Security Analytics
Course Overview

Purdue University

Prof. Ninghui Li

Based on slides by Prof. Jenifer Neville and Chris Clifton

Relationship to Other Security Courses

- This Fall
  - 526 Information Security
  - 555 Cryptography
- Spring 2018
  - 523 Social Econ Legal Asp Of Sec
  - 527 Software Security (?)
  - 528 Network Security
  - 590-DSP: Data Security and Privacy

Relationship to Other Courses

- CS 573 Data Mining
- CS 578 Statistical Machine Learning
- CS 690-DPL Deep Learning

Plan for the Course

- Applied data mining and machine learning techniques, using security problems as examples
- Security and privacy issues in Machine Learning

Topics (1)

- Intro to data mining and machine learning
  - Backgrounds on Probability + Using Python for Data Analytics
- Several algorithms for classification
  - Naïve Bayes, Logistic Regression, SVM, Random Forest
- Neural networks
  - Back propagation, CNN, RNN, tensorflow
- Dealing with large datasets
  - MapReduce, PageRank, and Apache Spark

Topics (2)

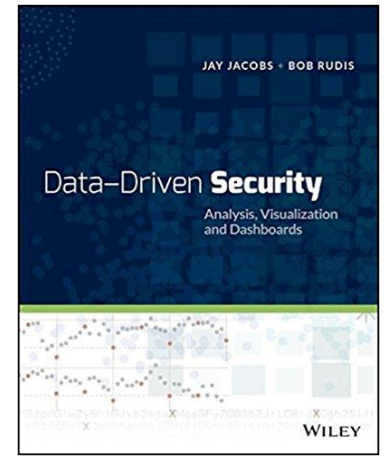
- Application of machine learning in Security
  - Phishing, malware classification, anomaly detection
- Adversarial machine learning
  - Adversarial examples, robustness of ML classifiers,
- Privacy in machine learning
  - Membership inferences, model inversion

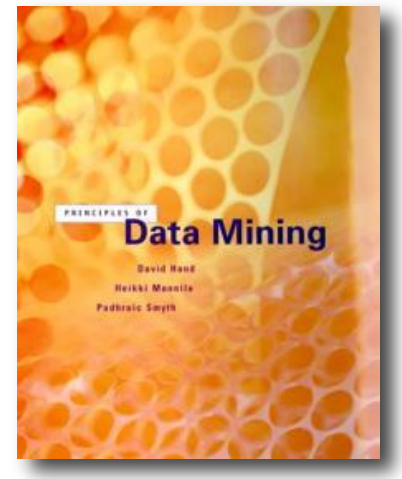
Logistics

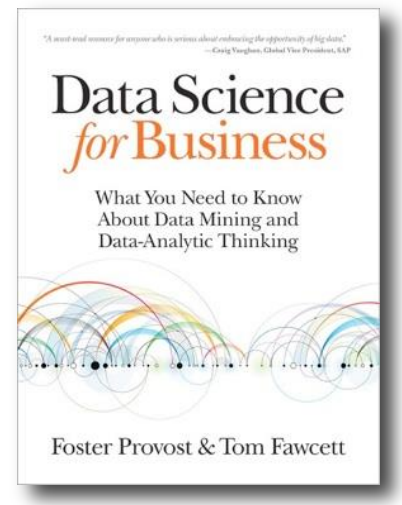
- Time and location: TTh 12:00-1:15pm, LWSN B134
- Instructor: Ninghui Li <ninghui@purdue.edu>,
  - LWSN 2142K, office hours: After lectures and Wed 1:30 to 2:30
- Teaching assistants: Wuwei Zhang <zhan1015@purdue.edu>
  - LWSN 2161, office hours TBD
- Webpage: http://www.cs.purdue.edu/~ninghui/courses/Fall18
- Piazza signup: piazza.com/purdue/fall2018/cs590sa0

