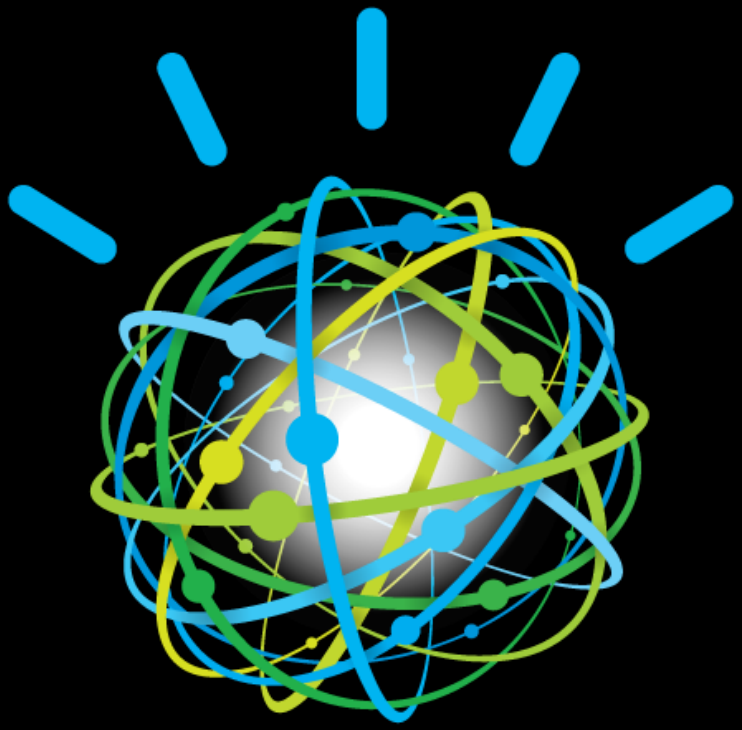
Putting Content to Work
ECM UserNet 2011

IBM

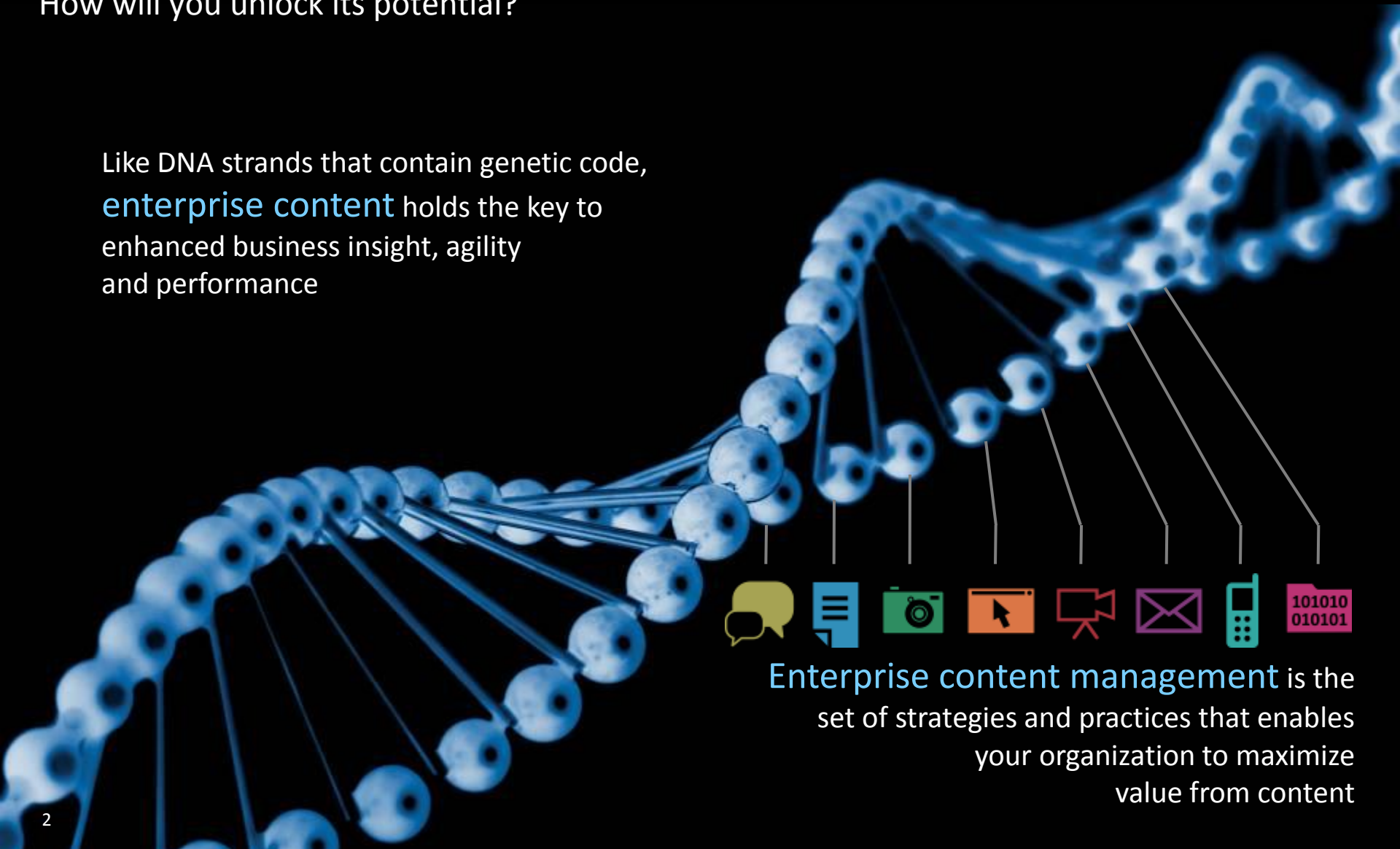


Answer Your Critical Business Questions Today: Leverage the Content Analytics Technology Behind IBM's Watson

Enterprise content: your organization's DNA

How will you unlock its potential?

Like DNA strands that contain genetic code, **enterprise content** holds the key to enhanced business insight, agility and performance



Enterprise content management is the set of strategies and practices that enables your organization to maximize value from content

To achieve competitive edge

Organizations need to be smarter, faster

- **77% of CEOs** say they do not have real-time information to make key business decisions
- **1 in 3 business leaders** frequently make business decisions based on information they don't have, or don't trust
- **1 in 2 business leaders** say they don't have access to the information they need to do their jobs

Companies that invest in business insight outperform their peers, showing 33% higher revenue growth, 12 times more profit growth, and 32% higher return on invested capital.

Sources: IBM 2010 CEO & CFO Studies, IBM 2010 Break Away With Business Analytics and Optimization Study

To enhance customer intimacy and employee collaboration

Business is becoming more social

- **69% of executives** report gaining measurable business benefits from social technologies
- **52% of organizations** plan to increase investment in social media and collaboration tools in 2010

"Social media has shifted control of the corporate message away from the organization and towards consumers and other stakeholders, and running away and hiding is no longer the safe option."

Burson-Marsteller
"The Global Social Media Check-up,"
February 2010

Sources: Gartner, 2010; McKinsey, 2009.

