25 July 1998

Crossing the Chasm: From Academic Machine Learning to Commercial Data Mining

Ronny Kohavi
Silicon Graphics, Inc.
ronnyk@sgi.com

Outline

- Two examples of technology transfer: CART and MLC++.
- Definitions and Market sizes. Can 70 data mining companies survive?
- Commercial problems and areas where research can help.
- What works (+ demo).
- Beer & Diapers and other stories.
- Summary

The CART Story

- 1970's: Breiman, Friedman, Olshen, Stone worked on trees. Wrote the book in 1982.
- 1980's: little impact until machine learning popularized it.
- 1984: Authors contributed $5K each, forming California Statistical Software.
- Mid 1980's: Hired Neville as programmer to cleanup and write version 2.0 (still Fortran). Nothing came out.
- SYSTAT compatible CART was developed.
- Early 1990's: Salford systems releases a PC version of CART. Overlay problems for 640K. Salford's revenues are mostly from consulting.
- Original authors still get some royalties.

The MLC++ Story

- 1993: several PhD students at Stanford are frustrated by reading many ML papers that tweak algorithms. Each time the author beats other algorithms on two out of three datasets.
- June 1993: MLC++ project started with the goals of supporting the following operations*:
  - Implement and test new ideas, variations on existing algorithms, hybrid algorithms, and hierarchical algorithms.
  - Generate performance statistics, such as accuracy, learning rates, confusion matrices.
  - Compare n algorithms on m different datasets.
  - View learned structures graphically.
  - View the learned concept and differences from target concept when it is known (GLD).

(*) Verbatim copy of MLC++ talk at Stanford from June 1993.

The MLC++ Story (II)

- Nilsson funded a master student during summer.
- ONR and NSF gave some money. NSF requested that it be made public domain and on the web.
- Over 15 students worked on project as course CS229B or independent projects at Stanford, several over multiple quarters.
- 1995: SGI adopts technology for MineSet™.
- 1996: MineSet releases with MLC++ algorithms.
- 1998: Over 6,000 MineSet licenses used at almost 1,000 universities.

Data Mining (Knowledge Discovery)

Knowledge discovery is iterative. As you uncover "nuggets" in the data, you learn to ask better questions

Generalize to the future

The non–trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

–– Fayyad, Piatetsky–Shapiro, Smyth [1996]

Not something we already know

Process leads to human insight.

For our task. Actionable

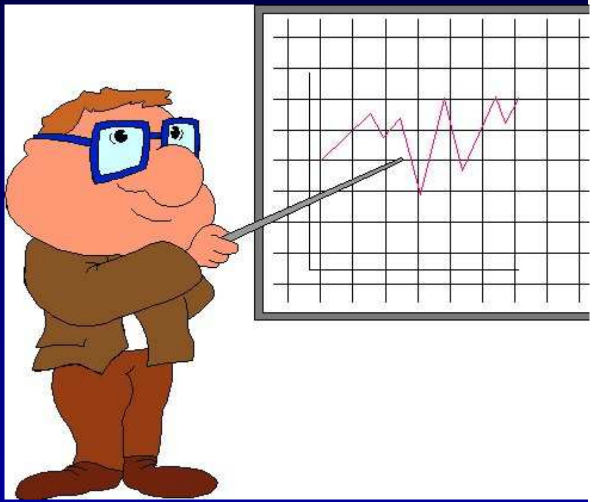
Pattern from Census Bureau data (adult at UCI repository):

If Relationship=Husband then Sex= male (prob=99.6)

Market Size Estimates

META Group

- **Based on a Meta Group report from May 1997**
  Data Mining Market Trends: A Multiclient Study



- **Data mining software market is \$206M in 1996, expected to grow to \$790M by 2000.**

- **Caveats:**
  - **Global 2000 enterprises surveyed–>skewed.**
  - **Includes SAS, which by most definitions did not have a data mining product in 1996.**
  - **Size in 1997 was lower than predicted.**
  - **Estimates are 2–4 times higher than IDC.**

Interesting Excerpts

- 80% of Global 2000 enterprises recognize that data mining will be a critical success factor by 1999.
- Data warehousing is no longer a secret weapon... Data mining represents the next formerly secret weapon to become commercialized and available to the corporate middle–class.

Database survey by Winter Corp (1998) showed UPS with 16TB (OLTP). Sears and Wall–mart have over 4TB for Decision Support.

Is this Market Big?

Is $790M for data mining tools a big market?

Same study claims that in year 2000 sizes for data mining related activities are:

- Data services (data providers): $4.7B
- System integrators (SW/HW/Prof serv): $1.8B
- Packaged solutions: $1.1B (70% CAGR).

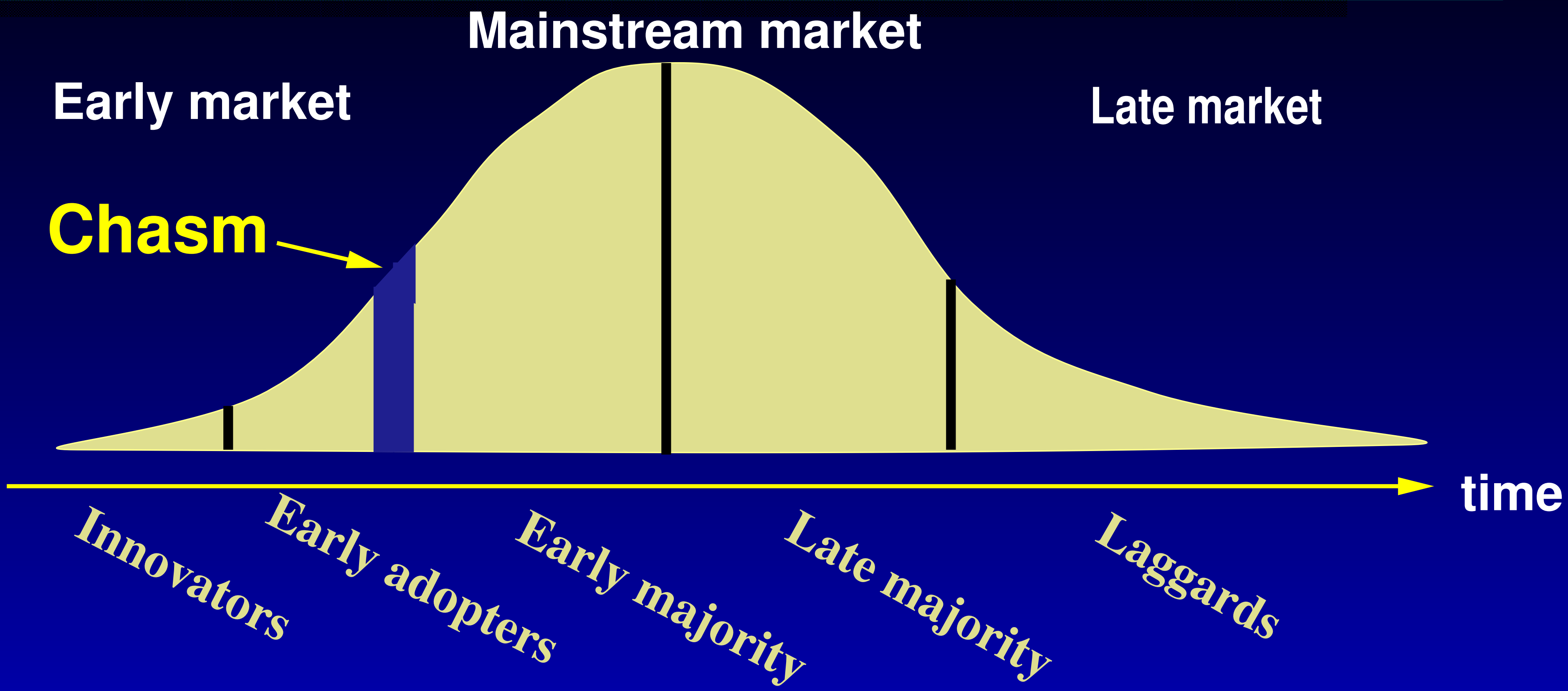
Palo Alto Management Group's study on large–scale database solutions claims that the VLDB market is $8.8B in 1996, growing to $70B in 2001.

Who is Attacking the Market?

- There are over 70 companies in the data mining space today.
- On data mining software alone, all but a couple are estimated to be losing money. Most are eating their VC money and consult to survive.

- Large companies (e.g., IBM, SGI) use data mining software as a loss–leader. The software pulls hardware/consulting, which is a very profitable business.

Technology Adoption Life Cycle



Innovators love technology. Early adopters are visionaries that see a major impact from the new technology. The Early majority needs a "whole" product, trusted suppliers, and clear ROI.

Crossing the TALC Chasm

To cross the chasm* you must have a value proposition to overcome aversion to new innovations, including *all* of:

1. Provide dramatic competitive advantage.
2. Radically improve productivity on a critical success factor.
3. Visibly, verifiably, and significantly reduce the total operating costs.

(*) Geoffrey Moore, Crossing the Chasm.

The Value Proposition

What is the value proposition of a data mining system and who is the end user?
Typical answers:

1. Decision support system for ad hoc analysis.
   User is the business user (sales/marketing).
   Problem: many questions are OLAP/reporting.

2. Prediction/scoring system.
   User is the analyst who builds the model.
   Problem: very hard to beat specialized systems and experienced analysts. Not enough users to make enough ale.

General Suggestions for the Entrepreneurs

- Team with business people that understand marketing, sales, distribution channels.
- Build a vertical solution (e.g., manufacturing, warranty DB) and use the end user's vocabulary. Build an application that maximizes profit instead of minimizing RMS.
- Build a GUI based on the task.
- Develop a scalable architecture (e.g., client/server).
- Integrate with other pieces of the process: databases, cleansing tools, reporting tools, post processing, deployment.

Data Mining Related Suggestions for Entrepreneurs

Don't start yet another horizontal data mining company that sells a machine learning algorithm!

- Few tool vendors will be left in a few years.
- Survivors lose on software R&D by increasing sales in other areas (HW/services).
- Large companies spend \$3–4M per year in data mining engineering effort alone.

Build a vertical solution for a specific industry.

- Establish leadership (beachhead) in one Mainstream market.
- Complete the chain around data mining (collection, cleaning, mining, acting, verifying).

Problems and Suggestions for R&D (I)

- Machine Learning algorithms are a small part of the food chain from data collection to decisions.
  - Suggestion: use technology for "smart" cleansing. The market is huge and could use ML algorithms.

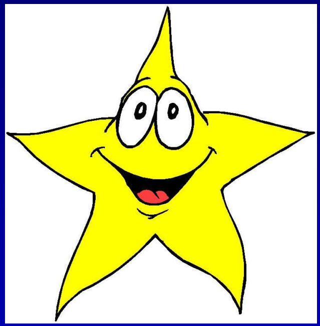
  - Suggestion: Integrate data mining technology with adjacent technologies, such as OLAP (on–line analytical processing) and reporting.

- The i.i.d. assumption commonly made is false. Suggestion: Develop methods to take into account feedback (e.g., marketing campaign).

Problems and Suggestions for R&D (II)

- Transformations are very hard to do today. Suggestion: develop a transformation algebra, similar to sigma–algebra for databases.

- Database theory for OLTP recommends entity relationship diagrams to normalize data. Data Warehouses use Kimball's star schemas. Our algorithms require one denormalized table.



  Suggestion: develop algorithms that take advantage of multiple tables without the need to materialize the join.

Problems and Suggestions for R&D (III)

- Many algorithms do not scale to large data sets.
  Suggestions: develop
  - Parallel/distributed algorithms
  - Active sampling.
  - Anytime algorithms (don't force me to wait hours before realizing that the root is bad).

  I'm pessimistic about out–of–core algorithms. Reading 100GB takes an hour. A multi–pass algorithm will be slow. Memory is cheap.

- Need more use of background knowledge.
  Suggestion: develop methods that utilize meta data (e.g., units) and take advantage of time variables. Harder: non–propositional (e.g., ILP).

What Works Today

- Transformation GUI (e.g., Clementine) Click
- Exploratory visualizations. Shneiderman’s overview, zoom, filter, details on demand.
- Model visualization and what–if analysis (demo).
- Exporting of models for real–time application of models, web distribution etc (deployment).
- Training users. Like driving, you need to learn to be effective and to avoid hitting brick walls.

Model Visualization

- Decision trees built from large datasets are usually large. Dynamic drill–down helps Click
- One of the simplest models is Naive–Bayes. Instead of showing a matrix, a visualization provides tremendous insight
  It also allows what–if analysis. Click
- Decision Tables show the OLAP cube.
  Because the number of low–level cubes grows exponentially fast, an overview with drill–down is important Click

Stories – Beer and Diapers

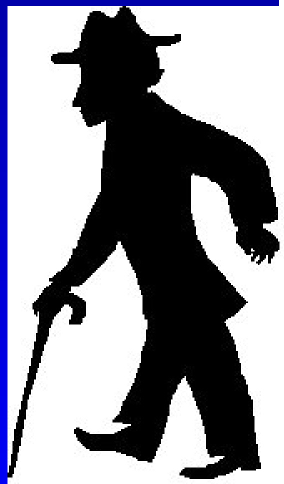
- Diapers and Beer. Most famous example of market basket analysis for the last few years. If you buy diapers, you tend to buy beer.
  - T. Blischok headed Terradata's Industry Consulting group.
  - K. Heath ran self joins in SQL (1990), trying to find two itemsets that have baby items, which are particularly profitable.
  - Found this pattern in their data of 50 stores/90 day period.
  - Unlikely to be significant, but it's a nice example that explains associations well.

Stories – Investment Institution

- A study of mailings from the investment institution showed that older people, particularly over 65, do not respond to IRA offers (Individual Retirement Accounts).

The VP who reviewed the work asked why he was paying good money for such obvious discoveries.

The consultant replied that it is the VP's institute that is sending these offers...



Stories – Cash Management

- A large bank asked to find the factors affecting churn of companies for which they manage cash. Specifically, to characterize which companies are likely to leave.

Among the strongest factors was...

If the account relationship manager is called <name>, over 50% of the clients leave!

Stories – Non–actionable Segment

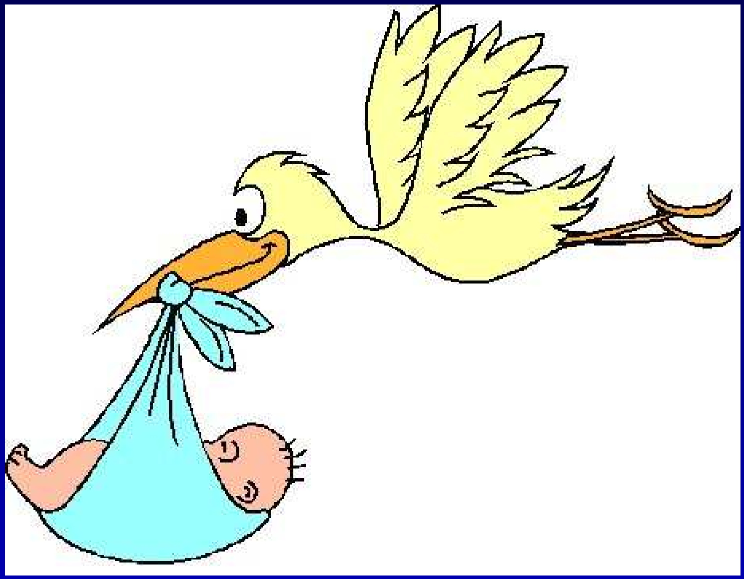
- A bank discovered a cluster of customers that have left the bank:
  - Older than the average customer.
  - Less likely to have a mortgage.
  - Less likely to have a credit card.

They were also...

(*) From Berry and Linoff's Data Mining techniques book.

Stories – The Common Birth Date

A bank discovered that almost 5% of their customers were born on 11 Nov 1911.



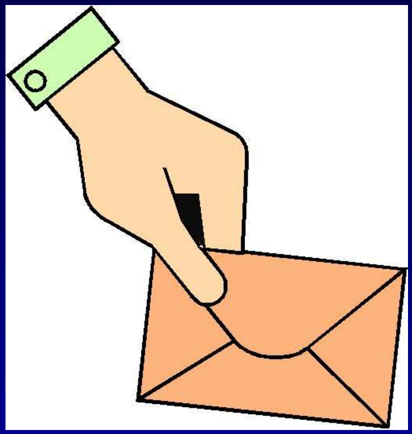
The field was mandatory in the entry system. Hitting 111111 was the easiest way to get to the next field.

Stories – Insurance for Californians

- A health insurance mailing campaign had 100% response rate from California.

  Reason: the mailing never went to California in the first place!

  – People who received the offers would pass them to their family members in other states.
  – Anyone from California that was in the dataset was there because s/he accepted the insurance.



The Dream of Data Mining(*)

Dear Mr. Jones:

We noticed you've not picked up any condoms at your local supermarket recently. (Your last purchase was 8 weeks ago.) Further, you have stopped buying feminine hygiene products, but have sharply increased your frozen pizza and dinners usage in the same time frame.

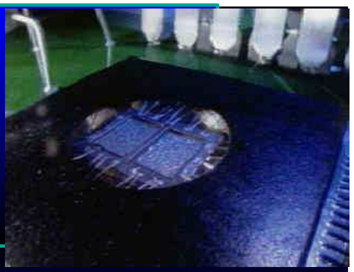
It's clear that Ms. Jody Sanders has dumped you... We confirmed this with the Post Office database––yep, she filed a change of address...

We at XXX International would like to offer you...

When will we be able to do this?

(*) Variant of a post on rec.humor.funny in 1995

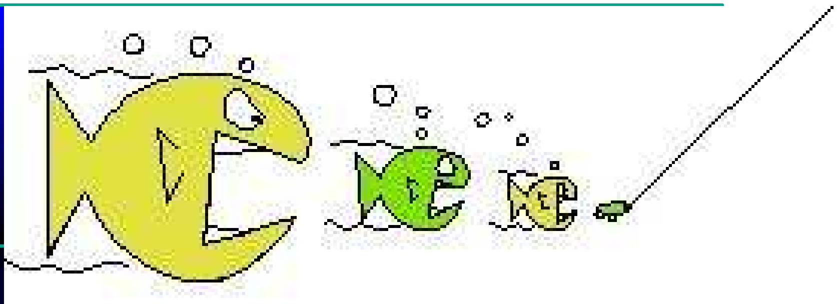
Summary



- Your PCs use 8086 technology not because it was better technology, but because of better marketing–––Intel built everything around the chip and convinced others to "design in" (integrate) with their product (*).
- Commercially, aim for large market share (over 15%) in a well protected market segment that is not too small. Beware of being *slightly better* in one area (e.g., technology) . You may be significantly worse in other areas important to customers.

(*) Bill Davidow, Marketing High Technology

Summary II



- Build a complete solution for the whole chain, not a single algorithm. Attack a vertical market with a solution using the end–user's vocabulary. Beware of starting yet another data mining company that builds decision trees from flat files.

- Think of applications of ML technology to data cleaning.

- For researchers: try to apply the technology to proximity areas, not just prediction: smart data cleansing, OLAP, reporting.

Acknowledgments

- I wish to thank
  - MineSet team
  - Lounette Dyer
  - Herb Edelstein
  - Jerry Friedman
  - Rob Gerritsen
  - George John
  - Karen Heath
  - Yael Kleefeld
  - Padraic Neville
  - Foster Provost
  - Evangelos Simoudis
  - Dan Steinberg
  - David Wolpert
- URL for this talk at
  http://reality.sgi.com/ronnyk/chasm.pdf