From Big Text To Big Knowledge!

Machine And Language Learning (MALL) Lab
SERC, Indian Institute of Science

What is Big Text?

Our lab deals only with text data!

Tweet Hand @dweebtweet · 16m
RT Niksomania: I must admit first 2 wins in #**CWC15** have made world Cup mus
see for almost every Indian..
#WontGiveItBack

Tweet Hand @dweebtweet · 16m
RT BalajeeRao: India Vs UAE World Cup AD: ift.tt/1vsyjCL; #**CWC15** extremely
funny #WontGiveItBack

ONE News - Sport @ONENewsSport · 17m
Aaron Finch: Australia not expecting any Black Caps surprises tvnz.co.nz/cricke
news/a… #**CWC15** #australia #NewZealand

1

View summa

Ejaz Rasool Chaudhry @EjazRC · 17m
Take a chill pill guys; IT'S JUST A GAME; try to focus on the real issues of/in the
country ...!
#LastTweet #**CWC15** #Pakistan #RealIssues

KOUSHIK BOGGARAPA @koushik04 · 17m
I don't think anyone can question DHONI
'LEADERSHIP' - VIRAT KOHLI #**cwc15**

1

twinklingtina @twinklingtina · 17m
There isn't a good match till Friday!! #**CWC15**

1

Explosion of Unstructured Text!

300 million new websites added in 2011 alone (a 117% growth)

500 million Tweets per day (circa Oct 2012)

Time to read for one person: 31years

Do you think this surplus of data is useful?

Why do you think this surplus of data is useful?

Yes, more data gives us a chance to do more!

Need to harvest knowledge from unstructured text data

What is Big Knowledge?

- Understanding what all this data means.

- Moving from strings to things!

Google

prime minister of india

Web News Images Videos Maps More Search tools

About 9,75,00,000 results (0.42 seconds)

Narendra Modi

India, Prime minister

Narendra Damodardas Modi is the 15th and current Prime Minister of India, in office since 26 May 2014. Modi, a leader of the Bharatiya Janata Party, previously served as the Chief Minister of Gujarat from 2001 to 2014. Wikipedia

Born: September 17, 1950 (age 64), Vadnagar

Spouse: Jashodaben Modi (m. 1968)

Party: Bharatiya Janata Party

Education: Gujarat University (1983), University of Delhi (1978)

Parents: Heeraben Modi, Damodardas Mulchand Modi

Siblings: Pankaj Modi, Prahlad Modi, Vasantiben Hasmukhlal Modi, Soma Modi

Say Hello to Knowledge Bases!

A **knowledge base** (KB) is a technology used to store complex structured and unstructured information used by a computer system.

Existing Data

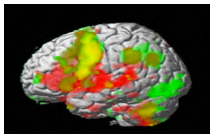
New Data

Construct KB

Maintain KB

Apply KB

Improve KB



Applications that use the KB

Automatically build, maintain, and make the KB available to intelligent applications (e.g., social media analysis, robotics, etc.)

Recent posts on Google+

Narendra Modi
23,57,736 followers • Shared publicly

Follow

While inaugurating Hazaribagh-Koderma rail line, spoke on need to transform railways with a focus on modernisation & better service delivery. We want to infuse ... 20 Feb 2015

People also search for

View 15+ more

Arvind

Amit Shah

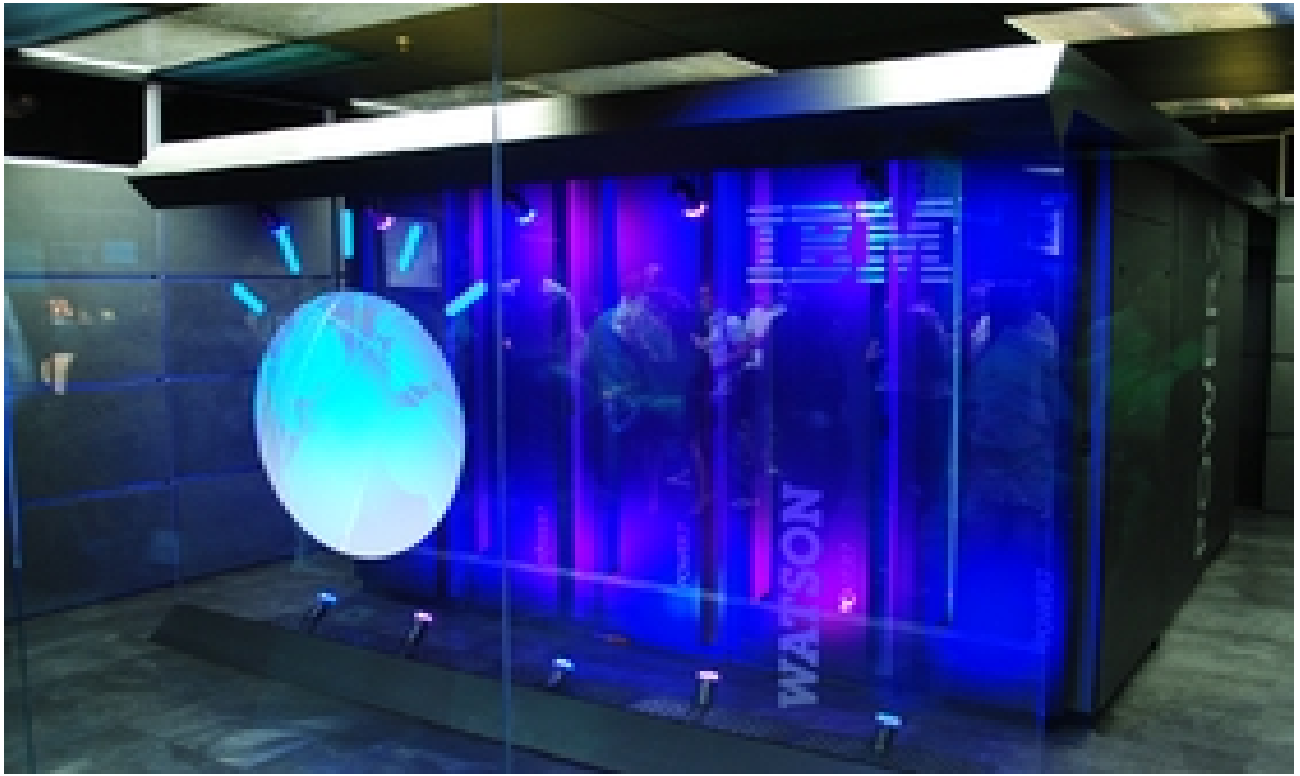
Rahul

Manmohan

Pranab

We can use the KB to find people similar to Narendra Modi!

It’s the IBM Watson!



What is Watson?

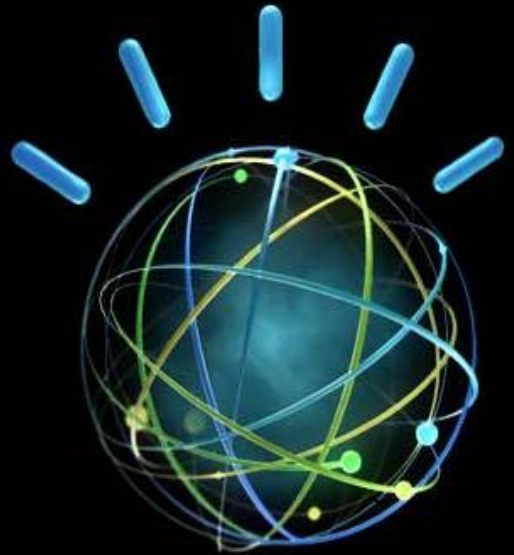
As a cognitive technology, Watson is a natural extension of what humans can do at their best.

Natural Language

Watson can read and understand natural language, important in analyzing unstructured data that make up as much as 80 percent of data today.

Hypothesis Generation

When asked a question, Watson relies on hypothesis generation and evaluation to rapidly parse relevant evidence and evaluate responses from disparate data.



