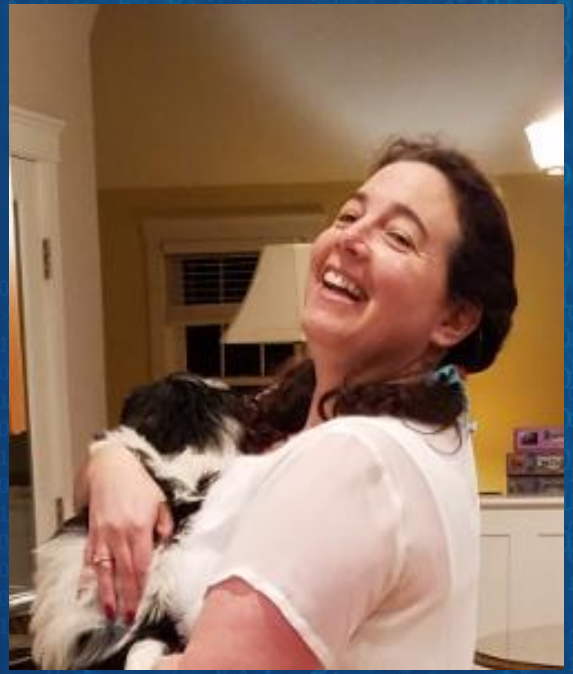
AlignConsulting

It’s 5 o’clock. Do you know where your info risks are?

Kate Pugh, Columbia, AlignConsulting
Sandra Serkes, Columbia, Valora Technologies

KM World November 5, 2019

We love love our furry friends...but we *really* love metadata



Sandra Serkes, CEO, Valora Technologies
Guest Lecturer, Columbia University Information and Knowledge Strategy Master's Program

Katrina Pugh, CEO, AlignConsulting
Faculty and former Academic Director, Columbia University Information and Knowledge Strategy Master's Program

What is content bloat costing us?

1 hour, 41 minutes/week spent searching (average worker)

13.2 billion hours across all US workers

$628 billion opportunity cost for the US

Assumes 2000 hr/year/worker, 158M workers, $23.63/hr for wages and overhead, 4.2% lost productivity searching for content. Sources: Nucleus Research “Content Bloat Drains Productivity by 8 percent,” June, 2016. Trading Economics US hourly wages. https://tradingeconomics.com/united-states/wages

What could info bloat cost us?

Breach of customer data

Proposed fine: **£183M**

Failing to protect customer data

Proposed fine: **£99M**

GDPR introduced hefty penalties per instance:

up to **€20M** or **4% of global turnover** – *whichever is higher*

Why else do we “take control of” unstructured data?

Compliance:
Retention, Legal Hold, GDPR, Privacy, Security

Identification:
ROT, Dupes, PII, DocType, Relevance, Privilege, Custodian, Date, IP potential and more

Disposition:
Quarantine, Defensible Deletion, Redaction, Migration, Remediation

Do any of these look like you?

Large Oil & Gas Manufacturer

- Frequent acquisitions of other companies
- Millions of documents
- Not enough SMEs
- Risky content..Legal hold…Data Privacy… ROIT

Multi-store, multi-country fast-food

- Independent, franchise & retail sites
- Managers autonomously hire, fire
- GDPR concerns

Info hound (bankruptcy data vault)

- Out run by content volume
- Regulatory reporting obligations
- Sensitive data
- Missing value-added revenue…?

Laying the groundwork for today

1. Metadata: Frame	2. Autoclassification: Analyze and Design	3. Info Gov: Control
Metadata tells us: ▪ What it is (e.g., contract, whitepaper) ▪ What it is about (e.g. topics, insights) ▪ Inherited labels (e.g., file type, date, author, last modified date) ▪ Impact, risk	Unitize, translate, transform, analyze, assess, triage groups of documents, based on metadata and context	Structures, policies, procedures, processes and controls implemented to help manage information at an enterprise level over time
Designing, contextualizing	*Processing, tagging*	*Planning for future*

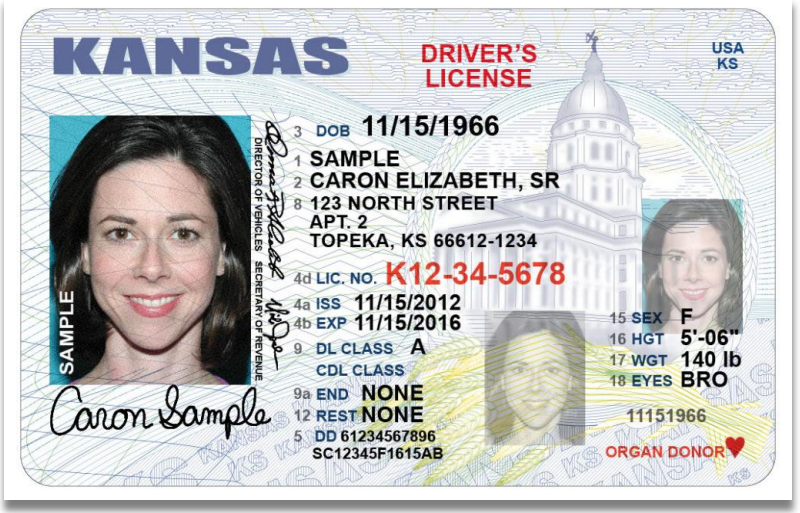
1. Metadata: Framing

Rich Metadata for Enterprise Content

What attributes & tags make sense for your organization?

+ “hidden” metadata

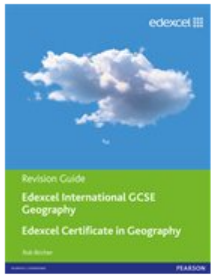
Metadata is “portable”



But what about similarly rich metadata (tags) for our enterprise content?

Isn’t rich metadata there…or can’t we just add it?

- Sort of..
  - Rich metadata tends to be common with structured data (databases), and uncommon with unstructured data (email, slide decks, reports, etc.)
  - Rich metadata is typically input by hand (via forms & wizards) and suffers from negligence, inconsistency, laziness, error, etc.

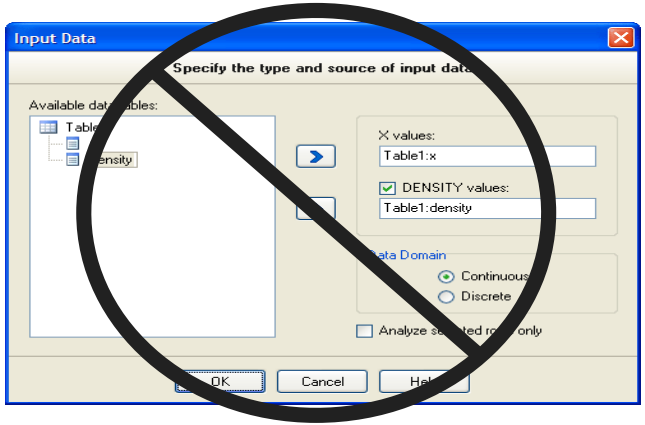


See larger version of cover

Price	£7.99 + £0.16 UK VAT
ISBN	9781446905777
Publication Date	March 2013
Format	BOOK

Jump to:

↓ Customer reviews



What does rich metadata get you?

