Si-lent night! Ho-ly night! All is calm, all is bright; Round yon Virgin Mother and Child!

Jingle Bells

Dashing through the snow
 In a one horse open sleigh
 O'er the fields we go
 Laughing all the way
 Bells on bob tails ring
 Making spirits bright
 What fun it is to laugh and sing
 A sleighing song tonight

Oh, jingle bells, jingle bells
 Jingle all the way
 Oh, what fun it is to ride
 In a one horse open sleigh
 Jingle bells, jingle bells
 Jingle all the way
 Oh, what fun it is to ride

- Unobserved variables
 - topic assignment of words
 - per-document topic proportions (final result)

Example 2

Example 2: Shakespeare's tragedies

$T = \{art, bear, call, day, dead, dear, death, die, eye, fair, father, fear, friend, god, good, heart, heaven, king, lady, lie, like, live, love, make, man, mean, men, must, night, queen, think, time\}$
 $= \{t_1, \dots, t_{32}\}$

T – special vocabulary selected among 100 most frequent words.

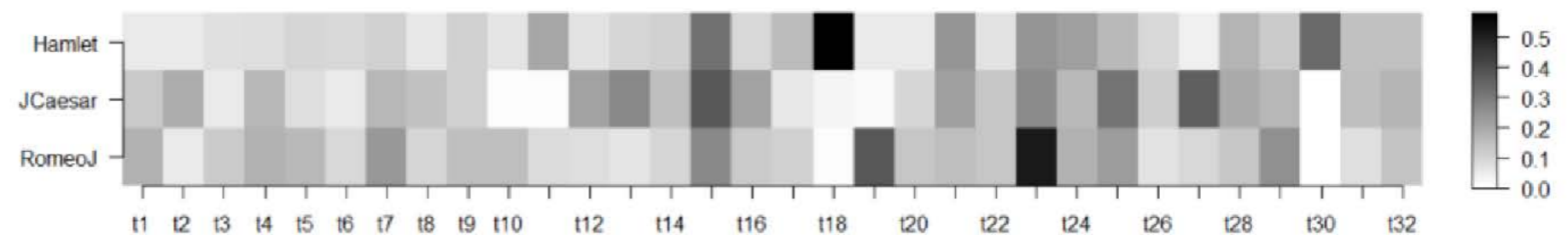


Figure 10: Heatmap of T in 3 tragedies

§ Doc 1: Hamlet (words 16769)

§ Doc 2: Julius Caesar (words 11003)

§ Doc 3: Romeo and Juliet (words 14237)

§ ? number of topics

Example 3

§ NASDAQ docs 2009 - 2017.1

§ Lemmatization, negation

§ ? number of topics

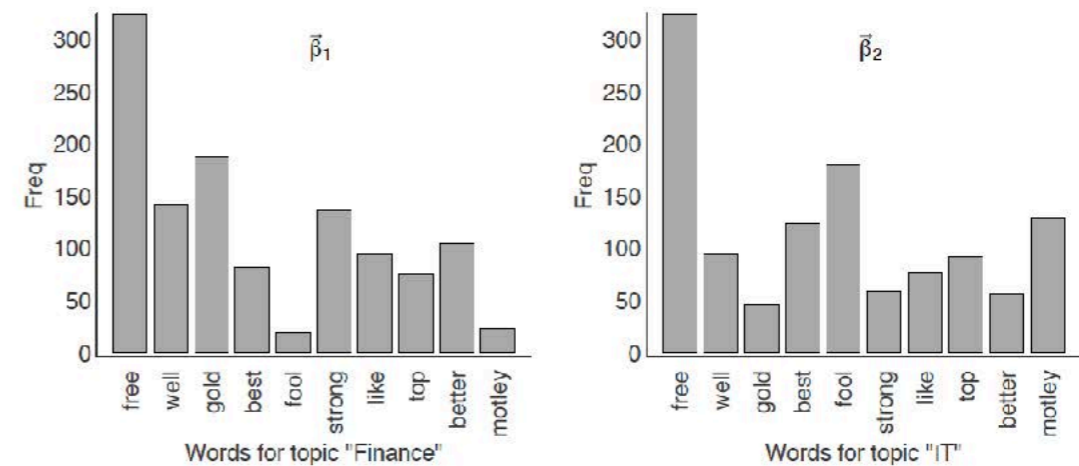


Figure 2: Distribution of words by topic ($\vec{\beta}_1$ and $\vec{\beta}_2$)

XFGdtmWDistr

Word	Freq. (in k)	Freq. for Top 5 Sectors
free	649	10
well	238	9
gold	235	1
best	207	9
fool	200	5
strong	196	5
like	172	5
top	167	3
better	162	0
motley	152	2

