OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

FORESIGHT AND UNDERSTANDING FROM SCIENTIFIC EXPOSITION (FUSE)

Office of Incisive Analysis

LEADING INTELLIGENCE INTEGRATION

Dewey Murdick, Program Manager
NIH Portfolio Management
24 July 2012

IARPA

BE THE FUTURE

"Invests in high-risk/high-payoff research programs that have the potential to provide our nation with an overwhelming intelligence advantage"
http://www.iarpa.gov/

Foresight and Understanding from Scientific Exposition (FUSE)

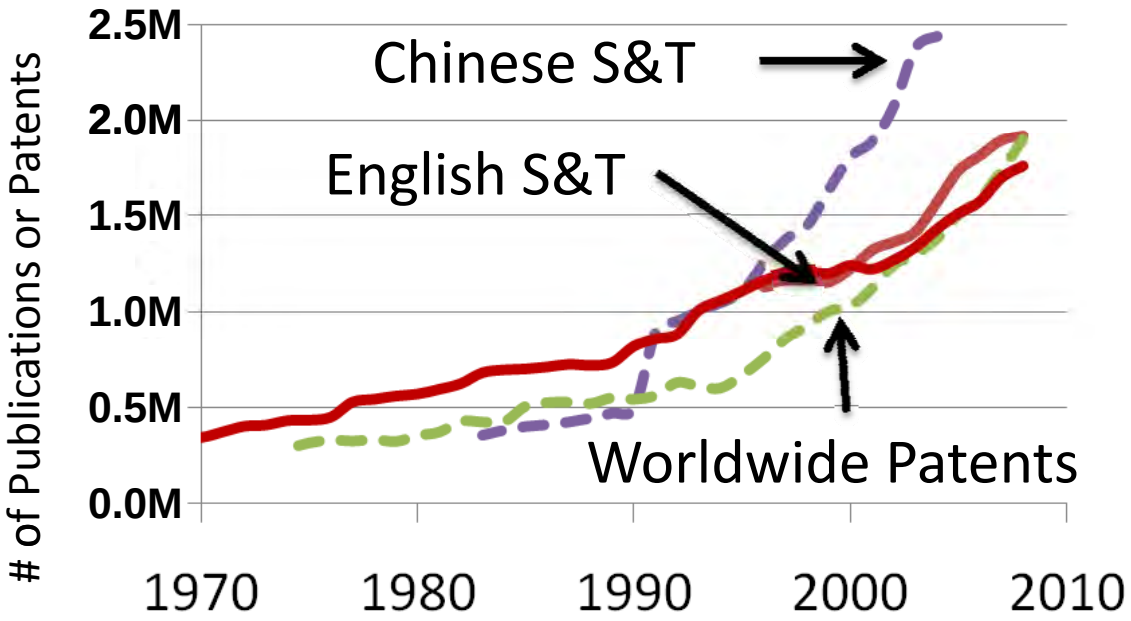
Enable reliable, early detection of emerging scientific and technical capabilities across disciplines and languages found within the full-text content of scientific, technical, and patent literature

Focus from the outset on **English, Chinese, German**, Japanese, Russian, *Korean*, and *Spanish*

Novelty	→	Discover patterns of emergence and connections between technical concepts at a speed, scale, and comprehensiveness that exceeds human capacity
Usage	→	Alert analyst of emerging technical areas with sufficient explanatory evidence to support further exploration

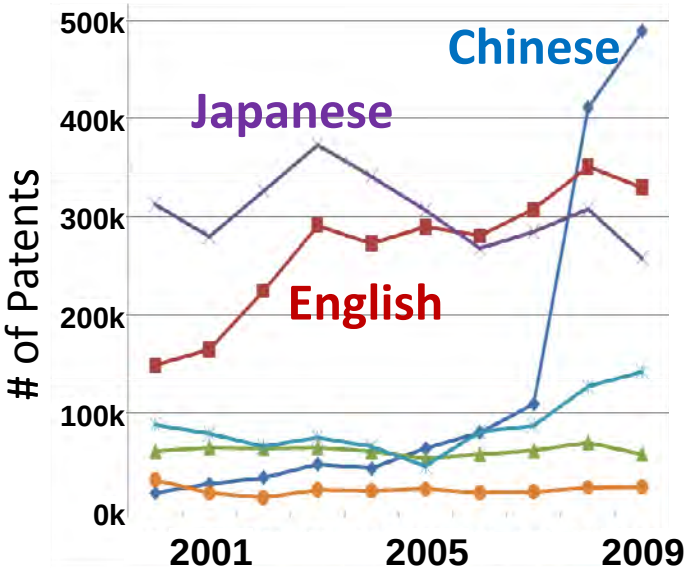
Worldwide Scientific and Patent Literature

Publications (by Language) and Patents



Growth in scientific and patent literature is estimated at 800k docs/month

Patents by Language



Other Languages

Korean (cyan)
German (green)
Russian (orange)

FUSE Approach

Today, *ad hoc* "technical horizon scanning" already consumes substantial expert time, is narrowly focused on a small number of topics, and is subject to limited systematic validation.

Analysts need a reliable and transparent capability to scan continually for signs of technical capability emergence.

Process full-text documents to identify, prioritize, and nominate technical areas for analysts with supporting evidence.

Evaluate progress with case studies, expert judgment, and quantitative tests.

Today	FUSE
Manual	Automatic
Selected coverage	"Complete" literature coverage
Updated infrequently	Updated monthly
Months to produce (for one technical area)	24hrs to produce (for all technical areas)
Ad hoc evaluation	Formal models of emergence

Complete, Continual, Unbiased

Key Technical Challenges

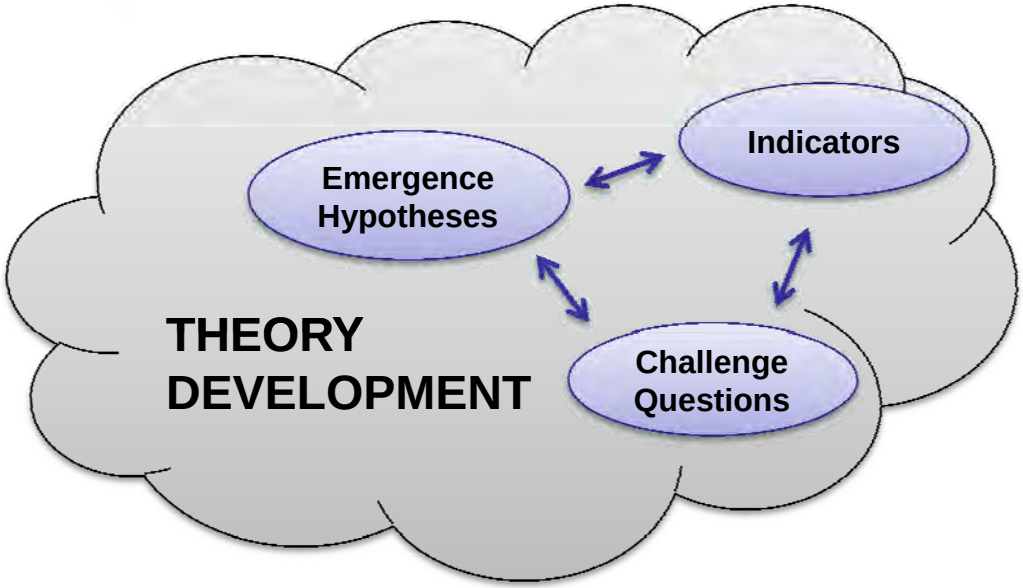
automated detection of emerging concepts, methods, technologies, …

- **Tech Emergence Theories:** Develop, test, and validate indicators and establish models / theories of technical emergence
- **Text-to-indicators:** Process multidiscipline, multilingual, and noisy full-text literature
  - Extract usable within-document and cross-document features *(e.g., rhetorical stance on citations, methods, applications, …)*
  - Connect features in context, create indicators, and learn patterns *(e.g., feature / indicator fusion, time series summarization & analysis, …)*
  - Generate groupings of documents that meaningfully represent "the what" that is emerging – Related Document Groups (RDGs)
- **Nomination, Evidence:** Identify, prioritize & nominate technical areas, and provide evidence with understandable explanations
- **System Integration and Engineering**: Effective virtualized service orchestration

Phase 1 Hypotheses:

1. Emergence of scientific and technical capability can be detected for selected case studies
2. Features derived from full-text are critical for detecting technical emergence signatures
3. Modules implemented will enable performers to develop prototype systems in Ph. 2 that can scan for technical emergence at scale, across time.

FUSE Theory Implementation



Theory Development

- Refers to the performer activity in response to BAA Section 1.B.1 calling for fundamental advances in our understanding of how the **real-world processes of technical emergence leave discernible traces** in the public scientific, technical, and patent literature, and how those traces can be detected.
- Theory **development will be ongoing** through the full FUSE program and is implicitly measured through key program metrics
- Highly variable, **team-specific process**; can capture essence and progress through *emergence hypotheses, indicators,* and *challenge questions.*

Emergence Hypotheses

- Concise statements that capture one or more properties of what it means for a concept to emerge
- Posited by each performer team; revision is expected throughout the program, but not on a continuous basis

Indicators

- Quantitative, computable, measures that relate to some aspect of an emergence hypothesis
- Posited by each performer team; expect many indicators to be proposed and implemented (~10-20 per team)

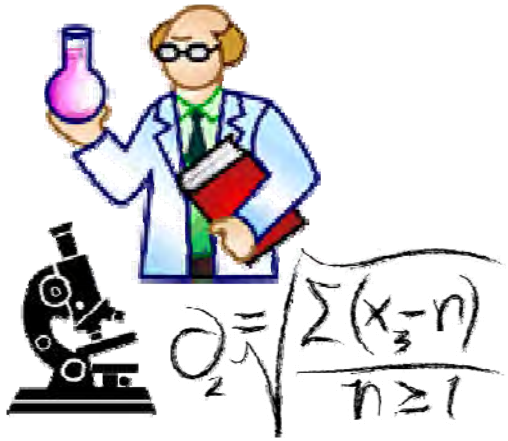
Challenge Questions

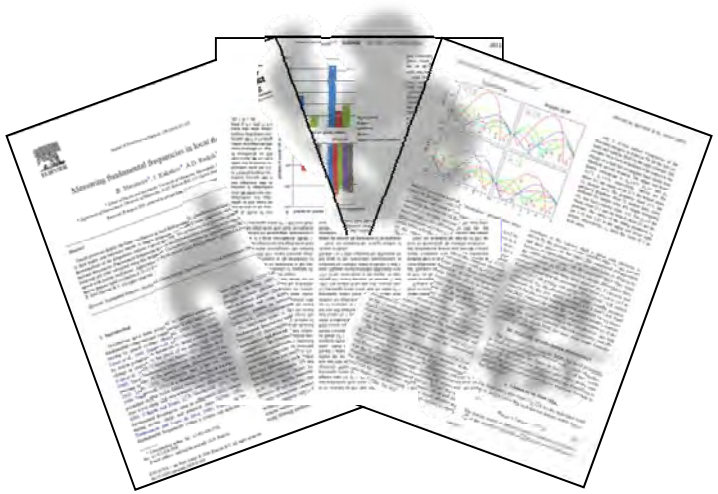
- Mechanism to test and evaluate the relationship between emergence hypotheses and indicators
- Common set defined by Government; performer-posited challenge questions are being reviewed by the Government team

What is technical emergence?

(*Current* "Emergence Hypotheses")

- A concept has emerged if it has been accepted by others within and beyond one's community. **~Columbia**
- A concept is emerging when its actant network is increasing in robustness. **~BAE**
- A concept has emerged when evidence has appeared that the concept is new and unexpected, noticeable and growing. **~Raytheon BBN**
- A concept is emerging when it is identifiable by its own practitioners, enables a capability that was not achievable previously, and persists. **~SRI**





How do we probe "Technical Emergence"?

Government-Defined Challenge Questions

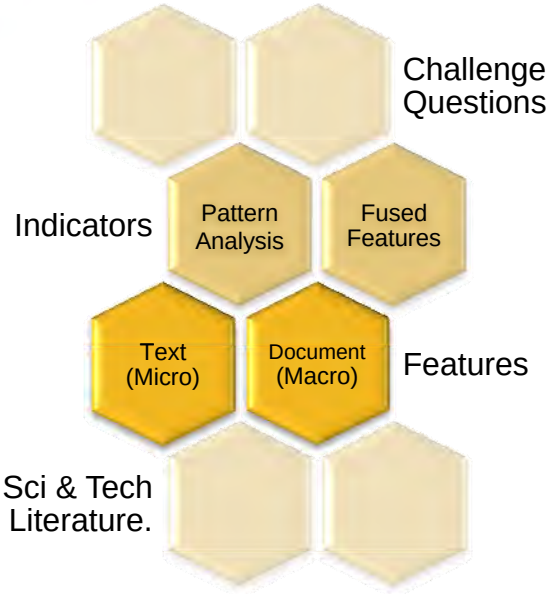
- Was there a **community of practice** around <concept> during <time period>?
- Were there **debates** within the scientific community on <concept> during <time period>?
- Was there a demonstration of **practical application** of <concept> during <time period>?
- Was <concept> considered an **alternative** to an established concept during <time period>?

- *Was there a demonstration of commercial application of <concept> during <time period>?*
- *Was the infrastructure required to perform research in <concept> readily available during <time period>?*

Performer-Defined Challenge Questions

- Was <concept> accepted during <time period>? ***Columbia***
- Did the acceptance of <concept> increase or decrease during <time period>? ***Columbia***
- How interdisciplinary was the scientific and technical knowledgebase around <concept> during <time period>? ***SRI***
- Did usage of new terminology describing <concept> increase in robustness during <time period>? ***BAE***
- How many citations of papers from <concept> published in <time period 1> would you expect to see in <time period 2>? ***Raytheon BBN***
- Does the <concept> dominate a thread during <time period>? ***Raytheon BBN***

Text to Indicator Implementation



Text to Indicators

- Refers to the performer activity in response to BAA Section 1.B.2 calling for **fundamental advances in the exploitation of the full-text** of technical documents, in addition to the document's metadata
- This research thrust captures the **bulk of the performer time and effort** with multiple subteams on each performer team
- **Feature specific evaluation is performer-centric** (defined in waypoints) and often leverages intrinsic evaluations; overall evaluation of research is at the Nomination Quality metric level and is extrinsic, relative to case studies

Text to Features

- All performer teams have demonstrated progress on FUSE specific development in identifying and extracting / generating features from the FUSE document repository
- All performer teams have efforts focused on the textual level (micro) and the document level (macro); relative balance of feature level varies between teams (Columbia = text focus; BAE = balanced; Raytheon BBN, SRI document focused)
- Little redundancy between teams; multiple teams are pursuing broad topics such as 'topic modeling' and 'citation context', but the details reveal significantly different motivations, approaches, and applications of these features

Nomination and Evidence Explanation – Ongoing Work

RNAi : 2006-2010 : CQ1

Was there a community of practice around RNAi during 2006-2010?

The answer is YES, with a confidence of 72%

Topic Summary

RNA interference (RNAi) is an RNA-dependent gene silencing process within living cells. The selective and robust effect of RNAi on gene expression makes it a valuable research tool, both in cell culture and in living organisms.

Justification and evidence for answer

Many indicators suggest a positive answer to the CQ, especially within the Coauthorship Graph, Time Series and Funding groups (see below).

Coauthorship Graph

Click for detailed view

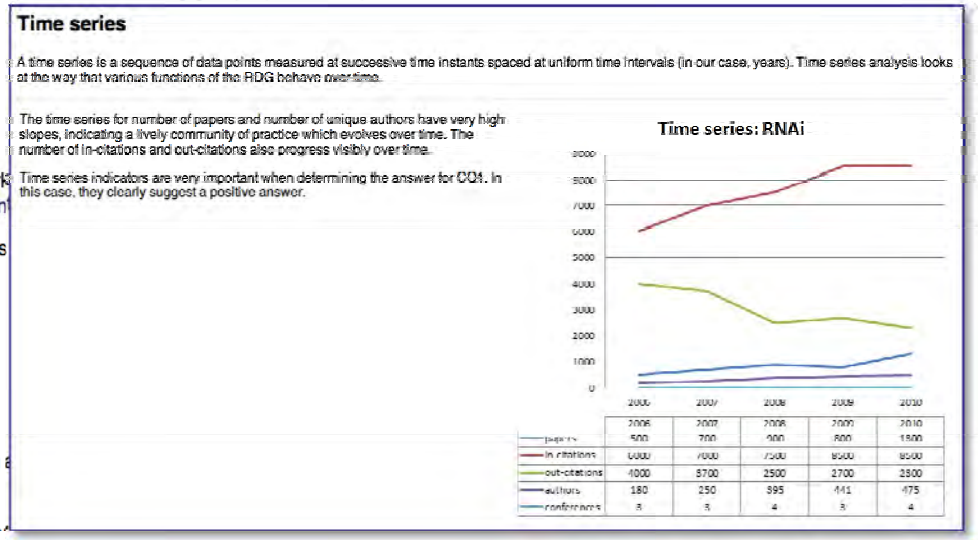
The coauthorship graph for RNAi spans 520 authors, and it has the properties of a small-world network connected network with a high clustering coefficient as well as a reasonably high power law coefficient

Coauthorship graph indicators are the most powerful when determining the answer for CQ1, and in this positive answer.

Time Series

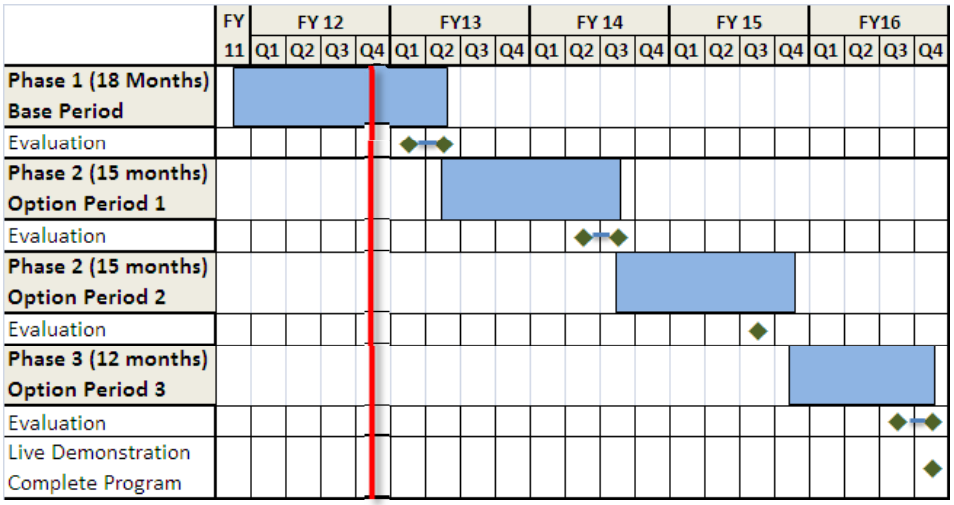
Click for detailed view

The time series for number of papers and number of unique authors have very high slopes, indicating a The number of in-citations and out-citations also progress visibly over time.



FUSE Status Update

- Five year, fundamental research program
- Teams under contract since August 2011
  - BAE Systems
  - Columbia University
  - Raytheon BBN
  - SRI International



- Formal test and evaluation to begin October 2012, three additional rounds of formal evaluation scheduled
  - Case Studies, Eight Examples: Tissue Engineering, Cold Fusion, RF Metamaterials, DNA Microarrays, Genetic Algorithms, RNAi, Steganography, Horizontal Gene Transfer

If you had a system that could

(a) reliably identify what technical capabilities are emerging
and
(b) provide humanly understandable evidence explanation
then

how would you use this tool?

How might an IARPA PM use FUSE?

- Idea development (a problem space)
  - Optimal emergence state for an R&D program
  - Hype versus reality
  - New enabling component capabilities
  - Signs of potential convergence
  - Overlap with past and existing efforts
  - Why is this innovative?
- Program impact assessment
  - Impact within the research community
  - Application development or potential

Would the components of FUSE enable a new capability or enhance an existing workflow?

Potential Component Services

- Lineage / Provenance of Acquired Data
- Data Transformation (e.g., PDF to XML)
- XML Enrichment (e.g., new extractions, zoning)
- Cross-Document "Feature" Linkage
- Generation of Document Groupings
- Indicators (e.g., Confirming, Dis-confirming)
- Nomination, Prioritization by Time Period
- Evidence Explanation and Justification for Nomination

A Few Unique Qualities of FUSE

- **Connecting disparate communities with lots of data sows the seeds for much discovery**
  - Social science, emergence theory, Natural Language Processing (NLP), etc.
- **Scientific, Technical, and Patent literatures are new genres**
  - Most Natural Language Processing (NLP) trained on news wire
  - Foreign language tokenization, etc., too
- **Novel extrinsic test for technical emergence**
  - Measuring detection of emergence (S&T)
  - Not measuring NLP metric X *(internal measures)*
- **Challenging scale, but tractable**
  - Concept extraction, within/cross-doc linkages
  - Currently growing on order 800k/mo. (sci lit + patents), FUSE has ~75-80% filed patents & ~10% sci lit (1980-2010)
- **Automated evidence explanation is really challenging, but exciting**
- **Synthesize & Scan "Horizon", not Search**

Anticipated Impact

- **Scientific & Technical Analysis Impact**
  - Relevant, timely, and bias-controlled analytic force multiplier to maintain technical vigilance, across all disciplines and multiple languages
  - Discover previously unknown emergence signals of interest at speed, scale, and comprehensiveness that exceeds human capacity
- **Technical Impact**
  - Generalized and validated theories of technical emergence
  - New cross-document conceptual feature extraction technologies
  - Progress in computer-generated evidence representations for human use
- **Secondary Impact**
  - Improved priority filter for USG investment strategies and policy

The FUSE Team

- 3 (+2) large businesses
- 3 (+4) small businesses
- 14 academic orgs
- 1 not-for-profit org
- Many data vendors
- Plus FFRDC & gov orgs

Questions

Dewey Murdick, Ph.D.
FUSE Program Manager, IARPA
dewey.murdick@iarpa.gov