Enterprise content management is evolving …

From Systems of Record to Systems of Engagement

Consideration	Systems of Record—Enterprise Content Management	Systems of Engagement—Social Business Systems
Focus	Transactions	Interactions
Governance	Command & Control	Collaboration
Core Elements	Facts, Dates, Commitments	Insights, Ideas, Nuances
Value	Single Source of the Truth	Open Forum for Discovery & Dialog
Performance Standard	Accuracy & Completeness	Immediacy and Accessibility
Content	Authored	Communal
Primary Record Type	Documents (Text, Graphics)	"Conversations" (Text-based, Images, Audio, Video)
Searchability	Easy	Hard
Usability	User gets trained on system and has access to follow-on support	User "knows" system from consumer experience
Accessibility	Regulated & Contained	Ad Hoc & Open
Retention	Permanent	Transient
Policy Focus	Security (Protect Assets)	Privacy (Protect Users)

*"**Conversations** - in a wide variety of forms and on a dizzying array of devices are now the challenge …*

the pressure by the business to implement is accelerating"

Source: Systems of Engagement and The Future of Enterprise IT - A Sea Change in Enterprise IT (AIIM and Geoffrey Moore, TCG Advisors)

Social business drives an unprecedented need for insight from natural language conversations

500 billion impressions
annually made about products and services **

770 million people
worldwide visited a social networking site *

44x information growth by 2020 ***

Public Social Media

Conversations about quality, experience, price, value, service …

- *Forums and Newsgroups*
- *Wikis, Blogs and Microblogs*
- *Social Networks*
- *Social Media News Aggregators*

Corporate Social Business

Conversations about strategy, projects, issues, risks, outcomes …

Microsoft SharePoint 2010

Lotus. software

In addition to conversations about quality, experience, price, value, service …

- *Wikis, RSS and Forums*
- *Email and Collaborative Content*
- *Call Center Notes and Recordings*
- *Customer and Employee Surveys*
- *Reports, Minutes and Research*

Sources: * comScore, Social Networking Phenomenon
** *Empowered,* a book by Josh Bernoff / Ted Schadler
*** IDC Digital Universe Study, May 2010

Truly understanding natural language is the next great computing challenge

- Over **80%** of information today is unstructured and based on **natural language**
- The impact of **Systems of Engagement** both inside and outside the firewall is dramatic … such masses of information **not easily understandable** by humans
- Legacy approaches have all failed; **"searching" not the right approach**
- A new approach is needed, leveraging **content analysis** and natural language processing

IBM WATSON

The Next Grand Challenge

http://www.youtube.com/watch?v=VjHMYuGkzlU

Real language is real hard

Chess

- A finite, mathematically well-defined search space
- Limited number of moves and states
- Grounded in **explicit**, **unambiguous** mathematical **rules**

Human Language

- Ambiguous, **contextual** and implicit
- Contains slang, riddles, idioms, abbreviations, acronyms and more
- Grounded only in **human cognition**
- Seemingly infinite number of ways to express the same concepts and meaning

The hard part: understanding natural language with confidence and accuracy

- Where was Einstein born?

Unstructured

One day, from among his city views of Ulm, Otto chose a watercolor to send to Albert Einstein as a remembrance of Einstein's birthplace.

Structured

Person	Born In
A. Einstein	Ulm

- Welch ran this?

If leadership is an art then surely Jack Welch has proved himself a master painter during his tenure at GE.

Person	Organization
J. Welch	GE

The Jeopardy! Challenge

5 key dimensions to drive the technology

- Broad/open domain
- Complex language
- High precision
- Accurate confidence
- High speed

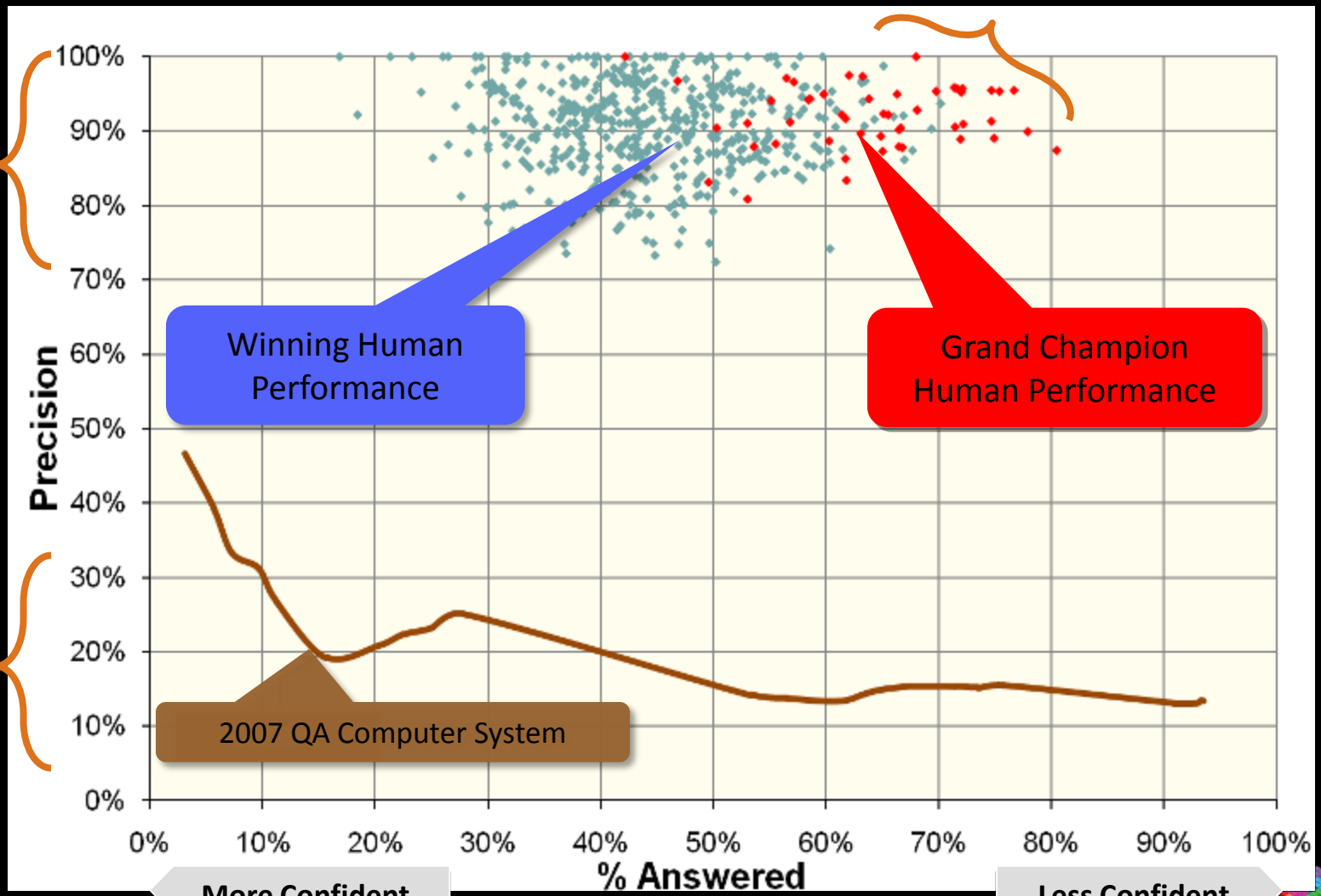
$200
If you're standing, it's the direction you should look to check out the wainscoting

$800
In cell division, mitosis splits the nucleus & cytokinesis splits this liquid *cushioning* the nucleus

