


# Information Distance From a Question to an Answer

Ming Li  
University of Waterloo



In this lecture, we propose a new theory, and present a system implementing this theory, for natural language processing.

In the 20<sup>th</sup> century, we have invented hi-tech:  
Replacing them: **Natural User Interface**  
Phones, TVs, Laptops



*For 3 million years, our hands have been tied by tools.  
It is time to free them, by natural interface.*

---





# But the reality is not here yet

---

- Let's ask Siri
  - Do fish sleep?
  - Where can I find dog food?
  - How hot is Sun's surface?
  - What does a cat eat?
  - Where is the Kalahari Desert?
- What is the problem?

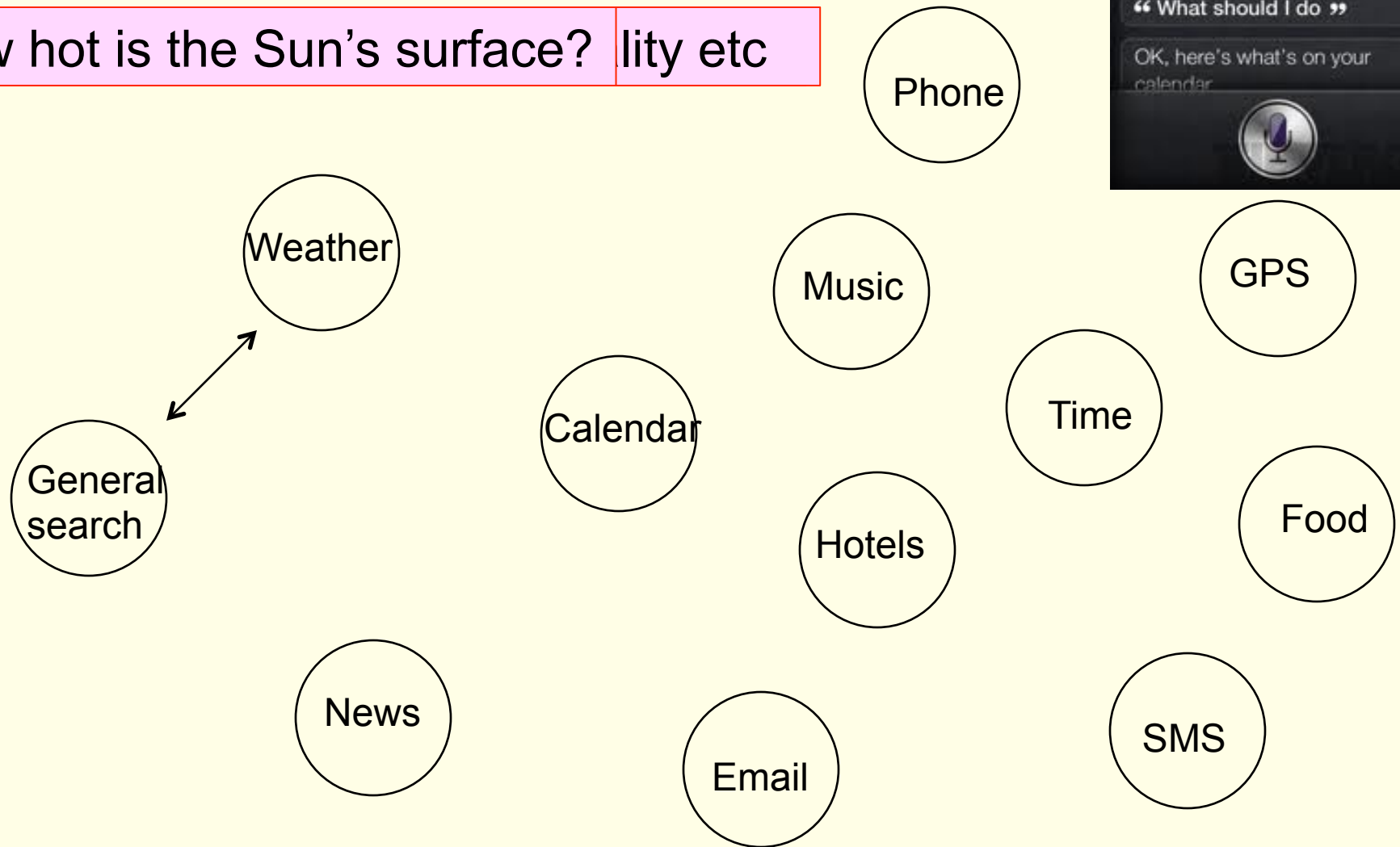
# Problem 1: keywords vs templates

---

- If you use keywords, like Siri, then you make mistakes like: “Do fish sleep? → seafood”
- If you use templates, like Evi, then you have trouble with even slight variations: “Who prime minister of Canada?” or “Who is da prime minister of Canada?”
- Second approach requires us to recognize variation distance.

# Problem 2: Domain classification

How hot is the Sun's surface? lity etc



# Problem 3: What I said is not what it heard

*To appear in CACM, July*

- Speech recognition system is not robust.
- Solution:
  - Use 40 million user asked questions, set  $Q$ .
  - Given voice recognition result  $\{q_1, q_2, q_3\}$ , we wish to find  $q$ , s.t.:
$$\{q_1, q_2, q_3\} \Leftrightarrow q \Leftrightarrow Q$$
is minimized.
- How to define the distances?

## Problem 4:

What it translates to is not what I meant

---

□ Translation systems are not ready for QA.

■ 蚂蚁几条腿? Google: Ants several legs.

□ Solution:

■ Use 40 million user asked questions, set  $Q$ .

■ Given the translation result  $q_1$ , we find  $q$ , s.t.:

$q_1 \Leftrightarrow q \Leftrightarrow Q$   
is minimized.

□ How do we define the distance?

