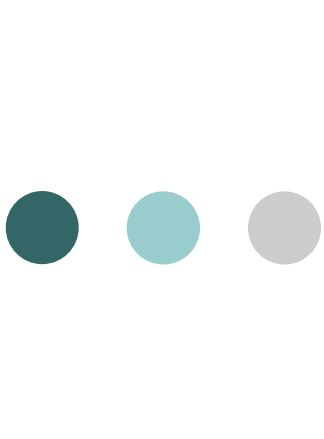


Social Data Mining

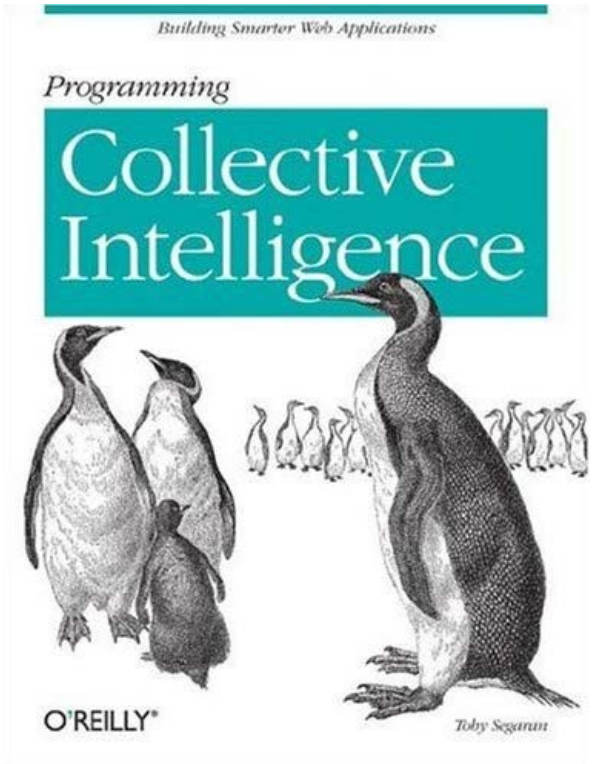
Toby Segaran

About Me

Genstruct
Knowledge Driven Discovery

collective intellect

PEOPLE-POWERED
SATISFACTION
CUSTOMER SERVICE



http://kiwitobes.com

What is data mining?

- Implicit
- Unknown
- Useful

What is data?

.1

COJ: COUNCIL 2004-06-...?
COJ : MAYORAL COMMITTEE 2004-06-10
COJ: FINANCE, STRATEGY & ECONOMIC DEVELOPMENT COMMITTEE 2004-06-07

(ECONOMIC DEVELOPMENT UNIT)

URBAN RENEWAL TAX INCENTIVE
SUBMISSION TO NATIONAL TREASURY

1. STRATEGIC THRUST

   Economic Growth and Development, Housing Delivery, Service Delivery Excellence, Inner City Regeneration and Good Governance

2. OBJECTIVE

   To inform committees on the City of Johannesburg's (CoJ) application to National Treasury for the designation of the Inner City as Johannesburg's Urban Development Zone (UDZ) in line with the National Treasury's urban renewal tax incentive.

3. BACKGROUND

   Legislative proposals dealing with the Urban Renewal Tax Incentive were signed into law in December 2003, i.e. "Revenue Laws Amendment Act", 2003: Insertion of section 13*quat* in Act 58 of 1962 (Income Tax Act): *"Deductions in respect of erection or improvement of buildings in urban development zones"*.

   The proposed tax incentive applies to the construction and refurbishment of both commercial and residential buildings in designated decayed or under-utilised inner city areas and will be available to owners or lessors of the properties. The incentive will take the form of a special depreciation allowance for the refurbishment of existing buildings (taxpayers will receive a 20% straight-line depreciation allowance over a 5-year period) and/or the construction of new buildings (taxpayers will receive a 17-year write-off period with a 20% write-off in the first year, and a 5% write-off thereafter).

4. JOHANNESBURG'S UDZ: THE INNER CITY

   In terms of giving effect to the legislation, the CoJ has prepared an application for submission to National Treasury (see Annexure A) by the end of June 2004. Key elements contained in the application are outlined next.

   4.1 MOTIVATION FOR EXTENT OF UDZ AREA BEING APPLIED FOR BY COJ

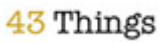
   The Inner City area is nearly 1,800 hectares in extent. The maximum permissible area allowable in terms of the legislation, related to Johannesburg's population, is 690 hectares. However, an application for a larger area can be made. A detailed motivation to the Minister of Finance for the declaration of the entire Inner City as Johannesburg's UDZ has been provided. Section A of the application describes the Inner City as a system of four main zones, each with different characteristics, i.e.:

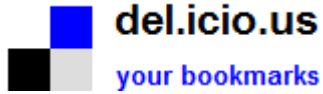
Data-mining traditional uses

Why it’s important now

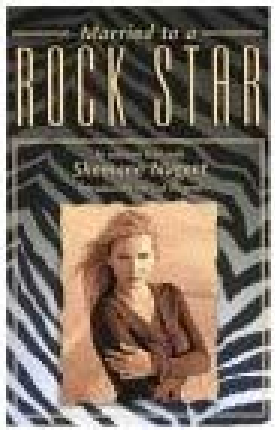
Data

Why it's important now





Why it’s important now



All products are actually sold on Amazon

Why it's important now

Facebook

Google

Sponsored Links

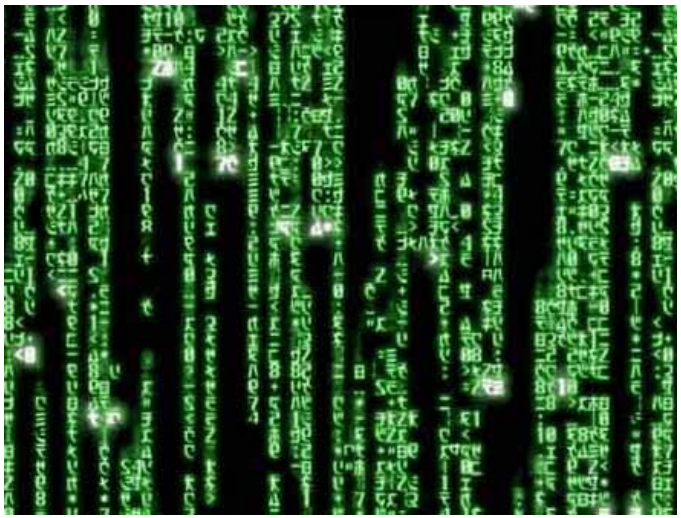
private **tours** of **Berlin**
English speaker will show you the sights of **Berlin** (or in Spanish)
www.afriendin**berlin**.de

Berlin Tours
Learn More about Destinations All Around Europe - Find Great Info!
www.BudgetTravel.com

Berlin Tours
Find and book 25+ things to do in **Berlin** on Viator.
www.viator.com/**berlin**

Berlin Sightseeing **Tours**
Tour Berlin & Surrounding Germany. Book Online & Save. 800-208-4421
www.**Berlin-Tours**.net

Why it’s important now



>

For Social Insight

Home Prices

Blogs and News

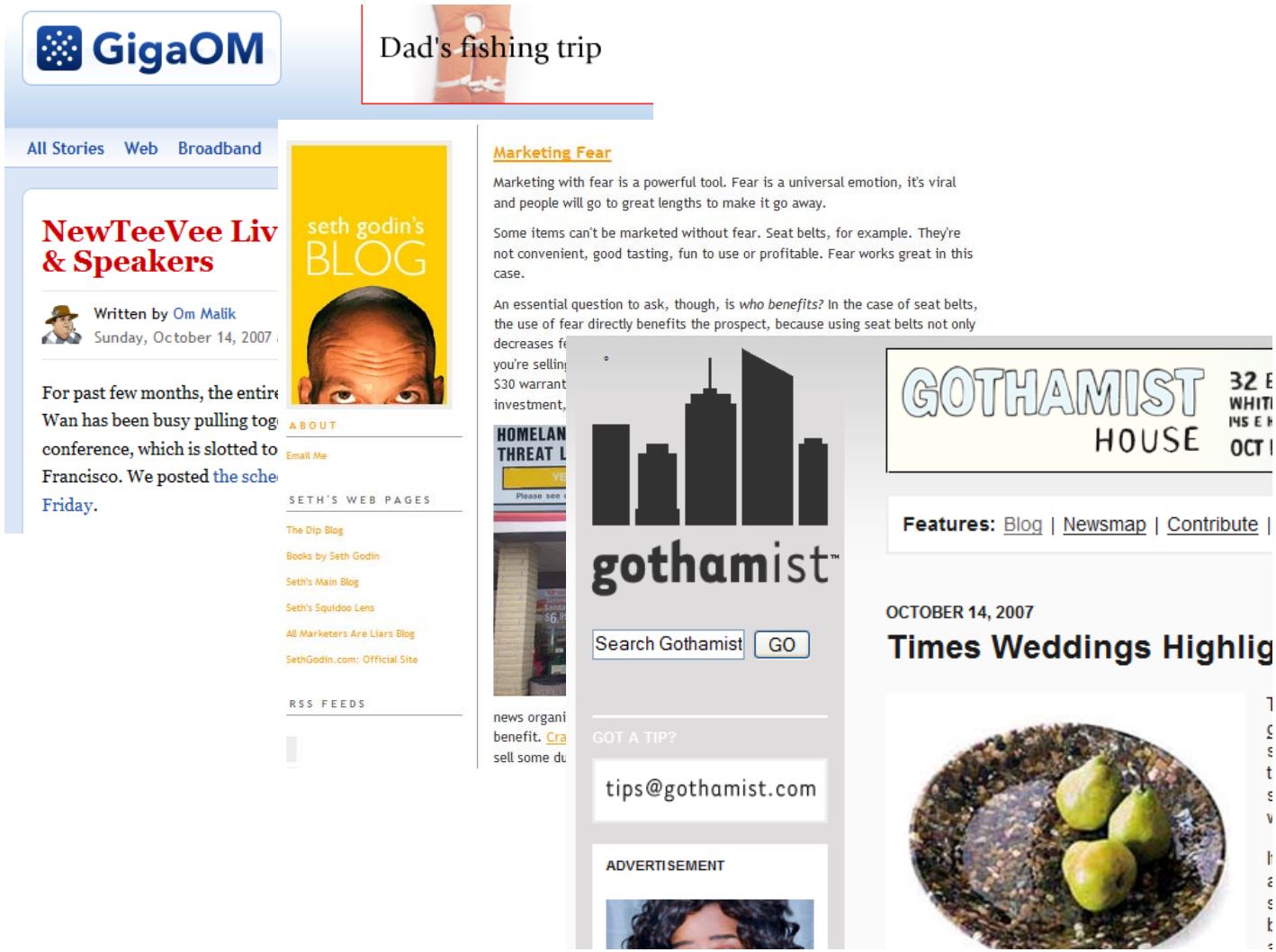
Movie Data

Fashion

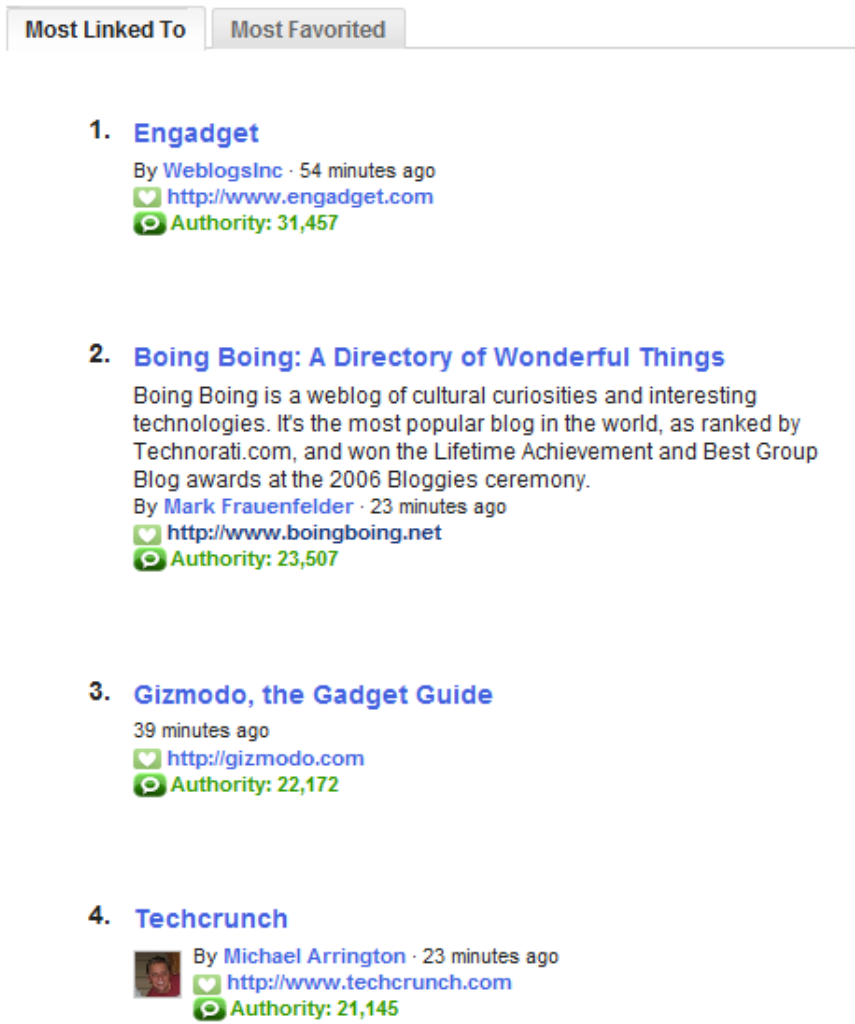
Product Prices

Hotties

Blogs…


GigaOM
Dad's fishing trip
All Stories
Web
Broadband
NewTeeVee Liv
& Speakers
Written by Om Malik
Sunday, October 14, 2007
For past few months, the entire
Wan has been busy pulling tog
conference, which is slotted to
Francisco. We posted the sche
Friday.
seth godin's
BLOG
ABOUT
Email Me
SETH'S WEB PAGES
The Dip Blog
Books by Seth Godin
Seth's Main Blog
Seth's Squidoo Lens
All Marketers Are Liars Blog
SethGodin.com: Official Site
RSS FEEDS
Marketing Fear
Marketing with fear is a powerful tool. Fear is a universal emotion, it's viral and people will go to great lengths to make it go away.
Some items can't be marketed without fear. Seat belts, for example. They're not convenient, good tasting, fun to use or profitable. Fear works great in this case.
An essential question to ask, though, is who benefits? In the case of seat belts, the use of fear directly benefits the prospect, because using seat belts not only
decreases f
you're sellin
$30 warrant
investment,
HOMELAN
THREAT L
news organi
benefit. Cra
sell some du
gothamist
Search Gothamist
GO
GOT A TIP?
tips@gothamist.com
ADVERTISEMENT
GOTHAMIST
HOUSE
Features: Blog | Newsmap | Contribute |
OCTOBER 14, 2007
Times Weddings Highlig

The Technorati Top 100



Getting the content

The

Six

Degrees

Hypothesis

Experienced

It

Is

When

You

Travel

Building a Word Matrix

The

Six

Degrees

Hypothesis

Experienced

It

Is

When

You

Travel

Six

Degrees

Hypothesis

Experienced

Travel

Six	3
Degrees	3
Hypothesis	1
Experienced	5
Travel	6

The Word Matrix

	“china”	**“kids”**	**“music”**	“travel”	**“yahoo”**
Gothamist	0	3	3	3	0
GigaOM	6	0	1	4	2
QuickOnlineTips	0	2	2	0	12
O’Reilly Radar	1	0	3	6	4

