移动计算时代NLP研究的挑战和机遇

李志飞

Google → Mobvoi

zfli@mobvoi.com

Outline

- **移动计算时代的来临**
  - **摘自Mary Meeker 的Internet Trends报告**
- 移动计算的技术特点
- NLP相关的应用分析
  - SIRI
  - Word Lens
  - Google Now
- Compact Data Structures for Mobiles
- 结束语

移动计算时代的来临

INTERNET TRENDS

D10 CONFERENCE
5/30/2012

Mary Meeker

http://www.kpcb.com/file/kpcb-internet-trends-2012

移动设备的普及

1.1B Global Mobile 3G Subscribers, 37% Growth, Q4 – @ Only 18% of Mobile Subscribers

Rank	Country	CQ4:11 3G Subs (MM)	3G Penetration	3G Sub Y/Y Growth	Rank	Country	CQ4:11 3G Subs (MM)	3G Penetration	3G Sub Y/Y Growth
1	USA	208	64%	31%	16	Canada	16	62%	34%
2	Japan	122	95	9	17	Taiwan	14	48	17
3	China	57	6	115	18	South Africa	13	21	49
4	Korea	45	85	10	19	Turkey	13	20	62
5	Italy	44	51	25	20	Portugal	13	78	19
6	UK	42	53	25	21	Vietnam	12	11	358
7	Brazil	41	17	99	22	Mexico	11	11	55
8	India	39	4	841	23	Malaysia	10	27	7
9	Germany	38	36	23	24	Sweden	10	73	25
10	Spain	33	57	21	25	Philippines	10	11	45
11	France	30	45	35	26	Saudi Arabia	10	19	17
12	Indonesia	29	11	27	27	Netherlands	9	44	34
13	Poland	28	57	17	28	Egypt	8	10	60
14	Australia	22	76	21	29	Austria	7	58	24
15	Russia	17	8	45	30	Nigeria	6	6	51

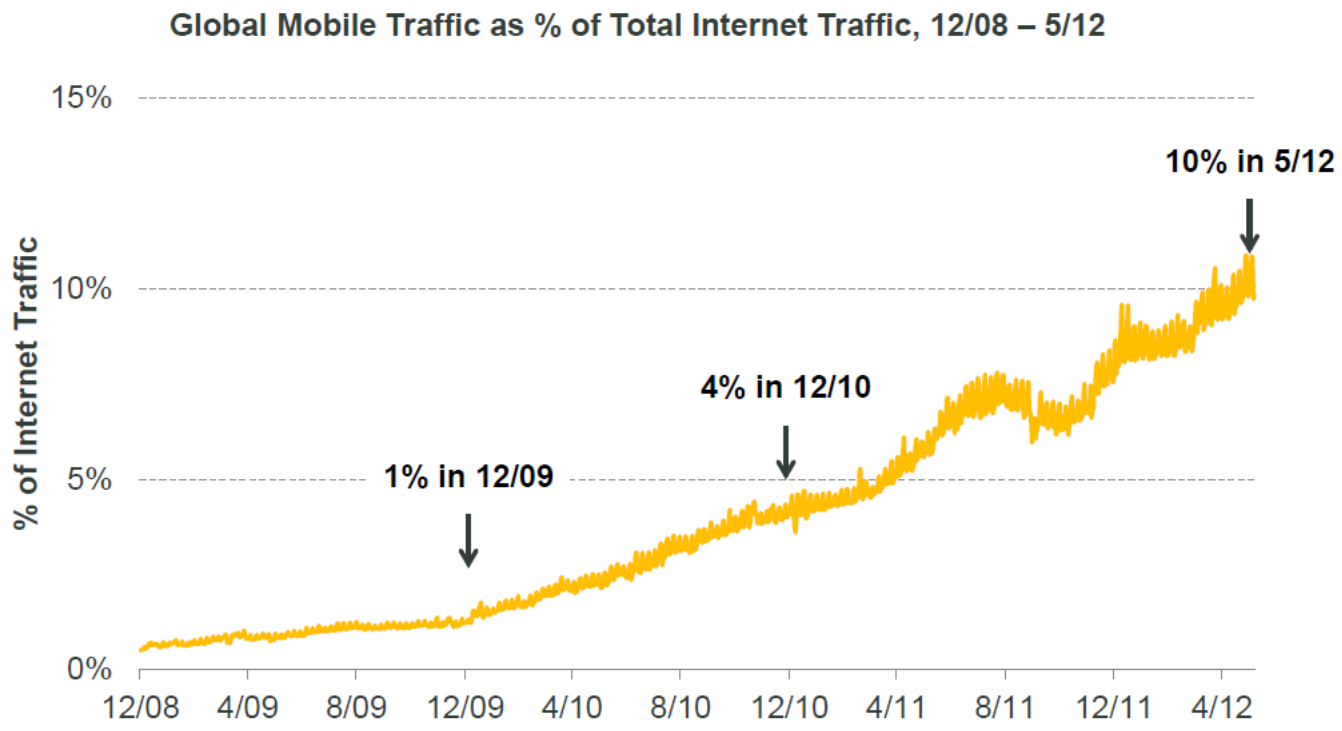
Global 3G Stats:	Subscribers = 1,098MM	Penetration = 18%	Growth = 37%

KPCB

Note: *3G includes CDMA 1x EV-DO and Rev. A/B, WCDMA, HSPA; One user may have multiple mobile subscriptions and may be counted as multiple subscriber. Source: Informa WCIS+.

移动流量快速增长

Global Mobile Traffic Growing Rapidly to 10% of Internet Traffic

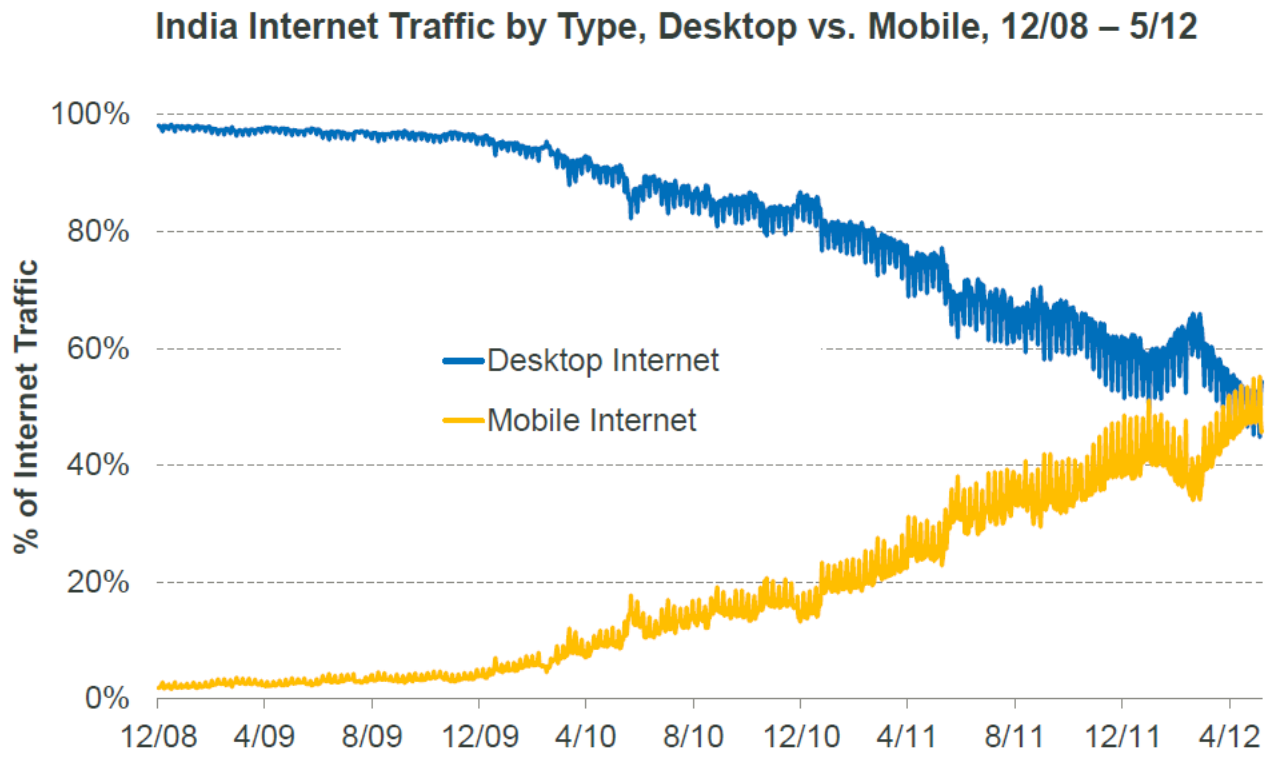


KPCB

Source: StatCounter Global Stats.

移动流量快速增长

Good / Bad News – Rapidly Growing Mobile Internet Usage Surpassed More Highly Monetized Desktop Internet Usage in May, 2012, in India

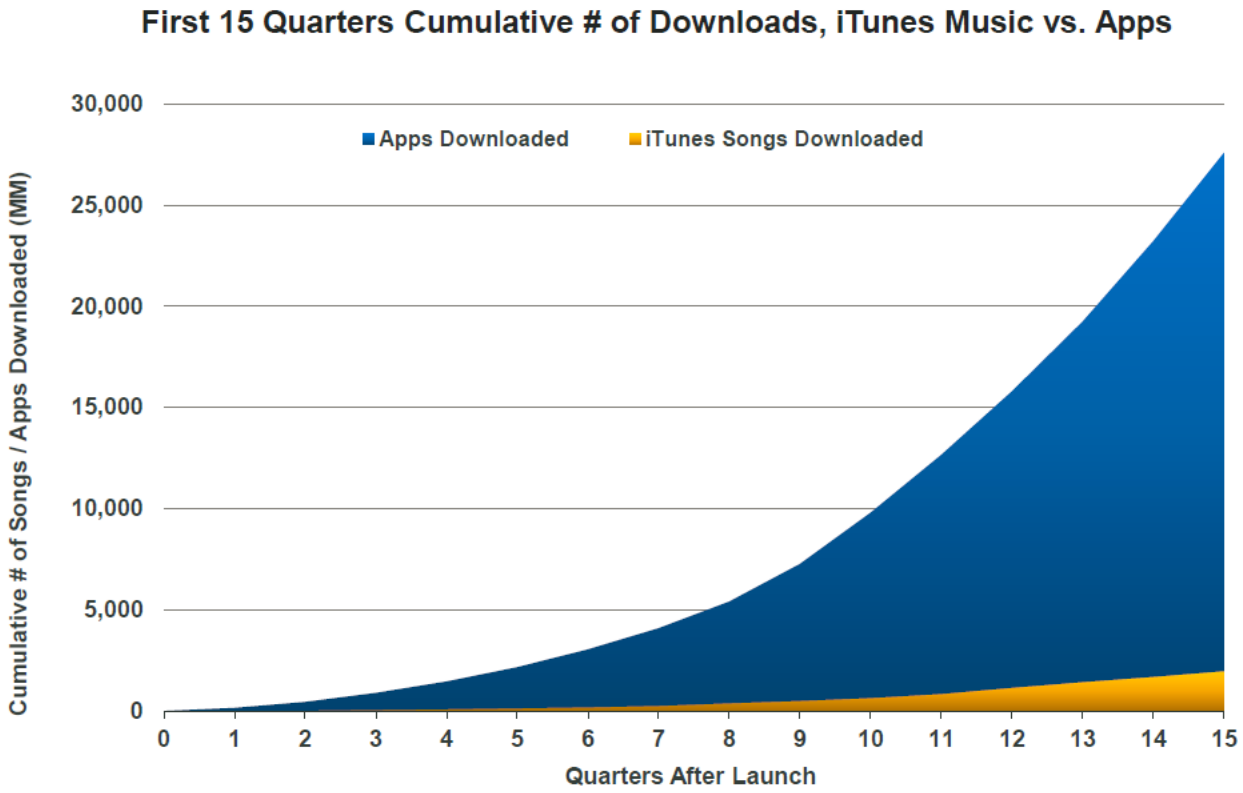


KPCB

Source: StatCounter Global Stats.

移动应用的快速增长

Apple App Store Distribution – iTunes App Store Driving 46MM+* Downloads per Day

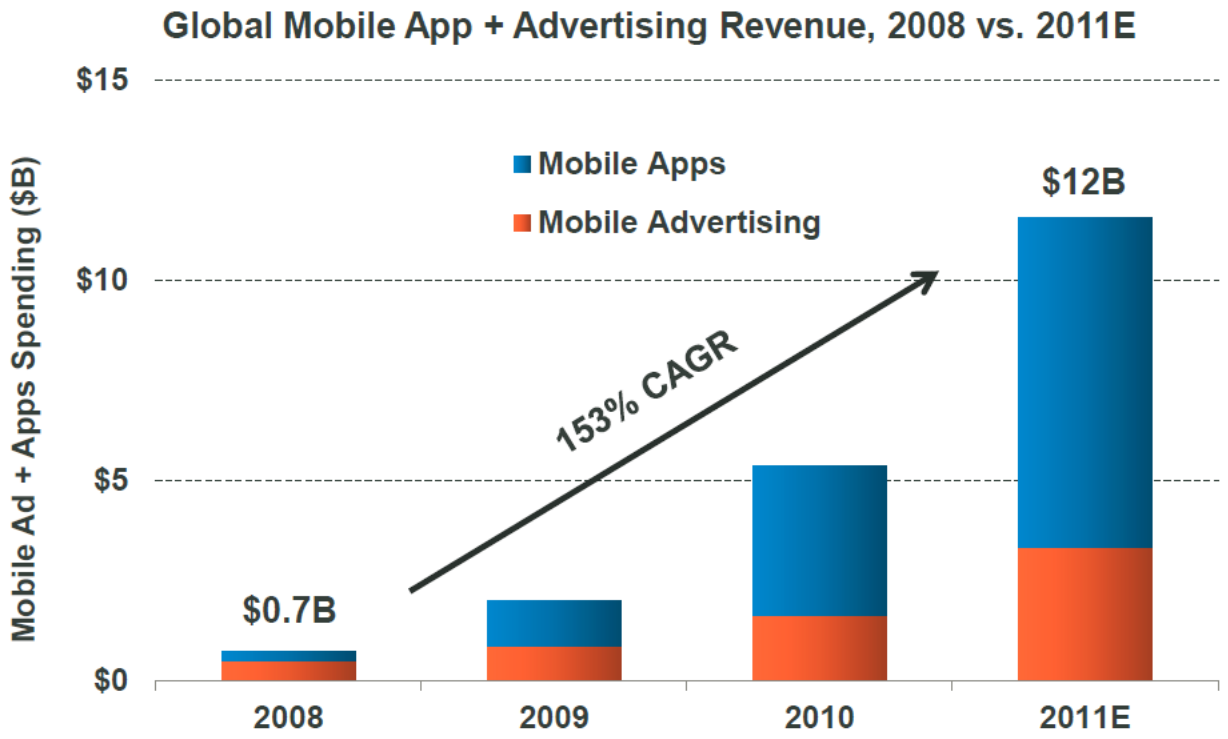


KPCB

Note: * 46MM daily app downloads calculation based on days between Apple announced milestones (18B downloads as of 10/4/11 and 25B downloads as of 3/5/12). iTunes Music store launched in CQ2:03, App Store launched in CQ3:08. Source: KPCB estimates based on Apple data, as of CQ1:12.

移动计算的商业成功

Mobile Monetization Growing Rapidly (71% Apps, 29% Ads)

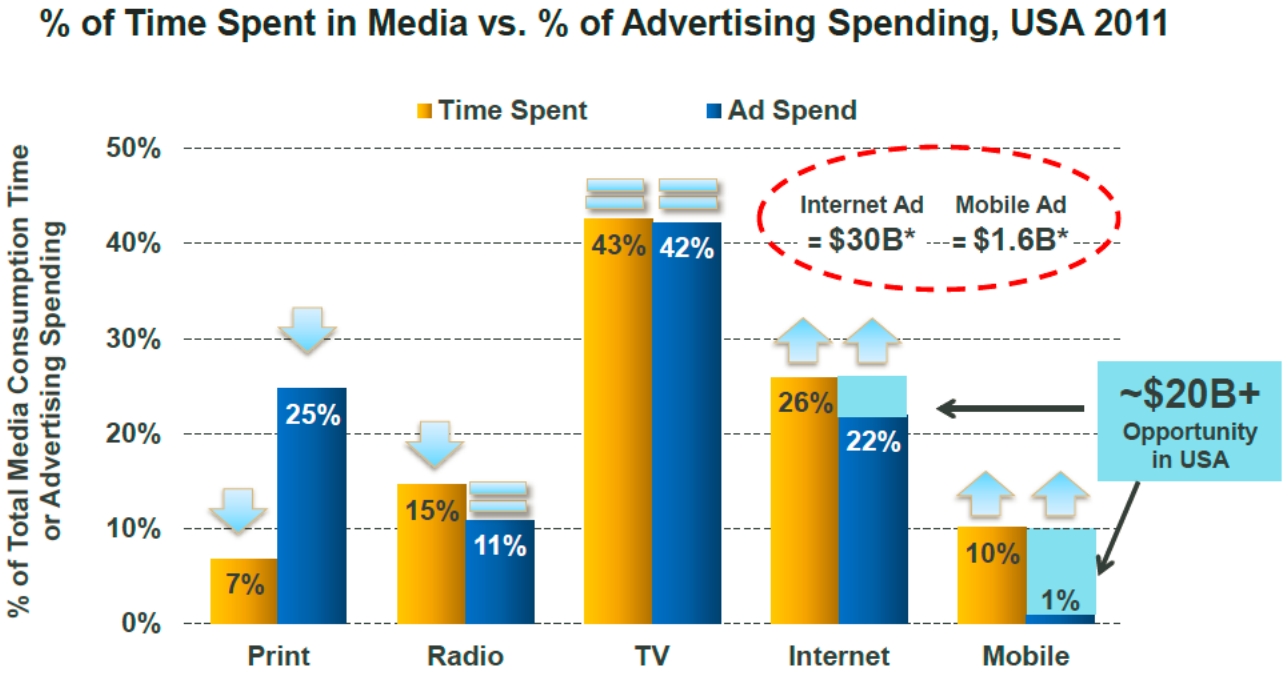


KPCB

Source: Gartner. CAGR is compound annual growth rate.
Note: Apple has paid >$3B $'s to developers as of 9/11, implying gross app market revenue of $4B in 3 years; Google indicated during CQ3 earnings call that it expects $2.5B mobile ad revenue in 2011E

移动计算的商业前景

Material Upside for Mobile Ad Spend vs. Mobile Usage



KPCB

Note: *Internet (excl. mobile) advertising reached $30B in USA in 2011 per IAB, Mobile advertising reached $1.6B per IAB. Print includes newspaper and magazine. $20B opportunity calculated assuming Internet and Mobile ad spend share equal their respective time spent share. Source: Time spent and ad spend share data eMarketer, 12/11, Internet and mobile ad dollar spent amount per IAB.

移动计算时代的
重新思考 (Re-imagination)

计算设备的变化

Re-Imagination of Computing Devices…

THEN…
(Desktops / Notebooks)

NOW…
(Tablets / Smartphones)

KPCB

电话通信的变化

Re-Imagination of Connectivity…

THEN…

NOW…

KPCB

杂志的变化

Re-Imagination of Magazines…

THEN…

Piles of Print Copies

NOW…

(Flipboard)

More Content / Always Up-To-Date / Personalized / Access Everywhere / Interactive (Video + Audio) / Share

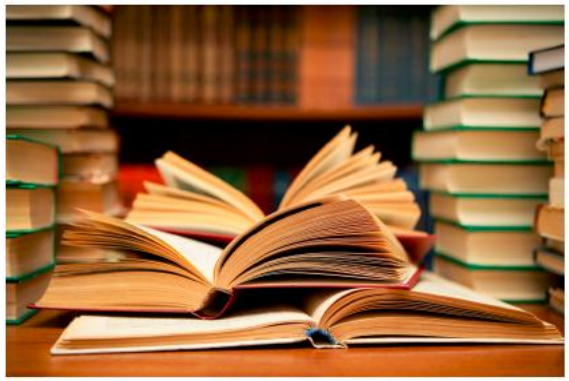
Your new Flipboard Instagram. Social search. Speed.

KPCB

书的变化

Re-Imagination of Books…

THEN…



NOW…
(Amazon Kindle / Apple iBooks)

KPCB

音乐的变化

Re-Imagination of Music…

THEN…

Buy Albums + CDs in Stores /
Playback via Dedicated Players

NOW…

(Spotify…)
Discovery of Music Through Friends + Experts /
Instant On-Demand Streaming on Internet-Enabled Devices

KPCB

录音的变化

Re-Imagination of Sound…

THEN…

Tape Recorder / Hard to Edit / Share

NOW…

(SoundCloud)

Record / Edit / Upload / Playback Anywhere / Anytime / On Any Device / Playlist sharing / Discovery

KPCB

视频的变化

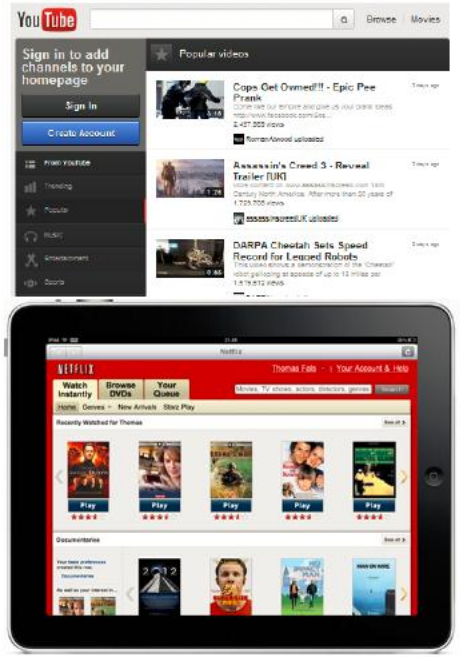
Re-Imagination of Video…

THEN…

Physical Retail / Rental Stores

NOW…

(YouTube / Netflix…)
On-Demand / Instant Streaming / Accessible Everywhere



KPCB

电话的变化

Re-Imagination of TV…

THEN…

Linear Programming / Pre-Set Channels /
Little Control Over Content

NOW…

(YouTube Channels / Bleacher Team Stream…)
On Demand Personalized Content on Big Screen

KPCB

通信的变化

Re-Imagination of Communication…

THEN…

Dedicated Devices / Limited Function & Range / Intrusive

NOW…

(Voxer…)

Push-To-Talk / Voice Message / Picture / Text / Location / Group Chat

KPCB

绘画的变化

Re-Imagination of Drawing…

THEN…

Dedicated Canvas / Paint Supplies / Studios / Limited Distribution

NOW…

(Paper by Fiftythree…)

Reusable Canvas (Screen) / Creating Art Anywhere Anytime / Digitally Enhanced Creation Tools / Instant Sharing

KPCB

摄影的变化

Re-Imagination of Photography…

THEN…

Dedicated Camera / Manually
Transfer Digital Files / Develop Films

NOW…

(Instagr.am / Camera+ / Hipstamatic…)
Always With You Camera (Smartphone) /
Instant Digital Effects / Share / Sync / Discover

KPCB

地图的变化

Re-Imagination of Navigation + Live Traffic Info…

THEN…

Physical Copies of Map in Car /
TV, Radio Reporting of Traffic Info

NOW…

(Waze)
User-Generated Digital Map /
Live Crowd-Sourced Traffic Data

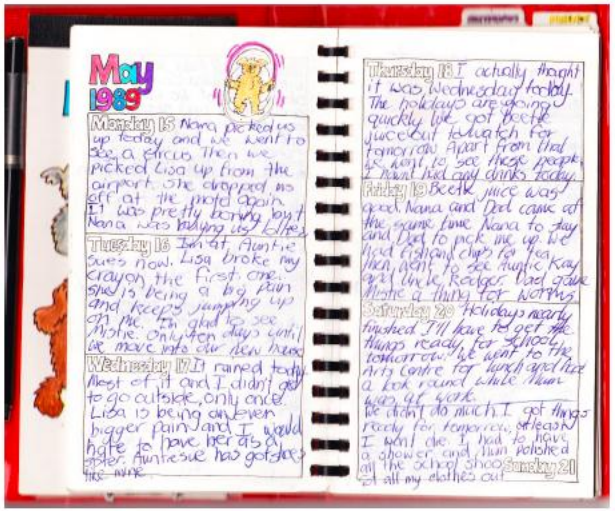
KPCB

日记的变化

Re-Imagination of Diaries…

THEN…

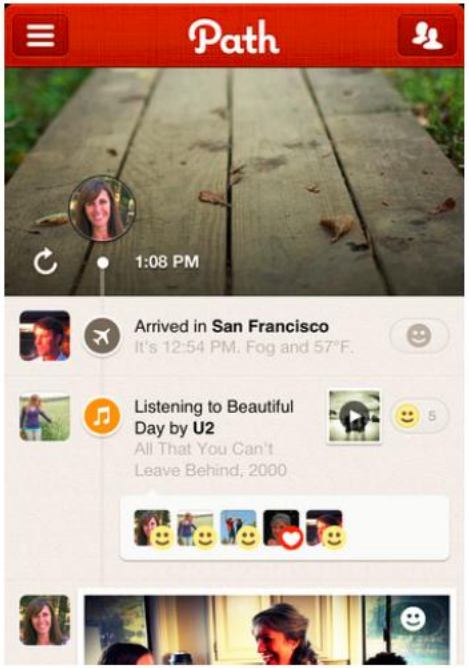
Hand-Written / Drawn



NOW…

(Path)

One-Tap to Add Entry / Multimedia / Location-Aware / Share / Search



KPCB

签名的变化

Re-Imagination of Signatures…

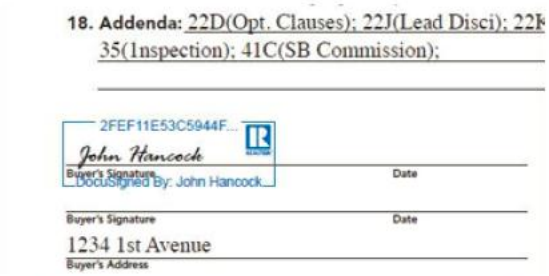
THEN…

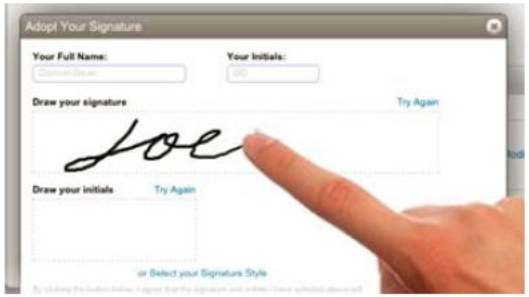
Scan / Fax / Mail to Return Signature Page

NOW…

(DocuSign)

Electronic Documents / Secure Audit Trail / Instant E-Signature





KPCB

学习的变化

Re-Imagination of Learning…

THEN…

NOW…

From learning by listening to learning by doing… Education and learning will become as much fun as videogames. And we call it 'full body learning.'

- Bing Gordon
Partner, KPCB

KPCB

信息量的变化

Re-Imagination of Data…

THEN…

Store Everything Because We Can Do It Inexpensively

SOON…

Data Obesity / Data Quality Issues
How To Find a Needle in a Haystack?

KPCB

Source: Mike Abbott, KPCB.

Outline

- 移动计算时代的来临
  - 主要摘自Mary Meeker 的Internet Trend报告
- **移动计算的技术特点**
- NLP相关的应用分析
  - SIRI
  - Word Lens
  - Google Now
- Compact Data Structures for Mobiles
- 结束语

移动设备的优点

- 移动设备配有各种传感器，就像人的眼睛，嘴巴，耳朵等。这为实现人工智能 **(AI)** 提供了前所未有的机遇
  - 地点定位 **(GPS)**
  - 摄像头，麦克风，耳机
  - 温度传感器，加速传感器
  - 触摸屏
- 随身携带性
  - 成为你生活的一部分
  - 知道你的
    - 地点 **(Location)**
    - 网络访问记录 **(History)**
    - 个人爱好 **(Interests)**
    - 社会关系网络 **(Social Networks)**

移动设备的局限性

- 有限的计算资源
  - CPU, Memory
- 输出输入不方便
  - 键盘输入不方便
  - 显示屏幕比较小
- 电池持续时间比较短
- 3G 网络还没有普及，带宽比较窄

Outline

- 移动计算时代的来临
  - 主要摘自Mary Meeker 的Internet Trend报告
- 移动计算的技术特点
- **NLP相关应用分析**
  - SIRI
  - Word Lens
  - Google Now
- Compact Data Structures for Mobiles
- 结束语

SIRI: 智能个人助理

- 苹果的智能个人助理
- 播放**SIRI**的视频

  http://www.apple.com/ios/siri/

SIRI: 语音控制命令


9:41 AM
“Meet with Jamie for coffee at two today”
Your meeting with Jamie is scheduled for 2 pm today.
12
Wednesday, September 2012
Coffee
2 PM
Cancel
Confirm
9:41 AM
“Make a reservation at seven hills tonight”
OK, Seven Hills has some tables for two around 7 pm.
SEVEN HILLS RESTAURANT
$$$, Closed
Italian
224 reviews
Tue - Thu 5:30 PM - 9:30 PM
Fri - Sat 5:30 PM - 10:00 PM
Sun 5:30 PM - 9:30 PM
Make Reservation
(415) 775-1550
9:41 AM
“Set the timer for 34 minutes”
OK, 34 minutes and counting...
33:57
36 35 34 33 32

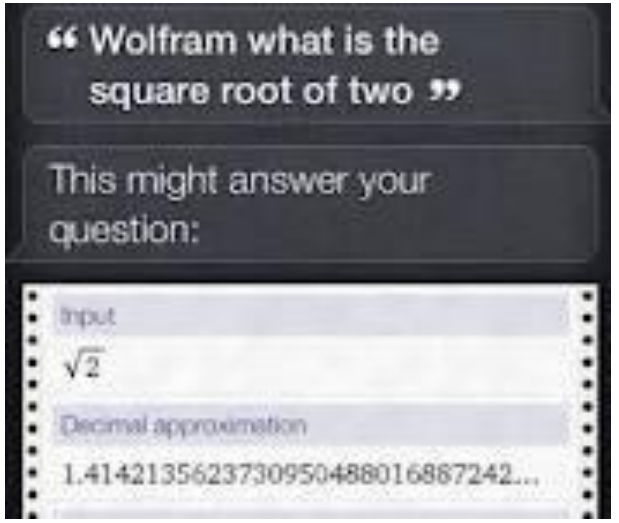
SIRI: 信息访问


9:41 AM
“What’s the weather like today”
The weather’s looking good today… up to 81° and partly sunny:
66°
H: 81°
L: 55°
12:00 PM 75°
1:00 PM 75°
2:00 PM 79°
3:00 PM 79°
9:41 AM
AAPL - Apple Inc. is at $676.81 today, up $13.59, or 2.05%.
APPLE INC.
AAPL 676.81
▲ 13.59 (2.05%)
Open 679.99
High 680.87
Low 673.54
Vol 10.74M
P/E 15.90
Mkt Cap 634.1B
52w High 680.87
52w Low 354.24
Avg Vol 14.08M
Yield 1.60
9:41 AM
I found quite a few movies that are playing around here.
NOW PLAYING
Cupertino
THE CAMPAIGN R
2:10 4:40 7:10 9:40 79%
RUBY SPARKS R
1:30 4:00 7:00 10:00 80%
PARANORMAN PG
11:40 2:10 4:40 7:10 9:40 79%
THE ODD LIFE OF TIMO… PG
1:45 4:35 7:25 10:15 79%
2 DAYS IN NEW YORK R
1:15 4:05 6:45 10:40 79%

SIRI: 问题回答


9:41 AM
“ How many miles are in a marathon ”
This might answer your question:
Input interpretation
convert
1 length of a marathon race to miles
Result
26.22 miles
Additional conversions
42.19 km (kilometers)
42195 meters
4.22 × 10^6 cm (centimeters)
22.78 nmi (nautical miles)
Comparison as length
= length of a marathon race


Bell 3G
11:59 AM
What can I help you with?
“ What does cogent mean ”
I found this for you:
Input interpretation
cogent (English word)
Definition
adjective powerfully persuasive
American pronunciation
k'ohjuhnt (IPA: k'oʊdʒənt)
Hyphenation
co-gent (6 letters | 2 syllables)


“ Wolfram what is the square root of two ”
This might answer your question:
Input
√2
Decimal approximation
1.41421356237309504880168872420...

技术上真的成熟了吗？


AT&T
10:35 AM
“ Call me an ambulance ”
From now on, I’ll call you ‘An Ambulance’. OK?
Cancel
Yes


Sprint 3G
12:34 PM
Reviews from yelp
“ I want to kidnap a child ”
I found 20 kid-friendly restaurants… 14 of them are fairly close to you:
Route 6 North
Holy Smoke
20 reviews
1.1 miles
Rte 6
Arturo’s Trattoria
7 reviews
2.3 miles
Rte 100

智能个人助理：Big Picture

- 用户输入
  - 语音
  - 键盘

- 与机器闲聊
- 访问手机功能

- **访问网络信息**

智能网络信息访问

- 系统自动分析用户的意图，然后做出不同的反应

	用户输入	执行动作
问题回答	让子弹飞的导演是谁	回答“姜文”
	周杰伦的生日是哪天	回答“1979年1月18日”
	2012年奥运会在哪里举行	回答“伦敦”
网络搜索	北京大学	提供关于“北京大学”搜索结果
	百度一下故宫	提供关于“故宫”搜索结果
	生命的意义是什么	提供关于“生命的意义是什么”搜索结果
网络浏览	打开新浪财经的网站	浏览finance.sina.com.cn
	www.sohu.com	浏览www.sohu.com

传统搜索和SIRI

- **传统搜索 (Search)**
  - **优点**：大部分情况下works (i.e. 提供相关链接)
  - **缺点**：
    - 不能直接提供答案 (比如直接回答问题: *现在walmart还开门吗？*)
    - 不能直接在搜索结果上做Action (比如搜索餐馆然后直接订餐)
    - 不能利用自然语言输入做更精准的回答

- **SIRI (智能个人助理)**
  - **优点:** 能直接提供答案和直接做动作 (如订餐，导航等)，能自然语言输入
  - **缺点:** 不可靠 (30%成功，70%失败)

相关的NLP技术

- 语音识别
- 机器学习
  - classification
- 信息提取
  - “帮我买一张从西安到北京明天晚上东航的飞机票”
- 命名实体识别
- 语义分析 (semantic parsing)
  - 对用户的输入进行分析
- 对话管理 (dialog management)
  - 明天北京天气怎么样？
  - 上海呢？
- 问题回答 (question answering)

Hard-core NLP problems!

Factoid Question Answering

- **答案是非常明确并简短的问题**
  - 比如答案是**时间，地点，或人名**。问题的例子有：
    - 毛泽东是哪天出生的？刘德华的生日是哪天？
    - 湖南省的省会在哪里？ 中南工业大学在哪个城市？长沙是哪个省的省会？希腊的首都在哪里？毛泽东出生在哪个省？巴尔的摩在美国哪个州？
    - 让子弹飞的导演是谁？谢霆锋的前妻是谁？湖南省岳阳市市委书记是谁?

为什么精准问题回答有意义?

- 解决用户在移动环境下急需知道答案的需求
  - 比如在开车或旅游时，没法慢慢浏览
- 显示更多的相关内容 (见下一页例子)
  - 比如当用户问李亚鹏的老婆是谁时，系统不但可以回答“王菲”，更进一步可以显示王菲的近照，最近的歌曲和八卦等相关内容
- 回答需要推理的复杂问题的必须步骤
  - 后天是中秋节吗？(系统必须先知道今年中秋节是哪天)
  - 毛泽东的故乡今天天气怎么样？(系统必须先知道毛泽东的故乡在哪里)
  - 显示几张王菲的老公的照片(系统必须先知道王菲的老公是谁)
- Empower interactive的人机交互方式
  - 问：李亚鹏的老婆是谁？
  - 答：王菲
  - 问：她最近有什么歌曲啊？
  - 答：因为爱你
  - 问：放给我听听
  - 答：播放歌曲因为爱你

更相关的搜索结果


10:50
Bai du 百度
网页 地图 贴吧 应用 更多
李亚鹏老婆是谁
百度一下
李亚鹏的老婆是谁_百度知道
李亚鹏的老婆 ...王菲啊~ ...就是就是 王菲啊 ...王菲...这你也不知道？王菲嘛！ ...王菲...王菲...王菲啊都有个小孩了，你还不知道啊！ ...王菲...
李亚鹏老婆是谁呀4个回答
李亚鹏版《笑傲江湖》里的岳不群老婆宁中则是谁演的?好..2个回答
网络春晚《法制故事》中扮演李亚鹏妻子的是谁?3个回答
更多知道相关问题»
zhidao.baidu.com
李亚鹏的老婆是谁_在线视频观看_土豆网视频 李亚鹏的老婆是谁
原创 李亚鹏的老婆是谁 李亚鹏的老婆是谁转贴到土豆推 更多加豆单私藏下载... 老婆！这个男的是谁？ 01:38 老婆，这个男的是谁？ 老婆，这个男的是谁？...
www.tudou.com 33k - 优化阅读
李亚鹏的老婆是谁-娱乐动态-东南电影网
家庭状况:父亲王佑林、母亲夏桂影(去世)、哥哥王弋、


10:53
Bai du 百度
网页 地图 贴吧 应用 更多
王菲
百度一下
王菲_百度百科
王菲，中国著名女歌手、影视演员，是1990年代初期至今华语乐坛最出色的女歌手之一，被公认为乐坛天后。她是极少数能同时被华语歌坛同行一致赞美、甚...
baike.baidu.com
王菲_百度MP3
《王菲2cd精选》 发行时间: 2009-07-14 曲目数：28 语种: 中文 唱片公司: 环球唱片 1 我愿意 2 容易受伤的女人 3 但愿人长久 4 执迷不悔 5 棋子...
mp3.baidu.com 127k - 优化阅读
王菲- 维基百科，自由的百科全书
王菲推出的多张唱片均有百万销量佳绩。截至2000年3月，20张专辑合计台湾、日本与香港有据可考的销量达970万张，获金氏世界纪录大全认证为华语乐坛粤语专辑销量最高的...
zh.wikipedia.org 127k - 电脑版
王菲的最新相关信息
窦靖童王菲刘嘉玲风格迥异 姐们儿个个都是范儿

Re-imagination of NLP for **Search**

- **输入**

键盘	→	语音，手写，OCR
关键词	→	自然语言
Pattern Match	→	Semantic Parsing

- **输出**

大屏幕	→	小屏幕
文字	→	声音 (TTS)
Links	→	Answers
IR	→	QA

IR → QA


10:50
Bai du 百度
网页
地图
贴吧
应用
更多
李亚鹏老婆是谁
百度一下
李亚鹏的老婆是谁_百度知道
李亚鹏的老婆 ...王菲啊~ ...就是就是 王菲啊 ...王菲...这你也不知道？王菲嘛！ ...王菲...王菲...王菲啊都有个小孩了，你还不知道啊！ ...王菲...
李亚鹏老婆是谁呀4个回答
李亚鹏版《笑傲江湖》里的岳不群老婆宁中则是谁演的?好..2个回答
网络春晚《法制故事》中扮演李亚鹏妻子的是谁?3个回答
更多知道相关问题»
zhidao.baidu.com
李亚鹏的老婆是谁_在线视频观看_土豆网视频
李亚鹏的老婆是谁
原创 李亚鹏的老婆是谁 李亚鹏的老婆是谁转贴到土豆推 更多加豆单私藏下载... 老婆！这个男的是谁？ 01:38 老婆，这个男的是谁？ 老婆，这个男的是谁？...
www.tudou.com 33k - 优化阅读
李亚鹏的老婆是谁-娱乐动态-东南电影网
家庭状况:父亲王佑林、母亲夏桂影(去世)、哥哥王弋、


7:48
李亚鹏老婆是谁
李亚鹏老婆是谁
王菲
点击说话

Re-imagination of **NLP Models**

- **Context-aware Model**

  Context = time, location, person, interests …

- **Source Model**

  P(current word | history) → P(current word | history, **context**)

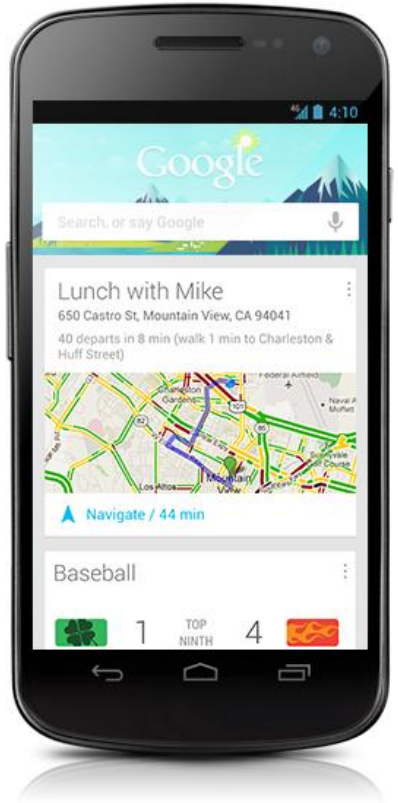
- **Channel Model**

  P(output | input) → P(output | input, **context**)

Google Now

- **Google的最新智能应用**
  - **Predictive UI**
  - **Factoid Question Answering**
- **播放Google Now的视频**

http://www.google.com/landing/now/



Word Lens

- 基于**OCR**的机器翻译应用
- 播放**Word Lens**的宣传视频

http://questvisual.com/us/

Joint decoding for OCR and translation?

Offline OCR and Translation?

Outline

- 移动计算时代的来临
  - 主要摘自Mary Meeker 的Internet Trend报告
- 移动计算的技术特点
- NLP相关的应用分析
  - SIRI
  - Word Lens
  - Google Now
- **Compact Data Structures for Mobiles**
- 结束语

Summary of Results

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0
Context encoding	2.5	1	1.285	3.95	39.5	O(2logV)	0
Implicit trie	2.5	1	1.285	3.95	39.5	O(2logV)	0
Bloomier	2	1	1.23	3.69	36.9	O(1)	1/(2^16)
MPH	2	1	1.108	3.325	33.25	O(1)	1/(2^16)

- Assumptions:
  - Each word ID consumes 2.5 bytes (i.e., V=2^20)
  - Probability value is quantized to one byte
  - Need to store 10M bigrams (i.e., N=10M)

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0

- Java String object
  - each java object has 8 bytes overhead
  - the size of an object has to be a multiple of 8
  - an empty string in java will take 40 bytes!
- HashMap<String, Integer>:
  - Key is a String, which needs 72 bytes.
    - assuming 17 chars (each char consumes 2 bytes)
    - 72 bytes = 8 + 4 + 4 + 4 + 4 + (8 + 4 + 17*2 + 2)
  - Value is an Integer
    - 16 bytes = (8 + 1 + 7)
  - Overhead ratio is 1.6
    - assuming a load factor 0.625 (so overhead ratio is 1/0.625=1.6)

```
public class String {
    char value[];
    int length;
    int offset;
    int hash;
}
```

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0

- HashMap uses an array to store the actual keys and values
  - The naive implementation store the keys and values as objects, which have high memory overhead
- To save memory:
  - Do not store the keys and values as java objects
  - Instead, store the keys and values in a bit array
    - So, a key (i.e., a bigram) will consume 5 bytes as each word Id consumes 2.5 bytes.

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0

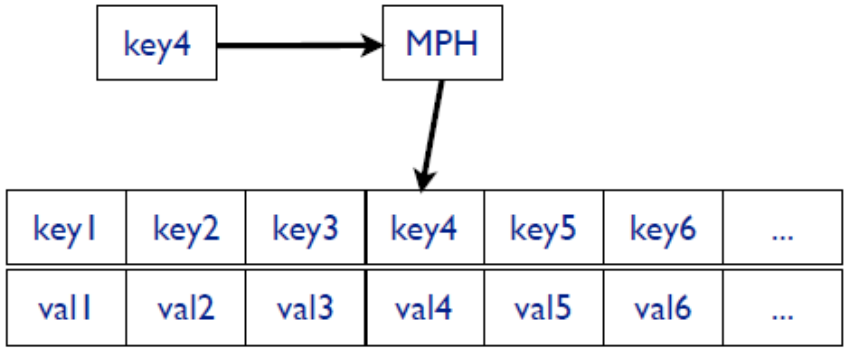
- To deal with hash collisions, HashMap needs to use an array that is much bigger than needed to store the actual keys and values
  - The benefit of this is constant lookup time.
- To save memory, we can:
  - Store the list of bigrams as a sorted array
  - Use binary search to look up a key
  - However, the look-up time will be O(logN) ,where N is the total number of bigrams

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0

- Use minimum perfect hash (MPH):
  - Store key and values in arrays
  - Use a MPH function to index the keys
- Lookup procedure:
  - First compute MPH of key
  - Then match **key** in the array
- MPH function consumes 2.6 extra bits for each key. This explains the overhead ratio 1.054

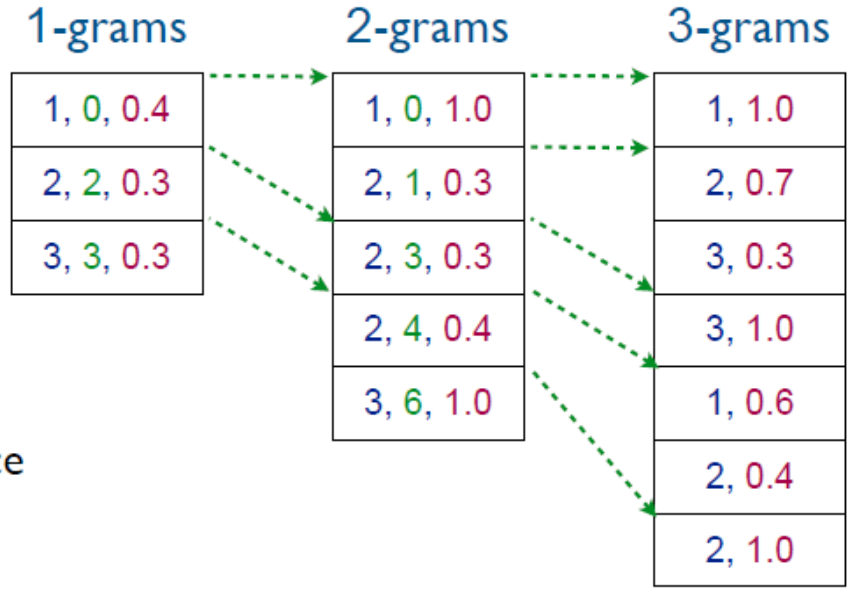
perfect: no collision

minimum: MPH outputs numbers in [1, N]



Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0
Implicit trie	2.5	1	1.285	3.95	39.5	O(2logV)	0

- Store the keys and values in arrays of Trie nodes
- Each Trie node includes:
  - word id
  - offset in the next layer
  - the value (if any)
- The overhead ratio (1.285) considers space for storing offsets in the unigram table



Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0
Implicit trie	2.5	1	1.285	3.95	39.5	O(2logV)	0
Context encoding	2.5	1	1.285	3.95	39.5	O(2logV)	0

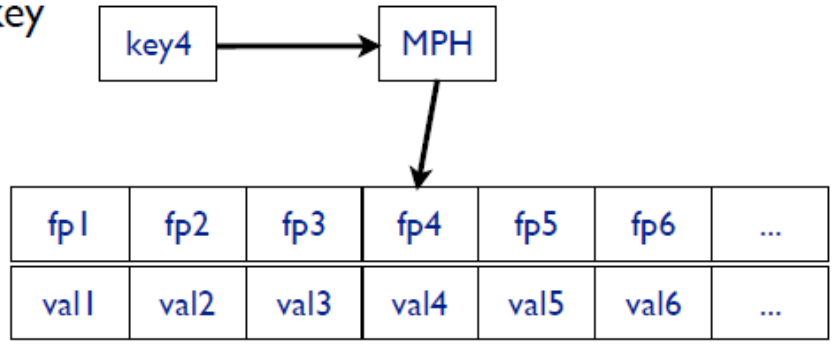
See (Pauls and Klein 2011) for details.

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0
Context encoding	2.5	1	1.285	3.95	39.5	O(2logV)	0
Implicit trie	2.5	1	1.285	3.95	39.5	O(2logV)	0
Bloomier	2	1	1.23	3.69	36.9	O(1)	1/(2^16)

- In comparison with MPH, Bloomier is perfect, but not minimum
- The overhead is 1.23.

Method	Key	Val	Overhead ratio	Bytes per bigram	Total (MB)	Time	P(Error)
Map (naive java)	72	16	1.6	141	1410	O(1)	0
Map	5	1	1.6	9.6	96	O(1)	0
Sorted array	5	1	1	6	60	O(log(N))	0
MPH	5	1	1.054	6.325	63.25	O(1)	0
Context encoding	2.5	1	1.285	3.95	39.5	O(2logV)	0
Implicit trie	2.5	1	1.285	3.95	39.5	O(2logV)	0
Bloomier	2	1	1.23	3.69	36.9	O(1)	1/(2^16)
MPH	2	1	1.108	3.325	33.25	O(1)	1/(2^16)

- Store a finger print of the key, not the actual key
- Lookup procedure:
  - First compute MPH of key
  - Then match **fingerprint** (fp) in the array
- MPH function consumes 2.6 extra bits for each key, explaining the overhead ratio 1.108



Compression of Arrays

- All the data structures presented require to use **array(s)** to store keys (or their fingerprints) and values
- So, compression of arrays will save memory
  - Note that the keys in the approximate maps (e.g., based on bloomier filter and MPH) can not be compressed as they are random
- Variable-length coding
  - Huffman code
  - Golomb codes
  - Boldi and Vigna (2005)
  - Elias-Fano codes for monotone sequences

Element Access of Compressed Array

- A common operation is to access the i-th element in the array (e.g. during the binary search)
  - Need to be fast
- Naive methods require to linear scan of the array in order to access the i-th element
- Common practices use (e.g. in Pauls and Klein, 2011) block encoding
  - still slow and space non-optimal
- New method:
  - use a bit array together with succinct data structures

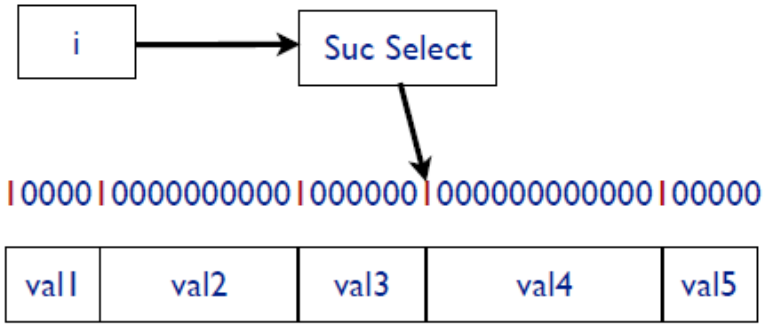
Constant Access with Succinct Data Structures

- Use whatever variable-length coding methods to encode elements
- Store the elements in a bit array
- Have an **additional** binary array to indicate the element boundaries
- Use succinct data structures to quickly identify the boundary (i.e. using select operation)
  - This requires about 13% additional space.

Two fundamental operations for a succinct data structure:

- select: find the position of the i-th one in the binary array

- rank: find how many ones before a given position in the binary array



Outline

- 移动计算时代的来临
  - 主要摘自Mary Meeker 的Internet Trend报告
- 移动计算的技术特点
- NLP相关的应用分析
  - SIRI
  - Word Lens
  - Google Now
- Compact Data Structures for Mobiles
- 结束语

Conclusions

- 移动计算时代已来临
- 移动计算下NLP研究面临新的机遇和挑战
- 移动智能应用将会越来越流行
  - SIRI, Google Now, Google Glass, Word Lens, etc
- 相关的NLP技术将会有巨大的商业应用
  - Efficient data structures for mobiles
  - Location/time-aware models
  - Semantic parsing
  - Natural language search
  - Question answering
  - Predictive models for personalization
  - Joint Decoding (for multi-modal inputs)

Google Glass

- **Google X的项目**
- **播放Google Glass的视频**

https://plus.**google**.com/+project**glass**

Mobvoi: we are hiring ☺

- 刚刚在上海成立的初创公司
- 获得世界知名VC的风险投资
- 专注于**基于NLP的智能移动搜索**
- 崇尚硅谷公司(如Google)工程师文化
- 寻求NLP Research Engineers

- 如有兴趣，请联系我: zfli@mobvoi.com