CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

Lecture 3: Relational Databases and SQL

Announcements

- Namecards! We want to get to know you!
- Homework 1 due this Sunday

Limitations of Spreadsheets

Data type

- Only on structured data

Data size

- Google sheets: 400,000 cells

Mechanics

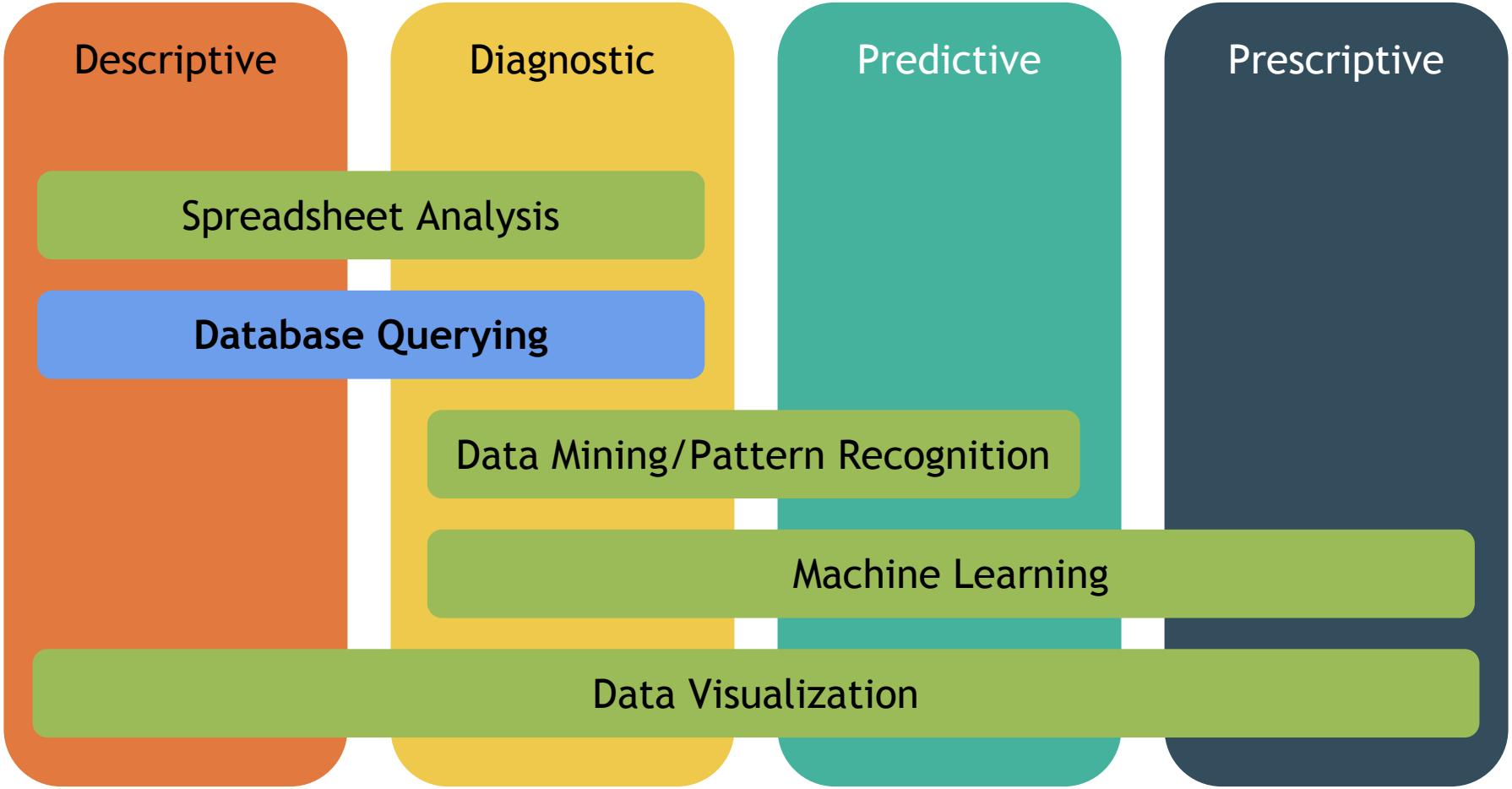
- Header rows, empty cells, strange behaviors, ...

