Good and evil data

Lech Rzedzicki
Lead Technologist, InnovateUK

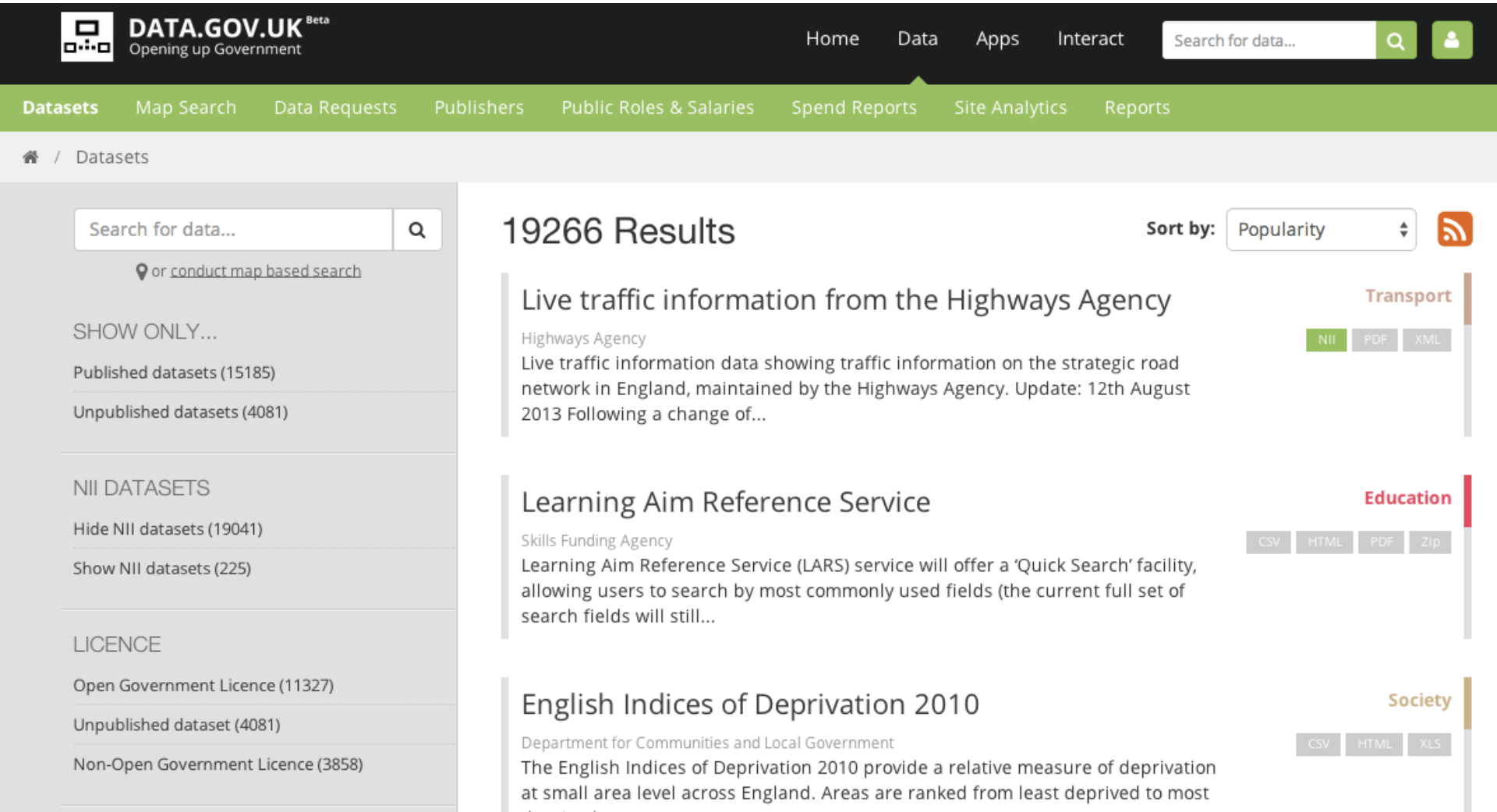
THE GOOD

Hans Rosling

IBM Watson

PIENSE
THINK
σκεψου
DENKE
PENSER
$300,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$1,000,000
Who is Bram Stoker?
$ 17,973
$200,000
WHO IS BRAM STOKER?
$5600

