

# Introduction to Data Science

## What is Data Science?

Gordon Anderson

# What is Data Science?

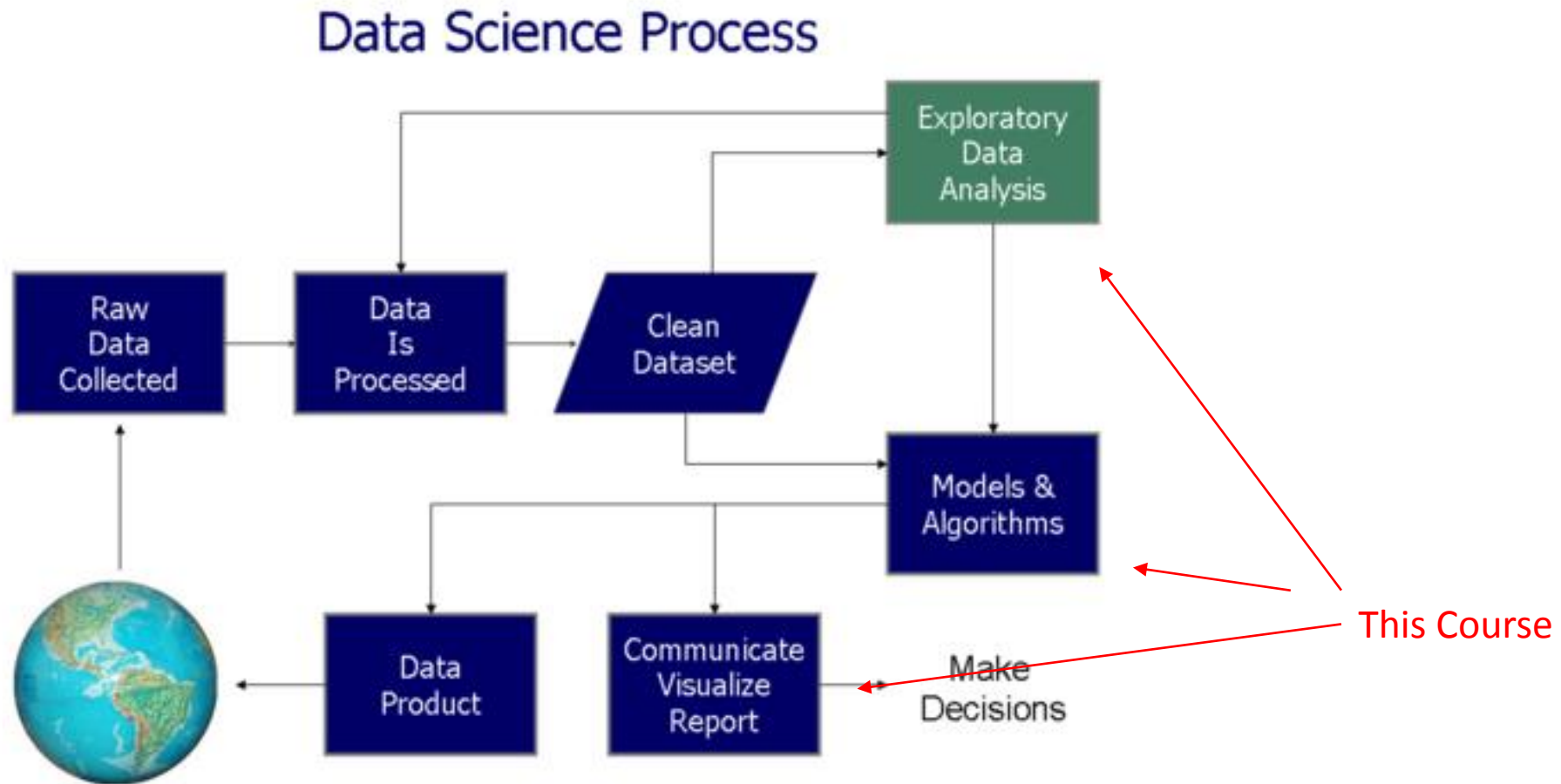
- "The key word in "Data Science" is not Data, it is Science".
- Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining and predictive analytics, as well as Knowledge Discovery in Databases (KDD).