Some analyses are difficult

- E.g., 2 restaurants closest to each other (easy in SQL)

Traceability

Tools & Techniques



Learning Goals

- Learn what are Relational Databases
- Learn what is a Data Scientist
- Familiarize yourself with Jupyter Notebook
- Get comfortable with SQL queries

Relational Databases and SQL

Relational Databases

- What are databases?
  - A large, integrated collection of data
  - Model of the real world
    - Entities (Students, Courses)
    - Relationships (Alice is taking CS102)

Example (Axess)

Students

Name	Year	Major
Alice	Senior	Undeclared
Bob	Junior	Sociology
Charlie	Grad	MS&E
..	...	...

Courses

Class	Instructor	Location	Days
CS102	Ethan, Lisa	Lathrop	T,Th
TAPS103	Dan Klein	Memaud	MW
ME101	Lillie	Quad	MWF
PE 3	Lillie	Track	F

Enrollment

Student Name	Class	Quarter	Grade
Alice	CS102	SP17	A
Alice	TAPS103	FA16	B
Charlie	ME101	FA16	C
Dory	CS102	SP17	A

Example (Axess)

Students

Name	Year	Major
Alice	Senior	Undeclared
Bob	Junior	Sociology
Charlie	Grad	MS&E
..	...	...

Courses

Class	Instructor	Location	Days
CS102	Ethan, Lisa	Lathrop	T,Th
TAPS103	Dan Klein	Memaud	MW
ME101	Lillie	Quad	MWF
PE 3	Lillie	Track	F

Enrollment

Student Name	Class	Quarter	Grade
Alice	CS102	SP17	A
Alice	TAPS103	FA16	B
Charlie	ME101	FA16	C
Dory	CS102	SP17	A

What is SQL?

- SQL is a standard language for:
  - Defining your data (update/insert/delete)
  - Manipulating your data (aggregation / joins / ...)
- SQL
  - Structured
  - Query
  - Language
- SQL is a **declarative** language
  - You tell the computer **what** you want (A Query)
  - Instead of telling it how to get what you want
    - Computer handles the logic on the backend

Why Learn Relational Databases?

- Relational database management systems (RDBMS) have been around for more than 40 years
- $30+ billion per year industry, increasing yearly

Why so successful?

- Simple model, high-level expressive query language, reliable and scalable systems.
- Dark times 5 years ago
  - Today's 'NoSQL' systems are starting to look more and more like RDBMSs.
  - Most "Big Data" systems now include SQL language

Why Learn SQL?

- One of oldest languages still in use (1978)
- Supported by all RDBMSs
- Standardized across products
- Interactive or embedded in programs

SQL Systems

- Commercial Proprietary systems:
  - Oracle
  - Microsoft SQL Server
  - IBM DB2
  - ..
- Open Source Systems
  - MySQL
  - **SQLite (using this for class)**
  - PostgreSQL
  - ...

Basic Concepts

Definition: Relation (Table)

Relation Name: Product

PName	Price	Manufacturer
Gizmo	$19.99	GizmoWorks
Powergizmo	$29.99	GizmoWorks
SingleTouch	$149.99	Canon
MultiTouch	$203.99	Hitachi

Adapted from CS145: Databases

Definition: Attribute (Column)

Product

PName	Price	Manufacturer
Gizmo	$19.99	GizmoWorks
Powergizmo	$29.99	GizmoWorks
SingleTouch	$149.99	Canon
MultiTouch	$203.99	Hitachi

Adapted from CS145: Databases

Definition: Tuple (Row)

Product

PName	Price	Manufacturer
Gizmo	$19.99	GizmoWorks
Powergizmo	$29.99	GizmoWorks
SingleTouch	$149.99	Canon
MultiTouch	$203.99	Hitachi

Adapted from CS145: Databases

SQL Types and Domains

Types

- VARCHAR / TEXT
- BOOLEAN
- DATE
- TIME
- ...