Table 1: Most frequent words used in NASDAQ articles

Example 4

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

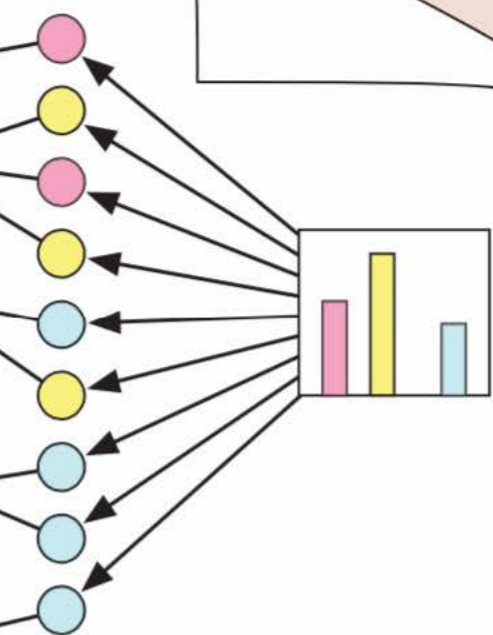
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

ADAPTED FROM NCBI

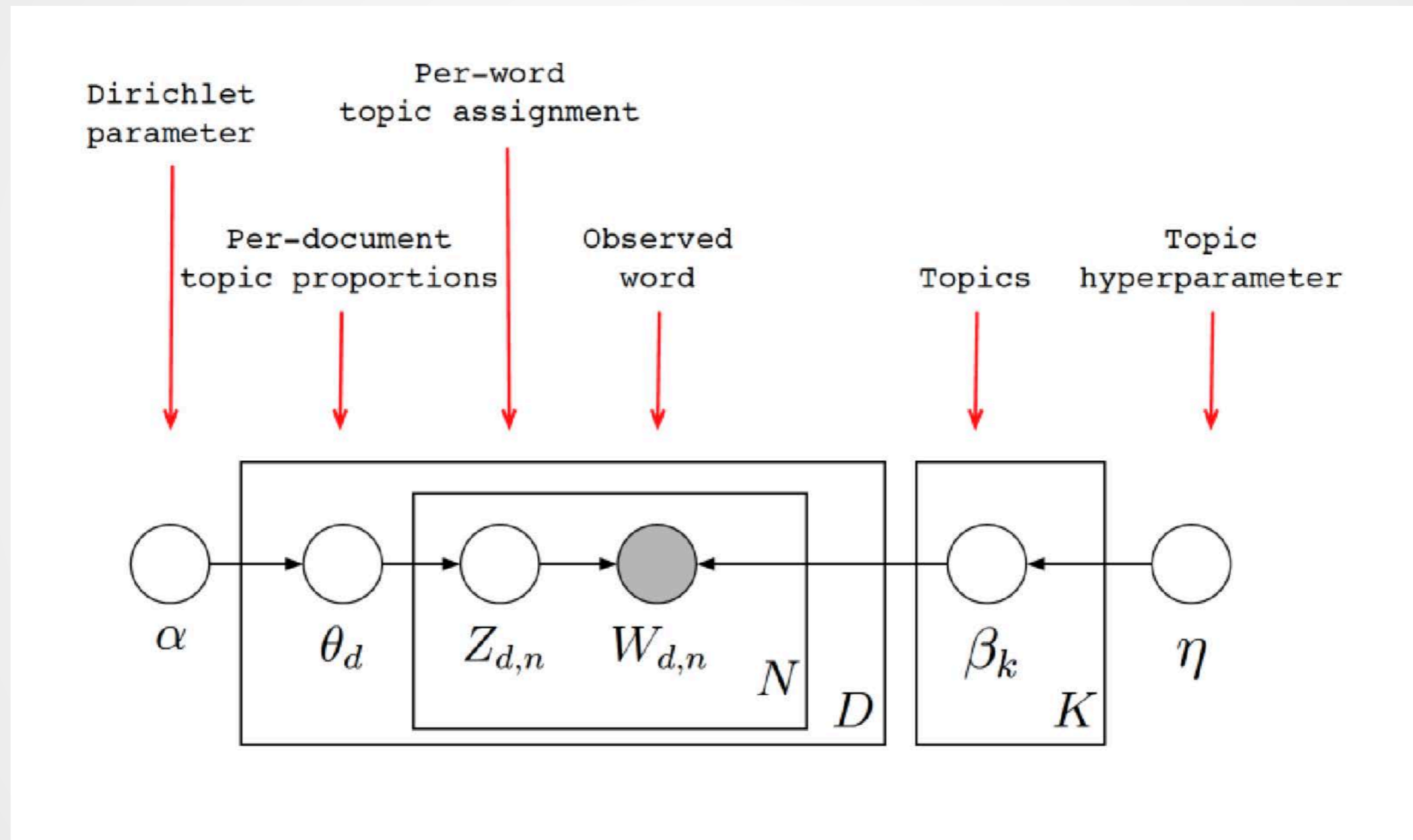
Topic proportions and assignments



§ Number of topics $K = 3$, number of documents $D = 1$

Topic Modelling

§ Schematic design for the document as a „bag of words“



α and η are LDA parameters; for further notation see next slide. Source: D.Blei (2012)

Joint distribution of words and topics

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \omega_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(\omega_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

$\beta_{1:K}$	topics
$\theta_{1:D}$	per doc topics proportion
$z_{1:D}$	per word topic assignment
$\omega_{1:D}$	observed words

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, | \omega_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, | \omega_{1:D})}{p(\omega_{1:D})}$$

Bayes' rule

§ If A and B are events, then Bayes' rule states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Six-sided die-tossing example:

- A: an odd number is rolled
- B: a number less than 4 is rolled
- A^c : an even number is rolled

$$P(A|B) = \frac{\frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}$$

Bayesian machinery

§ For point or interval estimation of a parameter in a model M based on data y ,

Bayesian inference is based on:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$p(\theta)$ is the prior density for the parameter,

$p(\theta|y)$ is the posterior density for the parameter,

$p(y|\theta)$ is the statistical model (or likelihood), and

$p(y)$ is the prior predictive density (or marginal likelihood)

Bayesian machinery

- § Main difference to classical (frequentist) statistics:
- § parameters are rv's and not fixed unknown quantities
- § combines prior information with data, within a decision theoretical framework
- § provides inferences that are conditional on the data and are exact,
- § obeys the likelihood principle

Bayesian machinery

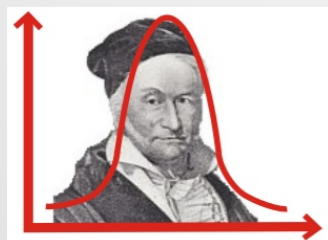
§ Disadvantages

§ selection of prior + posterior depends on priors

§ high computational cost

§ hyperparameters need to be checked

§ requires scenario „robustness tests“



Bayes

Dirichlet distribution

The Dirichlet distribution is defined on a $(k - 1)$ dimensional simplex

$$\Delta_k = \left\{ q \in \mathbb{R}^k : \sum_{i=1}^k q_i = 1, q_i \geq 0, i = 1, 2, \dots, k \right\}. \quad (1)$$

It can be thought of as a distribution of random probability mass/density functions (pdf). An excellent example based introduction can be found in [Frigyik et al. \(2010\)](#).

Definition 1 Let Q be a real value in Δ_k and suppose that $\alpha \in \mathbb{R}^k, \alpha_i > 0$ and define $\alpha_0 \stackrel{\text{def}}{=} \alpha^T \mathbf{1}$. Then Q has a $Dir(\alpha)$ distribution with pdf $f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k q_i^{\alpha_i - 1}$.

$$f(x; a, b) = \frac{\Gamma(a + b)}{\Gamma(a) + \Gamma(b)} x^{a-1} (1 - x)^{b-1}.$$

For $\alpha = (a, b)^T$ with $Q = (X, 1 - X) \sim Dir(\alpha)$ for $X \sim Beta(a, b)$.

