Technology, Big Data, and Mixed Methods

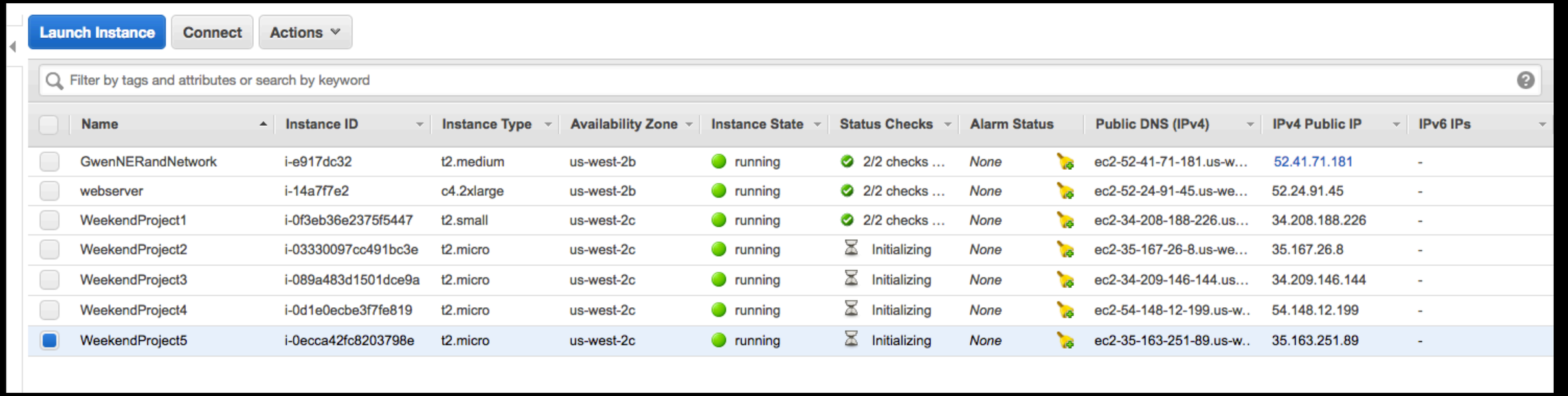
Merlin Chowkwanyun

Big Data

Computing Horsepower

Volume


MyOSG
Home
Resource Group
Virtual Organization
Status Map
Misc
Search OSG
Site (None)
ALBERTA
MANITOBA
BRITISH COLUMBIA
Edmonton
SASKATCHEWAN
NEWFOUNDLAND AND LABRADOR
Calgary
ONTARIO
QUEBEC
Winnipeg
Vancouver
NORTH DAKOTA
MONTANA
MINNESOTA
NEW BRUNSWICK
PE
Ottawa
Montreal
NOVA SCOTIA
MAINE
SOUTH DAKOTA
WISCONSIN
Toronto
OREGON
IDAHO
WYOMING
IOWA
United States
NEVADA
UTAH
COLORADO
KANSAS
MISSOURI
NJ
DE
Las Vegas
ARIZONA
NEW MEXICO
Phoenix
TENNESSEE
NORTH CAROLINA
SOUTH CAROLINA
ALABAMA
GEORGIA
TEXAS
San Antonio
LOUISIANA
Houston
CENTER FOR HIGH THROUGHPUT COMPUTING
Open Science Grid


Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
Alarm Status
Public DNS (IPv4)
IPv4 Public IP
IPv6 IPs
GwenNERandNetwork i-e917dc32 t2.medium us-west-2b running 2/2 checks ... None ec2-52-41-71-181.us-w... 52.41.71.181 -
webserver i-14a7f7e2 c4.2xlarge us-west-2b running 2/2 checks ... None ec2-52-24-91-45.us-we... 52.24.91.45 -
WeekendProject1 i-0f3eb36e2375f5447 t2.small us-west-2c running 2/2 checks ... None ec2-34-208-188-226.us... 34.208.188.226 -
WeekendProject2 i-03330097cc491bc3e t2.micro us-west-2c running Initializing None ec2-35-167-26-8.us-we... 35.167.26.8 -
WeekendProject3 i-089a483d1501dce9a t2.micro us-west-2c running Initializing None ec2-34-209-146-144.us... 34.209.146.144 -
WeekendProject4 i-0d1e0ecbe3f7fe819 t2.micro us-west-2c running Initializing None ec2-54-148-12-199.us-w.. 54.148.12.199 -
WeekendProject5 i-0ecca42fc8203798e t2.micro us-west-2c running Initializing None ec2-35-163-251-89.us-w.. 35.163.251.89 -

