TADA
practicalities & more on DM

24 April 2014

More on Data Mining as a Science

DM as method development

- Data mining develops methods for scientists
  - C.f. mathematics or statistics
- The research of DM in universities doesn’t follow the scientific paradigm
  - But that doesn’t make it a voodoo science
  - …the applications of DM are another story

Of DM, ML, and Stat

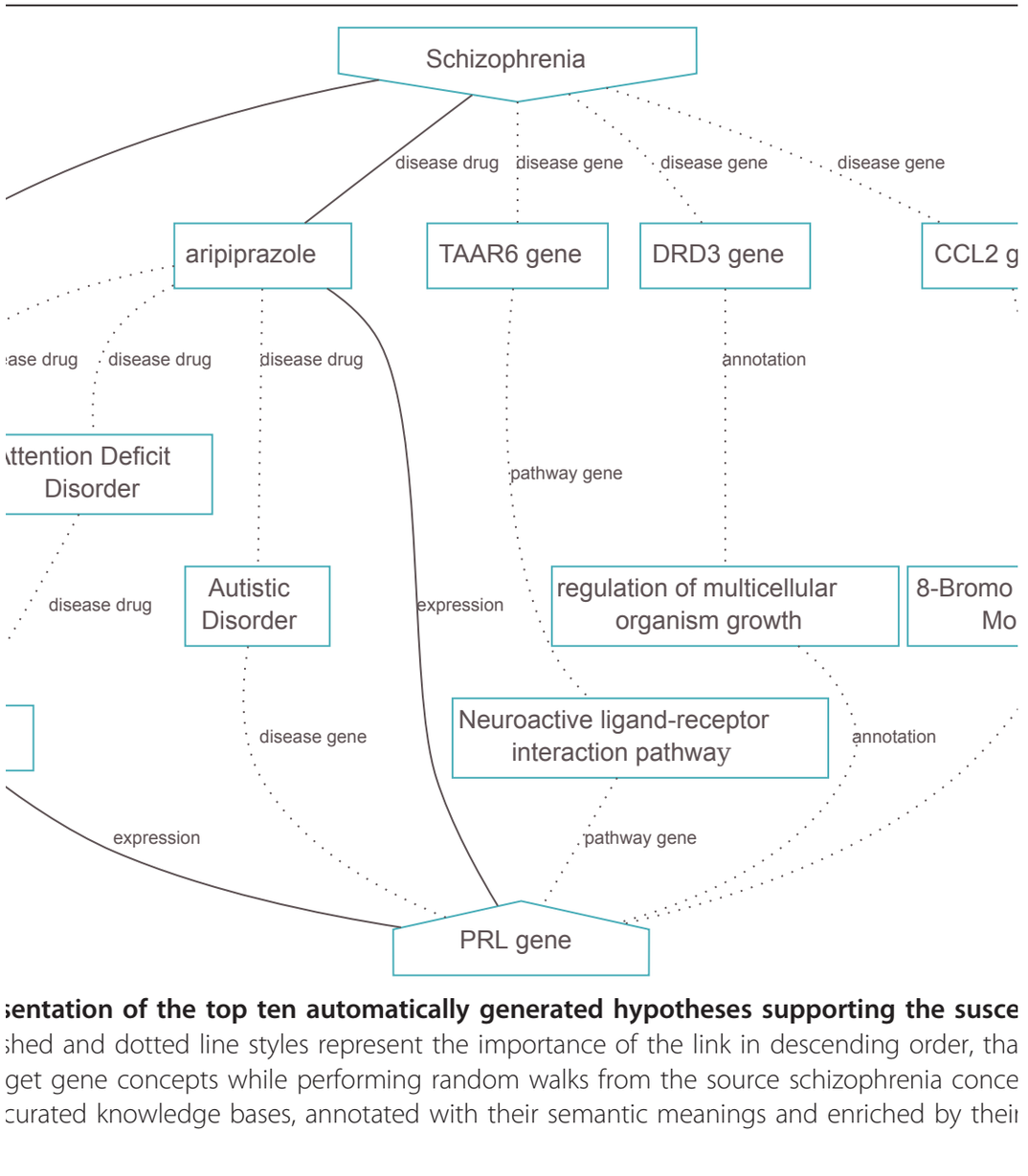
- One trichotomy:
  - Statistics studies how reliable inferences can be drawn from imperfect data
  - ML develops technology of automated induction
  - DM is the art of extracting useful patterns from large bodies of data

http://www.stat.cmu.edu/~cshalizi/350/, http://geomblog.blogspot.de/2014/03/data-mining-machine-learning-and.html