Dirichlet distribution

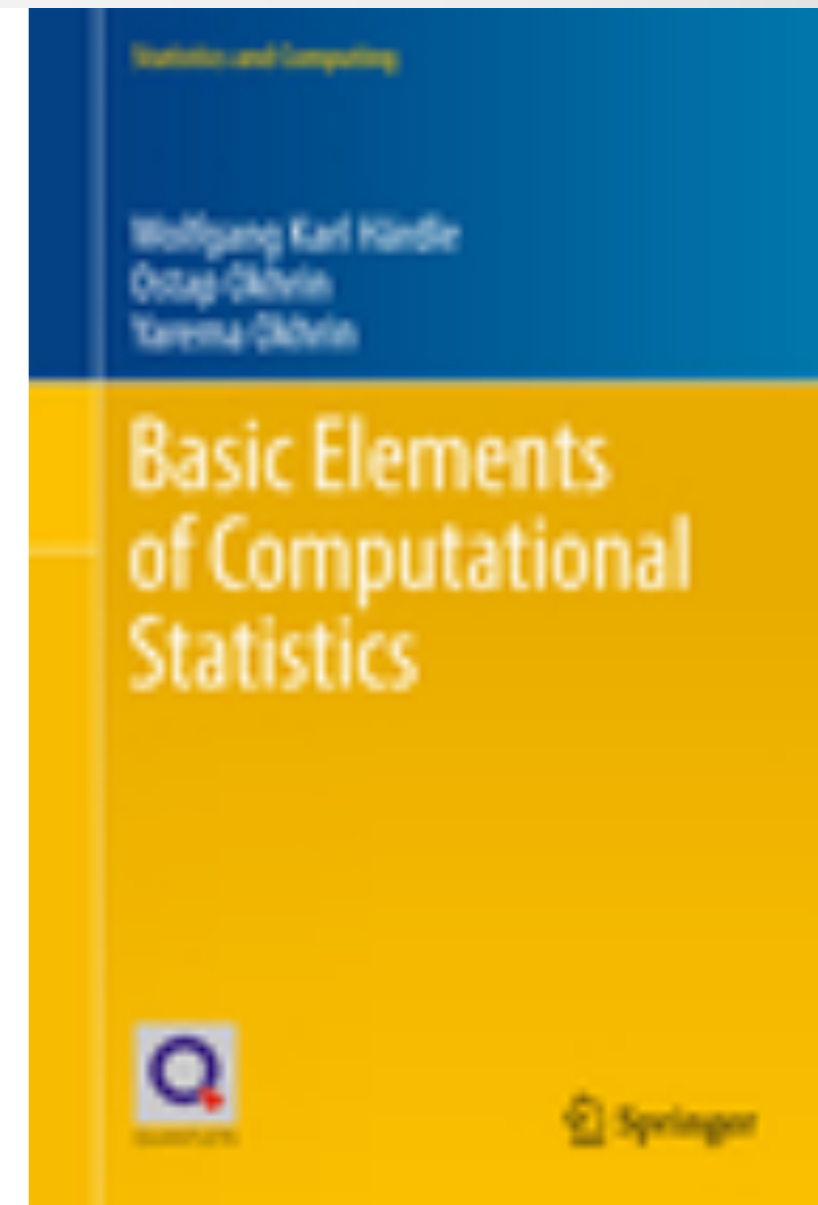
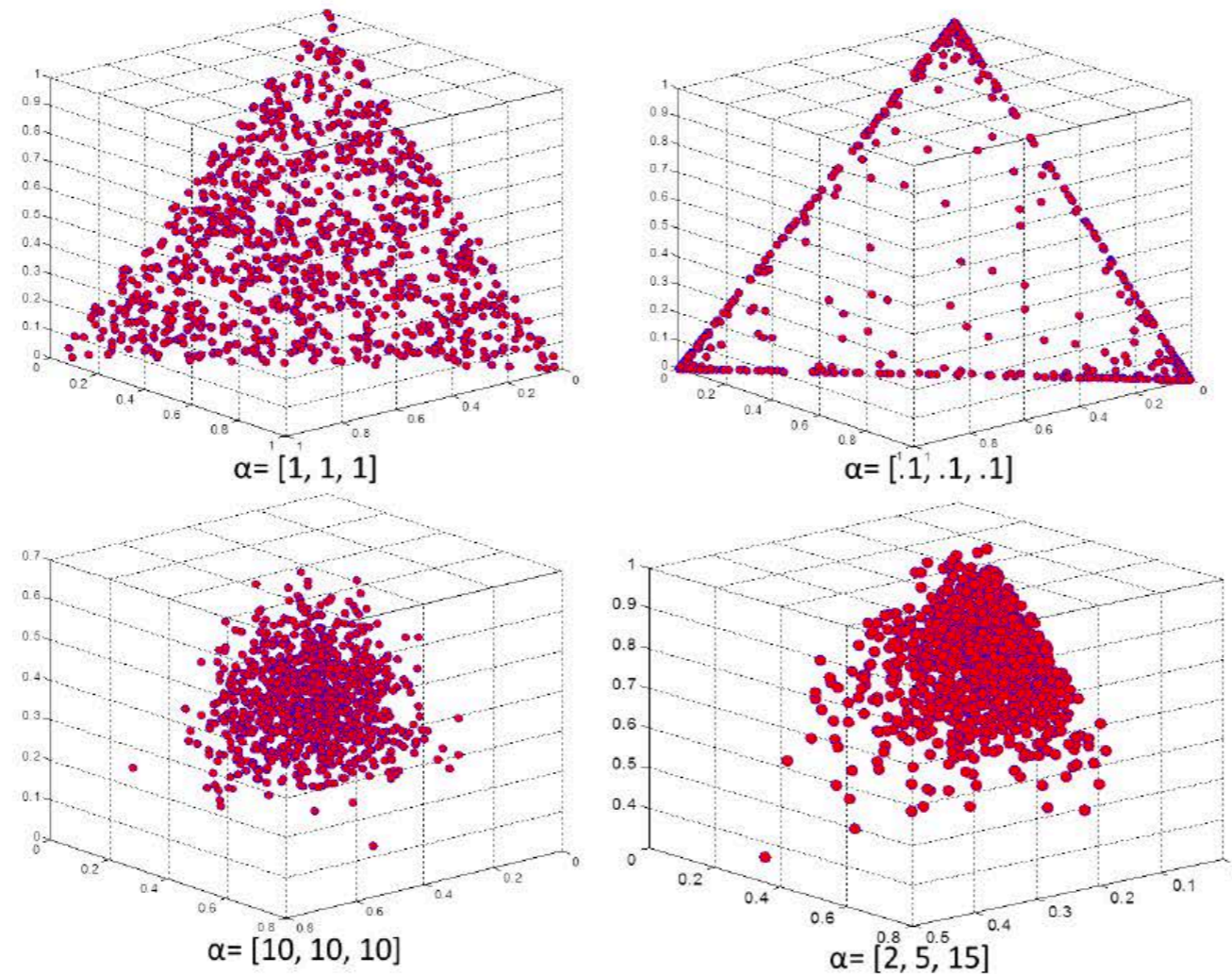


Figure 1.1: Plots of sample pmfs drawn from Dirichlet distributions for various values of α



XFGdtmDirichlet

LDA OK for more than topicing Christmas songs?

