InSciTe™ system based on Bigdata Analysis

November 2013
Dr. Sa-kwang Song
and Dr. Jangwon Gim

Dept. of Computer Intelligence Research
Korea Institute of Science and Technology Information

Introduction – KISTI

Youngseo Park
President

- National Super Computing Institute
- Advanced Information Institute
- Information Analysis Institute

Sunhwa Han

- Information Service Center
- NTIS Center
- SW Research Center

Wonkyung Sung

- Dept. of Scientific Bigdata Research

SangHwan Lee

- Dept. of Computer Intelligence Research

Hanmin Jung

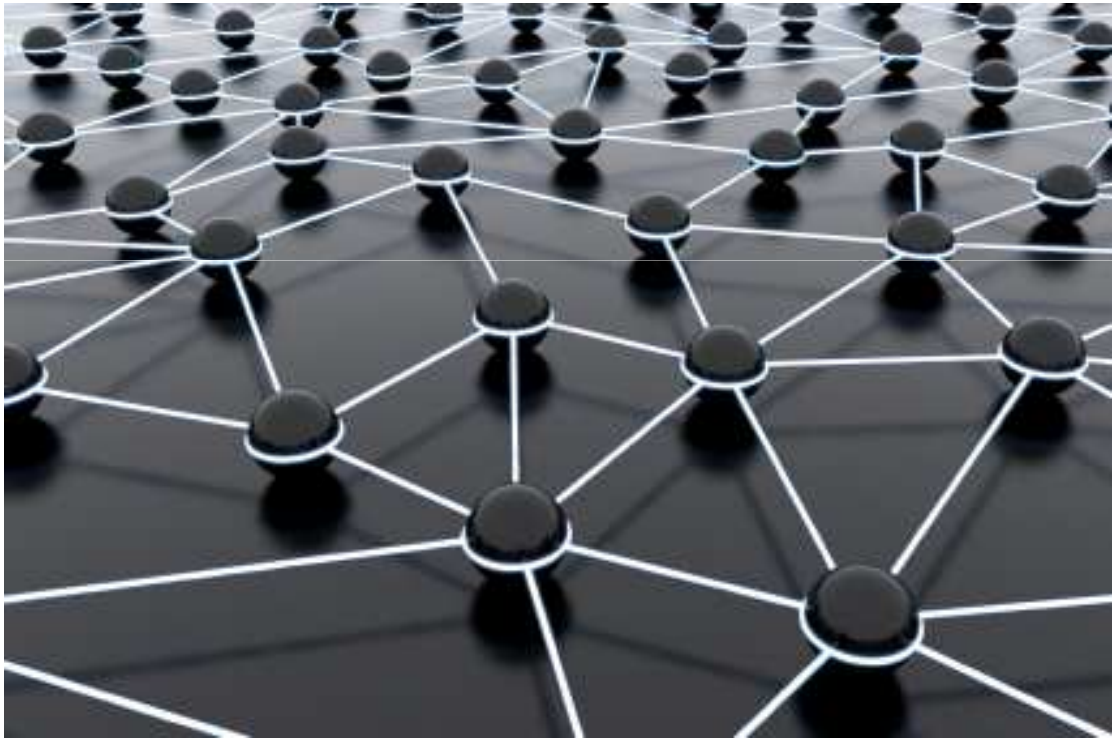
Introduction – Dept. of Computer Intelligence Research

Dept. of Computer Intelligence Research

Outline

- Introduction
- InSciTe Advanced System (2012)
- InSciTe Advisory System (2013)
- Demonstration

Introduction



Introduction – Dept. of Computer Intelligence Research

Paper/Patents

Web news/SNS

TEXT MINING

Semantic Web

Big Data Analytics

Computer Intelligence

Technology Intelligence

Introduction - Technology Intelligence(TI)

- Activity that enables companies to identify technological opportunities and developments that could affect the future growth and survival of their businesses.
- Aims to capture and disseminate the technological information needed for strategic planning and decision making.
- Effective TI capabilities are becoming increasingly important, as technology life cycles are shorten and business becomes more globalized.
- So, we have been developing a TI system, InSciTe, since 2010.
- InSciTe: Intelligence in Science and Technology

TI System - InSciTe™

- We have been developing InSciTe System since 2010.
  - InSciTe (2010)
    - The First TI system using Semantic Text Mining Techniques
  - InSciTe Advanced (2011)
    - Focused on predictive analysis based on TLCD(Technology Life-Cycle Decision) Model
  - InSciTe Adaptive (2012)
    - Focused on adaptivity as well as insight on each service.
    - Mobile application (Android platform)
  - **InSciTe Advisory (2013)**
    - **Focused on prescriptive analysis.**
    - **Mobile and Web based application.**

http://inscite.kisti.re.kr

Similar Projects

- CUBIST(Combining and Uniting Business Intelligence with Semantic Technologies) – EU
  - Led by Sheffield Hallam Univ., Ontotext, and SAP
    - 1st CUBIST workshop in July, 2011
  - For better Semantic Web search
  - Aims to develop new ways to interrogate not only the massive volume data on the Internet, but also analyze the different formats it exist in – such as blogs, wikis, and videos.

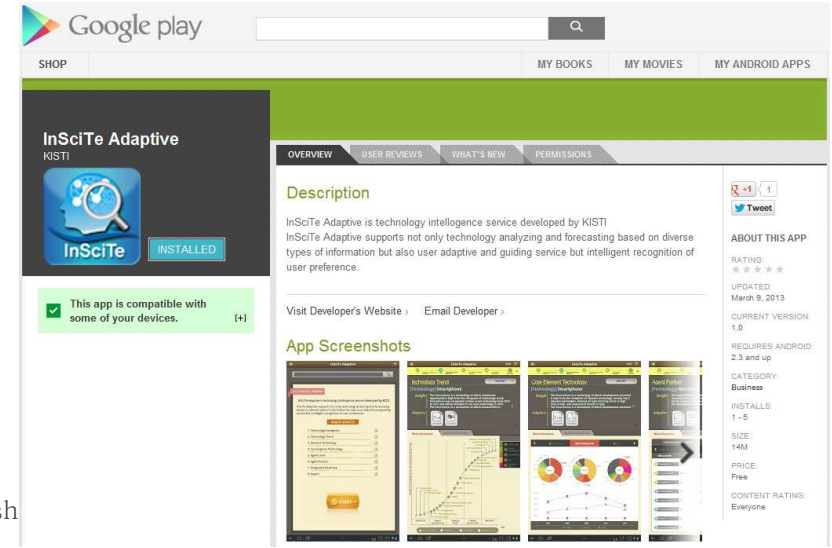
Similar Projects

- FUSE (Foresight and Understanding from Scientific Exposition) - USA
  - Broad Agency Announcement (BAA) Released by IARPA at Sep. 14, 2010
    - Kick off meeting in summer, 2011
  - Seeks to develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technical emergence using information found in the published scientific, technical, and patent literature

InSciTe Adaptive (2012)

InSciTe Adaptive

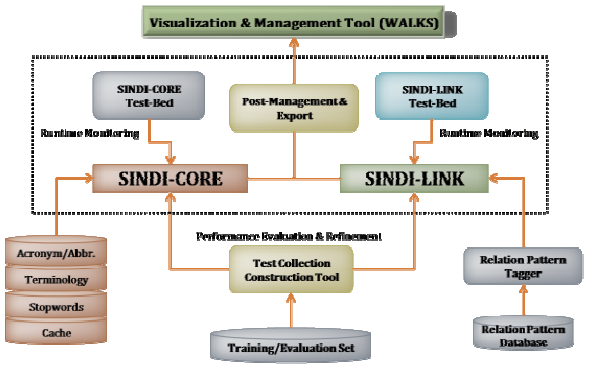
- InSciTe Adaptive
  - Supports R&D strategy planning based on analysis of technologies and organizations from textual documents.
  - Is *adaptive* to users and gives *insight* whenever users act.
  - Consists of 8 services including an automatic report generating service.
  - Can be downloaded at Google play.



