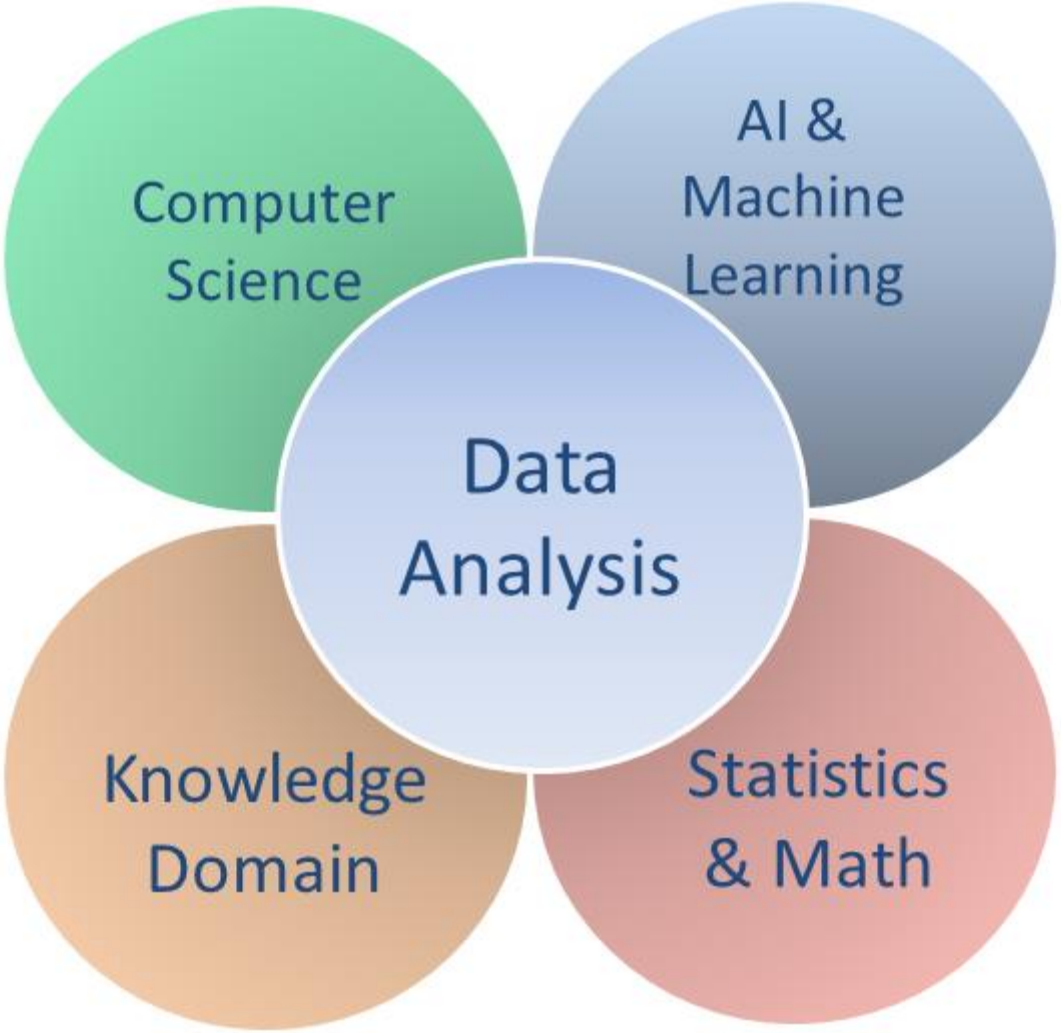
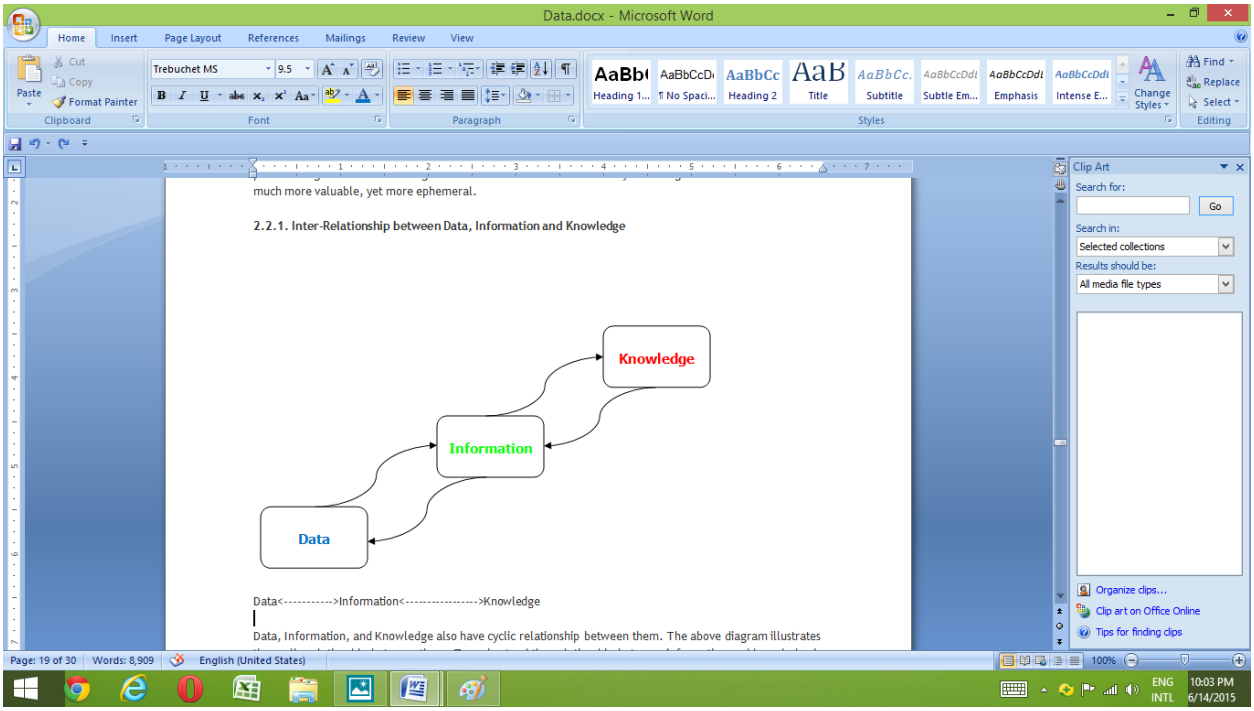
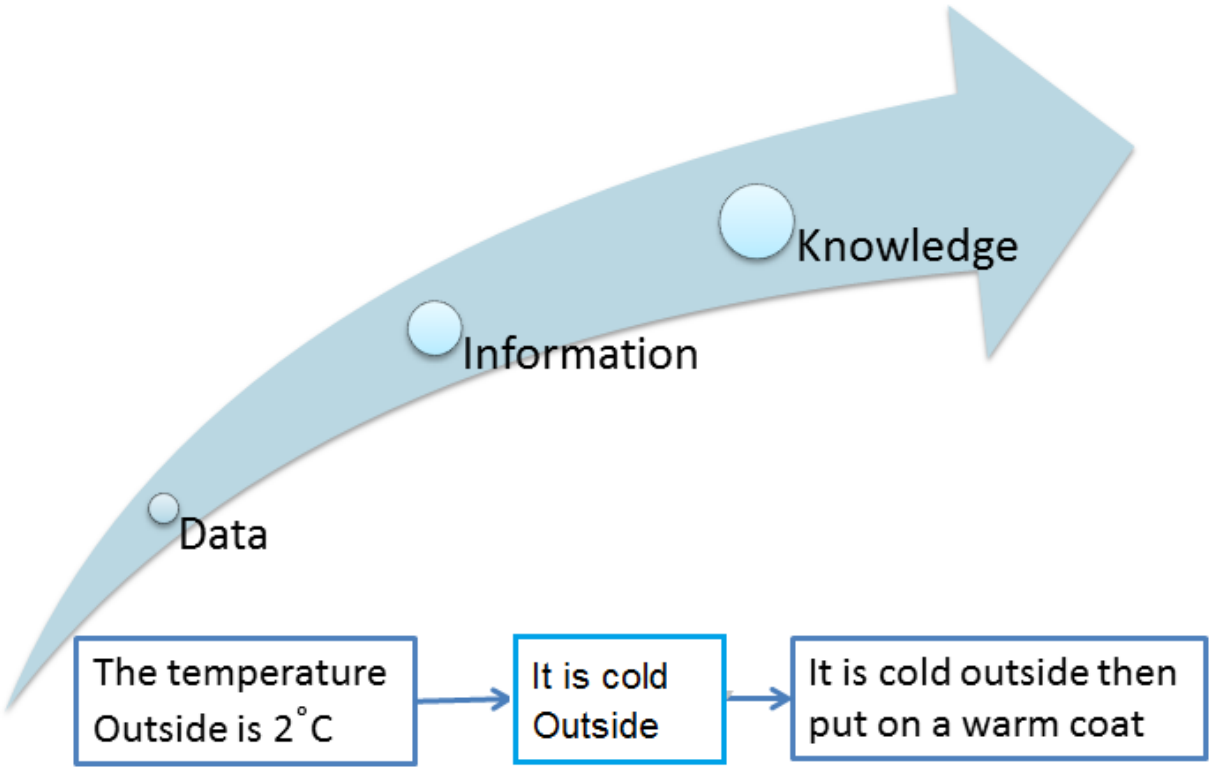
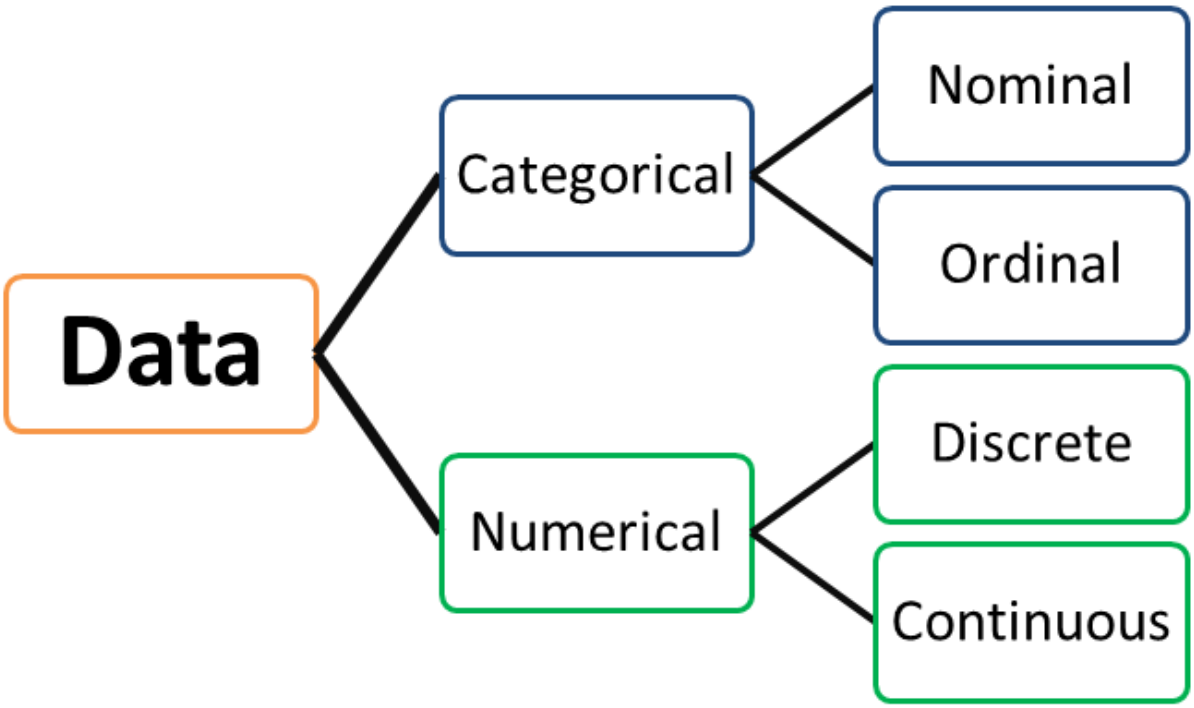
Chapter 1: Getting Started

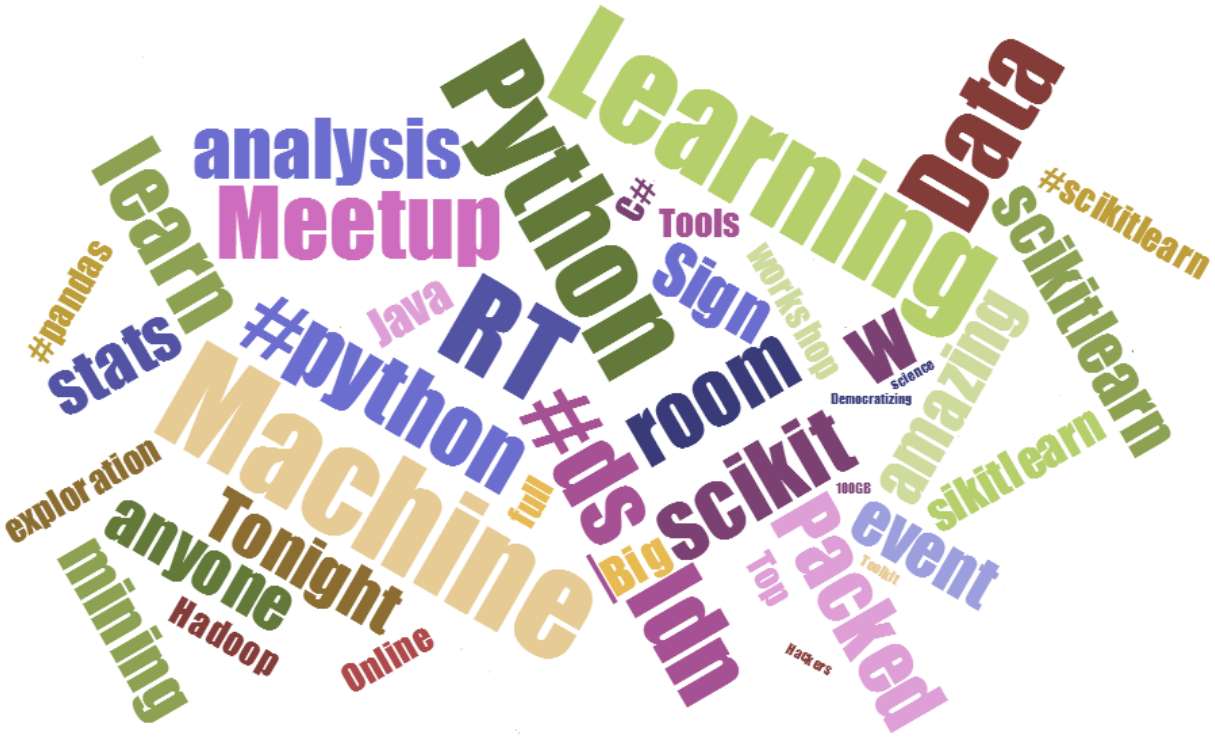


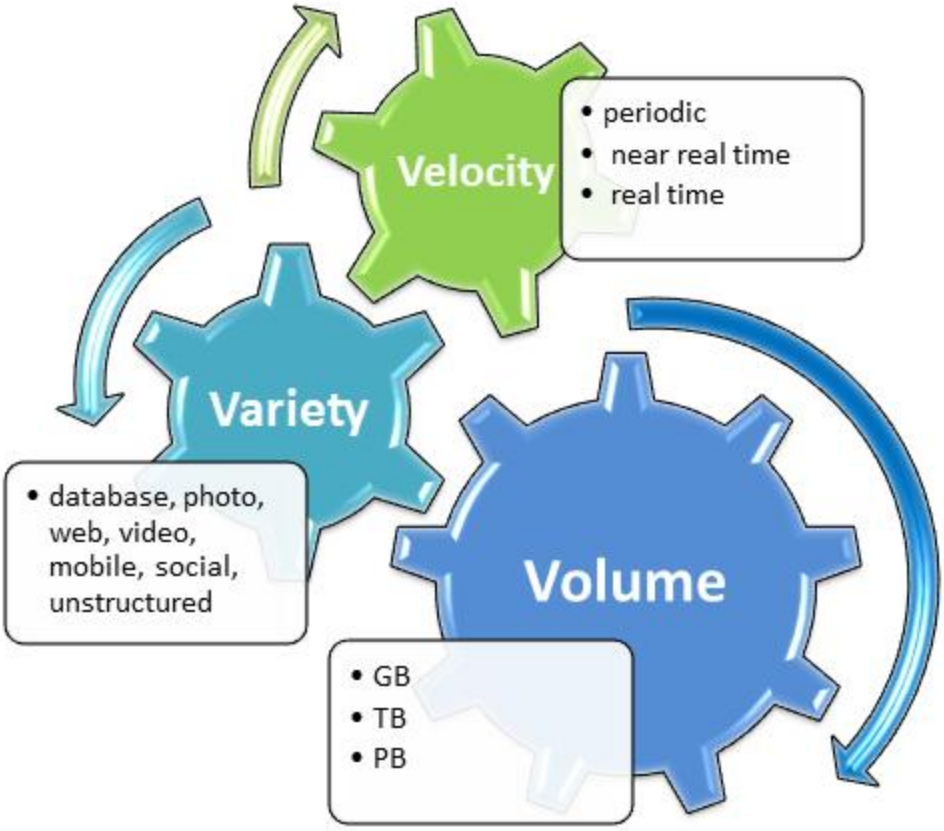

Data.docx - Microsoft Word
much more valuable, yet more ephemeral.
2.2.1. Inter-Relationship between Data, Information and Knowledge
Knowledge
Information
Data
Data<---------->Information<---------->Knowledge
Data, Information, and Knowledge also have cyclic relationship between them. The above diagram illustrates


Knowledge
Information
Data
The temperature Outside is 2°C
It is cold Outside
It is cold outside then put on a warm coat

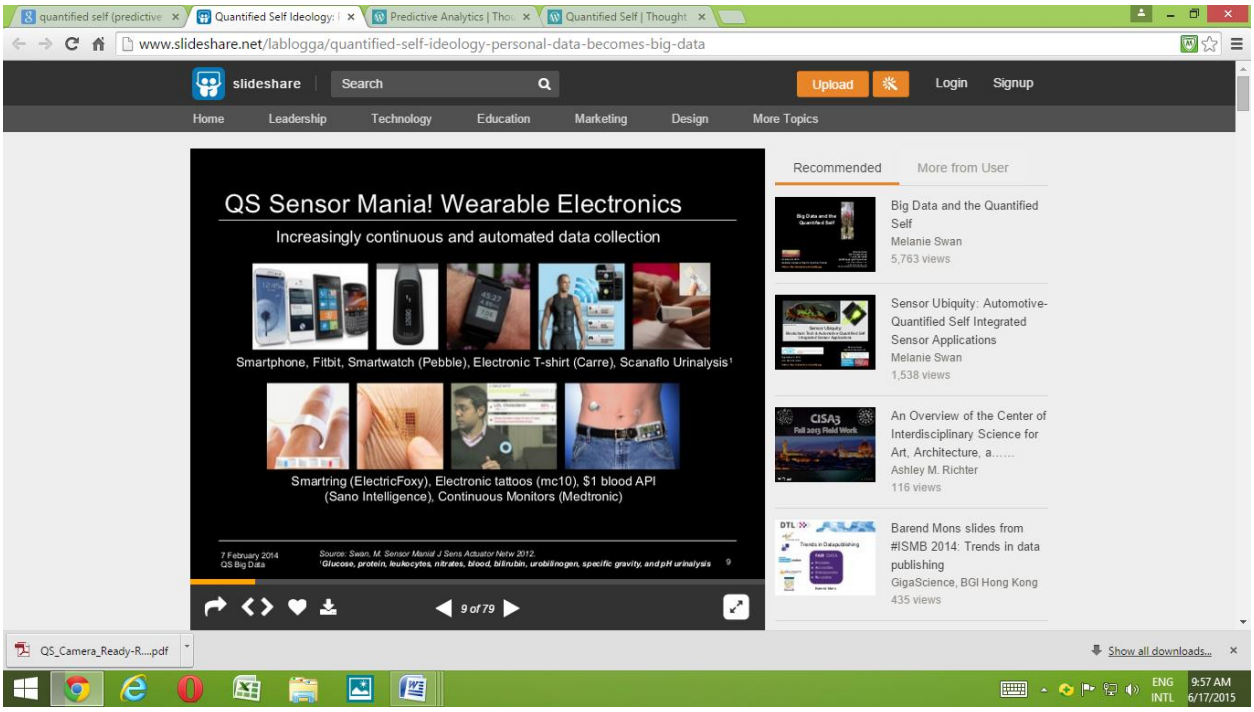


Quantitative		Qualitative
Structured Data Statistical analysis Objective conclusions	VS.	Unstructured data Summary Subjective conclusions

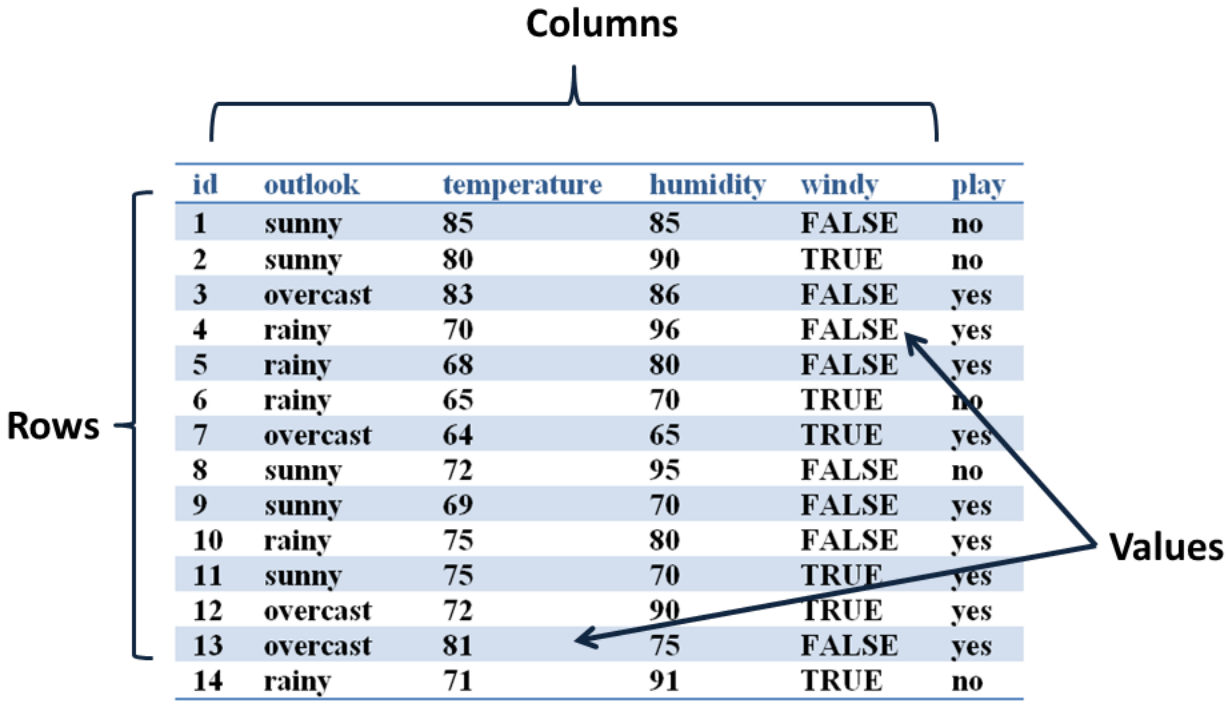

Learning
Data
Python
analysis
Meetup
learn
#scikitlearn
Scikitlearn
C#
Tools
#pandas
Java
RT
Sign
workshop
W
amazing
Stats
#python
room
Machine
#ds
scikit
sikitlearn
exploration
full
Tonight
anyone
Big
Idn
Packed
event
mining
Hadoop
Online
Top
Hackers


Velocity
• periodic
• near real time
• real time
Variety
• database, photo, web, video, mobile, social, unstructured
Volume
• GB
• TB
• PB

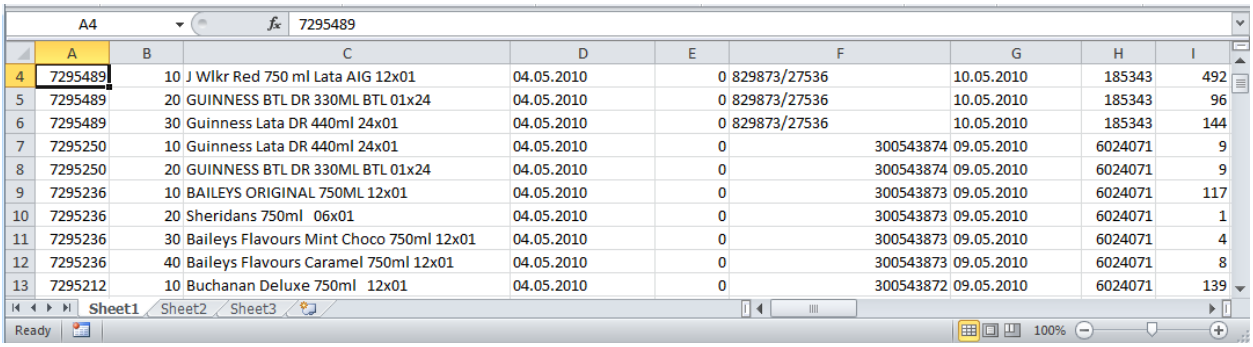
Better Algorithms **Vs.** **More Data** **Vs.** **Better Data**



Chapter 2: Preprocessing the Data



id	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no



A4 | 7295489

	A	B	C	D	E	F	G	H	I
4	7295489	10	J Wlkr Red 750 ml Lata AIG 12x01	04.05.2010	0	829873/27536	10.05.2010	185343	492
5	7295489	20	GUINNESS BTL DR 330ML BTL 01x24	04.05.2010	0	829873/27536	10.05.2010	185343	96
6	7295489	30	Guinness Lata DR 440ml 24x01	04.05.2010	0	829873/27536	10.05.2010	185343	144
7	7295250	10	Guinness Lata DR 440ml 24x01	04.05.2010	0	300543874	09.05.2010	6024071	9
8	7295250	20	GUINNESS BTL DR 330ML BTL 01x24	04.05.2010	0	300543874	09.05.2010	6024071	9
9	7295236	10	BAILEYS ORIGINAL 750ML 12x01	04.05.2010	0	300543873	09.05.2010	6024071	117
10	7295236	20	Sheridans 750ml 06x01	04.05.2010	0	300543873	09.05.2010	6024071	1
11	7295236	30	Baileys Flavours Mint Choco 750ml 12x01	04.05.2010	0	300543873	09.05.2010	6024071	4
12	7295236	40	Baileys Flavours Caramel 750ml 12x01	04.05.2010	0	300543873	09.05.2010	6024071	8
13	7295212	10	Buchanan Deluxe 750ml 12x01	04.05.2010	0	300543872	09.05.2010	6024071	139

Sheet1 Sheet2 Sheet3

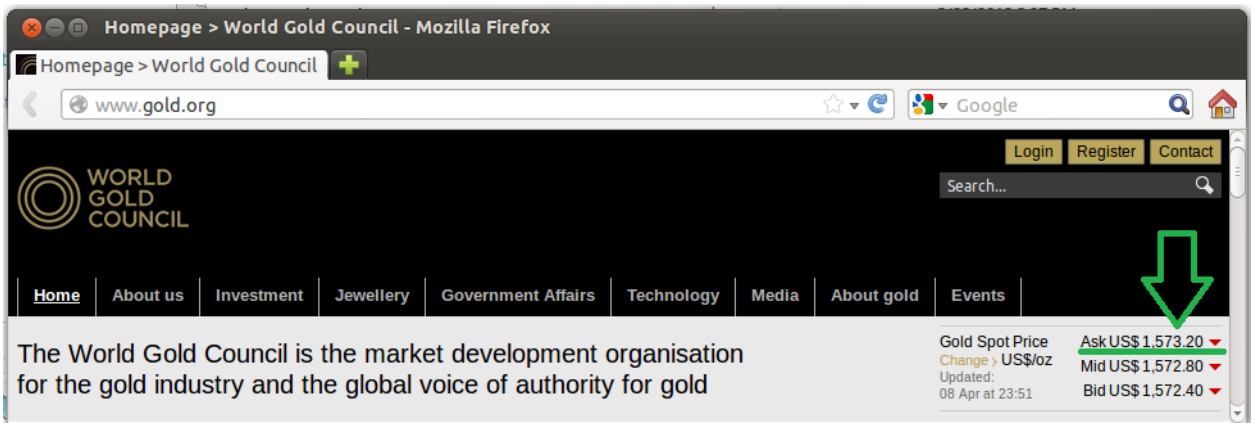
Ready 100%

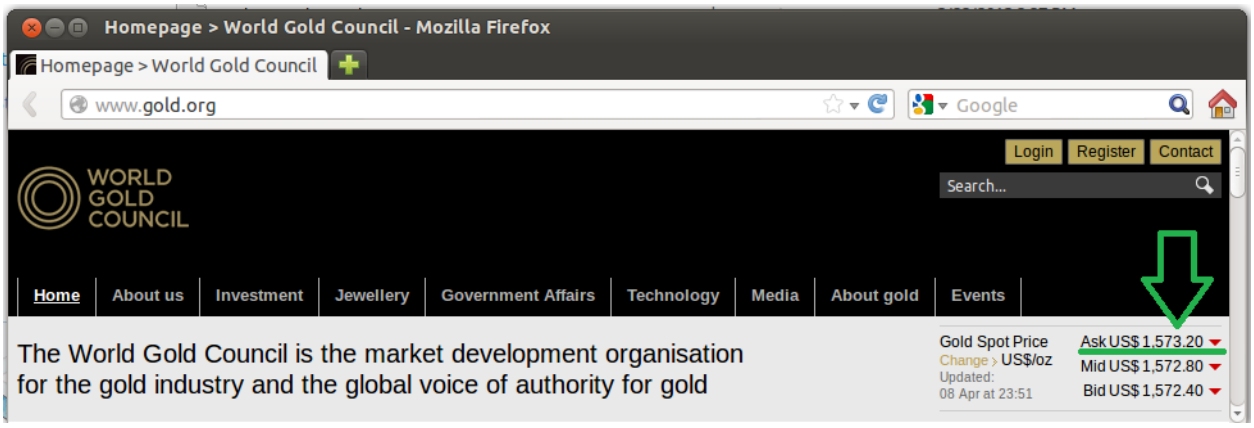
SQLQuery2.sql - LEVIATHAN-PC\....))*

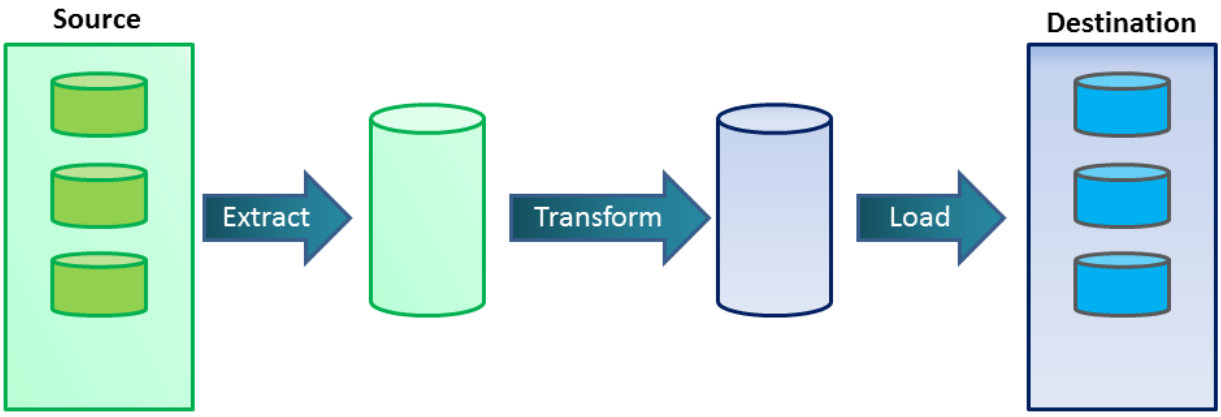
```
SELECT * FROM [test].[dbo].[weather]
```

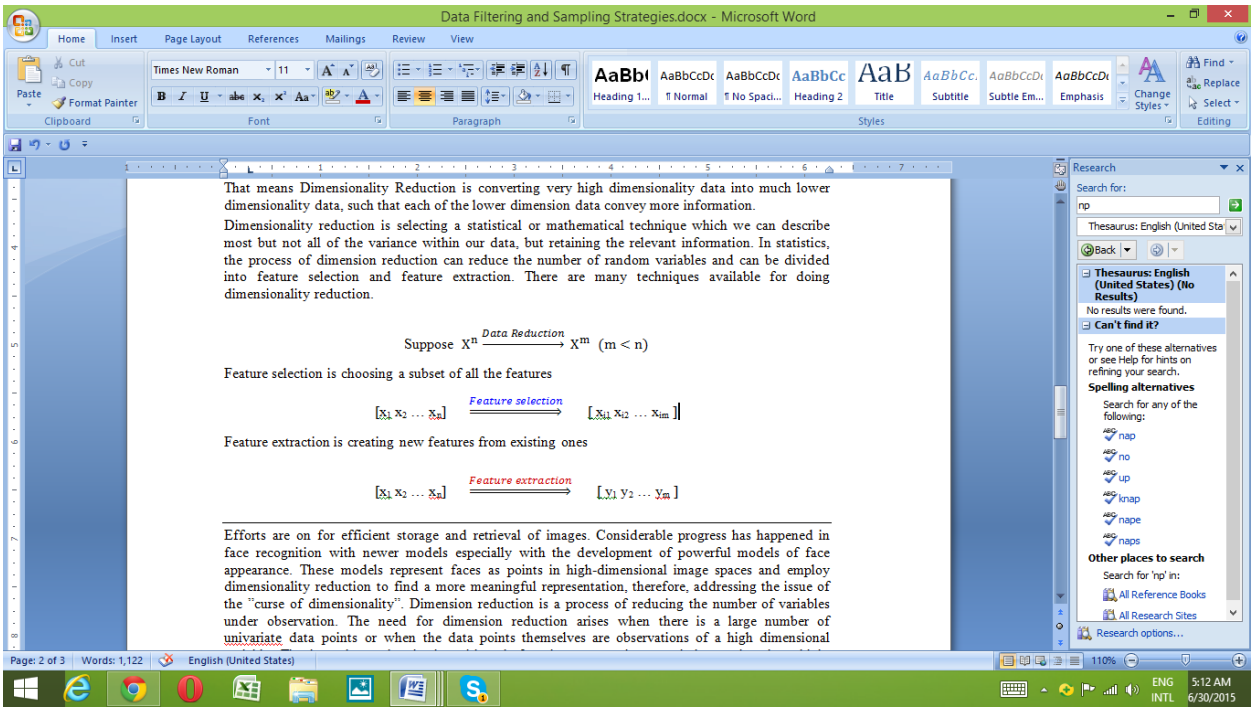
Results | Messages

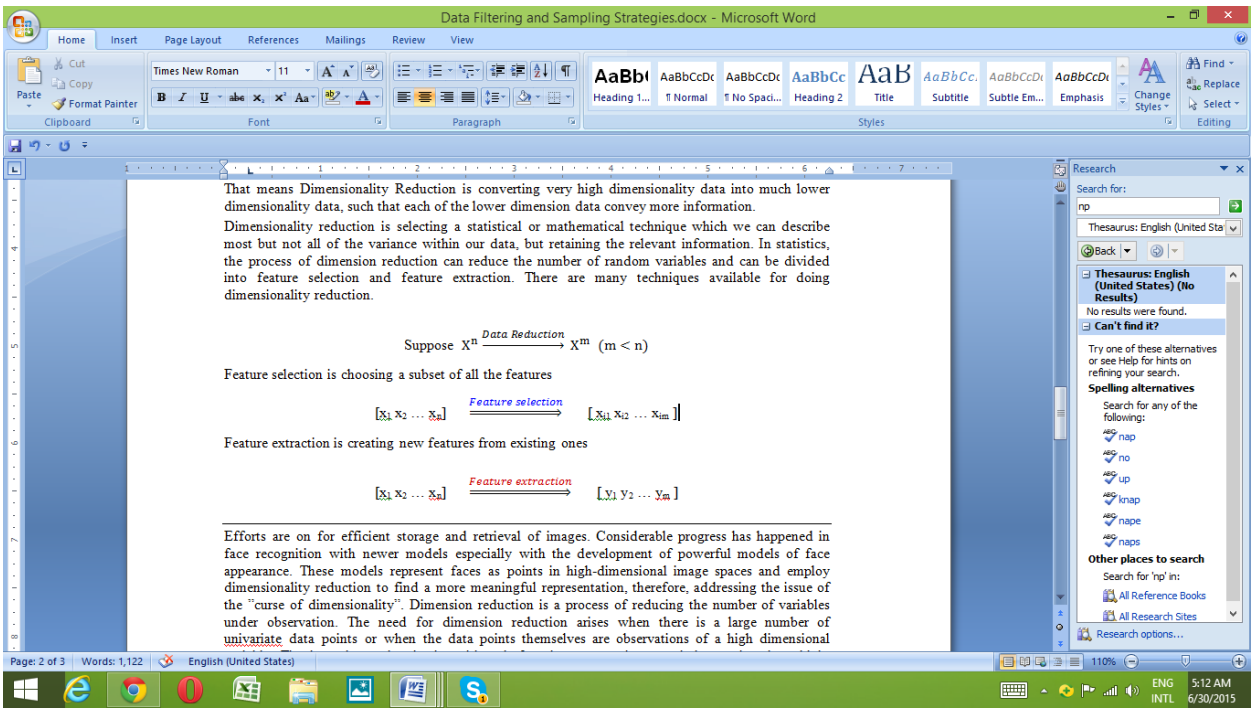
	id	outlook	temperature	humidity	windy	play
1	1	sunny	85	85	FALSE	no
2	2	sunny	80	90	TRUE	no
3	3	overcast	83	86	FALSE	yes
4	4	rainy	70	96	FALSE	yes
5	5	rainy	68	80	FALSE	yes
6	6	rainy	65	70	TRUE	no
7	7	overcast	64	65	TRUE	yes
8	8	sunny	72	95	FALSE	no
9	9	sunny	69	70	FALSE	yes
10	1..	rainy	75	80	FALSE	yes
11	1..	sunny	75	70	TRUE	yes
12	1..	overcast	72	90	TRUE	yes
13	1..	overcast	81	75	FALSE	yes
14	1..	rainy	71	91	TRUE	no

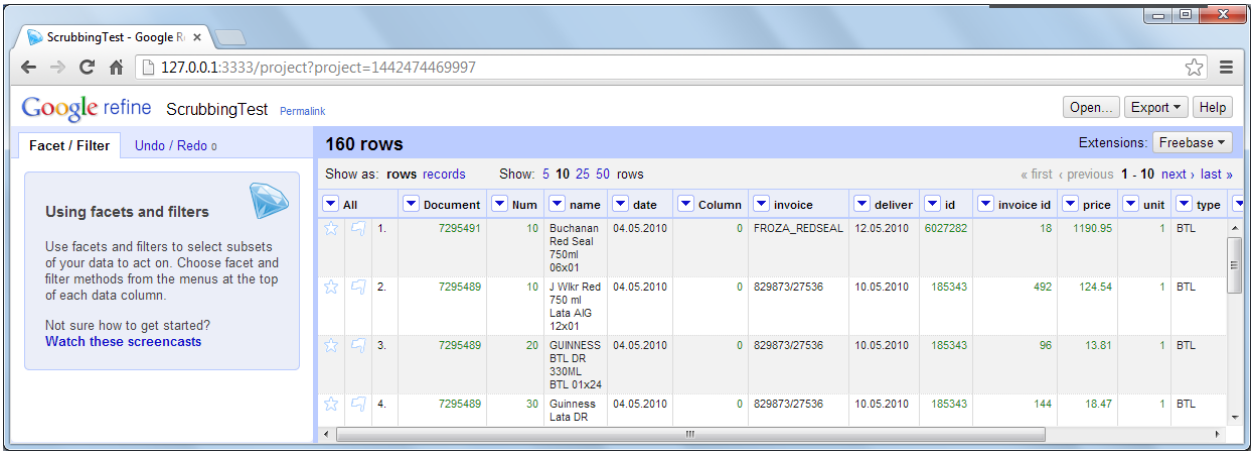



Homepage > World Gold Council - Mozilla Firefox
Homepage > World Gold Council
www.gold.org
Google
WORLD GOLD COUNCIL
Login
Register
Contact
Search...
Home
About us
Investment
Jewellery
Government Affairs
Technology
Media
About gold
Events
The World Gold Council is the market development organisation for the gold industry and the global voice of authority for gold
Gold Spot Price
Change › US$/oz
Updated:
08 Apr at 23:51
Ask US$ 1,573.20
Mid US$ 1,572.80
Bid US$ 1,572.40

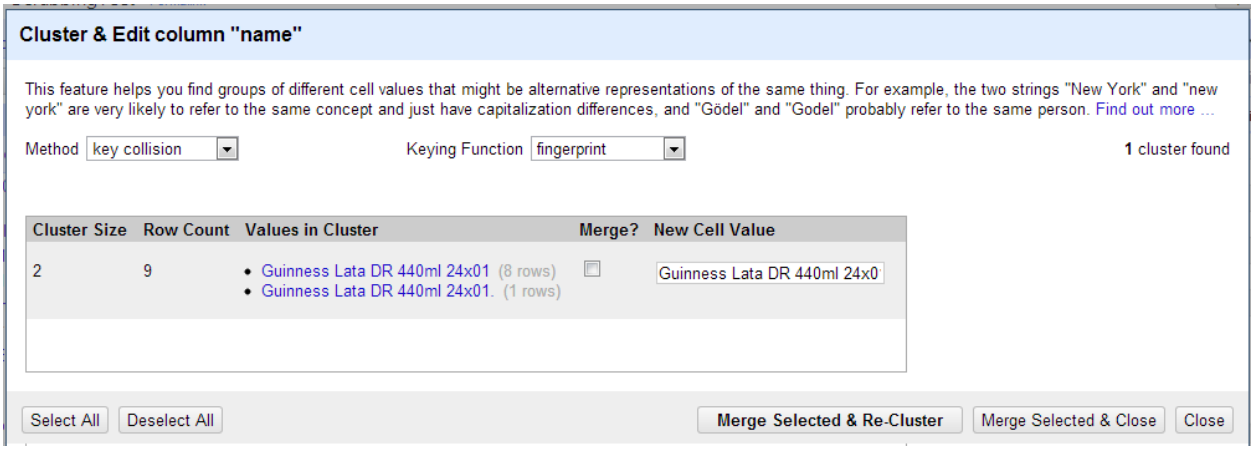

Source
Extract
Transform
Load
Destination

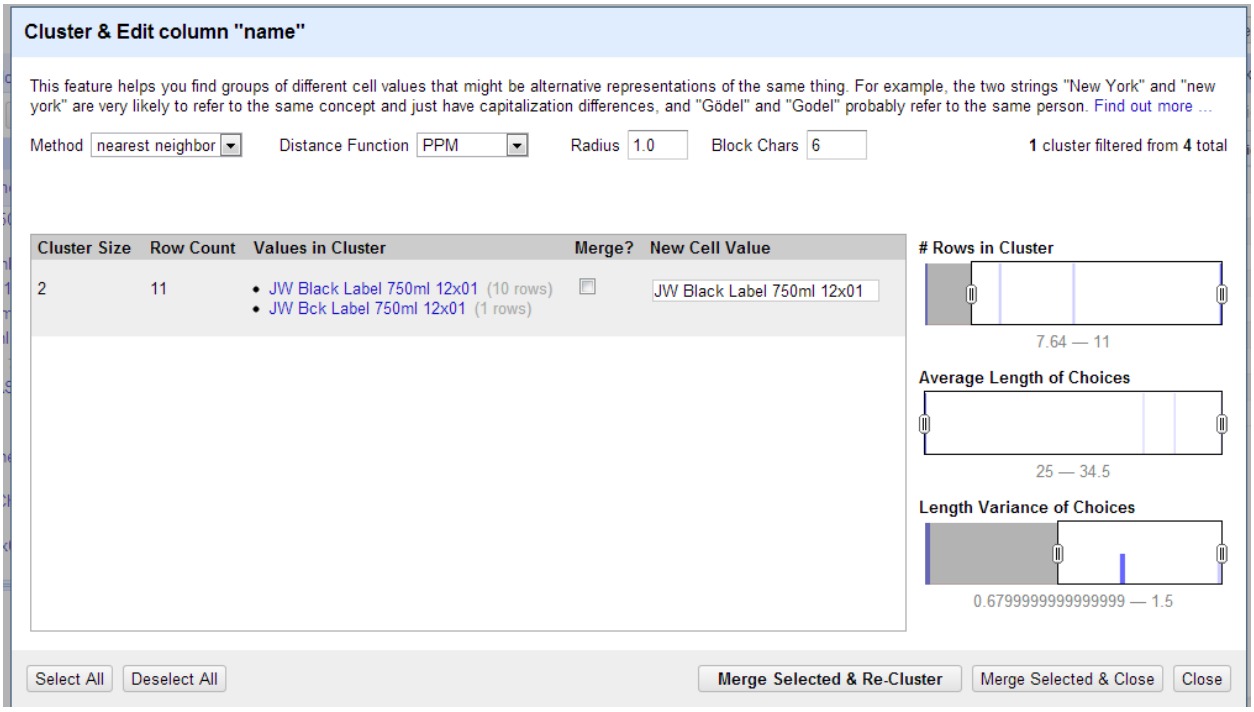

Data Filtering and Sampling Strategies.docx - Microsoft Word
That means Dimensionality Reduction is converting very high dimensionality data into much lower dimensionality data, such that each of the lower dimension data convey more information.
Dimensionality reduction is selecting a statistical or mathematical technique which we can describe most but not all of the variance within our data, but retaining the relevant information. In statistics, the process of dimension reduction can reduce the number of random variables and can be divided into feature selection and feature extraction. There are many techniques available for doing dimensionality reduction.
Suppose $X^n \xrightarrow{Data\ Reduction} X^m \quad (m < n)$
Feature selection is choosing a subset of all the features
$[x_1\ x_2 \dots x_n] \xrightarrow{Feature\ selection} [x_{i1}\ x_{i2} \dots x_{im}]$
Feature extraction is creating new features from existing ones
$[x_1\ x_2 \dots x_n] \xrightarrow{Feature\ extraction} [y_1\ y_2 \dots y_m]$
Efforts are on for efficient storage and retrieval of images. Considerable progress has happened in face recognition with newer models especially with the development of powerful models of face appearance. These models represent faces as points in high-dimensional image spaces and employ dimensionality reduction to find a more meaningful representation, therefore, addressing the issue of the "curse of dimensionality". Dimension reduction is a process of reducing the number of variables under observation. The need for dimension reduction arises when there is a large number of univariate data points or when the data points themselves are observations of a high dimensional
Research
np
Thesaurus: English (United States) (No Results)
No results were found.
Can't find it?
Spelling alternatives
nap
no
up
knap
nape
naps
Other places to search
All Reference Books
All Research Sites
Research options...
Page: 2 of 3
Words: 1,122
English (United States)
110%
5:12 AM
6/30/2015

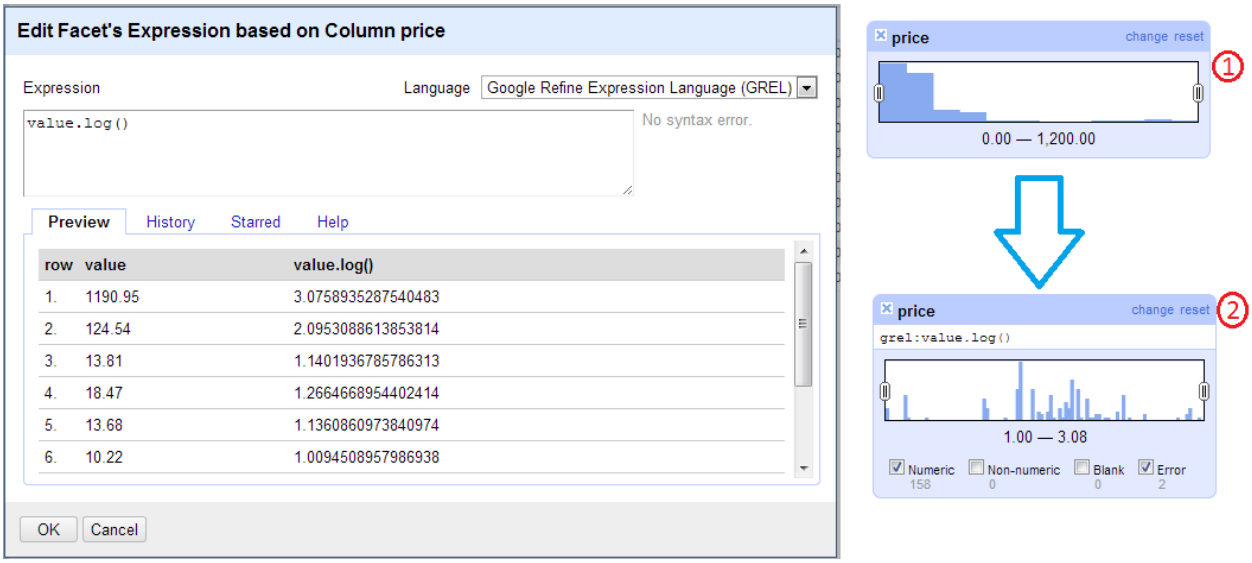

Data Filtering and Sampling Strategies.docx - Microsoft Word
That means Dimensionality Reduction is converting very high dimensionality data into much lower dimensionality data, such that each of the lower dimension data convey more information.
Dimensionality reduction is selecting a statistical or mathematical technique which we can describe most but not all of the variance within our data, but retaining the relevant information. In statistics, the process of dimension reduction can reduce the number of random variables and can be divided into feature selection and feature extraction. There are many techniques available for doing dimensionality reduction.
Suppose $X^n \xrightarrow{Data\ Reduction} X^m \quad (m < n)$
Feature selection is choosing a subset of all the features
$[x_1\ x_2 \dots x_n] \xrightarrow{Feature\ selection} [x_{i1}\ x_{i2} \dots x_{im}]$
Feature extraction is creating new features from existing ones
$[x_1\ x_2 \dots x_n] \xrightarrow{Feature\ extraction} [y_1\ y_2 \dots y_m]$
Efforts are on for efficient storage and retrieval of images. Considerable progress has happened in face recognition with newer models especially with the development of powerful models of face appearance. These models represent faces as points in high-dimensional image spaces and employ dimensionality reduction to find a more meaningful representation, therefore, addressing the issue of the "curse of dimensionality". Dimension reduction is a process of reducing the number of variables under observation. The need for dimension reduction arises when there is a large number of univariate data points or when the data points themselves are observations of a high dimensional
Research
np
Thesaurus: English (United States) (No Results)
No results were found.
Can't find it?
Spelling alternatives
nap
no
up
knap
nape
naps
Other places to search
All Reference Books
All Research Sites
Research options...
Page: 2 of 3
Words: 1,122
English (United States)
110%
5:12 AM
6/30/2015

