Thank you!

C. Faloutsos
CMU

Large Graph Mining

C. Faloutsos

CMU

~~Large Graph Mining~~
Data Mining for fun (and profit)

C. Faloutsos

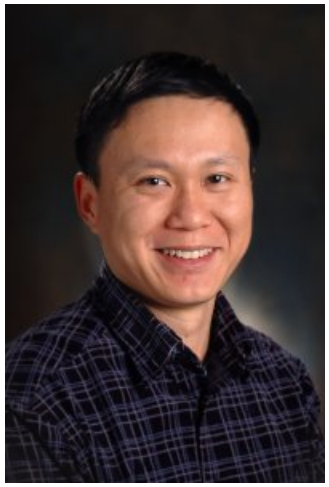
CMU

Outline

- Credit where credit is due
- Technical part – Data mining
  - Can it be automated?
  - Research challenges
- Non-technical part: `Listen'
  - To the data
  - To non-experts

Nominator

- Jian Pei



Endorsers

- Charu C. Aggarwal (IBM Research)
- Ricardo Baeza-Yates (Yahoo! Research)
- Albert-Laszlo Barabasi (Northeastern University)
- Denilson Barbosa (University of Alberta)
- Yixin Chen (Washington University at St. Louis)

Endorsers, cont'd

- William Cohen (Carnegie Mellon University)
- Diane J. Cook (Washington State University)
- Gautam Das (University of Texas at Arlington)
- Inderjit S. Dhillon (University of Texas at Austin)
- Chris H. Q. Ding (University of Texas at Arlington)

Endorsers, cont'd

- Petros Drineas (Rensselaer Polytechnic Institute)
- Tina Eliassi-Rad (Lawrence Livermore National Laboratory)
- Greg Ganger (Carnegie Mellon University)
- Minos Garofalakis (Technical University of Crete)
- James Garrett (Carnegie Mellon University)

Endorsers, cont'd

- Dimitrios Gunopulos (University of Athens)
- Xiaofei He (Zhejiang University)
- Panagiotis G. Ipeirotis (New York University)
- Eamonn Keogh (UCR)
- Hiroyuki Kitagawa (University of Tsukuba)
- Tamara Kolda (Sandia Nat. Labs)

Endorsers, cont'd

- Flip Korn (AT&T Research)
- Nick Koudas (University of Toronto)
- Hans-Peter Kriegel
- Ravi Kumar (Yahoo! Research)
- Laks Lakshmanan (UBC)
- Jure Leskovec (Stanford University)

Endorsers, cont'd

- Nikos Mamoulis (Hong Kong University)
- Heikki Manilla (Aalto University,
- Dharmendra S. Modha (IBM Research)
- Mario Nascimento (University of Alberta)
- Jennifer Neville (Purdue University)
- Beng Chin Ooi (National University of Singapore)

Endorsers, cont'd

- Dimitris Papadias (Hong Kong University of Science and Technology)
- Spiros Papadimitriou (IBM Research)
- Jian Pei (Simon Fraser University)
- Foster Provost (New York University)
- Oliver Schulte (Simon Fraser University)
- Dennis Shasha (New York University)
- Srinivasan Parthasarathy (OSU)

Endorsers, cont'd

- Jimeng Sun (IBM Research)
- Dacheng Tao (Nanyang University of Technology)
- Yufei Tao (The Chinese University of Hong Kong)
- Evimaria Terzi (Boston University)
- Alex Thomo (University of Victoria)
- Andrew Tomkins (Google Research)

Endorsers, cont'd

- Caetano Traina (University of Sao Paulo)
- Vassilis Tsotras (University of California, Riverside)
- Alex Tuzhilin (New York University)
- Haixun Wang (Microsoft Research)

Endorsers, cont'd

- Wei Wang (University of North Carolina at Chapel Hill)
- Philip S. Yu (University of Illinois, Chicago)
- Zhongfei Zhang (Binghamton University, State University of New York)

KDD committee

- Ramasamy Uthurusamy, Chair
- Robert Grossman (University of Illinois at Chicago)
- Jiawei Han (University of Illinois at Urbana-Champaign)
- Tom Mitchell (Carnegie Mellon University)
- Gregory Piatetsky-Shapiro (KDnuggets)

KDD committee cnt'd

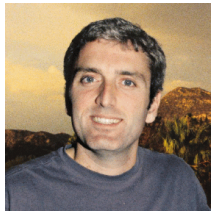
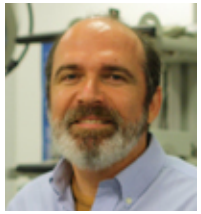
- Raghu Ramakrishnan (Yahoo! Research)
- Sunita Sarawagi (Indian Institute of Technology, Bombay)
- Padhraic Smyth (University of California at Irvine)
- Ramakrishnan Srikant (Google Research)

KDD committee cnt'd

- Xindong Wu (University of Vermont)
- Mohammed J. Zaki (Rensselaer Polytechnic Institute)

Family

- Parents Nikos & Sophia
- Siblings Michalis*, Petros*, Maria


- Wife Christina#

(*) : and co-authors

(#) : and research impact evaluator ('grandpa' test - see later…)

Academic ‘parents’

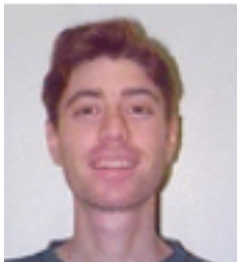
- Christodoulakis, Stavros (T.U.C.)

- Sevcik, Ken (U of T)

- Roussopoulos, Nick (UMD)

Academic 'children'

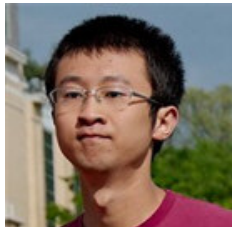
- King-Ip (David) Lin →
- ← Ibrahim Kamel
- Flip Korn →
- ← Byoung-Kee Yi
- Leejay Wu →
- ← Deepayan Chakrabarti



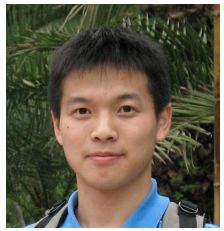
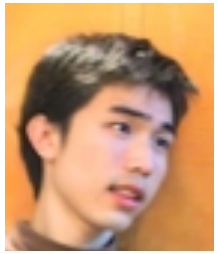
Academic 'children'

- Jia-Yu (Tim) Pan →
- ← Spiros Papadimitriou
- Jimeng Sun →
- ← Jure Leskovec
- Hanghang Tong →

Academic 'children'



- Mary McGlohon →
- ← Fan Guo
- Lei Li →
- ← Leman Akoglu
- Dueng Horng (Polo) Chau →
- ← Aditya Prakash
- U Kang →


CMU colleagues

- Tom Mitchell
- Garth Gibson
- Greg Ganger
- M. (Satya) Satyanarayanan
- Howard Wactlar
- Jeannette Wing
- + +

Co-authors

- [dblp 7/2010:] All 300 of you

- Agma J. M. Traina (22)
- Caetano Traina Jr. (20)
- …

Funding agencies

- NSF (Maria Zemankova, Frank Olken, ++)
- DARPA, LLNL, PITA
- IBM, MS, HP, INTEL, Y!, Google, Symantec, Sony, Fujitsu, …

Outline

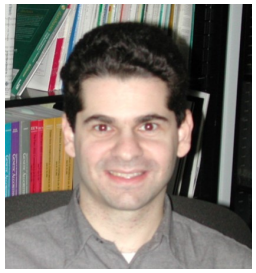
- Credit where credit is due
- Technical part – Data mining
  - Can it be automated?
  - Research challenges
- Non-technical part: `Listen'
  - To the data
  - To non-experts

Data mining = compression & …

$$1111\ldots1111 \rightarrow 1^{100}$$

$$1010\ldots1010 \rightarrow (10)^{50}$$

$$11.00100100\ldots \rightarrow$$



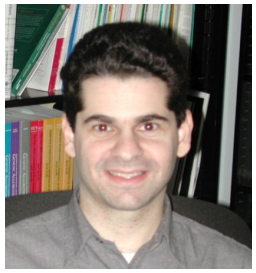
Christos Faloutsos, Vasileios Megalooikonomou: *On data mining, compression, and Kolmogorov complexity*. Data Min. Knowl. Discov. 15(1): 3-20 (2007)

Data mining = compression & …

$$1111\ldots1111 \rightarrow 1^{100}$$

$$1010\ldots1010 \rightarrow (10)^{50}$$

$$11.00100100\ldots \rightarrow 3.14159265\ldots \approx \pi$$



Christos Faloutsos, Vasileios Megalooikonomou: *On data mining, compression, and Kolmogorov complexity*. Data Min. Knowl. Discov. 15(1): 3-20 (2007)

Data mining = compression & …

$$1111\ldots1111 \rightarrow 1^{100}$$

$$1010\ldots1010 \rightarrow (10)^{50}$$

$$11.00100100\ldots \rightarrow 3.14159265\ldots \approx \pi$$

But: how can compression

- do forecasting?
- spot outliers?
- do classification?

Data mining = compression & …

$$1111\ldots1111 \rightarrow 1^{100}$$

$$1010\ldots1010 \rightarrow (10)^{50}$$

$$11.00100100\ldots \rightarrow 3.14159265\ldots \approx \pi$$

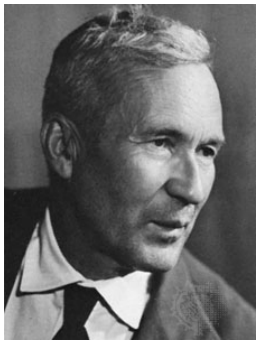
OK – then, isn't compression a solved problem (gzip, LZ)?

… compression is undecidable!

Theorem*: for an arbitrary string *x*, computing its Kolmogorov complexity *K(x)* is undecidable

↑

EVEN WORSE than NP-hard!



A.N. Kolmogorov

(*) E.g., [T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons,1991, section 7.7]

… compression is undecidable!

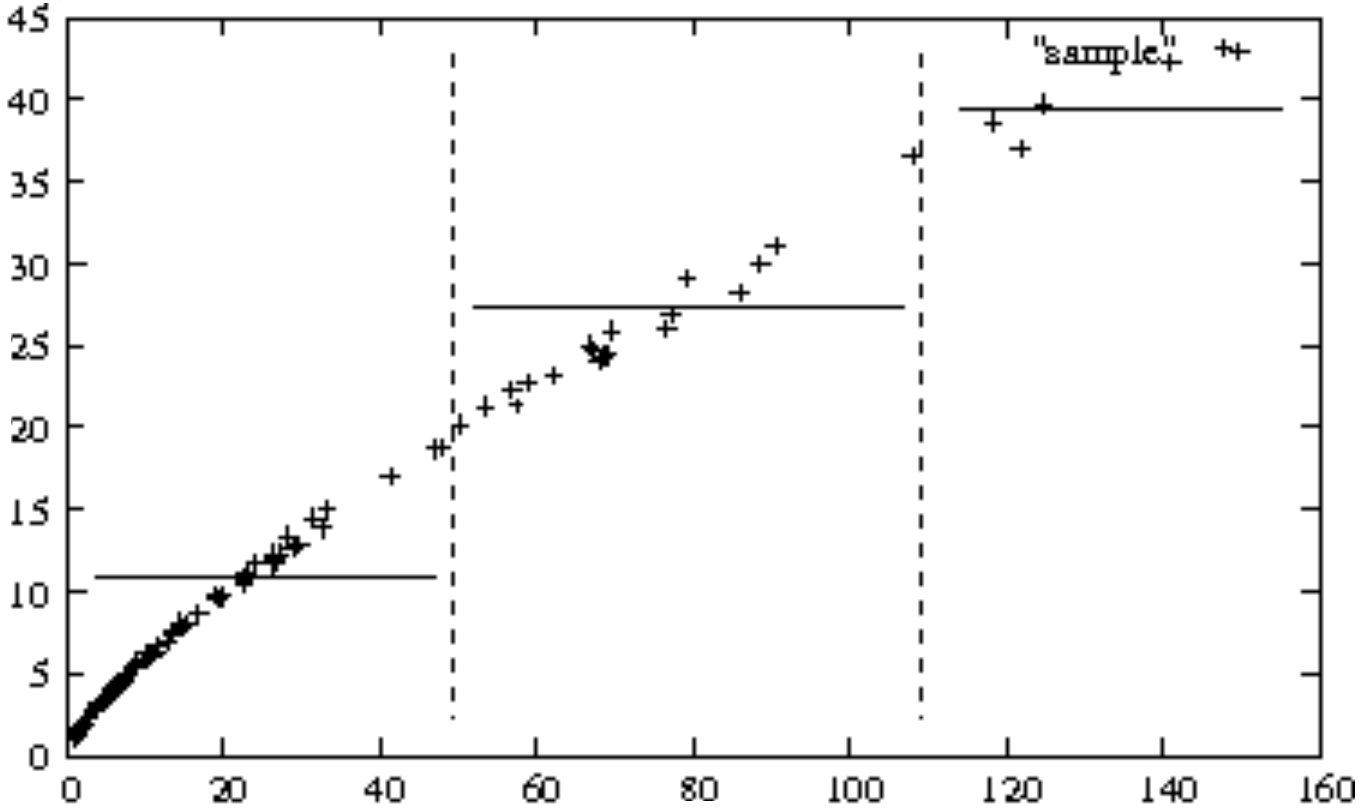
…which means there will always be better data mining tools/models/patterns to be discovered

