Data Munging

Week 4

Huanfa Chen

CASA0006: Spatial Data Capture and Analysis

CASA0009: Data Science for Spatial Systems

CASA0006

1. Introduction to Databases
2. Introduction to SQL
3. Advanced SQL
4. Data Munging
5. Advanced Clustering
6. Advanced Regression
7. Classification
8. Dimension Reduction
9. Unstructured Data
10. Analysis Workflow

CASA0009

1. Introduction to Databases
2. Introduction to SQL
3. Advanced SQL
4. Data Munging
5. Advanced Clustering
6. Advanced Regression
7. Interactive Viz 1: HTML + CSS
8. Interactive Viz 2: Javascript
9. Server Side Coding: Node.JS
10. Real-time data visualisation

Recap

What we've learned so far

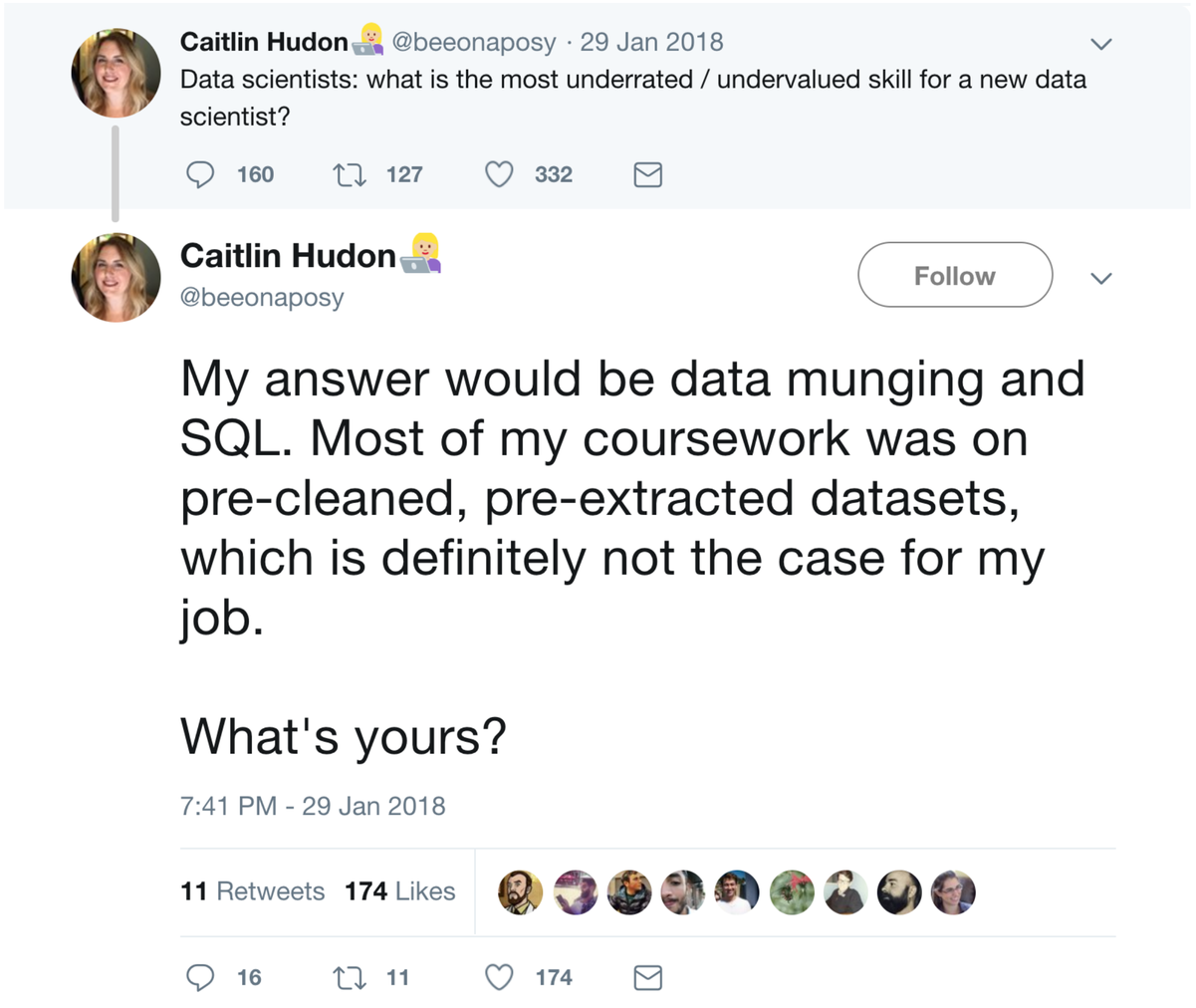
SQL – We now know lots about SQL!

We can connect to a database and use a database IDE.

We can also **query** data, **join** tables, **amend** tables and **update** our database.

But SQL will only help you so much in the wild, wild west of real world data analysis

Today we learn how to manage data, how to clean it up, organise it properly, and check whether it is actually useable



https://twitter.com/beeonaposy/status/958062683118624768

Data Handling

What's the Motivation

Bad Data in then Bad Data Out

- Data does not always arrive in the format you need it in
- In order to analyse it, you need to learn how to work with it
- Increasing numbers of tools and procedures are aiding this process, but ultimately you'll need to work hard to derive meaningful analysis of the data
- This is the least glamorous side of data science

Data Formats

Types and Structures, Advantages and Disadvantages

Classification of data files (Text vs. Binary)

Text Files

- Examples: TXT, CSV, TSV, XML, SQL, JSON
- Can be quickly explored using command-line tools, such as Unix utilities (grep, less, awk, tail, head)
- Can be edited using text editors (notepad, notepad++, vim)
- Be careful of the file encoding: ASCII, UTF-8, Unicode (always use UTF-8 if possible)

Binary Files

- Examples: XLSX (Excel data), Video/audio/photo files, zip files, sas files
- You need a library (e.g. Python) to read

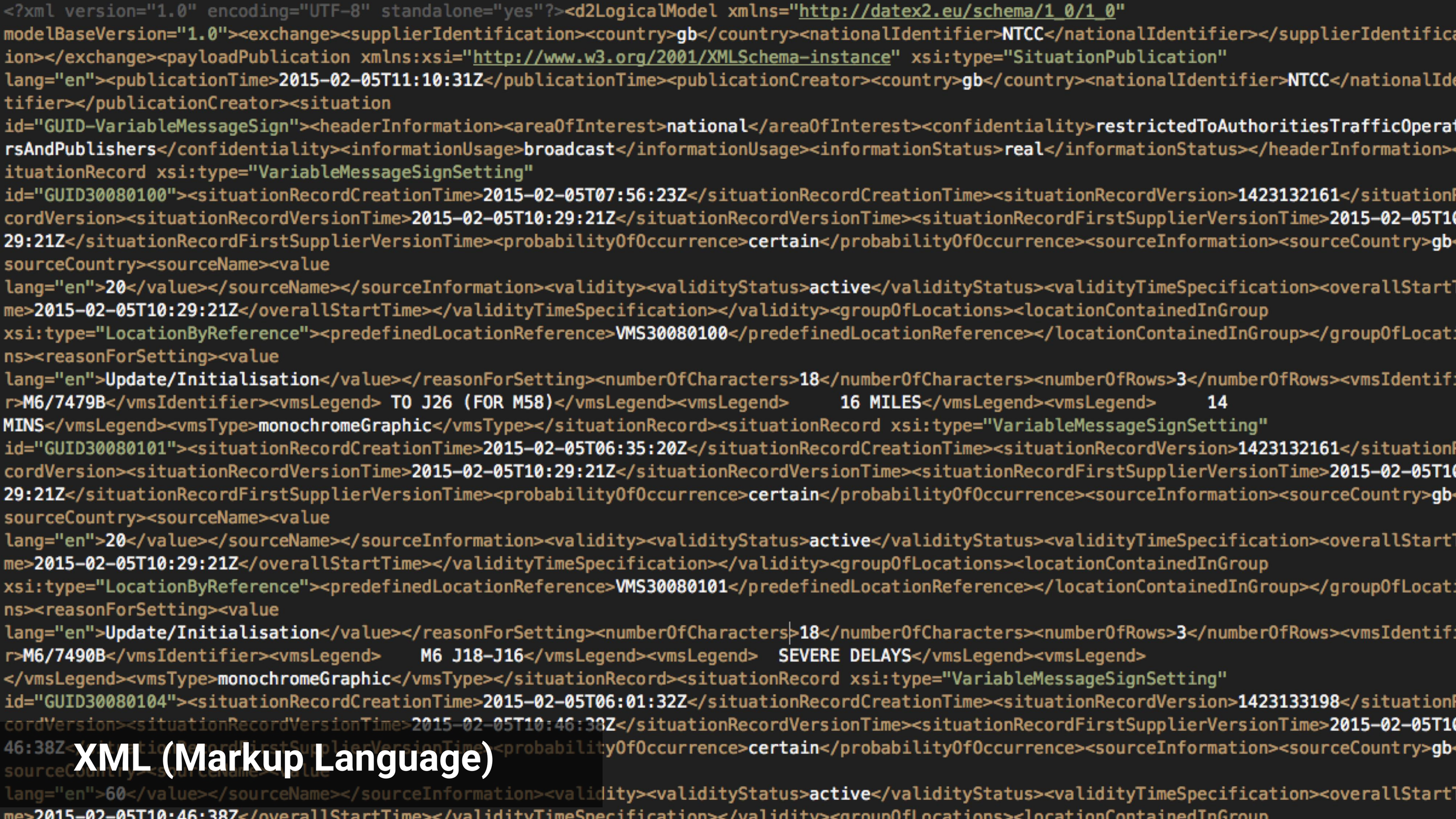
air-quality-london-monthly-averages.csv

```
Month,London Mean Roadside:Nitric Oxide (ug/m3),London Mean Roadside:Nitrogen Dioxide (ug/m3),London Mean Roadside:Oxides of Nitrogen (ug/m3),London Mean Roadside:Ozone (ug/
m3),London Mean Roadside:PM10 Particulate (ug/m3),London Mean Roadside:PM2.5 Particulate (ug/m3),London Mean Roadside:Sulphur Dioxide (ug/m3),London Mean Background:Nitric Oxid
m3),London Mean Background:Nitrogen Dioxide (ug/m3),London Mean Background:Oxides of Nitrogen (ug/m3),London Mean Background:Ozone (ug/m3),London Mean Background:PM10 Particula
m3),London Mean Background:PM2.5 Particulate (ug/m3),London Mean Background:Sulphur Dioxide (ug/m3)
Jan-08,,55.50268817,,29.51209677,24.96908602,14.67876344,4.217741935,,42.33870968,,36.9422043,18.8172043,.,3.572580645
Feb-08,,75.92241379,,20.31752874,39.47701149,28.77298851,7.55316092,,60.23706897,,26.42528736,31.89655172,.,6.734195402
Mar-08,,55.61021505,,40.10349462,21.56989247,12.30013459,3.86827957,,39.80107527,,50.22715054,15.47715054,.,2.286290323
Apr-08,,61.75694444,,37.88472222,28.74027778,20.46111111,4.475,,44.00972222,,50.13333333,21.72916667,.,3.236111111
May-08,,62.90322581,,46.26612903,34.61155914,27.50806452,4.634408602,,44.14112903,,60.51209677,29.54569892,16.5768262,4.25
Jun-08,,49.16111111,,39.83611111,23.19861111,16.01005747,3.593055556,,31.24166667,,51.32638889,18.25,12.61376404,2.547222222
Jul-08,,48.44489247,,34.98252688,22.95833333,14.2405914,3.100806452,,31.21639785,,46.62365591,17.20430108,11.89919355,2.49327957
Aug-08,,41.07258065,,30.02150538,20.69354839,11.45295699,2.155913978,,27.85080645,,37.09408602,15.50806452,11.20137931,2.088709677
Sep-08,,54.08055556,,22.375,28.22777778,17.97916667,3.748611111,,41.21527778,,28.88611111,22.24444444,15.32451253,3.056944444
Oct-08,,56.65860215,,19.33736559,23.00268817,12.91801075,4.305107527,,43.81317204,,25.42741935,16.46908602,11.62096774,2.838709677
Nov-08,,57.92916667,,18.80972222,21.85416667,14.71388889,3.2875,,42.31111111,,22.58194444,16.75277778,12.89027778,2.672222222
Dec-08,,65.10349462,,14.48655914,27.29704301,20.51612903,3.791666667,,47.50806452,,16.67069892,21.51075269,18.09543011,3.5
Jan-09,,68.52284946,,18.78494624,31.11290323,23.17741935,5.880376344,,51.98521505,,21.47177419,25.27553763,20.65994624,4.842741935
Feb-09,,71.92559524,,15.07440476,28.01041667,19.07589286,4.680059524,,52.66815476,,20.33779762,20.8452381,17.23809524,3.026785714
Mar-09,,65.15456989,,32.32930108,29.97177419,19.70698925,3.266129032,,45.34408602,,40.16801075,24.31586022,17.27956989,2.920698925
Apr-09,,63.175,,35.7375,32.59861111,26.08472222,2.691666667,,40.14583333,,46.03888889,27.69027778,21.61666667,2.7375
May-09,,48.61021505,,45.38575269,24.87768817,14.30510753,2.819892473,,28.08064516,,55.11290323,19.74865591,11.81451613,2.436827957
Jun-09,,56.07638889,,40.12222222,24.68055556,15.9375,3.626388889,,33.13194444,,52.51527778,19.85277778,13.11388889,2.516666667
Jul-09,,44.36827957,,32.77150538,20.22446237,11.2311828,2.422043011,,24.85752688,,40.43951613,15.49462366,9.841397849,2.15188172
Aug-09,,48.65994624,,28.60349462,21.69623656,13.25672043,2.258064516,,29.20833333,,35.95295699,16.58333333,11.38978495,2.478494624
Sep-09,,51.55972222,,26.32083333,24.25555556,14.44305556,2.404166667,,32.67361111,,33.11666667,19.11111111,11.61666667,3.009722222
Oct-09,,62.9422043,,15.03763441,27.24731183,18.0188172,4.255376344,,44.29032258,,20.27822581,21.51344086,14.38306452,4.025537634
Nov-09,,53.65555556,,29.9,21.94583333,12.89583333,3.378623188,,36.63888889,,36.52916667,17.54305556,11.39166667,2.829166667
Dec-09,,65.18010753,,18.88037634,23.09811828,16.27956989,4.127688172,,49.95564516,,22.26478495,18.58736559,13.94892473,3.642473118
Jan-10,100.5037634,65.68024194,166.4126344,14.03225806,30.17271505,24.34274194,4.729973118,45.19879032,51.90295699,97.19637097,17.35080645,25.0391129,22.0844086,4.204569892
Feb-10,78.78794643,63.8702381,142.7508929,23.81309524,25.19434524,18.33511905,4.586904762,29.2891369,47.6547619,76.99598214,29.82738095,19.40416667,15.94598214,3.230952381
Mar-10,72.37553763,59.83548387,120.6416667,26.42473118,25.61290323,15.86034946,3.429301075,22.99905914,40.81626344,63.82459677,42.51935484,19.39637097,14.07782258,2.696908602
Apr-10,63.91111111,60.15444444,124.0469444,37.17333333,29.75736111,18.81652778,4.640277778,20.9075,41.90416667,62.83097222,50.20833333,24.33041667,16.26958333,4.123611111
May-10,60.8608871,53.23010753,114.0969086,35.09126344,24.53534946,16.19112903,3.883467742,17.32392473,33.09704301,50.47930108,47.2327957,19.85994624,13.2313172,3.175134409
Jun-10,55.21847222,53.85861111,109.1,34.00972222,26.51805556,17.44194444,4.414583333,13.35805556,33.68222222,47.08291667,50.17597222,21.97083333,15.07541667,3.183194444
Jul-10,58.70712366,46.50860215,105.2418011,28.56680108,20.90107527,12.45336022,3.138978495,11.53413978,25.10887097,36.62634409,42.75228495,16.52365591,8.934408602,2.623521505
Aug-10,60.60376344,46.61518817,107.2458333,24.25107527,18.14099462,10.97446237,3.255913978,12.95591398,26.13763441,39.0641129,35.89879032,15.36129032,8.944623656,1.986290323
Sep-10,78.13791667,57.15319444,135.3459722,20.34319444,23.82527778,14.45958333,3.593888889,23.93791667,35.42777778,59.37722222,31.80902778,14.48555556,11.43555556,2.681388889
Oct-10,82.04825269,59.5686828,141.5930108,19.22755376,25.24086022,15.9061828,3.549865591,30.71223118,38.48481183,69.20147849,27.72997312,19.95241935,12.9655914,2.742741935
Nov-10,116.5559722,65.67013889,182.2393056,15.44180556,28.23222222,21.89291667,4.167083333,60.60083333,46.87958333,107.4583333,20.41666667,22.53055556,18.1975,3.685416667
Dec-10,111.1717742,71.32150538,182.4034946,10.65819892,27.19301075,22.61370968,3.904704301,54.85215054,53.8922043,108.8001344,13.86948925,21.69180108,19.24260753,3.683602151
Jan-11,78.13669355,61.80927419,139.9653226,21.48602151,24.35564516,16.98158602,3.381317204,31.90537634,44.37271505,76.39516129,26.48642473,19.39086022,14.64193548,3.461827957
Feb-11,[illegible]357,2.898958333,25.78303571,41.19122024,66.6110119,31.33630952,26.51845238,19.87157738,2.341666667
Mar-11,[illegible]957,3.803897849,36.55430108,49.69798387,86.27284946,32.50201613,36.93266129,29.91236559,4.123790323
Apr-11,64.41958333,63.40041667,127.7566667,38.70138889,39.22972222,28.22736111,3.123055556,17.33555556,40.38388889,57.69708333,52.86583333,34.19138889,25.63902778,4.414722222
May-11,50.15336022,49.42836022,99.51491935,42.35443548,23.68360215,13.0358871,2.404973118,9.665725806,26.24569892,35.9172043,57.42258065,19.12271505,11.31048387,3.215725806
Jun-11,46.36125,50.09527778,111.035,30.27319444,20.685,12.02555556,2.25,12.18291667,25.50986111,37.70125,43.06791667,15.73722222,10.53125,2.789166667
```