Big Data

Computing Horsepower

Big Data

Computing Horsepower

Volume

ID	Characteristic A	Characteristic B	Characteristic C	Characteristic D	Characteristic E	Characteristic F

SQL

ID	Characteristic A	Characteristic B	Characteristic C	Characteristic D	Characteristic E	Characteristic F

Non-Relational Flexible Databases


```
"retweet_count": 2301,
    "retweeted": false,
    "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "text": "It is being reported by virtually everyone, and is a fact,  that the media pile on against me is the worst in American political history!",
    "truncated": false,
    "user": {
        "contributors_enabled": false,
        "created_at": "Wed Mar 18 13:46:38 +0000 2009",
        "default_profile": false,
        "default_profile_image": false,
        "description": "#TrumpPence16",
        "entities": {
            "description": {
                "urls": []
            },
            "url": {
                "urls": [
                    {
                        "display_url": "DonaldJTrump.com",
                        "expanded_url": "http://www.DonaldJTrump.com",
                        "indices": [
                            0,
                            23
                        ],
                        "url": "https://t.co/mZB2hymxC9"
                    }
                ]
            }
        },
        "favourites_count": 35,
```

Non-Relational Flexible Databases


```
"retweet_count": 2301,
    "retweeted": false,
    "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "text": "It is being reported by virtually everyone, and is a fact,  that the media pile on against me is the worst in American political history!",
    "ideology: "conservative"
    "truncated": false,
    "user": {
        "contributors_enabled": false,
        "created_at": "Wed Mar 18 13:46:38 +0000 2009",
        "default_profile": false,
        "default_profile_image": false,
        "description": "#TrumpPence16",
        "entities": {
            "description": {
                "urls": []
            },
            "url": {
                "urls": [
                    {
                        "display_url": "DonaldJTrump.com",
                        "expanded_url": "http://www.DonaldJTrump.com",
                        "indices": [
                            0,
                            23
                        ],
                        "url": "https://t.co/mZB2hymxC9"
                    }
                ]
            }
        },
        "favourites_count": 35,
```

INTRODUCING PROJECT TOXICDOCS ALPHA

Columbia University and the City University of New York

Millions of Previously Classified Documents on Industrial Poisons (and Counting)

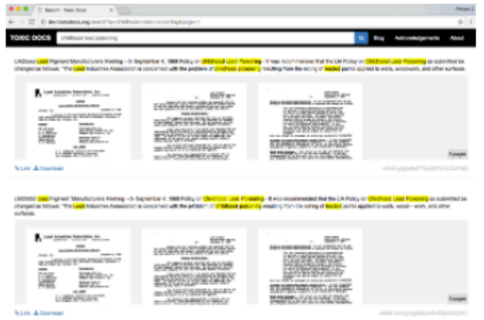
Type anything here

Advanced Search

Blazingly fast searches of once-secret industry documents

Columbia University's Center for the History and Ethics of Public Health, located at its Mailman School of Public Health, and the City University of New York's Graduate Center are proud to jointly present Toxic Docs. This dataset and website contain millions of pages of previously secret documents about toxic substances. They include secret internal memoranda, emails, slides, board minutes, unpublished scientific studies, and expert witness reports -- among other kinds of documents -- that emerged in recent toxic tort litigation.

Over the next couple years, we'll be constantly adding material from lawsuits involving lead, asbestos, silica, and PCBs, among other dangerous substances. Innovations in parallel and cloud computing have made conversion of these documents into machine-readable, searchable text a far faster process than would have been the case just a decade ago.