-> job security ☺

-> job satisfaction ☺

Let's see some examples of models

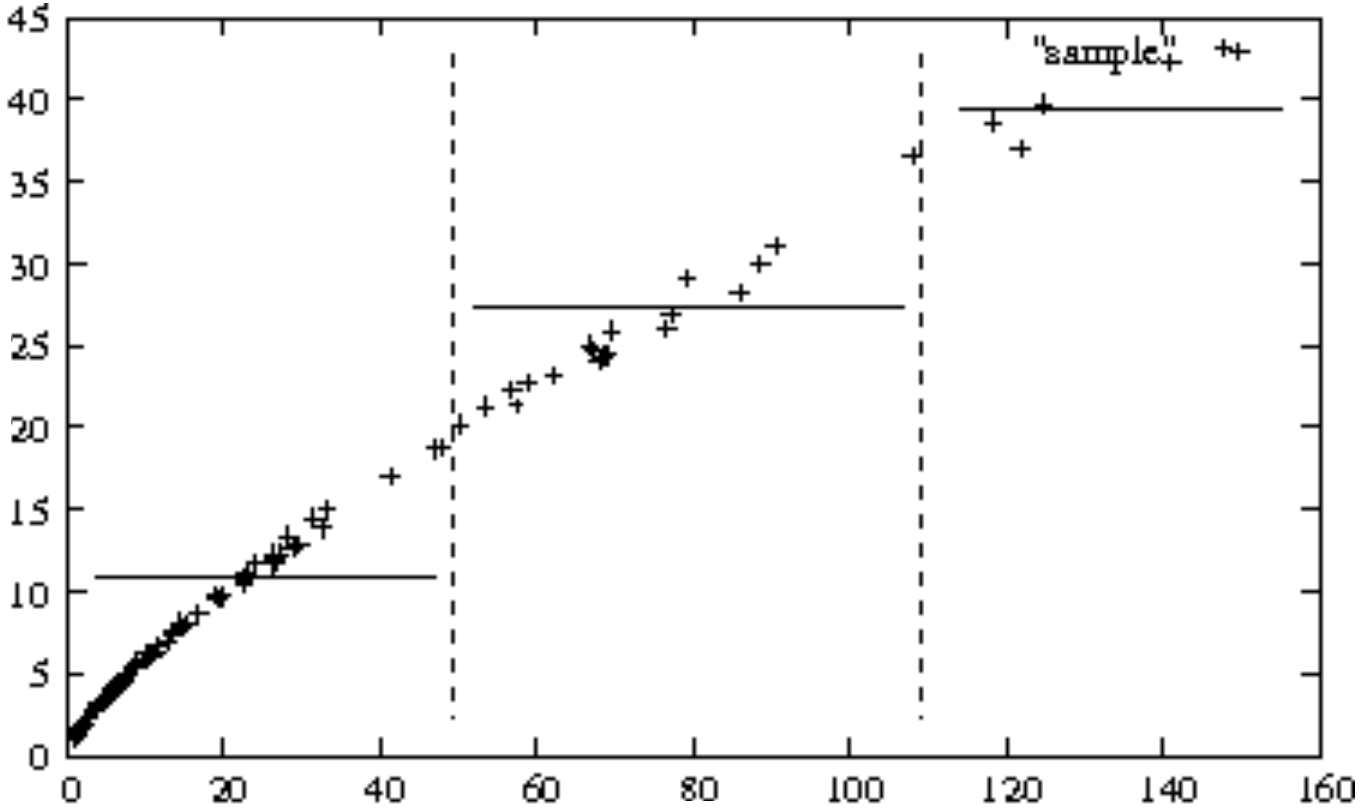
Response to new drug



Body weight

Let's see some examples of models

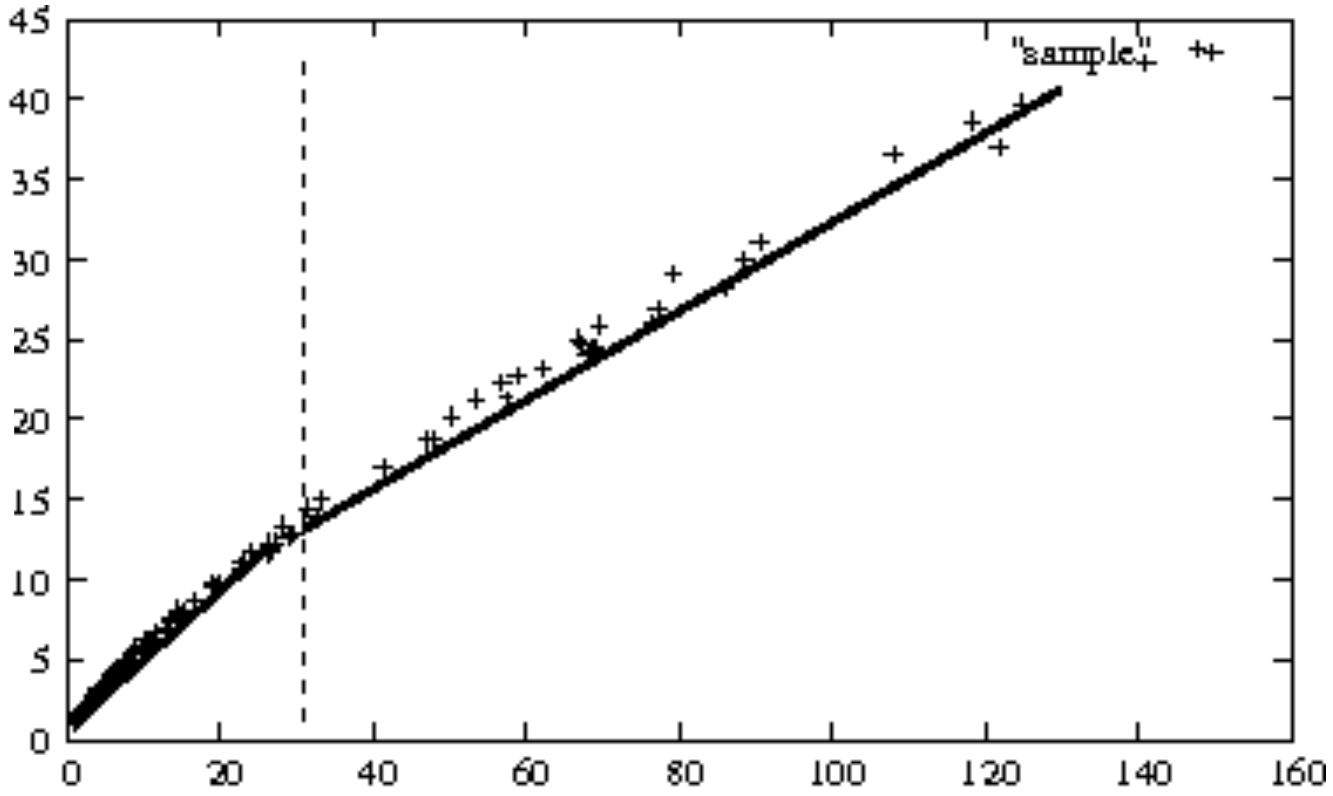
$ spent



income

Let's see some examples of models

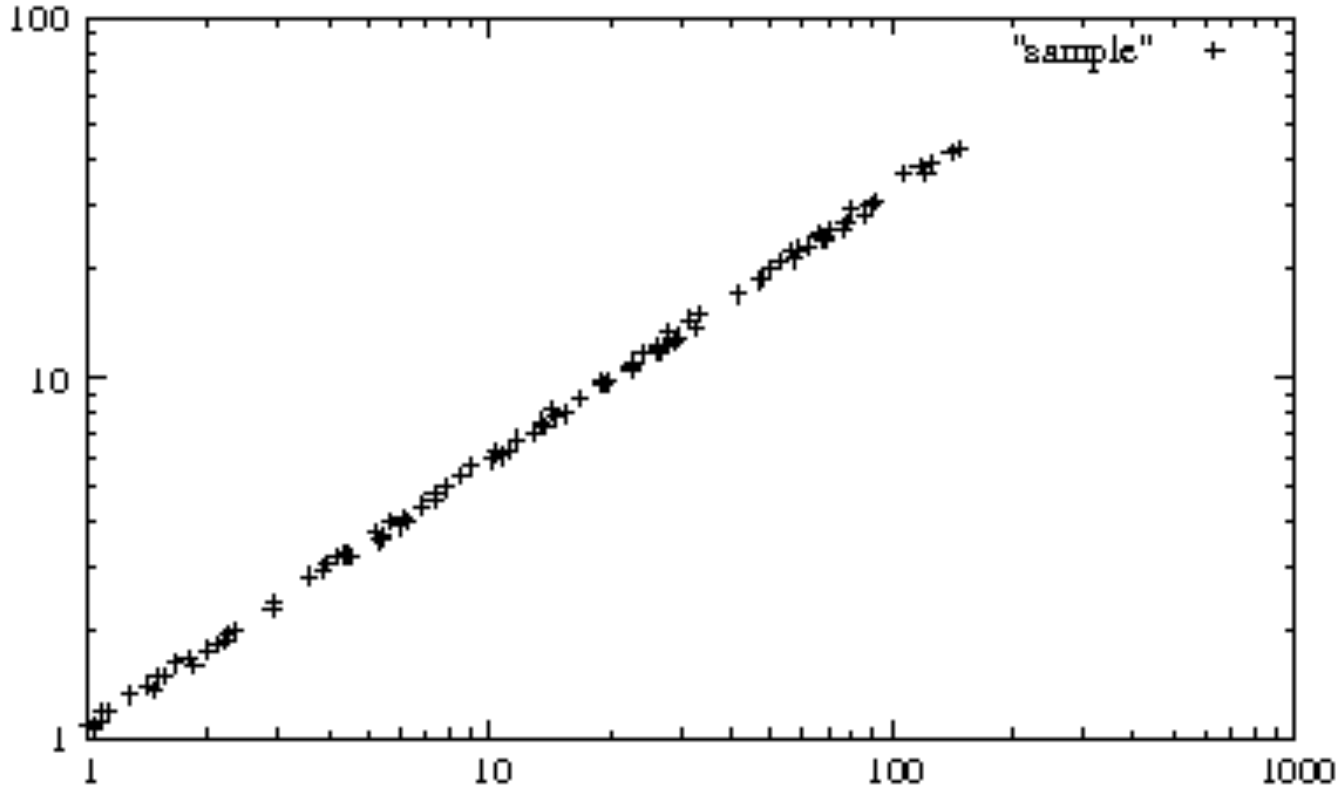
$ spent



income

Let's see some examples of models

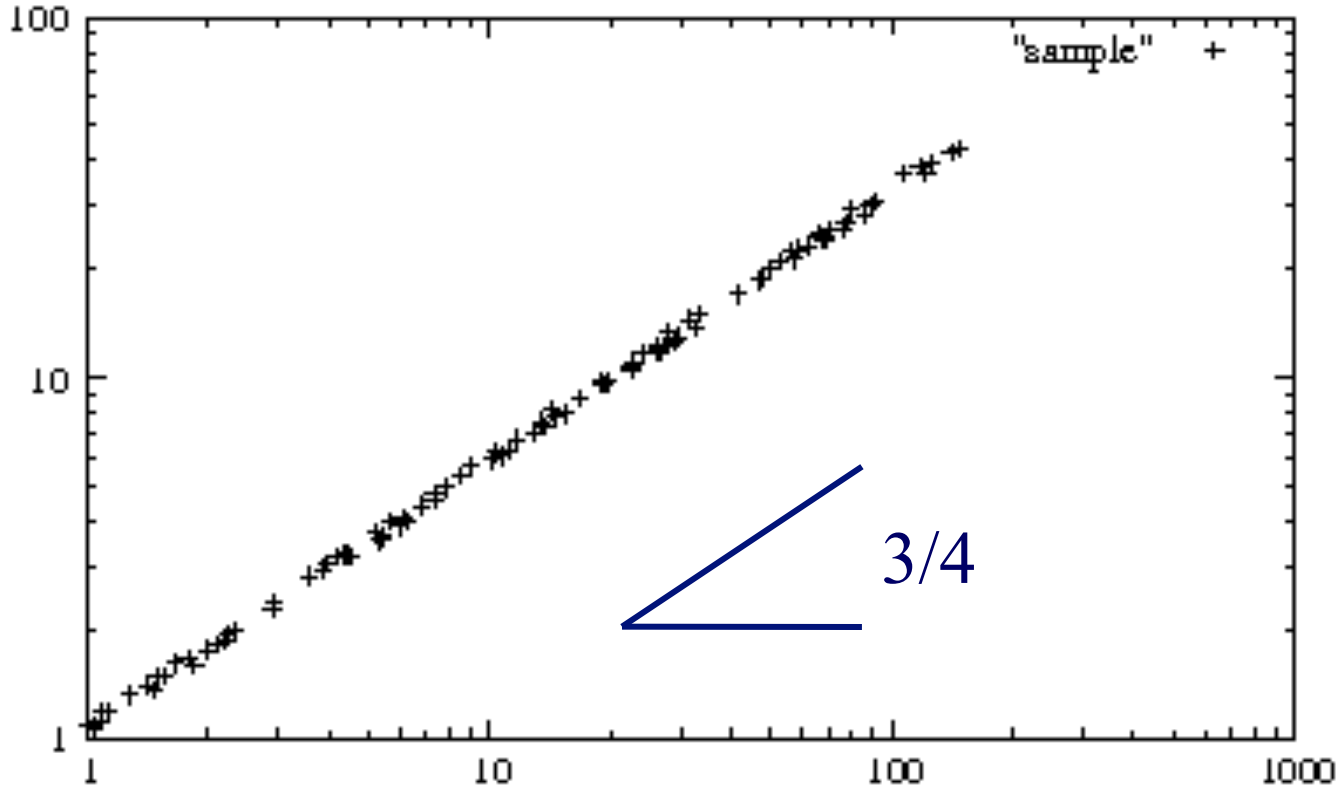