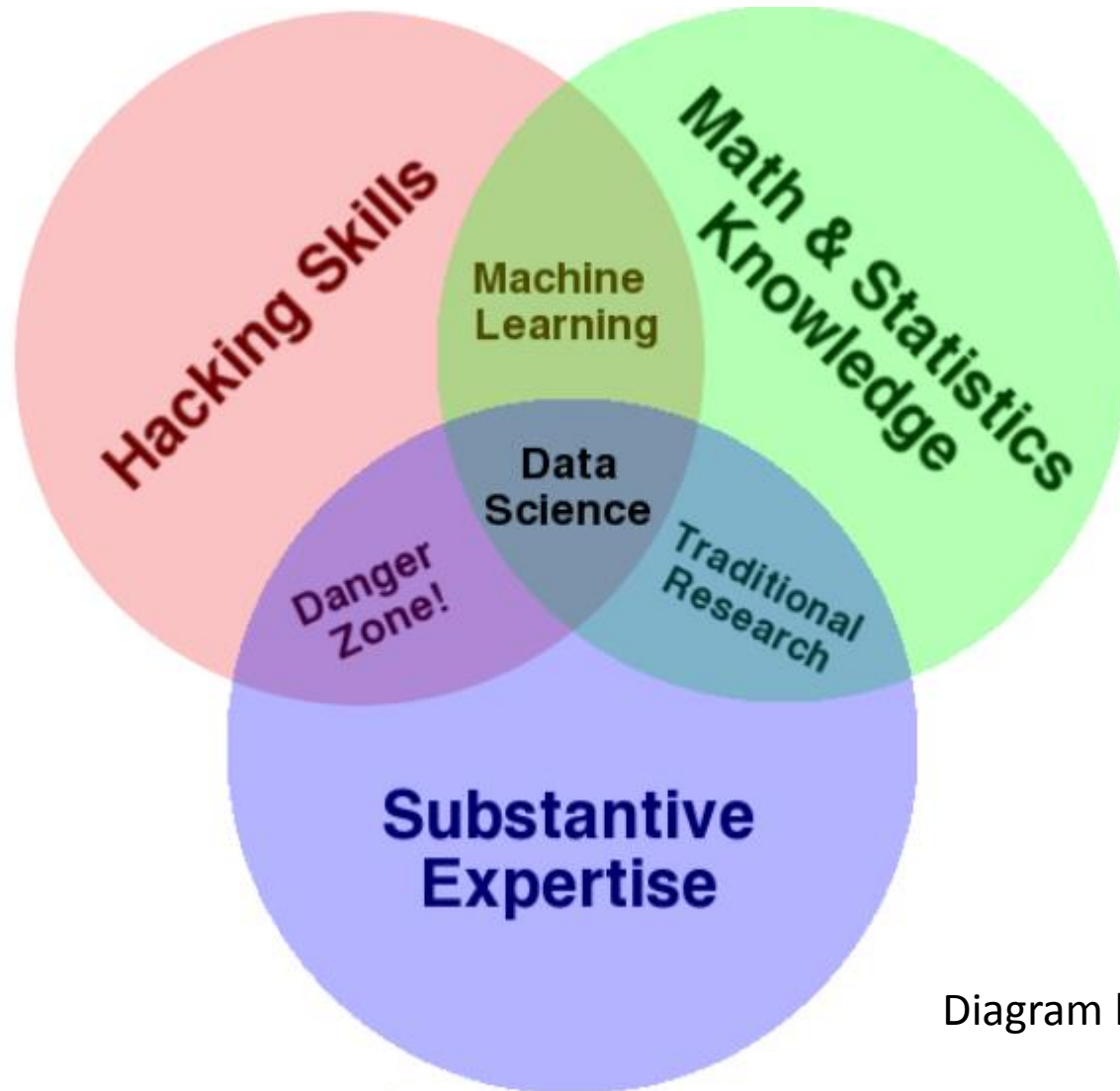
-Dhar and Leek 2013

# 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  $\neq$  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

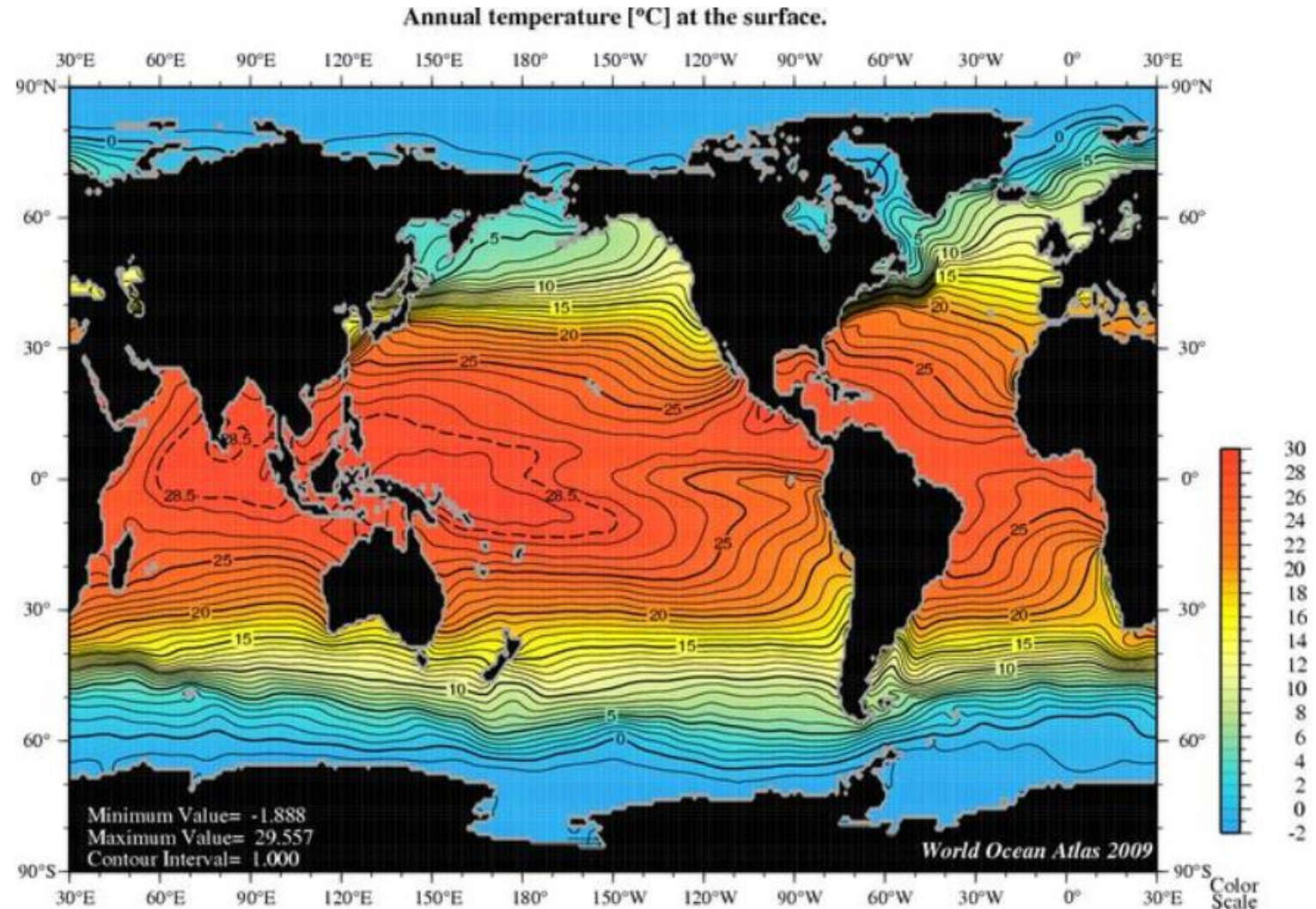
These data are real-valued numbers, a type of *numeric* data. They are, apparently, recordings of some kind of measurement. Is this information? Without some further clues we do not know what these data tell us, and are therefore not informative.

# 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 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
FL	8277	4845	1.3	70.66	10.7	53.6	44	54000

# 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*.

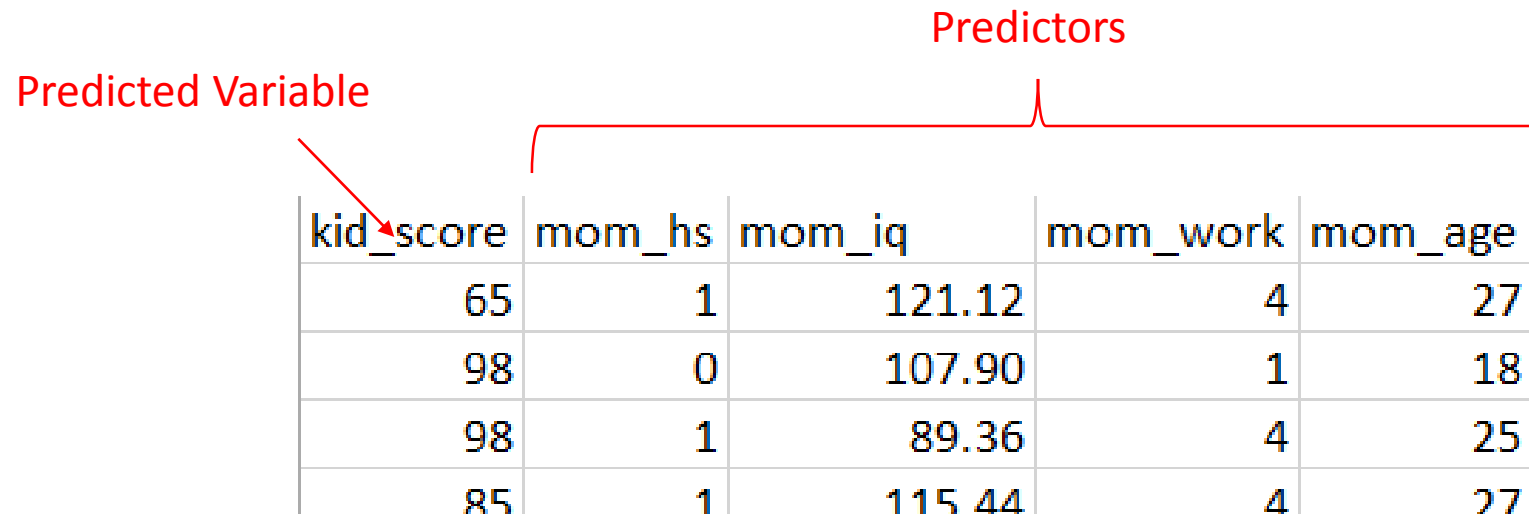
We could assign *labels* to these levels, namely “female” and “male”.

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.