$1000
Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

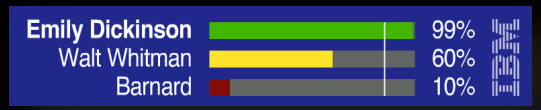
The Jeopardy! winner's cloud

Best human performance



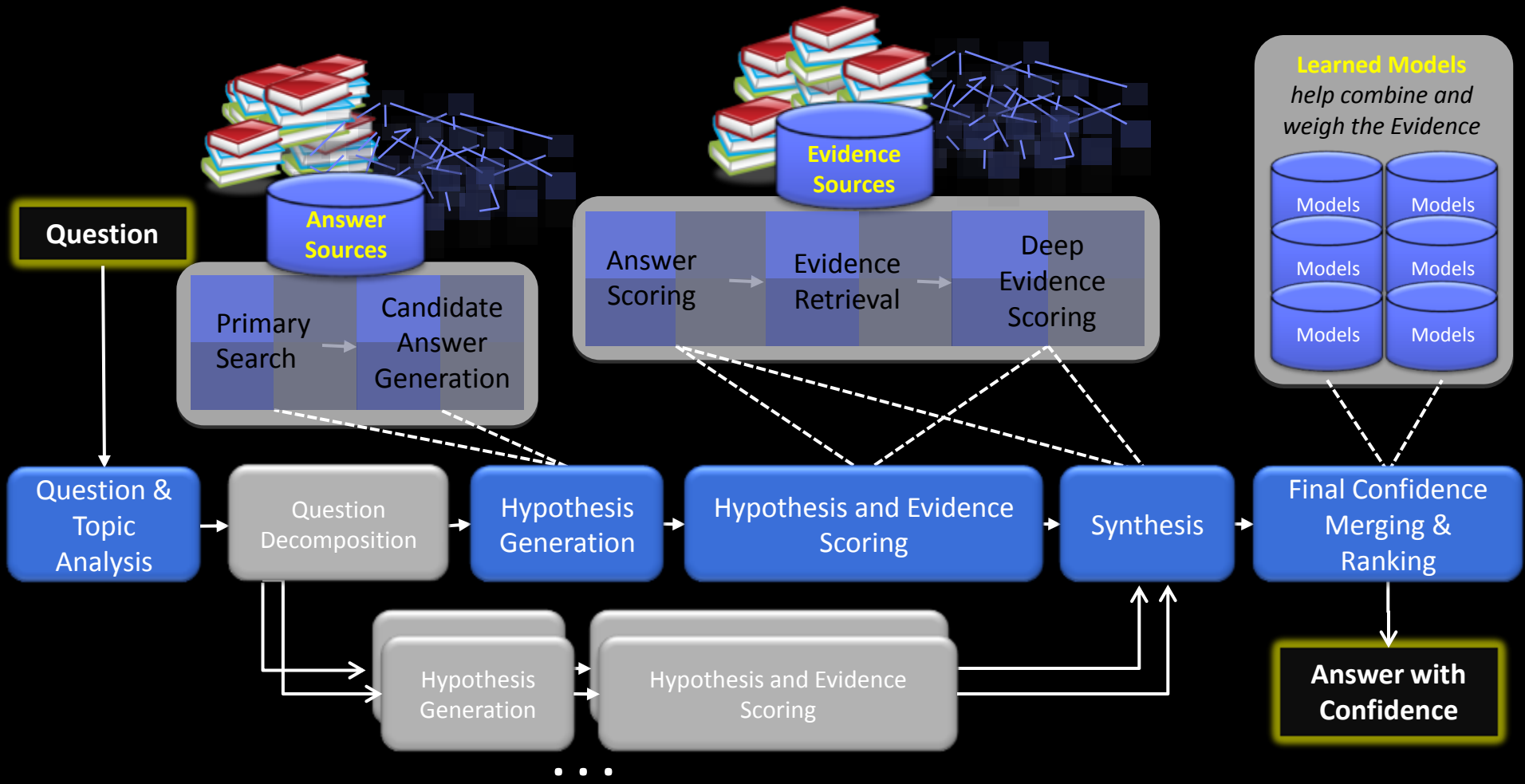
The Big Idea: Evidence-Based Reasoning over Natural Language Content

- **Deep Analysis** of clues / questions and stored knowledge (content)
  - Search for many possible answers based on different interpretations of question
  - Possible answers depend on stored and available knowledge (content)
- **Find, Analyze and Score Evidence** from many different sources (not just one document)
  - For each possible answer using many advanced NLP and reasoning algorithms
- **Combine Evidence** and compute a confidence value for each possibility using statistical machine learning
  - Ranks possible answers based on confidence
  - If confidence is above the threshold – then buzz in to answer