$ spent



income

Let's see some examples of models

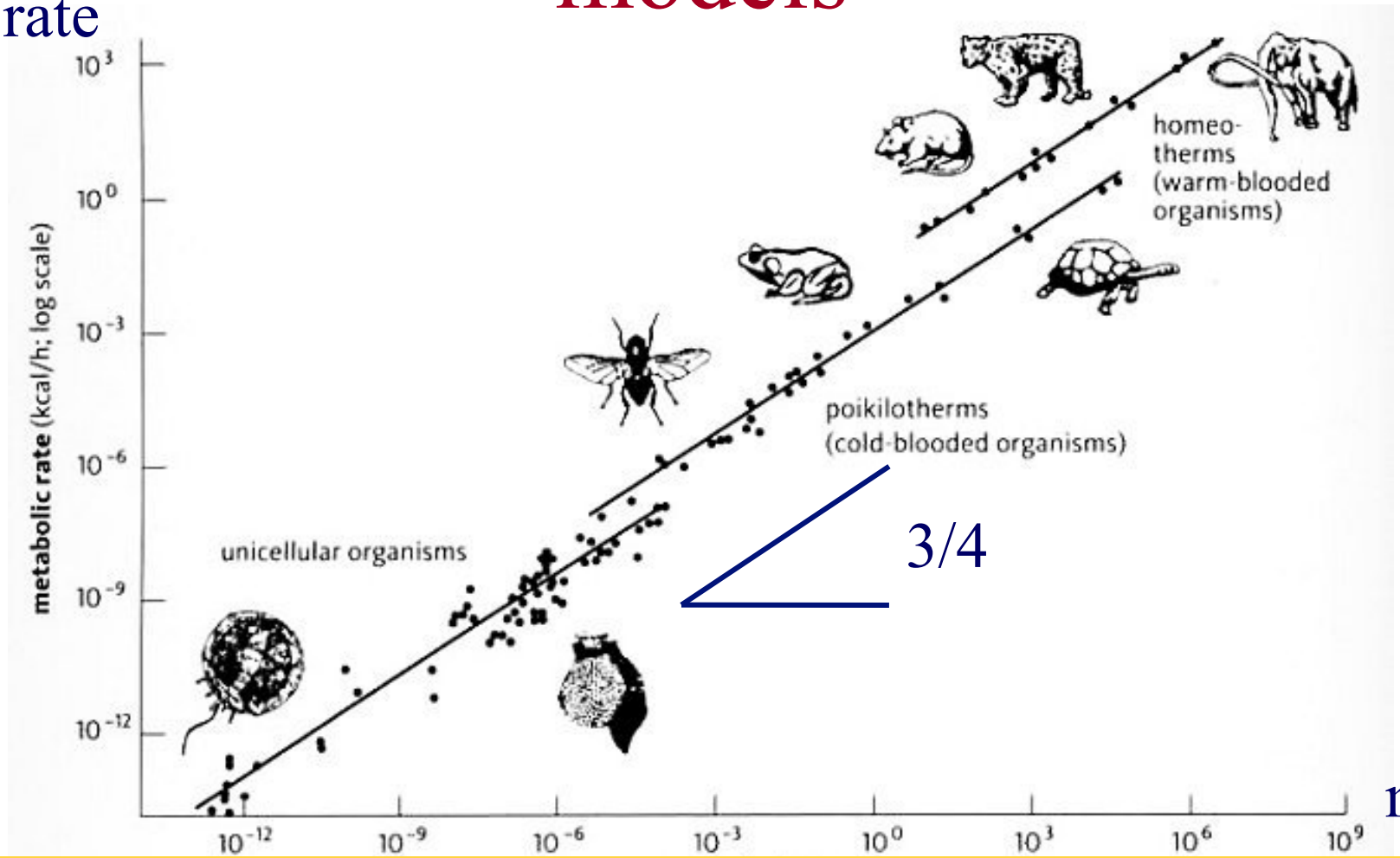
$ spent



income

Let's see some examples of models

Metabolic rate

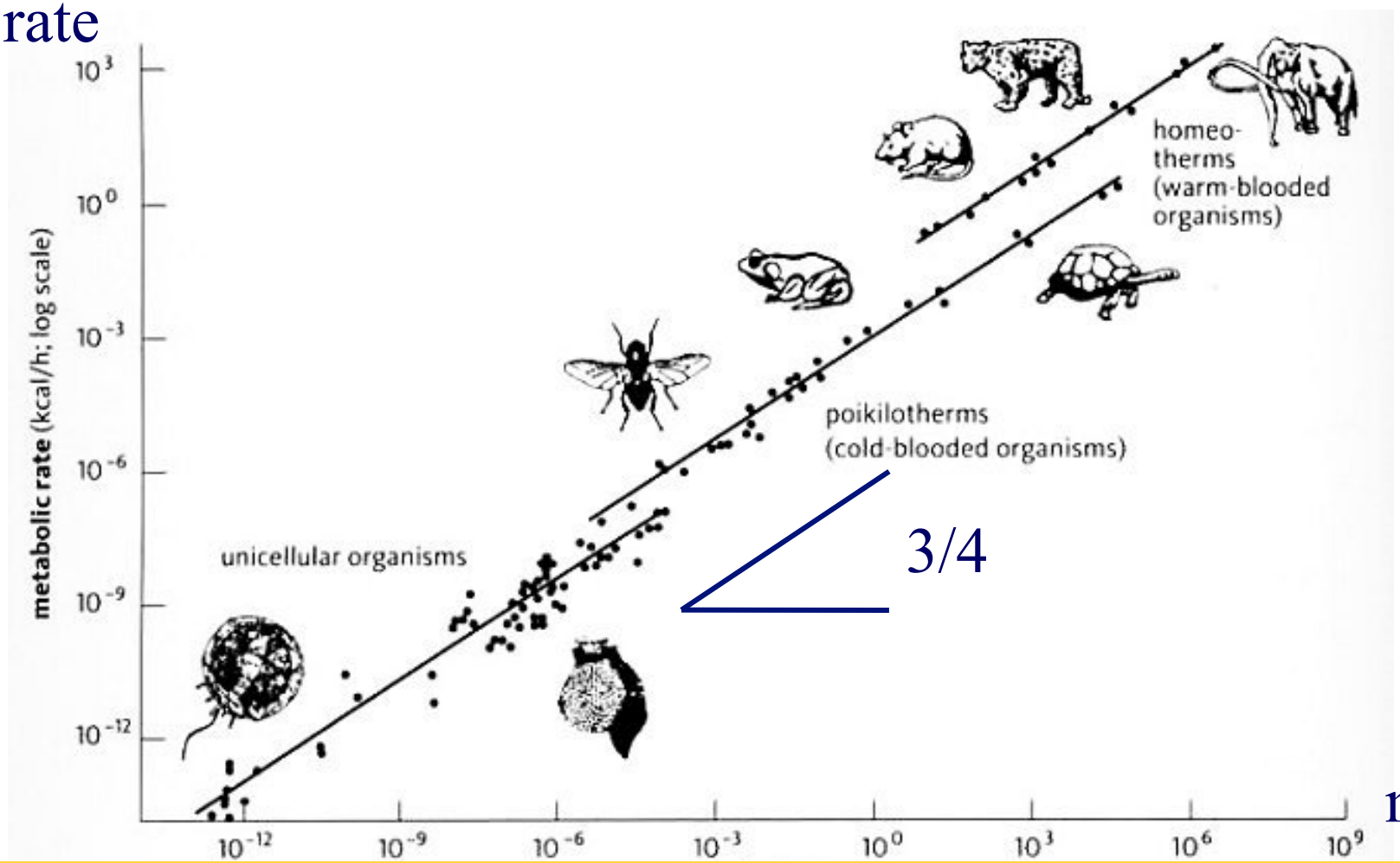


mass

http://universe-review.ca /R10-35-metabolic.htm

Kleiberg's law

Metabolic rate



mass

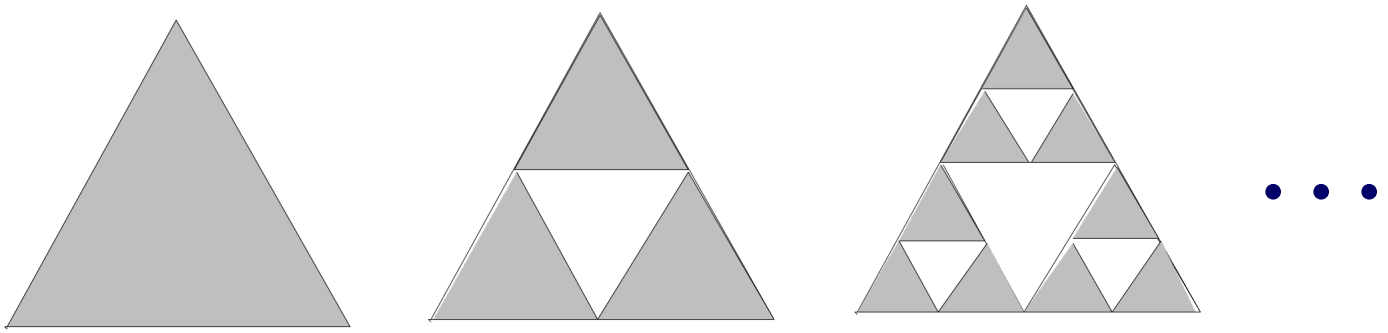
http://universe-review.ca /R10-35-metabolic.htm

