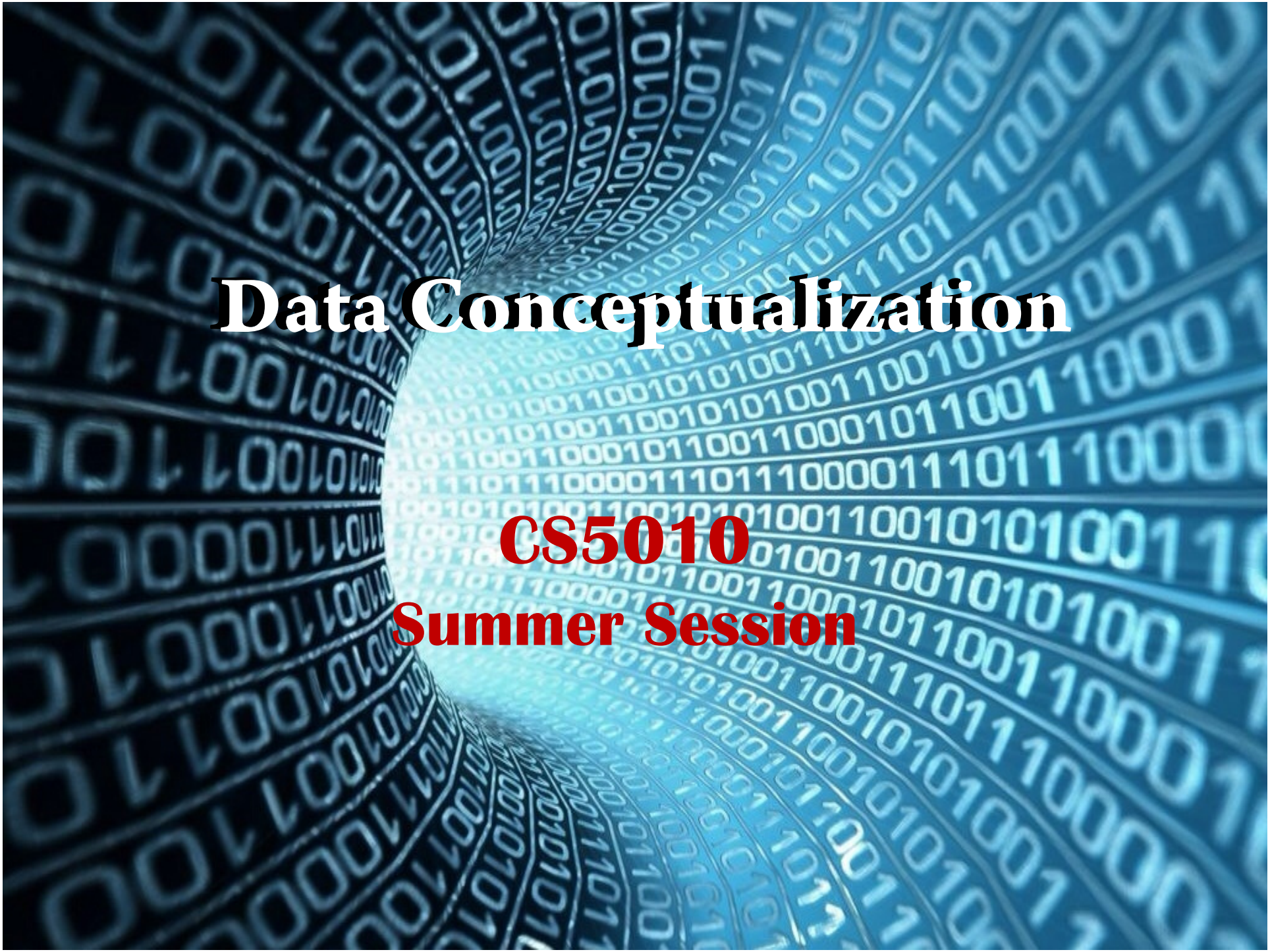

Data Conceptualization
CS5010
Summer Session

Welcome!

MS in Data Science program!

Data, Data Everywhere…!

"Water, water, every where,
And all the boards did shrink;
Water, water, every where,
Nor any drop to drink."

*-- excerpt from "**The Rime of the Ancient Mariner**"*
by Samuel Taylor Coleridge

- One of the poem's most famous lines – "Water, water, every where, Nor any drop to drink" could easily describe the state of **data** today

Data, Data Everywhere…!

- "Data, Data Everywhere, and Not A Drop of Value"
- "Data, Data Everywhere, and Nor Any Byte to Think"
- "Like the seven seas, today's volume of data is almost unfathomably vast" ~ J. Sorofman, Gartner.com
- "Drowning in Data and Starving for Information!"

Discussion: What is Data?

What is Data?