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



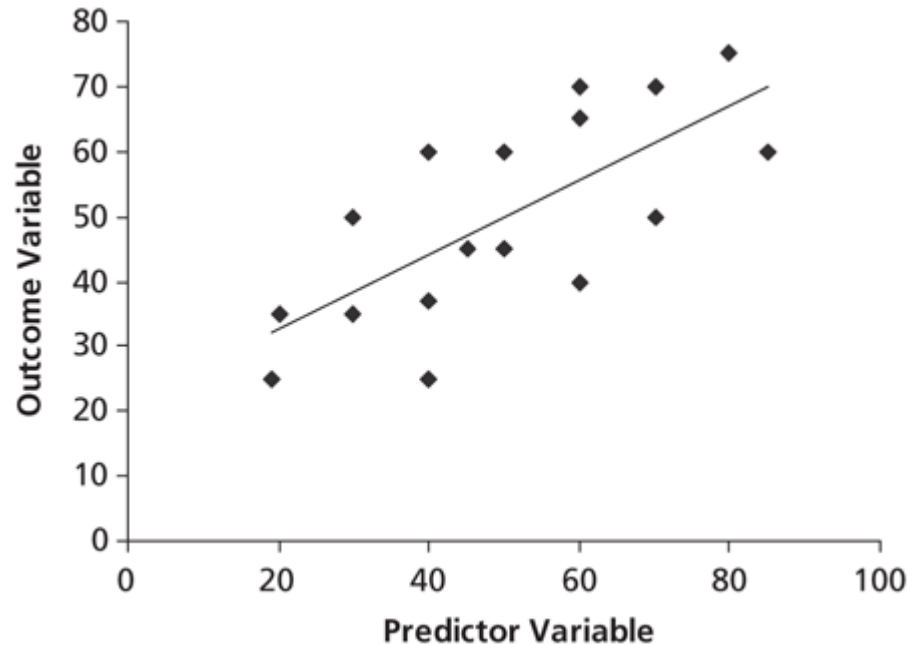
# Machine Learning Analysis

- 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

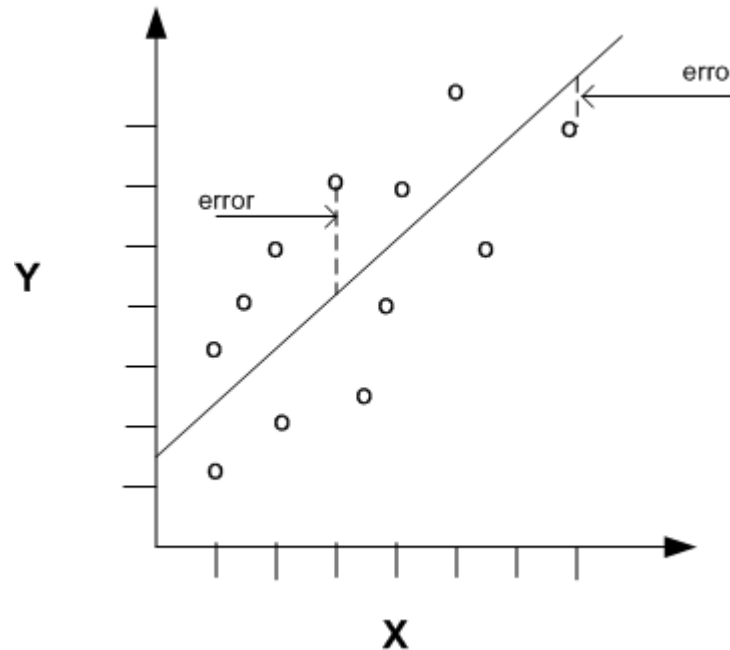
A model generalizes a theory from the observed data.



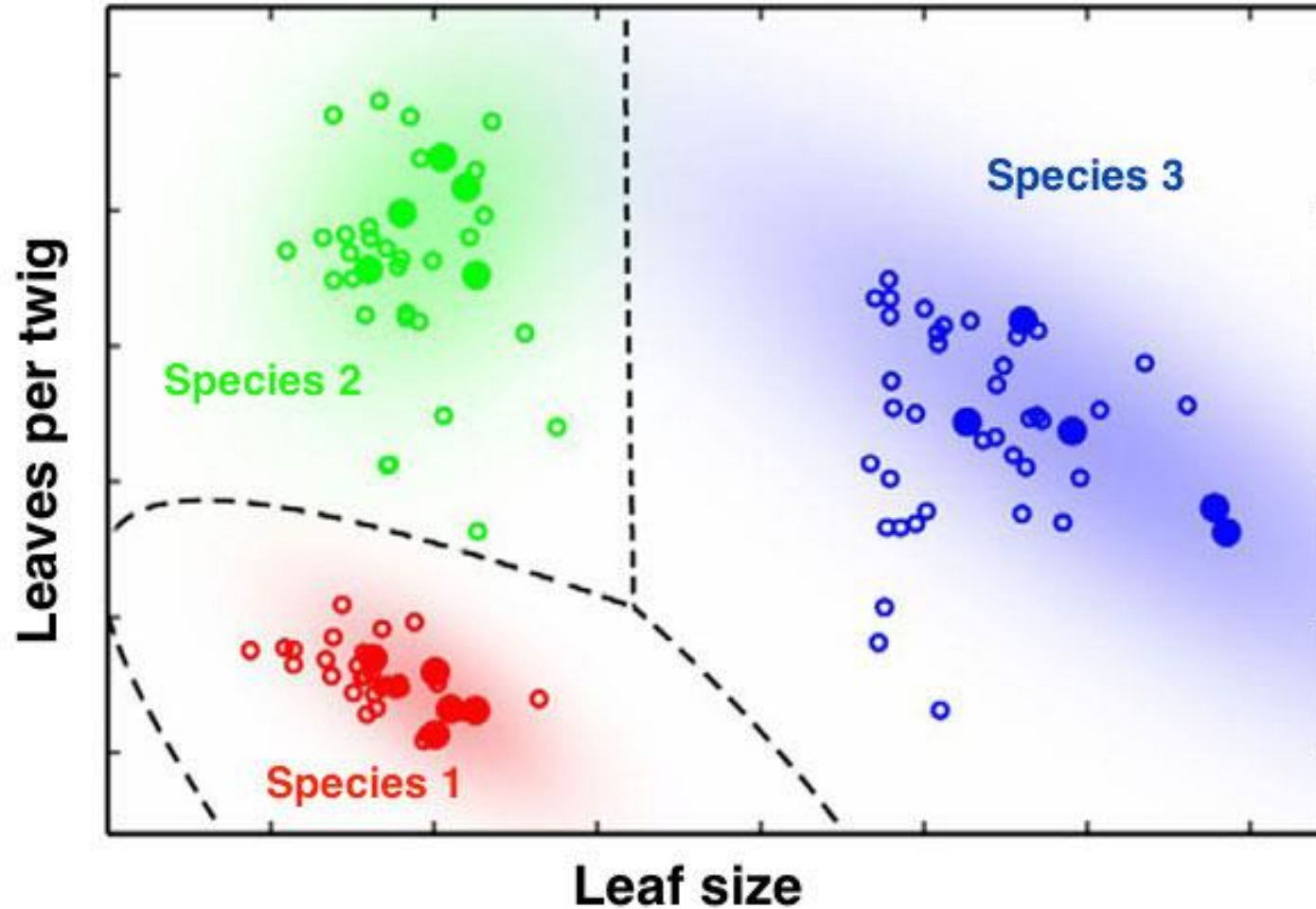
Assume a linear model can describe the data:

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

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

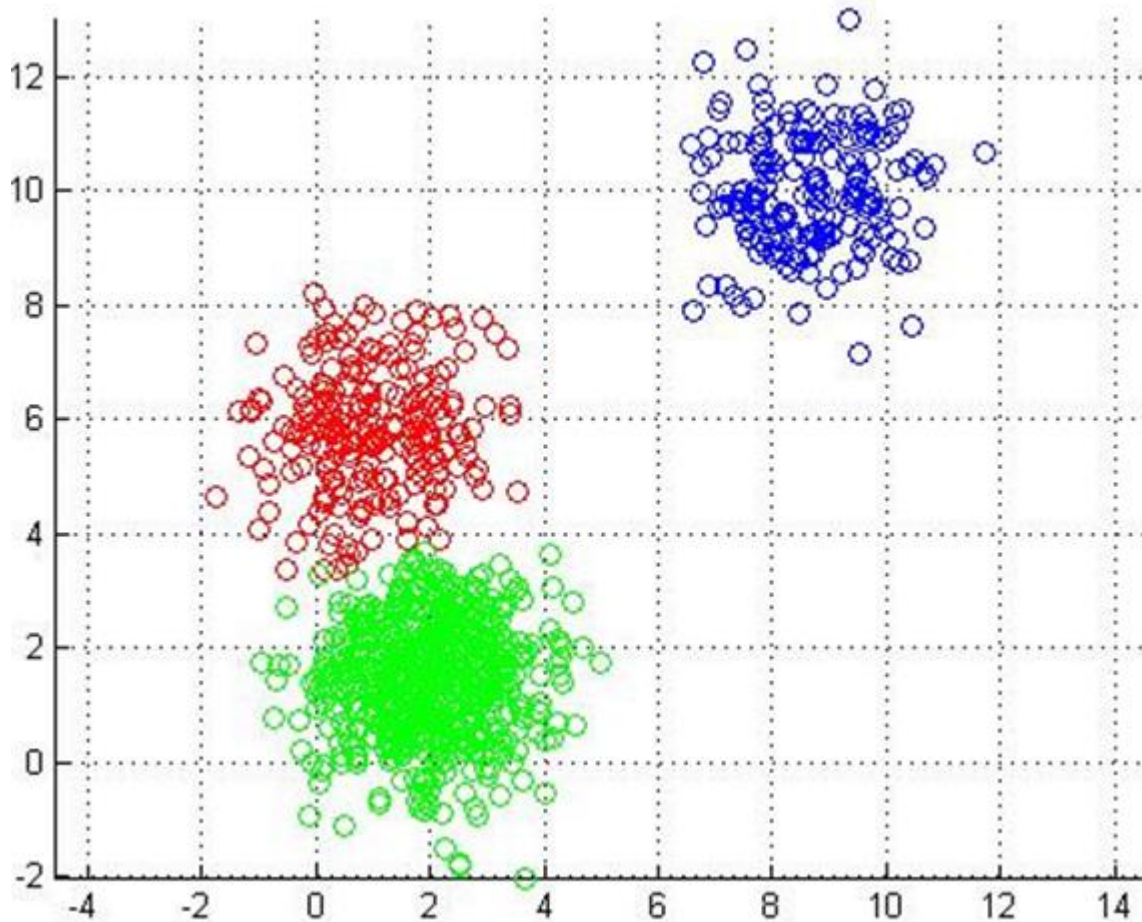


# Classification



Predict the species given  
leaf size and leaves per twig.

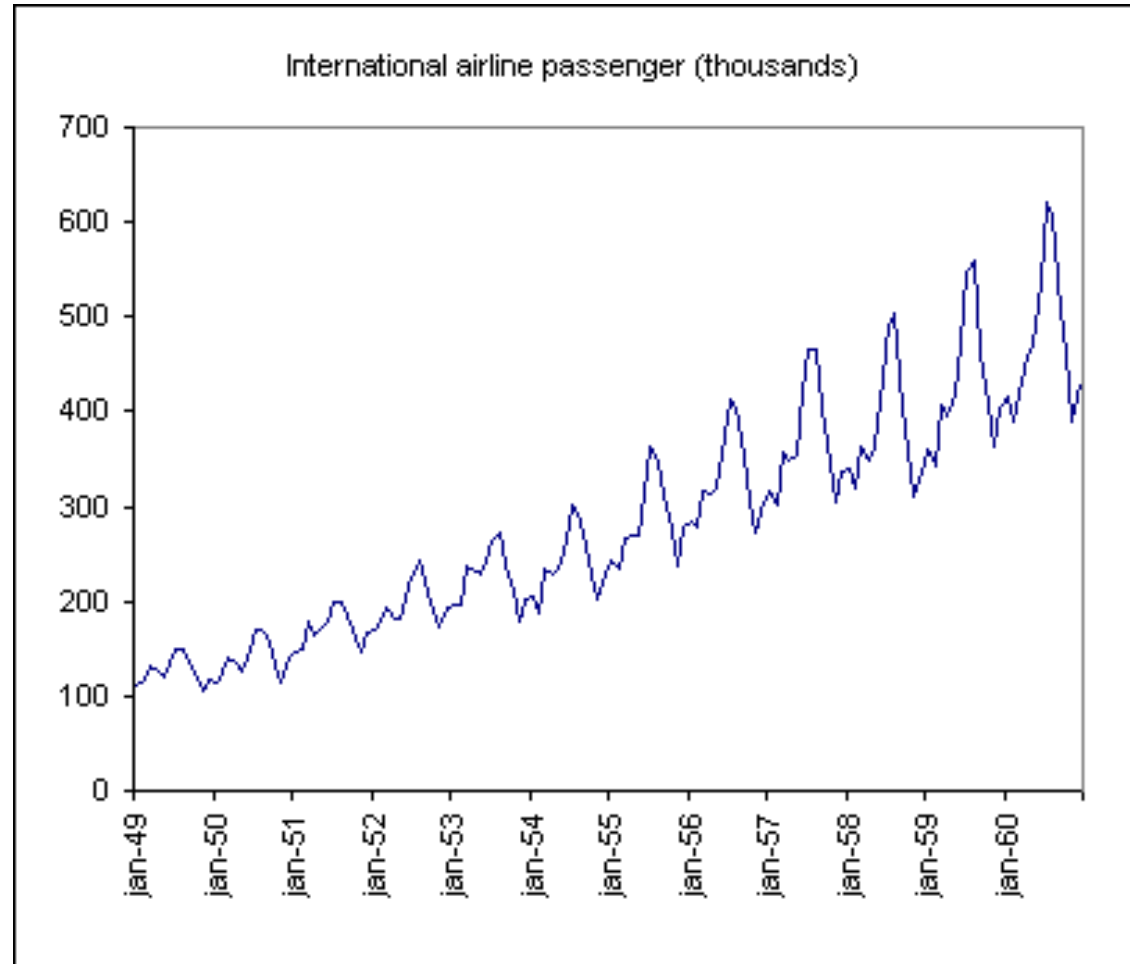
# Clustering Data



Are there groups in the data?