- § Dynamic Topic Modelling for Cryptocurrency Community Forums
- § „an indicator for fraudulent schemes constructed using DTM, text mining and unsupervised machine learning“
- § „study how opinions and the evolution of topics are connected with big events in the cryptocurrency universe“

Words are the new numbers

- § VH Larseny & LA Thorsrud (Norges Bank) LDA decompose a business newspaper
- § Simple hypothesis: the more intensive a given topic is represented at a given time, the more likely it is that this topic represents something important for the economy
- § Result: topics have predictive power for key variables and, especially noteworthy, for asset prices

LDA and Topic detection

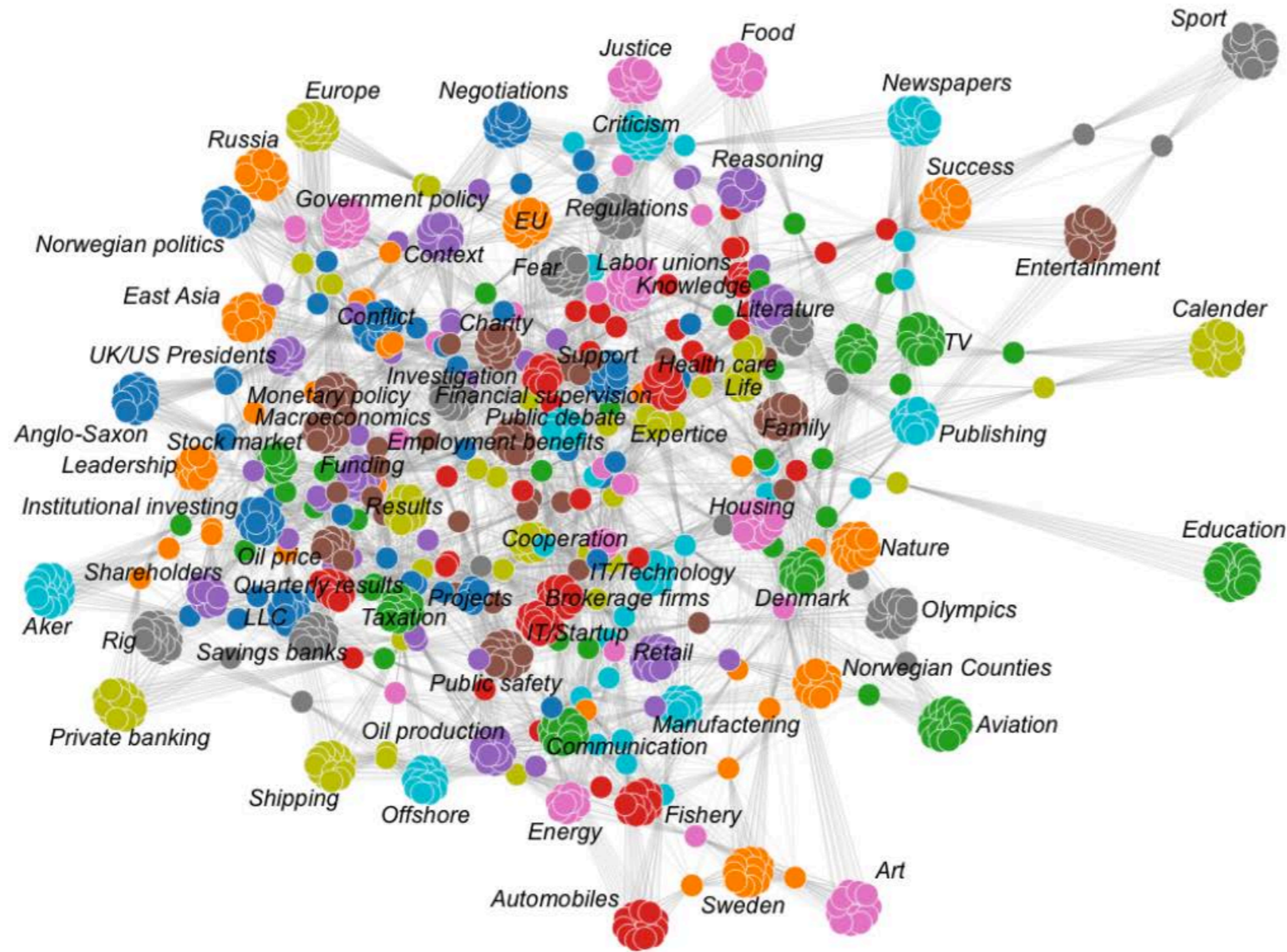


Figure 1. DN visualized using a topic net. Different colors are used for words in different topics. Words that belong to the same topic have an edge between them.

New ideas

- § Use the approach to measure inflation expectations (IE)
- § Do news have predictive power for IE?
- § Use text mining to detect news relevant for inflation expectations, perception of the central bank announcements by the media etc
- § Compare inflation expectation measure based on news with other measures such as T-bond differences etc
- § For options, Chen CYH, Fengler M, Härdle WKH, Liu YC (2017)

How to do LDA dynamically?

