False Positives:
Risks and Opportunities to Big Data

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca

Every purchase you make with a credit card, every magazine subscription you buy and medical prescription you fill, every Web site you visit and e-mail you send or receive, every academic grade you receive, every bank deposit you make, every trip you book and every event you attend -- all these transactions and communications will go into what the Defense Department describes as "a virtual, centralized grand database."

Safire, “You Are a Suspect,” NYT, about the Total Information Awareness project)

in the near future. [redacted] (S)

Stored Information Applications

- Real-time access to and decryption of stored electronic information secured by hardware-based encryption could be performed utilizing the "Clipper" technique.

[redacted] (S)

[redacted] (S)

[redacted] (S)

b1 b3 b4 b5 per NSA

- Technical solutions, such as they are, will only work if they are incorporated into all encryption products. To ensure that this occurs, legislation mandating the use of Government-approved encryption products or adherence to Government encryption criteria is required.

The attached chart provides a graphic representation of the information and summary provided above.

2 types of data

- Information at Rest
  - Large corporate databases
  - Large information indexes (Google)
  - Crowdsourced databases (YouTube & Faceook)
  - *Market Basket analysis ("Beer and diapers")*
- Information in Motion
  - Signal traffic / communications traffic
  - *Filtering and Sifting*

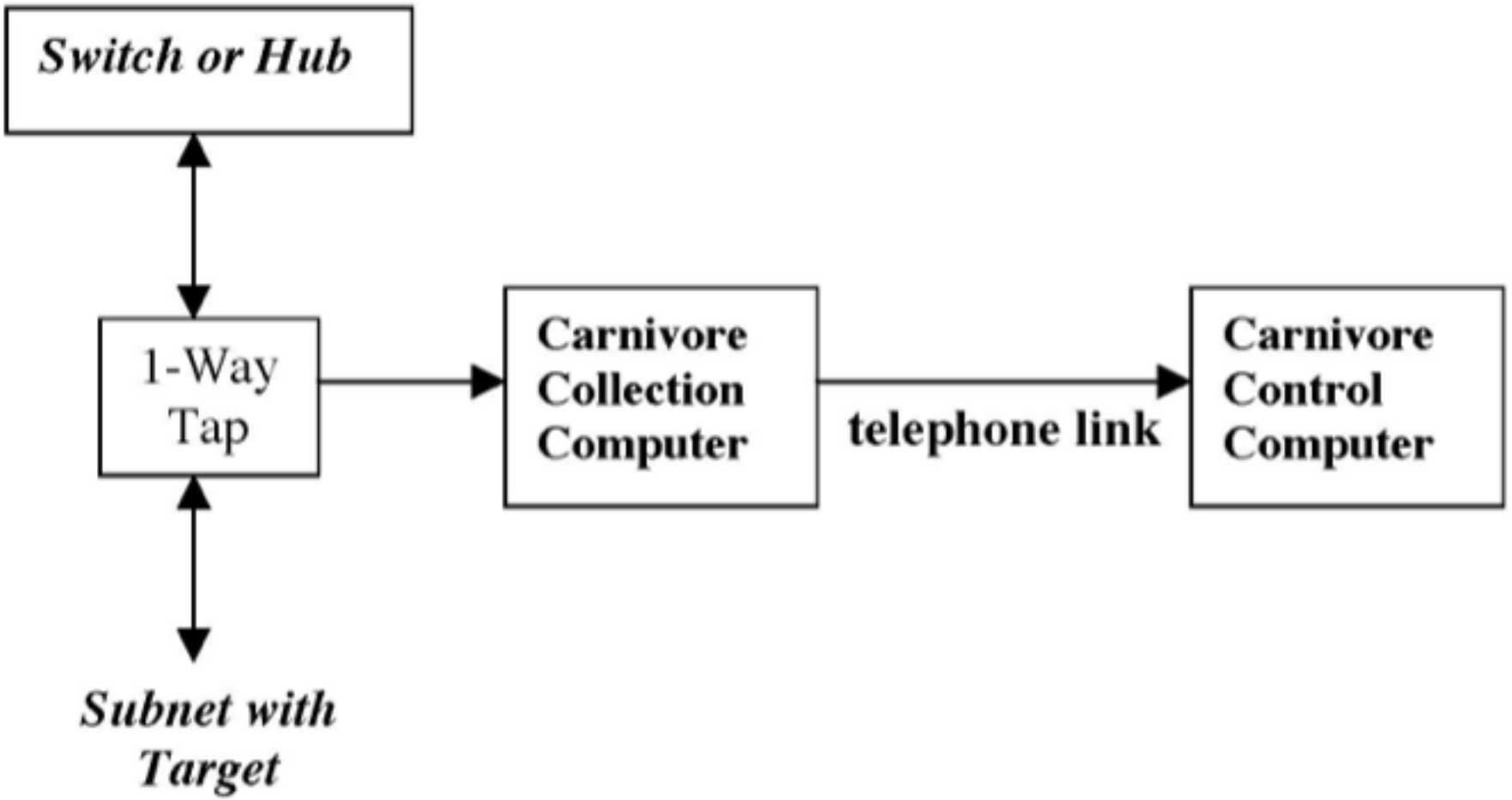
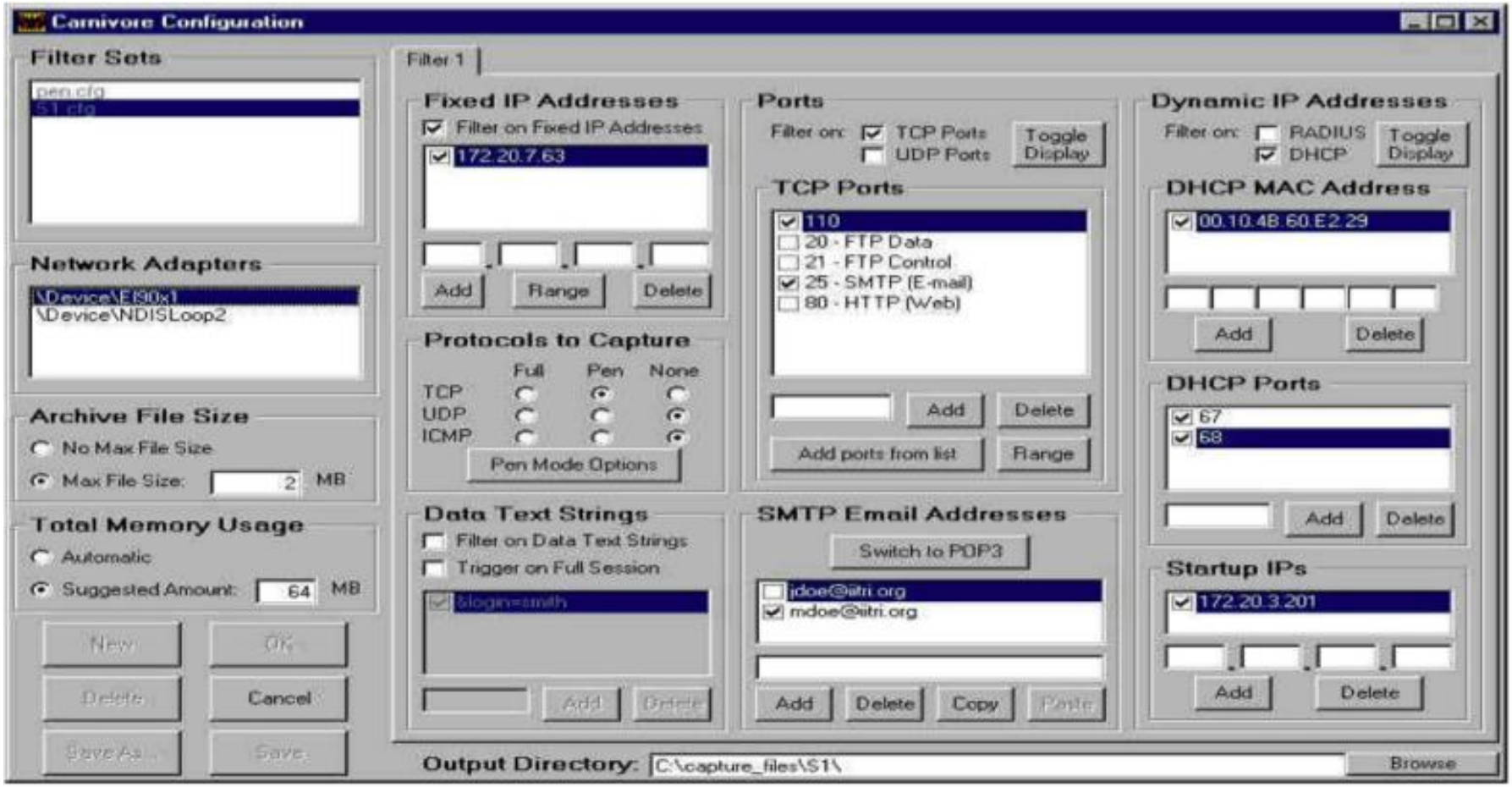
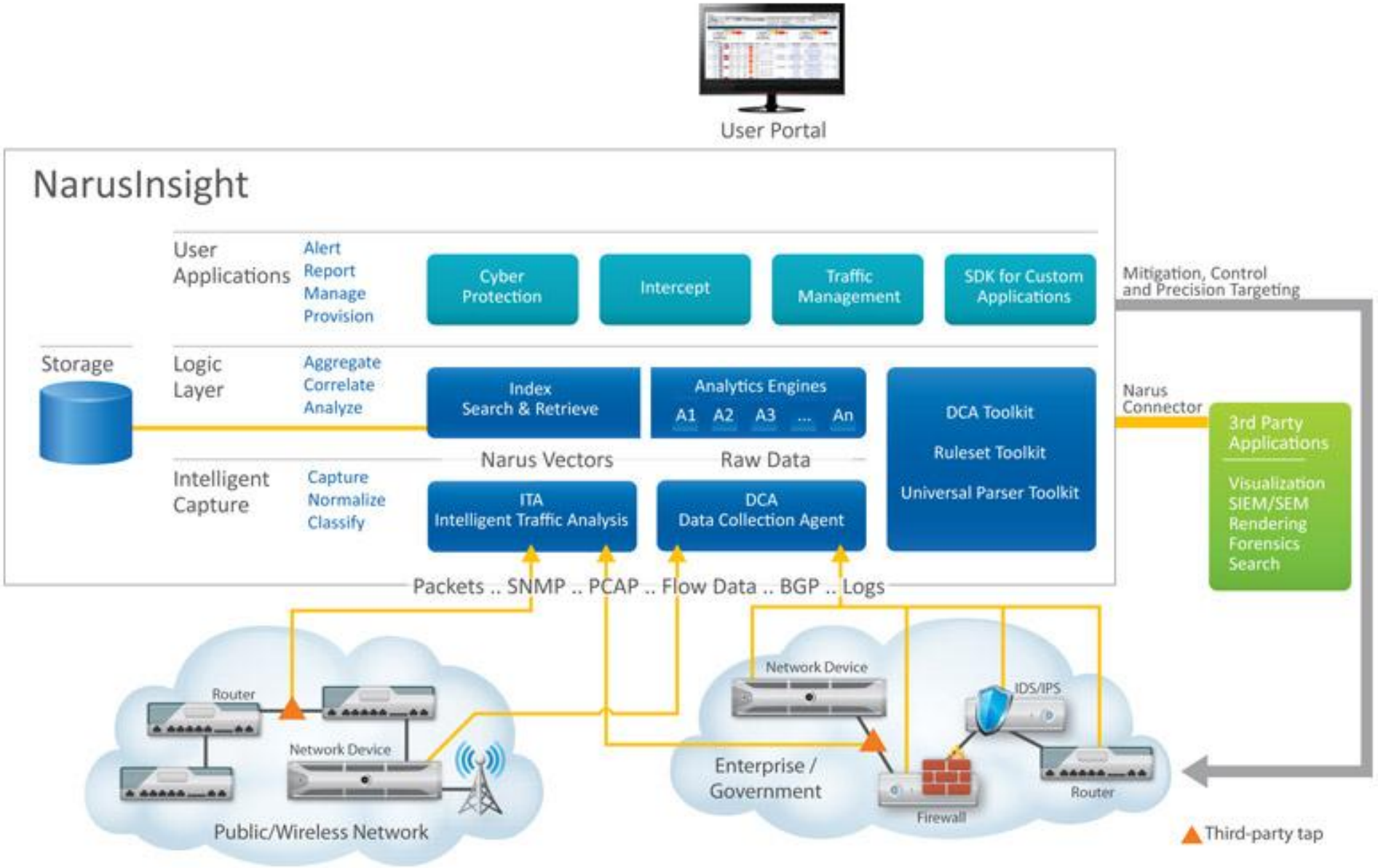


Figure ES-1. Carnivore Architecture

Carnivore Filter




User Portal
NarusInsight
User Applications
Alert
Report
Manage
Provision
Cyber Protection
Intercept
Traffic Management
SDK for Custom Applications
Mitigation, Control and Precision Targeting
Storage
Logic Layer
Aggregate
Correlate
Analyze
Index Search & Retrieve
Analytics Engines
A1 A2 A3 ... An
DCA Toolkit
Ruleset Toolkit
Universal Parser Toolkit
Narus Connector
3rd Party Applications
Visualization
SIEM/SEM
Rendering
Forensics
Search
Narus Vectors
Raw Data
Intelligent Capture
Capture
Normalize
Classify
ITA
Intelligent Traffic Analysis
DCA
Data Collection Agent
Packets .. SNMP .. PCAP .. Flow Data .. BGP .. Logs
Router
Network Device
Public/Wireless Network
Network Device
IDS/IPS
Enterprise / Government
Firewall
Router
Third-party tap

