Data Analysis and Data Science

CPS352: Database Systems

Simon Miner
Gordon College
Last Revised: 4/29/15

Agenda

- Check-in
- Online Analytical Processing
- Data Science
- Homework 8

Check-in

Online Analytical Processing

Online Transaction Processing (OLTP)

- Transactional data – database concerned with maintaining single focused end-user interactions
  - Examples
    - Customer placing an order on an e-commerce website
    - Account holder making a deposit at a bank
  - Can be comprised of several rows/records of data
    - Example: An order has records for the order itself, each line item, address, payment method, etc.
  - Lots of data can accumulate quickly for numerous transactions
    - Needed for its own sake (i.e. shipping orders, order history, monthly account statements, etc.
    - Also useful for analysis…
- OLTP databases built and optimized for speed of transactions (both in the ACID and interaction contexts)
  - i.e. Provisioned with smaller block sizes to facilitate more precise (and maybe quicker) read and write operations

Online Analytical Processing (OLAP)

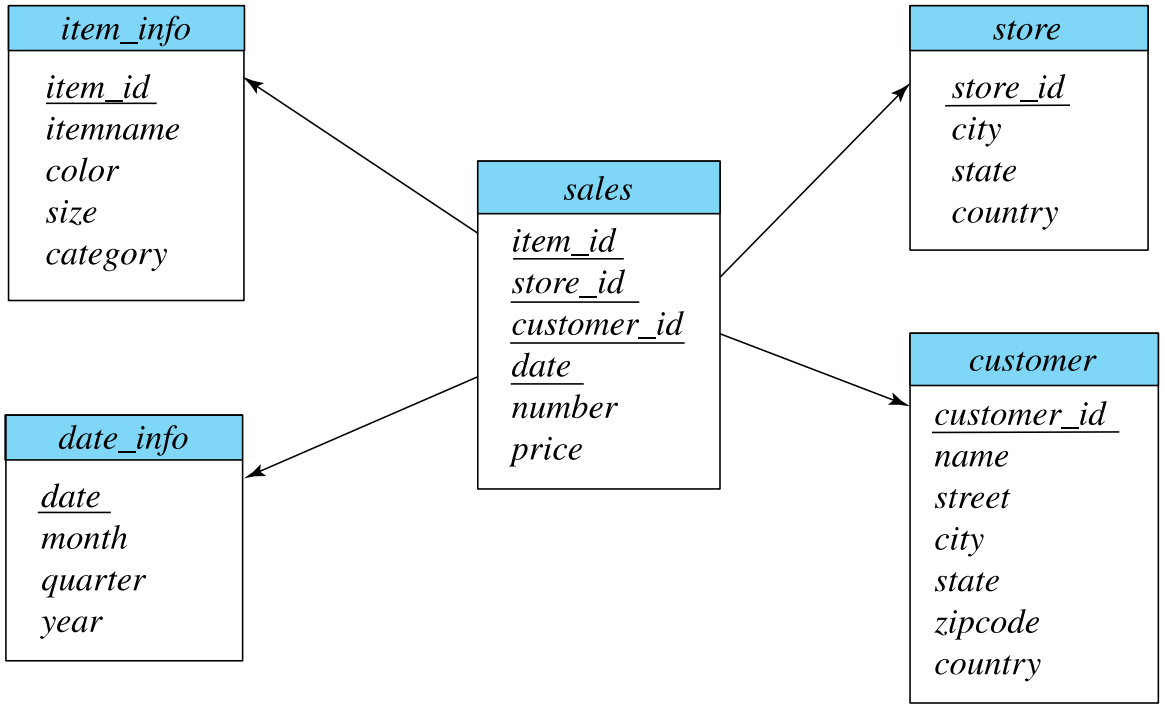
- Decision support systems (DSS) to help organizations determine longer term courses of action
  - Example: Not many orders for a certain product, so adjust product offerings to better match customer desires
  - Work with summaries and aggregations of raw transaction data
- OLAP user needs to have specific queries in mind
  - Example: Give me a cross-tab of item type vs. color …
- Data mining – automated process to reveal patterns in data and system usage
- OLAP databases designed to handle large amounts of data
  - i.e. Provisioned with larger block sizes to store and retrieve more data in read and write operations

Data Warehouse

- Unified repository for an organization's historical OLAP data
  - Supports trending, analysis, and decision making
- Gathered from numerous disparate sources via ETL processes
  - Extract – get data from individual source(s) owned or managed by various parties
  - Transform – manipulate data so that it fits into the data warehousing schema – i.e. de-duplication, summarization
  - Load – store the transformed data in the data warehouse
- Data is loaded at regular intervals
  - Slightly out of date, which is fine for analytical tasks the data warehouse is used for

Data Warehouse Schema

- Dimension values are usually encoded using small integers and mapped to full values via dimension tables
- Star schema
- Snowflake schema



A Data Warehouse in the Clouds

"Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Amazon Redshift offers you fast query performance when analyzing virtually any size data set using the same SQL-based tools and business intelligence applications you use today. With a few clicks in the AWS Management Console, you can launch a Redshift cluster, starting with a few hundred gigabytes of data and scaling to a petabyte or more, for under $1,000 per terabyte per year."

OLAP Concepts

- Attribute types
  - Dimension attribute – values to analyze on
    - Explicit – color, size, price, customer type, etc.
    - Derived – age (computed from DOB), ranges (years of experience)
  - Measurement attribute – value summarized or aggregated over various dimensions (sum, count, average, etc.)
- Cross-tab (pivot table) – tool allowing easy analysis of data along various dimensions
  - Available in tools like spreadsheets
  - Basic SQL is not an effective tool to produce this kind of structure (lots of dynamic "group by" queries needed)

Cross Tabulation of *sales* by *item-name* and *color*

size: **all**

item-name	*color* dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

- The table above is an example of a **cross-tabulation** (**cross-tab**), also referred to as a **pivot-table**.
  - Values for one of the dimension attributes form the row headers
  - Values for another dimension attribute form the column headers
  - Other dimension attributes are listed on top
  - Values in individual cells are (aggregates of) the values of the dimension attributes that specify the cell.