Domains

- Constraint for each type
  - E.g. VARCHAR (256)
    - Maximum of 256 characters of text

Differences Between Table and Spreadsheet

Table (Relation)

- Name is significant
  - Watch out for spaces and capitalization of letters!
- Order is not significant
  - can change on re-open
- Regular structure, more “row-oriented”
- Data elements always values, not formulas

Creating and Loading Data

System-dependent, but can nearly always start with .CSV file or similar

SQL Query (Basics)

SQL Query

- Basic form (there are many many more bells and whistles)

```
SELECT <attributes>
FROM   <one or more relations>
WHERE  <conditions>
```

Call this a SFW query.

Adapted from CS145: Databases

Selecting all Columns (Select *)

PName	Price	Category	Manufacturer
Gizmo	$19.99	Gadgets	GizmoWorks
Powergizmo	$29.99	Gadgets	GizmoWorks
SingleTouch	$149.99	Photography	Canon
MultiTouch	$203.99	Household	Hitachi

```
SELECT *
FROM   Product
WHERE  Category = 'Gadgets'
```

PName	Price	Category	Manufacturer
Gizmo	$19.99	Gadgets	GizmoWorks
Powergizmo	$29.99	Gadgets	GizmoWorks

Adapted from CS145: Databases

Selecting specific Columns

PName	Price	Category	Manufacturer
Gizmo	$19.99	Gadgets	GizmoWorks
Powergizmo	$29.99	Gadgets	GizmoWorks
SingleTouch	$149.99	Photography	Canon
MultiTouch	$203.99	Household	Hitachi

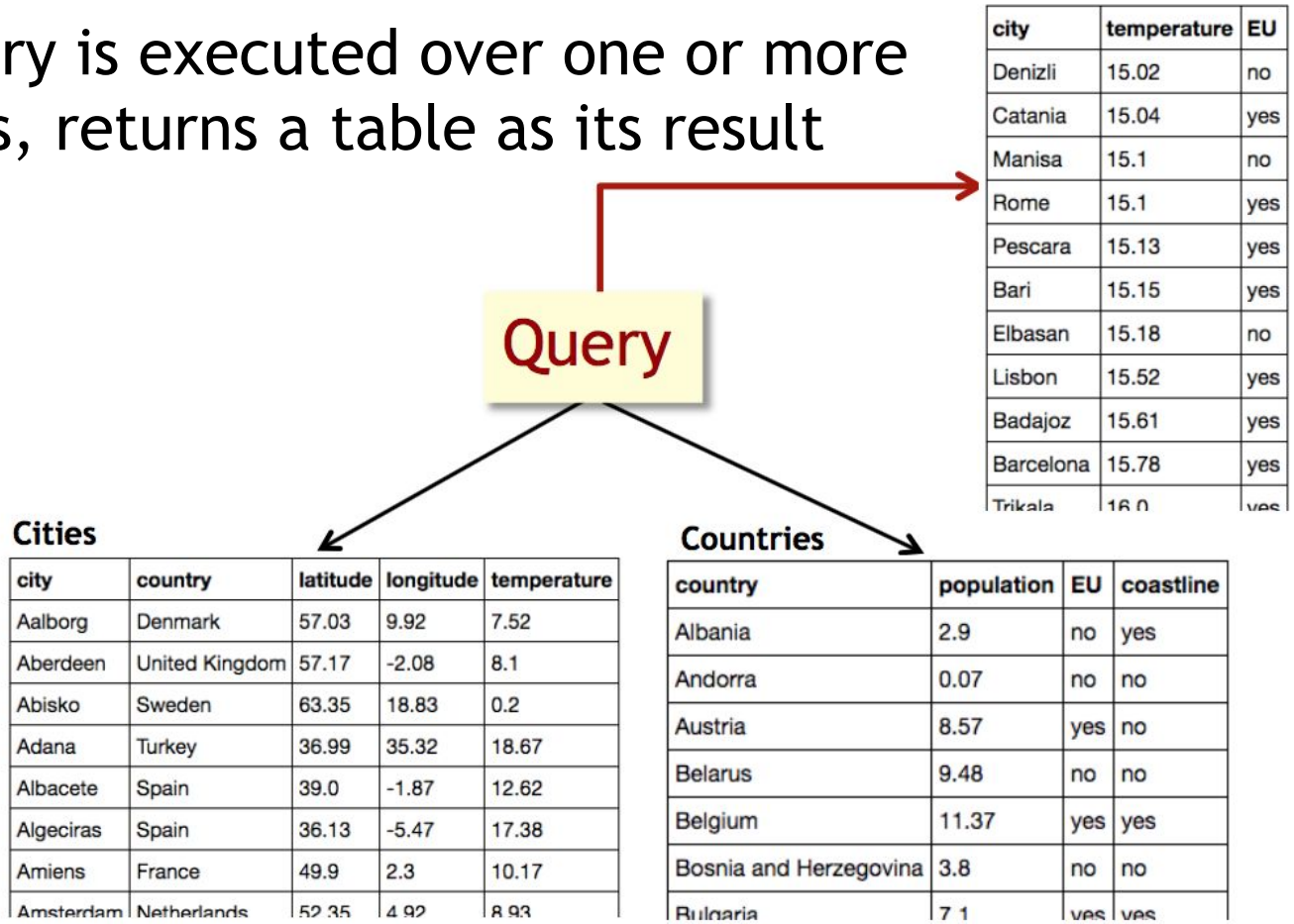
```
SELECT Pname, Price, Manufacturer
FROM   Product
WHERE  Category = 'Gadgets'
```