CSV (Comma Separated Values)

segment_key	count	avg_time	sd_time	max_time	min_time	source	target	cv	type
3587	391205	9.1149	6.7805	359	4	741	669	0.74389187	Directed
984	360170	9.586	9.7758	358	4	669	741	1.01979971	Directed
68	353014	25.6862	7.526	358	0	5426	5355	0.2929978	Directed
953	348145	14.7908	9.1934	359	8	5595	5598	0.62156205	Directed
65	334673	25.5393	7.6754	359	11	5355	5426	0.3005329	Directed
3352	324713	22.8086	10.7298	359	0	5598	5578	0.47042782	Directed
472	318177	14.6108	11.3782	359	8	5598	5595	0.7787527	Directed
1100	317666	14.9379	9.6449	359	10	852	796	0.64566639	Directed
954	297374	14.644	5.2328	358	9	796	852	0.35733406	Directed
335	284135	13.2136	9.4644	359	6	5595	5426	0.71626203	Directed
221	279761	23.5598	6.7942	357	0	5603	8909	0.28838106	Directed
650	278326	7.498	4.5546	357	2	513	747	0.60744198	Directed
4563	276120	22.8318	10.4154	358	13	5578	5598	0.45617954	Directed
731	269510	11.475	9.2281	359	7	852	635	0.80419172	Directed
364	265523	8.7586	4.9081	342	3	747	513	0.56037495	Directed
336	257254	13.3257	10.0377	359	6	5426	5595	0.75325874	Directed
225	253118	22.4641	8.2601	358	13	8909	5603	0.36770224	Directed
711	241084	11.5406	7.5678	359	6	635	852	0.65575447	Directed
362	234970	16.8952	4.9723	338	0	778	669	0.29430252	Directed
2229	224768	28.9935	7.3229	354	17	5570	8909	0.2525704	Directed
628	205687	15.347	6.1701	358	10	635	524	0.40203949	Directed
2421	203149	12.4098	5.0728	358	8	777	669	0.40877371	Directed
1617	200008	17.1539	6.6322	352	12	669	778	0.38662928	Directed
748	197022	24.7427	7.2107	355	14	5355	6336	0.29142737	Directed
2230	196840	27.7016	8.0205	351	16	8909	5570	0.28953201	Directed
2244	190282	24.3941	8.4964	352	14	6336	5355	0.34829733	Directed
2756	186073	9.4696	13.7077	359	4	635	796	1.44754794	Directed
3019	183795	30.4862	7.9075	337	16	1402	1444	0.25937965	Directed
3236	182898	16.8841	12.5801	359	9	5595	8909	0.74508561	Directed

TSV (Tab Separated Values)



XML (Markup Language)

```
<ROWSET>
  <ROW>
    <Area_Code>00AA</Area_Code>
    <Area_Name>City of London</Area_Name>
    <All_households>4338</All_households>
    <Households_-_No_adults_in_employment_-_with_dependent_children>118</Households_-_No_adults_in_employment_-_with_dependent_children>
    <Households_-_No_adults_in_employment_-_without_dependent_children>1128</Households_-_No_adults_in_employment_-_without_dependent_children>
    <Households_-_With_dependent_children_-_All_ages>455</Households_-_With_dependent_children_-_All_ages>
    <Households_-_With_dependent_children_-_Aged_0_-_4>198</Households_-_With_dependent_children_-_Aged_0_-_4>
    <Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>838</Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>
  </ROW>
  <ROW>
    <Area_Code>00AB</Area_Code>
    <Area_Name>Barking and Dagenham</Area_Name>
    <All_households>67273</All_households>
    <Households_-_No_adults_in_employment_-_with_dependent_children>6555</Households_-_No_adults_in_employment_-_with_dependent_children>
    <Households_-_No_adults_in_employment_-_without_dependent_children>21036</Households_-_No_adults_in_employment_-_without_dependent_children>
    <Households_-_With_dependent_children_-_All_ages>22816</Households_-_With_dependent_children_-_All_ages>
    <Households_-_With_dependent_children_-_Aged_0_-_4>10129</Households_-_With_dependent_children_-_Aged_0_-_4>
    <Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>25988</Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>
  </ROW>
  <ROW>
    <Area_Code>00AC</Area_Code>
    <Area_Name>Barnet</Area_Name>
    <All_households>126944</All_households>
    <Households_-_No_adults_in_employment_-_with_dependent_children>6627</Households_-_No_adults_in_employment_-_with_dependent_children>
    <Households_-_No_adults_in_employment_-_without_dependent_children>33008</Households_-_No_adults_in_employment_-_without_dependent_children>
    <Households_-_With_dependent_children_-_All_ages>39180</Households_-_With_dependent_children_-_All_ages>
    <Households_-_With_dependent_children_-_Aged_0_-_4>15875</Households_-_With_dependent_children_-_Aged_0_-_4>
    <Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>36097</Households_-_With_one_or_more_person_with_a_limiting_long-term_illness>
  </ROW>
  <ROW>
    <Area_Code>00AD</Area_Code>
    <All_households>89451</All_households>
    <Households_-_No_adults_in_employment_-_with_dependent_children>3936</Households_-_No_adults_in_employment_-_with_dependent_children>
```


XML (Markup Language)