Figure 18.1 in book

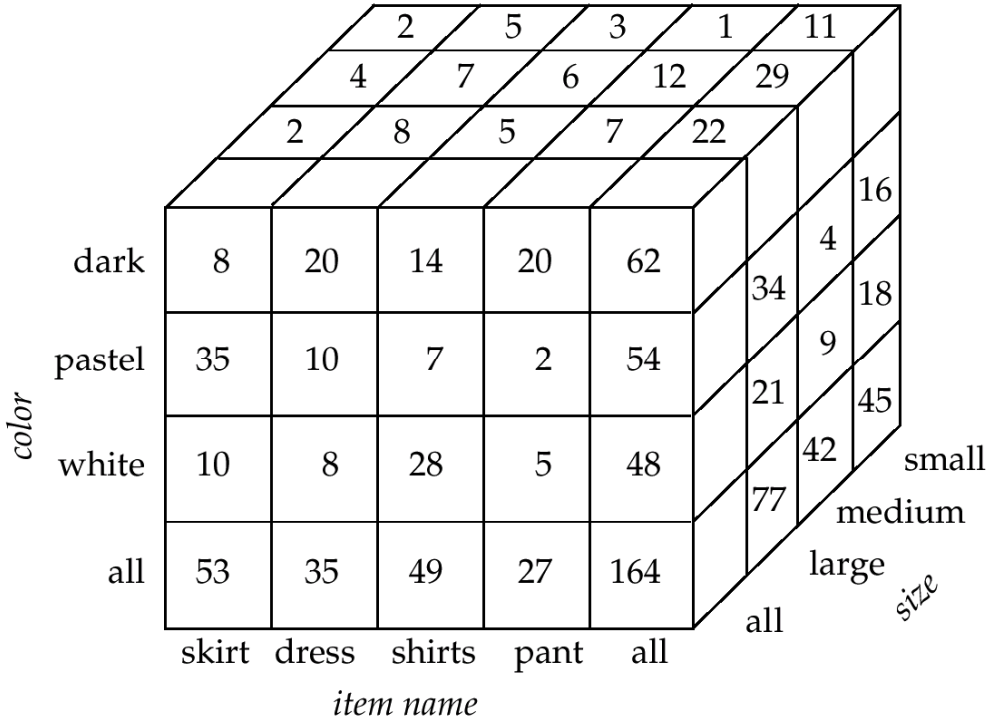
OLAP Operations

- Basic SQL
  - Aggregate functions, like sum(), count(), average()
  - Group by / having clause
- SQL-99 added support for operations to support analytics processing
  - Cube
  - Rollup
  - Rank / dense rank

Cube

- Structure to aggregate a single measurement attribute across numerous dimensions
  - Includes all possible combinations of dimension values
  - Number of cube dimensions = number of dimensional attributes
  - Each dimension “row” includes a summary value for the aggregate of all possible values of that dimension
- User slices cube for specific dimension values

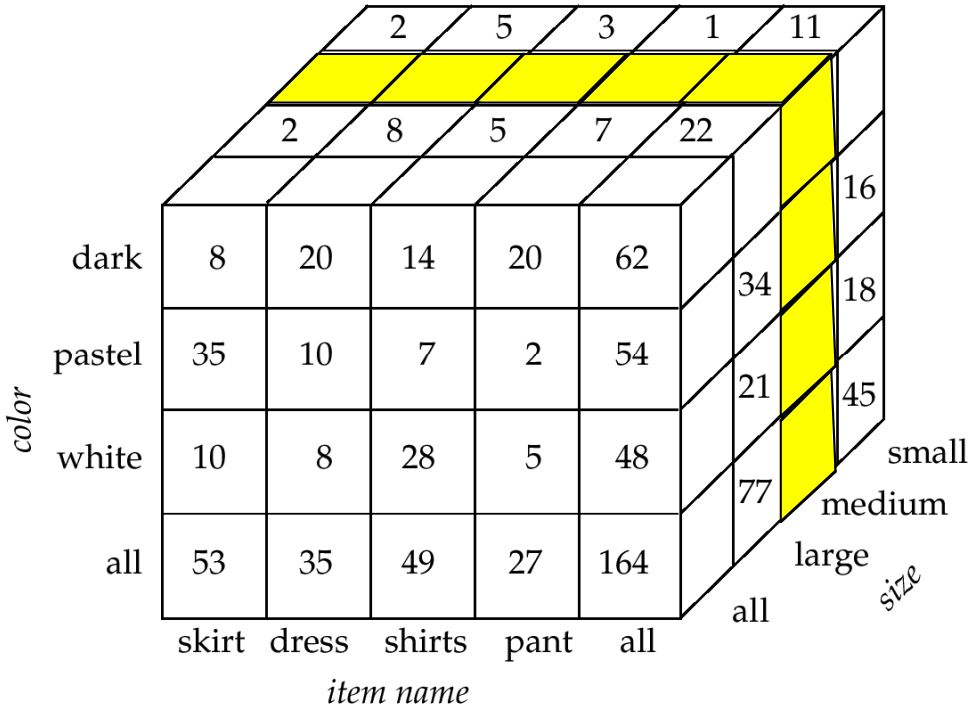
Cube Example



Cube showing sales for various combinations of item_name, color and size - including summaries for all item_names, colors, and/or sizes