PName	Price	Manufacturer
Gizmo	$19.99	GizmoWorks
Powergizmo	$29.99	GizmoWorks

Adapted from CS145: Databases

Querying

- A query is executed over one or more tables, returns a table as its result



city	temperature	EU
Denizli	15.02	no
Catania	15.04	yes
Manisa	15.1	no
Rome	15.1	yes
Pescara	15.13	yes
Bari	15.15	yes
Elbasan	15.18	no
Lisbon	15.52	yes
Badajoz	15.61	yes
Barcelona	15.78	yes
Trikala	16.0	yes

city	country	latitude	longitude	temperature
Aalborg	Denmark	57.03	9.92	7.52
Aberdeen	United Kingdom	57.17	-2.08	8.1
Abisko	Sweden	63.35	18.83	0.2
Adana	Turkey	36.99	35.32	18.67
Albacete	Spain	39.0	-1.87	12.62
Algeciras	Spain	36.13	-5.47	17.38
Amiens	France	49.9	2.3	10.17
Amsterdam	Netherlands	52.35	4.92	8.93

country	population	EU	coastline
Albania	2.9	no	yes
Andorra	0.07	no	no
Austria	8.57	yes	no
Belarus	9.48	no	no
Belgium	11.37	yes	yes
Bosnia and Herzegovina	3.8	no	no
Bulgaria	7.1	yes	yes

Tidbit of the day..

What is a Data Scientist?

Data Scientist

"A data scientist is a statistician who lives in San Francisco."

"Person who is better at statistics than any software engineer and better at software engineering than any statistician."

Data Scientist

Someone who

asks the right questions

+

collects the data

+

cleans and processes data

+

makes sense of it

+

tells a story

Data Scientist (tools)

Someone who

asks the right questions

+

collects the data (SQL/Python/..)

+

cleans and processes data (Python)

+

makes sense of it (SQL / Python / R..)

+

tells a story (Charts / Tableau)

Data Scientist (Example)

Someone who

- asks the right questions
  - "Which customers are most likely to leave my website?"
- collects the data
  - Talks to various teams within the firm to get the data
- cleans and processes data
  - Most of time the data is always incomplete / messy
- makes sense of it
  - Build a machine learning model to predict who will leave
- tells a story
  - Tells the sales/marketing team who to focus on

Data Scientist (Example)

Someone who

asks the right questions

"Which customers are most likely to leave my website?"

collects the data

Talks to various teams within the firm to get the data

cleans and processes data

Most of time the data is always incomplete / messy

makes sense of it

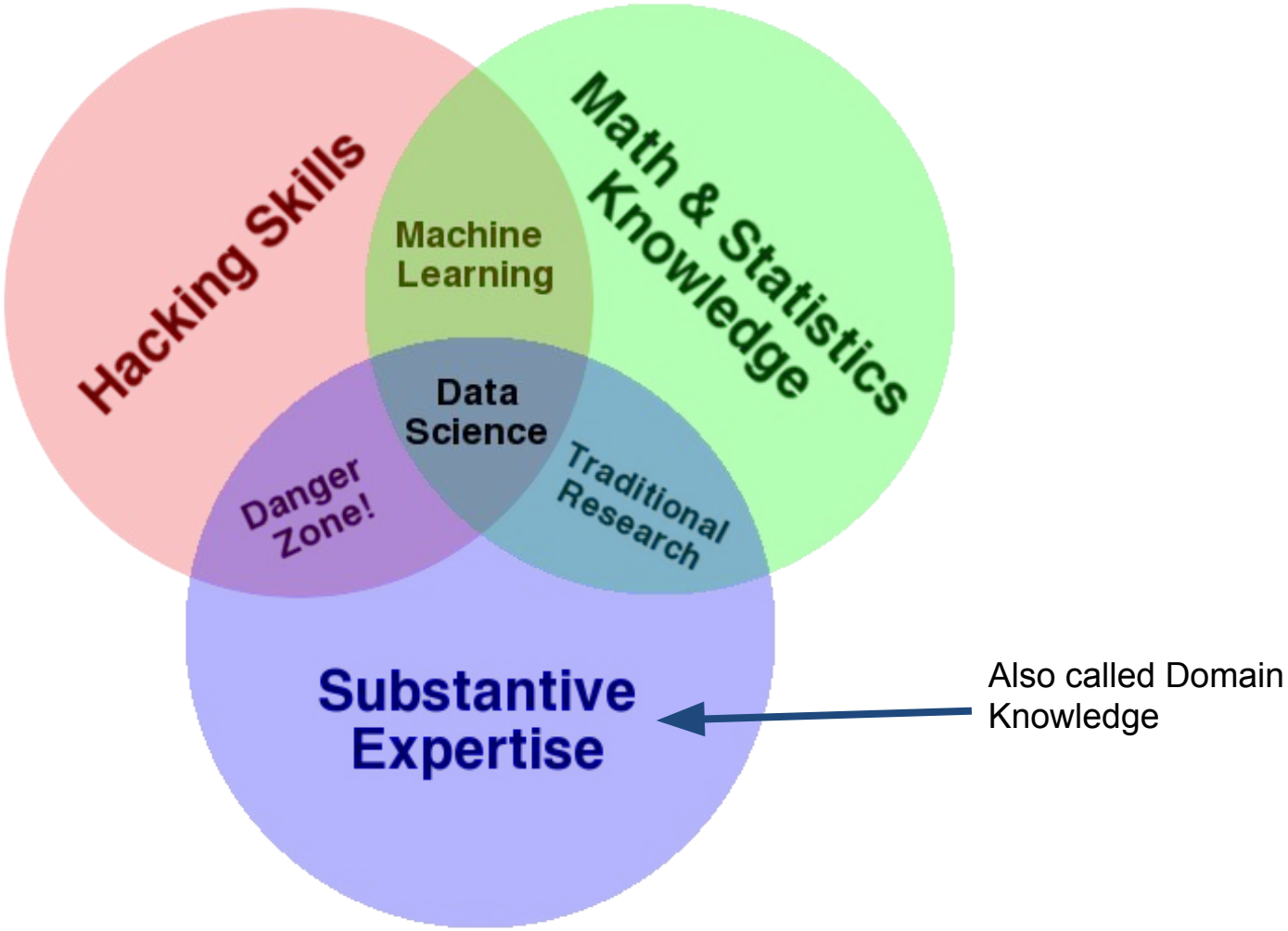
Build a machine learning model to predict who will leave