narus.com

DataSift

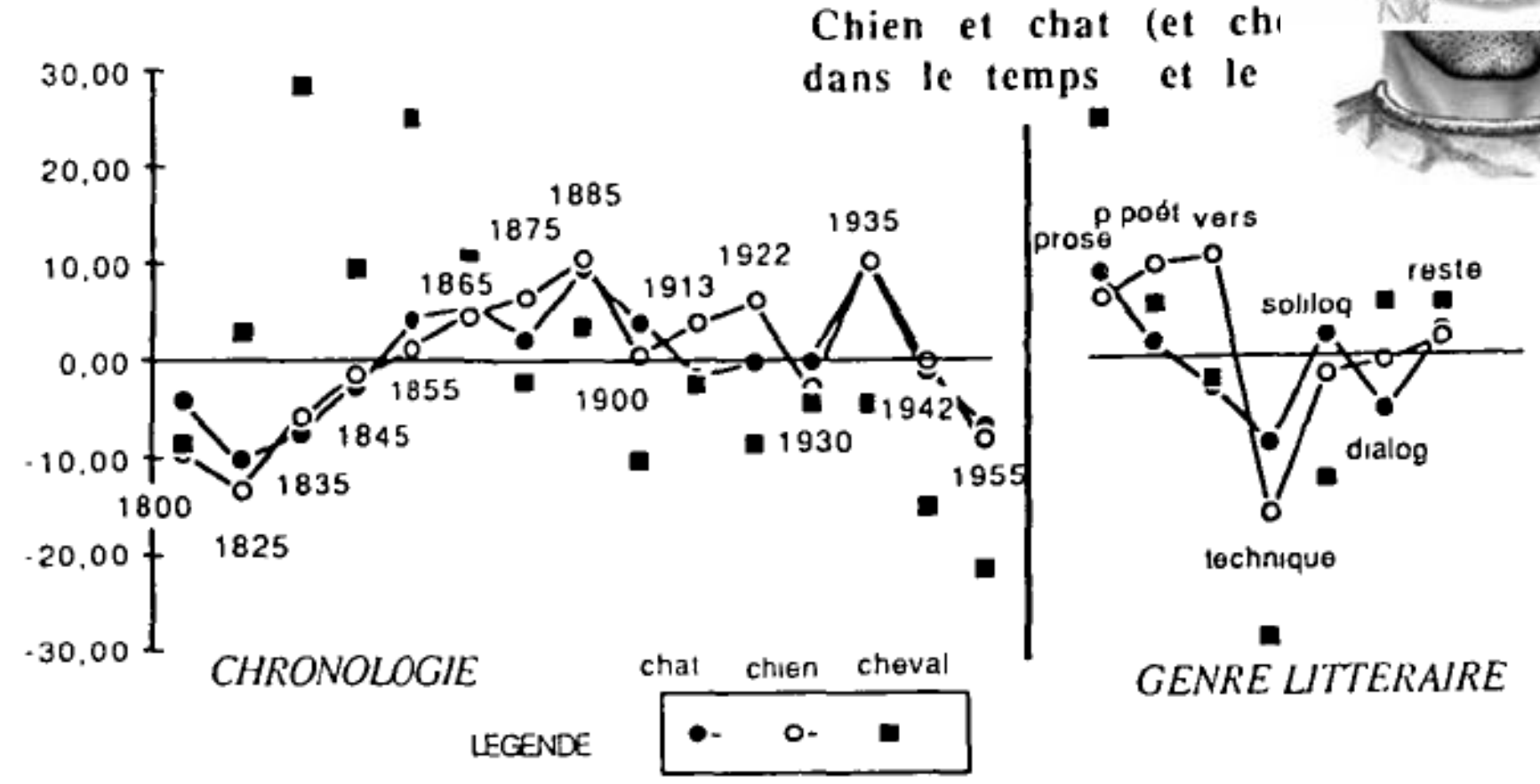
DataSift is a real-time media curation platform, allowing you to mine the Twitter Firehose for tweets matching the specific criteria of your choice. DataSift's custom Curation Stream Definition Language allows you to filter based on any meta data within a tweet, in addition to a number of other data providers from around the Web.

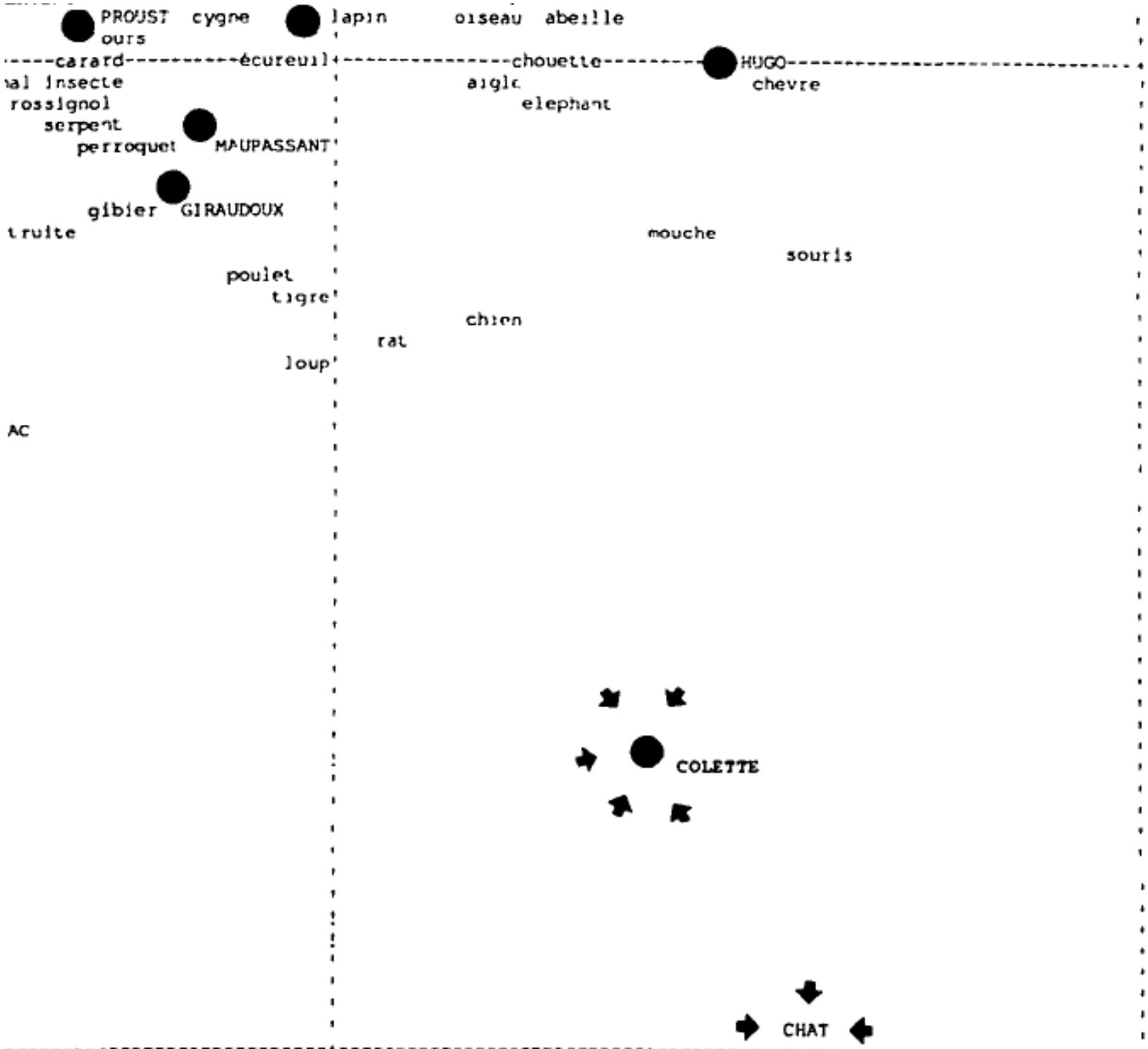
dev.twitter.com/docs/twitter-data-providers

Bush, “As We May Think”

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear.