Figure 18.3 in book

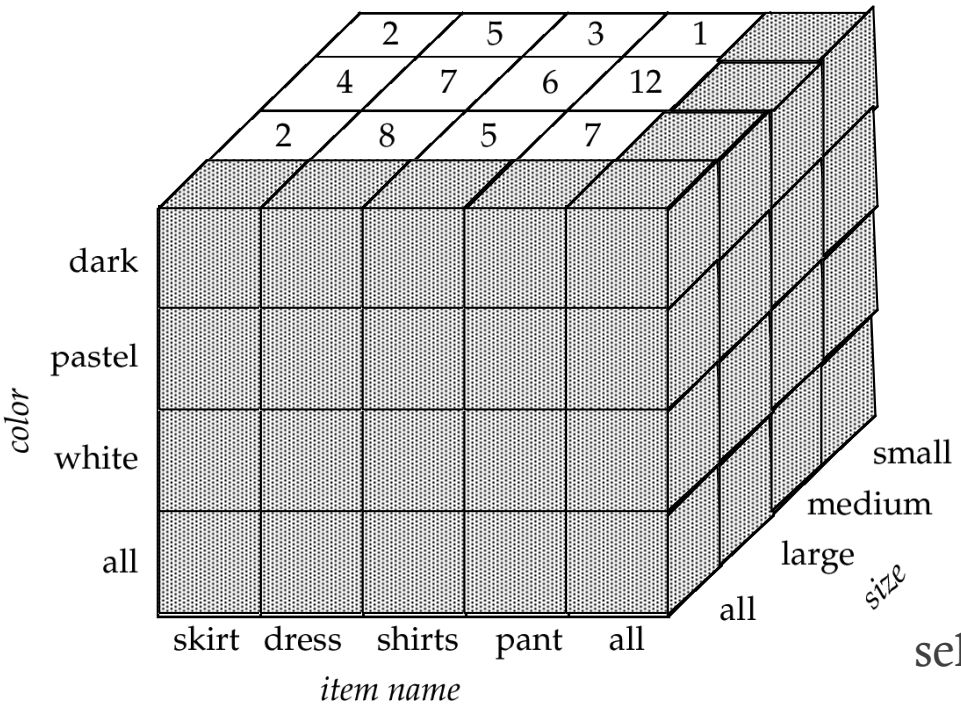
Cube Slice Example



Slice showing sales for various combinations of item_name and color for size = medium

Slice from figure 18.3 in book

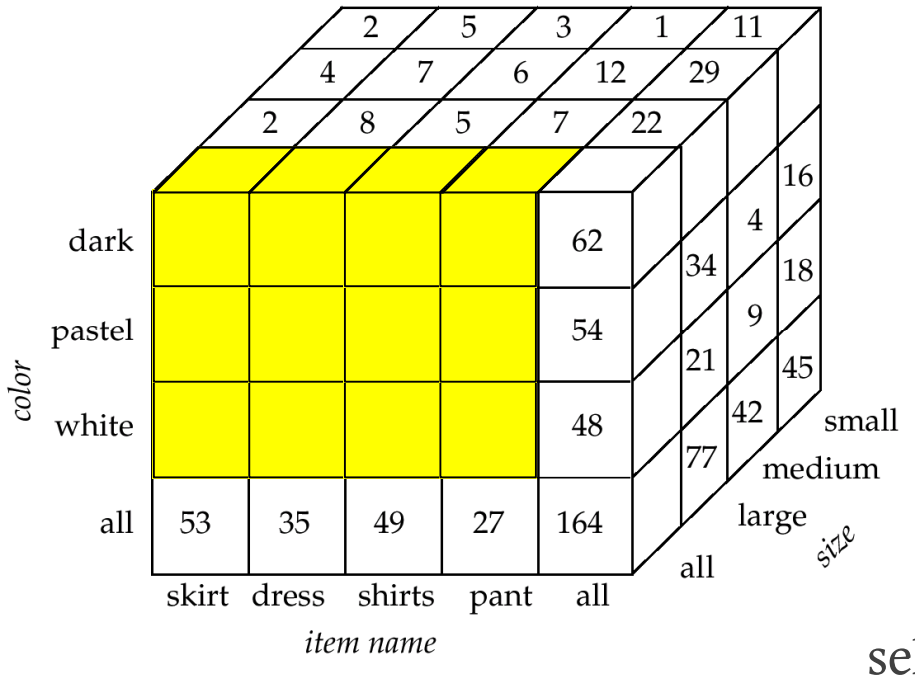
Cube Without Summaries




```
select item_name, color, size, sum(number)
    from sales
    group by item_name, color, size;
```

Sales for all possible combinations of item_name, color, size without summaries - everything but shaded squares and squares on bottom. (What would result from a standard SQL query using `group by`).