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

# Time Series

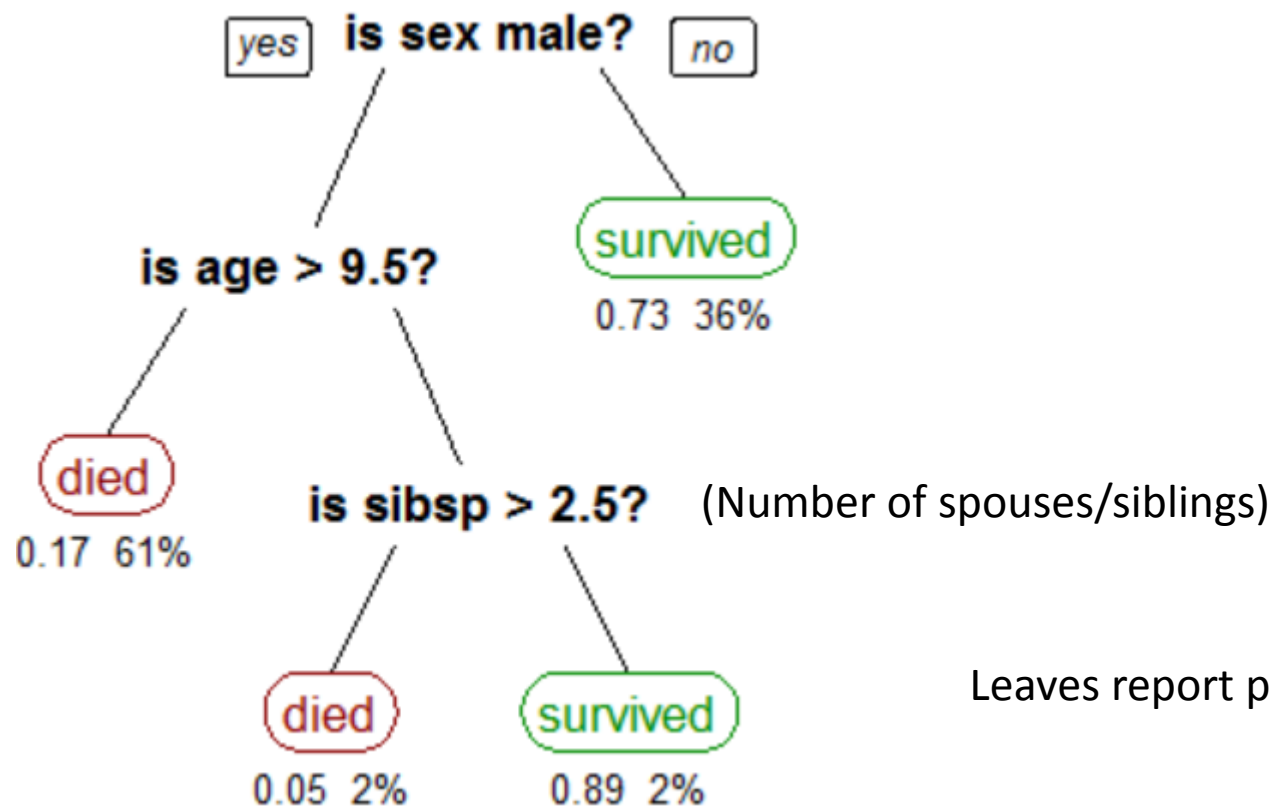


What are the cyclical and trend components?

Predict future trends..

# Decision Trees

Predict survival of passengers on the Titanic



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

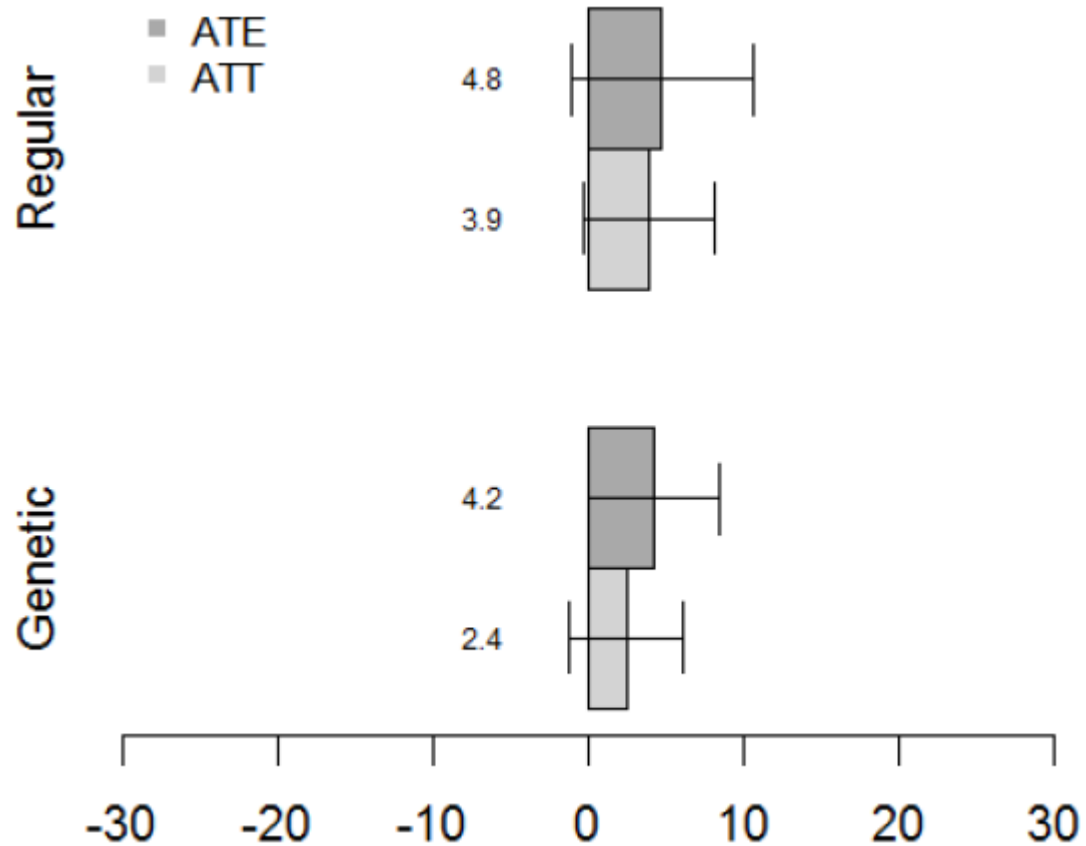
Each node is picked based on how well  
It partitions the data.

Max info gain: a 50-50 split.

Leaves report probability of survival/ percent of data

Visualizing Data...





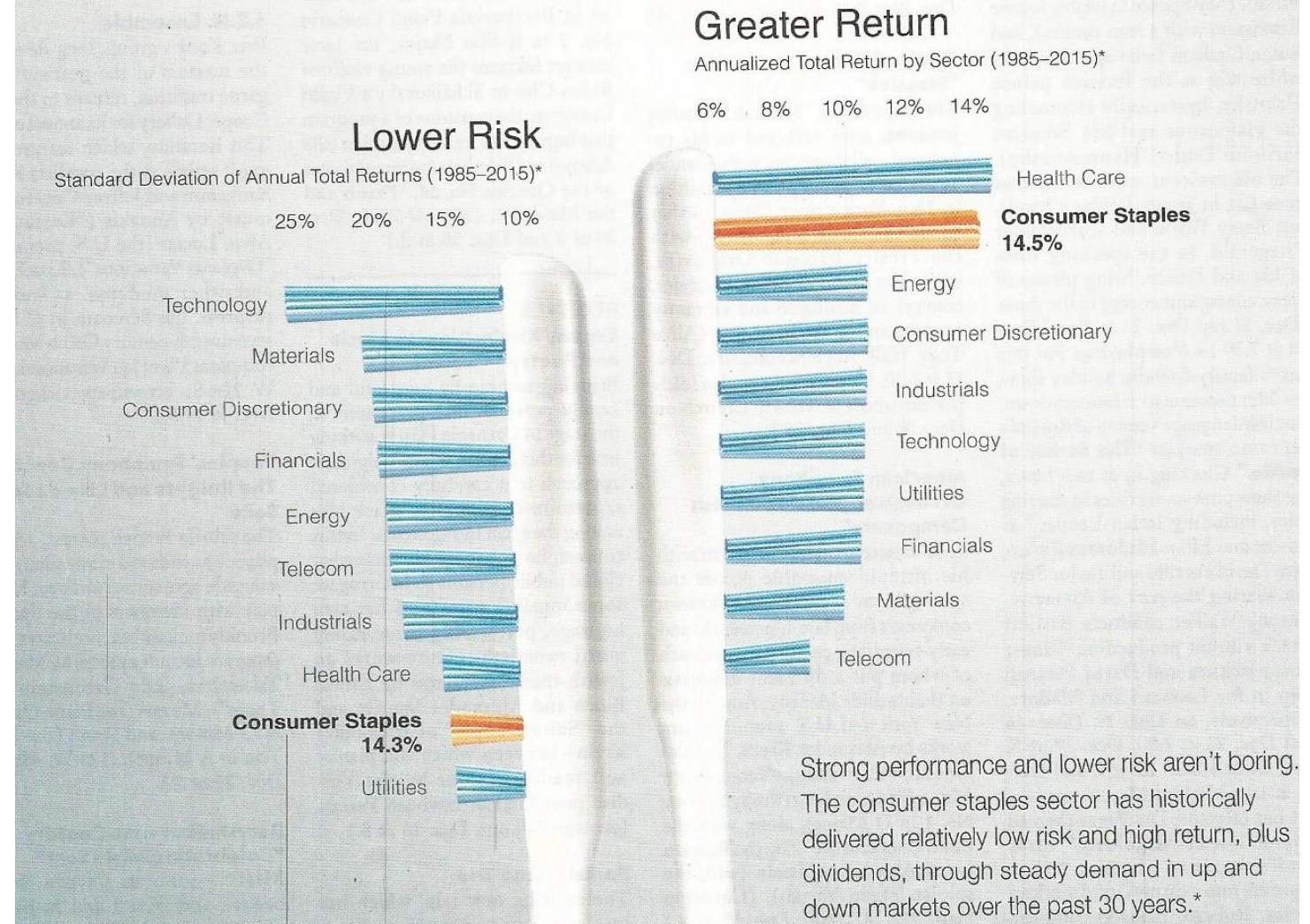
Educational data:  
The effect of reading a textbook  
before attending lecture on exam scores.