- § Make the params dynamic
- § Employ a state space technique

$$\beta_{t,k} | \beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I)$$

$$\alpha_{t,k} | \alpha_{t-1,k} \sim N(\alpha_{t-1,k}, \delta^2 I)$$

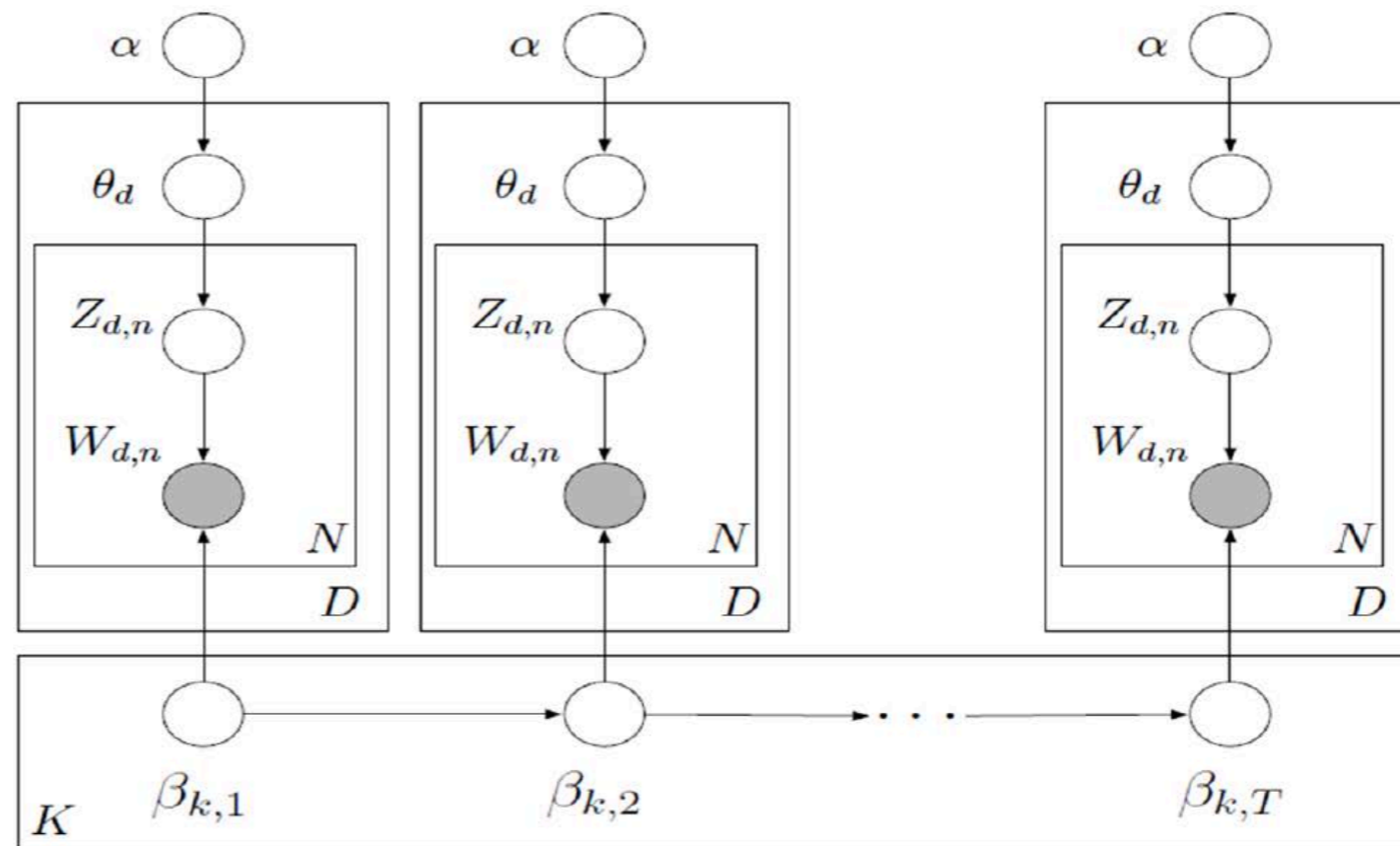


Figure 3: State space diagram of the dynamic topic model

DTM for Bitcoin message boards

- § Observe sub forums Economics, Bitcoin Discussion, Altcoin Discussion, Speculation, Scam Accusations
- § Run with weekly data from 20091122 to 20160806 period
- § Notable Topics:

Topic Number	Most Probable Words
1	value, gold, bar, dollar, rate, demand, interest, asset
2	business, casino, house, trust, gambling, run, strategy, player
5	government, control, criminal, law, study, regulation, state, rule
7	use, service, option, cash, good, spend, fiat, convert
12	account, payment, fund, card, paypal, party, merchant, credit
18	score, online, pay, shop, bill, product, purchase, phone
20	wallet, key, paper, computer, storage, code, data, secure
23	price, trade, market, trader, drop, volume, sell, stock
24	trading, term, hold, buy, pump, dump, earn, gamble
30	exchange, bitfinex, lesson, cryptocurrency, crash, platform, altcoins, popularity
32	investment, risk, invest, aim, impact, salary, making, way
33	year, altcoins, end, today, adoption, prediction, happen, trend
35	transaction, block, fee, chain, confirmation, hour, minute, hardfork
38	altcoin, company, loss, hack, scam, hacker, scammer, road
42	bank, system, security, fiat, banking, role, function, institution
45	ethereum, split, advantage, issue, side, change, fork, core
48	forum, post, topic, member, bitcointalk, thread, index, php
50	mining, miner, network, power, pool, cost, reward, electricity

Table 2: Notable topics from 50 topic model on Bitcoin Discussion subforum from 2016/07/31 to 2016/08/06

DTM for Bitcoin message boards

- § Bitcoins produced by CPU mining: worth less than cost to mine
- § GPU mining came into play 2010
- § ASIC application specific integrated circuits
- § First ASICs („AvalonProject“) in 2012, 2013.1 release of first chip
- § Antminer, a brand of ASICS, current top of the line