Brunet and the Grand Corpus

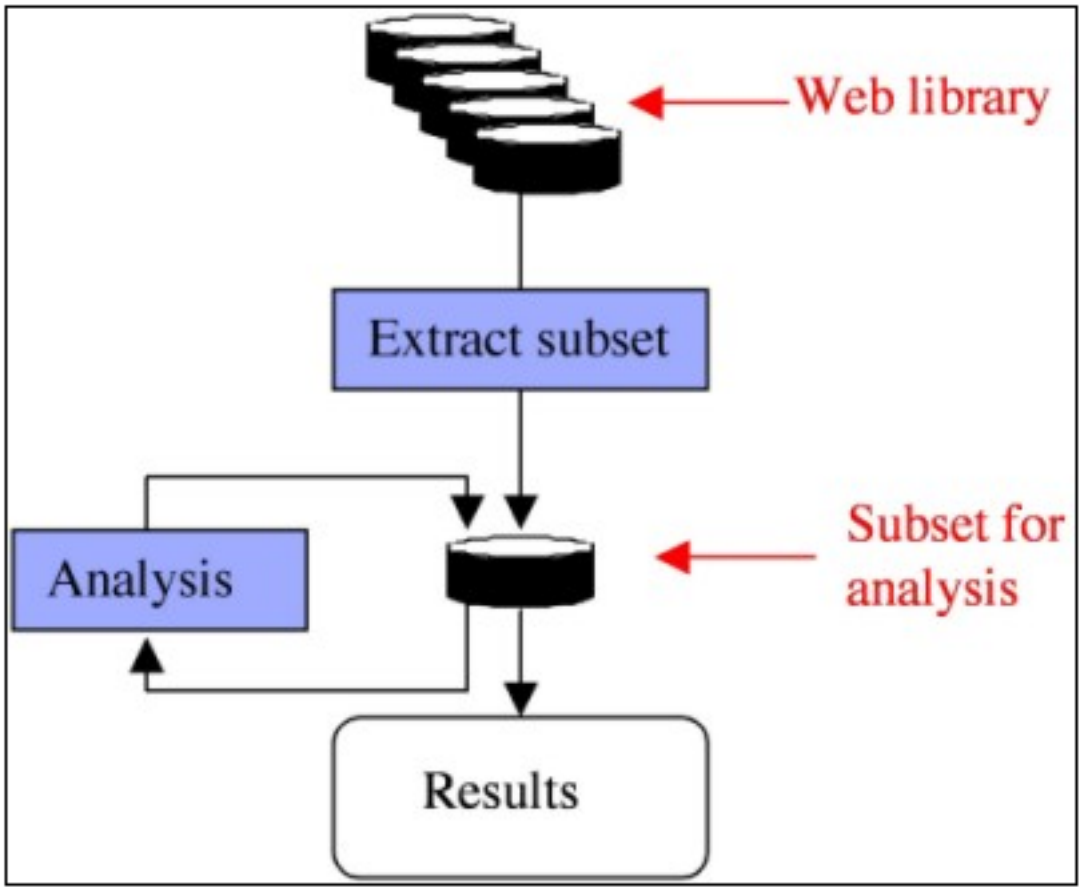



PROUST
cygne
lapin
oiseau
abeille
ours
carard
écureuil
chouette
HUGO
al insecte
aigle
chevre
rossignol
elephant
serpent
perroquet
MAUPASSANT
gibier
GIRAUDOUX
truite
mouche
souris
poulet
tigre
chien
rat
loup
AC
COLETTE
CHAT

Opportunities

- Filtering and Subsetting – extracting useful subsets
- Enrichment – automatically adding value
- Sequence Alignment – tracking permutation of passages
- Diachronic Analysis – studying human activity over time
- Classification and Clustering – exploring correlations and finding similar documents
- Social Network Analysis – studying networks of people, ideas and organizations
- Life Hacking – “Know thyself”