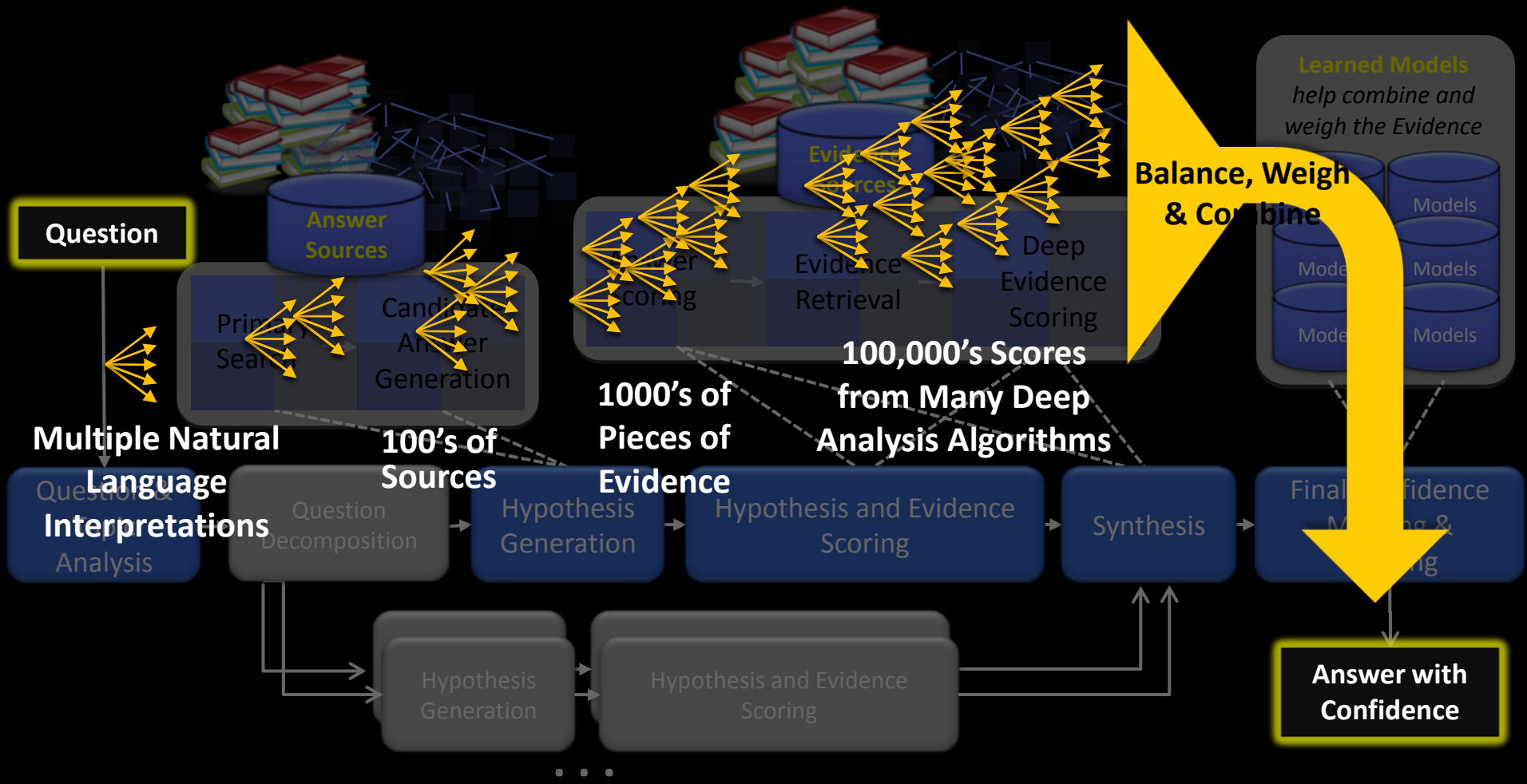
The technology behind IBM Watson

How it Really Works with Content



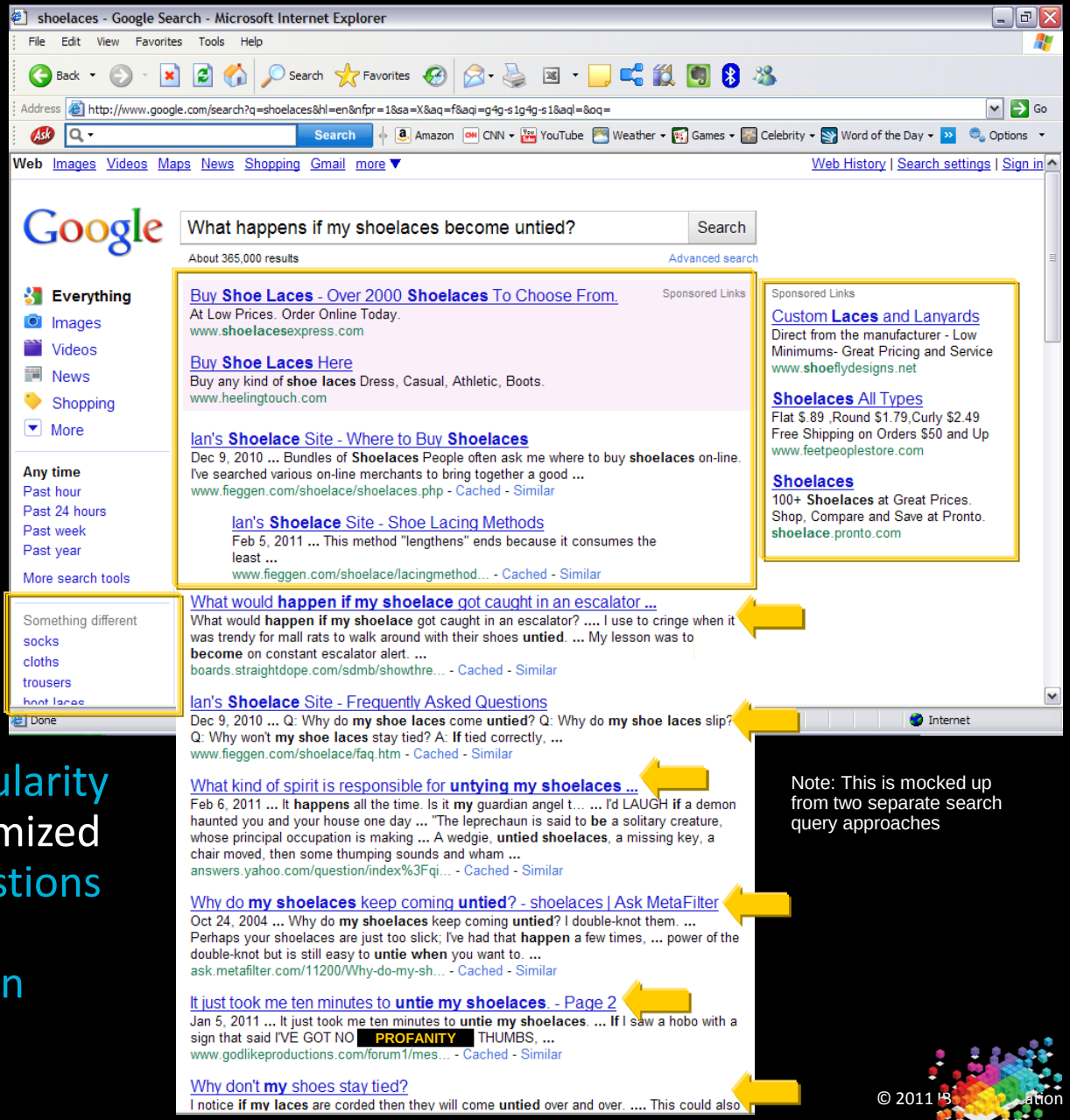
The technology behind IBM Watson

How it Really Works with Content



Isn't this just like search?

Question:
What happens if my shoelaces become untied?



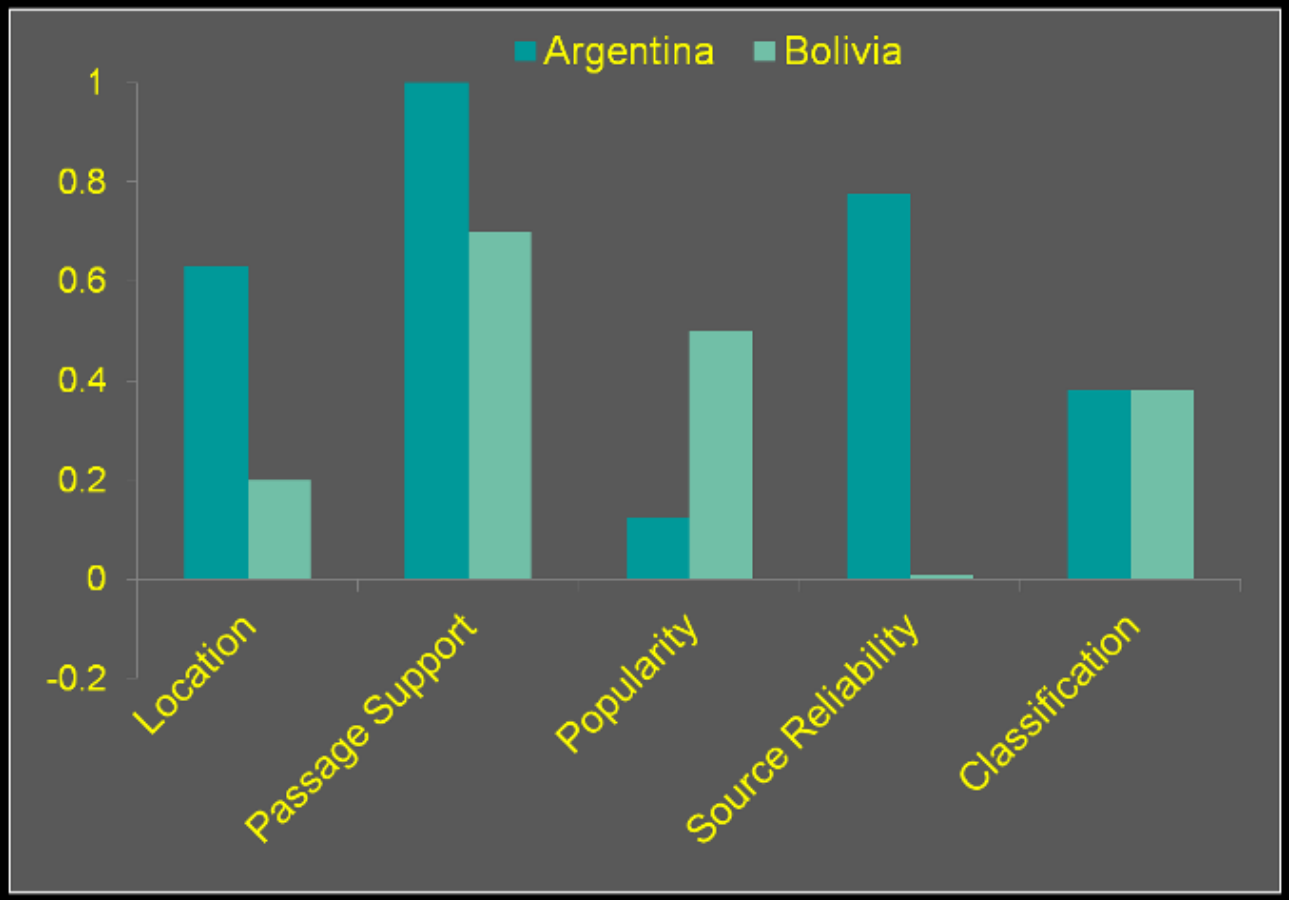
Search only results:

- Based on keyword popularity and search engine optimized
- Lots of shopping suggestions
- Results prove it didn't understand the question
- Can include profanity

Note: This is mocked up from two separate search query approaches

Evidence Profiles summarize evidence analysis across many sources

Clue: Chile shares its longest land border with this country.



Bolivia is more popular due to a commonly discussed border dispute but Argentina has more reliable sources

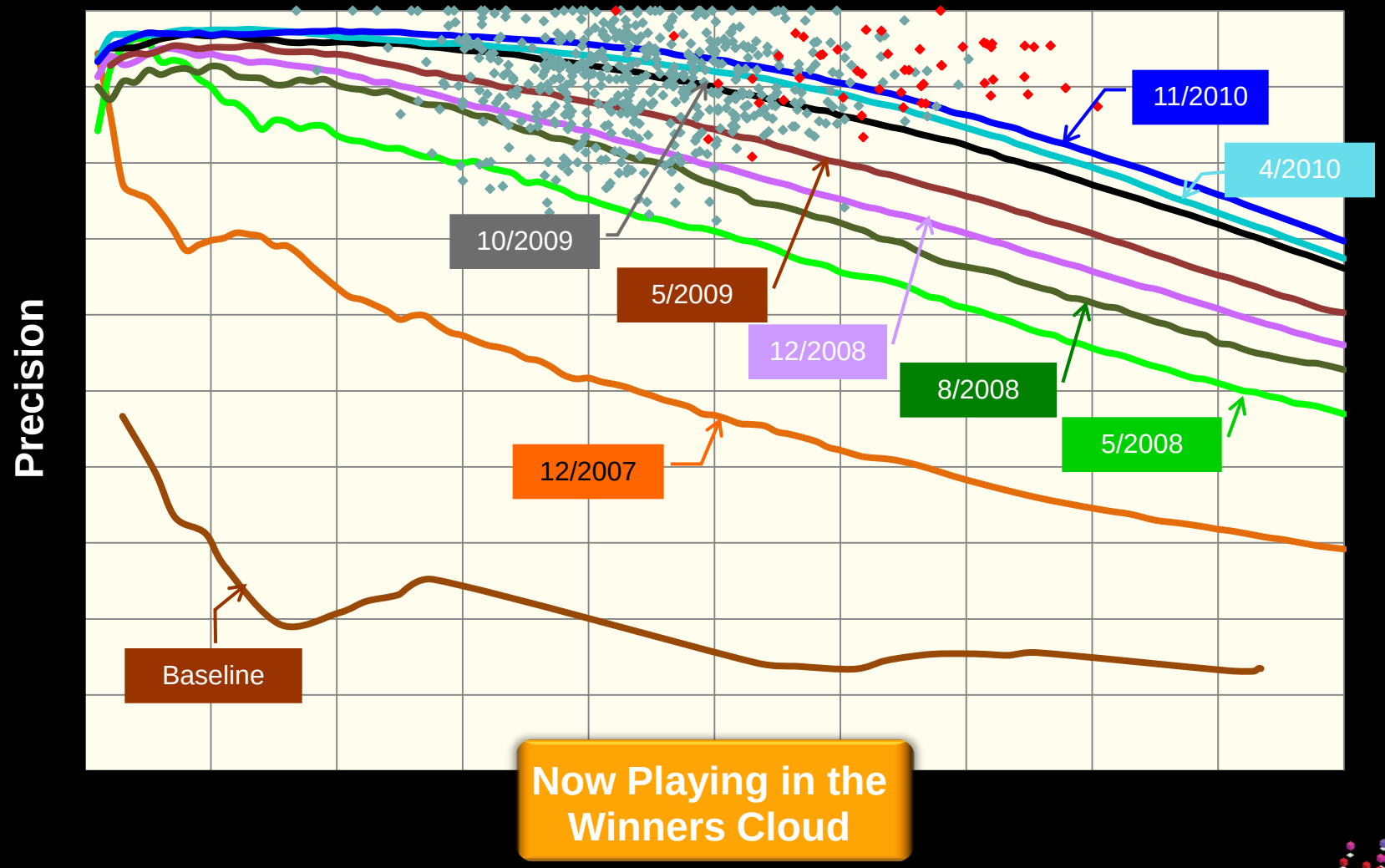
Correct Answer: Argentina

Using Statistical Machine Learning different classes of evidence earn different weights

For example, Watson uses statistical machine learning to discover that Jeopardy! categories are **weak indicators** of the answer type

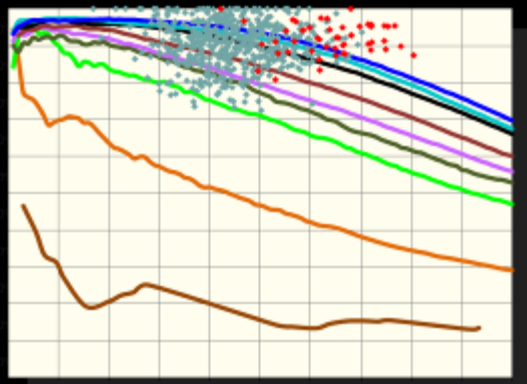
U.S. Cities	Country Clubs	Authors
St. Petersburg is home to Florida's annual tournament in this game popular on shipdecks **(Shuffleboard)**	From India, the shashpar was a multi-bladed version of this spiked club **(a mace)**	Archibald MacLeish based his verse play "J.B." on this book of the Bible **(Job)**
Rochester, New York grew because of its location on this **(the Erie Canal)**	A French riot policeman may wield this, simply the French word for "stick" **(a baton)**	In 1928 Elie Wiesel was born in Sighet, a Transylvanian village in this country **(Romania)**

DeepQA: Incremental progress in precision and confidence
(Period: 6/2007 - 11/2010)



Precision, confidence and speed

- **Deep Analytics:** We achieved champion-levels of precision and confidence over a huge variety of expression

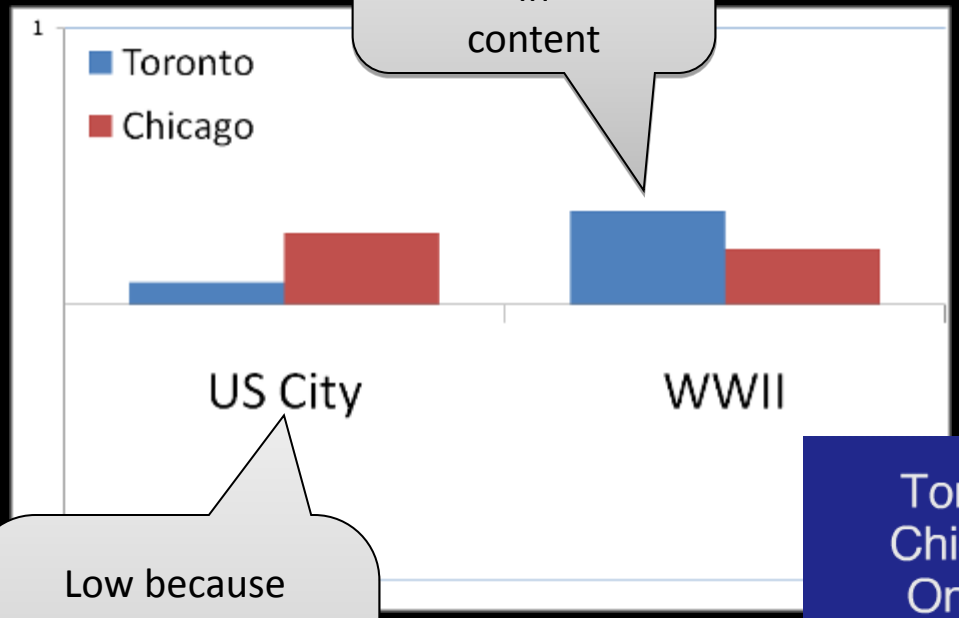


- **Speed:** By optimizing Watson's computation for Jeopardy! on over 2,800 POWER7 processing cores we went from 2 hours per question on a single CPU to an average of just 3 seconds – fast enough to compete with the best.

- **Results:** in 55 real-time sparring games against former Tournament of Champion Players last year, Watson put on a very competitive performance in all games and winning 71% of the them!

Toronto vs. Chicago

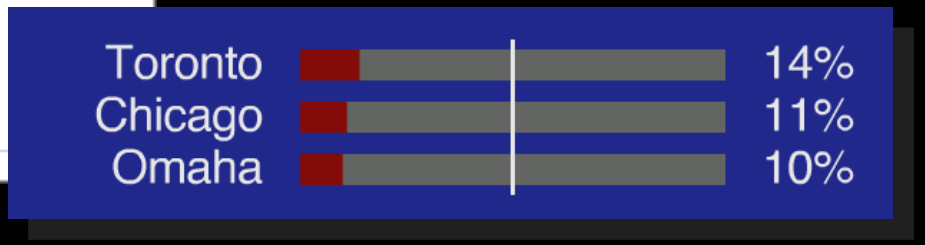
Low because of weak evidence in content



Low because being a **US City** is not a strong requirement simply based on Jeopardy! category

US CITIES

Its largest airport is named for a World War II hero; its second largest, for a World War II battle



- Overall confidence was below threshold for both answers
- You must answer in Final Jeopardy! so Watson guessed

Potential Watson Business Applications

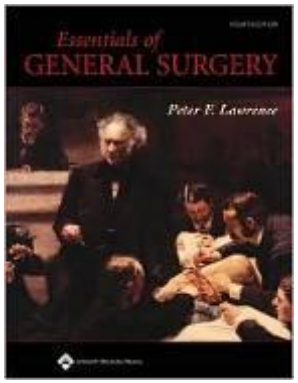
- **Healthcare / Life Sciences:** Diagnostic Assistance, Evidence-Based and Collaborative Medicine
- **Tech Support:** Self Service Help-Desk, **Contact Centers**
- **Enterprise Knowledge Management** and **Business Intelligence**
- **Government:** Improved Information Sharing and **Education**
- **Legal:** eDiscovery, Evidence Based Sentencing and Patent Research
- More to come …

Applying Watson to the Real World

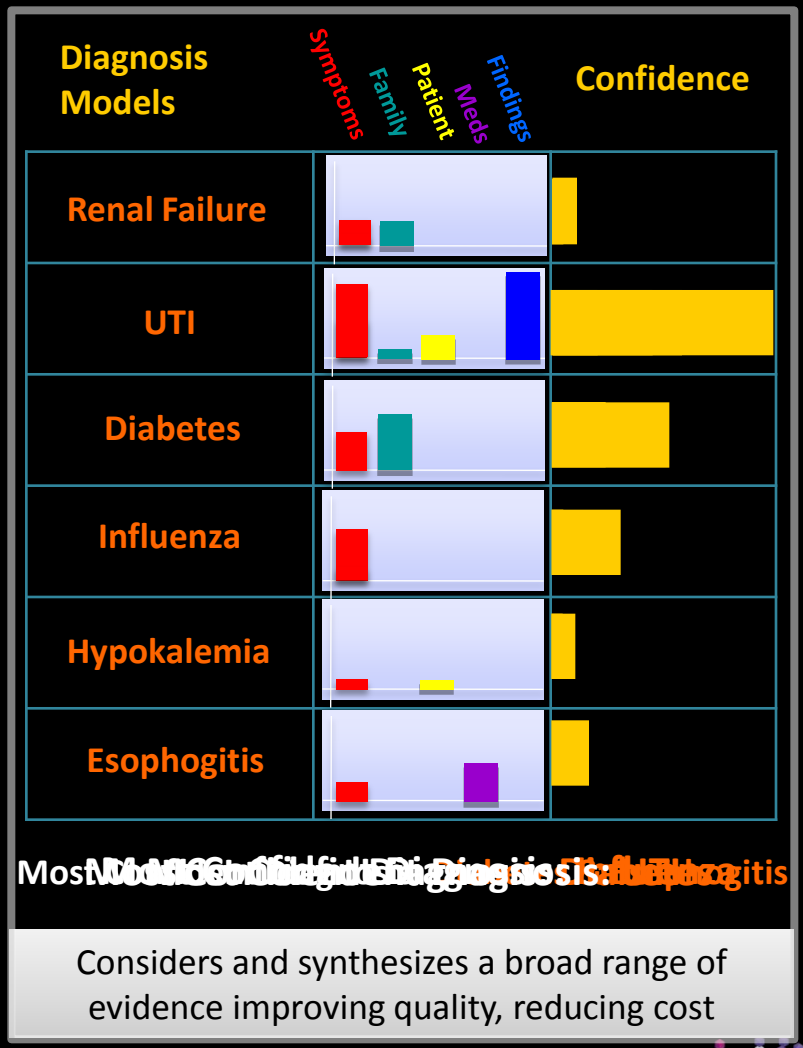
Continuous Evidence-Based Diagnostic Analysis

Answer Sources

- Symptoms
- Family History
- Patient History
- Medications
- Tests / Findings
- Notes / Hypotheses
- Huge Volumes of Texts, Journals, References, Databases. etc.



"anesthesia should be avoided if possible"



Considers and synthesizes a broad range of evidence improving quality, reducing cost

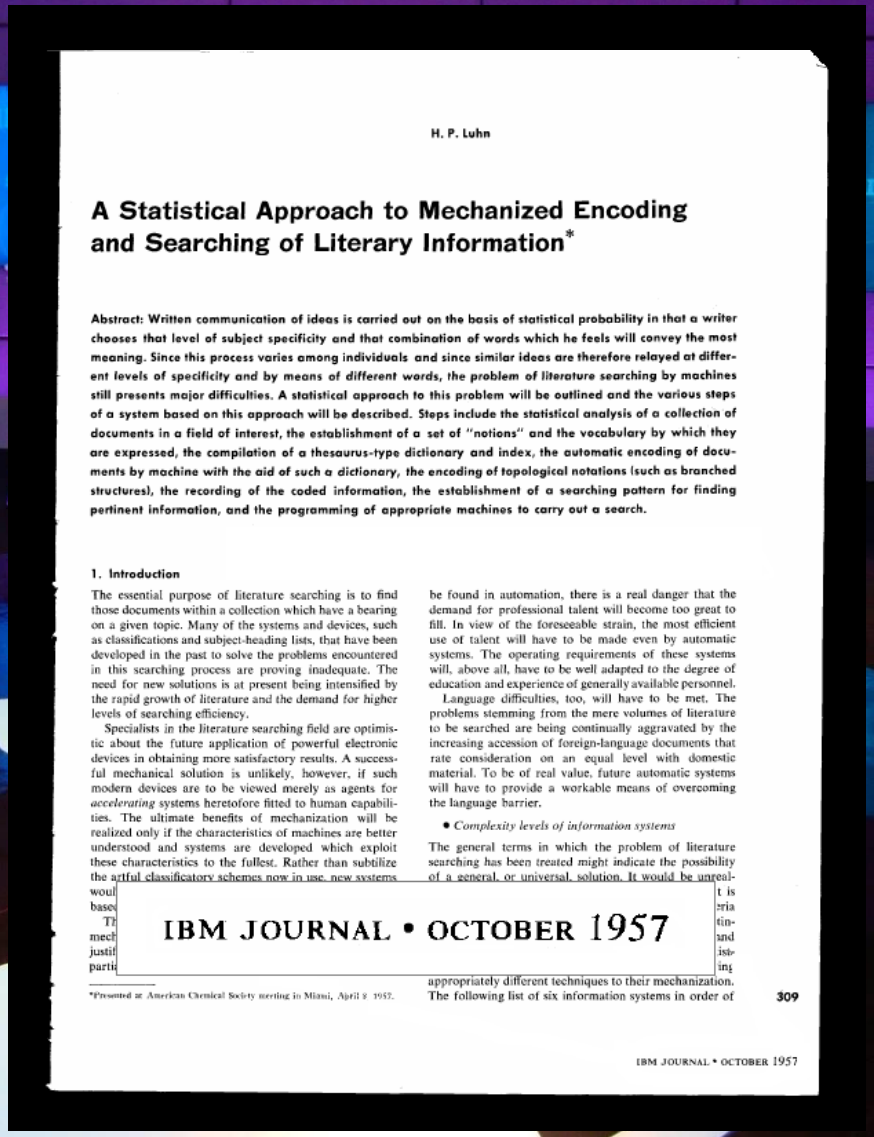
Watson and IBM ECM Today

- **Natural Language Processing (NLP)** is the cornerstone to translate interactions between computers and human (natural) languages
  - Watson uses IBM Content Analytics to perform critical NLP functions
- **Unstructured Information Management Architecture (UIMA)** is an open framework for processing text and building analytic solutions
  - Several IBM ECM products leverage UIMA text analytics processing:
    - IBM Content Analytics
    - OmniFind Enterprise Edition
    - IBM Classification Module
    - IBM eDiscovery Analyzer

IBM at 100: ECM Innovation for Over 50 Years

Beginning in 1957 …

Searching and Classifying



H. P. Luhn

A Statistical Approach to Mechanized Encoding and Searching of Literary Information*

Abstract: Written communication of ideas is carried out on the basis of statistical probability in that a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning. Since this process varies among individuals and since similar ideas are therefore relayed at different levels of specificity and by means of different words, the problem of literature searching by machines still presents major difficulties. A statistical approach to this problem will be outlined and the various steps of a system based on this approach will be described. Steps include the statistical analysis of a collection of documents in a field of interest, the establishment of a set of "notions" and the vocabulary by which they are expressed, the compilation of a thesaurus-type dictionary and index, the automatic encoding of documents by machine with the aid of such a dictionary, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

1. Introduction

The essential purpose of literature searching is to find those documents within a collection which have a bearing on a given topic. Many of the systems and devices, such as classifications and subject-heading lists, that have been developed in the past to solve the problems encountered in this searching process are proving inadequate. The need for new solutions is at present being intensified by the rapid growth of literature and the demand for higher levels of searching efficiency.

Specialists in the literature searching field are optimistic about the future application of powerful electronic devices in obtaining more satisfactory results. A successful mechanical solution is unlikely, however, if such modern devices are to be viewed merely as agents for *accelerating* systems heretofore fitted to human capabilities. The ultimate benefits of mechanization will be realized only if the characteristics of machines are better understood and systems are developed which exploit these characteristics to the fullest. Rather than subtilize the artful classificatory schemes now in use, new systems

IBM JOURNAL • OCTOBER 1957

be found in automation, there is a real danger that the demand for professional talent will become too great to fill. In view of the foreseeable strain, the most efficient use of talent will have to be made even by automatic systems. The operating requirements of these systems will, above all, have to be well adapted to the degree of education and experience of generally available personnel.

Language difficulties, too, will have to be met. The problems stemming from the mere volumes of literature to be searched are being continually aggravated by the increasing accession of foreign-language documents that rate consideration on an equal level with domestic material. To be of real value, future automatic systems will have to provide a workable means of overcoming the language barrier.

• *Complexity levels of information systems*

The general terms in which the problem of literature searching has been treated might indicate the possibility of a general, or universal, solution. It would be unreal-

t is
eria
tin-
and
ist-
ing

appropriately different techniques to their mechanization. The following list of six information systems in order of

*Presented at American Chemical Society meeting in Miami, April 8 1957.

Unlock valuable insight from content

What our clients are doing with Content Analytics

Understand what customers want before they ask.

Detect fraudulent claims before they are paid.

Dynamically deploy resources to the areas of greatest threat.

Save lives by quickly identifying critical safety defects.

Are you unlocking the value of your unstructured content?

NTT Docomo

Smart is: reducing customer churn

"As a result, we can easily identify trends and patterns from customer voices across our organization and provide better customer service."
— Makoto Ichise, Manager of Information System Department Group, NTT Docomo

Industry context: telecommunications
Value driver: improve customer service
Solution onramp: content analytics

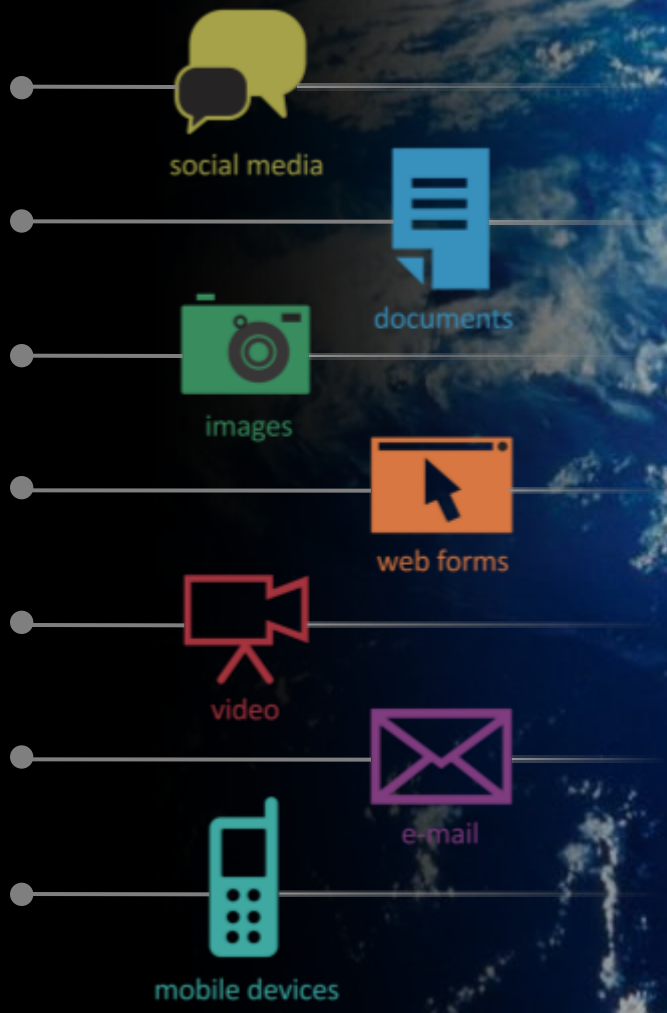
Business Challenge
Adopt a customer-oriented business strategy to offer highly satisfying products and services based on real voice of customers (VoC).

What's Smart?
They process call center notes and customer emails to detect likely candidates for customer churn. A rules-based text analysis engine in IBM Content Analyzer detects the customer churn candidates. An alerting engine then automatically sends reports to a department that deals specifically with customer churn situations.

Smarter Business Outcomes
Improved rates for model and service upgrades to loyal customers. Started new Premium Club points program based on VoC. Set initial parameters of mobile phones based on VoC.

Separate the signal from the noise

Leveraging content requires the ability to **search**, **assess** and **analyze** large volumes of **text** in order to understand and determine relevant **insight** quickly … from multiple information sources **inside and outside the firewall.**



Is this your content reality?

Information retrieval and understanding is poor …

- Can't find the right content when needed; decisions are being made based on the wrong information … the search and hope model is broken.
- The keep everything forever model has failed; it's well beyond human capacity to address … driving up costs and governance risks by storing unnecessary content.

Business decisions not leveraging text based information …

- Business Intelligence and Data Warehouse initiatives limited to data only; no visibility into 80 percent of needed information (content) for effective decisions.
- Current text analysis systems are too complex, require model building and take months to deploy; no ability to easily respond to changing conditions.

Smart is: leveraging analyzed content

Uncovering new insights

Business Challenge
How to derive insight from billions of arrest, complaint, summonses, homicide and shooting records to solve crimes faster.

What's Smart?
Implemented IBM Content Analytics to create a crime warehouse that combines structured and unstructured information.

Smarter Business Outcomes
Information reaches detectives in minutes, not days or weeks. Previously unknown relationships between suspects automatically uncovered.

Finding what you need

Business Challenge
Securely connecting 13,000 scientists and engineers to millions of documents to enable technical innovation.

What's Smart?
In four months implemented secure semantic text analytics and search solution for internal and external facing portals.

Smarter Business Outcomes
Scientists and engineers worldwide are now securely connected to the most relevant research assets, driving new innovations.

Going from raw information to rapid insight

Uncover business insight through unique visual-based approach

Aggregate and extract from multiple sources

… to form large **text**-based collections from multiple internal and external sources (and types), including ECM repositories, structured data, social media and more.

Organize, analyze and visualize

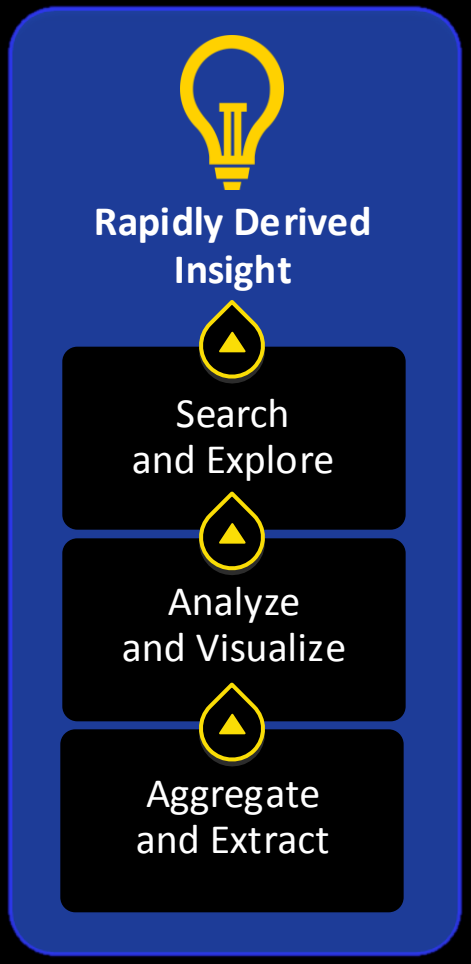
… enterprise **content** (and data) by identifying trends, patterns, correlations, anomalies and business context from collections.

