Inspired by Data: Yes, you can!

(some personal reflections)

Wojtek Kowalczyk

wojtek@cs.vu.nl

Computational Intelligence Group

&

Fraud Detection Expertise Center

www.fdec.nl

vrije Universiteit amsterdam

Instead of an introduction...

Assignment:

- In a moment you will see some cutting-edge applications of data mining
- How could you use similar solutions in your work?
- After the presentation I will ask you the following questions:
  - what problem do you want to solve?
  - what data could be used for solving the problem?
  - where do you see biggest obstacles?

Agenda

- Fraud Detection:
  - fraud with credit cards: profiles
  - skimming: group behavior
- Recommender Systems:
  - Netflix
  - ECI
- Cows and Navigation Systems: what do they have in common?
- Review of the ASSIGNMENT

Credit Card Fraud in UK (2005)

- 2.000.000 Euro's lost each day !
- A fraud transaction every 9 seconds
- 33% of cardholders affected by fraud
- “ONLY” 0.141% fraudulent transactions

Challenge:

build an intelligent, self-learning system that detects fraud in real-time!

Fraud is difficult to spot:

- No “universal fraud patterns”

(what is normal for one cardholder is unusual for another)

- Fraud patterns changing dynamically

(thieves are clever: action => reaction)

- Huge volumes of data

(hundreds of transactions per second, millions of accounts)

A perfect fraud detection system:

- Tuned to every cardholder
  (each cardholder treated individually)
- Adaptive
  (evolve with slow/small changes in cardholder behavior)
- Fast (real-time)
- High accuracy

A system based on profiles !!!

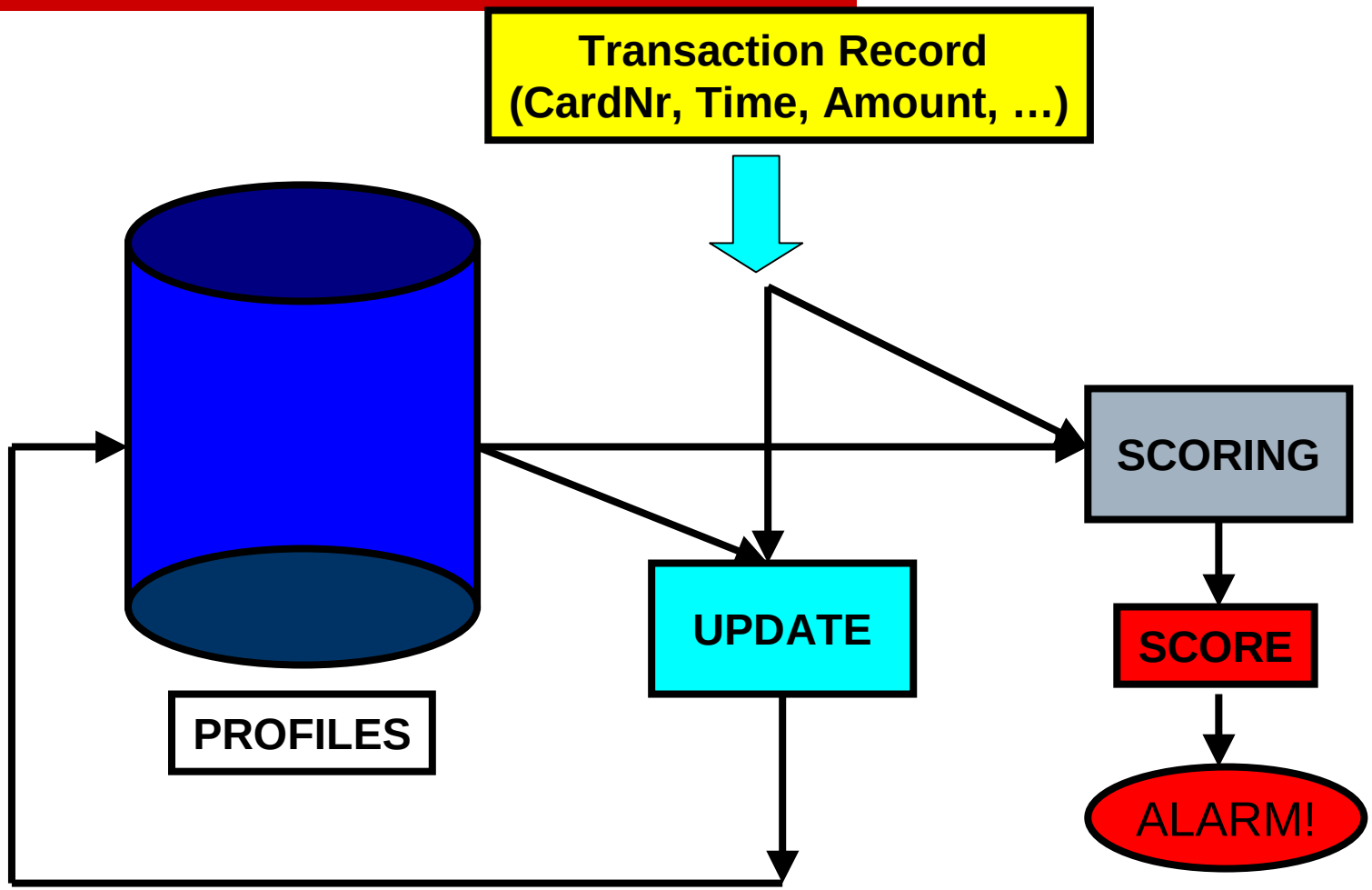
- Every cardholder gets a vector of parameters that describe his/her behavior:
  an "average-behavior" profile