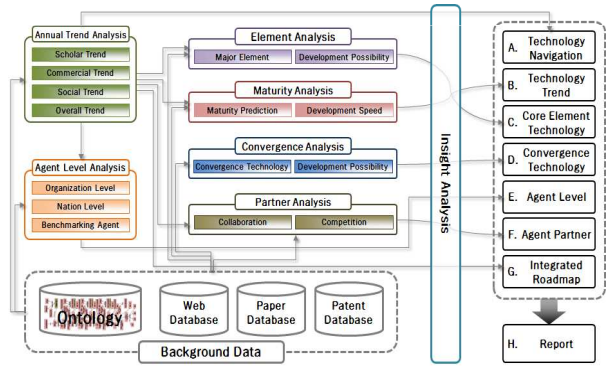
InSciTe System

Technology Intelligence Services

Text Analysis System



Semantic Analysis System

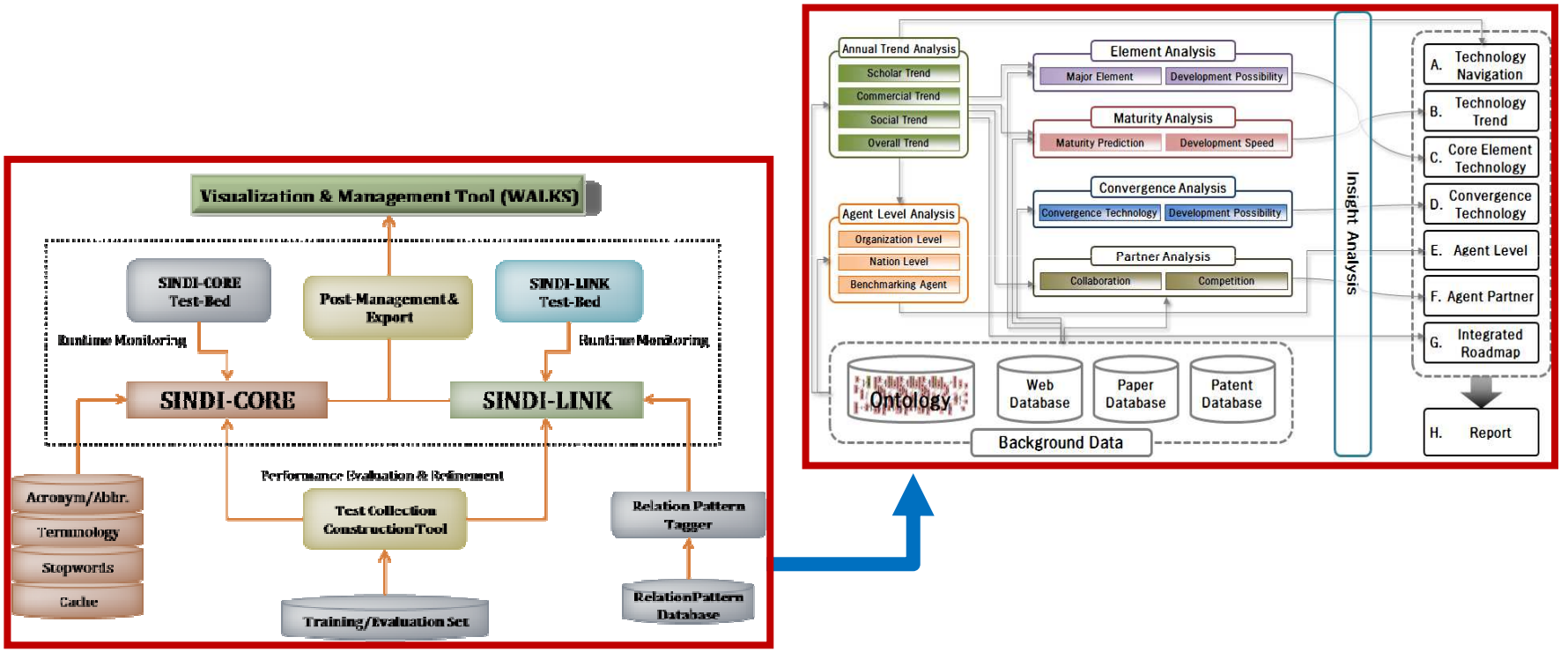


Hadoop Echo-system

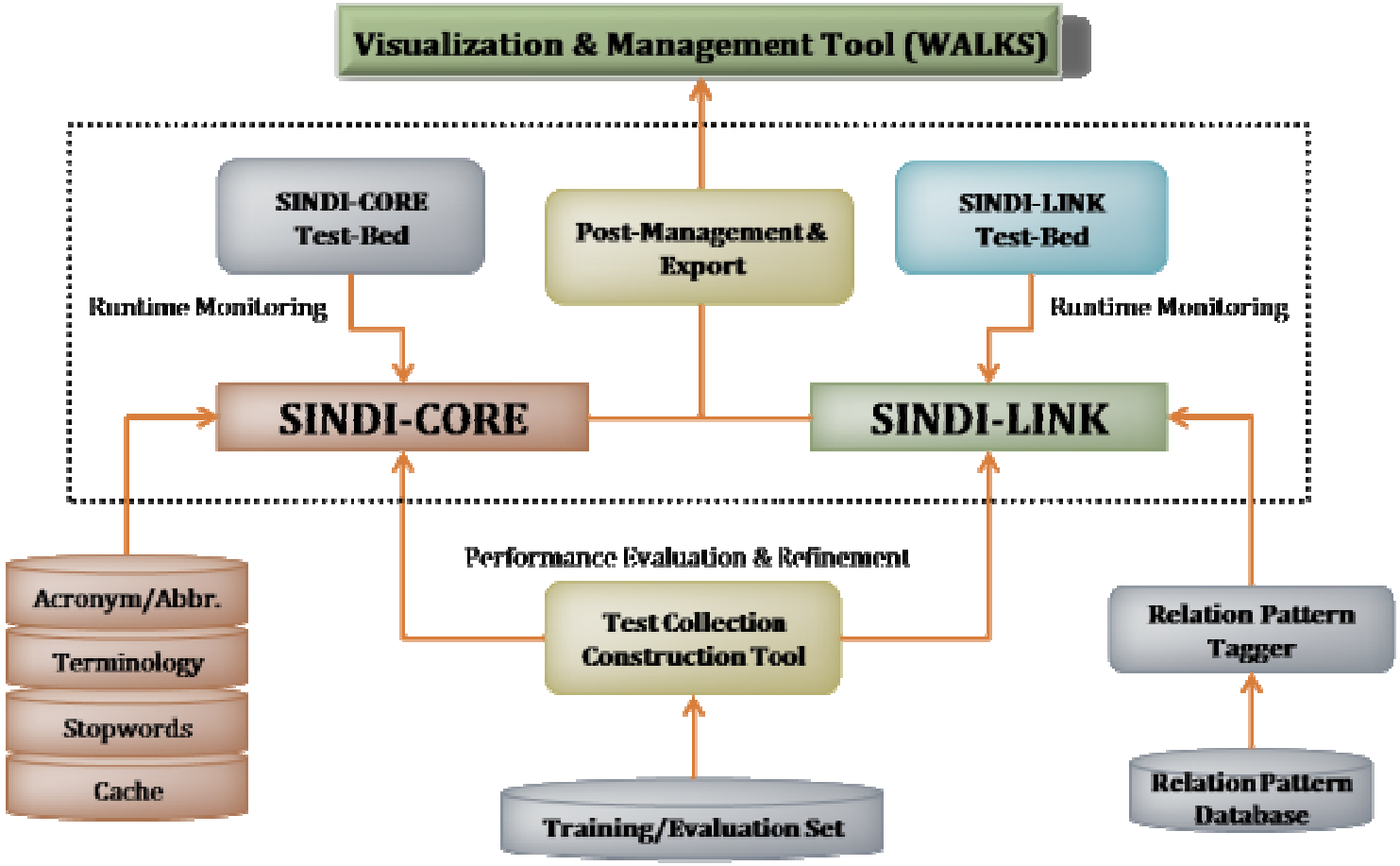
MySQL · hadoop · APACHE HBASE · OOZIE · HIVE · FLUME · R

Internal Structure

- Text Analytics + Semantic Analytics



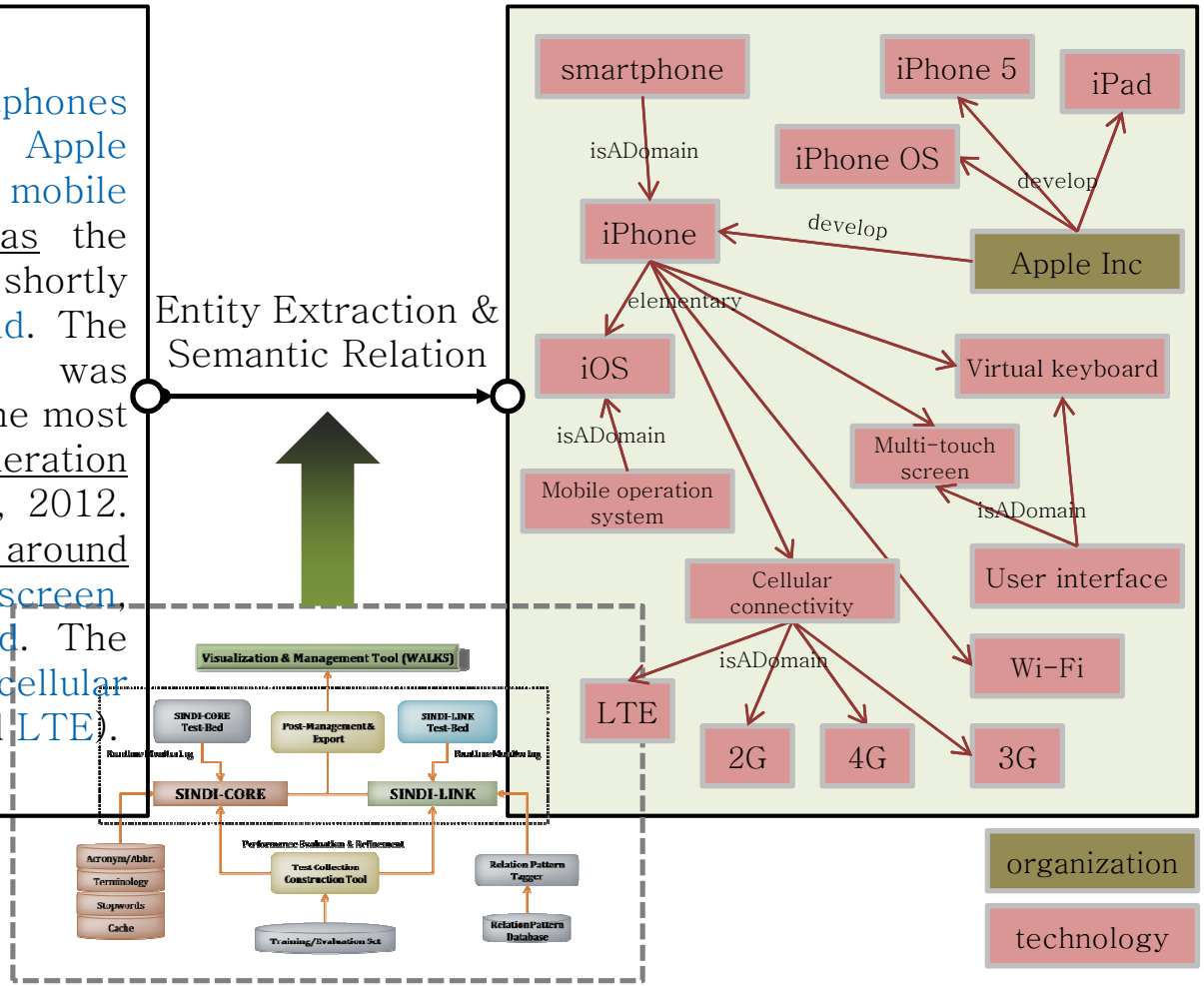
Text Analysis System



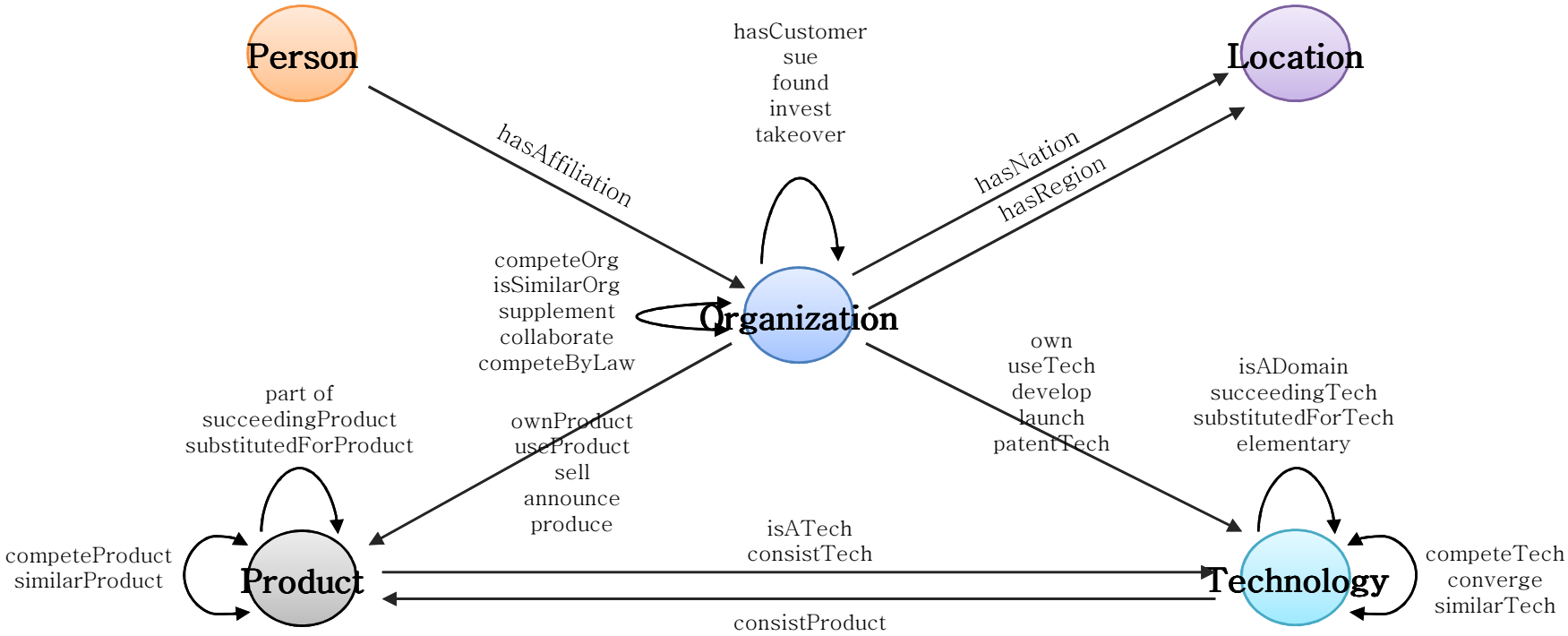
Text Analysis System

The iPhone is a line of smartphones designed and marketed by Apple Inc. It runs Apple's iOS mobile operating system, known as the "iPhone OS" until mid-2010, shortly after the release of the iPad. The first generation iPhone was released on June 29, 2007; the most recent iPhone, the sixth-generation iPhone 5, on September 21, 2012. The user interface is built around the device's multi-touch screen, including a virtual keyboard. The iPhone has Wi-Fi and cellular connectivity (2G, 3G, 4G, and LTE).

Entity Extraction & Semantic Relation



Ontology Scheme

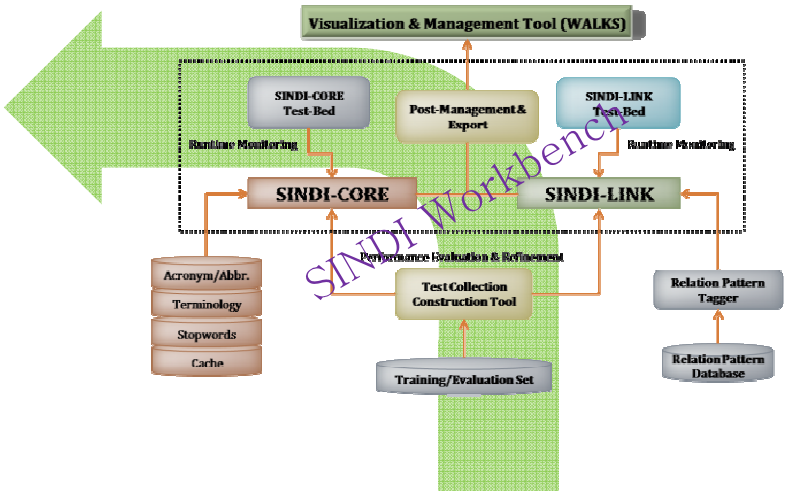


Statistics of Source Documents

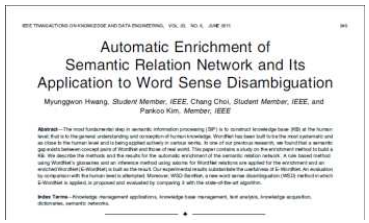
- Text Data to Ontology

498,361,449 Triples

technical terms	43,201,941
products	62,327,156
persons	25,110,360
organizations	40,993,708
locations	50,350,884



9,765,199
Scientific Papers
(2001 ~ 2011)