The Problem of Volume

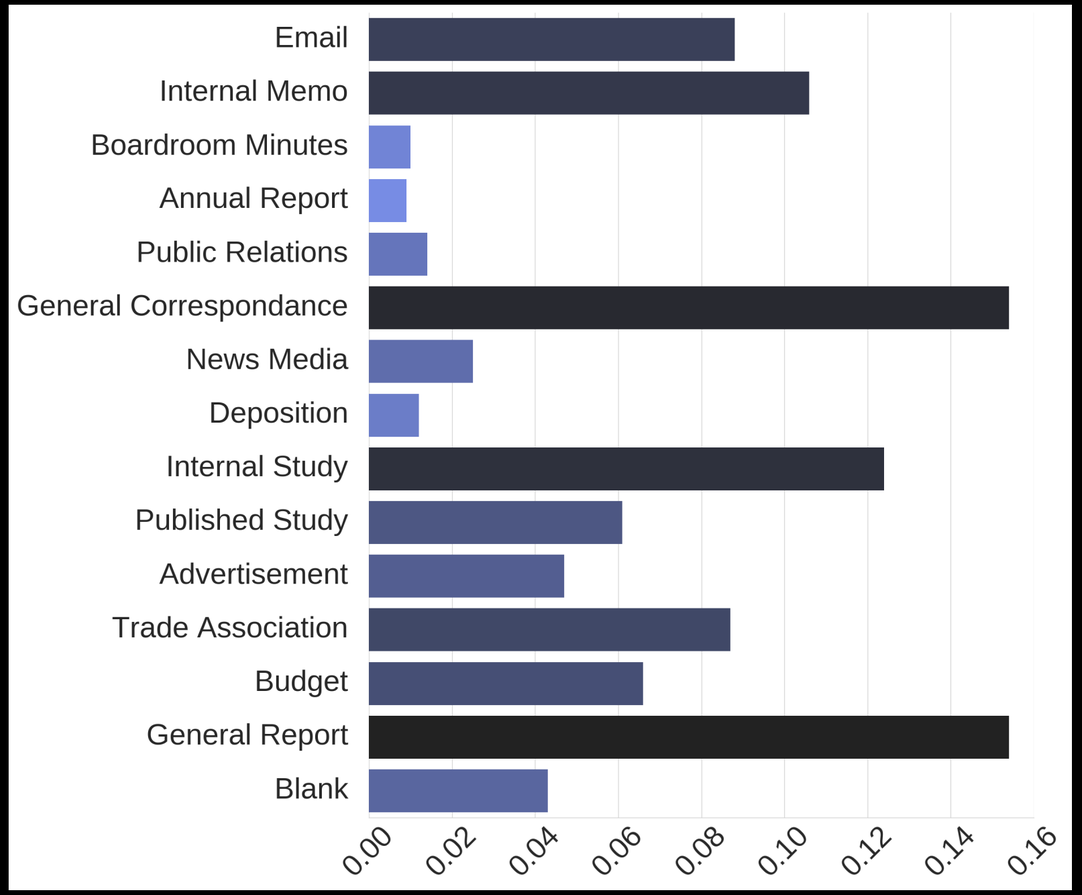
1 computer
4 million pages
=
6 months

4,228 servers
4 million pages
=
1 day

Huge document stash

1k manually labeled

15 “genres”/categories



Board Minutes Example

"AGENDA"

"MINUTES"

Attendee Lists

AGENDA
REGIONAL MEETING OF THE MCA BOARD OF DIRECTORS
9:30 a.m. (PST), Wednesday, March 15, 1972
Del Monte Lodge, Pebble Beach, California

I. OPENING REMARKS AND INTRODUCTION OF GUESTS.

II. MINUTES OF FEBRUARY 8, 1972, MEETING.

III. FINANCIAL REPORT. (Enclosure)

IV. BOARD OF DIRECTORS.

(a) Appointment of Alternate: C. A. Cash, Diamond Shamrock Corporation (for James A. Hughes)

(b) Proposed Rules of Organization and Procedure for the Economic Policy Review Committee. (Enclosure)

(c) Proposed Amendments to the Association's Bylaws. (Enclosure)

(d) Reports of Board Liaison Committees:
Environmental Quality and Control -- Carl A. Gerstacker
Government Relations -- George W. Russell
Other Technical Committees -- Lee V. Dauler
Public Relations and Education -- Luther S. Roehm

V. COMMITTEE APPOINTMENTS. (Enclosure)

VI. STAFF REPORT. (Appendix Enclosed)

VII. ADJOURNMENT.

See this shows membership

Next Meeting of the Dir ctors -- The Madison, Washington, D. C., 8:30 a.m., (EST), Tuesday, April 11, 1972 (Breakfast Meeting).

