# Wrapping It Up

Pauli Miettinen
Jilles Vreeken

24 July 2014 (TADA)

UNIVERSITÄT DES SAARLANDES

CLUSTER OF EXCELLENCE

max planck institut informatik

# What did we do?

**Introduction**

**Tensors**

**Information Theory**

**Mixed Grill**

**Wrap-up + <*ask-us-anything*>**

# Take Home: overall

Overview of the **hot** topics in data mining that Pauli and Jilles think are **cool**

**strongly biased** sample – by interest and available time

We wanted to give a general picture of what data mining is, what makes it special, and **what's currently happening** at the edge of human knowledge

# Key Take-Home Message

Data mining is **descriptive** not **predictive**
the goal is to give you insight into your data,
to offer (parts of) candidate hypotheses,
*what you do with those is up to you*.

# Take Home: Tensors

**Multi-way extensions**

of matrices

Anything you can do with matrices
you can do with tensors…
…only harder
…and taking into account
**multi-way relationships**

# Take Home: Decompositions

Different
**tensor decompositions**
reveal different types of patterns

The choice of correct
decomposition
must be based on
**application's needs**;
there's no golden bullet

# Take Home: Information Theory

**Exploratory data analysis**
wandering around your data,
looking for interesting things,
***without*** being asked questions
you cannot know the answer of.

Questions like:

*What distribution should we assume?*

*How many clusters/factors/patterns do you want?*

*Please parameterize this Bayesian network?*

# Take Home: Interestingness

**Interestingness**

is ultimately **subjective**

Still, to have algorithms that can find potentially interesting things we somehow **need** to formalize it

# Take Home: Information Theory

Information Theory is a branch of statistics, concerned with measuring **information**

$$information = reduction\ of\ uncertainty$$

Uncertainty can be quantified in **bits**

*Everything **new** you learn about your data allows you to **compress** it **better***

# Take Home: MDL

The Minimum Description Length (MDL) principle

given a set of models $\mathcal{M}$, the best model $M \in \mathcal{M}$
is that $M$ that minimizes

$$L(M) + L(D|M)$$

in which

$L(M)$ is the length, in bits, of the description of $M$

$L(D|M)$ is the length, in bits, of the description of the data when encoded using $M$

# Take Home: Maximum Entropy

The principle of Maximum Entropy

given a set of testable statistics $B$,
the best distribution $p^*$ is that $p$ that satisfies

$$\int_S p(x) f_i(x) dx = \alpha_i \quad \text{for } (f_i, \alpha_i) \in B \quad .$$

while maximizing

$$H(p)$$

$p^*$ is the **most** uniform, **least** biased distribution
that corresponds with belief set $B$
it models **your** expectation – assuming you use $B$ optimally

# Take Home: Graph Mining

Most graph mining approaches
are **global** and **predictive**

*'Explain everything in one go'*

real graphs are too complex for that

Taking a **local** and **descriptive** approach allows
for more detailed results, richer problems,
easier formalization, efficient solutions

very little done so far, many cool open problems

# Take Home: Redescriptions

**Redescriptions** explain
the same thing many times

Emerging topic that has
not yet fully broken into the
data mining canon

Can be seen as **translation**
*within* a dataset

# Take Home: Dynamic Data

Data is rarely
**static**
even though many algorithms expect that

**Streaming algorithms** work when
data is too big to fit anywhere
while **dynamic algorithms** aim
to adjust the answer with the changing data

# Take Home: Assignments

*"What the hell*
*where they thinking??"*

We wanted you to learn

to **read** scientific papers without getting lost in details

quickly forming **high level** pictures of **complex ideas**

read **critically**, seeing through scientific sales-pitches

show **independent thinking**, make ideas your **own**

We were not disappointed.

# Take Home: TADA

**Data analysis**
is important, **upcoming**, but still very young

aims to tackle **impossible problems**,
such as finding ***interesting*** things
in **enormous search spaces**

is a **weird** mix of theory and practice:
likes to be **foundational**, yet not afraid of **ad hoc**

and, not unimportant, ***it's lots of fun.***

# Exam dates

## The Exam

type: oral

when: September 11th

time: individual

where: E1.3 room 0.16

what: all material discussed in the lectures, plus one assignment (your choice) per topic

*If you know you can't or just won't attend, please let us* ***know***

## The Re-Exam

type: oral

when: October 1st

time: individual

where: E1.3 room 001

# Evaluation: I did not like

*"Slides are not detailed enough for revision"*

# Evaluation: Suggestions

*"More ways for discussing assignment solution"*

More ways for understanding the suggestion?