Ability to know what you have, where it is, and how to manage it

- Lifecycle management tags
- Protected status tags (legal hold, privacy, access, etc.)
- Context for searches & decision making

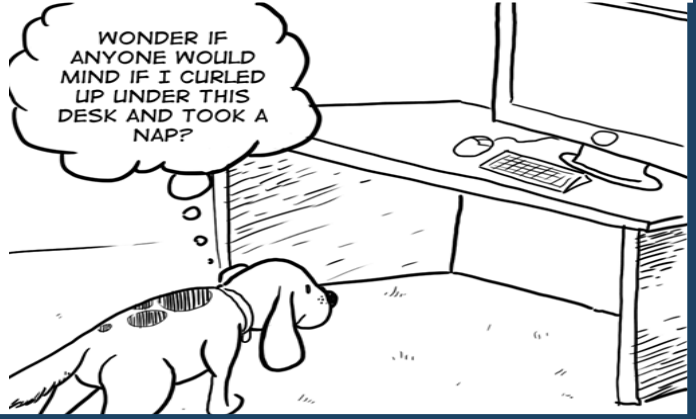
Ability to see across data repositories

- Normalized tags
- Data visualization, presentation and reporting

Ability to operate by facet

- Manage by content, not by time or other "simplistic" approaches

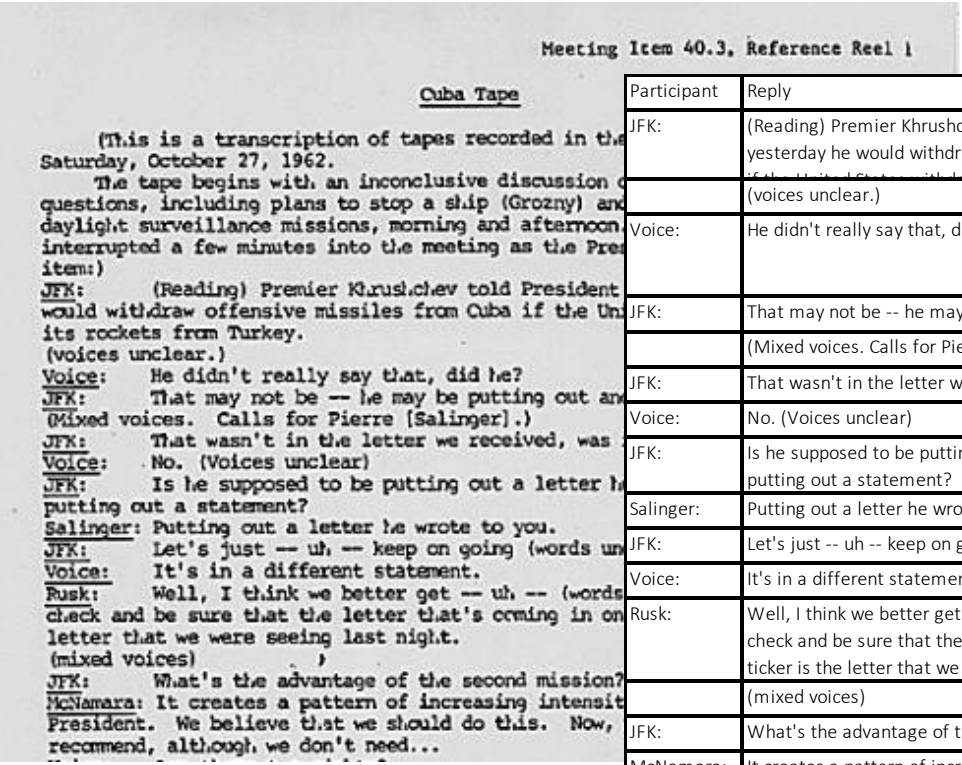
“Wonder if anyone would mind...?”



Context matters as we define metadata!

The slow way: Hand classifying Cuban Missile Crisis transcript

Ingestion (OCR) - Automated



Meeting Item 40.3, Reference Reel 1

Cuba Tape

(This is a transcription of tapes recorded in the
Saturday, October 27, 1962.
The tape begins with an inconclusive discussion
questions, including plans to stop a ship (Grozny) and
daylight surveillance missions, morning and afternoon.
interrupted a few minutes into the meeting as the Pres
item:)
JFK: (Reading) Premier Khrushchev told President
would withdraw offensive missiles from Cuba if the Un
its rockets from Turkey.
(voices unclear.)
Voice: He didn't really say that, did he?
JFK: That may not be -- he may be putting out an
(Mixed voices. Calls for Pierre [Salinger].)
JFK: That wasn't in the letter we received, was
Voice: No. (Voices unclear)
JFK: Is he supposed to be putting out a letter h
putting out a statement?
Salinger: Putting out a letter he wrote to you.
JFK: Let's just -- uh, -- keep on going (words un
Voice: It's in a different statement.
Rusk: Well, I think we better get -- uh, -- (words
check and be sure that the letter that's coming in on
letter that we were seeing last night.
(mixed voices)
JFK: What's the advantage of the second mission?
McNamara: It creates a pattern of increasing intensit
President. We believe that we should do this. Now,
recommend, although we don't need...

https://catalog.archives.gov/id/193723

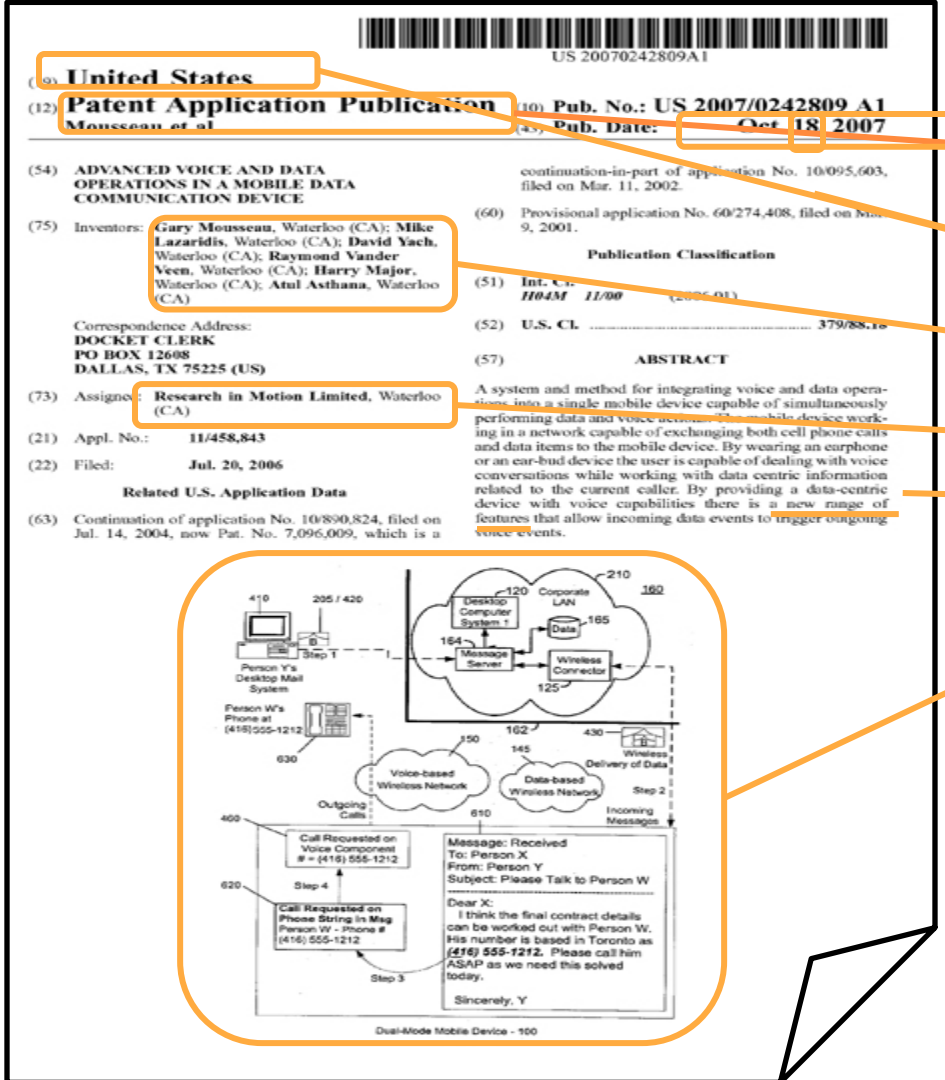
Coding – 1 Hour/page (= 84 hours)

Participant	Reply	Coding
JFK:	(Reading) Premier Khrushchev told President Kennedy yesterday he would withdraw offensive missiles from Cuba if the United States with	[Reading]
	(voices unclear.)	
Voice:	He didn't really say that, did he?	*Anti-Integrity*
JFK:	That may not be -- he may be puttir	
	(Mixed voices. Calls for Pierre [Salin	
JFK:	That wasn't in the letter we received	
Voice:	No. (Voices unclear)	
JFK:	Is he supposed to be putting out a le putting out a statement?	
Salinger:	Putting out a letter he wrote to you.	
JFK:	Let's just -- uh -- keep on going (wor	
Voice:	It's in a different statement.	
Rusk:	Well, I think we better get -- uh -- (V check and be sure that the letter tha ticker is the letter that we were see	
	(mixed voices)	
JFK:	What's the advantage of the second	
McNamara:	It creates a pattern of increasing int Mr. President. We believe that we s	

Preliminary analytics (= 1 hr. Messy, not repeatable)

Row Labels	(Irritated)	(Note: in this example there is no question mark.)	*anti-courtesy*	*Anti-courtesy* (interrupting)	*anti-integrity*	*Anti-Integrity* (this is a statement in disguise.)	*Courtesy*	*courtesy* (Agreeing, compliment)	*Inclusion* (clarification)	*Integrity*	*Integrity Q*	*Integrity-Q"	*Integrity-on*	*Translati	[Reading]	"anti-courtesy" (interruption)	"Anti Integ
Ball:										4							
Bundy:			1				1			12		4					
Bundy:										1							
Bundy: ...														1			
Bundy: (underlined)			1														
Dillion:										1							
Dillon:											1						
JFK:										1		2					
JFK:			1				1	1		17		4		4			
JFK:			1							2		1					
Low Voice:																1	
McCone:										1							
McNamara:			1							16							
Nitze:										5				1			
Participant																	
RFK:										9		1		1			
Rusk:			1							10		3					
Rusk:			1							2							
Salinge										1							

AutoClassifying an attachment (patent application)



US 20070242809A1

(19) United States

(12) Patent Application Publication

Mousseau et al.

(10) Pub. No.: US 2007/0242809 A1

(43) Pub. Date: Oct. 18, 2007

(54) ADVANCED VOICE AND DATA OPERATIONS IN A MOBILE DATA COMMUNICATION DEVICE

(75) Inventors: Gary Mousseau, Waterloo (CA); Mike Lazaridis, Waterloo (CA); David Yach, Waterloo (CA); Raymond Vander Veen, Waterloo (CA); Harry Major, Waterloo (CA); Atul Asthana, Waterloo (CA)

Correspondence Address:
DOCKET CLERK
PO BOX 12608
DALLAS, TX 75225 (US)

(73) Assignee: Research in Motion Limited, Waterloo (CA)

(21) Appl. No.: 11/458,843

(22) Filed: Jul. 20, 2006

Related U.S. Application Data

(63) Continuation of application No. 10/890,824, filed on Jul. 14, 2004, now Pat. No. 7,096,009, which is a continuation-in-part of application No. 10/095,603, filed on Mar. 11, 2002.

(60) Provisional application No. 60/274,408, filed on Mar. 9, 2001.

Publication Classification

(51) Int. Cl.
H04M 11/00 (2006.01)

(52) U.S. Cl. 379/88.18

(57) ABSTRACT

A system and method for integrating voice and data operations into a single mobile device capable of simultaneously performing data and voice operations. The mobile device working in a network capable of exchanging both cell phone calls and data items to the mobile device. By wearing an earphone or an ear-bud device the user is capable of dealing with voice conversations while working with data centric information related to the current caller. By providing a data-centric device with voice capabilities there is a new range of features that allow incoming data events to trigger outgoing voice events.

DocType = Patent Application

Date = 10/18/2007

Date Format = US

Author = Patent Authors, Author City, Author Country

Assignee = RIM

Tone = Neutral to slightly positive

Embedded Graphic with Title

Other Data Capturable Data Elements:

- Patent Number
- Filing Date
- Key Phrases & Terms
- Managing PTO
- Implied/Attached Docs
- Bar Code Present
- And many more . . .

Implied status:
Responsive, Nonprivileged

2. Auto-Classification: Analyze and Design

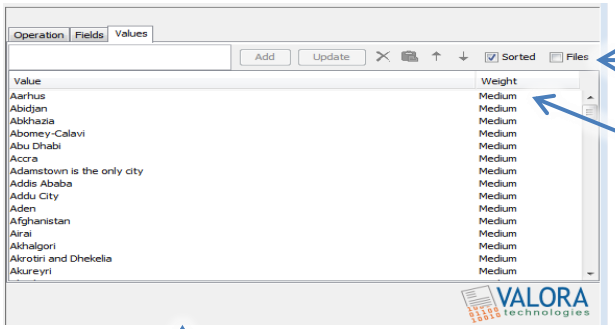
What is the secret sauce of auto-classification design?

What are Rules?

- Software algorithms that determine the disposition of content
- Typically follow an IF-THEN format, often nested
  - Ex: IF the *DocumentType* = Contract AND the *Date* < 1/1/2020, THEN Mark File as ACTIVE (Retain)

So, if you could do any of these rules, which ones should you do first? [Hint: could you do multiple *simultaneously*?]

Rules Engine illustration



Support for values *files*

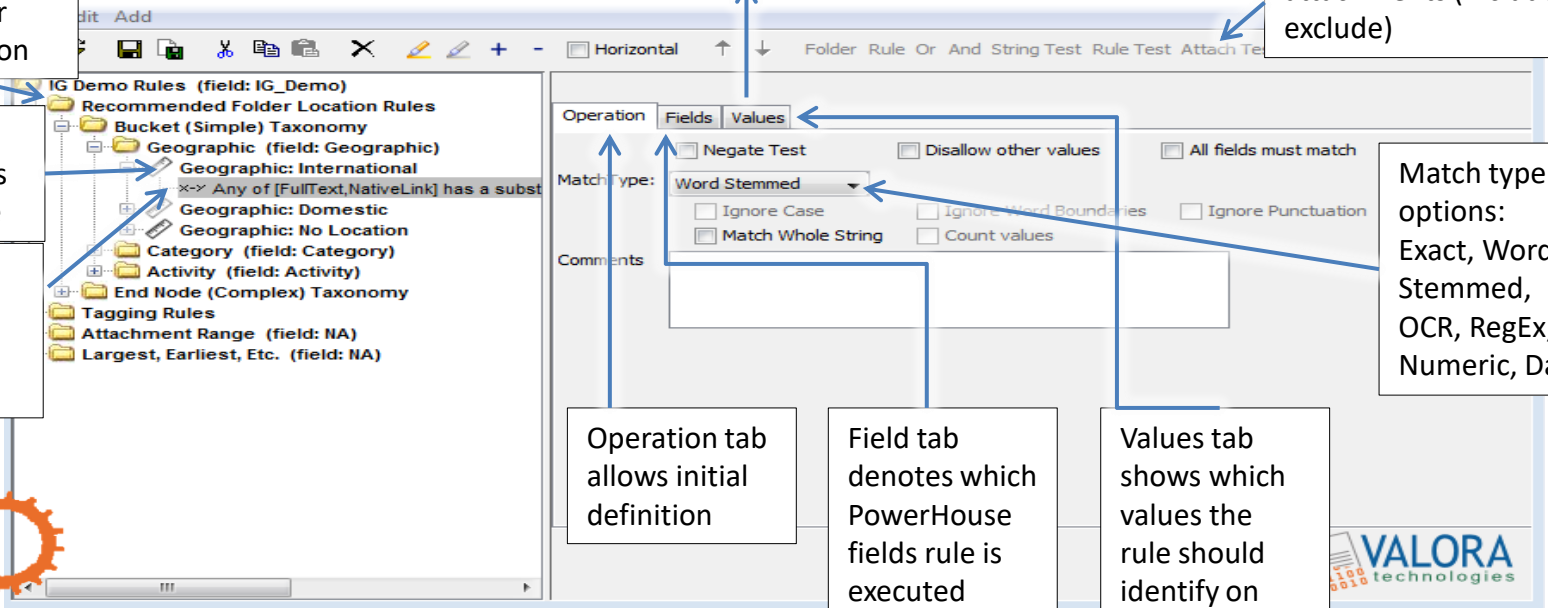
Weights & confidences create hierarchical decisioning options

Ability to test on family ranges and attachments (include & exclude)

Folders for organization

Ruler represents single rule

String test (can have multiple per rule)



Match type options: Exact, Word Stemmed, OCR, RegEx, Numeric, Date

Operation tab allows initial definition

Field tab denotes which PowerHouse fields rule is executed against

Values tab shows which values the rule should identify on

PowerHouse Rules Editor

Data Privacy

Identifying & AutoClassifying PII

Uniform Residential Loan Application

This application is designed to be completed by the Applicant(s) with the Lender's assistance. Applicants should complete this form as "Borrower" or "Co-Borrower" as applicable. Co-borrower information must also be provided (and the appropriate box checked) when [X] the income or assets of a person other than the Borrower (including the Borrower's spouse) will be used as a basis for loan qualification or [ ] the income or assets of the Borrower's Spouse will not be used as a basis for loan qualification, but his or her liabilities must be considered because the Borrower resides in a community property state, or the security property is located in a community property state, or the Borrower is relying on other property located in a community property state as a basis for repayment of the loan.

If this is an application for joint credit, Borrower and Co-Borrower each agree that we intend to apply for joint credit (sign below):

Borrower | Co-Borrower | Interest Only

I. TYPE OF MORTGAGE AND TERMS OF LOAN

Mortgage Applied for: [ ] V.A. [ ] Conventional [ ] Other (explain): [X] FHA [ ] USDA/Rural Housing Service

Agency Case Number | Lender Case Number

Amount: $ 625,000.00 | Interest Rate: 6.5% | No. of Months: 360

Amortization Type: [X] Fixed Rate [ ] GPM [ ] Other (explain): [ ] ARM (type):

II. PROPERTY INFORMATION AND PURPOSE OF LOAN

Subject Property Address (street, city, state, ZIP): 5421 Talahassee Place , Perry FL 32347 | No. of Units

Legal Description of Subject Property (attach description, if necessary) | Year Built: 1984

Purpose of Loan: [ ] Purchase [X] Construction [ ] Other (explain): [ ] Refinance [ ] Construction-permanent

Property will be: [X] Primary Residence [ ] Secondary Residence [ ] Investment

Complete this line if construction or construction-permanent loan

Year Lot Acquired | Original Cost | Amount Existing Liens | (a) Present Value of Lot | (b) Cost of Improvements | Total (a + b)

Complete this line if this is a refinance loan

Year Acquired: 1996 | Original Cost: $ 375,000 | Amount Existing Liens: $ 402,500 | Purpose of Refinance: Cash-Out/Home Improvement | Describe Improvements: [ ] made [X] to be made: Add-on two-car garage Cost: $20,000

Title will be held in what Name(s): Brady and Wilma Schleshenger | Manner in which Title will be held: Tenants in common | Estate will be held in: [X] Fee Simple [ ] Leasehold (show expiration date)

Source of Down Payment, Settlement Charges and/or Subordinate Financing (explain)

III. BORROWER INFORMATION

Borrower	Co-Borrower
Borrower's Name (include Jr. or Sr. if applicable): Brady G. Schleshenger	Co-Borrower's Name (include Jr. or Sr. if applicable): Wilma B. Schleshenger
Social Security Number: 562-55-5595	Home Phone (incl. area code): (956) 986-9985
DOB (MM/DD/YYYY): 04/11/1969	Yrs. School: 2
Social Security Number: 548-32-9855	Home Phone (incl. area code): (956) 986-9985
DOB (MM/DD/YYYY): 02/16/1978	Yrs. School: 15
[X] Married [ ] Separated [ ] Unmarried (include single, divorced, widowed)	Dependents (not listed by Co-Borrower): no. 1 ages 17
[X] Married [ ] Separated [ ] Unmarried (include single, divorced, widowed)	Dependents (not listed by Borrower): no. ages

Clear PII: SSN

Likely PII: ("warning sign")

Clear PII: Home Phone Number

Not PII: Interest Rate

Implied classification: Active PII, needs protection & redaction

3. Information Governance: Control

The information governance lifecycle

New content

Processing

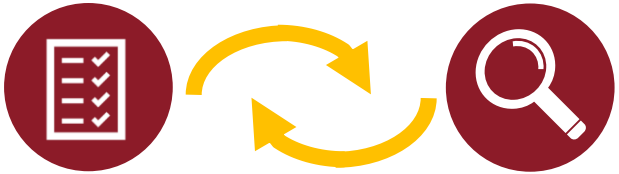
Ingests data or processes files in place

- Creating OCR for scanned images
- Transcriing voice and video
- Extracting text for native files & email
- Speech to text for audio/video files
- Translating content to English
- Re-ordering or re-aligning pages
- Applying redactions
- Unitization

Tagging

Extracts key Info& attributes about each document (aka Coding, Indexing)

- Document Type, Important Dates
- Key Names & Phrases
- Topics, Keywords & Themes
- File, Content and DocType attributes
- Relation to other documents (duplicate, related, attached, contradictory, etc.)



Disposition

(Rules) create a destination, action or status for each document

- Retention status & duration
- Privacy & security protections
- Folder (taxonomy) location
- Labelling & keywords display
- Activity notification
- Legal Hold

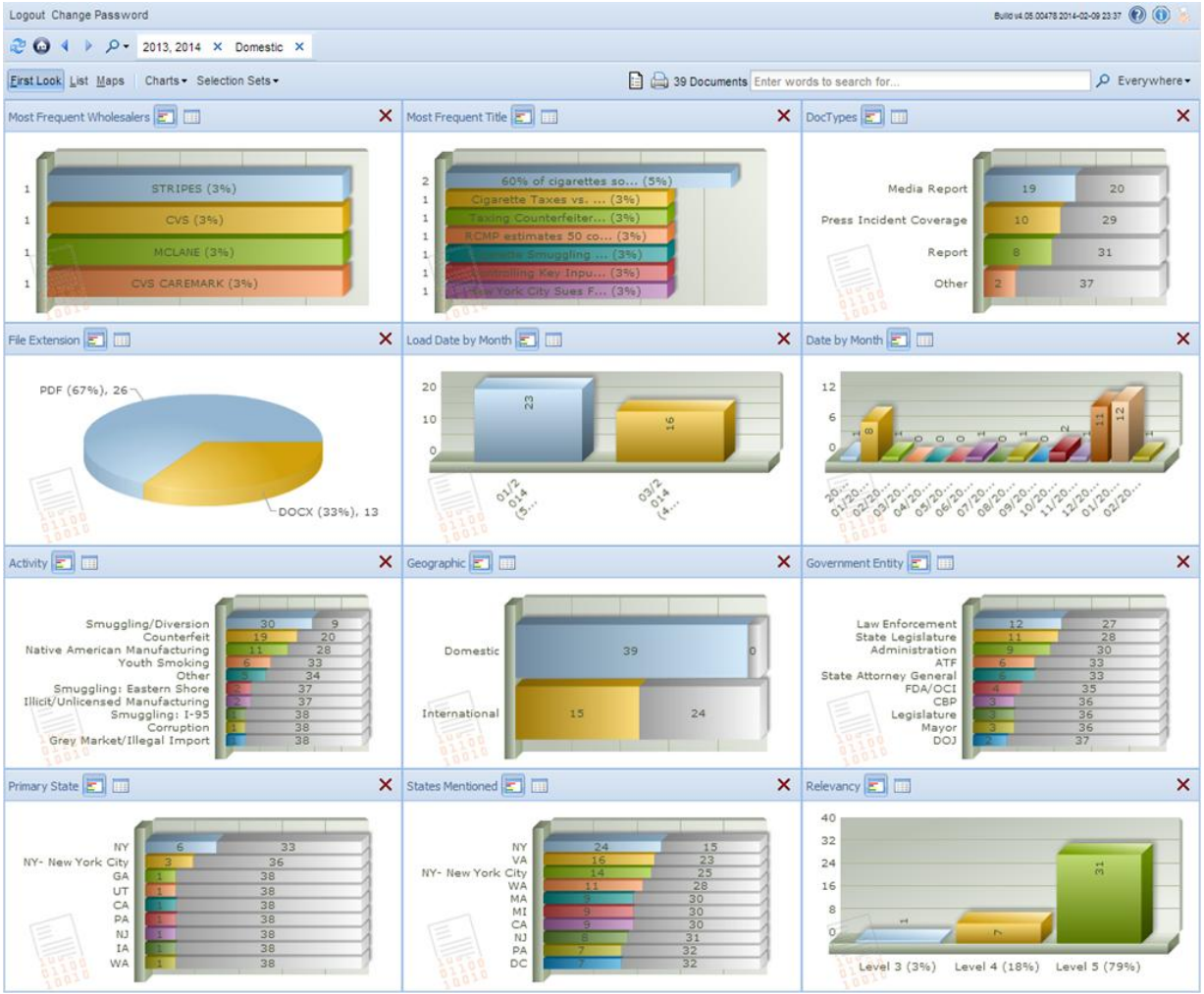
Business Assess't

AI and Machine learning help assess risks to reputation, ID new IP

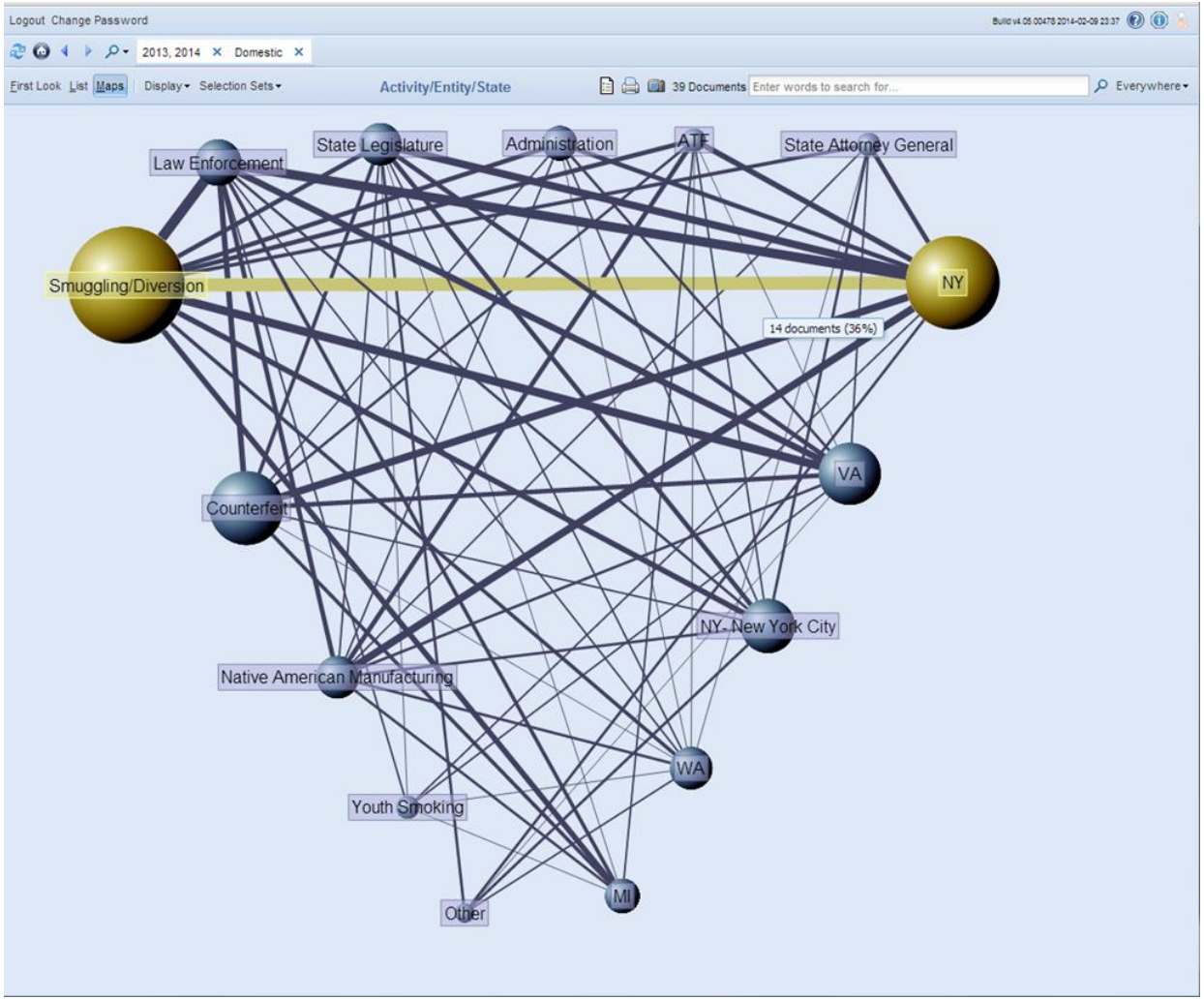
- Relevancy/ prioritization
- Contract inconsistencies
- Conflicts of interest
- New IP
- New Product or feature

Old content

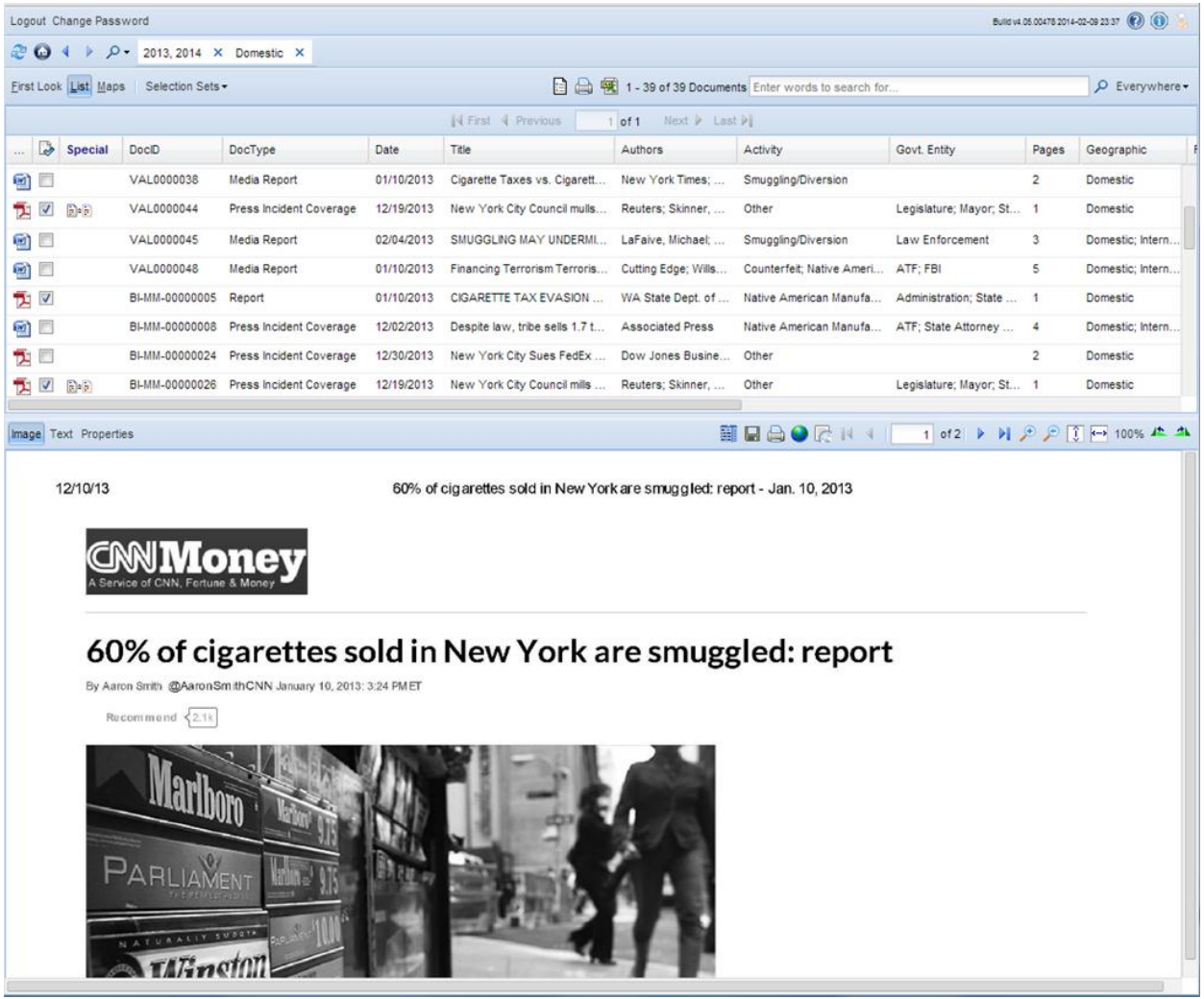
Data Visualization of Auto Classification (Data dashboard)



Data Visualization of Auto Classification (Graph Theory) (cont.)

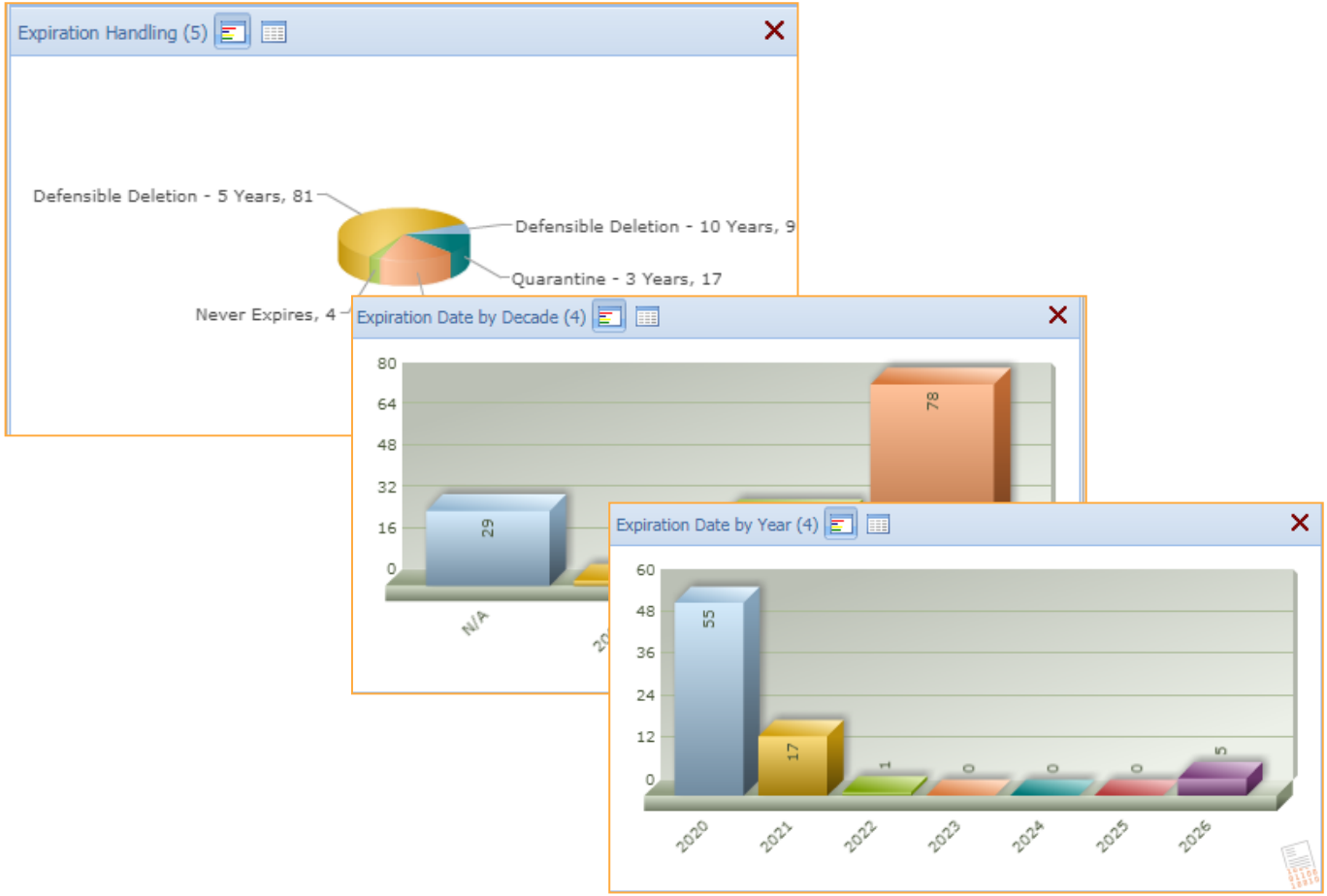


Data Visualization of Auto Classification



Retention

Automated Retention Handling



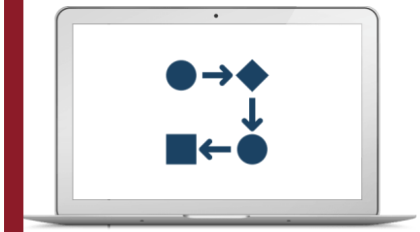
Automated Retention Handling (cont.)

...	Special	DocID	DocType	Date	Expiration Date	Expiration Handling	Title
		IG-00028283	Purchase Document			Policy Review	CLOSING STATEMENT
		IG-00028284	Press Incident Cov...	03/09/2015	03/09/2020	Defensible Deletion - 5 Years	Counterfeit auto parts are threat to auto safety
		IG-00028287	Press Incident Cov...	07/24/2015	07/24/2020	Defensible Deletion - 5 Years	Counterfeit LED TVs from China confiscated at Miami Seaport
		IG-00028289	Press Incident Cov...	08/23/2016	08/23/2021	Defensible Deletion - 5 Years	Counterfeit pain pills likely came to Prince illegally
		IG-00028293	Press Incident Cov...	06/25/2015	06/25/2020	Defensible Deletion - 5 Years	Counterfeit products drive fashion industry to improve, study suggests
		IG-00028294	Report	03/2016		Policy Review	UNREGULATED E-WASTE EXPORTS FUEL COUNTERFEIT ELECT...
		IG-00028298	Purchase Document	04/30/2015	04/30/2018	Quarantine - 3 Years	Your Credit Card Account Statement
		IG-00028299	Press Incident Cov...	05/11/2015	05/11/2020	Defensible Deletion - 5 Years	Does Etsy Have A Problem With Fake Goods?
		IG-00028301	Report	03/2014		Policy Review	ARREST SUMMARY REPORT - January - February - March 2014
		IG-00028304	Civil Court Docume...	06/13/2014		Never Expires	UNRESOLVED ISSUE
		IG-00028314	Press Incident Cov...	07/30/2015	07/30/2020	Defensible Deletion - 5 Years	Etsy Investors Sue for Fraud over Risk of Counterfeit Goods Etsy Inc. ...
		IG-00028308	Report	03/21/2014		Policy Review	EXECUTIVE SUMMARY

“State of the State” of Artificial Intelligence in auto-classification

Reactive machines

Does not use past experiences to make decisions



Limited memory

Learns by monitoring actions & behaviours over time

Theory of mind

Has thoughts and emotions that affect their own behavior

Self-awareness

Has consciousness, empathy and a sense of being

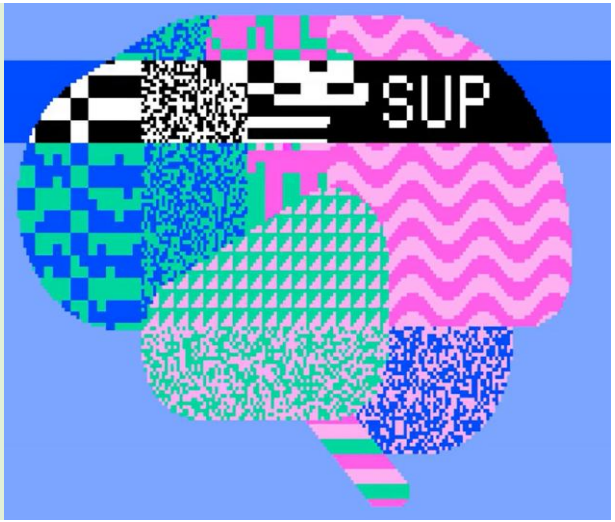
Processing: Full spectrum of human communication

Scientists have found a way to decode brain signals into speech

It's a step towards a system that would let people send texts straight from their brains.

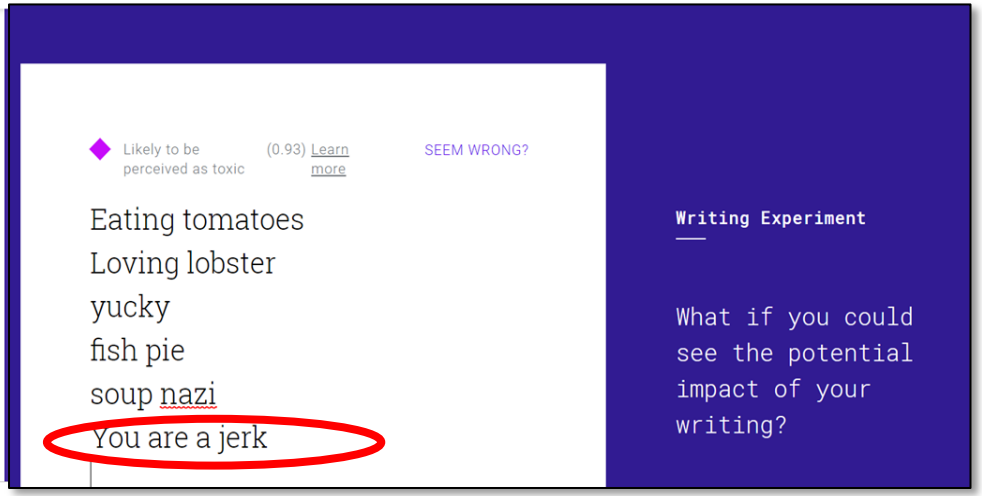
by Antonio Regalado

MIT Technology Review, April 24, 2019



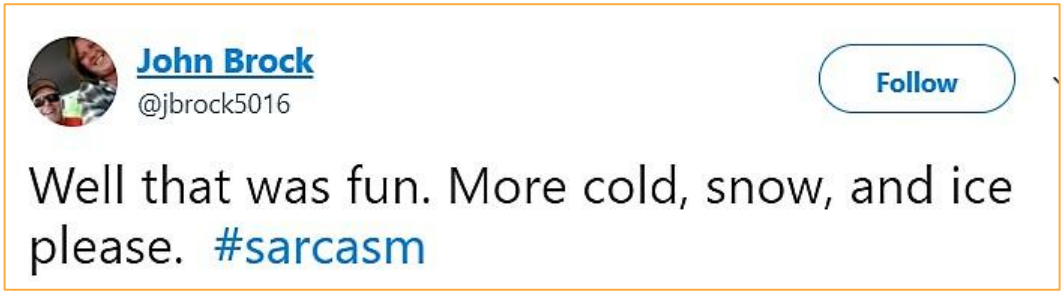
Tagging: AI recognizes tone, slang

- Google Perspective API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give real-time feedback to commenters or help moderators



https://www.perspectiveapi.com/#/

Tagging: Intent recognition example: Sarcasm



Humans don't always recognize it! It requires context (speaker, situation, world)

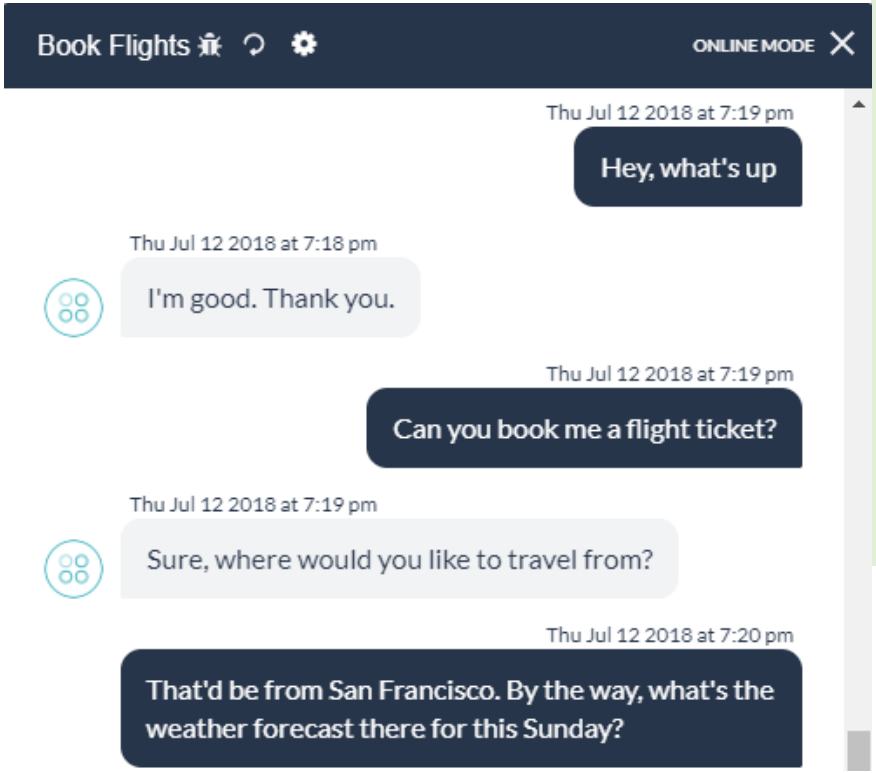
Approaches:

1. Machine learning (volume, then phrase matching / information retrieval)
2. Rules (e.g., unexpected juxtaposition, exaggeration)
3. Deep learning

Check out Indian Institute of Technology research Sarcasm Suite

Source: Quartz (2016) and NVDia

Tagging: Context switching



"How do you really carry context throughout a dialogue? This is the biggest challenge. A lot of how you understand what I'm saying depends on what I said maybe five sentences ago, or fifteen sentences ago. You're building up a state of the conversation."

Satinder Singh, Director of the Artificial Intelligence Lab at the University of Michigan.

Source: IBM

(Hold and resume concept (Kore.ai))

Source: Kore.ai

4. Case Studies

Recap: Do any of these look like you?

Large Oil & Gas Manufacturer

- Frequent acquisitions of other companies
- Millions of documents
- Not enough SMEs
- Risky content..Legal hold…Data Privacy… ROIT

Multi-store, multi-country fast-food

- Independent, franchise & retail sites
- Managers autonomously hire, fire
- GDPR concerns

Info hound (bankruptcy data vault)

- Out run by content volume
- Regulatory reporting obligations
- Sensitive data
- Missing value-added revenue…?

Case Study #1: Oil & Gas

Acquiring a company without most of the people. 200 GB of data and none of the tacit knowledge - who owns this? Should we keep it?	Legal Hold Retention Privacy/PII Contract anomalies Animal testing compliance	Custodian Matter PII/PHI Record Type Expiry Handling Dates	Retention ROT Data Privacy Classification Legal Hold Duplication	Contract consolidation Compliance watch Fileshare cleanup DSARs
Burning Platform	Scope	Critical Metadata	Analytics/ Intelligence	InfoGov Applications

“Don’t buy it before you know the info risk!”

Case Study #2: Multi-store/int'l fast food retailer

Burning Platform	Scope	Critical Metadata	Analytics/ Intelligence	InfoGov Applications
Data privacy concerns on the verge of moving to the cloud	~10K employees >100 stores ~20 repositories 80 terabytes of records Content analysis & migration Build an IG strategy	Custodian Document Type PII/PHI Record Type Expiry Handling Dates	Retention ROT Data Privacy Classification Rec File Loc. Duplication	Data privacy Fileshare cleanup DSARs IG maturity

"It's not about migration. It's about gov and access."

Case Study #3: Info Hound (bankruptcy data vault)

Too much data to assess, evaluate, prioritize in tight timelines	Federal, state, local bankruptcy court filings 1705 Juris-dictions 300 Formats 100K-200K cases/yr	Debtor Name & specifics Custom fields per DocType Jurisdictional Dates Financial	Bankruptcy terms violations? Defrauding regulators? Criminal checks? Redundancies? Expirations?	Compliance reporting Risk analysis
Burning Platform	Scope	Critical Metadata	Analytics/ Intelligence	InfoGov Applications

“The 2008 meltdown forced us all to rethink & retool.”

Three Cases' lessons learned

1. You **lay bare** facts and patterns ("It's like walking with all of the labels on your clothes on the outside")

2. Building **the rules engine** – how you label it, and what you do with it -- **takes longer** than you'd think. ("You learn when to dig deeper and when to back off.")

3. It's not just "Document Intelligence," or "eDiscovery." It's "**info governance**"

Discussion

Thanks for attending!

VALORA technologies

www.ValoraTech.com

info@ValoraTech.com

+1 781.229.2265

AlignConsulting

www.alignconsulting.com

katepugh@alum.mit.edu

+1 617-967-3910

Appendix

Rules for Protecting Consumer Data

Protecting PII is at the core of privacy regulations

Privacy Regulations

require:

- Do you know ***where*** it is?
- Do you know ***what*** it is?
- Is it ***sensitive***?
- Is it ***protected and secure***?
- Is it ***accessible*** when needed?

AutoClassification

allows you to:

- **Locate it:** email, shared drive, ECM
- **Identify it:** SSN, DOB, drivers license
- **Secure it:** set security access controls
- **Redact it:** grant permission levels
- **Action it:** DSAR, Right to be Forgotten

Rules for eDiscovery

Note: If you've got a good IG practice through AutoClassification, eDiscovery is much simpler, faster & lower cost

eDiscovery

requires us to:

- ? Look for content
- ? Filter through tons of data
- ? Relevant & irrelevant, Privileged & Non-priv
- ? Document review
- ? Document production

AutoClassification

you've already:

- ✓ **Located it:** across all data stores
- ✓ **Identified it:** for Early Data Assessment
- ✓ **Cleaned it:** only dealing with relevant content
- ✓ **Secured it:** who has access to what
- ✓ **Actioned it:** Retention, archived, etc.

Rules or Legal Hold

Note: You can still dynamically govern content under Legal Hold

Legal Hold

asks us:

- ? Which data? Whose data?
- ? For how long?
- ? What's relevant? What's not?
- ? Access for inside counsel? Outside counsel?
- ? What happens when Hold lifts?

AutoClassification

you've already:

- ✓ **Located it:** we haven't missed anything/where
- ✓ **Cleaned it:** already excluded irrelevant content
- ✓ **Identified type & custodian:** we know content, context & owner
- ✓ **Identified retention dates:** schedule already exists
- ✓ **Set it & forget it:** ongoing perpetual analysis