Dynamic Learning

Through repeated use, Watson literally gets smarter by tracking feedback from its users and learning from both successes and failures.

What makes it so smart?

Background Knowledge!

Lots of it!

Background world knowledge is key to Intelligent Decision Making

Where do we get knowledge base (KB) of world facts?

- manual entry doesn't scale
- our approach: read the web on a never-ending basis

Fragment of a KB
(automatically extracted)

A simple example!

Document 1

... Luke Ravenstahl is the

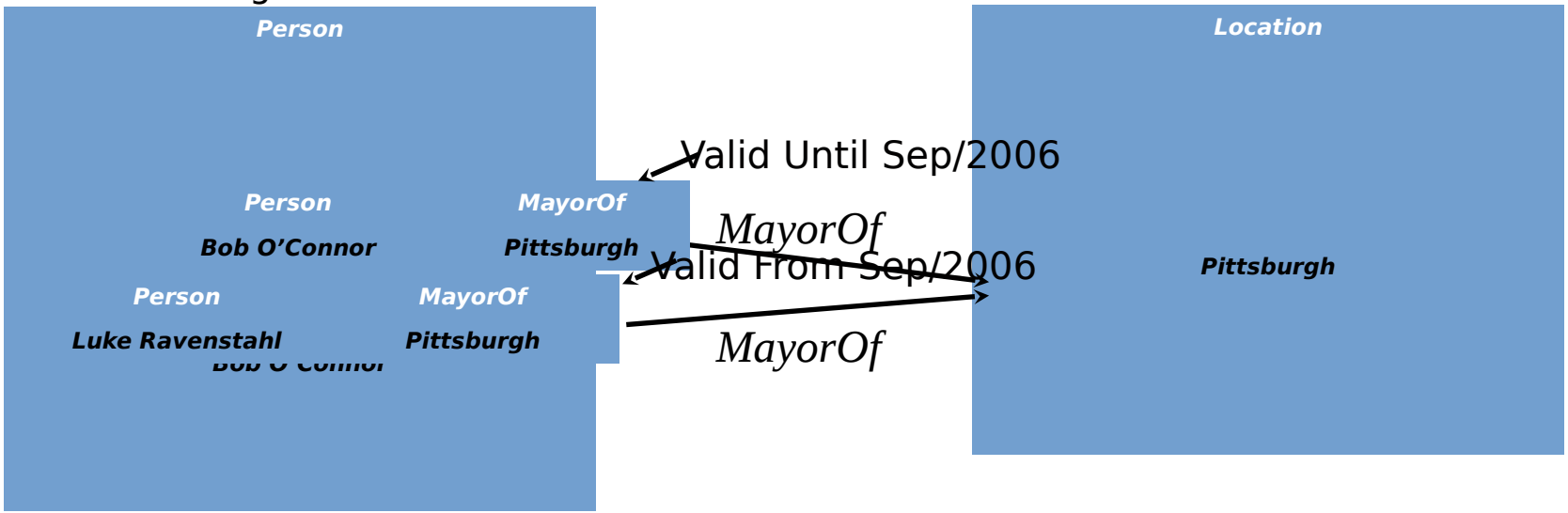
current Mayor of

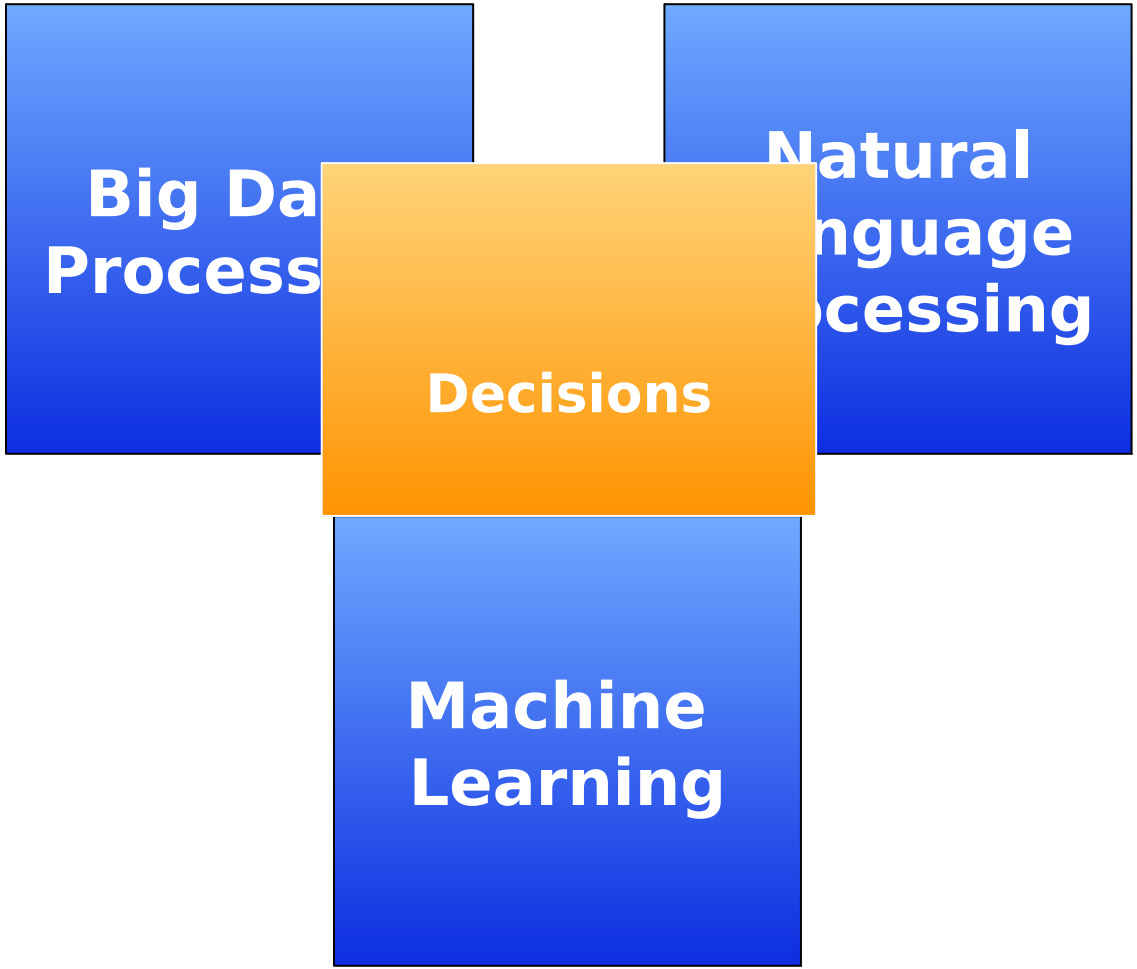
Pittsburgh ...

Document 2

... After the death of then-mayor Bob O'Connor, Luke

Ravenstahl became the mayor in September 2006 ...




Big Da
Process
Natural
nguage
ocessing
Decisions
Machine
Learning

Why is Knowledge Harvesting Hard?

- Lots of semantic classes, how to assign them to entities?
- How to discover relationship among entities?
- How to determine when those relationships are true?

We address these challenges by:

- learning from limited human supervision
  - to correct the computer when it isn't really sure due to lack of data
- exploiting redundancy of information
- developing scalable algorithms

Never Ending Learning Agent!

- Persistent software individual (runs 24x7x365)
- Learns many functions / knowledge types
- Learns easier things first, then more difficult
- The more it learns, the more it can learn next
- Learns from experience and from advice

NELL: Never Ending Language Learner

Inputs:

- initial ontology
  - ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist (planets revolve around a star)
- few seed examples of each ontology predicate (Earth revolves around the Sun)
- the web
- occasional interaction with human trainers

NELL: Never Ending Language Learner

The task:

- run 24x7, forever
- each day:
- extract more facts from the web
- learn to read (perform previous step) better than yesterday

Ongoing Research at MALL Lab

- Goal-directed KB expansion
- Continuous KB evaluation
- Temporal Micro Reading
- Representation Learning
- Knowledge on Demand

Thank You!