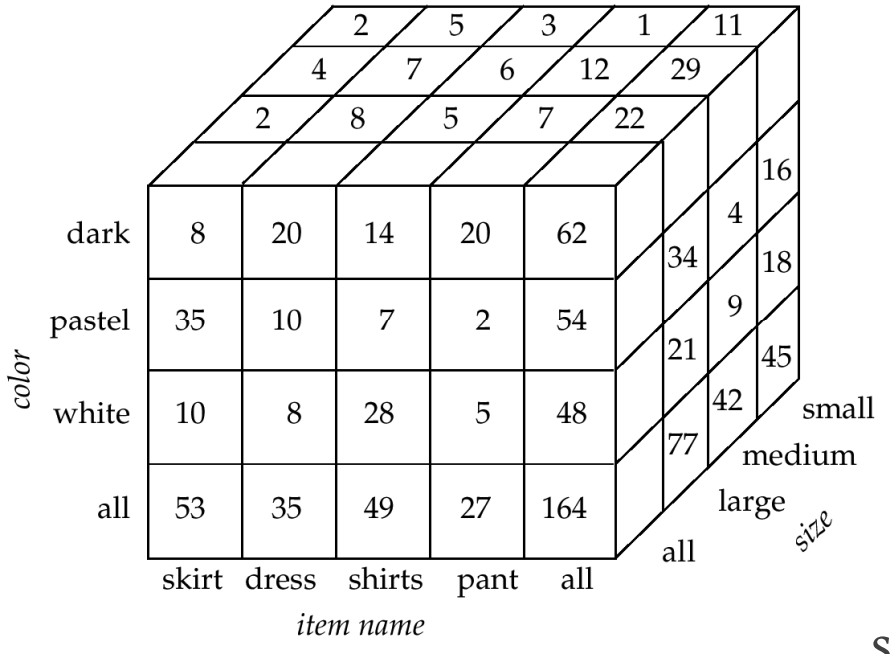
Another Cube Slice



Sales by item_name and color - for all sizes

```
select item_name, color, sum(number)
    from sales
    group by item_name, color;
```

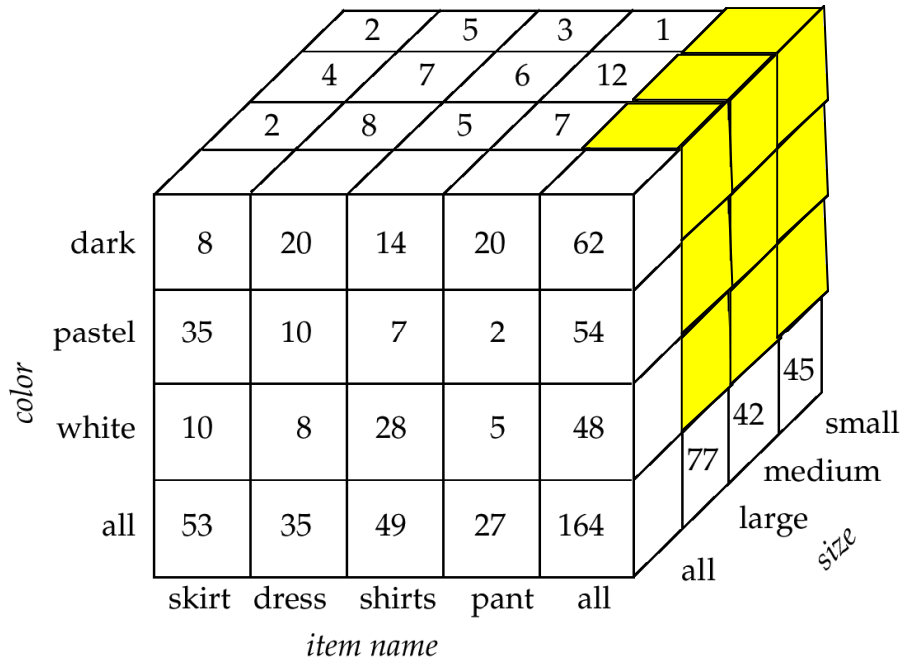
Still Another Cube Slice



Sales by item_name and size - for all colors
(not visible - all cells on bottom of cube,
except front and right side)

```
select item_name, size, sum(number)
    from sales
    group by item_name, size;
```

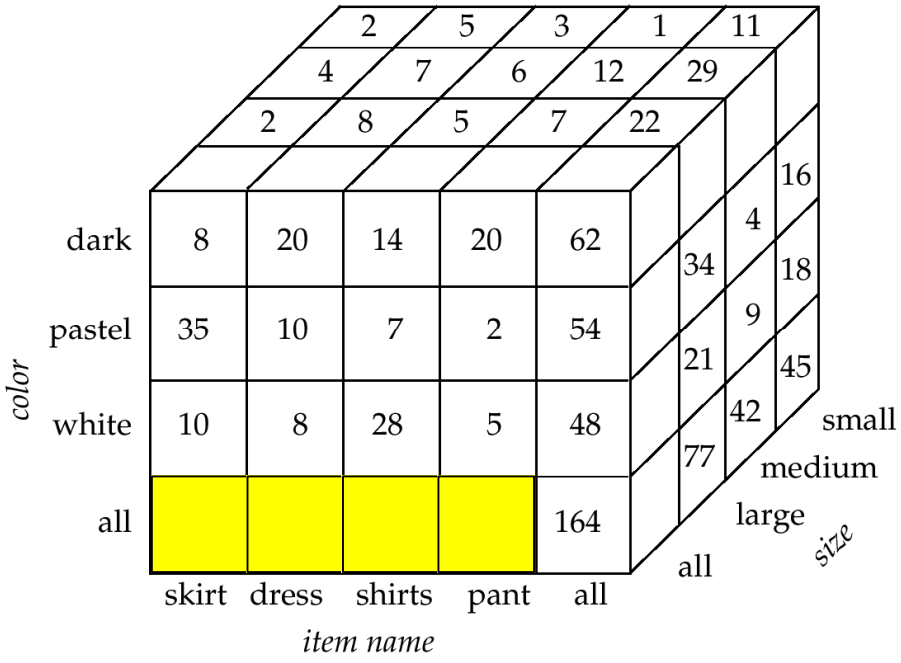
Yet Another Cube Slice



Sales by color and size - for all item_names

```
select color, size, sum(number)
    from sales
    group by color, size;
```

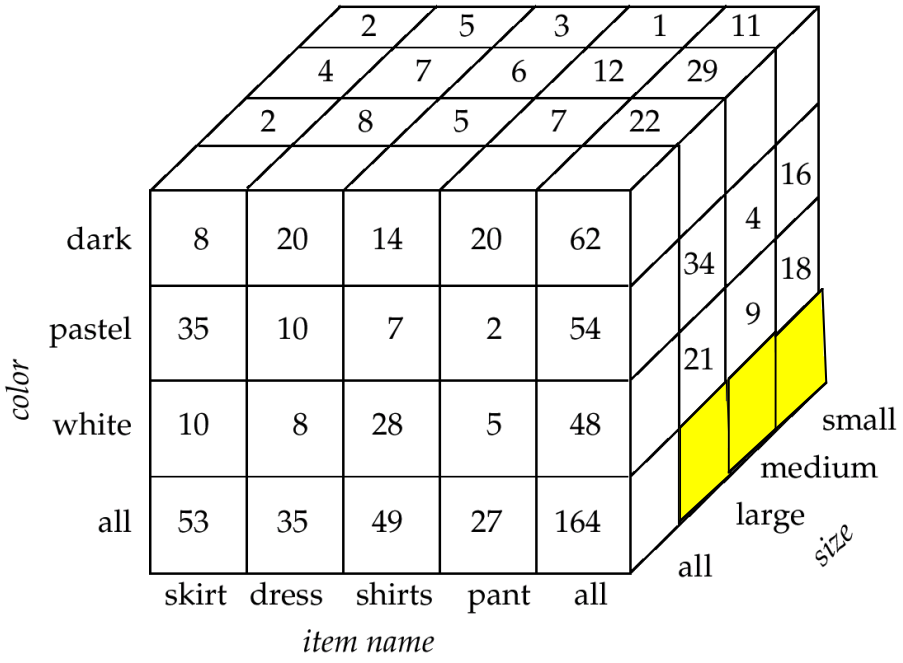
A Slice of a Cube Slice



Sales by item_name - for all colors and sizes

```
select item_name, sum(number)
    from sales
    group by item_name;
```

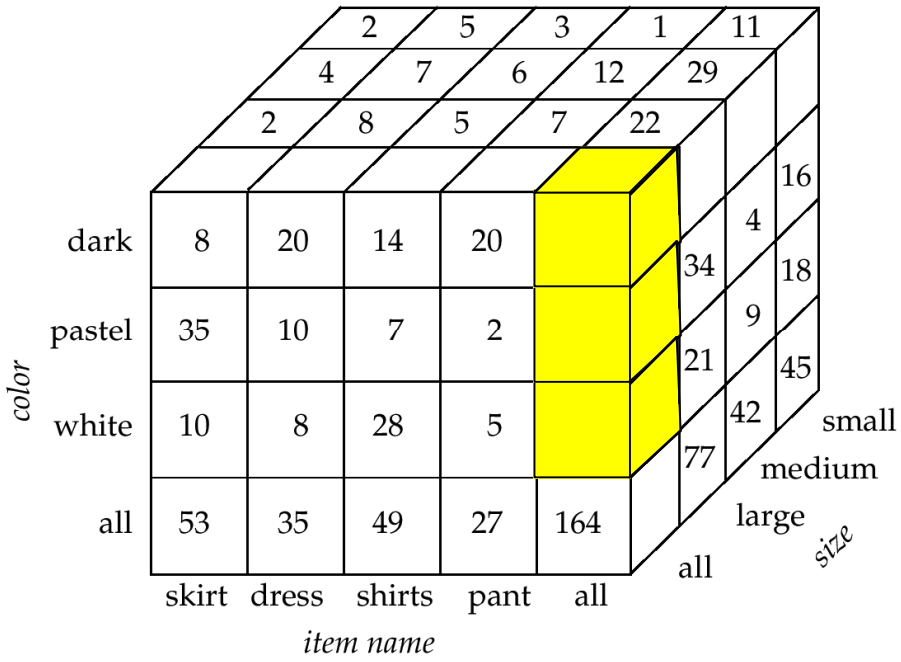
How About a Slice with Your Slice of Cube?



Sales by size - for all item_names and colors

```
select size, sum(number)
    from sales
    group by size, color;
```

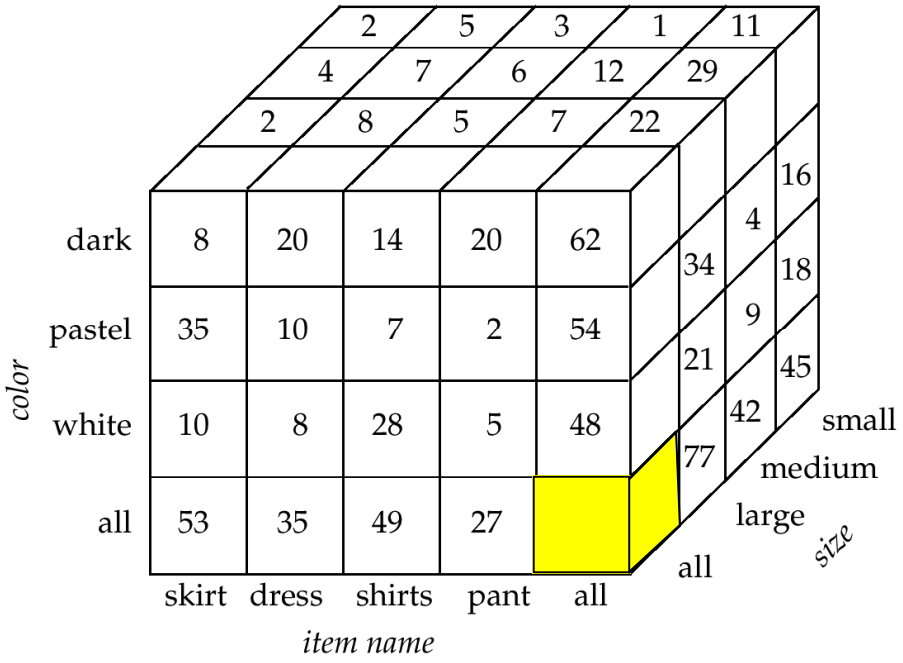
Aggregates by the Slice



Sales by color - for all item_names and sizes

```
select color, sum(number)
    from sales
    group by color;
```

