Introduction to Data Science

What is Data Science?

Gordon Anderson

What is Data Science?

- "The key word in "Data Science" is not Data, it is Science".

- Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining and predictive analytics, as well as Knowledge Discovery in Databases (KDD).

-Dhar and Leek 2013

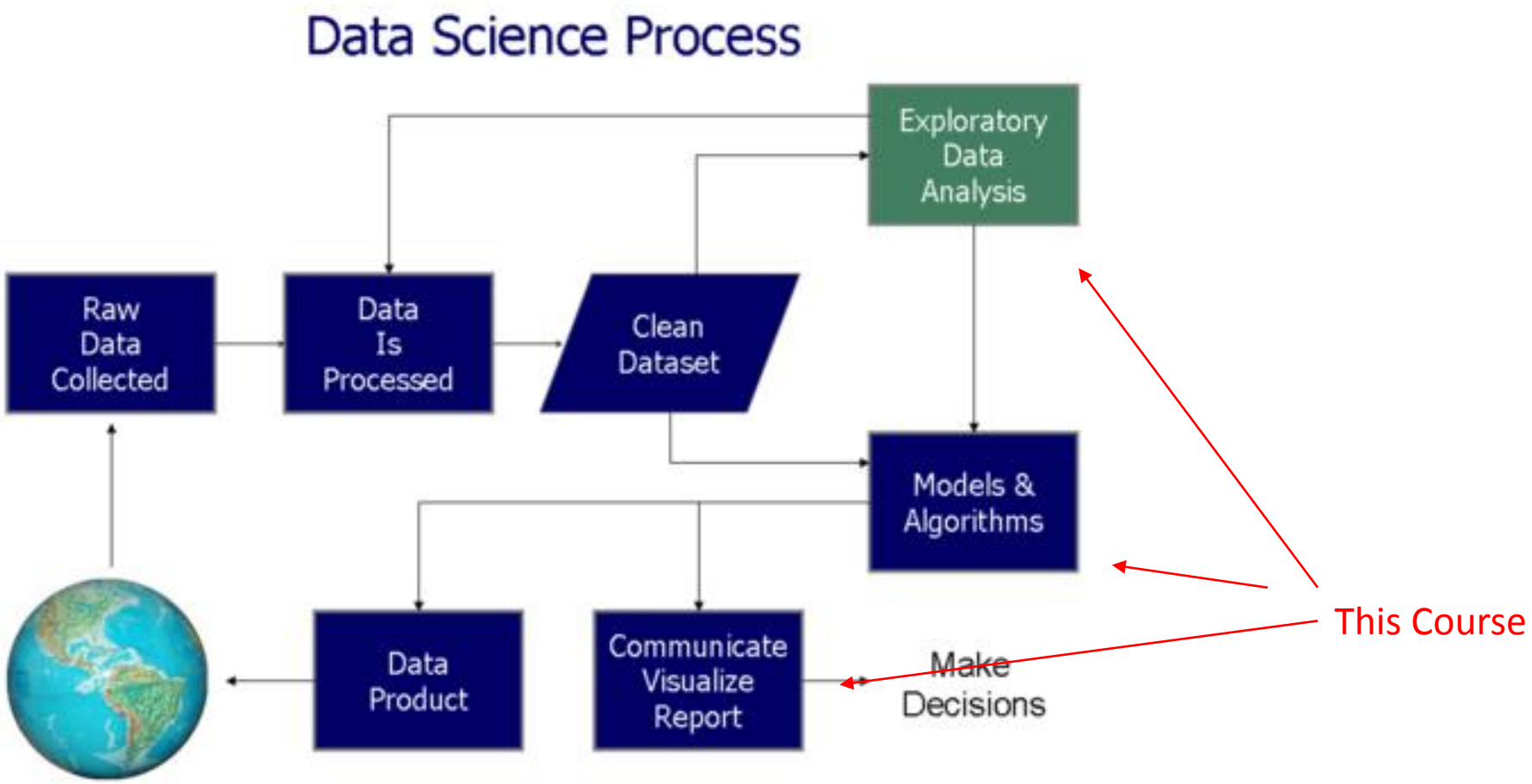
What is Data Science?

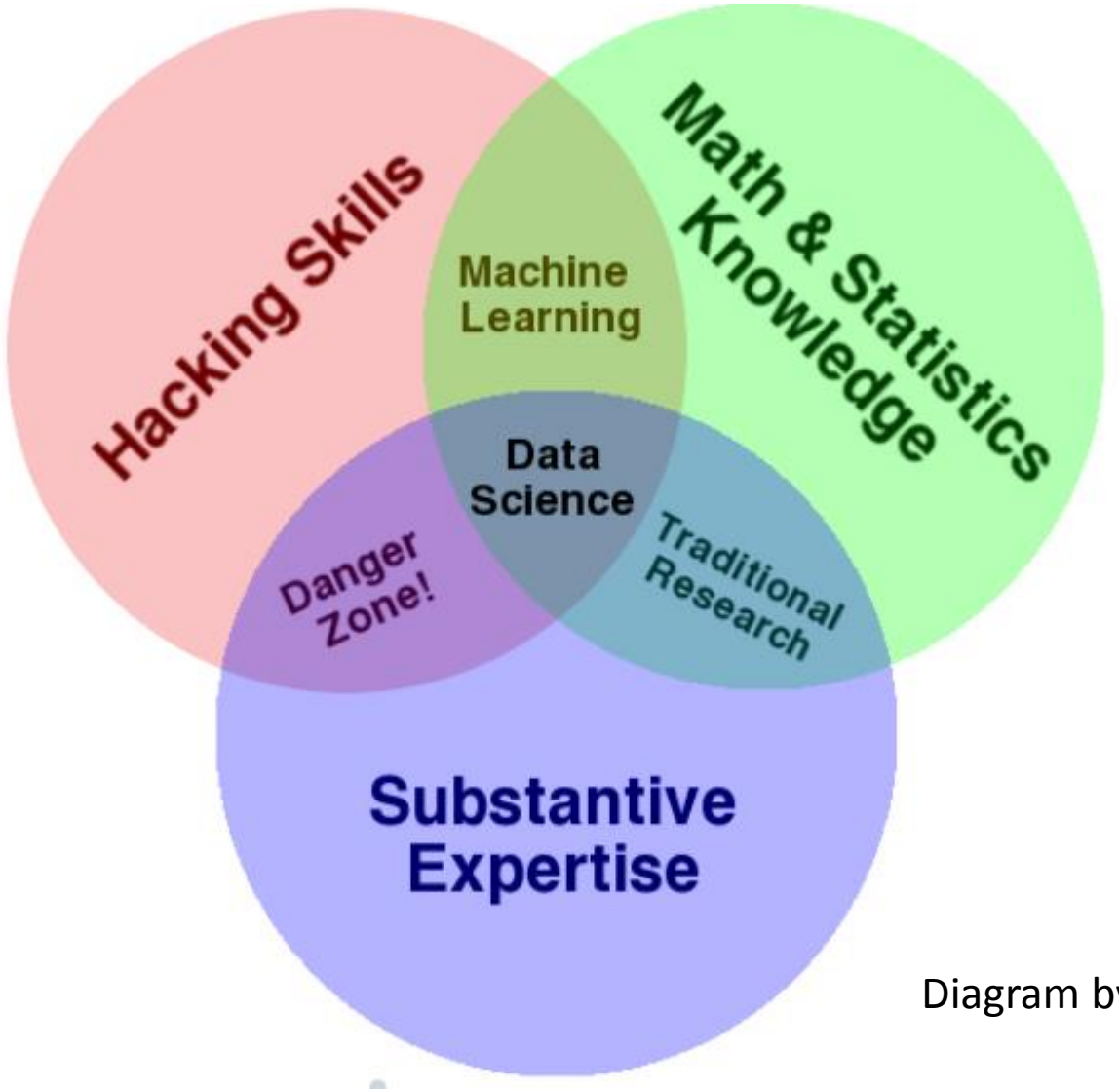
- Data science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.

  -National Institute of Standards and Technology (NIST)

- OR: The creation of data products.
  - Data Product: Any tool or view created with the help of data to make a more informed decision.
    - Descriptive
    - Predictive
    - Prescriptive

What is Data Science?





Data Scientist (n.) - Person who is better at statistics than any software engineer and better at software engineering than any statistician.

-josh_wills 2012

A Data Scientist is a statistician who lives in San Francisco

- unattributed

Diagram by Drew Conway 2013

What is Data Science?

- Big Data- ability to collect massive amounts of data.
- Data Science ≠ big data.
- Data Science doesn’t need big data.
- Data Science has been done for decades. It has now become highly popular *because* of Big Data.

Let’s Look at the Data in Data Science…

Measuring the Size of Data: Data Units

- Bit
  - 1 or 0, on or off, true or false
- Byte
  - How many bits?
  - How many unique values can it represent?
- Kilobyte 10^3 bytes
- Megabyte 10^6 bytes
- Gigabyte 10^9 bytes
- Terabyte 10^{12} bytes
- Petabyte 10^{15} bytes

Data Units

- Terabyte
  - In January 2010, the database of Wikipedia consists of a 5.87 terabyte SQL dataset.
  - The IBM computer Watson, against which Jeopardy! contestants competed in February 2011, has 16 terabytes of RAM.
- Petabyte
  - One petabyte of average MP3-encoded songs (for mobile, roughly one megabyte per minute), would require 2000 years to play.
  - As of January 2013, Facebook users had uploaded a total of 960 billion images and an estimated 357 petabytes of storage.

Data is ***not*** Information

Remember that Data Science produces knowledge/information from data.

Here's some data:

38.5
37.2
29.7
30.6
20.6
26.7
43.9
32.6
30.5
24.9
17.7
20.2

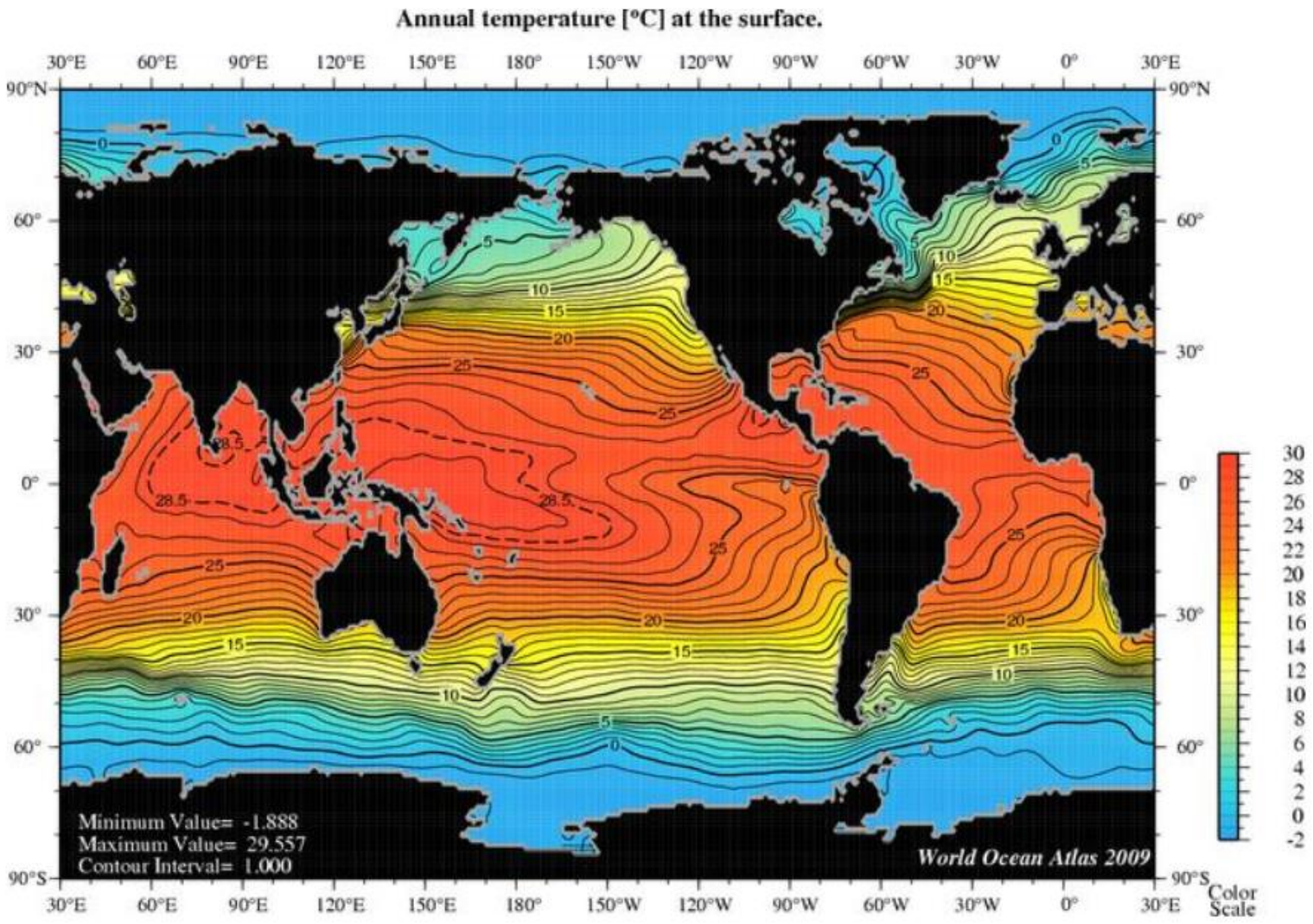
These data are real-valued numbers, a type of ***numeric*** data. They are, apparently, recordings of some kind of measurement. Is this information? Without some further clues we do not know what these data tell us, and are therefore not informative.

Data vs. Information

38.5
37.2
29.7
30.6
20.6
26.7
43.9
32.6
30.5
24.9
17.7
20.2

This is data

This is information



Data vs. Information

In most dictionaries and in popular usage, the term ***data*** is defined as "information".

Although data and information are closely related, they are not at all the same. We define data as a recorded observation of a quantifiable value.

Information is made by analyzing data. Information provides us with the means to make decisions. We cannot make decisions on pure data alone.

Semi-Structured Data- a hierarchical model

To bring meaning to data,
we add ***structure*** to the data.
This structure informs us
about the context of the data.

This is called ***meta-data***.

An XML example:
- HTML is another.

These are examples of semi-structured data. The more structure the better control we have over the data ***integrity***.

```xml
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author>Gambardella, Matthew</author>
      <title>XML Developer's Guide</title>
      <genre>Computer</genre>
      <price>44.95</price>
      <publish_date>2000-10-01</publish_date>
      <description>An in-depth look at creating applications
      with XML.</description>
   </book>
   <book id="bk102">
      <author>Ralls, Kim</author>
      <title>Midnight Rain</title>
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-12-16</publish_date>
      <description>A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen
      of the world.</description>
   </book>
   <book id="bk103">
      <author>Corets, Eva</author>
```

Structured Data-tabular format

A table contains rows and columns.

Each row is a ***tuple*** of related observations.

Each column is a collection of observations about a single aspect of the world we are studying.

Columns have ***data types***, such as numeric or categorical.

Most of our analysis
Will deal with data in this form

State	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
AL	3615	3624	2.1	69.05	15.1	41.3	20	50708
AK	365	6315	1.5	69.31	11.3	66.7	152	566432
AZ	2212	4530	1.8	70.55	7.8	58.1	15	113417
AR	2110	3378	1.9	70.66	10.1	39.9	65	51945
CA	21198	5114	1.1	71.71	10.3	62.6	20	156361
CO	2541	4884	0.7	72.06	6.8	63.9	166	103766
CT	3100	5348	1.1	72.48	3.1	56	139	4862
DE	579	4809	0.9	70.06	6.2	54.6	103	1982

Structured Data-tabular format

Numeric

Categorical

Age	Sex	ChestPain	RestBP	Ca	Thal	AHD
63	1	typical	145	0	fixed	No
67	1	asymptom	160	3	normal	Yes
67	1	asymptom	120	2	reversable	Yes
37	1	nonangina	130	0	normal	No
41	0	nontypical	130	0	normal	No
56	1	nontypical	120	0	normal	No
62	0	asymptom	140	2	normal	Yes
57	0	asymptom	120	0	normal	No
63	1	asymptom	130	1	reversable	Yes
53	1	asymptom	140	0	reversable	Yes
57	1	asymptom	140	0	fixed	No
56	0	nontypical	140	0	normal	No

Why is the “Sex” column Categorical?

The numbers take on a few discrete values, namely 0 and 1.
We say that Sex has two *levels.*
We could assign *labels* to these levels, namely “female” and “male”.

The Ca column has levels 0, 1, 2, 3.
It is also categorical.

Data Analysis and Modeling

- One variable is the dependent variable, or the *predicted variable*.
- The independent variables are also called the *predictors*.
- Example: predict cognitive scores of children based on mother's characteristics.

Predictors

Predicted Variable

kid_score	mom_hs	mom_iq	mom_work	mom_age
65	1	121.12	4	27
98	0	107.90	1	18
98	1	89.36	4	25
85	1	115.44	4	27

Machine Learning Analysis

- ML: given data, *learn* the parameters that create a model of the "trends" in the data.
- Model can be used for predicting a value given new data, or for understanding how the variables in the data interact.
- Supervised learning: examples of data and outcomes provided to the algorithm.
- Unsupervised learning: no outcomes provided.

A brief look at some of the
analysis techniques we'll be working with…

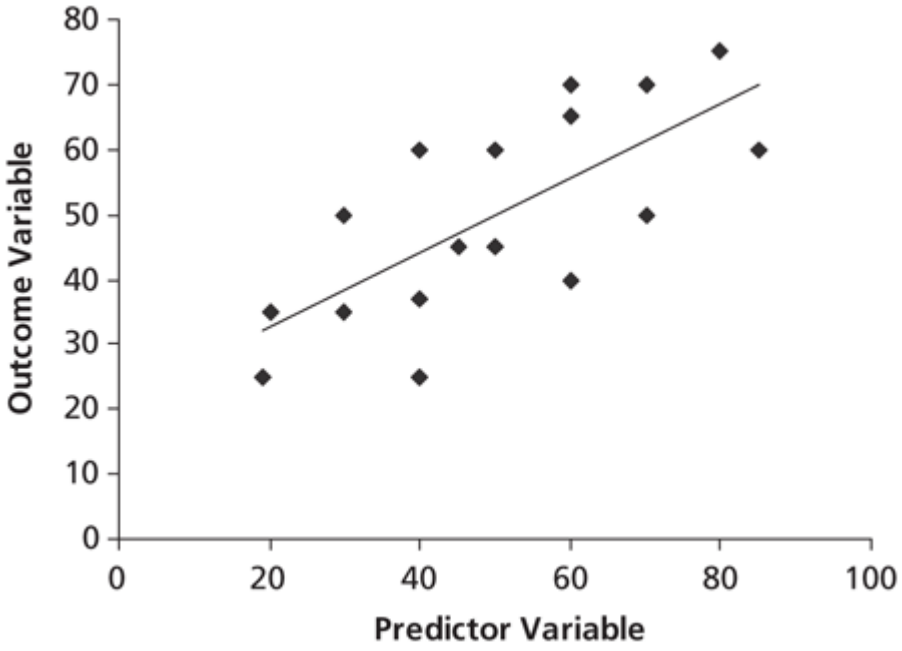
Linear Regression

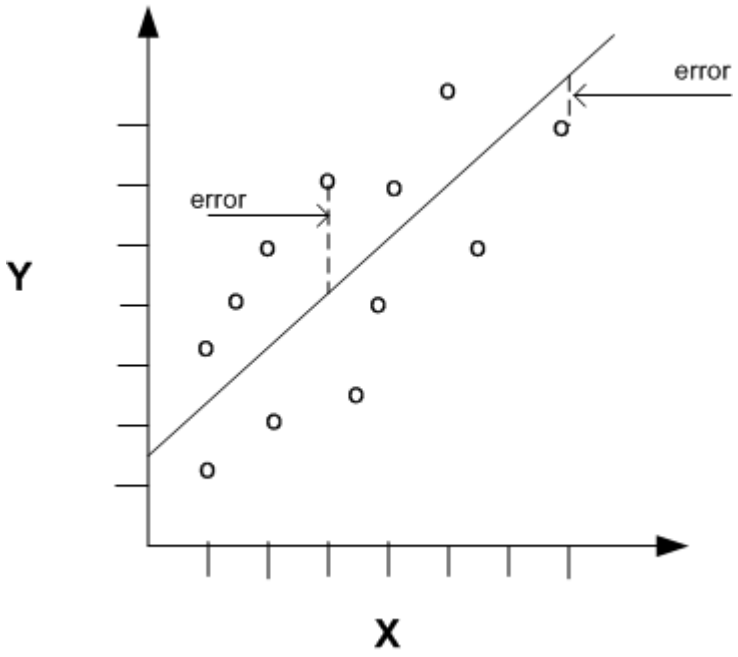
A model generalizes a theory from the observed data.

Assume a linear model can describe the data:

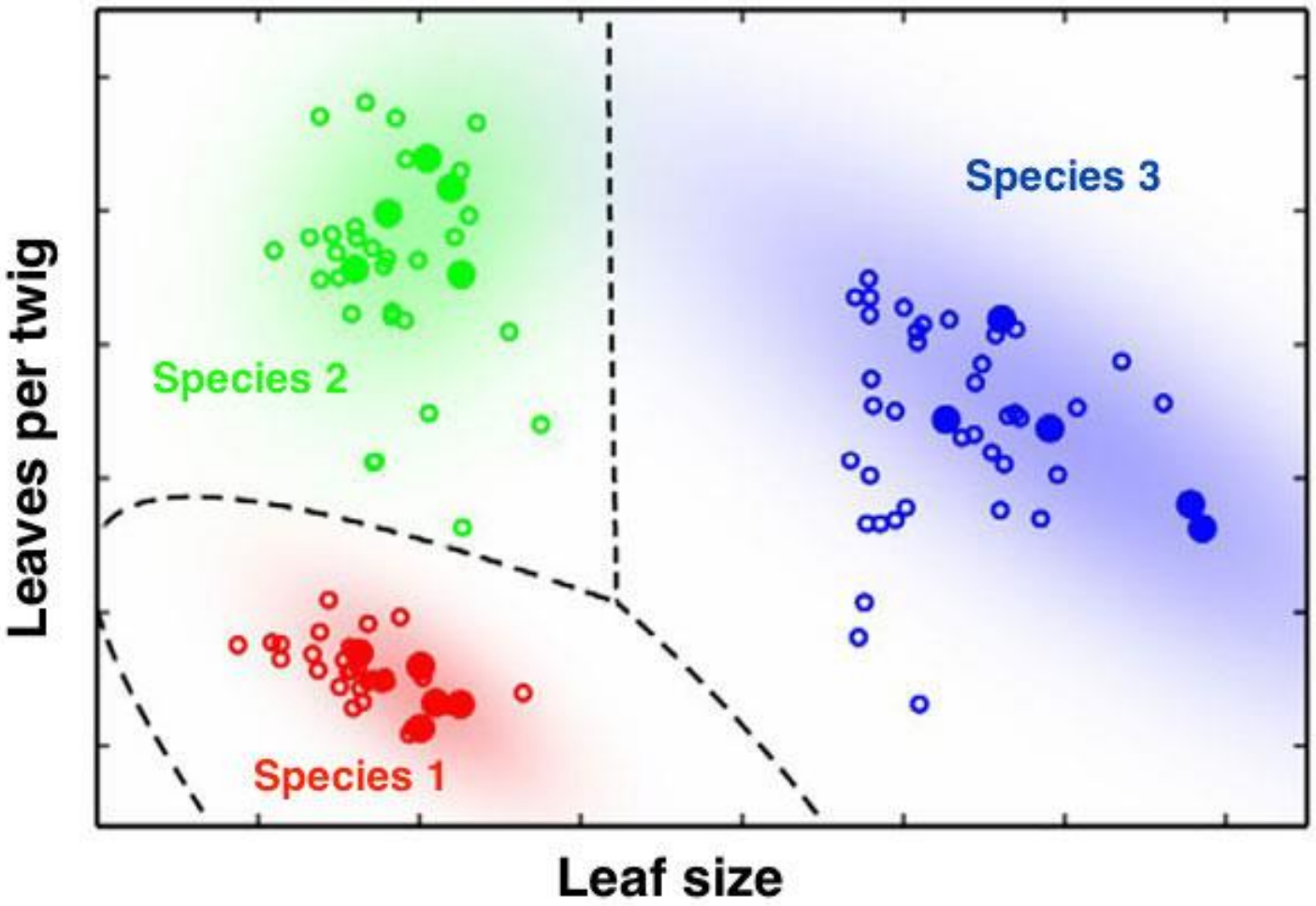
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Can use the model for prediction or description on trends in data



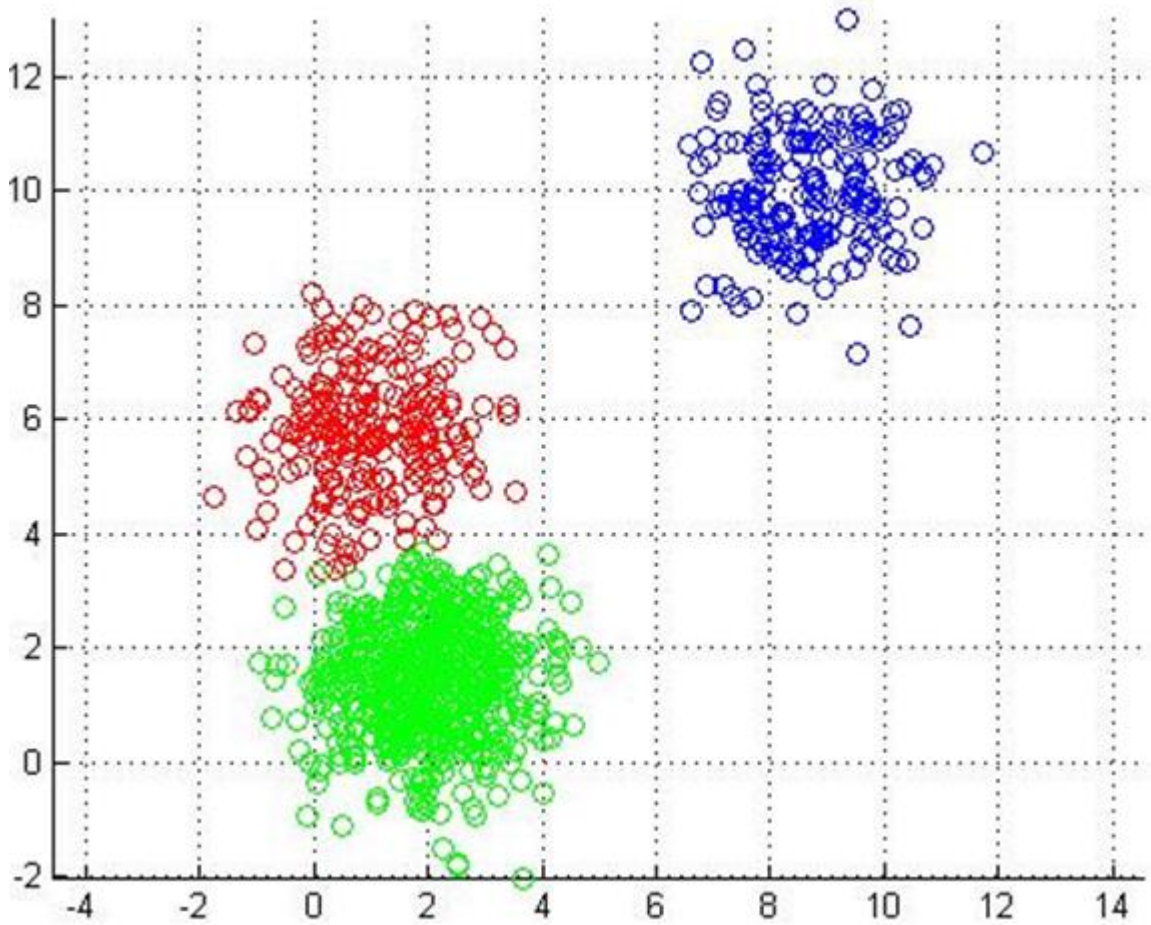


Classification



Predict the species given leaf size and leaves per twig.

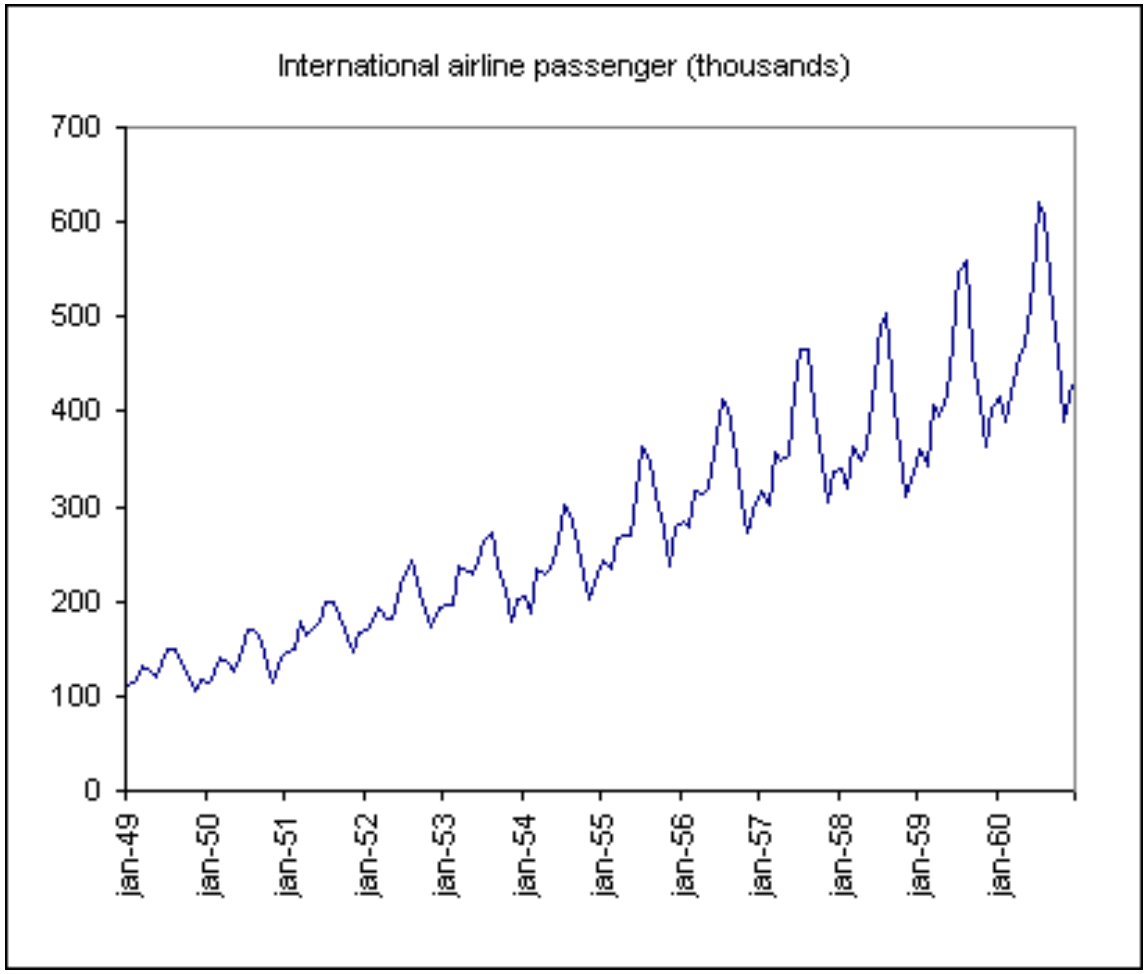
Clustering Data



Are there groups in the data?

If so, what can we say about members of a group?

Time Series

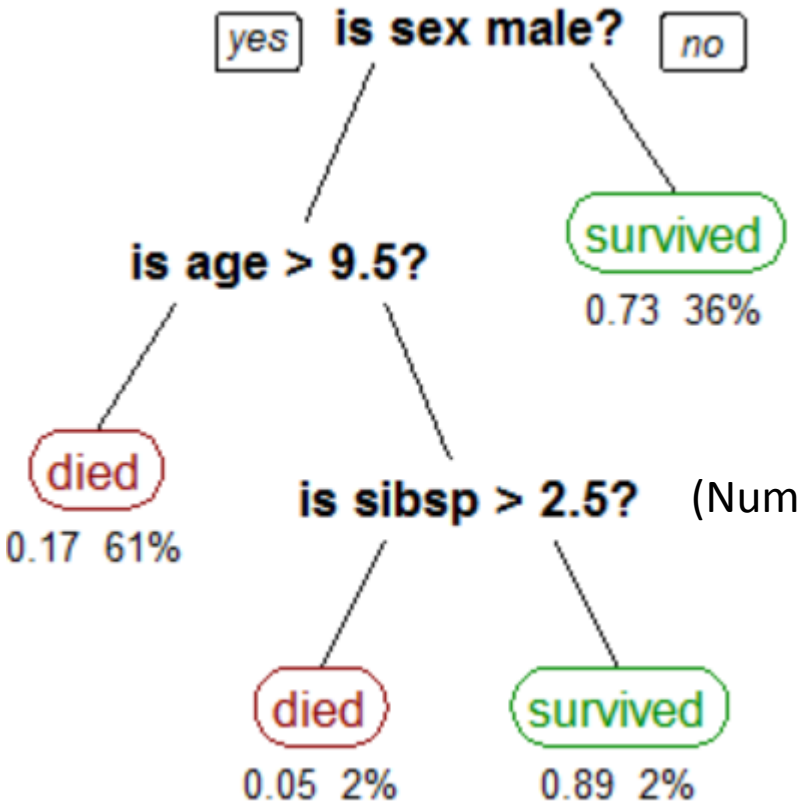


What are the cyclical and trend components?

Predict future trends..

Decision Trees

Predict survival of passengers on the Titanic



(Number of spouses/siblings)

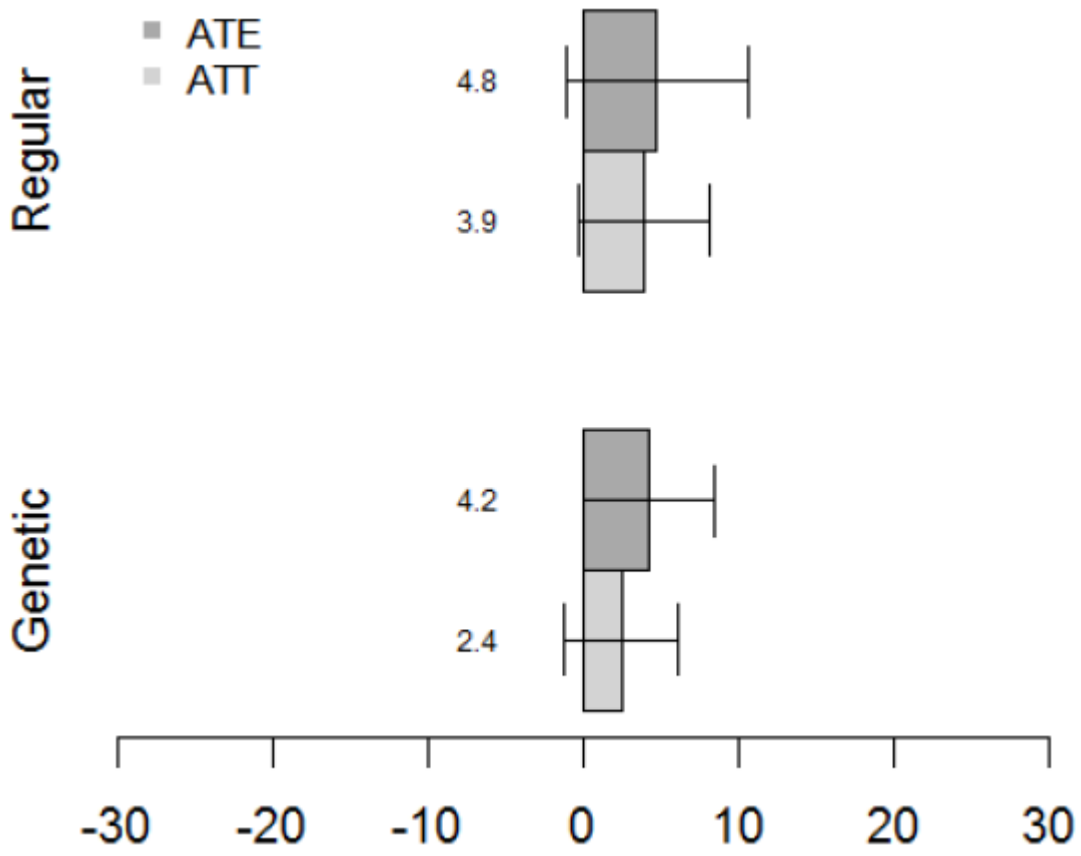
Trees are a non-linear model.
Often more robust to noise.

Each node is picked based on how well
If partitions the data.

Max info gain: a 50-50 split.

Leaves report probability of survival/ percent of data

Visualizing Data…

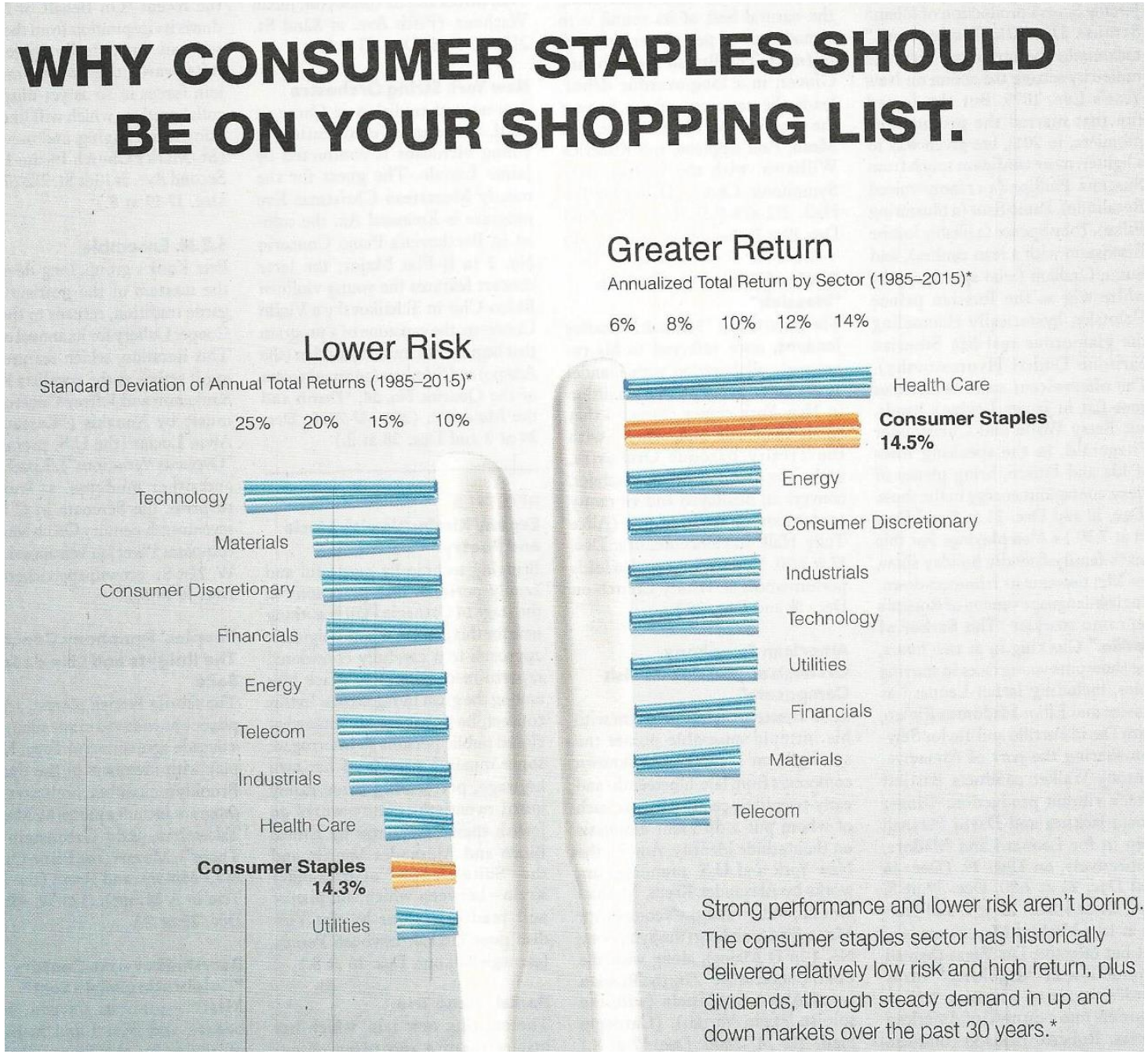


Educational data:
The effect of reading a textbook
before attending lecture on exam scores.

Two matching algorithms used.
95% confidence intervals shown.

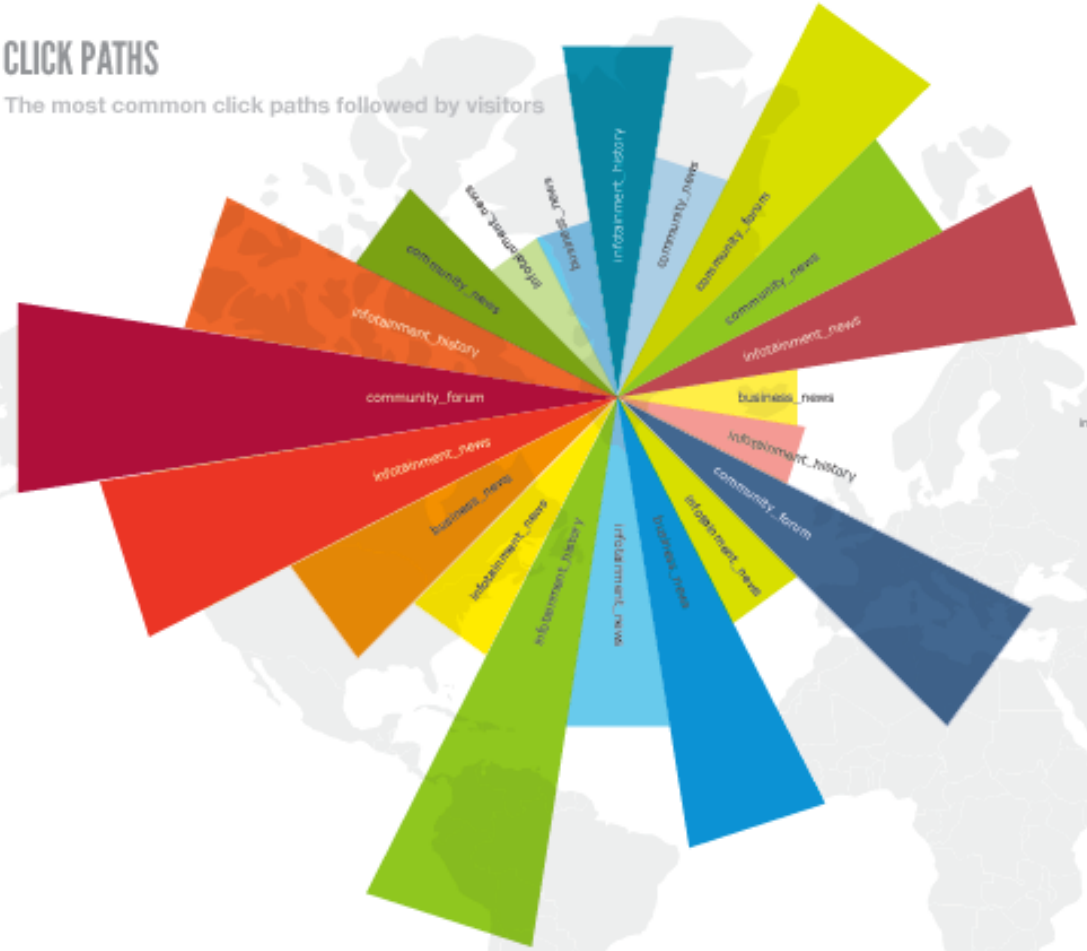
Saw this in a magazine
While waiting for the dentist.
I had to have it!

(Ad for Fidelity Investments).



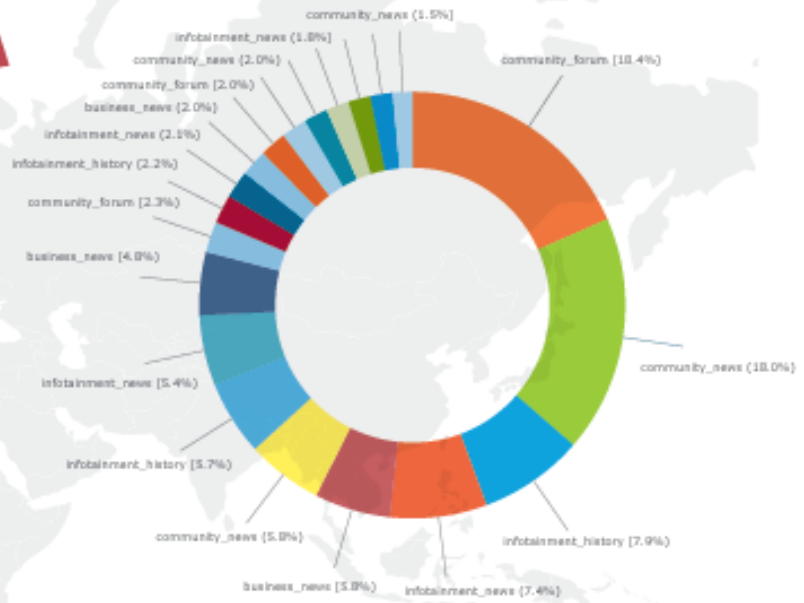
CLICK PATHS

The most common click paths followed by visitors



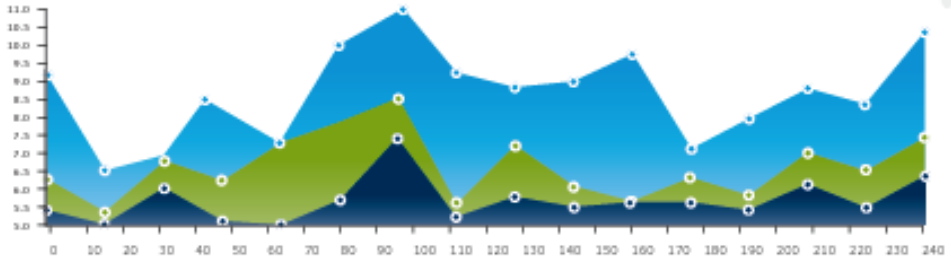
PAGEVIEWS

The most popular pages, represented in the tag cloud by page views this month...