Two matching algorithms used.  
95% confidence intervals shown.

# WHY CONSUMER STAPLES SHOULD BE ON YOUR SHOPPING LIST.

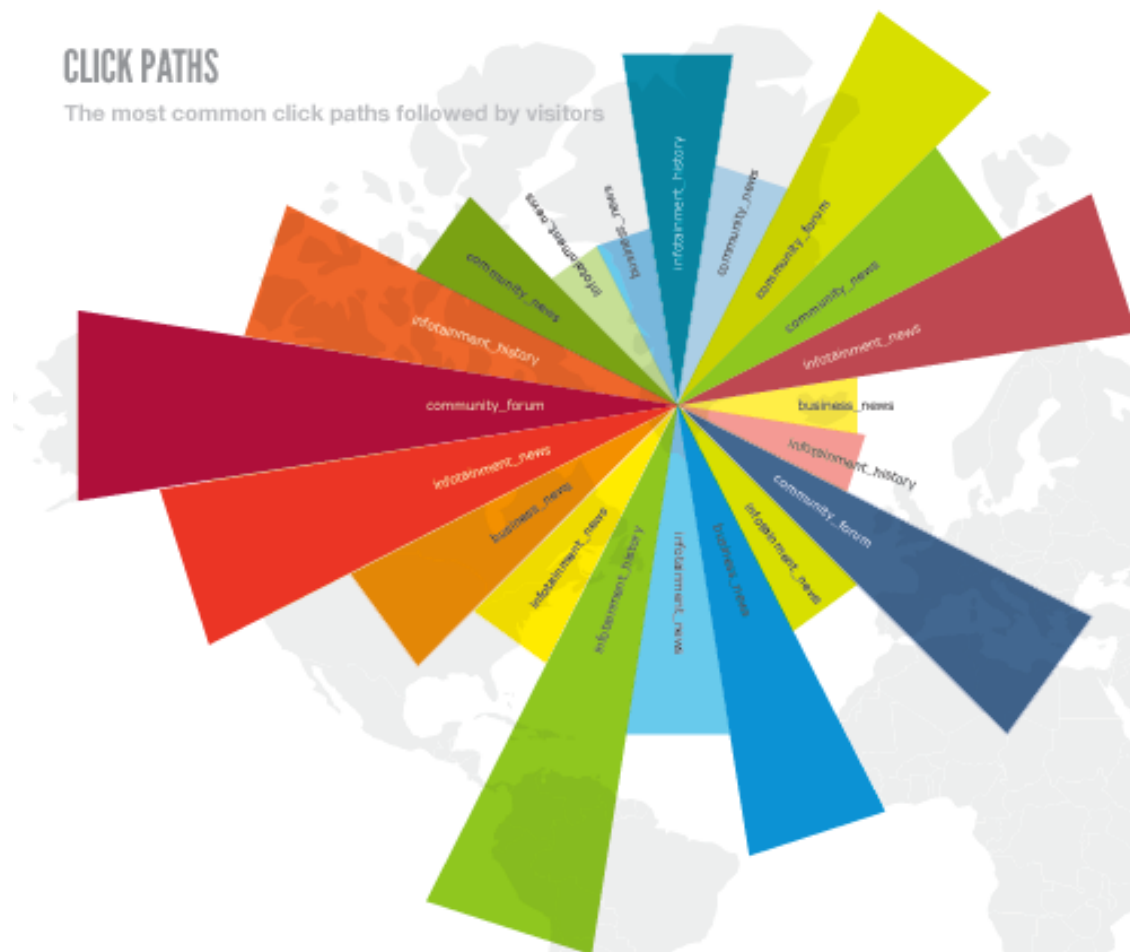
Saw this in a magazine  
While waiting for the dentist.  
I had to have it!

(Ad for Fidelity Investments).