Determining distance

	"china"	"kids"	"music"	"yahoo"
Gothamist	0	3	3	0
GigaOM	6	0	1	2
Quick Online Tips	0	2	2	12

Euclidean "as the crow flies"

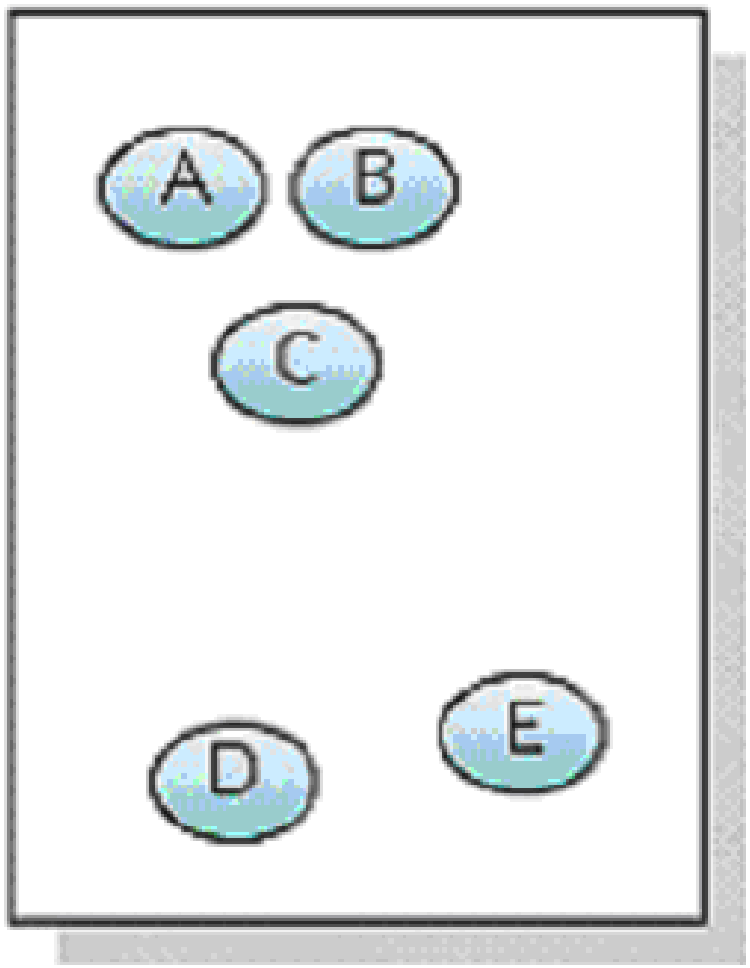
$$\sqrt{(6-0)^2+(0-2)^2+(1-2)^2+(2-12)^2}$$

= 12 (approx)