Data Mining success stories

Bioinformatics

- BioGraph provides automated inference of functional hypotheses
- E.g. which genes are most potential to be associated with certain diseases



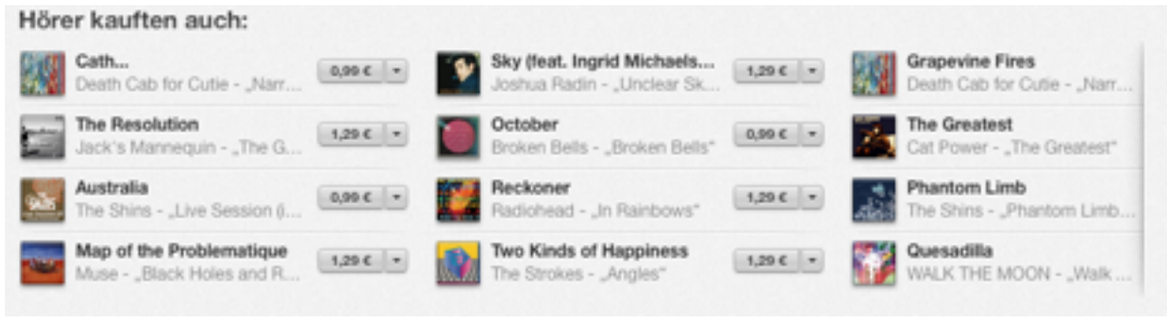
sentation of the top ten automatically generated hypotheses supporting the susce
shed and dotted line styles represent the importance of the link in descending order, tha
get gene concepts while performing random walks from the source schizophrenia conce
curated knowledge bases, annotated with their semantic meanings and enriched by thei

Liekens et al.: BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation, *Genome Biology*, 2011

Making money

- “Recommended for you”

- “Others often bought also”