# Problem 5: Which one is the answer?

---

- Given a question, a QA system finds many answers
- Which one is the “closest” to the question?
- Need a distance to define “closeness”

# Talk plan

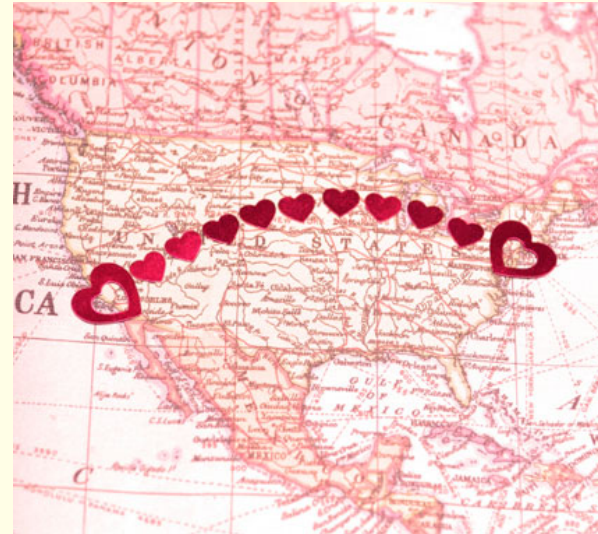
---

- Define the ultimate distance
- Apply it to solve problems 1-5, focusing on Problems 1 and 2.



# What is the “distance”?

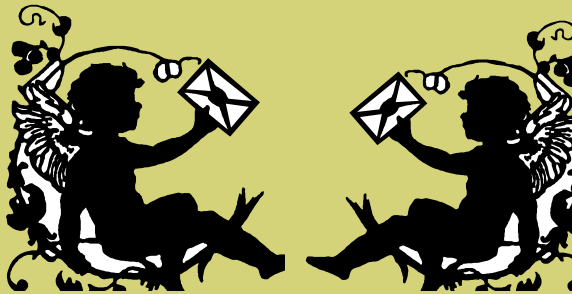
- In physical space:



- What is the distance between two information carrying entities: web-pages, genomes, abstract concepts, books, vertical domains, a question and an answer?
- We want a theory:
  - Derived from the first principles;
  - Provably better than “all” other theories;
  - Usable.

# The classical approaches do not work

- For all the distances we know: Euclidean distance, Hamming distance (sum of # of pixels that differ), nothing works. For example, they do not reflect our intuition on:



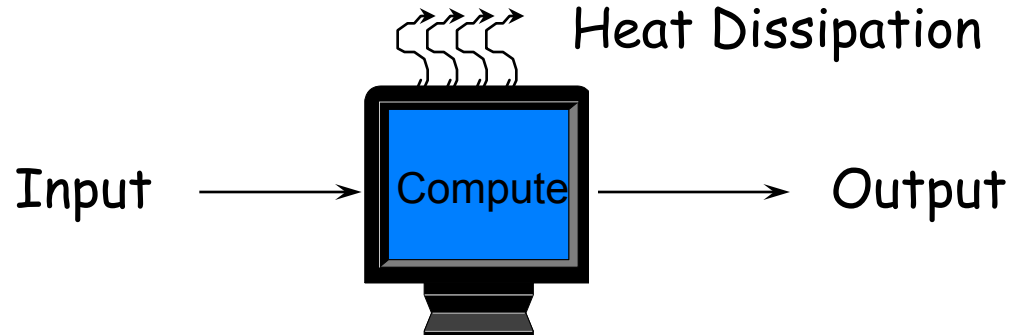
Austria



Byelorussia 1991-95

- But from where shall we start?
- We will start from first principles of physics and make no more assumptions. We wish to derive a general theory of information distance.

# Thermodynamics of Computing



Von Neumann, 1950

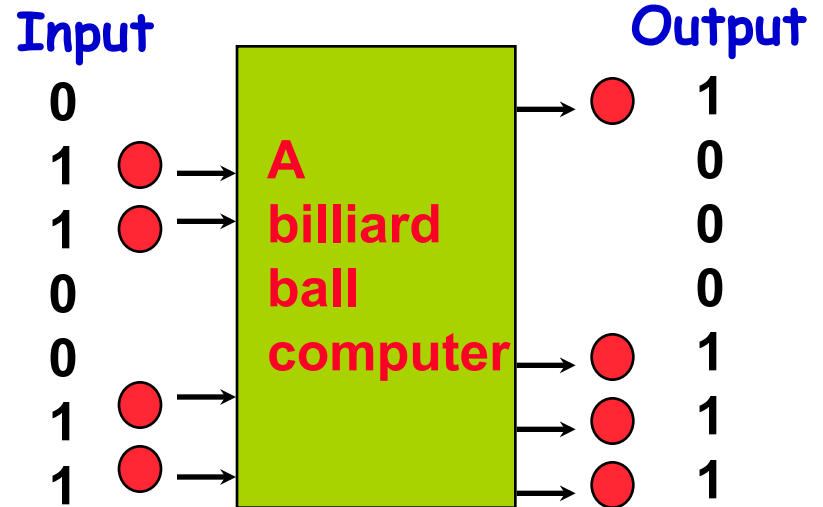
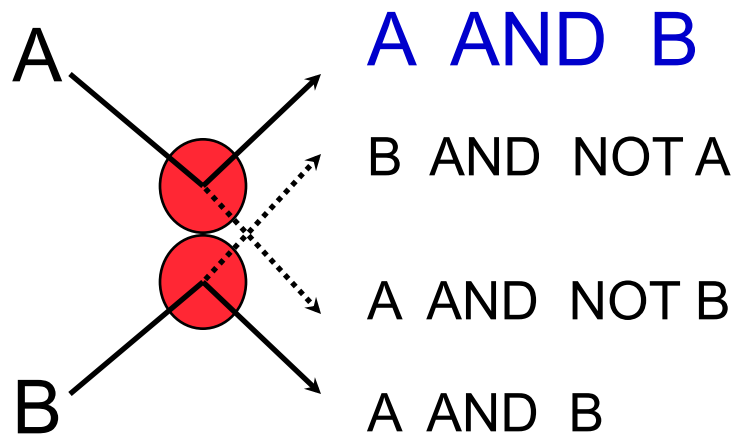
Physical Law:  $1kT$  is needed to  
(irreversibly) process 1 bit.

Landauer



# Reversible computation is free

- A billiard ball computer.

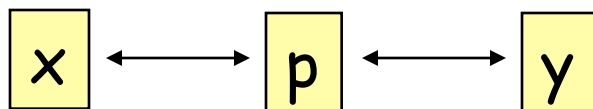


# Deriving the theory ...

---

Cost of conversion between  $x$  and  $y$  is:

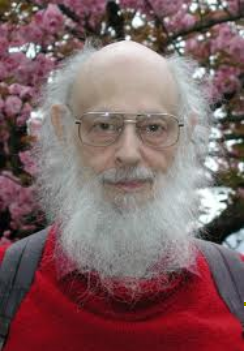
*$E(x,y)$  = smallest number of bits needed to convert reversibly between  $x$  and  $y$ .*



**Fundamental Theorem:**

$$E(x,y) = \max\{ K(x/y), K(y/x) \}$$

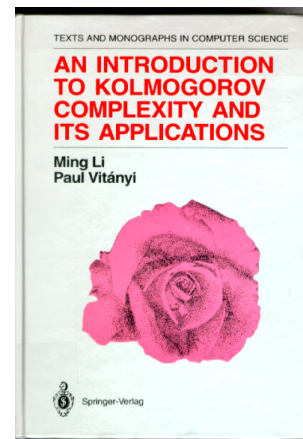
Bennett, Gacs, Li, Vitanyi, Zurek, STOC'93.



# Kolmogorov complexity



- Kolmogorov complexity was invented in the 1960's by Solomonoff, Kolmogorov, and Chaitin.
- Kolmogorov complexity of a string  $x$  condition on  $y$ ,  $K(x|y)$ , is the length of shortest program that given  $y$  prints  $x$ .  $K(x) = K(x|\epsilon)$ .
- If  $K(x) \geq |x|$ , then we say  $x$  is random.



# Proving $E(x,y) \leq \max\{K(x|y), K(y|x)\}$ .

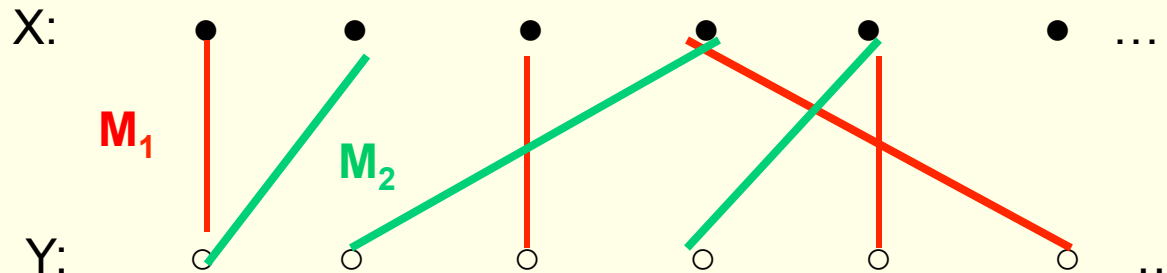
**Proof.** Define graph  $G=\{XUY, E\}$ , and let  $k_1=K(x|y)$ ,  $k_2=K(y|x)$ , assuming  $k_1 \leq k_2$

- where  $X=\{0,1\}^*x\{0\}$

- and  $Y=\{0,1\}^*x\{1\}$

- $E=\{(u,v): u \text{ in } X, v \text{ in } Y, K(u|v) \leq k_1, K(v|u) \leq k_2\}$

degree  $\leq 2^{k_2+1}$



degree  $\leq 2^{k_1+1}$

- We can partition  $E$  into at most  $2^{k_2+2}$  matchings.

- For each  $(u,v)$  in  $E$ , node  $u$  has most  $2^{k_2+1}$  edges hence belonging to at most  $2^{k_2+1}$  matchings, similarly node  $v$  belongs to at most  $2^{k_1+2}$  matchings. Thus, edge  $(u,v)$  can be put in an unused matching.

- Program P: has  $k_2, i$ , where  $M_i$  contains edge  $(x,y)$

- Generate  $M_i$  (by enumeration)

- From  $M_i, x \rightarrow y$ , from  $M_i, y \rightarrow x$ .

QED



# Information distance:

---

$$D(x,y) = \max\{K(x|y), K(y|x)\}$$

*Theorem: For any other “reasonable”  $D'$ , there is a constant  $C$ , such that for all  $x, y$ ,*

$$D(x,y) \leq D'(x,y) + C$$

# Inferring the history of chain letters:

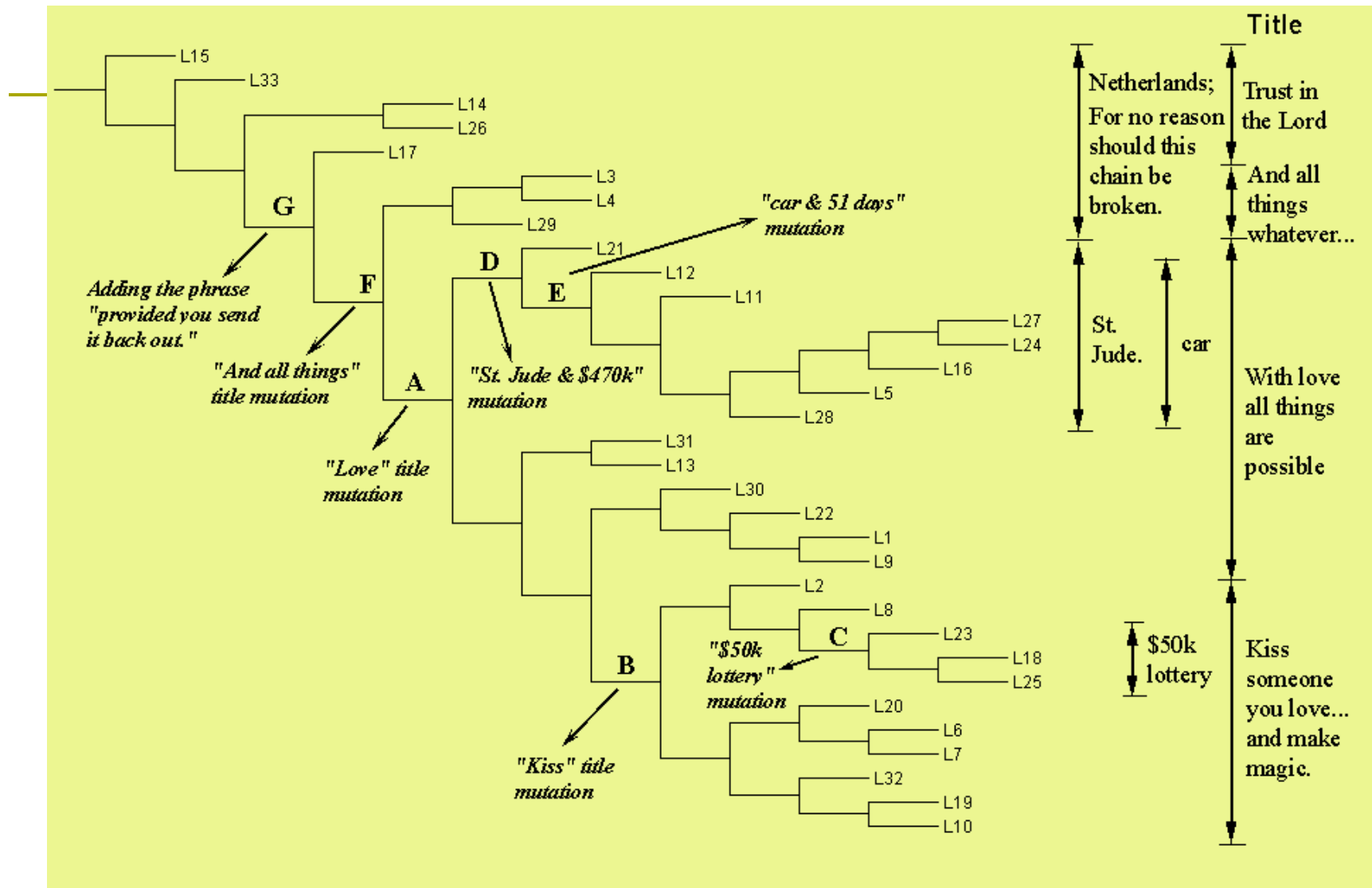
---

- For each pair of chain letters  $(x, y)$  we estimated  $D(x, y)$  by a compression program.
- Construct their evolutionary history based on  $D(x, y)$  distance matrix.
- The resulting tree is a perfect phylogeny: distinct features are all grouped together.

C. Bennett, M. Li and B. Ma, Chain letters and evolutionary histories.  
*Scientific American*, 288:6(June 2003) (feature article), 76-81.



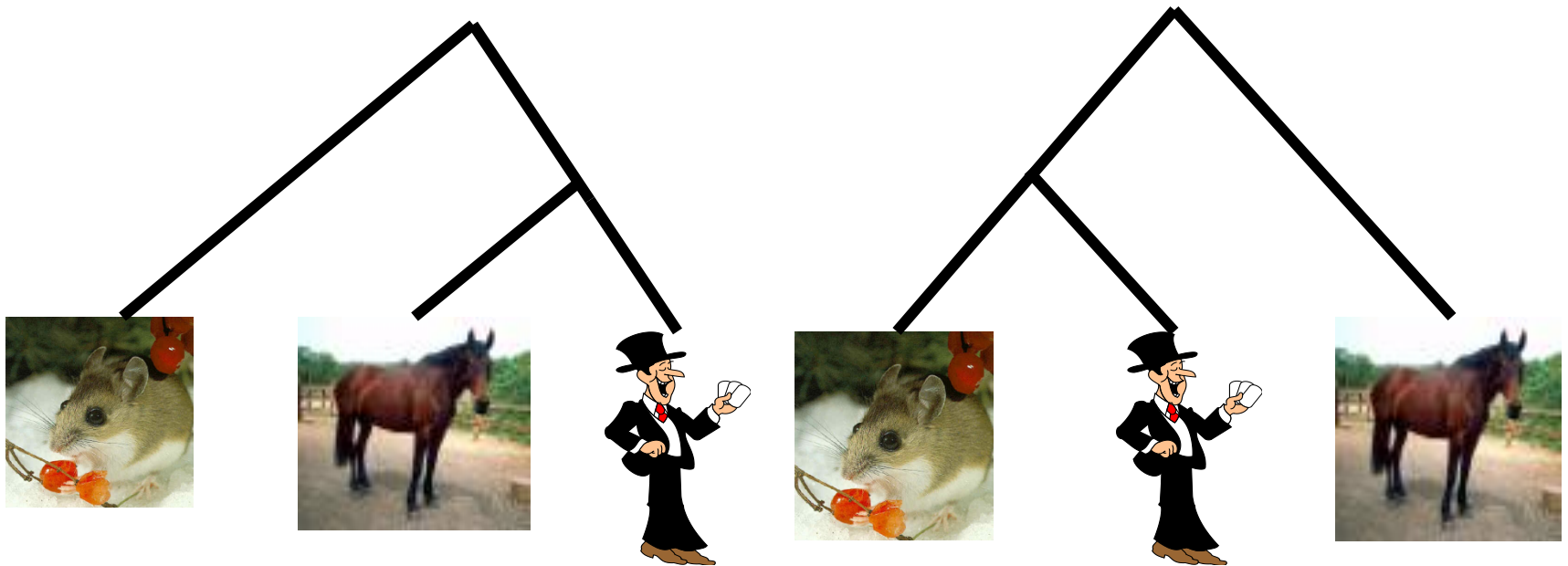
# Phylogeny of 33 Chain Letters



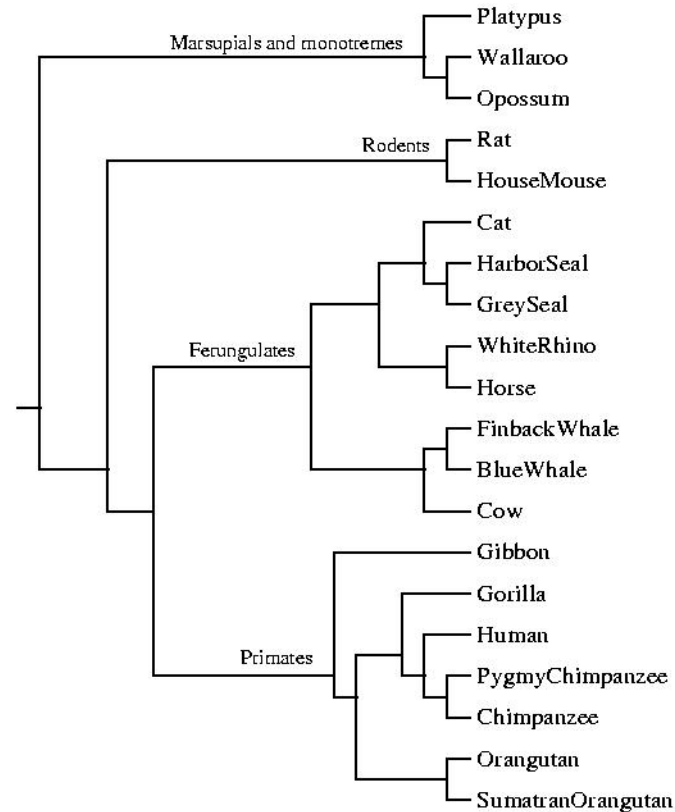
Confirmed by VanArsdale's study, answers an open question

In biology, we are often interested in finding the “**phylogenetic tree**” of species. For example, a problem is “**Eutherian Order**”: Who is our closer relative?

---



# Evolutionary History of Mammals



# This method has been applied to 100's of applications

---

- ❑ Molecular evolution
- ❑ Plagiarism detection
- ❑ Language evolution
- ❑ Image registry
- ❑ Music classification
- ❑ Hurricane risk assessment
- ❑ Protein sequence classification
- ❑ Fetal heart rate detection
- ❑ Authorship, topic, domain identification
- ❑ Network traffic analysis
- ❑ Software engineering
- ❑ Internet search
- ❑ Speech recognition

# Better than other methods

---

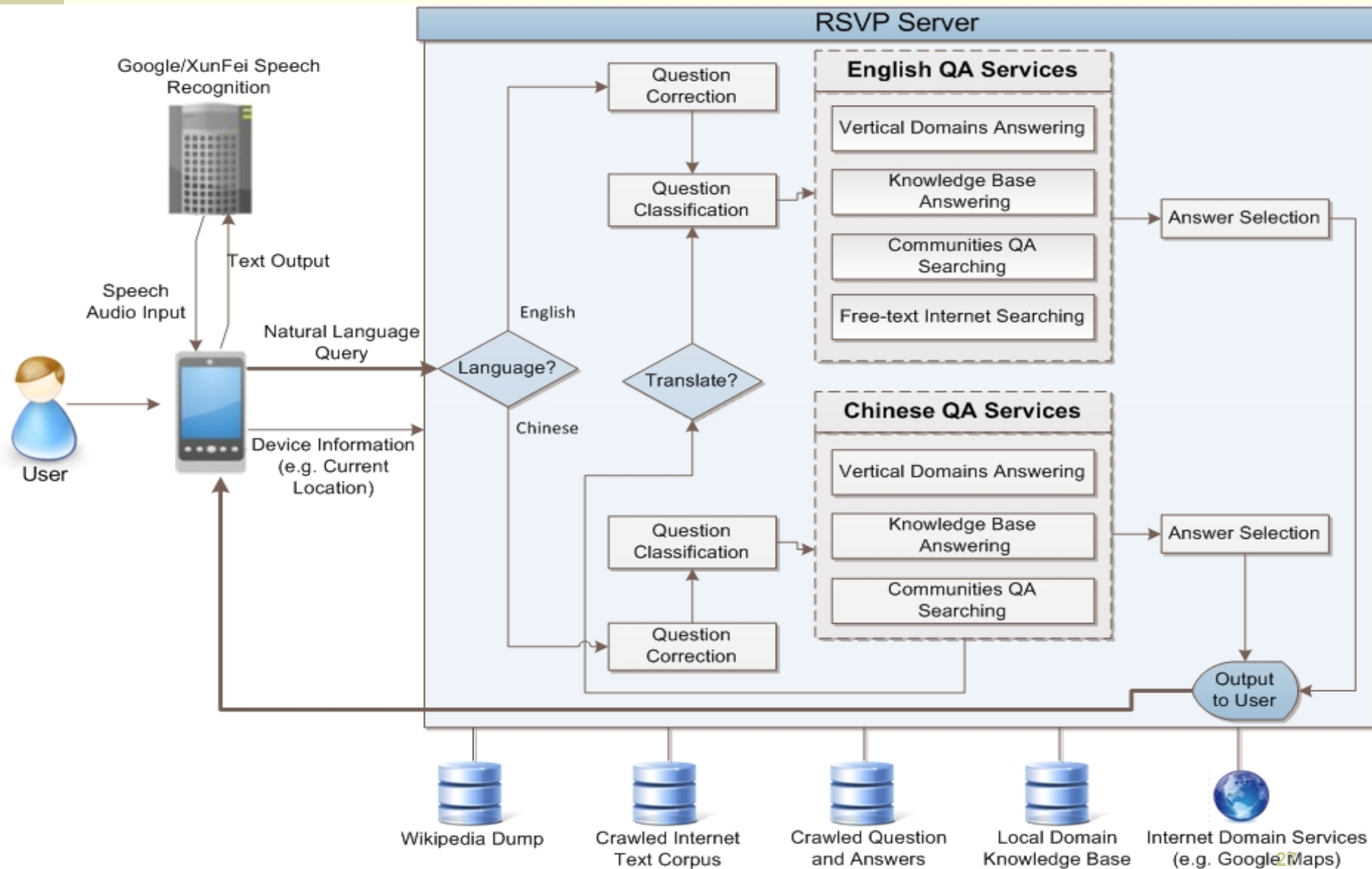
- Keogh-Lonardi-Ratananmahatana, KDD-04
  - Tested our approach against 51 other methods for classifying time series from top conferences in the field: KDD, SIGMOD, ICDM, ICDE, SSDB, VLDB, PKDD, PAKDD
  - They have concluded that our Information Distance approach performs the best, most robust, blind to applications avoiding over tuning.



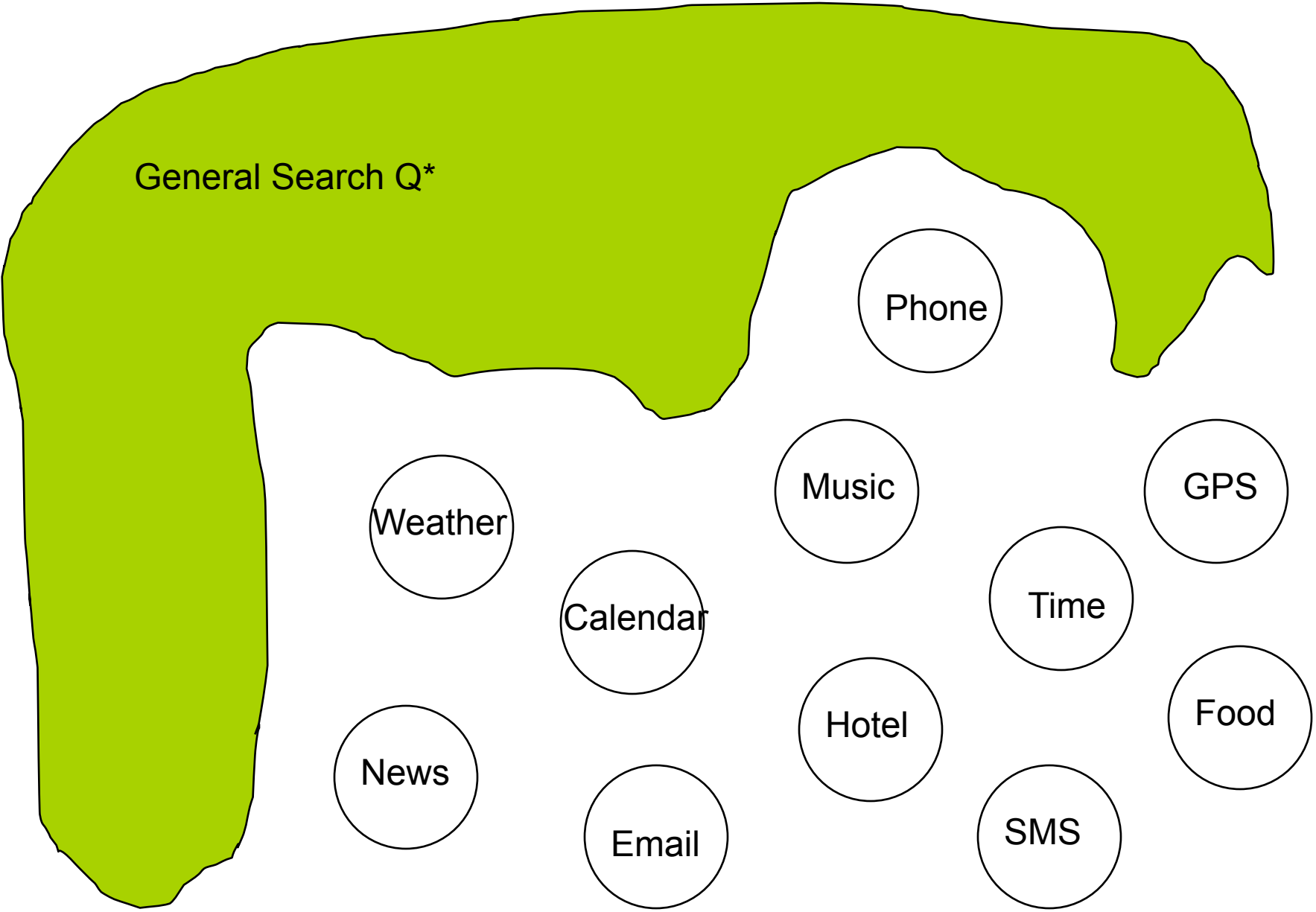
# RSVP: Natural language QA Engine

- Originally, for cross language SMS service:
  - Funded by Canada's IDRC, for developing world
  - Natural language question answering.
  - For people who are not on the internet.
- Then the project has evolved to a full fledged cross-language QA search engine.

# RSVP QA Engine Architecture



A typical “personal assistant” system



# Problem 1. Template variation

---

What is weather like in Fredericton tomorrow?

What is weather in Fredericton tomorrow?

Tomorrow what is weather like in Fredericton?

In Fredericton what will be weather like tomorrow?

How is weather in Fredericton tomorrow?

I wish to know the weather in Fredericton tomorrow?

They all mean the same –  
and they have very small information distance to each other!

# Approximating semantics

---

- ½ century of research of computational linguistics did not lead to “understanding”
- **Let's take a new path**: equate information distance with semantic distance.

# Semantic Encoding

---

- Thus, we are implementing an information distance encoding system.
- Anything with small information distance gets the same answer.

# Problem 2. Domain Classification

---

- ❑ Weather domain positive/negative samples:
  - What should I wear today?
  - May I wear a T-shirt today?
  - What was the temperature 2 weeks ago?
  - Shall I bring an umbrella today?
  - Do I need sunscreen tomorrow?
  - What is the temperature on the surface of the Sun?
  - How hot is the sun
  - Should I wear warm clothes today?
  - What is the weather like last Christmas?

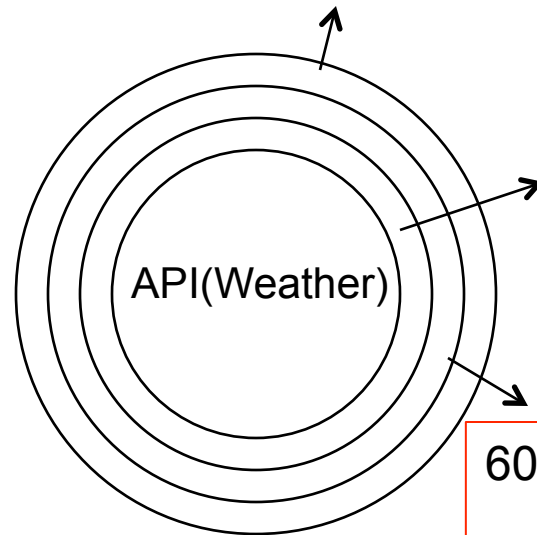


# To build up a weather domain systematically

Clusters:

What is the weather like [location phrase] ?

What is the temp [time phrase] [location phrase] ?



Keywords: weather, city, time, rain, temperature, hot, cold, wind, snow, umbrella, T-shirt,

There are ~3000 negative examples:

What is the temperature of the sun?

What is the temperature of the boiling water?

6000 questions extracted from Q:

What is the weather like?

What is the weather like today?

What is the weather like in Paris?

What is the temperature today?

What is the temperature in Paris?

# Comparison of RSVP, Siri, S-Voice on 100 typical weather “related” questions:

---

	Siri	S-Voice	RSVP
Number of Correct	46	22	94
Number of Errors	54	78	6
Success rate	46%	22%	94%

- What weather is good for an apple tree?
- What is the temperature on Jupiter?

Question	Siri	S-Voice	RSVP
What is weather like in Toronto?	Y	Y	Y
What is weather like in Sydney Australia?	Y	N	Y
What is weather like the day after tomorrow?	Y	N	Y
Do I need an umbrella tomorrow?	Y	N	Y
What is the temperature on the surface of the sun?	N	N	Y
What should I wear today?	N	N	Y
Did it snow last Christmas?	N	N	Y
What is the temperature of boiling water?	Y	N	Y
What is the temperature of melting point of gold?	N	N	Y
How hot is the boiling water?	N	N	Y
Is it going to rain tomorrow?	Y	Y	Y
What is the lowest temperature in the next few days?	Y	N	Y
Will it rain in Buffalo, New York tomorrow?	Y	Y	Y
What is the weather forecast tomorrow?	Y	Y	Y
What is the weather conditions on Saturn?	N	N	Y
Is it hot outside today?	Y	N	Y
How cold is the temperature in Canada?	N	N	Y
What was the weather forecast May 16, 2009 in Rochester NY?	Y	Y	Y
What is the weather report today?	Y	Y	Y
What is the high temperature today?	Y	N	Y
What will be the high and low temperatures tomorrow?	N	N	N
What is the temperature of the ocean in San Diego today?	N	N	N
Is it cold today?	Y	N	Y
Is the cold war still going on today?	Y	Y	Y
How hot is the great basin desert?	N	N	Y
Is it going to be nice weather tomorrow in London England	Y	N	Y
Weather conditions when the great wall was built.	N	N	Y
How is the weather today in French?	N	N	Y
What is the humidity today?	Y	N	Y
How bad is the weather today?	Y	N	N
How hot is the temperature in Israel?	N	N	Y
How hot is the temperature today?	N	N	Y
Is it going to be hot or cold tomorrow	N	Y	Y
What is the cold war period	Y	Y	Y
What is the weather like for the Superbowl?	N	N	Y
What is the weather like for the Olympics?	N	N	Y
How cold is the great salt lake	N	Y	N
Why is the weather so bad in Liverpool?	N	Y	Y
How do weather forecast help a fisherman?	Y	Y	Y
What is the temperature on Jupiter?	Y	N	Y
What is the hot weather condition?	N	N	Y
Why do you have good weather when you have high temperature?	N	N	Y
What jobs need to know the weather forecast?	N	N	Y
What would be an accurate source for a weather forecast?	N	N	Y
What is the weather on planet Jupiter?	N	N	Y
What is the weather like in the pacific ocean in December?	N	N	Y
What is the temperature on the surface of the sun?	N	N	Y
What is the surface temperature of planet Jupiter?	Y	N	Y
What is the low temperature on Mars?	N	N	Y
What is the great white shark's body temperature?	Y	N	Y

Question	Siri	S-Voice	RSVP
What is the average temperature for a dog?	Y	N	Y
What is a good temperature to keep a freezer at?	Y	Y	Y
What causes the San Francisco weather?	N	N	N
What are the wind tunnels used for today?	N	N	Y
Is rising temperature a great threat?	Y	N	Y
Is black bad to wear in hot weather?	Y	N	Y
Is a 36 degrees body temperature bad?	Y	N	Y
How to encourage your horse to use his shelter in bad weather	N	N	Y
How hot should a 350 Chevy engine run?	N	N	Y
How is hot dog prepared?	N	N	Y
How does weather forecast help the captain of a ship?	Y	N	Y
How does human body take care of itself in the cold weather?	Y	N	Y
How does humidity affect air temperature?	N	N	Y
How do you survive bad weather?	Y	N	Y
How does rain help us today?	N	N	Y
How cold can ice temperature go?	N	N	Y
How cold is Jupiter?	N	Y	Y
How are sports influenced by bad weather?	N	N	Y
what is good temperature for hamsters?	Y	Y	Y
What is temperature for hamsters?	N	N	Y
How does the sun affect the weather?	N	N	Y
How cold can it be outside for a dog?	N	N	Y
How cold weather can cats deal with outside?	N	N	Y
Is it always hot in New York?	Y	N	Y
Is the weather bad?	Y	Y	N
What is good temperature to set refrigerator?	Y	N	Y
What is the highest temperature on Mars?	N	Y	Y
What is the high temperature on Mars?	N	N	Y
What is the weather like in the sun?	N	N	Y
Who is Good Morning America's weekend weather person?	Y	Y	Y
How does weather forecast help farmers?	Y	N	Y
What weather is good for apple tree?	N	N	Y
Does acid rain occur in New York?	Y	Y	Y
How cold does it have to be for a dog to freeze?	N	N	Y
How are air temperature and humidity related?	Y	N	Y
How do scientists make weather forecast?	Y	N	Y
How do people protect themselves from bad weather?	N	N	Y
how is bad weather started?	N	N	Y
In what temperature do the human body perform best in?	Y	N	Y
Is it bad to fish in hot weather?	N	N	Y
Is Ford Fusion good for winter weather?	N	N	Y
What is a good site to find weather in Japan?	Y	N	Y
What seabird is regarded as an omen for bad weather?	N	N	Y
What weather conditions are good for Cabernet Franc?	N	N	Y
What is the worst weather condition for tire traction?	N	N	Y
Is silk good to wear in hot weather?	N	N	Y
Will it be rainy tomorrow?	Y	Y	Y
Weather forecast for Bangkok tomorrow.	Y	N	Y
How hot is the weather in New Hampshire?	Y	Y	Y
How hot is it in Africa today?	Y	Y	Y

# Problem 3. Speech improvement

To appear in *Comm. ACM*, July, 2013

---

- Original question: Are there any known aliens?
- Voice recognition result
  - Are there any loans deviance
  - Are there any loans aliens
  - Are there any known deviance
- RSVP outputs: Are there any known aliens

# Problem 4. Translation

To appear in *Comm. of ACM*, July, 2013

---

- 从深圳到北京坐飞机多长时间？
  - Google translation: Long will it take to fly from Shenzhen to Beijing?
  - Bing Translation: From Shenzhen to Beijing by plane to how long?
  - RSVP translation: How long does it take to fly from Shenzhen to Beijing
- 恐龙是什么时候灭绝的？
  - Google: Dinosaur extinction when?
  - RSVP: What time is the extinction of the dinosaurs?

# Translation experiments:

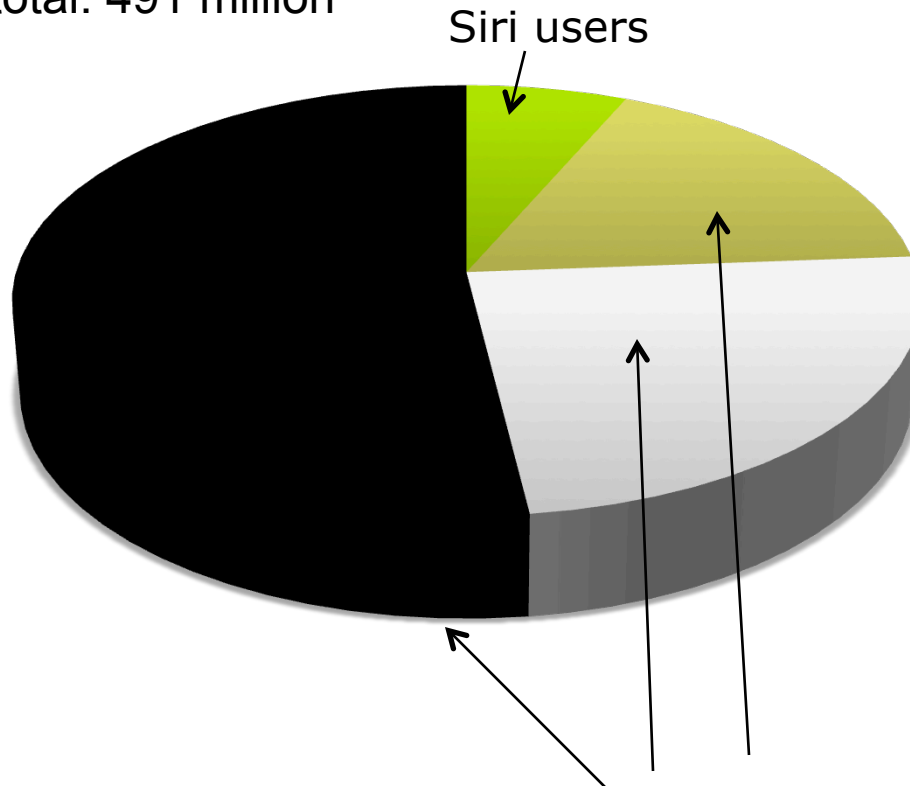
---

Table 3. Experimental results of translation in QA domain

Experiment	# questions	CC	WC	CW	WW	Base Translator Accuracy	RSVP Accuracy
Google as base translator	428	112	211	6	99	27.5%	75.6%
Microsoft as base translator	428	116	207	11	94	29.6%	75.6%
Google as base translator	52	21	15	0	16	40%	69%
Google as base translator	114	44	49	1	20	38%	81.6%

# Importance of Translation

Smartphones: 2<sup>nd</sup> Quarter 2012:  
US: 23.8 million (down from 24)  
China: 44.4 million (up from 24)  
2011 world total: 491 million



Can we reach these people?



- Native English Speakers, 375 million
- Non Native English Speakers, a billion
- Chinese speakers, 1.4 billion
- Others



# Conclusion

---

- Why is all these useful?
- A case study...

# Collaborators:

---

- ❖ Information distance: C. Bennett, P. Gacs, P. Vitanyi, W. Zurek
- ❖ RSVP system: B. Ma, J.B. Wang, Y. Tang, D. Wang, K. Xiong, X. Cui, C. Sun, J. Bai, Z. Zhu, G.Y. Feng.
- ❖ Financial support: Canada's IDRC, PDA, Killam Prize, C4-POP, CFI, NSERC.

# Experiments summary

---

**Table 2. Experimental results for speech correction**

Experiment	Total No. of questions	CC	WC	CW	WW
1.	164	105	39	5	15
2.	300	219	25	6	50
3.	300	222	15	5	58
4.	257	141	41	7	68
5.	181	100	26	4	51
6.	214	125	29	10	50
7.	206	145	19	8	34
8	298	180	12	4	102
9	131	77	14	0	40
10	57	28	4	1	24
11	63	35	9	1	18
12	107	62	9	2	34