*"Bit heavy course for 5 ECTS"*

Yes.

*"More details for practical stuff, like how and why"*

Maybe. Maybe not here.

*"More lectures with both lecturers"*

Really?

# Things to do

## Master thesis projects

- in principle: yes!
- in practice: depending background, motivation, interests, and grades --- plus, on whether we have time
- interested? mail Pauli and/or Jilles

## Student Research Assistant (HiWi) positions

- in principle: maybe!
- in practice: depends on background, grades, and in particular your **motivation** and **interests**
- interested? mail Jilles and/or Pauli, include CV and grades

# Sample Topics – JV

**Graphs**

- characterising viruses
- realistic graph generators
- mining interesting sub graphs
- patterns in tweets

**Causality**

- did X cause Y?
- mining causal graphs
- what's the cause of *this*?
- predicting the future

**Useful Patterns**

- the Difference & the Norm
- privacy & data generation
- pattern-based indexing
- noise reduction

**Rich Data & Text**

- pattern-based topic models
- grammar & compression
- rich MaxEnt modelling
- outliers in rich data

# Sample Topics – PM

**Matrices**

– tropical algebras
– Boolean algebras
– efficient algorithms
– good applications

**Tensors**

– new decompositions
– efficient algorithms
– applications

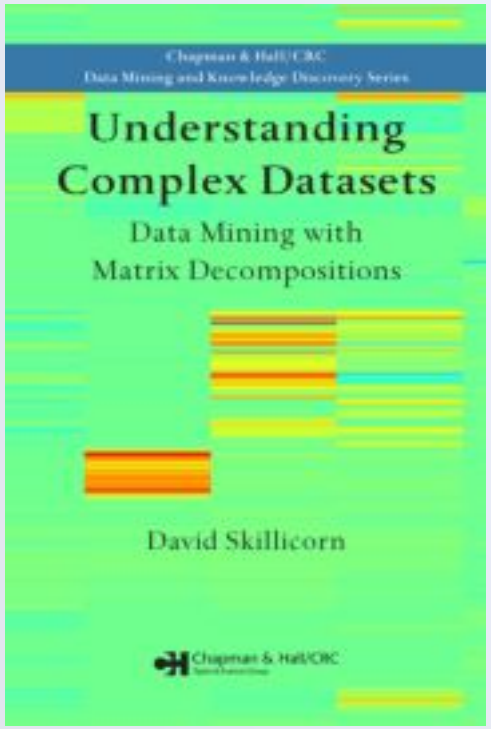
**Theory**

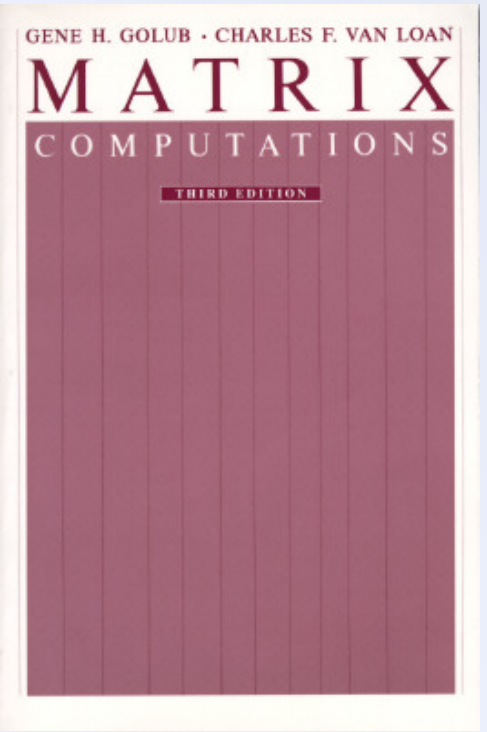
– approximability
– computational complexity
– practical results
– DM motivated