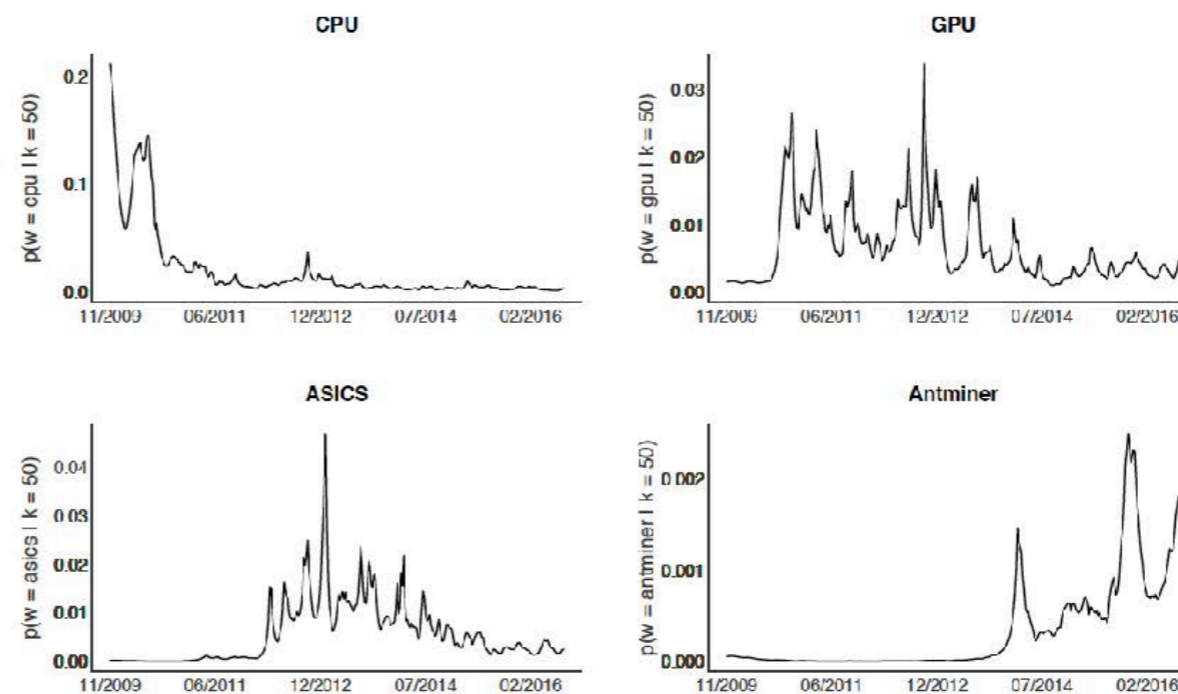


Figure 4: Comparison of word evolution for different mining technologies 22/11/2009 - 06/08/2016

XFGdtmMining



DTM for Bitcoin message boards

- § **The event:** insolvency of the MtGox Bitcoin exchange in 2014.
- § **MtGox** started 2007 trading Magic: The Gathering Online trading cards which is where it got its name (**M**agic: **T**he **G**athering **eX**change).
- § 2010, rebranded as exchange for Bitcoins.
- § MtGoX grew gradually and watched BTC go from USD 0.1 in 2010 to parity with the USD in 2011.
- § Then the owner of MtGox decided to sell the exchange in order to dedicate himself to `other projects'.

DTM for Bitcoin message boards

- § internal email (after the sale) revealed: already 80,000 BTCs (worth over 60K USD) had been missing
- § Three months later a major event occurred: 60,000 accounts were exposed publicly and a compromised MtGox auditors account was used to create huge sell orders and crash the BTC price from 17.51 USD to 0.01 USD.
- § Many of the exposed accounts were used to steal coins from other bitcoin services due to password reuse. However, unlike many other BTC services, MtGox managed to recover its reputation and became the largest BTC exchange, handling 70% of all trades worldwide.

DTM for Bitcoin message boards

- § Fast forward 2013: real problems began, in June withdrawals of USD were suspended and even though a couple of weeks later in July it had been announced that withdrawals had fully resumed, as of September few withdrawals had successfully been completed.
- § Complaints piled up over the next few months and on 7 February 2014 all BTC withdrawals had been suspended for good. On the 24th of February all activities had halted, the website went offline and a leaked internal crisis management document claimed that 744,408 BTC (worth almost half a billion dollars) had been lost and the company was insolvent.

DTM for Bitcoin message boards

§ Topic 23 = Bitcoin trading

§ Topic 38 = scams and hacks

§ Topic 23: peak in 2013.6, transaction issues first occurred; trailing off when MtGox starts to gain momentum in topic 38