JEOPARDY_QUESTIONS1.json

[{"category": "HISTORY", "air_date": "2004-12-31", "question": "'For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory'", "value": "$200", "answer": "Copernicus", "round": "Jeopardy!", "show_number": "4680"}, {"category": "ESPN's TOP 10 ALL-TIME ATHLETES", "air_date": "2004-12-31", "question": "'No. 2: 1912 Olympian; football star at Carlisle Indian School; 6 MLB seasons with the Reds, Giants & Braves'", "value": "$200", "answer": "Jim Thorpe", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EVERYBODY TALKS ABOUT IT...", "air_date": "2004-12-31", "question": "'The city of Yuma in this state has a record average of 4,055 hours of sunshine each year'", "value": "$200", "answer": "Arizona", "round": "Jeopardy!", "show_number": "4680"}, {"category": "THE COMPANY LINE", "air_date": "2004-12-31", "question": "'In 1963, live on \"The Art Linkletter Show\", this company served its billionth burger'", "value": "$200", "answer": "McDonald\\'s", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EPITAPHS & TRIBUTES", "air_date": "2004-12-31", "question": "'Signer of the Dec. of Indep., framer of the Constitution of Mass., second President of the United States'", "value": "$200", "answer": "John Adams", "round": "Jeopardy!", "show_number": "4680"}, {"category": "3-LETTER WORDS", "air_date": "2004-12-31", "question": "'In the title of an Aesop fable, this insect shared billing with a grasshopper'", "value": "$200", "answer": "the ant", "round": "Jeopardy!", "show_number": "4680"}, {"category": "HISTORY", "air_date": "2004-12-31", "question": "'Built in 312 B.C. to link Rome & the South of Italy, it's still in use today'", "value": "$400", "answer": "the Appian Way", "round": "Jeopardy!", "show_number": "4680"}, {"category": "ESPN's TOP 10 ALL-TIME ATHLETES", "air_date": "2004-12-31", "question": "'No. 8: 30 steals for the Birmingham Barons; 2,306 steals for the Bulls'", "value": "$400", "answer": "Michael Jordan", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EVERYBODY TALKS ABOUT IT...", "air_date": "2004-12-31", "question": "'In the winter of 1971-72, a record 1,122 inches of snow fell at Rainier Paradise Ranger Station in this state'", "value": "$400", "answer": "Washington", "round": "Jeopardy!", "show_number": "4680"}, {"category": "THE COMPANY LINE", "air_date": "2004-12-31", "question": "'This housewares store was named for the packaging its merchandise came in & was first displayed on'", "value": "$400", "answer": "Crate & Barrel", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EPITAPHS & TRIBUTES", "air_date": "2004-12-31", "question": "'\"And away we go\"'", "value": "$400", "answer": "Jackie Gleason", "round": "Jeopardy!", "show_number": "4680"}, {"category": "3-LETTER WORDS", "air_date": "2004-12-31", "question": "'Cows regurgitate this from the first stomach to the mouth & chew it again'", "value": "$400", "answer": "the cud", "round": "Jeopardy!", "show_number": "4680"}, {"category": "HISTORY", "air_date": "2004-12-31", "question": "'In 1000 Rajaraja I of the Cholas battled to take this Indian Ocean island now known for its tea'", "value": "$600", "answer": "Ceylon (or Sri Lanka)", "round": "Jeopardy!", "show_number": "4680"}, {"category": "ESPN's TOP 10 ALL-TIME ATHLETES", "air_date": "2004-12-31", "question": "'No. 1: Lettered in hoops, football & lacrosse at Syracuse & if you think he couldn't act, ask his 11 \"unclean\" buddies'", "value": "$600", "answer": "Jim Brown", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EVERYBODY TALKS ABOUT IT...", "air_date": "2004-12-31", "question": "'On June 28, 1994 the nat'l weather service began issuing this index that rates the intensity of the sun's radiation'", "value": "$600", "answer": "the UV index", "round": "Jeopardy!", "show_number": "4680"}, {"category": "THE COMPANY LINE", "air_date": "2004-12-31", "question": "'This company's Accutron watch, introduced in 1960, had a guarantee of accuracy to within one minute a month'", "value": "$600", "answer": "Bulova", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EPITAPHS & TRIBUTES", "air_date": "2004-12-31", "question": "'Outlaw: \"Murdered by a traitor and a coward whose name is not worthy to appear here\"'", "value": "$600", "answer": "Jesse James", "round": "Jeopardy!", "show_number": "4680"}, {"category": "3-LETTER WORDS", "air_date": "2004-12-31", "question": "'A small demon, or a mischievous child (who might be a little demon!)'", "value": "$600", "answer": "imp", "round": "Jeopardy!", "show_number": "4680"}, {"category": "HISTORY", "air_date": "2004-12-31", "question": "'Karl led the first of these Marxist organizational efforts; the second one began in 1889'", "value": "$800", "answer": "the International", "round": "Jeopardy!", "show_number": "4680"}, {"category": "ESPN's TOP 10 ALL-TIME ATHLETES", "air_date": "2004-12-31", "question": "'No. 10: FB/LB for Columbia U. in the 1920s; MVP for the Yankees in '27 & '36; \"Gibraltar in Cleats\"'", "value": "$800", "answer": "(Lou) Gehrig", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EVERYBODY TALKS ABOUT IT...", "air_date": "2004-12-31", "question": "'Africa's lowest temperature was 11 degrees below zero in 1935 at Ifrane, just south of Fez in this country'", "value": "$800", "answer": "Morocco", "round": "Jeopardy!", "show_number": "4680"}, {"category": "THE COMPANY LINE", "air_date": "2004-12-31", "question": "'Edward Teller & this man partnered in 1898 to sell high fashions to women'", "value": "$800", "answer": "(Paul) Bonwit", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EPITAPHS & TRIBUTES", "air_date": "2004-12-31", "question": "'1939 Oscar winner: \"...you are a credit to your craft, your race and to your family\"'", "value": "$2,000", "answer": "Hattie McDaniel (for her role in Gone with the Wind)", "round": "Jeopardy!", "show_number": "4680"}, {"category": "3-LETTER WORDS", "air_date": "2004-12-31", "question": "'In geologic time one of these, shorter than an eon, is divided into periods & subdivided into epochs'", "value": "$800", "answer": "era", "round": "Jeopardy!", "show_number": "4680"}, {"category": "HISTORY", "air_date": "2004-12-31", "question": "'This Asian political party was founded in 1885 with \"Indian National\" as part of its name'", "value": "$1000", "answer": "the Congress Party", "round": "Jeopardy!", "show_number": "4680"}, {"category": "ESPN's TOP 10 ALL-TIME ATHLETES", "air_date": "2004-12-31", "question": "'No. 5: Only center to lead the NBA in assists; track scholarship to Kansas U.; marathoner; volleyballer'", "value": "$1000", "answer": "(Wilt) Chamberlain", "round": "Jeopardy!", "show_number": "4680"}, {"category": "THE COMPANY LINE", "air_date": "2004-12-31", "question": "'The Kirschner brothers, Don & Bill, named this ski company for themselves & the second-highest mountain'", "value": "$1000", "answer": "K2", "round": "Jeopardy!", "show_number": "4680"}, {"category": "EPITAPHS & TRIBUTES", "air_date": "2004-12-31", "question": "'Revolutionary War hero: \"His spirit is in Vermont now\"'", "value": "$1000", "answer": "Ethan Allen", "round": "Jeopardy!", "show_number": "4680"}, {"category": "3-LETTER WORDS", "air_date": "2004-12-31", "question": "'A single layer of paper, or to perform one's craft diligently'", "value": "$1000", "answer": "ply", "round": "Jeopardy!", "show_number": "4680"}, {"category": "DR. SEUSS AT THE MULTIPLEX", "air_date": "2004-12-31", "question": "'<a href=\"http://www.j-archive.com/media/[illegible] you won't believe the ending when he \"Hatches the Egg\"'", "value": "$400", "answer": "Horton", "round": "Double Jeopardy!", "show_number": "4680"}, {"category": "PRESIDENTIAL STATES OF BIRTH", "air_date": "2004-12-31", "question": "'California'", "value": "$400", "answer": "Nixon", "round": "Double Jeopardy!", "show_number": "4680"}, {"category": "AIRLINE TRAVEL", "air_date": "2004-12-31", "question": "'It can be a place to leave your puppy when you take a trip, or a carrier for him that fits under an airplane seat'", "value": "$400", "answer": "a kennel", "round": "Double Jeopardy!", "show_number": "4680"},
[illegible]

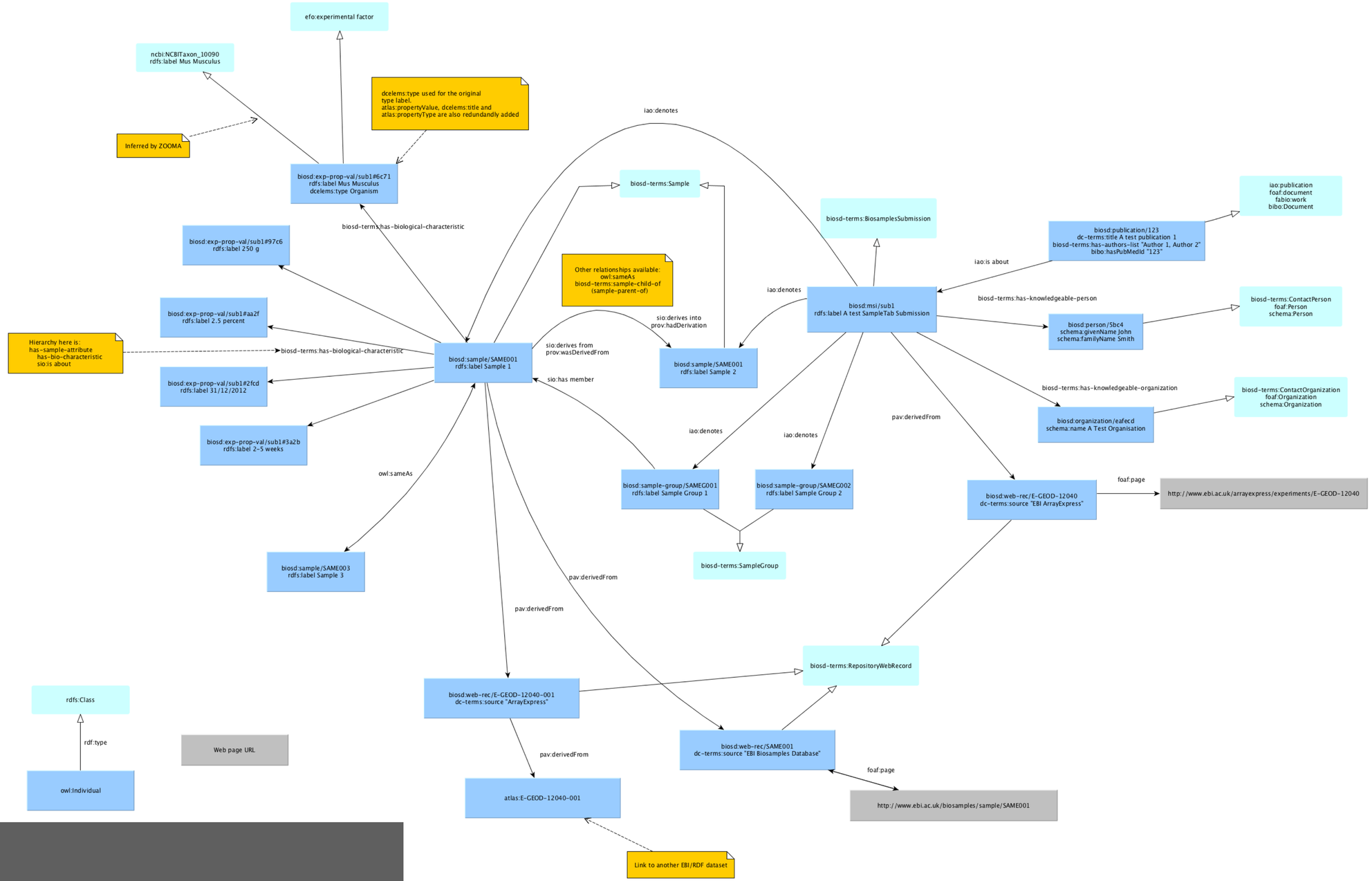
JSON (Javascript Object notation)

```
[{
    "category": "HISTORY",
    "air_date": "2004-12-31",
    "question": "'For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory'",
    "value": "$200",
    "answer": "Copernicus",
    "round": "Jeopardy!",
    "show_number": "4680"
}, {
    "category": "ESPN's TOP 10 ALL-TIME ATHLETES",
    "air_date": "2004-12-31",
    "question": "'No. 2: 1912 Olympian; football star at Carlisle Indian School; 6 MLB seasons with the Reds, Giants & Braves'",
    "value": "$200",
    "answer": "Jim Thorpe",
    "round": "Jeopardy!",
    "show_number": "4680"
}, {
    "category": "EVERYBODY TALKS ABOUT IT...",
    "air_date": "2004-12-31",
    "question": "'The city of Yuma in this state has a record average of 4,055 hours of sunshine each year'",
    "value": "$200",
    "answer": "Arizona",
    "round": "Jeopardy!",
    "show_number": "4680"
}, {
    "category": "THE COMPANY LINE",
    "air_date": "2004-12-31",
    "question": "'In 1963, live on \"The Art Linkletter Show\", this company served its billionth burger'",
    "value": "$200",
    "answer": "McDonald\\'s",
    "round": "Jeopardy!",
    "show_number": "4680"
}, {
    "category": "EPITAPHS & TRIBUTES",
    "air_date": "2004-12-31",
    "question": "'Signed the Declaration of Independence, wrote the Constitution of Mass., second President of the United States'",
    "value": "$200",
    "answer": "John Adams",
```