Outline

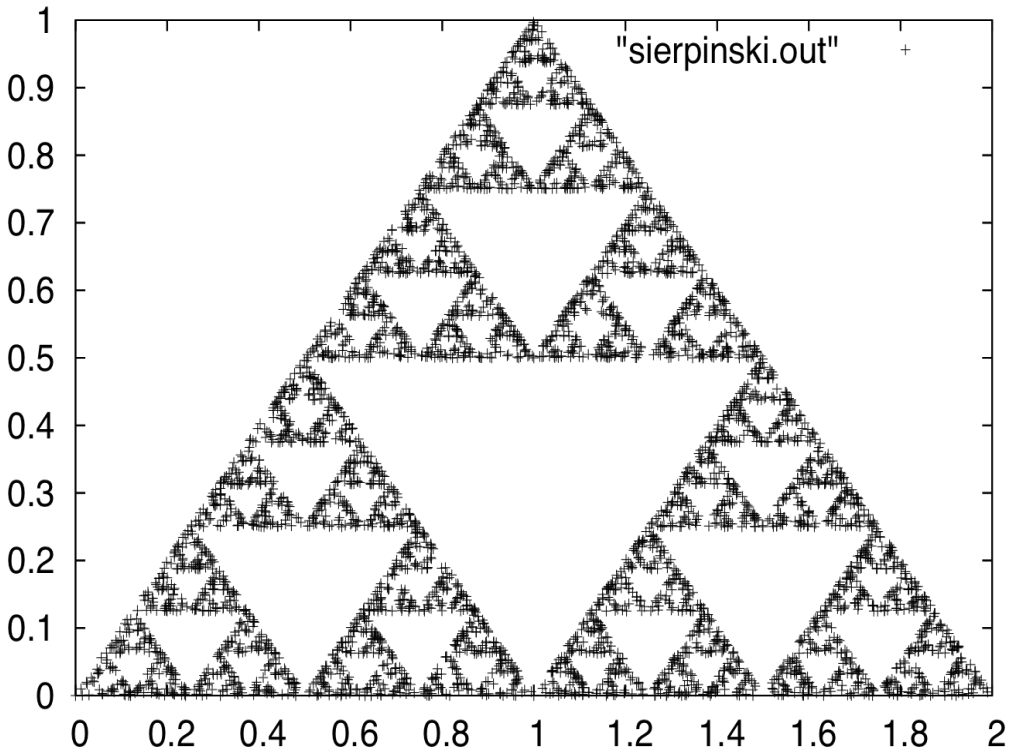
- Credit where credit is due
- Technical part – Data mining
  - Can it be automated? NO!
    - Always room for better models
  - Research challenges
- Non-technical part: `Listen'
  - To the data
  - To non-experts

Always room for better models

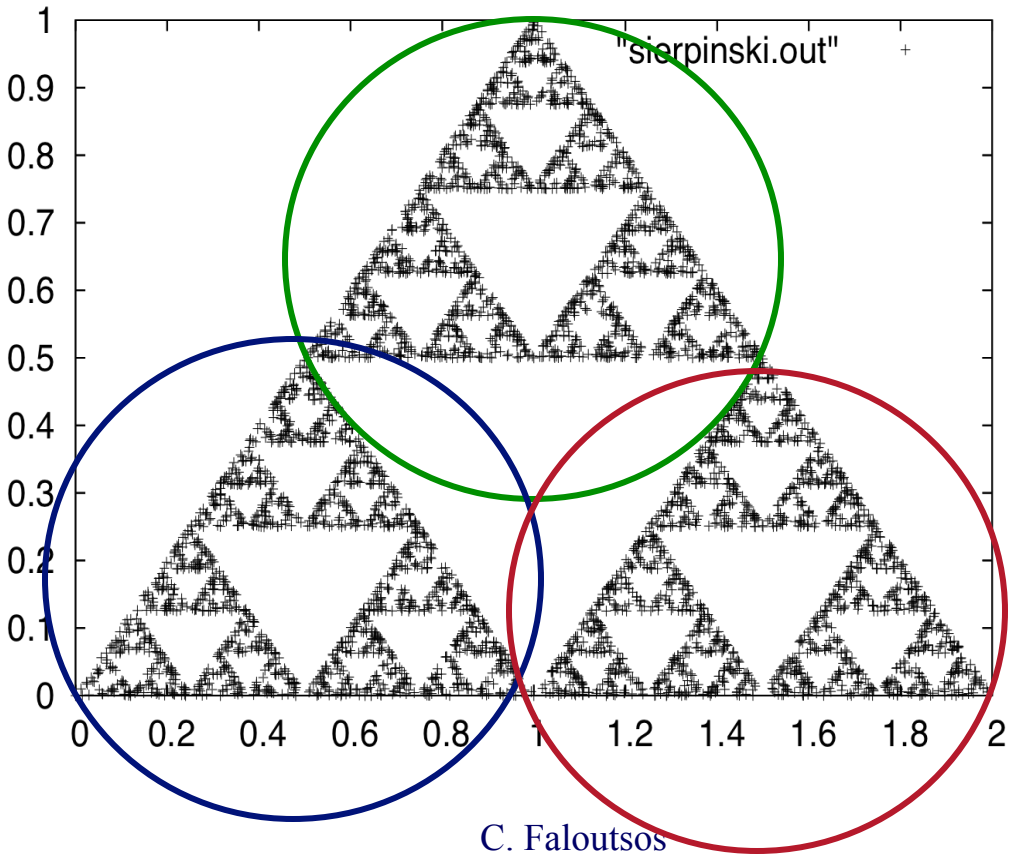
- Eg.: clustering – k-means (or our favorite clustering algo)
- How many clusters are in the Sierpinski triangle?



Always room for better models

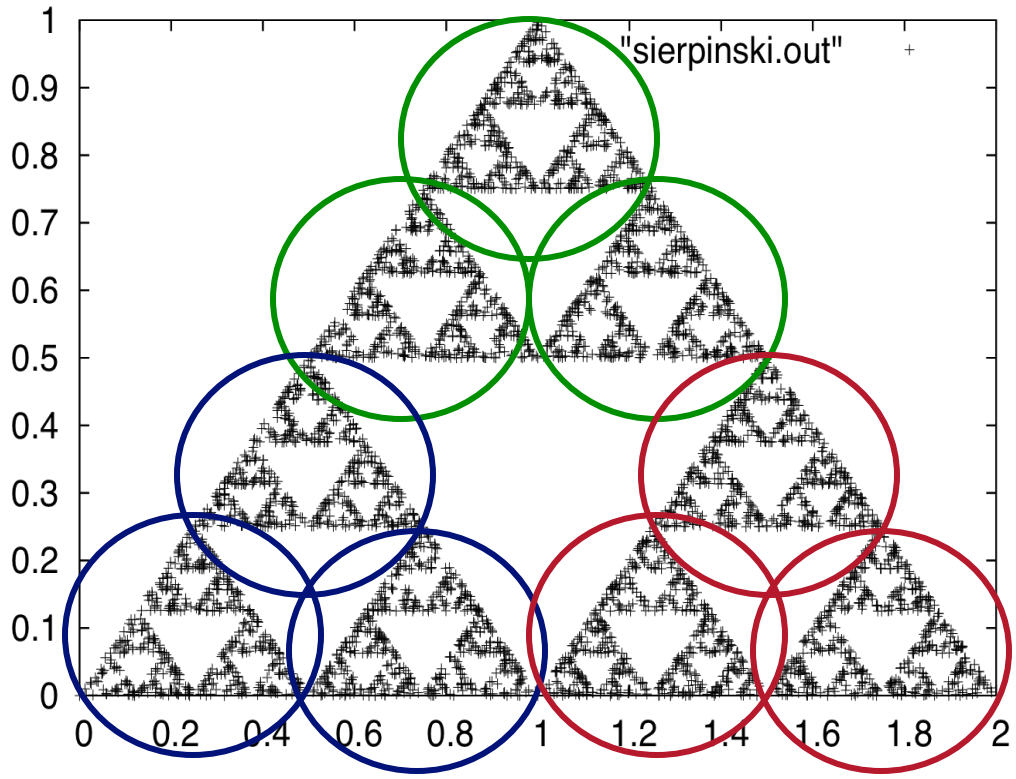


Always room for better models



K=3 clusters?

Always room for better models

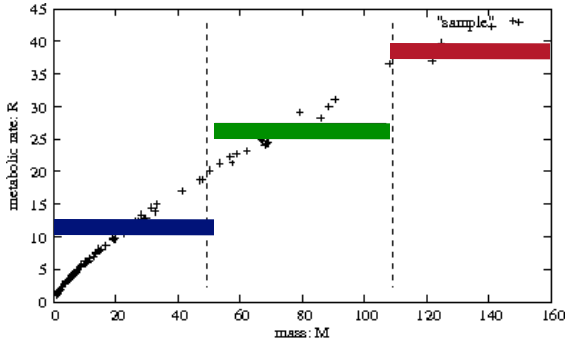


K=3 clusters?

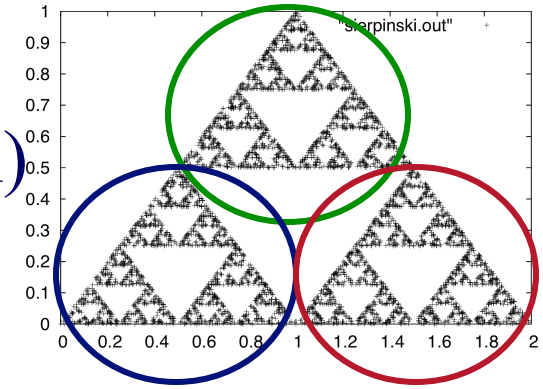
K=9 clusters?

Always room for better models

Piece-wise flat

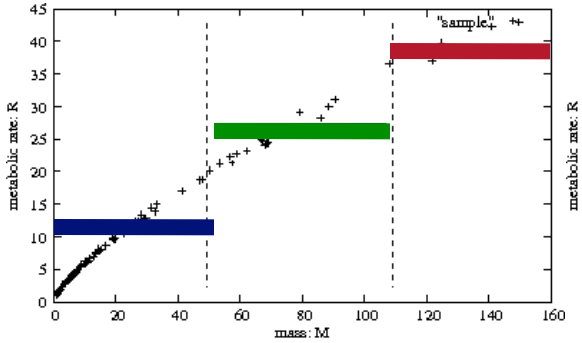


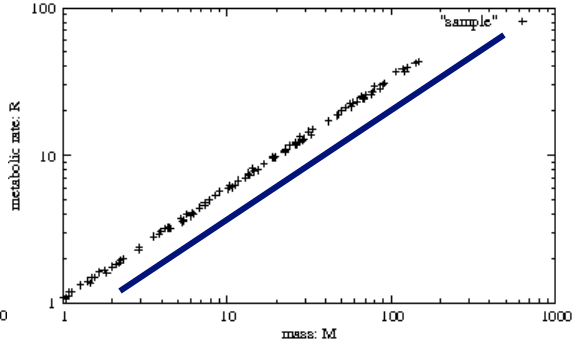
Mixture of (Gaussian) clusters



Always room for better models

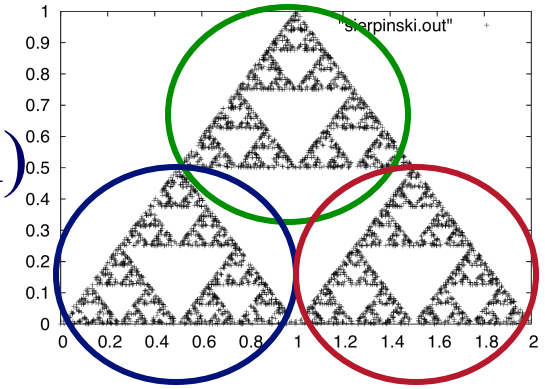
Piece-wise flat

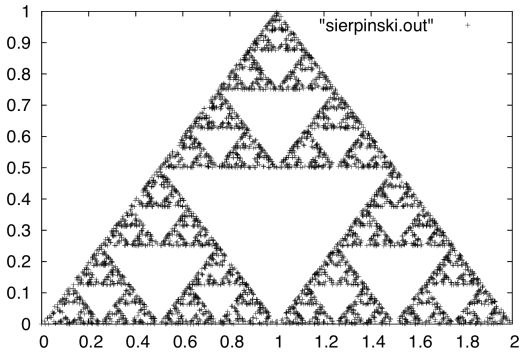




¾ Power law

Mixture of (Gaussian) clusters

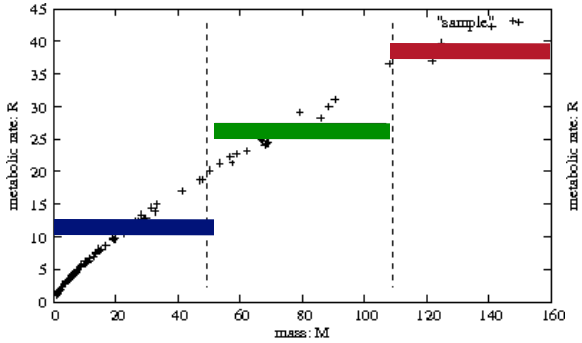


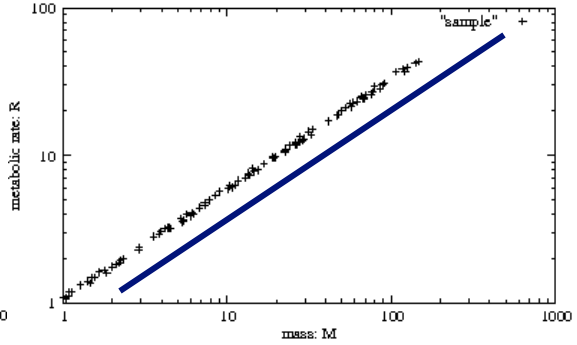


??