- All of modern targeted advertisement is based on some type of data mining

Google AdWords

Obama’s re-election

- Data of electorate was used to target the campaing efforts where they count

- DM was also used to optimize fund-raising from small donations

Church uses Big Data

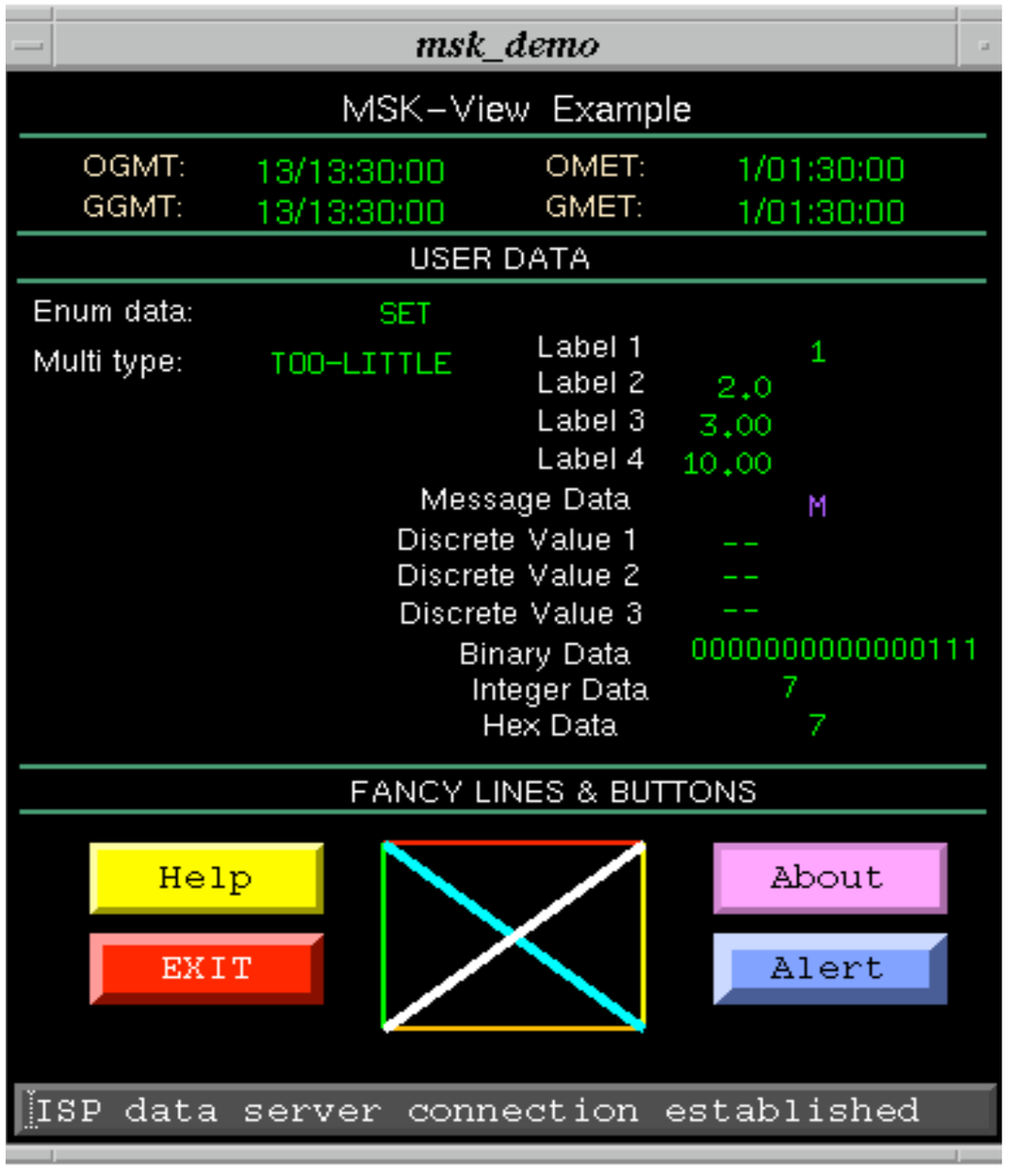
- Evangelical Lutheran Church of Finland uses data mining to study its parishes

- What type of people live in which geographical areas?