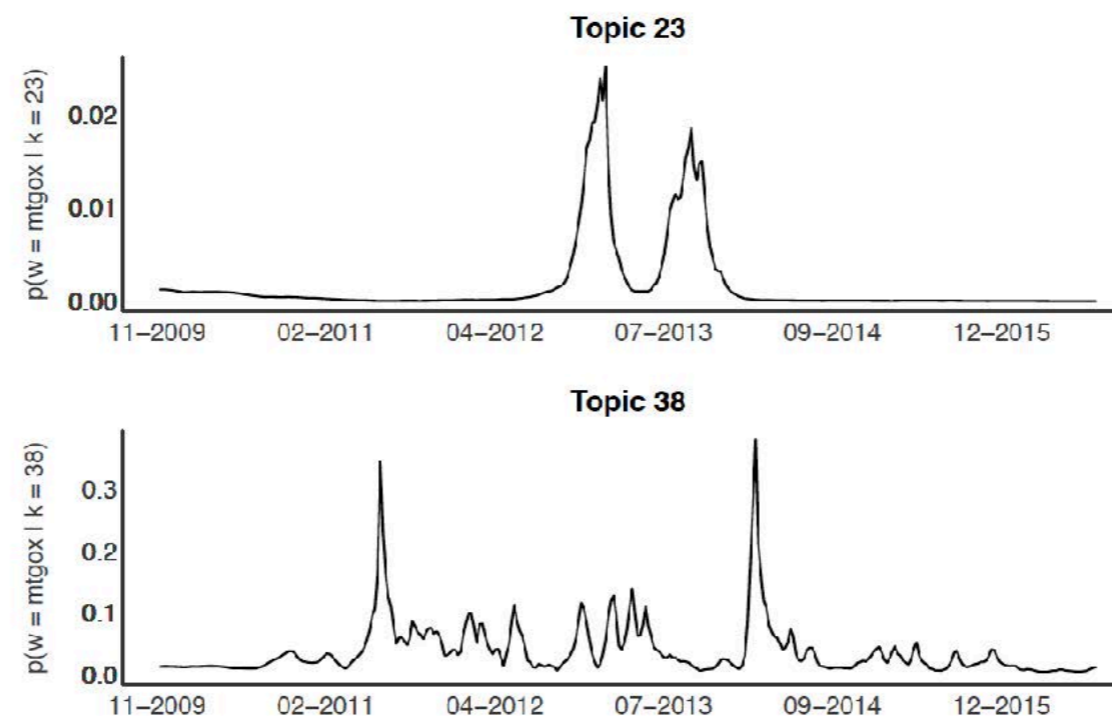


Figure 5: MtGox word evolution 22/11/2009 - 06/08/2016

XFGdtmMtGox



DTM for more events in the Bitcoin universe

§ Evaluate the effectiveness of topic models

§ list of Bitcoin services which have been victims of hacks:

<https://bitcointalk.org/index.php?topic=576337.0>

§ basis for event discovery validation

DTM for more events in the Bitcoin universe

§ List of 37 events on 33 different Bitcoin services. For each word determine which topics the word achieves a topic prominence larger than a low threshold.

§ Topic prominence of the words conditioned on topics through time and determine a significant event.

Event	Dates	Topic
Ubitex* (1,138b)	2011-04 to 2011-07	None
Allinvain	2011-06-13	23
MtGox	2011-06-19	23
Mybitcoin	2011-06-20, 2011-07	23
Bitomat	2011-07-26	23
Mooncoin	2011-09-11	23
Bitscalper	2012-01 to 2012-03	23
Linode	2012-03-01	23
Betcoin* (3,171b)	2012-04-11	None
Bitcoinica	2012-04-12, 2012-07-13	23
Btc-c	2012-07-13	12
Kronos	2012-08	23
Bitcoin Savings and Trusts	2012-08-28	23
Bitfloor	2012-09-04	23
Btcguild* (1,254b)	2013-03-10	None
OkPay (main victim of 2013 Fork)	2013-03-11	30
Ziggap* (1,708b)	2013-02 to 2013-04	None
Just-Dice	2013-07-15	23
Basic-Mining* (2,131b)	2013-10	None
Silkroad2	2013-10-02	23
Vircurex* (1,454b)	2013-10-05	None
GBL	2013-10-26	12
Bips* (1,294b)	2013-11-17	None
Picostocks* (5,896b)	2013-11-29	None
MtGox	2014-02-24	23
Flexcoin	2014-03-02	23
Cryptorush	2014-03-11	23
Mintpal	2014-10-14	23
Silkroad2	2014-11-06	23
Bitstamp	2015-01-04	23, 25
Bter	2015-02-14	23
Cryptsy	2016-01-01	23
Shapeshift	2016-04	23
Gatecoin*	2016-05-13	None
Bitfinex	2016-08-03	12

Wrapping up

- § A good tool for dealing with unstructured dark data
- § Not always possible to compute, so use the best approximations
- § Limitations: order of words („bag“ of words), order of documents, correlation

What have you seen?

- § An exploration of a popular cryptocurrency forum
- § Captured the effect of high profile scamming and hacking
- § # topics parameter optimal for event detection
- § An RDC dataset has been created from user posts on bitcointalk.org by using web scraping
- § In addition, the constructed model partitions almost all of the events above a certain severity in a single topic.

Dynamic Topic Modeling for Bitcoin Message fora

Marco Linton

Ernie Gin Swee Teo

Elisabeth Bommers

Wolfgang K. Härdle

Cathy Yi-Hsuan Chen

Ladislaus von Bortkiewicz Chair of Statistics

Sim Kee Boon Institute for Financial Economics

International Research Training Group

Humboldt-Universität zu Berlin

lvb.wiwi.hu-berlin.de

www.case.hu-berlin.de

irtg1792.hu-berlin.de