Always room for better models

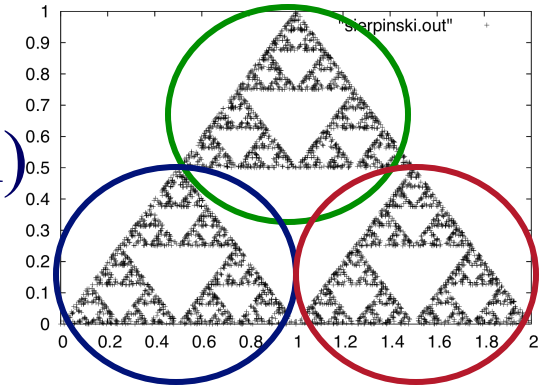
Piece-wise flat

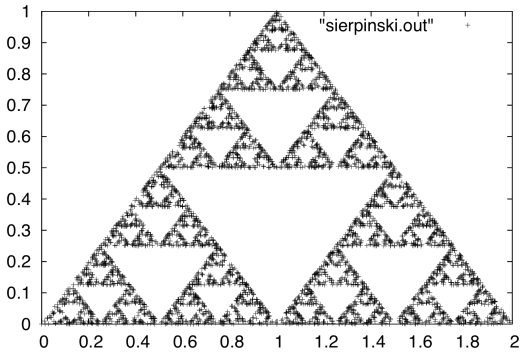




¾ Power law

Mixture of (Gaussian) clusters

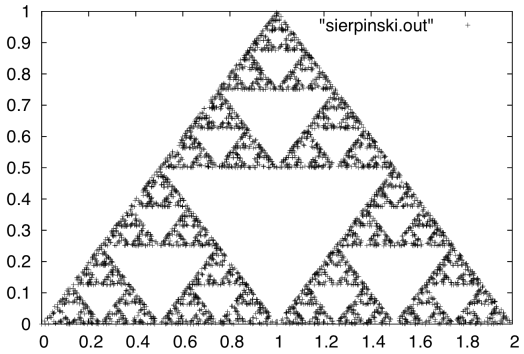




ONE, but Self-similar ‘cluster’

Always room for better models

- Barnsley's method of IFS (iterated function systems) can easily generate it [Barnsley +Sloan, BYTE, 1988]

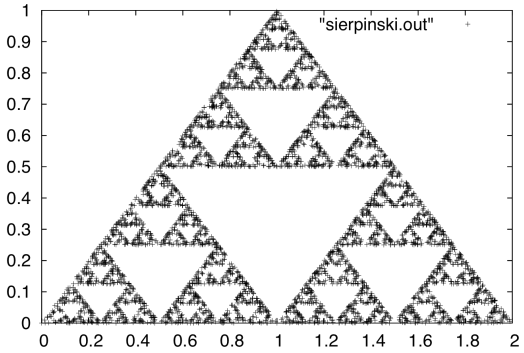


ONE, but Self-similar 'cluster'

~100 lines of C code: www.cs.cmu.edu~/christos/www/SRC/ifs.tar

Always room for better models

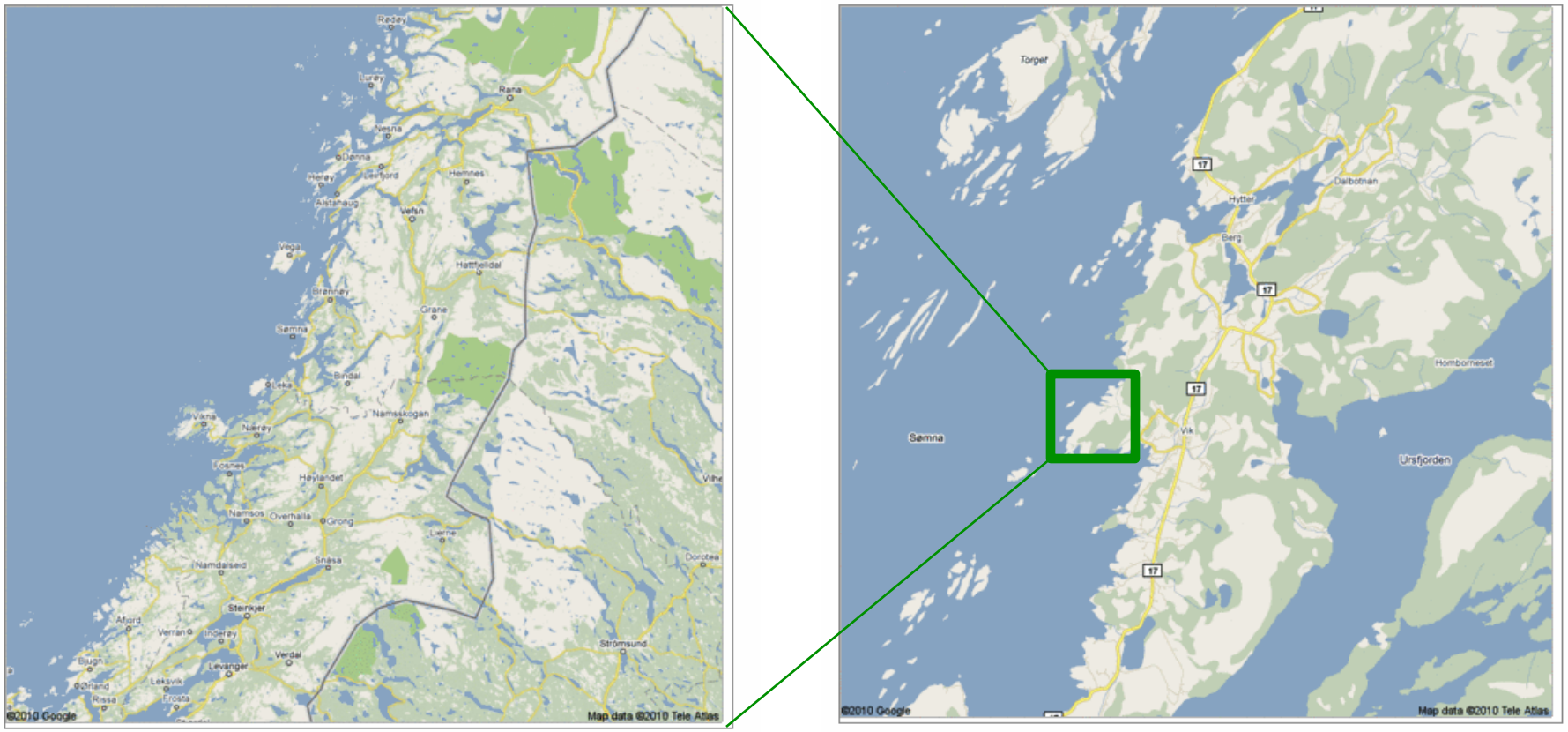
- But, does self-similarity appear in real life?



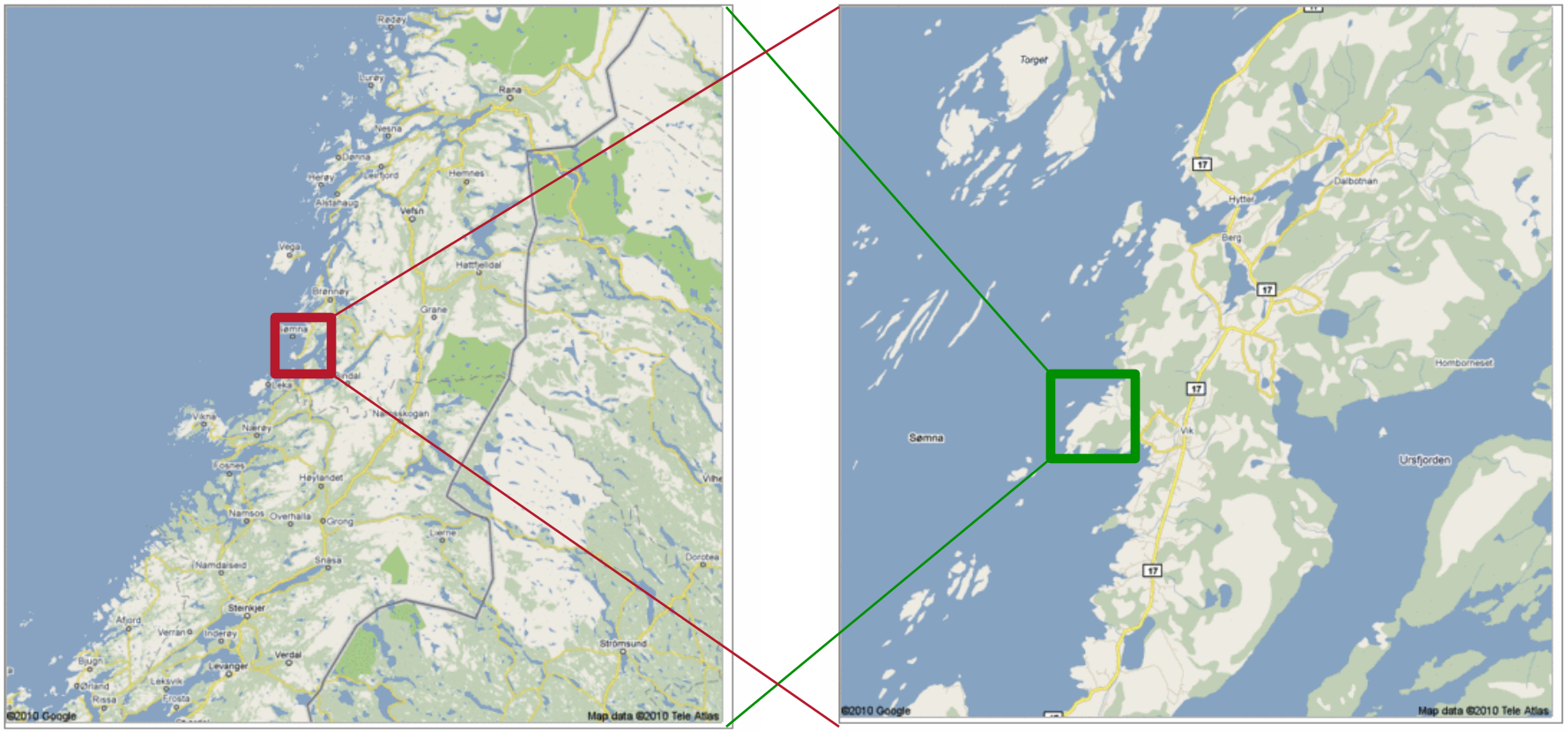
Real, self similar dataset

Real, self similar dataset

Real, self similar dataset



Real, self similar dataset



- the red is true
- origin: Norway
- but most other coastlines are 'self-similar', too!

How can we find better models?