Subsetting: Cornel WebLab

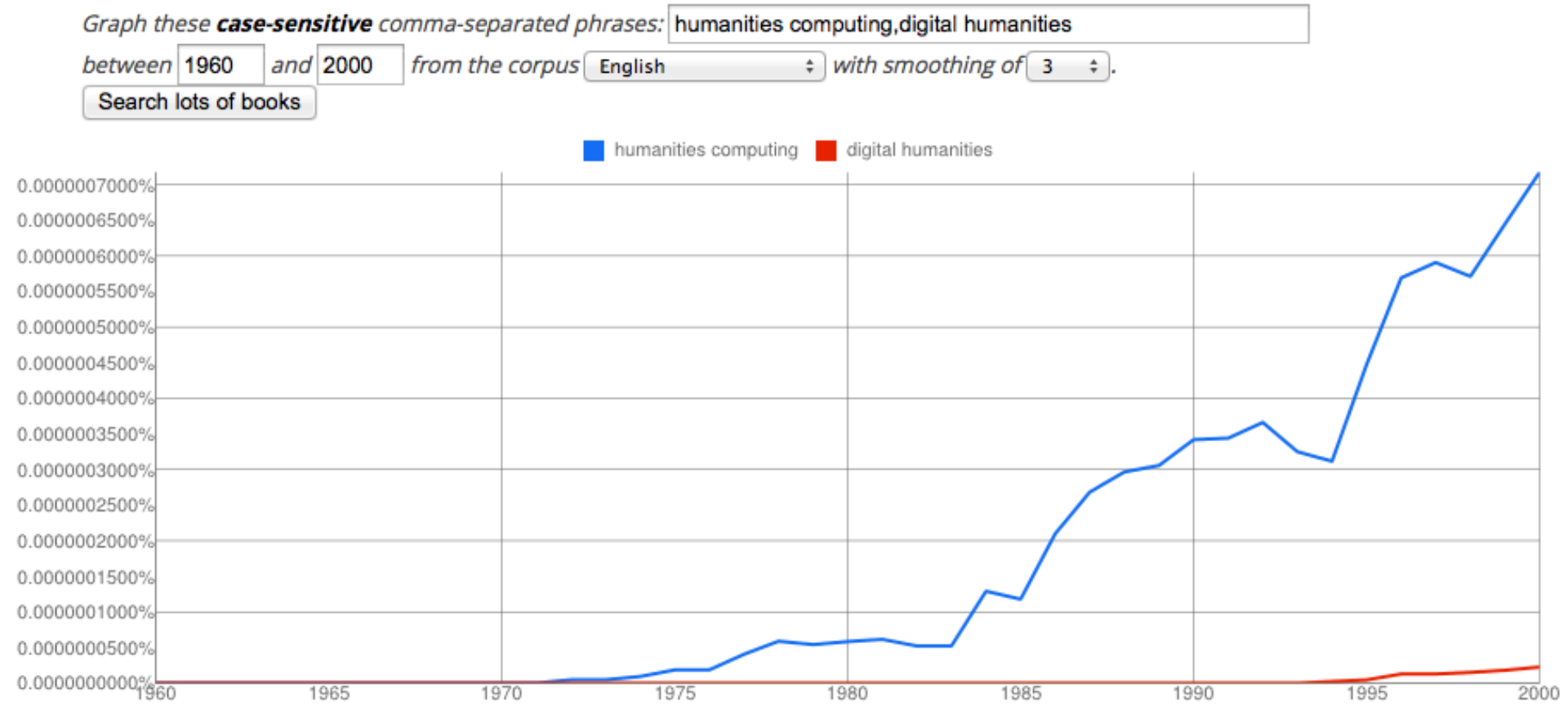


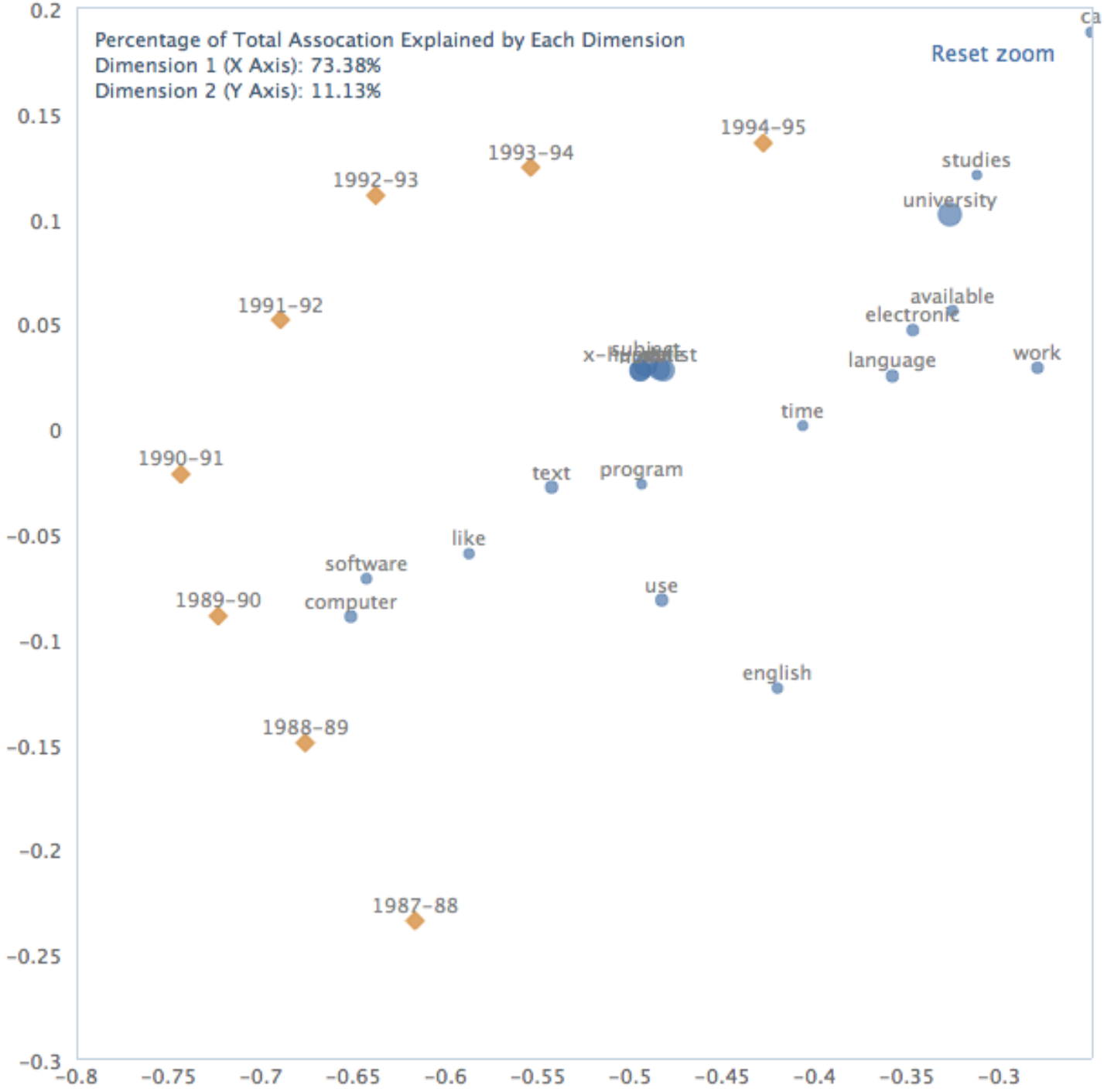
Sequence Alignment

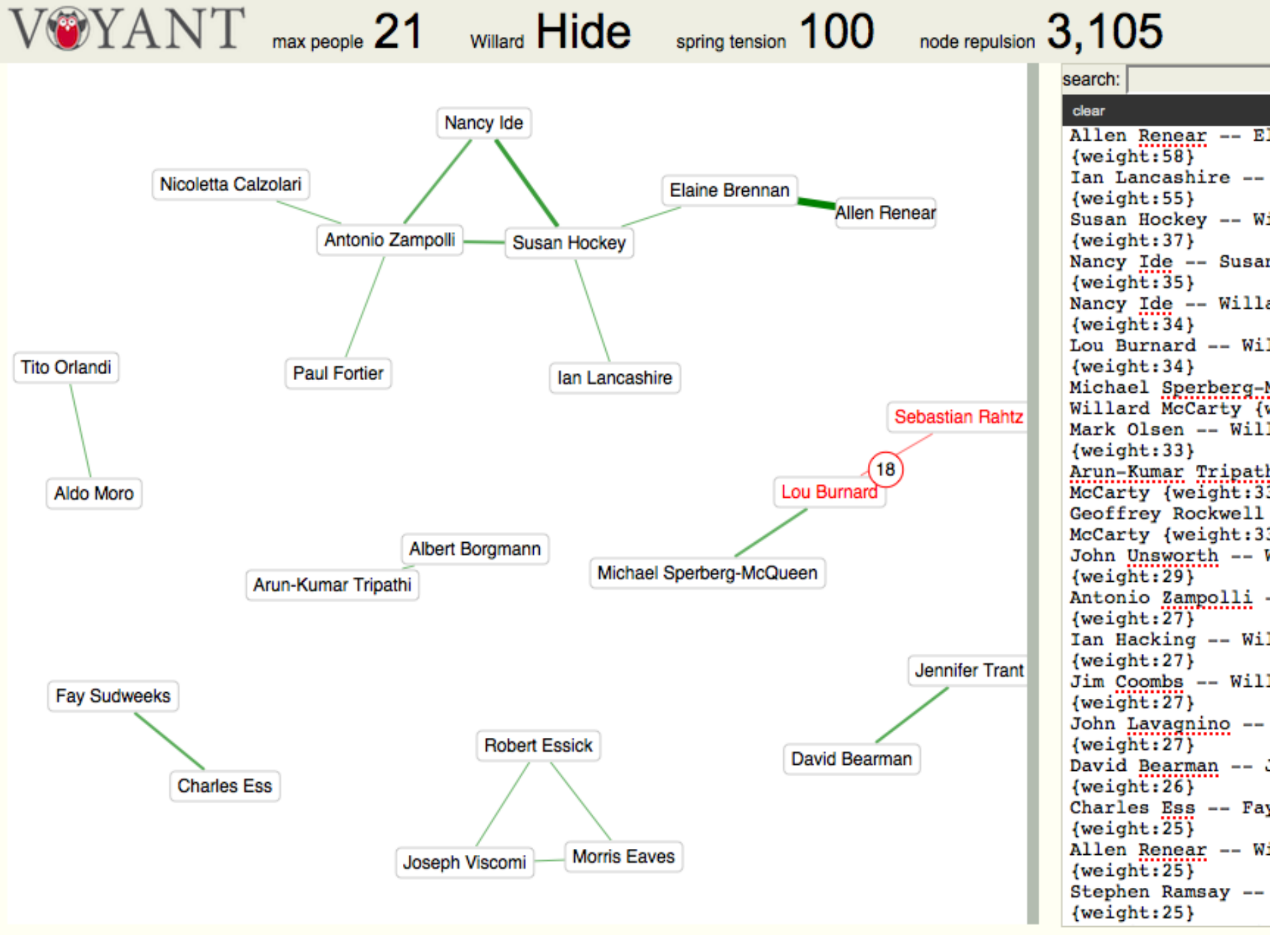
Les chinois rapportent qu'il disoit souvent: *c'est dans l'occident que se trouve le veritable saint*. Et cette sentence estoit tellement gravée dans l'esprit des sçavans, que soixante-cinq ans aprés la naissance de nostre seigneur, l'empereur Mimti touché de ces paroles, et déterminé par l'image d'un homme qui se presenta à luy durant le sommeil venant de l'occident, envoya de ce costé-là des ambassadeurs, avec ordre de continuer leur voyage jusqu'à ce qu'ils eussent rencontré le saint que le ciel luy avoit fait connoistre. C'estoit à peu prés le temps auquel Saint Thomas preschoit dans les Indes la loy chrétienne; et si ces mandarins eussent suivi leurs ordres, peut-estre que la Chine auroit profité de la prédication de cet apostre. Mais les dangers de la mer, qu'ils craignirent, les obligea de s'arrester à la premiere isle, où ils trouverent l'idole Fo ou Foe, qui avoit déja corrompu les Indes plusieurs siecles auparavant, de son execrable doctrine. (LeComte [1696])	Né 478 ans avant le Christ, il disait souvent, tel un prophète : Dans l'Occident se trouve le véritable saint. Soixante-cinq ans après la naissance du Christ, l'Empereur Mimti, interprétant cette parole du Maître, et sollicité par un songe, envoya vers l'Occident des ambassadeurs, avec ordre de continuer leur voyage jusqu'à ce qu'ils eussent rencontré le saint. En ce temps-là, saint Thomas prêchait dans les Indes la foi chrétienne ; et si ces mandarins s'étaient acquittés de leur mission, au lieu de s'arrêter dans la première île à cause du danger de la mer, peut-être la Chine aurait-elle fait partie de l'Église romaine... (Hazard [1935])

Horton, Olsen and Roe, “Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections.” *Digital Studies* 2:1