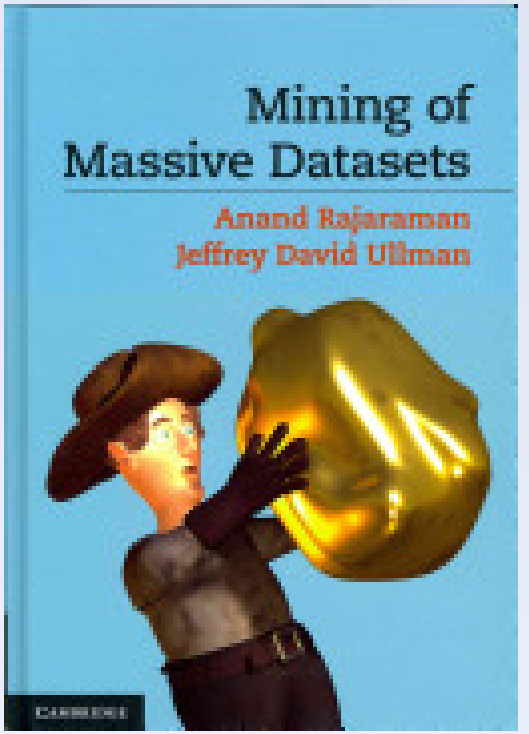
**Redescriptions**

– new algorithms
– new applications
– new formulations

# Good reads – PM



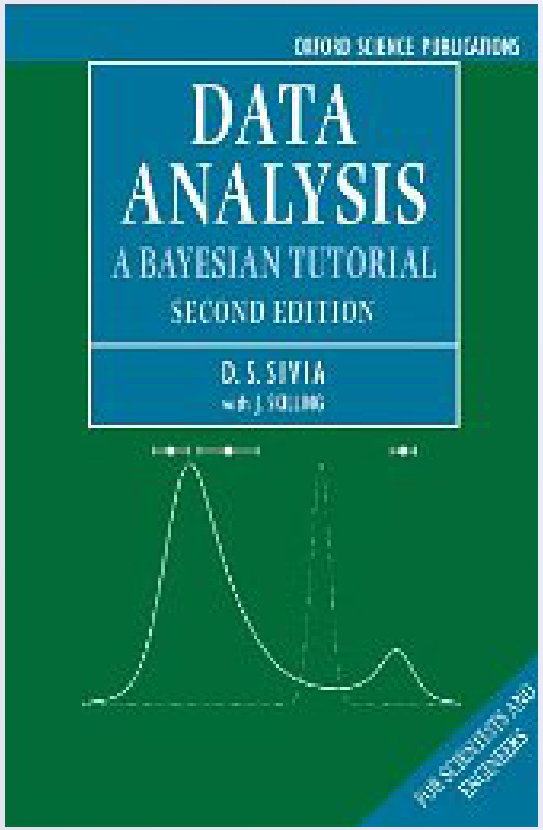




*Understanding Complex Datasets*
D. Skillicorn
(light reading on matrix and tensor decomps.)

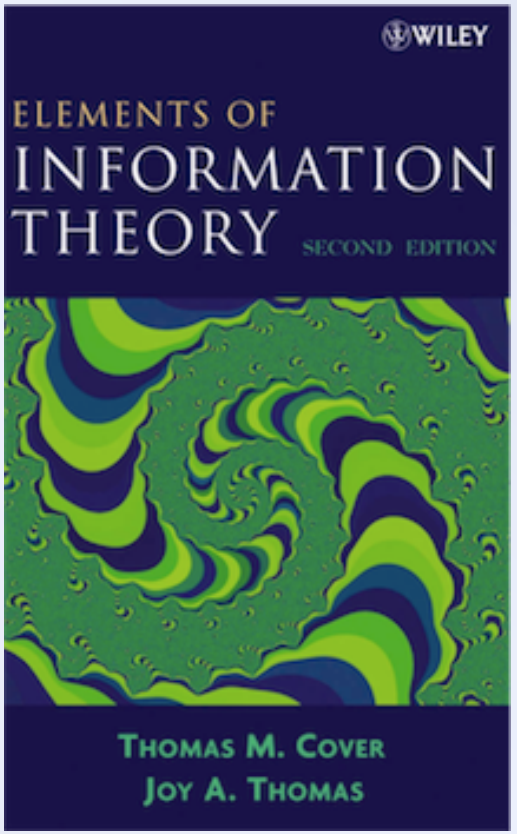
*Matrix Computations*
G.H. Golub & C. Van Loan
(anything-but-light, reference book)