- Obviously, an art ('undecidable'!)
- Helps if we
  - Listen to domain experts and
  - Listen to the data (next)

Outline

- Credit where credit is due
- Technical part – Data mining
  - Can it be automated? NO!
  - Research challenges
    - Listen to the data (the more, the better!)
- Non-technical part: `Listen'
  - To the data
  - To non-experts

Scalability

- Google: > 450,000 processors in clusters of ~2000 processors each [Barroso, Dean, Hölzle, "Web Search for a Planet: The Google Cluster Architecture" IEEE Micro 2003]
- Yahoo: ~5Pb of data [Fayyad'07]
- 'M45': 4K proc's, 3Tb RAM, 1.5 Pb disk

Promising research direction: scalability

- challenges
  - Vast amounts of data; storing; cooling (!); …
- … and opportunities:
  - DATA: Easier to collect (clickstreams, sensors etc)
  - S/W: Hadoop, hbase, pig, … : open source
  - H/W: 1Tb disk: ~ US$ 100

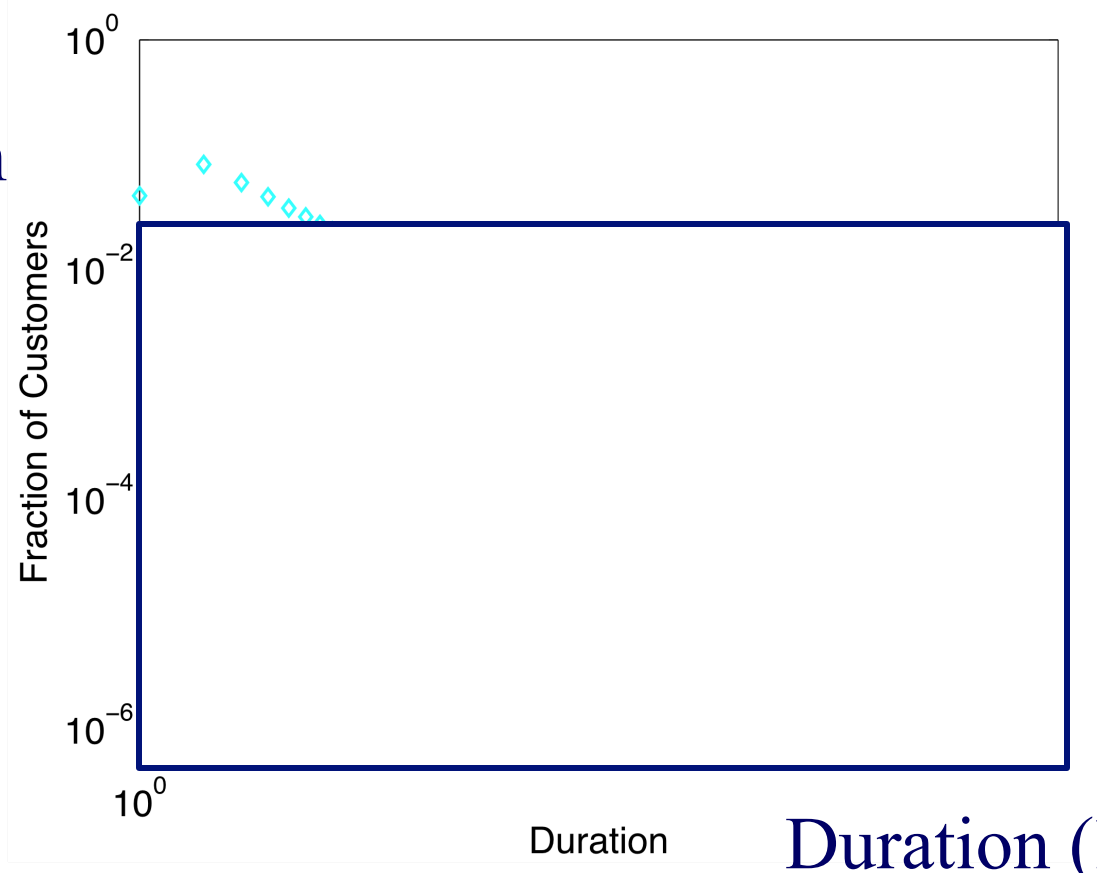
Promising research direction

- The more data, the more subtle patterns we may discover

- Examples of subtle patterns:

More data, more subtle patterns

PDF: fraction of customers (log scale)

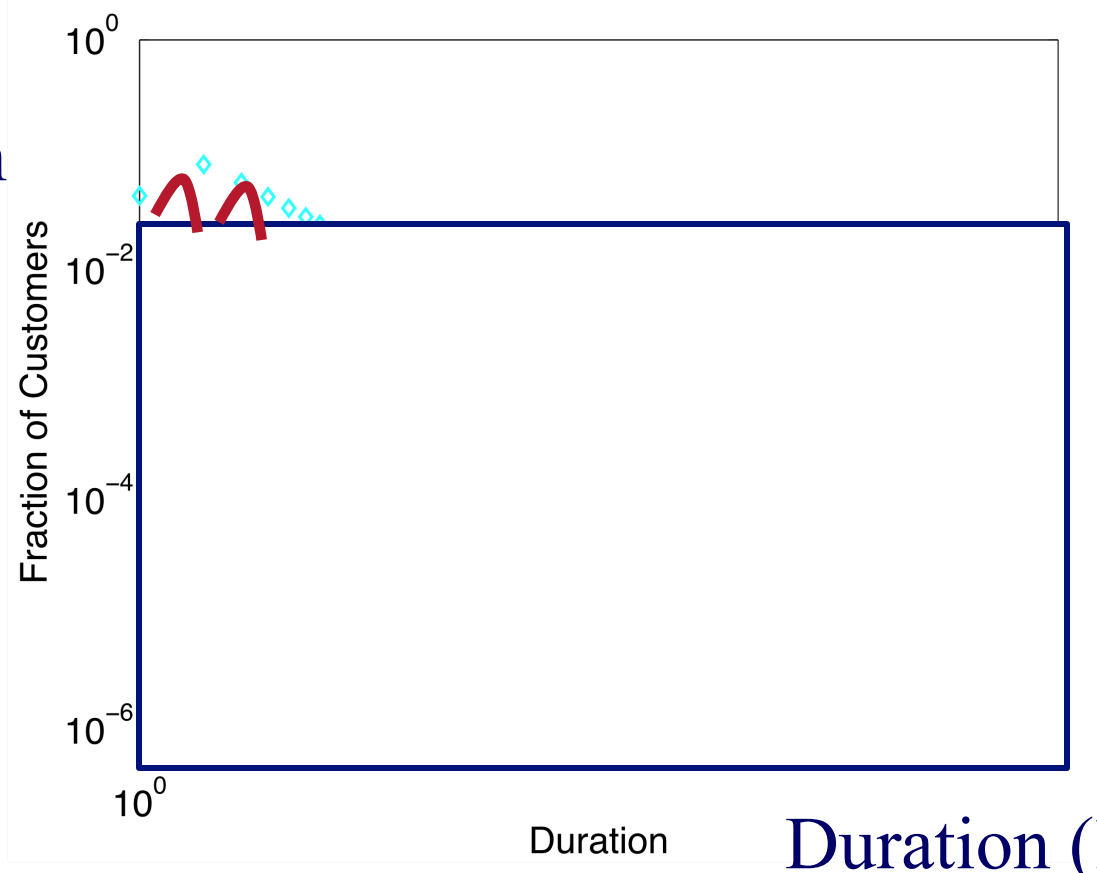


Duration (log scale)

Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: distributions*. KDD 2008: 596-604

More data, more subtle patterns

PDF: fraction of customers (log scale)

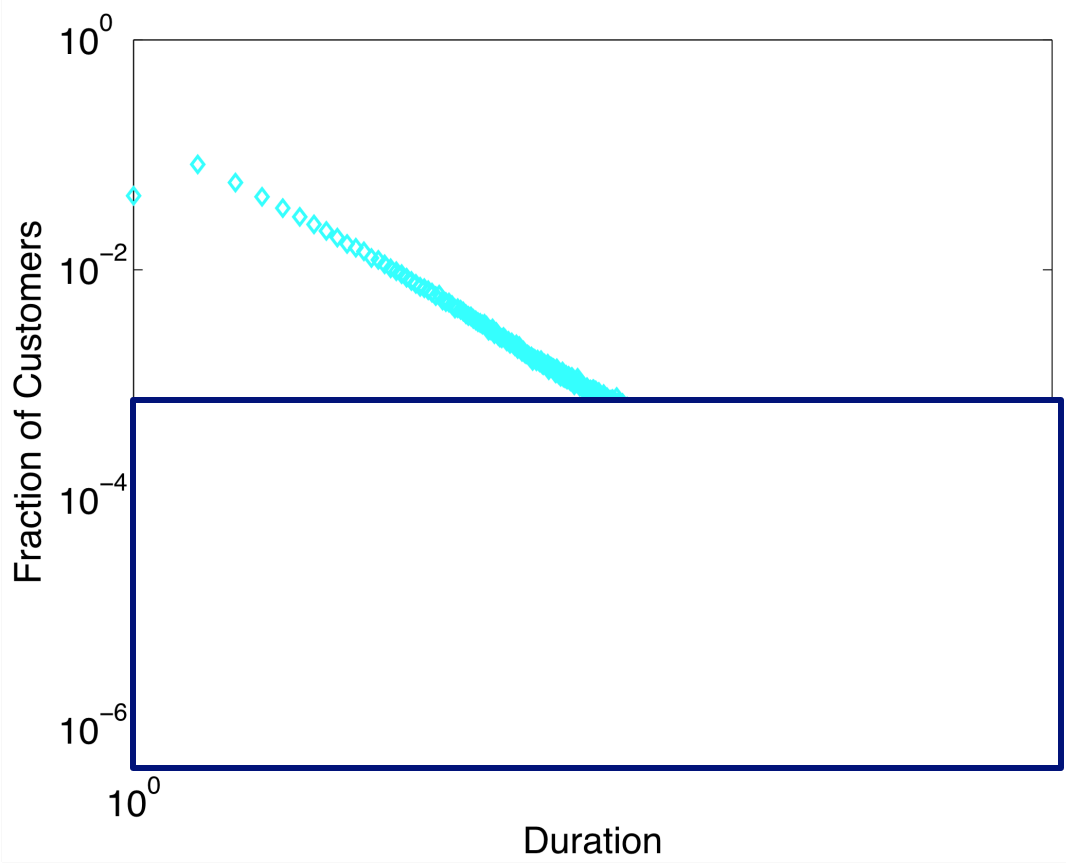


(mixture of) Gaussians

Duration (log scale)

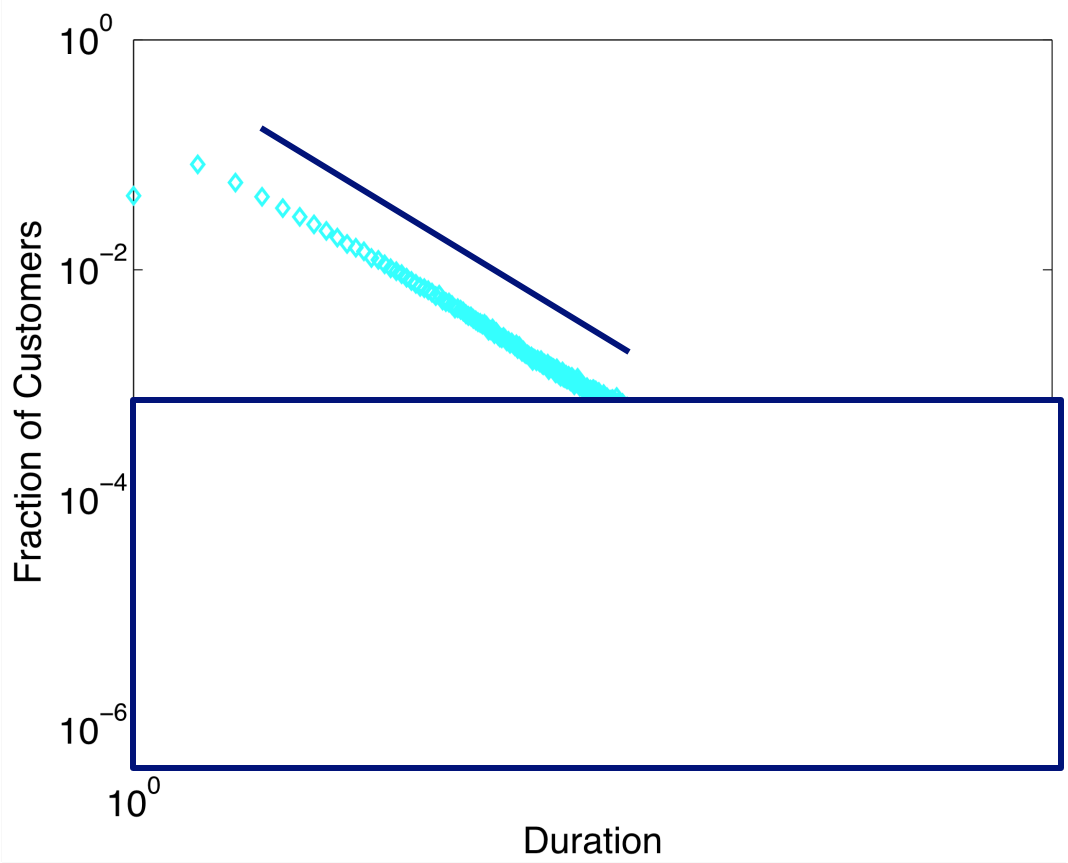
Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: distributions*. KDD 2008: 596-604