- The system constantly compares this "long-term" profile with the **recent behavior of cardholder**

- Transactions that **do not fit** into cardholder's profile are flagged as suspicious (or are blocked)

- **Profiles are updated with every single transaction**, so the system constantly adopts to (slow and small) changes in cardholders' behavior

How does it work?



Data Mining Challenges

- Identification of profile variables
- Sampling (very skewed distributions)
- Development of the scoring model
- Optimization criterion: what do we optimize
    - # detected fraud transactions?
    - # detected fraud cards?
    - amount of money saved?
    - ...

Yes, we can!

- A powerful system
- No tuning needed (self-learning!)
- Unlimited scalability & speed

(could serve whole China on a single PC!)

- Could be even implemented on card chip!

What about skimming fraud ???

Skimming: card reader + video camera + mobile phone

Fighting Skimming Fraud: traditional approach

- An ATM/POS terminal is compromised
- Cards are skimmed and copied
- Criminals start cleaning compromised accounts
- Some victims realize it and alert banks
- Banks analyze data to find a Common Point of Purchase (a CPP): a single terminal on which reported cards were used together on the same day
- All other cards used on the CPP are blocked
- Usually it's too late ...

More than 200 million Euro's lost per year !!!
(in Europe only)

Challenge: real-time detection!

- Monitor in real time all POS/ATM transactions
- Detect unusual patters and block compromised cards as quickly as possible
- Ideally: block compromised cards **before** fraud is discovered!
- A big question: ***can we do it ???***

- Some numbers:
  - 3000.000.000 transactions per year
  - up to 15.000.000 transactions per day
  - up to 400 transactions per second (peak hours)
  - 100.000.000 cards

Speed is the key !!!

- Maintain a sliding buffer of the last billion transactions in RAM (fast memory)
- Organize the transactions in such a way that some queries could be executed very fast
- Develop some *clever algorithms* that operate on this data structure
- **Will it work??? Yes, it will !!!**

Vooruitdenken in betalingsverkeer

Antoon Kuijpers

Met vestigingen in vier landen – Nederland, Duitsland, Italië en Finland – en de levering van betaaldiensten in acht Europese landen is Equens hard op weg om uit te groeien tot een van de grootste Europese betalingsverwerkers. Vooruitdenken is het credo.

'Het klinkt misschien gek, maar voor ons geldt dat de prijsstelling van onze dienstverlening, onze kwaliteit, betrouwbaarheid en veiligheid 'slechts' randvoorwaarden zijn", vertelt Antoon Kuijpers, Member of the Board of Directors van Equens. Met een jaarlijks volume van bijna negen miljard giraal transacties en een dikke drie miljard rentbankbetalingen en geldopnames – goed voor een marktaandeel van meer dan 12,5% in Europa – kan dat ook haast niet anders.