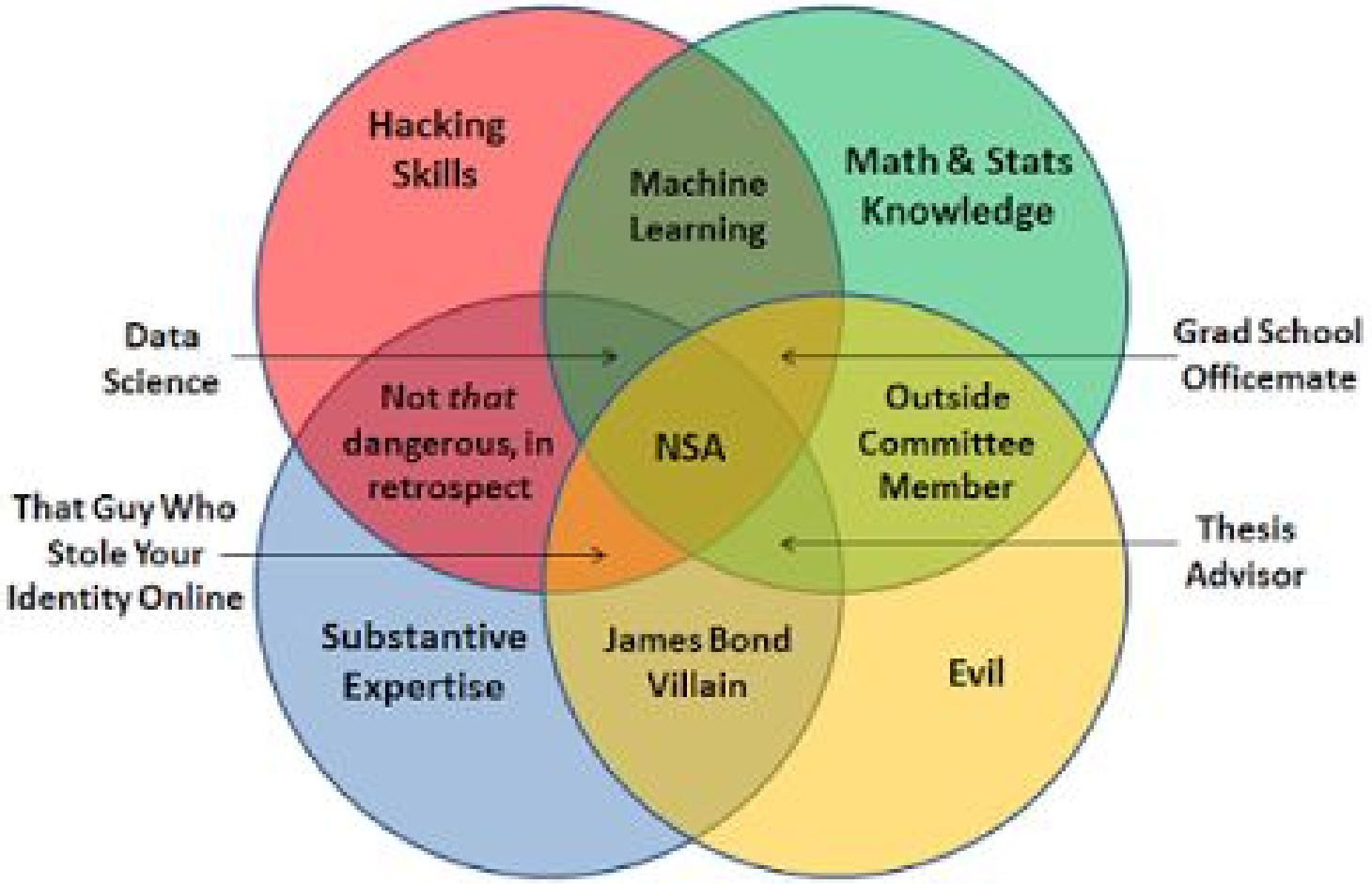
tells a story

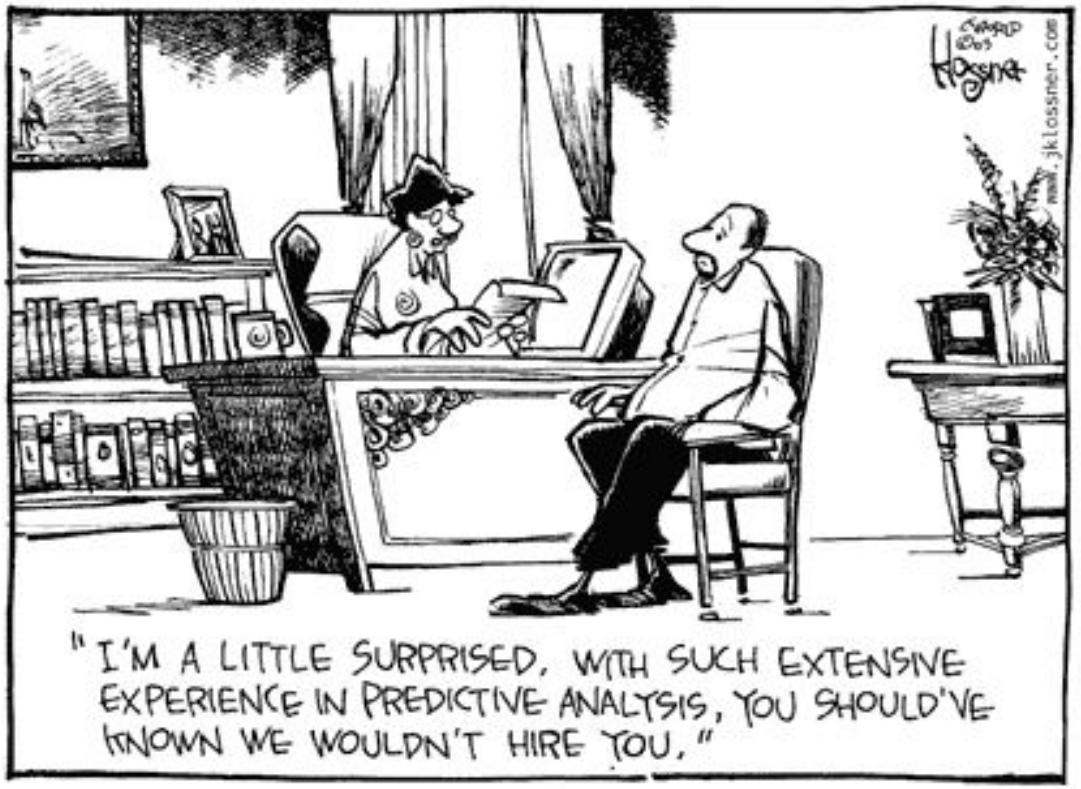
Tells the sales/marketing team who to focus on

~80% of your time

Data Scientist Venn Diagram




Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil


"I'M A LITTLE SURPRISED, WITH SUCH EXTENSIVE EXPERIENCE IN PREDICTIVE ANALYSIS, YOU SHOULD'VE KNOWN WE WOULDN'T HIRE YOU."

Jupyter Notebooks

Jupyter Notebooks

(formerly iPython notebooks)

- Modelled after “laboratory notebooks”
- can combine text boxes (“markdown”) with boxes containing executable code
- Can run/re-run boxes (cells) individually, or run/re-run entire notebook
- Rapid adoption in many sectors
- Notebooks run in a web browser (no internet needed)

In Class Demo

About the Data we're using

1. CityTemps (city, state, lat, lng, temp)

city	state	lat	lng	temp
Mobile	Alabama	31.2	88.5	44
Montgomery	Alabama	32.9	86.8	38
Phoenix	Arizona	33.6	112.5	35
Little Rock	Arkansas	35.4	92.8	31
Los Angeles	California	34.3	118.7	47

2. Regions (state, region, coastal)

state	region	coastal
Maine	Northeast	Y
Vermont	Northeast	N
Rhode Island	Northeast	Y
New York	Midatlantic	Y

Launch the Jupyter Notebook!

1. Download the following files (ensure both of them are in the same folder):
   a. SQLLecture.ipynb
   b. Weather.db
2. Run Jupyter notebook to open SQLLecture.ipynb
3. SQL Lecture notes posted online

SQL Features not Covered

- Set Operators
  - Union, Intersect, Except
- Keys
  - Designated column that must have unique value in each row
  - Or designated set of columns
- Null values
  - Special value usually denoting unknown or undefined
  - Not included in aggregations, =, <, etc.
  - Example: ... where temp <= 10 or temp > 10
- Outer joins

Please give us feedback here:
http://bit.ly/cs102_feedback

End.