- Raw data (*unprocessed* data)
  - a collection of numbers, characters,
  - a set of values of qualitative or quantitative variables
  - is collected, the result of (a) measurement(s)
  - sometimes called facts
  - as an abstract concept, can be viewed as the lowest level of abstraction
  - Data is *uninterpreted* information

Language: Data or Datum?

- Note: The word "**data**" used to be considered as the plural of "**datum**", but now it is generally used in the singular, as a mass noun
- Datum: "the datum is very high"
- Data (as a collection): "the data is available" [common]
- Data (as plural): "the data are available" [ok, too]

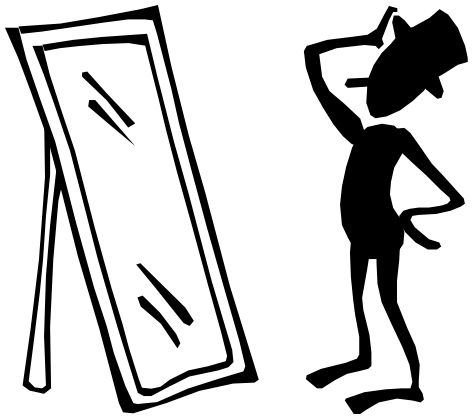
- Bottom line: decide on how you're going to use the word and *stick with one way!*

Data

- Even unremarkable objects such as golf balls can have a lot of data attached to them
- How about you? What data is associated with **you**? (What attributes do you possess?)

Data

- What about you?:
  - You have a name (most people have first and last names)
  - UVa computing ID
  - Home address
  - Date of birth
  - Weight
  - Height
  - Eye color
  - Nationality
  - …
- All these things are data



Golf Balls–What can we say about these?

- They're golf balls, so used in the sport Golf

  Category → Sport - Golf
- Color → white
- Texture → rough
- Condition → used
- …
- All have a *size*, there are a certain *number* of them, they probably have some *monetary value*, … etc.

There are different types of data

- Qualitative Data - *What "qualities" does it have?*

Is descriptive information (describes something).
Related to the *quality* of something

A description of colors, texture and feel, … etc of an object

A description of experiences

A **red**, **smooth** ball

There are different types of data

- Quantitative Data

Is numerical information (refers to numbers)

Number of "things"

Size of an object

Price of an item

Score on a test

…

Four kittens

There are different types of data

- Quantitative Data → Discrete Data

Numerical data that can only take certain values (e.g. whole numbers) ("has gaps") Is often "counted"

The *count* of the number of items – there can only be whole numbers of golf balls *(there is no such thing as 10.3 golf balls!)*

Scores in tests (e.g. 8/10)

Shoe sizes

There are different types of data

- Quantitative Data → Continuous Data

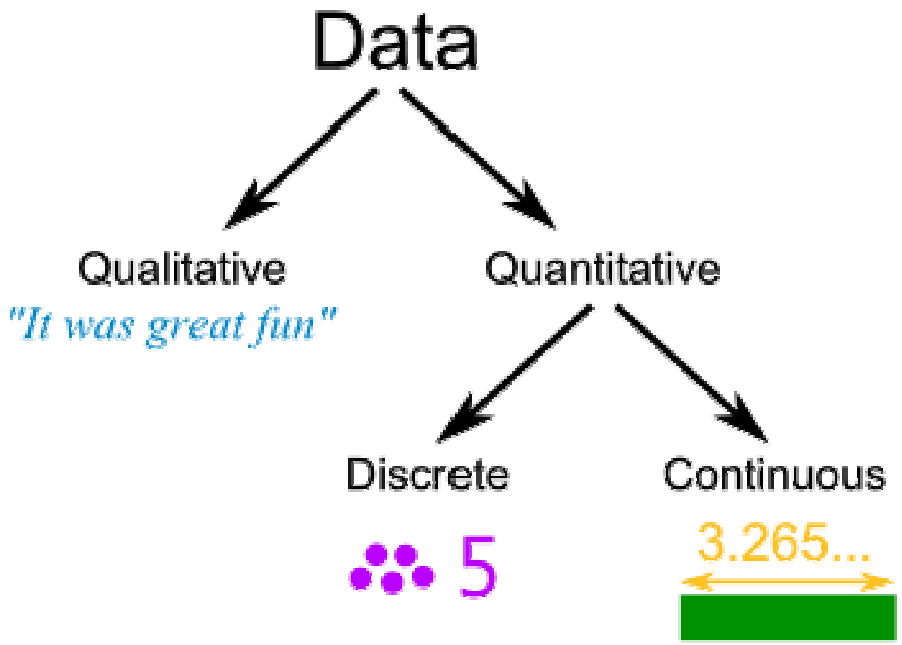
Numerical data with a continuous range. All values are possible with no gaps in between. Is often "measured"

Size of golf balls (e.g. 10.53mm, or 10.54mm, or 10.536mm)

Length of your foot (vs. your shoe size, which is discrete)

Time in a race (measured to fractions of a second)

Qualitative vs. Quantitative



There are different types of data

- Categorical Data

Puts the item you are describing into a **category** (giving an item a "label"). A variable that can take on a limited (and usually fixed) number of possible values

ITEM CONDITION: "new" / "used" / "broken"

A KIND OF RISK: "high" / "medium" / "low"

BLOOD TYPE: "A" / "B" / "AB" / "O" / …

TYPE OF ROCK: "igneous" / "sedimentary" / "metamorphic."

Some properties of data

- Consider these factors about the data you are using:
  1. Source: Where does the data come from?
     - Trustworthy/reliable/unbiased source?
  2. Quality: Is the data Accurate?
     - Not only how accurate data is, but do the providers have a process to ***maintain*** data accuracy?
     - Related: Is the data "clean" – can have different meanings depending on context, but generally: no missing data, all attributes clearly labeled (nothing ambiguous), data types are uniform (for a given attribute), all data in a useable form …

K Malik, eXelate at http://www.adexchanger.com/

Some properties of data

- Consider these factors about the data you are using:
  3. Scale: Is there enough data to draw meaningful conclusions
     - Is the data representative of the domain in question? Is there more data that can be obtained?
  4. Variety: What types of data are available?
     - Business/advertising example for audience targeting:

       Can it provide, for example, offline, online, personal, anonymous, mobile, PC, branded, blended, demo, interest and/or intent data?
     - Depending on context/situation this may be on of the more important factors …

K Malik, eXelate at http://www.adexchanger.com/

Data by itself is useless

- Data (big, medium, or small) has no value in and of itself. The value of data is extracted through **context** and **presentation**

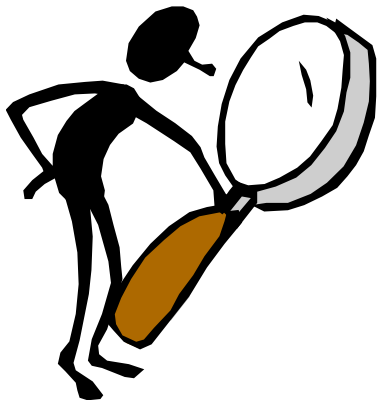
Two case studies highlighting context and presentation of data:

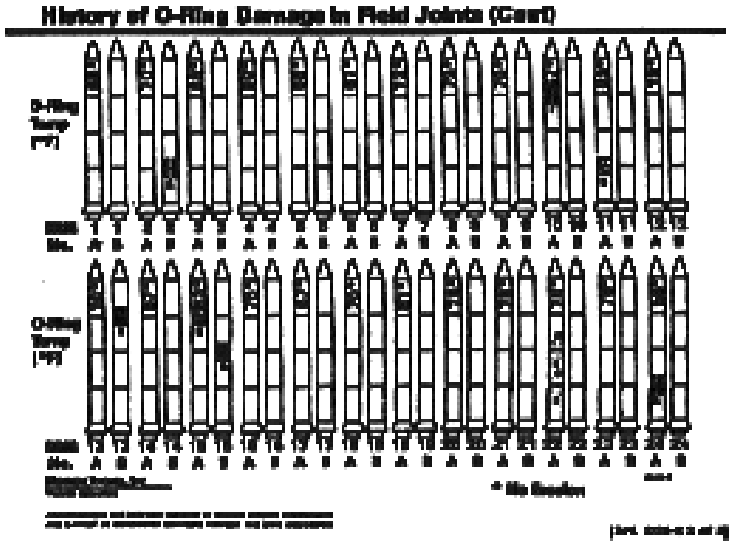
1. Edward Tufte's redesign of the Space Shuttle Challenger data
   - Why did the space shuttle Challenger explode? People speculate it was due to poorly functioning "**O rings**" on the booster rocket. However, these O rings didn't send that ship up on a **cold** winter's morn… people did (due to limiting their view of the data)!
2. A study to measure the effect of data presentation on the consumption of sugary soda
   - Altering the context and presentation of the data had a major impact on the buying behavior of teens! This study was led by Sara Bleich, an assistant professor of health policy at the Johns Hopkins Bloomberg School of Public Health

Bryce.vc, "Data data everywhere"

Case Study 1: Challenger Data

- The Engineers looked at graphs portraying tiny pictures of each shuttle booster, lined up in *chronological* order, showing launch temperatures and any O ring damage
- *Effectively looked like so many crayons in a box!*