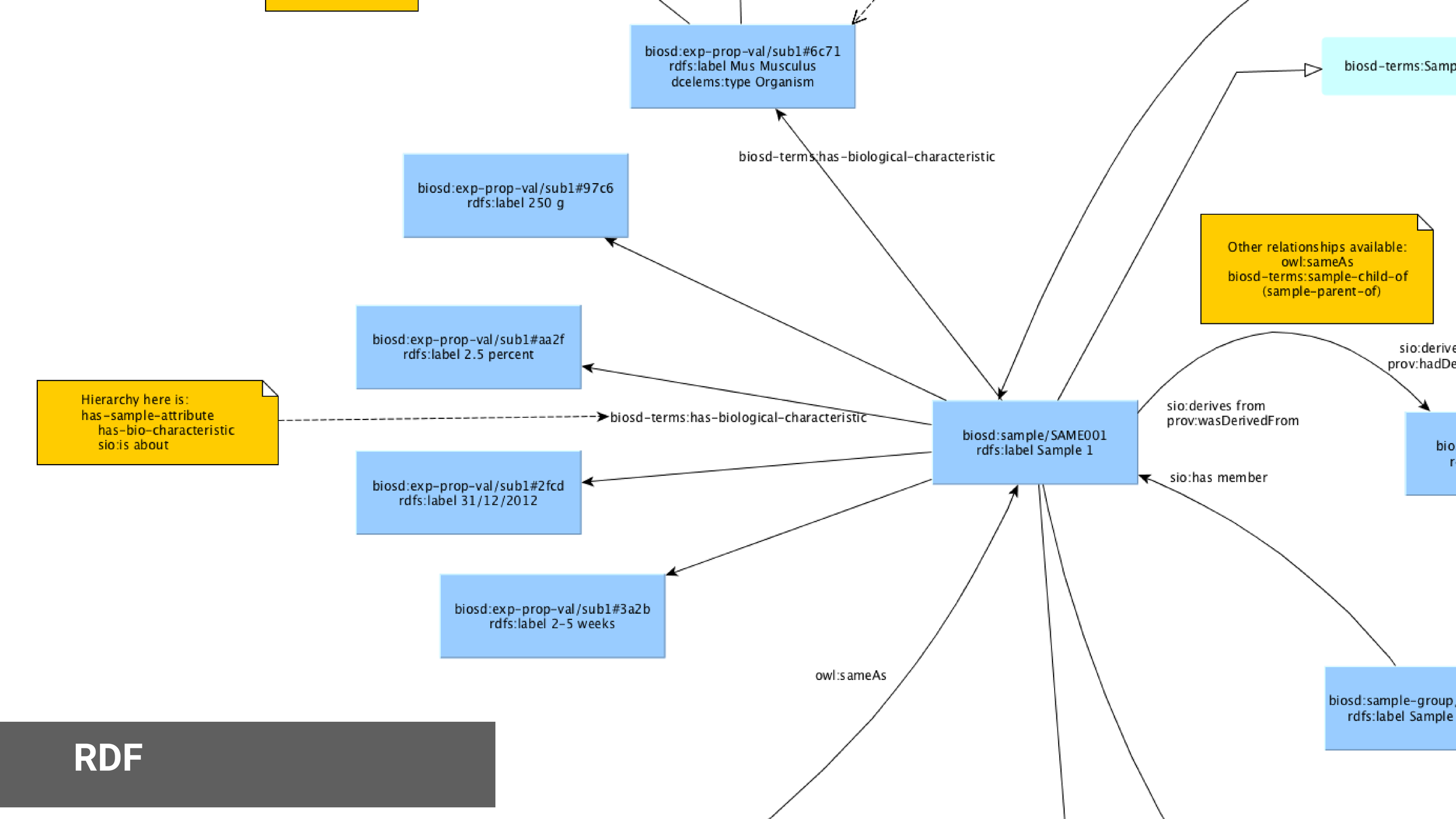

JSON (Javascript Object notation)

```
{
"type": "FeatureCollection",
"crs": { "type": "name", "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" } },

"features": [
{ "type": "Feature", "properties": { "id": 1, "lines": [ { "name": "Victoria", "network": "Tube", "colour":
"#0098D4" } ], "seg_desc": "Victoria" }, "geometry": { "type": "LineString", "coordinates": [ [ -
0.1148720, 51.4626062 ], [ -0.1154612, 51.4627201 ], [ -0.1167776, 51.4629816 ], [ -0.121132,
51.4638465 ], [ -0.1215893, 51.463931 ], [ -0.1223392, 51.4641384 ], [ -0.1234773, 51.4645792 ], [ -
0.1240446, 51.4648628 ], [ -0.1245532, 51.4652346 ], [ -0.1248932, 51.4655457 ], [ -0.1253925,
51.4661992 ], [ -0.1256804, 51.4668713 ], [ -0.125783, 064019 ], [ -0.1431186, 51.5065574 ], [ -
0.1428352, 51.5067094 ], [ -0.1418986, 51.5071603 ], . . .
```


GeoJSON


efo:experimental factor
ncbi:NCBITaxon_10090
rdfs:label Mus Musculus
dcelems:type used for the original
type label.
atlas:propertyValue, dcelems:title and
atlas:propertyType are also redundandly added
Inferred by ZOOMA
biosd:exp-prop-val/sub1#6c71
rdfs:label Mus Musculus
dcelems:type Organism
iao:denotes
biosd-terms:Sample
biosd-terms:BiosamplesSubmission
iao:publication
foaf:document
fabio:work
bibo:Document
biosd:publication/123
dc-terms:title A test publication 1
biosd-terms:has-authors-list "Author 1, Author 2"
bibo:hasPubMedId "123"
iao:is about
biosd-terms:has-biological-characteristic
biosd:exp-prop-val/sub1#97c6
rdfs:label 250 g
Other relationships available:
owl:sameAs
biosd-terms:sample-child-of
(sample-parent-of)
iao:denotes
biosd:msi/sub1
rdfs:label A test SampleTab Submission
biosd-terms:has-knowledgeable-person
biosd-terms:ContactPerson
foaf:Person
schema:Person
biosd:person/5bc4
schema:givenName John
schema:familyName Smith
biosd:exp-prop-val/sub1#aa2f
rdfs:label 2.5 percent
sio:derives into
prov:hadDerivation
Hierarchy here is:
has-sample-attribute
has-bio-characteristic
sio:is about
biosd-terms:has-biological-characteristic
biosd:sample/SAME001
rdfs:label Sample 1
sio:derives from
prov:wasDerivedFrom
biosd:sample/SAME001
rdfs:label Sample 2
biosd:exp-prop-val/sub1#2fcd
rdfs:label 31/12/2012
sio:has member
biosd-terms:has-knowledgeable-organization
biosd-terms:ContactOrganization
foaf:Organization
schema:Organization
biosd:organization/eafecd
schema:name A Test Organisation
biosd:exp-prop-val/sub1#3a2b
rdfs:label 2-5 weeks
iao:denotes
iao:denotes
pav:derivedFrom
owl:sameAs
biosd:sample-group/SAMEG001
rdfs:label Sample Group 1
biosd:sample-group/SAMEG002
rdfs:label Sample Group 2
biosd:web-rec/E-GEOD-12040
dc-terms:source "EBI ArrayExpress"
foaf:page
http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-12040
biosd:sample/SAME003
rdfs:label Sample 3
biosd-terms:SampleGroup
pav:derivedFrom
pav:derivedFrom
biosd-terms:RepositoryWebRecord
biosd:web-rec/E-GEOD-12040-001
dc-terms:source "ArrayExpress"
rdfs:Class
rdf:type
Web page URL
biosd:web-rec/SAME001
dc-terms:source "EBI Biosamples Database"
pav:derivedFrom
foaf:page
owl:Individual
atlas:E-GEOD-12040-001
http://www.ebi.ac.uk/biosamples/sample/SAME001
Link to another EBI/RDF dataset

RDF


biosd:exp-prop-val/sub1#6c71
rdfs:label Mus Musculus
dcelems:type Organism
biosd-terms:Samp
biosd-terms:has-biological-characteristic
biosd:exp-prop-val/sub1#97c6
rdfs:label 250 g
Other relationships available:
owl:sameAs
biosd-terms:sample-child-of
(sample-parent-of)
biosd:exp-prop-val/sub1#aa2f
rdfs:label 2.5 percent
sio:derive
prov:hadDe
Hierarchy here is:
has-sample-attribute
has-bio-characteristic
sio:is about
biosd-terms:has-biological-characteristic
biosd:sample/SAME001
rdfs:label Sample 1
sio:derives from
prov:wasDerivedFrom
bio
biosd:exp-prop-val/sub1#2fcd
rdfs:label 31/12/2012
sio:has member
biosd:exp-prop-val/sub1#3a2b
rdfs:label 2-5 weeks
owl:sameAs
biosd:sample-group
rdfs:label Sample