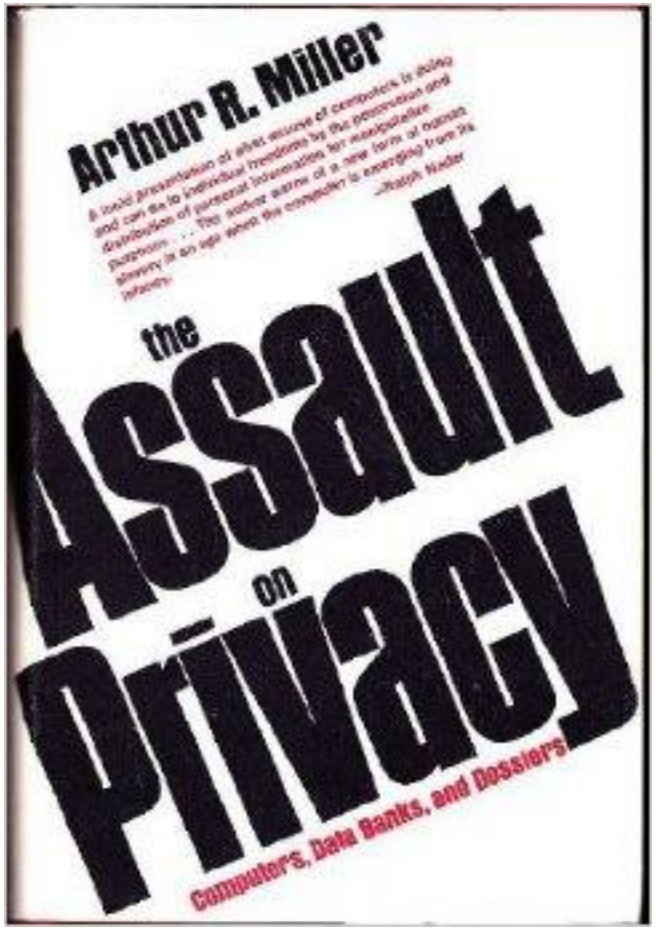
Open Data


DATA.GOV.UK Beta
Opening up Government
Home
Data
Apps
Interact
Search for data...
Datasets
Map Search
Data Requests
Publishers
Public Roles & Salaries
Spend Reports
Site Analytics
Reports
/ Datasets
Search for data...
or conduct map based search
SHOW ONLY...
Published datasets (15185)
Unpublished datasets (4081)
NII DATASETS
Hide NII datasets (19041)
Show NII datasets (225)
LICENCE
Open Government Licence (11327)
Unpublished dataset (4081)
Non-Open Government Licence (3858)
19266 Results
Sort by:
Popularity
Live traffic information from the Highways Agency
Transport
Highways Agency
NII
PDF
XML
Live traffic information data showing traffic information on the strategic road network in England, maintained by the Highways Agency. Update: 12th August 2013 Following a change of...
Learning Aim Reference Service
Education
Skills Funding Agency
CSV
HTML
PDF
Zip
Learning Aim Reference Service (LARS) service will offer a 'Quick Search' facility, allowing users to search by most commonly used fields (the current full set of search fields will still...
English Indices of Deprivation 2010
Society
Department for Communities and Local Government
CSV
HTML
XLS
The English Indices of Deprivation 2010 provide a relative measure of deprivation at small area level across England. Areas are ranked from least deprived to most

Chromogram 1.2b
Choose file: Pearle
View: Timeline Block
Color by: Title
Search:
May 28, 2005
0 50 100 150 200 250 300 350 400 450
28 28 28 28 29 31 Jun 1 1 2 2 2 2 2 3 21 21 22 23 23 24 Jul 1 3 3 3 4 5 6 Aug 20 20 22 23 23 24 24 24 25 25 25 25 26 26 26 28 Sep 4 5 11 15 15
wpdms
uss
his
history
hist...
poli...
rus
battl...
davi...
geor...
james
list
robe...
the
the
wil
env
universiti...
museum...
buil...
bui
Total 106776
a 131
battle 176
big 108
black 113
buildings 100
community 324
david 156
east 382
fort 182
george 125
grand 110
green 89
history 371
indian 103
james 129
john 332
la 225
lake 387
list 1123
little 101
michael 92
mount 233
music 129
national 286
new 578
north 299
operation 105
pages 344
paul 97
political 168
robert 119
south 98
st 94
st. 132
the 1298
towns 89
u.s. 154
united 194
uss 197
west 114
wikifun 101
wikipedia 123
wikipedians 90
william 156

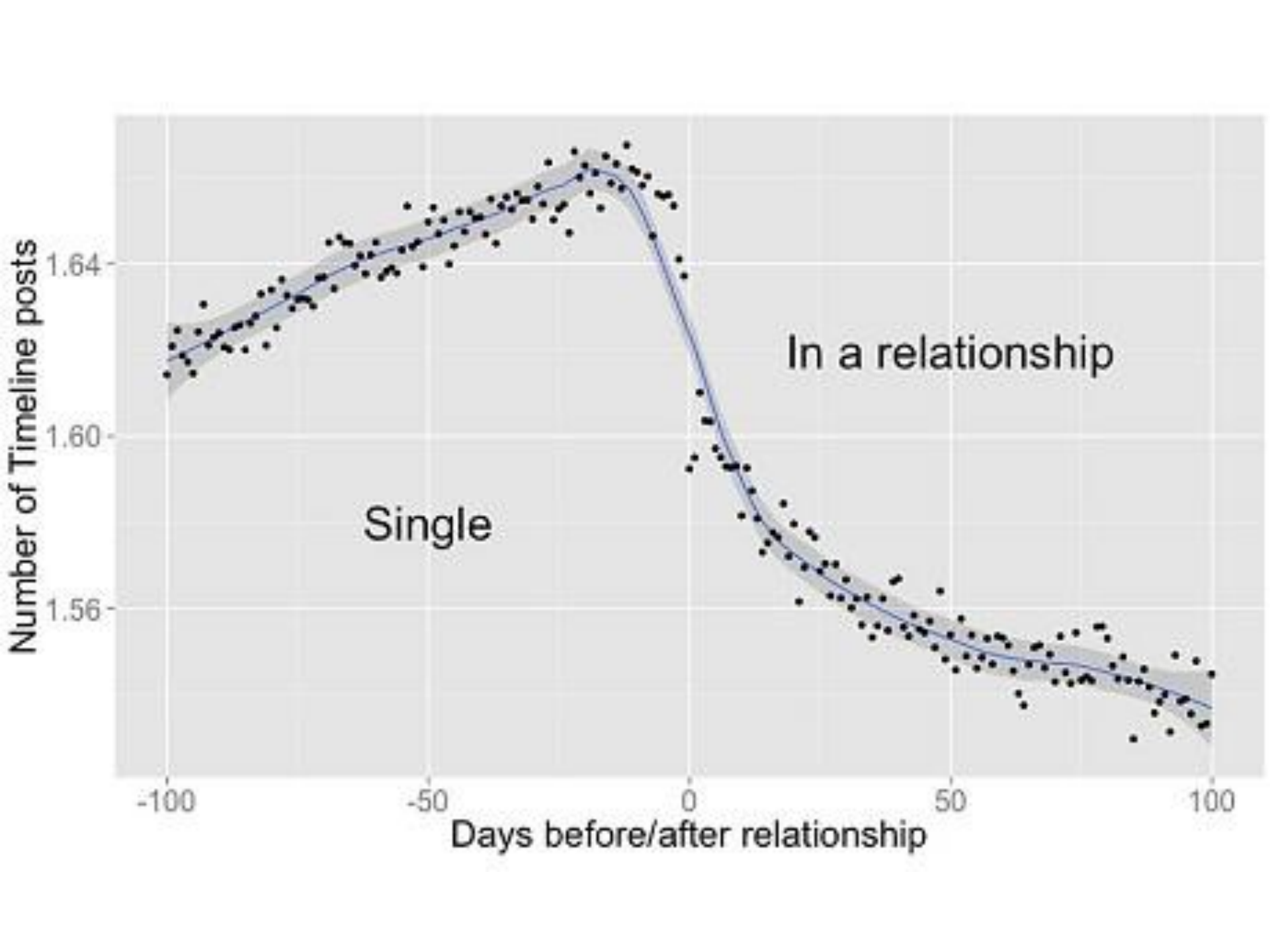
THE BAD

1972


Arthur R. Miller
the
Assault
on
Privacy
Computers, Data Banks, and Dossiers

“In Korea, Email Is Only For Old People”

“My Dear. Greetings, I know this mail may come to you as a surprise, I am Barrister Mussa Issah the solicitor/counsel to the late Sanni Abacha who was then before his death, the President/Head of State of the Federal Republic of Nigeria.”


Number of Timeline posts
1.64
1.60
1.56
Single
In a relationship
-100
-50
0
50
100
Days before/after relationship

$$PV_{Fixed} = \sum_{i=1}^{t} N * \frac{\left(\frac{CDS}{4}\right)}{10{,}000} * (1 - \sum_{i=0}^{i} pd_i) * df_i + \frac{N * \frac{\left(\frac{CDS}{4}\right)}{10{,}000} * pd_i * df_i}{2}$$

Where,

PV = Present Value

N = Notional Amount of the swap

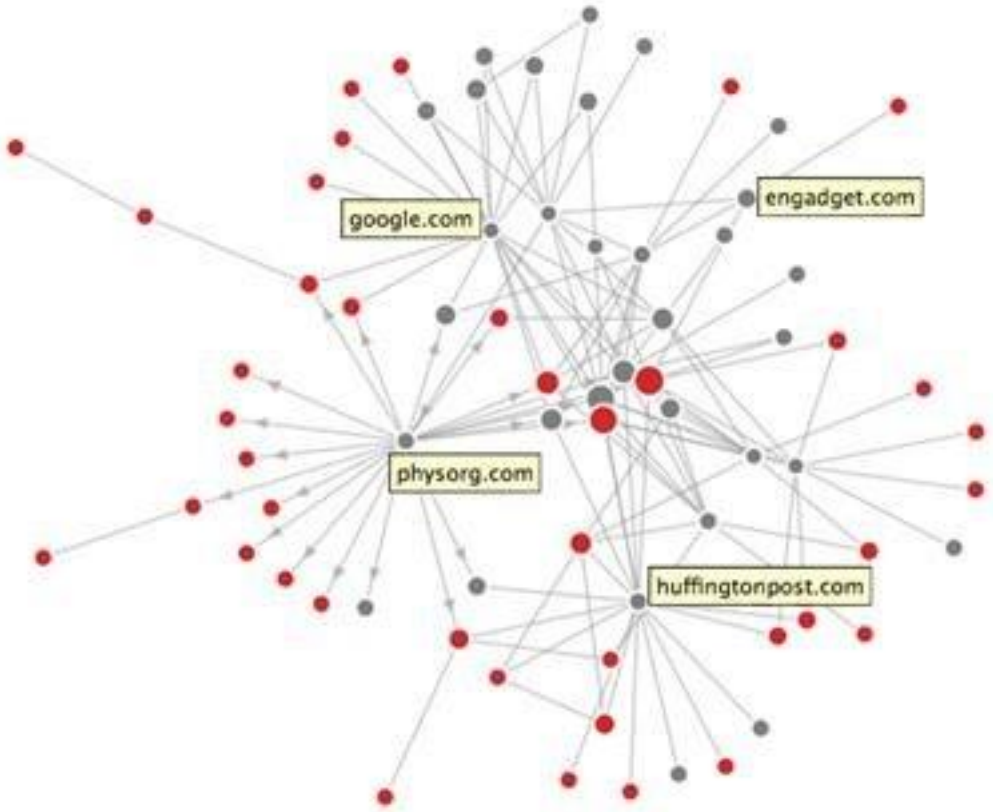
CDS = Credit Default Spread (in annual basis point terms)

pd = probability of default

df = discount factor for time i

(1 - $\sum_{i=0}^{i} pd_i$)= the probability the swap has survived up to the current quarter