Readings

- No required text, readings will be announced/distributed on course webpage.
- Recommended texts
  - Data-Driven Security: Analysis, Visualization and Dashboards by Jay Jacobs, Bob Rudis
  - *Principles of Data Mining*, Hand, Mannila, and Smyth, MIT Press, 2001.
    Available as e-book through Purdue library:
    http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6267275
  - *Data Science for Business*, F. Provost and T. Fawcett, O'Reilly Media, 2013.
    http://data-science-for-biz.com







More Readings

- Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville:
  - https://www.deeplearningbook.org/

- Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman
  - http://www.mmds.org/

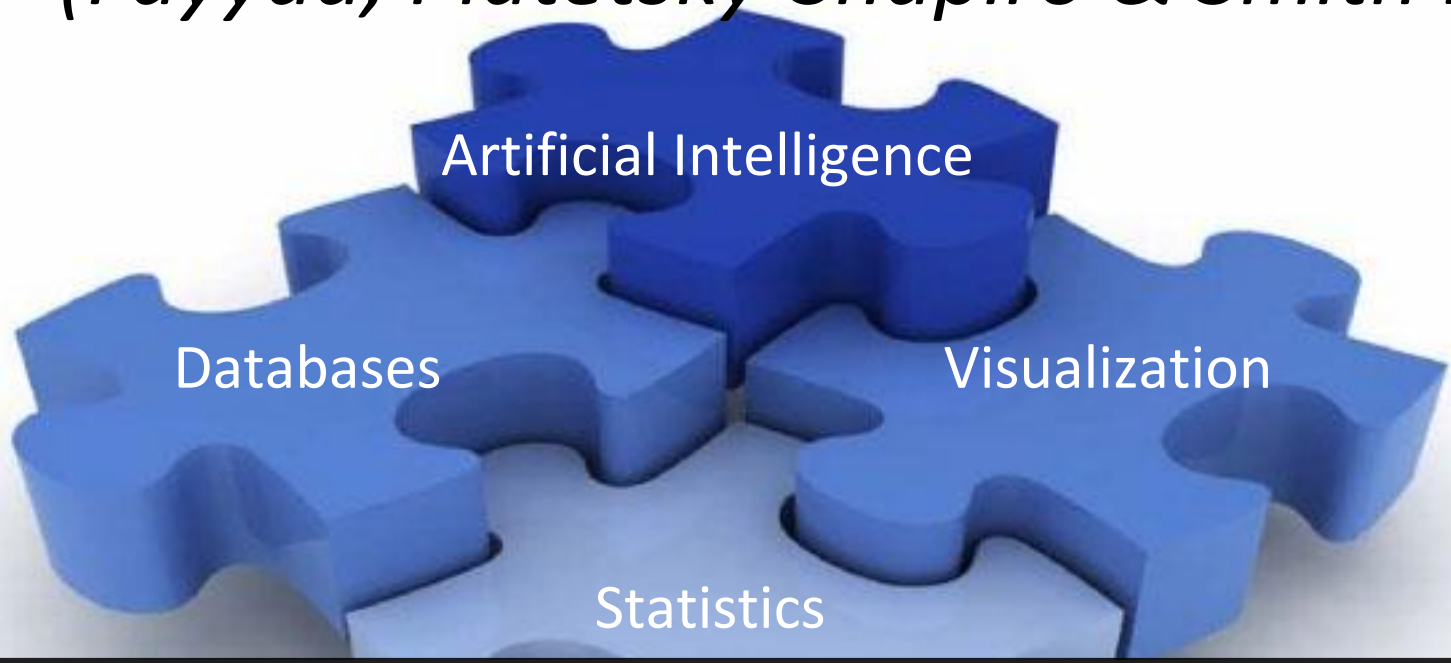
Workload

- Homeworks
  - About 6 assignments, which will be either written assignments, or small projects that require programming
  - Late policy: Five extension days to be used at your discretion
    - Must be stated explicitly in header of work being turned in
    - No fractional days
    - May not be used to extend submission past last day of class.
- Exams
  - 4 (in-class) quizs during the semester
  - A mid-term exam on Oct 16
  - Final exam