More data, more subtle patterns



Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

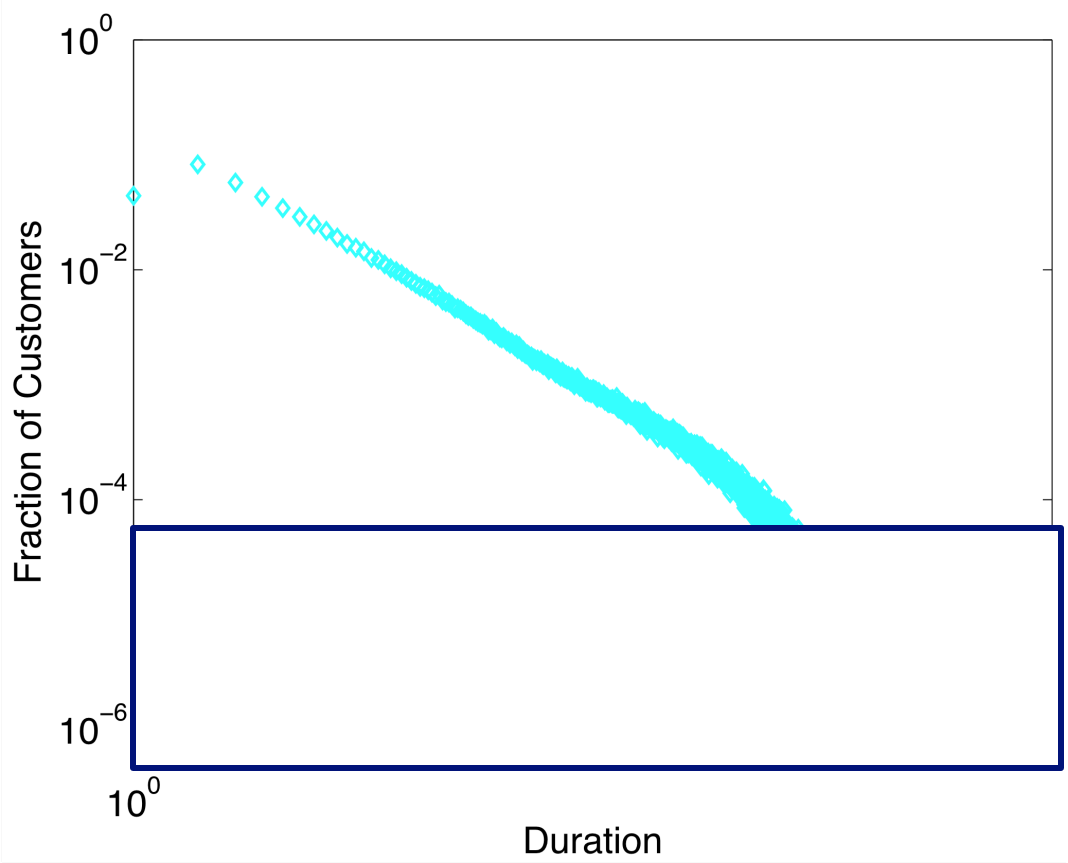
More data, more subtle patterns



Zipf
(Pareto,
Power-law)

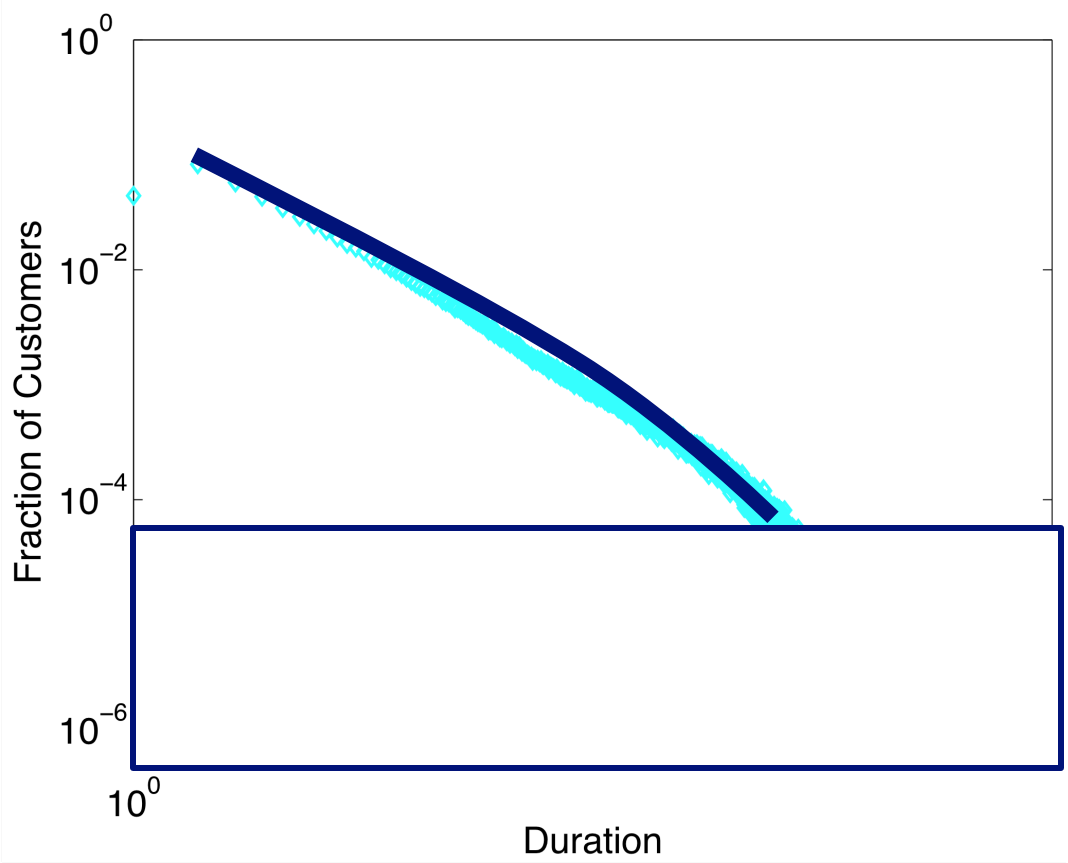
Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

More data, more subtle patterns



Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

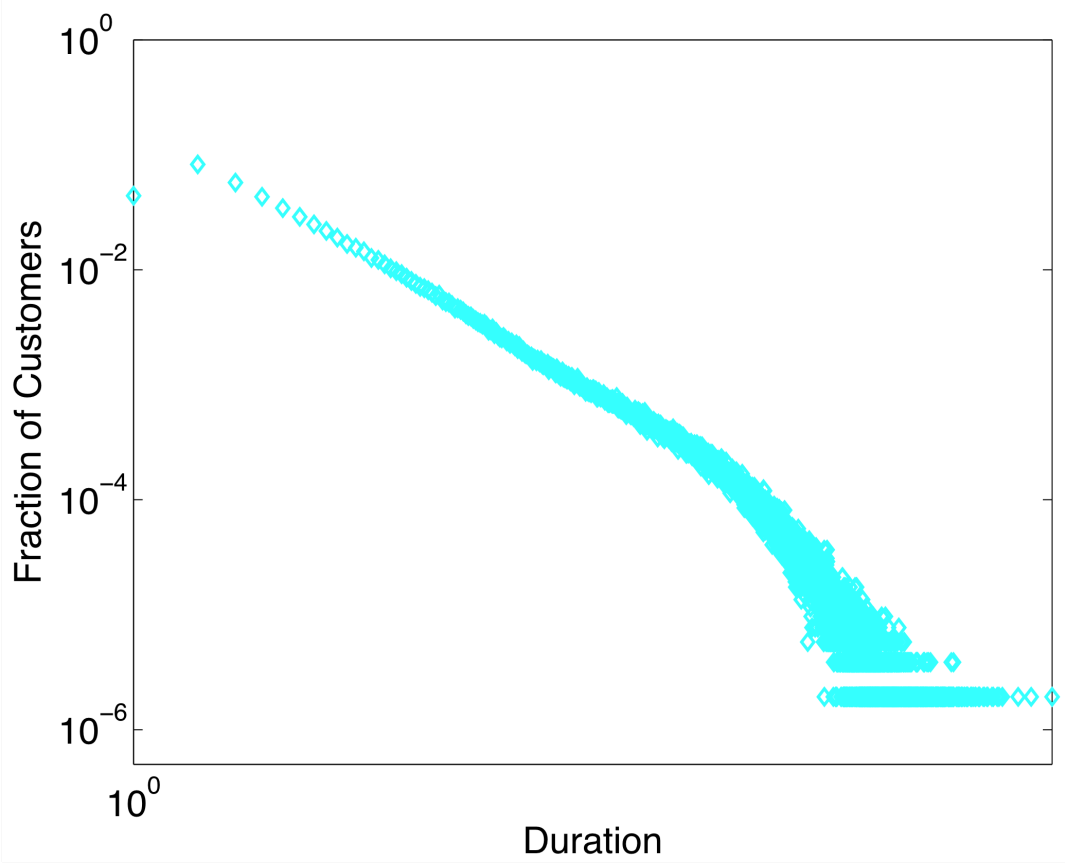
More data, more subtle patterns



lognormal

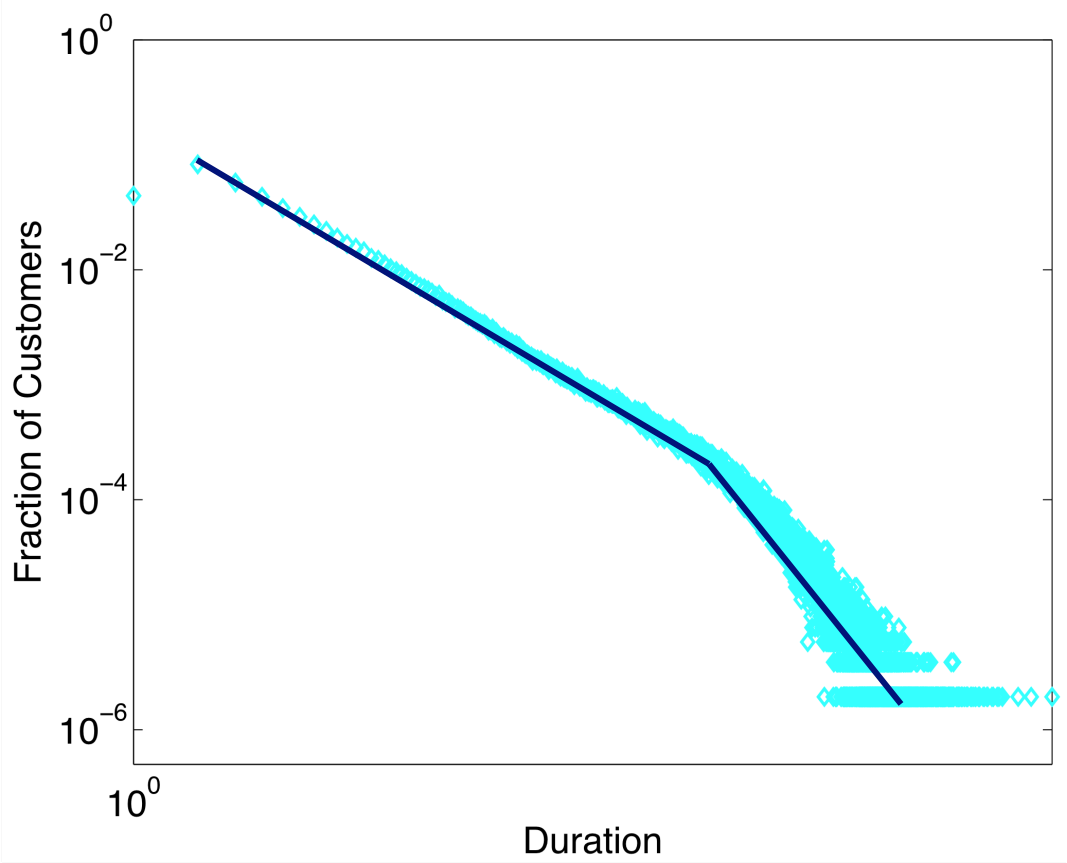
Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