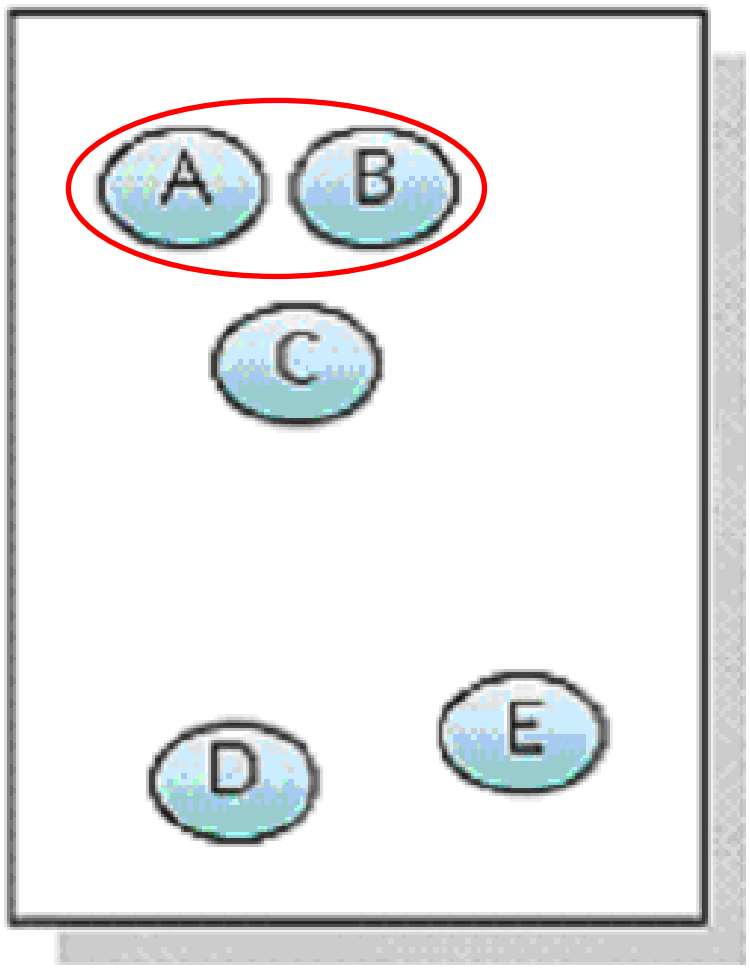
Hierarchical Clustering

- Find the two closest item
- Combine them into a single item
- Repeat…

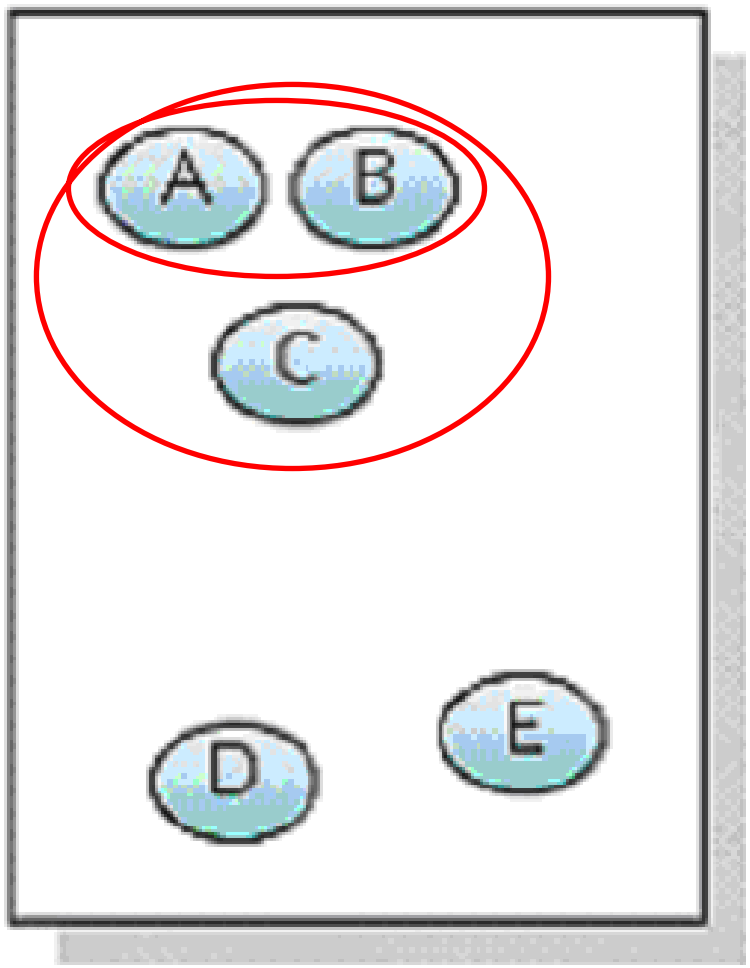
Hierarchical Algorithm



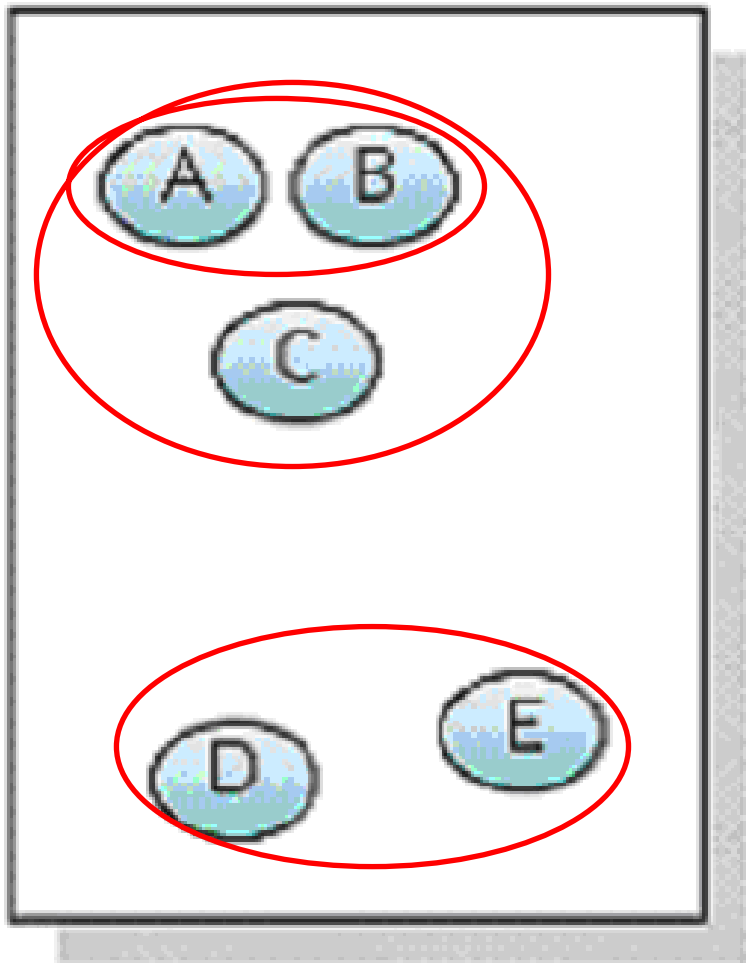
Hierarchical Algorithm



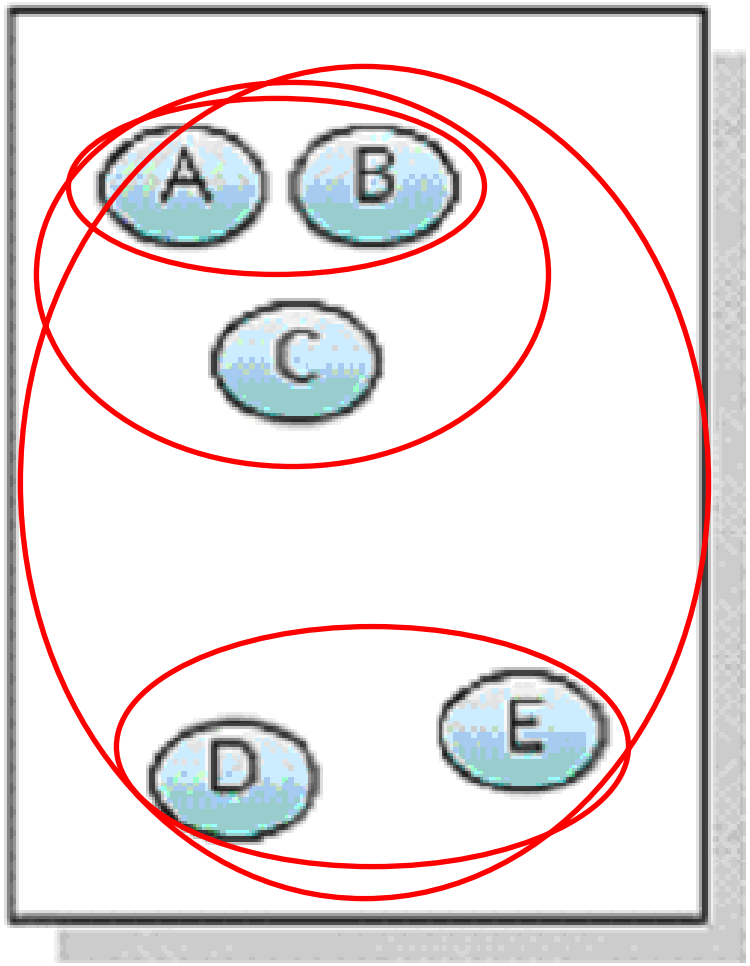
Hierarchical Algorithm



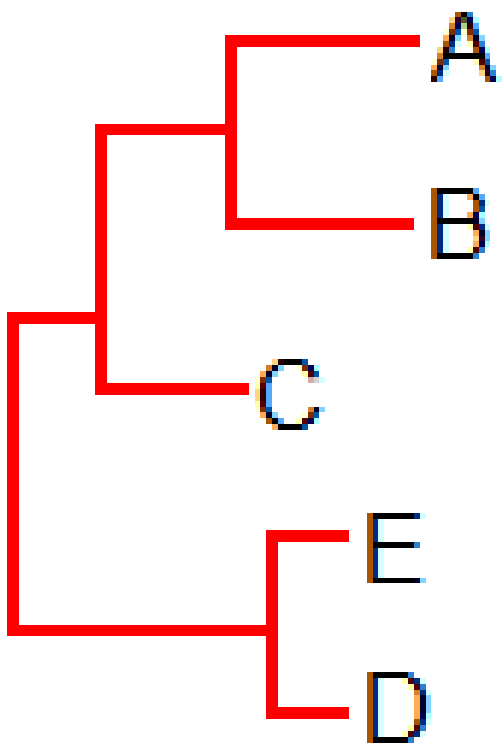
Hierarchical Algorithm



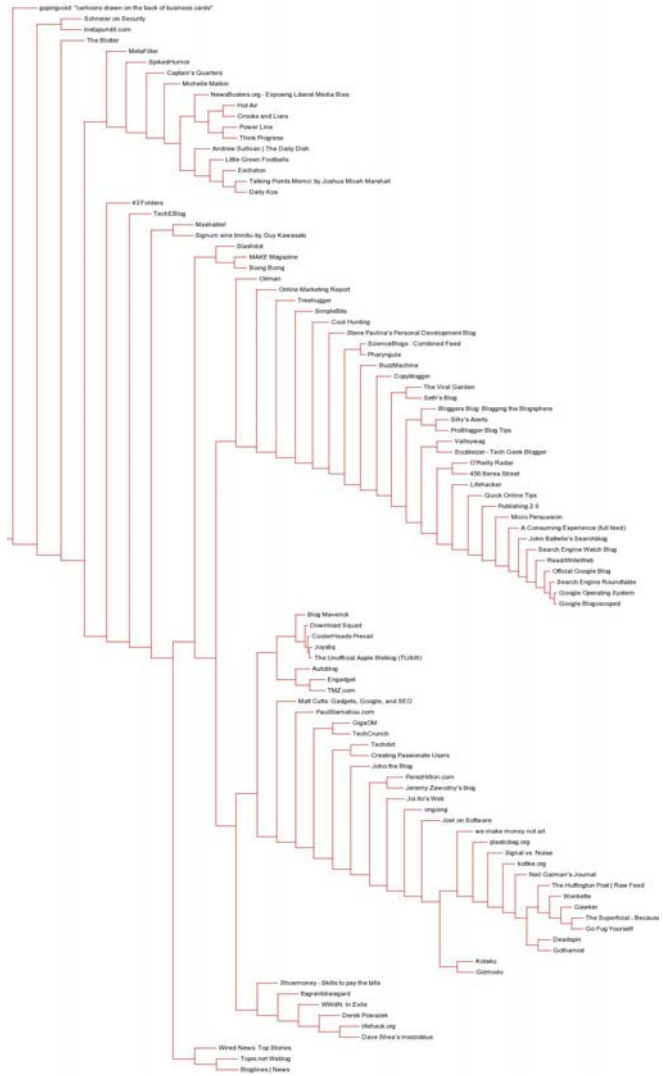
Hierarchical Algorithm



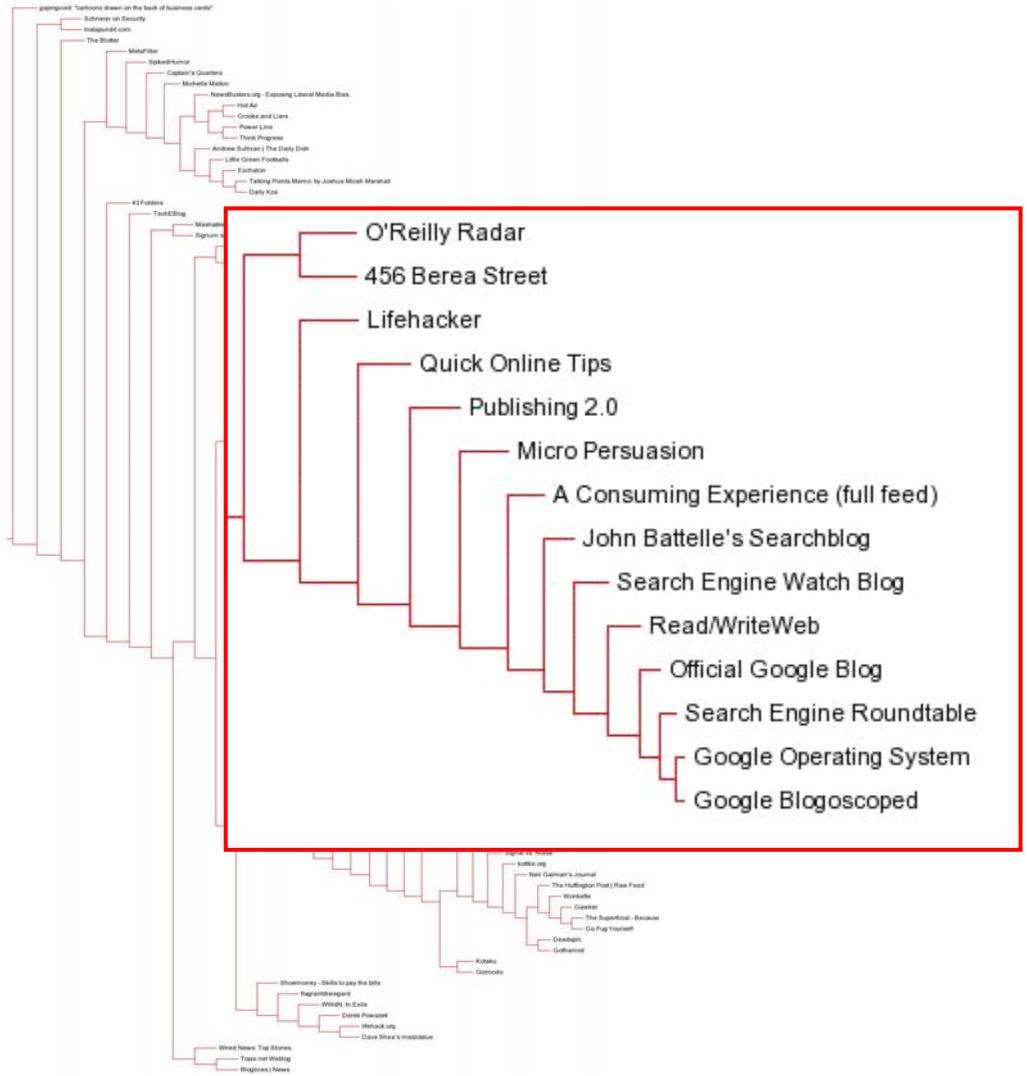
Dendrogram



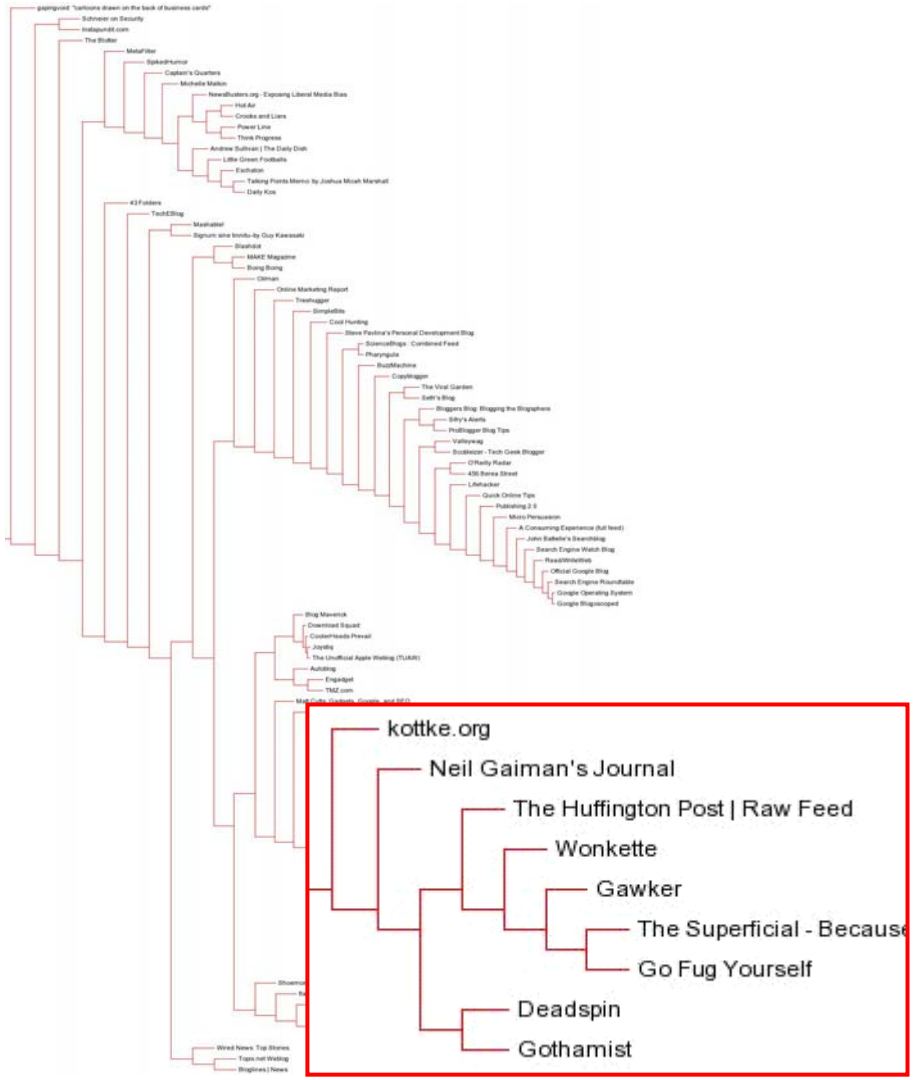
Hierarchical Blog Clusters



Hierarchical Blog Clusters



Hierarchical Blog Clusters

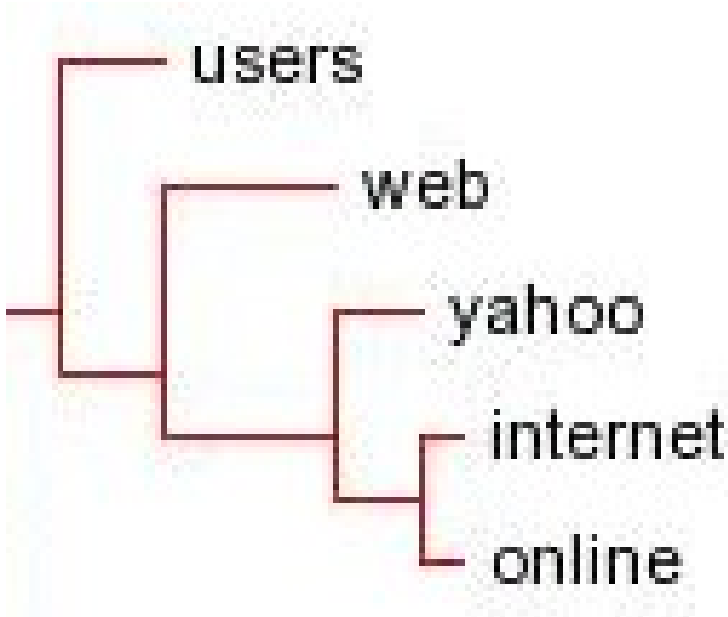


Rotating the Matrix

Words in a blog -> blogs containing each word

	Gothamist	GigaOM	Quick Onl
china	0	6	0
kids	3	0	2
music	3	1	2
Yahoo	0	2	12

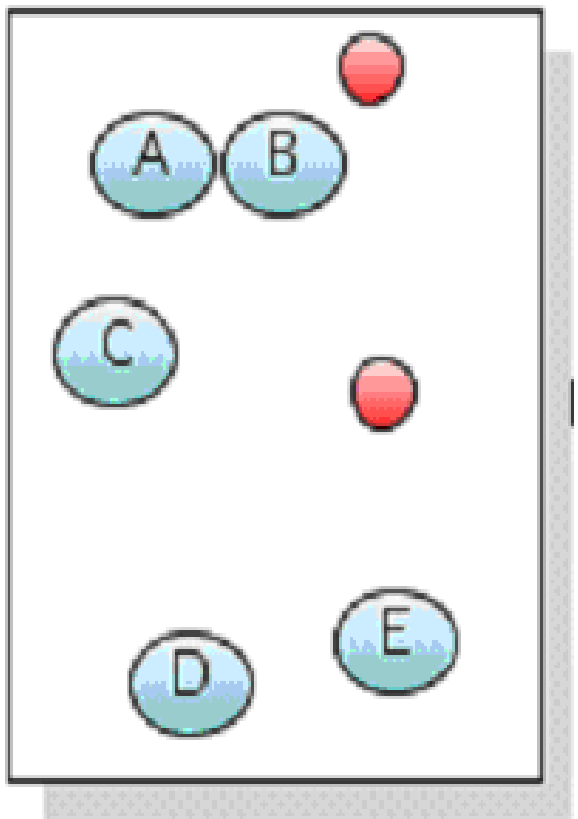
Hierarchical Word Clusters



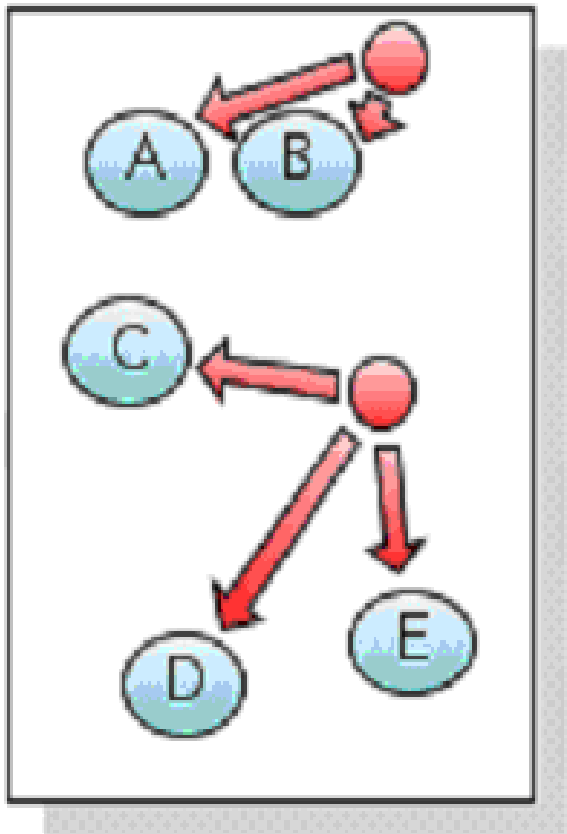
K-Means Clustering

- Divides data into distinct clusters
- User determines how many
- Algorithm
  - Start with arbitrary centroids
  - Assign points to centroids
  - Move the centroids
  - Repeat

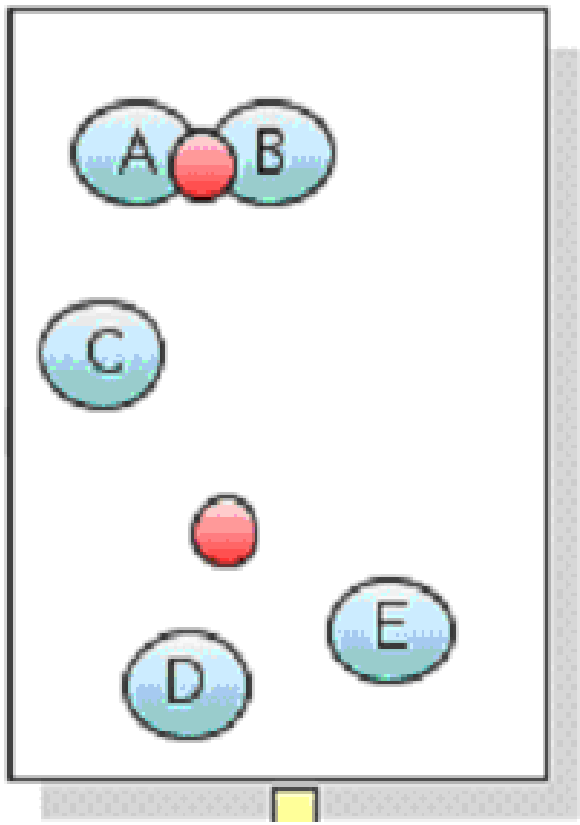
K-Means Algorithm



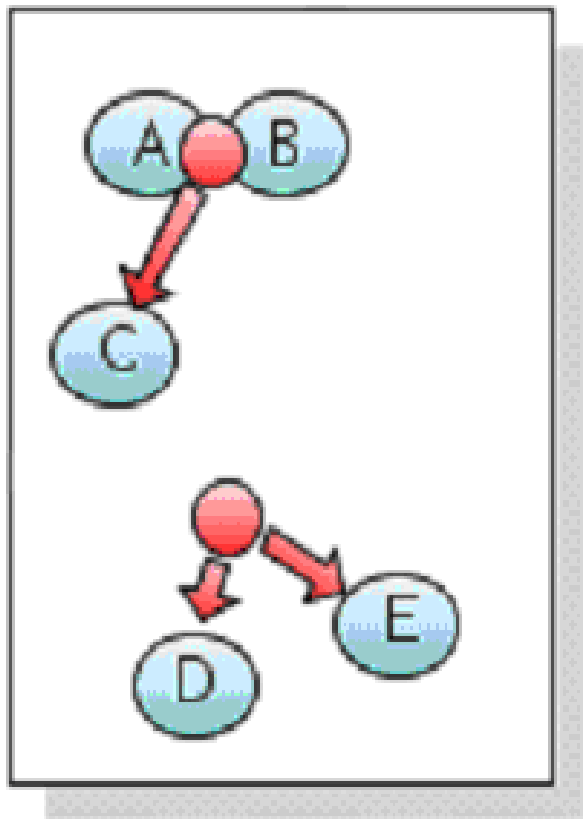
K-Means Algorithm



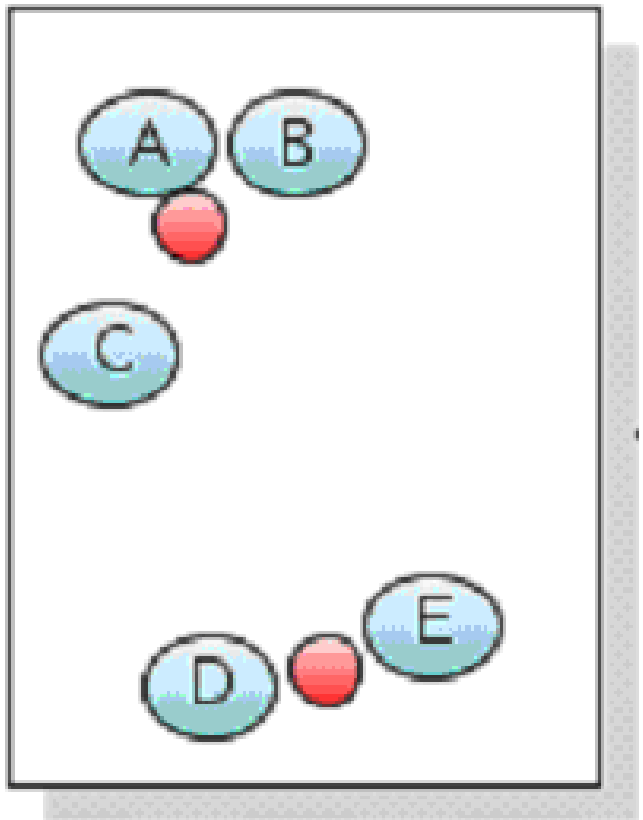
K-Means Algorithm



K-Means Algorithm



K-Means Algorithm



K-Means Results

1

The Viral Garden
Copyblogger
Creating Passionate Users
Oilman
ProBlogger Blog Tips
Seth's Blog

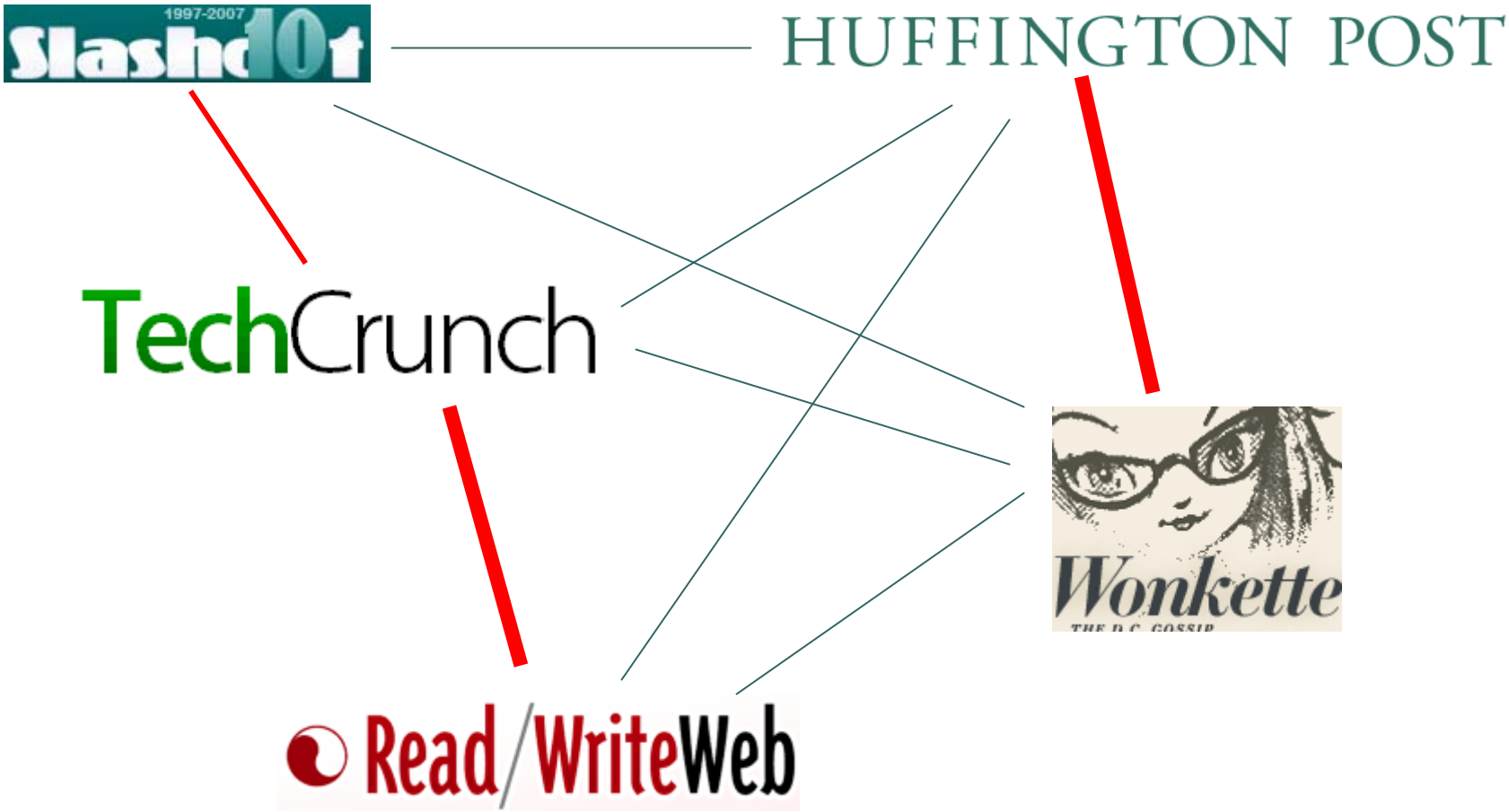
2

Wonkette
Gawker
Gothamist
Huffington Post

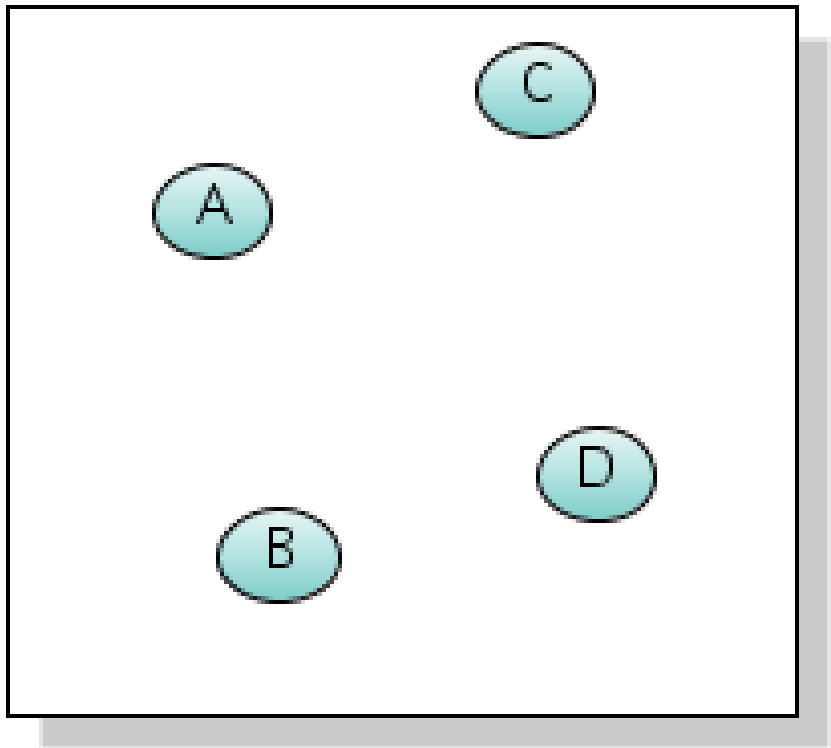
2D Visualizations

- Instead of Clusters, a 2D Map
- Goals
  - Preserve distances as much as possible
  - Draw in two dimensions
- Dimension Reduction
  - Principal Components Analysis
  - Multidimensional Scaling

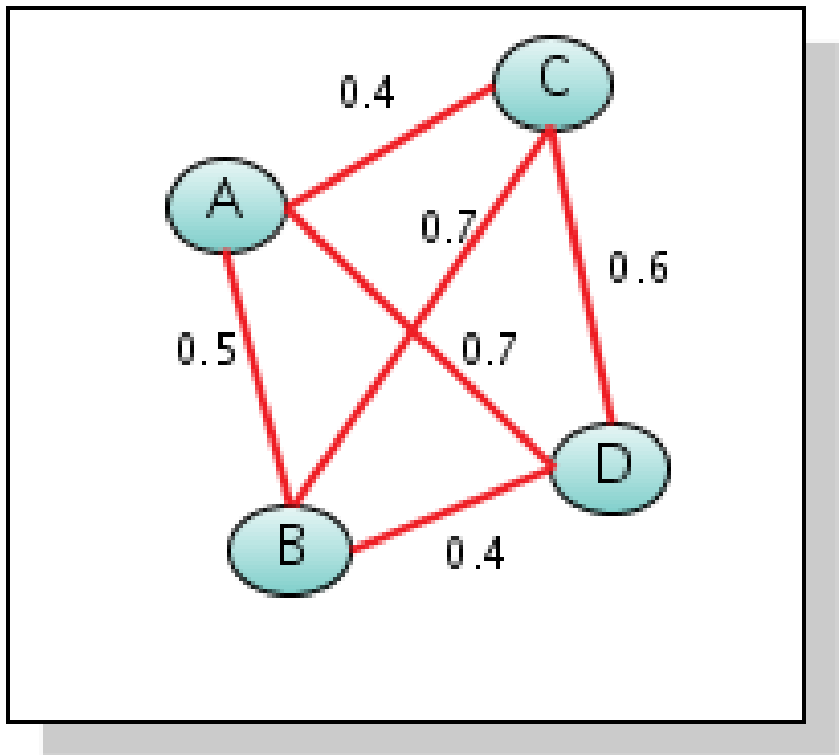
Multidimensional Scaling



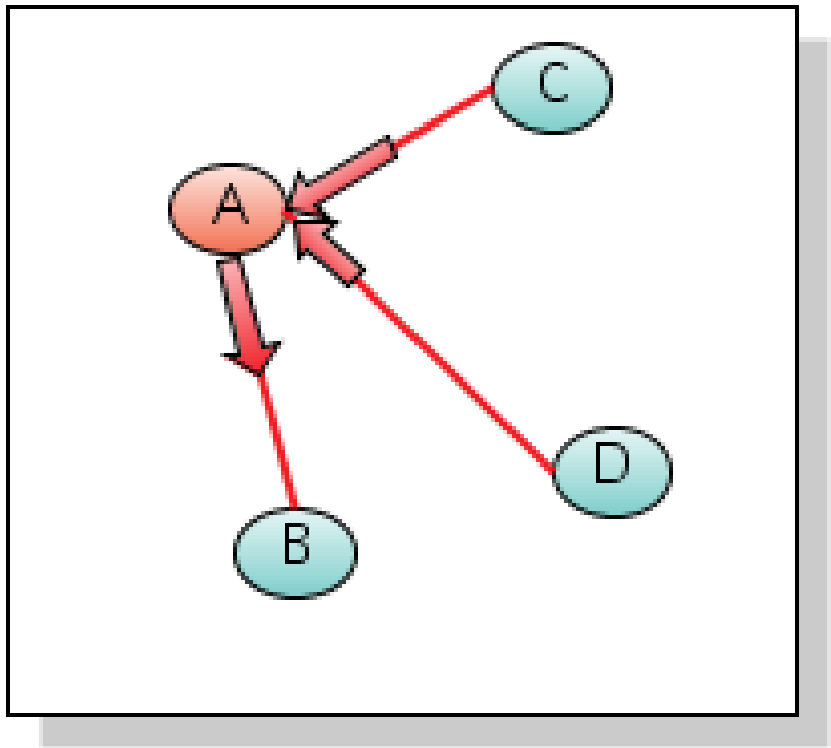
Multidimensional Scaling

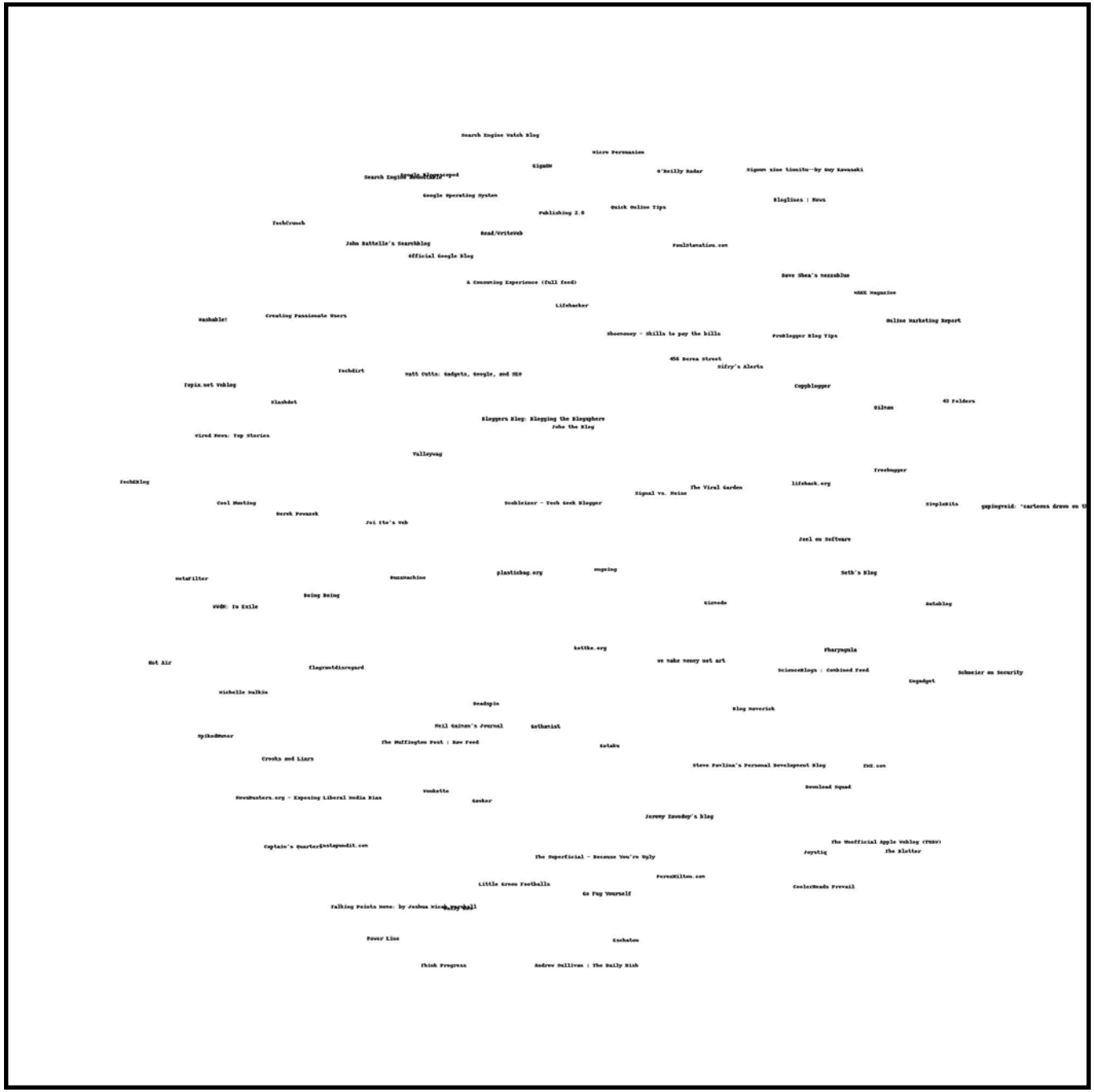


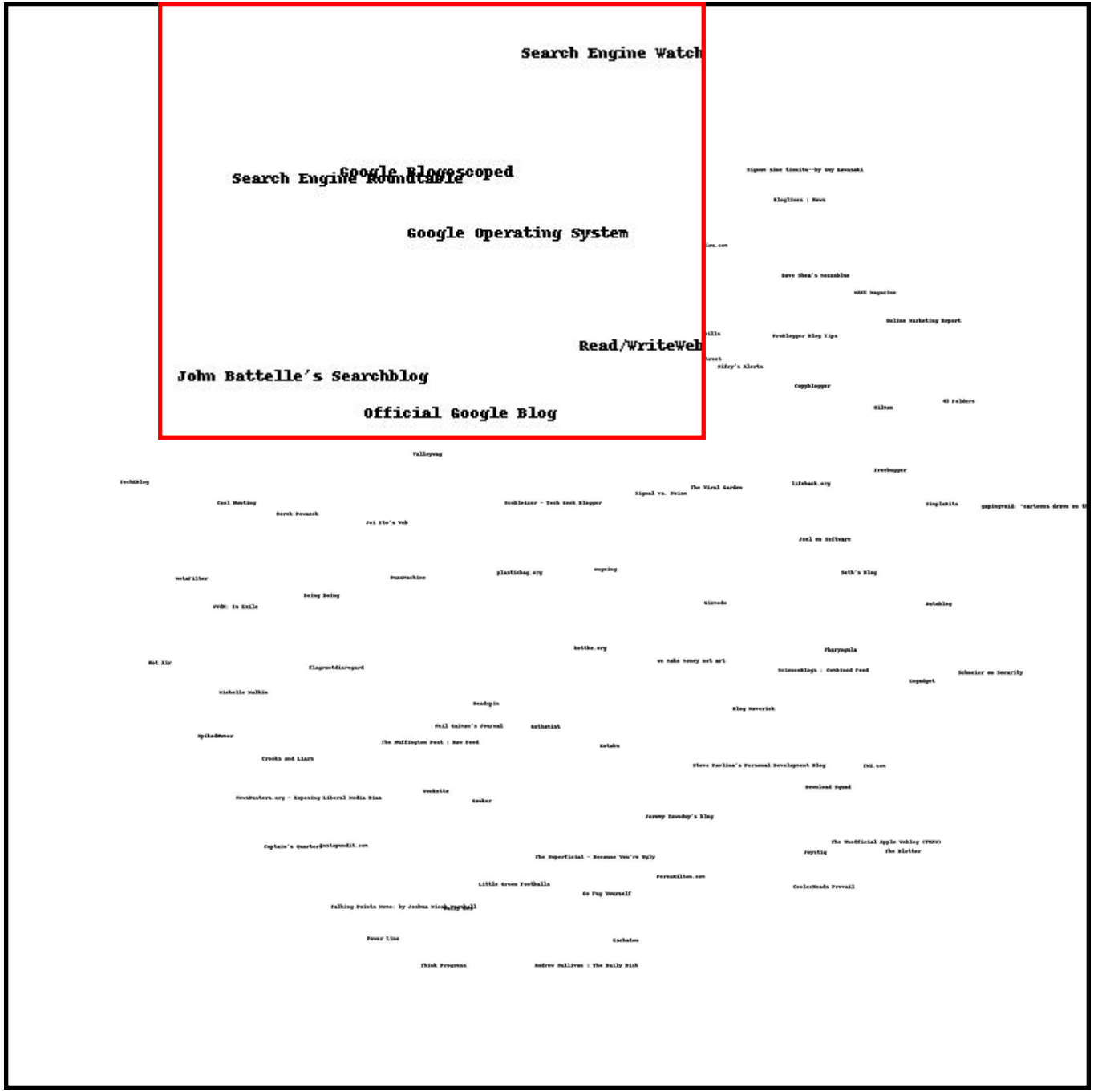
Multidimensional Scaling

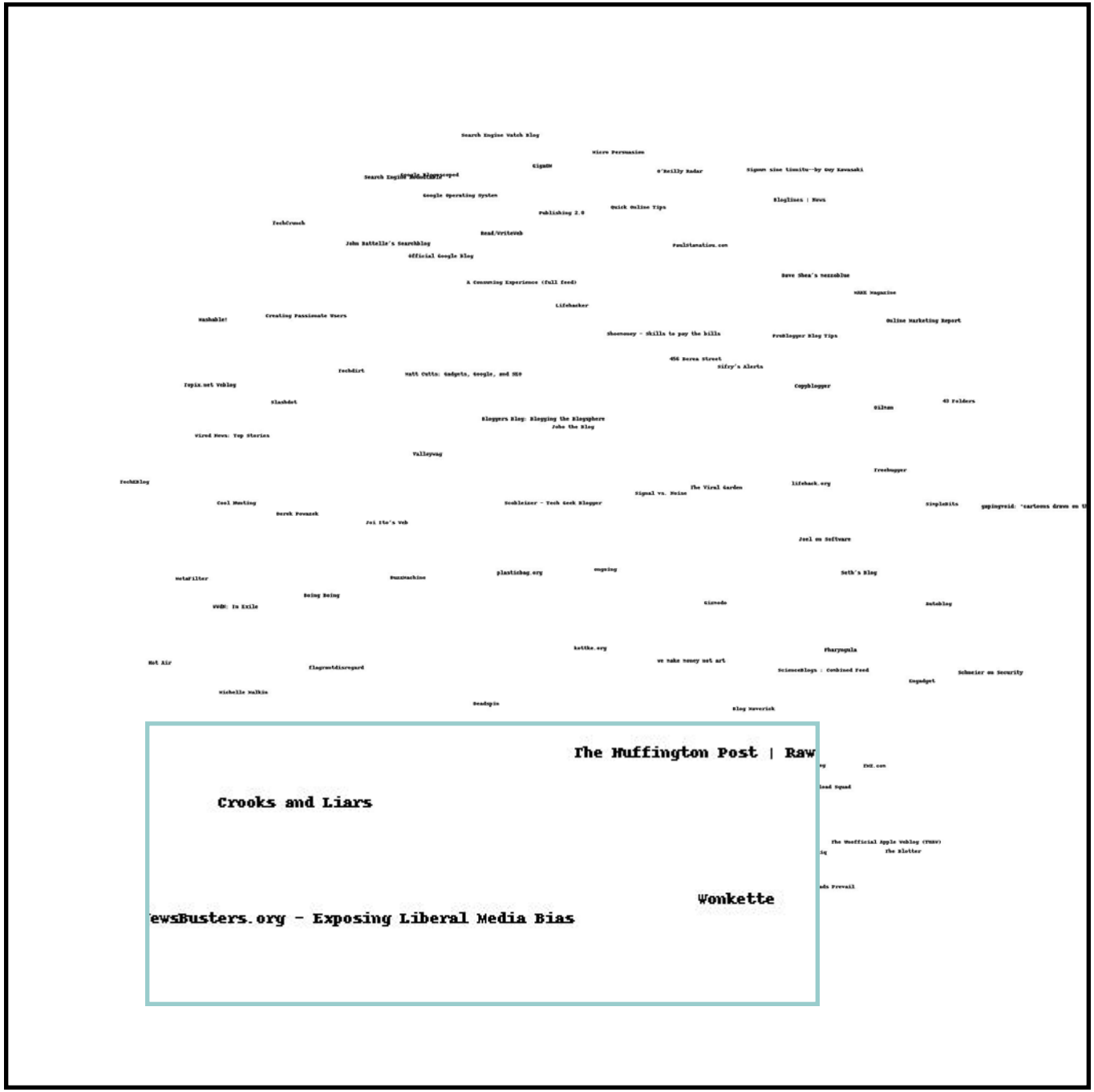


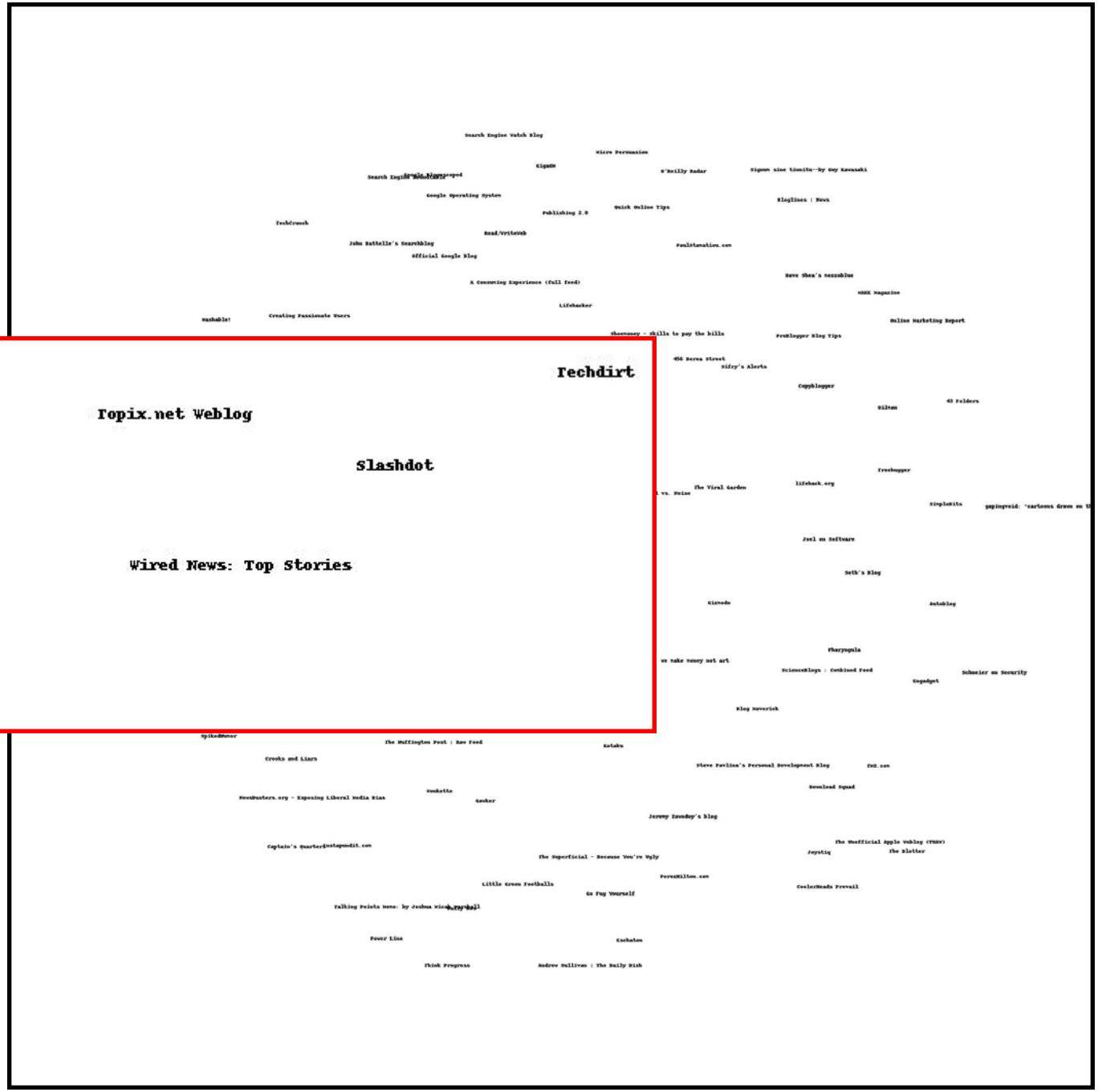
Multidimensional Scaling




Search Engine Watch Blog
Wired Persuasion
GigaOM
O'Reilly Radar
Search Engine Watch Blog
Google Operating System
Quick Online Tips
Publishing 2.0
TechCrunch
Read/WriteWeb
John Battelle's Searchblog
Official Google Blog
A Consuming Experience (full feed)
Lifehacker
Mashable!
Creating Passionate Users
Shoemoney - Skills to pay the bills
Problogger Blog Tips
456 Berea Street
Techdirt
Matt Cutts: Gadgets, Google, and SEO
Sifry's Alerts
Copyblogger
Slashdot
Bloggers Blog: Blogging the Blogsphere
John the Blog
Wired News: Top Stories
Valleywag
TechCrunch
The Viral Garden
lifehack.org
Cool Meeting
Scobleizer - Tech Geek Blogger
Signal vs. Noise
Derek Powazek
Joi Ito's Web
Joel on Software
plasticbag.org
Seth's Blog
BoingBoing
Gizmodo
Autoblog
Hot Air
kottke.org
Pharyngula
We Make Money not Art
flagrantdisregard
ScienceBlogs : Combined Feed
Schneier on Security
Engadget
Michelle Malkin
Deadspin
Blog Maverick
Neil Gaiman's Journal
Gothamist
Instapundit.com
The Huffington Post | Raw Feed
Kotaku
Crooks and Liars
Steve Pavlina's Personal Development Blog
Download Squad
Wonkette
Geeker
Jeremy Zawodny's blog
The Unofficial Apple Weblog (TUAW)
Captain's Quarters
The Superficial - Because You're Ugly
Joystiq
The Sletter
Little Green Footballs
Go Fug Yourself
Power Line
Eschaton
Think Progress
Andrew Sullivan | The Daily Dish


Search Engine Watch
Google Blogoscoped
Search Engine Roundtable
Google Operating System
Read/WriteWeb
John Battelle's Searchblog
Official Google Blog


The Huffington Post | Raw
Crooks and Liars
Wonkette
ewsBusters.org - Exposing Liberal Media Bias


Techdirt
Topix.net Weblog
Slashdot
Wired News: Top Stories
TechEBlog

Zillow

Home Facts

Public Facts		Owner's Facts
Residence:	Multi family	**The owner has not edited home facts or created an estimate.**
Bedrooms:	5	**Are you the owner?**
Bathrooms:	3.5	Edit home facts
Sq ft:	2,474	Learn more
Lot size:	2,819 sq ft / 0.06 acres	Create an estimate
Year built:	1902	Learn more
Year updated:	--	After you're done, your edited home facts will appear here. You can also make your estimate public or keep it private.
# Stories:	3	
# Units:	3	
Total rooms:	11	
Zestimate: $557,447		

Show all home facts

The Zillow API

- Allows querying by address
- Returns information about the property
  - Bedrooms
  - Bathrooms
  - Zip Code
  - Price Estimate
  - Last Sale Price

A home price dataset

House	Zip	Bathrooms	Bedrooms	Built	Type	Price
A	02138	1.5	2	1847	Single	505296
B	02139	3.5	9	1916	Triplex	776378
C	02140	3.5	4	1894	Duplex	595027
D	02139	2.5	4	1854	Duplex	552213
E	02138	3.5	5	1909	Duplex	947528
F	02138	3.5	4	1930	Single	2107871
etc..						

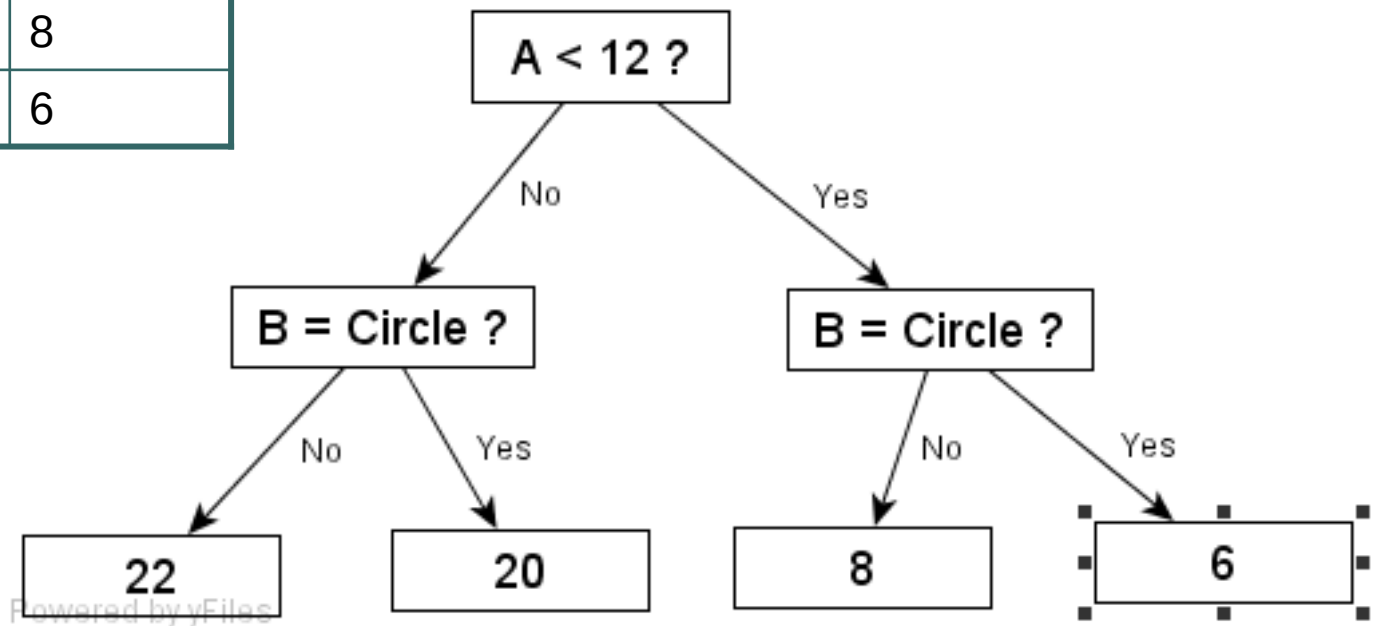
What can we learn?

- A made-up houses price
- How important is Zip Code?
- What are the important attributes?

- Can we do better than averages?

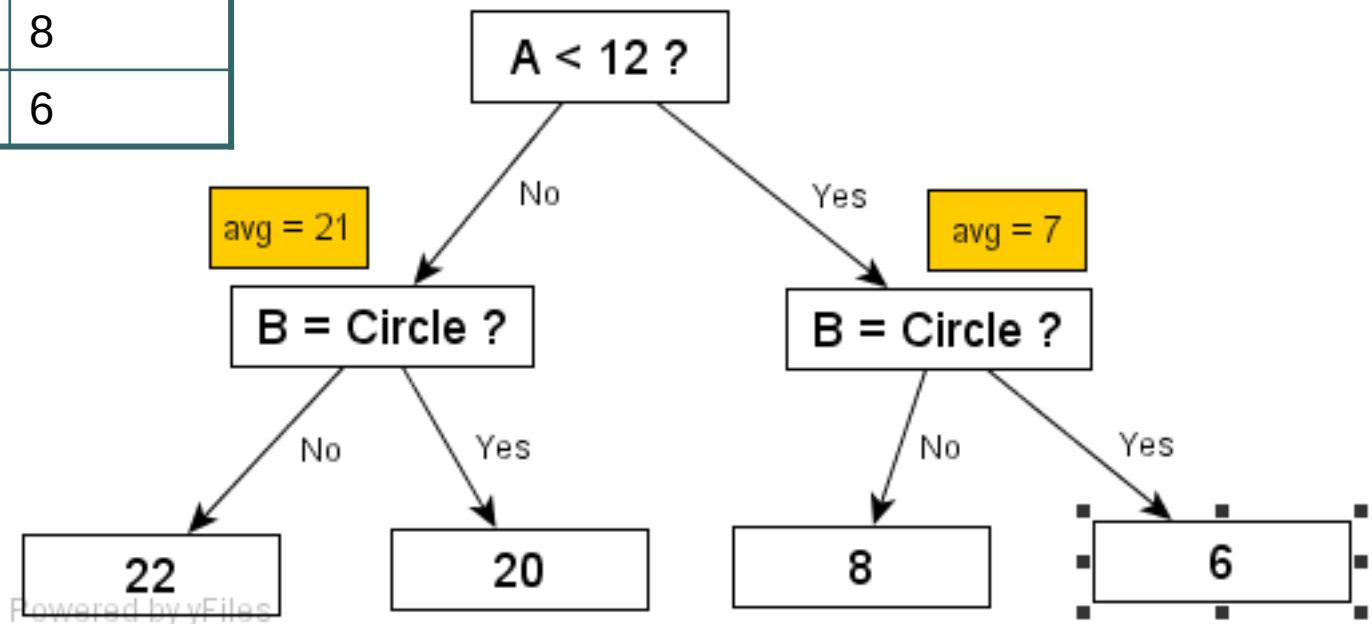
Introducing Regression Trees

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6



Introducing Regression Trees

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6



Minimizing deviation

- Standard deviation is the “spread” of results
- Try all possible divisions
- Choose the division that decreases deviation the most

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6

Initially

Average = 14

Standard Deviation = 8.2

Minimizing deviation

- Standard deviation is the “spread” of results
- Try all possible divisions
- Choose the division that decreases deviation the most

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6

B = Circle

Average = 13

Standard Deviation = 9.9

B = Square

Average = 15

Standard Deviation = 9.9

Minimizing deviation

- Standard deviation is the “spread” of results
- Try all possible divisions
- Choose the division that decreases deviation the most

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6

A > 18

Average = 8

Standard Deviation = 0

A <= 20

Average = 16

Standard Deviation = 8.7

Minimizing deviation

- Standard deviation is the “spread” of results
- Try all possible divisions
- Choose the division that decreases deviation the most

A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6

A > 11

Average = 7

Standard Deviation = 1.4

A <= 11

Average = 21

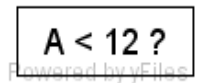
Standard Deviation = 1.4

CART Algoritm

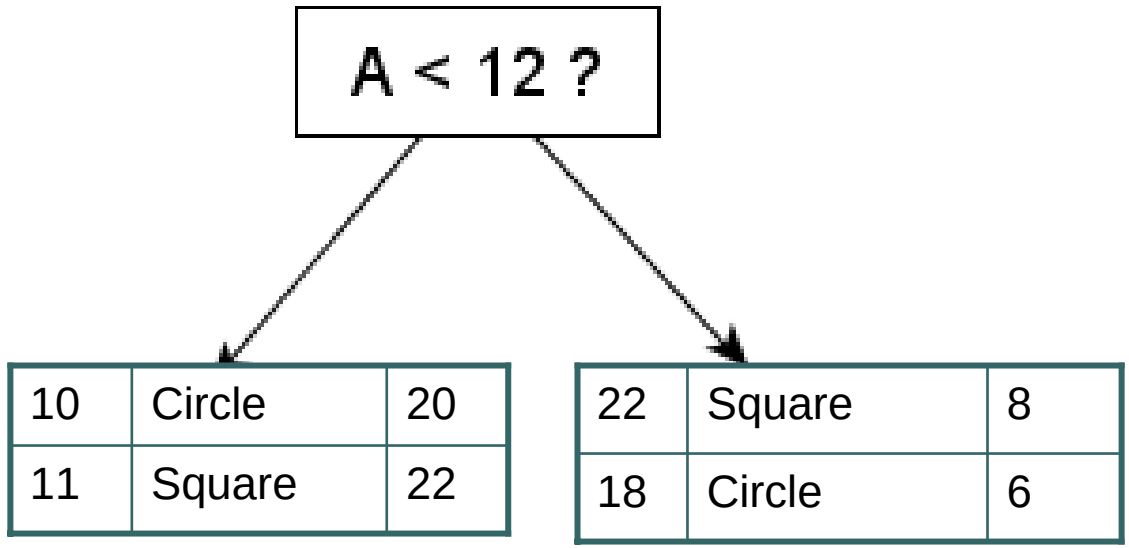
A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6

CART Algoritm

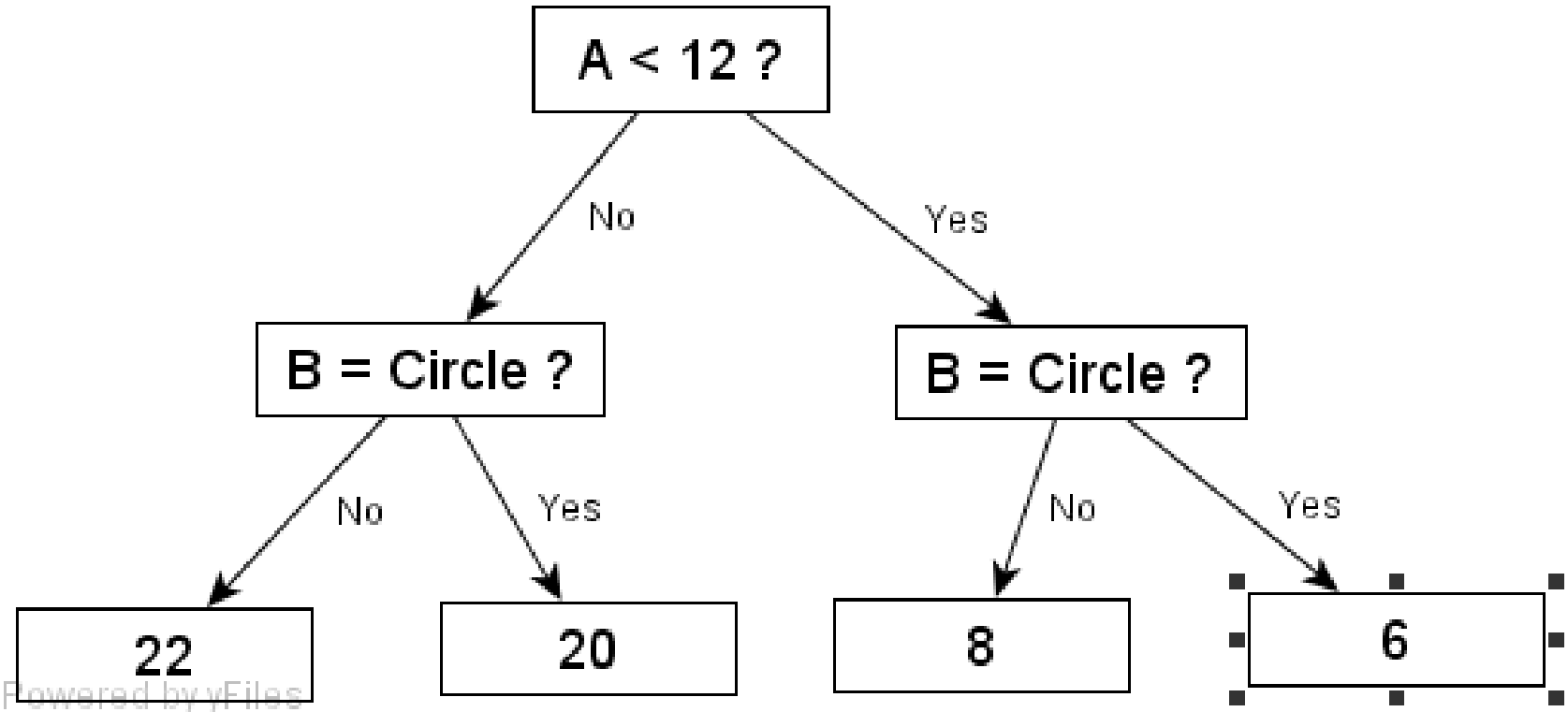
A	B	Value
10	Circle	20
11	Square	22
22	Square	8
18	Circle	6



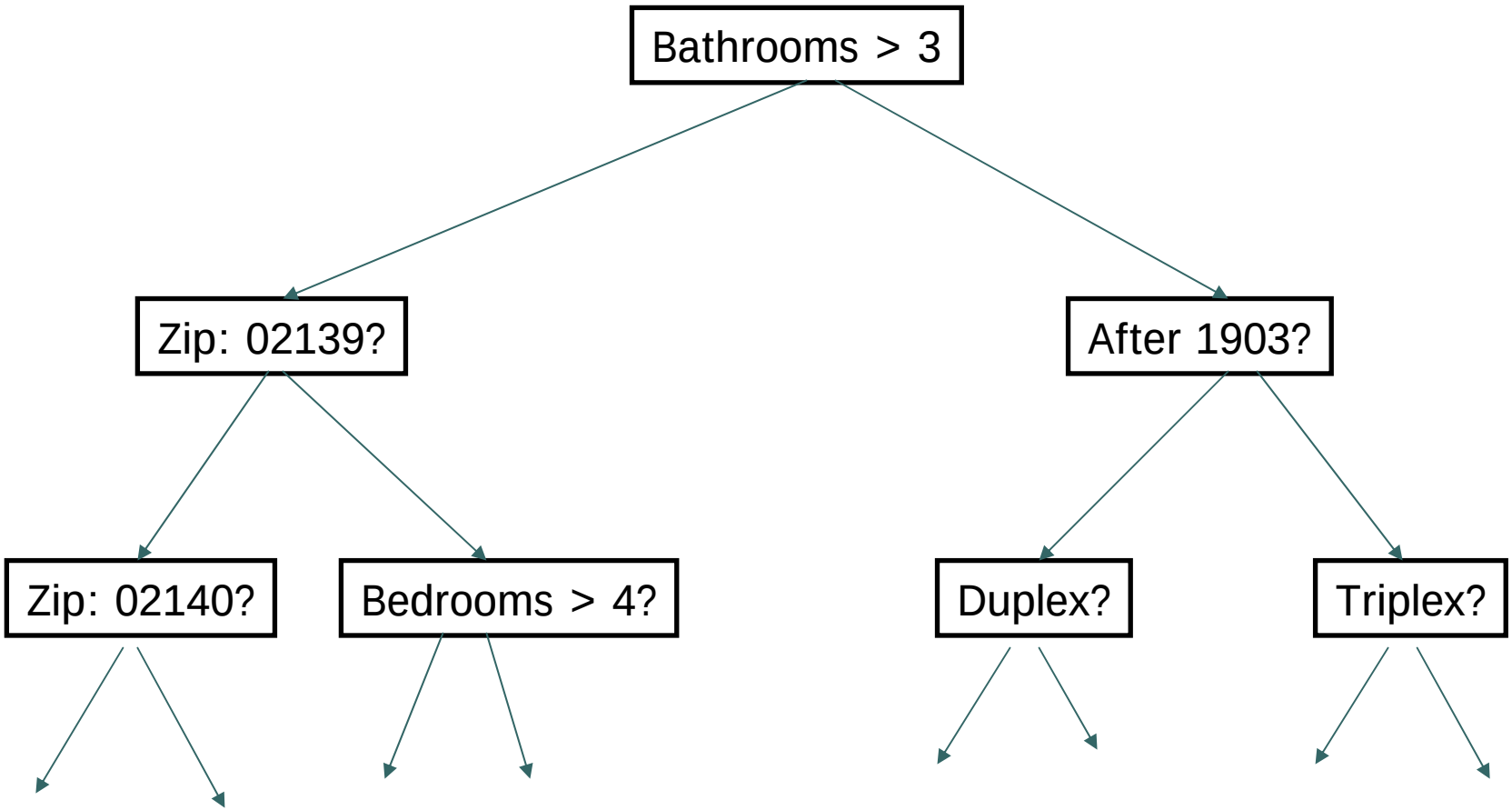
CART Algoritm



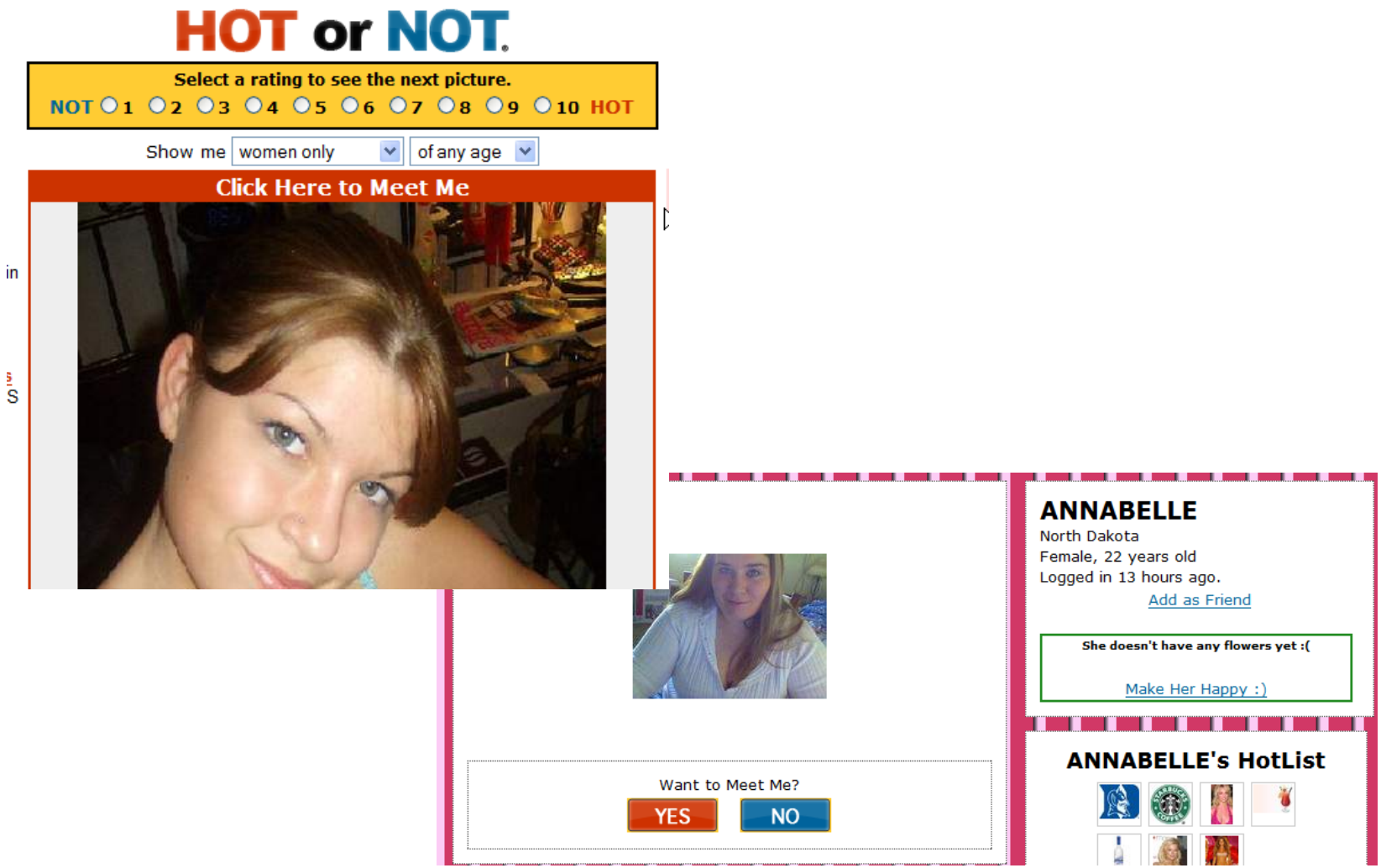
CART Algoritm



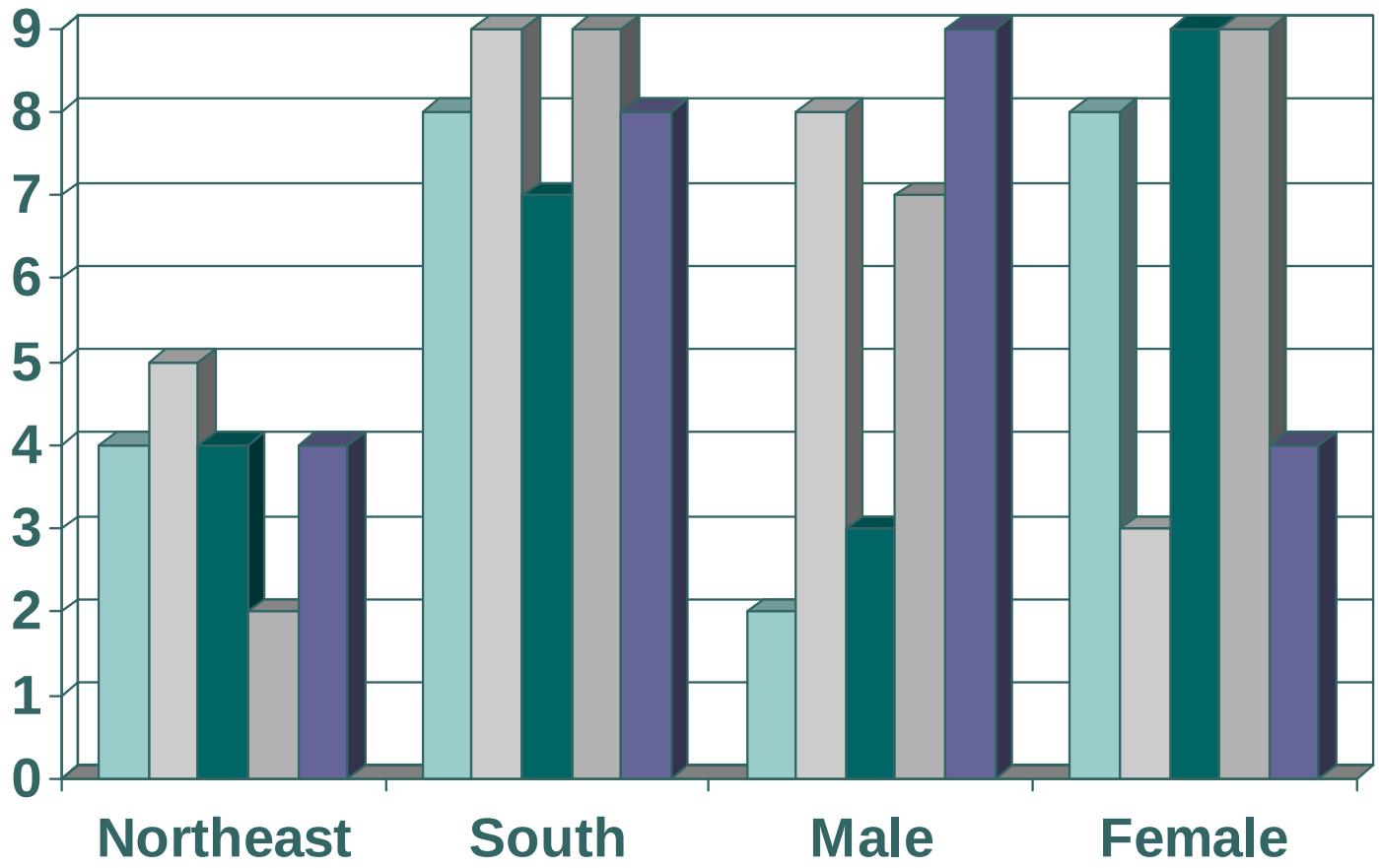
Zillow Results



Just for Fun… Hot or Not



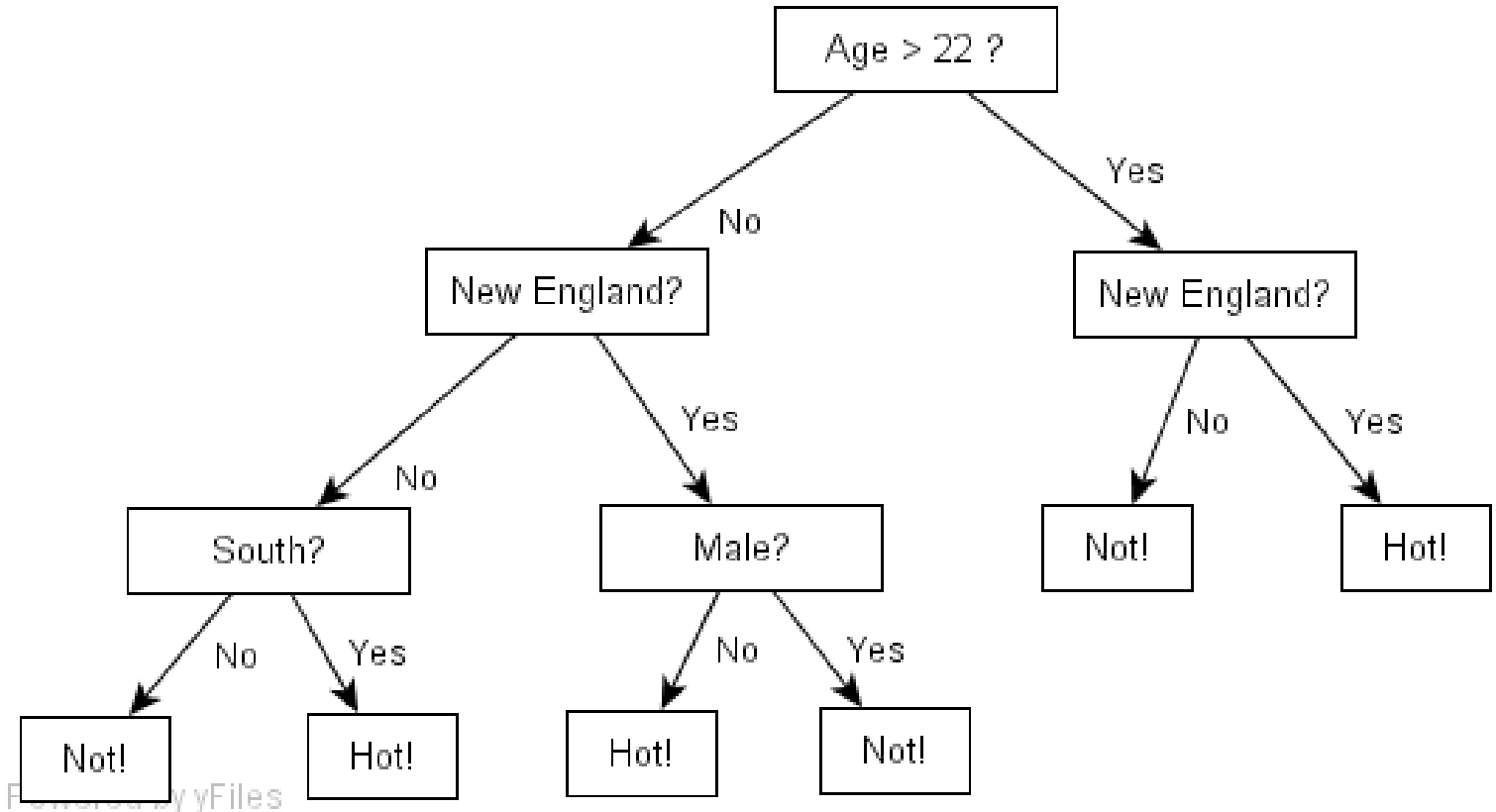
Variance dividers



Low Variance Split

High Variance Split

Just for Fun… Hot or Not



Supervised and Unsupervised

- Clustering methods are *unsupervised*
  - There are no answers
  - Methods just characterize the data
  - Show interesting patterns
- Regression Trees are *supervised*
  - “answers” are in the dataset
  - Tree models predict answers

Personal Ads

The Analysis

Five Cities

W4M Personal Ads

Bayesian filter

If you listen to **NPR**, watch **Hardball**, and love the **Red Sox**, you may be the guy for me.

Please email me back.

I'm a **professional** with a **grad school degree** who has a sense of **humor** and loves the **Sox**.

Boston

Sox	0.4
Red	0.35
Grad	0.2
Professional	0.1
Humor	0.1

Bayesian filter

$$P(C \mid W) = P (C \& W) / P (W)$$

How often do the word and the city appear together?

How often does the word appear overall…

Rank these, and you have a list of the words most particular to a given city

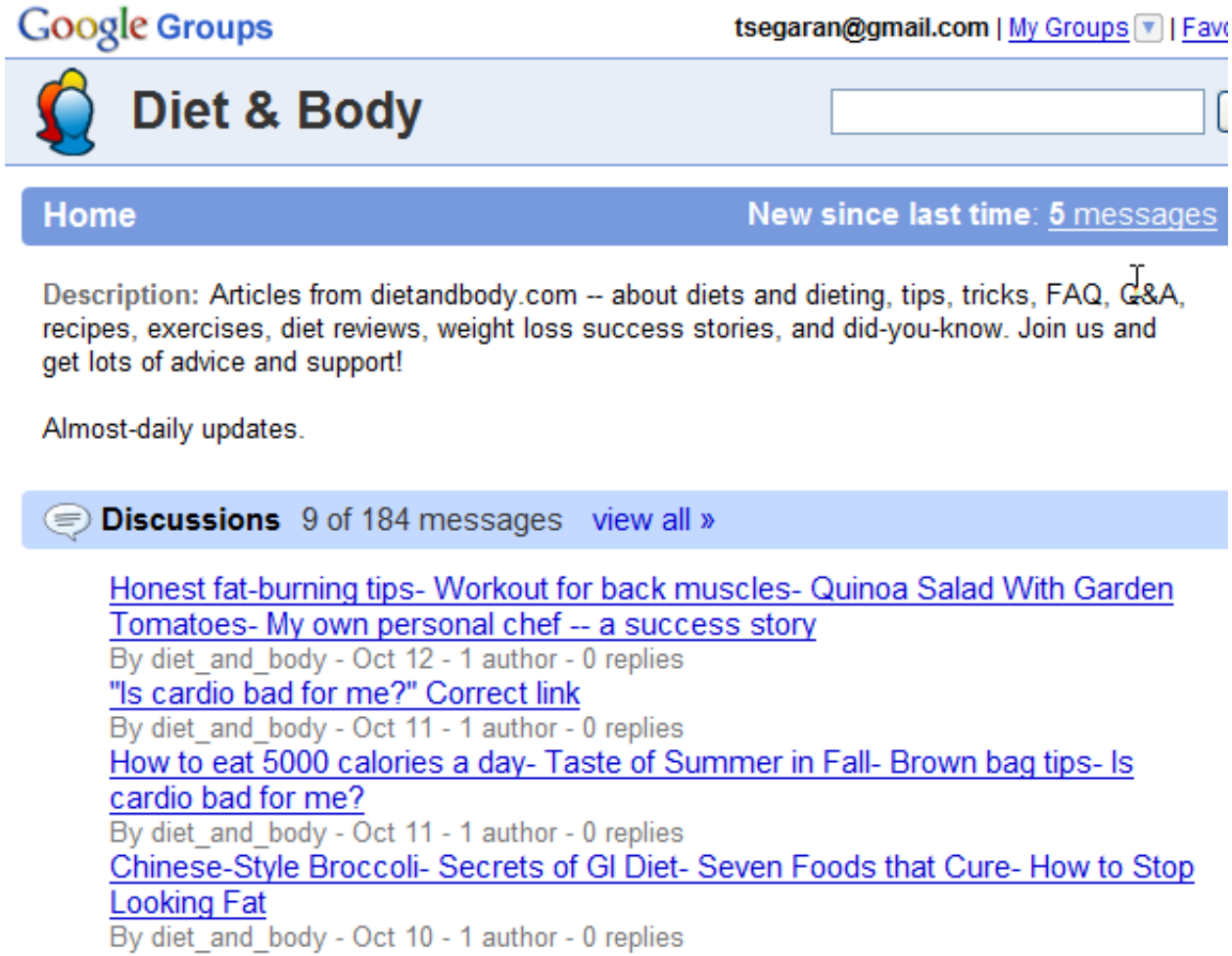
Results

New York	Boston	Chicago
Mets	Pink	**Cubs**
Lounges	**Sox**	Burbs
Offense	Poetry	**Bears**
Desires	Intellectually	Girlie
Musical	Punk	Insecure
Submissive	Appreciation	Cheat
Create	Exercise	Importance
Song	**Winter**	**Blunt**
Oral	**Education**	Mouth

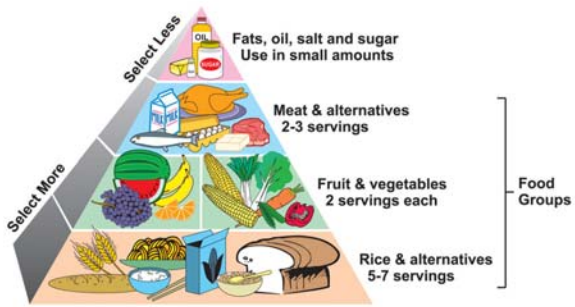
Results

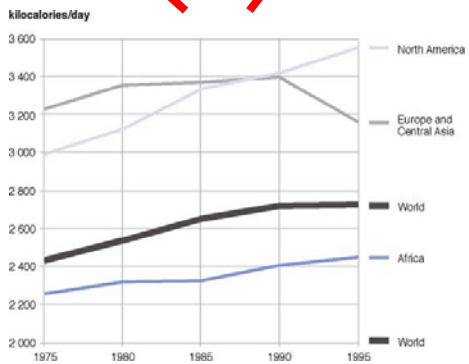
Los Angeles	San Francisco
Excellent	Tee
Vegas	Employment
Meaningful	Picnic
Star	STD
Lame	Tasting
Industry	Hikes
Heat	French
Fitness	.com
Entertainment	Kayaking
Latino	Cycling

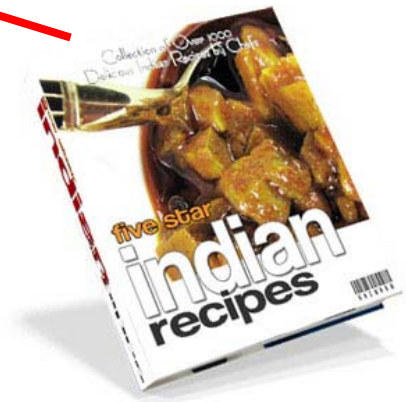
Newsgroup Discussion



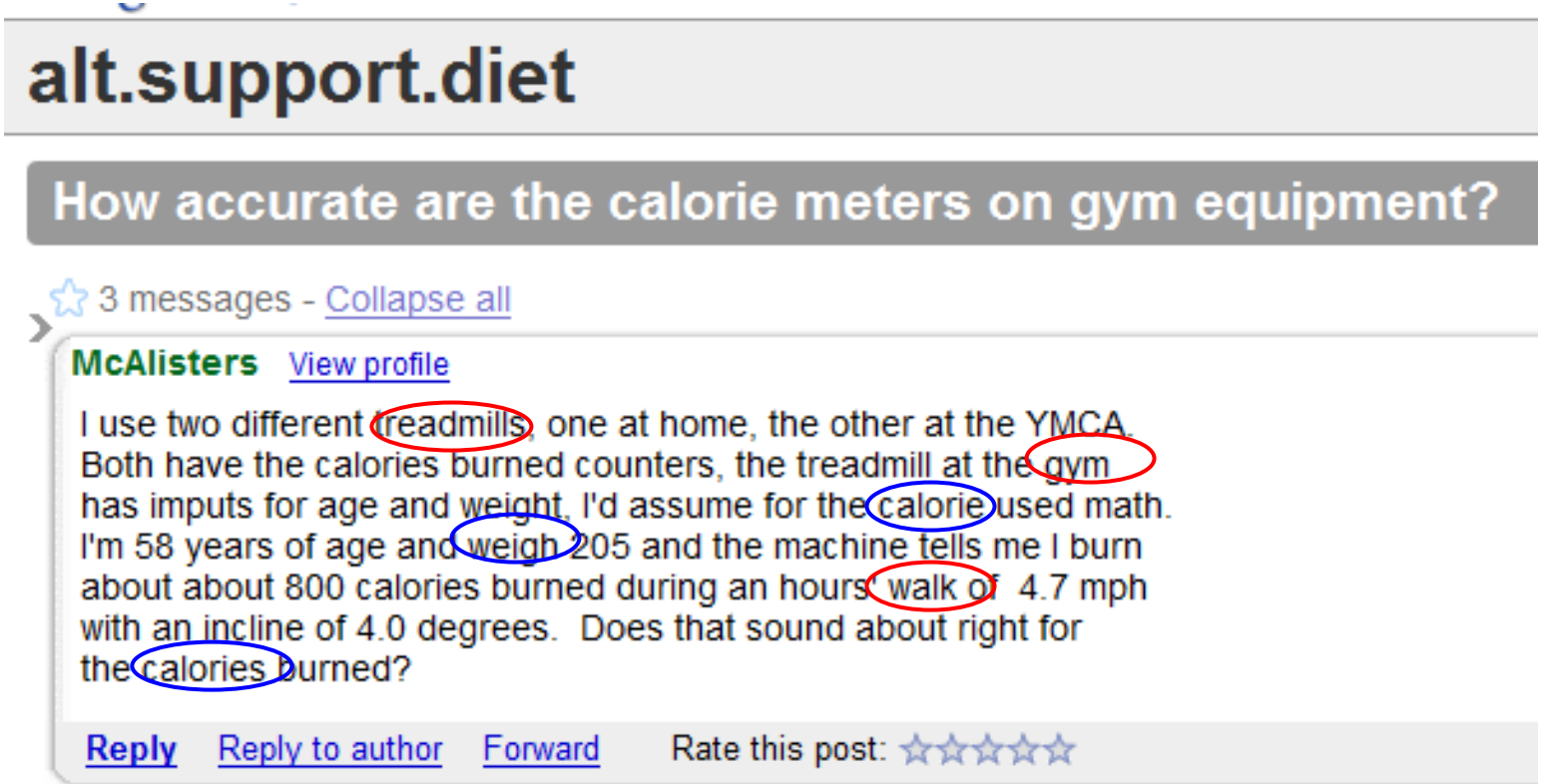
Overlapping themes







Themes in a document



alt.support.diet

How accurate are the calorie meters on gym equipment?

3 messages - Collapse all

McAlisters View profile

I use two different treadmills, one at home, the other at the YMCA. Both have the calories burned counters, the treadmill at the gym has imputs for age and weight, I'd assume for the calorie used math. I'm 58 years of age and weigh 205 and the machine tells me I burn about about 800 calories burned during an hours' walk of 4.7 mph with an incline of 4.0 degrees. Does that sound about right for the calories burned?

Reply Reply to author Forward Rate this post: ☆☆☆☆☆

Another word matrix

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	2	0	0	3	0
Calorie	0	2	1	1	3
Weigh	1	0	2	0	0
Carbs	0	3	0	0	2
Treadmill	1	0	0	2	0

Actual Matrix

Weights and features

	F1	F2	F3
Gym	0	1	2
Calorie	2	0	1
Weigh	2	2	1
Carbs	1	0	3
Treadmill	0	1	2

Features Matrix

X

	Msg1	M2	M3	M4	M5
F1	1	0	2	3	0
F2	0	2	1	1	3
F3	1	0	2	0	0

Weight Matrix

Matrix factorization

	F1	F2	F3
Gym	0	1	2
Calorie	2	0	1
Weigh	2	2	1
Carbs	1	0	3
Treadmill	0	1	2

Features Matrix

X

	Msg1	M2	M3	M4	M5
F1	1	0	2	3	0
F2	0	2	1	1	3
F3	1	0	2	0	0

Weight Matrix

=

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	1	3	3	0	1
Calorie	0	2	4	1	3
Weigh	2	3	1	0	1
Carbs	0	1	1	0	2
Treadmill	3	2	0	2	2

Current Guess

Matrix factorization

	F1	F2	F3
Gym	0	1	2
Calorie	2	0	1
Weigh	2	2	1
Carbs	1	0	3
Treadmill	0	1	2

Features Matrix

X

	Msg1	M2	M3	M4	M5
F1	1	0	2	3	0
F2	0	2	1	1	3
F3	1	0	2	0	0

Weight Matrix

=

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	1	3	3	0	1
Calorie	0	2	4	1	3
Weigh	2	3	1	0	1
Carbs	0	1	1	0	2
Treadmill	3	2	0	2	2

Current Guess

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	2	0	0	3	0
Calorie	0	2	1	1	3
Weigh	1	0	2	0	0
Carbs	0	3	0	0	2
Treadmill	1	0	0	2	0

Target Result

Matrix factorization

	F1	F2	F3
Gym	1	0	0
Calorie	0	1	1
Weigh	0	0	2
Carbs	0	1	0
Treadmill	1	0	0

Features Matrix

X

	Msg1	M2	M3	M4	M5
F1	2	0	0	1	0
F2	0	2	0	1	3
F3	1	0	1	0	0

Weight Matrix

=

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	2	0	0	3	0
Calorie	0	2	1	1	3
Weigh	1	0	2	0	0
Carbs	0	3	0	0	2
Treadmill	1	0	0	2	0

Current Guess

	Msg1	Msg2	Msg3	Msg4	Msg5
Gym	2	0	0	3	0
Calorie	0	2	1	1	3
Weigh	1	0	2	0	0
Carbs	0	3	0	0	2
Treadmill	1	0	0	2	0

Target Result

Interpreting Features

	F1	F2	F3
Gym	1	0	0
Calorie	0	1	1
Weigh	0	0	2
Carbs	0	1	0
Treadmill	1	0	0

Features Matrix

Theme 1	**Theme 2**	**Theme 3**
Gym	Calorie	Weigh
Treadmill	Carbs	Calorie

	Msg1	M2	M3	M4	M5
F1	2	0	0	1	0
F2	0	2	0	1	3
F3	1	0	1	0	0

Weight Matrix

Msg1	**Msg2**	**Msg3 etc.**
Theme 1	Theme 2	Theme 3
Theme 3		

“Diet and body” themes

Atkins
Induction
South
Beach
Carbs

Chocolate
Black
Coffee
Olive
Broccoli

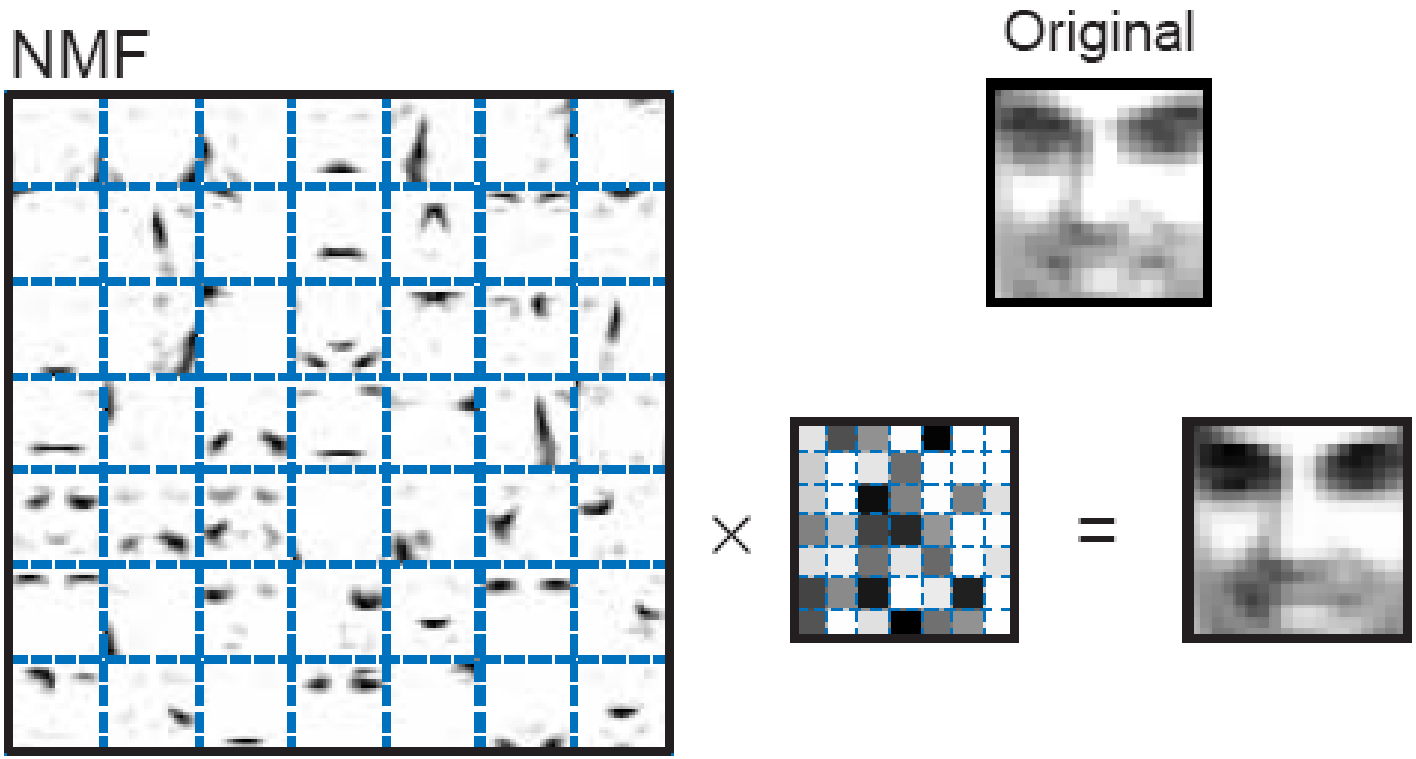
Calories
Weight
Fats
Protein
Cholesterol

Gym
Weights
Exercise
Running
Injured

Cook
Recipe
Fried
Home

Money
Organic
Want
Best

Side note: NMF for faces



Methods covered

- Regression trees
- Hierarchical clustering
- k-means clustering
- Multidimensional scaling
- Bayesian Classifier
- Non-negative Matrix Factorization

Other ideas

- Finance
  - Analysts already drowning in info
  - Stories sometimes broken on blogs
  - Message boards show sentiment

  - Extremely low signal-to-noise ratio

Other ideas

- Product problems/ideas
  - Use support message boards
  - Extract themes
  - Understand recurring issues
  - Learn what features people want

Other ideas

- Entertainment
  - How much buzz is a movie generating?
  - What psychographic profiles like this type of movie?
  - Of interest to studios and media investors