CS 217 Data Management and Information Processing

01-Introduction

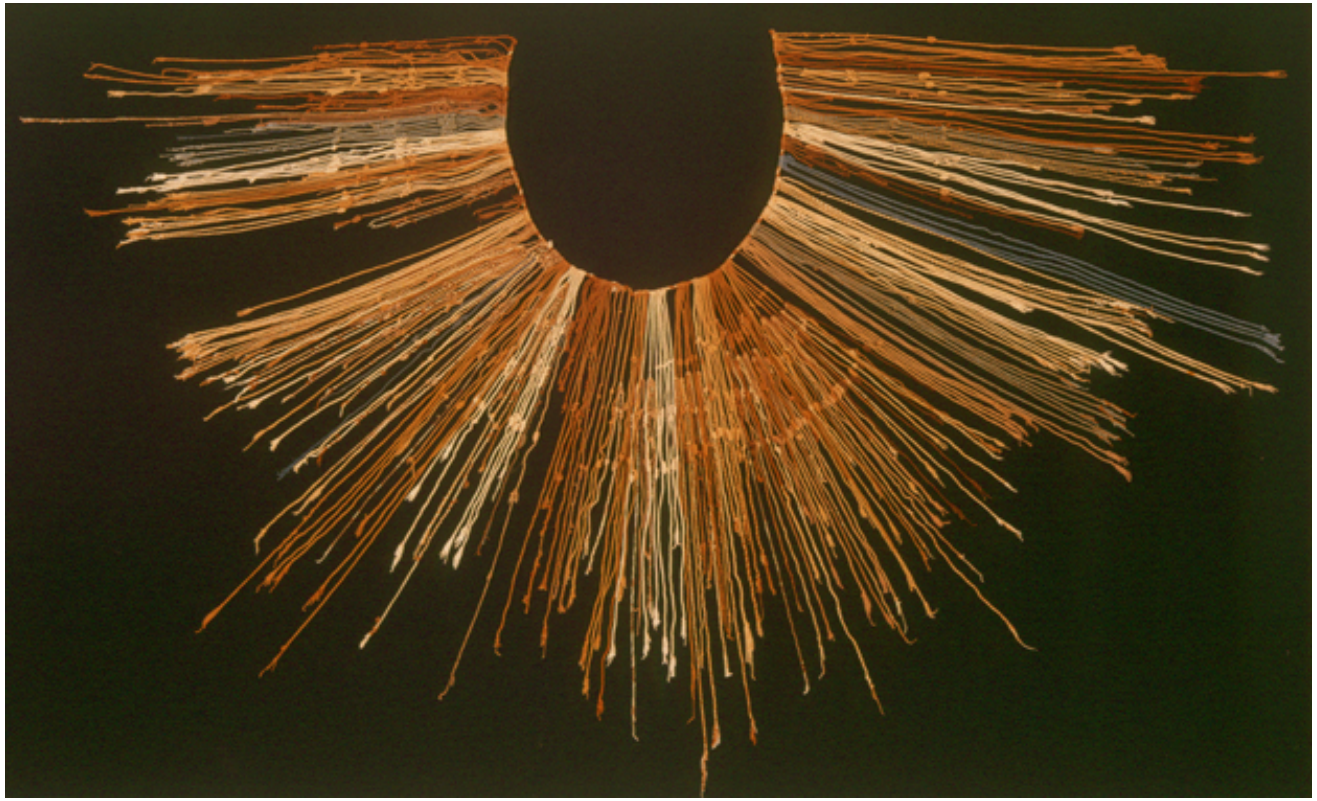
Instructor: Huiling Hu, Ph.D.

Spring 2020

Data Management and Information Processing

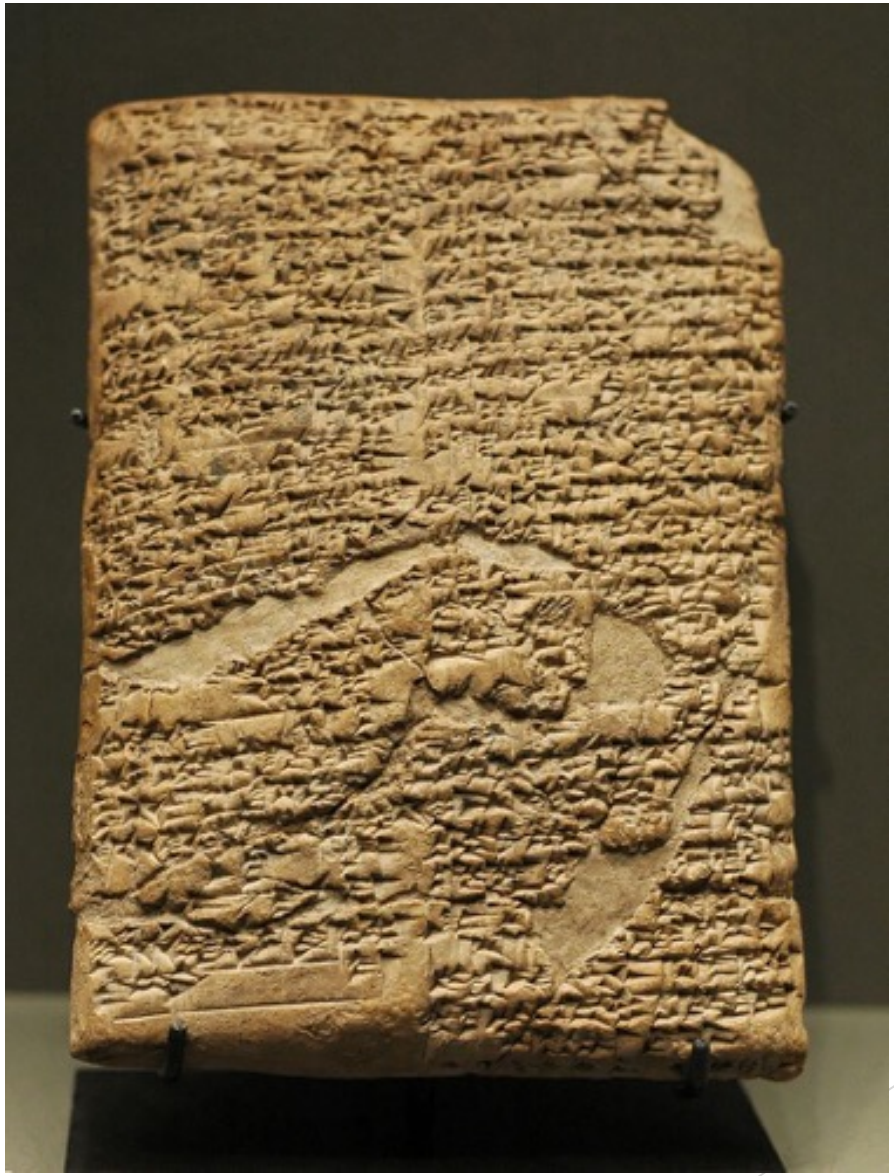
Data Management and Information Processing

What is/are Data?



Quipu (5000 years ago)
https://en.wikipedia.org/wiki/Quipu

What is/are Data?



Code of Hammurabi (~3500 years ago)

What is/are Data?

Books (~2000 years ago)

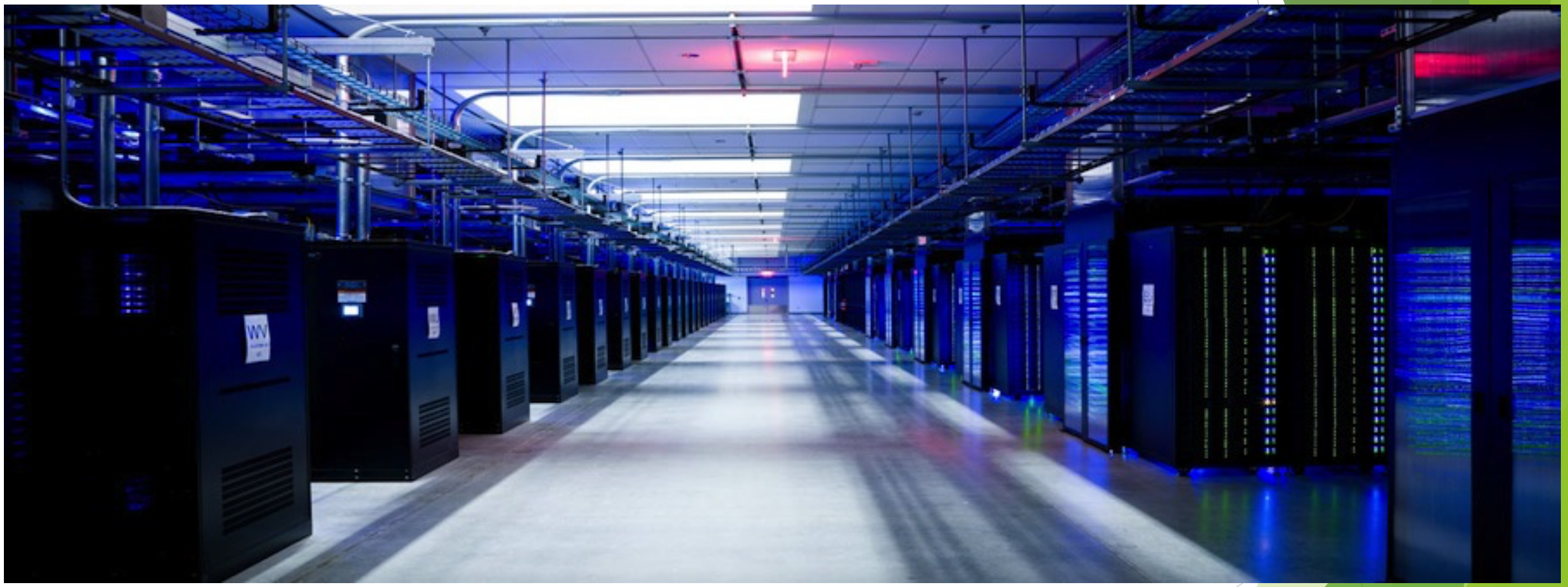
What is/are Data?

3.5-inch disk (~30 years ago)

Wha


Windows 8.1
with Internet Explorer11
Version 6.3
Disk 1 of 3711
起動ディスク - 1
Microsoft
セットアップの方法:
1. コンピュータの電源を切ります
2. [起動ディスク]をフロッピーディスクドライブに挿入します
3. コンピュータを再起動します
4. 画面に表示される指示に従い、操作します
©Copyright 1985 - 2013 Microsoft Corporation. All rights reserved

What is/are Data?



Data Warehouse (now)

90% of all data has been created in the last two years.

Data Management and Information Processing

To Understand Data Managing and Processing…

- How data is stored?
- How data is represented?
- How to process data?
- How to query data?
- How to design the “data relation”?

Part I: Data Fundamentals

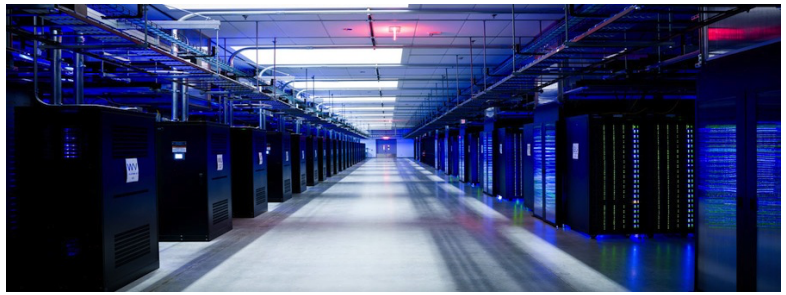
Store

Represent

Data

Data Representation and Storage

- All data ends up being stored as 0's and 1's.
  - Include numbers, text, …
- Data grows in large scale.



- We need to build structure in data!

Part II: Simple Data Processing using Pandas

- Very lightweight, intuitive to use
- Similar to array access, allow indexing
- Very easy to hook up with other components
- Sufficient for most small-scale projects

```
import pandas
flights = pandas.read_csv("flights.csv")
flights = flights[flights['dest'] == "ORD"]
flights = flights[['carrier', 'dep_delay', 'arr_delay']]
```

There are things you *cannot* do with it

- Complex data models
  - Every row in a table has a fixed number of columns
  - Can’t model one-to-many and many-to-many relationships
  - You can try using multiple spreadsheet tabs or multiple matrices for different types of data, but linking them is difficult
- Enforce data integrity constraints
- Keep data and analysis separate

Insight: data are not just numbers

- "Simple" data sets are just arrays or matrices of numbers:
  - Time-series of stock price data
  - Matrix of pixel colors in an image
  - 3D "matrix" of atmospheric temperatures in a weather simulation.

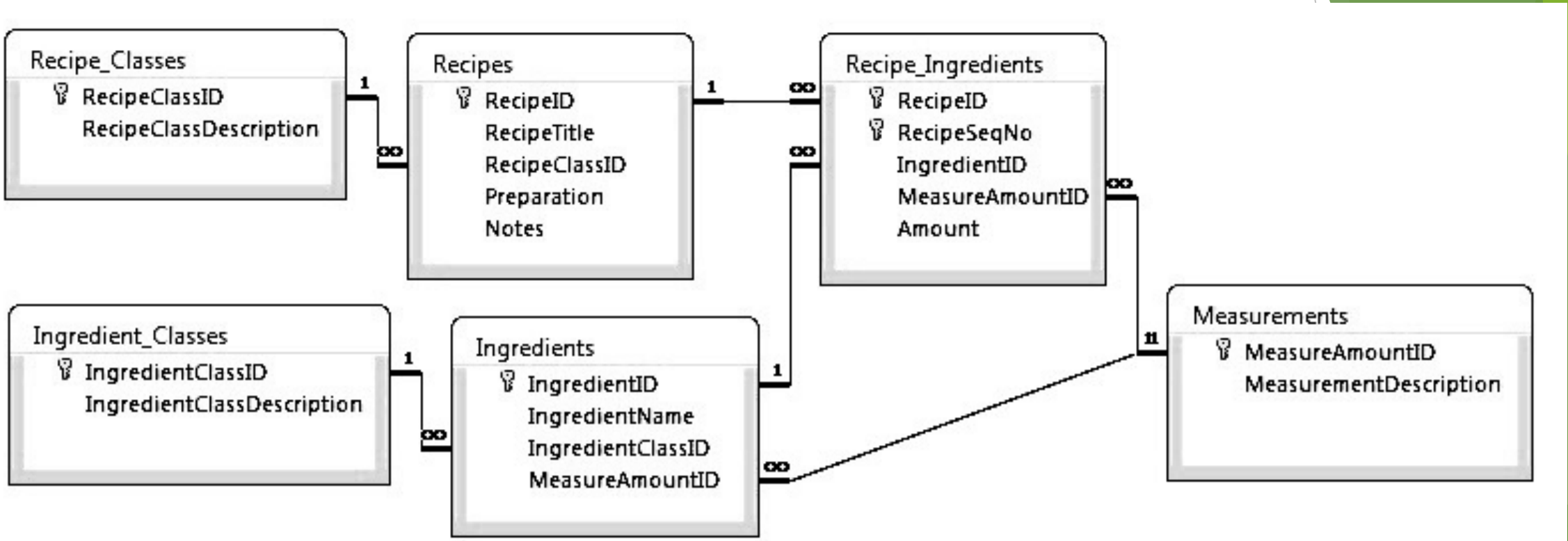
- Complex data also represent **relationships**
  - For example, the course scheduling information at Northwestern
  - It's not just a sequence of numbers.
  - It's a complex web of students, professors, courses, classrooms, grades, etc.
  - This course will teach you how to handle such data.

Part III: SQL

- More structured data
- Easy to manage complicated relationship in data

- Focus of this course.
  - More than half of the lectures will be on SQL

SQL database example



- This is the data **schema**.
  - Refers to how it's organized, not the recipe data itself.
- First design the structure of the data, then fill it in.

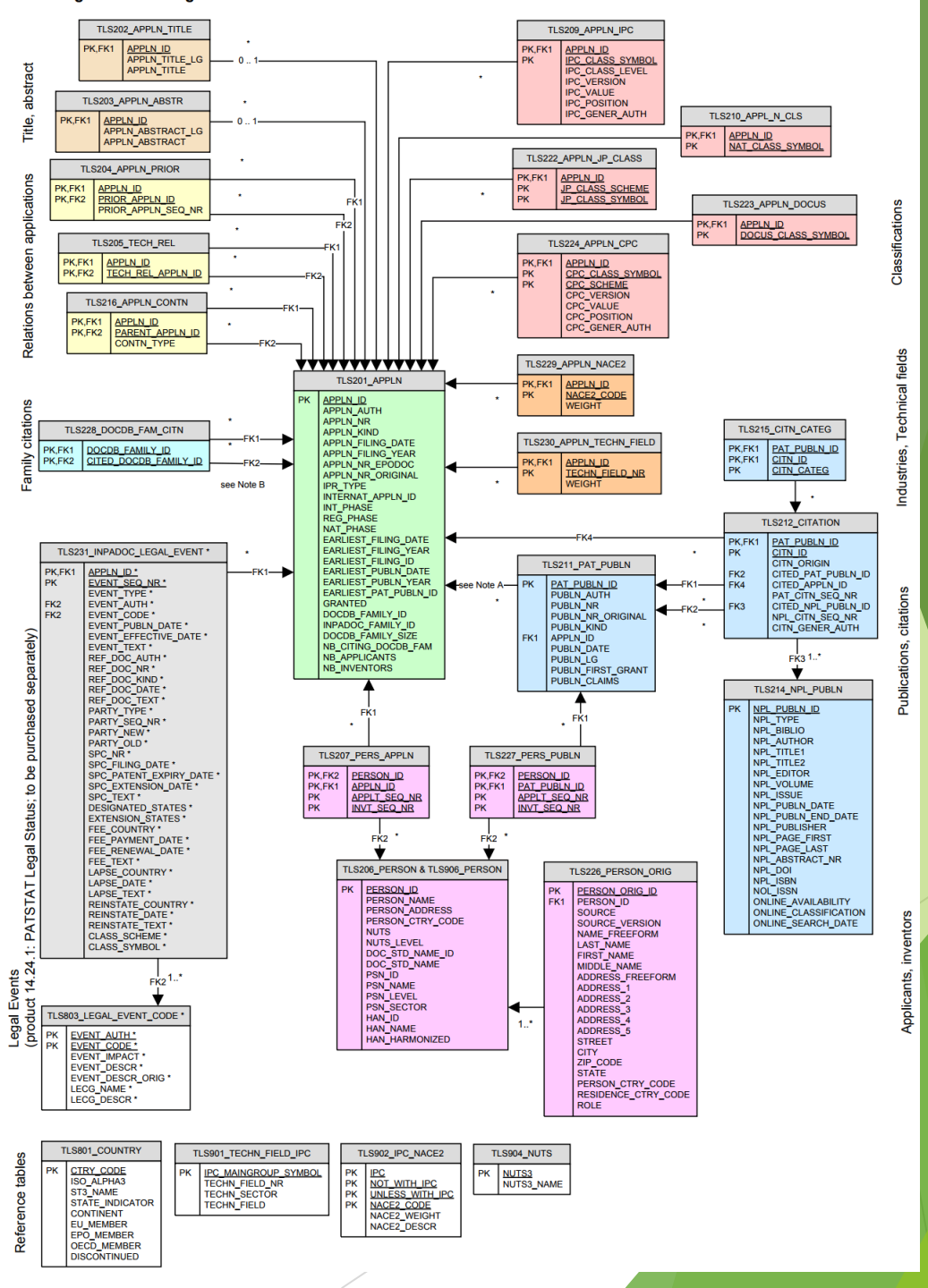
Questions to be answered from the recipe DB

- How many steps are in the “Chocolate chip cookie” recipe?
- What are the titles of the recipes that have seafood ingredients?
- Do any recipes use the same ingredient twice?
- Which recipe has the greatest number of steps?
- Etc.

PATSTAT: European Patent Office's International Patent Database

- 28 cross-referenced tables
- 6 DVDs of data
- 119GB of CSV files after unzipping

- This example has both complex structure and lots of data entries.



The Goal: Easy & Clean Descriptive Analytics

Answer a wide variety of complex questions using the same database:

- Where did our 10 biggest customers live in 2007?

```
SELECT customer.name, customer.city FROM
  customer JOIN order ON order.customer == customer.id
           JOIN order_item ON order_item.order == order.id
           JOIN product ON order_item.product = product.id
  WHERE order.placed >= “2007-01-01” AND order.placed < “2008-01-
01”
  GROUP BY customer.id
  ORDER BY SUM(order_item.qty * product.price) DESC
  LIMIT 10;
```

This is code in the SQL language.

- How many widgets are left in stock?
- What is the average price of the chairs we sell?

After taking this course

- You will be able to have a better understanding on how computer store and process data.
- You will be able to use appropriate tools to manage and process data, quickly.