More data, more subtle patterns



Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

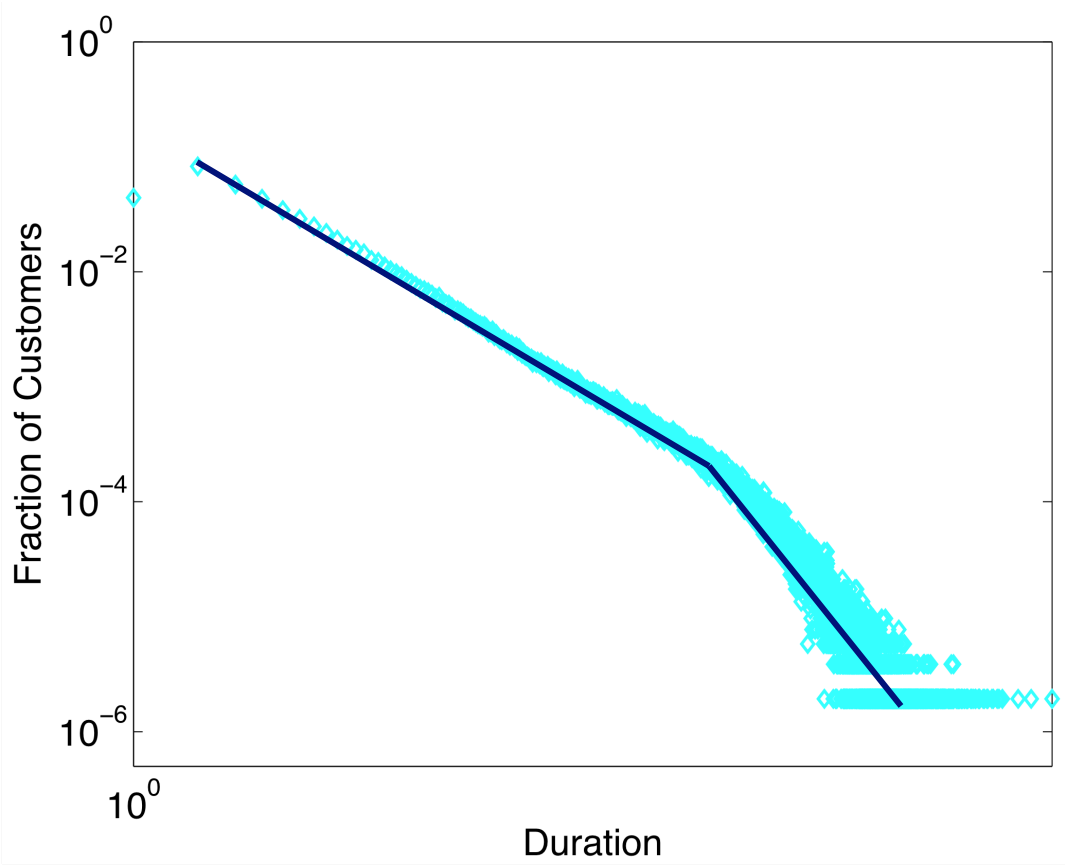
More data, more subtle patterns



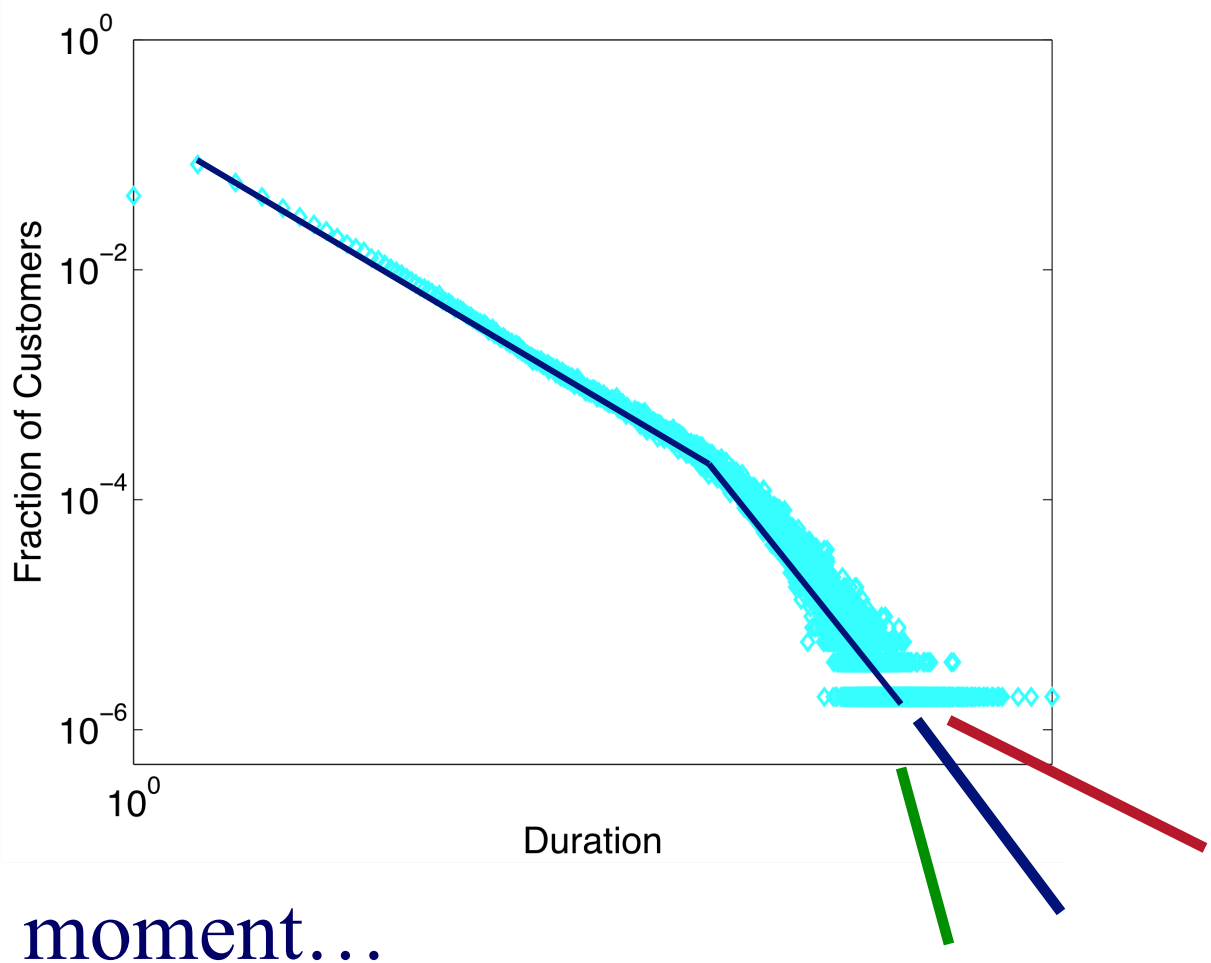
dPln
(=doubly
Pareto
Lognormal)

Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, Jure Leskovec: *Mobile call graphs: beyond power-law and lognormal distributions*. KDD 2008: 596-604

So, dPln is the answer?

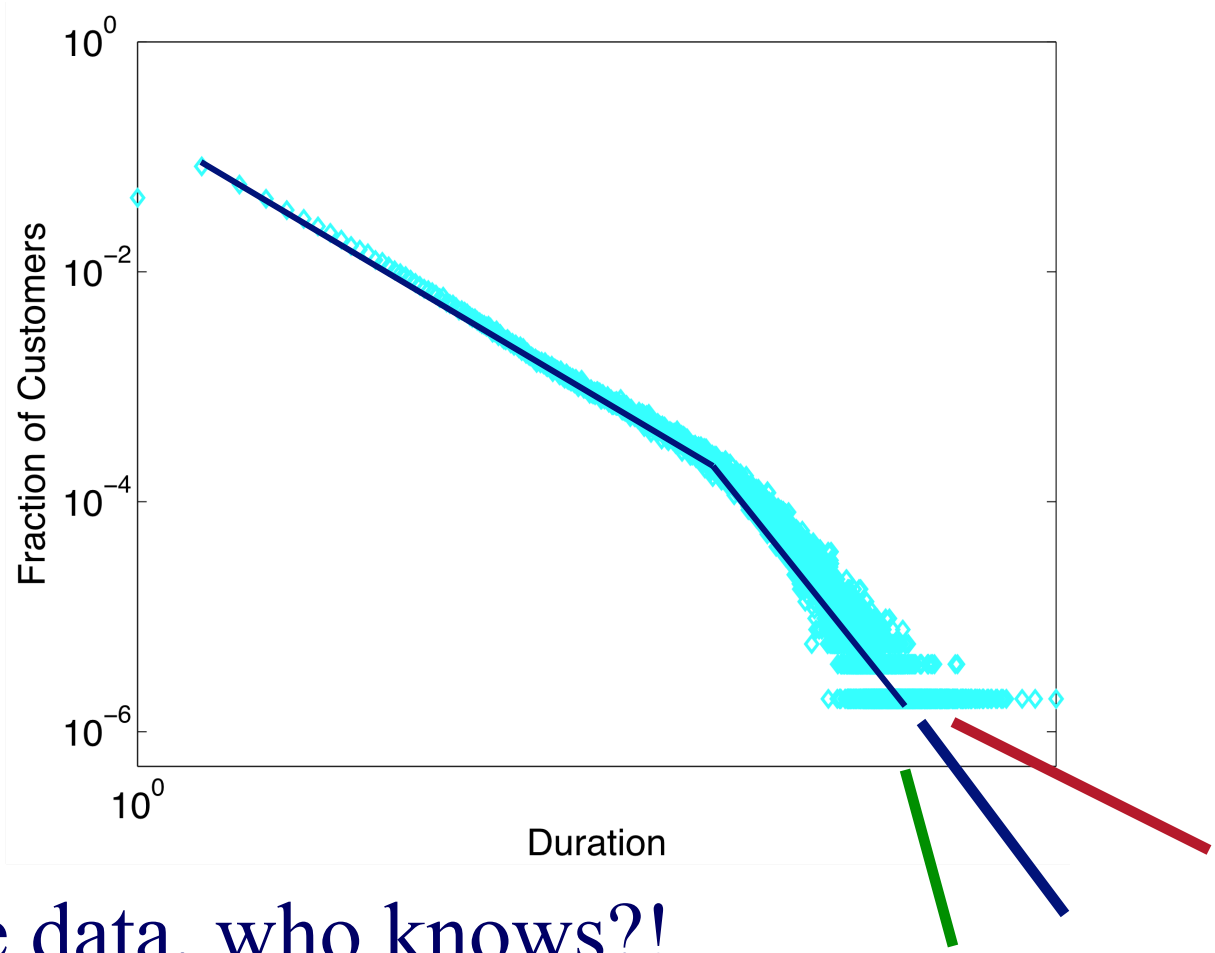


So, dPln is the answer?



Yes, for the moment…

So, dPln is the answer?



With more data, who knows?!

Outline

- Credit where credit is due
- Technical part – Data mining
  - Can it be automated? NO!
  - Research challenges
    - Listen to the data (the more, the better!)
- Non-technical part: 'Listen'
  - To the data
  - To non-experts

Listen to non-experts

- Explain ‘why’, to a non-expert (‘grandpa’)
- (and, even harder, explain ‘how’ – e.g.:
  - Frobenious Perron for irreducible MC

Listen to non-experts

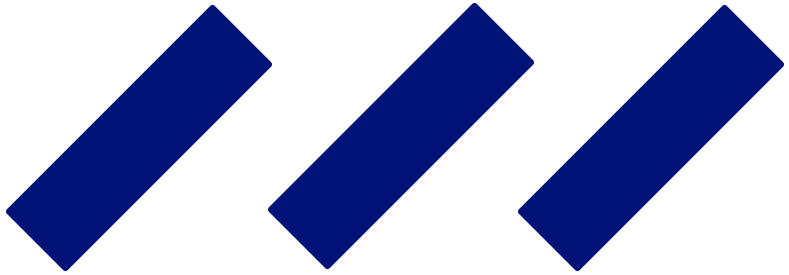
- Explain ‘why’, to a non-expert (‘grandpa’)
- (and, even harder, explain ‘how’ – e.g.:
  - Frobenious Perron for irreducible MC -> pageRank -> random surfer

Summary

- Data mining = compression = undecidable = job security ☺
- Hence: always room for better models/ patterns
  - Listen to the data (Gb, Tb and Pb of them!)
  - Listen to domain experts (e.g., ¾ Kleiberg's law)
- Listen to non-experts ('explain to grandpa')

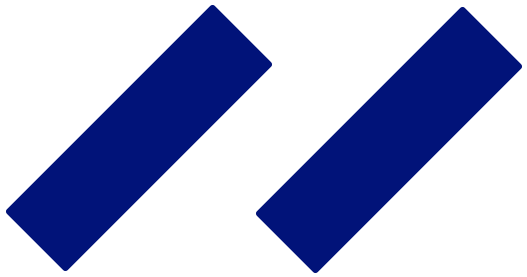
Compression, fun, recursion

- The shortest, recursive joke:
- There are 3 types of data miners



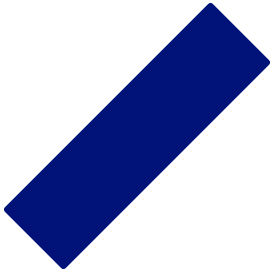
Compression, fun, recursion

- The shortest, recursive joke:
- There are 3 types of data miners
  - Those who can count



Compression, fun, recursion

- The shortest, recursive joke:
- There are 3 types of data miners
  - Those who can count
  - And those who can not



Thank you!

For the honor,
and for making this wonderful research community