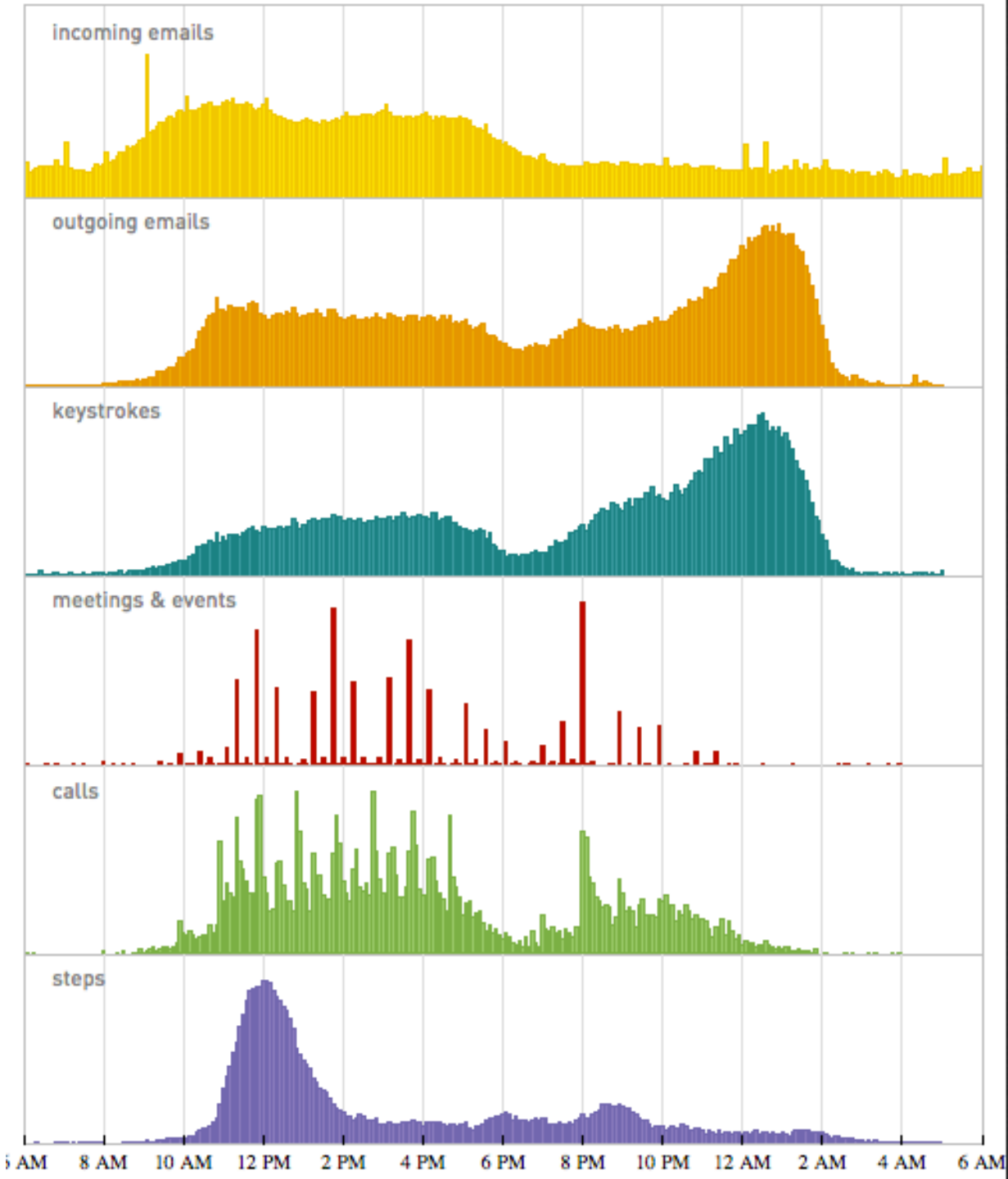
Diachronic Analysis




Percentage of Total Assocation Explained by Each Dimension
Dimension 1 (X Axis): 73.38%
Dimension 2 (Y Axis): 11.13%
Reset zoom
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.8
-0.75
-0.7
-0.65
-0.6
-0.55
-0.5
-0.45
-0.4
-0.35
-0.3
ca
1994-95
1993-94
1992-93
studies
university
available
electronic
1991-92
language
work
time
1990-91
text
program
like
software
use
1989-90
computer
english
1988-89
1987-88

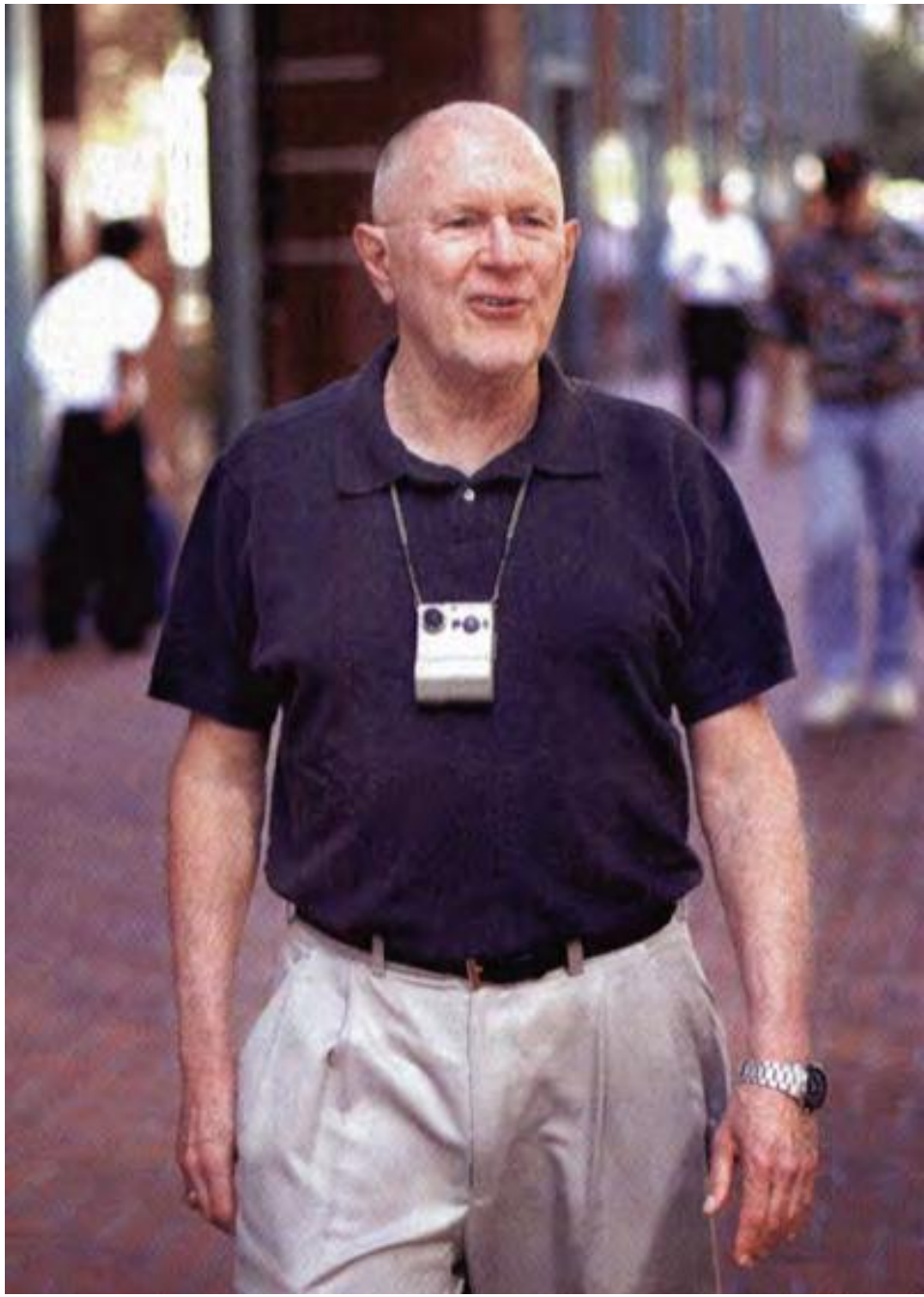

VOYANT
max people 21
Willard Hide
spring tension 100
node repulsion 3,105
search:
clear
Nancy Ide
Nicoletta Calzolari
Elaine Brennan
Allen Renear
Antonio Zampolli
Susan Hockey
Tito Orlandi
Paul Fortier
Ian Lancashire
Sebastian Rahtz
18
Aldo Moro
Lou Burnard
Albert Borgmann
Arun-Kumar Tripathi
Michael Sperberg-McQueen
Jennifer Trant
Fay Sudweeks
Robert Essick
David Bearman
Charles Ess
Joseph Viscomi
Morris Eaves
Allen Renear -- E
{weight:58}
Ian Lancashire --
{weight:55}
Susan Hockey -- Wi
{weight:37}
Nancy Ide -- Susan
{weight:35}
Nancy Ide -- Willa
{weight:34}
Lou Burnard -- Wil
{weight:34}
Michael Sperberg-M
Willard McCarty {w
Mark Olsen -- Will
{weight:33}
Arun-Kumar Tripath
McCarty {weight:3
Geoffrey Rockwell
McCarty {weight:3
John Unsworth -- W
{weight:29}
Antonio Zampolli -
{weight:27}
Ian Hacking -- Wil
{weight:27}
Jim Coombs -- Will
{weight:27}
John Lavagnino --
{weight:27}
David Bearman -- J
{weight:26}
Charles Ess -- Fay
{weight:25}
Allen Renear -- Wi
{weight:25}
Stephen Ramsay --
{weight:25}

