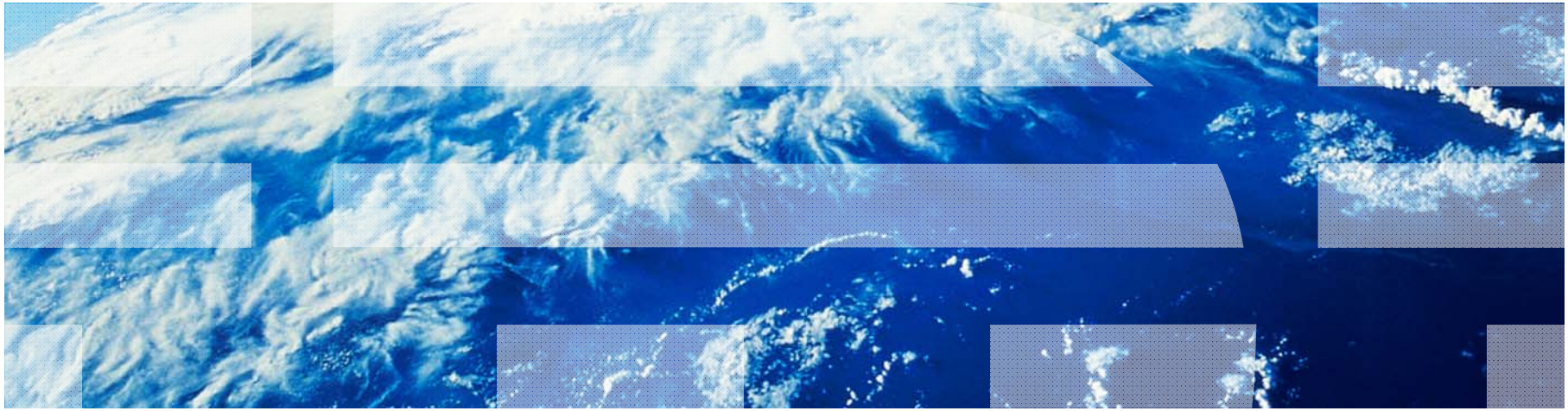
A Strategic Approach to Unlock the Opportunities from Big Data

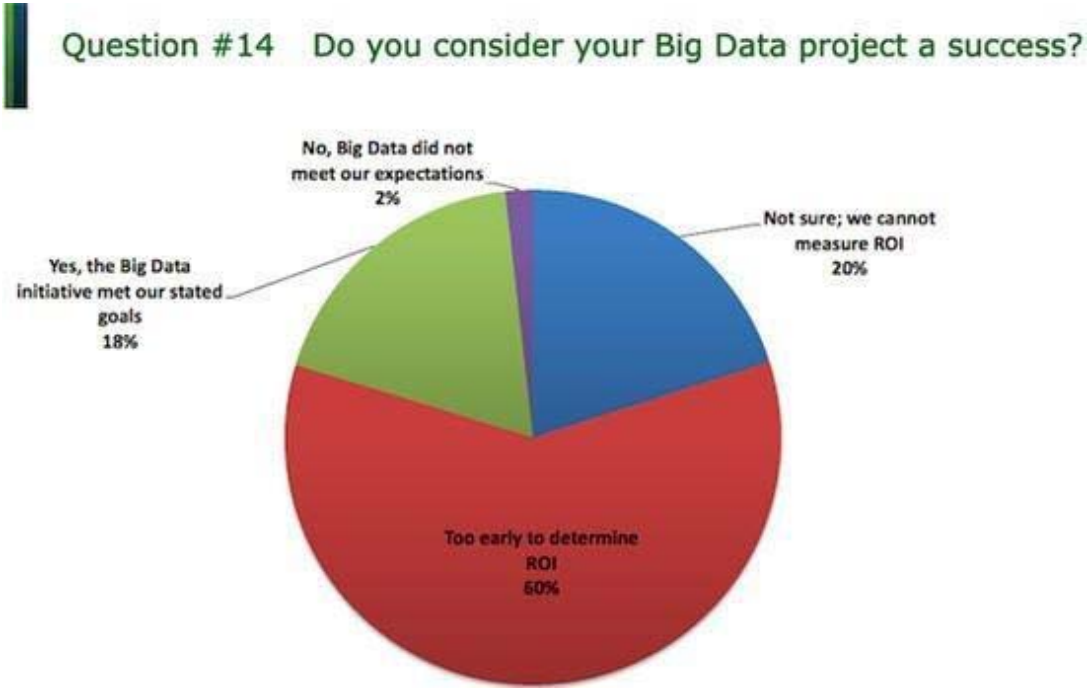
Yue Pan, Chief Scientist for Information Management and Healthcare
IBM Research - China
[contacts: panyue@cn.ibm.com]



Big Data or Big Illusion?

Much of the focus on the big data zoo has missed one key point: big or small, it's still data. It must be managed and integrated across the entire enterprise to extract its full value, to ensure its consistent use.

Barry Devlin, "The Big Data Zoo --- Taming the Beasts "



*Source: Gartner, "…… "

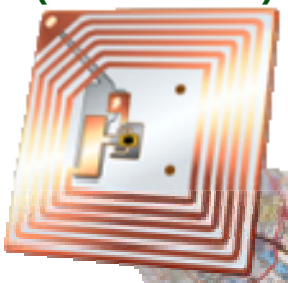
A Bird's Eye View of Big Data

12+ TBs **of tweet data every day**

? TBs **of data every day**

25+ TBs **of log data every day**

30 billion **RFID tags today (1.3B in 2005)**



4.6 billion **camera phones world wide**

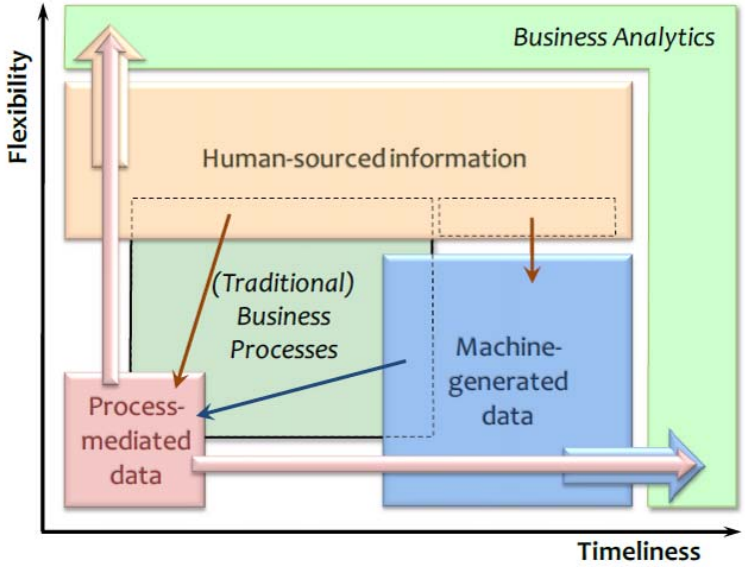
100s of millions of GPS enabled **devices sold annually**

76 million **smart meters in 2009… 200M by 2014**

2+ billion **people on the Web by end 2011**

A Bird's Eye View of Big Data

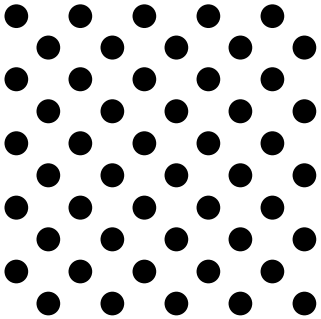
*The three domains of information**



*Source: Barry Devlin, "The Big Data Zoo --- Taming the Beasts "

The fourth dimension of Big Data: Veracity – handling data in doubt

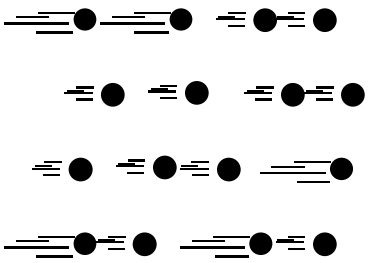
Volume



Data at Rest

Terabytes to exabytes of existing data to process

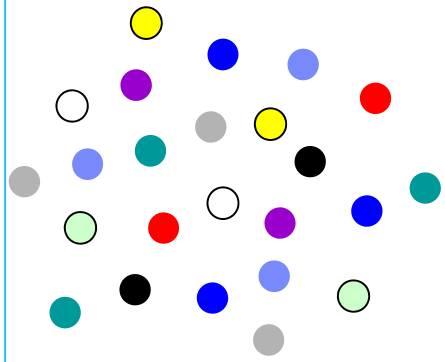
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

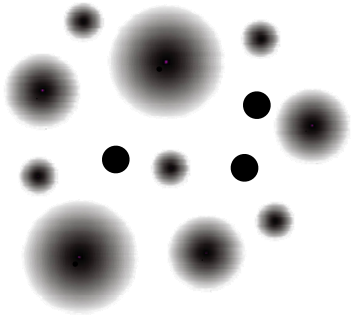
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

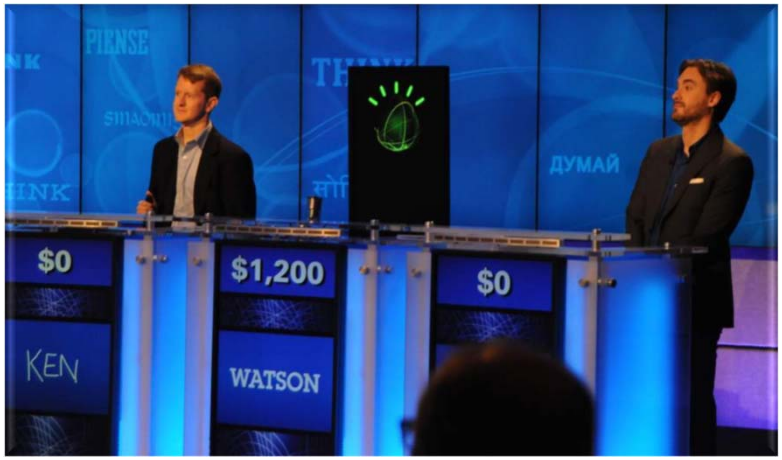


Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

* Truthfulness, accuracy or precision, correctness

Tame Big Data, Turn into Insight - Example: IBM Watson

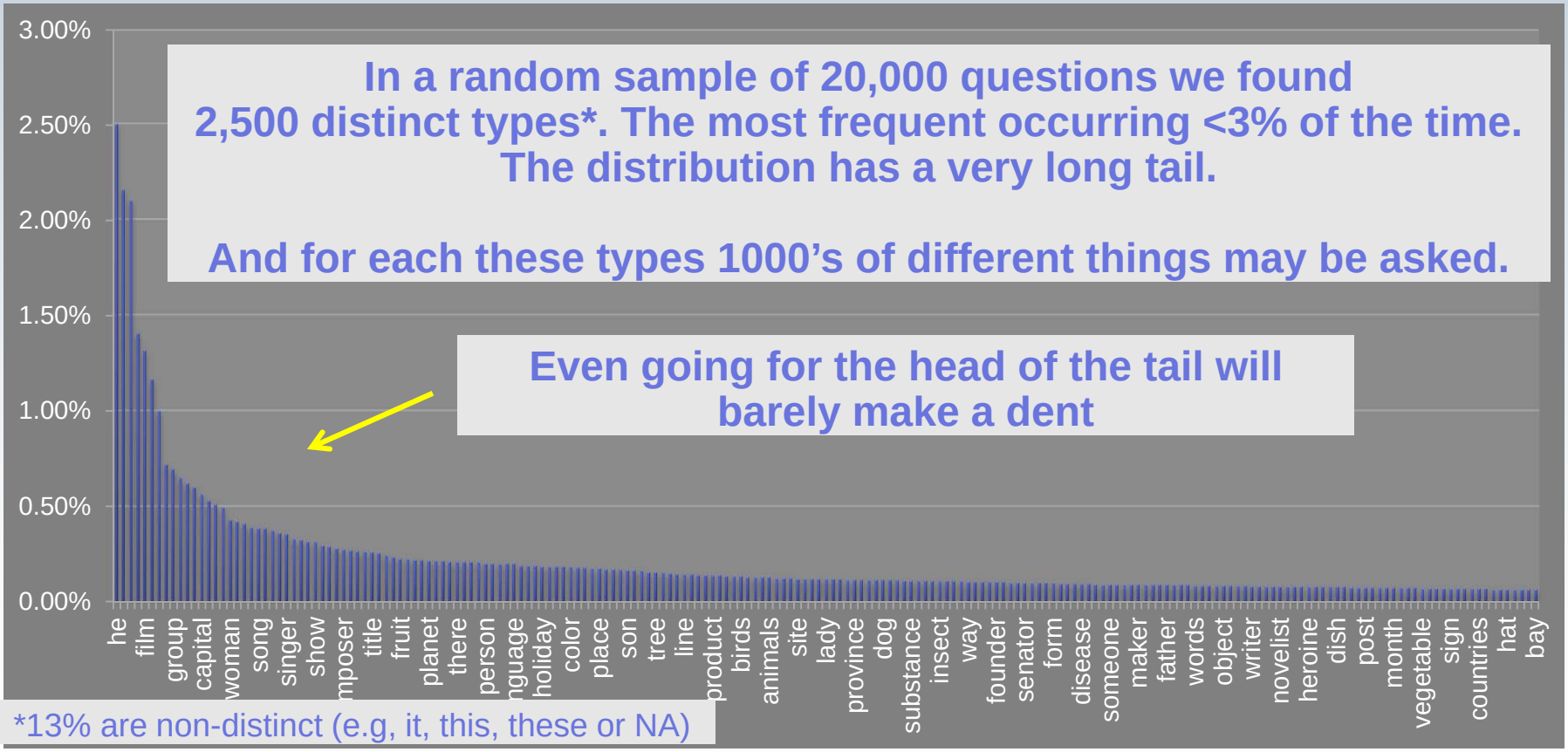


Watson's advanced analytic capabilities sort through the equivalent of 200 MILLION pages of data to uncover an answer in 3 SECONDS.

Jeopardy Challenge – the Broad Domain

We do NOT attempt to anticipate all questions and build databases.

We do NOT try to build a formal model of the world



In a random sample of 20,000 questions we found 2,500 distinct types*. The most frequent occurring <3% of the time. The distribution has a very long tail.

And for each these types 1000's of different things may be asked.

Even going for the head of the tail will barely make a dent

*13% are non-distinct (e.g, it, this, these or NA)

Our Focus is on reusable NLP technology for analyzing vast volumes of *as-is* text.
Structured sources (DBs and KBs) provide background knowledge for interpreting the text.

Algorithms built in Watson

- **Question and Content Analysis**
  - ESG Tokenizer and Deep Parser
  - R2 Named Entity Detector
  - Predicate Argument Structure Annotator (Logical Forms)
  - Shallow Semantic Relation Detector
  - Focus and LAT detection
  - Sentence & Intra-paragraph Anaphora Resolution (co-ref)
  - Question decomposition and classification
  - Deep Semantic Relation Detector
  - Large Scale Relation Detection
  - True-Casing
  - Text Alignment
- **Search**
  - Indri Document Search (short and long docs)
  - Indri Passage Search (regular and TIC)
  - Lucene Passage Search
  - RDF/KB Search
  - N-Gram Search
  - Language Mining (Frame Structure Queries)
- **Hypothesis Generation**
  - Document Title (for title-oriented sources)
  - Anchor Text, Title Matching
  - Quoted Text
  - Sentence Completion
  - Spreading Activation and Missing Link
  - Type-Based
  - Predicate Argument
  - Question Inversion
  - KB
- **Structured Inference**
  - Geo-Spatial
  - Temporal
  - Domain Specific Inference
- **Evidence Scoring**
  - LAT to Semantic Type Matching
  - Mined Lists, Thesauri and Folksonomy Matching
  - Statistical Context Typing
  - Introduction Typing
  - Identity and Gender Typing
  - Target Ontology Typing
  - Ngrams
  - Spreading Activation
  - IDFScorer – relative frequency in corpus
  - Doc Term Match
  - Textual Alignment
  - Passage Term Match
  - Backlink Scorer, Title Fraction –
  - Sattack – scores answer-bearing sources (source reliability)
  - Graph Matching/ Logical Form Analysis
  - Pattern Based and Statistical Transformation Logic
- **Supporting Evidence Retrieval**
  - Supporting Passage Search
  - Knowledge-Base and Spreading Activation
- **Final Merge**
  - Type-Based Answer Merging (x-doc, co-reference)
  - Ranking and Confidence Estimation – Logistic Regression
  - Learning Model over features
  - Multiple Models

Most Client Use Cases Combine Multiple Technologies

Pre-processing

Ingest and analyze unstructured data types and convert to structured data

Combine structured and unstructured analysis

Augment data warehouse with additional external sources, such as social media

Combine high velocity and historical analysis

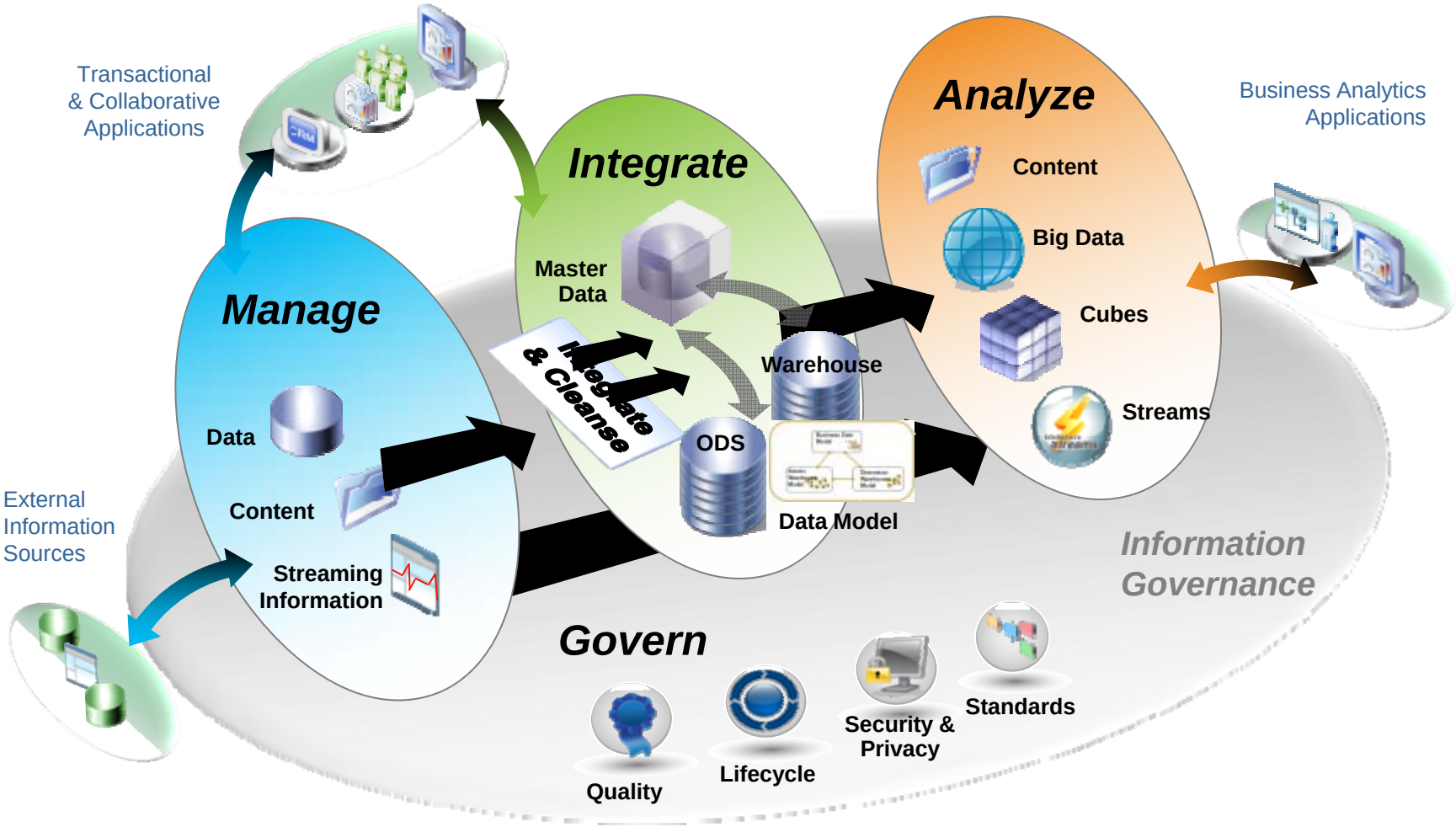
Analyze and react to data in motion; adjust models with deep historical analysis

Reuse structured data for exploratory analysis

Experimentation and ad-hoc analysis with structured data

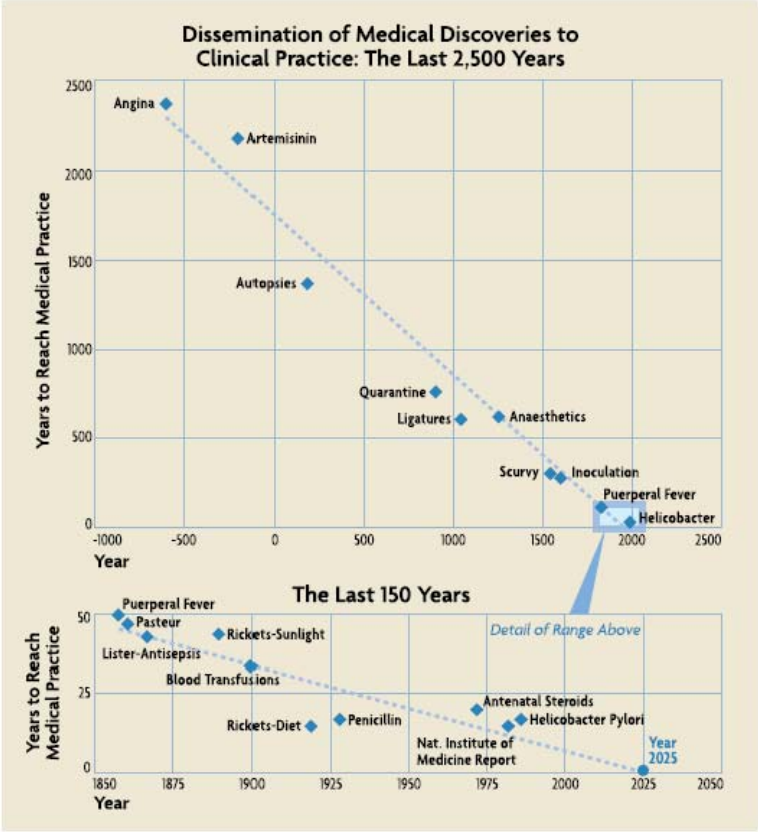
Advanced analytics requires a robust, comprehensive information platform

Trusted ◆ Relevant ◆ Governed

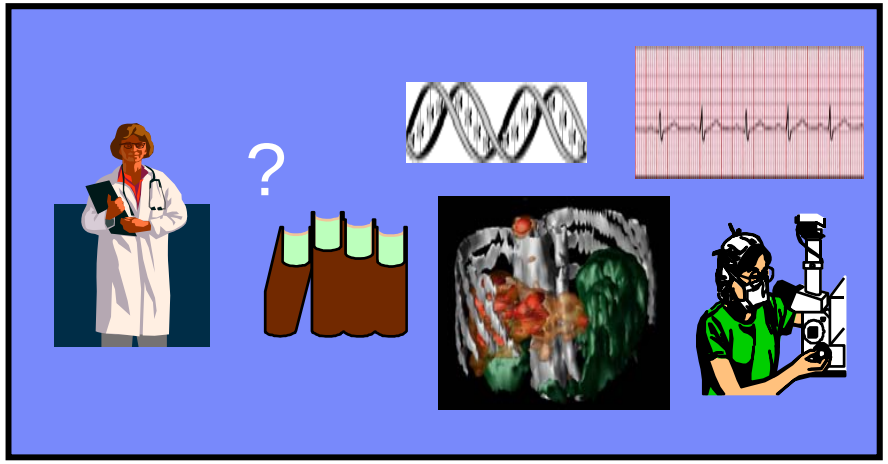


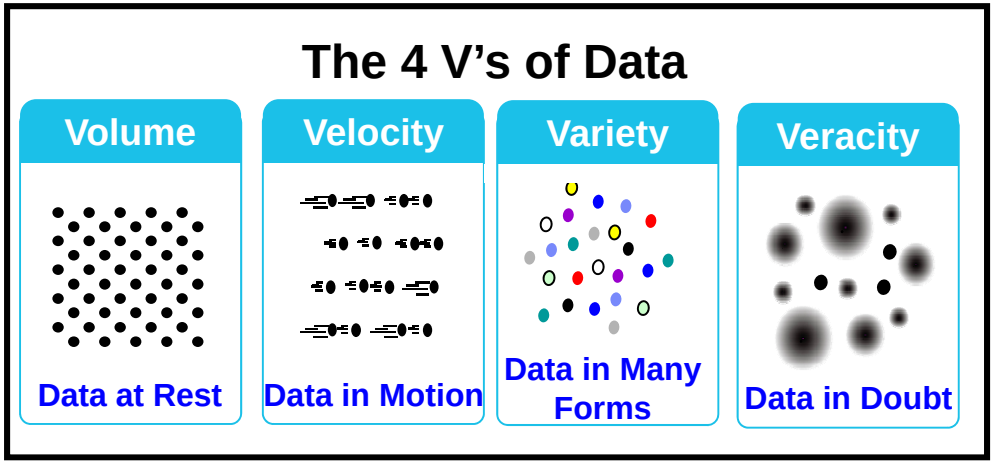
Big Data for Research and Innovation

- Based on empirical research or simulation results
- Exploit intensive computation and big data technology
- Combine domain expert's knowledge and data scientist's skills



Research: the road from data to foresight is long and expensive

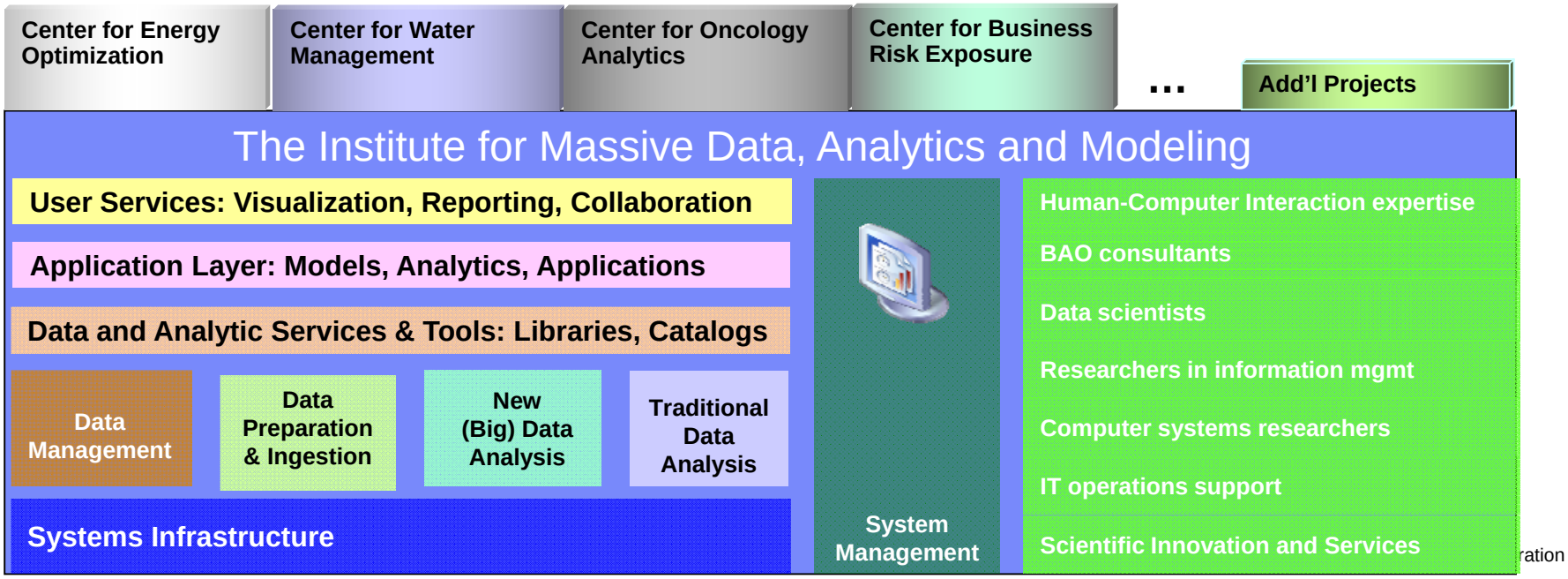




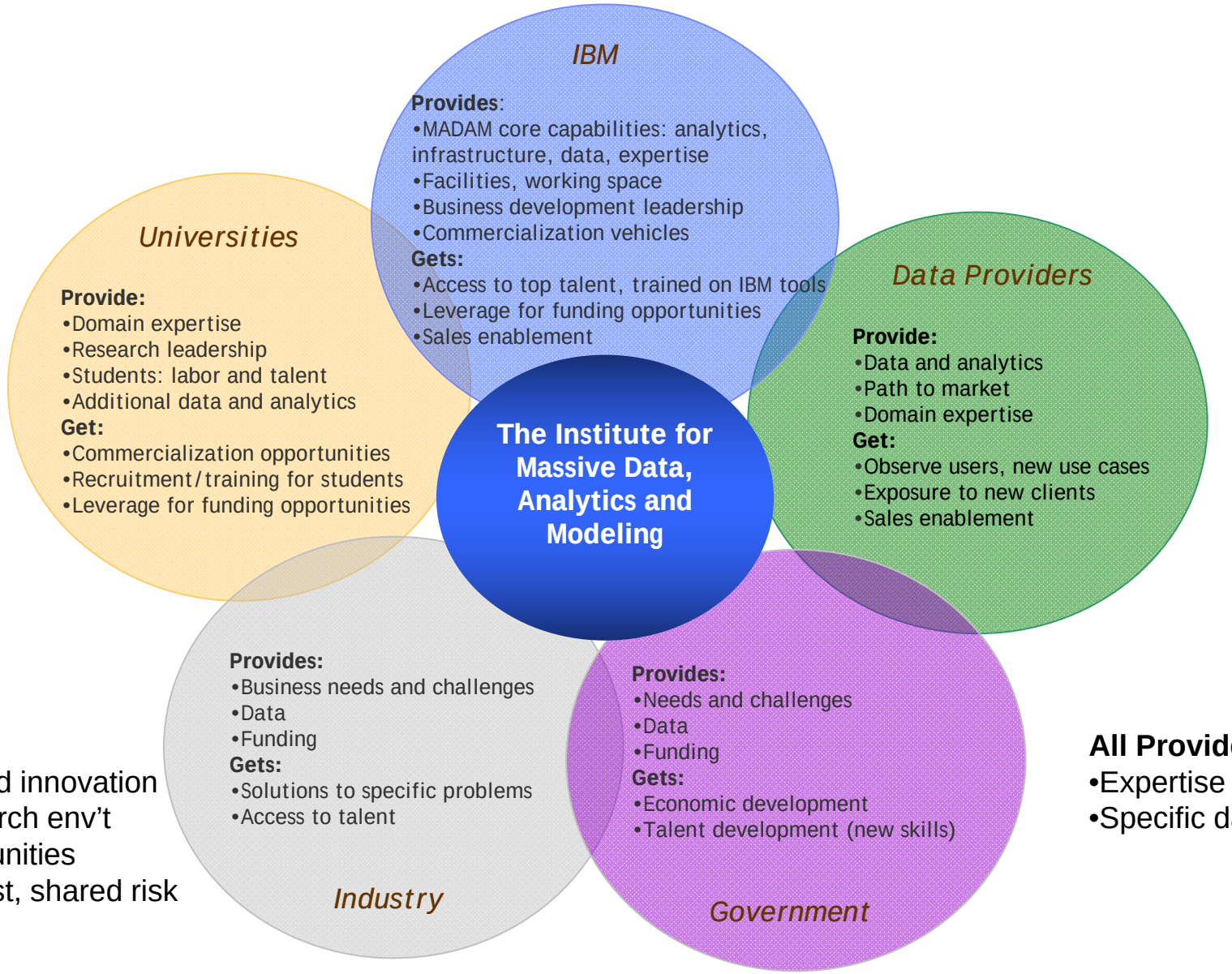
- Must acquire, integrate, enhance and align
- Must deal with missing and incomplete data
- Must store, protect, and manage
- Must create models and other analytics and test them
- Must run these analyses efficiently over large data volumes
- Must understand and share results
- **Requires significant EXPERTISE in data management, systems, analytics, and the domain**
- **Takes TIME and MONEY**

A Plug-and-Play environment could reduce cost and risk

- The Institute for Massive Data, Analytics and Modeling will unlock the value of data by providing a plug-and-play environment for exploring massive data
  - Pre-integrated **data** sets to provide **context**
  - Powerful **infrastructure** for data management and analytics
  - Rich collection of **analytics** and tools for analysis
  - **Expertise** in all aspects of the process
- Lets the domain expert focus on *their* strengths; we handle the data challenges
- Leverage these capabilities across multiple domains, and multiple investigations, to solve **important problems** for people, industry and the world at large
- Reduce costs, risk, and time to value!



The Institute as an Ecosystem: Vision



All Get:
- Accelerated innovation
- Rich research env't
- PR opportunities
- Shared cost, shared risk

All Provide:
- Expertise
- Specific data and IP

"Enabling the Benefits of Big Data"

Conclusion

- Big Data doesn’t operate in a silo.
- Most Client Use Cases Combine Multiple Technologies
- Big Data Platform and Open Collaboration could reduce cost and risk

Thank you!