http://www.hs.fi/talous/Iso+data+auttaa+pappia+saarnassa/a1397539201451

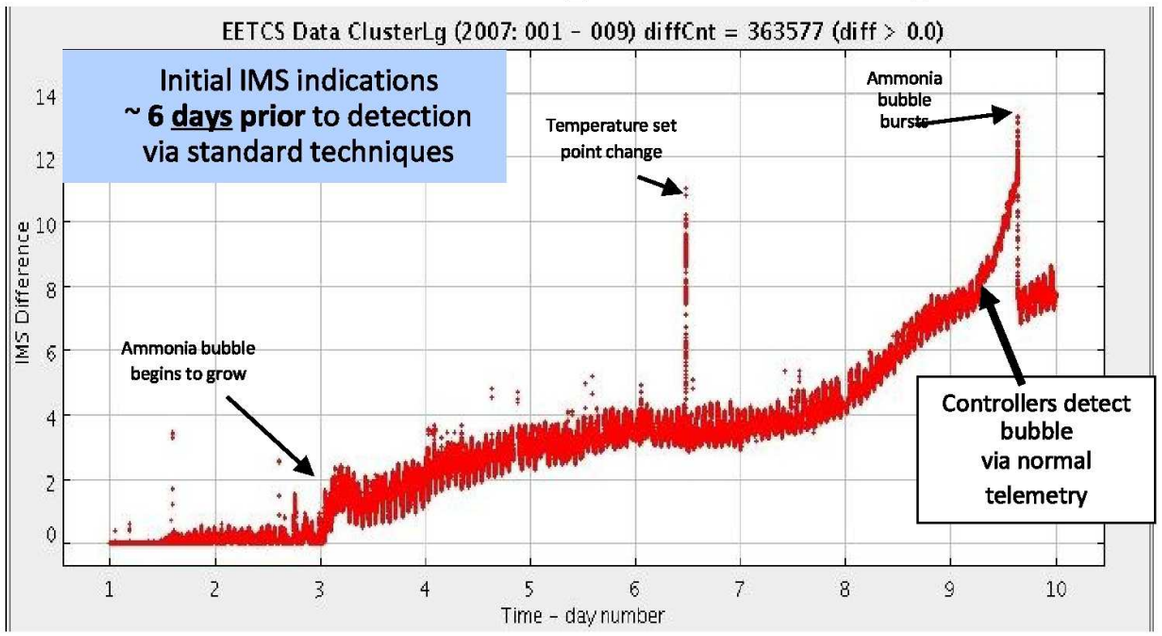
Space program safety

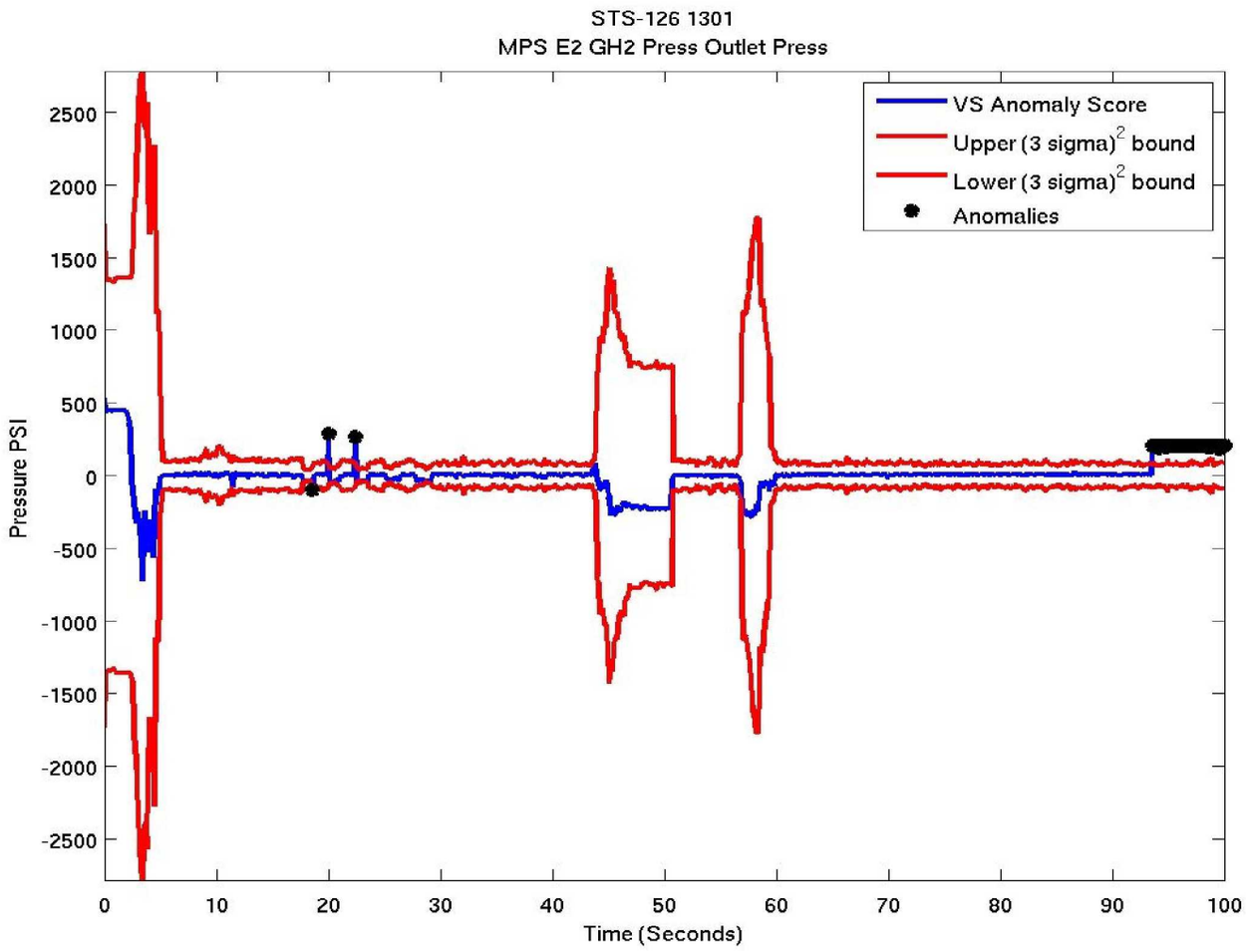
- ORCA searches outliers from sensor readings by comparing parameter-value vectors to their neighbors
- IMS builds a model of normal variance of sensor readings to detect anomalies