Kuijpers: "Niet zo belangrijk is het om steeds vooruit te denken. Op welke manier kunnen we in de behoefte van onze klanten, de Europese banken, voorzien? Wat doen onze concurrenten én op welke manier kunnen we ons van hen onderscheiden? Of hoe kunnen we onze ervaring in het betalingsverkeer toepassen op andere sectoren, zoals op dit moment in de gezondheidszorg en het verzekeringswezen?" Antoon Kuijpers, sinds 1994 werkzaam voor Equens (destijds Interpay) en onder meer verantwoordelijk voor de introductie van de Chipknip in Nederland, benadrukt graag hoe dynamisch zijn werkterrein is. "Nagenoeg iedereen binnen Equens is zich bewust van de rol die wij vervullen in het Europese betalingsverkeer. Daarom noemt Business Continuity een belangrijke plaats in. Niet alleen met het oog op een calamiteit, maar bovenal om nog beter te kunnen sturen op de bedrijfsdoelstellingen. Natuurlijk heeft Business Continuity ook haar weerslag op de producten en diensten die klanten afnemen. Zijn er sluitende afspraken gemaakt met onze klanten en worden de klantverwachtingen op een juiste manier gemanaged?" Vooruitdenken is vooruit regeren. Als bedrijf moet je dan ook tijdig maatregelen treffen om je klanten te faciliteren. Kuijpers benadrukt dit: "Equens voldoet aan verschillende ISO-standaards. Zoals de internationaal vermaarde standaard BS25999-2:2007, die wordt uitgegeven door het British Standards Institute. Op het gebied van databescherming behalen we het opgestelde protocol (Directive 95/46 EC) en voor informatiebeveiliging voldoen we aan ISO/IEC 27001:2005."

Slimme technieken

Behalve in Business Continuity speelt Equens al jaren een belangrijke rol bij het opsporen van fraude in het betalingsverkeer. Banken melden fraude van pashouders aan Equens. In opdracht van de banken voert Equens vervolgens een onderzoek uit, waarbij op basis van de meldingen van de verschillende banken een scherpe analyse kan worden gemaakt. De bevindingen worden daarna gemeld aan de individuele banken, zodat zij passende maatregelen kunnen treffen. Deze centrale aanpak blijkt tot nu toe effectief en levert aanzienlijke resultaten op.

Bij zo'n onderzoek en bij het detecteren van fraude maakt Equens gebruik van slimme technieken. Vooruitdenken wordt ook hier beloond, omdat je door het inzetten van nieuwe technieken criminelen een stapje voor kunt zijn. "Wij werken al jaren uitsluitend samen met de Vrije Universiteit van Amsterdam", zegt Kuijpers. De samenwerking heeft betrekking op het in een zeer vroeg stadium detecteren van fraude met betaalpassen, zoals bankpassen en creditcards. Er wordt gebruikgemaakt van neurale netwerken, een techniek die is gebaseerd op de werking van neuronen in de hersenen. Kuijpers: "De Vrije Universiteit heeft enorm bijgedragen aan een nieuw, state of the art en uiterst krachtig systeem voor het vaststellen van fraude." De detectie van fraude richt zich niet alleen op creditcards, maar ook op skimming van bankpassen, en is gebaseerd op kenmerken van de kaarthouder, kenmerken van de transactie, de winkel en de onderlinge samenhang.

'Centrale aanpak van fraude door banken en Equens blijkt effectief'

Miljard transacties

Door de uitgebreide inzet van informatietechnologie is het nu mogelijk om permanent de activiteiten op betaal- en geldautomaten te volgen, gecombineerd zijn dat maar liefst één miljard realtime beschikbare transacties waarvan Equens geavanceerde analyses kan uitvoeren. Het systeem reageert op afwijkende transacties, waardoor passen zeer snel kunnen worden geblokkeerd. Zo wordt het 'window of opportunity' voor criminelen teruggebracht van enkele dagen tot enkele minuten, waardoor mogelijke schade enorm wordt beperkt. En niet te vergeten het ongemak voor een pashouder. Kuijpers: "Binnen Equens volgen we vierentwintig uur per dag, zeven dagen in de week nationale en internationale ontwikkelingen. Overigens niet alleen op het gebied van fraudedetectie. Ook fungeren we als liaison voor de politie bij de opsporing van fraude en zijn we nauw betrokken bij de daadwerkelijke vervolging van daders via justitie en, waar mogelijk, het verhalen van de geleden schade. Onze specialisten nemen onze klanten daardoor veel werk uit handen."