CMA 036174

CONFIDENTIAL
Subject to Protective Order in Ross v. Conoco, Inc., No. 90-4837
14th Judicial District Court
Calcasieu Parish, Louisiana

1386

MINUTES of the two hundred seventeenth meeting of the Directors of the Manufacturing Chemists' Association, Inc., held at the Pinnacle Club, New York City, Monday, November 20, 1972, at 4:00 p.m. There were present:

Directors:

F. Perry Wilson, Chairman
*H. Harold Bible
Frederick L. Bissinger
Werner C. Brown
*Harry W. Buchanan
Fletcher L. Byrom
C. C. Candee
Herschel H. Cudd
Lee V. Dauler
Edward J. Donley
William C. Douce
*William P. Drake
William J. Driver
Carl A. Gerstacker
*James M. Gill
John M. Henske
H. E. Hirschland
James A. Hughes
Edward R. Kane
Philip F. Kirk
Lloyd G. Lillico
Robert H. Malott
*Robert M. Morris
Charles S. Munson
Peter C. Reilly
Luther S. Roehm
George W. Russell
Harold E. Thayer
*Jesse Werner
James R. Carnes

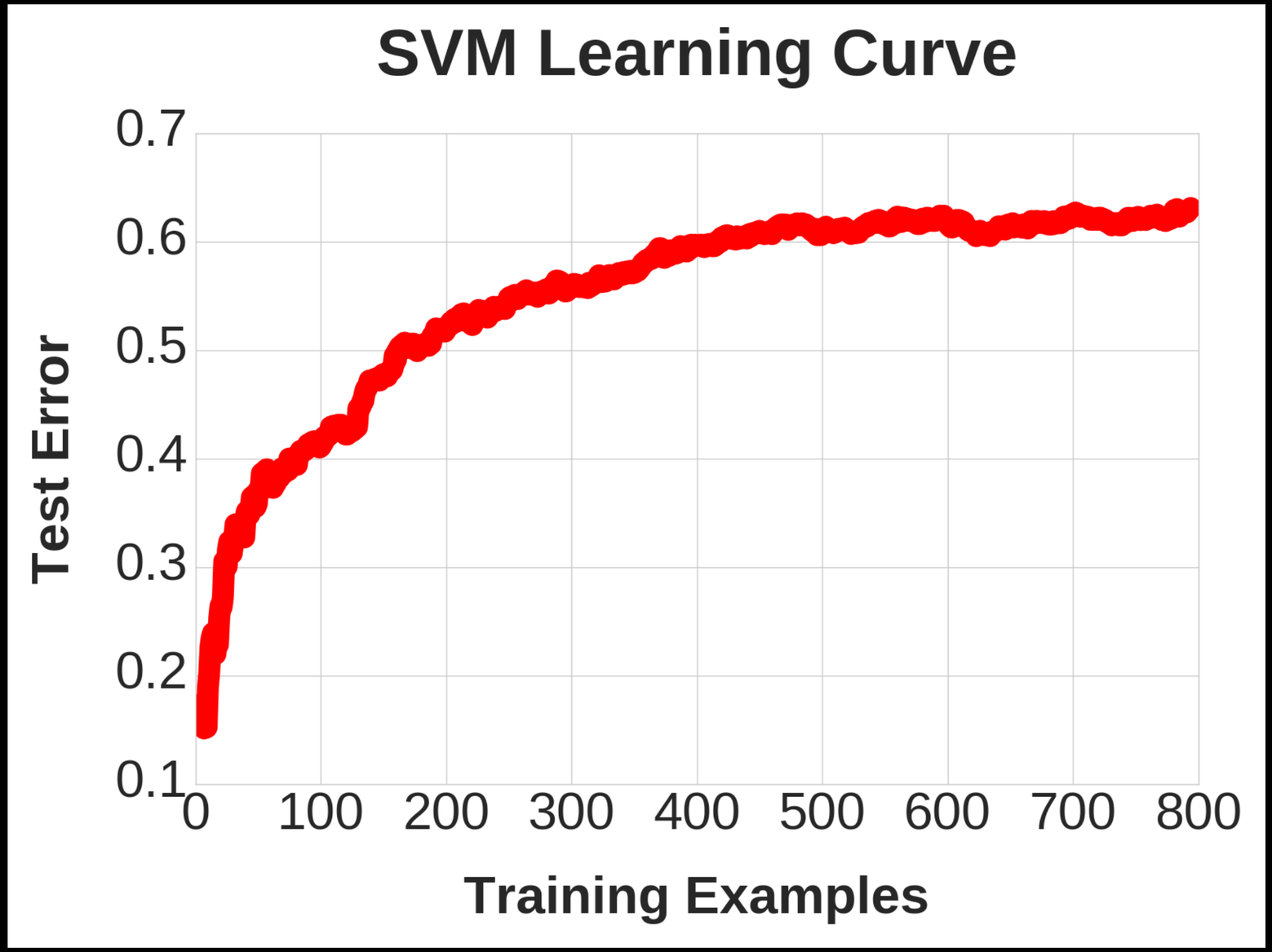
Alternates:

Warren M. Anderson (for F. Perry Wilson)
*Stanley H. Anonsen (for Harold E. Thayer)
C. A. Cash (for James A. Hughes)
*David N. Clark (for Lee V. Dauler)
*Frank J. Connor (for H. E. Hirschland)
John T. Connor (for Frederick L. Bissinger)
*Philip B. Dalton (for Jesse Werner)
*H. N. Fiaccone (for Luther S. Roehm)
*Richard Fleming (for Edward J. Donley)
John L. Gillis (for H. Harold Bible)
*H. C. Hollands (for Lloyd G. Lillico)
L. H. Johnstone (for William C. Douce)
Gordon Kiddoo (for Donald G. Stevens)
*James McWhirter (for William P. Drake)
*Robert L. Mitchell (for John W. Brooks)
Richard M. Morrow (for Herschel H. Cudd)
Earl C. Ray (for Philip F. Kirk)
Thomas E. Reilly, Jr. (for Peter C. Reilly)
Dickson L. Whitney (for C. C. Candee)
*Richard N. Williams (for John M. Henske)

General Counsel: Lloyd Symington

By invitation: Bruce M. Barackman, MCA
George E. Best, MCA

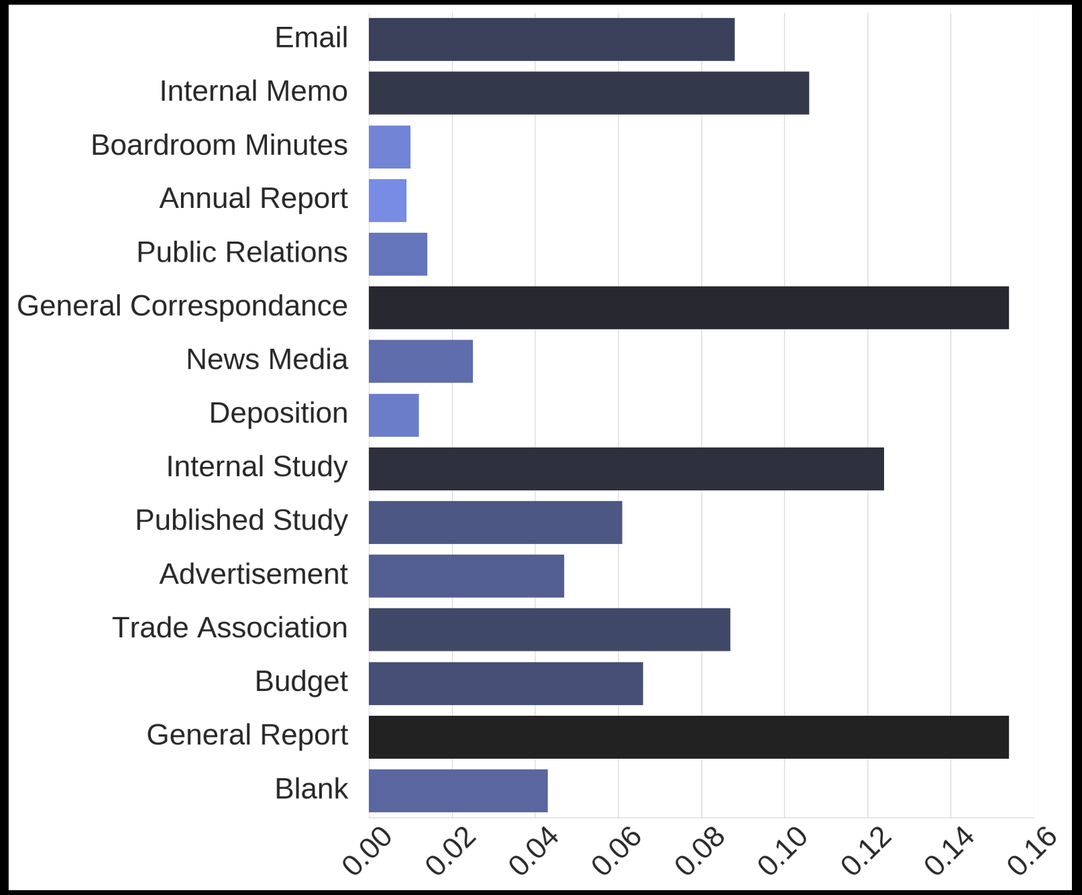
CMA 012526


SVM Learning Curve
Test Error
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
600
700
800
Training Examples