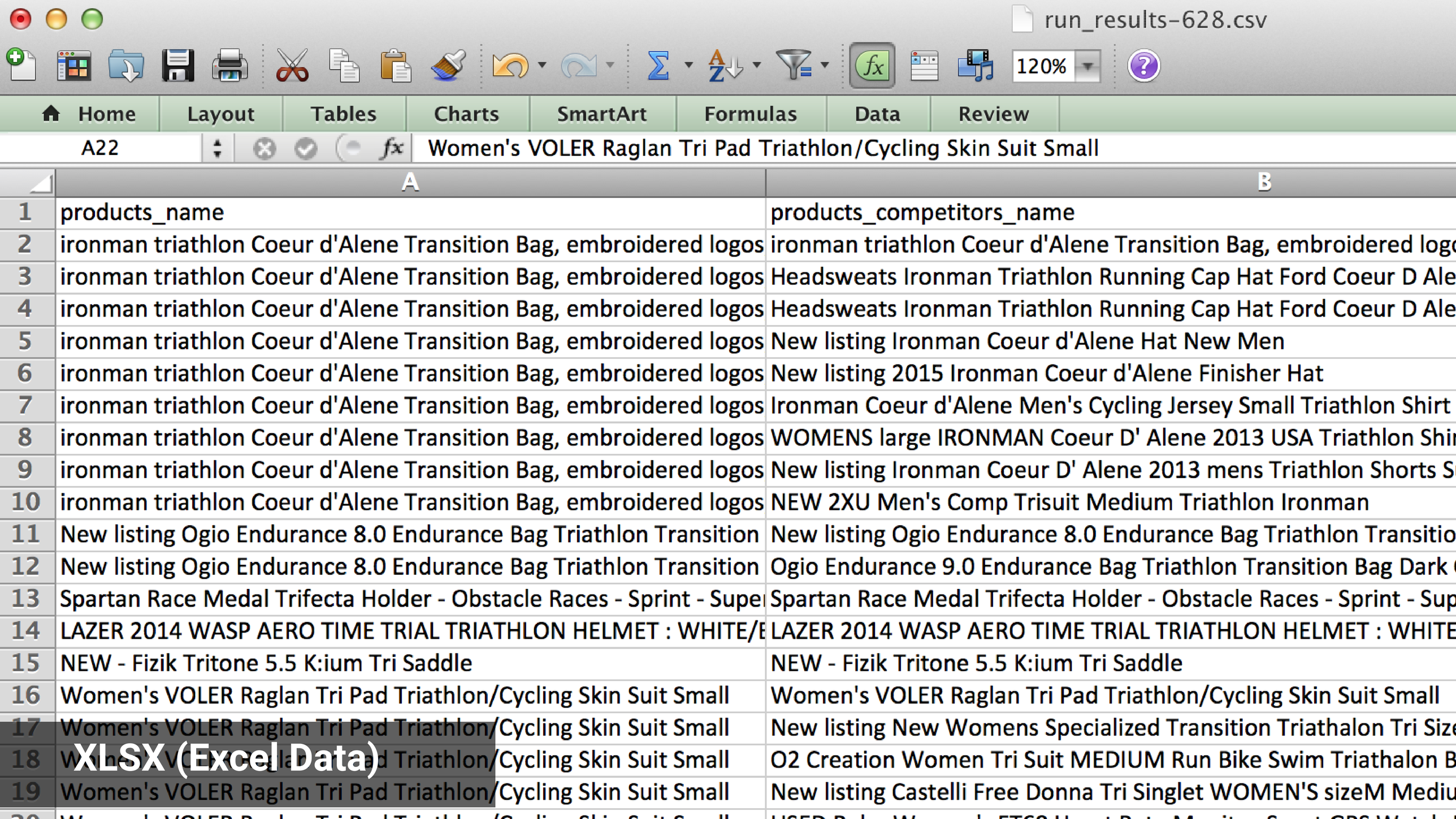
RDF

```
<rdf:RDF
    xmlns:api="http://purl.org/linked-data/api/vocab#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:os="http://a9.com/-/spec/opensearch/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:common="http://landregistry.data.gov.uk/def/common/"
    xmlns:xhv="http://www.w3.org/1999/xhtml/vocab#"
    xmlns:ppi="http://landregistry.data.gov.uk/def/ppi/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <ppi:TransactionRecord rdf:about="http://landregistry.data.gov.uk/data/ppi/transaction/D92BCBED-19E1-4060-916F-88EAB4372520/2014-06-38638">
    <ppi:hasTransaction>
      <ppi:Transaction rdf:about="http://landregistry.data.gov.uk/data/ppi/transaction/D92BCBED-19E1-4060-916F-88EAB4372520">
        <ppi:hasTransactionRecord rdf:resource="http://landregistry.data.gov.uk/data/ppi/transaction/D92BCBED-19E1-4060-916F-88EAB4372520/2014-06
        <ppi:transactionId rdf:datatype="http://landregistry.data.gov.uk/def/ppi/TransactionIdDatatype"
        >D92BCBED-19E1-4060-916F-88EAB4372520</ppi:transactionId>
      </ppi:Transaction>
    </ppi:hasTransaction>
    <ppi:pricePaid rdf:datatype="http://www.w3.org/2001/XMLSchema#integer"
    >232000</ppi:pricePaid>
    <ppi:recordStatus>
      <rdf:Description rdf:about="http://landregistry.data.gov.uk/def/ppi/add">
        <skos:prefLabel xml:lang="en">Add</skos:prefLabel>
        <rdfs:label xml:lang="en">Add</rdfs:label>
      </rdf:Description>
    </ppi:recordStatus>
    <ppi:estateType>
      <rdf:Description rdf:about="http://landregistry.data.gov.uk/def/common/freehold">
        <skos:prefLabel xml:lang="en">Freehold</skos:prefLabel>
        <rdfs:label xml:lang="en">Freehold</rdfs:label>
      </rdf:Description>
    </ppi:estateType>
    <ppi:newBuild rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >false</ppi:newBuild>
    <ppi:propertyAddress>
```

RDF

```
/*!40101 SET @OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET @OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET @OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
/*!40101 SET NAMES utf8 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;


# Dump of table line_seq_dim
# ------------------------------------------------------------

DROP TABLE IF EXISTS `line_seq_dim`;

CREATE TABLE `line_seq_dim` (
  `station_nm` varchar(150) NOT NULL DEFAULT '',
  `nlc` int(5) unsigned NOT NULL,
  `line_nm` varchar(25) NOT NULL DEFAULT '',
  `line_seq_num` int(3) unsigned NOT NULL,
  `branch_nm` varchar(50) NOT NULL DEFAULT '',
  PRIMARY KEY (`nlc`,`line_nm`,`line_seq_num`,`branch_nm`),
  KEY `station_idx` (`station_nm`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8;

LOCK TABLES `line_seq_dim` WRITE;
/*!40000 ALTER TABLE `line_seq_dim` DISABLE KEYS */;

INSERT INTO `line_seq_dim` (`station_nm`, `nlc`, `line_nm`, `line_seq_num`, `branch_nm`)
VALUES
```

SQL Files


run_results-628.csv
120%
Home
Layout
Tables
Charts
SmartArt
Formulas
Data
Review
A22
Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small
A
B
1 products_name products_competitors_name
2 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos ironman triathlon Coeur d'Alene Transition Bag, embroidered logo
3 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos Headsweats Ironman Triathlon Running Cap Hat Ford Coeur D Ale
4 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos Headsweats Ironman Triathlon Running Cap Hat Ford Coeur D Ale
5 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos New listing Ironman Coeur d'Alene Hat New Men
6 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos New listing 2015 Ironman Coeur d'Alene Finisher Hat
7 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos Ironman Coeur d'Alene Men's Cycling Jersey Small Triathlon Shirt
8 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos WOMENS large IRONMAN Coeur D' Alene 2013 USA Triathlon Shi
9 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos New listing Ironman Coeur D' Alene 2013 mens Triathlon Shorts S
10 ironman triathlon Coeur d'Alene Transition Bag, embroidered logos NEW 2XU Men's Comp Trisuit Medium Triathlon Ironman
11 New listing Ogio Endurance 8.0 Endurance Bag Triathlon Transition New listing Ogio Endurance 8.0 Endurance Bag Triathlon Transitio
12 New listing Ogio Endurance 8.0 Endurance Bag Triathlon Transition Ogio Endurance 9.0 Endurance Bag Triathlon Transition Bag Dark
13 Spartan Race Medal Trifecta Holder - Obstacle Races - Sprint - Super Spartan Race Medal Trifecta Holder - Obstacle Races - Sprint - Sup
14 LAZER 2014 WASP AERO TIME TRIAL TRIATHLON HELMET : WHITE/E LAZER 2014 WASP AERO TIME TRIAL TRIATHLON HELMET : WHITE
15 NEW - Fizik Tritone 5.5 K:ium Tri Saddle NEW - Fizik Tritone 5.5 K:ium Tri Saddle
16 Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small
17 Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small New listing New Womens Specialized Transition Triathalon Tri Size
18 Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small O2 Creation Women Tri Suit MEDIUM Run Bike Swim Triathalon B
19 Women's VOLER Raglan Tri Pad Triathlon/Cycling Skin Suit Small New listing Castelli Free Donna Tri Singlet WOMEN'S sizeM Mediu
XLSX (Excel Data)

a. Gene set enrichment analysis

Database	Gene set	Number of genes (mapped to MAGENTA)	95th percentile enrichment cutoff			75th percentile enrichment cutoff		
			P	FDR	Expected (observed) number of genes	*P*	FDR	Expected (observed) number of genes
GOTERM	Positive regulation of glycogen biosynthetic process	10 (10)	5.6×10^{-5}	0.005	1 (5)	3.6×10^{-3}	0.18	3 (7)
GOTERM	Insulin-like growth factor receptor binding	13 (13)	2.4×10^{-5}	0.006	1 (6)	0.02	0.35	3 (7)
GOTERM	Positive regulation of glucose import	22 (22)	1.0×10^{-4}	0.019	1 (7)	0.02	0.36	6 (10)
GOTERM	Insulin receptor signalling pathway	35 (34)	2.8×10^{-5}	0.022	2 (9)	4.3×10^{-3}	0.27	9 (16)
GOTERM	Chromatin remodelling complex	11 (9)	9.0×10^{-4}	0.036	0 (4)	0.16	0.55	2 (4)
KEGG	Glycosphyngolipid biosynthesis globo-series	14 (13)	2.6×10^{-3}	0.037	1 (4)	0.21	0.48	3 (5)
KEGG	Melanoma	71 (67)	1.6×10^{-3}	0.037	3 (10)	0.05	0.35	17 (23)
KEGG	Terpenoid backbone biosynthesis	15 (15)	5.9×10^{-3}	0.039	1 (1)	0.15	0.44	4 (6)
KEGG	Type 2 Diabetes Mellitus	47 (45)	2.2×10^{-3}	0.040	2 (8)	0.14	0.46	11 (15)
Panther	Cholesterol biosynthesis	11 (11)	1.8×10^{-3}	0.040	1 (4)	0.29	0.64	3 (4)
BIOCARTA	Growth hormone pathway	28 (27)	3.0×10^{-4}	0.044	1 (7)	0.11	0.25	7 (10)
KEGG	Oocyte meiosis	114 (108)	1.0×10^{-3}	0.048	5 (14)	0.07	0.45	27 (34)
Custom gene set of imprinted genes								
GTEX	Imprinting genes (All)	77 (72)	1.9×10^{-4}	-	4 (12)	0.11	-	18 (23)
GTEX	Imprinting genes (Primary)	38 (35)	6.9×10^{-3}	-	2 (6)	0.14	-	9 (12)
GTEX	Imprinting genes (Primary + Suggestive)	55 (50)	0.010	-	3 (7)	0.25	-	13 (15)

b. Protein-protein interaction analysis

Database	Pathway	Number of genes (overlapped with PPI network)	z score	*P*	adjusted *P*[a]
GOTERM	epidermal growth factor receptor signalling pathway	198 (31)	7.97	3.3×10^{-10}	1.4×10^{-7}
GOTERM	insulin receptor signalling pathway	151 (26)	7.90	1.1×10^{-9}	2.9×10^{-7}
GOTERM	stimulatory C-type lectin receptor signalling pathway	121 (22)	7.59	7.5×10^{-9}	1.2×10^{-6}
GOTERM	negative regulation of canonical Wnt signalling pathway	152 (25)	7.46	6.2×10^{-9}	1.2×10^{-6}
GOTERM	Notch signalling pathway	129 (22)	7.21	2.6×10^{-8}	3.3×10^{-6}
GOTERM	cellular response to insulin stimulus	71 (16)	7.62	3.7×10^{-8}	4.1×10^{-6}
GOTERM	positive regulation of glycogen biosynthetic process	15 (8)	9.39	5.3×10^{-8}	5.1×10^{-6}
GOTERM	positive regulation of protein phosphorylation	114 (20)	7.03	6.8×10^{-8}	5.9×10^{-6}
GOTERM	positive regulation of glucose import	27 (10)	8.42	8.3×10^{-8}	6.5×10^{-6}
GOTERM	Fc-epsilon receptor signalling pathway	186 (26)	6.58	9.6×10^{-8}	6.8×10^{-6}

[a] *P*-value is adjusted for multiple correction using Benjamini and Hochberg method.

The 2016 Summer Olympic program featured 28 sports with 41 disciplines, and a total of 306 events, tentatively resulting in 306 medal sets to be distributed. Athletes from 87 countries won medals, and 59 of them won at least one gold medal. Both of these categories set new records. 120 countries did not win a medal.

Two gold medals were awarded for a first-place tie in the women's 100 metre freestyle swimming event. No silver medal was awarded as a consequence.

Three silver medals were awarded for a second-place tie in the men's 100 metre butterfly swimming event. No bronze medal was awarded as a consequence.

In boxing (13 disciplines), judo (14), taekwondo (8), and wrestling (18), two bronze medals are awarded in each event (53 additional bronze medals total). Additionally, two bronze medals were awarded for a third-place tie in the women's 100 metre backstroke swimming and in the men's K-1 200 metres canoeing events.

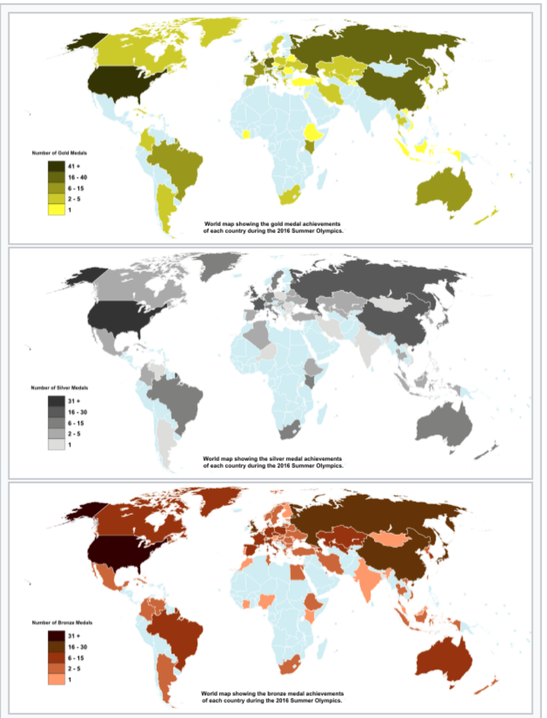
Key

* Host nation (Brazil)

2016 Summer Olympics medal table

Rank	NOC	Gold	Silver	Bronze	Total
1	United States (USA)	46	37	38	121
2	Great Britain (GBR)	27	23	17	67
3	China (CHN)	26	18	26	70
4	Russia (RUS)	19	17	20	56
5	Germany (GER)	17	10	15	42
6	Japan (JPN)	12	8	21	41
7	France (FRA)	10	18	14	42
8	South Korea (KOR)	9	3	9	21
9	Italy (ITA)	8	12	8	28
10	Australia (AUS)	8	11	10	29
11	Netherlands (NED)	8	7	4	19
12	Hungary (HUN)	8	3	4	15
13	Brazil (BRA)*	7	6	6	19
14	Spain (ESP)	7	4	6	17
15	Kenya (KEN)	6	6	1	13
16	Jamaica (JAM)	6	3	2	11
17	Croatia (CRO)	5	3	2	10
18	Cuba (CUB)	5	2	4	11
19	New Zealand (NZL)	4	9	5	18
20	Canada (CAN)	4	3	15	22

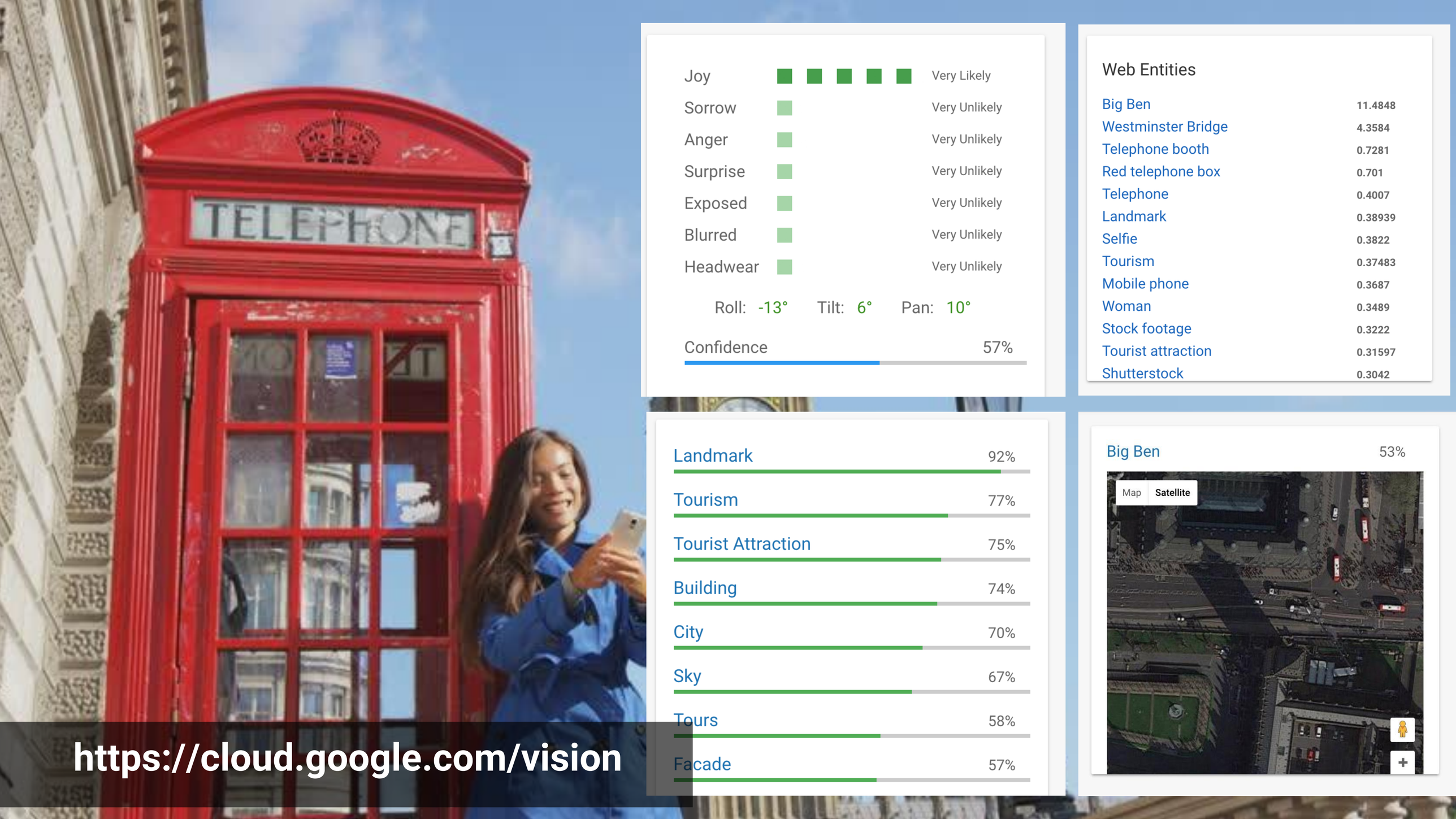
Representative set of the Olympic medals



World maps showing the gold, silver and bronze medal achievements of each country during the 2016 Summer Olympics.

The American final team (Adrian, Held, Phelps, and Dressel), after winning the 4 × 100 m freestyle relay at 20[illegible]

Images (.png, .jpg, .tif etc.)



Joy	Very Likely
Sorrow	Very Unlikely
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely

Roll: -13° Tilt: 6° Pan: 10°

Confidence 57%

Web Entities

Big Ben	11.4848
Westminster Bridge	4.3584
Telephone booth	0.7281
Red telephone box	0.701
Telephone	0.4007
Landmark	0.38939
Selfie	0.3822
Tourism	0.37483
Mobile phone	0.3687
Woman	0.3489
Stock footage	0.3222
Tourist attraction	0.31597
Shutterstock	0.3042

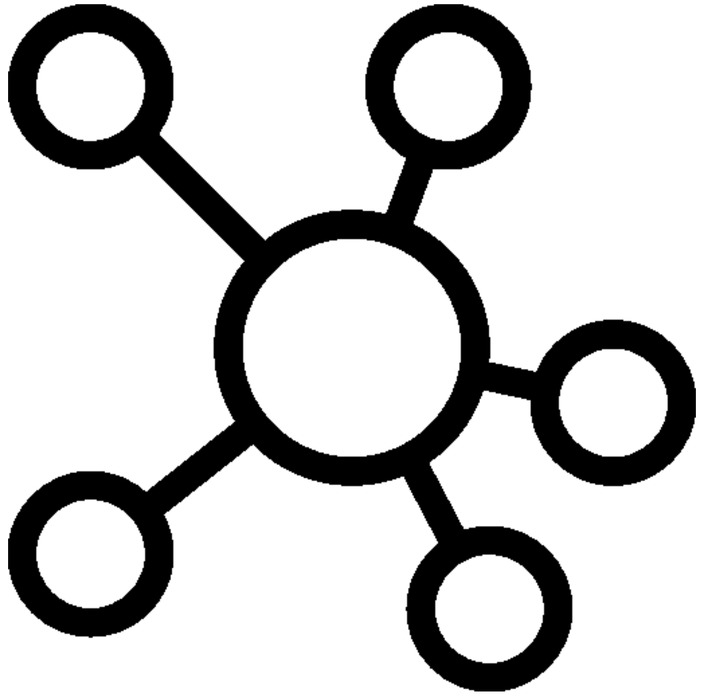
Landmark	92%
Tourism	77%
Tourist Attraction	75%
Building	74%
City	70%
Sky	67%
Tours	58%
Facade	57%

Big Ben 53%

https://cloud.google.com/vision

Where can we get Data From?

Where can we get Data From?



Urban Open Data Portals
  – London – GLA and Transport for London
  – New York – NYC Open Data (Socrata)
  – Seattle – Includes near real-time feeds
- Sectoral, National and International Open Data Initiatives
- Kaggle and KDNuggets
- Flickr, FourSquare, Twitter, Yelp, Airbnb, TripAdvisor, Orange, etc.
- Reddit /r/datasets

    https://www.reddit.com/r/bigquery/wiki/datasets
- WhatDoTheyKnow?
- GitHub lists (most notably this one or this one)
- …or just go and ask someone, a company or organisation
- ...or search more widely, datasets and public APIs are everywhere

What’s an API

Quick Guide

APPLICATION PROGRAMING INTERFACE

Allows applications (and machines) to pass data between each other

Application Programming Interfaces (APIs)

Calls

APIs are access points for interacting with online services

As a client, you'll generally request a service from an API, sending it the parameters and authentication details required to complete the operation

The structure in which the call is made is strict, but always documented

Many useful online APIs available, providing access to highly sophisticated services. However, the service is limited and you'll have to pay to get high volume access

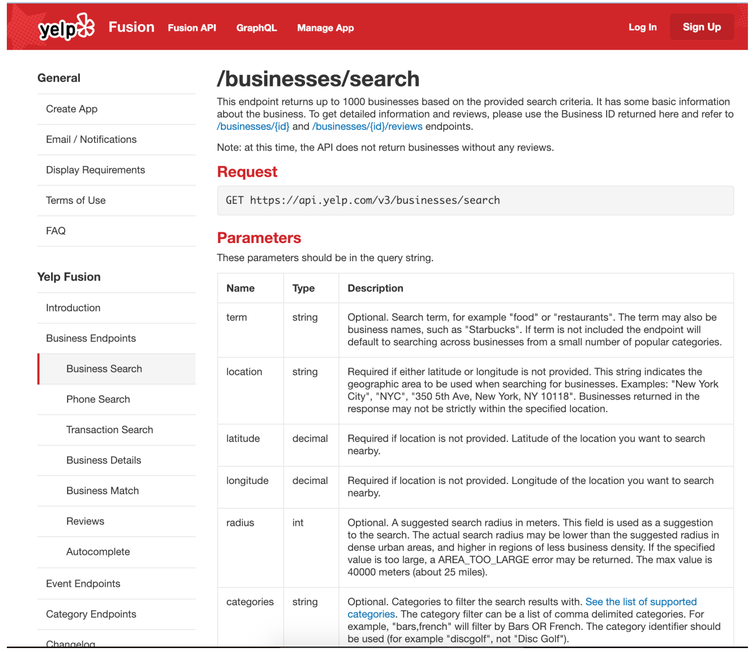
```
https://api.endpoint.com/service-details?
param-name-1=your-parameters-1&
param-name-2=your-parameters-2&
output-format=format&
auth=authentication-key
```

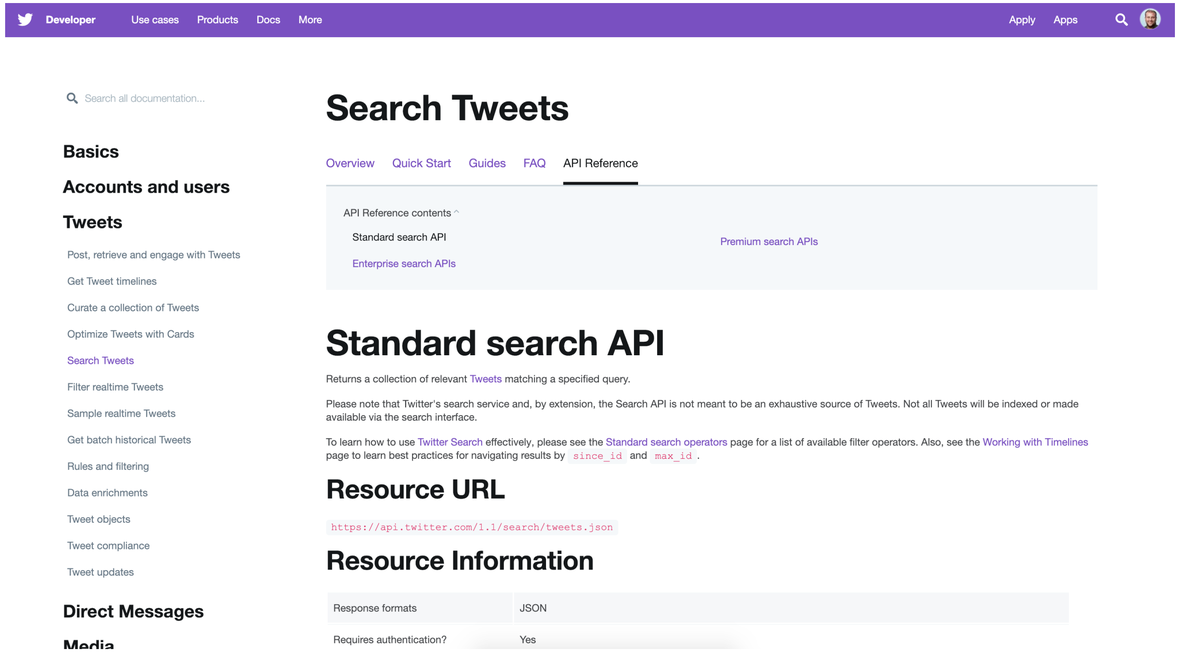
Application Programming Interfaces (APIs)

Handling Responses

An API call will result in a file being returned – typically XML or JSON – and you often get a choice around the data format

The response will follow a documented structure, allowing you to develop code to handle the result automatically





What’s an API

Application Programming Interfaces

Application

Send Data

Pass some variables via a URL

Get the Data back

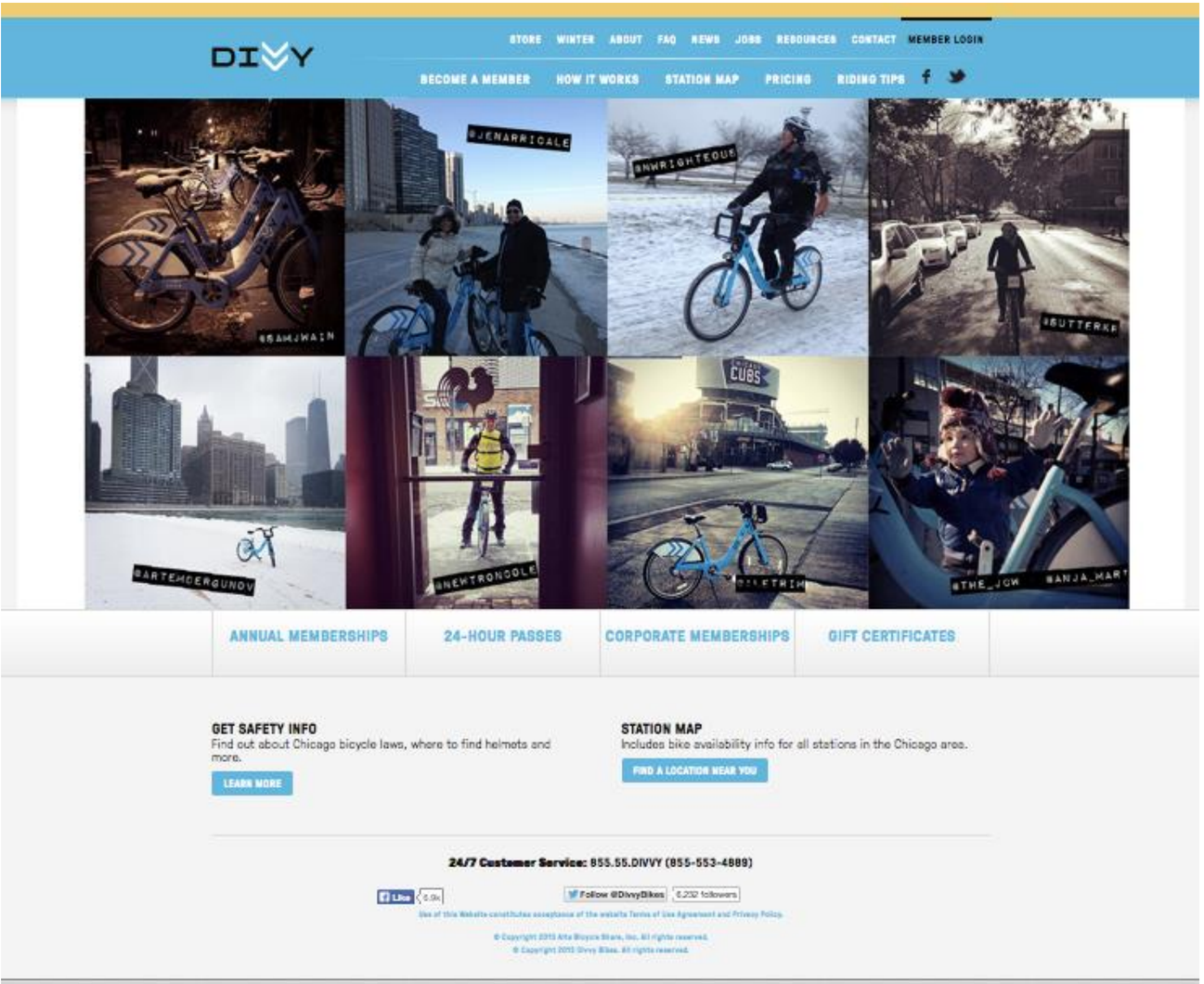
Server send results back

Server

Using API's

Chicago Bike API

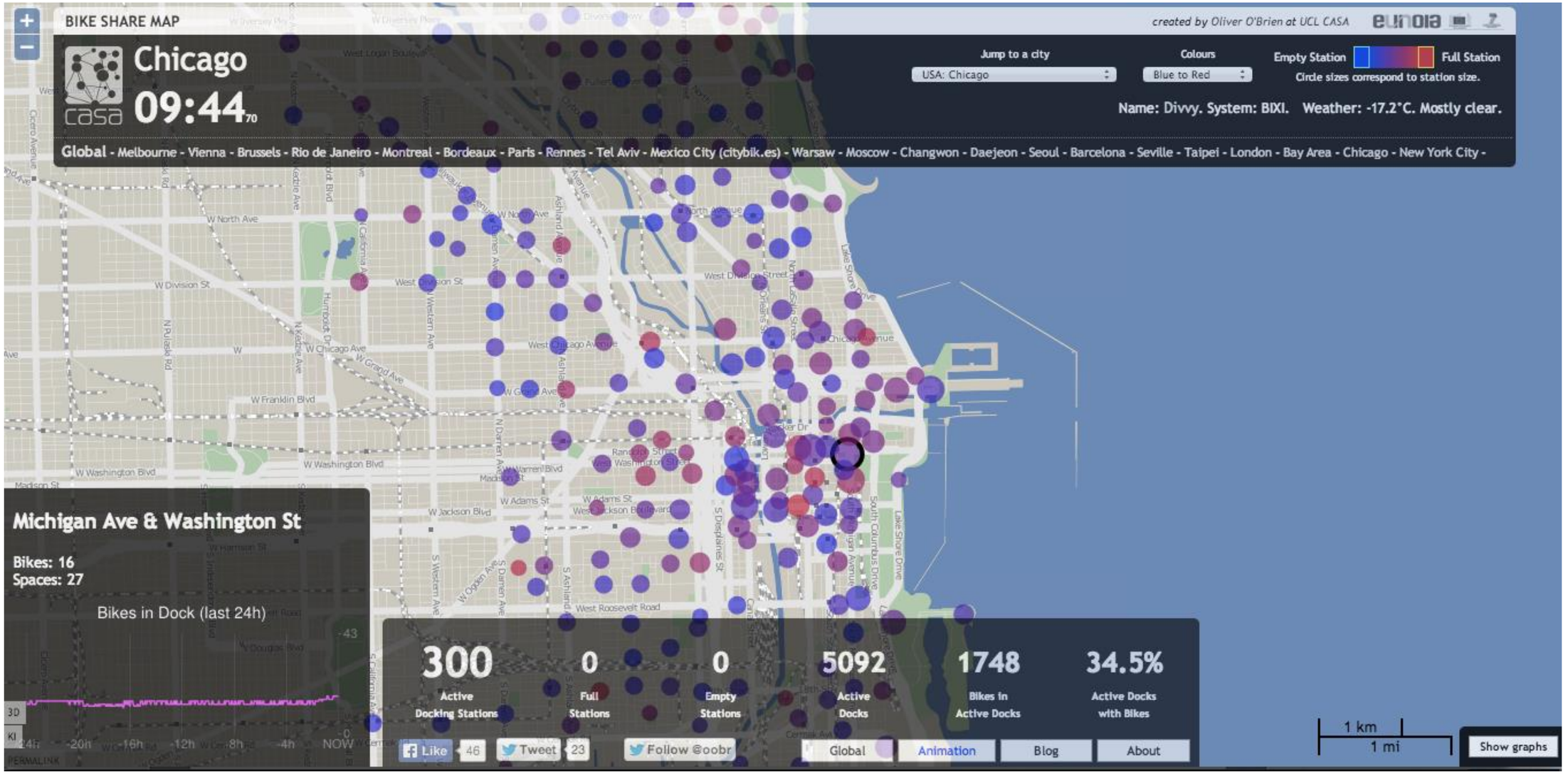
http://divvybikes.com/stations/json




```
{
   executionTime: "2014-02-06 07:44:02 AM",
 - stationBeanList: [
     - {
           id: 5,
           stationName: "State St & Harrison St",
           availableDocks: 17,
           totalDocks: 19,
           latitude: 41.8739580629,
           longitude: -87.6277394859,
           statusValue: "In Service",
           statusKey: 1,
           availableBikes: 2,
           stAddress1: "State St & Harrison St",
           stAddress2: "",
           city: "",
           postalCode: "",
           location: "620 S. State St.",
           altitude: "",
           testStation: false,
           lastCommunicationTime: null,
           landMark: "030"
       },
     - {
           id: 13,
           stationName: "Wilton Ave & Diversey Pkwy",
           availableDocks: 14,
```

Using API’s

Chicago Bike API -> http://bikes.oobrien.com/chicago



Quality and Munging

Ensuring Data Quality, Transforming and Enriching Data

Data Quality

Checking the Data and Trends

Validity of text data

- Alignment with regular expression (e.g. postcodes), or membership within fixed category range.

Validity of numeric data

- Alignment with expected or reasonable mean, standard deviation, etc. Location of data within an acceptable and anticipated range.
- Outlier identification and handling (retaining or removal)
- ‘Eyeball’ it – chart it (and/or map it) – does the distribution align with expectations?