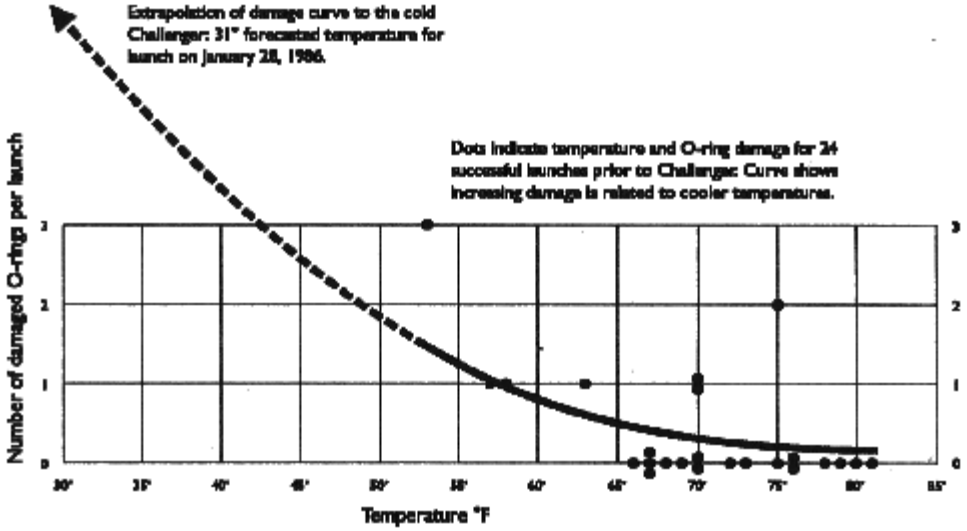




The Challenger: An Information Disaster - http://www.asktog.com/books/challengerExerpt.html

Case Study 1: Challenger Data

- Edward Tufte's reorganization of the data. He took the same information but presented it in a form that even a child could understand. In a single glance anyone can see that cold temperature strongly correlates with an increase in the number of damaged O-rings per launch!



Number of damaged O rings per launch

Temperature *F

The Challenger: An Information Disaster - http://www.asktog.com/books/challengerExerpt.html

Case Study 2: Exercise vs Calorie Count

- But what if you knew that it would take 50 minutes of jogging to burn off one soda?

- When researchers taped signs saying just that on the drink coolers in four inner-city neighborhood stores, sales of sugary beverages to teenagers dropped by 50 percent.

- That tactic was more effective than a sign saying that the drinks had 250 calories each, or a sign saying that a soft drink accounts for 11 percent of recommended daily calories.

Exercise Info, Not Calorie Counts, Helps Teens Drop Sodas - http://www.npr.org/blogs/thesalt/2011/12/16/143790349/exercise-info-not-calorie-counts-helps-teens-drop-sodas

Case Study 2: Exercise vs Calorie Count

- Merely listing the calories (250) in a drink (*which is already listed on the bottle*) seemed to have ***no effect.***
- Instead of buying soda or fruit juice, many kids who read the sign picked water instead

- The underlying data itself didn't change. The context and presentation of the data changed
- And as a result, so did ***behavior!***

Exercise Info, Not Calorie Counts, Helps Teens Drop Sodas - http://www.npr.org/blogs/thesalt/2011/12/16/143790349/exercise-info-not-calorie-counts-helps-teens-drop-sodas

Discussion

- Have you come across any advertising or other presented information (journal, news, magazine, flyer …) that *made you change your behavior/attitude?*
- Explain how the way the data was presented resulted in a change in your behavior and/or attitude

Simple example: Buying Sunglasses

- Would you buy sunglasses based on:

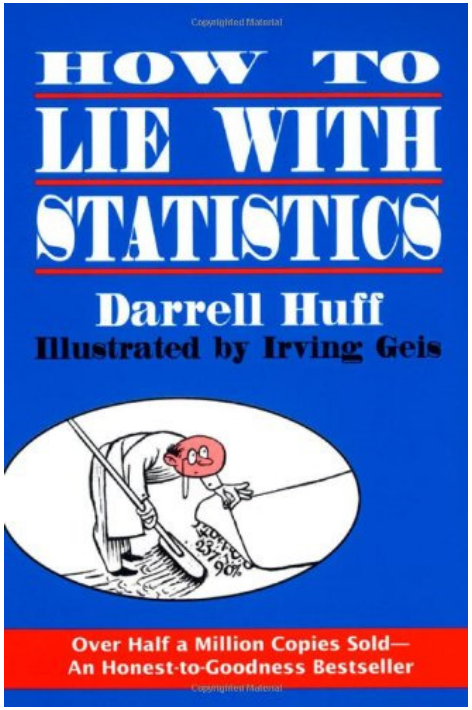
Simple example: Buying Sunglasses

- Or…

UV Protection never looked so cool! Protect your eyes in style!

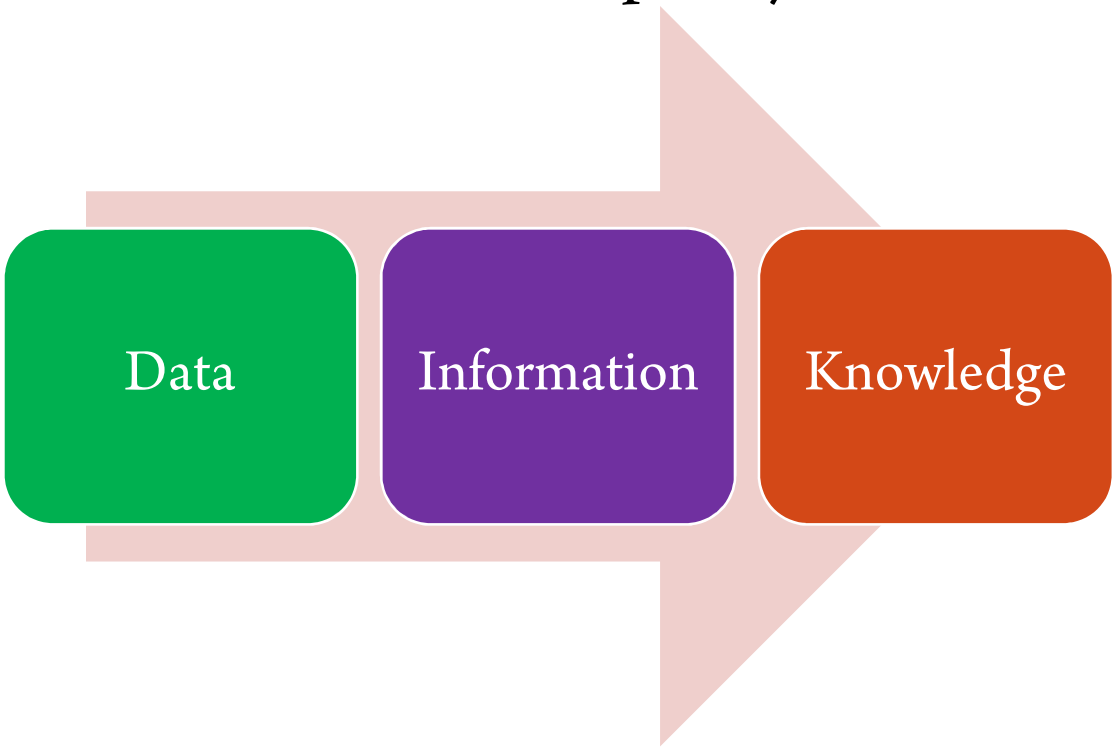
!!!Caution!!!

- Data presentation…
  - Can mislead
  - Can misinform
  - Can even be totally incorrect



Data – Information - Knowledge

- "Data", "information", and "knowledge" are closely related terms. Some people mistakenly use them interchangeably. This is ***not*** correct
- These three terms mean completely different things



Data by itself is useless

- Data, when collected and structured suddenly becomes a lot more useful. Let's revisit the golf ball

Color	White
Category	Sport – Golf
Condition	Used
Diameter	43mm
Price (per ball)	$0.5 (AUD)

- But still, each of the data values is still rather meaningless by itself. To create information out of data, we need to interpret that data

SchoolOfData.org, 'What is Data?'

From Data to Information to Knowledge

Color	White
Category	Sport – Golf
Condition	Used
Diameter	43mm
Price (per ball)	$0.5 (AUD)

- Let's take the size: A diameter of 43mm doesn't tell us much (**Data**). It is only meaningful when we compare/associate it to other things. In sports there are often size regulations for equipment. The minimum size for a competition golf ball is 42.67mm. Good, we can use that golf ball in a competition. This is **information**. But it still is not knowledge. **Knowledge** is created when the information is learned, applied and understood

SchoolOfData.org, 'What is Data?'

Data – Information - Knowledge

- ***Data*** [recall] – a collection of facts, such as values or measurements. Can be numbers, words, characters, measurements, observations, or even just descriptions of things
- ***Information*** – the patterns, associations, or relationships among the data can provide information
  - E.g. analysis of retail point of scale transaction data can yield information on which products are selling and when
- ***Knowledge*** – information can be converted into knowledge
  - Through extensive experience, insight, understanding, and contextualized information. Knowledge enables us to act
  - E.g. information converted into knowledge about historical patterns and future trends. Summary information on retail supermarket sales can be analyzed along with promotional efforts to provide knowledge of consumer buying behavior
  - Thus a manufacturer or retailer could determine which items are most susceptible to promotional efforts

Another Set of Definitions

- **Data**

Facts and *figures* which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc. Definition by Thierauf (1999): "*unstructured facts and figures that have the least impact.*"

- **Information**

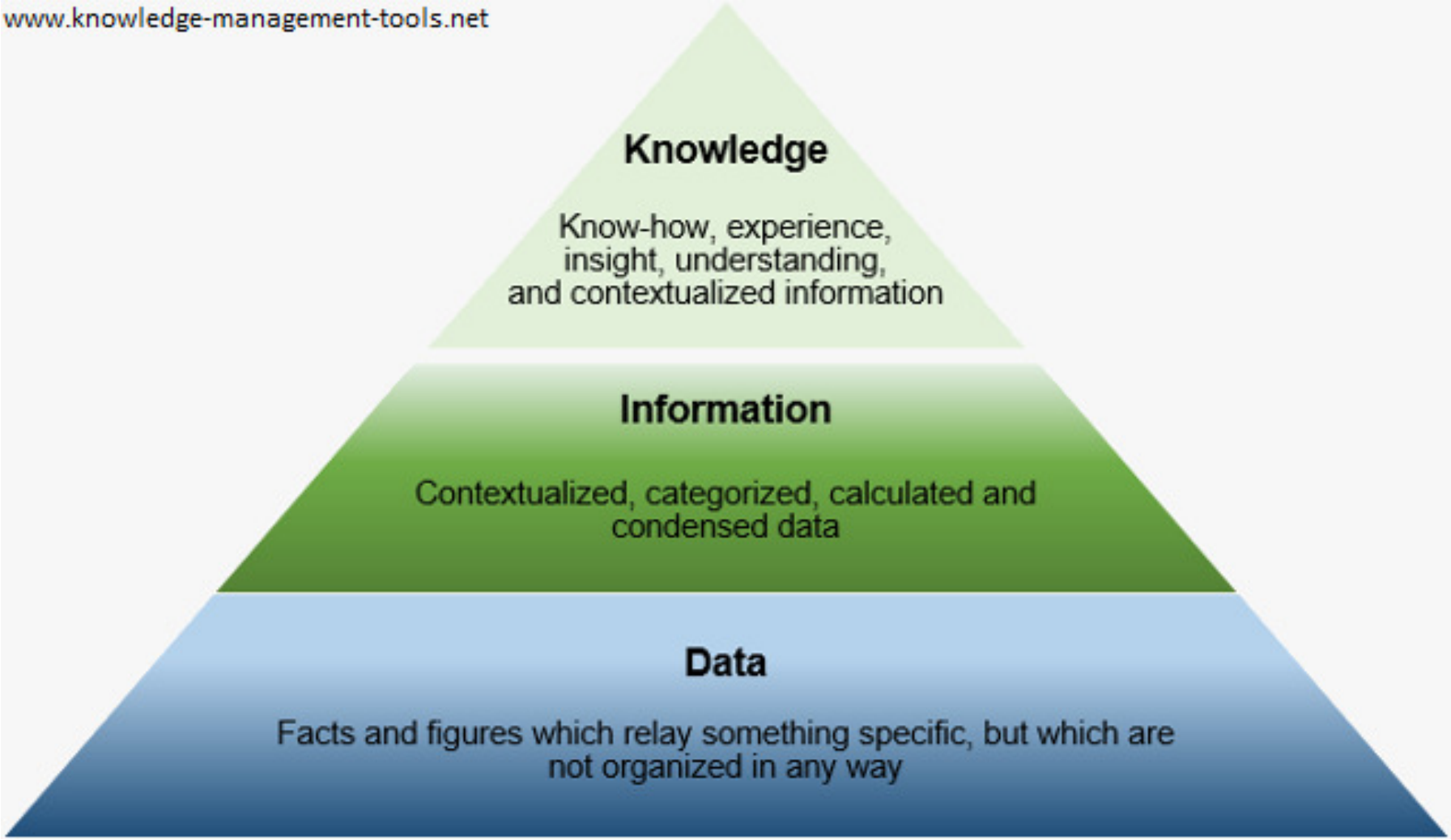
For data to become information, it must be *contextualized*, *categorized*, *calculated* and *condensed* (Davenport & Prusak 2000). Information thus paints a bigger picture; it is data with *relevance* and purpose (Bali et al 2009). It may convey a *trend* in the environment, or perhaps indicate a *pattern* of sales for a given period of time.

Another Set of Definitions

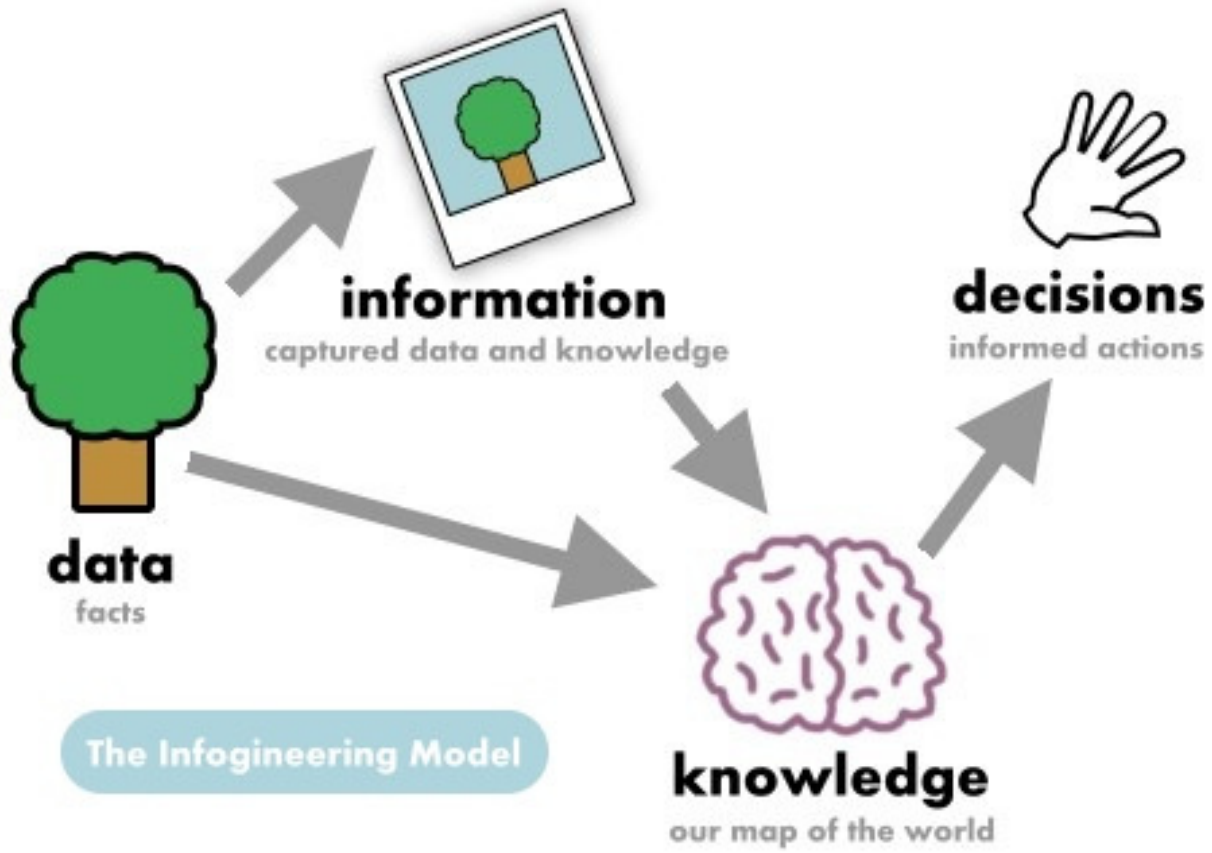
- **Knowledge**

Definition by Gamble and Blackwell (2001): "Knowledge is a fluid mix of *framed experience*, *values*, contextual information, *expert insight*, and grounded *intuition* that provides an environment and framework for *evaluating* and incorporating new experiences and information. It originates and is applied in the mind of the knowers. In organizations it often becomes embedded not only in documents or repositories, but also in organizational *routines*, *practices* and *norms*."

Data – Information - Knowledge



Data – Information - Knowledge



[www.infogineering.net]

Data – Information - Knowledge

An example:

- Data becomes information by interpretation; e.g., the *height of Mt. Everest* is generally considered as "data“
- A *book on Mt. Everest geological characteristics* may be considered as "information"
- A *report* containing practical information on the *best way to reach Mt. Everest's peak* may be considered as "knowledge"

Data – Information - Knowledge

Another small example:

- **Data** – a list of numbers having no context
  - `51, 77, 58, 82, 64, 70` *–What could this mean?*
- **Information** – the list of numbers are student test scores
  - Now that some context is included, the data starts to make more sense: *each number is a test score belonging to a student*
  - Further information can be obtained by carrying out some analysis of the data – e.g. average score = 67. Student 1 did particularly badly in the test because they were so far below the average mark
- **Knowledge**
  - The class teacher can apply the rule "*If a score is much lower than the average, then discuss it with the student to see what needs to be done to improve it.*"

Data Science

- Merely using data isn't what data science is about
- A data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product
- Therefore, one of the things data science can do is enable the creation of **data products**

- Discussion: two small case studies illustrating data products
  - CDDB
  - Google Swine Flu tracking

Case Study 1: CDDB

- An early data product on the Web was the **CDDB database**

- Developers realized that every CD had a *unique signature*, based on the exact length (in samples) of each track on the CD
- Gracenote built a database of track lengths, and coupled it to a database of album metadata (track titles, artists, album titles)
- If you ever used iTunes to rip a CD, you've taken advantage of this database. Before it does anything else, iTunes reads the length of every track, sends it to CDDB, and gets back the track titles

Case Study 1: CDDB

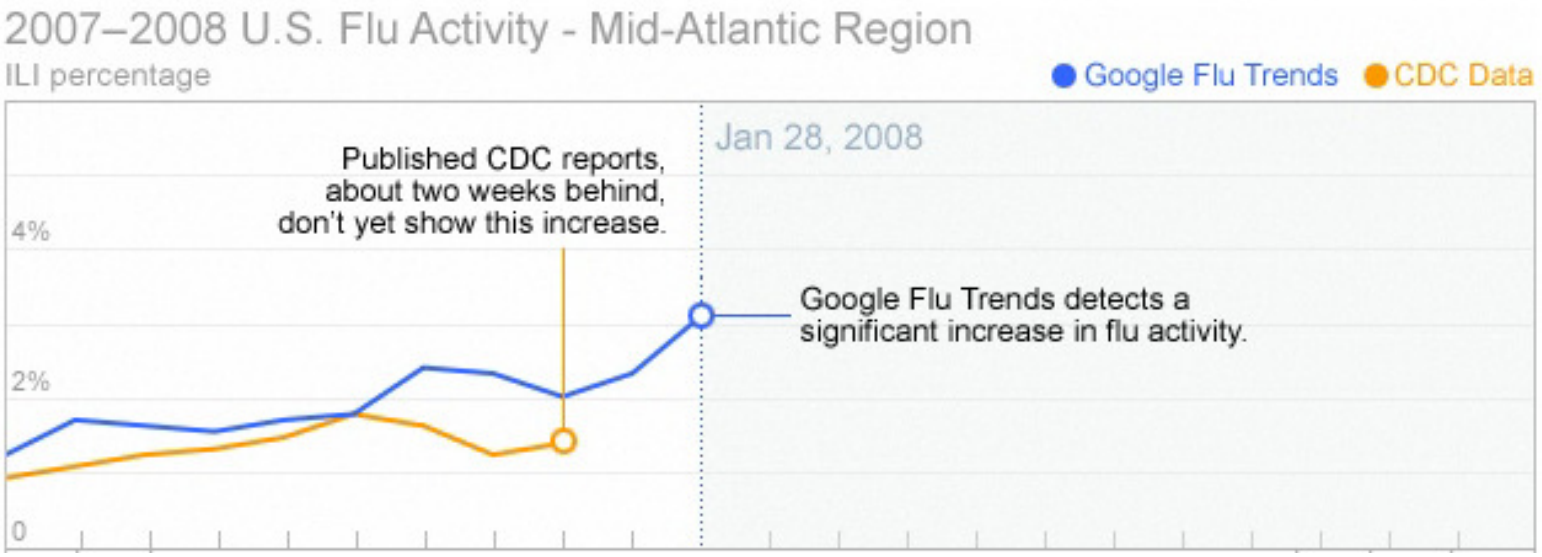
- CDDB views music as **data**, not as audio, and creates new value in doing so
- Their business is fundamentally different from selling music, sharing music, or analyzing musical tastes (though these can also be "data products")

- CDDB arises entirely from viewing a musical problem as a DATA PROBLEM

Case Study 2: Google Flu Tracking

- Around 2009 there was a Swine Flu epidemic
- The CDC (Center for Disease Control) tries to track of this flu activity by estimating the percentage of infected people
- Google was able to track the progress of the epidemic by following searches for flu-related topics. Google spotted trends in the Swine Flu epidemic roughly two weeks before the CDC by analyzing searches that people were making in different regions of the country

Case Study 2: Google Flu Tracking



- **Google** certainly isn't the only company that knows how to use data.
- **Facebook** and **LinkedIn** use "patterns of friendship relationships to suggest other people you may know, or should know, with sometimes frightening accuracy"

O'Reilly Radar – Data Science

Case Study 2: Google Flu Tracking

- **Amazon** also saves your searches, "correlates what you search for with what other users search for, and uses it to create surprisingly appropriate recommendations."
- These recommendations are "data products" that help to drive Amazon's retail business. Amazon understands the value of a customer and that customers "generate a trail of 'data exhaust' that can be mined and put to use"

Some Quotes To End…

Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution. They are inherently interdisciplinary. They can tackle all aspects of a problem, from initial data collection and data conditioning to drawing conclusions. They can think outside the box to come up with new ways to view the problem, or to work with very broadly defined problems: *"here's a lot of data, what can you make from it?"*

-- **M. Loukides**, at Radar O'Reilly

Some Quotes To End…

The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades.

-- **Hal Varian**, professor of information sciences, business, and economics at the University of California at Berkeley (Also, Google's chief economist)

We hope you all will be successful Data Scientists!