Data mining

The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

(Fayyad, Piatetsky-Shapiro & Smith 1996)



Machine learning: How can we build computer systems that automatically improve with experience? *(Mitchell 2006)*

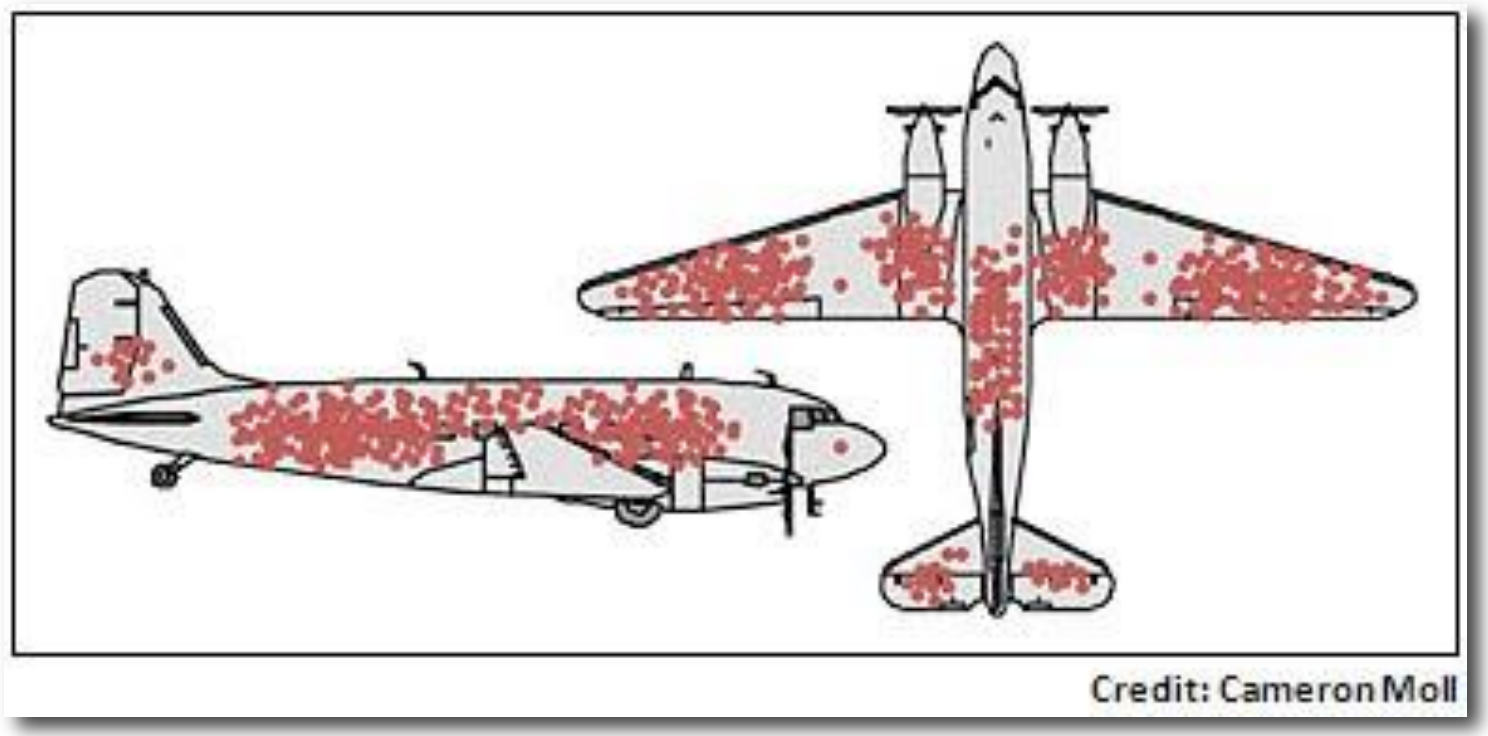
- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.
- The relationships and summaries derived through a data mining exercise are often referred to as *models* or *patterns*. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.
- While novelty is an important property of the relationships we seek, it is not sufficient to qualify a relationship as being worth finding. In particular, the relationships must also be understandable.