THE UGLY


google.com
engadget.com
physorg.com
huffingtonpost.com

(NSA) Metadata

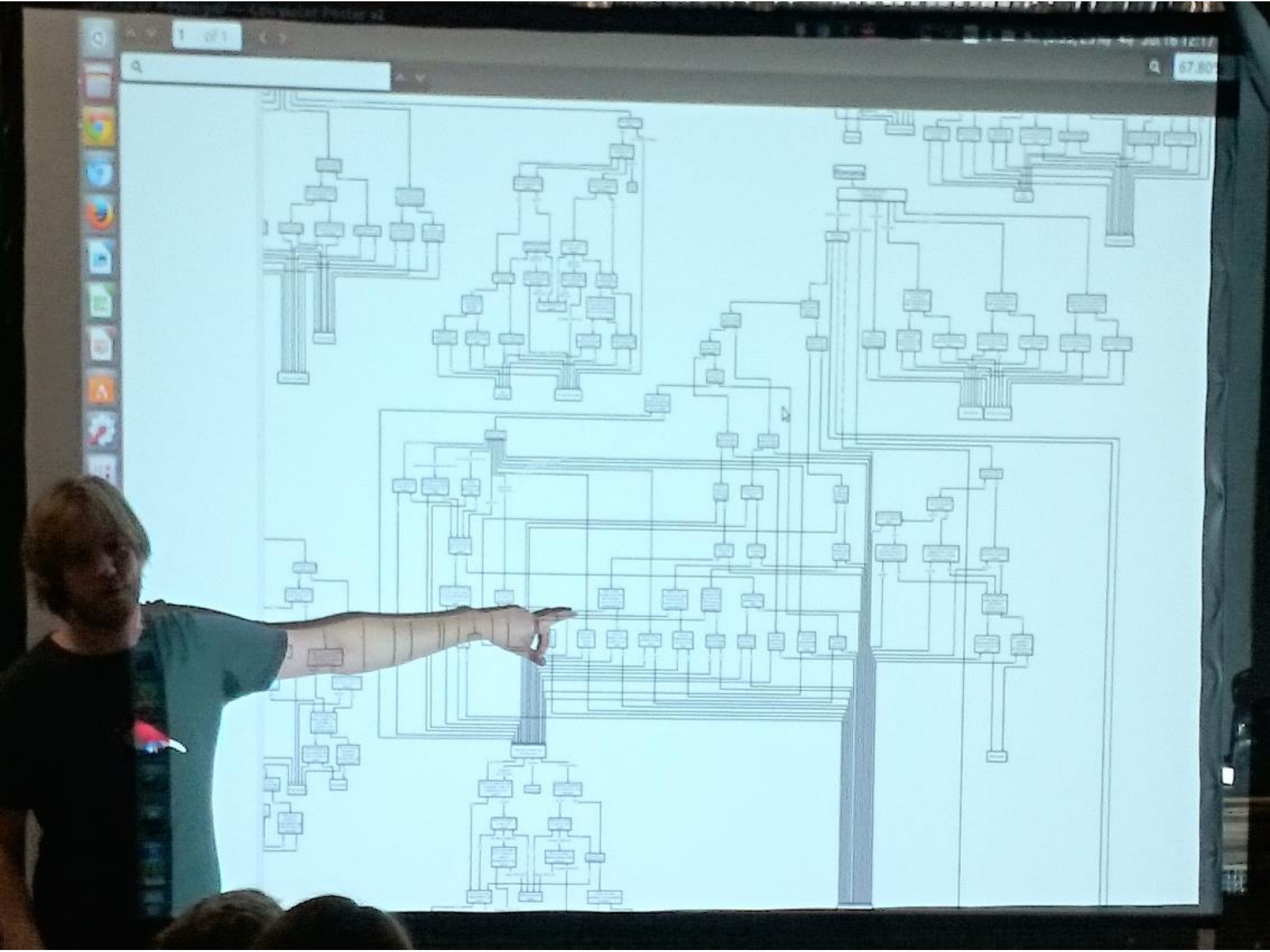
They know you rang a phone sex service at 2:24 am and spoke for 18 minutes. But they don't know what you talked about.

They know you called the suicide prevention hotline from the Golden Gate Bridge. But the topic of the call remains a secret.

They know you spoke with an HIV testing service, then your doctor, then your health insurance company in the same hour. But they don't know what was discussed.

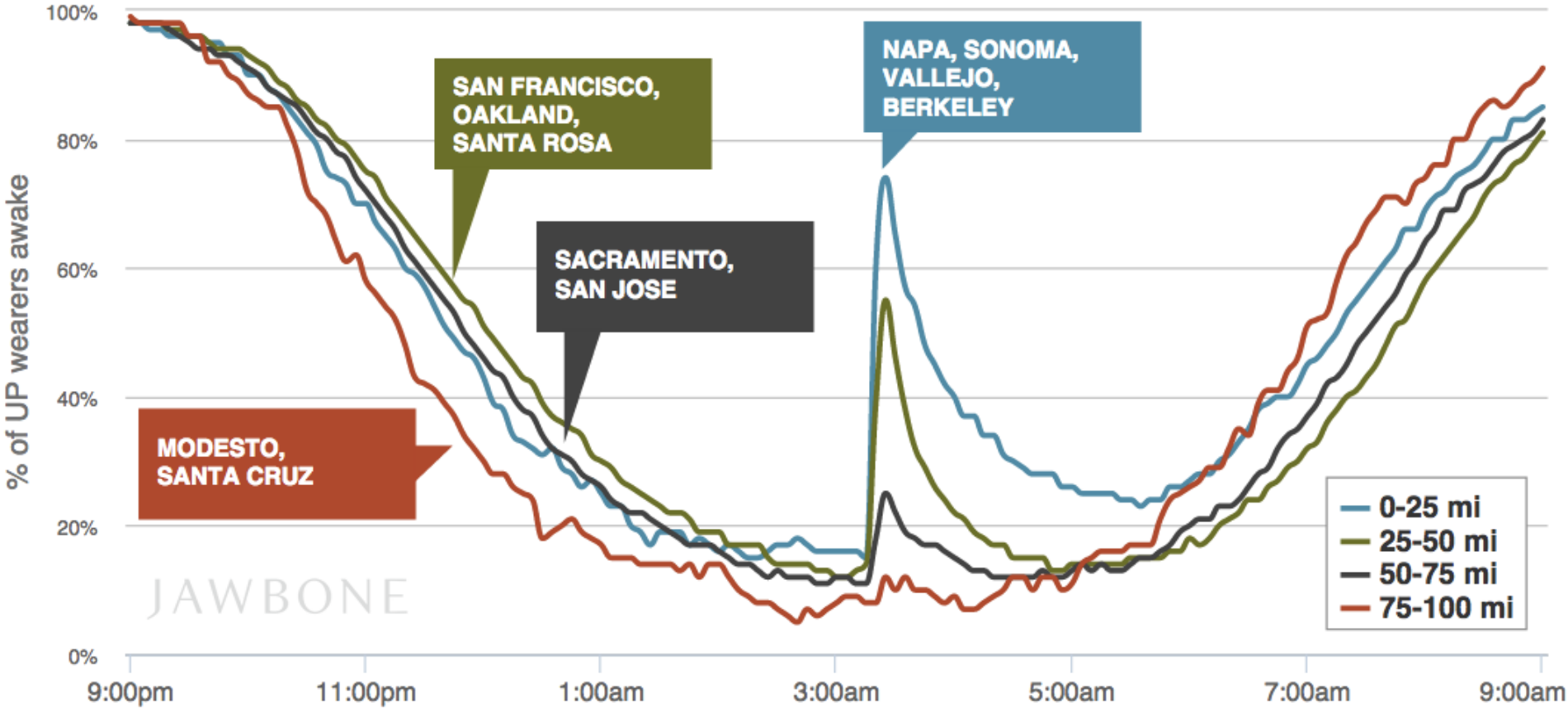
They know you called a gynecologist, spoke for a half hour, and then called the local Planned Parenthood's number later that day. But nobody knows what you spoke about.

1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naïve Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Using Mahout
Using Weka
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
80%
Using ETL
How much Data?
Google OpenRefine
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran



The future

- Easy to use data tools
- New devices and interfaces
- Invisible technology
- Initiatives from EU, InnovateUK and yours truly and you!


100%
80%
60%
40%
20%
0%
% of UP wearers awake
9:00pm
11:00pm
1:00am
3:00am
5:00am
7:00am
9:00am
SAN FRANCISCO, OAKLAND, SANTA ROSA
NAPA, SONOMA, VALLEJO, BERKELEY
SACRAMENTO, SAN JOSE
MODESTO, SANTA CRUZ
JAWBONE
0-25 mi
25-50 mi
50-75 mi
75-100 mi

Oculus

Thank you!!!

@xchaotic
lech.rzedzicki@innovateuk.gov.uk