

CIS501 – Data Mining: Finding the Data and Models that Create Value

Woon Wei Lee

Fall 2013, CIS Program, EECS

10:00am-11:15am, Sundays and Wednesdays



For today:

- Introduction (Logistics, Ethics, Rules, etc.)
- Data Mining, what is it, etc..

What is Data Mining?

- **Composite term for describing confluence of ideas from:**

- Statistics (machine learning) and computer science (databases)
- Applied to large databases in science, engineering and business.

- **Gartner group definition:**

“Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques”

- **In a state of flux:**

- Many definitions, lot of debate about what it is and what it is not.
- Terminology not standard
 - e.g. bias, classification, prediction,
 - Feature = independent variable,
 - target = dependent variable,
 - case = exemplar = row.
 - Etc, etc..



Cont'd...

- **Different perspectives:**
 - **Broad Definition** includes traditional statistical methods, applied to large data sets
 - **Narrow Definition** emphasizes automated and heuristic methods (business users)
- **Alternative terms:**
 - “**Big Data**”
 - Analytics
 - Knowledge Discovery in Databases (KDD) (popular amongst AI community)
 - “Data dredging, fishing expeditions” (carries negative connotations)

But, what is it in this course?

- **In part, you get to decide! But..**
 - In general, I will be taking a more statistical/AI perspective on data mining
 - i.e. the “Broad” definition.
 - Not so concerned about business or “infrastructural” aspects
- **Coding will be emphasized (and, generally, is expected)**
 - However, no background on data mining is expected
 - Foundation statistics and calculus also expected
- **Emphasis will be on application of data mining, not development of algorithms**
 - Not the same as a basic understanding

Drivers

- Market: From focus on product/service to focus on customer
- IT: From focus on up-to-date balances to focus on patterns in transactions - Data Warehouses - OLAP
- Dramatic drop in storage costs : Huge databases
 - e.g. Walmart: 20 million transactions/day, 10 terabyte database, Blockbuster: 36 million households
- Automatic Data Capture of Transactions
 - e.g. Bar Codes , POS devices, Mouse clicks, Location data (GPS, cell phones)
- Internet: Personalized interactions, longitudinal data
- **Q: This list is from 2003.. any new ones?**

Four dimensions: Volume, Velocity, Variety, and Veracity.

- **Volume:** Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.
 - Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
 - Convert 350 billion annual meter readings to better predict power consumption
- **Velocity:** Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
 - Scrutinize 5 million trade events created each day to identify potential fraud
 - Analyze 500 million daily call detail records in real-time to predict customer churn faster
- **Variety:** Structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more.
 - Monitor 100's of live video feeds from surveillance cameras to target points of interest
 - Exploit the 80% data growth in images, video and documents to improve customer satisfaction
- **Veracity:** 1 in 3 business leaders don't trust the information they use to make decisions. Establishing trust in data presents a huge challenge as the variety and number of sources grows.

Applications

- Business applications
 - Customer Relationship Management (CRM)
 - Attrition Prediction/Churn Analysis
 - Credit Scoring
 - Targeted Marketing
 - Customer segmentation
 - **Q: Can anyone think of a very, very prominent example?**
 - Fraud detection

- Technical/Scientific applications
 - Health
 - Genome analysis
 - Medical records
 - Disease prediction
 - Engineering
 - Fault detection/diagnosis
 - Security
 - Environmental
 - Patterns in climate data
 - Remote sensing

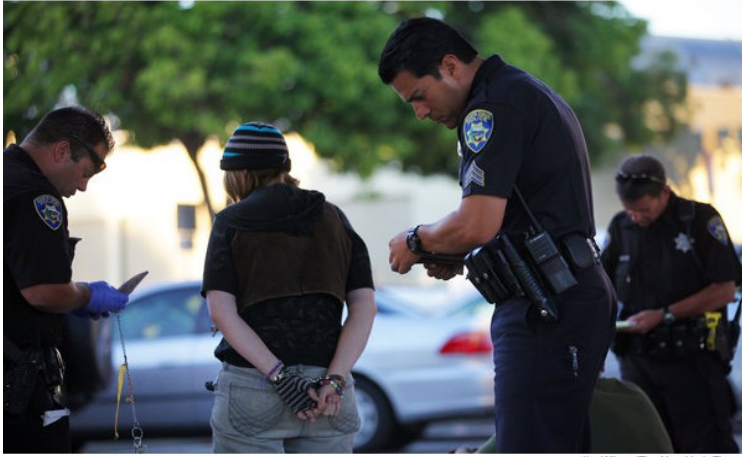
Q: Other applications?

In the news (one good, one bad!)

- Two potential car thieves arrested after police trap in parking garage
- Computer program had predicted that car thefts were likely in that parking structure on that particular day!
- Based on past crime occurrence patterns

Example of an upcoming Deal

Sending the Police Before There's a Crime



Log in to friends + nytimes

What's
Anna H
Hunger
India R
5th Day

Jim Wilson/The New York Times


Officers at a location flagged by a computer program as a place where car burglaries were especially likely put a woman in custody.

by ERICA GOODE
Published: August 15, 2011

The arrests were routine. Two women were taken into custody after they were discovered peering into cars in a downtown parking garage in Santa Cruz, Calif. One woman was found to have outstanding warrants; the other was carrying illegal drugs.

But the presence of the police officers in the garage that Friday afternoon in July was anything but ordinary: They were directed to the parking structure

RECOMMEND
TWITTER
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Times! 

Related
Times Topic: Police
[Enlarge This Image](#)

MARTHA

US\$234.00 - One Year via Credit Card

SUBSCRIBE

NewScientist Tech

Home News In-Depth Articles Blogs Opinion TV Galleries Topic Guides Last Word Subscribe Dating

SPACE TECH ENVIRONMENT HEALTH LIFE PHYSICS&MATH SCIENCE IN SOCIETY

Home | Tech | News

Smartphone jiggles reveal your private data

16 August 2011 by Jacob Aron
Magazine issue 2825. [Subscribe and save](#)


THE slight movements of your smartphone every time you tap on the touchscreen could be giving away what you are typing.

Eavesdropping on a computer user's keyboard input is called keylogging. Often the culprit is software that runs invisibly, tracking what you type and reporting back to the attacker who installed it - making it easy to steal passwords or bank details.

Keylogging is much harder to pull off on smartphones because most mobile operating systems allow only whatever app is on screen to access what you are typing, says security researcher [Hao Chen](#) of the University of California, Davis.

However, Chen and his colleague [Liang Cai](#) have got around that hurdle and created a keylogger that runs on [Android smartphones](#). It uses the phone's motion sensors to detect vibrations from tapping the screen. Since mobile operating systems do not treat the motion-sensor output as private or in need of protection, it presents a target for hackers wanting to create an innocent-looking app that secretly monitors phone users.

PRINT SEND



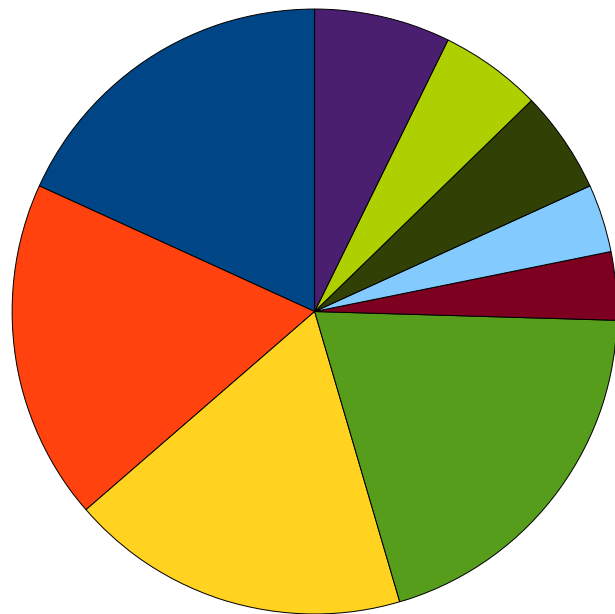
Tapped for information (Image: M. Docher/Plainpicture)

ADVERTISEMENT

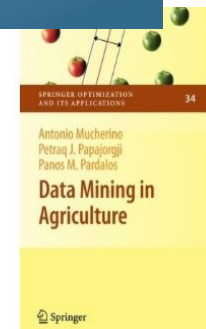
- Creation of key-loggers for Android phones
- Data mining algorithms used to classify subtle movements of the phone, to predict key pressed.

Data mining “Zeitgeist”

- Very unscientific but.... here we go →



■ Health/Biology	■ Science and Engineering
■ Text mining	■ Financial/Business
■ Security	■ Renewable Energy
■ Multimedia	■ Population/Social Studies
■ Web Mining	



What this means..

- Diversity – data is everywhere, and so is data mining!
- Data mining frequently plays role of an *enabling technology*
- Opportunities
 - Lots of applications on Science/Engineering but..
 - relatively few on Renewable energy/Sustainability
- And of course, a little marketing: knowledge and skills in Data Mining are currently in demand 😊
 - “Top ten tech skills” - Data mining is #7 - *NetworkWorld*
 - "The world revolves around data. Anything you can do to develop data analysis, data mining and information on demand skills is incredibly critical" - *Kevin Faughnan, IBM*