D.L. Iverson: System Health Monitoring for Space Mission Operations, 2008 IEEE Aerospace conference

More on IMS





- In early January 2007, ISS Early External Thermal Control System developed an ammonia gas bubble
- Bubble noted by ISS controllers only ~9 hours before it "burst" and dissipated back into liquid

Ashok N. Srivastava: Data Mining at NASA: from Theory to Applications, KDD 2009

Practicalities

Schedule

Month	Day	Lecture topic	Assignments
April	17	Intro	
	24	Practicalities & where DM is used	1st assignment given out
May	1	*No lecture* (First of May)	
	8	Intro to Tensors	1st assignment DL, 2nd assignment given out
	15	Tensors in DM	
	22	Special topics in tensors	
	29	*No lecture* (Ascension day)	
June	5	MDL for pattern mining	2nd assignment DL, 3rd assignment given out
	12	Maximum entropy & iterative data mining	
	19	*No lecture* (Corpus Christi)	
	26	Kolmogorov complexity, cumulative entropy, and causality	
July	3	Graphs I	3rd assignment DL, 4th assignment given out
	10	Graphs II	
	17	Graphs III	
	24	Wrap-up	4th assignment DL

On Exam

- Day and place TBA
  - Most likely in early September
- Type TBA
- Final grade is based on the final exam and the assignments
  - Assignments also determine the eligibility to sit the final exam

On assignments: general

- 4 assignments
- Grading: fail, pass, excellent
- You can fail one assignment
  - 2 fails ⇒ course failed
- Every excellent gives 1/3 point improvement on the final exam grade
  - But maximum of 1 full point (3 ex’s)
- **You must pass the final exam to pass the course**

On assignments: requirements

- Assignments are to be written in proper academic-style English
- **Proper citations**
  - You are given sources, but you can also use outside sources
    - Naturally must be mentioned
    - Plagiarism ⇒ failed assignment

On assignments: format

- Assignments need to be returned as PDF files by email
  - **No** .doc(x), .odt, .rtf, .txt, .xml, .html, .pages, .ps, .wp, or anything else
- No lenght limits — use the space you need
  - Probably most will need 3–4 pages…
- All PDFs must have name, matriculation number, email address, and clearly state the topic

On assignments: returning

- The assignments are returned by email to tada14@mpi-inf.mpg.de
- DL is 1600 hours on the stated day
  - **No** delays, no excuses, time based on the mail time stamp
- We'll acknowledge the submission that we receive before the lecture on the DL day

On assignments: grading

- Assignments are **not** for repeating what the papers say
  - We've read the papers already
- We expect you to discuss **and criticize** the sources, build connections, point out differences, provide new insights, etc.
- Some assignments are marked *hard*
  - This is taken into account when grading

First assignments

1. Did Tukey invent Data Minig?
2. (Don’t) Believe the Hype
3. Big Data: The Best Thing since Sliced Bread or just Another Bottle of Snake Oil?
4. Where did the Candidates Go? (*Hard*)

http://resources.mpi-inf.mpg.de/d5/teaching/ss14/tada/assignments/1.html