# Text Analytics Webinar

Tom Reamy
Chief Knowledge Architect
KAPS Group
http://www.kapsgroup.com
Author: Deep Text

# Agenda

- Introduction
  - The Book – Background
  - What is Text Analytics – Definition & Elements
  - Current State of Text Analytics
  - Value of Text Analytics & Obstacles
- Applications – Enterprise Search, Info Apps, Social
- Development - Approaches
- Getting Started with Text Analytics
- Questions / Discussions

# DEEP TEXT

USING TEXT ANALYTICS
to Conquer Information Overload,
Get Real Value from Social Media,
and Add Big(ger) Text to Big Data

TOM REAMY

# Introduction:
# Deep Text: The Book

- The only book on text analytics
- 5 sections, 3 chapters each
  - Text Analytics Basics
  - Getting Started in Text Analytics (Smart Start)
  - Text Analytics Development
  - Text Analytics Applications
  - ETA – Enterprise Text Analytics as a Platform
- A treasure trove of technical detail, likely to become a definitive source on text analytics. – Kirkus Reviews
- This book will give you all the answers and is the definitive book on the business possibilities of the technology**.** - Martin White

# Introduction:
# Deep Text: The Book – Who Am I?

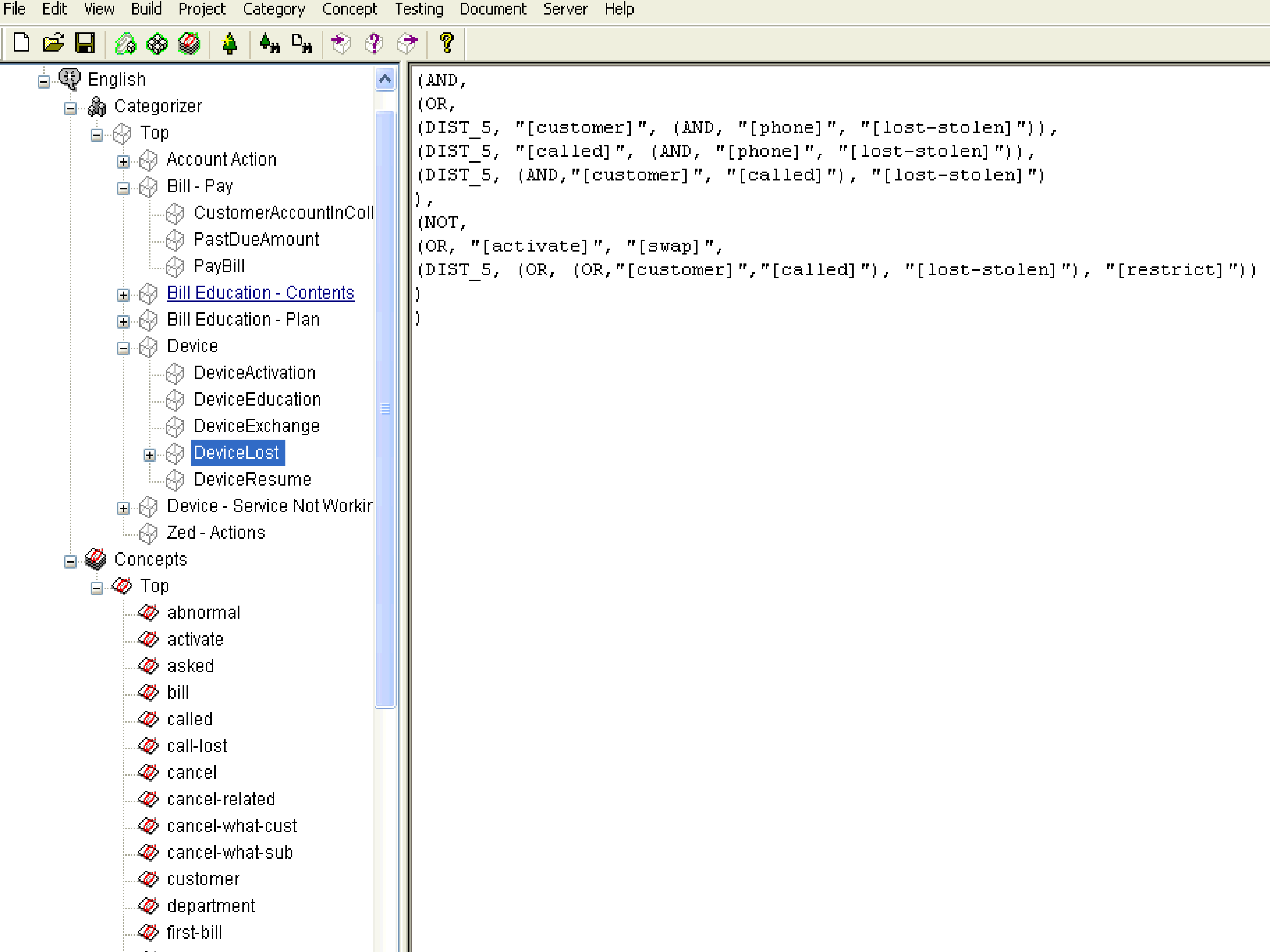
- Professional student / independent consultant – all but 6 years
- History of Ideas to Programmer – AI (Only 2 years away)
- Games – Galactic Gladiators/Adventures – still available
- KAPS Group – 13 years, Network of consultants
  - Taxonomy to text analytics
  - Consulting, development – platform and applications
  - Strategy, Smart Start, Search, Smart Social Media
  - Partners – SAS, IBM, Synaptica, Expert System, Smartlogic, etc.
  - Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, Dept. of Transportation, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com

# Introduction:
# What is Text Analytics?

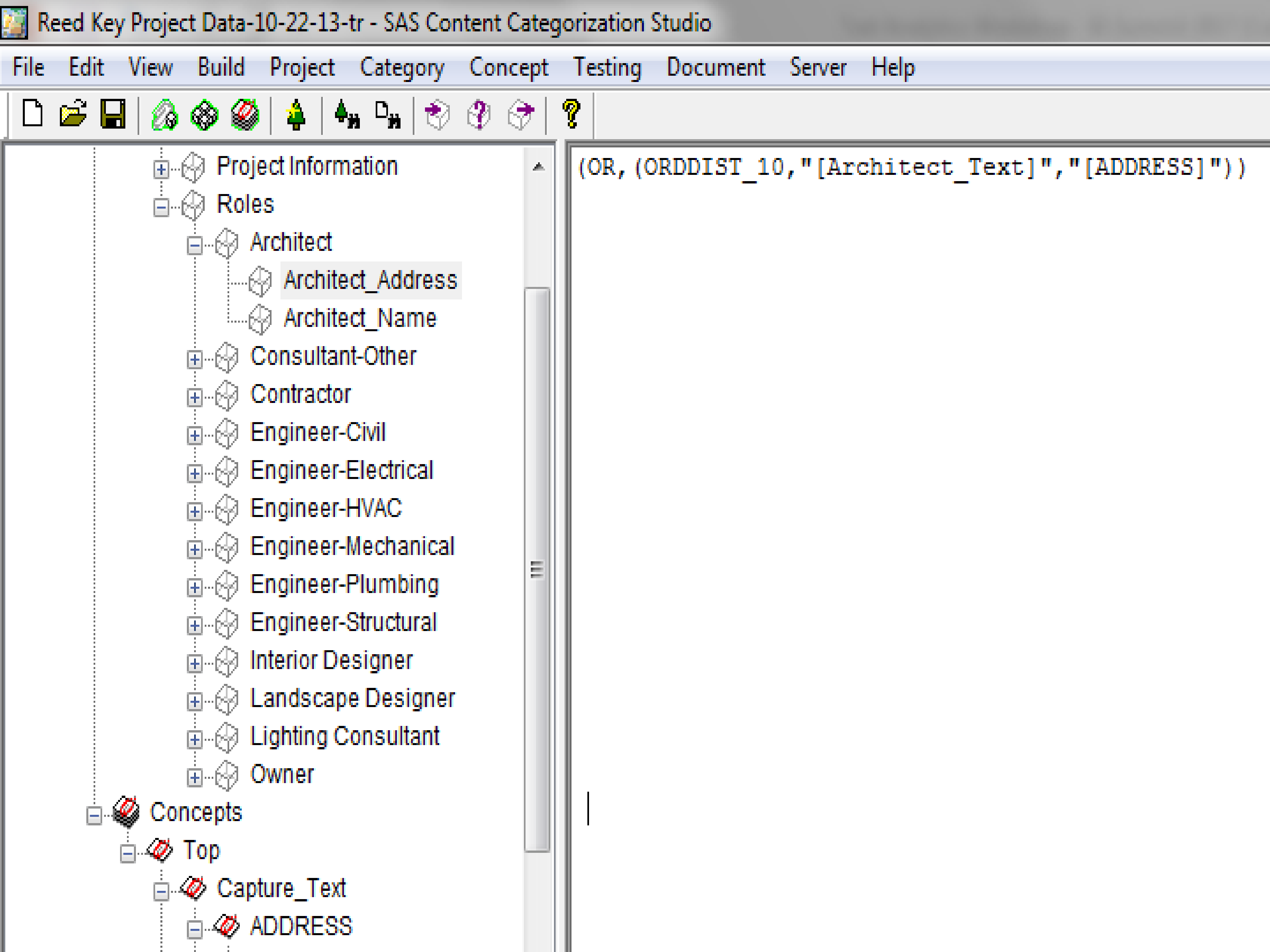
- Text analytics is the use of software and knowledge models to analyze and add structure to unstructured text.
- Text Mining – NLP, statistical, predictive, machine learning
    - Different skills, mind set, Math & data not language
- Annotation/Extraction – entities and facts – known and unknown, concepts, events - catalogs with variants, rule based
- Sentiment Analysis
    - Entities and sentiment words – statistics & rules
- Summarization
    - Dynamic – based on a search query term
    - Document – based on primary topics, position in document

# Introduction:
# What is Text Analytics?

- Auto-categorization = the brains of the outfit
- Training sets – Bayesian, Vector space
- Terms – literal strings, stemming, dictionary of related terms
- Rules – simple – position in text (Title, body, url)
- Boolean– Full search syntax – AND, OR, NOT
- Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE



```
(AND,
(OR,
(DIST_5, "[customer]", (AND, "[phone]", "[lost-stolen]")),
(DIST_5, "[called]", (AND, "[phone]", "[lost-stolen]")),
(DIST_5, (AND,"[customer]", "[called]"), "[lost-stolen]")
),
(NOT,
(OR, "[activate]", "[swap]",
(DIST_5, (OR, (OR,"[customer]","[called]"), "[lost-stolen]"), "[restrict]"))
)
)
```



```
(OR,(ORDDIST_10,"[Architect_Text]","[ADDRESS]"))
```

# Deep Text Webinar
# Introduction: Text Analytics

- History – Inxight - Moved TA from academic and NLP to enterprise - auto-categorization, entity extraction, and Search-Meta Data
- Shift to sentiment analysis - easier to do, obvious pay off
  - Backlash – Real business value?
- Current Market: 2016 – exceed $1 Bil for text analytics (10% of total Analytics)
- Growing 20% a year, search is 33% of total market
- Fragmented market place – full platform, social media, open source, taxonomy management, extraction & analytics, embedded in applications (BI, etc.), CM, Search
- No clear leader.

# Deep Text Webinar
## Benefits of Text Analytics

- What is the ROI of text analytics?
  - Wrong question?
  - What is ROI of organizing your company
- Benefits in 3 areas:
  - Search – IDC -20K per employee per year – Time & Quality
  - Social Media – understand what customers are saying
    - Lead generation, early warning, brand management
  - Multiple Info Apps
    - Range of applications – almost unlimited
- Selling the benefits – numbers, stories, need education

# Deep Text Webinar
# Primary Obstacle:  Complexity

- Usability of software is one element
- More important is difficulty of conceptual-document models
  - Language is easy to learn , hard to understand and model
- Need to add more intelligence (semantic resources) and ways for the system to learn – social feedback
- Customization – Text Analytics– heavily context dependent
  - Level of specificity – Telecommunications
- New approaches can solve much of this – Fall?

# Deep Text Webinar
# Applications

- 3 Main Types:
  - Search – An Enterprise Platform
  - Info Apps – Unstructured Text is Everywhere
  - Social Media – Fastest Growing Area

## Deep Text Webinar
## Enterprise Search Still Sucks

- Documents deal in language BUT it's all chicken scratches to Search
- Relevance – requires meaning
- Faceted Navigation has become the basic/ norm
  - Facets require huge amounts of metadata - tagging
- Auto-categorization – aboutness, subject facets
  - This is still fundamental to search experience
- Hybrid Model: publish Document -> Text Analytics analysis -> suggestions for categorization, entities, metadata - > present to author
- Content type rules:
  - No such thing as unstructured text – poly-structured
  - Sections – Specific - "Abstract" to Function "Evidence"

# Deep Text Webinar
# Enterprise Info Apps

- Focus on business value, cost cutting, new revenues
- Applications require sophisticated rules, not just categorization by similarity
- Business Intelligence – products, competitors
  - It is a growing field with revenues of $13.1 billion in 2015.
- Financial Services - Combine structured transaction data (what) with unstructured text (why)
  - Customer Relationship Management, Fraud Detection
  - Stock Market Prediction , eDiscovery, Text Assisted Review, HR resumes, automatic summaries, Expertise analysis, etc., etc.

## Social Media Applications Characteristics

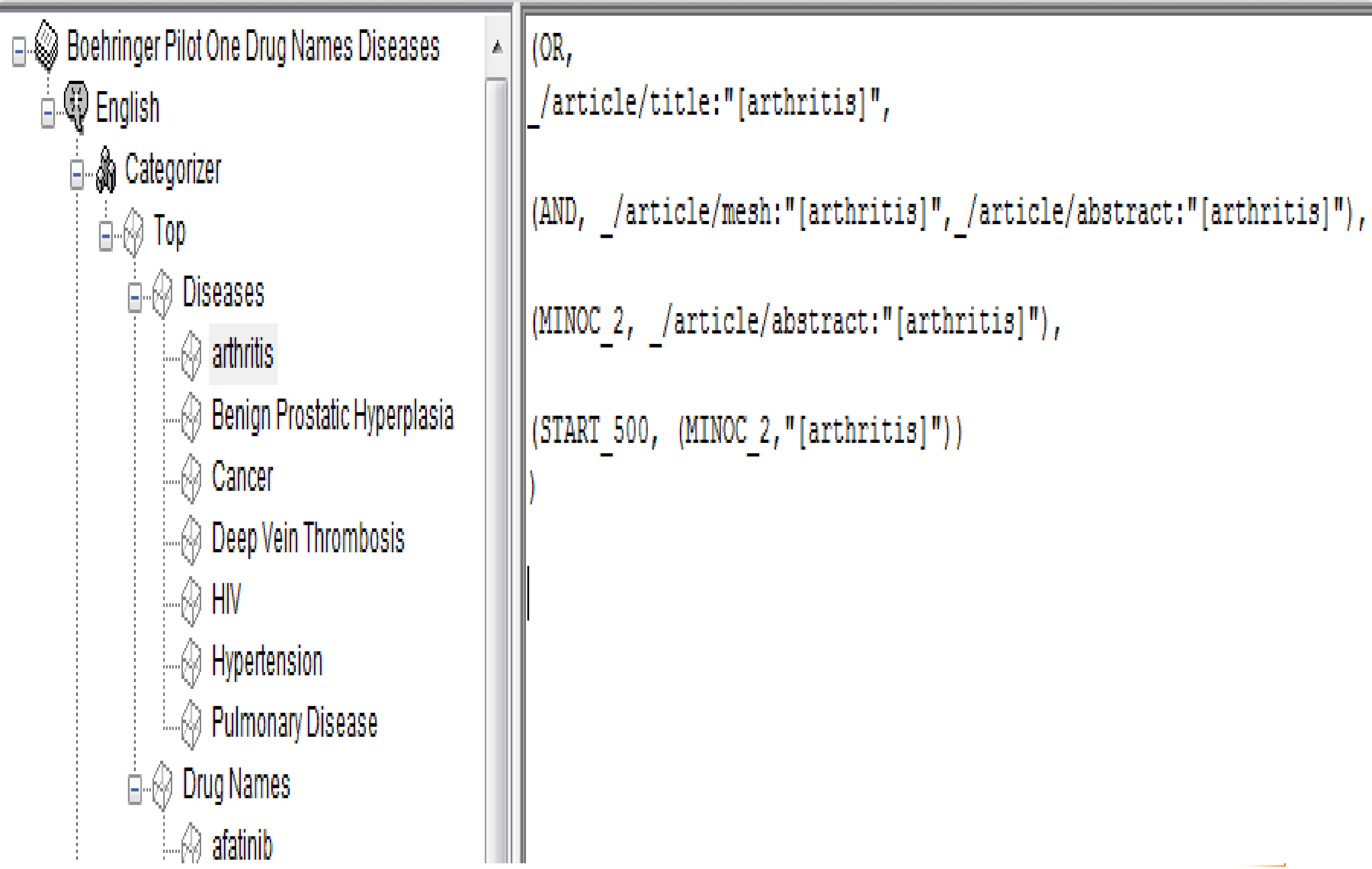
- Scale = Huge!  100’s of Millions / Billions
- Poor Quality of the Text
- Conversations, not stand alone documents
  - Issues of co-reference, who is speaking
- Direct Business Value
  - Customers, competitors, fix products, new products
- New techniques beyond counting pos. & neg.
  - Context, intensity, new models of emotions
  - New conceptual models, models of users, communities

## Social Media Applications

- Voice of the Customer-Employee-Voter
- Detection of a recurring problem categorized by subject, customer, client, product, parts, or by representative.
- Subscriber mood before and after a call – and why
- Political – conservative and liberal minds/texts
  - Disgust, shame, cooperation, openness
- Behavior Prediction – customer likely to cancel
- Fraud detection – lies in text have different patterns
- Areas:  sex, age, power-status, personality

# Deep Text Webinar
# Development: Deep Text vs. Deep Learning

- Two Schools – Language Rules vs. Math / Patterns
  - Depth & Intelligence vs. Speed & Power
- Deep Learning
  - Neural Networks – from 1980's, new = size and speed
  - Strongest in areas like image recognition, fact lookup
  - Weakest – concepts, subjects, deep language, metaphors, etc.
- Deep Text – Language, concepts, symbols
  - Categorization – most basic to human cognition
    - Beyond Categorization – making everything else smarter
  - Natural level categories: Mammal – Dog – Golden Retriever
  - Rules = higher accuracy – 98% - Rules brittle?



```
(OR,
_/article/title:"[arthritis]",

(AND, _/article/mesh:"[arthritis]",_/article/abstract:"[arthritis]"),

(MINOC_2, _/article/abstract:"[arthritis]"),

(START_500, (MINOC_2,"[arthritis]"))
)
```

# Deep Text Webinar
# Deep Text vs. Deep Learning

- Deep Learning is a Dead End - accuracy – 60-70%
  - Black Box – don’t know how to improve except indirect manipulation of input – “We don’t know how or why it works”
  - Domain Specific, tricks not deep understanding
  - No common sense  and no strategy to get there
  - Major – loss of quality – who is training who?
- Extra Benefits of a Deep Text Approach – Multiple InfoApps
- Future = Interpenetration of Opposites
  - Make Deep Learning smarter, add learning to Deep Text