Slice Cubed



Total sales for all item_names, colors, and sizes

```
select sum(number)
from sales;
```

Slicing with SQL

- 2^n SQL queries needed to generate all summary representations for a cube (where n = number of dimensions)
  - For item_name, color, and size (3 dimensions), $2^3 = 8$ queries

```
select item_name, color, size, sum(number)
   from sales
   group by item_name, color, size;
```

```
select item_name, color, sum(number)
    from sales
   group by item_name, color;
```

```
select item_name, size, sum(number)
   from sales
   group by item_name, size;
```

```
select color, size, sum(number)
   from sales
   group by color, size;
```

```
select item_name, sum(number)
   from sales
   group by item_name;
```

```
select color, sum(number)
   from sales
   group by color;
```

```
select size, sum(number)
   from sales
   group by size;
```

```
select sum(number)
   from sales;
```

SQL Cube Function

- cube ($dimension_1$, $dimension_2$, … $dimension$)
  - Used in the group by clause
  - Produces all summary representations in the cube
- Examples
  - select item_name, color, size, sum(number)
    from sales
    group by cube(item_name, color, size)
  - select job, education, sex, avg(salary)
    from (select job, case
    when edlevel >= 18 then 'GRADUATE'
    when edlevel >= 16 then 'COLLEGE'
    else 'HIGH SCHOOL'
    end as education, sex, salary
    from employee) as e
    group by cube(job, education, sex)
    order by job, education, sex;
    - Report on average salary based on job, education level, and gender.

Rollup

- Summarize data based on the first listed dimension
  - Similar to cube (which yields 2^n groups) for n dimensions
    - Includes all possible combinations of various dimensions and "all"
  - Yields n+1 groups for n dimensions
    - All the dimensions
    - All dimensions except the last
    - All the dimensions except the last and second to last
- rollup(*dimension*$_1$, *dimension*$_2$, … *dimension*) j --in group by clause
  - "Rolling up" dimensions from right to left…
- Example
  - select job, education, sex, avg(salary)
    from (select job, case
    when edlevel >= 18 then 'GRADUATE'
    when edlevel >= 16 then 'COLLEGE'
    else 'HIGH SCHOOL'
    end as education, sex, salary
    from employee) as e
    group by rollup(job, education, sex)
    order by job, education, sex;

Rank

- Rank records over dimension attributes
- rank() over (order by *dimension sort_direction*)
  - Used in select clause
  - Lower numbers mean higher rank (rank = 1 being highest)
- Example
  - select firstnme, lastname, salary,
    rank() over (order by edlevel desc) as edrank
    from employee
    order by edrank;
  - Examine the relationship between salary and educationlevel

Dense Rank

- Ranking function without skipping numbers
- dense_rank() over (order by *dimension sort_direction*)
  - Used in select clause
  - Lower numbers mean higher rank (rank = 1 being highest)
- Example
  - select firstnme, lastname, salary,
    dense_rank() over (order by edlevel desc) as edrank
    from employee
    order by edrank

Datatabases

...of the Bible

Law Breaker Analysis

Deuteronomy 6:1-10

Data Science

What is Data Science?

- "The science of systematically discovering patterns in very large data sets to extract useful knowledge and predict something of value." Udacity
- Data scientist has been called "the sexiest job of the 21st century."
- Organizations are producing a tremendous amount of data, and want to analyze and derive value from it.
  - Where to start? What questions to ask? What to look for?
  - Enter "Data Science"

What Does a Data Scientist Do?

- Data Wrangling/Munging
- Data Analysis
- Communication

Data Wrangling/Munging

- Obtain data from potentially disparate sources
  - From files, databases, APIs, spreadsheets, etc…
- Organize data
  - In some (large) storage solution (RDBMS, NoSQL, etc.)
  - So that it's easy to work with
- Clean data
  - Missing or incorrect values
  - Converting information to standard format
  - ETL
- Something of an art from
  - Get the data into a consistent format that lends itself to analysis

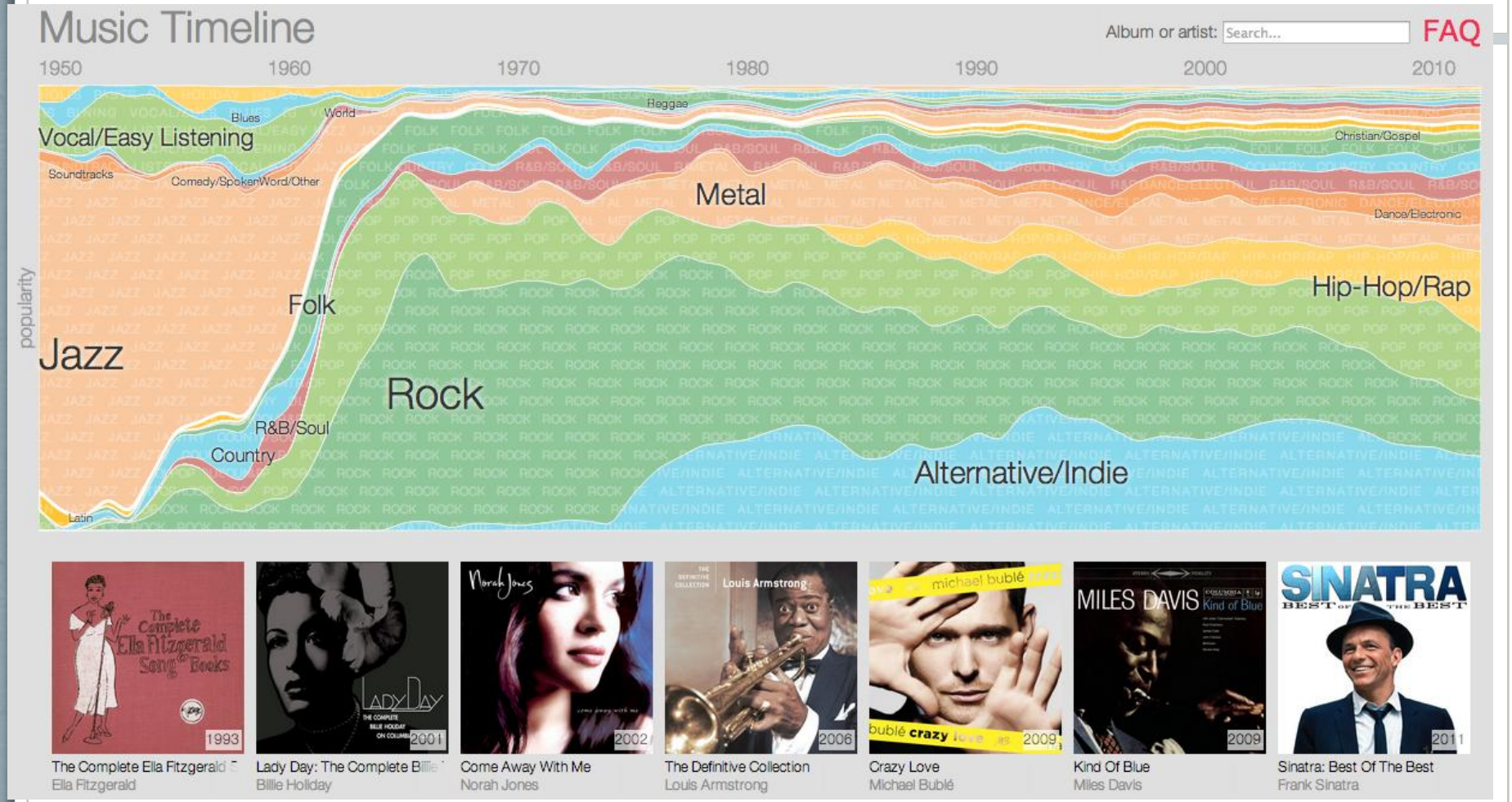
Data Analysis

- Explore and experiment – discover patterns in the data
- Create and apply algorithms to the data
  - Multivariate calculus and linear algebra
  - Statistics
  - Machine learning
- Interpret results and make predictions