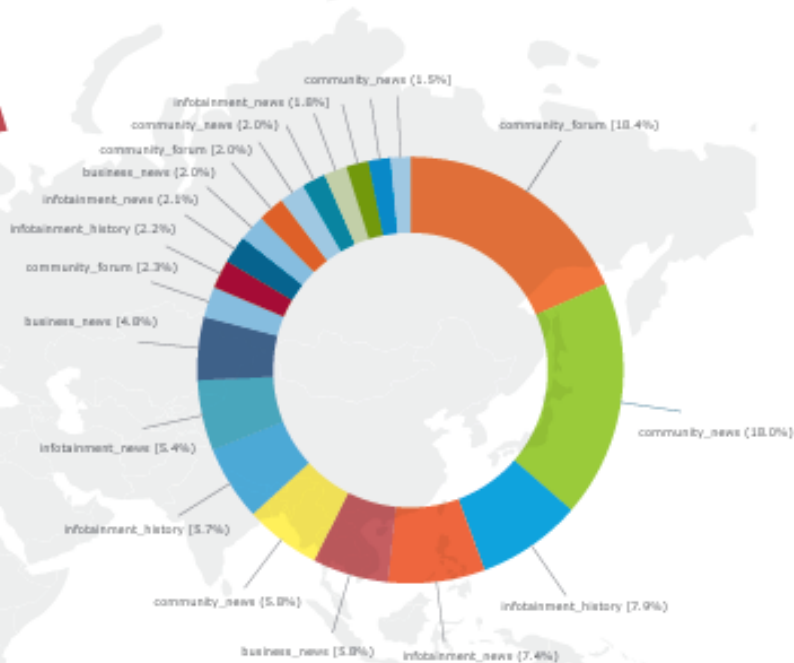
## CLICK PATHS

The most common click paths followed by visitors



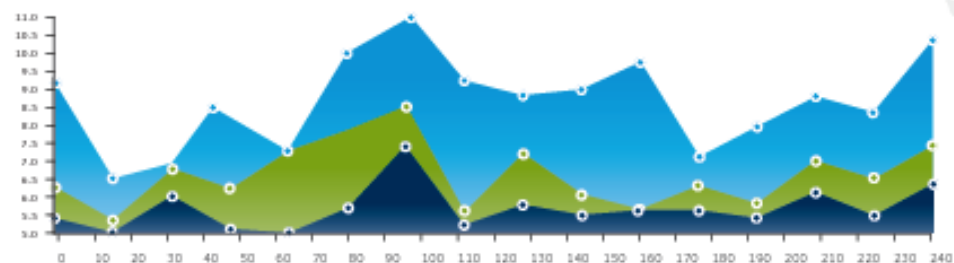
## PAGEVIEWS

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

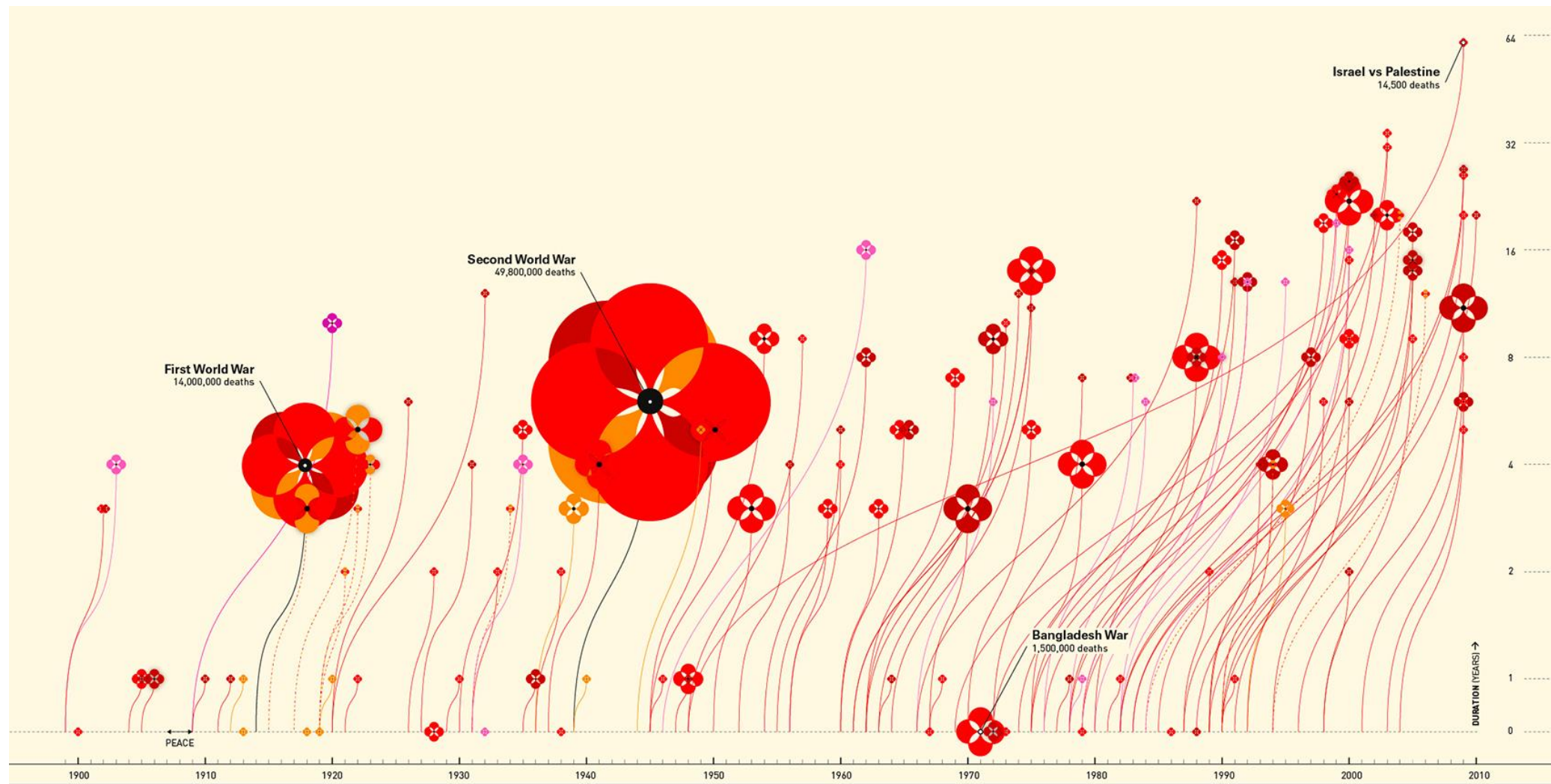




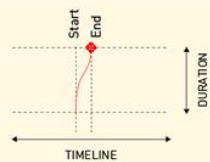
Latest data published by the US Energy Information Administration provides a unique picture of economic growth - and decline. China has sped ahead of the US as shown by this map, which resizes each country according to CO2 emissions. And, for the first time, world emissions have gone down.



City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank	City	2008 Rank	2009 Rank	2010 Rank
1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	7	8	8	8	
9	9	9	9	10	10	10	10	11	11	11	11	12	12	12	12	13	13	13	13	14	14	14	14	15	15	15	15	16	16	16	
17	17	17	17	18	18	18	18	19	19	19	19	20	20	20	20	21	21	21	21	22	22	22	22	23	23	23	23	24	24	24	
25	25	25	25	26	26	26	26	27	27	27	27	28	28	28	28	29	29	29	29	30	30	30	30	31	31	31	31	32	32	32	
33	33	33	33	34	34	34	34	35	35	35	35	36	36	36	36	37	37	37	37	38	38	38	38	39	39	39	39	40	40	40	
41	41	41	41	42	42	42	42	43	43	43	43	44	44	44	44	45	45	45	45	46	46	46	46	47	47	47	47	48	48	48	
49	49	49	49	50	50	50	50	51	51	51	51	52	52	52	52	53	53	53	53	54	54	54	54	55	55	55	55	56	56	56	
57	57	57	57	58	58	58	58	59	59	59	59	60	60	60	60	61	61	61	61	62	62	62	62	63	63	63	63	64	64	64	
65	65	65	65	66	66	66	66	67	67	67	67	68	68	68	68	69	69	69	69	70	70	70	70	71	71	71	71	72	72	72	
73	73	73	73	74	74	74	74	75	75	75	75	76	76	76	76	77	77	77	77	78	78	78	78	79	79	79	79	80	80	80	
81	81	81	81	82	82	82	82	83	83	83	83	84	84	84	84	85	85	85	85	86	86	86	86	87	87	87	87	88	88	88	
89	89	89	89	90	90	90	90	91	91	91	91	92	92	92	92	93	93	93	93	94	94	94	94	95	95	95	95	96	96	96	
97	97	97	97	98	98	98	98	99	99	99	99	100	100	100	100	101	101	101	101	102	102	102	102	103	103	103	103	104	104	104	
105	105	105	105	106	106	106	106	107	107	107	107	108	108	108	108	109	109	109	109	110	110	110	110	111	111	111	111	112	112	112	
113	113	113	113	114	114	114	114	115	115	115	115	116	116	116	116	117	117	117	117	118	118	118	118	119	119	119	119	120	120	120	
121	121	121	121	122	122	122	122	123	123	123	123	124	124	124	124	125	125	125	125	126	126	126	126	127	127	127	127	128	128	128	
129	129	129	129	130	130	130	130	131	131	131	131	132	132	132	132	133	133	133	133	134	134	134	134	135	135	135	135	136	136	136	
137	137	137	137	138	138	138	138	139	139	139	139	140	140	140	140	141	141	141	141	142	142	142	142	143	143	143	143	144	144	144	
145	145	145	145	146	146	146	146	147	147	147	147	148	148	148	148	149	149	149	149	150	150	150	150	151	151	151	151	152	152	152	
153	153	153	153	154	154	154	154	155	155	155	155	156	156	156	156	157	157	157	157	158	158	158	158	159	159	159	159	160	160	160	
161	161	161	161	162	162	162	162	163	163	163	163	164	164	164	164	165	165	165	165	166	166	166	166	167	167	167	167	168	168	168	
169	169	169	169	170	170	170	170	171	171	171	171	172	172	172	172	173	173	173	173	174	174	174	174	175	175	175	175	176	176	176	
177	177	177	177	178	178	178	178	179	179	179	179	180	180	180	180	181	181	181	181	182	182	182	182	183	183	183	183	184	184	184	
185	185	185	185	186	186	186	186	187	187	187	187	188	188	188	188	189	189	189	189	190	190	190	190	191	191	191	191	192	192	192	
193	193	193	193	194	194	194	194	195	195	195	195	196	196	196	196	197	197	197	197	198	198	198	198	199	199	199	199	200	200	200	
201	201	201	201	202	202	202	202	203	203	203	203	204	204	204	204	205	205	205	205	206	206	206	206	207	207	207	207	208	208	208	
209	209	209	209	210	210	210	210	211	211	211	211	212	212	212	212	213	213	213	213	214	214	214	214	215	215	215	215	216	216	216	
217	217	217	217	218	218	218	218	219	219	219	219	220	220	220	220	221	221	221	221	222	222	222	222	223	223	223	223	224	224	224	
225	225	225	225	226	226	226	226	227	227	227	227	228	228	228	228	229	229	229	229	230	230	230	230	231	231	231	231	232	232	232	
233	233	233	233	234	234	234	234	235	235	235	235	236	236	236	236	237	237	237	237	238	238	238	238	239	239	239	239	240	240	240	
241	241	241	241	242	242	242	242	243	243	243	243	244	244	244	244	245	245	245	245	246	246	246	246	247	247	247	247	248	248	248	
249	249	249	249	250	250	250	250	251	251	251	251	252	252	252	252	253	253	253	253	254	254	254	254	255	255	255	255	256	256	256	
257	257	257	257	258	258	258	258	259	259	259	259	260	260	260	260	261	261	261	261	262	262	262	262	263	263	263	263	264	264	264	
265	265	265	265	266	266	266	266	267	267	267	267	268	268	268	268	269	269	269	269	270	270	270	270	271	271	271	271	272	272	272	
273	273	273	273	274	274	274	274	275	275	275	275	276	276	276	276	277	277	277	277	278	278	278	278	279	279	279	279	280	280	280	
281	281	281	281	282	282	282	282	283	283	283	283	284	284	284	284	285	285	285	285	286	286	286	286	287	287	287	287	288	288	288	
289	289	289	289	290	290	290	290	291	291	291	291	292	292	292	292	293	293	293	293	294	294	294	294	295	295	295	295	296	296	296	
297	297	297	297	298	298	298	298	299	299	299	299	300	300	300	300	301	301	301	301	302	302	302	302	303	303	303	303	304	304	304	
305	305	305	305	306	306	306	306	307	307	307	307	308	308	308	308	309	309	309	309	310	310	310	310	311	311	311	311	312	312	312	
313	313	313	313	314	314	314	314	315	315	315	315	316	316	316	316	317	317	317	317	318	318	318	318	319	319	319	319	320	320	320	
321	321	321	321	322	322	322	322	323	323	323	323	324	324	324	324	325	325	325	325	326	326	326	326	327	327	327	327	328	328	328	
329	329	329	329	330	330	330	330	331	331	331	331	332	332	332	332	333	333	333	333	334	334	334	334	335	335	335	335	336	336	336	
337	337	337	337	338	338	338	338	339	339	339	339	340	340	340	340	341	341	341	341	342	342	342	342	343	343	343	343	344	344	344	
345	345	345	345	346	346	346	346	347	347	347	347	348	348	348	348	349	349	349	349	350	350	350	350	351	351	351	351	352	352	352	
353	353	353	353	354	354	354	354	355	355	355	355	356	356	356	356	357	357	357	357	358	358	358	358	359	359	359	359	360	360	360	
361	361	361	361	362	362	362	362	363	363	363	363	364	364	364	364	365	365	365	365	366	366	366	366	367	367	367	367	368	368	368	
369	369	369	369	370	370	370	370	371	371	371	371	372	372	372	372	373	373	373	373	374	374	374	374	375	375	375	375	376	376	376	
377	377	377	377	378	378	378	378	379	379	379	379	380	380	380	380	381	381	381	381	382	382	382	382	383	383	383	383	384	384	384	
385	385	385	385	386	386	386	386	387	387	387	387	388	388	388	388	389	389	389	389	390	390	390	390	391	391	391	391	392	392	392	
393	393	393	393	394	394	394	394	395	395	395	395	396	396																		



#### 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)



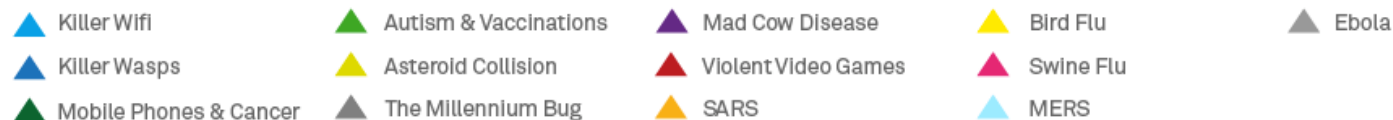
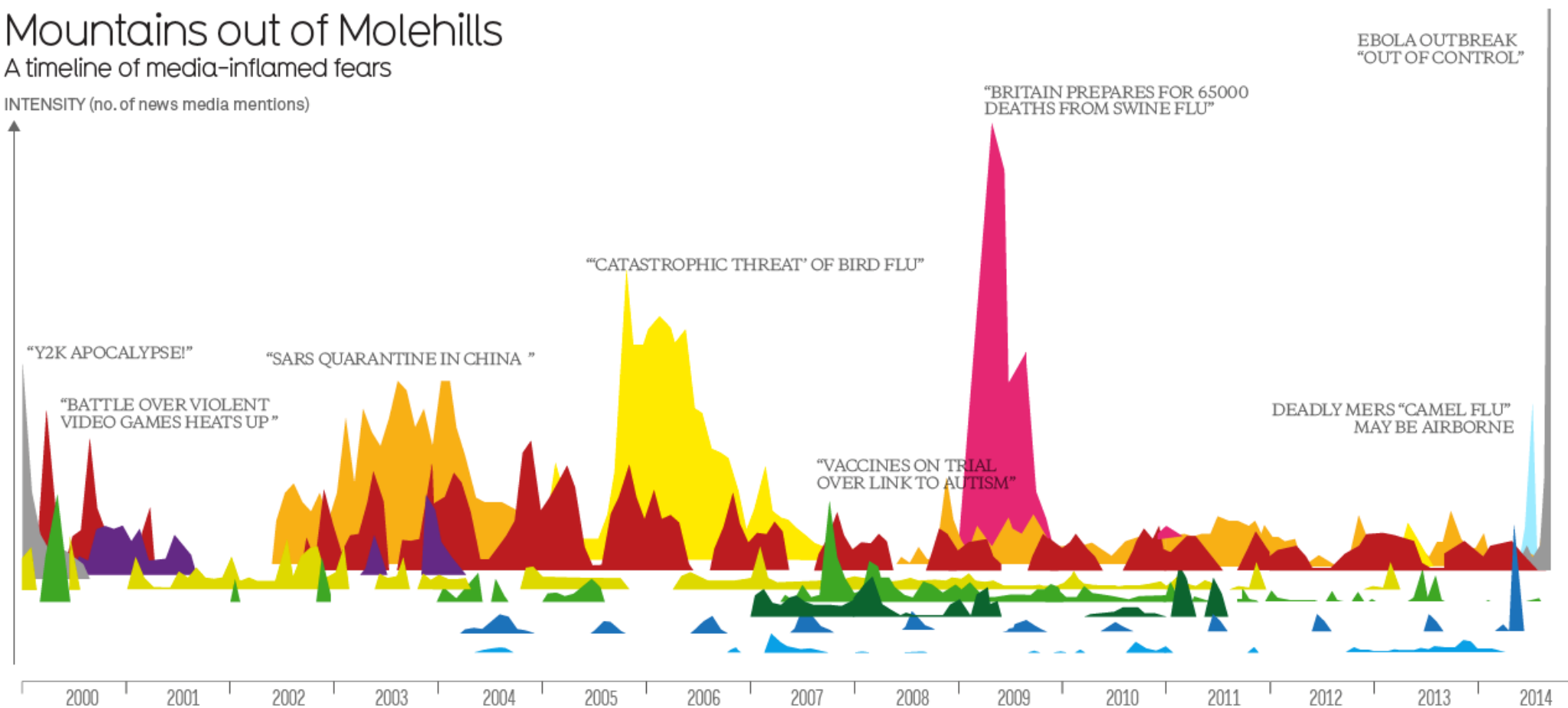
#### REGIONS INVOLVED IN WARS (POPPY'S COLOUR)



# Mountains out of Molehills

A timeline of media-inflamed fears

INTENSITY (no. of news media mentions)



rollover to scale by deaths

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