# Text Analytics Development: Categorization Process
## Start with Taxonomy and Content

- Starter Taxonomy
  - If no taxonomy, develop (steal) initial high level
  - Library of semantic resources – templates, catalogs, data
- Analysis of taxonomy – suitable for categorization
  - Structure – not too flat, not too large, orthogonal categories
- Content Selection
  - Map of all anticipated content, Selection of training sets
- Start: taxonomy as initial categorization
- Term building – from content – basic set of terms that appear often / important to content
  - Auto-suggested and/or human generated
- Cycles: test set, recall, precision -> more content
- Rule templates, sectionize documents

# Deep Text Webinar
# Development: Entity Extraction Process

- Facet Design – from Knowledge Audit, K Map
- Find and Convert catalogs:
  - Organization – internal resources
  - People – corporate yellow pages, HR
  - Include variants
  - Scripts to convert catalogs – programming resource
- Build initial rules – follow categorization process
  - Differences – scale, threshold – application dependent
  - Recall – Precision – balance set by application
  - Issue – disambiguation – Ford company, person, car
- Unknown entities – NLP rules – “cap cap said”

## Deep Text Webinar
## Getting Started with Text Analytics

- Text Analytics is weird, a bit academic, and not very practical
  - It involves language and thinking and really messy stuff
- On the other hand, it is really difficult to do right (Rocket Science)
- Organizations don’t know what text analytics is and what it is for
- False Model – all you need is our software and your SME’s
  - Categorization is not a skill that SME’s have
- Companies get stuck – know the software but not how to really use it well, leads to abandoned projects

# Deep Text Webinar
## Smart Start: Think Big, Start Small, Scale Fast

- Think Big: Strategic Vision
  - K Audit – content, people, technology, KOS
  - Establish infrastructure – faster project development
  - Avoid expensive mistakes – dead end technology, etc.
- Start Small: Pilot or POC
  - Immediate value and learn by doing
  - Easier to get Management Buy-In
- Scale Fast: Multiple applications
  - Infrastructure – technical and semantic
  - Semantic Resources – catonomies, ontologies
  - First Project + 10%, Subsequent Projects – 50%

# Text Analytics and Information Architecture

- Text Analytics can provide deeper information structure
  - Similar to library and database resources
- Current use of taxonomy to next level
  - Extra – taxonomies and ontologies that DO!
- Adds the dimension of meaning
- Richer sets of relationships – ontologies, graph databases
- KA Audit is like a standard IA content inventory plus meaning
- Output – IA – site map
- Output – KA Audit – site map for entire enterprise, plus people, technology, information needs and behaviors
- Text analytics as tool for IA research – Text mining, explore content

# Deep Text Webinar
# Conclusions : Text Analytics:

- Is an infrastructure platform technology
- Makes everything smarter
- Is a great partner for IA – mutual enrichment
- Is a great partner for AI – mutual enrichment
- Needs a strategic vision
  - But also concrete and quick application to drive acceptance
- Future is Deep Text and Deep Learning integration
  - Text + Data, Language + Math, Social + Enterprise

# Questions?

Learn More:

- SLA – 6/16-20 -Phoenix
- Sentiment Symposium – 6/27-28-New York
- Taxonomy Boot Camp – 10/17-18-London
- Internet Librarian – 10/22-25-Monterey
- Text Analytics Forum – 11/6-9 –DC

# Resources

- Books
  - Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data
    - Tom Reamy
  - Women, Fire, and Dangerous Things
  - Don't Think of an Elephant
    - George Lakoff
  - Knowledge, Concepts, and Categories
    - Koen Lamberts and David Shanks
  - Thinking Fast and Slow
    - Daniel Kahneman
  - Any cognitive science book written after 2010

# Resources

- Conferences – Web Sites
  - Text Analytics Forum - All aspects of text analytics
    - http://www.textanalyticsforum.com
  - Semtech
    - http://www.semanticweb.com
  - Dataversity Conferences
  - http://www.dataversity.net/
  - Sentiment Analysis Symposium
    - www.sentimentsymposium.com

# Resources

- LinkedIn Groups:
  - Text Analytics
  - Text Analytics Forum
  - Taxonomy Community of Practice
  - Sentiment Analysis
  - Text and Social Analytics
  - Metadata Management
  - Semantic Technologies, Semantic Web
  - Association for Information Science & Technology