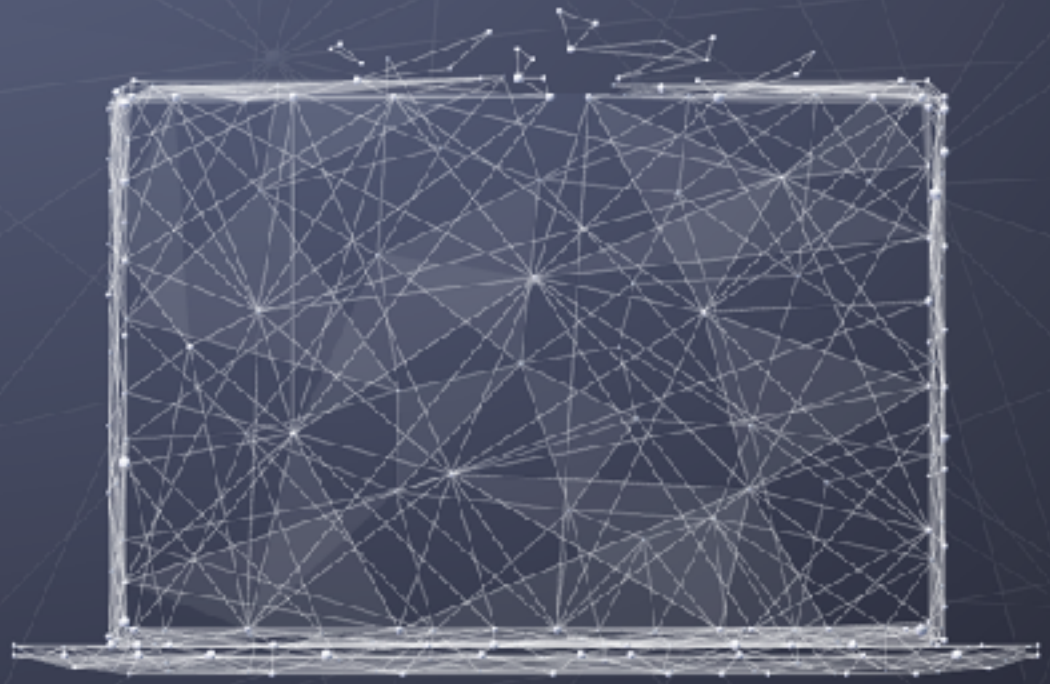
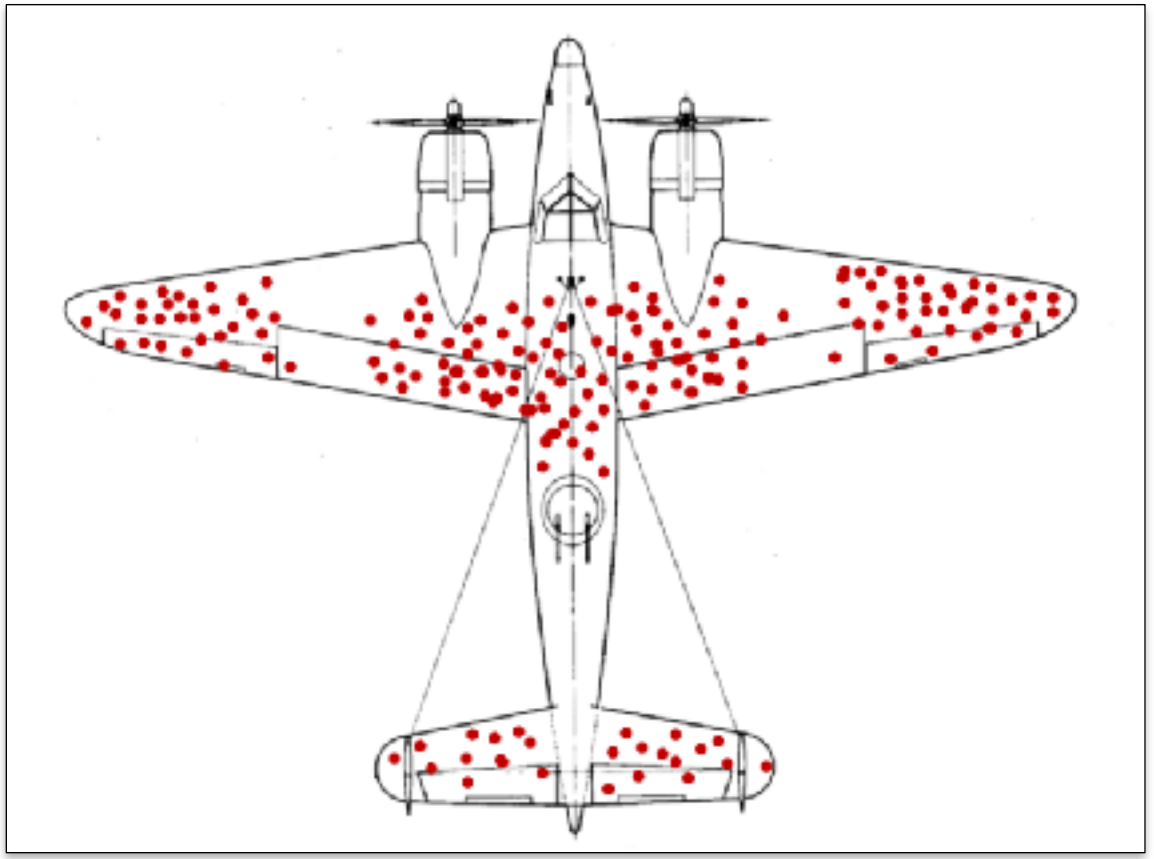
Data Science
Data Engineering I
Introduction to data science



PURDUE UNIVERSITY | College of Science

Early data science

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



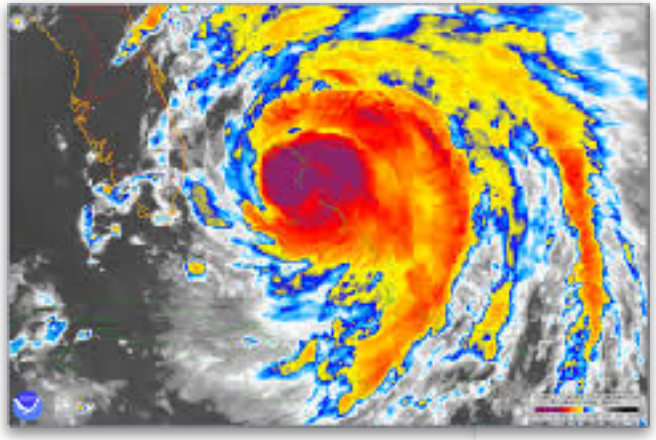
This image depicts the bullets holes observed on planes that returned from the front

Example: Automating science

Sky image cataloging (Fayyad et al 1993)

- 3 TB of image data from Palomar Observatory Sky Study
- Task: Automatically categorize 100 million stellar objects (e.g., star/galaxy)
- 94% accuracy using 1700 examples and 40 features
- Also identified objects too faint for humans to assess

Example: Supply chain



SECTIONS HOME SEARCH

The New York Times


States Push Back After Net Neutrality Repeal


Investors Spooked at Specter of Central Banks Halting Bond-Buying Spree


Purge of Kevin Spacey Tests 'All the Money in the World' a Big Problem


Wielding Data, Women Force a Reckoning Over Bias in the Economics Field


Canada Attacks U.S. Tariffs by Taking Case to World Trade Organization

PAID POST: Quebec City A... Favorite Culture

BUSINESS DAY

What Wal-Mart Knows About Customers' Habits

By CONSTANCE L. HAYS SEP. 14, 2004

Correction Appended

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's computer network, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

The experts mined the data and found that the stores would indeed need certain products -- and not just the usual flashlights. "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," Ms. Dillman said in a recent interview. "And the pre-hurricane top-selling item was beer."

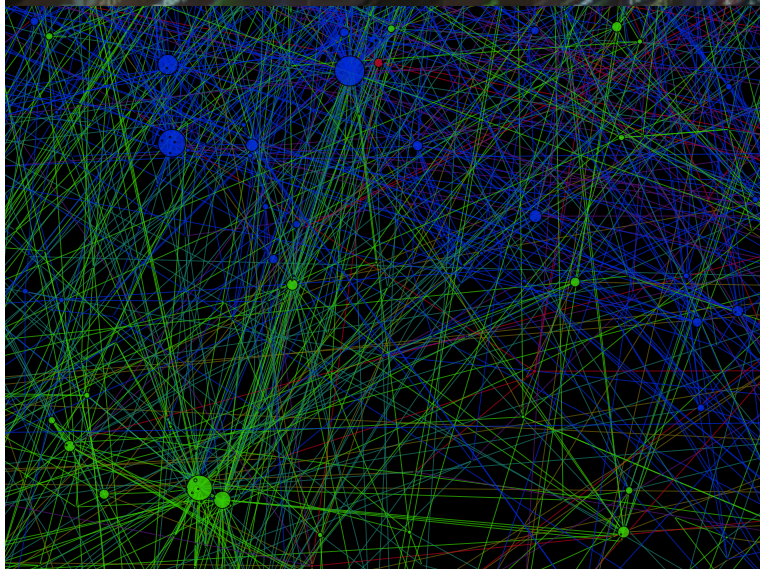
Thanks to those insights, trucks filled with toaster pastries and six-packs were seen speeding down Interstate 95 toward Wal-Marts in the path of Frances. Most of the products that were stocked for the storm sold quickly, the company said.

PURDUE UNIVERSITY | College of Science

Example: Fraud detection

Broker malfeasance (Neville et al 2005)

- 600,000+ stock brokers are registered trade on the NASDAQ
- FINRA investigates 6,000 brokers per year to identify and limit fraud/malfeasance
- Task: Automatically predict brokers likely to be fraudulent, using organizational network information
- ML methods identified new cases that previous hand-crafted rules wouldn't have caught



Example: Targeted marketing

Magazine

How Companies Learn Your Secrets

Antonio Bolfo/Reportage for The New York Times

Andrew Pole had just started working as a statistician for Target when two colleagues from the marketing department stopped by ask an odd question: "If we wanted to figure out if a customer is even if she didn't want us to know, can you do that?"

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."

PURDUE UNIVERSITY

Example: Healthcare

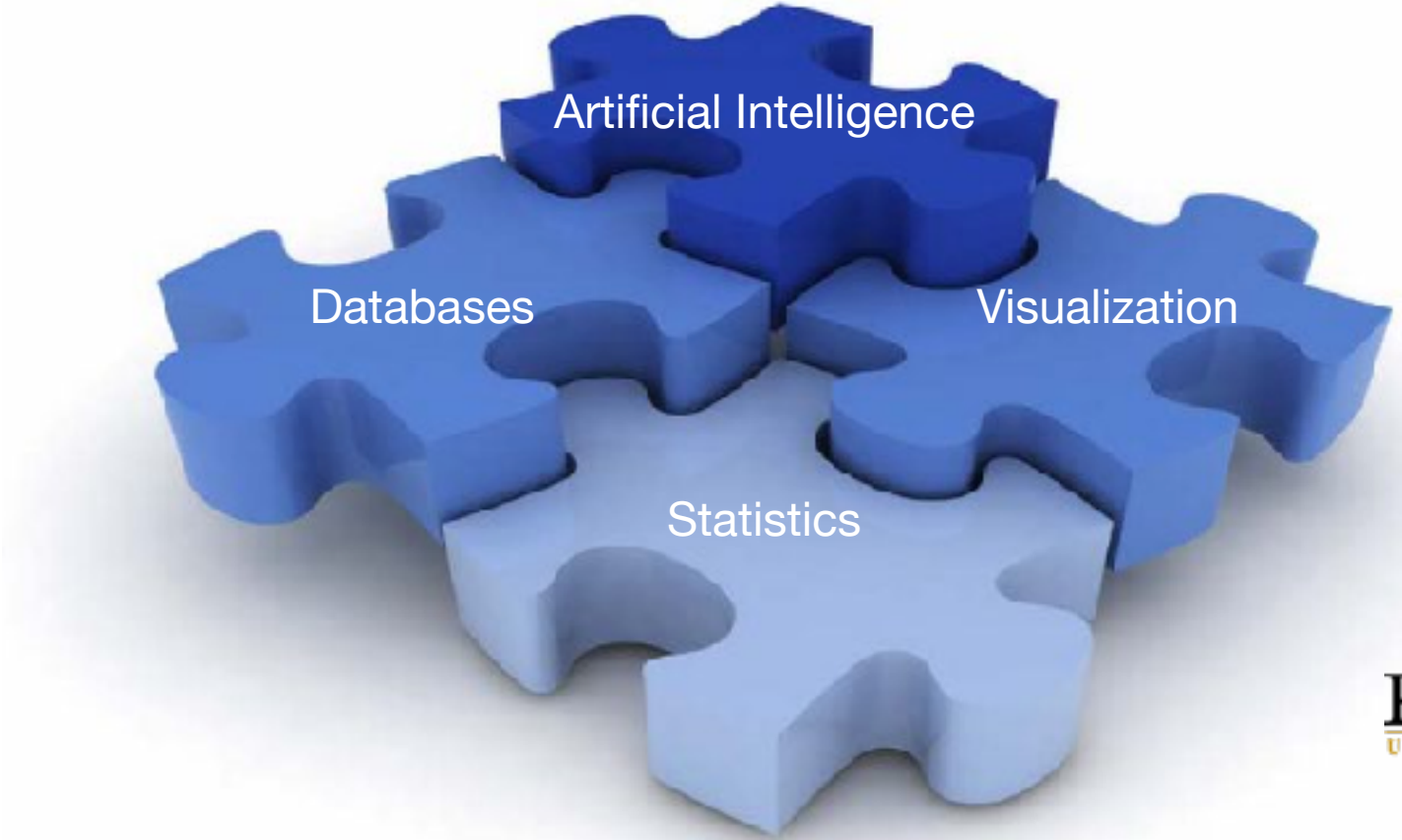
Early detection of sepsis (Henry et al. 2015)

- Electronic health records of 16,234 patients
- Cox proportional hazards model. Uses 27 factors, based on routinely collected inputs
- ML method correctly predicts septic shock in 85% cases, without increasing false positives
- More than 2/3 of the time, ML method correctly predicted septic shock before any organ dysfunction — a 60% improvement over existing screening protocols

What is data science?

Definition of data mining

- The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro & Smith 1996)



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with IaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY by Krzysztof Zawadzki

PURDUE UNIVERSITY | College of Science

BIG DATA ANALYTICS PIPELINE & STAKEHOLDERS

DATA MANAGEMENT

DATA PREPARATION, ORGANIZATION, META DATA, DATA PROVENANCE, AND DATA LIFE CYCLE MANAGEMENT

ANALYTICS AND INSIGHTS

FEATURE SELECTION, MODEL GENERATION, STATISTICAL ANALYSIS, MACHINE LEARNING, DIAGNOSTICS, PROGNOSTICS, PREDICTIVE ANALYTICS, AND PRESCRIPTIVE ANALYTICS AND OPTIMIZATION

DOMAIN KNOWLEDGE

BUSINESS ANALYSIS AND APPLICATION INTEGRATION

COLLABORATION & DECISION SUPPORT

BUSINESS FOCUSED INSIGHTS, APPLICATIONS, AND DECISION MAKING

ACTIVITY STREAMS

RELATIONAL STORES/EDWs

UNSTRUCTURED RICH MEDIA

ETL/ELT

DATABASES
DATA
WAREHOUSES

INSIGHTS

BUSINESS APPLICATION

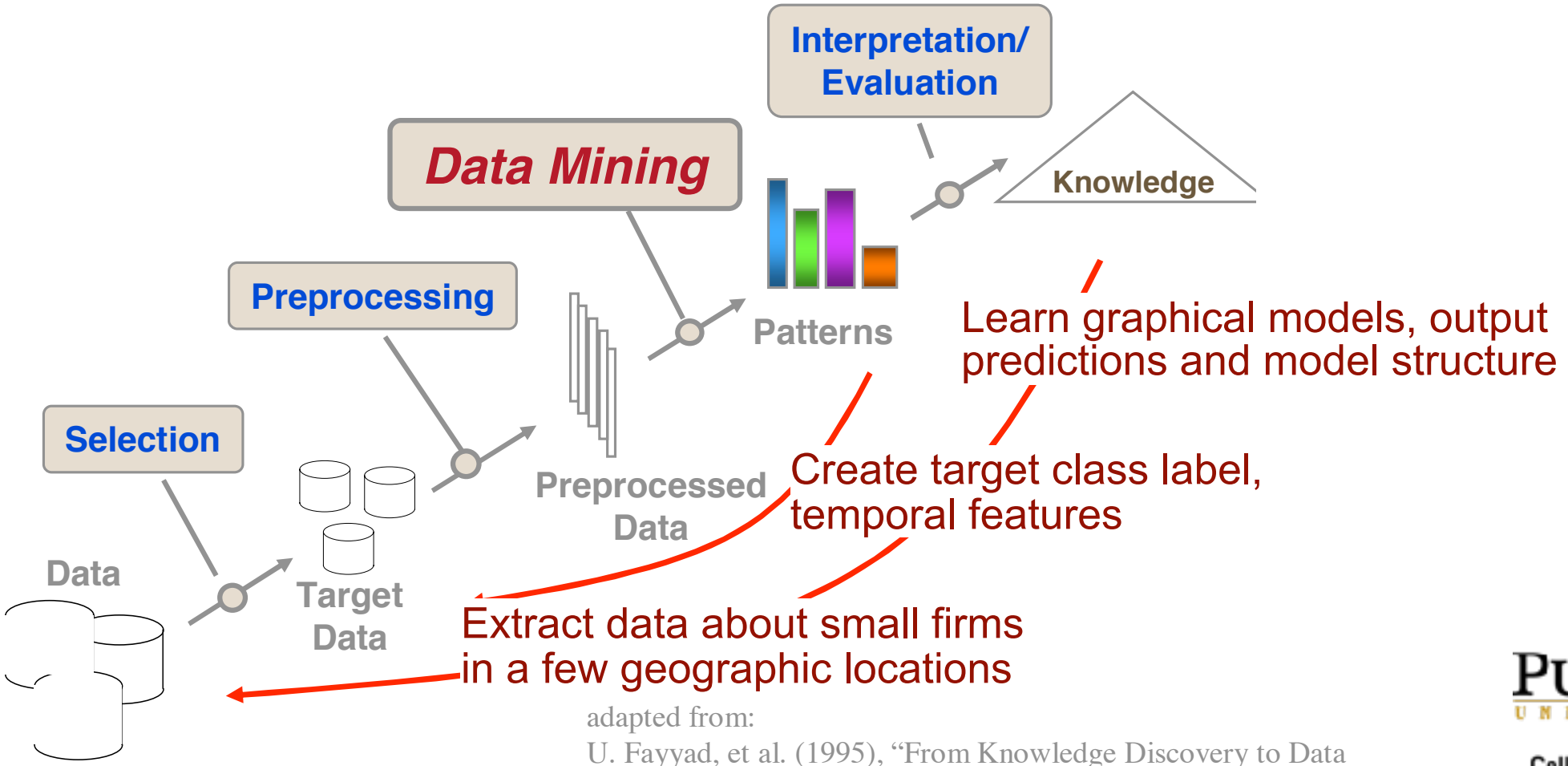
DATA INTEGRATORS
CURATORS/STEWARDS
APPLICATION DEVELOPERS
SYSTEM INTEGRATORS

DATA SCIENTISTS

BUSINESS ANALYSTS

BUSINESS USERS

Data science process



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

PURDUE UNIVERSITY
College of Science