WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.

EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!

IF THAT’S THE SAME THING AS SPAM, WE’RE HAVING A GOOD MEETING HERE.

Slides by Michael Hahsler
What is Information Engineering?

"Information engineering (IE) or information engineering methodology (IEM) is a software engineering approach to designing and developing information systems. It can also be considered as the generation, distribution, analysis and use of information in systems."

[Wikipedia]

"Information Engineering is the incorporation of an engineering approach and discipline to the generation of information and the promotion of the better use of information and resources."

[Steven A. Demurj Jan, CSE, UConn]
What is Analytics?

- Analytics is the discovery and communication of meaningful patterns in data.
- Analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.
- Analytics often favors data visualization to communicate insight.
Why do companies care?

Businesses collect and warehouse lots of **data**.
- Bank/credit card transactions
- Web data, e-commerce
- Social media
- Internet of things (IOT)

**Computers** are cheaper and more powerful.
- SaaS/IaaS/PaaS

**Competition** to provide better services.
- Mass customization and recommendation systems
- Targeted advertising
- Improved logistics
Some companies have built their very businesses on their ability to collect, analyze, and act on data.

Every company can learn from what these firms do.

by Thomas H. Davenport
**THINGS YOU CAN COUNT ON**

Analytics competitors make expert use of statistics and modeling to improve a wide variety of functions. Here are some common applications:

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>DESCRIPTION</th>
<th>EXEMPLARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply chain</td>
<td>Simulate and optimize supply chain flows; reduce inventory and stock-outs.</td>
<td>Dell, Wal-Mart, Amazon</td>
</tr>
<tr>
<td>Customer selection, loyalty, and service</td>
<td>Identify customers with the greatest profit potential; increase likelihood that they will want the product or service offering; retain their loyalty.</td>
<td>Harrah’s, Capital One, Barclays</td>
</tr>
<tr>
<td>Pricing</td>
<td>Identify the price that will maximize yield, or profit.</td>
<td>Progressive, Marriott</td>
</tr>
<tr>
<td>Human capital</td>
<td>Select the best employees for particular tasks or jobs, at particular compensation levels.</td>
<td>New England Patriots, Oakland A's, Boston Red Sox</td>
</tr>
<tr>
<td>Product and service quality</td>
<td>Detect quality problems early and minimize them.</td>
<td>Honda, Intel</td>
</tr>
<tr>
<td>Financial performance</td>
<td>Better understand the drivers of financial performance and the effects of nonfinancial factors.</td>
<td>MCI, Verizon</td>
</tr>
<tr>
<td>Research and development</td>
<td>Improve quality, efficacy, and, where applicable, safety of products and services.</td>
<td>Novartis, Amazon, Yahoo</td>
</tr>
</tbody>
</table>

_Havard Business Review, 2006_
Types of Analytics

- Stochastic Optimization
- Optimization
- Predictive modeling
- Forecasting
- Simulation
- Alerts
- Query/drill down
- Ad hoc reporting
- Standard Reporting

**Prescriptive**

- How can we achieve the best outcome including the effects of variability?
- How can we achieve the best outcome?
- What will happen next if?
- What if these trends continue?
- What could happen…?
- What actions are needed?
- What exactly is the problem?
- How many, how often, where?
- What happened?

**Predictive**

- Data Mining / Stats
- Statistics
- Machine Learning

**Descriptive**

- DB / CS

Based on: Competing on Analytics, Davenport and Harris, 2007
Who does all this?
And who gets the big paycheck?
Who does all this?
And who gets the big paycheck?

Of course! That weird DATA SCIENTIST living in an overpriced house in Silicon Valley!
Who is a data scientist?

- The perfect data scientist from Kolassa’s Venn diagram is a mythical sexy unicorn ninja rockstar who can transform a business just by thinking about its problems.
- A person who is better at statistics than any software engineer and better at software engineering than any statistician.
- Data scientist is now widely used for people working with data.

https://yanirseroussi.com/2016/08/04/is-data-scientist-a-useless-job-title/
What will we learn in this course?

And where can you learn more?

From where do we get data?

- SQL
- XML
- Data Warehouses

→ Get also a CS major/minor

Data Preparation

- SQL
- Code

→ EMIS 7332

Describe Data

- Simple statistics
- Statistical test
- Visualization

→ EMIS 3340

Model Data

- Regression
- Classification
- Forecasting

→ EMIS 7332

Optimization

→ EMIS 3360

Decision or Decision Support Tool
How to do an analytics project?
Remember this from EMIS 2360?

Decision-Making Process

1. Recognize Problem / Opportunity
2. Define Goals/Objectives
3. Assemble Relevant Data
4. Identify Feasible Alts
5. Select the Criterion
6. Construct a Model
7. Predict Alts’ Outcomes
8. Choose the Best Alt.
9. Audit the Results

Overall Mission / Objectives
How to do an analytics project?