Wolfram: Quantified Self

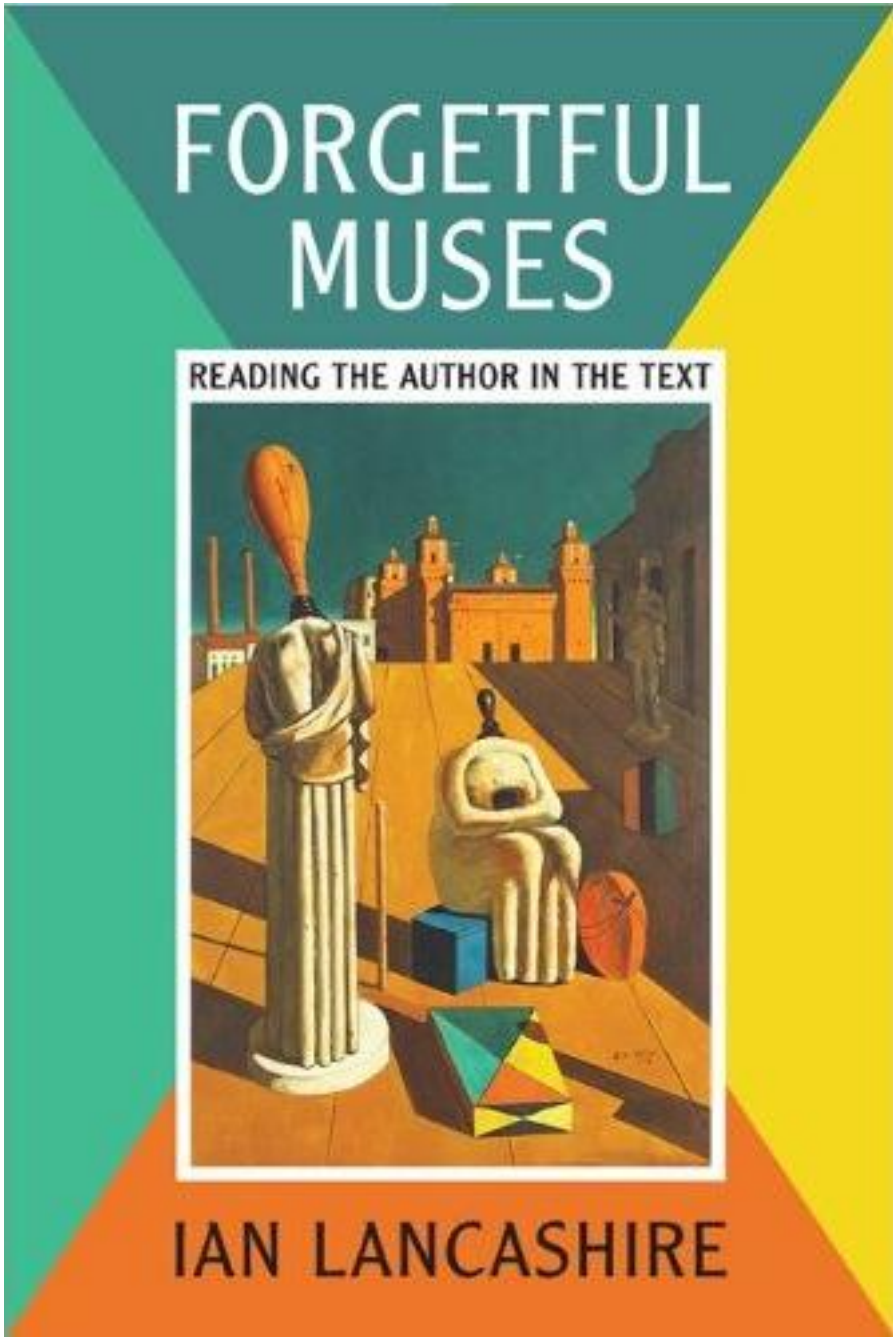


Gordon Bell: MyLifeBits

Wearing SenseCam

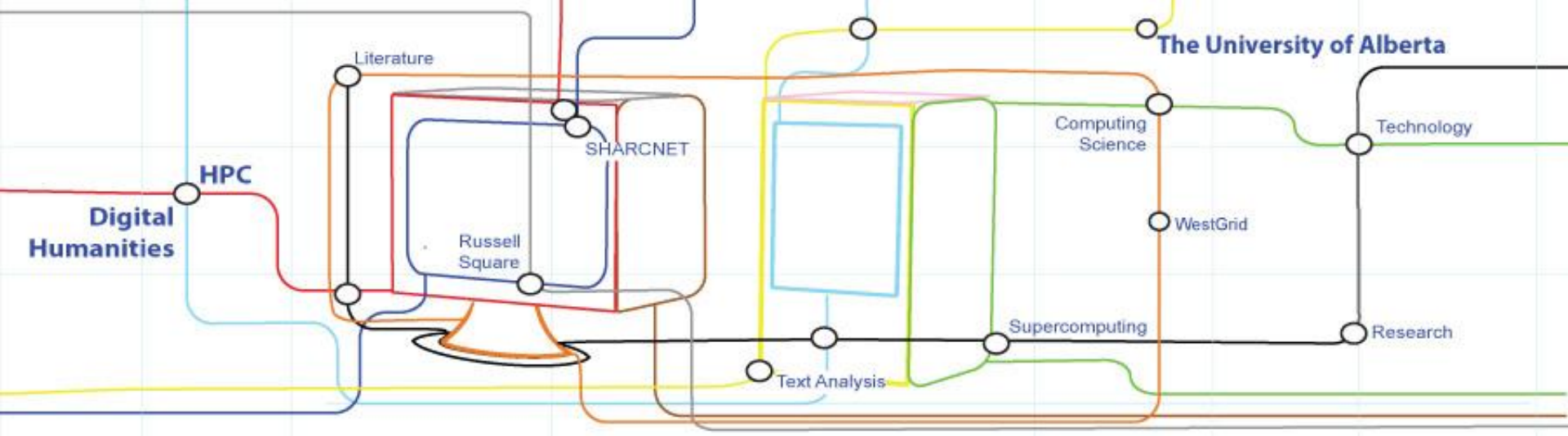


Lancashire, Forgetful Muses



Predictive Data Mining

It would be unfortunate if data mining for terrorism discovery had currency within national security, law enforcement, and technology circles because pursuing this use of data mining would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community.
(Executive Summary, p. 1)



- *Conjecturator*

  *Feature **A** appears more/less often in the group of texts **B** than in group **C** that are distinguished by structural feature **D***

- 87,000 Conjectures
  - 19th Century Eng. Fiction divided in to 10-year chunks
  - Thesaurus used for word group “Features”

https://twitter.com/conjecturator

Tweets

Conjecturator @conjecturator 20 Jun 10

Why does the word group ground appear more in Proc. Royal Geographical Society and Monthly Record of Geography than in Am. J. Philology?

Expand

Conjecturator @conjecturator 18 Jun 10

Why does the word group instant appear more in The Old Testament Student than in Hermes?

Expand

Conjecturator @conjecturator 17 Jun 10

Why does the word group reincarnation appear more in The Auk than in J. Ethnological Society of London (1869-1870)?

Expand

Thanks
tapor.ca
voyant-tools.ca

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca