

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE



Finding Patterns of Emergence – Foresight and Understanding from Scientific Exposition (FUSE)

L E A D I N G I N T E L L I G E N C E I N T E G R A T I O N

Dewey Murdick, IARPA
9 January 2014



I A R P A

BE THE FUTURE

“Invests in high-risk/high-payoff research programs that have the potential to provide our nation with an overwhelming intelligence advantage over our future adversaries.”

<http://www.iarpa.gov/>



Goal: Validated, early detection of technical emergence

Reduce “technical surprise” via 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

Special focus from the outset on multiple languages, Phase 2 focus on **English** and **Chinese**

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



What is technical emergence?

Hypotheses from Phase 1

- 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**

Many ways to probe technical emergence

- Community of Practice
- Practical Application
- Debates
- Alternative
- **Acceptance**
- **Interdisciplinarity**
- **Attention (Citation)**
- **Prediction**
- **Dominant sub-topic within set**
- **Commercial Application**
- **Infrastructure**



Multiple research goals are being explored

- Develop refined theories of technical area emergence
 - as expressed by hypotheses and implemented in quantifiable indicators
 - demonstrate indicator viability in English and Chinese across disciplines
- Automatically process and prioritize English-language surrogate entities, and nominate, with evidential support, those exhibiting technical emergence
 - Across multiple disciplines (i.e., ten disciplines are explored for the document-based forecasting challenge)
 - Across multiple entity classes (e.g., people, terms, documents)
 - For a defined set of surrogates of emergence
- Demonstrate a stable, working, and robust capability for Chinese to process and prioritize surrogate entities
- *Demonstrate through small scale experiments whether or not full-text scientific article and patent body content enhances predictive capability (over metadata-only sources) of technical emergence in English*



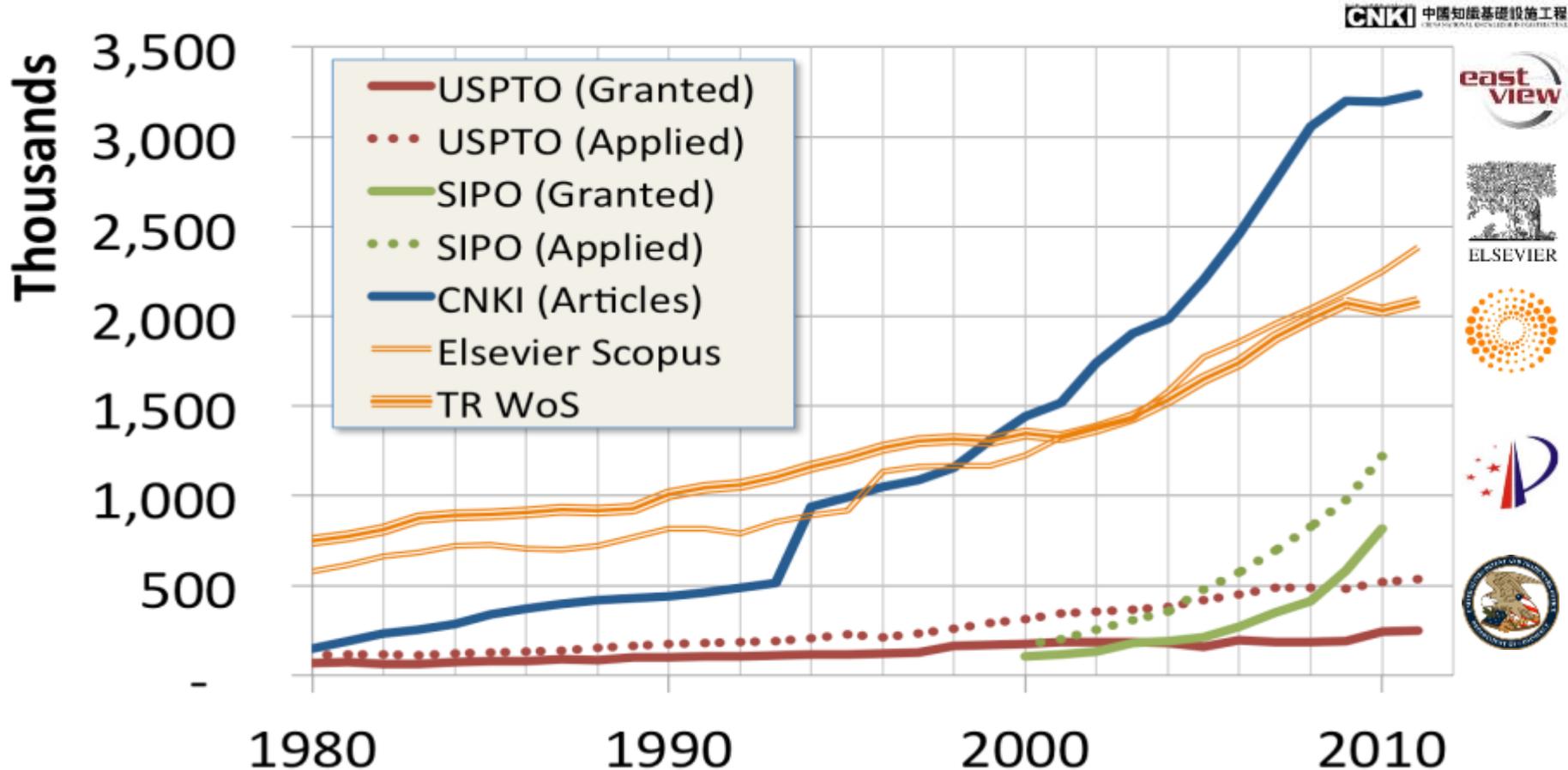
FUSE Status Summary

- Four year, fundamental research program with four teams under contract since August 2011
- Phases for exploring technical emergence...
 - Phase 1: Detect within a consistent theoretical construct / All teams met or exceeded metrics and targets
 - Phase 2 Option Period 1: Forecasts (limited models) at scale in English scientific and patent literatures; Chinese patents
 - Phase 2 Option Period 2: Forecasts (rich set of models) at scale in English and Chinese scientific and patent literatures

	FY	Year 1 (FY11)				Year 2 (FY 12)				Year 3 (FY13)				Year 4 (FY 14)				Year 5 (FY 15)			
	10	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Solicitation/Award																					
Phase 1 (18 Months) Base Period																					
Evaluation																					
Phase 2 (15 months) Option Period 1																					
Evaluation																					
Phase 2 (15 months) Option Period 2																					
Evaluation																					
Live Demonstration																					



Scientific and Patent Literature



Growth estimated at ~35k unique docs/month for FUSE; worldwide ~800k docs/month



FUSE: Example Nominations

String	Translation	Pe
北京北方微电子基地设备工艺研究中心有限责任公司	Beijing North Micro Base Equipment Technology Research Center, LLC	0.967
鸿准精密工业股份有限公司	Foxconn Technology Co., Ltd.	0.962
布劳恩股份有限公司	Braun Corporation	0.916
瑞尼斯豪公司	Raines Howe Company	0.892
东洋制罐株式会社	Toyo Seikan Corporation	0.880
塔工程有限公司	Tower Engineering Co., Ltd.	0.865
大王制纸株式会社	King Paper Co., Ltd.	0.857
ut 斯达康通讯有限公司	ut Starcom Communications Limited	0.830
安德烈亚斯.斯蒂尔两合公司	Andreas Stihl co.	0.810
南方医科大学	Southern Medical University	0.802

Three year forecast for most prominent filers in Chinese Patent Office

N-gram	Pe
germline stem cell	0.688
atp-powered	0.625
genetic or pharmacological	0.625
properties of amyloid	0.571
knockout mice showed	0.571
laser capture microdissection	0.547
strain of influenza	0.519
jnk phosphorylation	0.500
ecological speciation	0.500
signalling(1,2	0.500

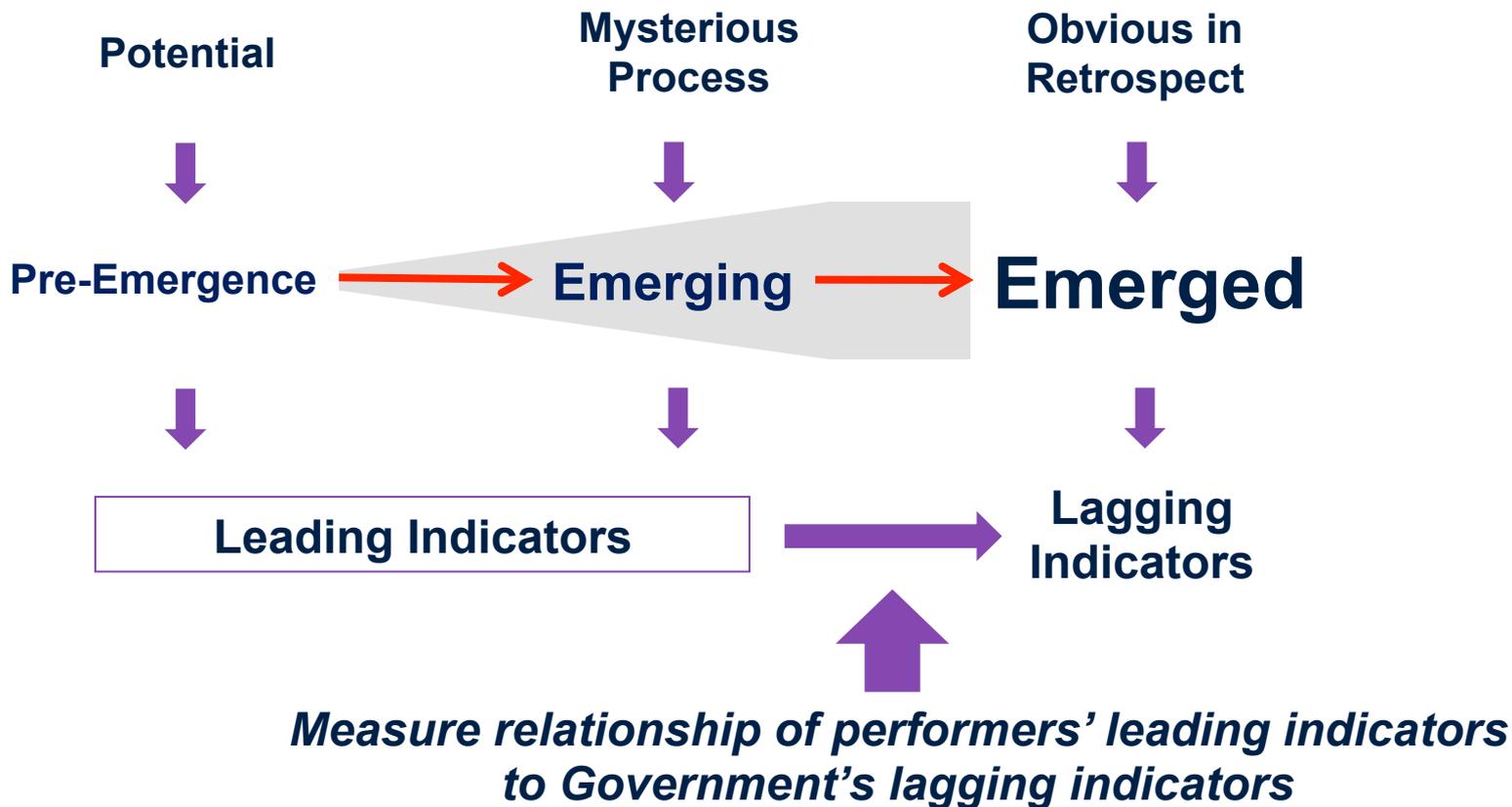
Two year forecast for most prominent terms in English-language scientific papers (Web of Science)

String	Translation	Pe
传递转矩	Transmission torque	0.839
柔性线路板	FPC	0.825
健康发展	Healthy development	0.821
部分耦合	Part of the coupling	0.803
纳米空心球	Hollow nanospheres	0.800
拓扑信息	Topology information	0.798

Three year forecast for most prominent patent term in Chinese Patent Office



Phase 2 T&E Foundations: Leading and Lagging Indicators



FUSE Research Thrusts

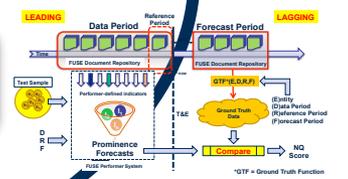
Theory & Hypothesis Development
Supports indicator development and explanation; a robust theory is unlikely

Document Features
Patents, S&T Lit

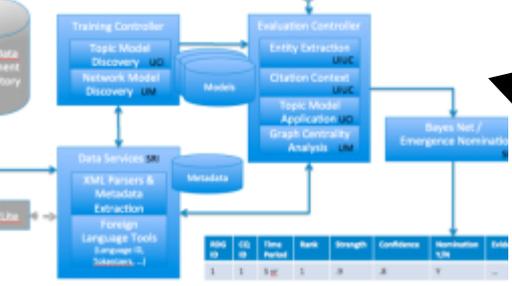
Indicator Development
Leading indicators

- clustering c
- power law
- number of

Nomination Quality
Forecast formulation



System Engineering



Evidence Representation

RNAI : 2006-2010 : CQ

Was there a community of practice around RNAI during 2006-2010?
 The answer is **YES**, with a confidence of **72%**

Many indicators suggest a positive answer to the CQ, especially within the **Coauth Coauthorship Graph**

Click for detailed view

The coauthorship graph for RNAI spans 520 authors, and it has the properties of a small-i communities. It is a fully connected network with a high clustering coefficient as well as a Coauthorship graph indicators are the most powerful when determining the answer for CQ the direction of a positive answer.

Prominence of surrogate entities of emergence

DOCS TERMS



Storyboard Example

Patent Forecast

Patent: 6539080 - Method and system for providing quick directions

Reference Period: 2004

Forecast Period: 2007

Predicted Prominence: 0.781417

Applicants

Assignees

Citations

<u>Evidence</u>	<u>Value</u> ▼	<u>Weight of Evidence</u> ▶
<u>Number of prior art references</u>	46.000000	0.082000
<u>Number of patent inventors</u>	4.000000	0.078439
<u>Qualification for Emerging Cluster</u>		
<u>Growth of the citation index of patents</u>		
<u>Slope of the regression line for countries per year</u>		

Organization Forecast

Organization: 重庆长安汽车股份有限公司

Reference Period: 2007

Forecast Period: 2009

Predicted Prominence: 0.950000

<u>Evidence</u>	<u>Value</u> ▼	<u>Weight of Evidence</u> ▶
<u>Slope of the Regression line for patent counts</u>	45.200001	0.739408
<u>Growth of annual patent counts</u>	7.167742	0.706099
<u>Slope of the regression line for technology areas the organization is prolific in</u>	1.300000	0.739408
<u>Growth of the technology areas the organization is prolific in</u>	0.700000	0.523528



Indicator Development and Testing Underway

Regular analysis and evaluation of each team's features (e.g., scientific noun phrases, topic models) and their portfolio of indicators (i.e., quantitatively measured aspects / patterns of technical emergence)

Promising Midterm Indicator Types

- Citation, Author Networks (All)
- Topic Diversity (SRI)
- Citation Context and Sentiment (SRI)
- Technology and application concept type evolution (SRI)
- Patent classification dynamics (SRI, BAE)
- Emerging cluster / hot patent status (BAE)
- Length of independent claims (BAE)
- Patent originality (BAE)
- Corporate, Academic patent authorship (BAE)
- Topic modeling across time, thread dynamics (BBN)
- Research levels (BBN)
- Time series analysis, extensive portfolio (COL)
- Temporal pattern classification, time-series clustering (COL)

Fundamental Research

- Argumentative Zoning (SRI, COL)
- Time-dependent term co-occurrence (SRI)
- Author-topic modeling (SRI)
- Operations on annotated graphs, e.g., scientific concepts, terms (SRI)
- Chinese patent indicators (BAE, BBN)
- Fine-grained topic models (BBN)
- Causality modeling framework (BBN)
- Primary concept mentions (COL)
- Citation sentiment (COL)



Technical Horizon Scanning - FUSE

Horizon Scanning systematically gathers a broad range of information about emerging issues and trends in an organization's political, economic, social, technological, or ecological environment.

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 to scan continually for signs of technical capability emergence.

Complete, Continual, Controlled Bias

Today	FUSE
Manual	Automatic
Selected coverage	“Complete” literature coverage
Updated infrequently	Updated with data
Ad hoc evaluation	Validated indicator hypotheses



Technical Horizon Scanning - ForeST

Forecasting Science and Technology – ForeST Program

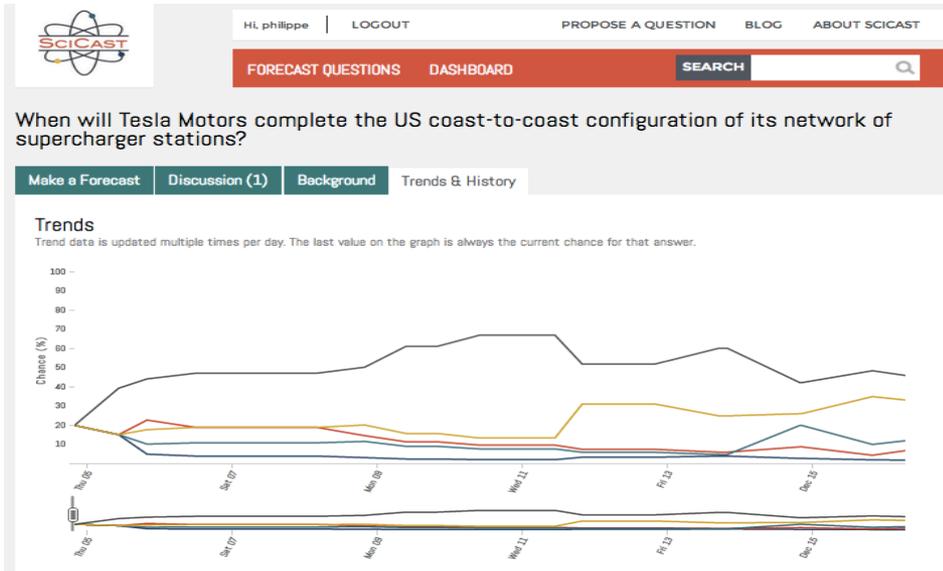
Goal: Generate precise, testable forecasts for S&T developments by combining the judgments of thousands of subject matter experts.

Approach: Build the world's largest prediction market for S&T events. Thousands of subject matter experts in dozens of countries will make nuanced conditional forecasts for ~1,000 S&T events per year. Data-driven (i.e., scientific and patent literatures) indicators will be used to generate questions and adjust forecasts, see scicast.org.

Evaluation: Forecasts are scored against actual events, as they occur, and against alternative forecasting methods used in academia, industry, and the IC.

Potential impact: Dramatically improve the IC's foresight of worldwide technical capabilities with actionable information.

Schedule: June 2013 - June 2015.





Technical Horizon Scanning - ForeST

Teams will generate questions, such as:

- What is the probability of a 10cm carbon nanotube being fabricated before 31 December 2014?
- Will the number of accepted articles for the 2015 International Conference on Machine Learning (ICML) conference that contain the term 'deep learning' in the title/abstract exceed those that contain the term 'support vector machine(s)' in the title/abstract?
- How many unique assignees will have at least two USPTO patent applications published using the term 'Type III Secretion System' in its title/abstract/background/claims between 1 October 2013 and 30 September 2014?
- By 31 December 2017, how many FDA-approved products will be based on RNA interference?
- Will there be reported shortages of technetium-99m in the US in 2015?



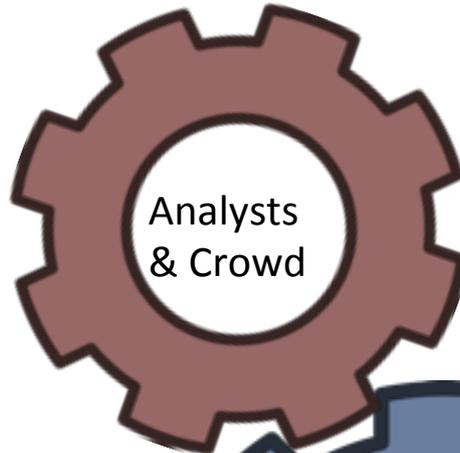
Horizon Scanning helps to anticipate surprise...

- Are new discoveries, advancements, applications, etc.:
 - Creating an opportunity to improve a functional capability for me? Others? (May require additional investment.)
 - Invalidating or outdated my assumptions? Others?
 - Changing the intent of an industry, a decision maker, influencer, or “the crowd?”
 - Improving or degrading a functional capability (relative or absolute) of an entity?
 - Forcing an action by an entity (legal, diplomatic, military, cyber, information sharing, economic, investments, ...)?
- Are changes in the environment (climate, resources, disease, infrastructure, ...) doing any of the above?



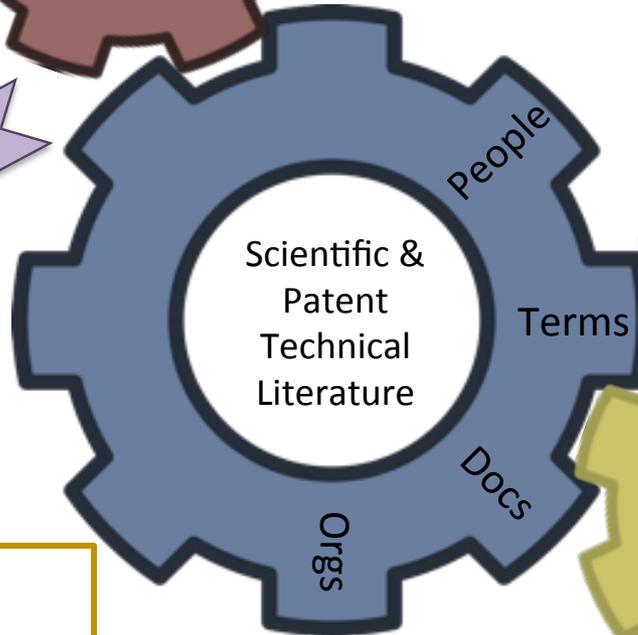
Horizon Scanning helps strategic planning...

- Resource allocation
- Hiring strategies
- Portfolio analysis
- R&D program development
- Competitive analysis
- Expert or claim credibility assessment
- R&D program impact assessment analysis
- Alternate futures analysis, foresight analysis
- And others...



Horizon Scanning integrates FUSE, ForeST, and additional analysis to forecast technical adoption, application, and impact

ForeST elicits crowd judgments about performance and applications



FUSE focuses on data-driven forecasting from S&T literature in generically applicable ways



Anticipated Impact

- **Scientific & Technical Analysis**
 - 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
- **Underlying Technologies**
 - Generalized and validated theories of technical emergence
 - New cross-document conceptual feature extraction technologies
 - Progress in computer-generated evidence representations for human use
- **Broader Applications**
 - Improved priority filter for USG investment strategies and policy



The FUSE Team

COLUMBIA ENGINEERING
The Fu Foundation School of Engineering and Applied Science



BAE SYSTEMS



NEW YORK UNIVERSITY



Rensselaer

Raytheon
BBN Technologies



UMASS
AMHERST



SciTech
Strategies



Penn
UNIVERSITY OF PENNSYLVANIA



UCIRVINE



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



MITRE



NORTHROP GRUMMAN

Booz | Allen | Hamilton



AVIAN
ENGINEERING LLC

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



Tarragon
CONSULTING CORPORATION



Questions



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