CRISP-DM Reference Model

- **Cross Industry Standard Process for Data Mining**
- De facto standard for conducting data mining and knowledge discovery projects.
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own process.
# Tasks in the CRISP-DM Model

<table>
<thead>
<tr>
<th>Business Understanding</th>
<th>Data Understanding</th>
<th>Data Preparation</th>
<th>Modeling</th>
<th>Evaluation</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Determine Business Objectives</strong></td>
<td><strong>Collect Initial Data</strong></td>
<td><strong>Select Data</strong></td>
<td><strong>Select Modeling Techniques</strong></td>
<td><strong>Evaluate Results</strong></td>
<td><strong>Plan Deployment</strong></td>
</tr>
<tr>
<td>Background&lt;br&gt;Business Objectives&lt;br&gt;Business Success Criteria</td>
<td>Initial Data Collection Report</td>
<td>Rationale for Inclusion/Exclusion</td>
<td><strong>Modeling Technique</strong>&lt;br&gt;<strong>Modeling Assumptions</strong></td>
<td>Assessment of Data Mining Results w.r.t. Business Success Criteria&lt;br&gt;Approved Models</td>
<td>Deployment Plan</td>
</tr>
<tr>
<td><strong>Assess Situation</strong></td>
<td><strong>Describe Data</strong></td>
<td><strong>Clean Data</strong></td>
<td><strong>Generate Test Design</strong>&lt;br&gt;<strong>Test Design</strong></td>
<td><strong>Review Process</strong>&lt;br&gt;<strong>Review of Process</strong></td>
<td><strong>Plan Monitoring and Maintenance</strong>&lt;br&gt;<strong>Monitoring and Maintenance Plan</strong></td>
</tr>
<tr>
<td>Inventory of Resources Requirements, Assumptions, and Constraints&lt;br&gt;Risks and Contingencies&lt;br&gt;Terminology&lt;br&gt;Costs and Benefits</td>
<td>Data Description Report</td>
<td>Data Cleaning Report</td>
<td><strong>Build Model</strong>&lt;br&gt;Parameter Settings Models&lt;br&gt;Model Descriptions</td>
<td><strong>Determine Next Steps</strong>&lt;br&gt;List of Possible Actions Decision</td>
<td><strong>Produce Final Report</strong>&lt;br&gt;Final Report&lt;br&gt;Final Presentation</td>
</tr>
<tr>
<td><strong>Determine Data Mining Goals</strong></td>
<td><strong>Explore Data</strong></td>
<td><strong>Construct Data</strong>&lt;br&gt;Derived Attributes Generated Records</td>
<td><strong>Assess Model</strong>&lt;br&gt;Model Assessment Revised Parameter Settings</td>
<td></td>
<td><strong>Review Project</strong>&lt;br&gt;Experience Documentation</td>
</tr>
<tr>
<td>Data Mining Goals&lt;br&gt;Data Mining Success Criteria</td>
<td>Data Exploration Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Produce Project Plan</strong></td>
<td><strong>Verify Data Quality</strong>&lt;br&gt;Data Quality Report</td>
<td><strong>Integrate Data</strong>&lt;br&gt;Merged Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Plan&lt;br&gt;Initial Assessment of Tools and Techniques</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model*
Example: How is POS data stored?

- Relational data base?
- How do the tables look like?
  → On Line Transaction Processing
- Has every store/region its own data base?

- What if I want to know how many units of product A were sold in the last three month in Texas?

- There must be an easier way!
Data Warehouse
ELT: Extract, Transform and Load

• **Extracting** data from outside sources
• **Transforming** it to fit analytical needs. E.g.,
  - Clean (missing data, wrong data)
  - Translate (1 → "female")
  - Join (from several sources)
  - Calculate and aggregate data
• **Loading** it into the end target (data warehouse)
Data Warehouse
Properties

- **Subject Oriented**: Data warehouses are designed to help you analyze data in a certain area (e.g., sales).

- **Integrated**: Integrates data from disparate sources into a consistent format.

- **Nonvolatile**: Data in the data warehouse are never overwritten or deleted.

- **Time Variant**: They maintain both historical and (nearly) current data.
OnLine Analytical Processing (OLAP)

- Stores data in "data cubes" for fast OLAP operations.
- Requires a special database structure (Snow-flake scheme)

Operations:
- Slice
- Dice
- Drill-down
- Roll-up
- Pivot
Data Visualization

- Infoviz is a field by its own.
- Napoleon's Army in Russia by Charles Minard (around 1850)
Eat fruits when they are in season!!!
Legal, Privacy and Security Issues

DID YOU EVER THINK ABOUT SELLING OUR CONFIDENTIAL DATABASE OF CUSTOMER INFORMATION?

IT WOULD BE MASSIVELY PROFITABLE WHILE VIRTUALLY UNDETECTABLE.

BUT HIGHLY UNETHICAL.

I DON'T KNOW YOU ANYMORE.

I'M YANKING YOUR CHAIN. WHEN DO WE START?
Legal, Privacy and Security Issues

- Are we allowed to collect the data?
- Are we allowed to use the data?
- Is privacy preserved in the process?
- Is it ethical to use and act on the data?

- Problem: Internet is global but legislation is local!
BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slinging squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were unaware that the game was storing their locations so that they could later be the targets of ads....
Pokémon Go’s constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

“Their privacy policy is vague,” Hong said. “I’d say deliberately vague, because of the lack of clarity on the business model.”

The agreement says Pokémon Go collects data about its users as a “business asset.” This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game.

If Niantic is ever sold, the agreement states, all that data can go to another company.