Consistency and reliability

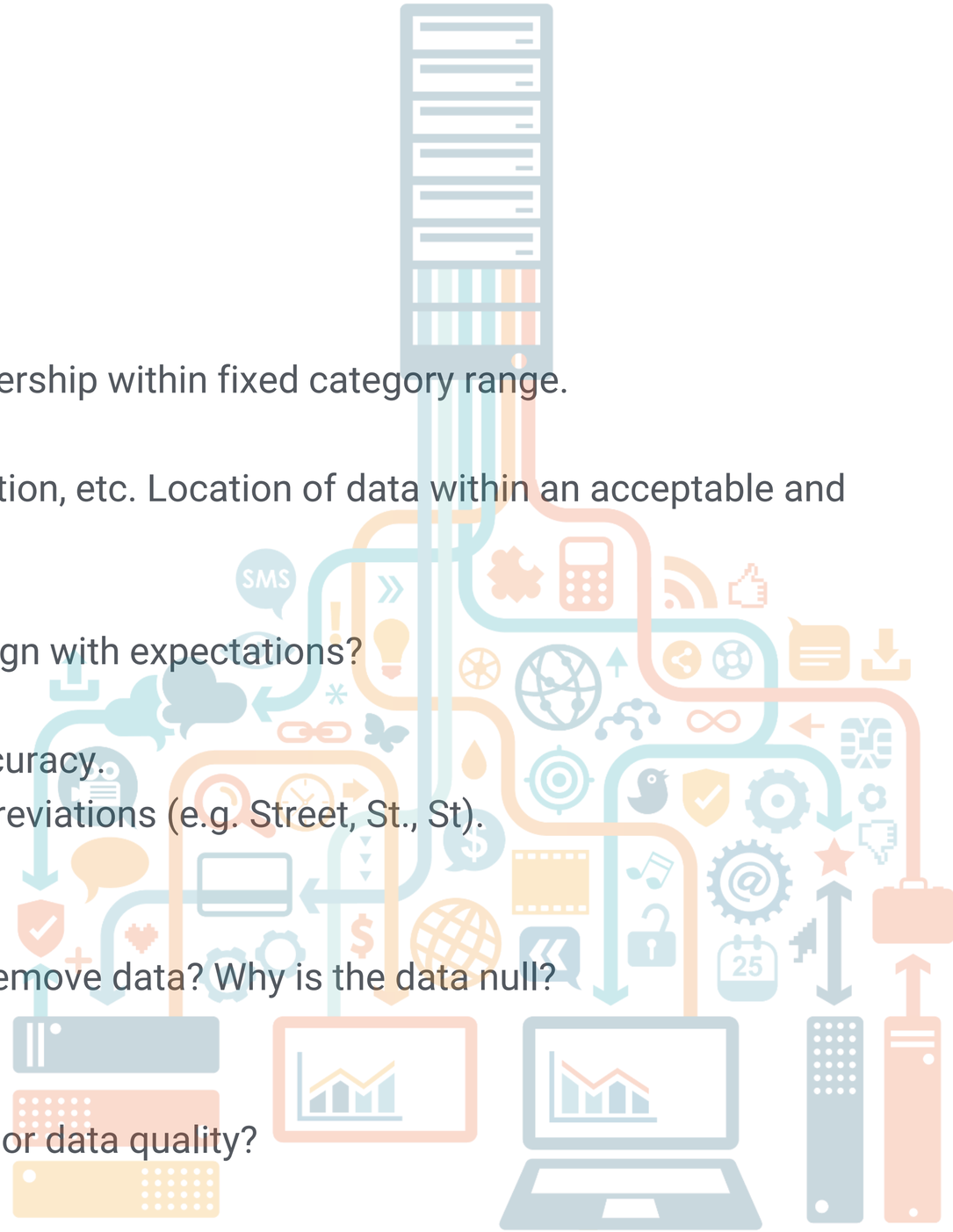
- Consistency in data type, measurement methodology and accuracy.
- Naming conventions and usage of spaces, full stops and abbreviations (e.g. Street, St., St).
- Alignment with other datasets and perspectives.

Completeness

- Handling of NULL, N/A and zero values. Allow null values or remove data? Why is the data null?
- Does zero mean zero?

Duplicate identification and elimination

- What should be considered a duplicate? Does this indicate poor data quality?



‘Mapping’ Data Types

Mapping the raw data to data types in coding

- Always refer to the metadata for data types of attributes
- Be critical of the default types given by Python
- Data types in Pandas and Numpy/Sklearn are not completely compatible

Input	Python types	Pandas types	Numpy Types
NULL, NA, ""	None or numpy.nan	None	numpy.nan
0, 1, 2 ….	int	int64	int_, int8, etc
0.1, 1.2	float	float64	float64, etc
True/False, Y/N, 1/0	bool	bool	bool_
R, G, B, etc.	int or str	int or object	int or string_
'3-FEB-2020', '10/25/20', etc.	datetime module (date, datetime, or time)	datetime64	datetime64

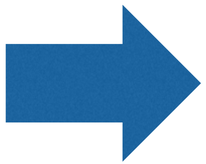
‘Mapping’ Data Types

Categorical data

- Categorical/nominal attributes are saved as ‘string’ in files, but should be mapped to an “int” before analysis (e.g. using scikit-learn)
- Rule of thumb: using one-hot encoding (e.g. sklearn.OneHotEncoder)

CSV

Color
R
G
B
B



Python (numpy.ndarray)

Color_R	Color_G	Color_B
1	0	0
0	1	0
0	0	1
0	0	1

Data Munging

Processes

Pruning and restructuring of data

- Removal of superfluous columns and rows
- Re-classing data types where needed (e.g. string to timestamp, int to boolean)

Replacing text and string values

- Using None and np.nan correctly
- Decide reporting and handling approach for missing values within analysis

Numeric transforms for skewed distributions

- Log transform – converts lognormal distribution to normal
- Reduces skew, enables use of wider set of statistical tools and clearer visualisation

Unstacking and deconstruction of large datasets

- Creating columns based on common attribute values (pivot tables)
- Creating lookup tables for repeated value combinations of data

Creation of categories based on grouped values

- You may need to create one-hot encoding for classification data

Data about Data

Meta Data and Linked Data

Metadata

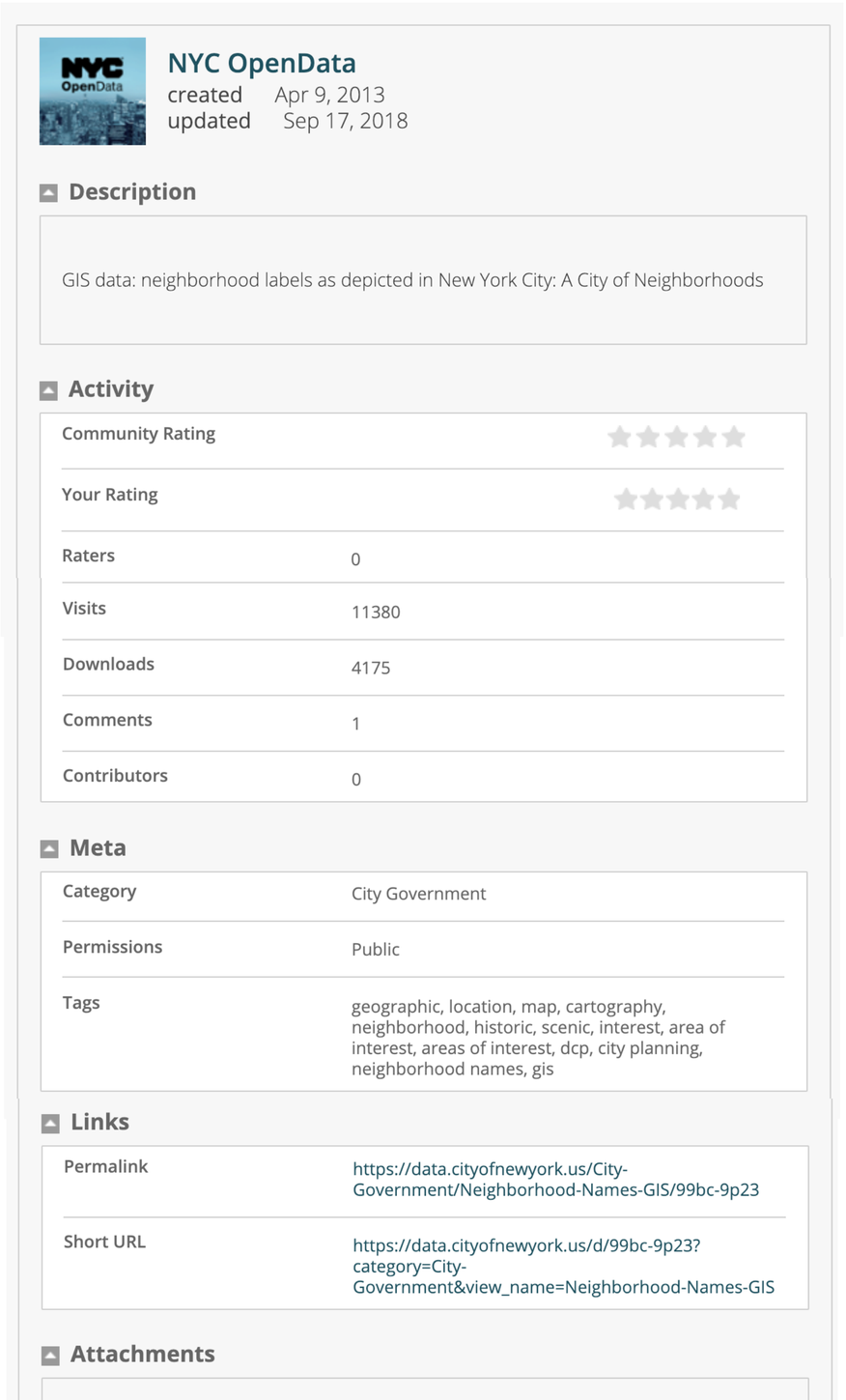
Data about Data

- Provides context around the nature of the data, allowing you to understand the relevance of your findings

- No fixed method or structure, but can include:

- Purpose of collection
- Method of collection
- Post-collection treatment methods
- Date and time extents
- Definition of null values
- Standards for data transfer
- Creator and contact details

- If you're not given it, you can always ask the data provider
  Don't Assume a Data Format!



Linked Data

Data linked to Data

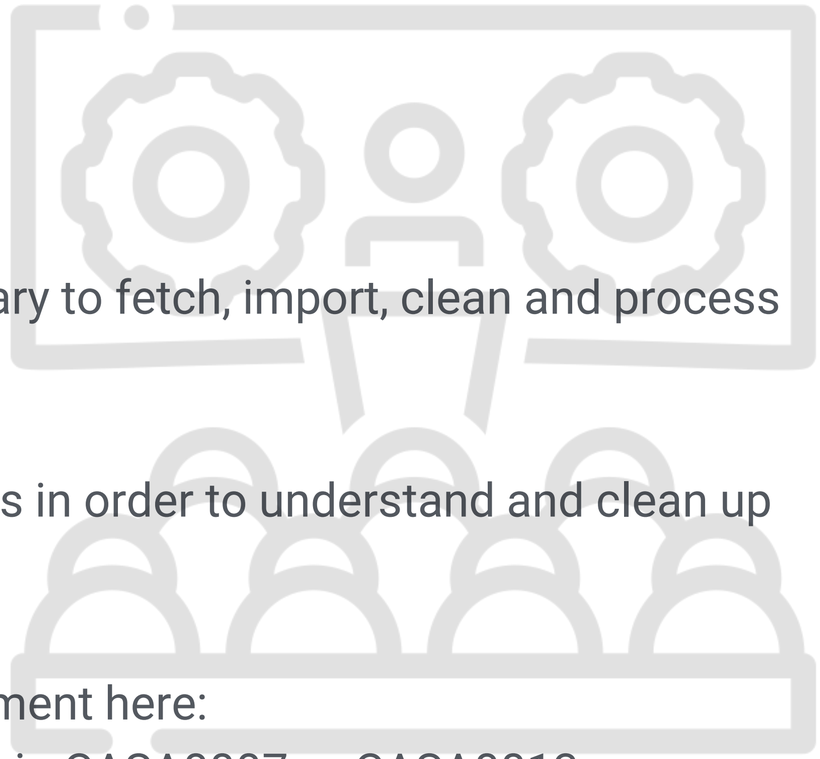
- Incorporates specific links between data, usually within a certain hierarchy or network
- Provides more context about data, and enables machine interpretation
- Part of the semantic web movement
- Potentially very useful in GIS applications and datasets

North Somerset is a Westminster Constituency.

Objects related to "North Somerset"	
Extent	24948-1
In European Region	South West
Within	South West
Touches	Bristol North West Bristol South North East Somerset Bristol West Weston-Super-Mare
Same As	E14000850
referenced via **Westminster Constituency** by	South West
referenced via **Contains** by	South West
referenced via **Primary Topic** by	Description of http://data.ordnancesurvey.co.uk/id/7000000000024948

Workshop

API and Data Munging



- The workshop will focus on using Python 3 and the Pandas library to fetch, import, clean and process a raw dataset

- You'll apply different methods and functions provided by Pandas in order to understand and clean up the data

- Before the workshop, please install the SDS computing environment here: https://github.com/jreades/sds_env/, which is the same one as in CASA0007 or CASA0013.

- Download this week's Jupyter Notebook from Moodle

Next Week

More Data Processing

5 Advanced Clustering

Questions?