EDM AND THE 4TH PARADIGM OF SCIENTIFIC DISCOVERY

Reflections On The 2010 KDD Cup Competition

John Stamper
Human-Computer Interaction Institute
Carnegie Mellon University

Technical Director
Pittsburgh Science of Learning Center DataShop

LearnLab
Pittsburgh Science of Learning Center

eScience

Jim Gray – the 4^{th} paradigm

Jim Gray (computer scientist)

From Wikipedia, the free encyclopedia

James Nicholas "Jim" Gray (born 12 January 1944, lost at sea 28 January 2007) was an American computer scientist who received the Turing Award in 1998 "for seminal contributions to database and transaction processing research and technical leadership in system implementation."

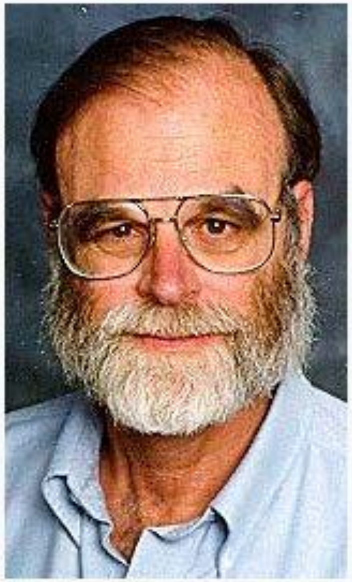
Contents [show]

Family and education [edit]

Gray was born in San Francisco, California, the second child of a mother who was a teacher and a father in the U.S. Army; the family moved to Rome where Gray spent most of the first three years of his life, learning to speak Italian before English.[2] The family then moved to Virginia, spending about four years there, until Gray's parents divorced, after which he returned to San Francisco with his mother.[2] His father, an amateur inventor, patented a design for a ribbon cartridge for typewriters that earned him a substantial royalty stream.[2]

After being turned down for the Air Force Academy he entered the University of California, Berkeley as a freshman in 1961, paying $67 per semester.[2] To help pay for college he worked as a co-op for General Dynamics, where he learned to use a Monroe calculator; discouraged by his chemistry grades, he left Berkeley for

James Nicholas "Jim" Gray



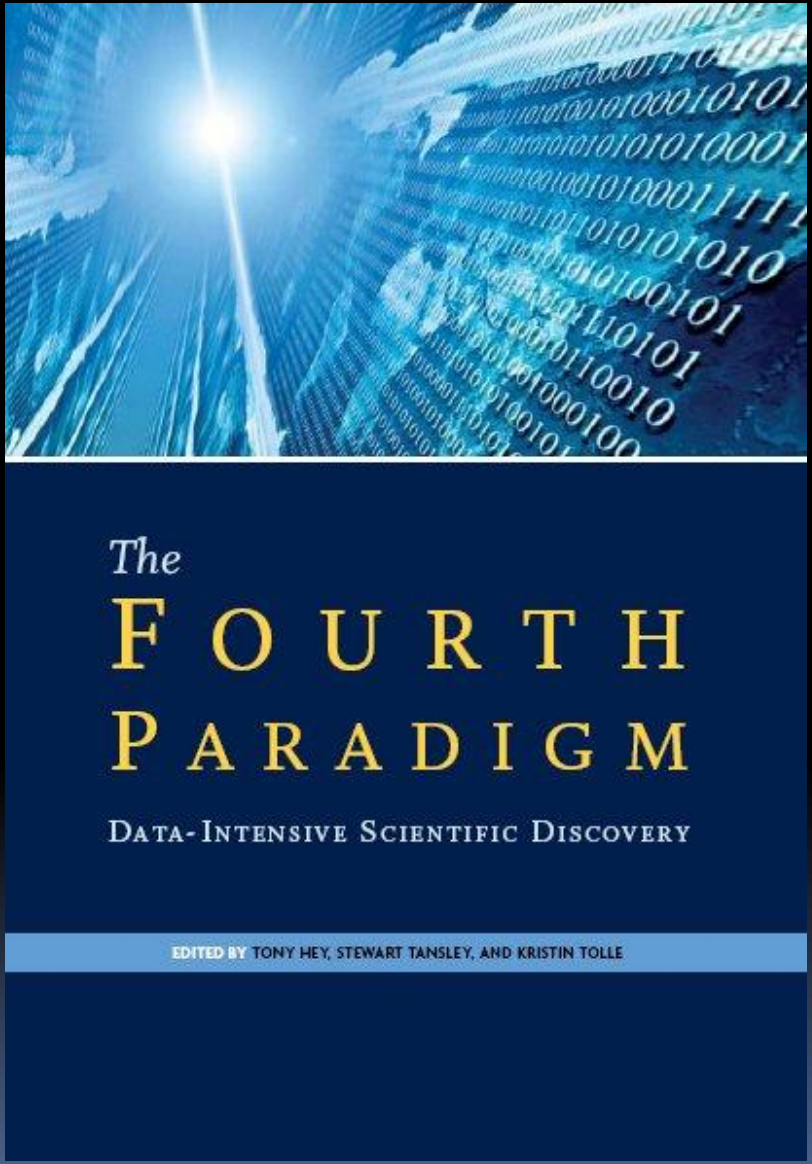
Born	January 12, 1944[1] San Francisco, California[2]
Died	**(lost at sea)** January 28, 2007
Nationality	American
Fields	Computer Science
Institutions	IBM, Tandem Computers, DEC, Microsoft
Alma mater	University of California, Berkeley
Doctoral advisor	Michael Harrison[2]

http://en.wikipedia.org/wiki/Jim_Gray_(computer_scientist)

Paradigms of Scientific Exploration

- Empirical – started thousands of years ago
- Theoretical – last few hundred years
- Computational – last 30 – 40 years
- Data Exploration (eScience)

The Book



http://www.fourthparadigm.com

Data Exploration

- Driven by the availability (or overabundance) of data
- Ties simulation with data analysis, highly statistical
- Requires tools to collect, analyze, and visualize large data sets

Data Exploration

Focus Areas

- Health (Medicine, DNA)
- Environmental (Global Warming)
- Astronomy (Galaxy Mapping)
- Physics (CERN)

Education is missing

http://www.fourthparadigm.com

Can EDM be part of eScience?

We need:

- Data
- Tools
- Ideas and methods

EDM Data Size

What is the right size for EDM discovery?

Data Granularity

Finest – Transaction ⟵ We are mostly here

Steps

Problems

Units

Tests

Class Grades

Class Avgs

Schools ⟵ Policy is being made here

Coarsest - ....

EDM Conference Data

2010

- Average 520 Students
- Median 148 Students
- Largest 172,000 Transactions

2009

- Average 1,168 Students
- Median 300 Students
- Largest 437,000 Transactions

How about 2011?

- Hypothesis – Average will be larger due mainly to a few large datasets

Trend towards larger data sets…

- … and they are coming!
- Carnegie Learning / Assistments
- Seeing a move from collecting data to secondary analysis
- This is good, but it has risks!

Risks of Secondary Analysis

- Misunderstanding the data
- Stagnation on a few datasets
- Privacy/Security

Minimizing the risks

- Misunderstanding the data – Standard formats
- Stagnation on a few datasets – turn on the flow
- Privacy/Security – must have reasonable procedures to protect student identity

Warning – Shameless Plug Ahead!!!

Standard Repositories

- Repositories like DataShop are one way to mitigate these issues and provide:
  - Standardization
  - Privacy/Security
  - Lots of data

DataShop Stats…

DataShop - How to increase awareness?

- Tutorials/Workshops
- Press/media
- Competitions

2010 KDD Cup Competition

- KDD Cup is the premier data mining challenge
- 2010 KDD Cup called "Educational Data Mining Challenge"
- Ran from April 2010 through June 2010

2010 KDD Cup Competition

- The challenge asked participants to predict student performance on mathematical problems from logs of student interaction with Intelligent Tutoring Systems.

facebook

Carnegie Learning

Association for Computing Machinery

PSLC DataShop
a data analysis service for the learning science community

ELSEVIER

IBM Research

KDD Cup Competition

Why do we care?

- Advances in prediction
- Advances modeling

Prediction

- Prediction of student performance is the reason for assessment.
- Tons of effort placed on Standardized Testing
- What if we could predict from student data better?

Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI). 19(3), pp. 243-266.

Modeling

- Student Models drive many of the decisions for adaptive instruction
- What level of granularity should these models be?
- Better Student Models should lead to faster learning

The Data

Data was provided by Carnegie Learning Inc

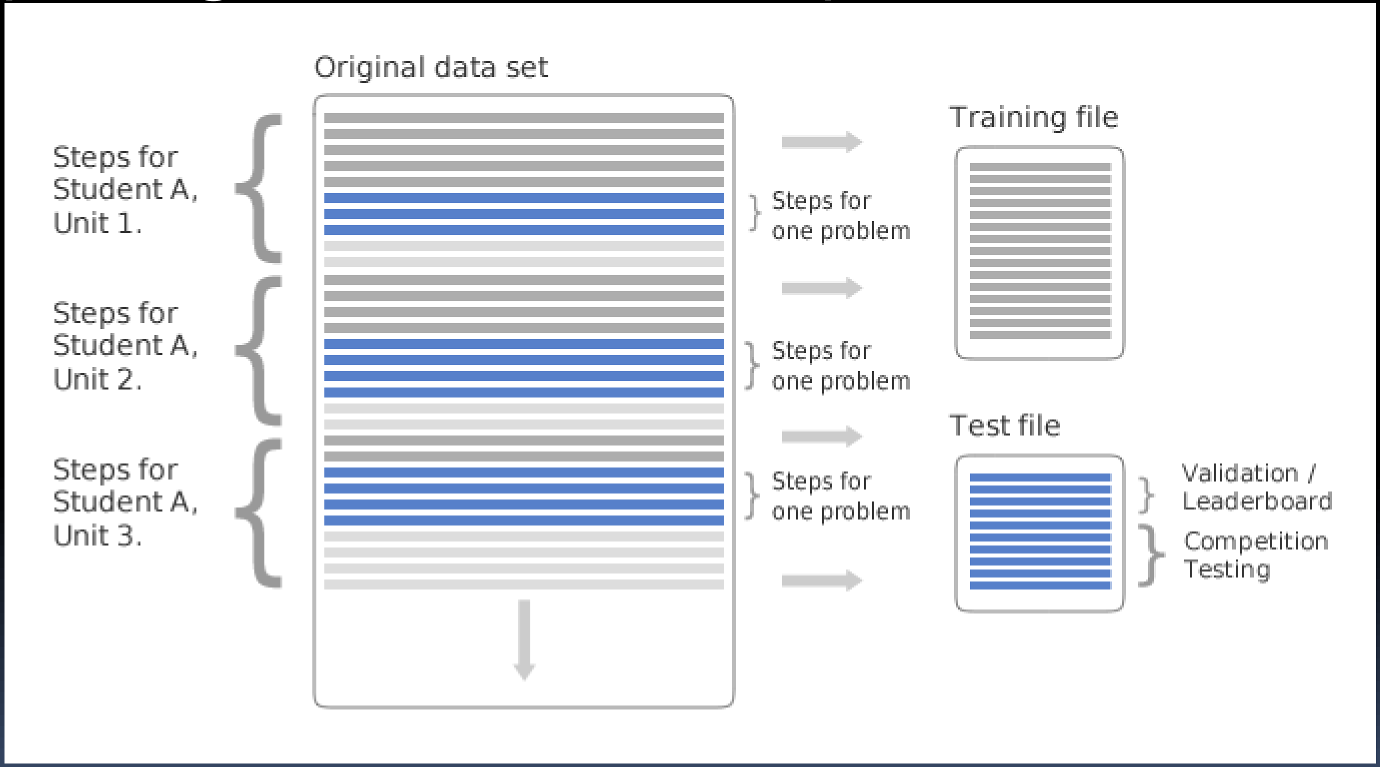
Dataset	Students	Steps	File size
Algebra I 2008-2009	3,310	9,426,966	3 GB
Bridge to Algebra 2008-2009	6,043	20,768,884	5.43 GB

Details on the Data

Row	Student	Problem	Step	Incorrects	Hints	Error Rate	Knowledge component	Opportunity Count
1	S01	WATERING_VEGGIES	(WATERED-AREA Q1)	0	0	0	Circle-Area	1
2	S01	WATERING_VEGGIES	(TOTAL-GARDEN Q1)	2	1	1	Rectangle-Area	1
3	S01	WATERING_VEGGIES	(UNWATERED-AREA Q1)	0	0	0	Compose-Areas	1
4	S01	WATERING_VEGGIES	DONE	0	0	0	Determine-Done	1
5	S01	MAKING-CANS	(POG-RADIUS Q1)	0	0	0	Enter-Given	1
6	S01	MAKING-CANS	(SQUARE-BASE Q1)	0	0	0	Enter-Given	2
7	S01	MAKING-CANS	(SQUARE-AREA Q1)	0	0	0	Square-Area	1
8	S01	MAKING-CANS	(POG-AREA Q1)	0	0	0	Circle-Area	2
9	S01	MAKING-CANS	(SCRAP-METAL-AREA Q1)	2	0	1	Compose-Areas	2
10	S01	MAKING-CANS	(POG-RADIUS Q2)	0	0	0	Enter-Given	3

Details on the Data

Splitting Data for the Competition



2010 KDD Cup Competition

- 655 registered participants
- 130 participants who submitted predictions
- 3,400 submissions

Final submissions of all teams with a fact sheet

Rank	Team Name	Cup Score	Leaderboard Score	Final Submission Time
1	National Taiwan University	0.272952	0.276803	2010-06-08 23:46:50
2	Zhang and Su	0.273692	0.276790	2010-06-08 23:39:35
3	BigChaos @ KDD	0.274556	0.279046	2010-06-07 03:48:20
4	Zach A. Pardos	0.276590	0.279695	2010-06-08 21:31:07
5	Old Dogs With New Tricks	0.277864	0.281163	2010-06-08 23:49:11
6	SCUT Data Mining	0.280476	0.284624	2010-06-08 23:25:27
7	pinta	0.284550	0.289200	2010-06-08 22:14:55
8	DMLab	0.285977	0.291296	2010-06-08 19:37:50

Solutions
1st National Taiwan University

- Used a DM course around 2010 KDD CUP
- Expanded features by various binarization and discretization techniques
- Resulting sparse feature sets are trained by logistic regression (using LIBLINEAR)
- Condensed features so that the number is less than 20.
- Final submission used ensemble by linear regression.

Solutions
2nd Zhang and Su

- Used combination of techniques
  - Gradient Boosting Machines
  - Singular Value Decomposition

- Combined results of multiple SVDs which is called Gradient Boosting.

Solutions
3rd Big Chaos @ KDD

- Used collaborative filtering techniques
  - Matrix Factorization
  - Factorize student/step/group relationships
- Other Baseline Predictions
- Neural network combines an ensemble of predictions
- Originally developed for the Netflix competition

Solutions
4^{th} Zach Pardos

- Used a novel Bayesian HMM
  - learns individualized student specific parameters (prior, learn rate, guess and slip)
  - uses these parameters to train skill specific models.
- The bagged decision tree classifier was the primary classifier
- Bayesian model was used in ensemble selection to generate extra features for decision tree classifier

What did we learn?

- The top teams used very different techniques to achieve similar results
- More work still needed to bring these techniques into the mainstream
- How good does the prediction have to be?

2010 KDD Cup Benefits

- Advances in prediction and student modeling
- Excitement in the KDD Community
- The datasets are now in the "wild" and showing up in non KDD conferences
- Competition site is still up and functioning! (including facts and papers from winning teams!)

2010 KDD Cup Competition

Next steps to continue momentum?

2012 EDM Cup Competition!

Goals

- Generate Excitement within the EDM Community
- Use as a bridge to connect KDD, LAKS, EC-TEL, AERA, etc.
- Make the competition annual
- Have each year build on knowledge gained from previous year
- Vary the questions and data

The Future of EDM

- More and more data will come
- It needs to be mined

- EDM as a community or conference?

EDM Data Size

- What is the right size for EDM Discovery?

PSLC DataShop

a data analysis service for the learning science community

Free Data is there,

Use it!

Make Discoveries!

http://pslcdatashop.org