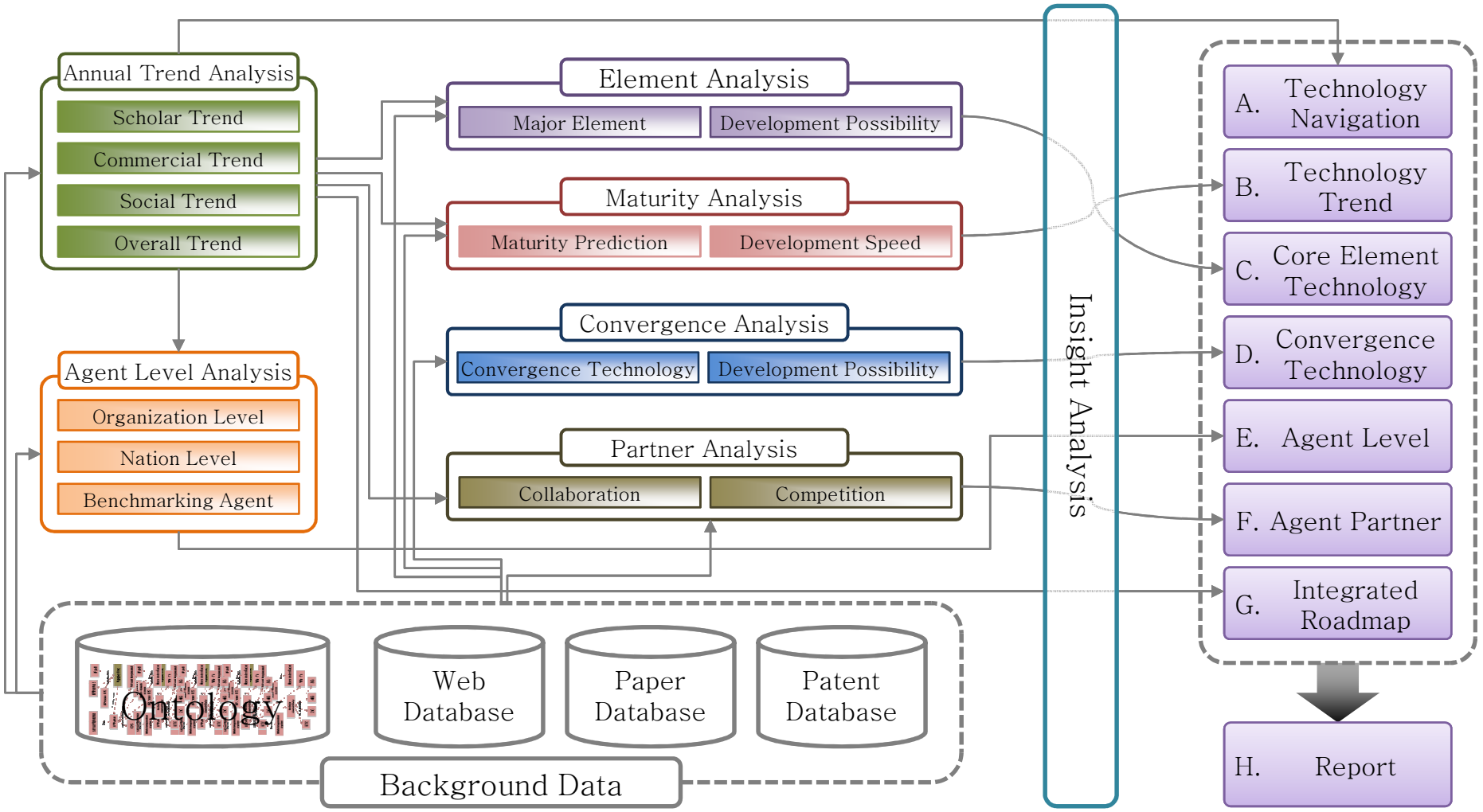

Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation

7,615,819
US/EU/PCT Patents
(2001 ~ 2011)


United States Patent

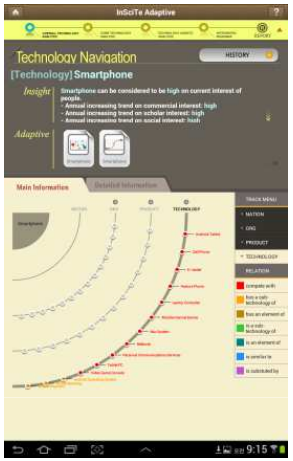
5,268,696
Web Documents
(as of Nov. 13, 2012)

Semantic Analysis System

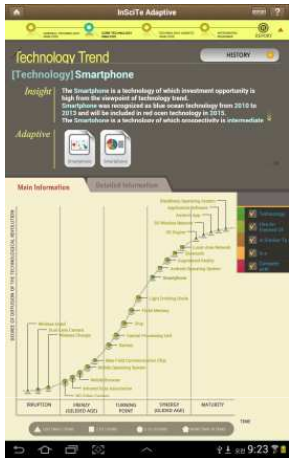


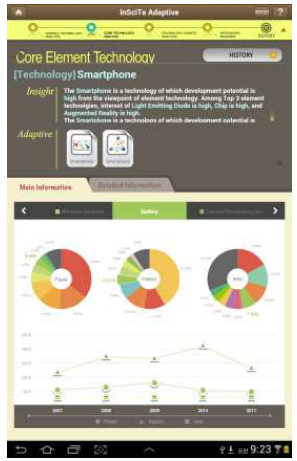
InSciTe Advanced services

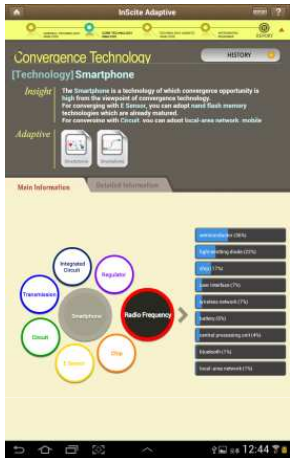
Overall Analysis



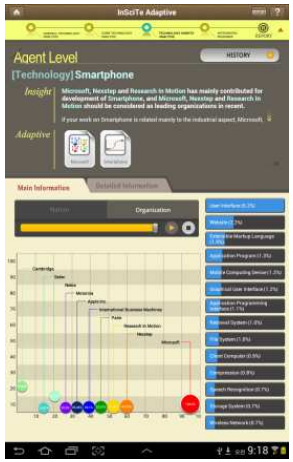
Technology Centric Analysis

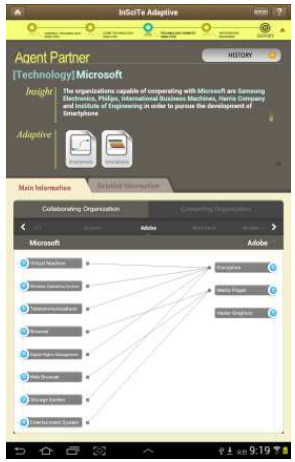




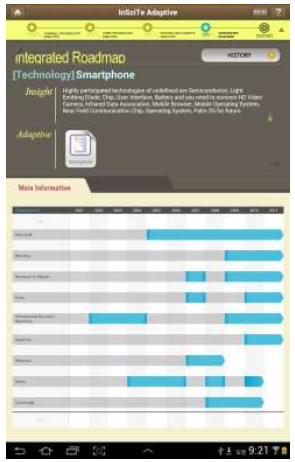


Agent Centric Analysis





Agent Trend and Report



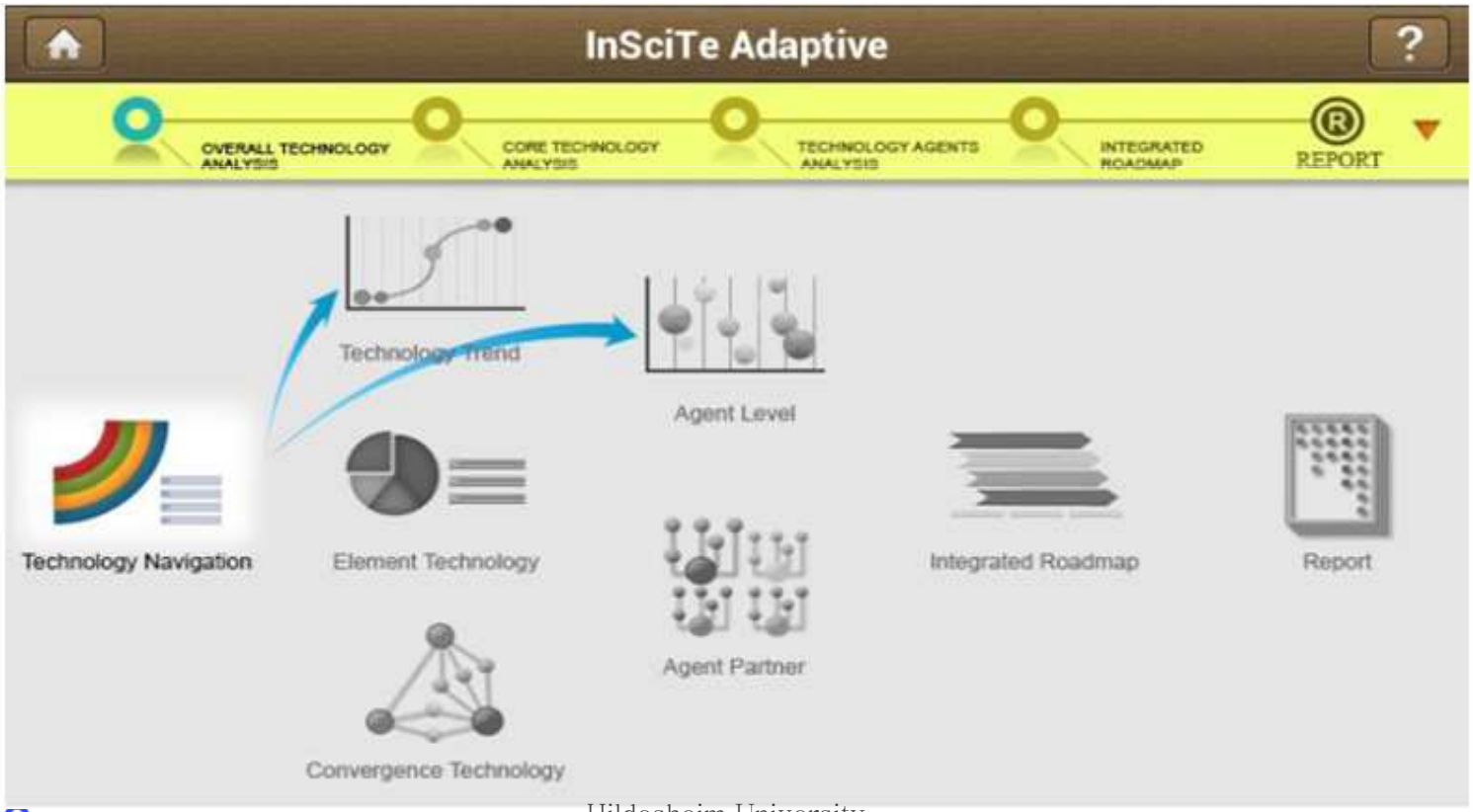
Developing a project using InSciTe

Exploring & finding interesting technologies

Deciding most interesting technology

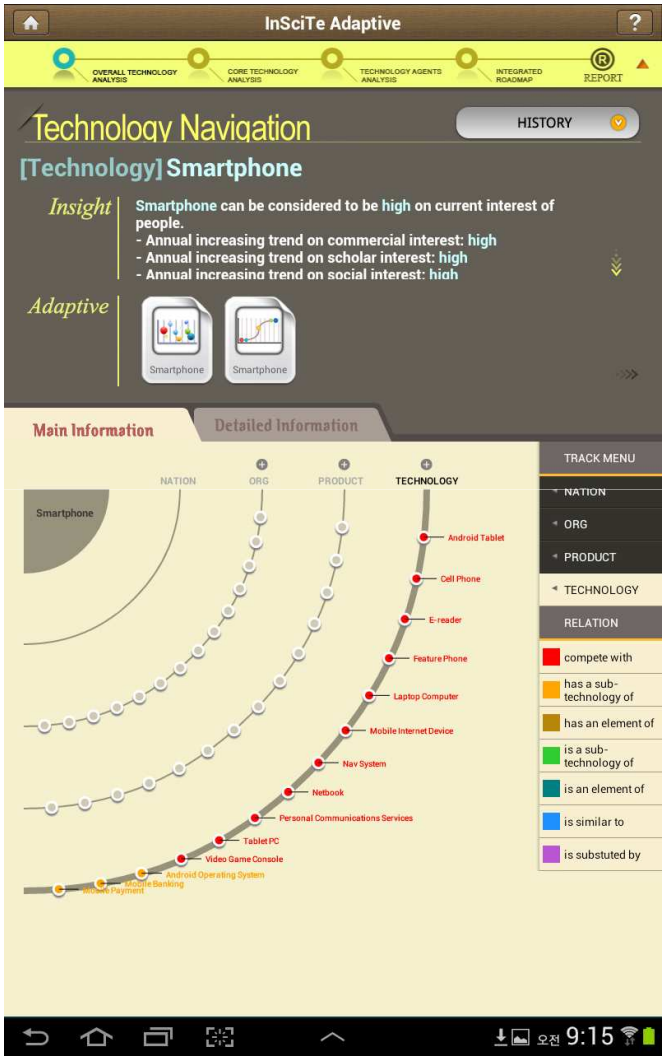
Analyzing competitors

Building Strategy and plan.



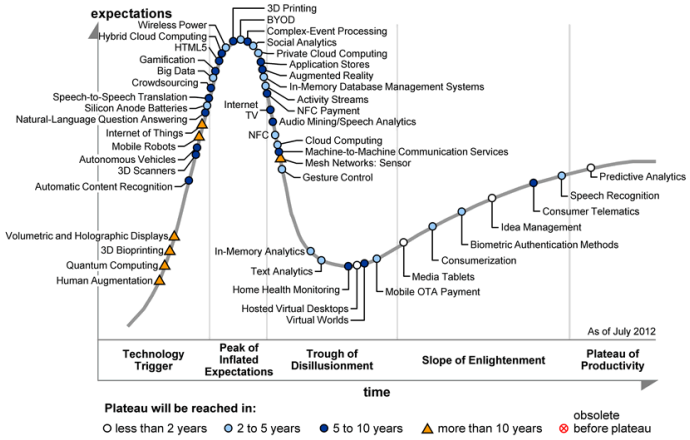
InSciTe service –Technology Navigation

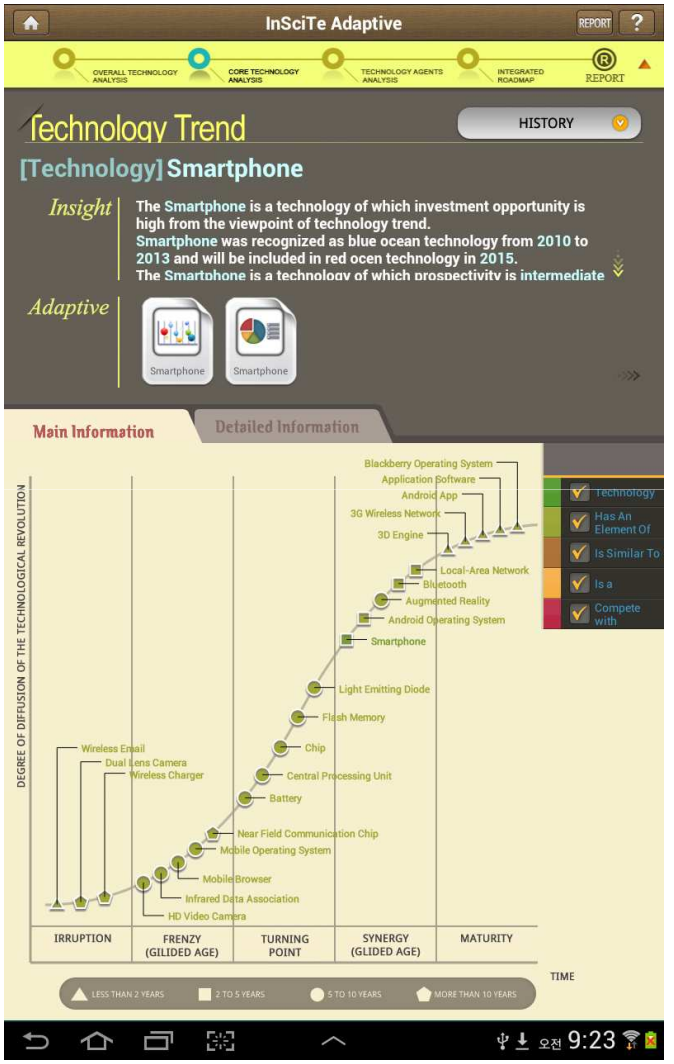
- Objective
  - General/overall analysis of user interested technology
- Insight
  - Provides users with trends on commercial, social, and scholar interest.



InSciTe service – Technology Trends

- Objective
  - analysis of current maturity level of technologies
- Insight
  - guiding user to technologies appropriate for investment and participation

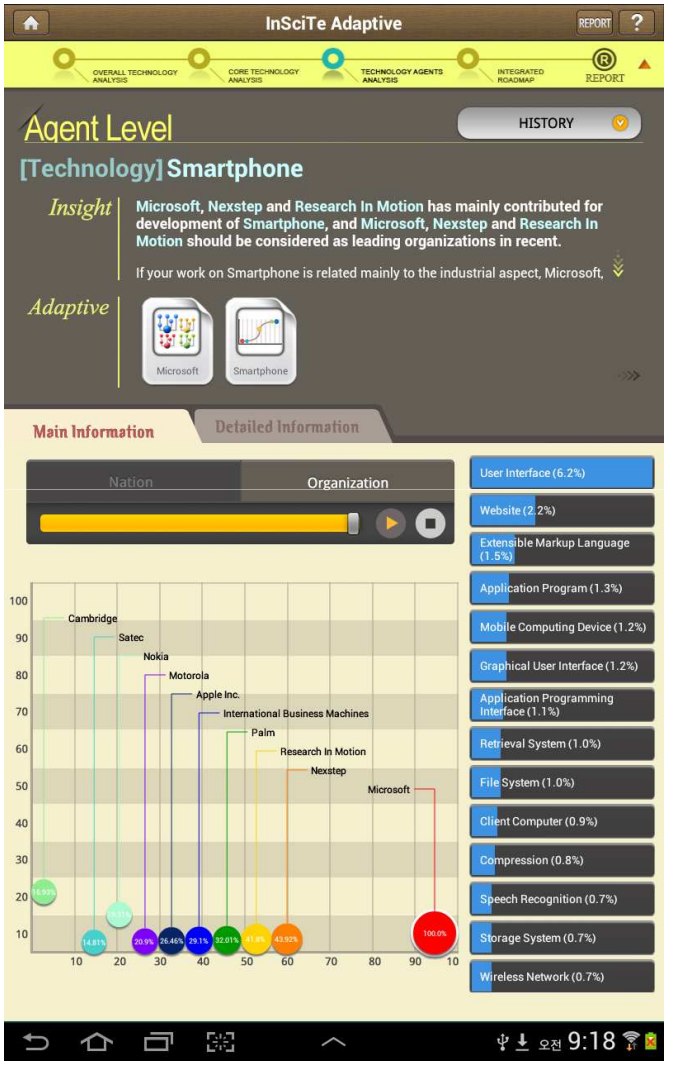




InSciTe service – Agent Level

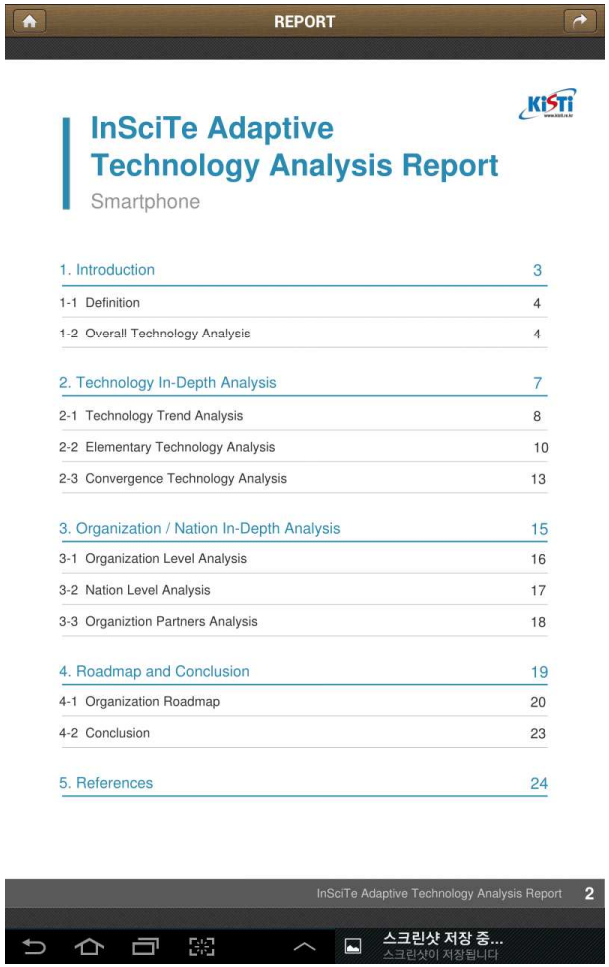
- Objective
  - analysis of major agents on academic and industrial point of view

- Insight
  - discovering benchmarking organizations for academic and industrial aspects

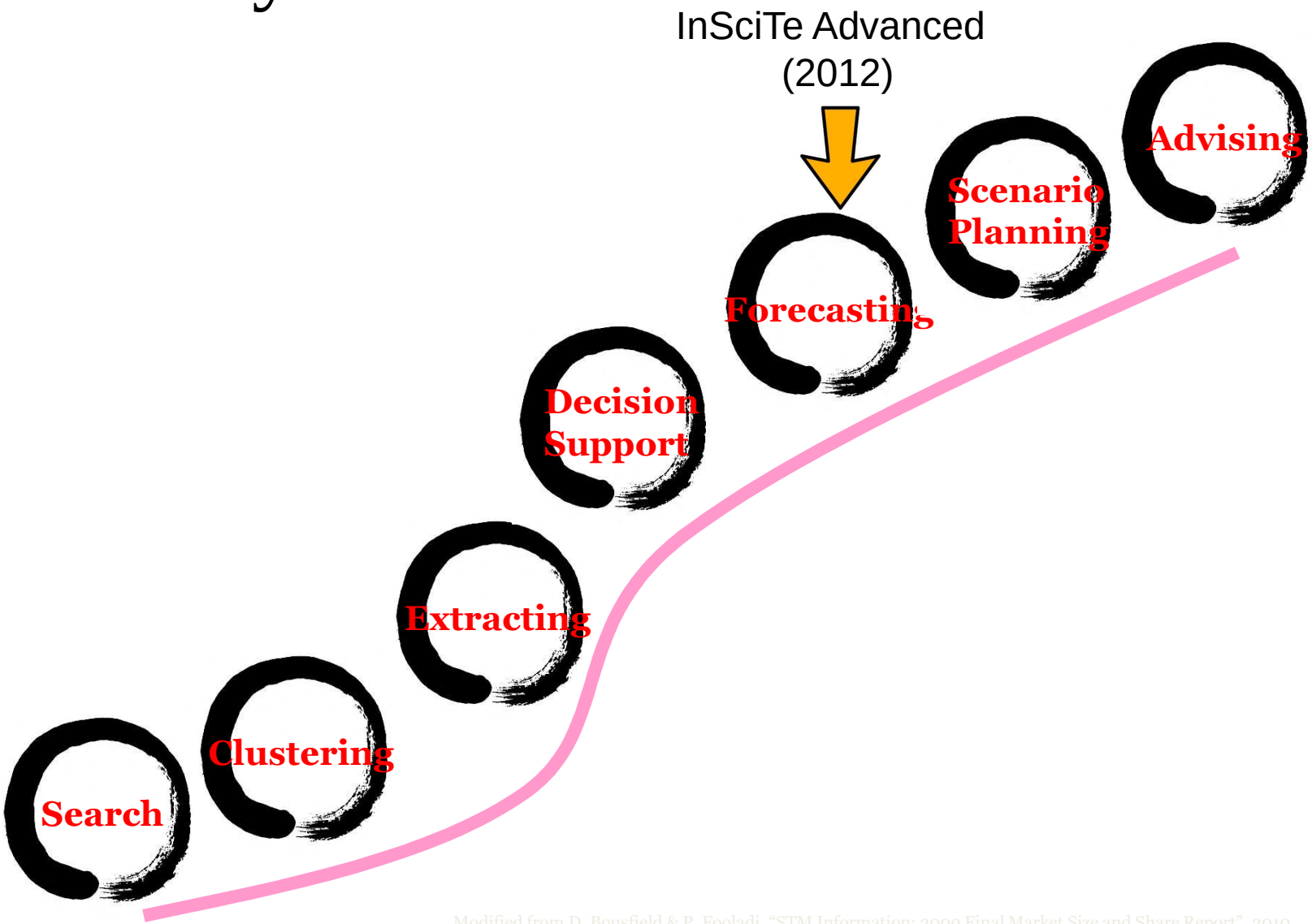


InSciTe service – Report

- Automatically generated report



Value Pyramid



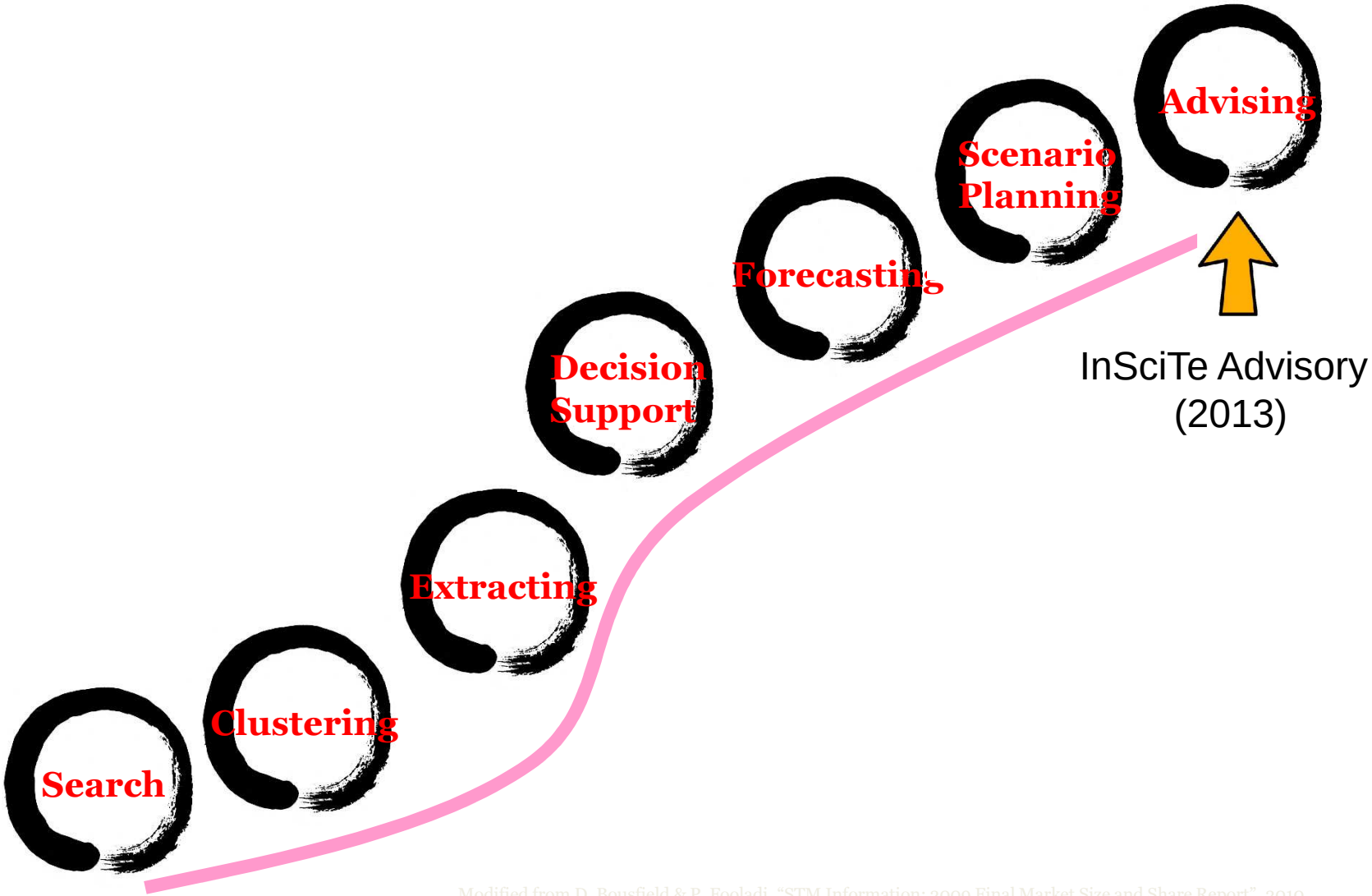
Modified from D. Bousfield & P. Fooladi, "STM Information: 2009 Final Market Size and Share Report", 2010.

InSciTe Advisory(2013)

InSciTe Advisory

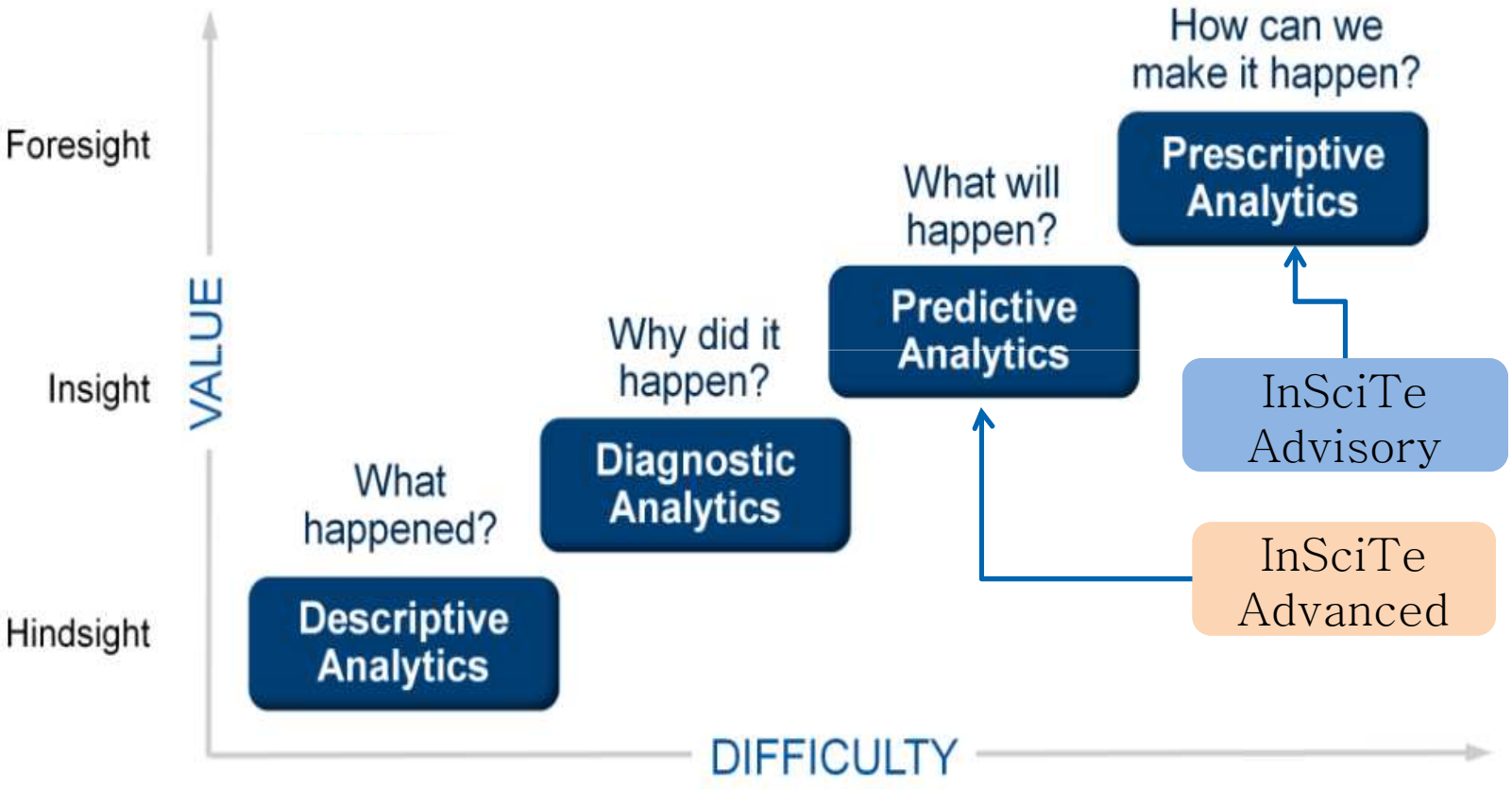
- InSciTe Advisory
  - Provides a selected researcher with how to improve his/her research competitiveness.
  - Recommends role model researchers who are advanced to and could be followed by the researcher.
  - Recommends future research plans to reach to the research level of the role model researchers, based on 5W1H questions.
  - Consists of two main services including descriptive analytics and prescriptive analytics.

Value Pyramid



Modified from D. Bousfield & P. Fooladi, "STM Information: 2009 Final Market Size and Share Report", 2010.

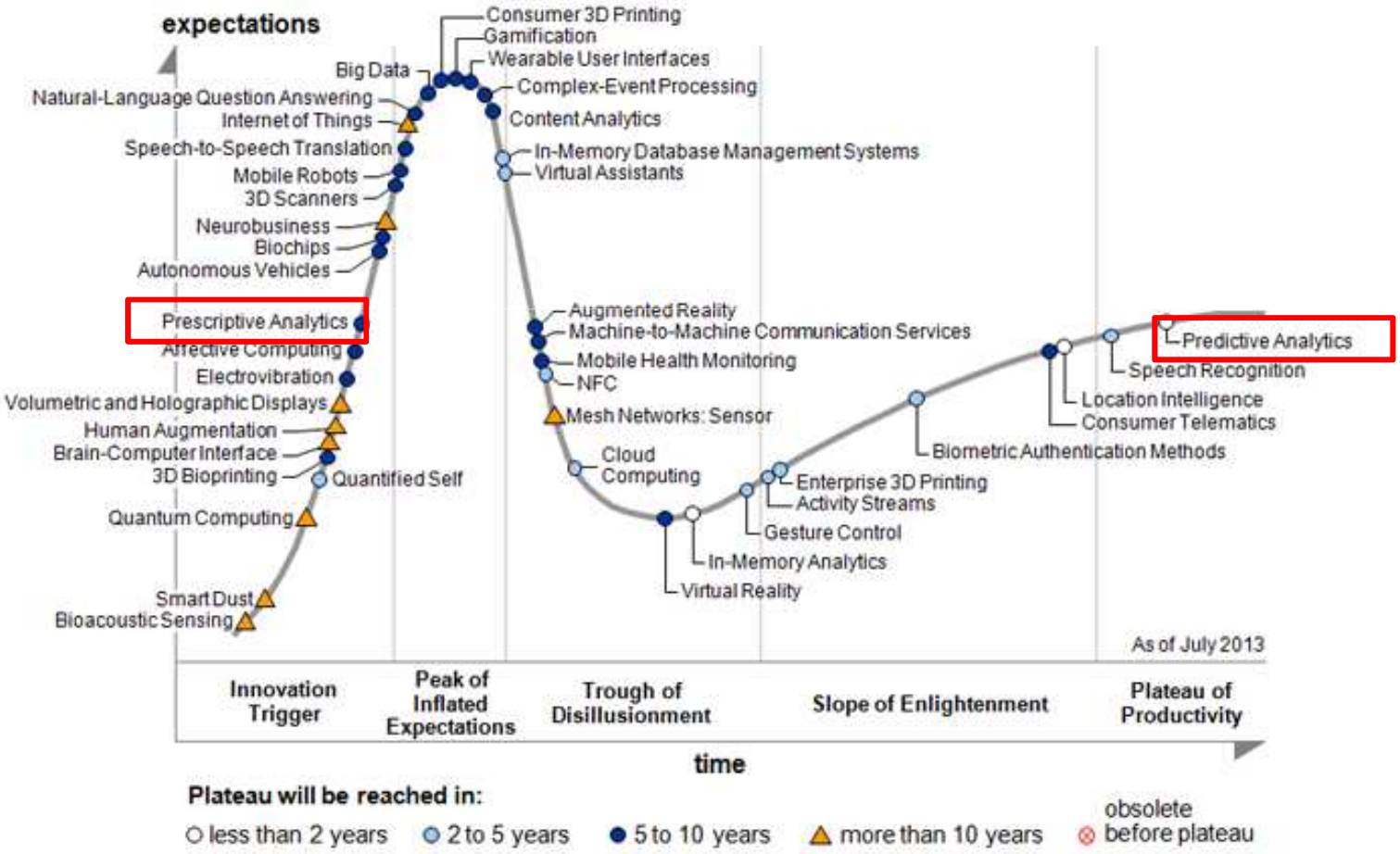
Value Pyramid



Data Analysis Steps for Business Intelligence

Technical Trends in 2013

- 2013 Hype Cycle for Emerging Technologies



3 Phases of Business Analytics

- **Descriptive Analytics**: A set of technologies and processes that use data to understand and analyze business performance

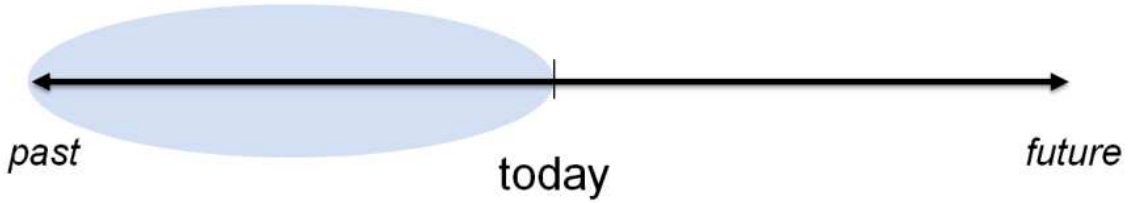
- **Predictive Analytics**: The extensive use of data and mathematical techniques to uncover explanatory and predictive models of business performance representing the inherit relationship between data inputs and outputs/outcomes.

- **Prescriptive Analytics**: A set of mathematical techniques that computationally determine a set of high-value alternative actions or decisions given a complex set of objectives, requirements, and constraints, with the goal of improving business performance.

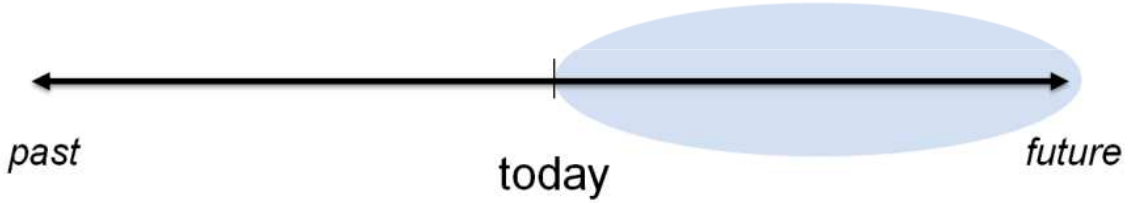
A cross-brand team from IBM (Irv Lustig, Brenda Dietrich, Christer Johnson and Christopher Dziekan)