Example: John Snow's London Cholera Outbreak Map

- London's 1854 cholera outbreak claimed 14,000 lives
- Two competing theories: Air pollution (Dr. William Farr) and Water Contamination by "special animal poison" (Dr. John Snow)
- Farr uses data with 8 explanatory variables showed relationship between elevation and deaths.
- Snow produced a graph with 13 wells and death tolls, showing concentration of death near one well

Full map: https://www1.udel.edu/johnmack/frec682/cholera/snow_map.png
https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map

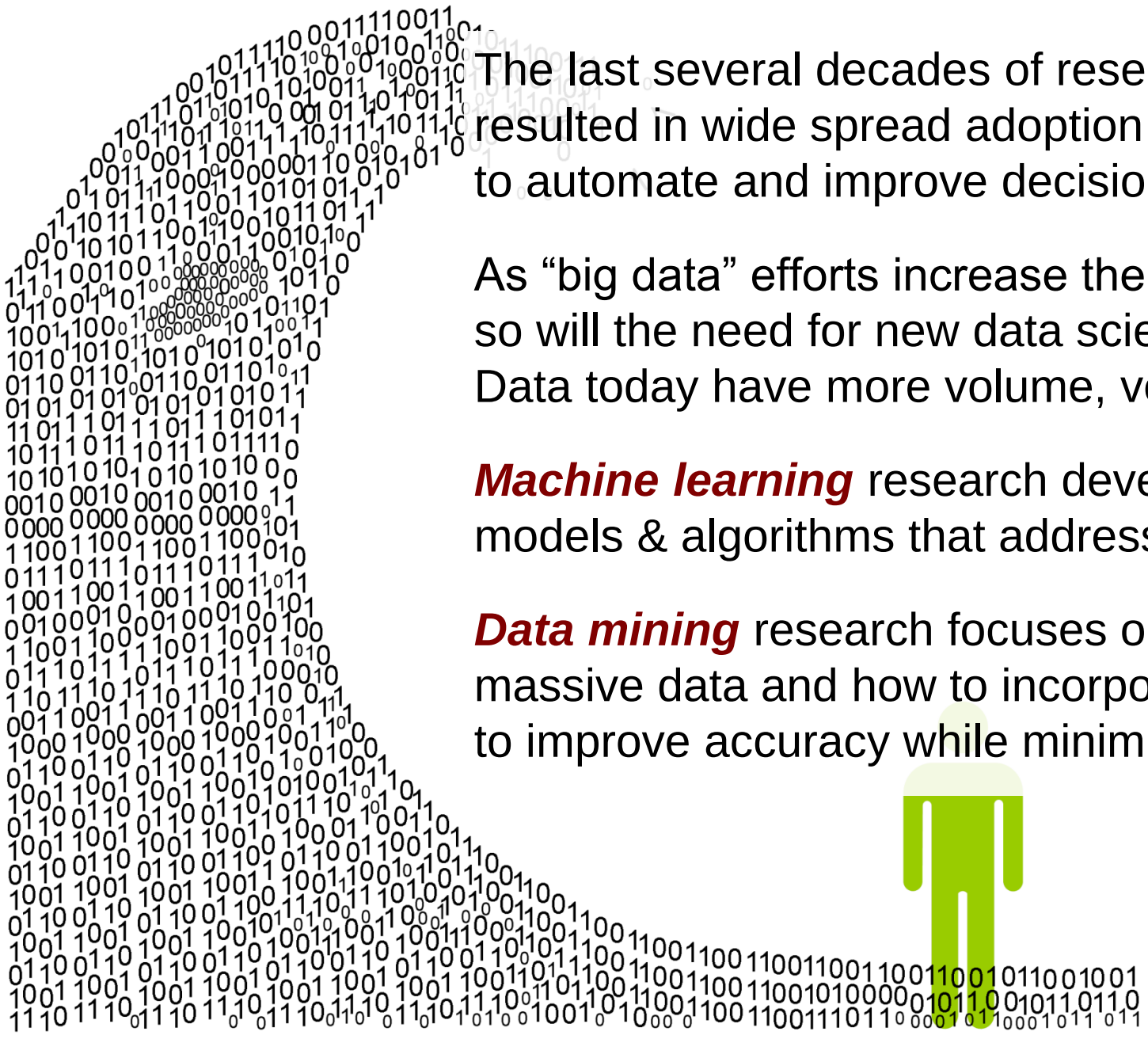
Example: Abraham Wald's Analysis of Planes



During WWII, statistician Abraham Wald was asked to help decide where to add armor to their planes

https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

The data revolution



The last several decades of research in ML/DM has resulted in wide spread adoption of predictive analytics to automate and improve decision making.

As “big data” efforts increase the collection of data… so will the need for new data science methodology. Data today have more volume, velocity, variety, etc.

Machine learning research develops statistical tools, models & algorithms that address these complexities.

Data mining research focuses on how to scale to massive data and how to incorporate feedback to improve accuracy while minimizing effort.

Bringing big data to the enterprise

#ibmbigdata

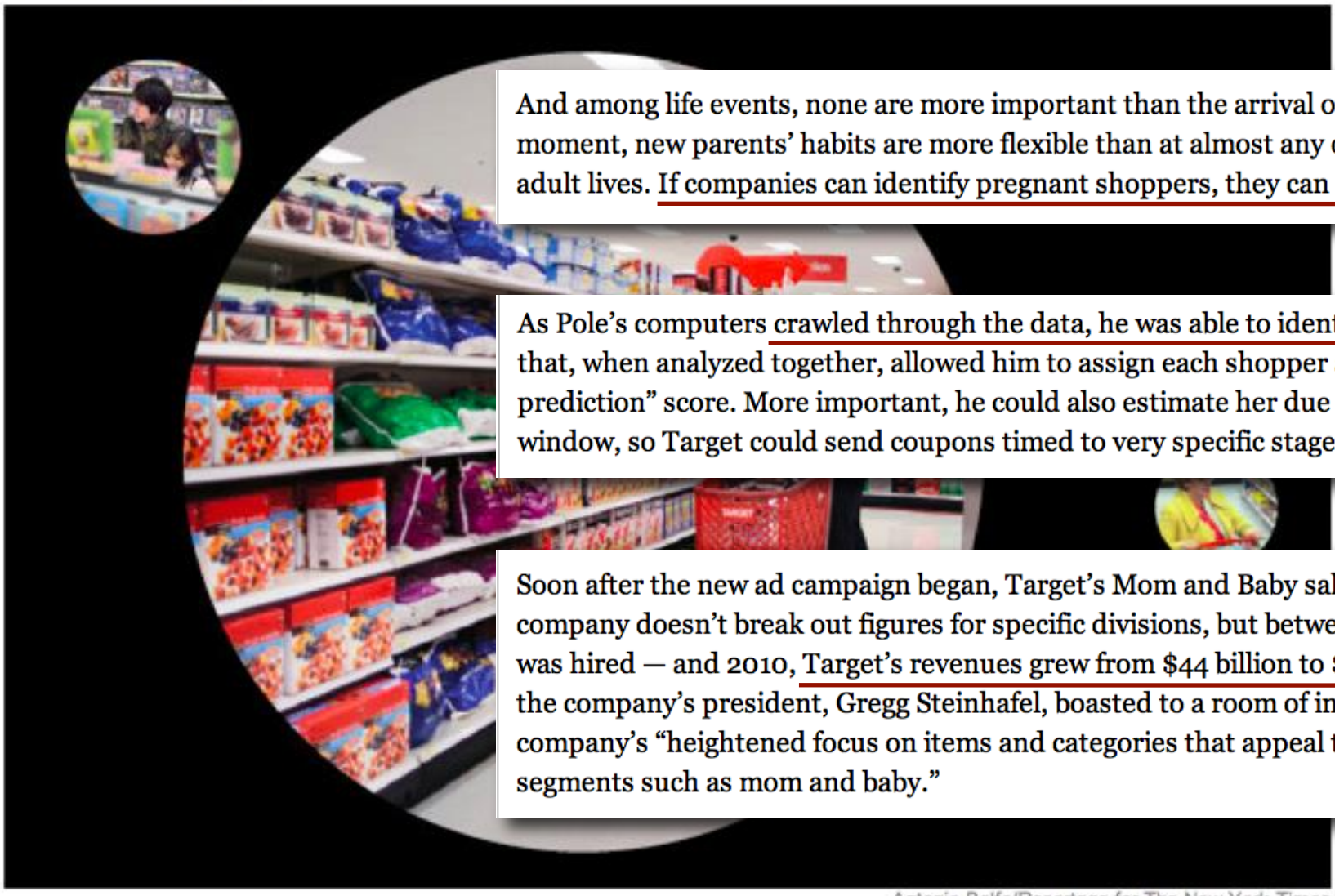
What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**

The New York Times

Magazine

How Companies Learn Your Secrets



Antonio Bolfo/Reportage for The New York Times

And among life events, none are more important than the arrival of a baby. At that moment, new parents' habits are more flexible than at almost any other time in their adult lives. If companies can identify pregnant shoppers, they can earn millions.

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Soon after the new ad campaign began, Target's Mom and Baby sales exploded. The company doesn't break out figures for specific divisions, but between 2002 — when Pole was hired — and 2010, Target's revenues grew from $44 billion to $67 billion. In 2005, the company's president, Gregg Steinhafel, boasted to a room of investors about the company's "heightened focus on items and categories that appeal to specific guest segments such as mom and baby."

By CHARLES DUHIGG

Published: February 16, 2012 | 570 Comments

Skills for a Data Scientist/Analyst

- Domain expertise
- Data management
- Programming
- Statistics
- Visualization

Where are Security Analytics Used by Enterprises?

- Assessing risk
- Identifying malicious behavior
- Meeting compliance mandates

Table 1. Systems, Services and Applications Used for Data Collection Today

Systems, Services and Applications	Response
Application information (event logs, audit logs)	86.3%
Network-based firewalls/IPS/IDS/UTM devices	82.5%
Vulnerability management tools (scanners, configuration and patch management, etc.)	77.6%
Endpoint protection (MDM, NAC, log collectors)	72.0%
Host-based anti-malware	70.6%
Dedicated log management platform	65.0%
Whois/DNS/Dig and other Internet lookup tools	62.4%
Security intelligence feeds from third-party services	60.9%
Network packet-based detection	60.3%
SIEM technologies and systems	59.8%
Intelligence from your security vendors	58.6%
Host-based IPS/IDS	57.1%
Relational database management systems (transactions, event logs, audit logs)	53.4%
ID/IAM (identity and access management) systems	50.1%
User behavior monitoring	41.7%
Network-based malware sandbox platforms	41.4%
Cloud activity/Security data	36.2%
Management systems for unstructured data sources (NoSQL, Hadoop)	24.8%
Other	4.7%

SANS 2016 Security Analytics Survey

What are Concrete Security Applications for Data Analytics?

- Intrusion detection
  - Network-based, host-based
  - Insider threats
- Malicious entity identification
  - Spam/phishing emails
  - Phishing websites/websites delivering malwares
  - Malwares
  - IP addresses controlled by malicious parties
- Enhance security technology (such as authentication)
- Situation awareness
- Identifying vulnerabilities in code, systems, etc.
- …

Security Analytics

- In Which Ways is Analytics in Security Different from Data Mining/Machine Learning
  - Against intelligent adversaries

Readings for Topic 1

- Chapter 1 of *Principles of Data Mining*
- Chapter 1 of Data-Driven Security