Search and explore to derive insight

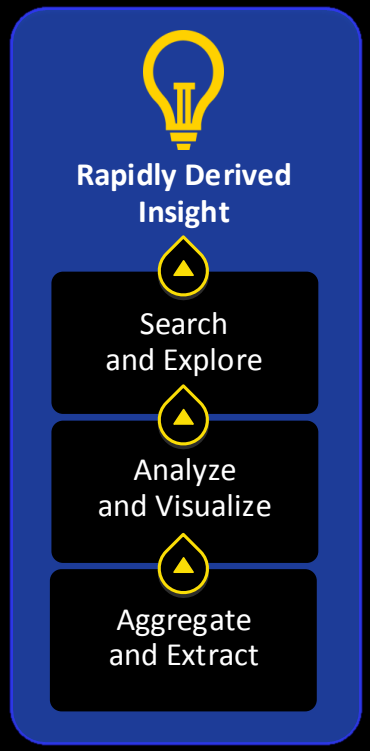
… from collections to confirm what is suspected or uncover something new without being forced to build models or deploy complex systems.

IBM Content Analytics is a platform to derive rapid insight

- Transform raw information into **business insight quickly** without building models or deploying complex systems.
- Derive insight in **hours** or **days** … not weeks or months.
- **Easy to use** for all knowledge workers to **search** and **explore** content.
- **Flexible** and **extensible** for deeper insights.



Enabling the power of rapid insight



- Find relevant enterprise content quickly and securely
- Assess enterprise content to decommission the unnecessary and govern the content that matters
- Customize rapid insight to industry and customer specific needs
  - IBM LanguageWare Tooling (included)
  - IBM Classification Module (optional)
  - IBM Text Analytics Group (services)
- Enable deeper insights through integration to other systems and solutions
  - IBM ECM and ACM solutions
  - IBM Cognos and SPSS Analytics Systems
  - IBM InfoSphere and Netezza Data Warehouse Systems

Fast Food Restaurant and Mindshare Technologies

Smart is: discovering emerging trends and patterns

Transform raw information into business insight quickly and easily

Industry context: retail customer surveys
Value driver: faster, deeper insights
Solution onramp: content analytics

Business Challenge
Provide immediate feedback from surveys, which is then translated into actionable intelligence through powerful and incisive reporting.

What's Smart?
Exposing additional insight currently not available through their manual survey review process

Smarter Business Outcomes
Mindshare Technologies, using IBM Content Analytics, received a prestigious partnership award from the company for discovery of new emerging treads/patterns and delivering key insights into new products, customer demographics and visit frequency.

IBM Content Analytics adds value to …

Healthcare Analytics
- ***Analyzing:*** *E-Medical records, hospital reports*
- ***For:*** *Clinical analysis; treatment protocol optimization*
- ***Benefits:*** *Better management of chronic diseases; optimized drug formularies; improved patient outcomes*

Crime Analytics
- ***Analyzing:*** *Case files, police records, 911 calls…*
- ***For:*** *Rapid crime solving & crime trend analysis*
- ***Benefits:*** *Safer communities & optimized force deployment*

Automotive Quality Insight
- ***Analyzing:*** *Tech notes, call logs, online media*
- ***For:*** *Warranty Analysis, Quality Assurance*
- ***Benefits:*** *Reduce warranty costs, improve customer satisfaction, marketing campaigns*

Customer Care
- ***Analyzing:*** *Call center logs, emails, online media*
- ***For:*** *Buyer Behavior, Churn prediction*
- ***Benefits:*** *Improve Customer satisfaction and retention, marketing campaigns, find new revenue opportunities*

Insurance Fraud
- ***Analyzing:*** *Insurance claims*
- ***For:*** *Detecting Fraudulent activity & patterns*
- ***Benefits:*** *Reduced losses, faster detection, more efficient claims processes*

Social Media for Marketing
- ***Analyzing:*** *Call center notes, SharePoint, multiple content repositories*
- ***For:*** *churn prediction, product/brand quality*
- ***Benefits:*** *Improve consumer satisfaction, marketing campaigns, find new revenue opportunities or product/brand quality issues*

BJC Healthcare and Washington University Partnership

Smart is: unlocking biomedical informatics answers

"We anticipate this solution to be a game changer in biomedical research and patient care. I believe that IBM Content Analytics will ultimately accelerate the pace of clinical and translational research through more rapid and accurate extraction of research relevant information from clinical documents"

Dr. Rakesh Nagarajan, M.D., Ph.D., Associate Professor, Department of Pathology and Immunology, Washington University.

Industry context: healthcare
Value driver: access to biomedical trends, insight
Solution onramp: content analytics

Business Challenge

Existing Biomedical Informatics (BMI) resources were disjointed and non-interoperable, available only to a small fraction of researchers, and frequently redundant. No capability to tap into the wealth of research information trapped in unstructured clinical notes, diagnostic reports, etc.

What's Smart?

Capitalizing on the untapped, unstructured information of clinical notes and reports by using IBM Content Analytics with IBM InfoSphere Warehouse.

Smarter Business Outcomes

Researchers now able to answer key questions previously unavailable. Examples include *Does the patient smoke?, How often and for how long?, If smoke free, how long? What home medications is the patient taking? What is the patient sent home with? What was the diagnosis and what procedures performed on patient?*

IBM

US Army and IBM Pilot Program

Smart is: intelligently classifying documents

"Consistent, reliable and automated configuration of content is critical."

Industry context: *government*
Value driver: *speed, accuracy of classification*
Solution onramp: *content analytics*

Business Challenge

With millions of email messages going through the Army's systems every year, the department needed to improve the accuracy and speed of its content categorization in order to meet NARA's regulations for accurate and effective records retention.

What's Smart?

The department is seeking to transform its manual, inaccurate human categorization process with automated classification technology. In its pilot, the Army resolved inconsistencies in content categorization using IBM Classification Module's contextual classification; replacing its over-burdened, labor-intensive content categorization process.

Smarter Business Outcomes

Improves visibility and access to accurately categorized email content . Provides more insight for records retention and legal discovery. Reduces storage required for email messages.

Text Analytics Group

Accelerate your time to value

- Text Analytics Group, an ECM services group, focuses on solving client needs with content analytics
- Provides solution engagements to quickly deliver deeper insights (or rapid insights)
  - Client-driven development and delivery of tailored, vertically-targeted solutions
  - Skilled professionals in text and content analysis
  - Proven track record implementing text and content analytic solutions that deliver smarter business outcomes
- Industry solution examples:

Medical Records Text Analysis (MRTA)

- Solution for healthcare payers and providers
- Analyzes medical records to enable better patient care, more efficient plans, accelerated research, trials and more

Intelligence Bundle

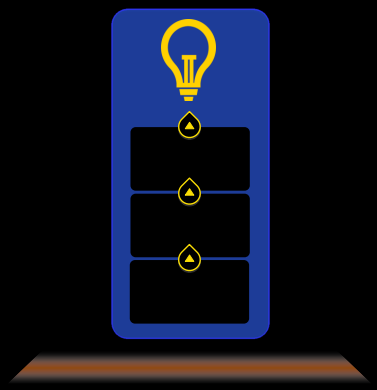
- Solution for police, security, intelligence, and customs agencies
- Combines IBM Content Analytics with key services assets to deliver unique (rapid, deeper, better) insights

Start unlocking the insight trapped in your content today

Uncover business insight *quickly to improve product quality and customer service, detect fraud, optimize decision making and more …*

IBM Content Analytics

- Find relevant enterprise content quickly and securely
- Assess enterprise content to decommission the unnecessary and govern the content that matters
- Customize rapid insight to industry and customer specific needs
- Enable deeper insights through integration to other systems and solutions



Thank you