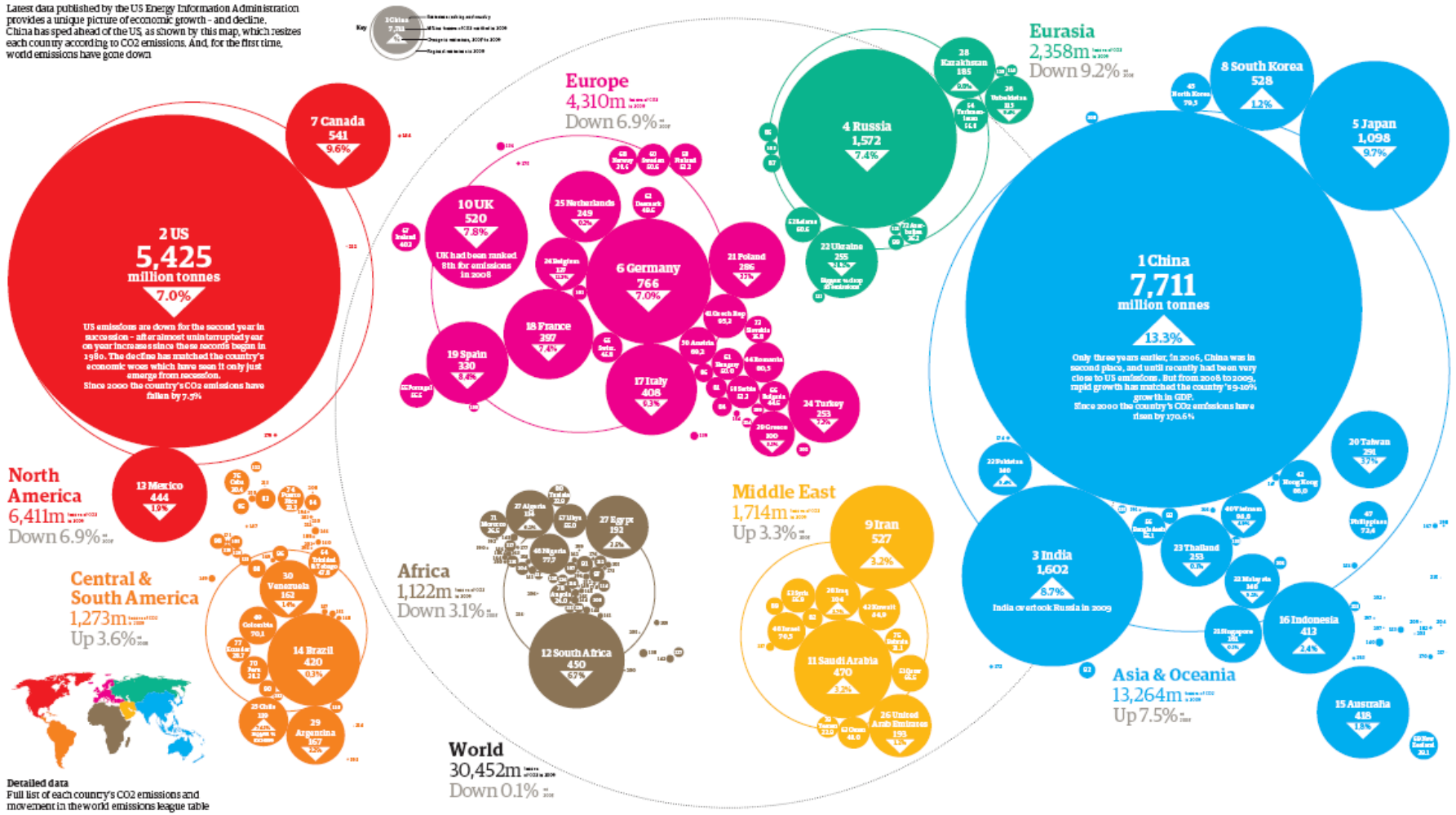
NEW VS. RETURNING VISITORS

Our stickiest web pages, measured by dwell time

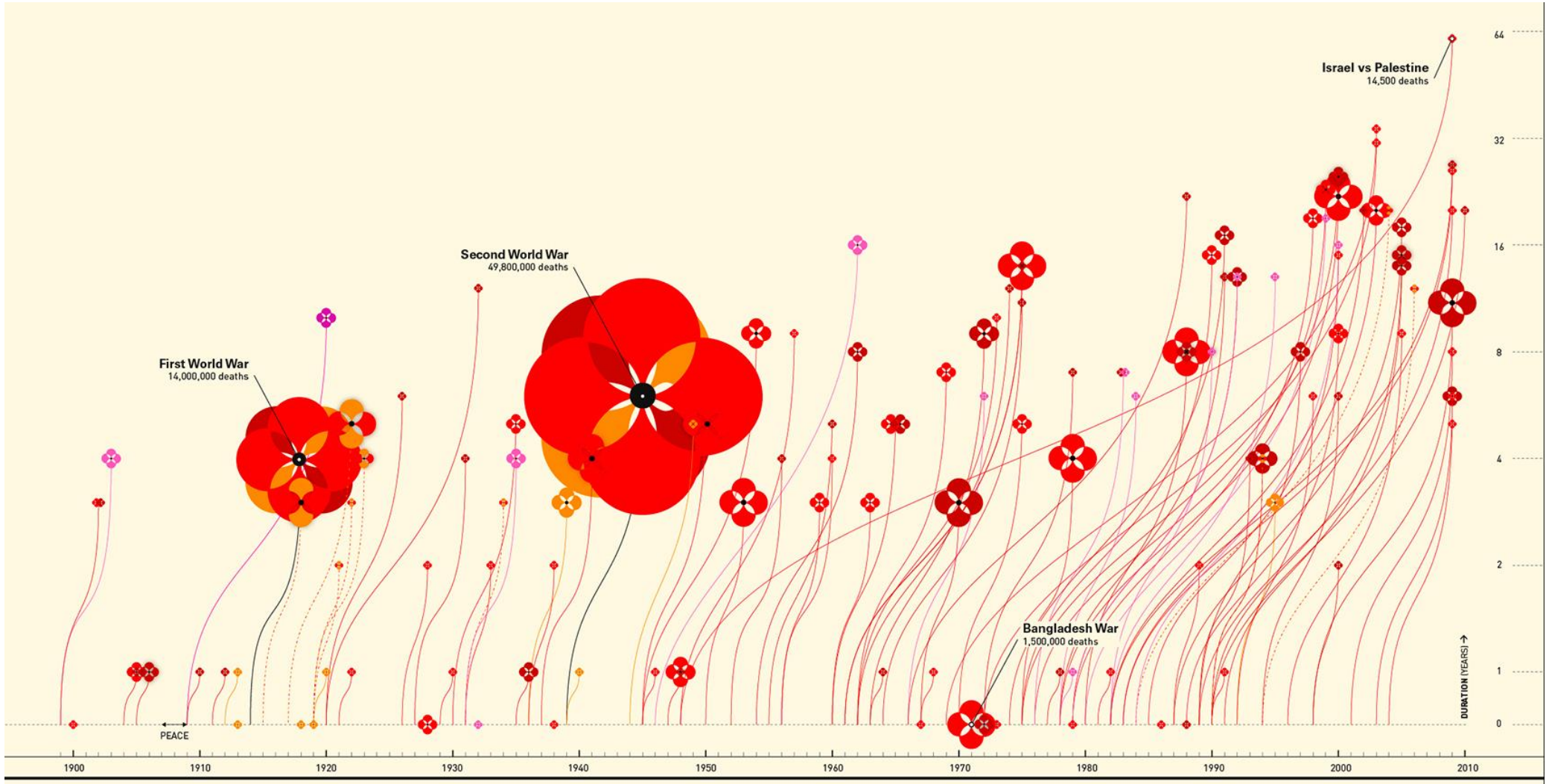


An atlas of pollution: the world in carbon dioxide emissions

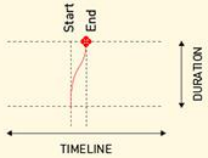
Latest data published by the US Energy Information Administration provides a unique picture of economic growth – and decline. China has sped ahead of the US, as shown by this map, which resizes each country according to CO2 emissions. And, for the first time, world emissions have gone down



Detailed data
Full list of each country's CO2 emissions and movement in the world emissions league table



POPPY DIAGRAM



The remembrance poppy commemorates soldiers who have died in war. Each poppy in the diagram depicts a war of the last century (with more than 10,000 deaths). The stem grows from the year when the war started. The poppy flowers in the year the war ended. Its size shows the number of deaths.

NUMBER of DEATHS IN THOUSANDS (POPPY'S SIZE)

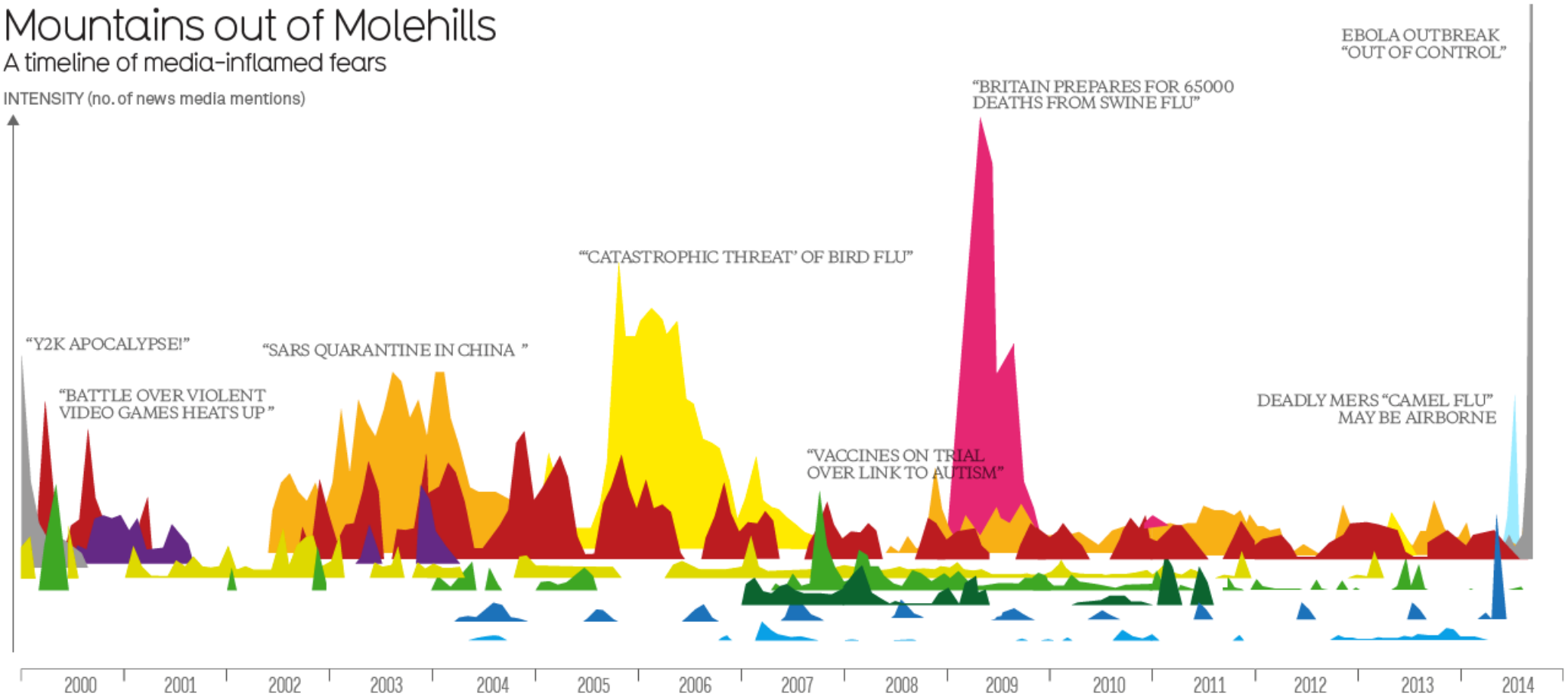
0–99 | 100–499 | 500–999 | 1,000–3,000

REGIONS INVOLVED IN WARS (POPPY'S COLOUR)

Africa | Asia | Asia/Europe | Europe | N. America | S. America | Global

Mountains out of Molehills

A timeline of media-inflamed fears



rollover to scale by deaths

- Killer Wifi
- Killer Wasps
- Mobile Phones & Cancer
- Autism & Vaccinations
- Asteroid Collision
- The Millennium Bug
- Mad Cow Disease
- Violent Video Games
- SARS
- Bird Flu
- Swine Flu
- MERS
- Ebola

source: Google Trends, Google News Timeline

design & concept: David McCandless, Aug 2014

informationisbeautiful.net

Summary:

Topics:

1. What is Data Science?
2. Aspects of Data
3. Storing data
4. Analysis/modeling
5. Visualizing data