Temporal Perspective

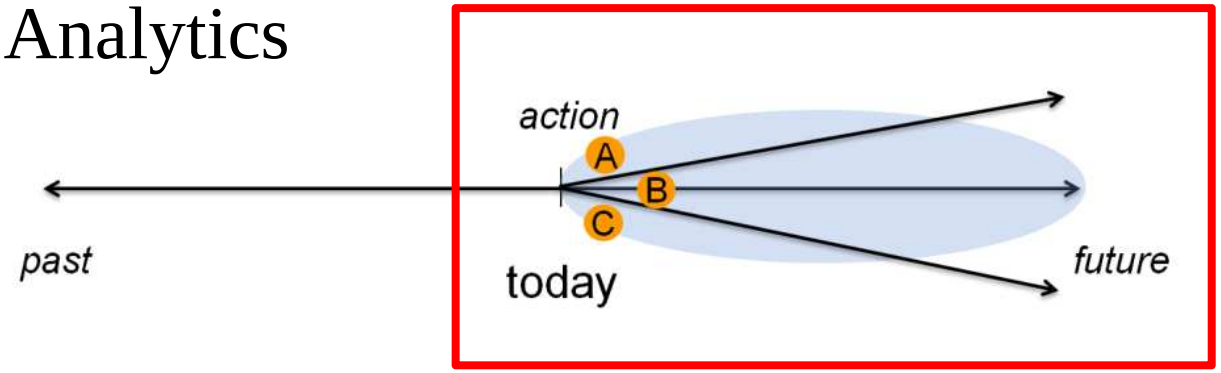
- Descriptive Analytics



- Predictive Analytics



- Prescriptive Analytics



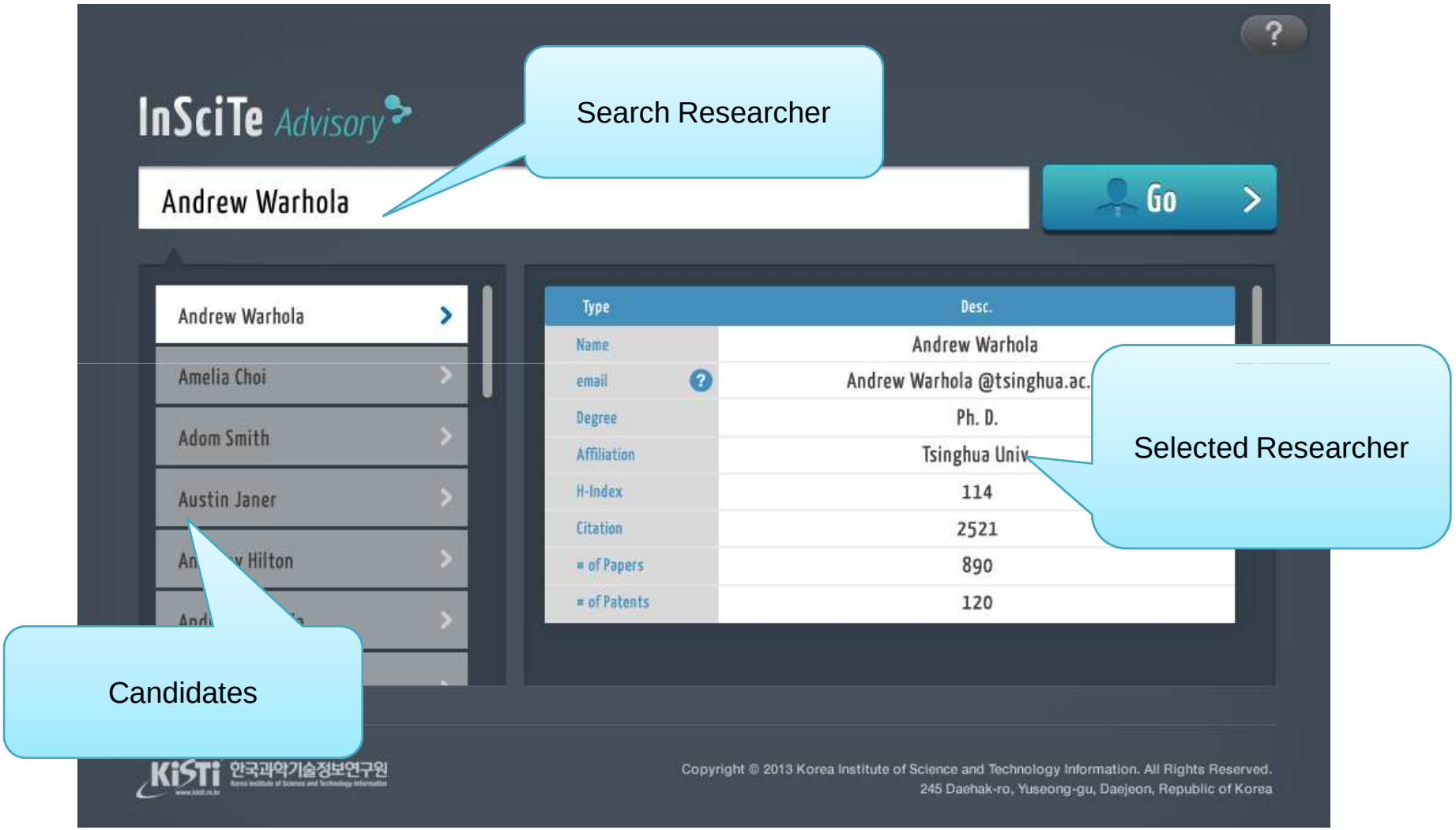
Gartner

Related Work

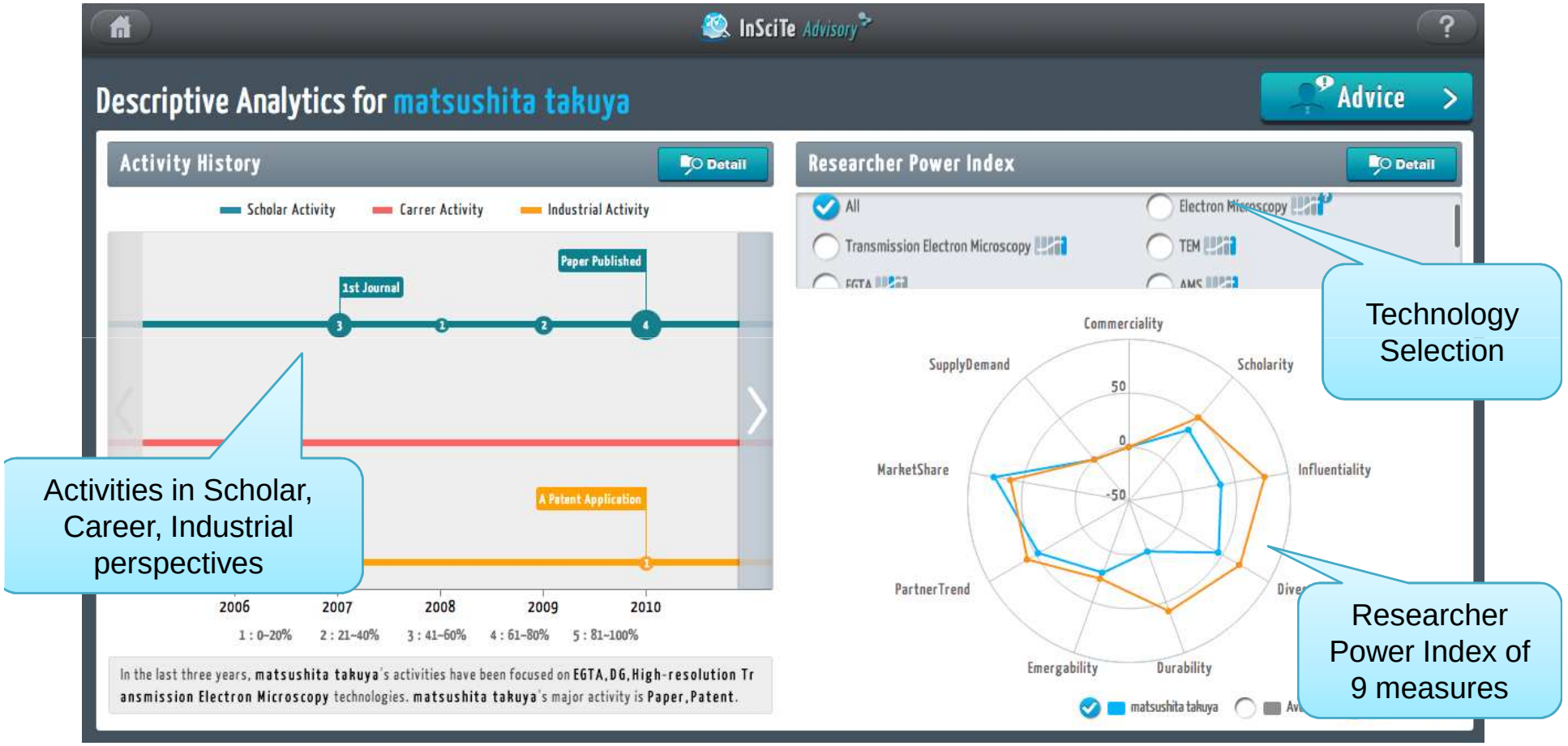
- IBM: Tax Revenue Collections in NY State (2010)
  - Problems: Dynamically changing situation for taxpayers of interest (Move or disappear, File bankruptcy)
  - Prescriptive analytics: creation of a set of optimal strategies that respond to the constantly changing information and uncertainty about delinquent taxpayers

- AYATA: Oil & Gas Companies (2013)
  - Problems: Billion-dollar decisions about when and where to drill, how many sites to develop, and how to produce from multiple wells.
    - Big impact from every decision, with the high cost of drilling and completing wells
  - Prescriptive analytics: showing the impact of each decision so operations managers can ensure future production output.

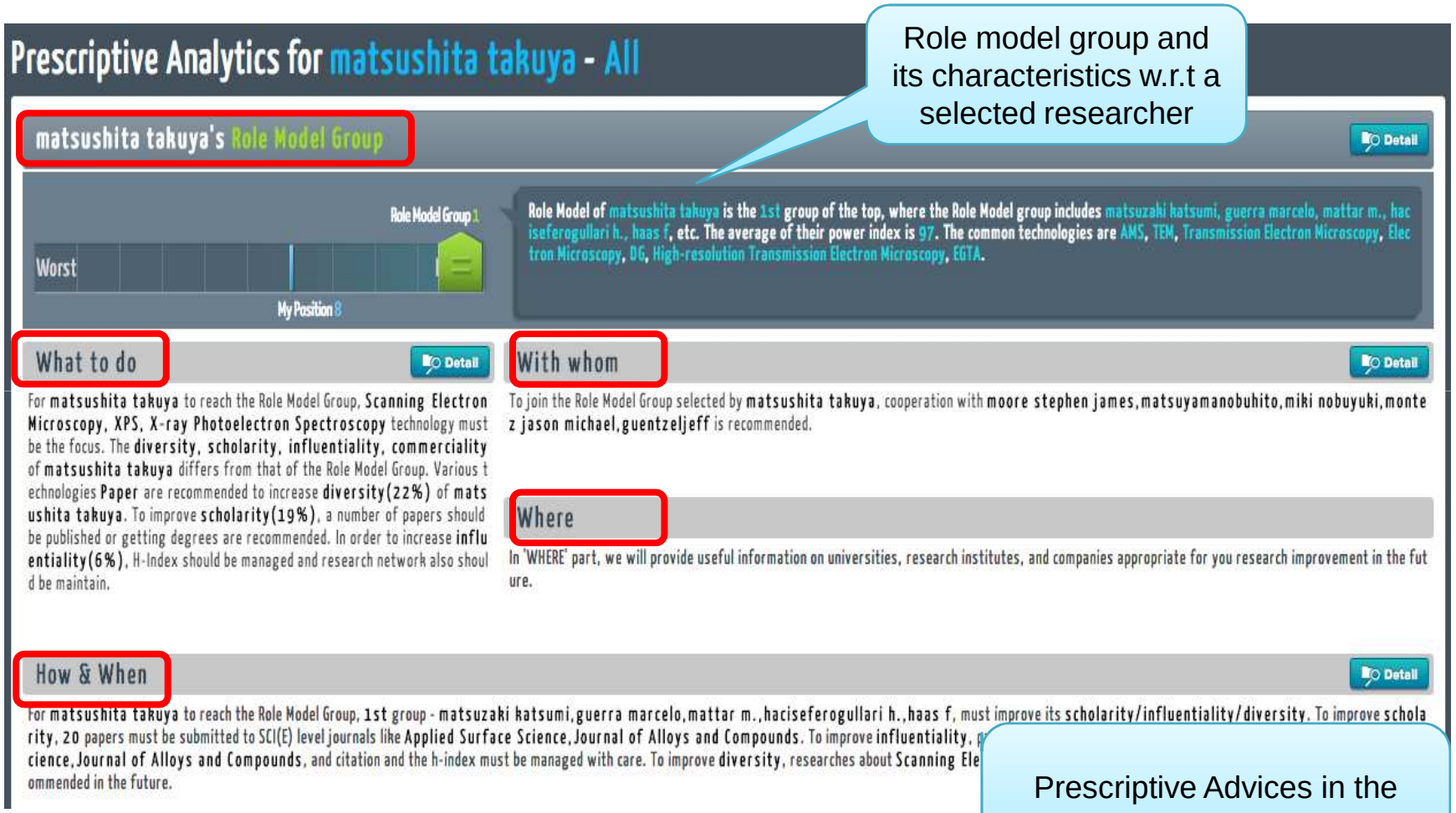
Screenshots: Researcher-centric Analysis



Screenshots: Descriptive Analytics



Screenshots: Prescriptive Analytics



Role model group and its characteristics w.r.t a selected researcher

Prescriptive Advices in the point of 5W1H.

Screenshots: Role Model Group

- a group of researchers whom the selected researcher should follow.
- Technological characteristics of the Role Model Group, Researcher Power index, and representative researchers of the group are described.
- Role Model group selection bar can be moved so that users can change their Role Model group.

Role Model Group selection bar

Role Model Group Summary

Prescription (5W1H)

- What to do
  - Describes the goal the user has to do to achieve the level of the Role Model Group researchers.
- With whom
  - Recommends researchers or organizations to work together.
- How & When
  - Explains how to do something and when to do it in detail, in order to achieve the goal.
- Where
  - Describes which organization is appropriate for the user in order to achieve the goal.
- Why
  - Included in other sections.

Automatically Generated Report

InSciTe Advisory REPORT

Hanmin Jung Aug. 2013.

Copyright (c) Korea Institute of Science and Technology Information. All rights reserved.

O-kwang Sung

SECTION 1:
Introduction

1-1 Biography

O-kwang Sung works as senior researcher of Computer Intelligence Research Laboratory at Korea Institute of Science and Technology Information (KISTI), Korea.

He received B.S. degree in Statistics in 1987 and M.S. degree in Computer Science in 1989 at Chungnam National University[1], Korea. He received Ph.D. degree in Computer Science at Korea Advanced Institute of Science and Technology (KAIST)[2], Korea. He had worked for Bioinformatics team at Electronics and Telecommunications Research Institute, Korea from 2006 to 2010. He joined Korea Institute of Science and Technology Information[3], Daejeon, Korea in 2010. His current research interest includes Business Intelligence[4], Text Mining[5], Semantic Web, Natural Language Processing, Information Retrieval, and Big Data.

His detailed information on the Internet can be found in the following links.

Personal Links
Google Scholar: http://scholar.google.com/citations?user=Tv0DDvYAAAAJ&hl=en
Linked In: http://www.linkedin.com/pub/sa-kwang-song/35/8a0/b00
DBLP: http://www.dblp.org/search/index.php#query=author:sa_kwang_song

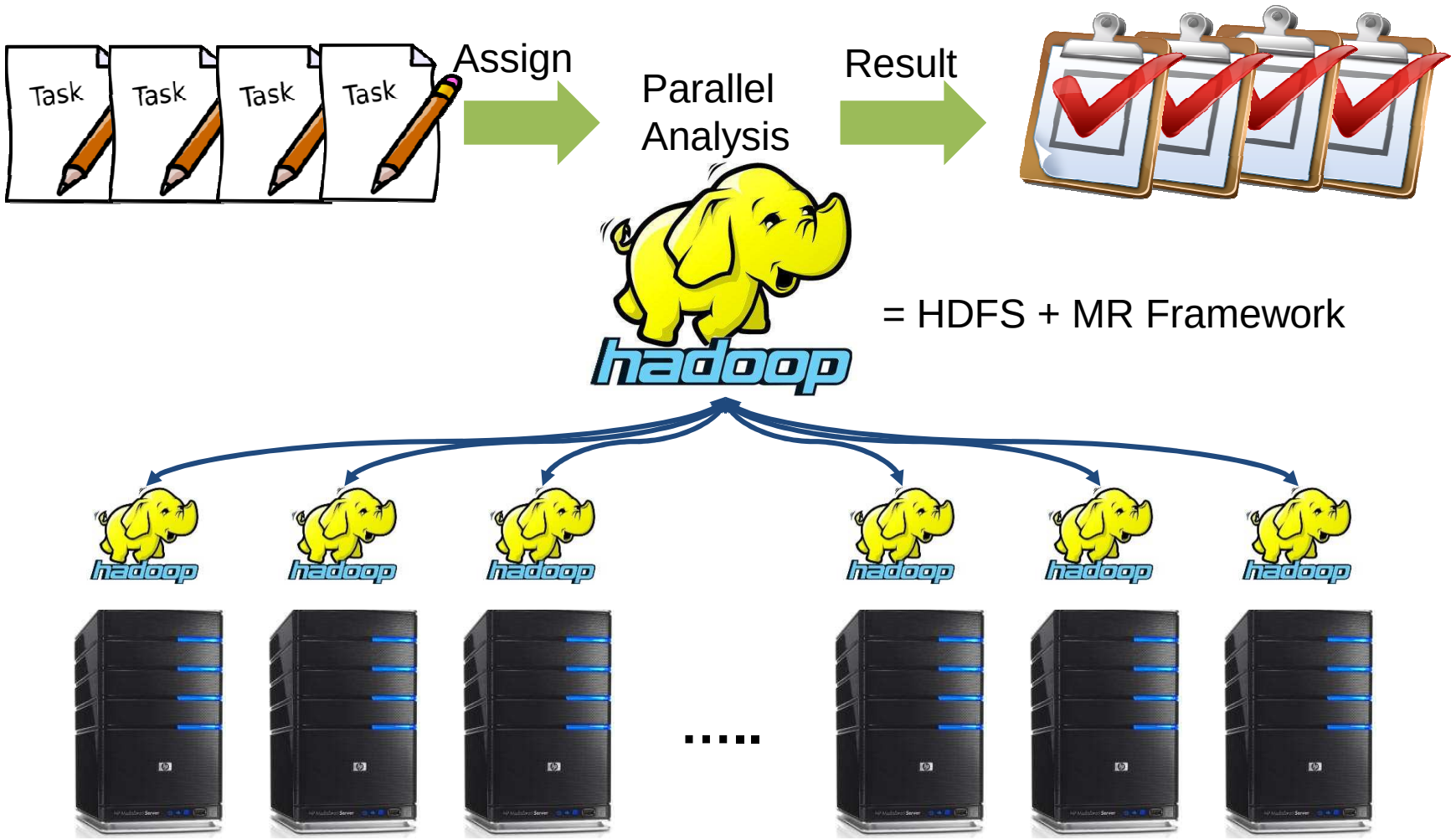
1-2 Research Levels

O-kwang Sung has been interested in the following major technology topics since 2001:

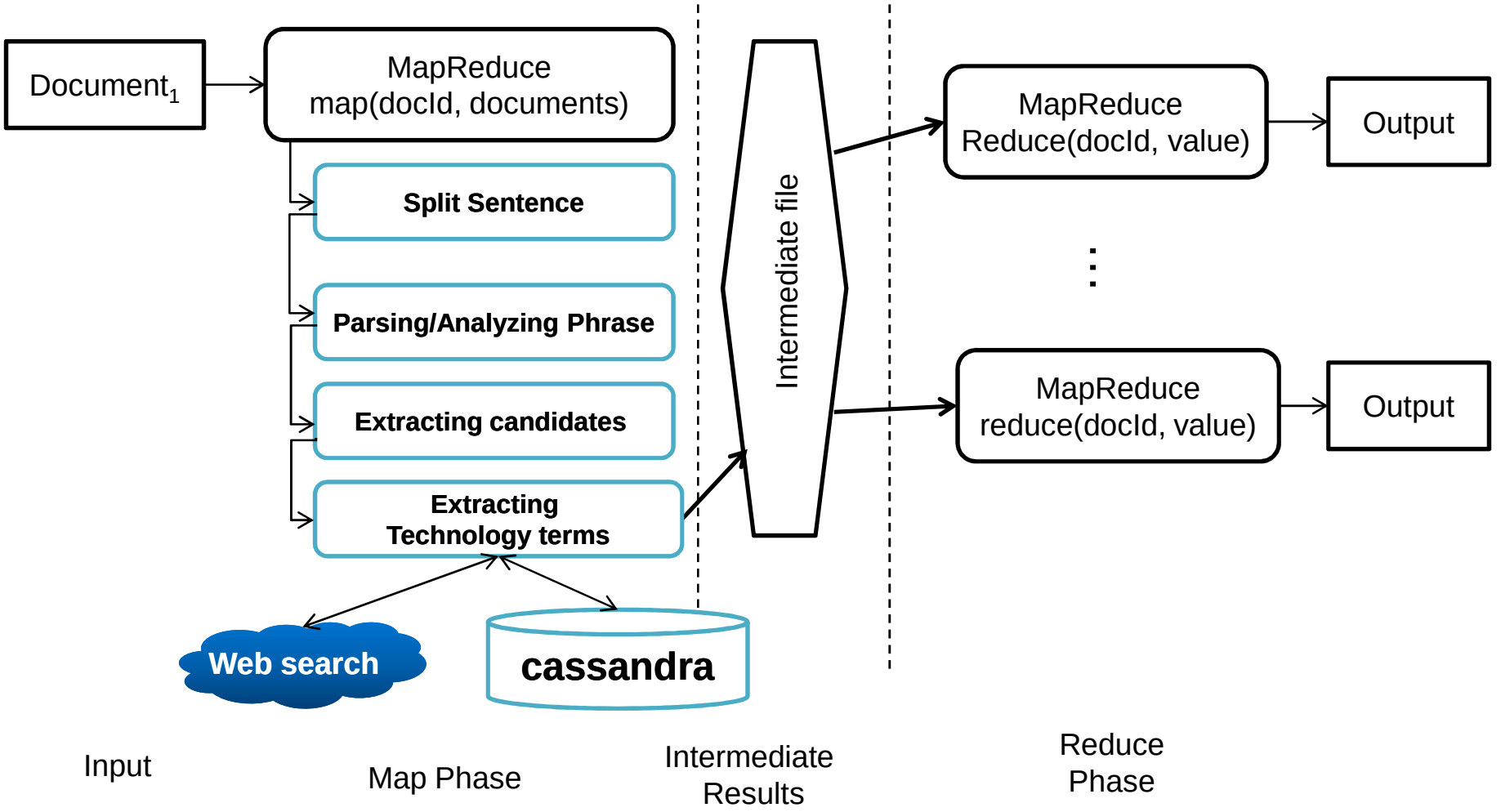
Technology Intelligence, Semantic Web, Text Mining, Business Intelligence, Big Data Analytics, Natural Language Processing, Information Retrieval.

Big Data Platform

Hadoop based Big Data Platform

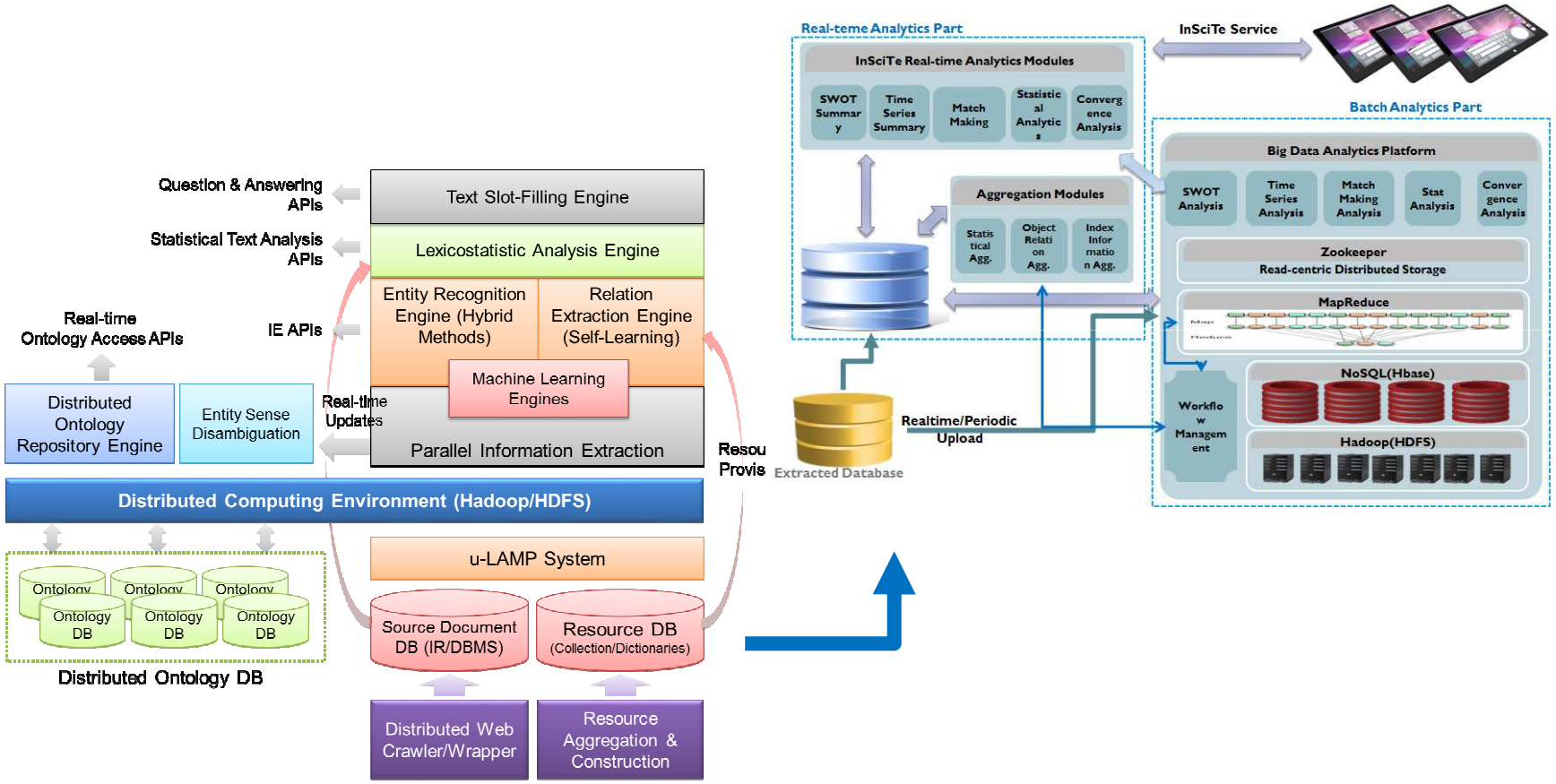


Information Extraction

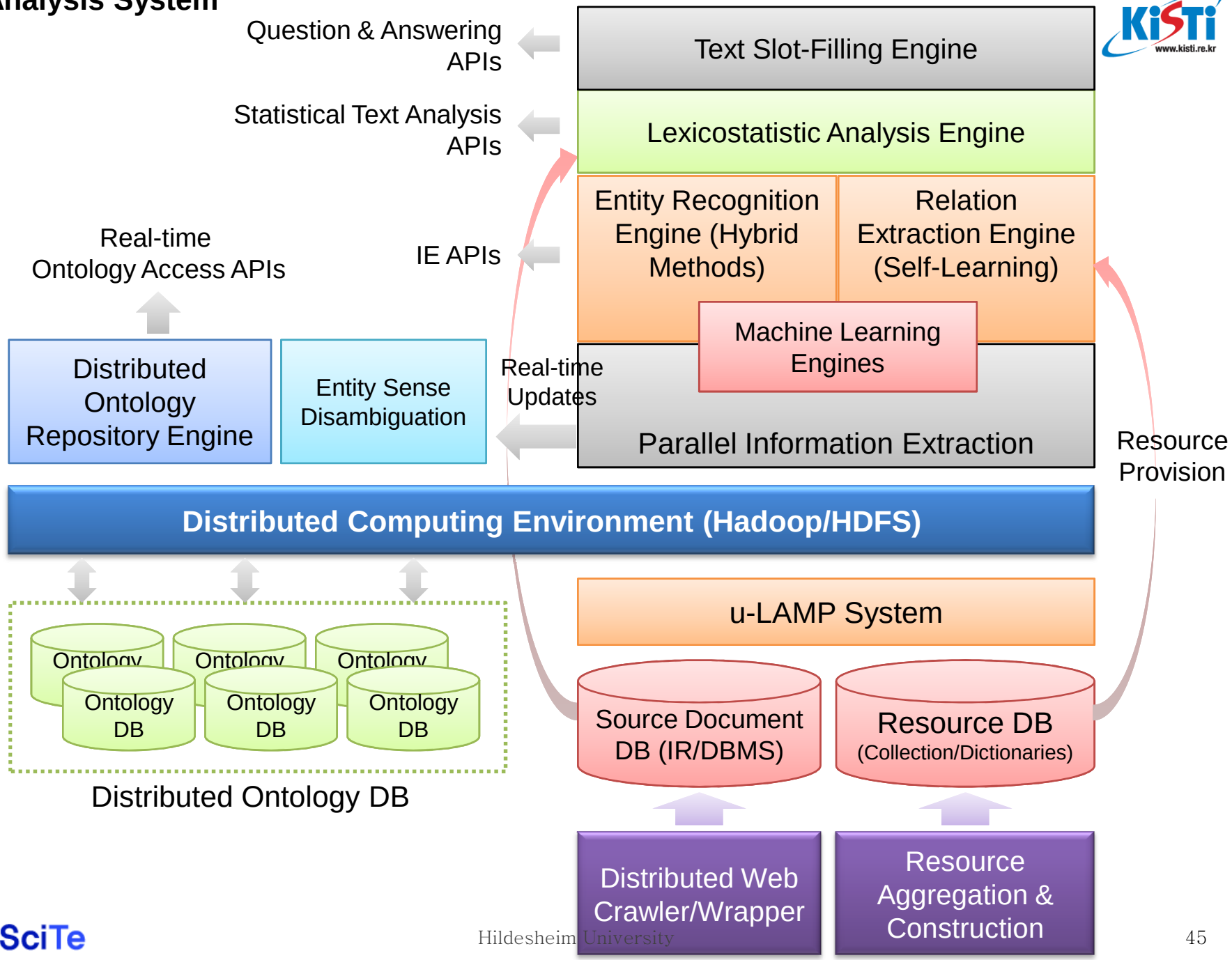


Internal Structure

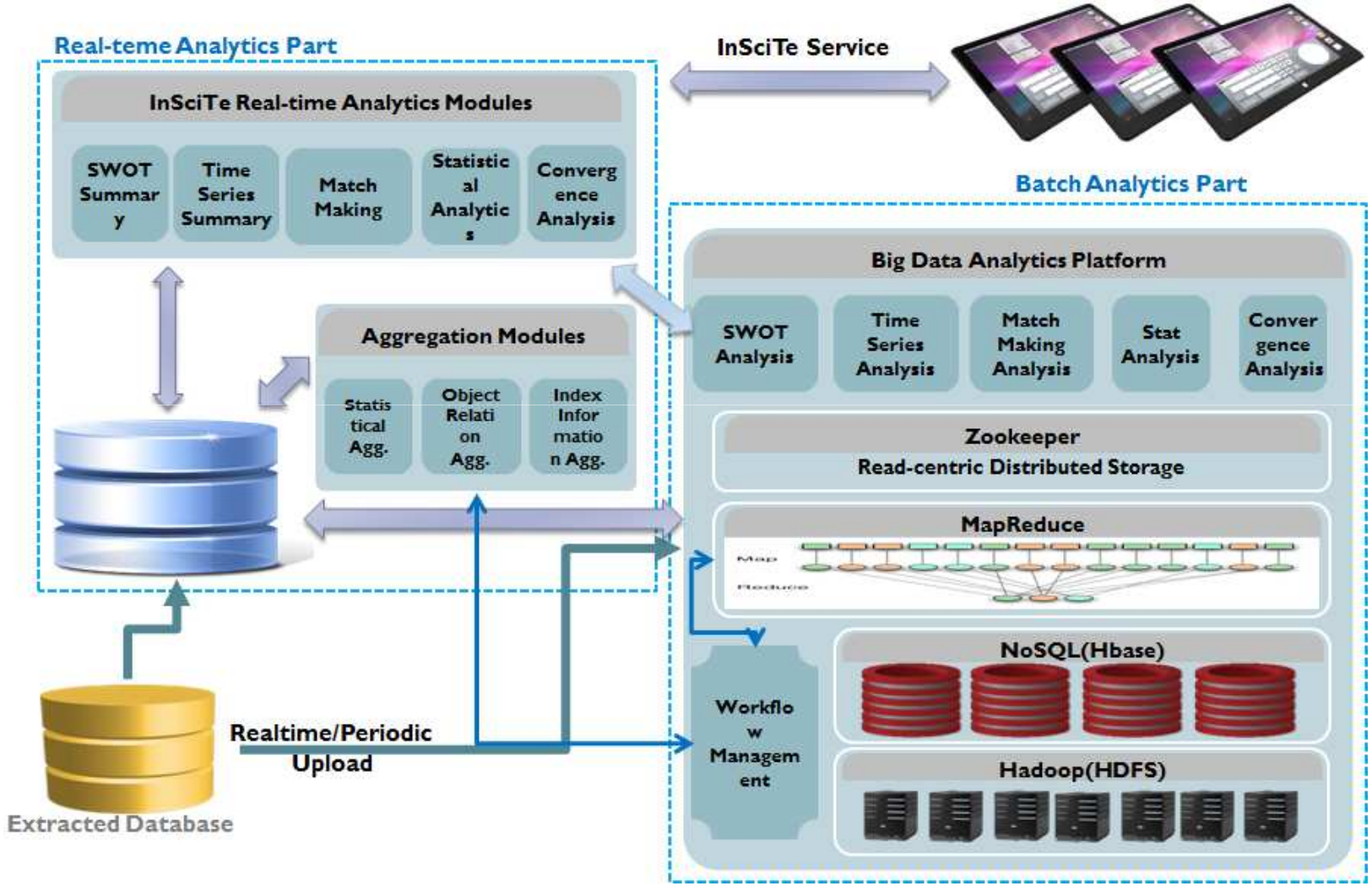
- Text Analysis + Semantic Analysis



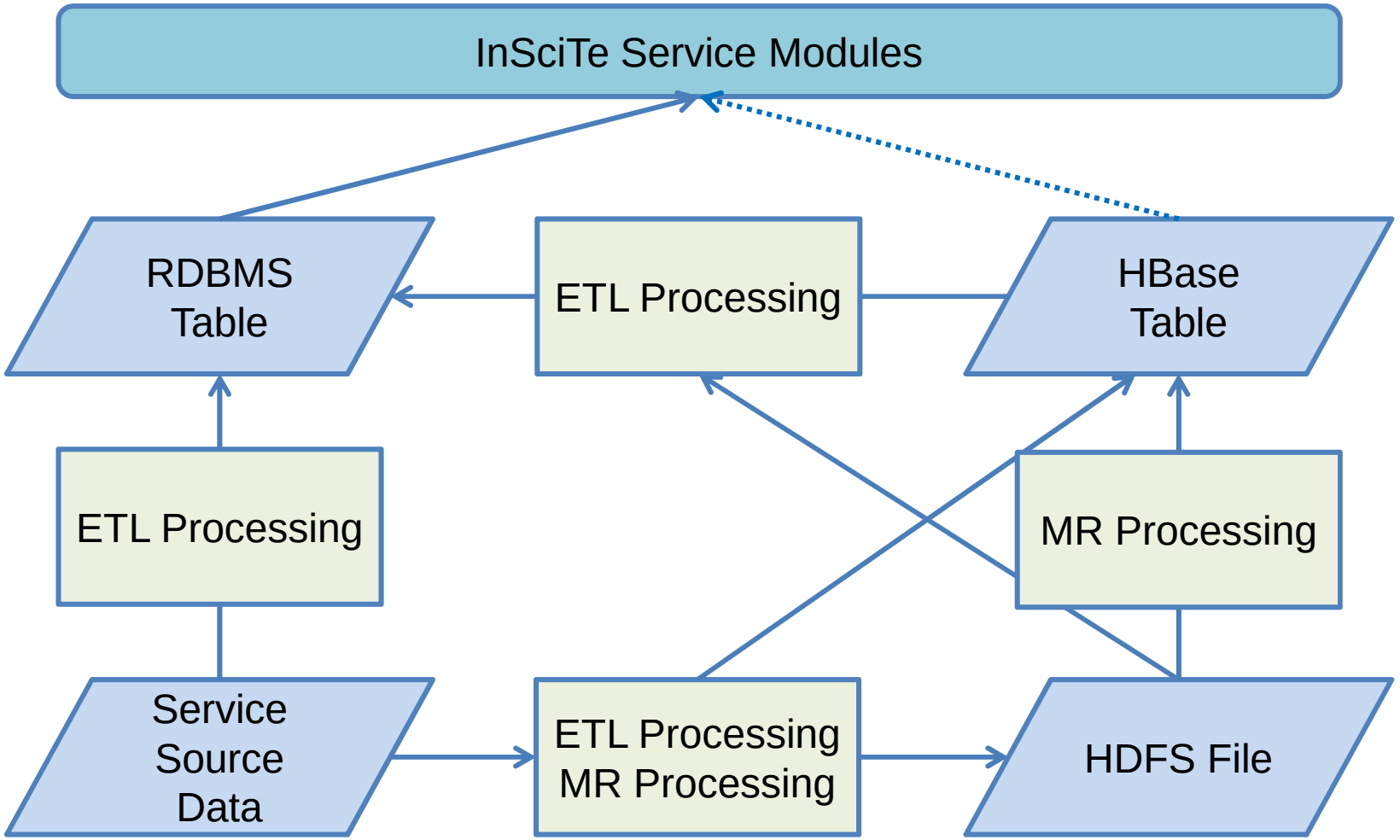
Text Analysis System



Semantic Analysis System



Data Processing with Hadoop echosystem



Layered Architecture of InSciTe System

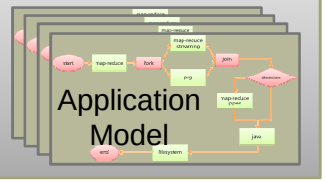
Application Layer

Application 1 | ..... | Service 1 | Service 2

Application Authoring Layer

Authoring Workbench

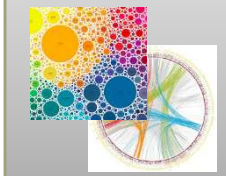
Model Verifying Engine | Model Executing Engine | Module Recommending Engine | Debugging Engine



Analytics Layer

Stat Modules	Relation Modules	SWOT Modules	Trend Analysis Modules	Match Making Processing Modules	...	Visualization Modules
Stat 1 ...	Rel 1 ...	SWOT 1 ...	TA 1 ...	MM 1 ...		Vz 1 ...
Stat 2 Stat N	Rel 2 Rel N	SWOT 2 SWOT N	TA 2 TA N	MM 2 MM N		Vz 2 Vz N

Visualization Layer



Prefuse, FLARE

Jgraph, Ubigraph, ...

Unstructured data

Structured data

Hadoop-ecosystem based Bigdata Platform Layer

Batch Processing Manager | (Near) Real-time Processing Manager

Internal Storage Manager

Inverted Index Manager | Feature Vector Manager | Co-occurrence Info. Manager | Matrix Manager | Triple Manager | Trie Manager | ...

Chukwa, Flume, Scribe	R, Mahout	Pig, Hive	Hcatalog, Avro	Graph DB(Giraph)	Memcached, MemBase
Sqoop, hiho	NoSQL	HDFS	MapReduce	Zookeeper	Esper, Storm, Hstreaming

RDB

InSciTe

Layered Architecture

- Big Data platform layer: various Hadoop related tools are included in addition to RDBMS in order to support the upper level layers
- Analytics layer: Grouped and modularized analytics modules including statistical analytics, relation analytics, SWOT analytics, trend analytics, match-making analytics,
- Visualization layer: Analytics module specialized in Big Data visualization.
- Application authoring layer: Application authoring workbench for building user's application composed of one or more analytics modules in the analytics layer.
- Application layer: applications or services based on the application models built in the application authoring layer.

GO


Demo
NO
TRIDENT
REPLACEMENT
Don't attack
Iran
Troops
Out
HANDS OFF
IRAN
Socialist Worker
Bush
Blair

Q&A

Sa-kwang Song
esmallj@kisti.re.kr
http://inscite.kisti.re.kr
Dept. of Computer Intelligence Research
Korea Institute of Science and Technology Information