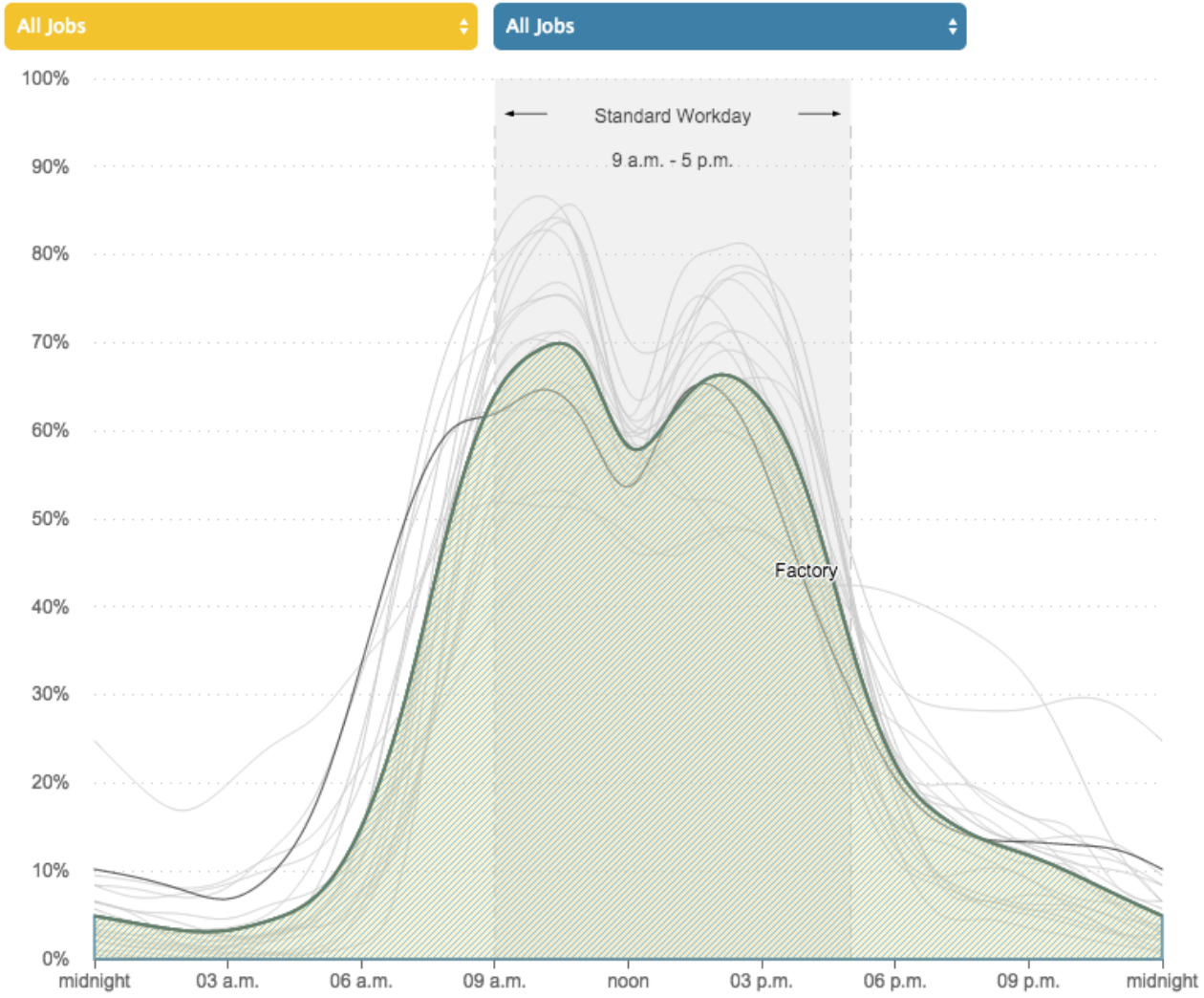
Communication

- Need to articulate complex findings in straightforward ways
  - Patterns in data
  - Algorithm results and interpretations
  - Recommendations
- Data visualization
  - Reports
  - Charts
  - Infographics

Sample Visualizations

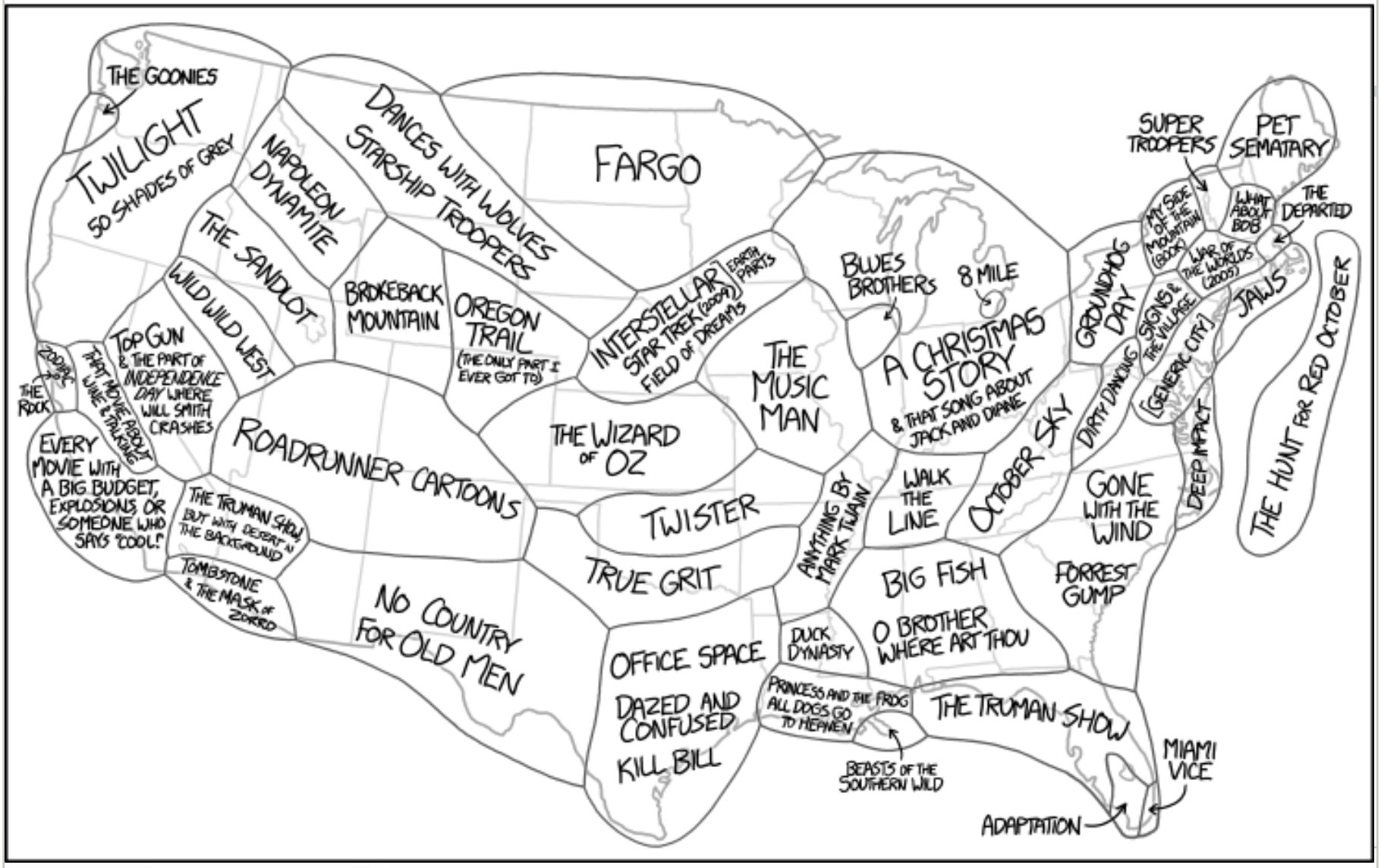


American Workday



A CHEAT SHEET FOR FIGURING OUT WHERE IN THE US YOU ARE BY RECOGNIZING THE BACKGROUND FROM MOVIES

(FOR USE BY GEOGUESSR PLAYERS AND CRASH-LANDED ASTRONAUTS)



R

- Programming language well-suited for manipulating and running computations on large data sets
  - Built-in types include vector, matrix, and data frame (like a spreadsheet)
    - Operations on these data structures are carried out on every data value in them
  - Built in functions to
    - Plot charts, histograms, etc.
    - Do advanced statistical operations (Chi-squared distribution, etc.)

Homework 8