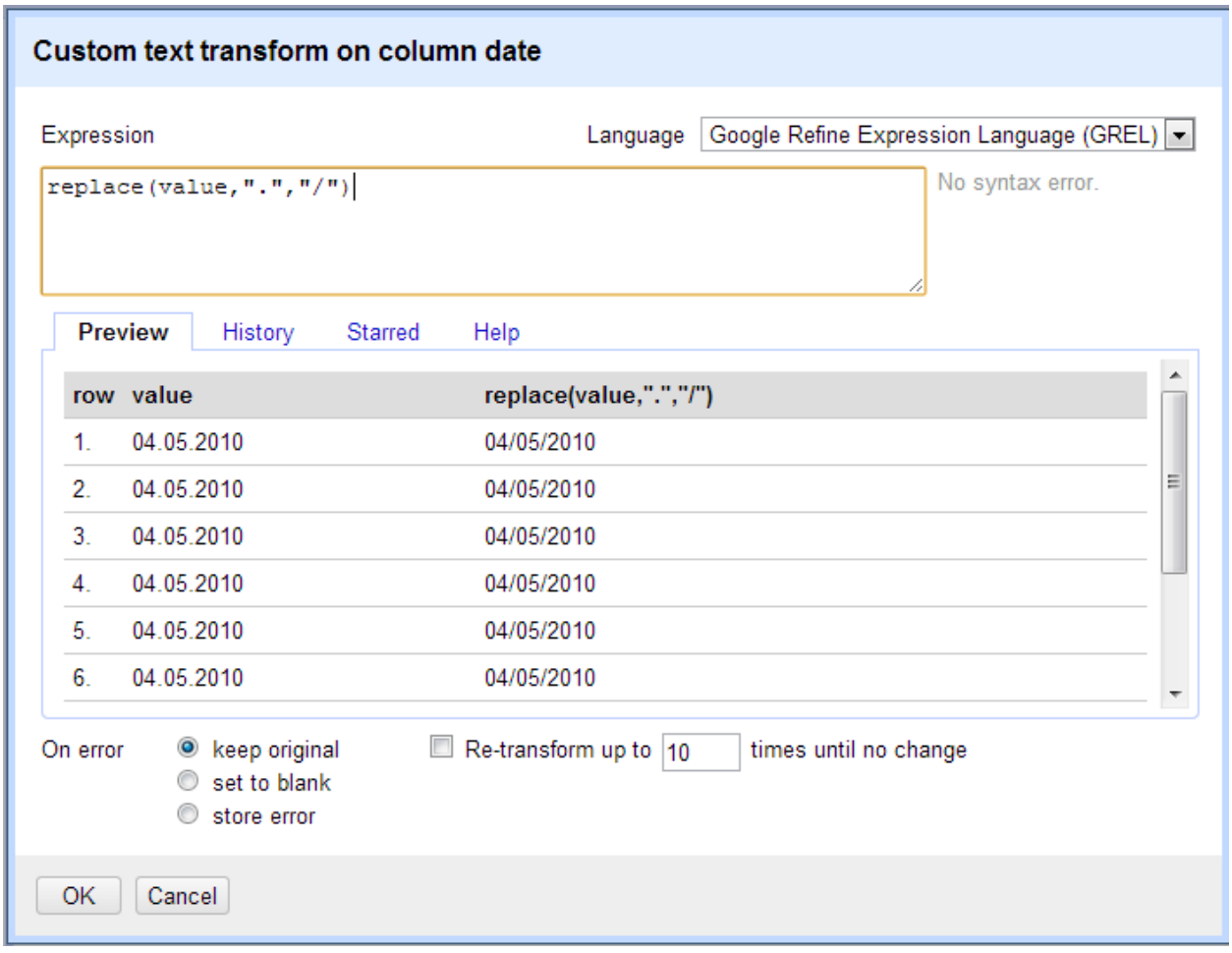

ScrubbingTest - Google R
127.0.0.1:3333/project?project=1442474469997
Google refine ScrubbingTest Permalink
Open... Export Help
Facet / Filter Undo / Redo 0
160 rows
Extensions: Freebase
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 10 next › last »
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
All
Document
Num
name
date
Column
invoice
deliver
id
invoice id
price
unit
type
1. 7295491 10 Buchanan Red Seal 750ml 06x01 04.05.2010 0 FROZA_REDSEAL 12.05.2010 6027282 18 1190.95 1 BTL
2. 7295489 10 J Wlkr Red 750 ml Lata AIG 12x01 04.05.2010 0 829873/27536 10.05.2010 185343 492 124.54 1 BTL
3. 7295489 20 GUINNESS BTL DR 330ML BTL 01x24 04.05.2010 0 829873/27536 10.05.2010 185343 96 13.81 1 BTL
4. 7295489 30 Guinness Lata DR 04.05.2010 0 829873/27536 10.05.2010 185343 144 18.47 1 BTL


name
change
43 choices Sort by: name count
Cluster
BAILEYS ORIGINAL 750ML 12x01 14
Buchanan Deluxe 750ml 12x01 12
JW Black Label 750ml 12x01 10
Buchanan Sr 18y 750ml 06x01 8
Guinness Lata DR 440ml 24x01 8
J&B Rare 750ml 12x01 7
JWlkr Red 750ml MIDAS C/IBC 12x01 7
Vat 69 750ml 12x01 7
Baileys Flavours Caramel 750ml 12x01 6
Baileys Flavours Mint Choco 750ml 12x01 5
CptMrg OSR 750ml 12x01 relay / Lgut 5


Cluster & Edit column "name"
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...
Method key collision
Keying Function fingerprint
1 cluster found
Cluster Size
Row Count
Values in Cluster
Merge?
New Cell Value
2
9
Guinness Lata DR 440ml 24x01 (8 rows)
Guinness Lata DR 440ml 24x01. (1 rows)
Guinness Lata DR 440ml 24x0
Select All
Deselect All
Merge Selected & Re-Cluster
Merge Selected & Close
Close


Cluster & Edit column "name"
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...
Method nearest neighbor
Distance Function PPM
Radius 1.0
Block Chars 6
1 cluster filtered from 4 total
Cluster Size
Row Count
Values in Cluster
Merge?
New Cell Value
2
11
JW Black Label 750ml 12x01 (10 rows)
JW Bck Label 750ml 12x01 (1 rows)
JW Black Label 750ml 12x01
Rows in Cluster
7.64 — 11
Average Length of Choices
25 — 34.5
Length Variance of Choices
0.6799999999999999 — 1.5
Select All
Deselect All
Merge Selected & Re-Cluster
Merge Selected & Close
Close


Edit Facet's Expression based on Column price
Expression
Language
Google Refine Expression Language (GREL)
value.log()
No syntax error.
Preview
History
Starred
Help
row value value.log()
1. 1190.95 3.0758935287540483
2. 124.54 2.0953088613853814
3. 13.81 1.1401936785786313
4. 18.47 1.2664668954402414
5. 13.68 1.1360860973840974
6. 10.22 1.0094508957986938
OK
Cancel
price
change reset
0.00 — 1,200.00
price
change reset
grel:value.log()
1.00 — 3.08
Numeric 158
Non-numeric 0
Blank 0
Error 2


Custom text transform on column date
Expression
Language
Google Refine Expression Language (GREL)
replace(value,".","/")
No syntax error.
Preview
History
Starred
Help
row value replace(value,".","/")
1. 04.05.2010 04/05/2010
2. 04.05.2010 04/05/2010
3. 04.05.2010 04/05/2010
4. 04.05.2010 04/05/2010
5. 04.05.2010 04/05/2010
6. 04.05.2010 04/05/2010
On error
keep original
set to blank
store error
Re-transform up to 10 times until no change
OK
Cancel

Templating Export

Prefix

```
{
  "rows" : [
```

Row Template

```
    {
      "Document" : {{jsonize(cells["Document"].valu
      "Num" : {{jsonize(cells["Num"].value)}},
      "name" : {{jsonize(cells["name"].value)}},
      "date" : {{jsonize(cells["date"].value)}},
      "Column" : {{jsonize(cells["Column"].value)}}
      "invoice" : {{jsonize(cells["invoice"].value)
      "deliver" : {{jsonize(cells["deliver"].value)
      "id" : {{jsonize(cells["id"].value)}},
      "invoice id" : {{jsonize(cells["invoice id"].
      "price" : {{jsonize(cells["price"].value)}},
      "unit" : {{jsonize(cells["unit"].value)}},
      "type" : {{jsonize(cells["type"].value)}},
      "value" : {{jsonize(cells["value"].value)}},
      "value2" : {{jsonize(cells["value2"].value)}}
      "amount" : {{jsonize(cells["amount"].value)}}
```

Row Separator

```
,
```

Suffix

```
  ]
}
```

```
{
  "rows" : [
    {
      "Document" : 7295491,
      "Num" : 10,
      "name" : "Buchanan Red Seal 750ml   06x01",
      "date" : "04/05/2010",
      "Column" : 0,
      "invoice" : "FROZA_REDSEAL",
      "deliver" : "12.05.2010",
      "id" : 6027282,
      "invoice id" : 18,
      "price" : 1190.95,
      "unit" : 1,
      "type" : "BTL",
      "value" : 21437.1,
      "value2" : 32,
      "amount" : 18
    },
    {
      "Document" : 7295489,
      "Num" : 10,
      "name" : "J Wlkr Red 750 ml Lata AIG 12x01",
      "date" : "04/05/2010",
      "Column" : 0,
      "invoice" : "829873/27536",
      "deliver" : "10.05.2010",
      "id" : 185343,
      "invoice id" : 492,
      "price" : 124.54,
      "unit" : 1,
      "type" : "BTL",
```

Reset Template | Export | Cancel

Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

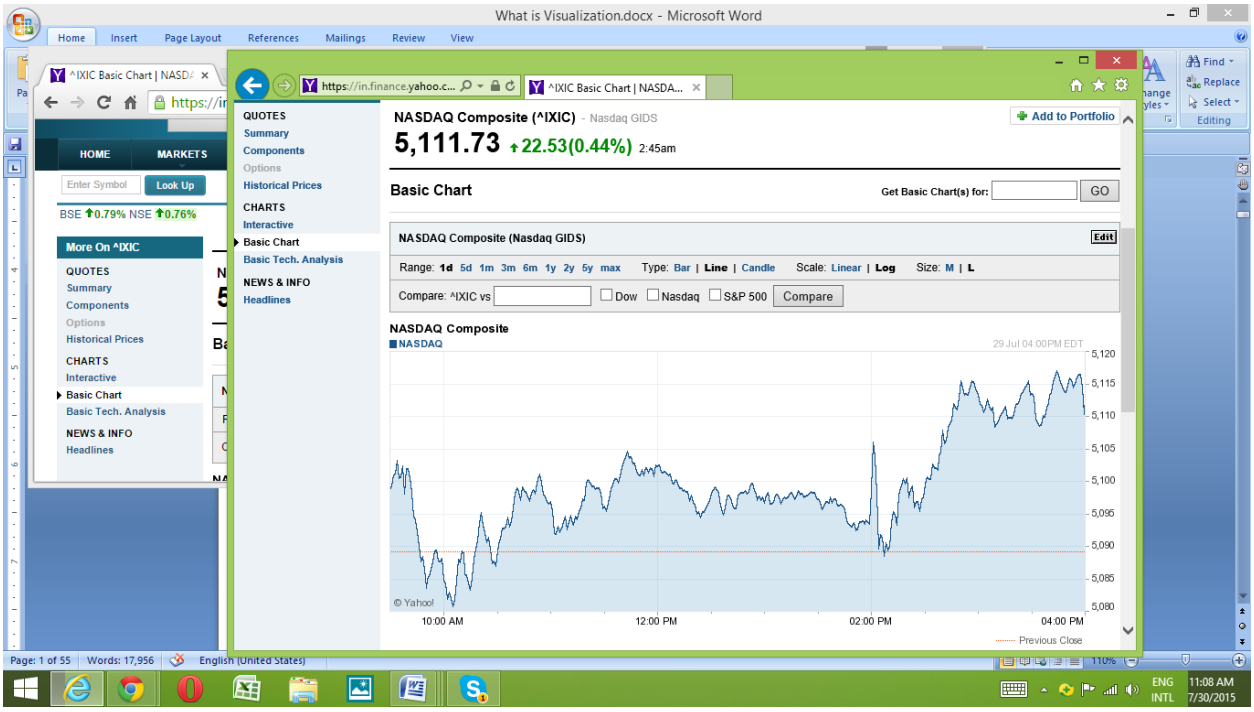
- [x] Mass edit cells in column name
- [x] Mass edit cells in column name
- [x] Text transform on cells in column date using expression value.toDate()
- [x] Text transform on cells in column date using expression grel:replace(value,".","/")
- [x] Text transform on cells in column amount using expression value.toNumber()
- [x] Text transform on cells in column invoice using expression value.toNumber()
- [x] Text transform on cells in column invoice using expression value.toString()

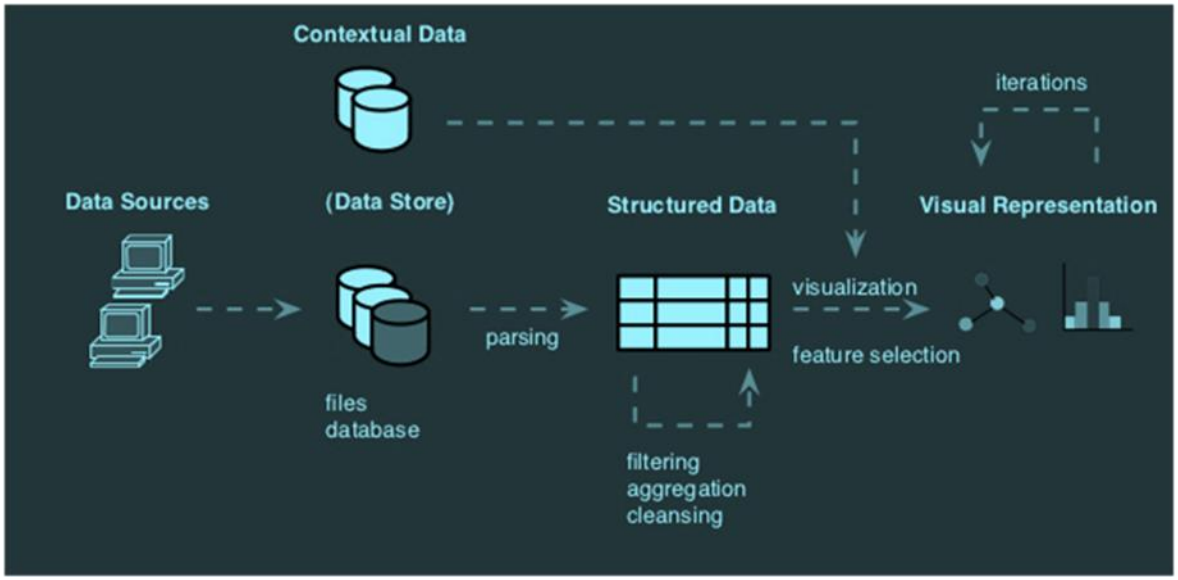
Select All | Unselect All

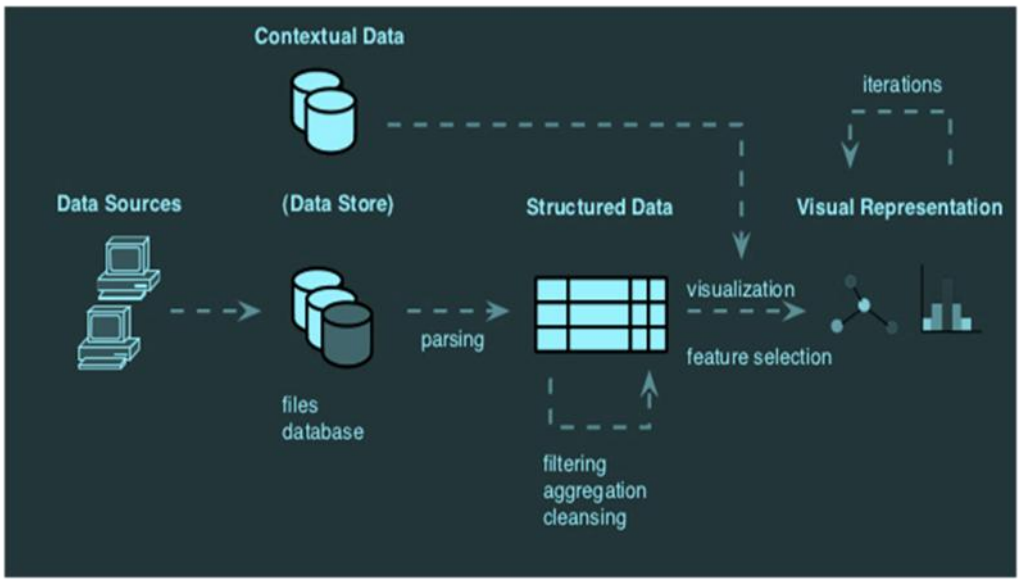

```
[
  {
    "op": "core/mass-edit",
    "description": "Mass edit cells in colum
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "name",
    "expression": "value",
    "edits": [
      {
        "fromBlank": false,
        "fromError": false,
        "from": [
          "Guinness Lata DR 440ml 24x01",
          "Guinness Lata DR 440ml 24x01."
        ],
        "to": "Guinness Lata DR 440ml 24x01"
      }
    ]
  },
  {
    "op": "core/mass-edit",
    "description": "Mass edit cells in colum
    "engineConfig": {
      "facets": [],
```


Close

Chapter 3: Getting to Grips with Visualization





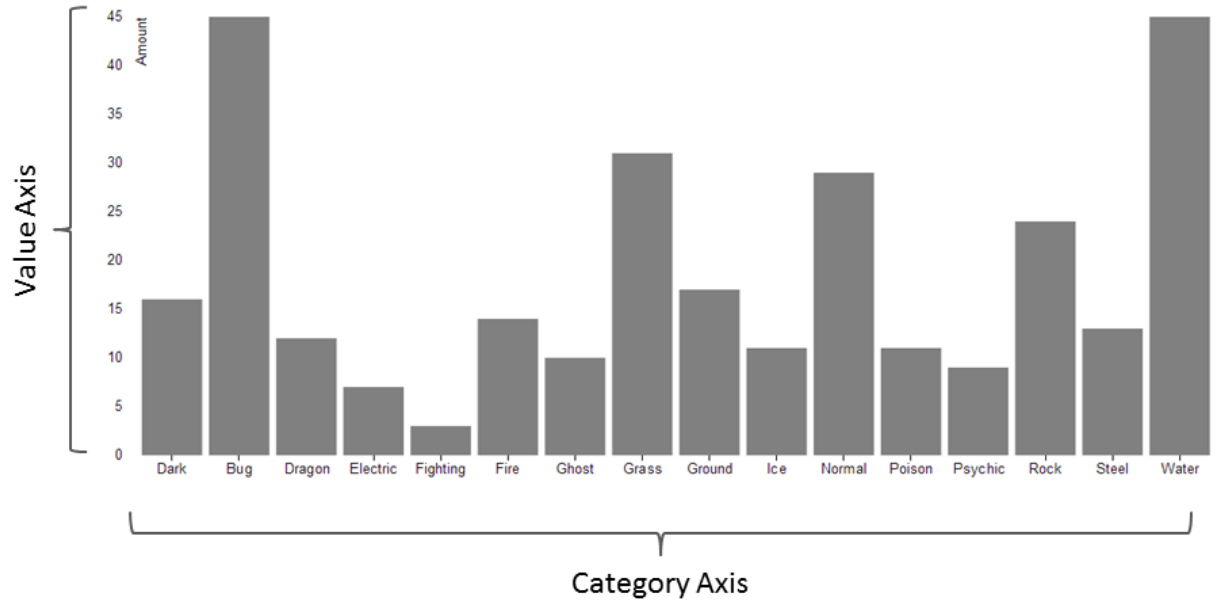


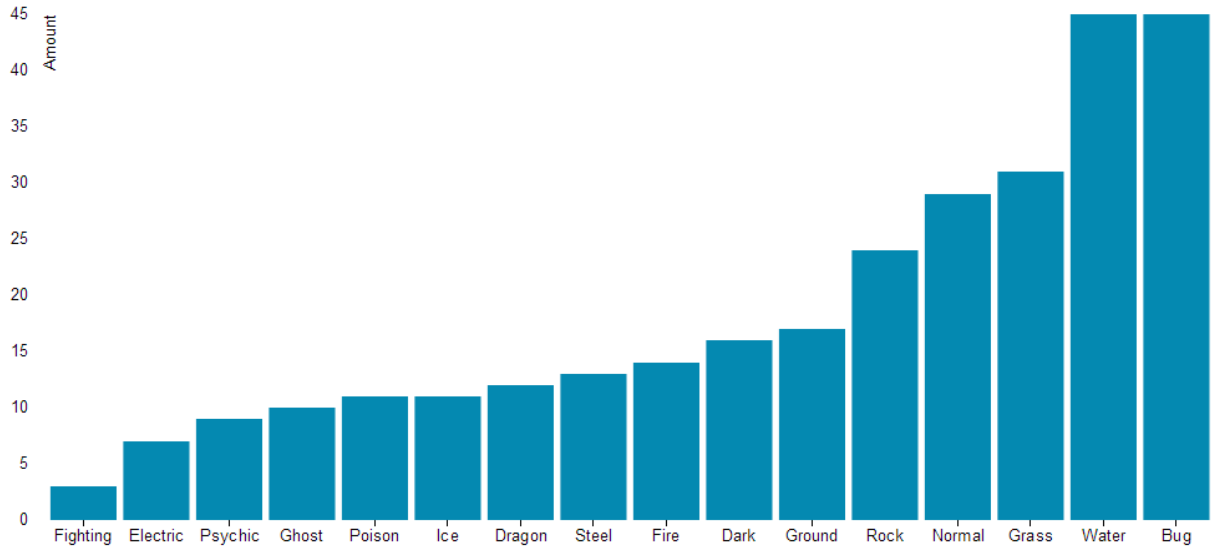

```
<!DOCTYPE html>
<html>
  <head>
    <title> HTML Hello World </title>
    <style>
    body {
      font: 10px arial;
    }
    </style>
  </head>
  <body>
    <p> My first paragraph! </p>
    <script>

        D3 + JavaScript Code
        . . .

    </script>

  </body>
</html>

```

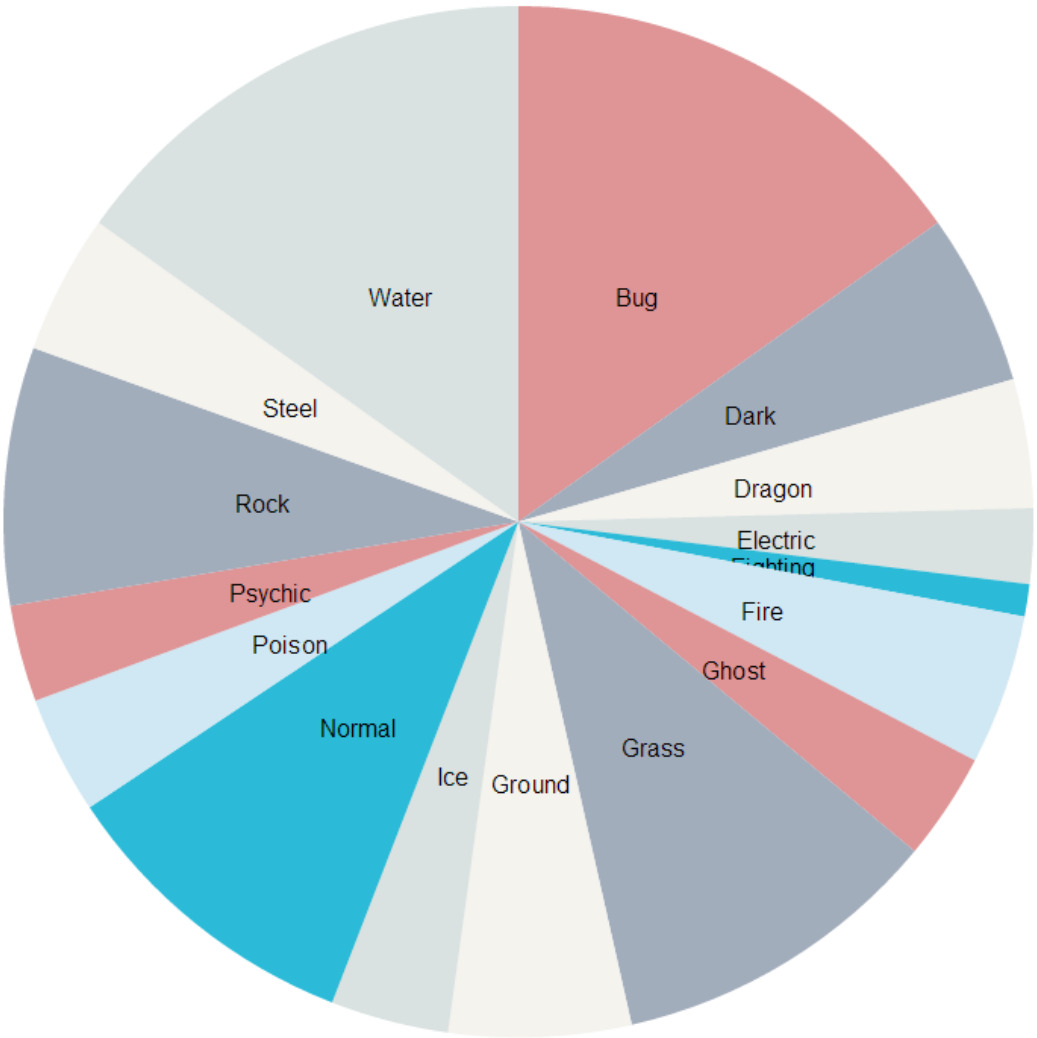
We will define the basic styles for our visualization in **CSS** into the **<style>** tag.

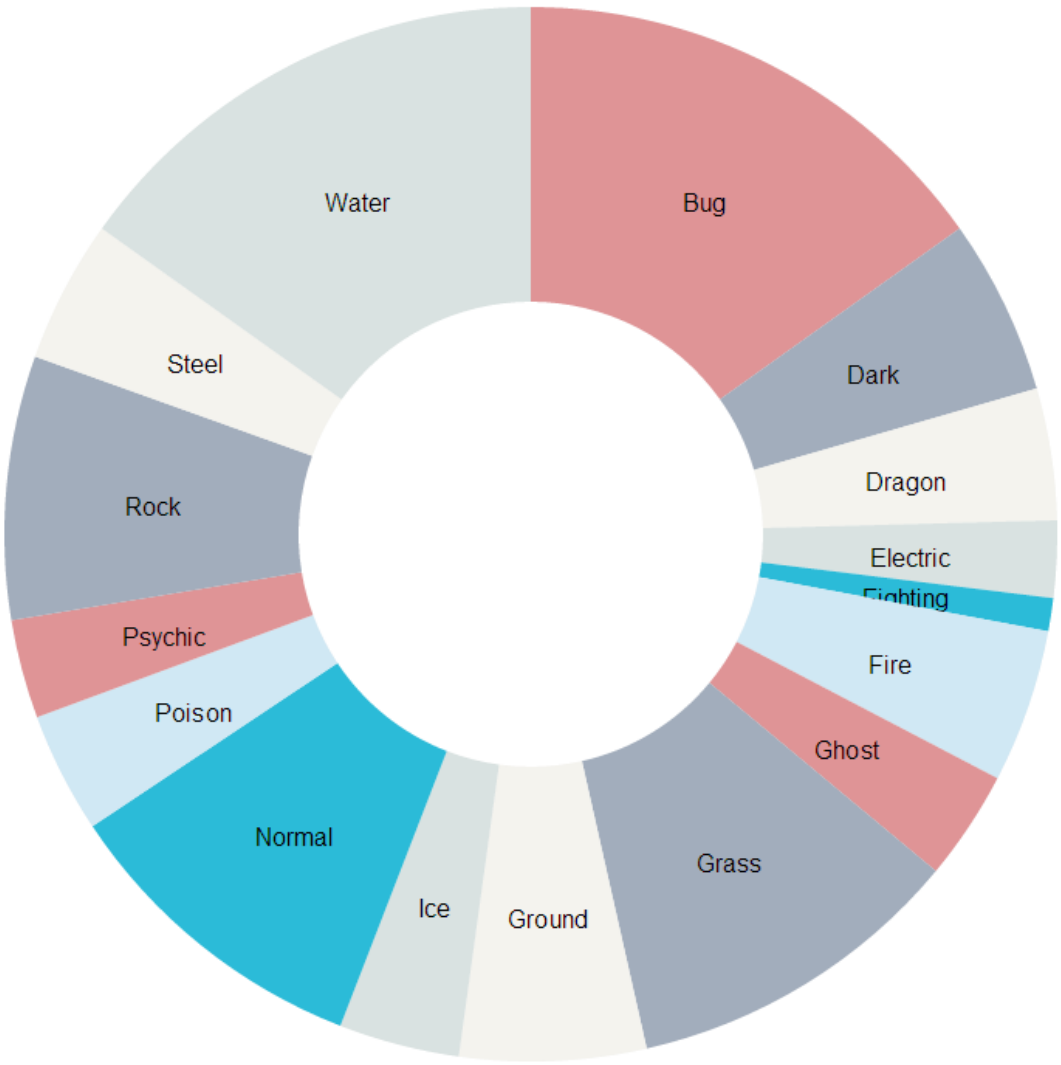
The **D3** code is written in **JavaScript** and will be contained in the **<script>** tag into the body of the **HTML** file

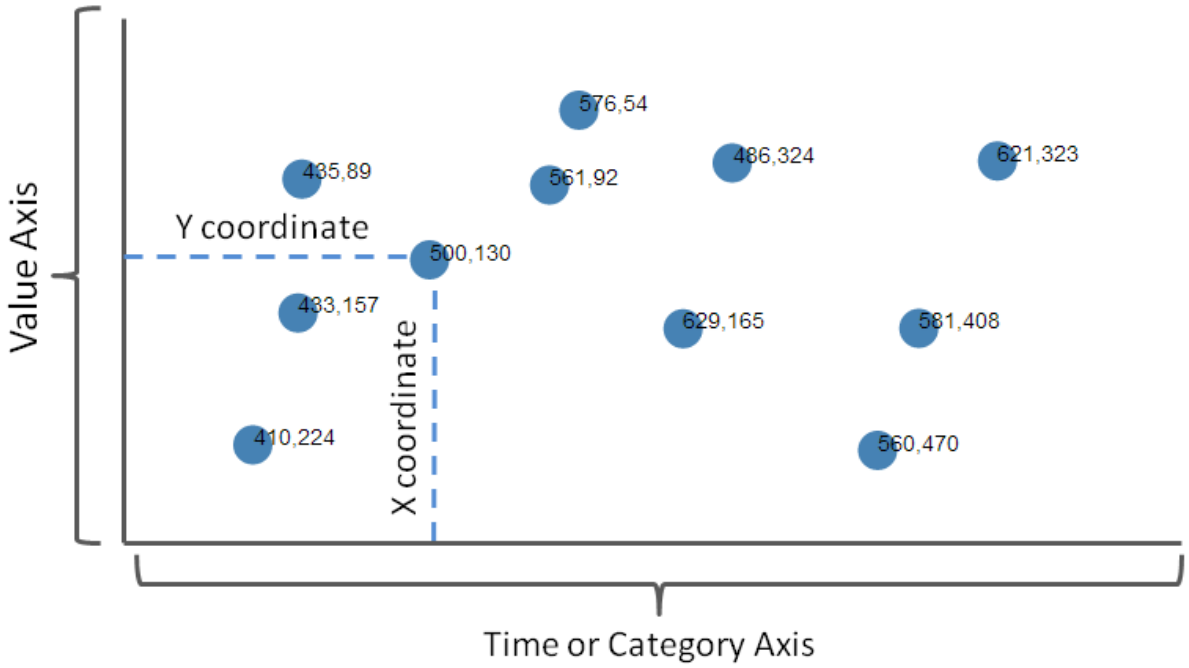

Directory listing for / - Mozilla Firefox
Directory listing for /
localhost:8000
Google
Directory listing for /
d3.v3/

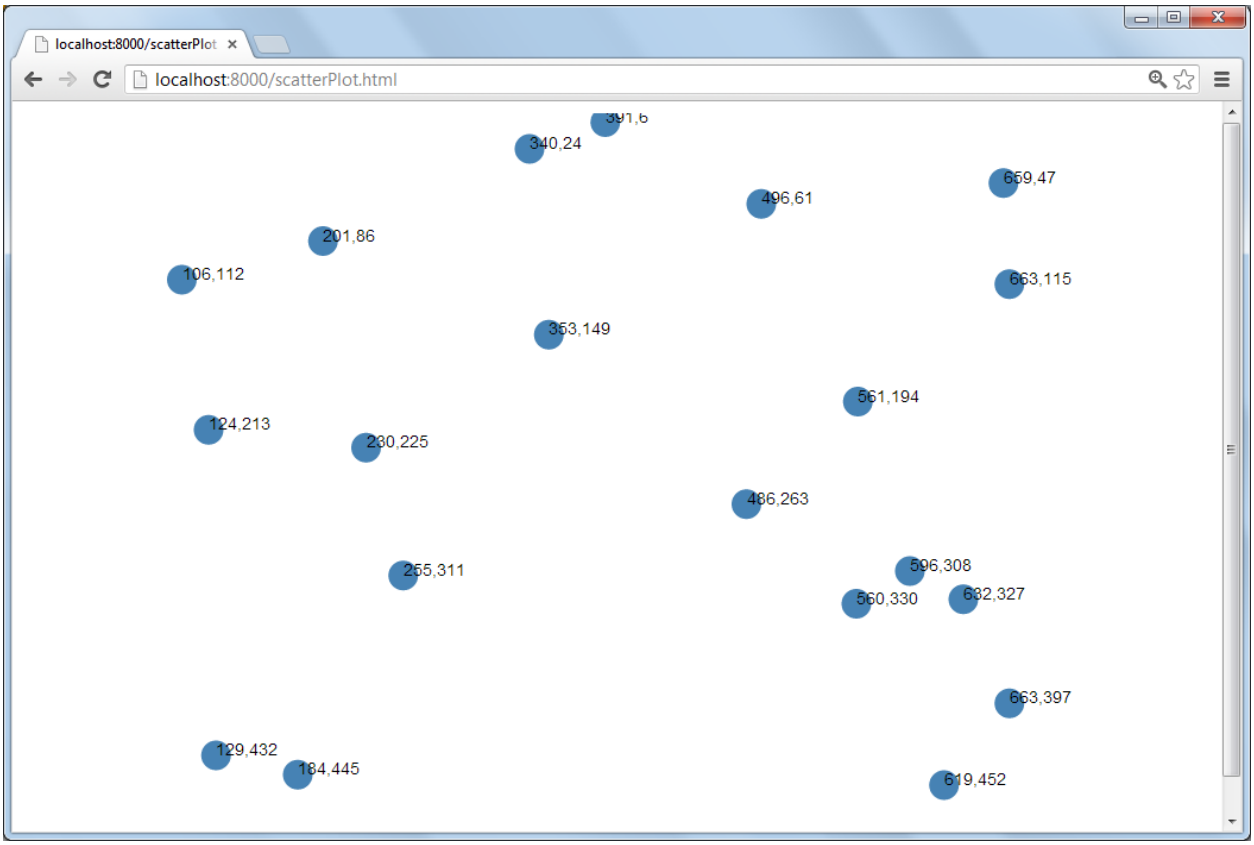

Value Axis
Amount
45
40
35
30
25
20
15
10
5
0
Dark
Bug
Dragon
Electric
Fighting
Fire
Ghost
Grass
Ground
Ice
Normal
Poison
Psychic
Rock
Steel
Water
Category Axis

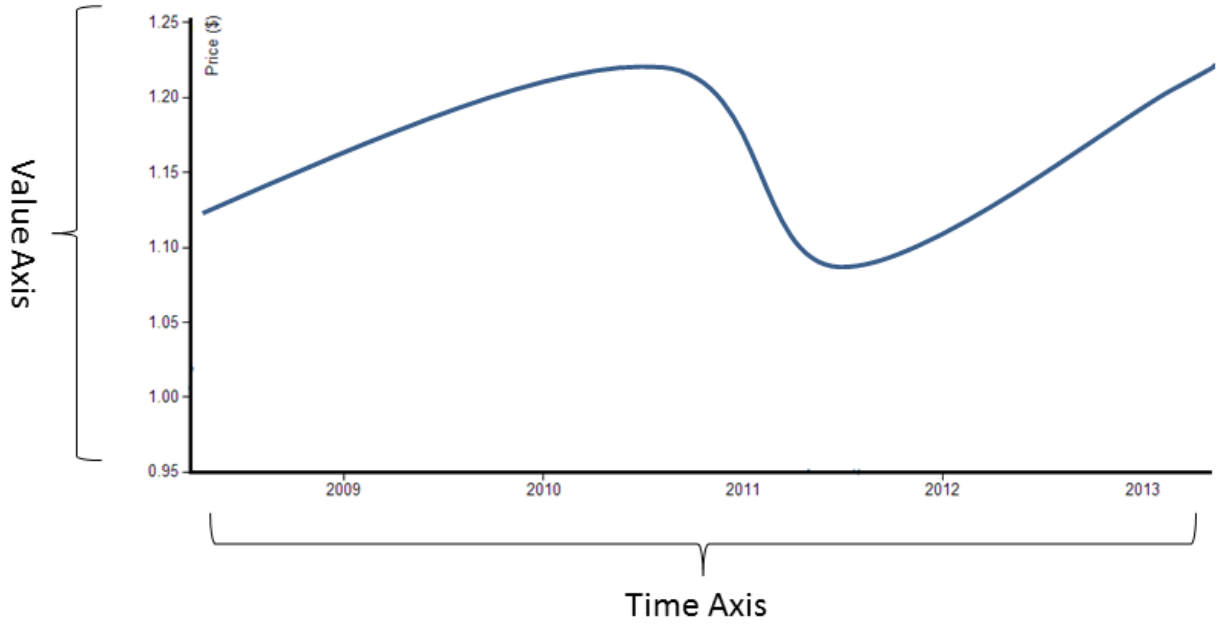

Amount
45
40
35
30
25
20
15
10
5
0
Fighting
Electric
Psychic
Ghost
Poison
Ice
Dragon
Steel
Fire
Dark
Ground
Rock
Normal
Grass
Water
Bug

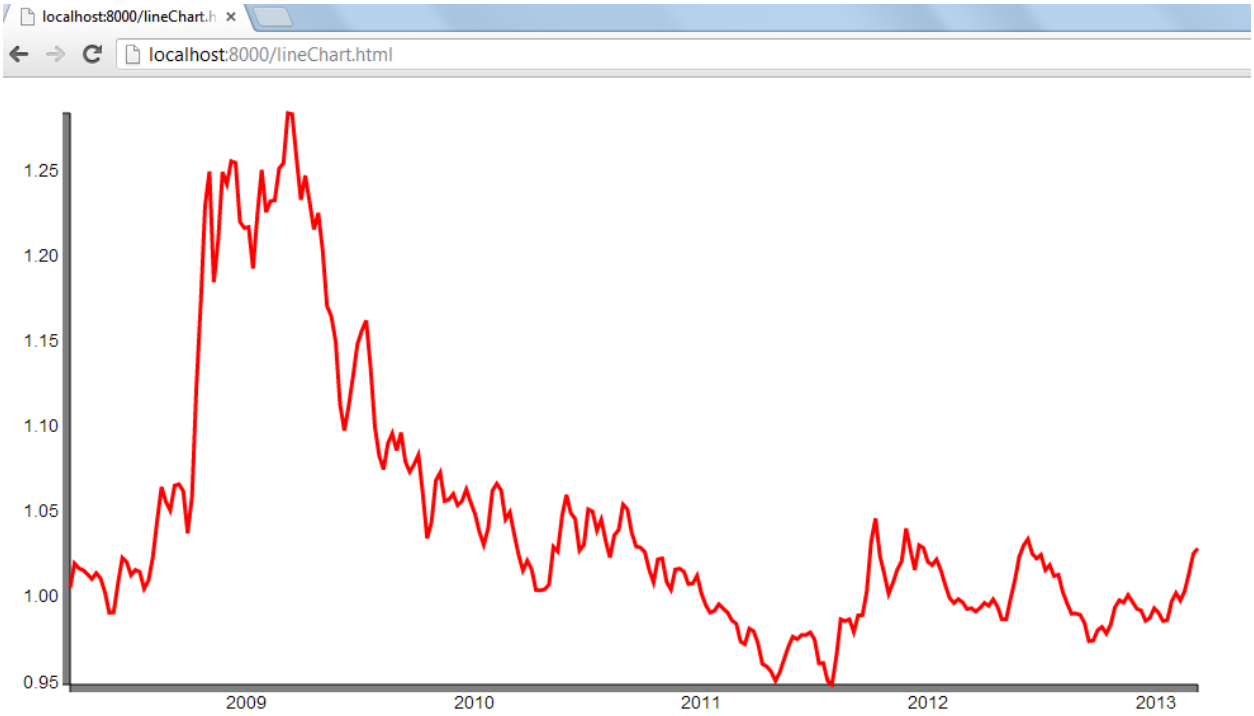

Water
Bug
Steel
Dark
Rock
Dragon
Electric
Psychic
Fire
Poison
Ghost
Normal
Grass
Ice
Ground

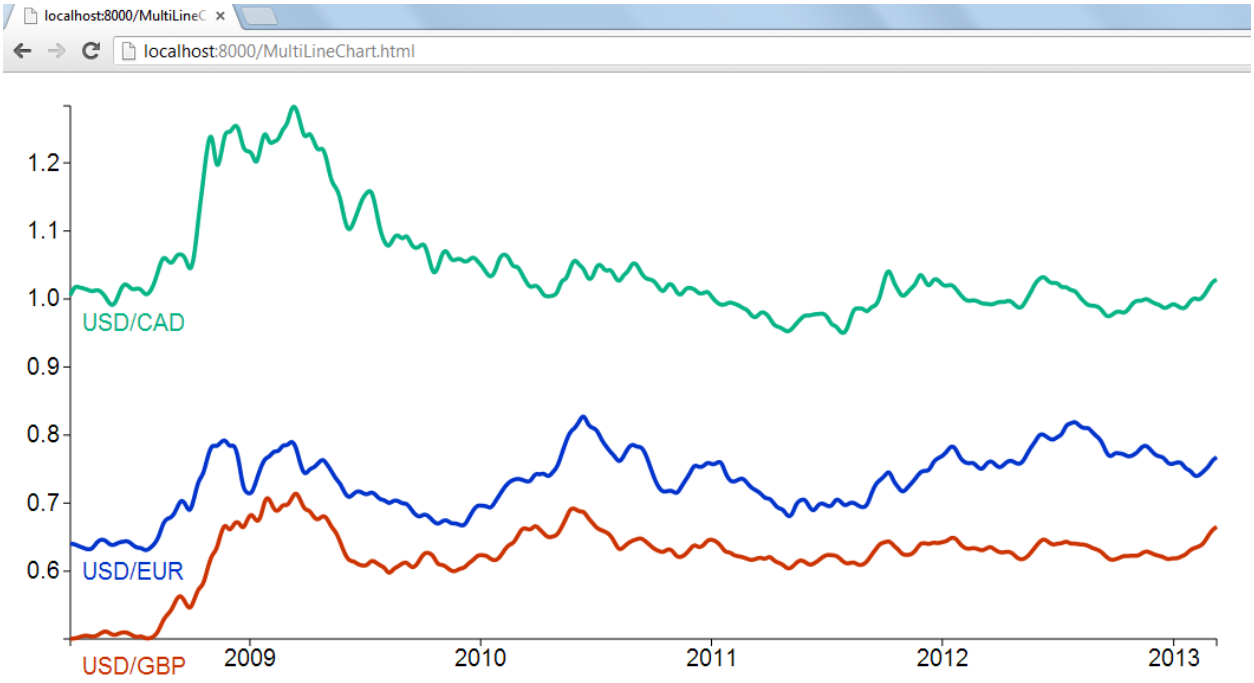

Water
Bug
Steel
Dark
Rock
Dragon
Electric
Fighting
Psychic
Fire
Poison
Ghost
Normal
Grass
Ice
Ground

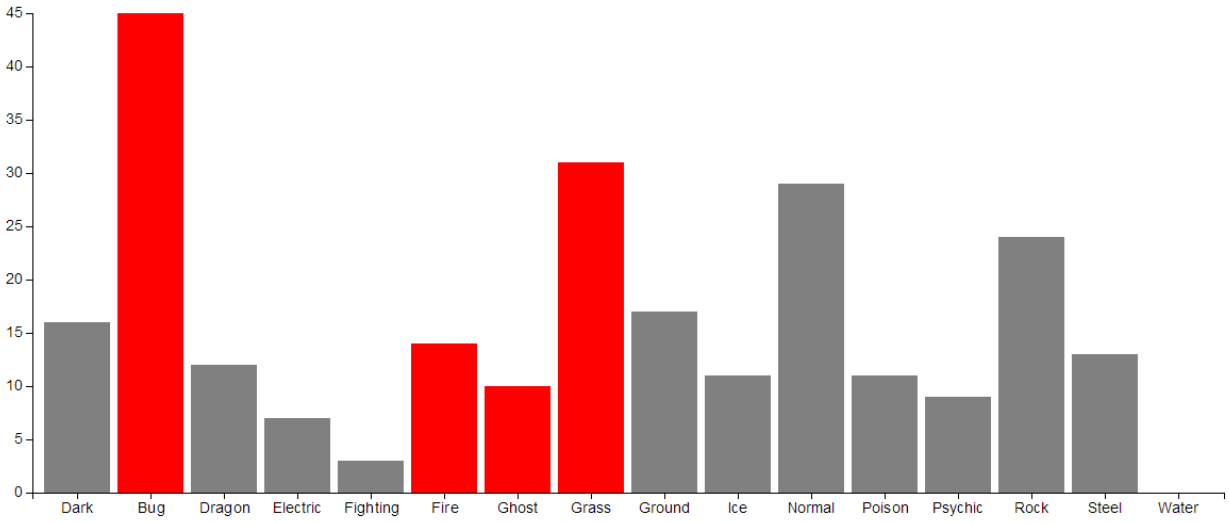

576,54
435,89
561,92
486,324
621,323
Y coordinate
500,130
433,157
629,165
581,408
X coordinate
410,224
560,470
Value Axis
Time or Category Axis


localhost:8000/scatterPlot
localhost:8000/scatterPlot.html
340,24
659,47
496,61
201,86
106,112
663,115
353,149
561,194
124,213
230,225
486,263
255,311
596,308
560,330
682,327
663,397
129,432
184,445
619,452

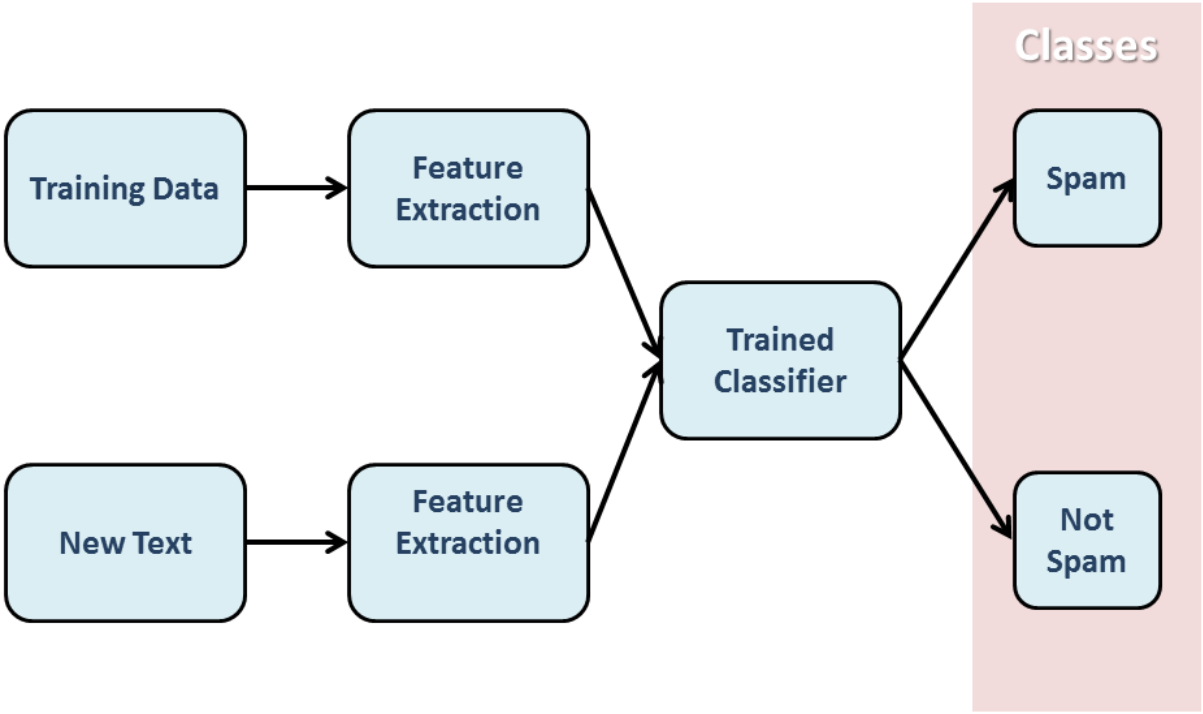

Value Axis
Price ($)
1.25
1.20
1.15
1.10
1.05
1.00
0.95
2009
2010
2011
2012
2013
Time Axis


localhost:8000/lineChart.h
localhost:8000/lineChart.html
1.25
1.20
1.15
1.10
1.05
1.00
0.95
2009
2010
2011
2012
2013


localhost:8000/MultiLineC
localhost:8000/MultiLineChart.html
1.2
1.1
1.0
0.9
0.8
0.7
0.6
USD/CAD
USD/EUR
USD/GBP
2009
2010
2011
2012
2013


45
40
35
30
25
20
15
10
5
0
Dark
Bug
Dragon
Electric
Fighting
Fire
Ghost
Grass
Ground
Ice
Normal
Poison
Psychic
Rock
Steel
Water

Chapter 4: Text Classification

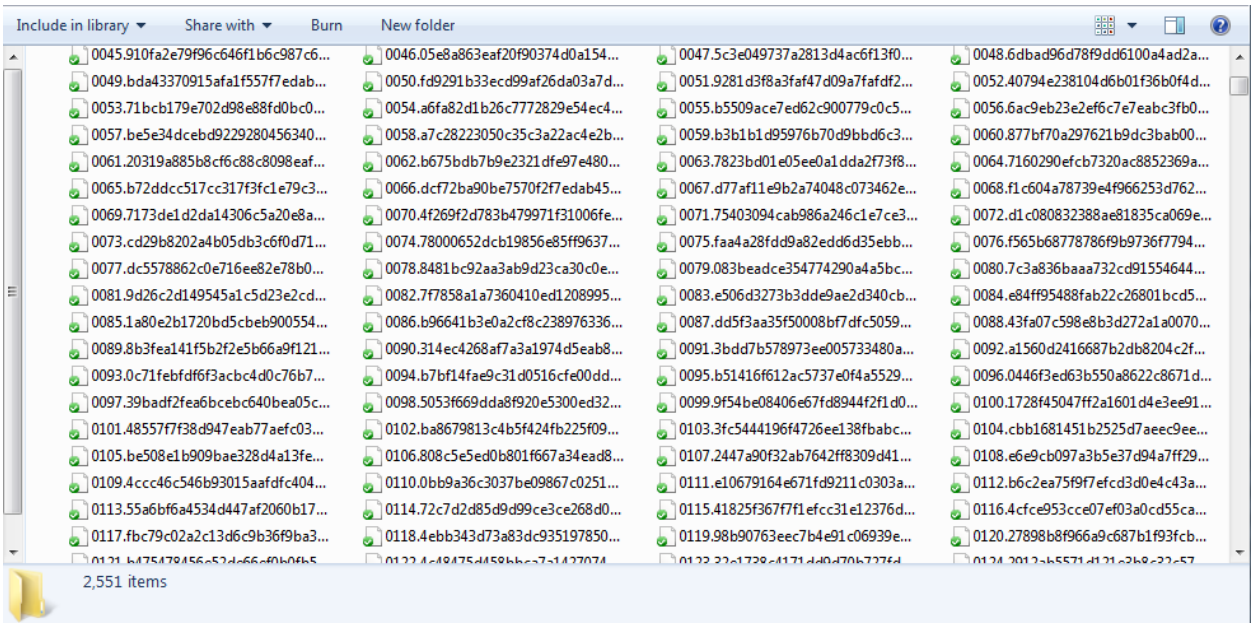


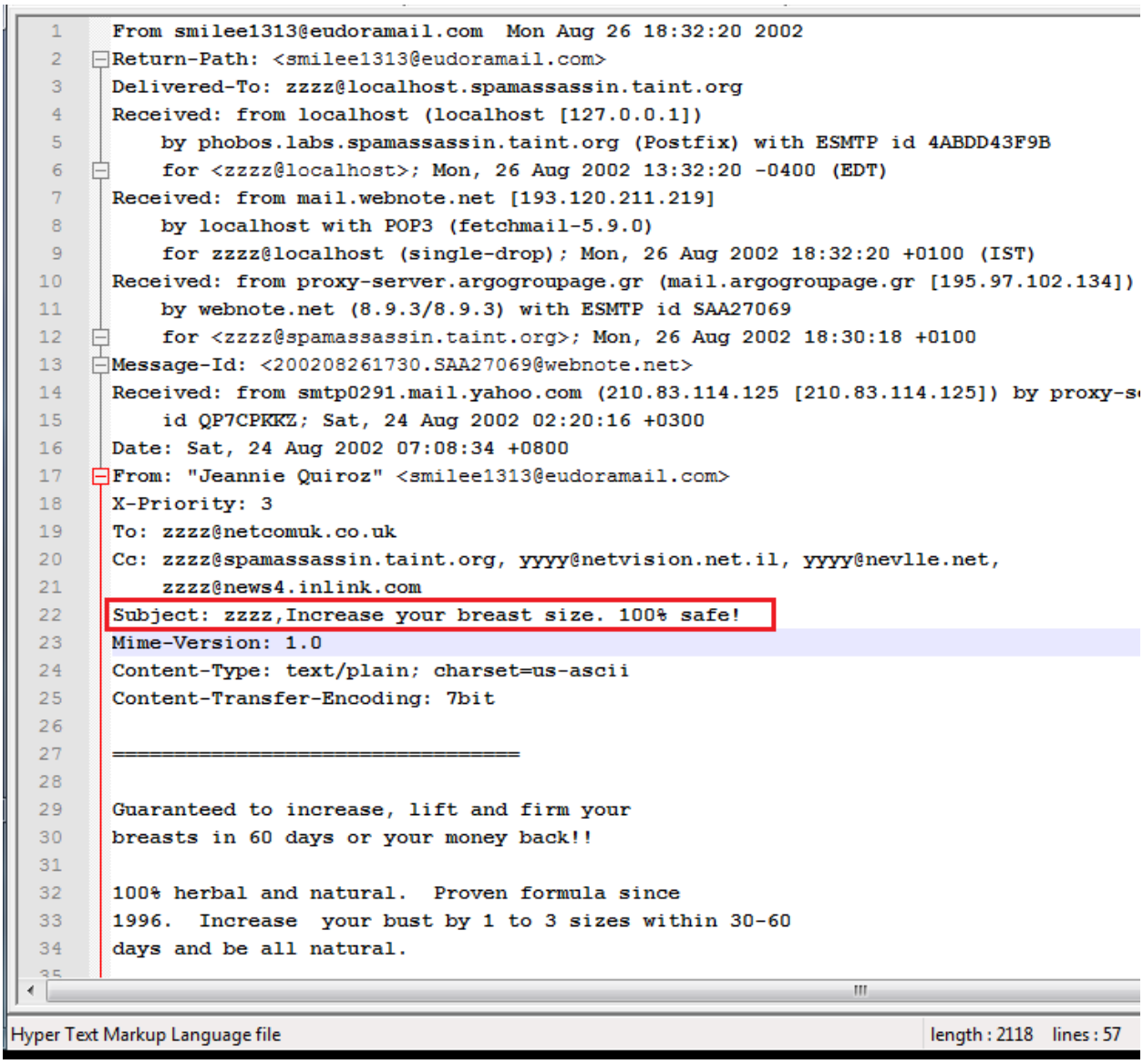
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$P(A \mid B)$ =The conditional probability of **A** given **B**

$$P(word \mid category) = \frac{P(category \mid word)P(word)}{P(category)}$$

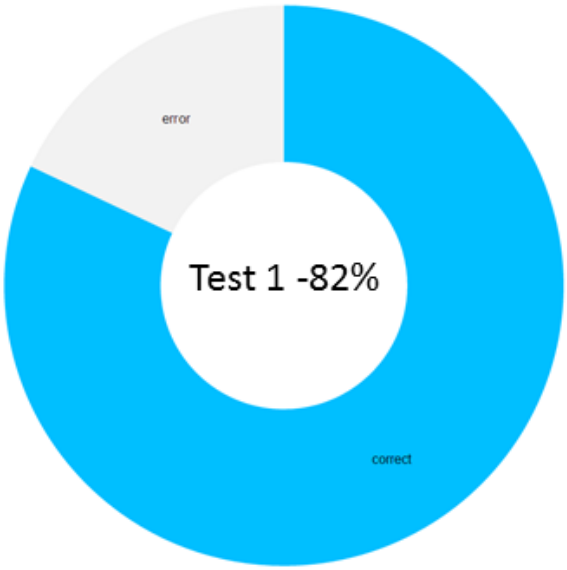
$$P(category \mid word_1, word_2, \ldots, word_n) = P(category) x \Pi_i P(word \mid category)$$

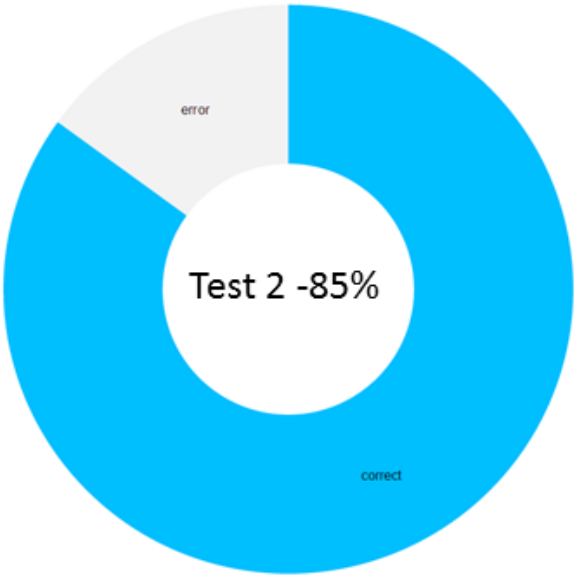

Include in library
Share with
Burn
New folder
0045.910fa2e79f96c646f1b6c987c6...
0046.05e8a863eaf20f90374d0a154...
0047.5c3e049737a2813d4ac6f13f0...
0048.6dbad96d78f9dd6100a4ad2a...
0049.bda43370915afa1f557f7edab...
0050.fd9291b33ecd99af26da03a7d...
0051.9281d3f8a3faf47d09a7fafdf2...
0052.40794e238104d6b01f36b0f4d...
0053.71bcb179e702d98e88fd0bc0...
0054.a6fa82d1b26c7772829e54ec4...
0055.b5509ace7ed62c900779c0c5...
0056.6ac9eb23e2ef6c7e7eabc3fb0...
0057.be5e34dcebd9229280456340...
0058.a7c28223050c35c3a22ac4e2b...
0059.b3b1b1d95976b70d9bbd6c3...
0060.877bf70a297621b9dc3bab00...
0061.20319a885b8cf6c88c8098eaf...
0062.b675bdb7b9e2321dfe97e480...
0063.7823bd01e05ee0a1dda2f73f8...
0064.7160290efcb7320ac8852369a...
0065.b72ddcc517cc317f3fc1e79c3...
0066.dcf72ba90be7570f2f7edab45...
0067.d77af11e9b2a74048c073462e...
0068.f1c604a78739e4f966253d762...
0069.7173de1d2da14306c5a20e8a...
0070.4f269f2d783b479971f31006fe...
0071.75403094cab986a246c1e7ce3...
0072.d1c080832388ae81835ca069e...
0073.cd29b8202a4b05db3c6f0d71...
0074.78000652dcb19856e85ff9637...
0075.faa4a28fdd9a82edd6d35ebb...
0076.f565b68778786f9b9736f7794...
0077.dc5578862c0e716ee82e78b0...
0078.8481bc92aa3ab9d23ca30c0e...
0079.083beadce354774290a4a5bc...
0080.7c3a836baaa732cd91554644...
0081.9d26c2d149545a1c5d23e2cd...
0082.7f7858a1a7360410ed1208995...
0083.e506d3273b3dde9ae2d340cb...
0084.e84ff95488fab22c26801bcd5...
0085.1a80e2b1720bd5cbeb900554...
0086.b96641b3e0a2cf8c238976336...
0087.dd5f3aa35f50008bf7dfc5059...
0088.43fa07c598e8b3d272a1a0070...
0089.8b3fea141f5b2f2e5b66a9f121...
0090.314ec4268af7a3a1974d5eab8...
0091.3bdd7b578973ee005733480a...
0092.a1560d2416687b2db8204c2f...
0093.0c71febfdf6f3acbc4d0c76b7...
0094.b7bf14fae9c31d0516cfe00dd...
0095.b51416f612ac5737e0f4a5529...
0096.0446f3ed63b550a8622c8671d...
0097.39badf2fea6bcebc640bea05c...
0098.5053f669dda8f920e5300ed32...
0099.9f54be08406e67fd8944f2f1d0...
0100.1728f45047ff2a1601d4e3ee91...
0101.48557f7f38d947eab77aefc03...
0102.ba8679813c4b5f424fb225f09...
0103.3fc5444196f4726ee138fbabc...
0104.cbb1681451b2525d7aeec9ee...
0105.be508e1b909bae328d4a13fe...
0106.808c5e5ed0b801f667a34ead8...
0107.2447a90f32ab7642ff8309d41...
0108.e6e9cb097a3b5e37d94a7ff29...
0109.4ccc46c546b93015aafdfc404...
0110.0bb9a36c3037be09867c0251...
0111.e10679164e671fd9211c0303a...
0112.b6c2ea75f9f7efcd3d0e4c43a...
0113.55a6bf6a4534d447af2060b17...
0114.72c7d2d85d9d99ce3ce268d0...
0115.41825f367f7f1efcc31e12376d...
0116.4cfce953cce07ef03a0cd55ca...
0117.fbc79c02a2c13d6c9b36f9ba3...
0118.4ebb343d73a83dc935197850...
0119.98b90763eec7b4e91c06939e...
0120.27898b8f966a9c687b1f93fcb...
2,551 items

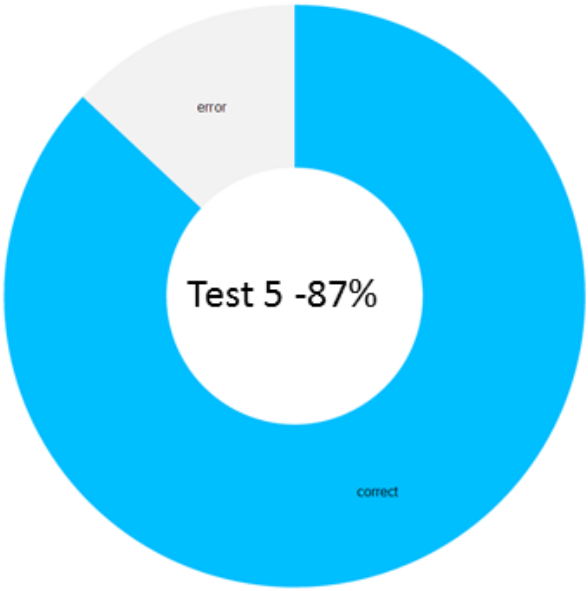


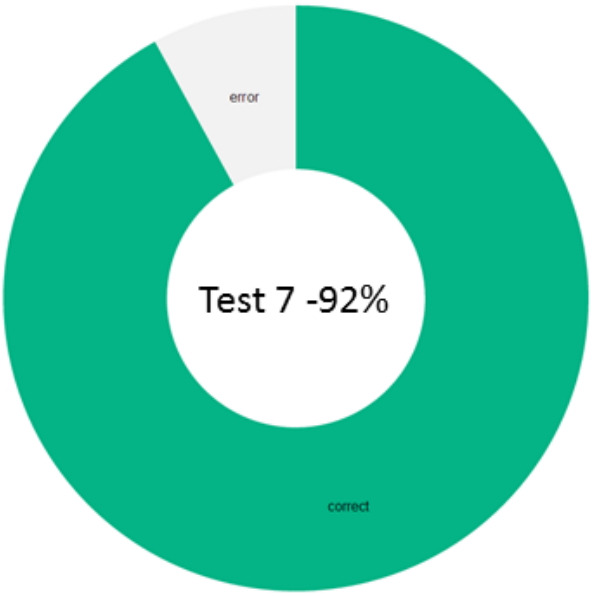
```
From smilee1313@eudoramail.com  Mon Aug 26 18:32:20 2002
Return-Path: <smilee1313@eudoramail.com>
Delivered-To: zzzz@localhost.spamassassin.taint.org
Received: from localhost (localhost [127.0.0.1])
    by phobos.labs.spamassassin.taint.org (Postfix) with ESMTP id 4ABDD43F9B
    for <zzzz@localhost>; Mon, 26 Aug 2002 13:32:20 -0400 (EDT)
Received: from mail.webnote.net [193.120.211.219]
    by localhost with POP3 (fetchmail-5.9.0)
    for zzzz@localhost (single-drop); Mon, 26 Aug 2002 18:32:20 +0100 (IST)
Received: from proxy-server.argogroupage.gr (mail.argogroupage.gr [195.97.102.134])
    by webnote.net (8.9.3/8.9.3) with ESMTP id SAA27069
    for <zzzz@spamassassin.taint.org>; Mon, 26 Aug 2002 18:30:18 +0100
Message-Id: <200208261730.SAA27069@webnote.net>
Received: from smtp0291.mail.yahoo.com (210.83.114.125 [210.83.114.125]) by proxy-s
    id QP7CPKKZ; Sat, 24 Aug 2002 02:20:16 +0300
Date: Sat, 24 Aug 2002 07:08:34 +0800
From: "Jeannie Quiroz" <smilee1313@eudoramail.com>
X-Priority: 3
To: zzzz@netcomuk.co.uk
Cc: zzzz@spamassassin.taint.org, yyyy@netvision.net.il, yyyy@nevlle.net,
    zzzz@news4.inlink.com
Subject: zzzz,Increase your breast size. 100% safe!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit

==================================

Guaranteed to increase, lift and firm your
breasts in 60 days or your money back!!

100% herbal and natural.  Proven formula since
1996.  Increase  your bust by 1 to 3 sizes within 30-60
days and be all natural.
```

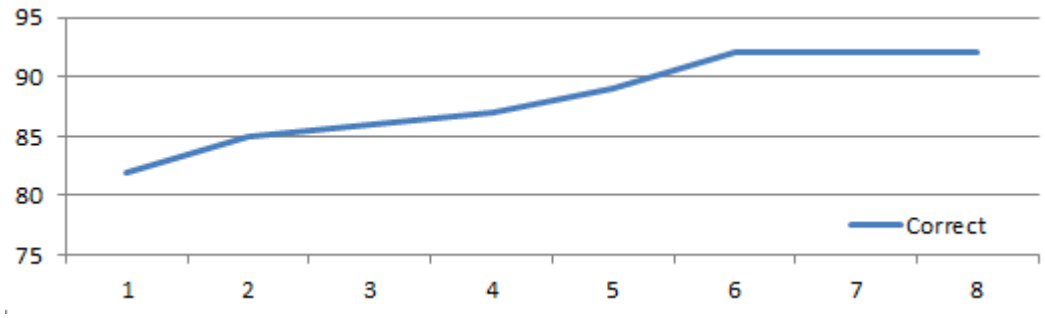
Hyper Text Markup Language file length : 2118 lines : 57


error
Test 1 -82%
correct


error
Test 2 -85%
correct

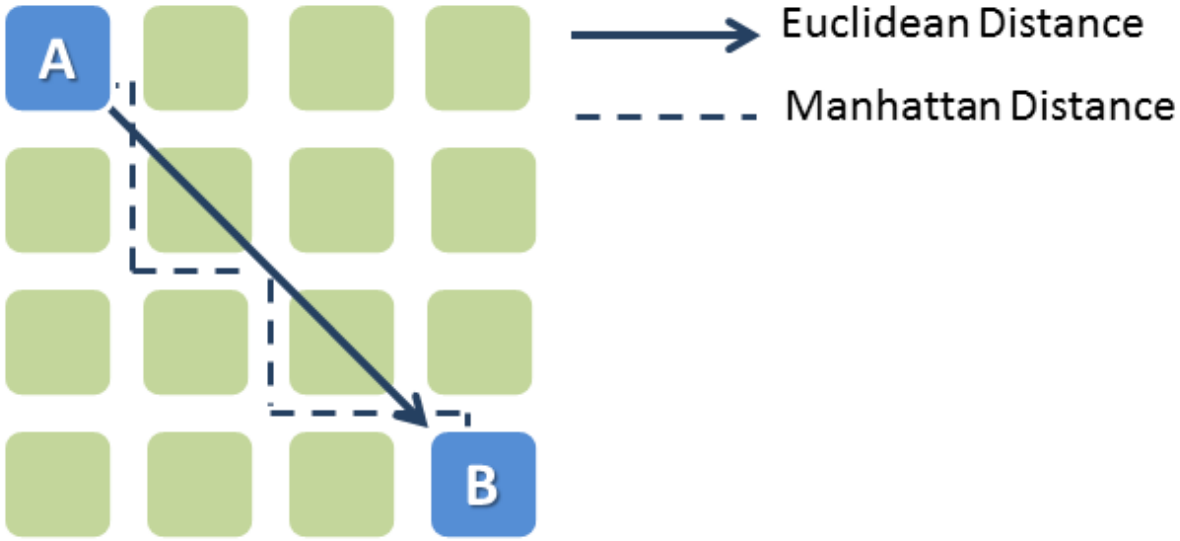

error
Test 5 -87%
correct

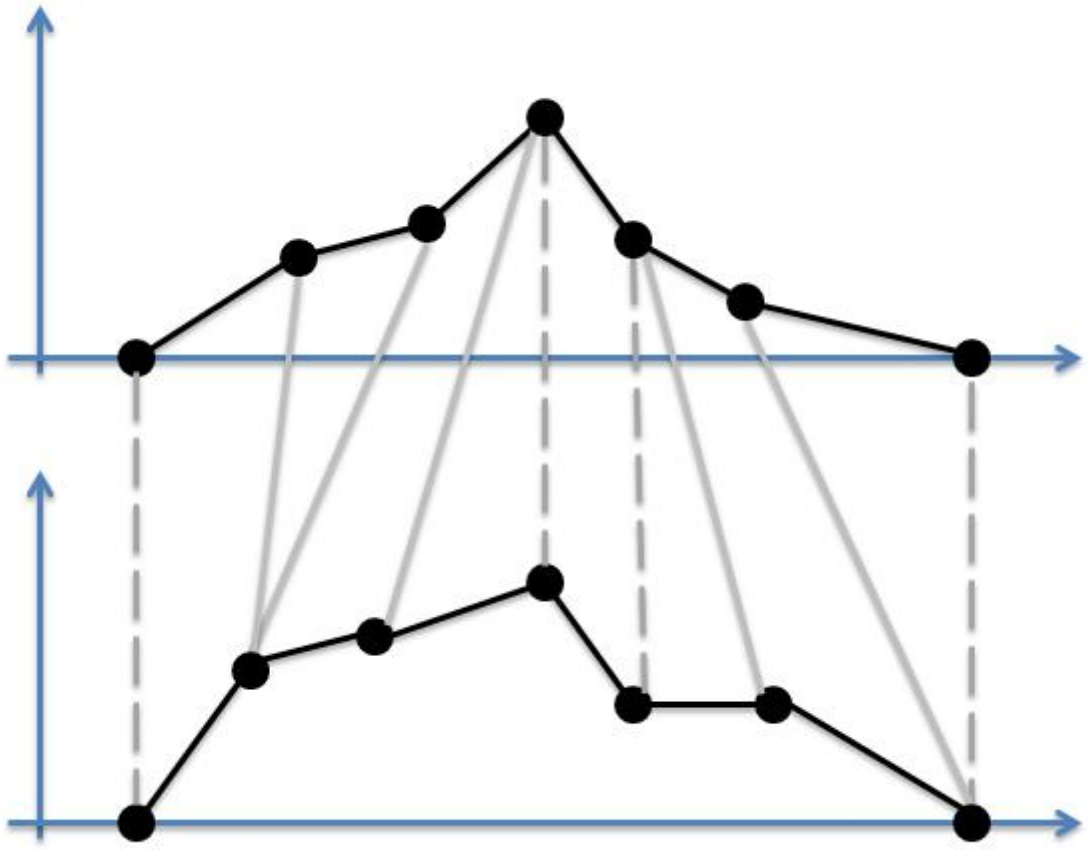

error
Test 7 -92%
correct

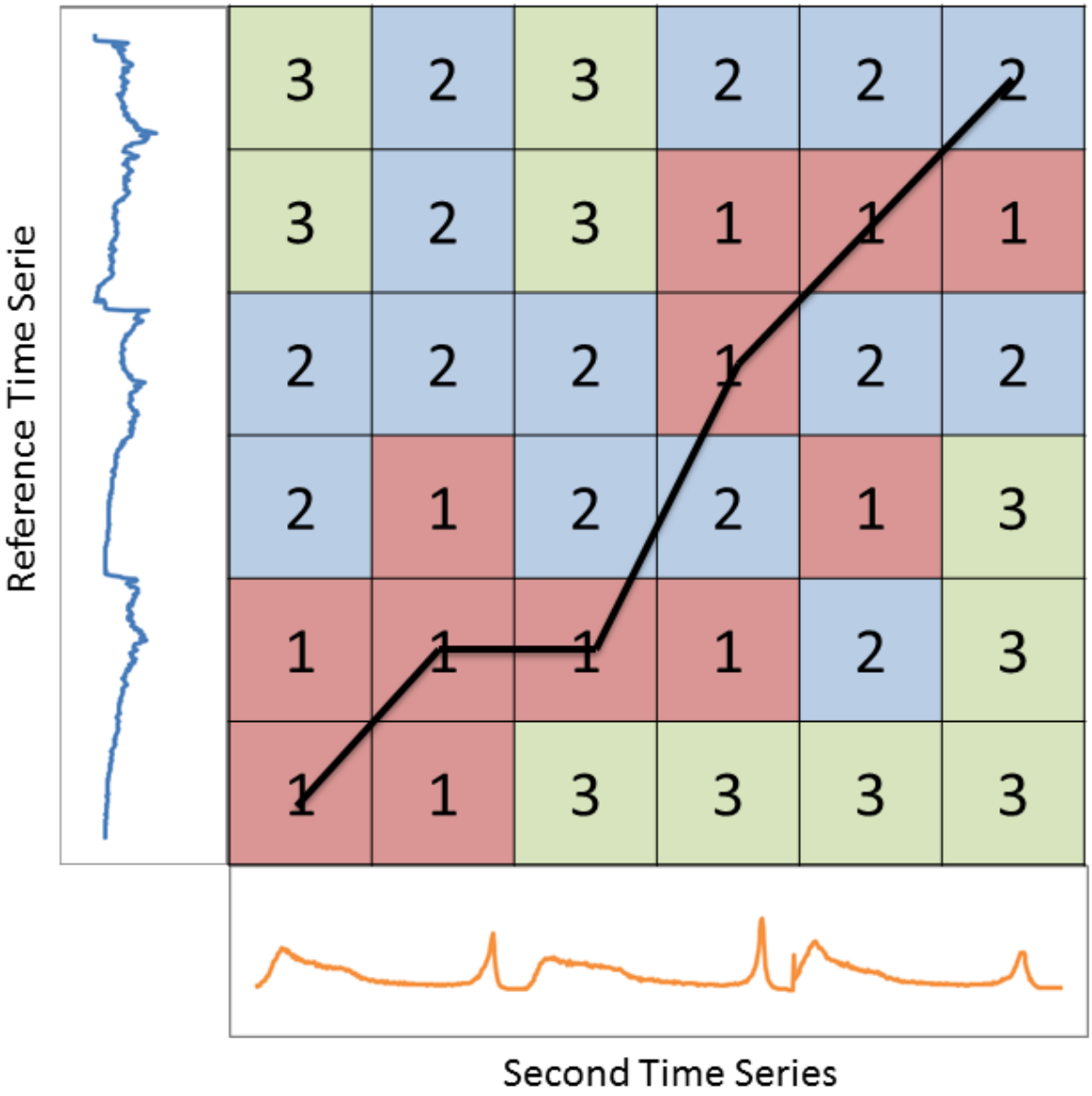

95
90
85
80
75
1
2
3
4
5
6
7
8
Correct

Chapter 5: Similarity-Based Image Retrieval

$$\wedge = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{m,1} & \cdots & d_{m,n} \end{pmatrix} \Rightarrow \vee = \{v_1 \ldots v_n\}$$






Reference Time Serie
3 2 3 2 2 2
3 2 3 1 1 1
2 2 2 1 2 2
2 1 2 2 1 3
1 1 1 1 2 3
1 1 3 3 3 3
Second Time Series

Python Shell

File Edit Shell Debug Options Windows Help

```
Python 3.2.3 (default, Apr 11 2012, 07:15:24) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
31-0.0
32-1589.0
29-1600.0
33-2321.0
35-2665.0
37-2767.0
26-3175.0
30-3182.0
36-3799.0
34-3991.0
28-4368.0
27-4396.0
21-19246.0
17-33044.0
22-34385.0
15-38277.0
45-45764.0
24-47578.0
13-53215.0
10-58143.0
5-59204.0
38-63817.0
39-66253.0
23-70798.0
44-72400.0
11-73442.0
19-78024.0
9-82699.0
3-84618.0
40-85208.0
25-86379.0
2-91017.0
1-91593.0
41-91969.0
43-92199.0
4-92280.0
```

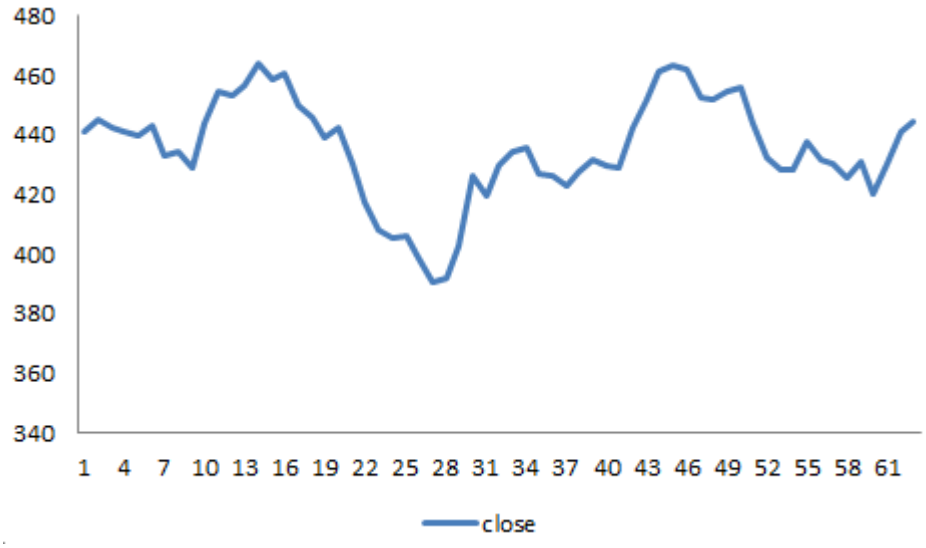
Ln: 28 Col: 10

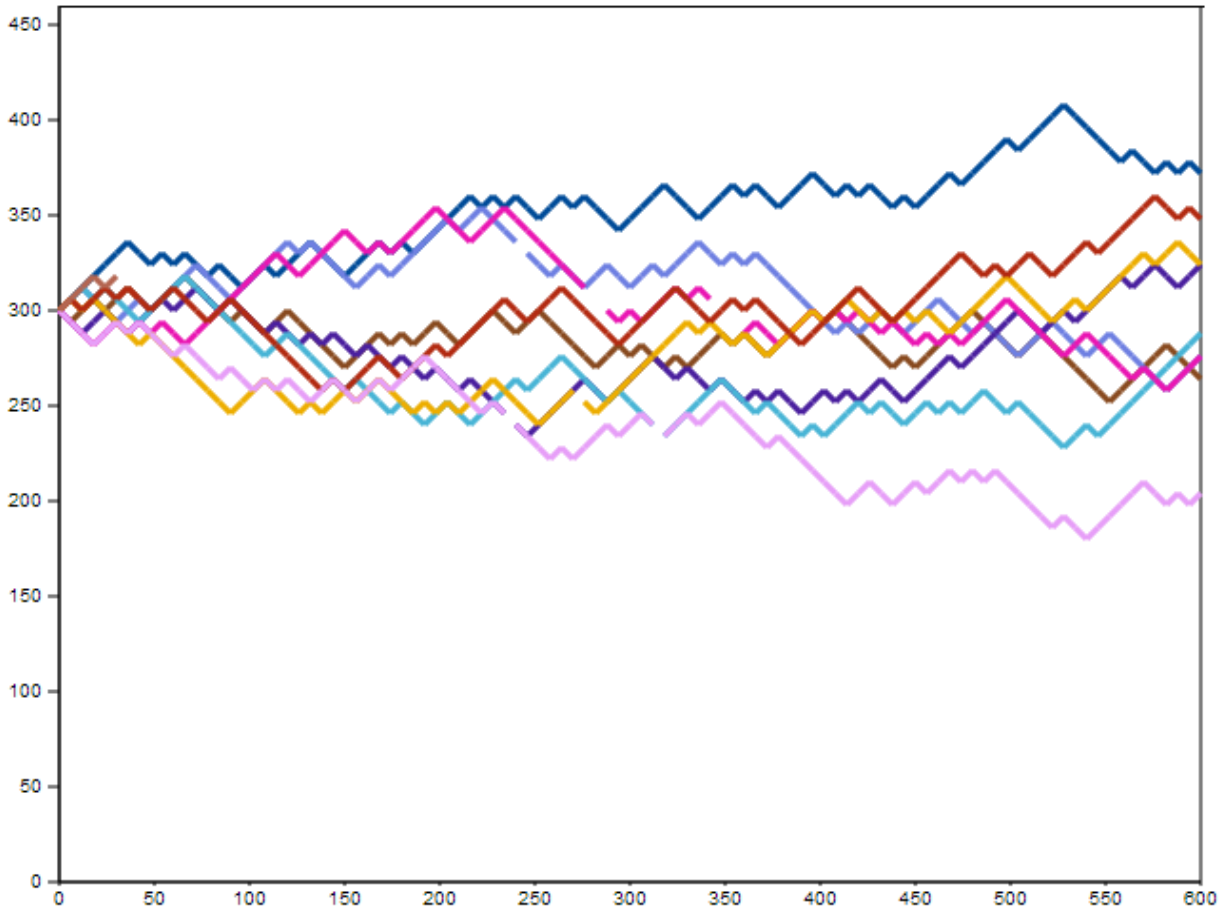
Reference Image	Top 3 Similar Images		
			
			
			

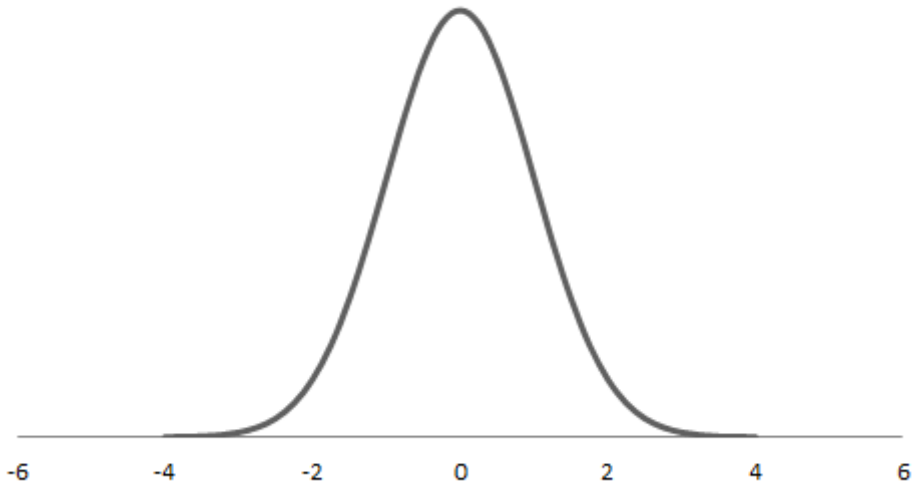
Reference Image	Top 3 Similar Images		
			
			
			

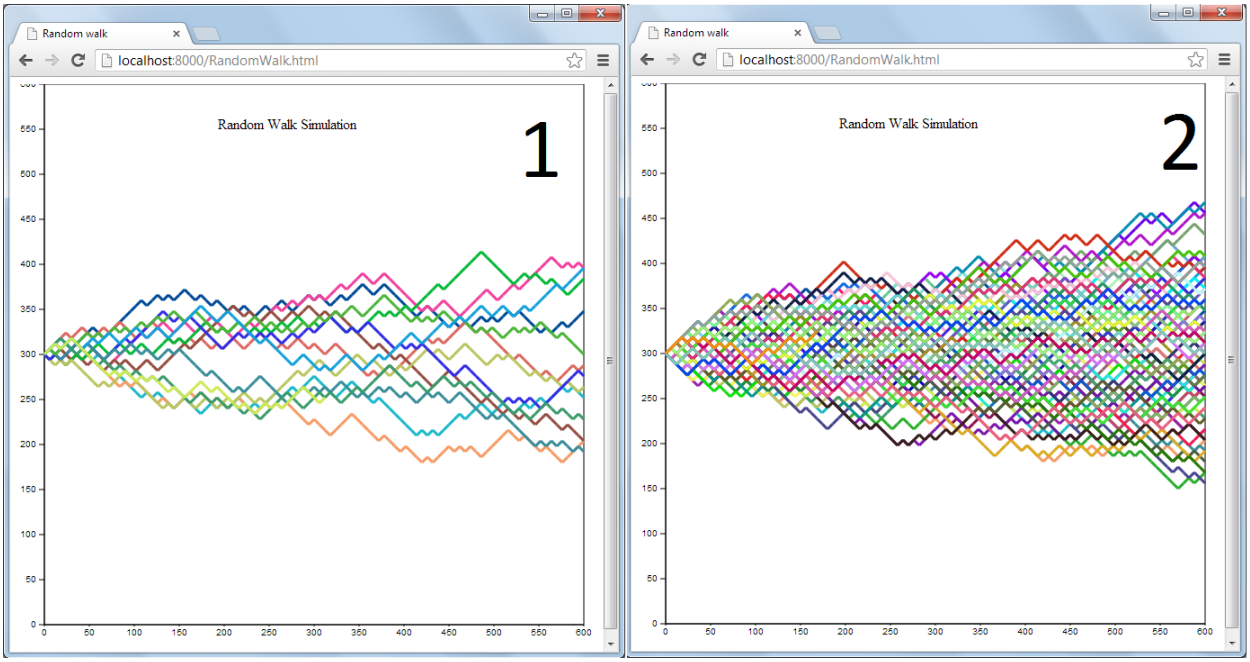
Reference Image	Top 3 Similar Images		
			

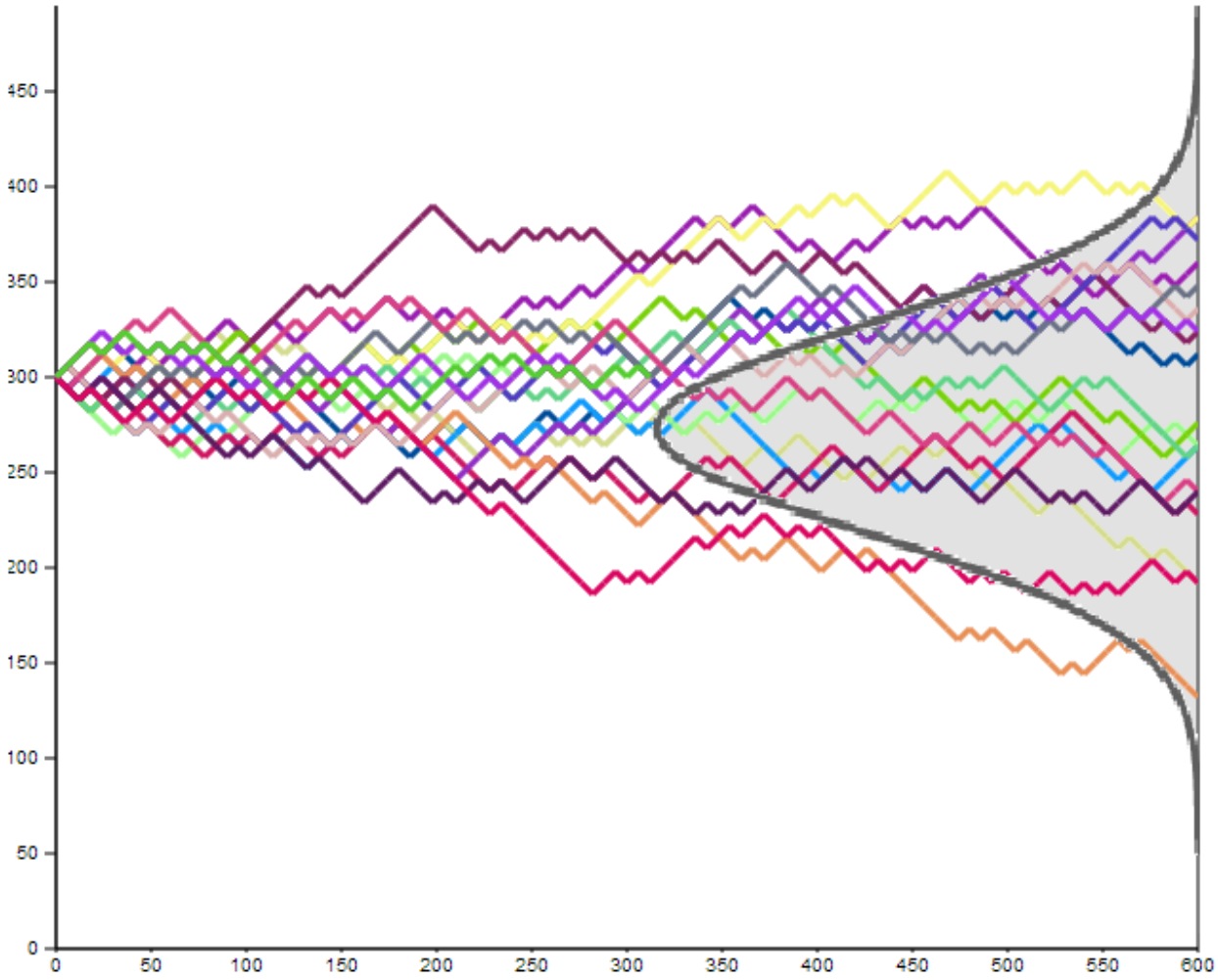
Chapter 6: Simulation of Stock Prices

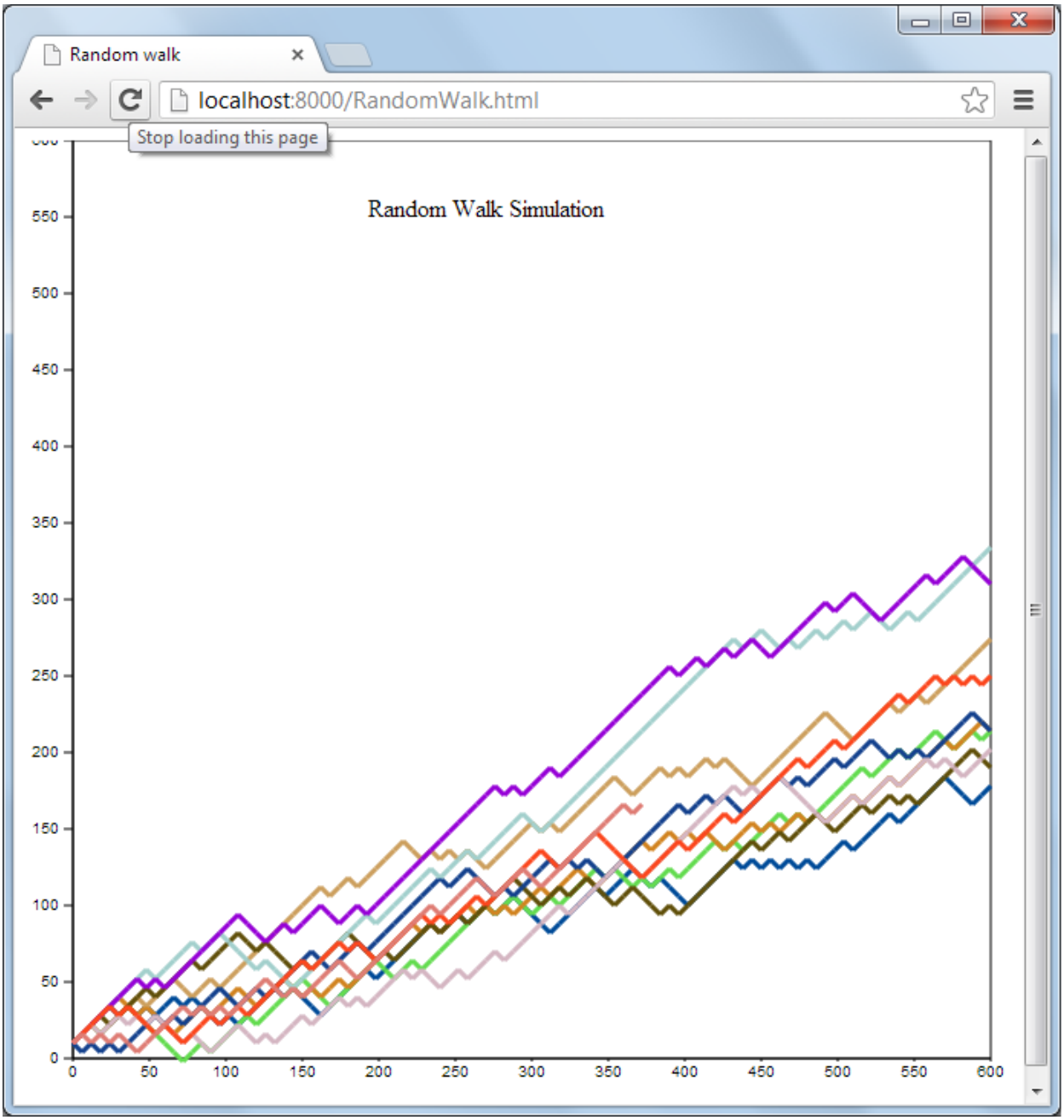




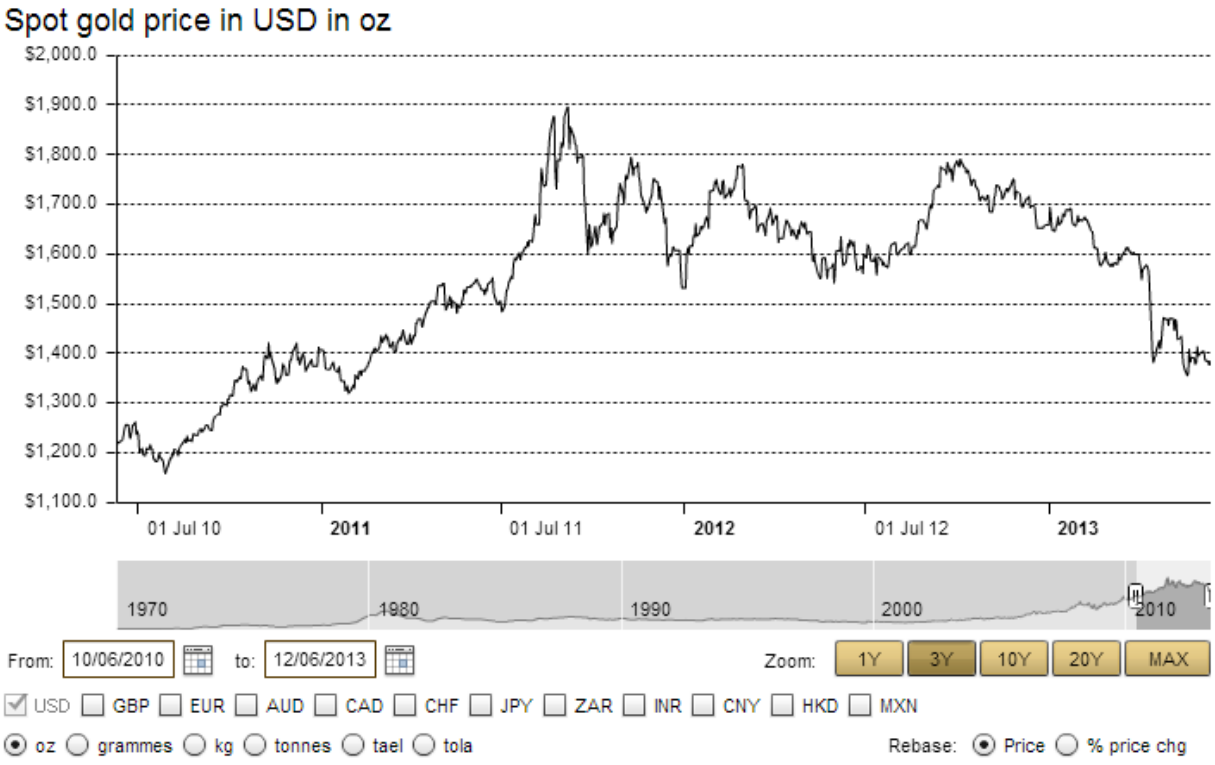

-6
-4
-2
0
2
4
6

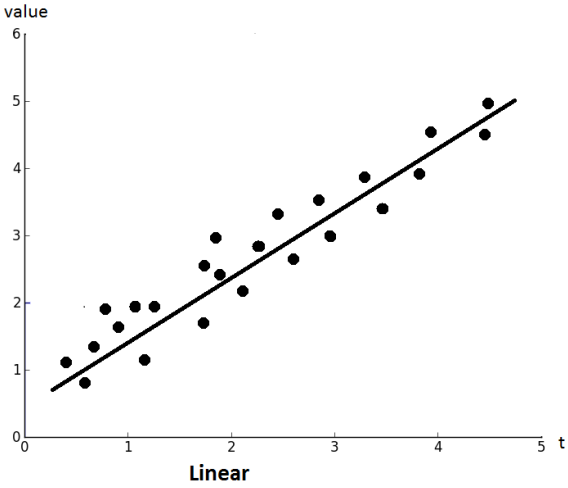

Random walk
localhost:8000/RandomWalk.html
Random Walk Simulation
1
Random walk
localhost:8000/RandomWalk.html
Random Walk Simulation
2

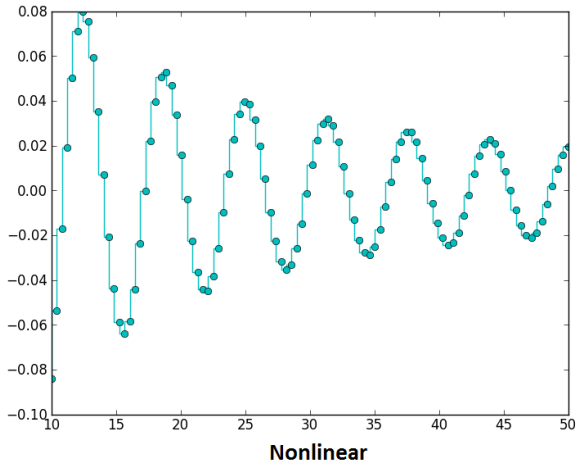

450
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
550
600


Random walk
localhost:8000/RandomWalk.html
Stop loading this page
Random Walk Simulation
550
500
450
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
550
600

Chapter 7: Predicting Gold Prices







$$X_t = T_t + S_t + V_t$$


$480.0
$460.0
$440.0
$420.0
$400.0
$380.0
$360.0
$340.0
$320.0
$300.0
$280.0
$260.0
2002
01 Jul 02
2003
01 Jul 03
2004
01 Jul 04

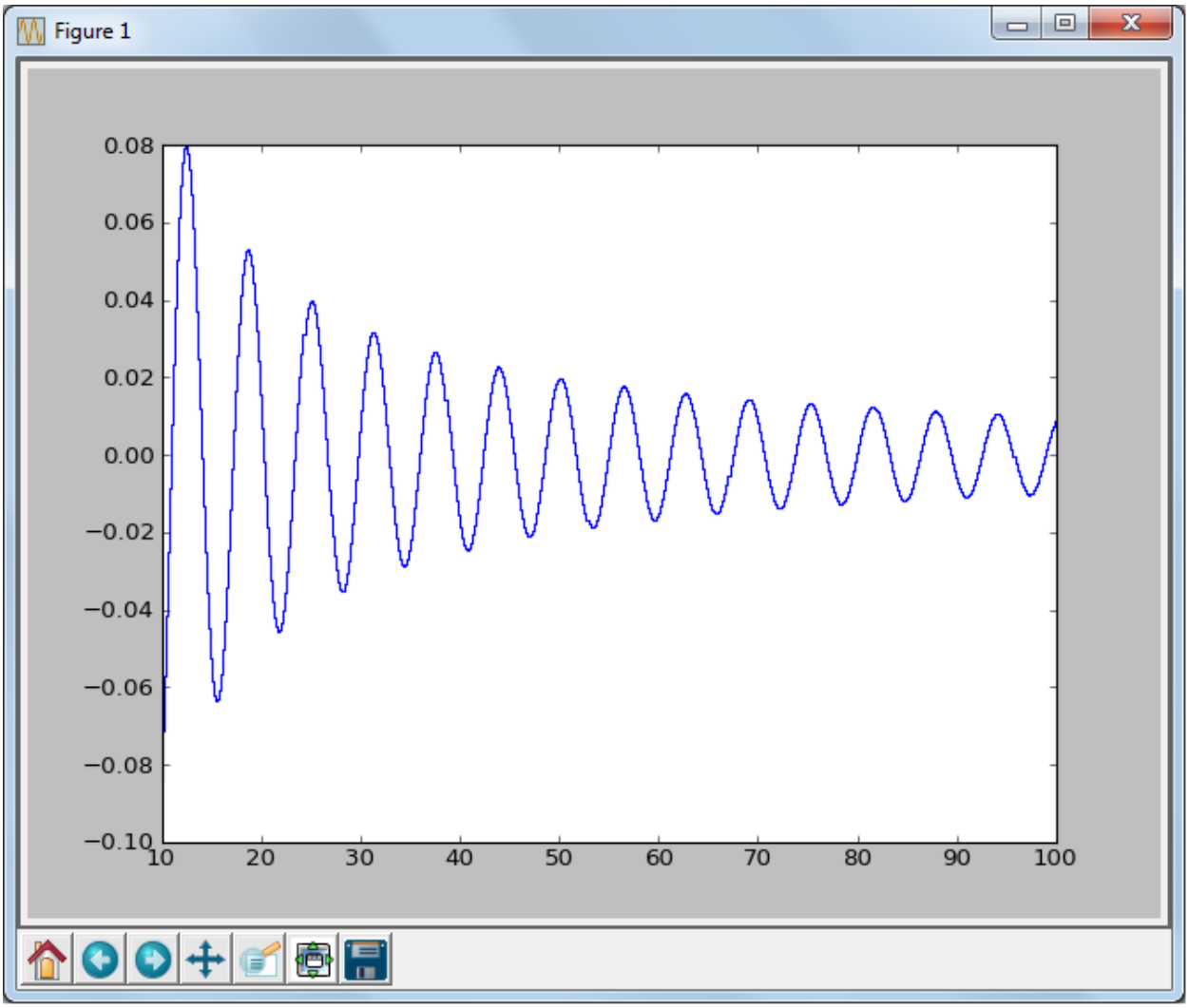
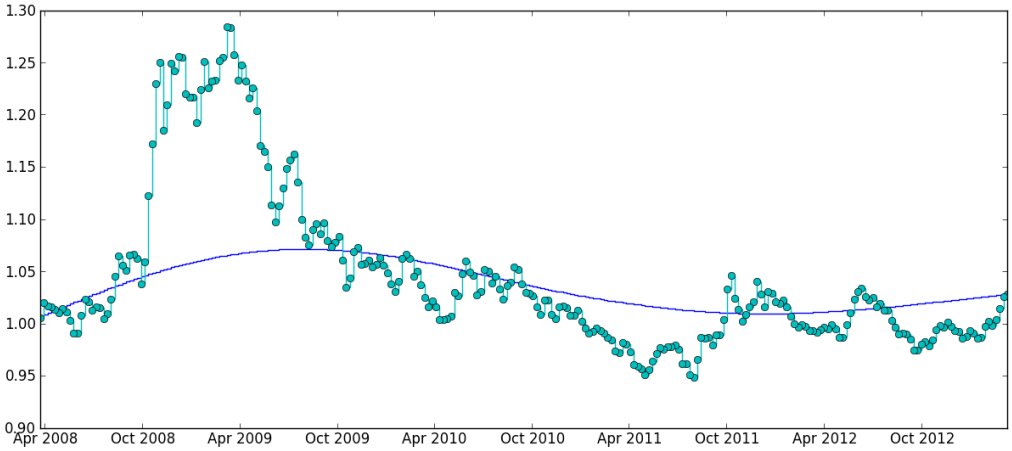
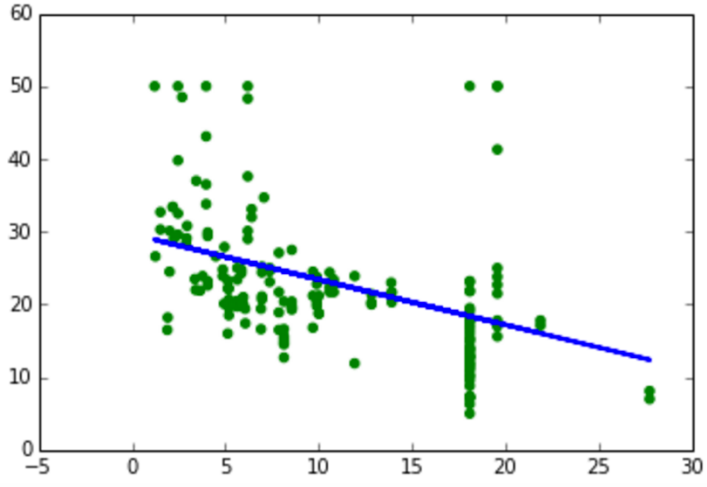
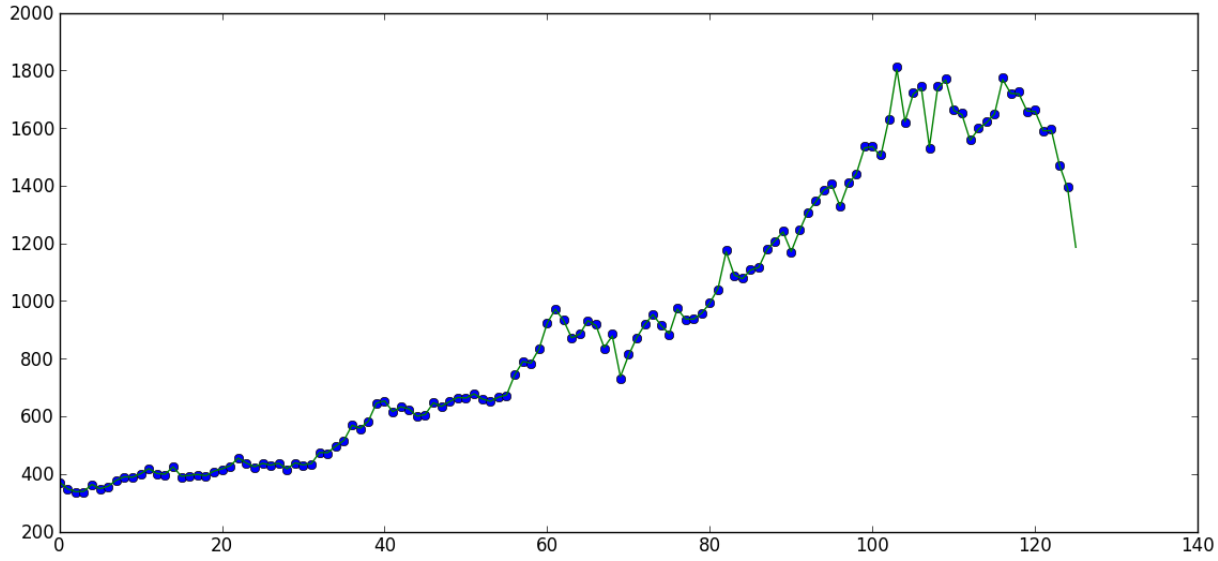


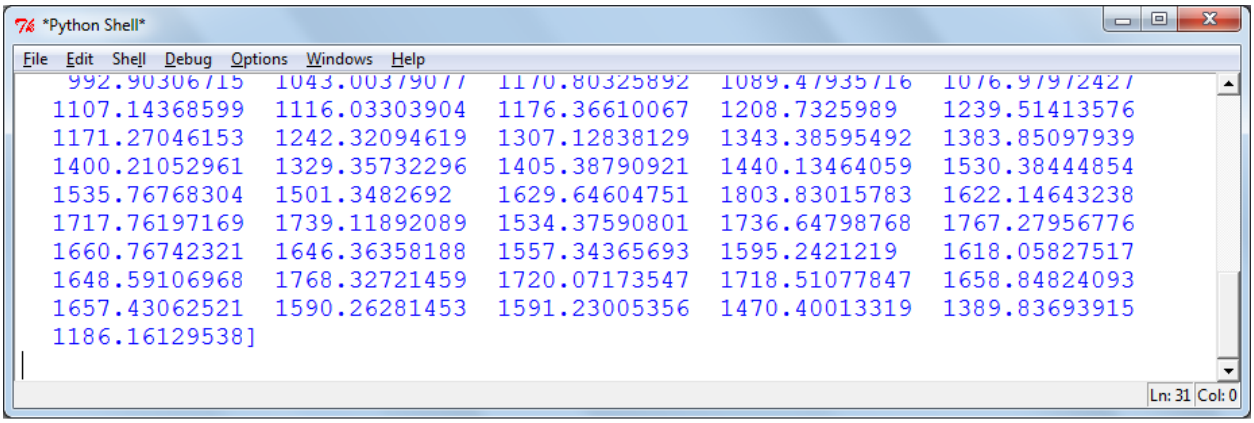
Figure 1
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
10
20
30
40
50
60
70
80
90
100

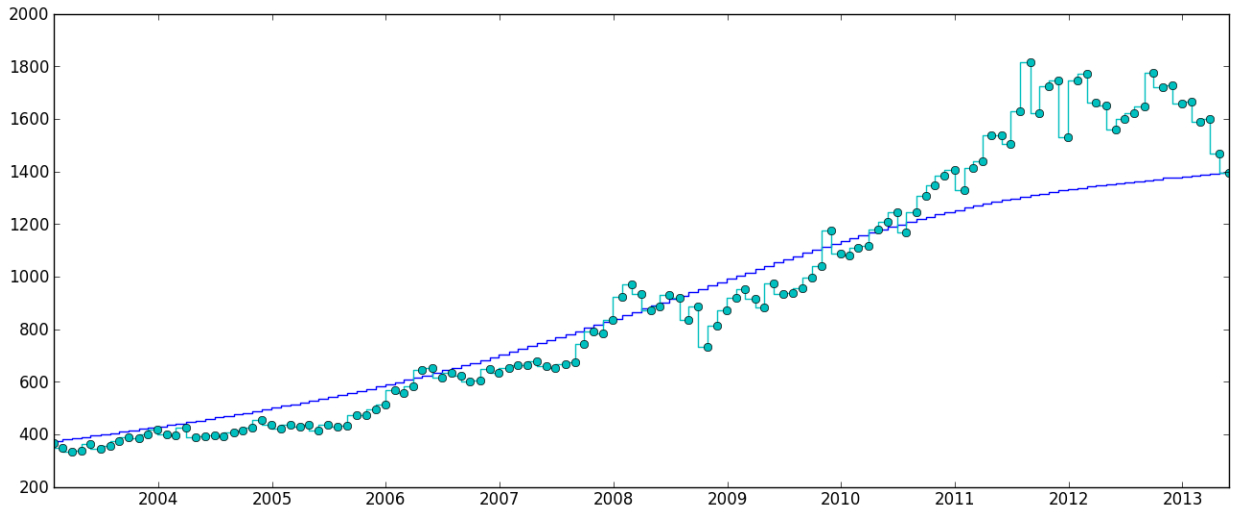

1.30
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
Apr 2008
Oct 2008
Apr 2009
Oct 2009
Apr 2010
Oct 2010
Apr 2011
Oct 2011
Apr 2012
Oct 2012

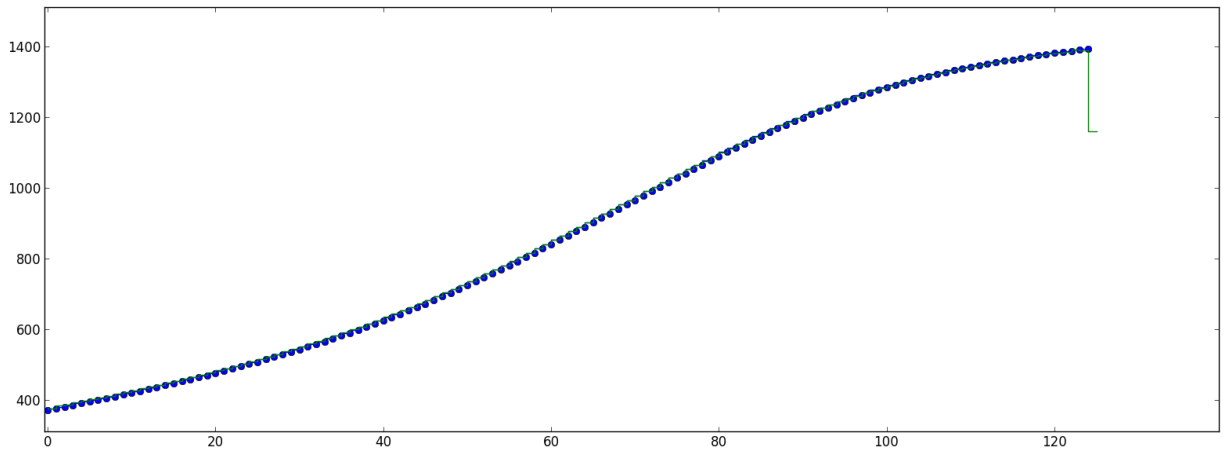
```
Populating the interactive namespace from numpy and matplotlib
```

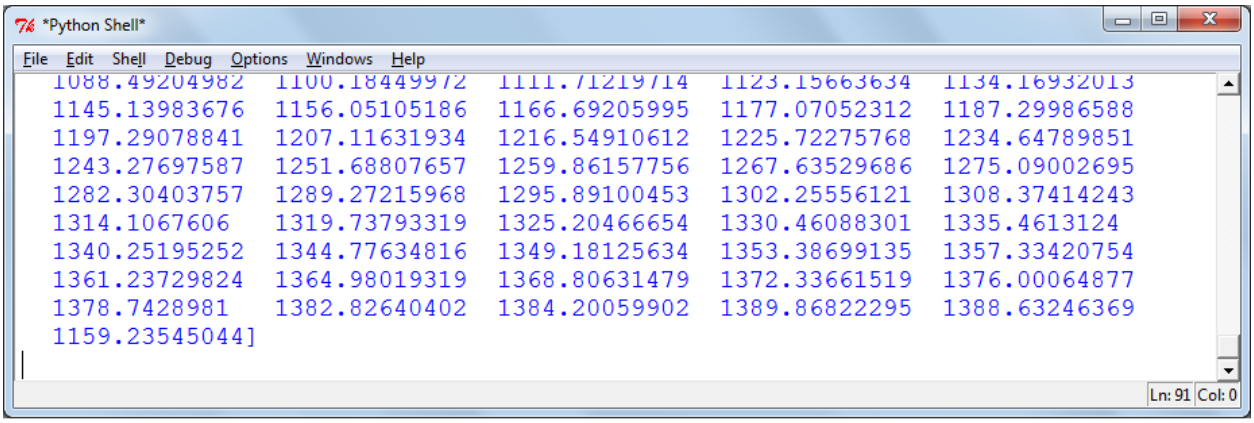


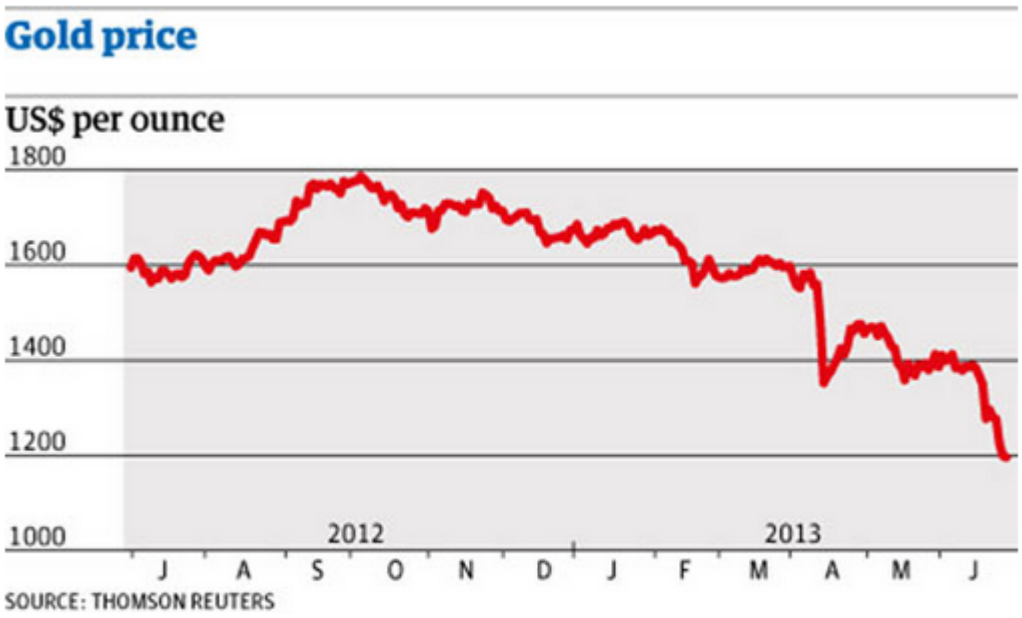



Python Shell
File Edit Shell Debug Options Windows Help
992.90306715 1043.00379077 1170.80325892 1089.47935716 1076.97972427
1107.14368599 1116.03303904 1176.36610067 1208.7325989 1239.51413576
1171.27046153 1242.32094619 1307.12838129 1343.38595492 1383.85097939
1400.21052961 1329.35732296 1405.38790921 1440.13464059 1530.38444854
1535.76768304 1501.3482692 1629.64604751 1803.83015783 1622.14643238
1717.76197169 1739.11892089 1534.37590801 1736.64798768 1767.27956776
1660.76742321 1646.36358188 1557.34365693 1595.2421219 1618.05827517
1648.59106968 1768.32721459 1720.07173547 1718.51077847 1658.84824093
1657.43062521 1590.26281453 1591.23005356 1470.40013319 1389.83693915
1186.16129538]
Ln: 31 Col: 0


2000
1800
1600
1400
1200
1000
800
600
400
200
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

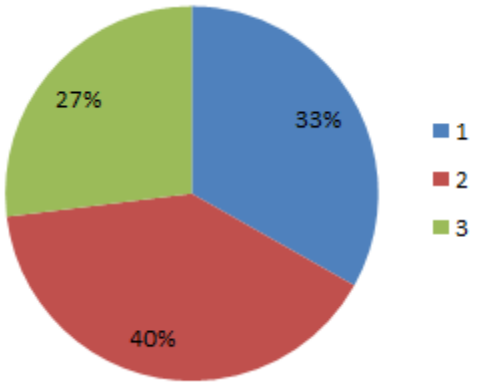

1400
1200
1000
800
600
400
0
20
40
60
80
100
120

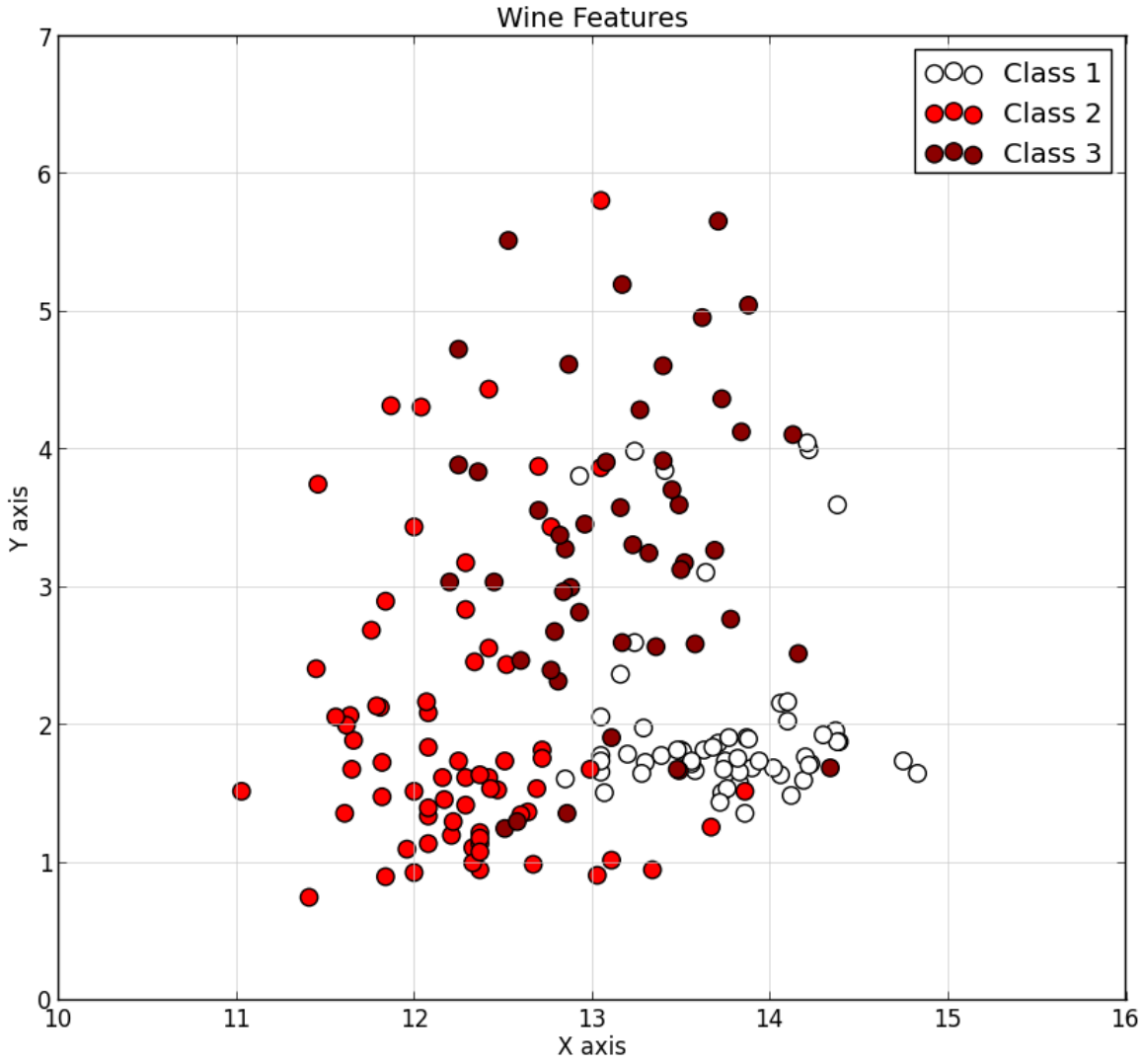


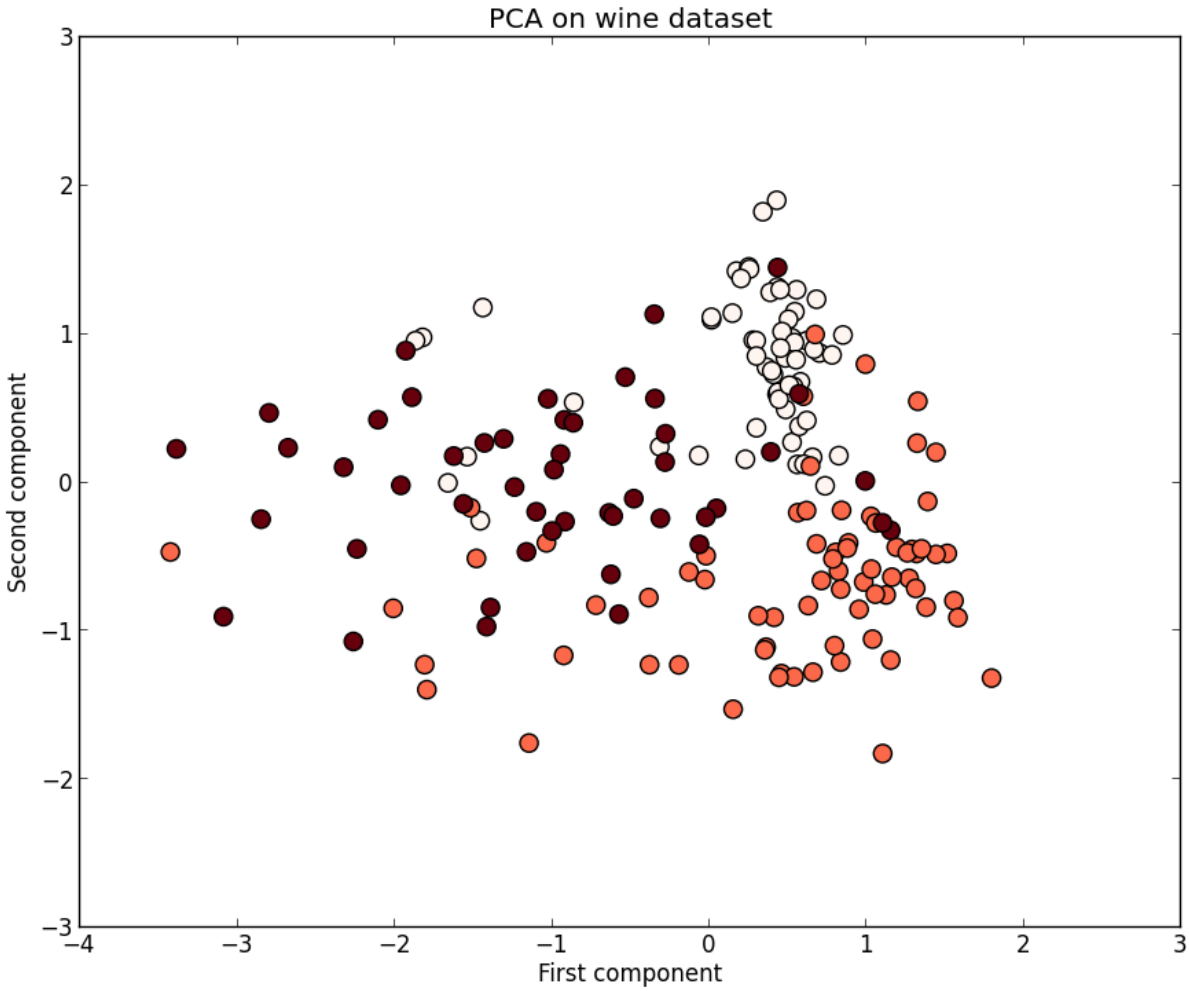


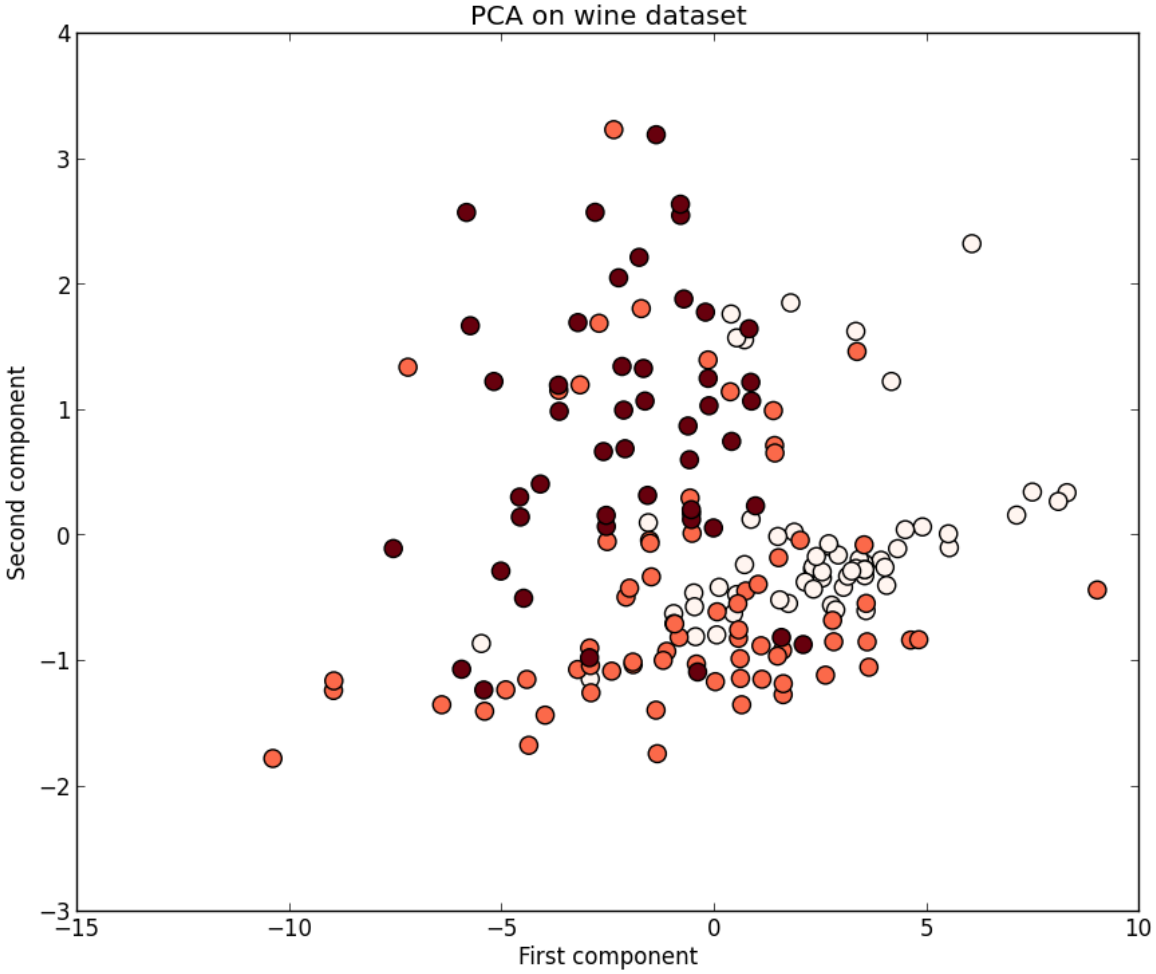
Gold prices at the end of June 2013 Photograph: Guardian/Thomson Reuters

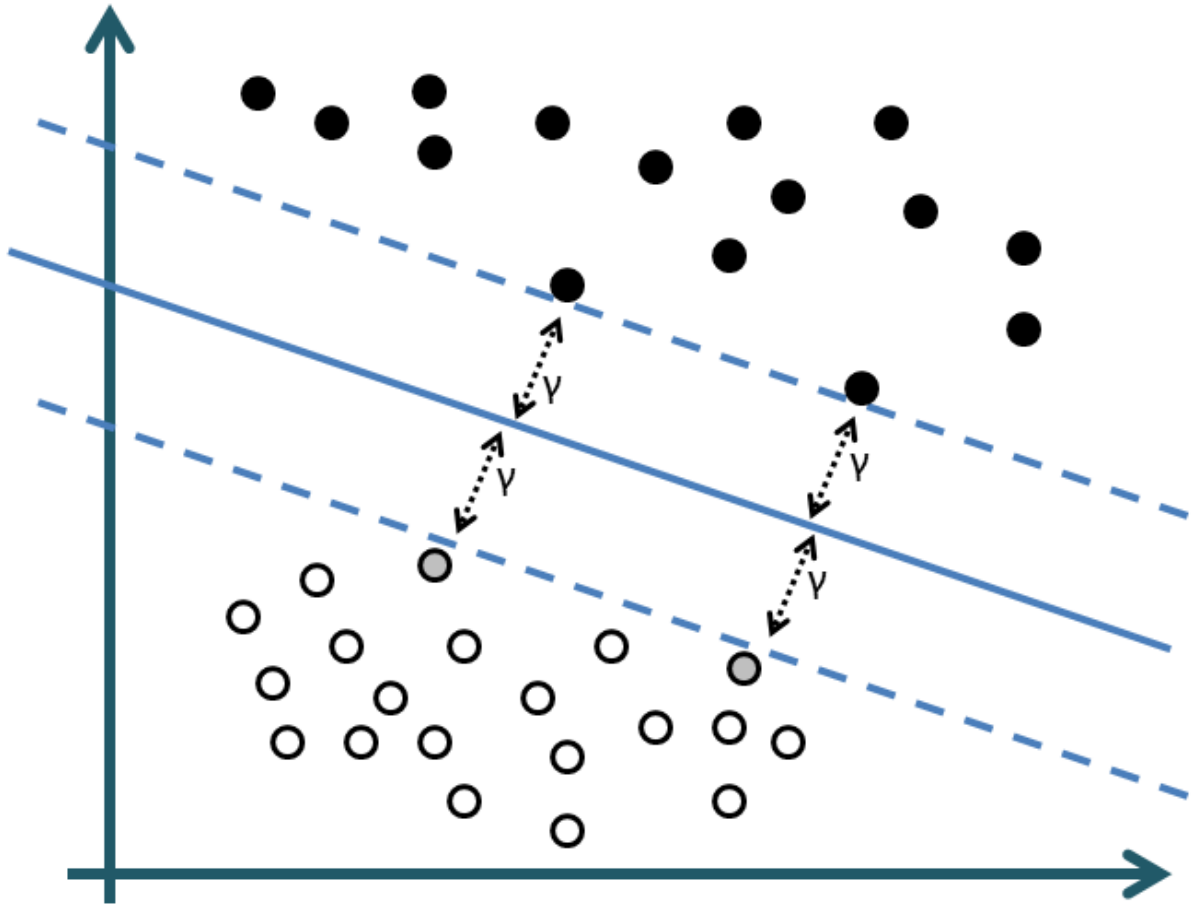
Chapter 8: Working with Support Vector Machines






PCA on wine dataset
Second component
First component
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3


PCA on wine dataset
Second component
First component
4
3
2
1
0
−1
−2
−3
−15
−10
−5
0
5
10



$$(< gamma * uT * v > + coef\,0)^{deg\,ree}$$

$$\exp\left(\frac{-(u-v)^2}{gamma}\right)$$

$$\exp\left(\frac{-(u-v)^2}{gamma}\right)$$

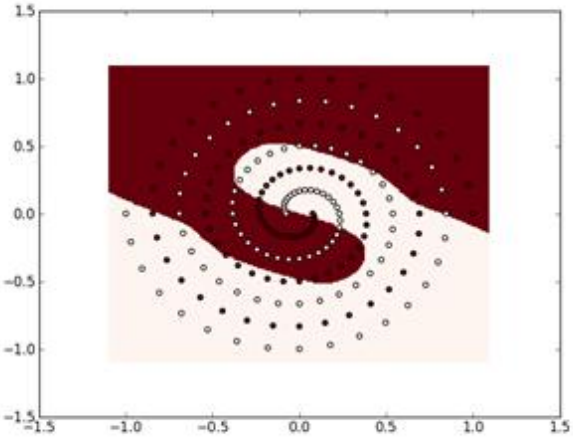
$$\tanh(gamma * uT * v + coef0)$$

$$\tanh(gamma * uT * v + coef0)$$

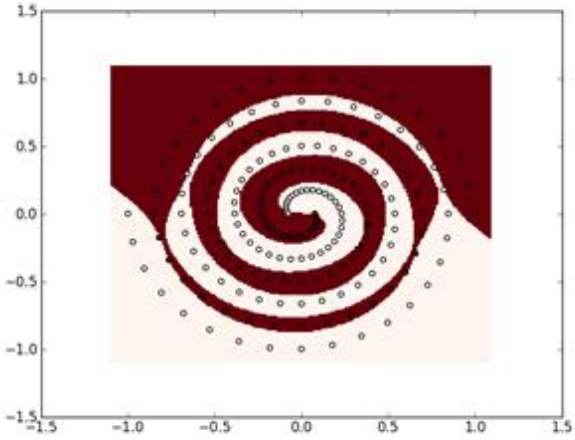
$$\frac{1}{\sqrt{(u - gamma)^2 + coef0^2}}$$

$$\frac{1}{\sqrt{(u - gamma)^2 + coef0^2}}$$

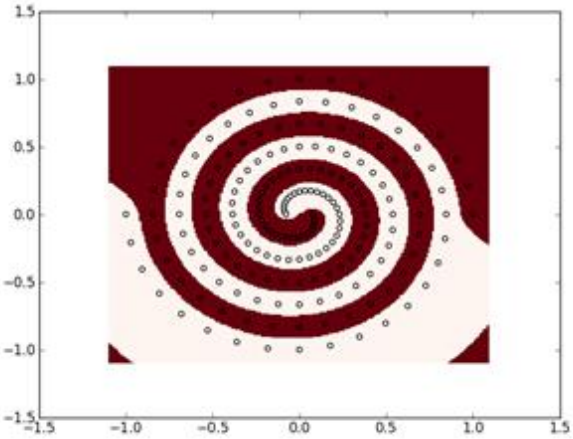
Gamma = 10



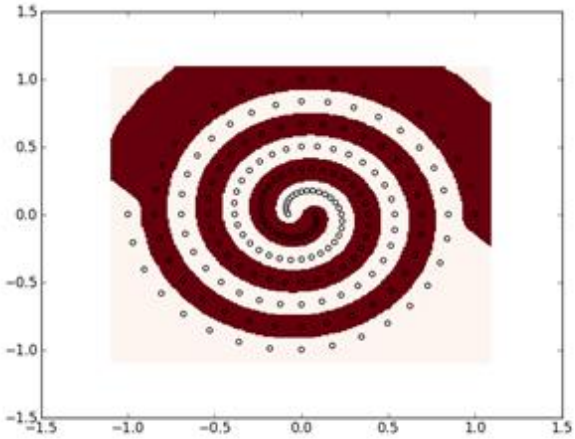
Gamma = 25

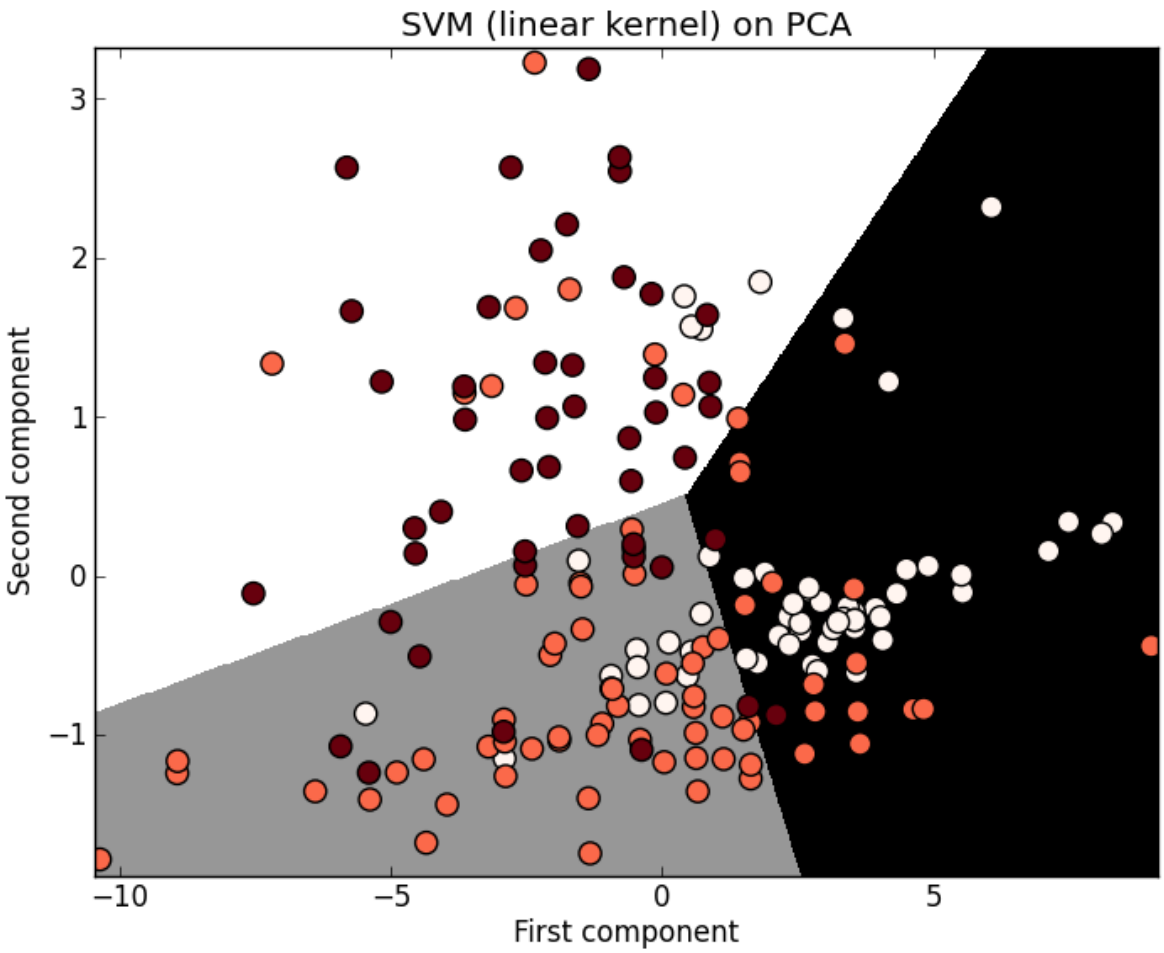


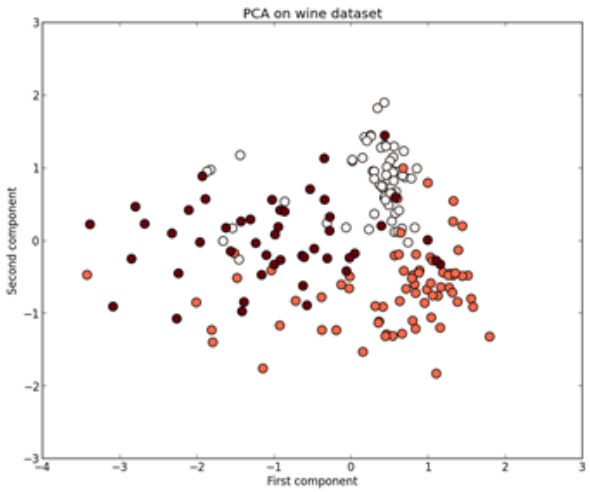
Gamma = 50

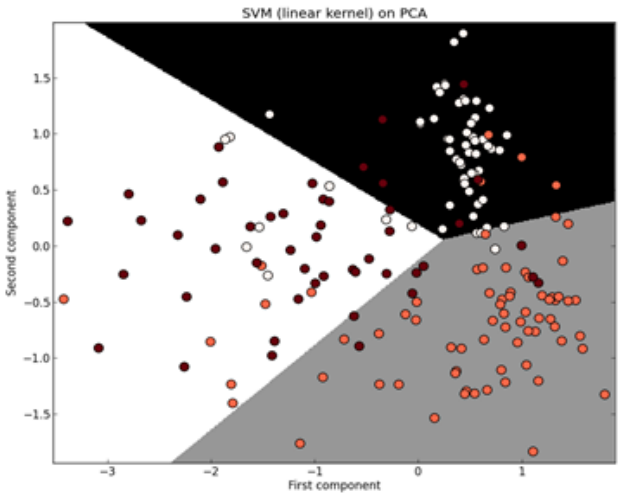


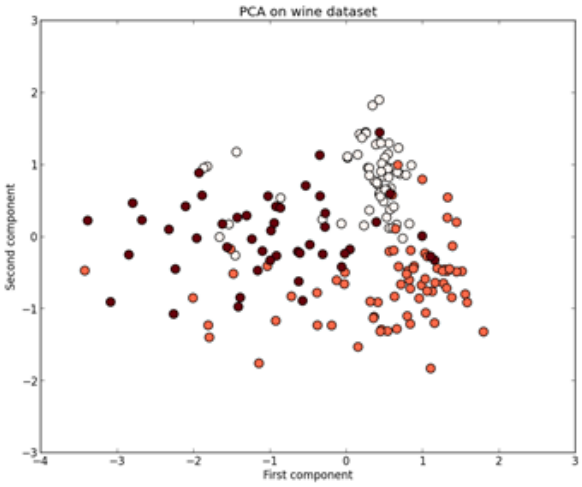
Gamma = 100

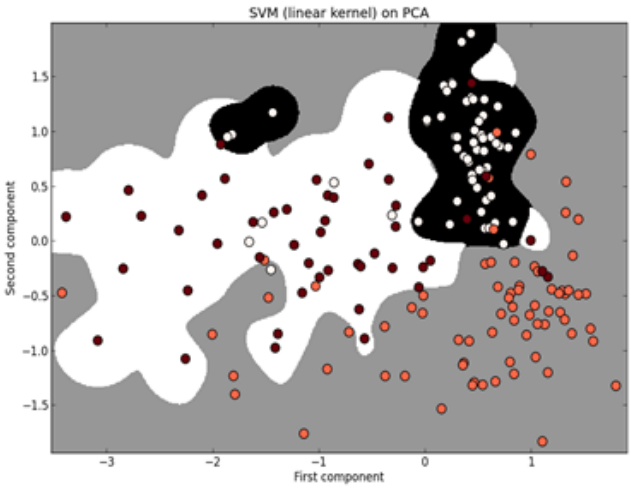



SVM (linear kernel) on PCA
Second component
First component
3
2
1
0
−1
−10
−5
0
5

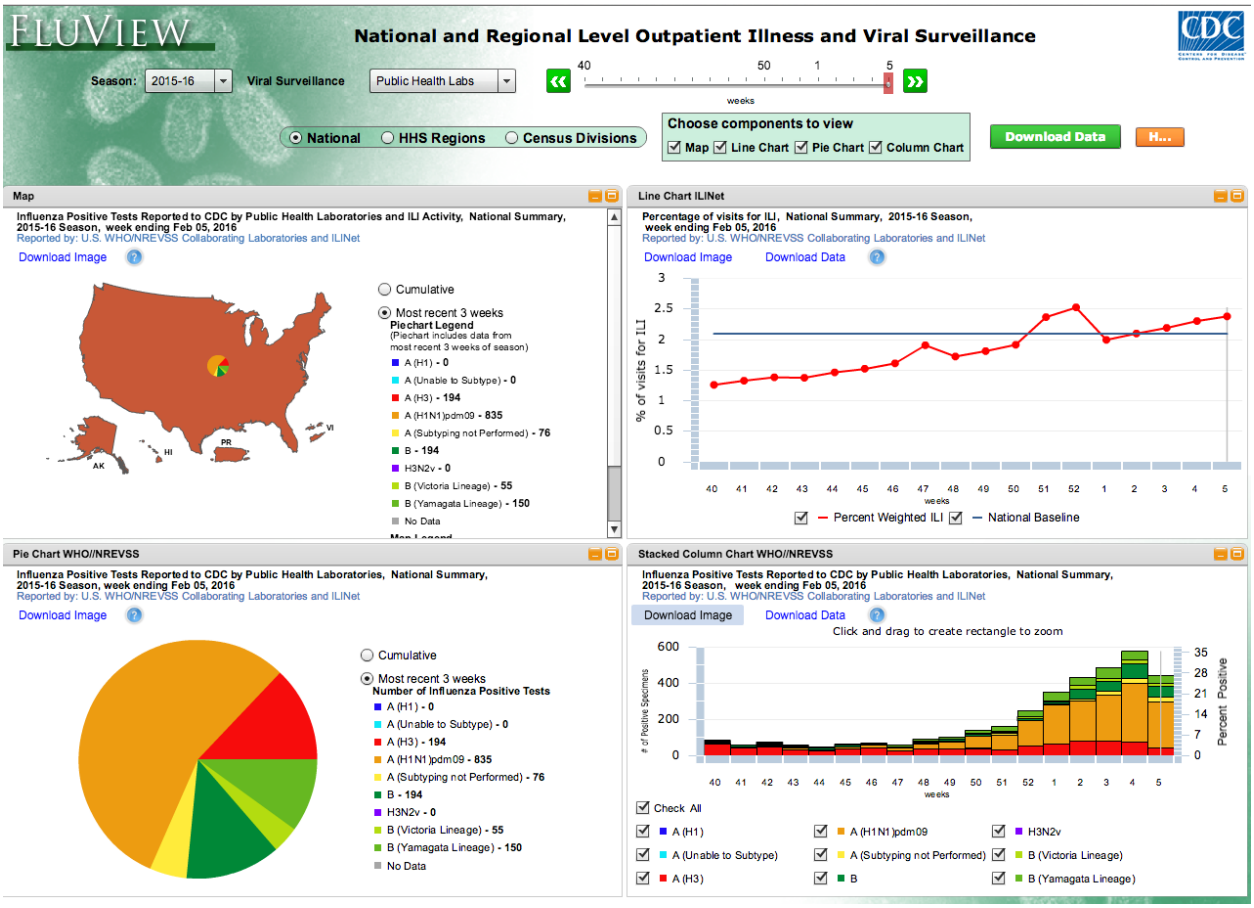

PCA on wine dataset
Second component
First component

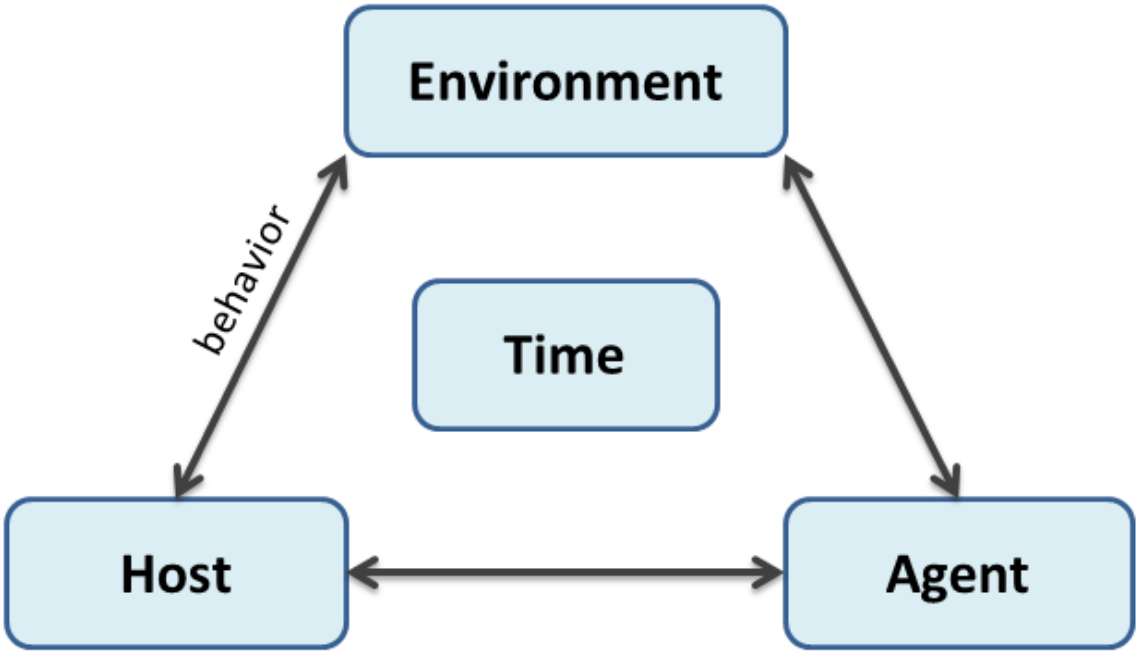

SVM (linear kernel) on PCA
Second component
First component

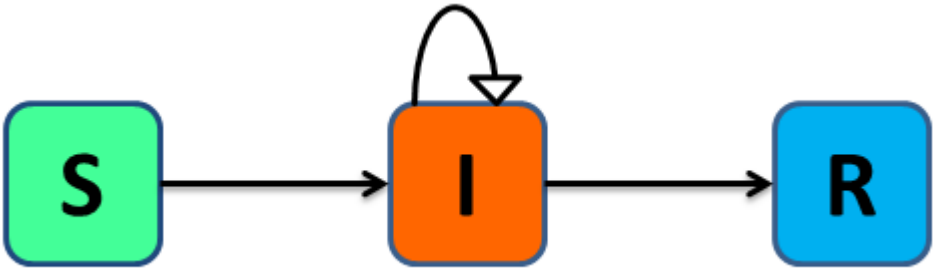

PCA on wine dataset
Second component
First component


SVM (linear kernel) on PCA
Second component
First component

Chapter 9: Modeling Infectious Diseases with Cellular Automata




Environment
behavior
Time
Host
Agent

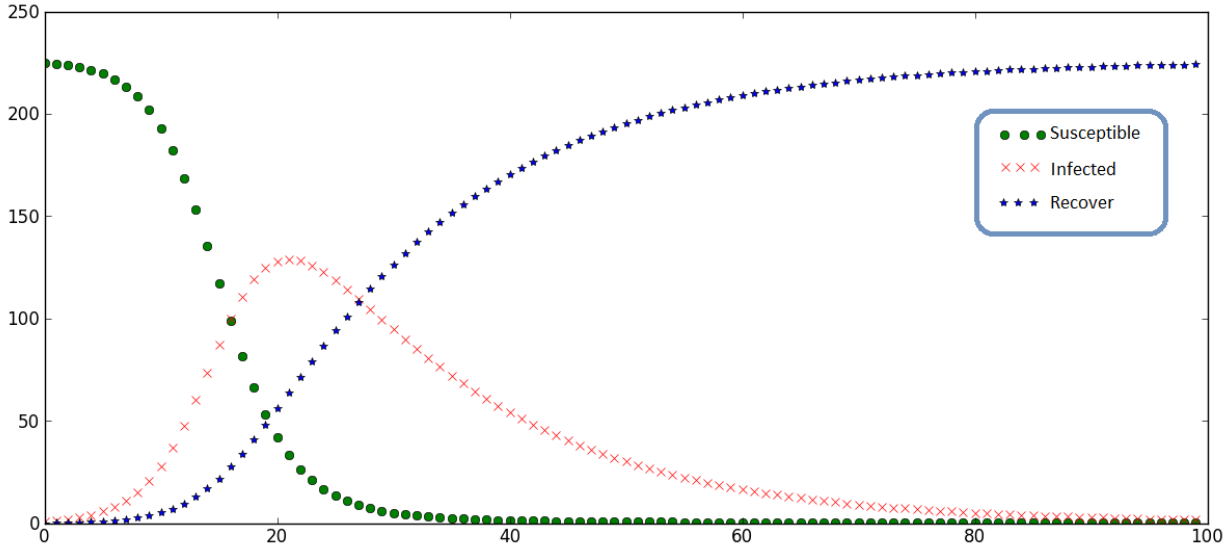

S
I
R

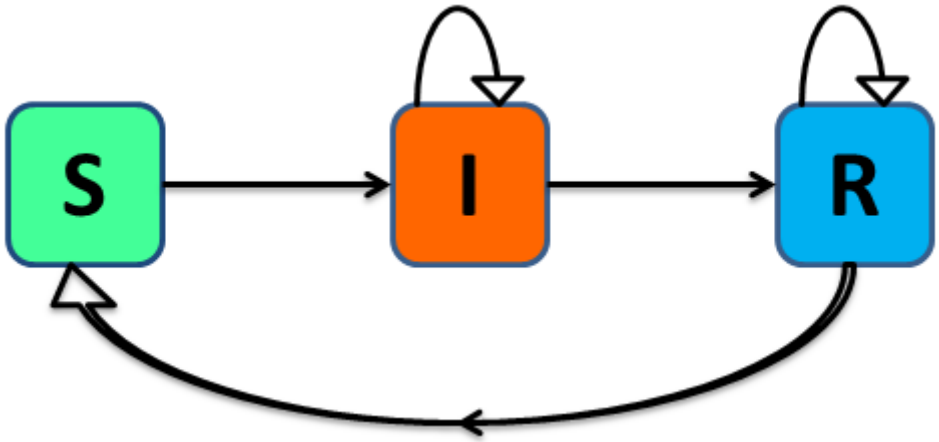
nm

(a) $$\frac{dS}{dt} = -\beta * S * I$$

(b) $$\frac{dI}{dt} = \beta * S * I - \gamma * I$$

(c) $$\frac{dR}{dt} = \gamma * I$$



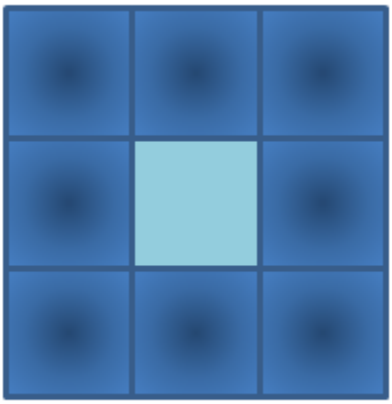


(a) $\frac{dS}{dt} = -\beta * S * I + \sigma * R$

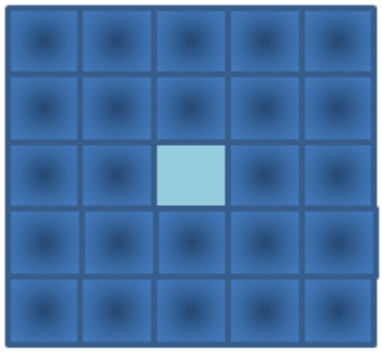
(b) $\frac{dI}{dt} = \beta * S * I - \gamma * I$

(c) $\frac{dR}{dt} = \gamma * I - \sigma * R$

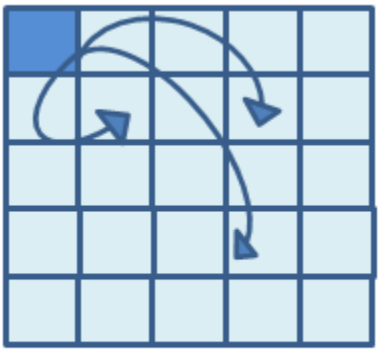
(a) Von Neumann



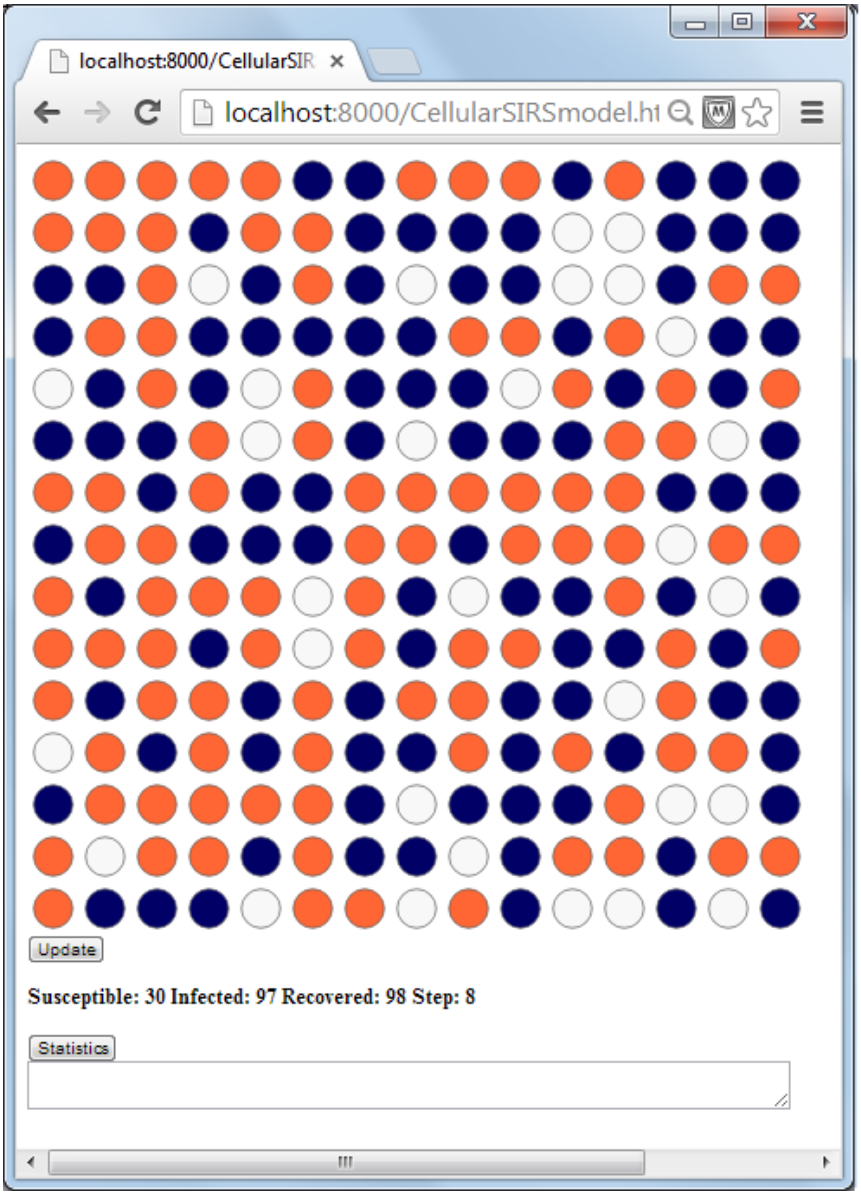
(b) Moore



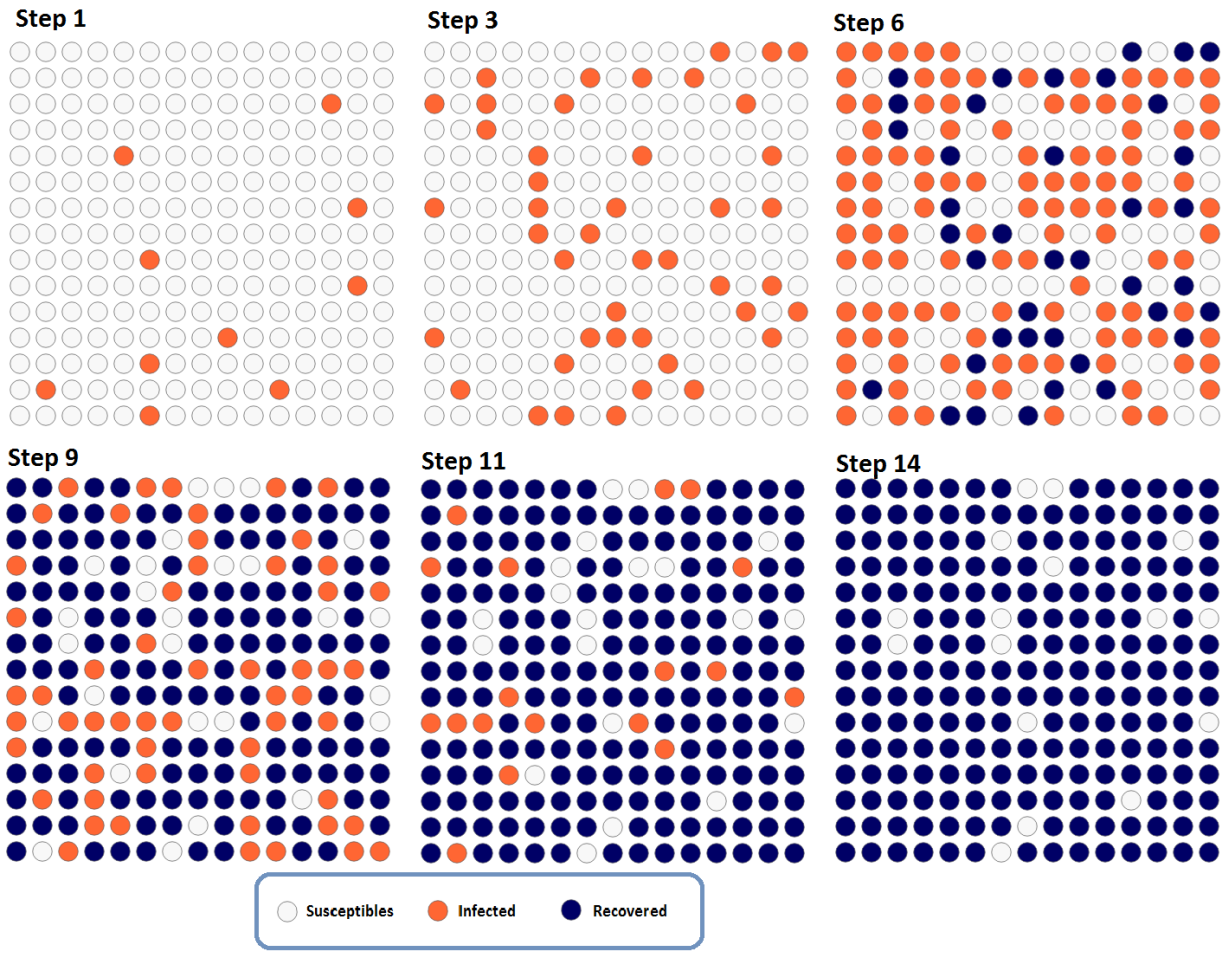
(c) Moore Extended

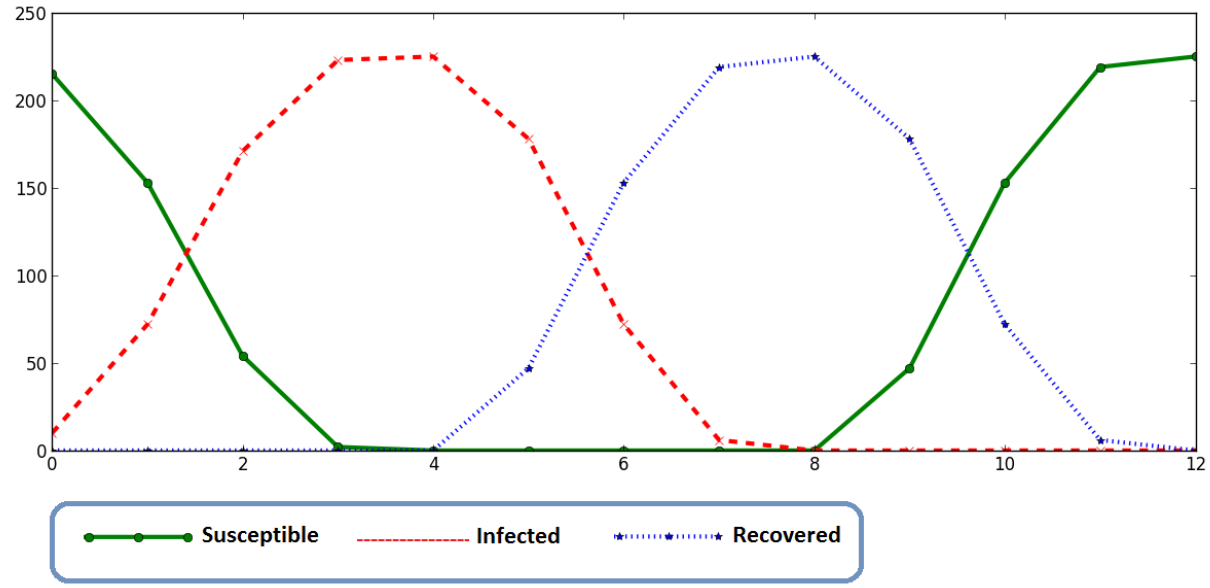


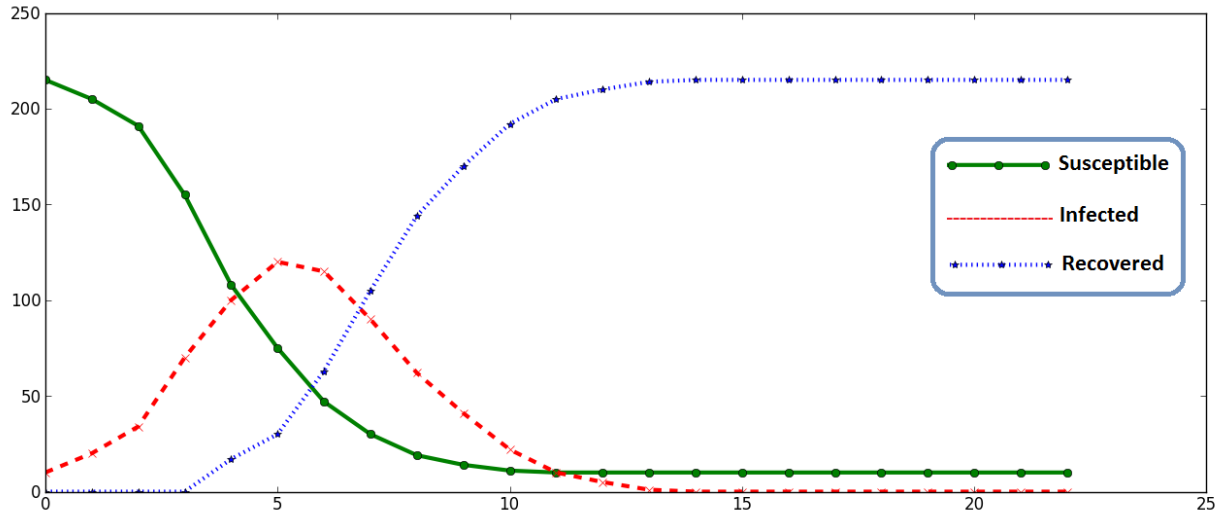
(d) Global


localhost:8000/CellularSIR
localhost:8000/CellularSIRSmodel.ht
Update
Susceptible: 30 Infected: 97 Recovered: 98 Step: 8
Statistics


Susceptibles
Infected
Recovered

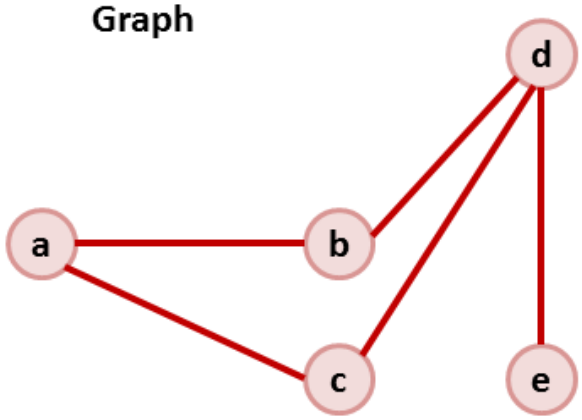

Step 1
Step 3
Step 6
Step 9
Step 11
Step 14
Susceptibles
Infected
Recovered


250
200
150
100
50
0
0
2
4
6
8
10
12
Susceptible
Infected
Recovered


250
200
150
100
50
0
0
5
10
15
20
25
Susceptible
Infected
Recovered

Chapter 10: Working with Social Graphs

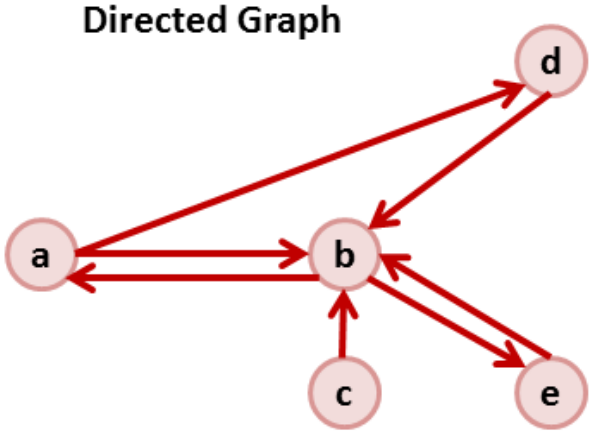
Graph



Adjacency matrix

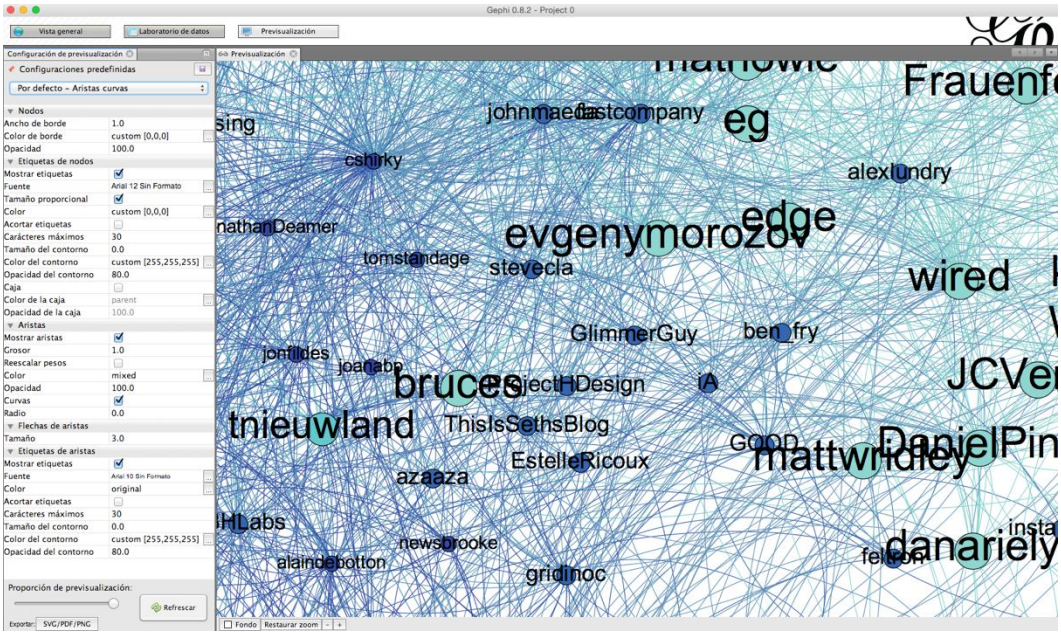
0	1	1	0	0
1	0	0	1	0
1	0	0	1	0
0	1	1	0	1
0	0	0	1	0

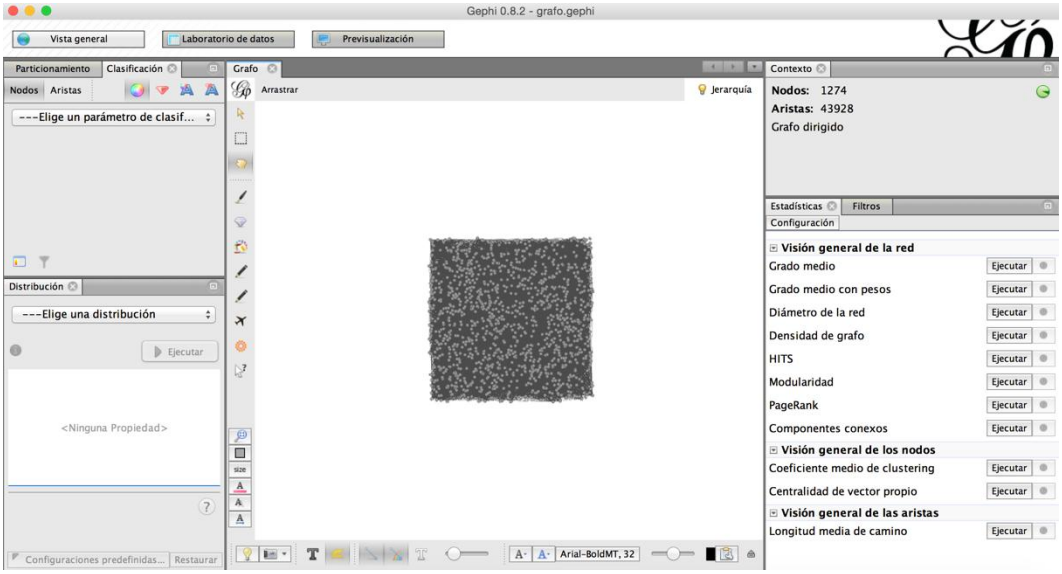
Directed Graph



Adjacency matrix

0	1	0	0	0
1	0	1	1	1
0	0	0	0	0
1	0	0	0	0
0	1	0	0	0


Gephi 0.8.2 - Project 0
Vista general
Laboratorio de datos
Previsualización
Configuración de previsualización
Configuraciones predefinidas
Por defecto – Aristas curvas
Nodos
Aristas
Etiquetas de nodos
Etiquetas de aristas
Flechas de aristas
Proporción de previsualización:
Refrescar
Exportar:
SVG/PDF/PNG
Fondo
Restaurar zoom
Frauenf
johnmaeda
fastcompany
eg
alexlundry
edge
evgenymorozov
wired
GlimmerGuy
ben_fry
JCVe
bruces
ObjectHDesign
tnieuwland
ThisIsSethsBlog
GOOD
mattwridley
DanielPin
EstelleRicoux
azaaza
newsbrooke
alaindebotton
gridinoc
danariely


Gephi 0.8.2 - grafo.gephi
Vista general
Laboratorio de datos
Previsualización
Particionamiento
Clasificación
Nodos
Aristas
---Elige un parámetro de clasif...
Distribución
---Elige una distribución
Ejecutar
<Ninguna Propiedad>
Configuraciones predefinidas...
Restaurar
Grafo
Arrastrar
Jerarquía
Arial-BoldMT, 32
Contexto
Nodos: 1274
Aristas: 43928
Grafo dirigido
Estadísticas
Filtros
Configuración
Visión general de la red
Grado medio
Grado medio con pesos
Diámetro de la red
Densidad de grafo
HITS
Modularidad
PageRank
Componentes conexos
Visión general de los nodos
Coeficiente medio de clustering
Centralidad de vector propio
Visión general de las aristas
Longitud media de camino
Ejecutar


Parámetros de Modularidad
Modularidad
Algoritmo de detección de comunidades.
Aleatorio
Produce una mejor descomposición pero auménta el tiempo de cómputo
Utilizar pesos
Utilizar peso de aristas
Resolución:
1.0
Menor para obtener más comunidades (más pequeñas) y mayor que 1.0 para obtener menos comunidades (más grandes).
Cancelar
Aceptar

HTML Report

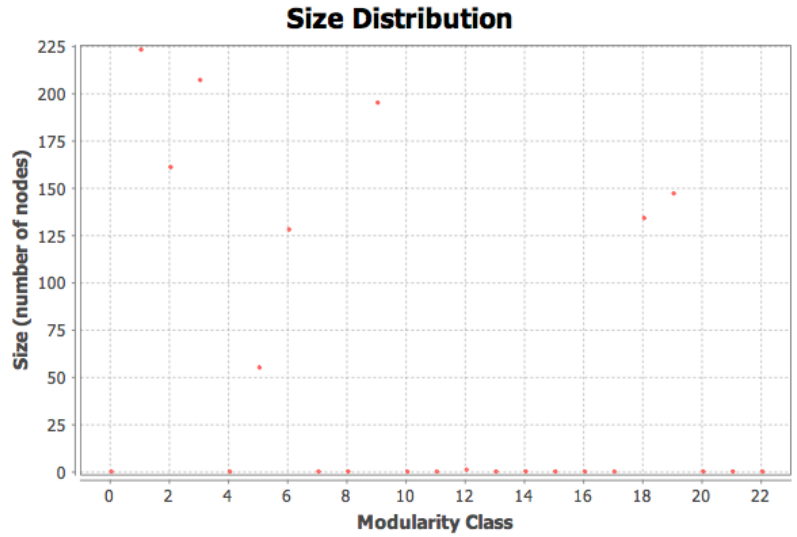
Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0,333
Modularity with resolution: 0,333
Number of Communities: 23



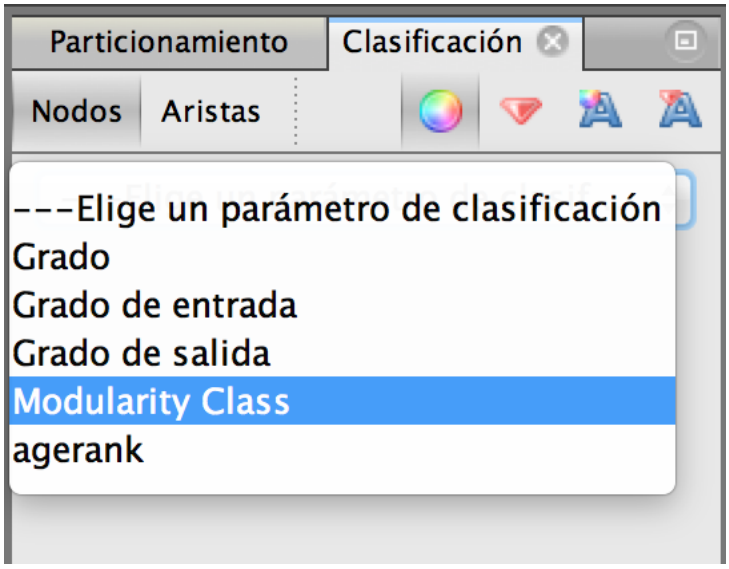
Algorithm:

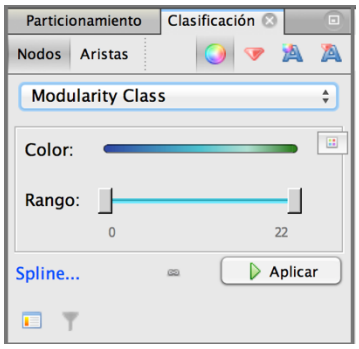
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

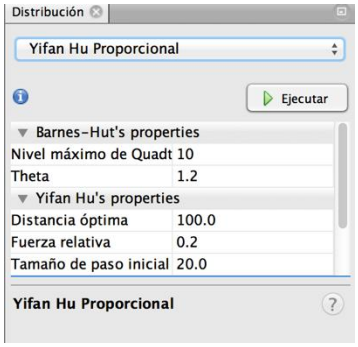
Resolution:

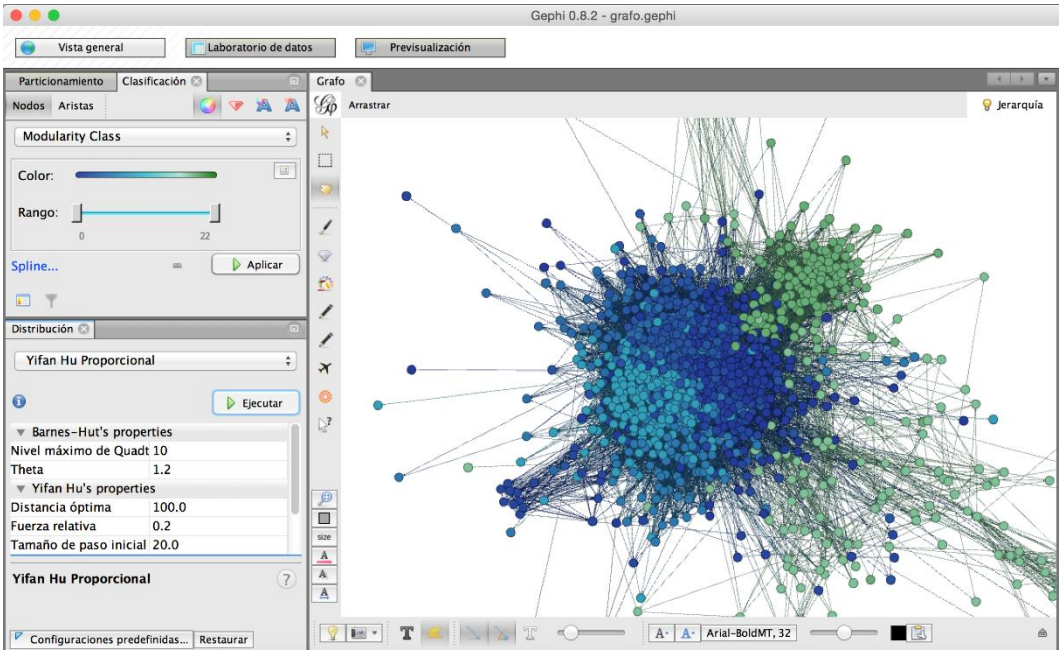
R. Lambiotte, J.-C. Delvenne, M. Barahona *Laplacian Dynamics and Multiscale Modular Structure in Networks 2009*

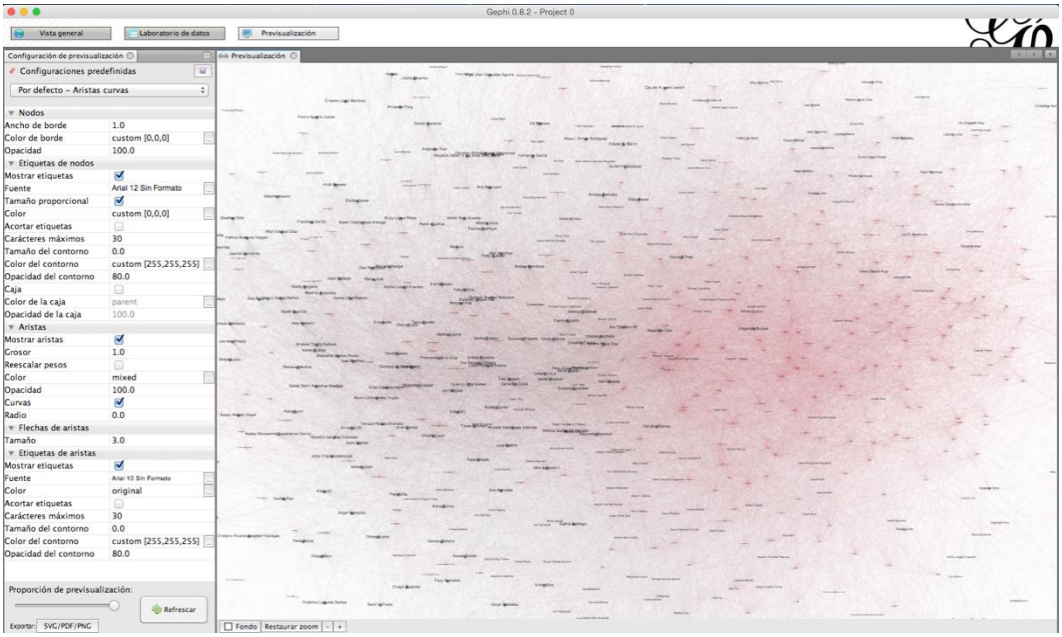
Imprimir | Copiar | Guardar | Cerrar

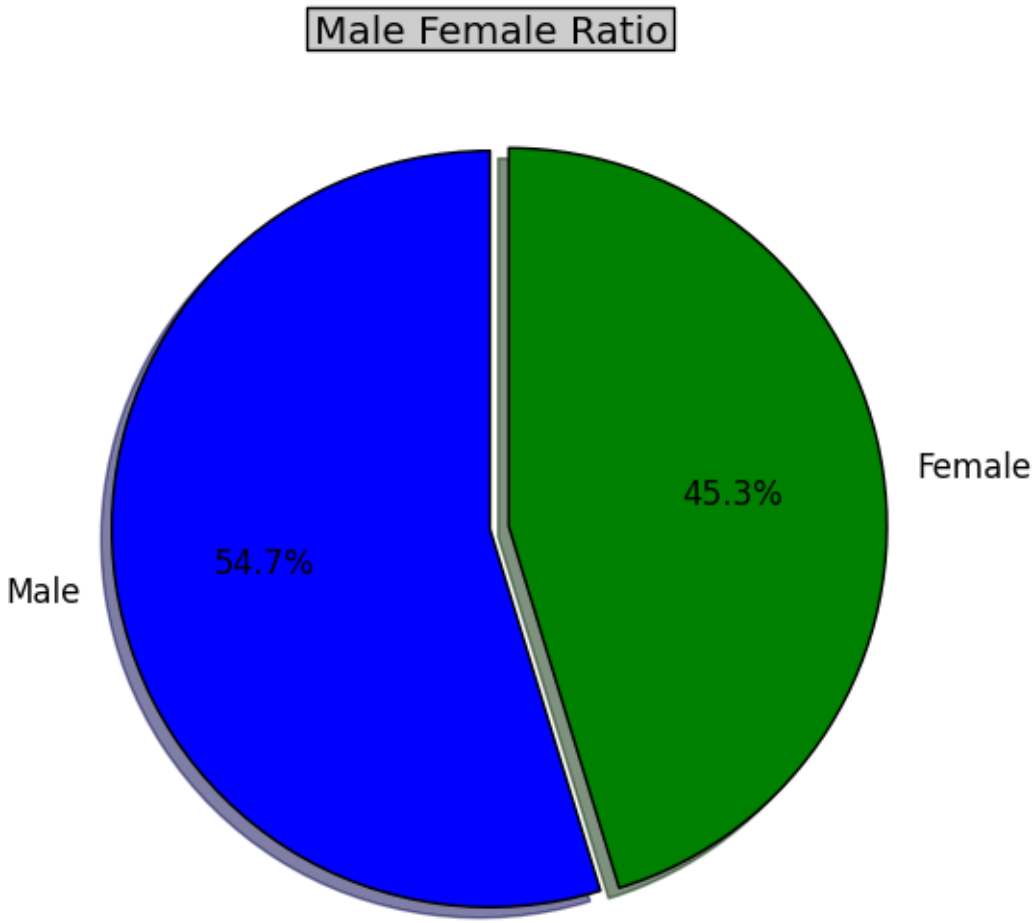

Particionamiento
Clasificación
Nodos
Aristas
---Elige un parámetro de clasificación
Grado
Grado de entrada
Grado de salida
Modularity Class
agerank

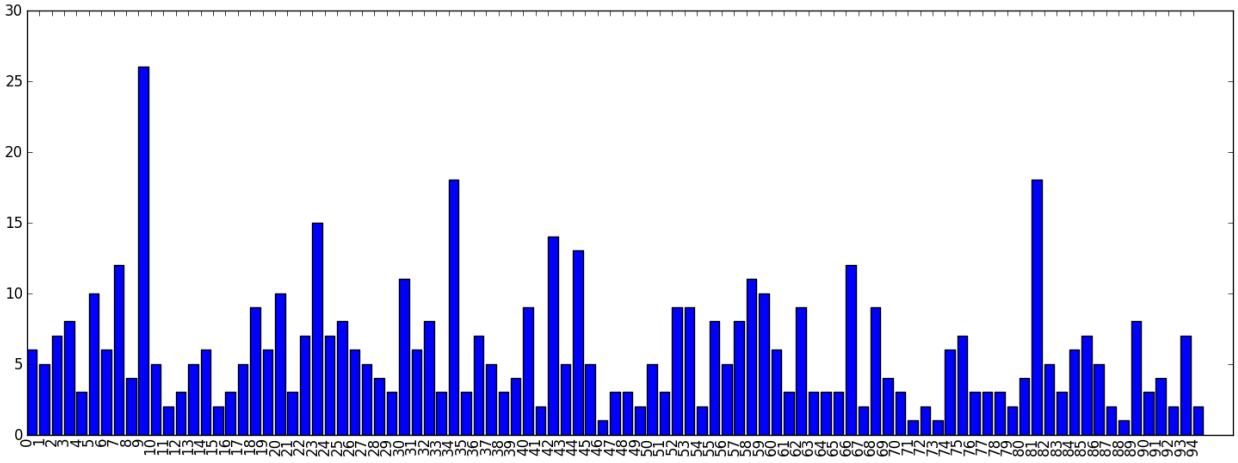

Particionamiento
Clasificación
Nodos
Aristas
Modularity Class
Color:
Rango:
0
22
Spline...
Aplicar

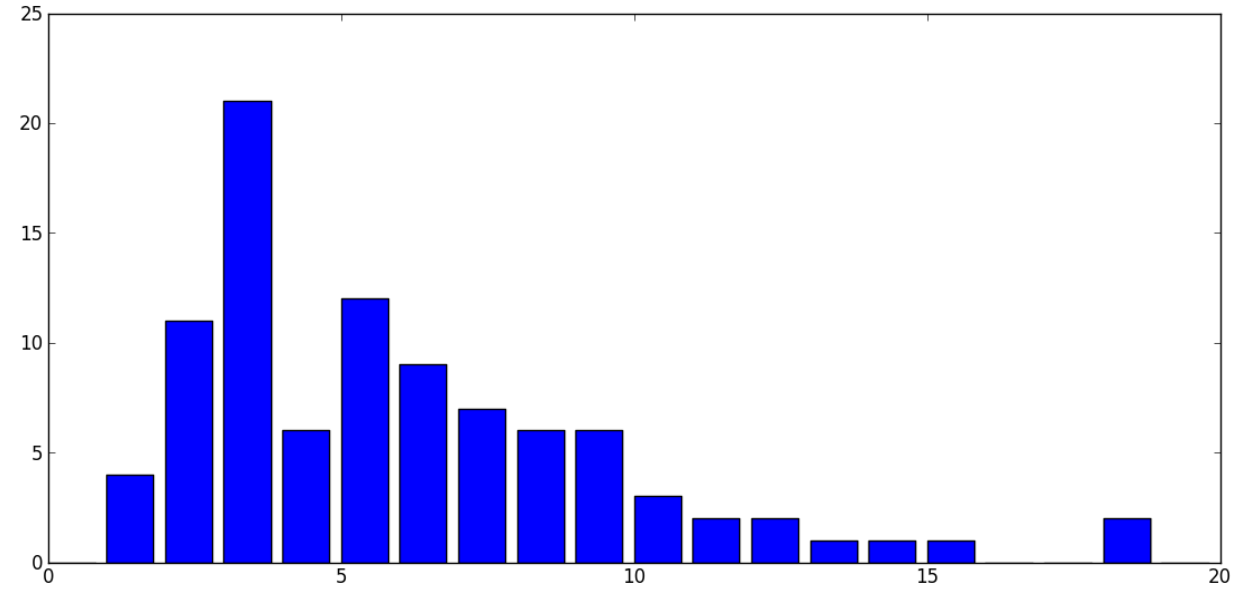

Distribución
Yifan Hu Proporcional
Ejecutar
Barnes-Hut's properties
Nivel máximo de Quadt 10
Theta 1.2
Yifan Hu's properties
Distancia óptima 100.0
Fuerza relativa 0.2
Tamaño de paso inicial 20.0
Yifan Hu Proporcional

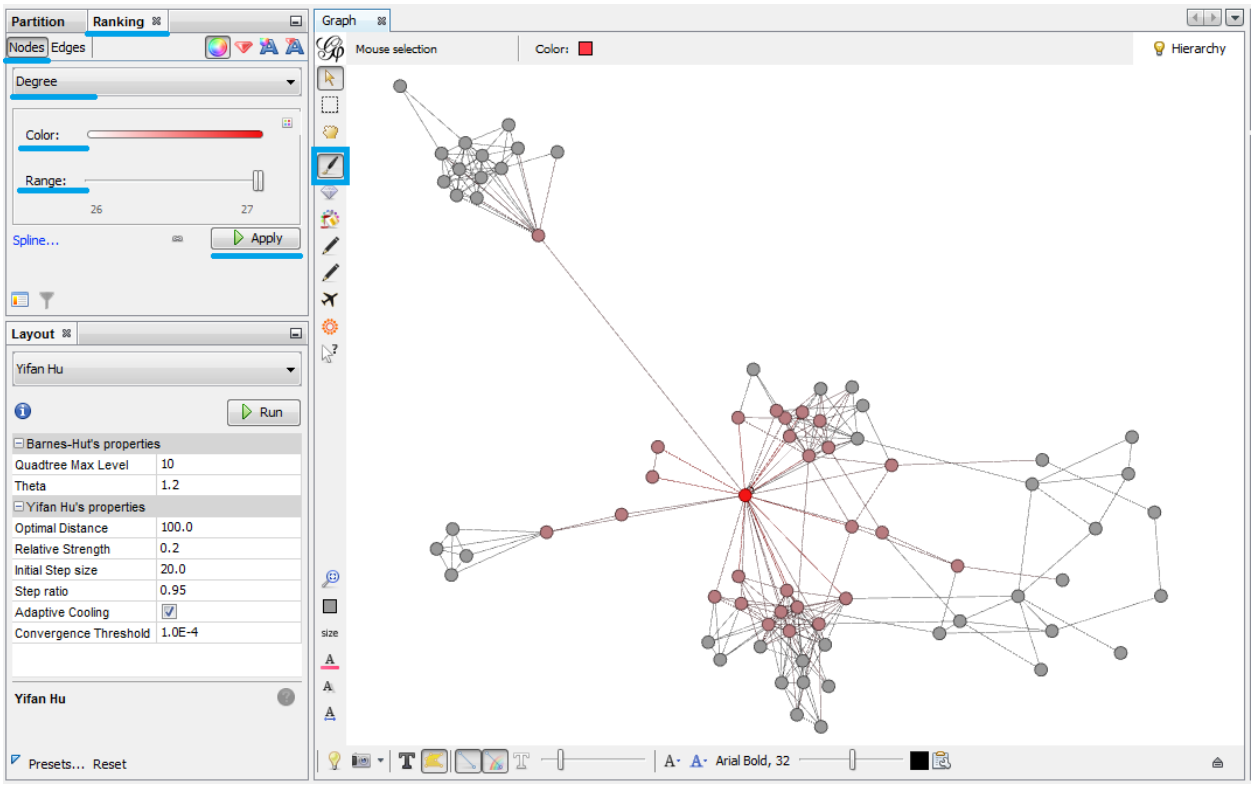

Gephi 0.8.2 - grafo.gephi
Vista general
Laboratorio de datos
Previsualización
Particionamiento
Clasificación
Nodos
Aristas
Modularity Class
Color:
Rango:
0
22
Spline...
Aplicar
Distribución
Yifan Hu Proporcional
Ejecutar
Barnes-Hut's properties
Nivel máximo de Quadt 10
Theta 1.2
Yifan Hu's properties
Distancia óptima 100.0
Fuerza relativa 0.2
Tamaño de paso inicial 20.0
Yifan Hu Proporcional
Configuraciones predefinidas...
Restaurar
Grafo
Arrastrar
Jerarquía
Arial-BoldMT, 32

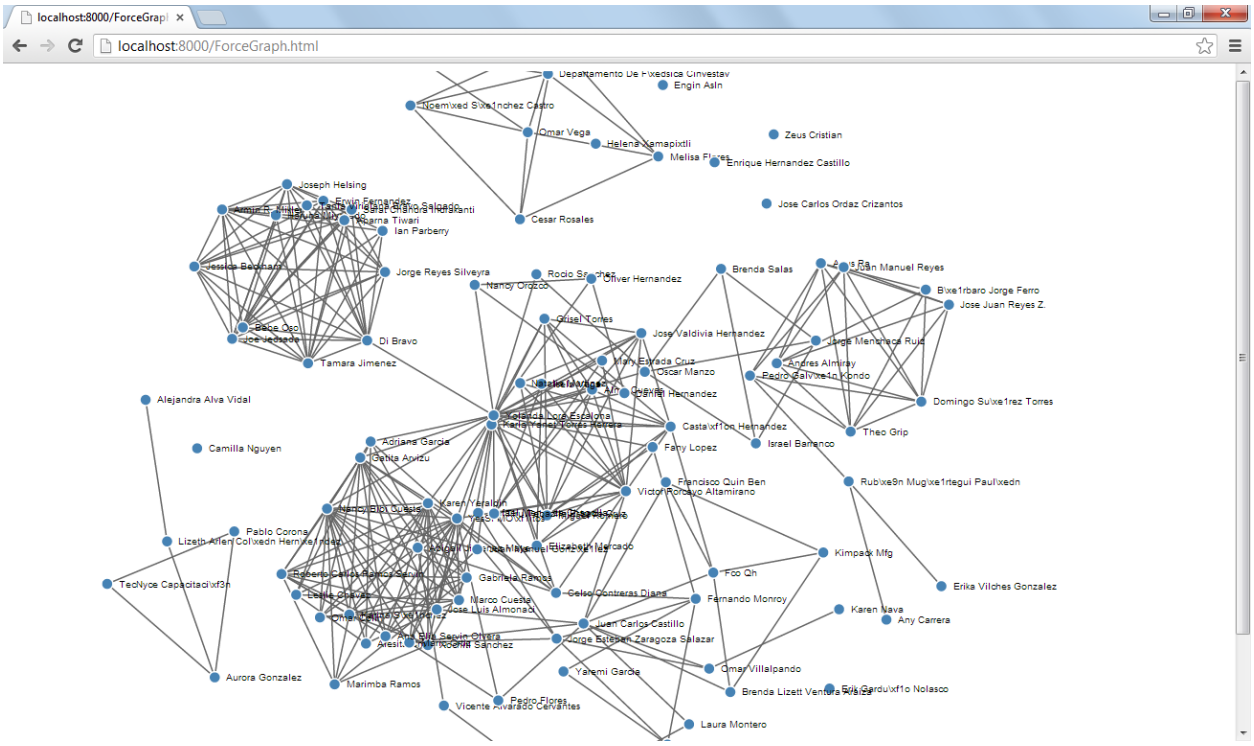

Gephi 0.8.2 - Project 0
Vista general
Laboratorio de datos
Previsualización
Configuración de previsualización
Configuraciones predefinidas
Por defecto – Aristas curvas
Nodos
Ancho de borde 1.0
Color de borde custom [0,0,0]
Opacidad 100.0
Etiquetas de nodos
Mostrar etiquetas
Fuente Arial 12 Sin Formato
Tamaño proporcional
Color custom [0,0,0]
Acortar etiquetas
Carácteres máximos 30
Tamaño del contorno 0.0
Color del contorno custom [255,255,255]
Opacidad del contorno 80.0
Caja
Color de la caja parent
Opacidad de la caja 100.0
Aristas
Mostrar aristas
Grosor 1.0
Reescalar pesos
Color mixed
Opacidad 100.0
Curvas
Radio 0.0
Flechas de aristas
Tamaño 3.0
Etiquetas de aristas
Mostrar etiquetas
Fuente Arial 10 Sin Formato
Color original
Acortar etiquetas
Carácteres máximos 30
Tamaño del contorno 0.0
Color del contorno custom [255,255,255]
Opacidad del contorno 80.0
Proporción de previsualización:
Refrescar
Exportar:
SVG/PDF/PNG
Previsualización
Fondo
Restaurar zoom


Male Female Ratio
Female
45.3%
54.7%
Male

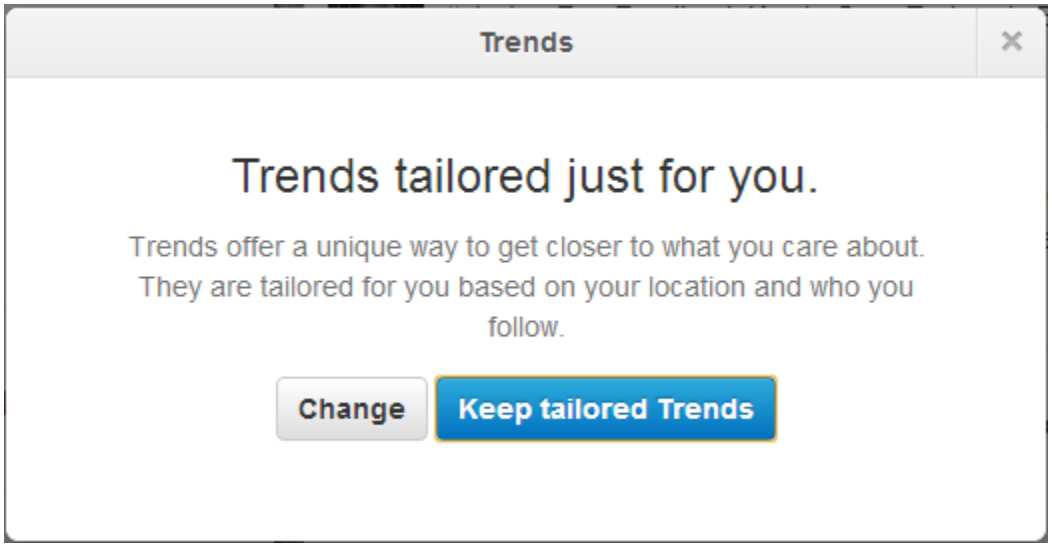

30
25
20
15
10
5
0


25
20
15
10
5
0
0
5
10
15
20


Partition
Ranking
Nodes
Edges
Degree
Color:
Range:
26
27
Spline...
Apply
Layout
Yifan Hu
Run
Barnes-Hut's properties
Quadtree Max Level
10
Theta
1.2
Yifan Hu's properties
Optimal Distance
100.0
Relative Strength
0.2
Initial Step size
20.0
Step ratio
0.95
Adaptive Cooling
Convergence Threshold
1.0E-4
Yifan Hu
Presets...
Reset
Graph
Mouse selection
Color:
Hierarchy
size
Arial Bold, 32


localhost:8000/ForceGrapl
localhost:8000/ForceGraph.html
Engin Asln
Zeus Cristian
Enrique Hernandez Castillo
Jose Carlos Ordaz Crizantos
Cesar Rosales
Joseph Helsing
Ian Parberry
Jorge Reyes Silveyra
Nancy Orozco
Oliver Hernandez
Brenda Salas
Jose Juan Reyes Z.
Jose Valdivia Hernandez
Oscar Manzo
Tamara Jimenez
Alejandra Alva Vidal
Domingo Sulxe1rez Torres
Theo Grip
Israel Barranco
Camilla Nguyen
Adriana Garcia
Fany Lopez
Francisco Quin Ben
Kimpack Mfg
Erika Vilches Gonzalez
Fernando Monroy
Karen Nava
Any Carrera
Juan Carlos Castillo
Omar Villalpando
Laura Montero
Aurora Gonzalez
Marimba Ramos
Pedro Flores

Chapter 11: Working with Twitter Data



Sign in with your Twitter account

Please log in to access that page.

Username: *

New to Twitter? Sign up

Password: *

Log in

My applications

Create a new application

Looks like you haven't created any applications yet!

OAuth settings

Your application's OAuth settings. Keep the "Consumer secret" a secret. This key should never be human-readable in your application.

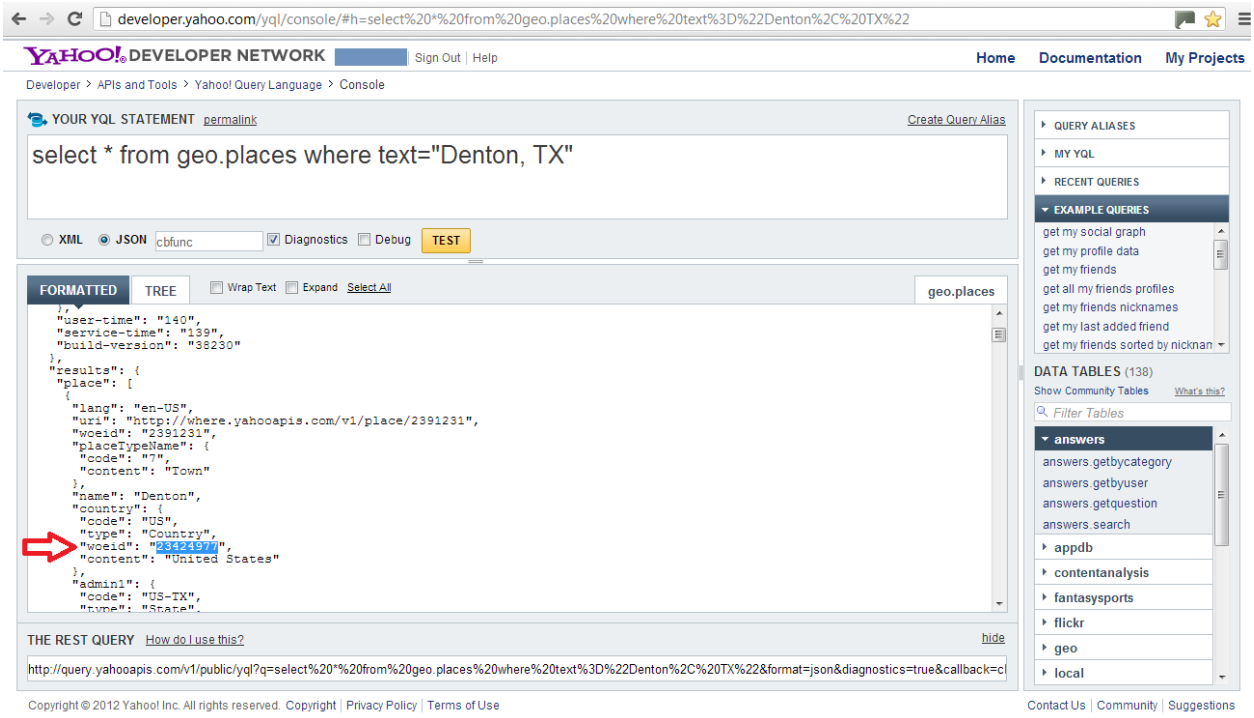
Access level	Read-only About the application permission model
Consumer key	[redacted]
Consumer secret	[redacted]
Request token URL	https://api.twitter.com/oauth/request_token
Authorize URL	https://api.twitter.com/oauth/authorize
Access token URL	https://api.twitter.com/oauth/access_token
Callback URL	None
Sign in with Twitter	No

Your access token

Use the access token string as your "oauth_token" and the access token secret as your "oauth_token_secret" to sign requests with your own Twitter account. Do not share your oauth_token_secret with anyone.

Access token	[redacted]
Access token secret	[redacted]
Access level	Read-only

```
MacBook-Pro-de-Hector-3:~ hectorc$ sudo pip install twython
The directory '/Users/hectorc/Library/Caches/pip/http' or its parent directory is not owned by
 the current user and the cache has been disabled. Please check the permissions and owner of t
hat directory. If executing pip with sudo, you may want sudo's -H flag.
The directory '/Users/hectorc/Library/Caches/pip' or its parent directory is not owned by the
current user and caching wheels has been disabled. check the permissions and owner of that dir
ectory. If executing pip with sudo, you may want sudo's -H flag.
Collecting twython
  Downloading twython-3.4.0.tar.gz
Collecting requests>=2.1.0 (from twython)
  Downloading requests-2.10.0-py2.py3-none-any.whl (506kB)
    100% |████████████████████████████████| 512kB 316kB/s
Collecting requests_oauthlib>=0.4.0 (from twython)
  Downloading requests_oauthlib-0.6.1-py2.py3-none-any.whl
Collecting oauthlib>=0.6.2 (from requests_oauthlib>=0.4.0->twython)
  Downloading oauthlib-1.1.2.tar.gz (111kB)
    100% |████████████████████████████████| 112kB 380kB/s
Installing collected packages: requests, oauthlib, requests-oauthlib, twython
  Running setup.py install for oauthlib ... done
  Running setup.py install for twython ... done
Successfully installed oauthlib-1.1.2 requests-2.10.0 requests-oauthlib-0.6.1 twython-3.4.0
```




```
twitter = Twython(ConsumerKey,
                  ConsumerSecret,
                  AccessToken,
                  AccessTokenSecret)
users = twitter.lookup_user(screen_name = "datacampmx")
for user in users:
    print "image: "      + user["profile_image_url_https"]
    print "twitts: "     + str(user["statuses_count"])
    print "followers: "  + str(user["followers_count"])
    print "followings: " + str(user["friends_count"])
    print "time zone: "  + user["time_zone"]
```

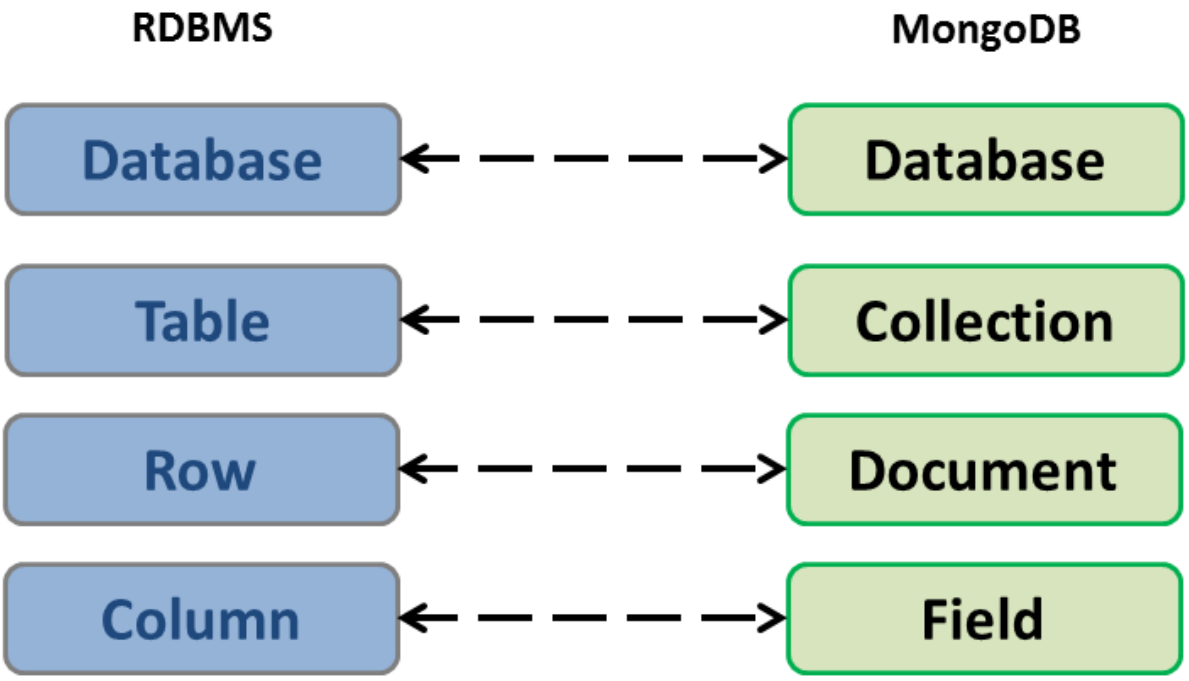
```
image: https://pbs.twimg.com/profile_images/744959798962094080/edTbc-mg_normal.jpg
twitts: 361
followers: 266
followings: 669
time zone: Pacific Time (US & Canada)
```

In [*]:

```
from twython import TwythonStreamer

ConsumerKey  = "41EltwaPLNgsn0Q4VS5g"
ConsumerSecret = "augwZxzQGsJuyfzLzGn0ASpherv2YgpeLTKEXXFk"
AccessToken = "141340589-WhKonOAcDmCX1MVJNpd3UEB2gvzZt2nmPBJfMy3o"
AccessTokenSecret = "LDsdOa9Mex2yZ0AMud4eUe0mlcqvvsycwp0yneSWQw"

class MyStreamer(TwythonStreamer):
    def on_success(self, data):
        if 'text' in data:
            print data['text'].encode('utf-8')

    def on_error(self, status_code, data):
        print status_code, data

stream = MyStreamer(ConsumerKey,
                    ConsumerSecret,
                    AccessToken,
                    AccessTokenSecret)

stream.statuses.filter(track='cnn')
```

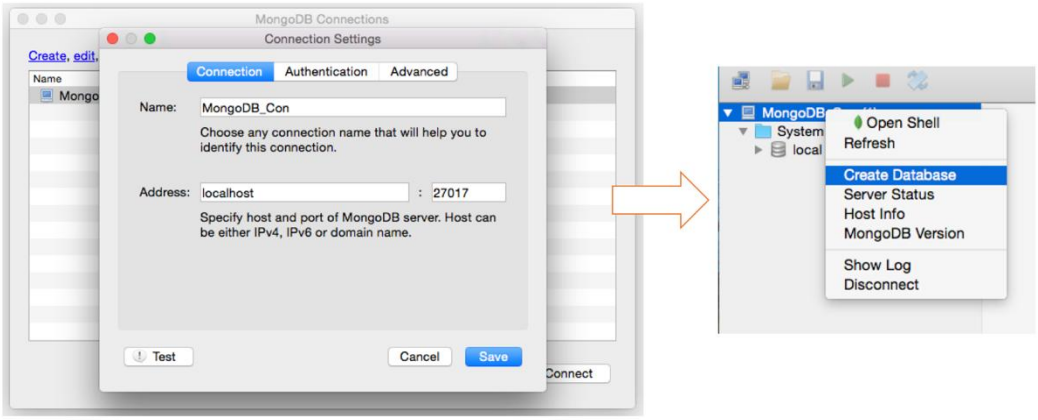
```
"Frustrated"??? More like mad. Angry. Righteous indignation. "Frustrated" is a traffic jam. This was a man's LIFE!!!
https://t.co/Wuwc6vI5OL
RT @ssirah: Capek2 ngitung ulang ktp 😅😅 https://t.co/K7JoTPyTCC
RT @magnifier661: 🔥CNN OFFICIAL BLACKOUT🔥
JULY 1ST - JULY14TH
@Toyota @etrade @WellsFargo @sprint @GEICO
#BlackOutCNN #Trump2016 https://t…
#USA#news Wimbledon 2016: Dresses too revealing?: Nike's flimsy attire for women's pros at Wimbledon is gener... http
s://t.co/9ZyME2XP8L
RT @cnnarabic: الداخلية المصرية تعلن مقتل كاهن كنيسة مارجرجس في #العريش برصاص مجهول.. و "#داعش" يتبنى
https://t.co/sJQYHH4hAE
#مصر #الكنيسة
RT @CNNPolitics: .@realDonaldTrump: "I totally disavow the Ku Klux Klan" https://t.co/PElISCG7gC https://t.co/DImhdKS
W6B
#DOMA# Wimbledon 2016: Dresses too revealing? https://t.co/9ZyME2XP8L
Feds: Stop driving these Honda models right now https://t.co/3caFJLe11f
Wimbledon 2016: Dresses too revealing? https://t.co/H3wCm47XIX
RT @CNN: President Obama on #Istanbul terror attack: “The prayers of the American people are with the people of Turke
y” https://t.co/dnDQyf…
@Brainiac_13 @CNN I am no more IPhones for me I've grown tired of them
```

Chapter 12: Data Processing and Aggregation with MongoDB




```
iMac-de-Hector:~ hectorcuesta1$ brew install mongodb
==> Downloading https://homebrew.bintray.com/bottles/mongodb-3.2.4.yosemite.bott
######################################################################## 100.0%
==> Pouring mongodb-3.2.4.yosemite.bottle.tar.gz
==> Caveats
To have launchd start mongodb at login:
  ln -sfv /usr/local/opt/mongodb/*.plist ~/Library/LaunchAgents
Then to load mongodb now:
  launchctl load ~/Library/LaunchAgents/homebrew.mxcl.mongodb.plist
Or, if you don't want/need launchctl, you can just run:
  mongod --config /usr/local/etc/mongod.conf
==> Summary
🍺  /usr/local/Cellar/mongodb/3.2.4: 17 files, 208.7M
```

```
iMac-de-Hector:~ hectorcuesta1$ mongod
2016-04-13T09:45:43.025-0500 I CONTROL  [initandlisten] MongoDB starting : pid=5541 port=27017 dbpath=/data/db
64-bit host=iMac-de-Hector.local
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] db version v3.2.4
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] git version: e2ee9ffcf9f5a94fad76802e28cc978718bb7a30
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] allocator: system
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] modules: none
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] build environment:
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten]     distarch: x86_64
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten]     target_arch: x86_64
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] options: {}
2016-04-13T09:45:43.026-0500 I STORAGE  [initandlisten] exception in initAndListen: 29 Data directory /data/db
not found., terminating
2016-04-13T09:45:43.026-0500 I CONTROL  [initandlisten] dbexit:  rc: 100
iMac-de-Hector:~ hectorcuesta1$ mongod
2016-04-13T09:46:59.909-0500 I CONTROL  [initandlisten] MongoDB starting : pid=5550 port=27017 dbpath=/data/db
64-bit host=iMac-de-Hector.local
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] db version v3.2.4
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] git version: e2ee9ffcf9f5a94fad76802e28cc978718bb7a30
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] allocator: system
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] modules: none
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] build environment:
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten]     distarch: x86_64
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten]     target_arch: x86_64
2016-04-13T09:46:59.910-0500 I CONTROL  [initandlisten] options: {}
2016-04-13T09:46:59.910-0500 I STORAGE  [initandlisten] wiredtiger_open config: create,cache_size=4G,session_ma
x=20000,eviction=(threads_max=4),config_base=false,statistics=(fast),log=(enabled=true,archive=true,path=journa
l,compressor=snappy),file_manager=(close_idle_time=100000),checkpoint=(wait=60,log_size=2GB),statistics_log=(wa
it=0),
2016-04-13T09:47:02.519-0500 I FTDC     [initandlisten] Initializing full-time diagnostic data capture with dir
ectory '/data/db/diagnostic.data'
2016-04-13T09:47:02.519-0500 I NETWORK  [HostnameCanonicalizationWorker] Starting hostname canonicalization wor
ker
2016-04-13T09:47:04.408-0500 I NETWORK  [initandlisten] waiting for connections on port 27017
2016-04-13T09:50:22.606-0500 I NETWORK  [initandlisten] connection accepted from 127.0.0.1:62468 #1 (1 connecti
on now open)
2016-04-13T09:50:22.612-0500 I NETWORK  [initandlisten] connection accepted from 127.0.0.1:62469 #2 (2 connecti
ons now open)
```




```
> show dbs
Corpus  0.0625GB
admin   (empty)
local   0.03125GB
test    0.0625GB
>
```

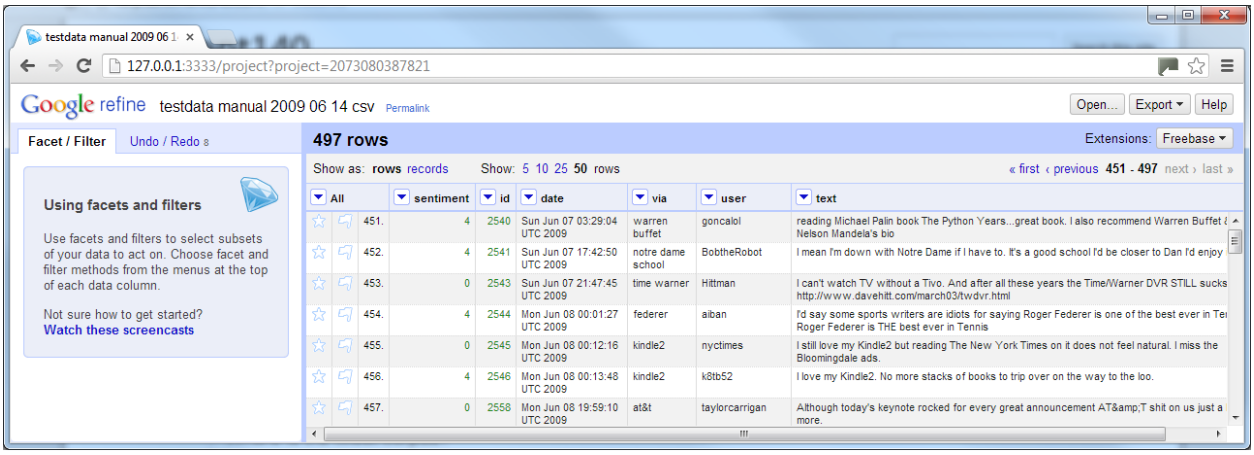
```
> use Corpus
switched to db Corpus
> show collections
system.indexes
tweets
>
```


```
> db.test.data.find()
{ "_id" : ObjectId("51eedee2d341516bbfdbc6ff"), "name" : { "first" : "Jan", "las
t" : "Smith" } }
{ "_id" : ObjectId("51eedf0cd341516bbfdbc700"), "name" : { "first" : "Damian", "
last" : "Cuesta" } }
{ "_id" : ObjectId("51eedf17d341516bbfdbc701"), "name" : { "first" : "Isaac", "l
ast" : "Cuesta" } }
> _
```

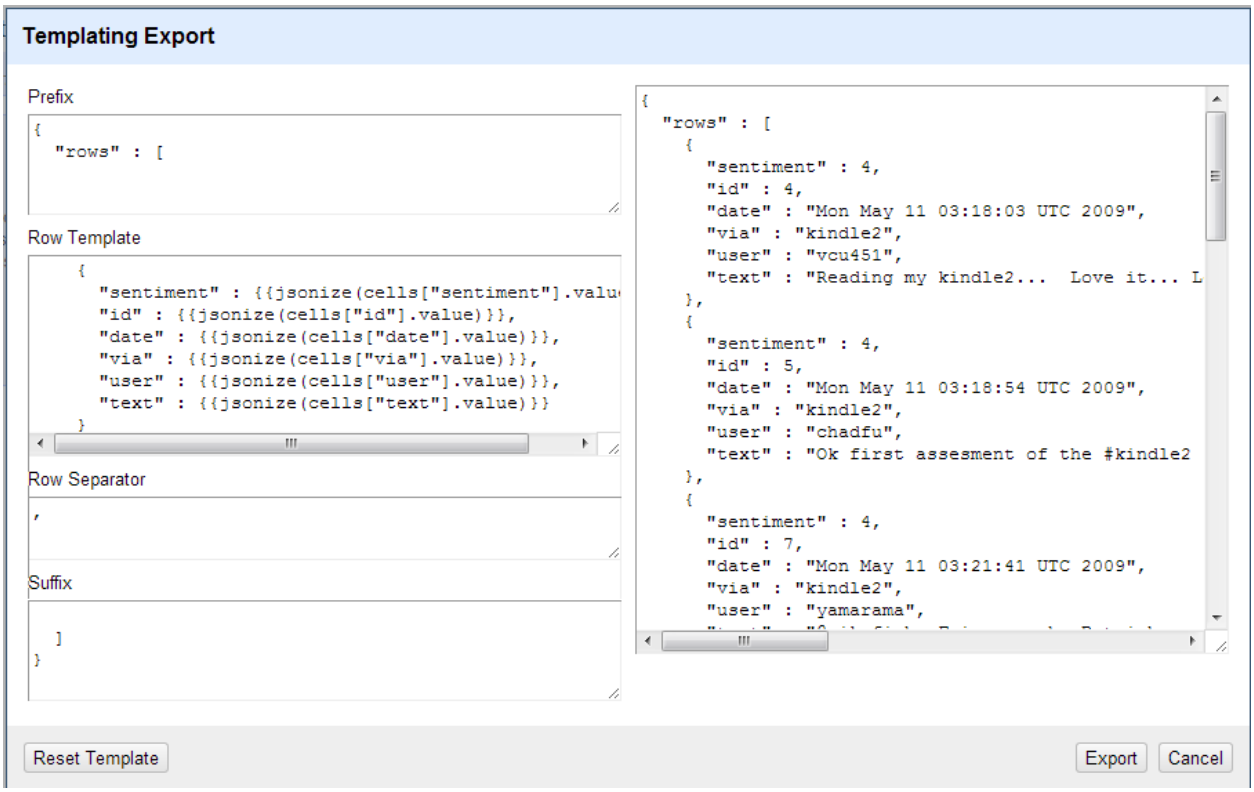
```
> db.test.data.find({"name.last":"Cuesta"})
{ "_id" : ObjectId("51eedf0cd341516bbfdbc700"), "name" : { "first" : "Damian", "
last" : "Cuesta" } }
{ "_id" : ObjectId("51eedf17d341516bbfdbc701"), "name" : { "first" : "Isaac", "l
ast" : "Cuesta" } }
>
```

```
> db.test.data.findOne()
{
        "_id" : ObjectId("51eedee2d341516bbfdbc6ff"),
        "name" : {
                "first" : "Jan",
                "last" : "Smith"
        }
}
```

```
> db.test.data.find({"name.last":"Cuesta"}).explain()
{
        "cursor" : "BasicCursor",
        "isMultiKey" : false,
        "n" : 2,
        "nscannedObjects" : 3,
        "nscanned" : 3,
        "nscannedObjectsAllPlans" : 3,
        "nscannedAllPlans" : 3,
        "scanAndOrder" : false,
        "indexOnly" : false,
        "nYields" : 0,
        "nChunkSkips" : 0,
        "millis" : 0,
        "indexBounds" : {

        },
        "server" : "Hadoop-PC:27017"
}
```



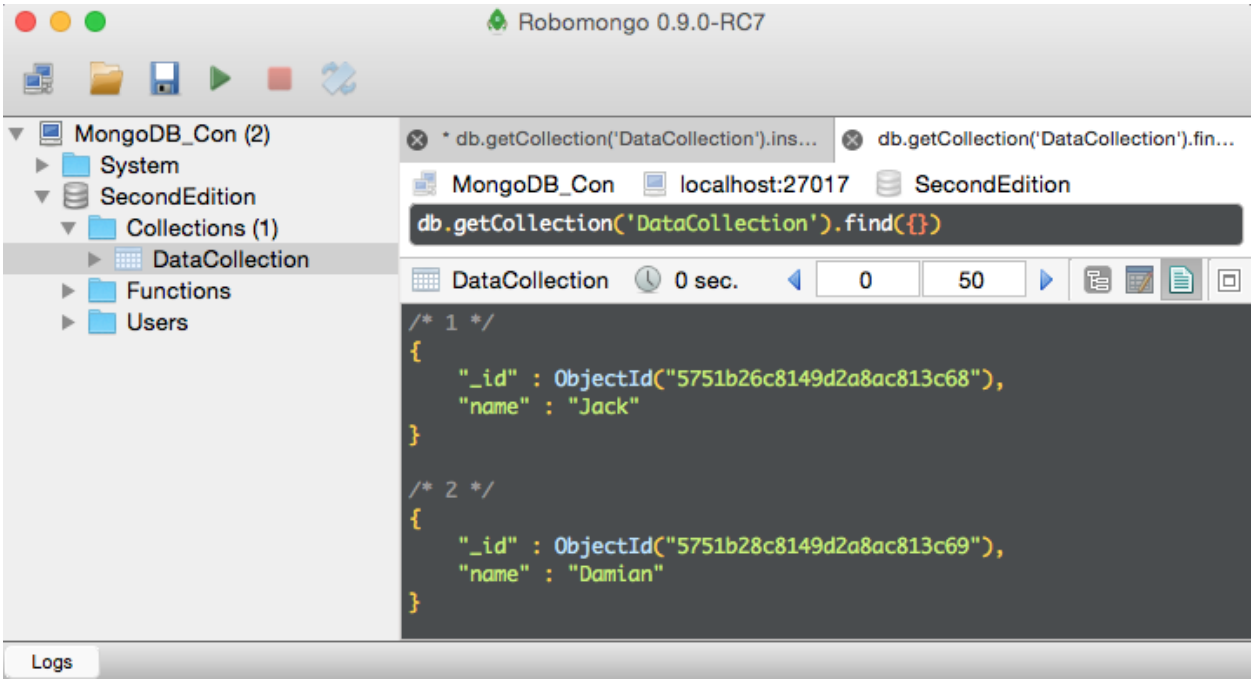

Custom text transform on column text
Expression
Language
Google Refine Expression Language (GREL)
value.replace(",","")
No syntax error.
Preview
History
Starred
Help
row
value
value.replace(",","")
1.
Reading my kindle2... Love it... Lee childs is good read.
Reading my kindle2... Love it... Lee childs is good read.
2.
Ok first assesment of the #kindle2 ...it fucking rocks!!!
Ok first assesment of the #kindle2 ...it fucking rocks!!!
3.
@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
4.
@mikefish Fair enough. But i have the Kindle2
@mikefish Fair enough. But i have the Kindle2
On error
keep original
set to blank
store error
Re-transform up to
10
times until no change
OK
Cancel




```
In [1]: !pip install pymongo
```

```
Collecting pymongo
  Downloading pymongo-3.2.2-cp27-none-macosx_10_10_intel.whl (262kB)
    100% |████████████████████████████████| 266kB 780kB/s
Installing collected packages: pymongo
Successfully installed pymongo-3.2.2
```

```
In [2]: import pymongo
```

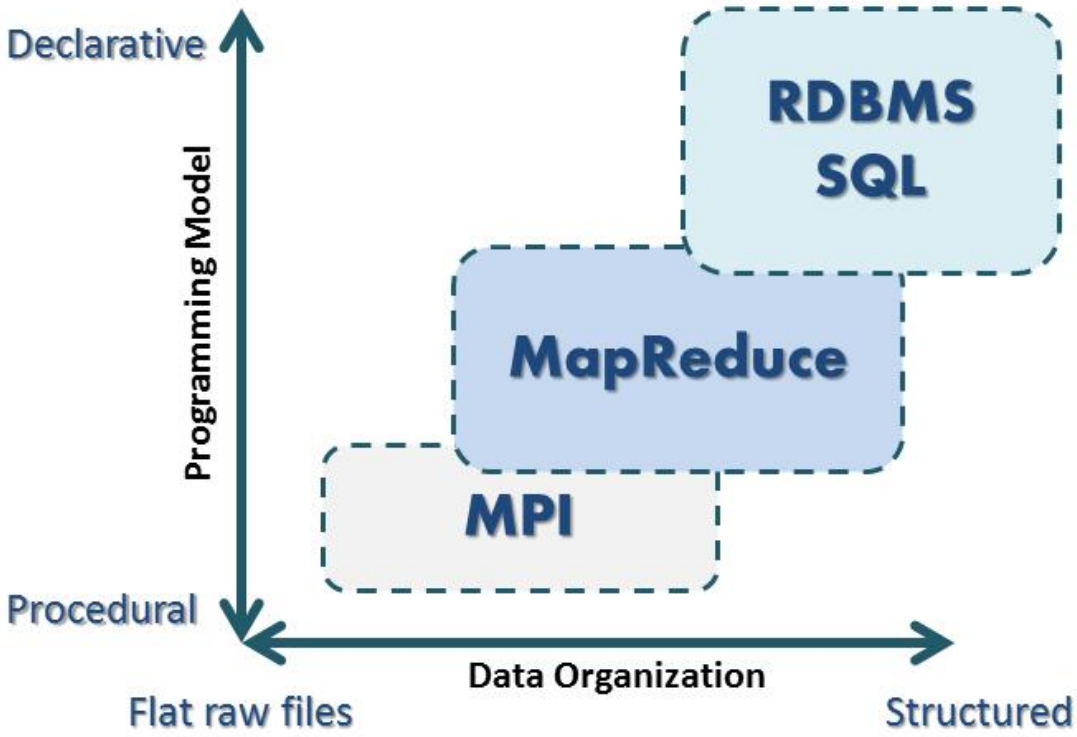


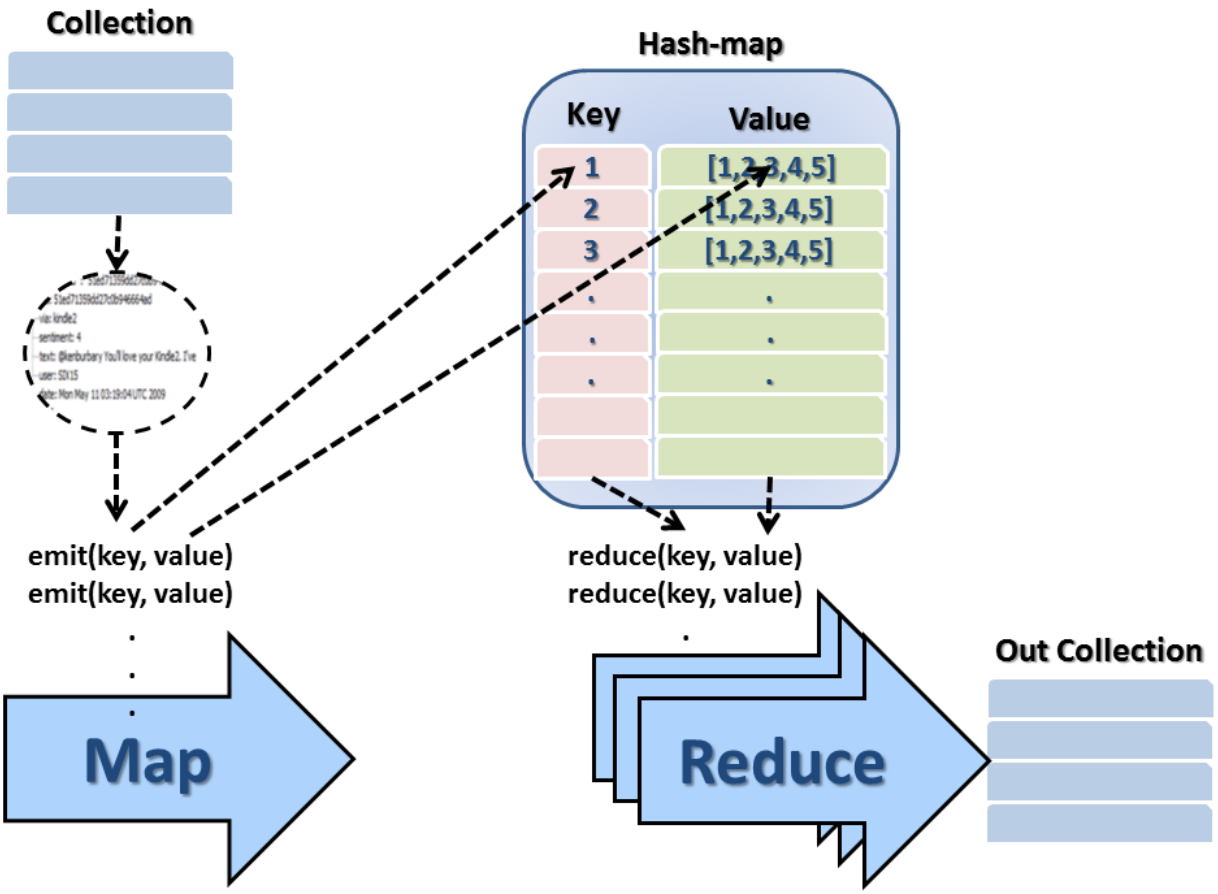

```
>>> ================================ RESTART ================================
>>>
{'count': 139, '_id': 2}
{'count': 177, '_id': 0}
{'count': 181, '_id': 4}
>>>
```

```
>>> ================================ RESTART ================================
>>>
{'count': 1, '_id': 'fred wilson'}
{'count': 8, '_id': 'warren buffet'}
{'count': 1, '_id': 'aapl'}
{'count': 2, '_id': 'mashable'}
{'count': 1, '_id': 'hitler'}
{'count': 1, '_id': 'yankees'}
{'count': 1, '_id': 'republican'}
{'count': 7, '_id': 'exam'}
{'count': 1, '_id': 'world cup'}
{'count': 5, '_id': 'viral marketing'}
>>>
```

```
>>> ================================ RESTART ================================
>>>
{'count': 7, 'avgId': 1065.857142857143, '_id': 'exam', 'maxId': 2195, 'minId': 218}
{'count': 1, 'avgId': 226.0, '_id': 'republican', 'maxId': 226, 'minId': 226}
{'count': 1, 'avgId': 1025.0, '_id': 'world cup', 'maxId': 1025, 'minId': 1025}
{'count': 1, 'avgId': 2398.0, '_id': 'yankees', 'maxId': 2398, 'minId': 2398}
{'count': 1, 'avgId': 14045.0, '_id': 'aapl', 'maxId': 14045, 'minId': 14045}
{'count': 1, 'avgId': 2296.0, '_id': 'hitler', 'maxId': 2296, 'minId': 2296}
 ...
```

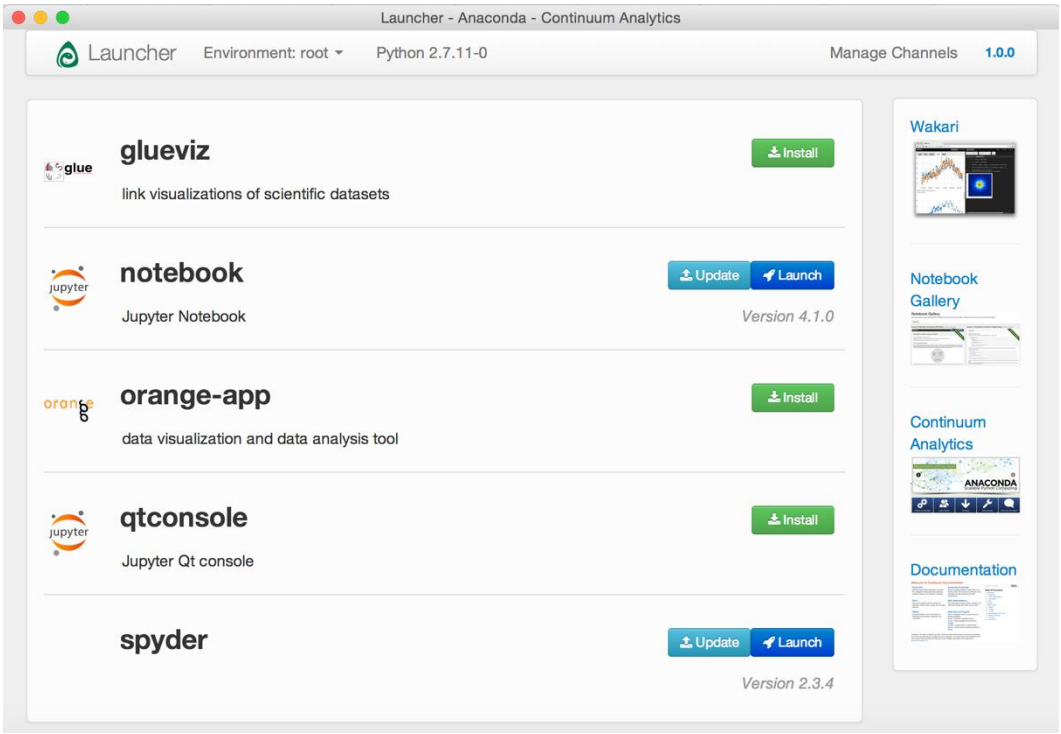
Chapter 13: Working with MapReduce

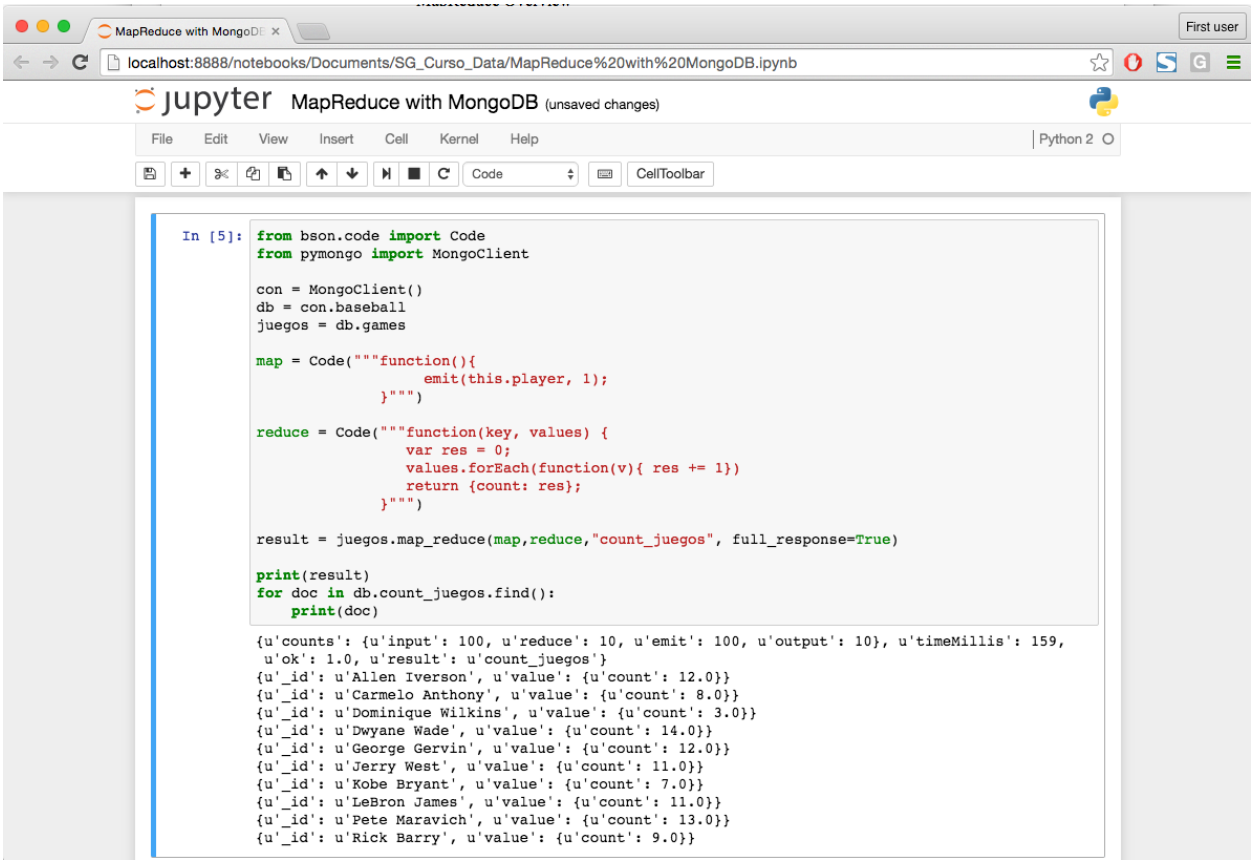


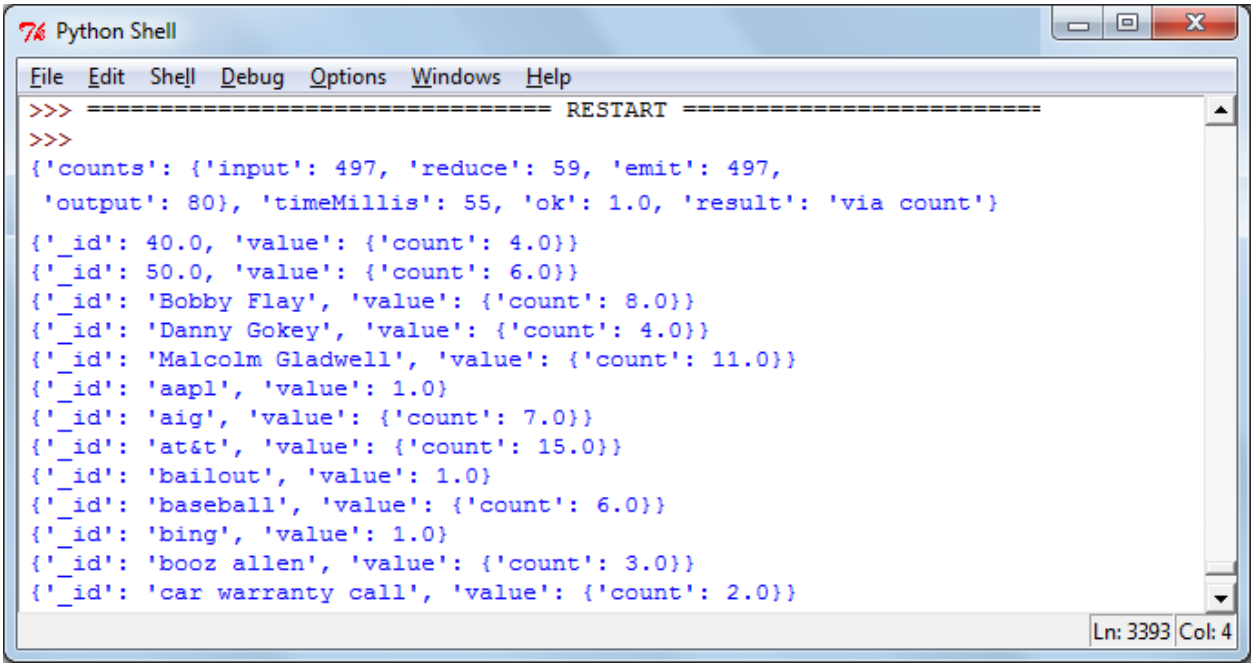



```
> mapTest = function(){ emit(this.via, 1); }
function (){ emit(this.via, 1); }
> reduceTest = function(key, values) {
...
... var res = 0;
... values.forEach(function(v){ res += 1})
...
... return {count: res};
... }
function (key, values) {

var res = 0;
values.forEach(function(v){ res += 1})

return {count: res};
}
>
```

```
> use Corpus
switched to db Corpus
> db.tweets.mapReduce(mapTest,reduceTest, {out:"results"})
{
        "result" : "results",
        "timeMillis" : 135,
        "counts" : {
                "input" : 497,
                "emit" : 497,
                "reduce" : 59,
                "output" : 80
        },
        "ok" : 1,
}
> db.results.find()
{ "_id" : 40, "value" : { "count" : 4 } }
{ "_id" : 50, "value" : { "count" : 6 } }
{ "_id" : "Bobby Flay", "value" : { "count" : 8 } }
{ "_id" : "Danny Gokey", "value" : { "count" : 4 } }
{ "_id" : "Malcolm Gladwell", "value" : { "count" : 11 } }
{ "_id" : "aapl", "value" : 1 }
{ "_id" : "aig", "value" : { "count" : 7 } }
{ "_id" : "at&t", "value" : { "count" : 15 } }
{ "_id" : "bailout", "value" : 1 }
{ "_id" : "baseball", "value" : { "count" : 6 } }
{ "_id" : "bing", "value" : 1 }
{ "_id" : "booz allen", "value" : { "count" : 3 } }
{ "_id" : "car warranty call", "value" : { "count" : 2 } }
{ "_id" : "cheney", "value" : { "count" : 5 } }
{ "_id" : "china", "value" : { "count" : 6 } }
{ "_id" : "comcast", "value" : { "count" : 4 } }
{ "_id" : "dentist", "value" : { "count" : 17 } }
{ "_id" : "driving", "value" : 1 }
{ "_id" : "east palo alto", "value" : { "count" : 4 } }
{ "_id" : "eating", "value" : { "count" : 12 } }
Type "it" for more
>
```

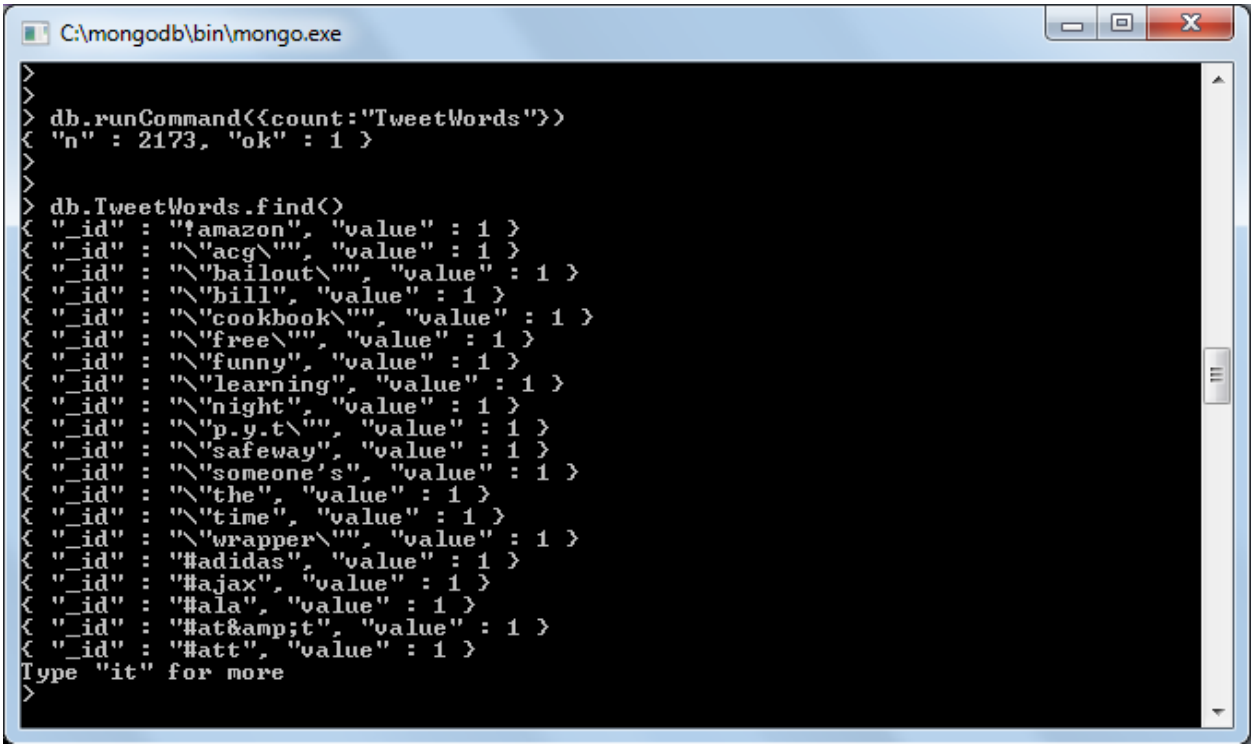


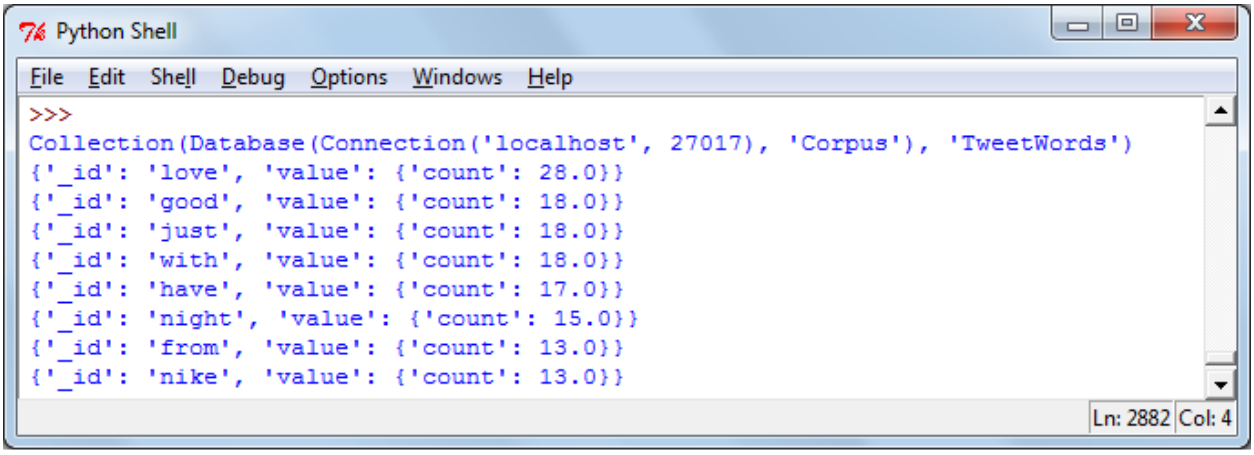

MapReduce with MongoDB
First user
localhost:8888/notebooks/Documents/SG_Curso_Data/MapReduce%20with%20MongoDB.ipynb
jupyter MapReduce with MongoDB (unsaved changes)
File Edit View Insert Cell Kernel Help
Python 2
Code
CellToolbar
In [5]:
from bson.code import Code
from pymongo import MongoClient
con = MongoClient()
db = con.baseball
juegos = db.games
map = Code("""function(){
emit(this.player, 1);
}""")
reduce = Code("""function(key, values) {
var res = 0;
values.forEach(function(v){ res += 1})
return {count: res};
}""")
result = juegos.map_reduce(map,reduce,"count_juegos", full_response=True)
print(result)
for doc in db.count_juegos.find():
print(doc)
{u'counts': {u'input': 100, u'reduce': 10, u'emit': 100, u'output': 10}, u'timeMillis': 159,
u'ok': 1.0, u'result': u'count_juegos'}
{u'_id': u'Allen Iverson', u'value': {u'count': 12.0}}
{u'_id': u'Carmelo Anthony', u'value': {u'count': 8.0}}
{u'_id': u'Dominique Wilkins', u'value': {u'count': 3.0}}
{u'_id': u'Dwyane Wade', u'value': {u'count': 14.0}}
{u'_id': u'George Gervin', u'value': {u'count': 12.0}}
{u'_id': u'Jerry West', u'value': {u'count': 11.0}}
{u'_id': u'Kobe Bryant', u'value': {u'count': 7.0}}
{u'_id': u'LeBron James', u'value': {u'count': 11.0}}
{u'_id': u'Pete Maravich', u'value': {u'count': 13.0}}
{u'_id': u'Rick Barry', u'value': {u'count': 9.0}}


Python Shell
File Edit Shell Debug Options Windows Help
>>> ========== RESTART ==========
>>>
{'counts': {'input': 497, 'reduce': 59, 'emit': 497,
'output': 80}, 'timeMillis': 55, 'ok': 1.0, 'result': 'via count'}
{'_id': 40.0, 'value': {'count': 4.0}}
{'_id': 50.0, 'value': {'count': 6.0}}
{'_id': 'Bobby Flay', 'value': {'count': 8.0}}
{'_id': 'Danny Gokey', 'value': {'count': 4.0}}
{'_id': 'Malcolm Gladwell', 'value': {'count': 11.0}}
{'_id': 'aapl', 'value': 1.0}
{'_id': 'aig', 'value': {'count': 7.0}}
{'_id': 'at&t', 'value': {'count': 15.0}}
{'_id': 'bailout', 'value': 1.0}
{'_id': 'baseball', 'value': {'count': 6.0}}
{'_id': 'bing', 'value': 1.0}
{'_id': 'booz allen', 'value': {'count': 3.0}}
{'_id': 'car warranty call', 'value': {'count': 2.0}}
Ln: 3393 Col: 4

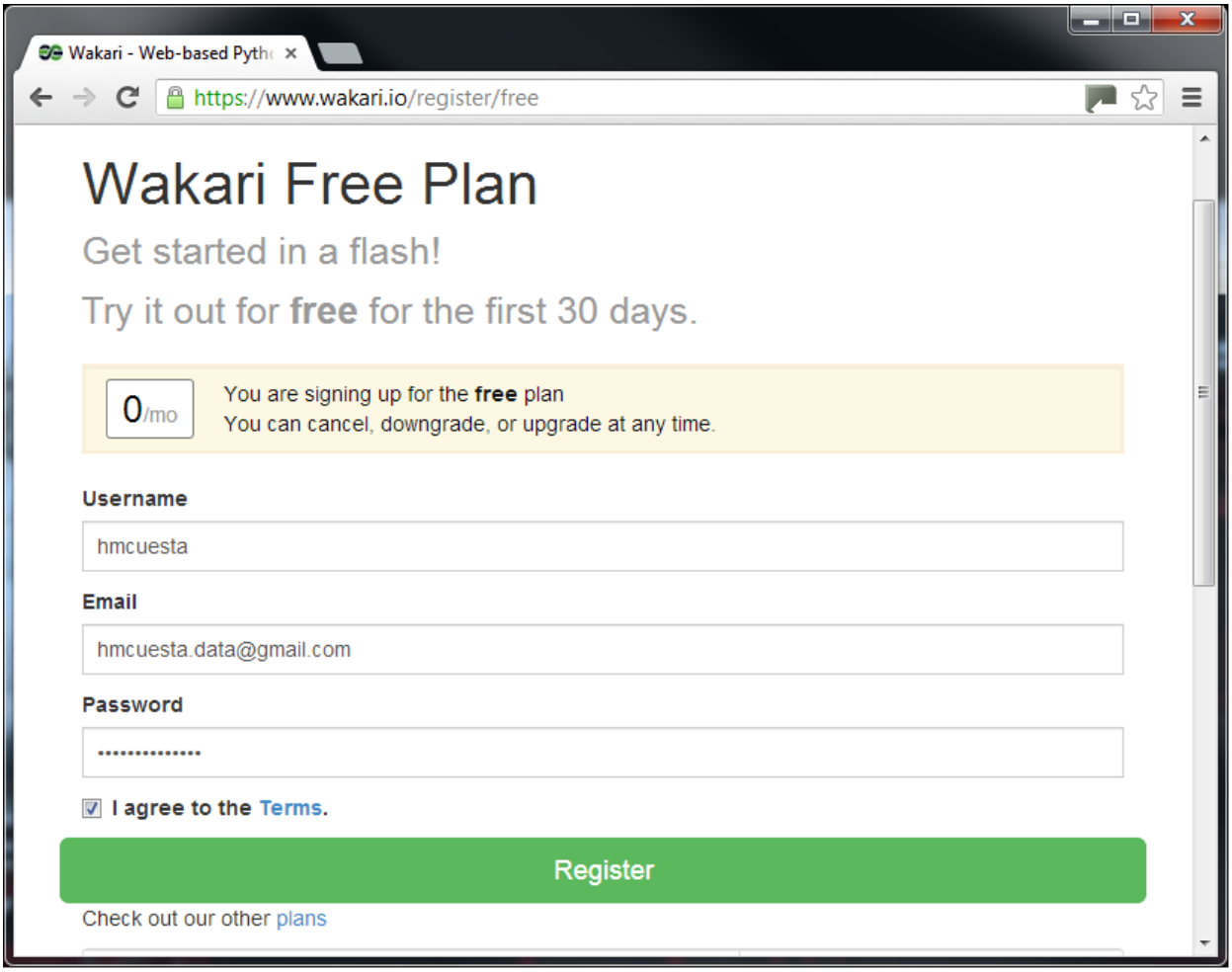
```
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d648"} , "player" : "LeBron James" , "points" : 40}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d649"} , "player" : "Rick Barry" , "points" : 6}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64a"} , "player" : "George Gervin" , "points" : 0}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64b"} , "player" : "Kobe Bryant" , "points" : 56}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64c"} , "player" : "Pete Maravich" , "points" : 4}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64d"} , "player" : "Dwyane Wade" , "points" : 65}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64e"} , "player" : "Pete Maravich" , "points" : 55}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d64f"} , "player" : "Dwyane Wade" , "points" : 45}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d650"} , "player" : "Allen Iverson" , "points" : 66}
{ "_id" : { "$oid" : "5206caef9dd27c1964b1d651"} , "player" : "Rick Barry" , "points" : 18}

  . . .
```

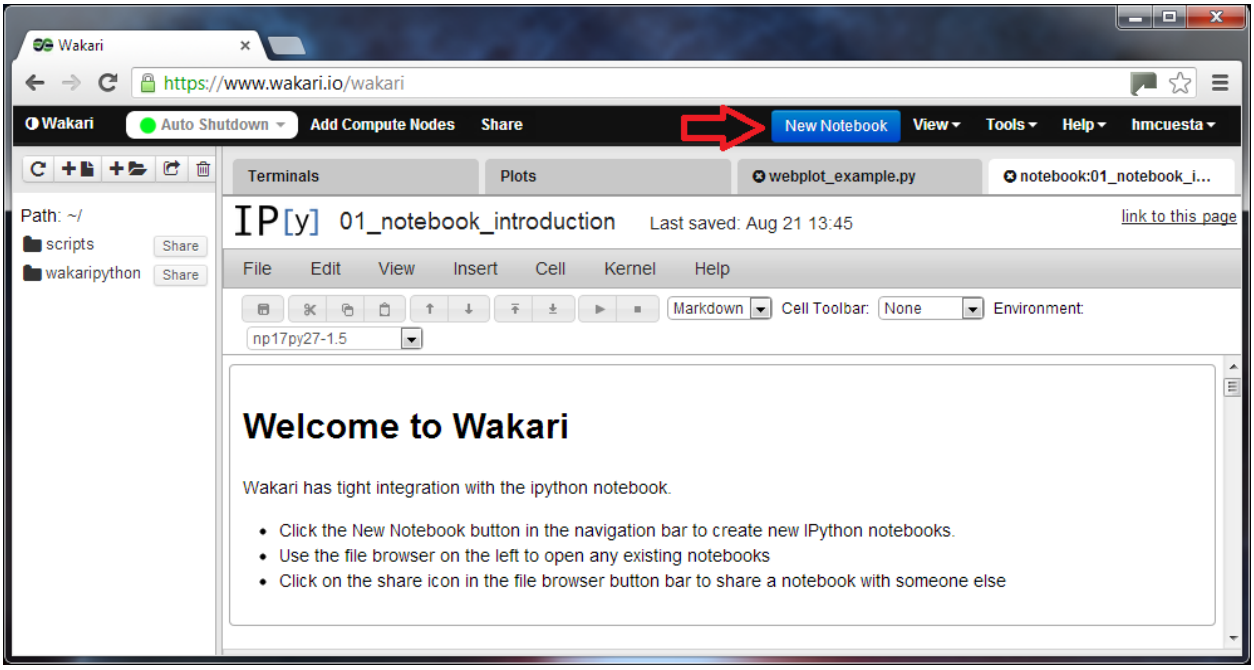

```
>>> ================================ RESTART ================================
>>>
Collection(Database(Connection('localhost', 27017), 'baseball'), '_result')
{'_id': 'Allen Iverson', 'value': {'max': 66.0, 'total': 310.0, 'avg': 34.44444444444444, 'min': 9.0}}
{'_id': 'Carmelo Anthony', 'value': {'max': 91.0, 'total': 473.0, 'avg': 47.3, 'min': 1.0}}
{'_id': 'Dominique Wilkins', 'value': {'max': 98.0, 'total': 545.0, 'avg': 60.55555555555556, 'min': 20.0}}
{'_id': 'Dwyane Wade', 'value': {'max': 95.0, 'total': 834.0, 'avg': 55.6, 'min': 15.0}}
{'_id': 'George Gervin', 'value': {'max': 81.0, 'total': 235.0, 'avg': 47.0, 'min': 0.0}}
{'_id': 'Jerry West', 'value': {'max': 98.0, 'total': 645.0, 'avg': 58.63636363636363, 'min': 9.0}}
{'_id': 'Kobe Bryant', 'value': {'max': 95.0, 'total': 497.0, 'avg': 45.18181818181818, 'min': 0.0}}
{'_id': 'LeBron James', 'value': {'max': 100.0, 'total': 546.0, 'avg': 49.63636363636363, 'min': 3.0}}
{'_id': 'Pete Maravich', 'value': {'max': 97.0, 'total': 562.0, 'avg': 43.23076923076923, 'min': 4.0}}
{'_id': 'Rick Barry', 'value': {'max': 98.0, 'total': 781.0, 'avg': 48.8125, 'min': 6.0}}
>>>
```

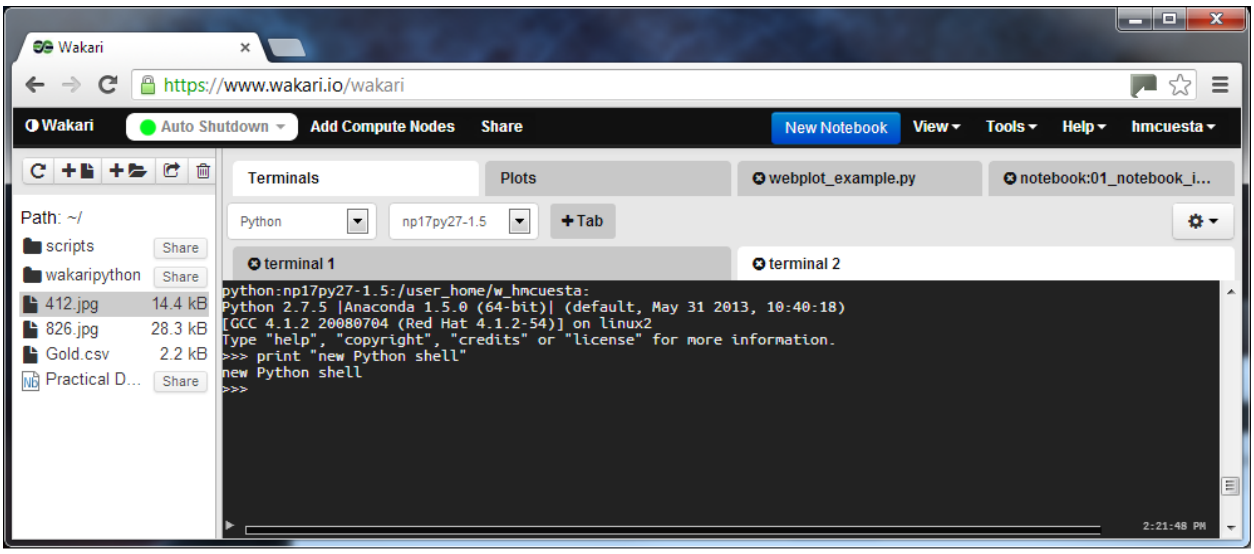


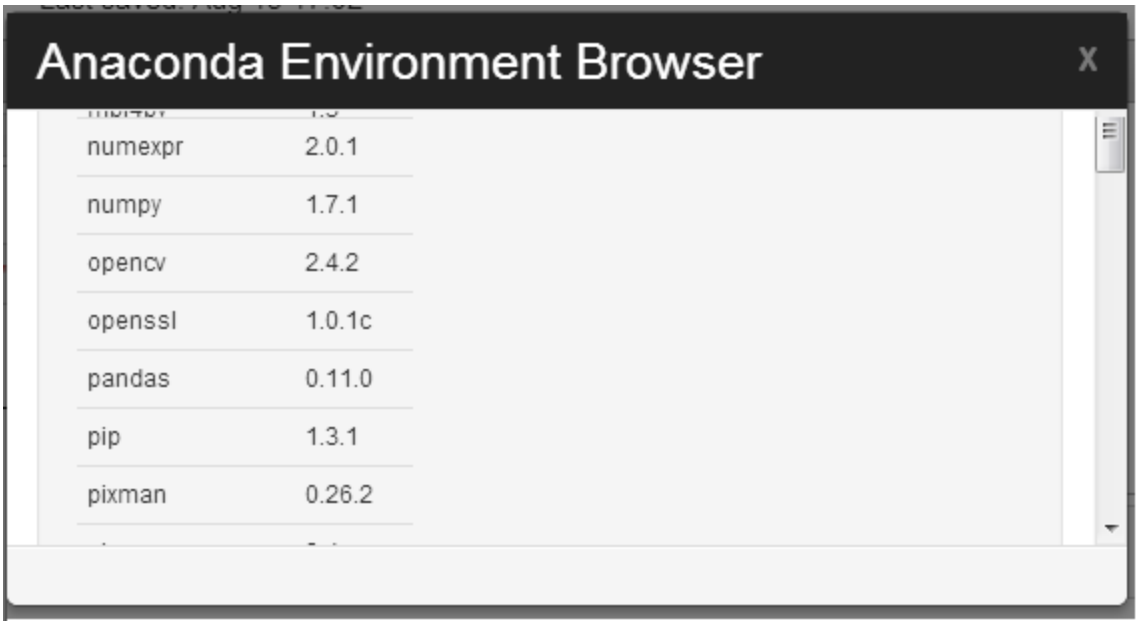

Python Shell
File Edit Shell Debug Options Windows Help
>>>
Collection(Database(Connection('localhost', 27017), 'Corpus'), 'TweetWords')
{'_id': 'love', 'value': {'count': 28.0}}
{'_id': 'good', 'value': {'count': 18.0}}
{'_id': 'just', 'value': {'count': 18.0}}
{'_id': 'with', 'value': {'count': 18.0}}
{'_id': 'have', 'value': {'count': 17.0}}
{'_id': 'night', 'value': {'count': 15.0}}
{'_id': 'from', 'value': {'count': 13.0}}
{'_id': 'nike', 'value': {'count': 13.0}}
Ln: 2882 Col: 4

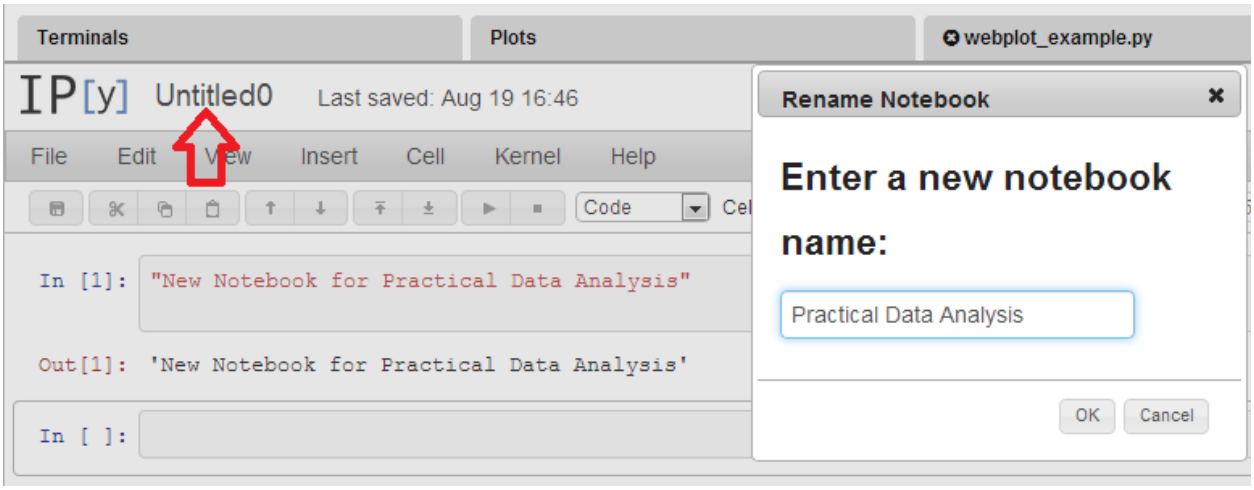
Chapter 14: On-line Data Analysis with Jupyter and Wakari

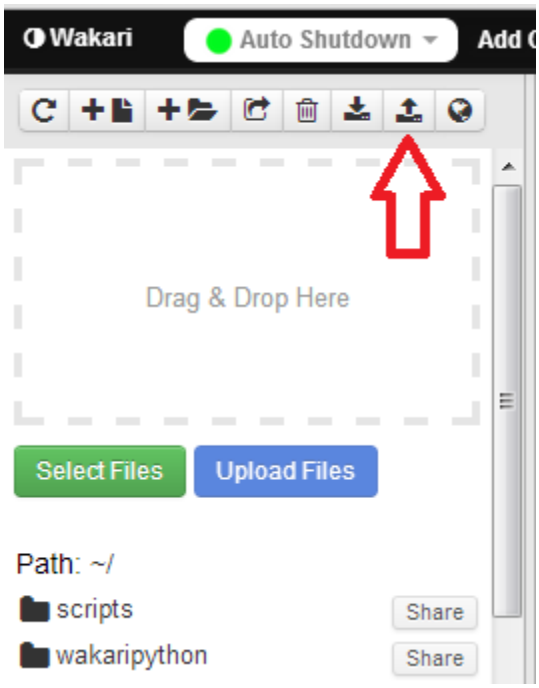


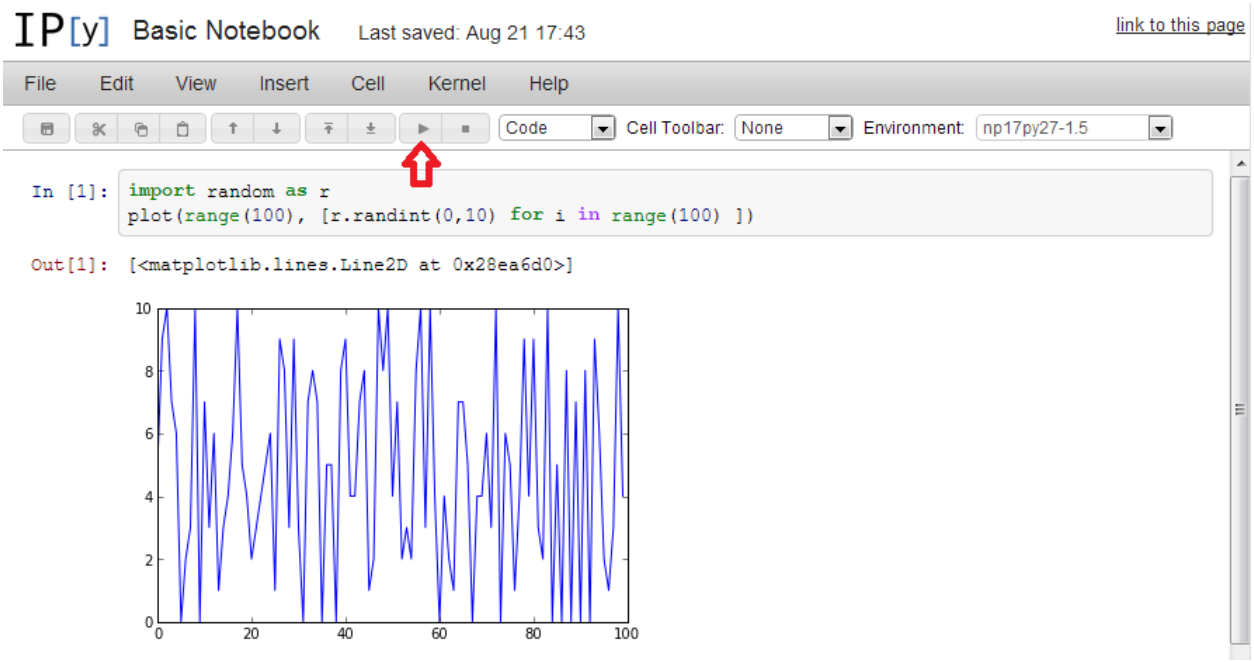

Wakari
https://www.wakari.io/wakari
Wakari
Auto Shutdown
Add Compute Nodes
Share
New Notebook
View
Tools
Help
hmcuesta
Terminals
Plots
webplot_example.py
notebook:01_notebook_i...
Path: ~/
scripts
Share
wakaripython
Share
IP[y] 01_notebook_introduction
Last saved: Aug 21 13:45
link to this page
File
Edit
View
Insert
Cell
Kernel
Help
Markdown
Cell Toolbar:
None
Environment:
np17py27-1.5
Welcome to Wakari
Wakari has tight integration with the ipython notebook.
Click the New Notebook button in the navigation bar to create new IPython notebooks.
Use the file browser on the left to open any existing notebooks
Click on the share icon in the file browser button bar to share a notebook with someone else


Wakari
https://www.wakari.io/wakari
Wakari
Auto Shutdown
Add Compute Nodes
Share
New Notebook
View
Tools
Help
hmcuesta
Terminals
Plots
webplot_example.py
notebook:01_notebook_i...
Path: ~/
scripts
Share
wakaripython
Share
412.jpg
14.4 kB
826.jpg
28.3 kB
Gold.csv
2.2 kB
Practical D...
Share
Python
np17py27-1.5
+Tab
terminal 1
terminal 2
python:np17py27-1.5:/user_home/w_hmcuesta:
Python 2.7.5 |Anaconda 1.5.0 (64-bit)| (default, May 31 2013, 10:40:18)
[GCC 4.1.2 20080704 (Red Hat 4.1.2-54)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> print "new Python shell"
new Python shell
>>>
2:21:48 PM


Anaconda Environment Browser
numexpr 2.0.1
numpy 1.7.1
opencv 2.4.2
openssl 1.0.1c
pandas 0.11.0
pip 1.3.1
pixman 0.26.2

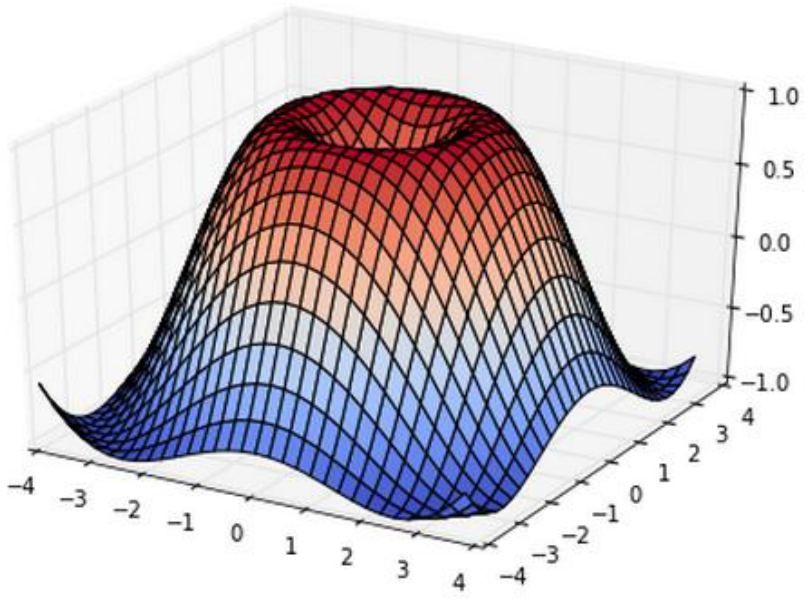

Terminals
Plots
webplot_example.py
IP[y] Untitled0
Last saved: Aug 19 16:46
File Edit View Insert Cell Kernel Help
Code
In [1]: "New Notebook for Practical Data Analysis"
Out[1]: 'New Notebook for Practical Data Analysis'
In [ ]:
Rename Notebook
Enter a new notebook name:
Practical Data Analysis
OK
Cancel


Wakari
Auto Shutdown
Add
Drag & Drop Here
Select Files
Upload Files
Path: ~/
scripts
Share
wakaripython
Share


IP[y] Basic Notebook
Last saved: Aug 21 17:43
link to this page
File
Edit
View
Insert
Cell
Kernel
Help
Code
Cell Toolbar:
None
Environment:
np17py27-1.5
In [1]:
import random as r
plot(range(100), [r.randint(0,10) for i in range(100)])
Out[1]:
[<matplotlib.lines.Line2D at 0x28ea6d0>]
0
2
4
6
8
10
0
20
40
60
80
100

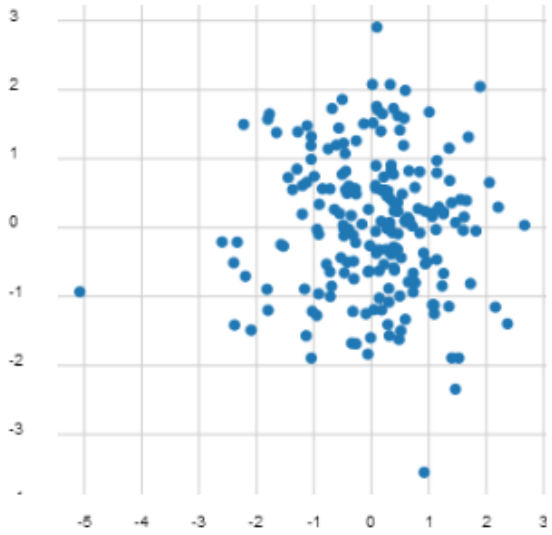
In [2]:
```
from pylab import *
from mpl_toolkits.mplot3d import Axes3D
fig = figure()
ax = Axes3D(fig)
X = np.arange(-4, 4, 0.25)
Y = np.arange(-4, 4, 0.25)
X, Y = np.meshgrid(X, Y)
R = np.sqrt(X**2 + Y**2)
Z = np.sin(R)

ax.plot_surface(X, Y, Z, rstride=1, cstride=1, cmap=cm.coolwarm)
```

Out[2]: <mpl_toolkits.mplot3d.art3d.Poly3DCollection at 0x27416d0>



In [3]:
```
from webplot import p
n = 200
X = np.random.normal(0,1,n)
Y = np.random.normal(0,1,n)

p.scatter(X,Y)
```

Out[3]: Plots

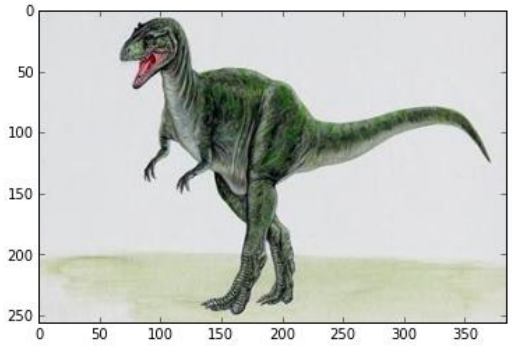


Path: ~/

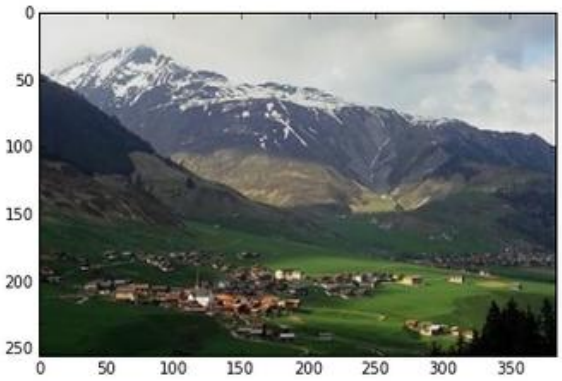
scripts	Share	
wakaripython	Share	
Practical Data Analysis.ipynb	Share	
412.jpg		14.4 kB
826.jpg		28.3 kB

In [3]:
```
from PIL import Image
import pylab
dino = Image.open("412.jpg")
pylab.imshow(dino)
```

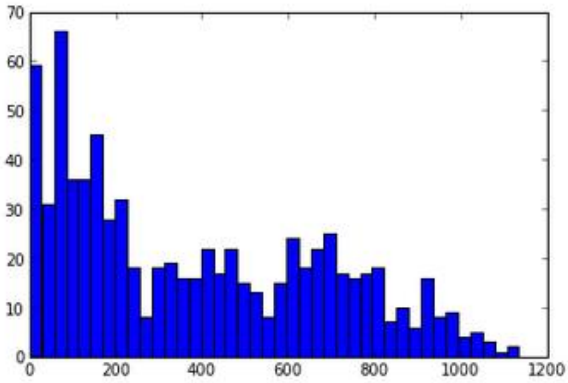
Out[3]: <matplotlib.image.AxesImage at 0x2c0c190>



In [8]:
```
land = Image.open("826.jpg")
pylab.imshow(land)
pylab.show()
hist = land.histogram()
pylab.hist(hist,bins=40)
```

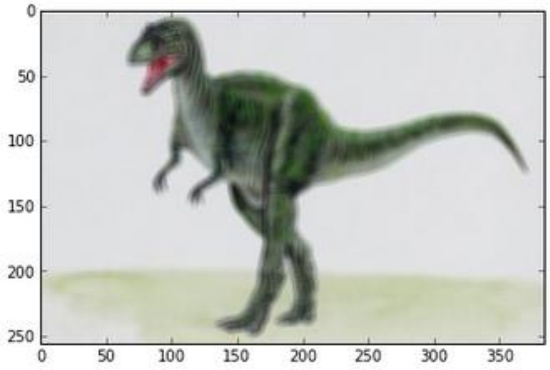


Out[8]:
```
(array([59, 31, 66, 36, 36, 45, 28, 32, 18,  8, 18, 19, 16, 16, 22, 17, 22,
       15, 13,  8, 15, 24, 18, 22, 25, 17, 16, 17, 18,  7, 10,  6, 16,  8,
        9,  4,  5,  3,  1,  2]),
 array([    0.   ,    28.375,    56.75 ,    85.125,   113.5  ,   141.875,
         170.25 ,   198.625,   227.   ,   255.375,   283.75 ,   312.125,
         340.5  ,   368.875,   397.25 ,   425.625,   454.   ,   482.375,
         510.75 ,   539.125,   567.5  ,   595.875,   624.25 ,   652.625,
         681.   ,   709.375,   737.75 ,   766.125,   794.5  ,   822.875,
         851.25 ,   879.625,   908.   ,   936.375,   964.75 ,   993.125,
        1021.5  ,  1049.875,  1078.25 ,  1106.625,  1135.   ]),
<a list of 40 Patch objects>)
```



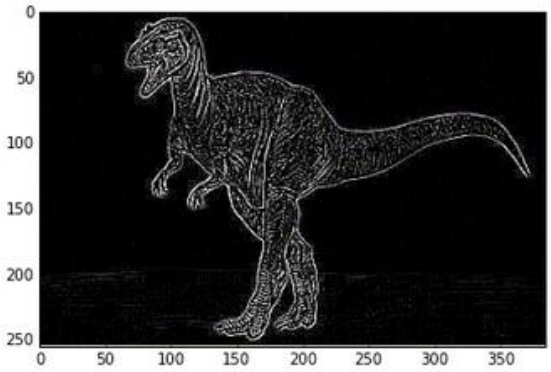
In [9]:
```
from PIL import ImageFilter
im1 = dino.filter(ImageFilter.BLUR)
pylab.imshow(im1)
```

Out[9]: <matplotlib.image.AxesImage at 0x3e61710>



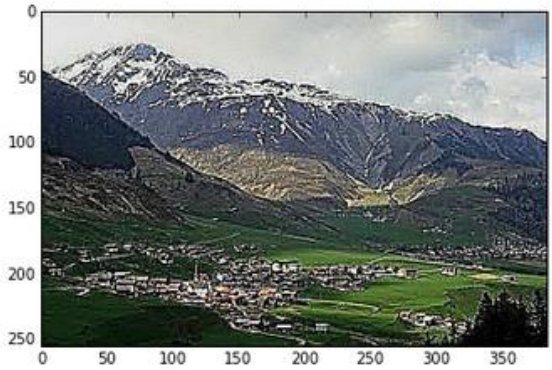
In [10]:
```
im2 = dino.filter(ImageFilter.FIND_EDGES)
pylab.imshow(im2)
```

Out[10]: <matplotlib.image.AxesImage at 0x41fd250>



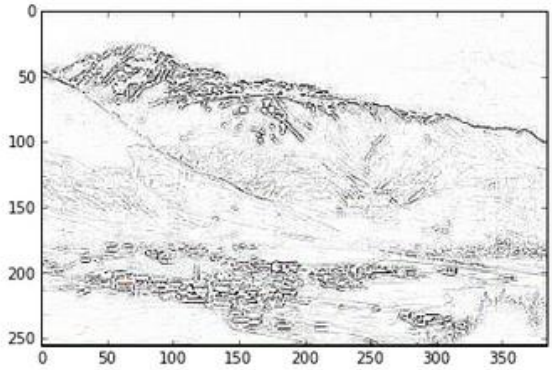
In [11]:
```
im3 = land.filter(ImageFilter.EDGE_ENHANCE_MORE)
pylab.imshow(im3)
```

Out[11]: <matplotlib.image.AxesImage at 0x4227e50>



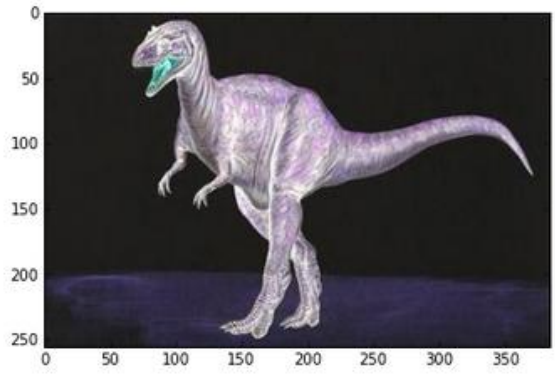
In [12]:
```
im4 = land.filter(ImageFilter.CONTOUR)
pylab.imshow(im4)
```

Out[12]: <matplotlib.image.AxesImage at 0x4c1aa90>



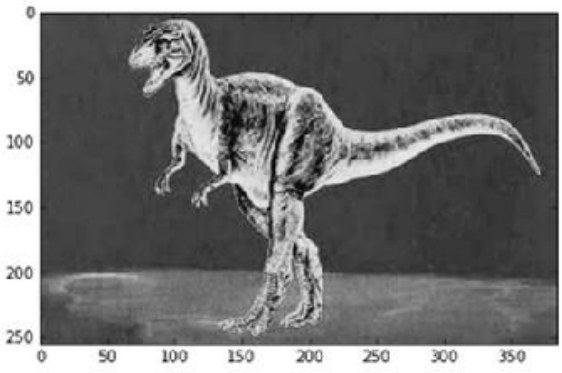
In [14]:
```
from PIL import ImageOps
im5 = ImageOps.invert(dino)
pylab.imshow(im5)
```

Out[14]: <matplotlib.image.AxesImage at 0x5080bd0>



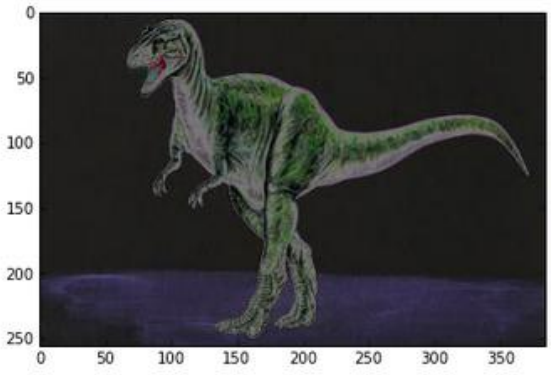
In [15]:
```
im6 = ImageOps.grayscale(dino)
pylab.imshow(im6)
```

Out[15]: <matplotlib.image.AxesImage at 0x50ba790>



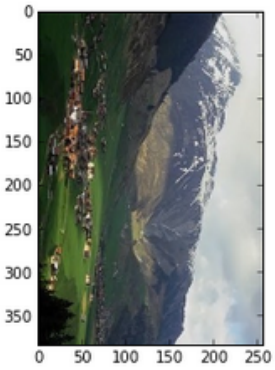
In [16]:
```
im7 = ImageOps.solarize(dino, threshold=128)
pylab.imshow(im7)
```

Out[16]: <matplotlib.image.AxesImage at 0x50dbc50>



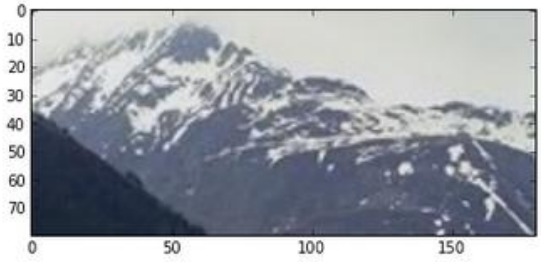
In [17]:
```
im8 = land.transpose(Image.ROTATE_270)
pylab.imshow(im8)
```

Out[17]: <matplotlib.image.AxesImage at 0x582f090>



In [31]:
```
im9 = land.crop((20, 20, 200, 100))
pylab.imshow(im9)
```

Out[31]:
```
<matplotlib.image.AxesImage at 0x78826d0>
```



In [1]:
```
import pandas as pd
ts = pd.read_csv('Gold.csv', index_col=0, parse_dates=True)
ts
```

Out[1]:
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 125 entries, 2003-01-31 00:00:00 to 2013-05-31 00:00:00
Data columns (total 1 columns):
price    125  non-null values
dtypes: float64(1)
```

In [2]:
```
ts.plot()
```

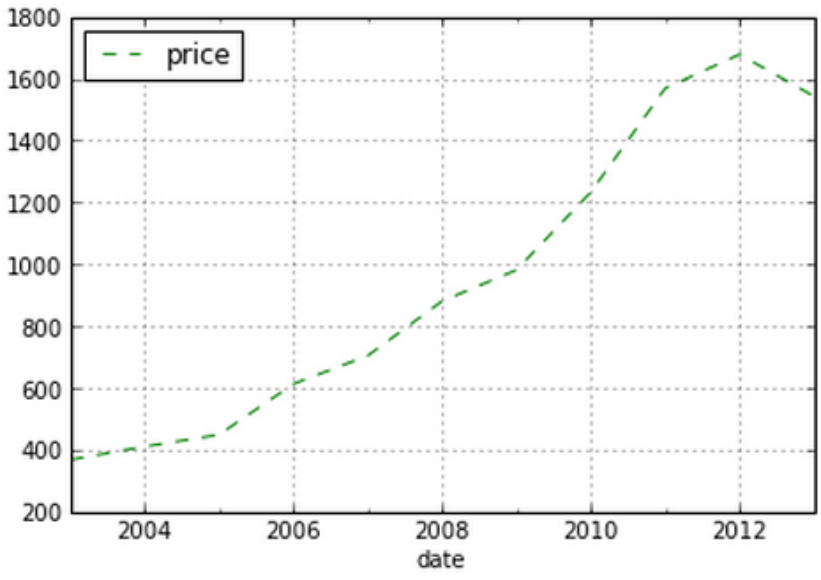
Out[2]:
```
<matplotlib.axes.AxesSubplot at 0x3a5f810>
```

In [3]:
```
ts["2006":"2007"].plot(color = "green")
```

Out[3]: <matplotlib.axes.AxesSubplot at 0x4464750>

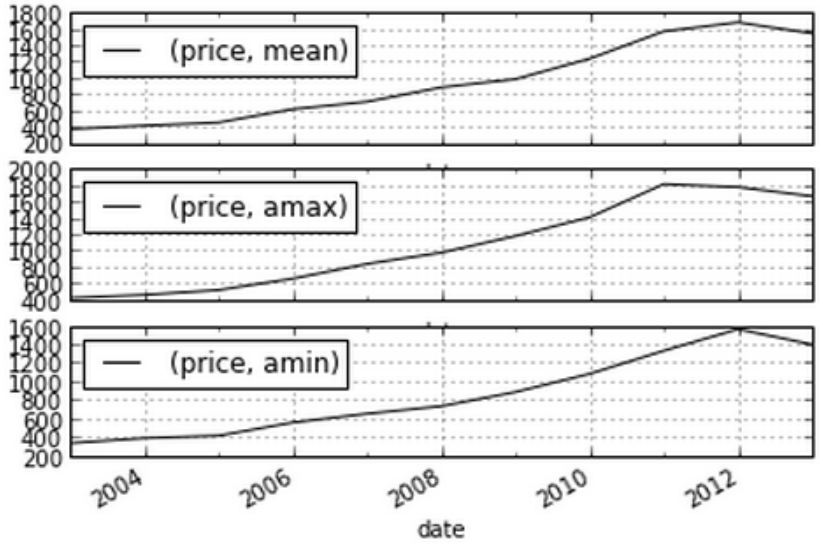
In [4]:
```
ts_res = ts.resample("A")
ts_res.plot(style='g--')
```

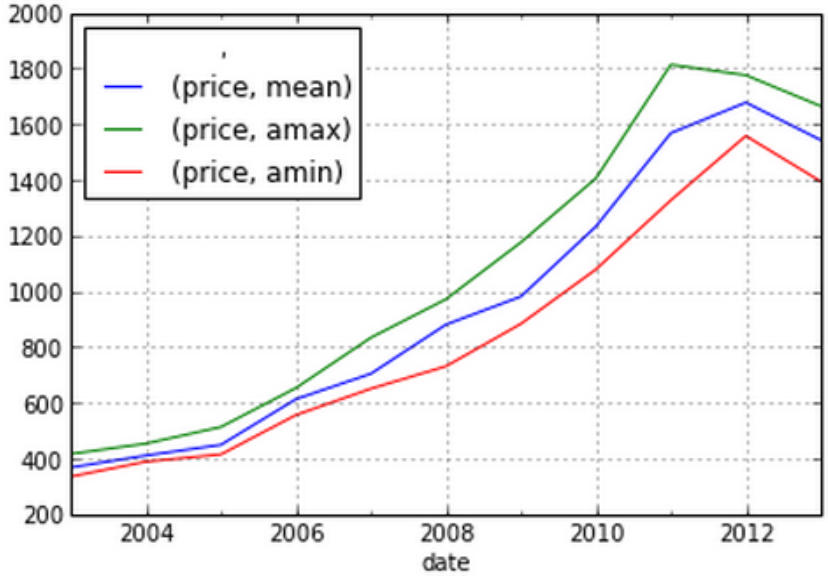
Out[4]: <matplotlib.axes.AxesSubplot at 0x42d8290>



In [5]:
```
ts_res = ts.resample("A", how=["mean", np.max, np.min])
ts_res.plot(subplots=True)
ts_res.plot()
```

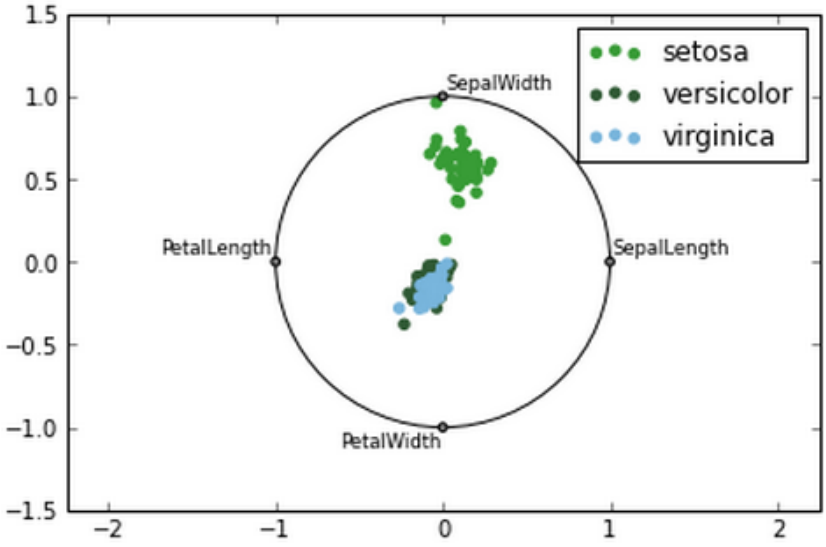
Out[5]: <matplotlib.axes.AxesSubplot at 0x4daf710>





In [6]:
```
iris = pd.read_csv("iris.csv")
pd.tools.plotting.radviz(iris, "name")
```

Out[6]: <matplotlib.axes.AxesSubplot at 0x4762490>



In [7]:
```
iris.head()
```

Out[7]:

	name	SepalLength	SepalWidth	PetalLength	PetalWidth
0	setosa	5.1	3.5	1.4	0.2
1	setosa	4.9	3.0	1.4	0.2
2	setosa	4.7	3.2	1.3	0.2
3	setosa	4.6	3.1	1.5	0.2
4	setosa	5.0	3.6	1.4	0.2

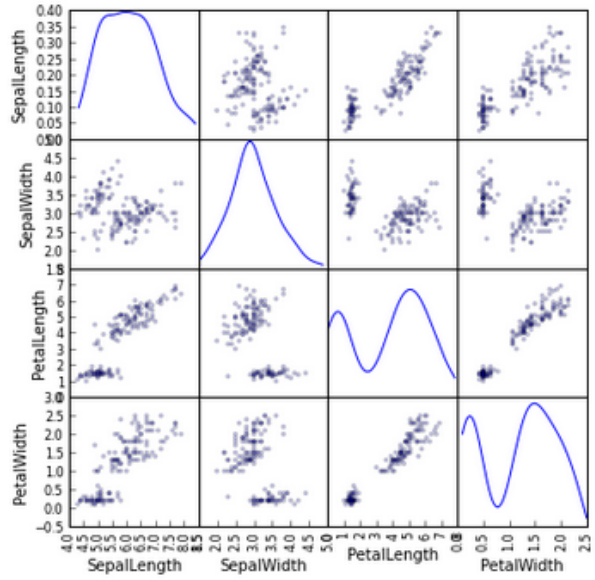
In [10]:
```
iris.describe()
```

Out[10]:

	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [6]:
```
from pandas.tools.plotting import scatter_matrix
scatter_matrix(iris, alpha=0.2, figsize=(6, 6), diagonal='kde')
```

Out[6]:
```
array([[<matplotlib.axes.AxesSubplot object at 0x337f450>,
        <matplotlib.axes.AxesSubplot object at 0x3b6cf50>,
        <matplotlib.axes.AxesSubplot object at 0x3dab550>,
        <matplotlib.axes.AxesSubplot object at 0x3d96e90>],
       [<matplotlib.axes.AxesSubplot object at 0x3f34b50>,
        <matplotlib.axes.AxesSubplot object at 0x3f57e90>,
        <matplotlib.axes.AxesSubplot object at 0x3f40410>,
        <matplotlib.axes.AxesSubplot object at 0x40d4490>],
       [<matplotlib.axes.AxesSubplot object at 0x42517d0>,
        <matplotlib.axes.AxesSubplot object at 0x40dbcd0>,
        <matplotlib.axes.AxesSubplot object at 0x43c8d90>,
        <matplotlib.axes.AxesSubplot object at 0x43eb110>],
       [<matplotlib.axes.AxesSubplot object at 0x43d2550>,
        <matplotlib.axes.AxesSubplot object at 0x45766d0>,
        <matplotlib.axes.AxesSubplot object at 0x4597a10>,
        <matplotlib.axes.AxesSubplot object at 0x457ce50>]], dtype=object)
```



In [11]:
```
iris.groupby("name").sum()
```

Out[11]:

	SepalLength	SepalWidth	PetalLength	PetalWidth
name				
setosa	250.3	171.4	73.1	12.3
versicolor	296.8	138.5	213.0	66.3
virginica	329.4	148.7	277.6	101.3

In [12]:
```
iris.groupby("name").max()
```

Out[12]:

	SepalLength	SepalWidth	PetalLength	PetalWidth
name				
setosa	5.8	4.4	1.9	0.6
versicolor	7.0	3.4	5.1	1.8
virginica	7.9	3.8	6.9	2.5

In [13]:
```
iris.groupby("name").min()
```

Out[13]:

	SepalLength	SepalWidth	PetalLength	PetalWidth
name				
setosa	4.3	2.3	1.0	0.1
versicolor	4.9	2.0	3.0	1.0
virginica	4.9	2.2	4.5	1.4

In [14]:
```
iris.groupby("name").describe()
```

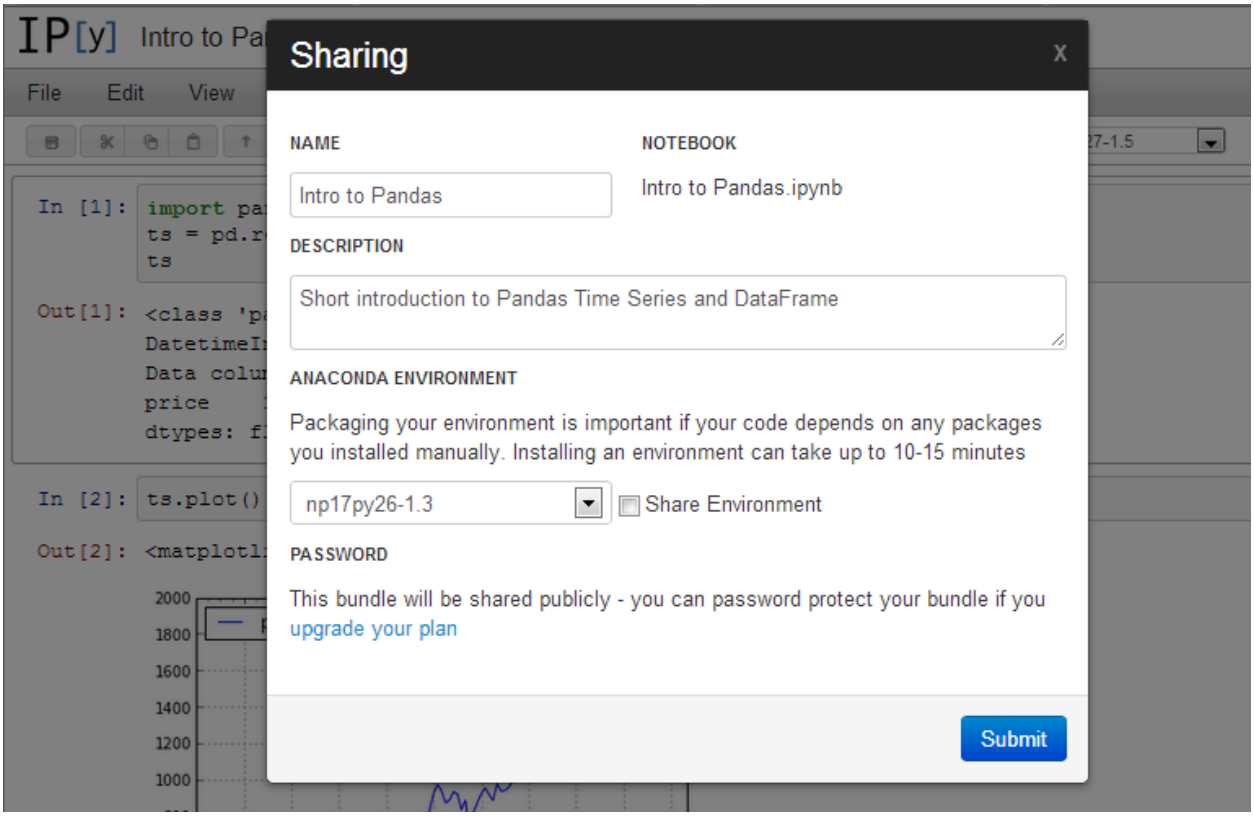
Out[14]:

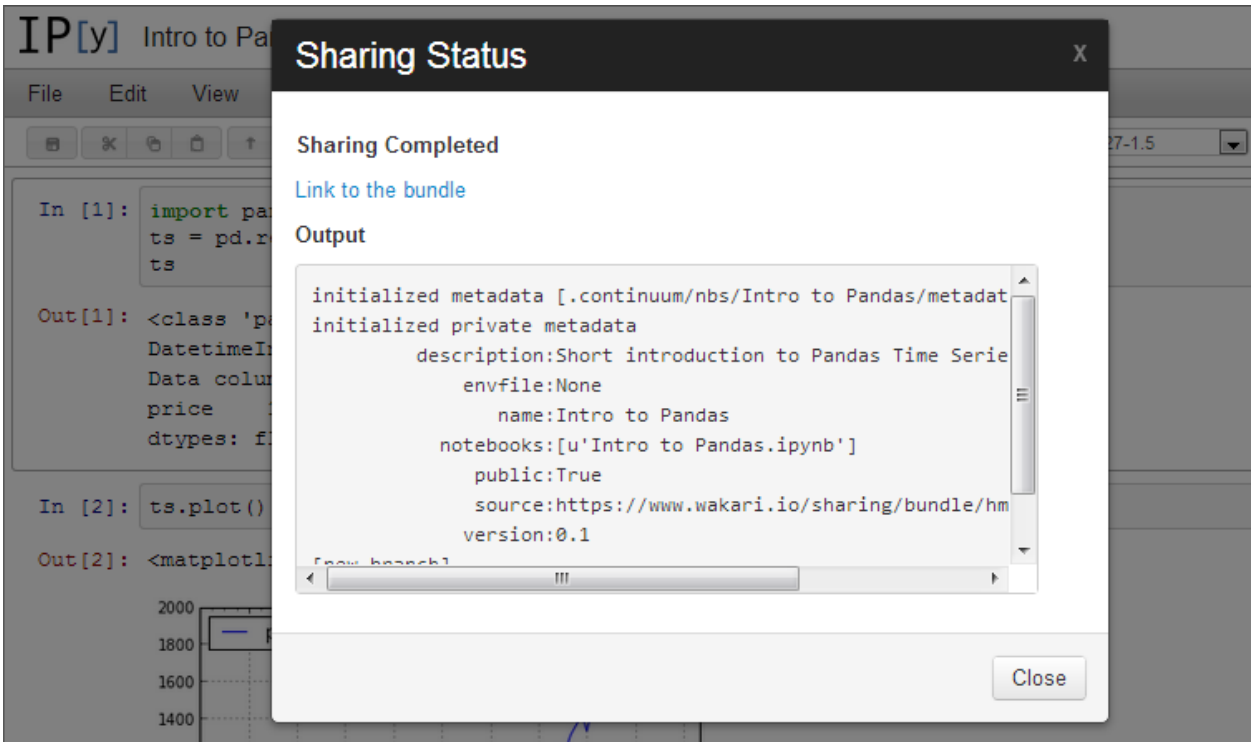
		SepalLength	SepalWidth	PetalLength	PetalWidth
name					
setosa	**count**	50.000000	50.000000	50.000000	50.000000
	mean	5.006000	3.428000	1.462000	0.246000
	std	0.352490	0.379064	0.173664	0.105386
	min	4.300000	2.300000	1.000000	0.100000
	25%	4.800000	3.200000	1.400000	0.200000
	50%	5.000000	3.400000	1.500000	0.200000
	75%	5.200000	3.675000	1.575000	0.300000
	max	5.800000	4.400000	1.900000	0.600000
versicolor	**count**	50.000000	50.000000	50.000000	50.000000
	mean	5.936000	2.770000	4.260000	1.326000
	std	0.516171	0.313798	0.469911	0.197753
	min	4.900000	2.000000	3.000000	1.000000
	25%	5.600000	2.525000	4.000000	1.200000
	50%	5.900000	2.800000	4.350000	1.300000
	75%	6.300000	3.000000	4.600000	1.500000
	max	7.000000	3.400000	5.100000	1.800000
virginica	**count**	50.000000	50.000000	50.000000	50.000000
	mean	6.588000	2.974000	5.552000	2.026000
	std	0.635880	0.322497	0.551895	0.274650
	min	4.900000	2.200000	4.500000	1.400000
	25%	6.225000	2.800000	5.100000	1.800000
	50%	6.500000	3.000000	5.550000	2.000000
	75%	6.900000	3.175000	5.875000	2.300000
	max	7.900000	3.800000	6.900000	2.500000

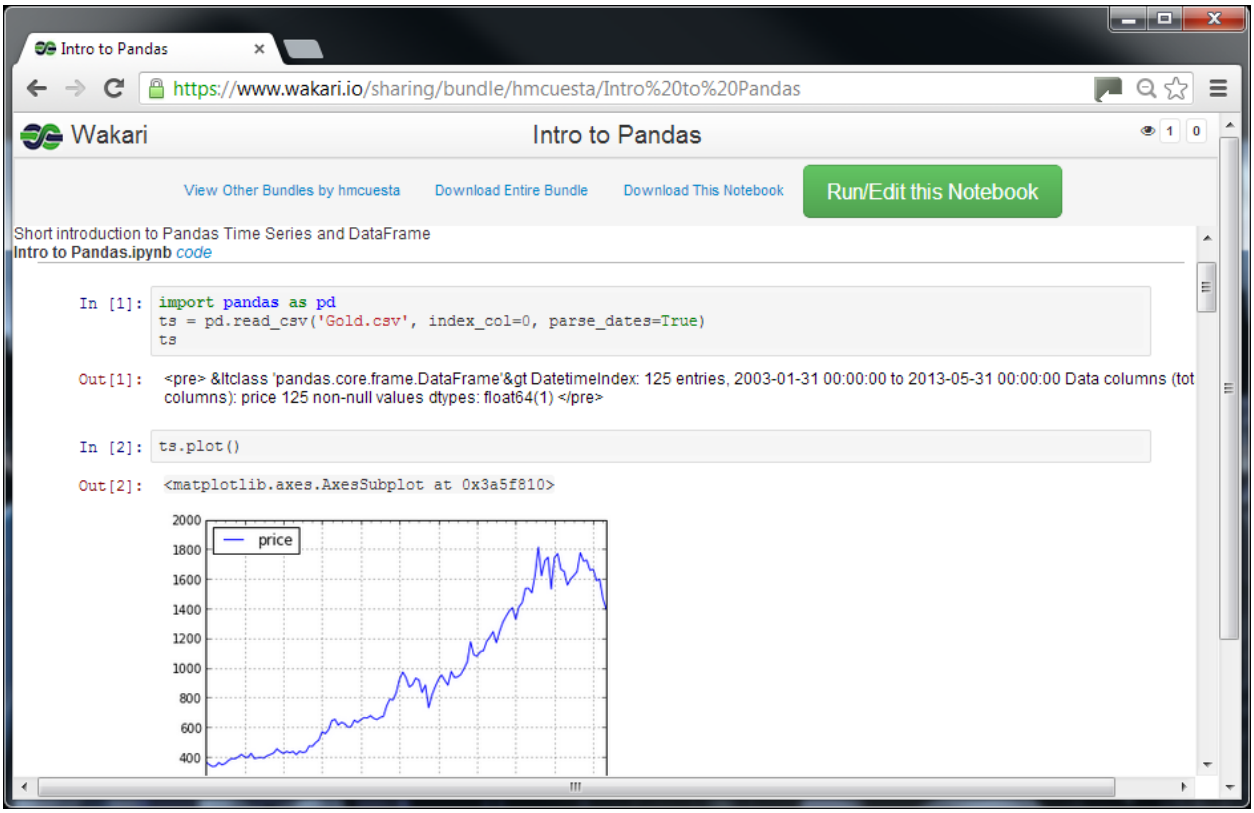
```
In [15]: iris["SepalLength"].corr(iris["PetalLength"])

Out[15]: 0.87175377588658287

In [16]: iris.corr()
```

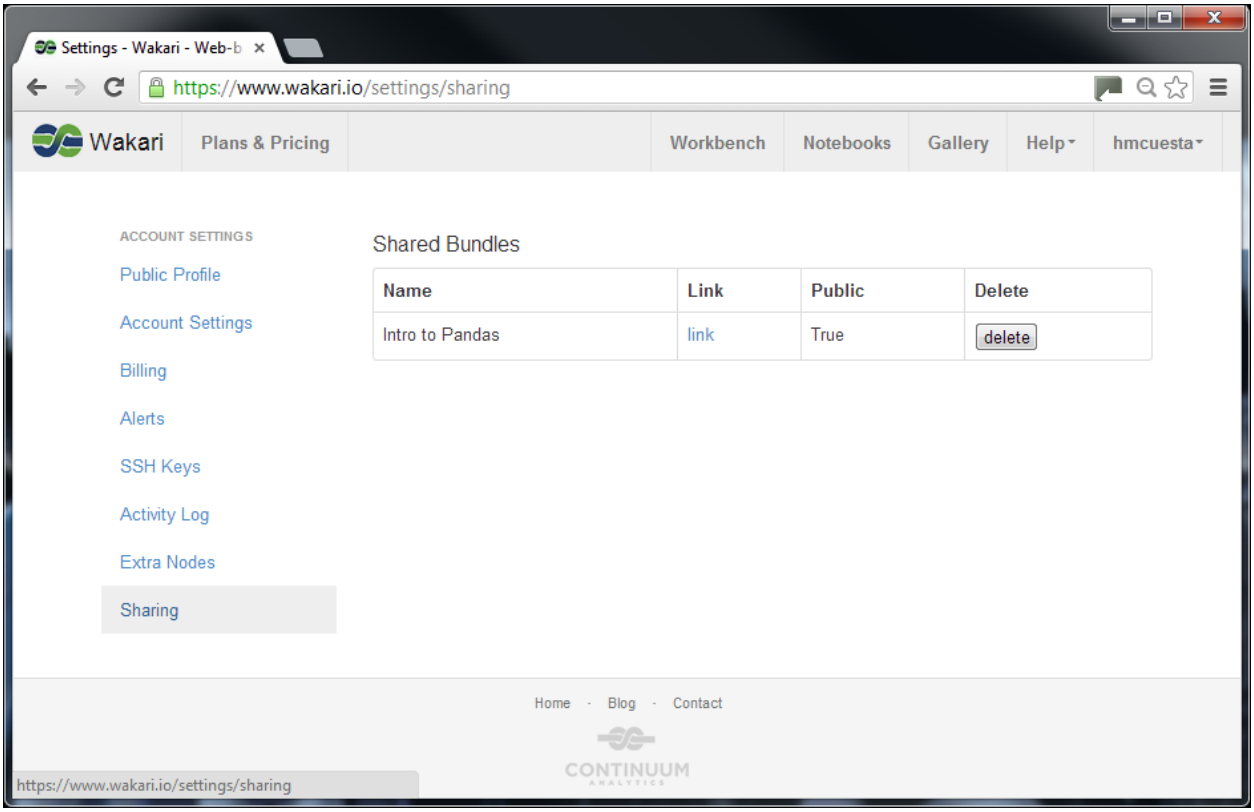
Out[16]:

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	1.000000	-0.117570	0.871754	0.817941
SepalWidth	-0.117570	1.000000	-0.428440	-0.366126
PetalLength	0.871754	-0.428440	1.000000	0.962865
PetalWidth	0.817941	-0.366126	0.962865	1.000000

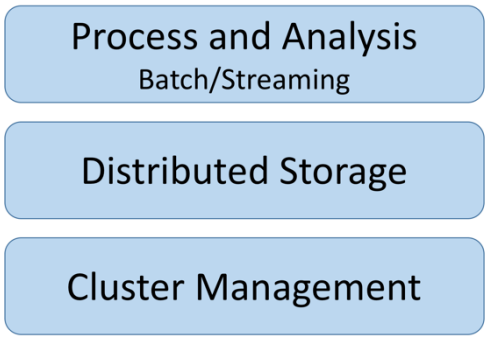


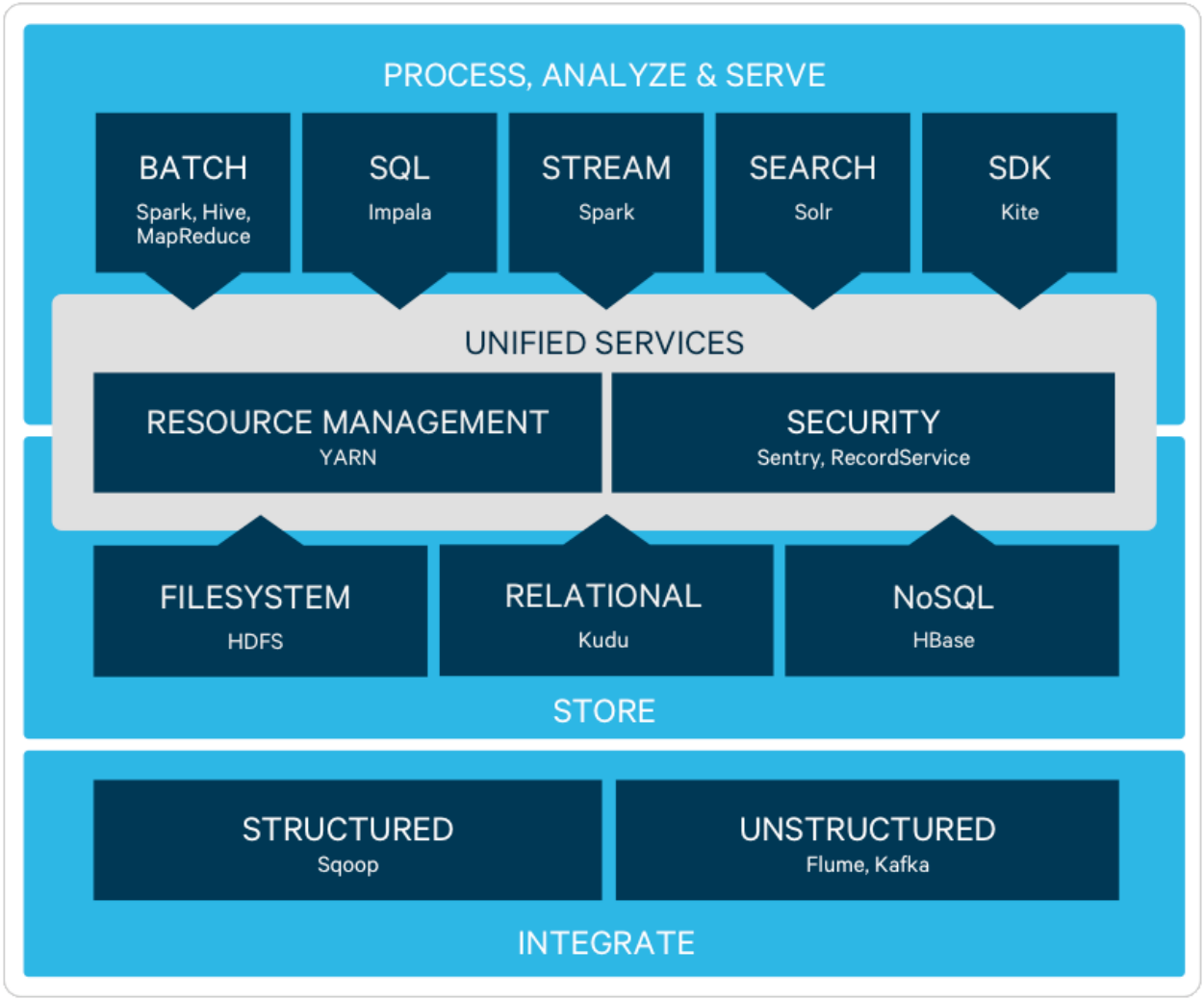

IP[y] Intro to Pa
File Edit View
Sharing Status
x
Sharing Completed
Link to the bundle
Output
initialized metadata [.continuum/nbs/Intro to Pandas/metadat
initialized private metadata
description:Short introduction to Pandas Time Serie
envfile:None
name:Intro to Pandas
notebooks:[u'Intro to Pandas.ipynb']
public:True
source:https://www.wakari.io/sharing/bundle/hm
version:0.1
Close
In [1]:
import pa
ts = pd.r
ts
Out[1]:
<class 'p
DatetimeI
Data colu
price
dtypes: f
In [2]:
ts.plot()
Out[2]:
<matplotl
2000
1800
1600
1400

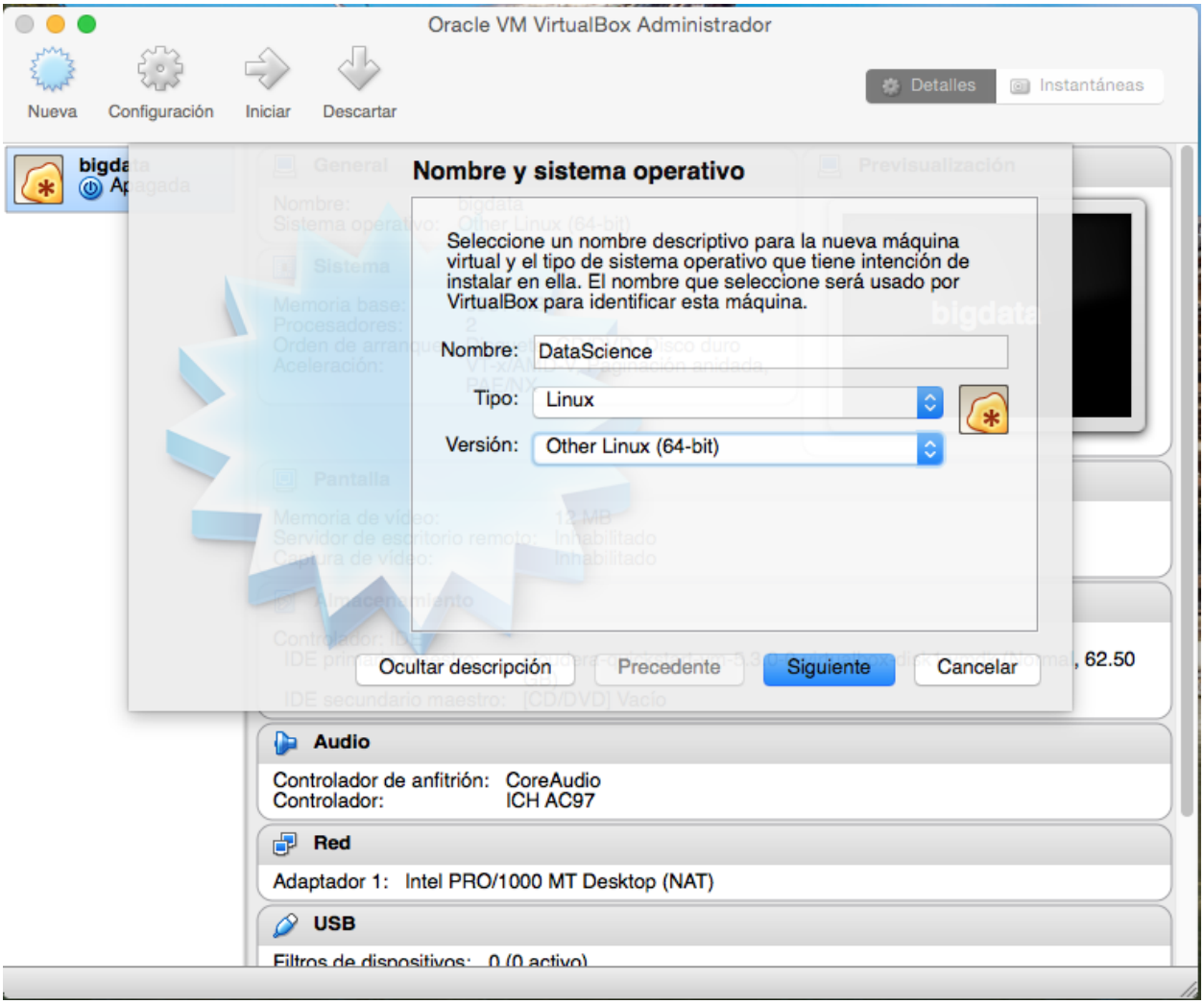

Intro to Pandas
https://www.wakari.io/sharing/bundle/hmcuesta/Intro%20to%20Pandas
Wakari
Intro to Pandas
View Other Bundles by hmcuesta
Download Entire Bundle
Download This Notebook
Run/Edit this Notebook
Short introduction to Pandas Time Series and DataFrame
Intro to Pandas.ipynb code
In [1]:
import pandas as pd
ts = pd.read_csv('Gold.csv', index_col=0, parse_dates=True)
ts
Out[1]:
<pre> &ltclass 'pandas.core.frame.DataFrame'&gt DatetimeIndex: 125 entries, 2003-01-31 00:00:00 to 2013-05-31 00:00:00 Data columns (tot
columns): price 125 non-null values dtypes: float64(1) </pre>
In [2]:
ts.plot()
Out[2]:
<matplotlib.axes.AxesSubplot at 0x3a5f810>
price
2000
1800
1600
1400
1200
1000
800
600
400

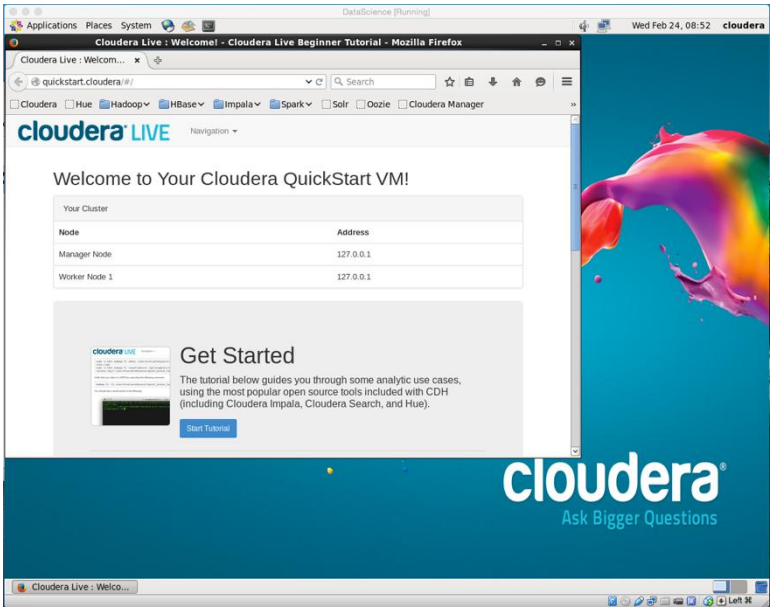

Settings - Wakari - Web-b
https://www.wakari.io/settings/sharing
Wakari
Plans & Pricing
Workbench
Notebooks
Gallery
Help
hmcuesta
ACCOUNT SETTINGS
Public Profile
Account Settings
Billing
Alerts
SSH Keys
Activity Log
Extra Nodes
Sharing
Shared Bundles
Name
Link
Public
Delete
Intro to Pandas
link
True
delete
Home · Blog · Contact
CONTINUUM
https://www.wakari.io/settings/sharing

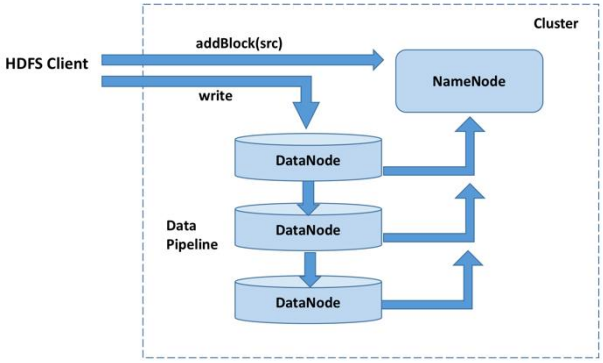
Chapter 15: Understanding Data Processing using Apache Spark

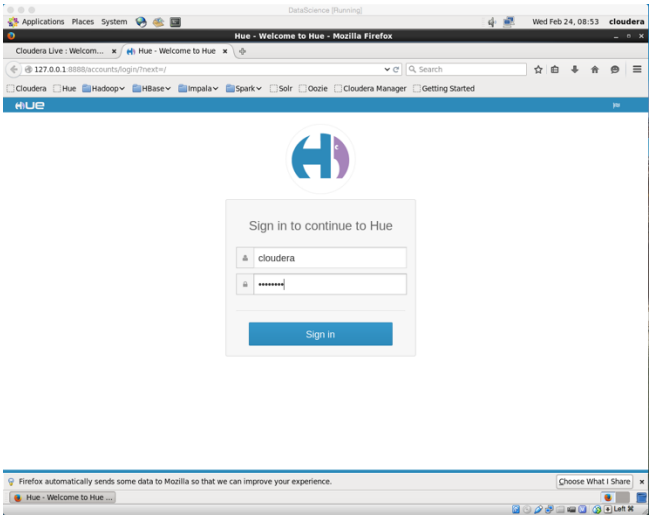


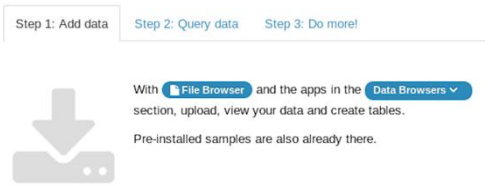


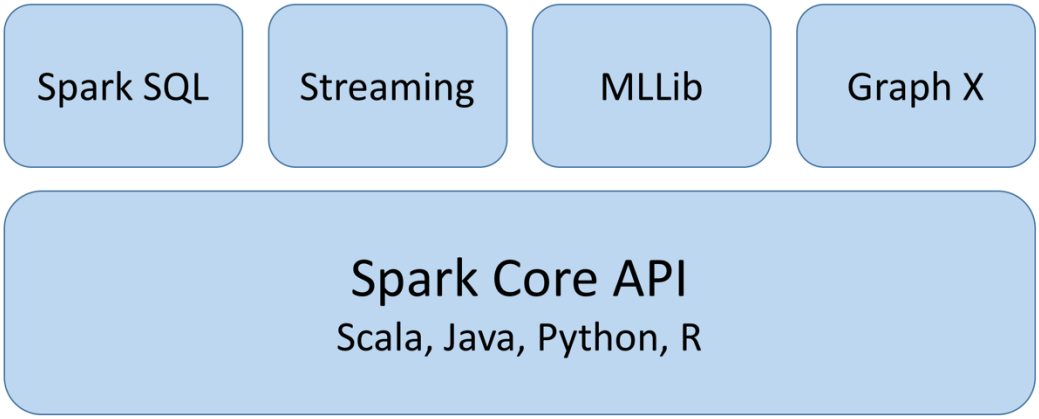

Oracle VM VirtualBox Administrador
Nueva
Configuración
Iniciar
Descartar
Detalles
Instantáneas
Nombre y sistema operativo
Seleccione un nombre descriptivo para la nueva máquina virtual y el tipo de sistema operativo que tiene intención de instalar en ella. El nombre que seleccione será usado por VirtualBox para identificar esta máquina.
Nombre: DataScience
Tipo: Linux
Versión: Other Linux (64-bit)
Ocultar descripción
Precedente
Siguiente
Cancelar
Audio
Controlador de anfitrión: CoreAudio
Controlador: ICH AC97
Red
Adaptador 1: Intel PRO/1000 MT Desktop (NAT)
USB

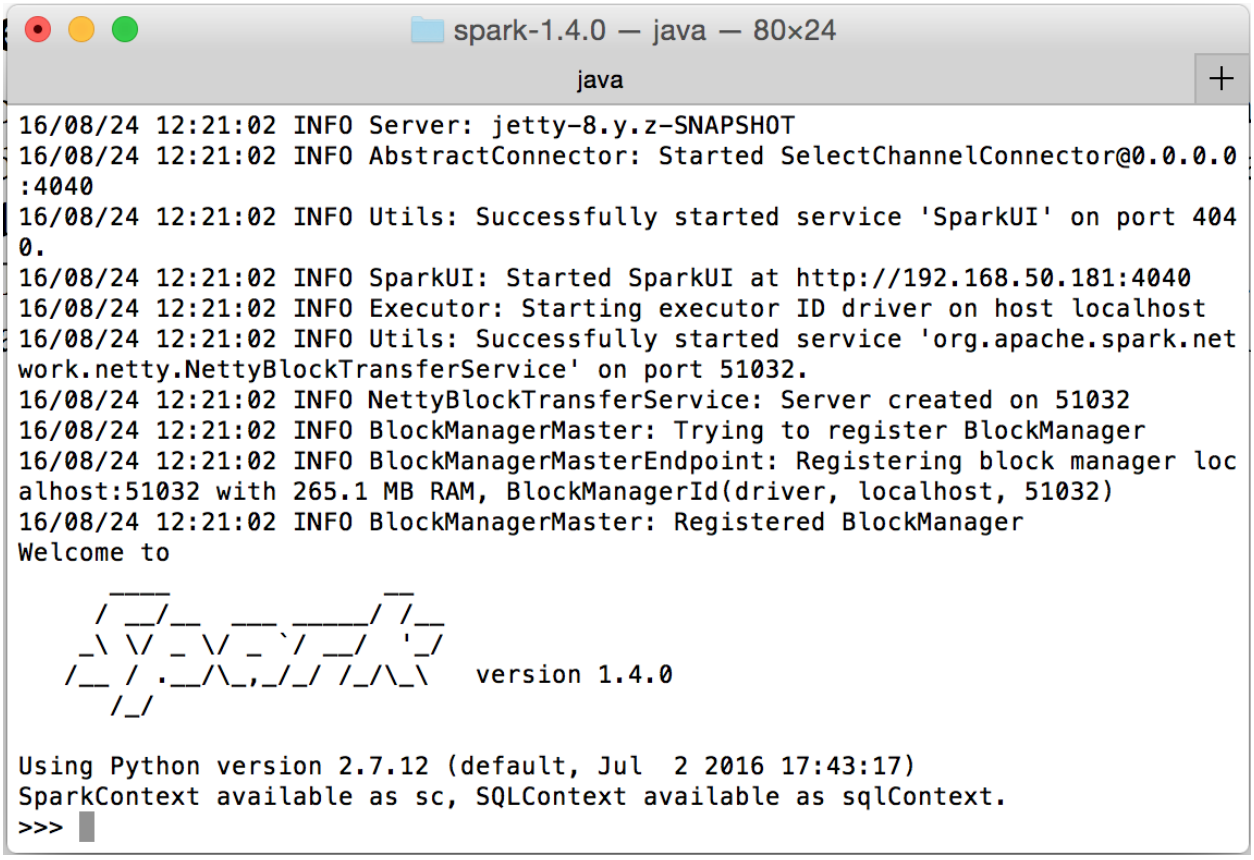

Applications Places System
Wed Feb 24, 08:52 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
Get Started
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Start Tutorial
cloudera
Ask Bigger Questions
Cloudera Live : Welco...

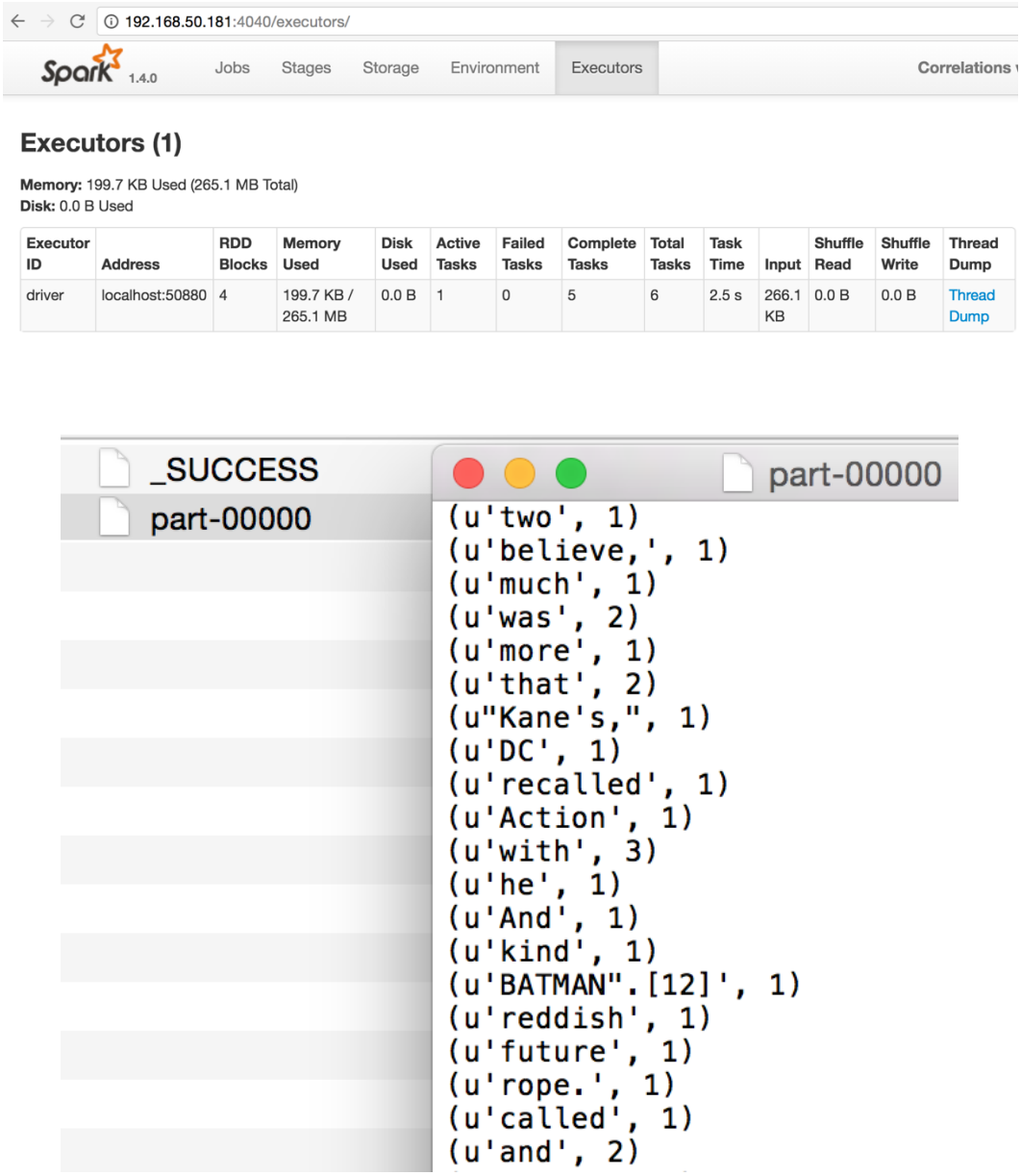

Cluster
HDFS Client
addBlock(src)
write
NameNode
DataNode
DataNode
DataNode
Data
Pipeline


Applications Places System
Wed Feb 24, 08:53 cloudera
Hue - Welcome to Hue - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Welcome to Hue
127.0.0.1:8888/accounts/login?next=/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Sign in to continue to Hue
cloudera
Sign in
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Hue - Welcome to Hue ...


Step 1: Add data
Step 2: Query data
Step 3: Do more!
With File Browser and the apps in the Data Browsers
section, upload, view your data and create tables.
Pre-installed samples are also already there.


Spark SQL
Streaming
MLLib
Graph X
Spark Core API
Scala, Java, Python, R


spark-1.4.0 — java — 80×24
java
16/08/24 12:21:02 INFO Server: jetty-8.y.z-SNAPSHOT
16/08/24 12:21:02 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0
:4040
16/08/24 12:21:02 INFO Utils: Successfully started service 'SparkUI' on port 404
0.
16/08/24 12:21:02 INFO SparkUI: Started SparkUI at http://192.168.50.181:4040
16/08/24 12:21:02 INFO Executor: Starting executor ID driver on host localhost
16/08/24 12:21:02 INFO Utils: Successfully started service 'org.apache.spark.net
work.netty.NettyBlockTransferService' on port 51032.
16/08/24 12:21:02 INFO NettyBlockTransferService: Server created on 51032
16/08/24 12:21:02 INFO BlockManagerMaster: Trying to register BlockManager
16/08/24 12:21:02 INFO BlockManagerMasterEndpoint: Registering block manager loc
alhost:51032 with 265.1 MB RAM, BlockManagerId(driver, localhost, 51032)
16/08/24 12:21:02 INFO BlockManagerMaster: Registered BlockManager
Welcome to
version 1.4.0
Using Python version 2.7.12 (default, Jul 2 2016 17:43:17)
SparkContext available as sc, SQLContext available as sqlContext.
>>>


192.168.50.181:4040/executors/
Spark 1.4.0
Jobs
Stages
Storage
Environment
Executors
Correlations
Executors (1)
Memory: 199.7 KB Used (265.1 MB Total)
Disk: 0.0 B Used
Executor ID
Address
RDD Blocks
Memory Used
Disk Used
Active Tasks
Failed Tasks
Complete Tasks
Total Tasks
Task Time
Input
Shuffle Read
Shuffle Write
Thread Dump
driver
localhost:50880
4
199.7 KB / 265.1 MB
0.0 B
1
0
5
6
2.5 s
266.1 KB
0.0 B
0.0 B
Thread Dump

_SUCCESS
part-00000
part-00000
(u'two', 1)
(u'believe,', 1)
(u'much', 1)
(u'was', 2)
(u'more', 1)
(u'that', 2)
(u"Kane's,", 1)
(u'DC', 1)
(u'recalled', 1)
(u'Action', 1)
(u'with', 3)
(u'he', 1)
(u'And', 1)
(u'kind', 1)
(u'BATMAN".[12]', 1)
(u'reddish', 1)
(u'future', 1)
(u'rope.', 1)
(u'called', 1)
(u'and', 2)