Huge document stash

1k manually labeled

15 “genres”/categories



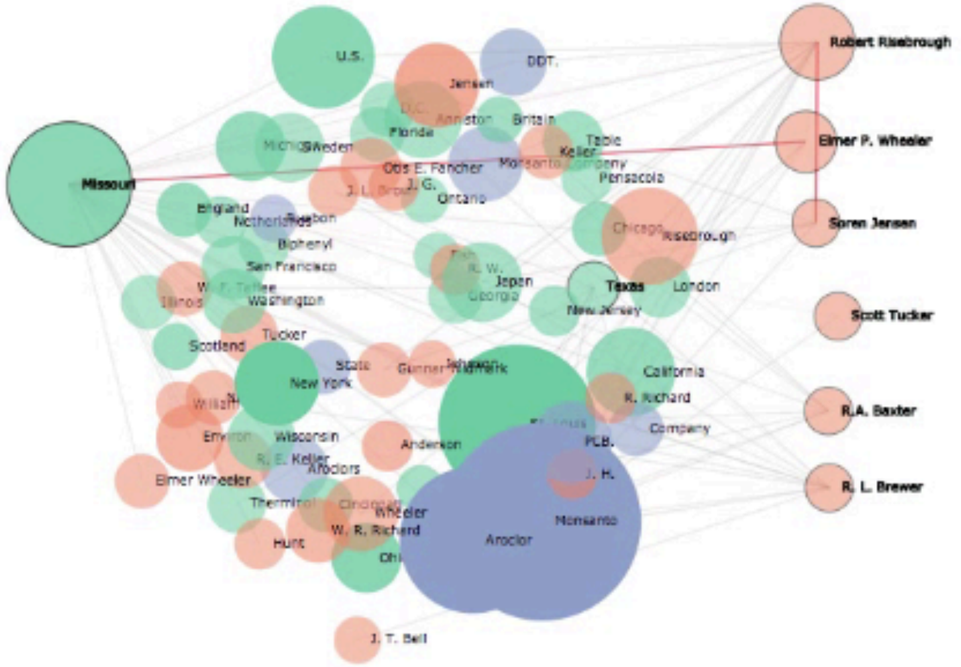
new topic

Threads in Forum : The Foyer — Forum Tools

	Thread / Thread Starter	Rating	Last Post	Replies	Views
Sticky Threads					
	Let's not bicker and argue about who killed who, ... Bernard		07-08-2015 09:30 AM by Bernard	4	15,359
	Read this before posting! Bernard		03-19-2005 05:13 PM by Bernard	0	19,476
Normal Threads					
	Hello everyone! Travis 71		Today 12:40 AM by Travis 71	0	1
	Wondering about keppra.... moldy73		04-26-2017 03:26 PM by Cint	3	87
	tongue spasms at night whisper		04-26-2017 12:16 PM by whisper	9	634
	My history... smitty1670		04-26-2017 12:14 PM by Sabbo	4	132
	Decision to concieve at 36 being diagnosed with Nocturnal Epilepsy.... Jnash23		04-26-2017 10:14 AM by Sabbo	5	119
	Hi! Everyone CathyW		04-25-2017 12:10 AM by acshuman	4	94
	Meds Increase smitty1670		04-24-2017 09:18 AM by Jeanbean	5	146