*Mining of Massive Datasets*
Rajaraman, Lescovec & Ullman
(work-in-progress textbook)

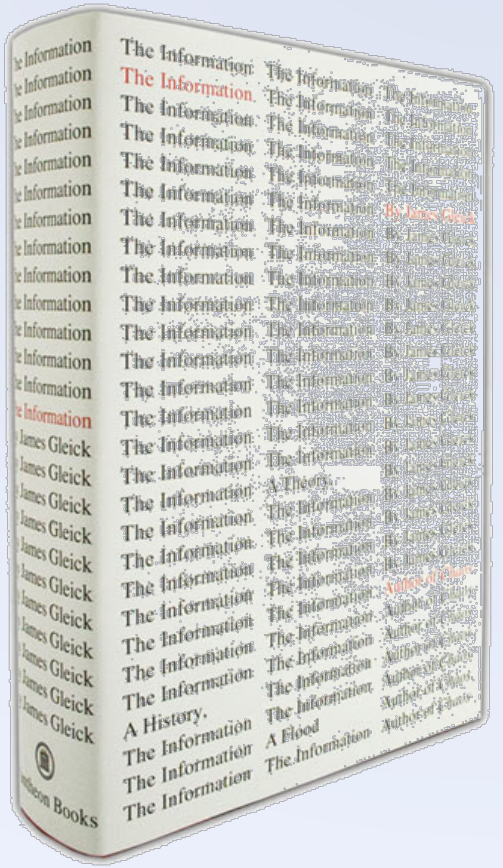
# Good Reads – JV




*Data Analysis: a Bayesian Tutorial*
D.S. Sivia & J. Skilling
(very good, but skip the MaxEnt stuff)




*Elements of Information Theory*
Thomas Cover & Joy Thomas
(very good textbook)




*The Information*
James Gleick
(great light reading)

# Teach us More!

Well, ok… but, we are still thinking what/if to teach next semester.

Options include:

| | |
|---|---|
| Information Theory | (regular course – JV) |
| Mining and Using Patterns | (seminar/discussion – JV) |
| Causal Inference | (seminar/discussion – JV) |
| Tensor Methods | (seminar/discussion – PM) |
| Redescription Mining | (seminar/discussion – PM) |
| Fixing It (or, Reproducible Science) | (seminar/practical – PM&JV) |
| Data Mining Lab | (practical – PM&JV) |

# Algorithmic Data Analysis Group

*…coming soon…*

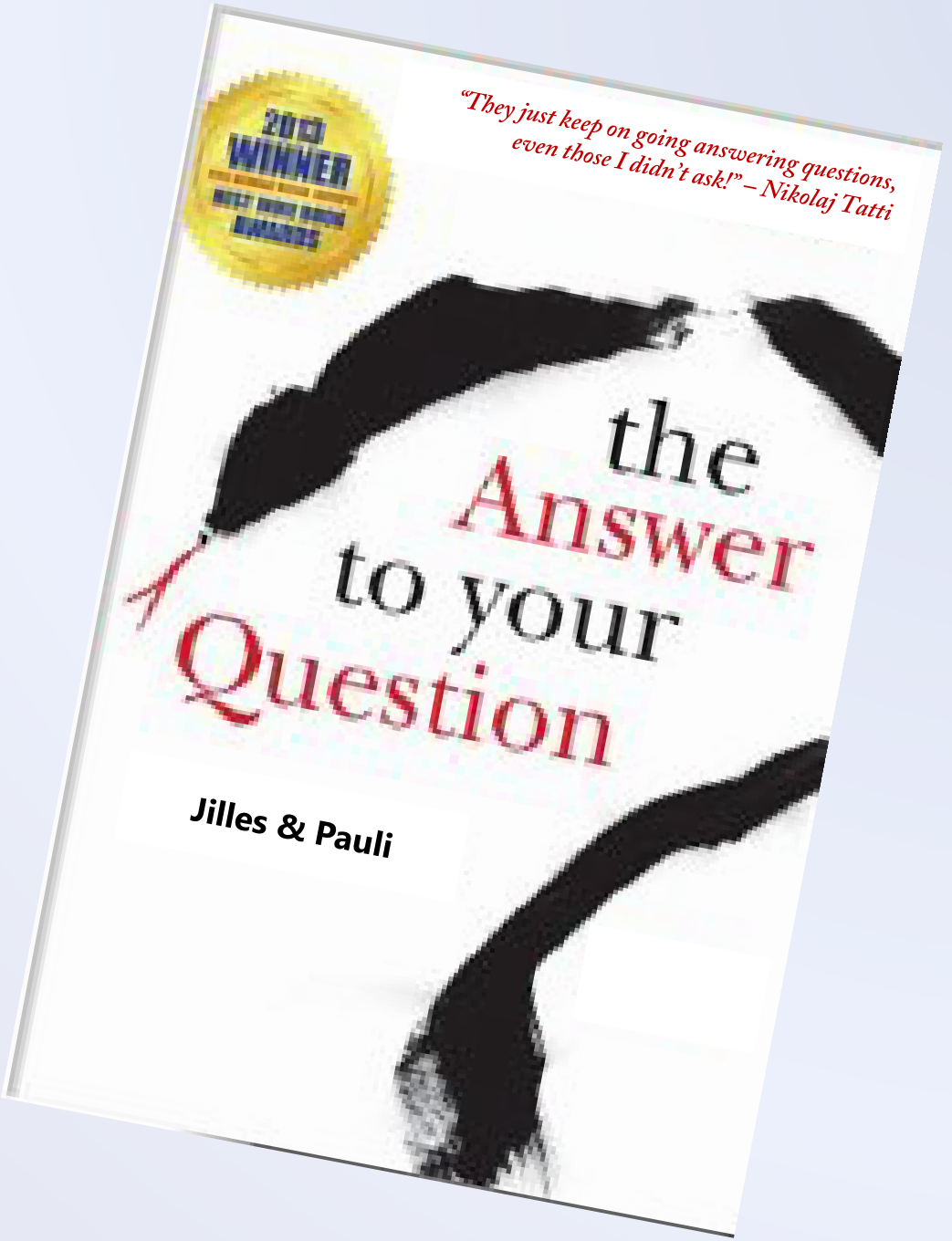
a joint-venture of the MPI groups on
Data Mining and Exploratory Data Analysis.

*ada.mpi-inf.mpg.de*

We'll include announcements of
relevant talks and events, and
cool new work by yours truly

(maybe even mailing list)

# Question Time!



# Privacy & Data Mining

*"What is your opinion on privacy preserving data mining? Have you ever worked with it? Do you think it is useful, or does it somehow contradicts 'the spirit' of data mining?"*

# Text Mining

*"Have you ever worked with text mining?*
*Do you think considering grammar is necessary,*
*or is mere statistics enough?"*

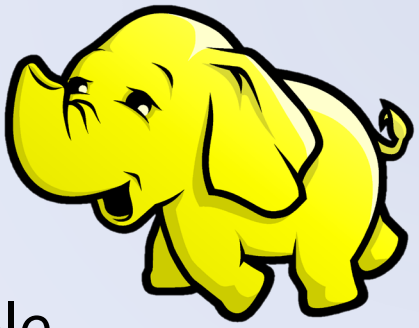
# Big Data

*"Does Big Data exist?"*

*"How big is Big Data?"*

*"When is the data Big enough?*
*Is more data always better?"*

# Mining Massive Data



Map Reduce, Hadoop, Big Table,
Cassandra, Spark, Dremel, etc, etc

**engineering** or **science**?

Essentially **tricks** – not magic –
that work well for certain specific problems

For KDD 2014, *at least* 25 out of 150 presentations
will be **specifically** aimed at 'large scale' stuff

# Mining the Cloud

*"How about data analytics in the cloud?"*

# Social Network Analysis

Many, many, many papers about
social network analysis

So far: lots of statistics, not much 'mining'
That is, most are about how to **model a graph probabilistically**, how to **fit a given distribution**.

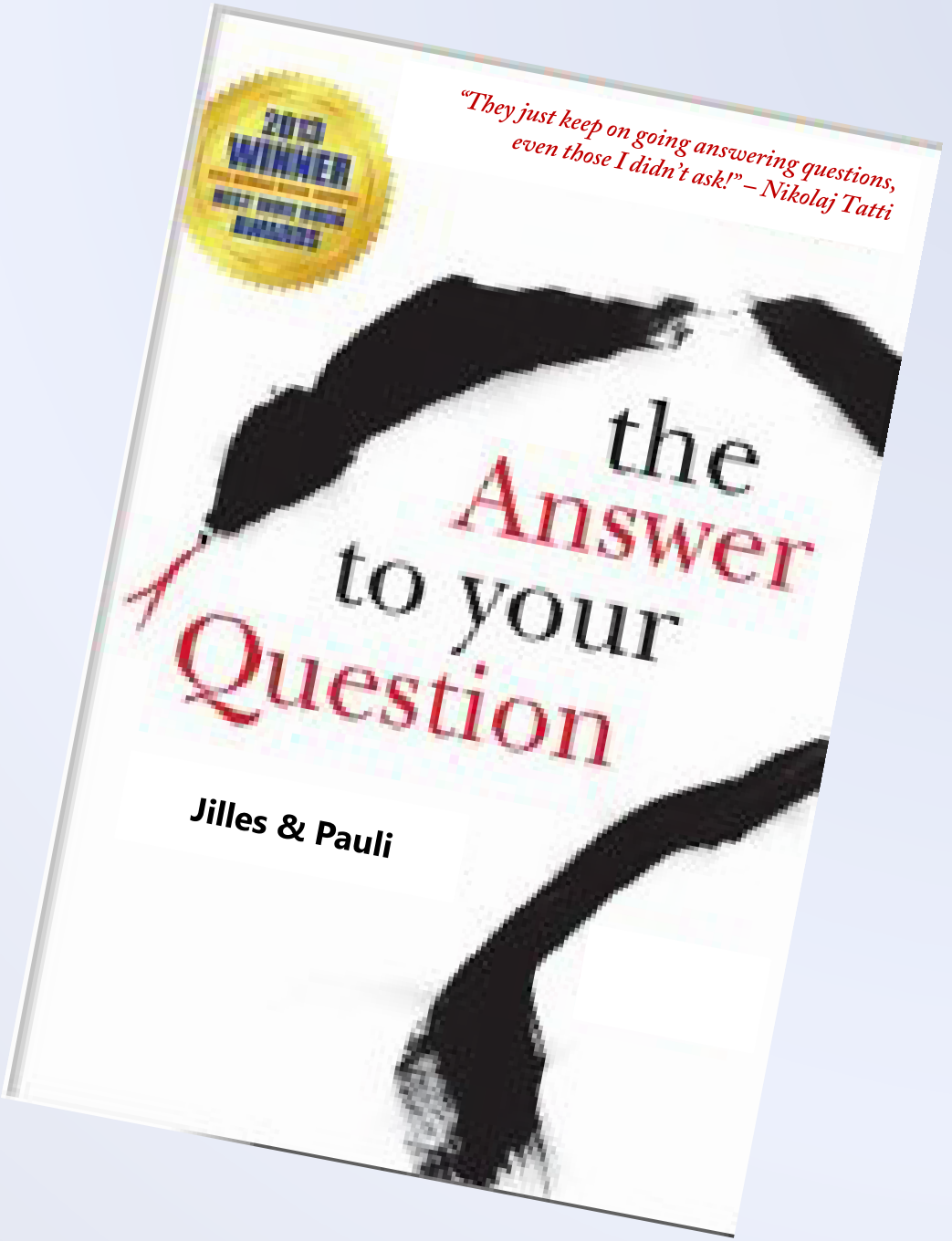
The Elephant in the Room:
what is the 'graph' distribution?

Nobody knows. Yet.

# Graph Mining

This is the part
where Pauli and Jilles
may or may not say something
about graphs.

Your Question Here!



# Conclusions

This concludes TADA'14.

We hope you enjoyed the ride.

# *Thank you!*

This concludes TADA'14.

We hope you enjoyed the ride.