Strategisch partnerschap

Steeds weer signaleren Equens nieuwe kansen in de markt. Terwijl er binnen Europa nog hard wordt gewerkt aan het tot stand brengen van een uniforme Europese betaalzone en het harmoniseren van betaalproducten, is Equens volop bezig met de volgende logische stap op weg naar verdere globalisering van de betaalmarkt. Zo sloten het bedrijf recentelijk een memorandum van overeenstemming met de Federal Reserve Banks voor de verwerking van betalingen tussen de Verenigde Staten en Europa. Met ingang van 2010 bieden beide partijen banken een kosteneffectieve mogelijkheid voor de verwerking van grensoverschrijdende betalingen in meerdere valuta's, waaronder de Amerikaanse dollar en de euro. "Voor Equens is zo'n strategisch partnerschap een belangrijke stap die goed aansluit op onze visie op een mondiale betaalmarkt. Een kwestie van vooruitdenken." ©

"Vooruitdenken in betalingsverkeer"
(Management Team Financials, October 2009; www.fdec.nl)

Antoon Kuijpers (Board of Directors Equens Europe):

"De Vrije Universiteit heeft enorm bijgedragen aan een nieuw, state of the art en uiterst krachtig systeem voor het vaststellen van fraude.

De detectie van fraude richt zich niet alleen op creditcards, maar ook op skimming van bankpassen, en is gebaseerd op kenmerken van de kaarthouder, kenmerken van de transactie, de winkel en de onderlinge samenhang.

Het systeem reageert op afwijkende transacties, waardoor passen zeer snel kunnen worden geblokkeerd. Zo wordt het ‘window of opportunity’ voor criminelen teruggebracht van enkele dagen tot enkele minuten, waardoor mogelijke schade enorm wordt beperkt."

NETFLIX Challenge (2006)

- Biggest on-line DVD movie rental company:
  - 5 million active customers
  - 80.000 movies to choose from
  - ship 2 million disks per day
- How people choose from 80.000 movies???
  - Get feedback: 3 million ratings per day
  - Analyze data and predict preferences:
    1 billion predictions per day
  - The CINEMATCH system: state-of-the-art (in 2006)

Welcome | How It Works | Browse Selection | Start Your FREE Trial | Free Trial Info

Browse Selection

We have virtually every DVD published - from classics and new releases to TV and cable series. You'll be able to choose from over 80,000 DVD titles. In addition, you'll be able to select from over 2,000 instant viewing movies (such as the Matrix, Super Size Me, and Zoolander) to watch instantly on your PC.

New Releases (see more)

Ghost Rider

Breach

The Messengers

Music and Lyrics

Blood Diamond

Action & Adventure (see more)

Casino Royale

Pirates of the Caribbean: Dead Man's Chest

Superman Returns

Mission: Impossible III

Inside Man

Drama (see more)

The Pursuit of Happyness

Babel

The Departed

The Guardian

The Illusionist

Start your FREE TRIAL

SEARCH FOR MOVIES

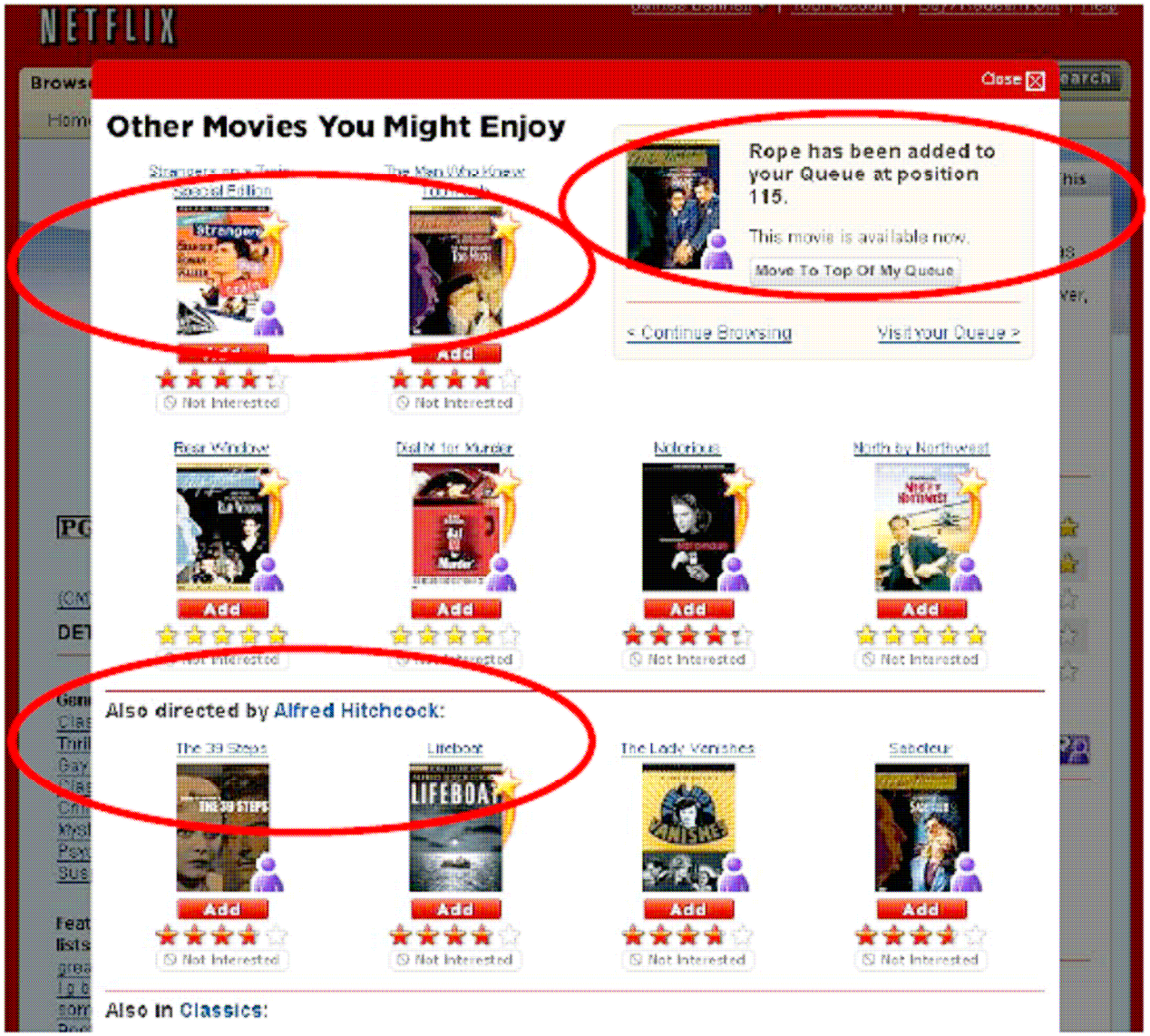
GO

80,000+ Titles
200+ Genres

Browse Our Selection

- New Releases
- Action & Adventure
- Anime & Animation
- Blu-ray
- Children & Family
- Classics
- Comedy
- Documentary
- Drama
- Faith & Spirituality
- Foreign
- Gay & Lesbian
- HD DVD
- Horror
- Independent
- Music & Musicals
- Romance
- Sci-Fi & Fantasy
- Special Interest
- Sports & Fitness
- Television
- Thrillers

Show Interest
Get Recommendations



NETFLIX

NETFLIX in 2010-2012:

- 20 million subscribers
- receive 10 million ratings a day
- generate 5 billion predictions per day
- Movies distributed via Internet
- Accuracy of predictions and speed of the system is crucial for maintaining competitive advantage!

➔ **Announce a $1.000.000 CHALLENGE !!!**

www.netflixprize.com

- **$ 1.000.000 Grand Prize** for a data miner who will improve the accuracy of Netflix recommendation system **by 10%** !!!
- Started in **October 2006**
- To be finished in or before **October 2011**
- **FINISHED in September 2009!!!**
- **A follow-up challenge will start soon…**

The Netflix Challenge

- **100.000.000** rating **records** collected over 1997-2005
- **rating record:**
  <customer_id, movie_id, date, rating>
- **500.000 customers**
- **18.000 movies**
- **rating** = an integer: **1, 2, 3, 4 or 5**
- Additionally, **3.000.000 test records**:
  <customer_id, movie_id, date, ? >

GOAL: fill in "?'s" with numbers, so the error is minimized!

RMSE and percents

- RMSE=Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted - true)^2}$$

- Netflix baseline: RMSE=0.9514
- 1% improvement: RMSE=0.9419
- 5% improvement: RMSE=0.9038
- 10% improvement: RMSE=0.8563

A brief history of the competition

- **October, 2006: Start of the Competition**
- **After 7 days:** Netflix base-line reached !!!
- **After 42 days:** 5% improvement (one page in C!!!)
- January: **6% improvement**
- May: **7% improvement**
- October 2007: **8.46%**
- October 2008: **9.4%**
- **August 2009: 10.04% ($1.000.000; a Team of Teams)**

Main Trick: feature vectors

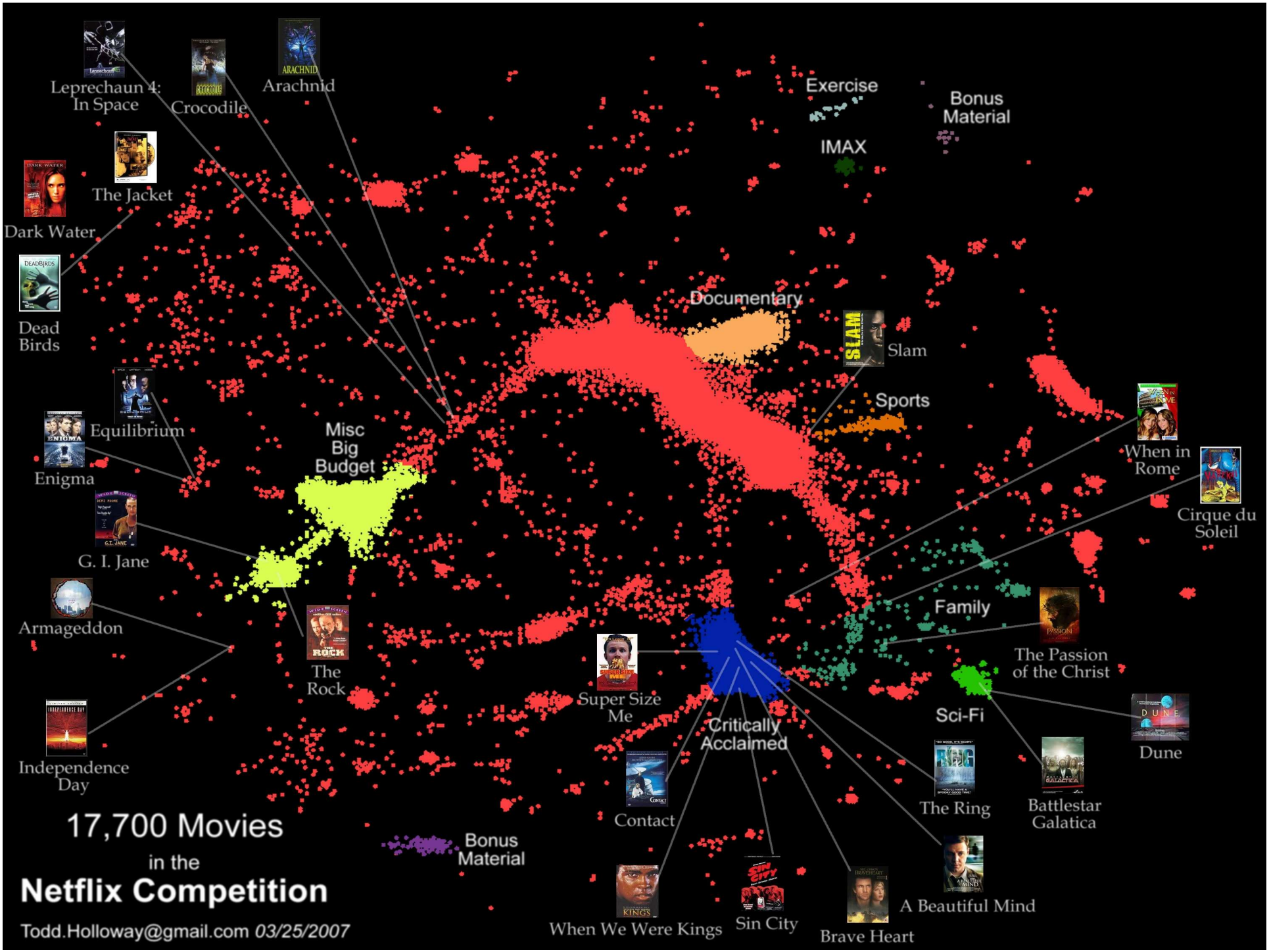
- Assume that each user can be described by 100 numbers (features) and that each movie can also be described by 100numbers, such that the user rating is given by the dot products of these (unknown) numbers:

```
Rating[user][movie] =
sum(userFeature[f][user]*movieFeature[f][movie])
```

- Find optimal values of these 50M unknowns by a simple optimization procedure!

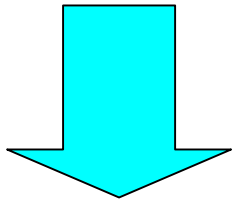
Simon Funk's trick

- Amazingly short code (essential part: 2 lines!)
- Can be run on a laptop with 1GB in a few hours
- Very good results (5% better than Netflix)
- Vector representations can be used for **clustering, visualization, interpretation**, etc.


Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Misc Big Budget
Enigma
When in Rome
Cirque du Soleil
G. I. Jane
Armageddon
The Rock
Family
The Passion of the Christ
Super Size Me
Critically Acclaimed
Sci-Fi
Dune
Independence Day
Contact
The Ring
Battlestar Galatica
17,700 Movies
in the
Netflix Competition
Bonus Material
When We Were Kings
Sin City
Brave Heart
A Beautiful Mind
Todd.Holloway@gmail.com 03/25/2007

Recommender systems:

- **Input:** 10.000.000 purchase records:
  - 500.000 customers
  - 50.000 products
  - purchase record: <customer, product>



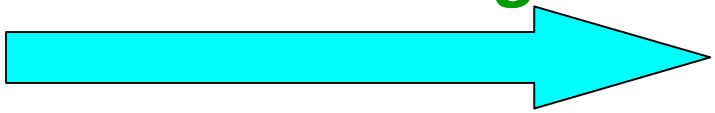
Recommender Algorithm

- **Output:** a matrix 500.000 x 50.000 numbers:
  - for every <customer, product> combination the *probability* that the customer will buy the product

How do they work?

Customer	Product
1	1193
1	661
1	914
1	3408
1	2355
1	1197
1	1287
1	2804
1	594
1	919
1	595
2	457
2	1096
2	920
2	459
2	1527
3	1385
3	3451
3	3095
3	780
500000	535
500000	2010
500000	2011
500000	3751
500000	2019
500000	541

Recommender Algorithm



Customer	Product	Prob
1	345	0.32
1	2364	0.15
1	743	0.07
2	354	0.12
2	6443	0.03
2	12334	0.01
3	432	0.44
3	74321	0.21
3	864	0.18
4	2345	0.16
4	532	0.04
4	2656	0.01
500000	345	0.31
500000	43378	0.3
500000	2350	0.26

What can you do with them?

1) one-to-one marketing

2) N-to-one marketing

3) expected demand for each product

4) best mailing selection for a product (mix)

5) best K-combinations of N-products

6) ...

"best chance of purchase"

or

"best monetary value"

Customer	Product	Prob
1	345	0.32
1	2364	0.15
1	743	0.07
2	354	0.12
2	6443	0.03
2	12334	0.01
3	432	0.44
3	74321	0.21
3	864	0.18
4	2345	0.16
4	532	0.04
4	2656	0.01
500000	345	0.31
500000	43378	0.3
500000	2350	0.26

Problem: Sell 5 Thrillers via Newsletter

- Select clients that will buy any of the 5 thrillers:
  - TOT HET VOORBIJ IS, FRENCH, NICCI
  - ONAANTASTBAAR, SLAUGHTER, KARIN
  - GENIAAL GEHEIM, BALDACCI, DAVID
  - DE ELFDE PLAAG, SMITH, WILBUR
  - DOOD DOOR SCHULD, BEISHUIZEN, TINEKE
- Three selections of about 20.000 clients:
  - Random
  - ECI
  - A Recommender System
- Evaluation:
  - two weeks after mailing, ECI analyzes response

31

Conclusion: The Big Winner!

- Our Recommender System gives better results:
  - Click Through Rate:

    12% (relative), 3.6% (absolute) better than ECI
  - Return per Mail:

    3.4 times better than random

    1.7 times better than ECI
- No background knowledge, no thinking, just data
- Fast (real-time predictions)
- Unlimited Scalability (billions of records)

True power still unleashed: 1-2-1 selections !!!

Assignment!!! (finally ...)

- **How could you use similar solutions in your situation?**
  - What problem do you want to solve?
  - What data could be used for solving the problem?
  - Where do you see biggest obstacles?

Thank you

&

good luck with data mining!