Home | **Relationship Visualization** | About | Contact

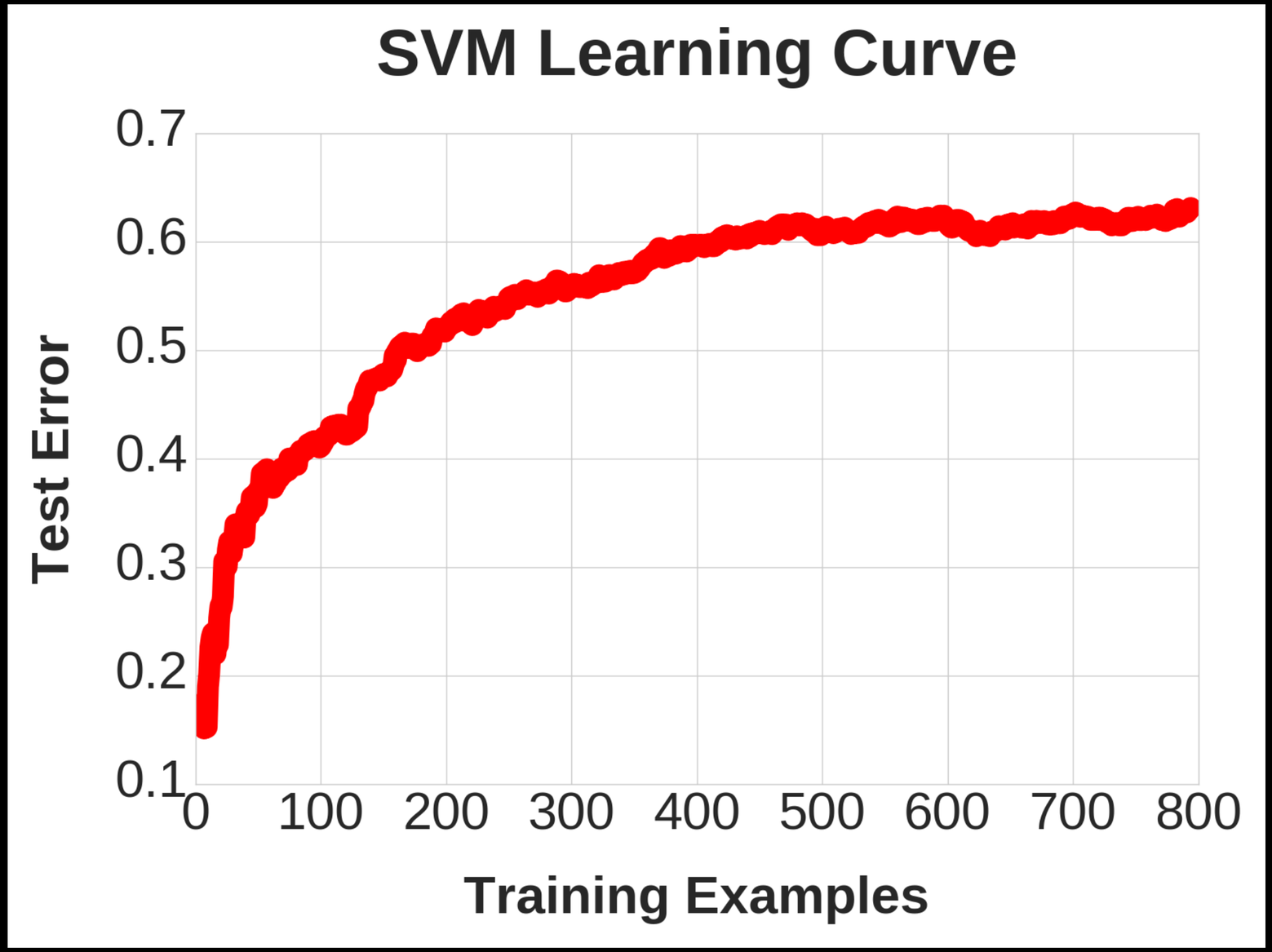
Entities found:



List of Topics

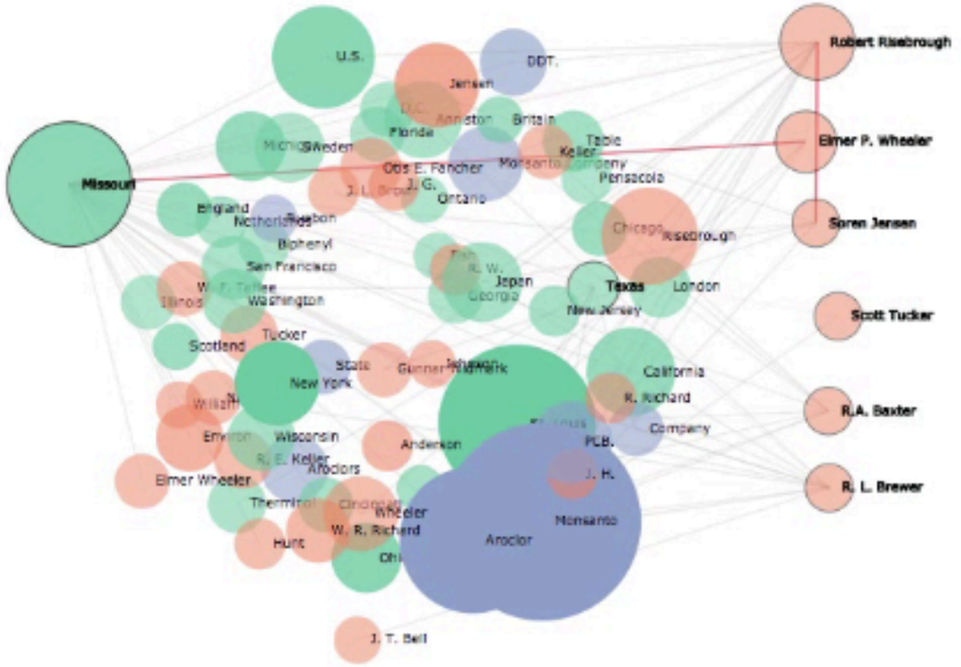
1. jobs economy american ve businesses work tax americans year make ECONOMIC ISSUES
2. isil people syria military terrorist attacks united terrorists american ve ISIS/SYRIA
3. ve people make president question don obama lot things good JUNK
4. america people nation country americans history years page time american AMERICAN VALUES
5. ve work country make back working page day home great ECONOMIC ISSUES
6. iran nuclear international russia united world deal states weapons sanctions WEAPONS: RUSSIA/SYRIA
7. united states people countries world africa change democracy progress global DEMOCRACY ABROAD
8. war security america iraq afghanistan troops ve forces states support IRAQ/AFGHANISTAN
9. world people rights peace nations freedom future human israel copyright AMERICAN ALLIES
10. veterans service nation men american families honor serve lost military VETERANS
11. law information american court justice national process intelligence issues case JUNK
12. president obama people ve country political made page question continue JUNK

```
"contributors": null,
    "coordinates": null,
    "created_at": "Tue Aug 23 12:56:46 +0000 2016",
    "entities": {
        "hashtags": [],
        "symbols": [],
        "urls": [],
        "user_mentions": [
            {
                "id": 1367531,
                "id_str": "1367531",
                "indices": [
                    72,
                    80
                ],
                "name": "Fox News",
                "screen_name": "FoxNews"
            }
        ]
    },
    "favorite_count": 7826,
    "favorited": false,
    "geo": null,
    "id": 768069472464666624,
    "id_str": "768069472464666624",
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "is_quote_status": false,
    "lang": "en",
    "place": null,
    "retweet_count": 2547,
    "retweeted": false,
    "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "text": "I am now in Texas doing a big fundraiser for the Republican Party and a @FoxNews Special on the BORDER and with victims of border crime!",
    "truncated": false,
    "user": {
        "contributors_enabled": false,
        "created_at": "Wed Mar 18 13:46:38 +0000 2009",
        "default_profile": false,
        "default_profile_image": false,
        "description": "#TrumpPence16",
        "entities": {
            "description": {
                "urls": []
            },
            "url": {
                "urls": [
                    {
                        "display_url": "DonaldJTrump.com",
                        "expanded_url": "http://www.DonaldJTrump.com",
                        "indices": [
                            0,
                            23
                        ],
                        "url": "https://t.co/mZB2hymxC9"
                    }
                ]
```


SVM Learning Curve
Test Error
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
600
700
800
Training Examples

Home | **Relationship Visualization** | About | Contact

Entities found:



OCR Accuracy

rate determination of a test oompound during a oyole
were as followB' A 20-1111. sample of the mixed liquor
(aotivated sludge + liquor) was withdrawn one hour after
feeding and at the end or the aeratlon period. The sample of
miXed liquor was then extraoted and the extraot concentrated
acoording to the procedure given in Analytlcal ChemiatZ7 Method
71-18. The ooncentration of test compounds was ~hen determined
either by an ultraviolet (UV) spectroscop10 or ~otron-capture
gaa chromatolraphic (EC-OC) method. Details of 1e UV methods
are given in Analytlcal Chem1atry Method 71-17 ~.~d the EC-OC
methods In Analytical Chea1stry Method 71-35. The disappearance

rate was oaloulated from the followlng equat10n: