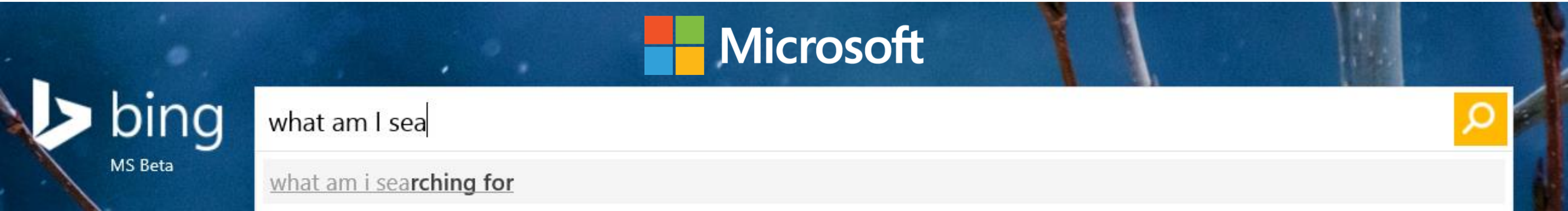
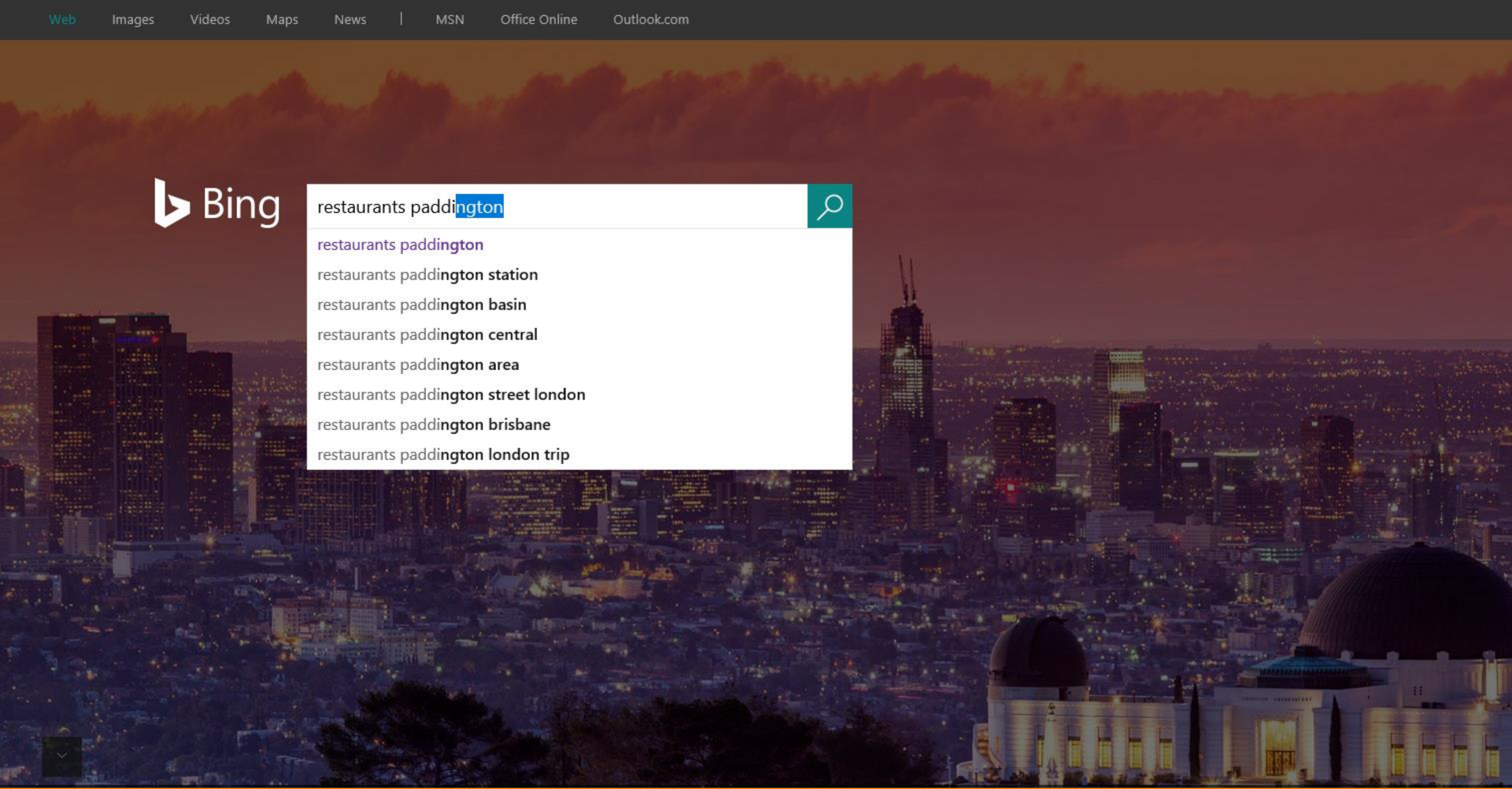
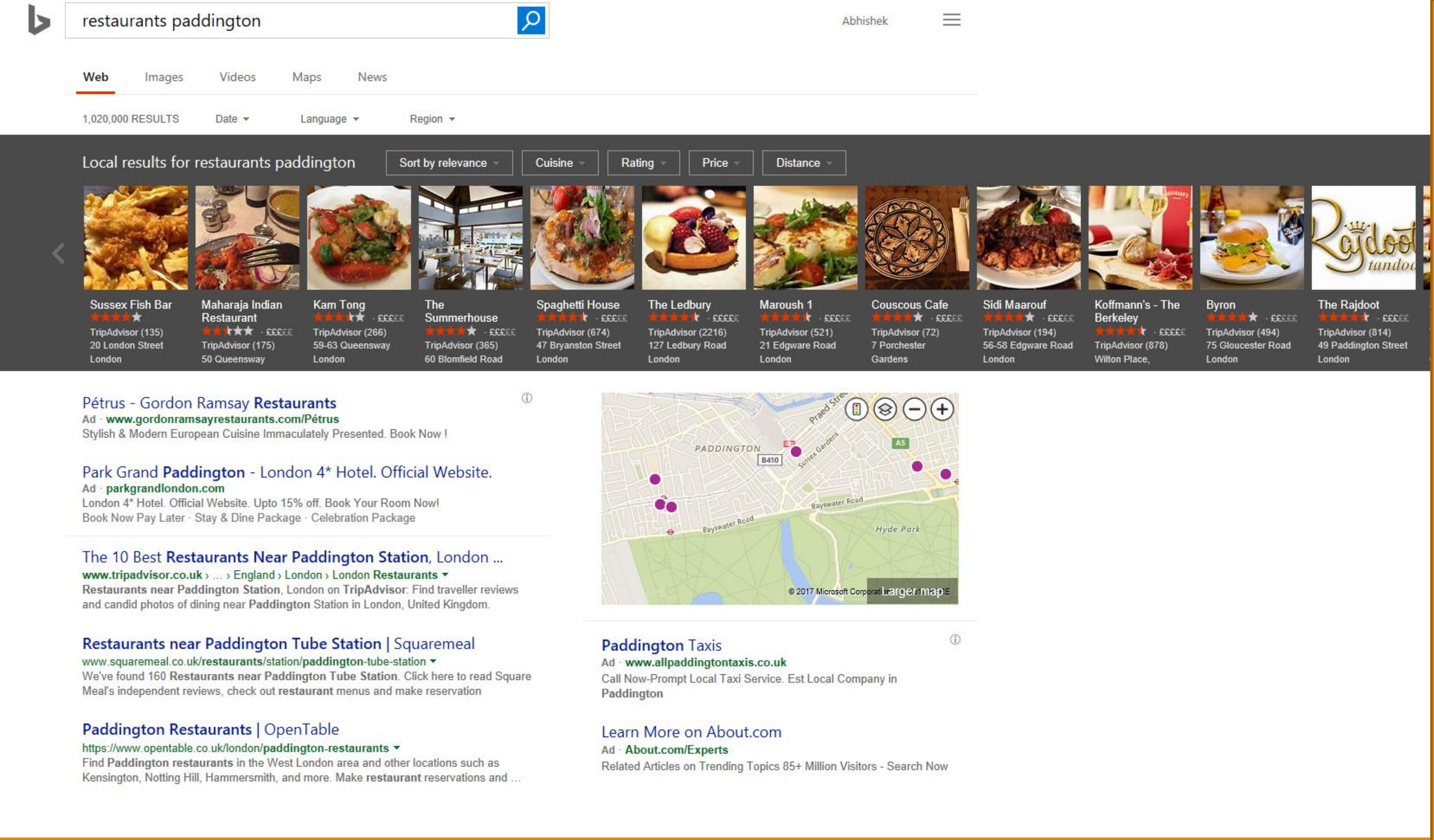
Engineering at Scale

THE CHALLENGES OF PREDICTING QUERIES IN WEB SEARCH ENGINES

Paul Baecke




Web
Images
Videos
Maps
News
MSN
Office Online
Outlook.com
Bing
restaurants paddington
restaurants paddington
restaurants paddington station
restaurants paddington basin
restaurants paddington central
restaurants paddington area
restaurants paddington street london
restaurants paddington brisbane
restaurants paddington london trip


restaurants paddington
Abhishek
Web
Images
Videos
Maps
News
1,020,000 RESULTS
Date
Language
Region
Local results for restaurants paddington
Sort by relevance
Cuisine
Rating
Price
Distance
Sussex Fish Bar
TripAdvisor (135)
20 London Street
London
Maharaja Indian Restaurant
TripAdvisor (175)
50 Queensway
Kam Tong
TripAdvisor (266)
59-63 Queensway
London
The Summerhouse
TripAdvisor (365)
60 Blomfield Road
Spaghetti House
TripAdvisor (674)
47 Bryanston Street
London
The Ledbury
TripAdvisor (2216)
127 Ledbury Road
London
Maroush 1
TripAdvisor (521)
21 Edgware Road
London
Couscous Cafe
TripAdvisor (72)
7 Porchester Gardens
Sidi Maarouf
TripAdvisor (194)
56-58 Edgware Road
London
Koffmann's - The Berkeley
TripAdvisor (878)
Wilton Place,
Byron
TripAdvisor (494)
75 Gloucester Road
London
The Rajdoot
TripAdvisor (814)
49 Paddington Street
London
Pétrus - Gordon Ramsay Restaurants
Ad · www.gordonramsayrestaurants.com/Pétrus
Stylish & Modern European Cuisine Immaculately Presented. Book Now !
Park Grand Paddington - London 4* Hotel. Official Website.
Ad · parkgrandlondon.com
London 4* Hotel. Official Website. Upto 15% off. Book Your Room Now!
Book Now Pay Later · Stay & Dine Package · Celebration Package
The 10 Best Restaurants Near Paddington Station, London ...
www.tripadvisor.co.uk › ... › England › London › London Restaurants
Restaurants near Paddington Station, London on TripAdvisor: Find traveller reviews and candid photos of dining near Paddington Station in London, United Kingdom.
Restaurants near Paddington Tube Station | Squaremeal
www.squaremeal.co.uk/restaurants/station/paddington-tube-station
We've found 160 Restaurants near Paddington Tube Station. Click here to read Square Meal's independent reviews, check out restaurant menus and make reservation
Paddington Restaurants | OpenTable
https://www.opentable.co.uk/london/paddington-restaurants
Find Paddington restaurants in the West London area and other locations such as Kensington, Notting Hill, Hammersmith, and more. Make restaurant reservations and ...
PADDINGTON
B410
A5
Bayswater Road
Hyde Park
© 2017 Microsoft Corporation
Larger map
Paddington Taxis
Ad · www.allpaddingtontaxis.co.uk
Call Now-Prompt Local Taxi Service. Est Local Company in Paddington
Learn More on About.com
Ad · About.com/Experts
Related Articles on Trending Topics 85+ Million Visitors - Search Now

Microsoft

Introduction

How is what we do ‘Extreme Computing’?

What is the product

Complexity online

Complexity offline

Complexity of systems

Some examples

Some numbers (online)

10^12 requests served per year

10^16 bytes of data logge

10^14 ms of CPU time used per

Thousands **Years of CPU Time per year!**

	QPS	Data	CPU
Average QPS	> 100k	> 1 GB	> 5million ms

More numbers (offline)

Some data is refreshed 12 times per day

All data is updated daily

Models updated weekly

Data scientists run 100s of experiments per week

Availability goal is > 99.995% uptime

Latency goal is < 50ms average

Why this matters

At this scale, every engineering decision matters

There is a deep focus on efficient data structures and algorithms

Every ms of CPU time saved, every byte of storage optimized:

Saves money & time

Allows for better experiences to be built

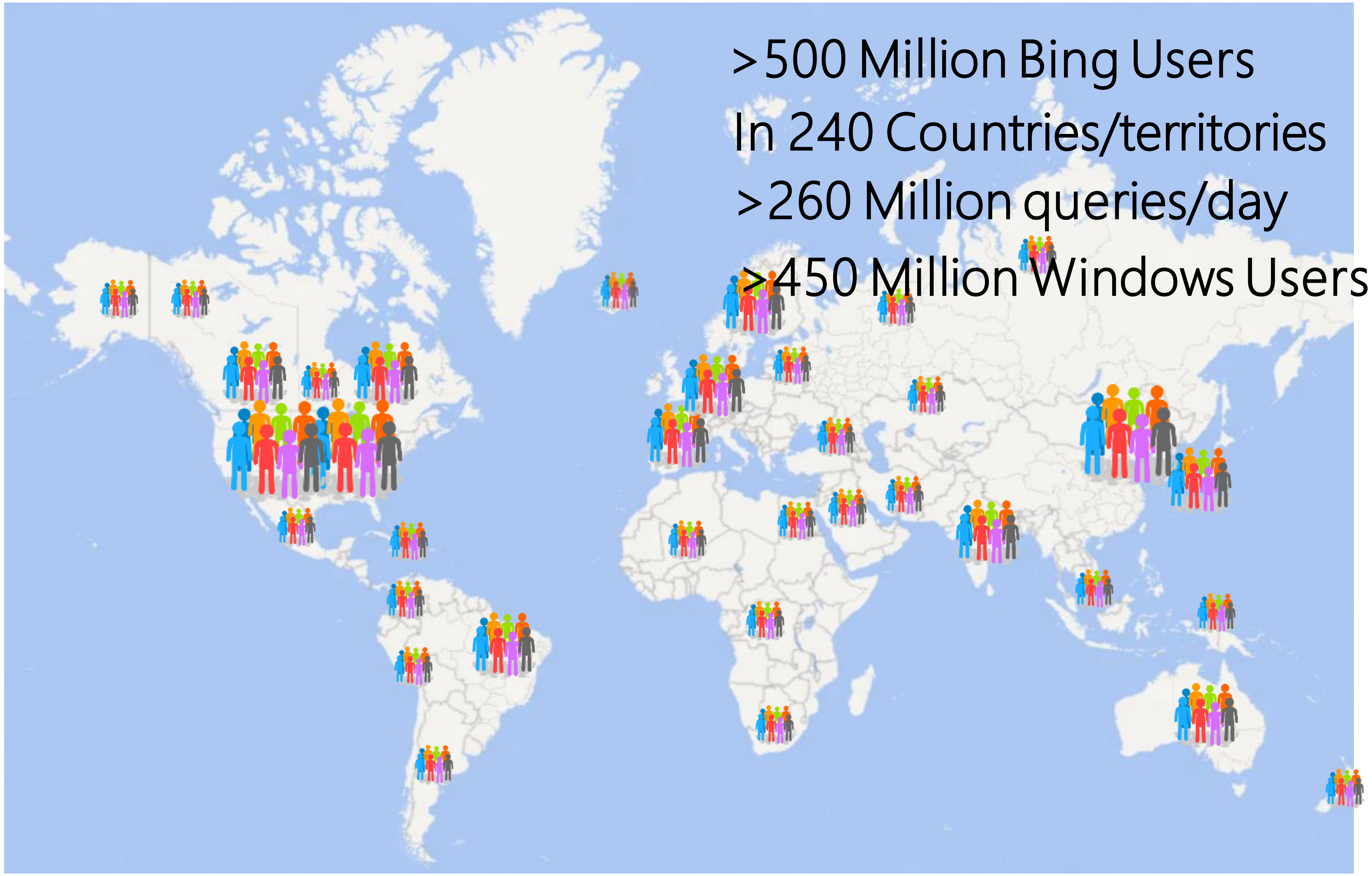
Makes users happier

Keeps our engineers on the cutting edge of research and best practices

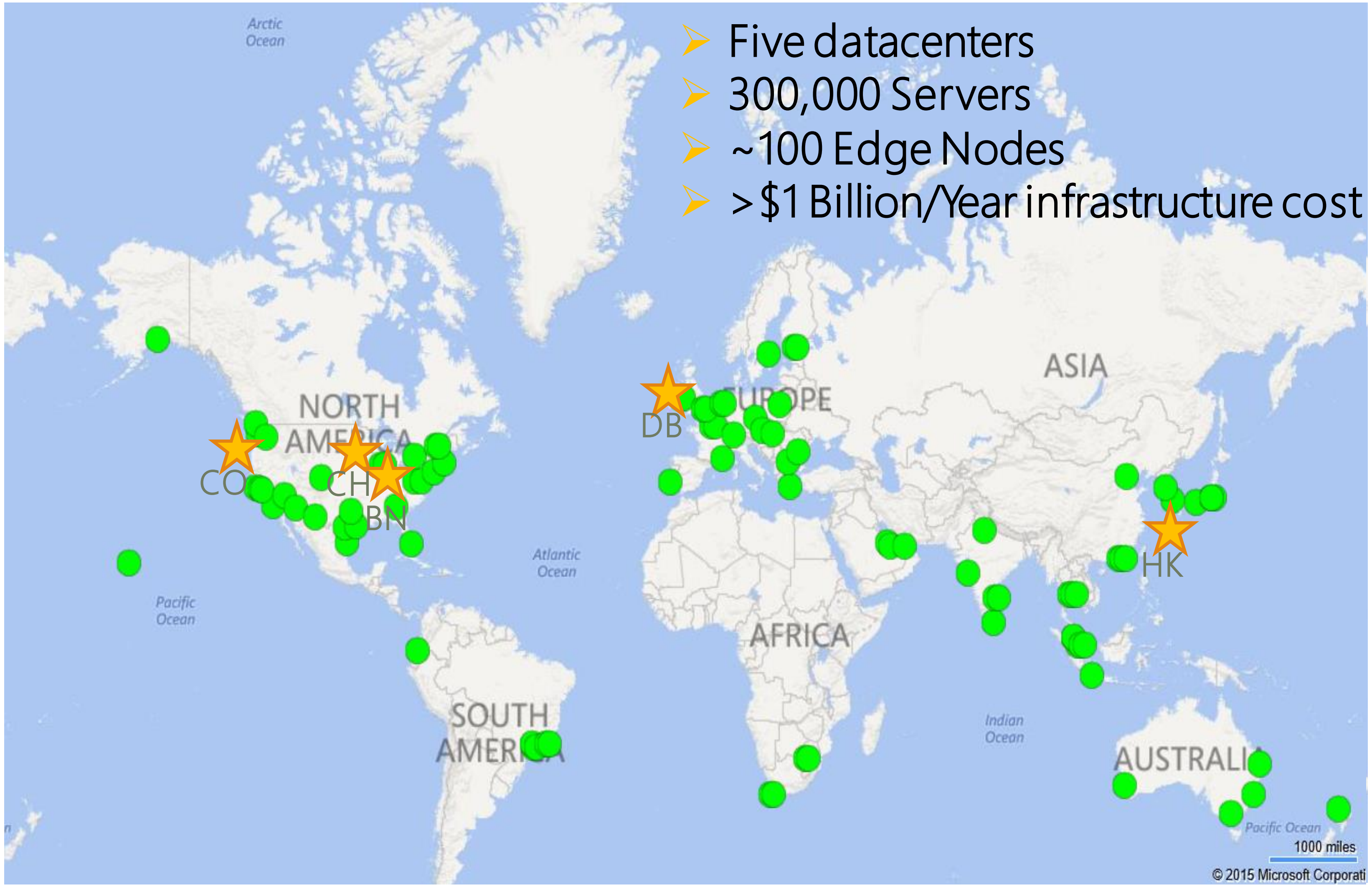
Bing & Autosuggest Infrastructure

WHAT IT TAKES TO SERVE BING & KEEP LIVE SITE HEALTHY

Bing Usage


>500 Million Bing Users
In 240 Countries/territories
>260 Million queries/day
>450 Million Windows Users

Serving 500M users requires massive scale



Bing AutoSuggest in Numbers

150k+	500+	~50ms	1.1B+	~30
Keystrokes/ Second	# AS servers	Server latency	Suggestions (USA)	Supported countries

Outages are Newsworthy

How long will big-name customers like Netflix put up with Amazon cloud outages?

Amazon's cloud went down, again, this time on Christmas Eve, for 12 hours , blacking out 7% of AWS customers. Will continued cloud outages erode confidence in the public cloud?

Facebook down? People across Midwest reporting problems

Published 10:30 AM CDT Jun 18, 2015

Facebook, Instagram briefly go down; Twitter freaks out

Post Nation

Twitter goes down, chaos and productivity ensue (UPDATE: It's back)

Why is Google silent on its outage?

Microsoft Searches for Cause of Bing Outage

FAILURE RELATED TO INTERNAL TESTING AFTER MAPS ROLLOUT

Global Amazon outage keeps some customers from placing orders

A customer service representative tells CNET the company's servers have been down since about 10 a.m. PT and may not be back up for several hours.

Microsoft Bing Outage Brings Down Yahoo Search, Cortana, and Siri

Microsoft's Bing search engine went down on Friday morning due to what seems to be a bad update rolled out by the Redmond-based software giant and which also affected a number of other services, including Yahoo Search, Windows Phone's personal assistant Cortana, and Apple's Siri.

Gmail Down: Google Outage Twice In 2 Days

Posted By: Michelle Jones Posted date: March 18, 2014 03:07:36 PM In: Technology No Comments

Amazon EC2 Outage Shows Risks of Cloud

Amazon's EC2 cloud went dark last week--knocking sites like Foursquare, Reddit, and Quora offline, and affecting hundreds of Amazon cloud customers. The outage is a black eye for the young cloud services industry and gives businesses a reason to think

CLOUD COMPUTING, DOWNTIME, MICROSOFT

Microsoft Says Config. Change Caused Azure Outage

BY YEVGENIY SVERDLIK ON NOVEMBER 20, 2014 ADD YOUR COMMENTS

Microsoft Azure team has published a post-mortem on the widespread outage the cloud went through Wednesday that affected about 20 services in most availability zones around the world.

ion change meant to make Blob storage (Azure's e service for unstructured data) perform better y sent Blob front ends "into an infinite loop,"

Twitter crashes hard, Internet freaks out

By Julianne Pepitone @CNNMoneyTech June 21, 2012: 3:34 PM ET

NEW YORK (CNNMoney) -- Cue the collective Internet freakout! Twitter went down for several hours on Thursday afternoon, depriving users of a place to complain that Twitter was down.

Is Facebook down? A history of outages

When Facebook goes down it's a serious issue: bored office workers are bereft of distractions, children are obliged to talk to their families, media executives cry over lost traffic and nobody gets poked … (no seriously, that still exists).

In addition, Facebook engineers frantically rush to get everything back online – but it wasn't always like that. In 2010 when Facebook's site broke down it could be fixed just by turning it off and on again, literally.

Bing is down: 'The page you want isn't available' (update: fixed)

Google, YouTube Outages Whip Twitter Into Frenzy: 'Is This a Global Emergency?'

Gmail Went Down And Everyone Panicked

The Huffington Post

It's not just you. Gmail went down, and everyone flipped out.

At around 2 p.m. EST, Google's email service became largely inaccessible, as confirmed by the website Down For Everyone Or Just Me. But you probably already know that if you've been checking Twitter:

Google Outage: Internet Traffic Plunges 40%

The web giant is refusing to discuss why all its services from Google Search to Gmail to YouTube stopped working across the world.

Yahoo's search goes dark following Bing outage

CNBC.com staff | @CNBC

Friday, 2 Jan 2015 | 3:22 PM ET

CNBC

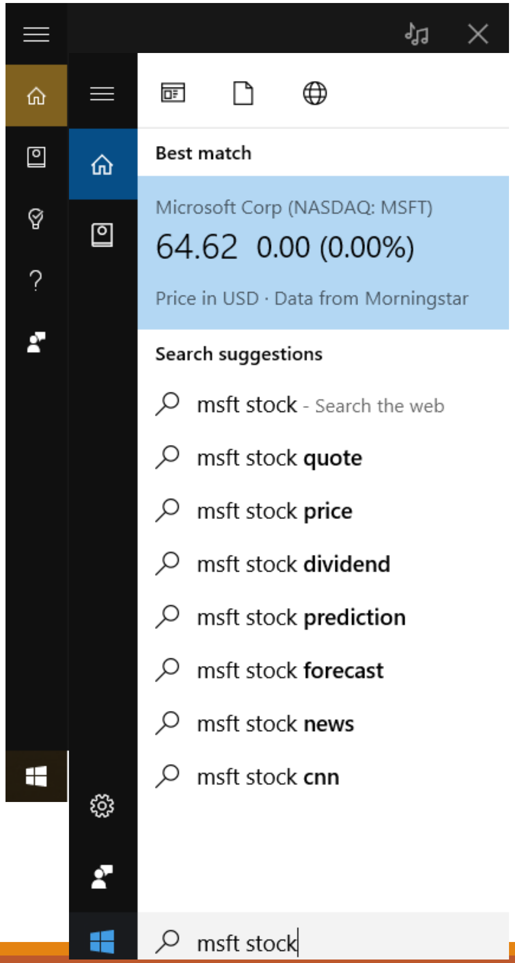
WORLD WIDE WEB

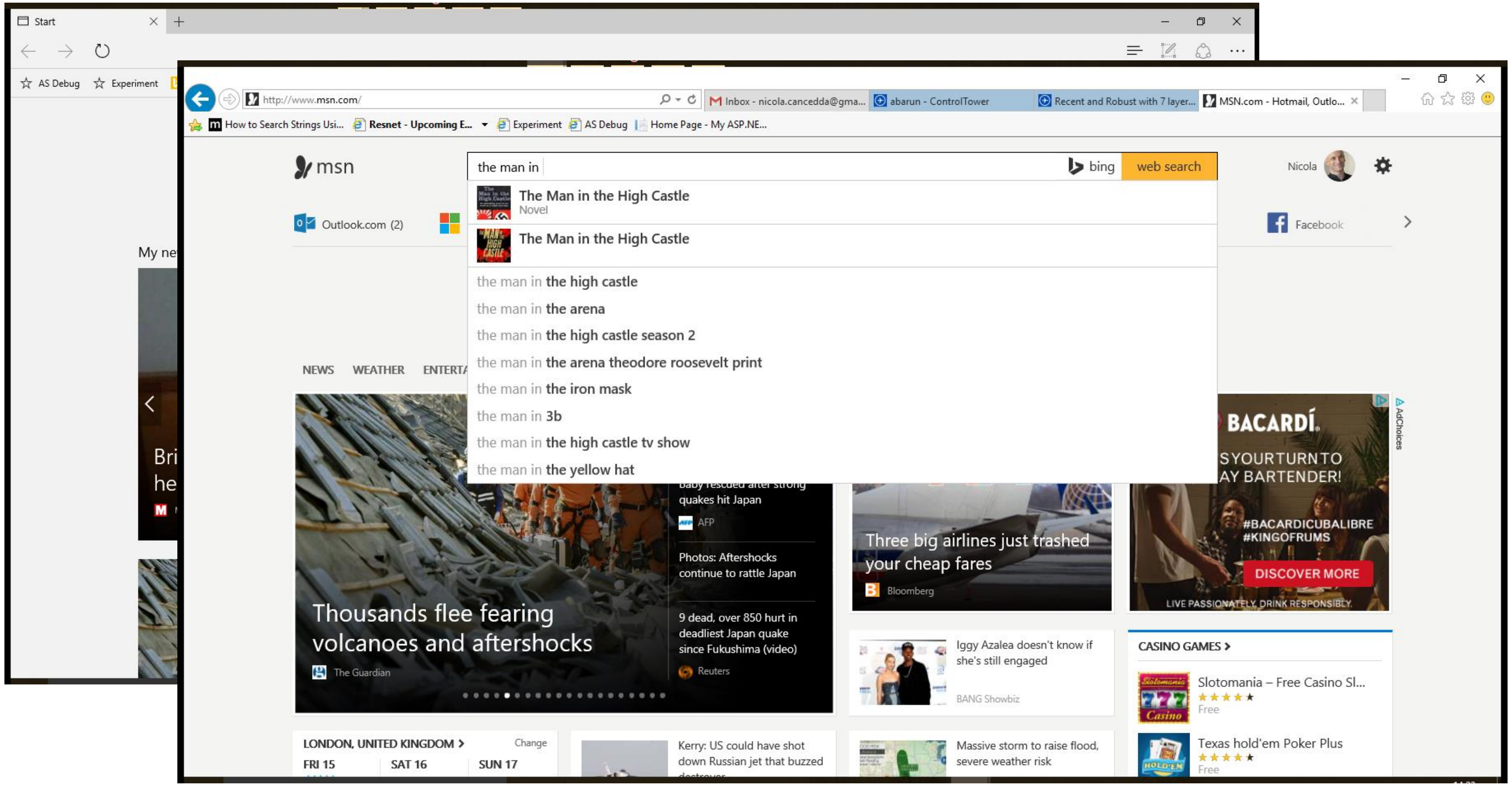
Gmail Down for 12 Hours, Google Says 'Sorry About That'

Posted September 24, 2013

Twitter apologizes for worldwide outage

Powered by Bing AutoSuggest





AutoSuggest

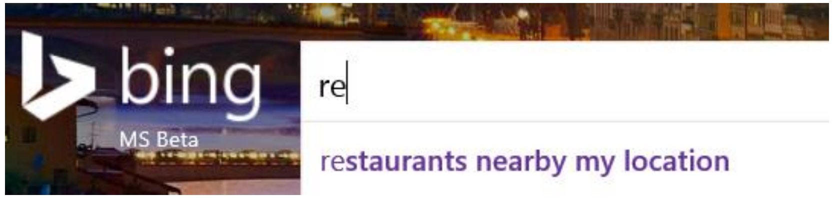
Predicting your query before you type it

What should we suggest?

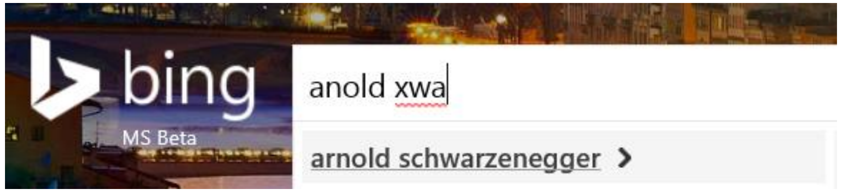
Alice	Bob	Charlie
User previous queries: - movie streaming - imdb ranking	User location: 	Day of week: Sunday Time of day: 12h30 Device: mobile
netflix	**new york times**	**nearby restaurants**

Why is it useful?

1. Reduce query formulation effort

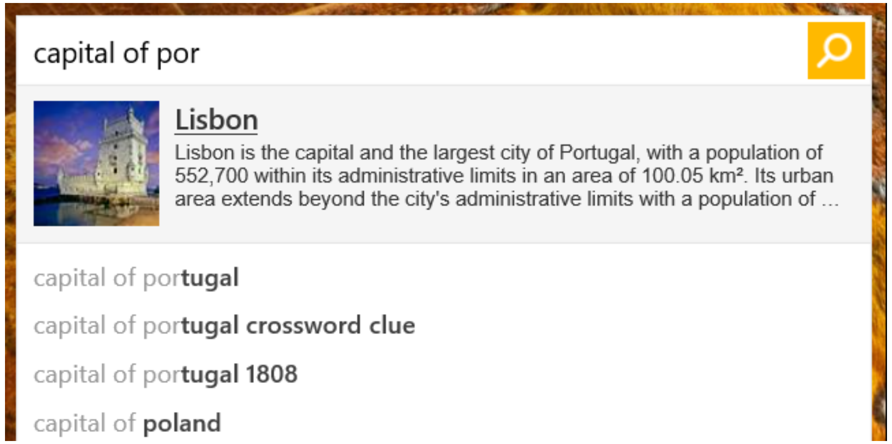


2. Prevent misspellings

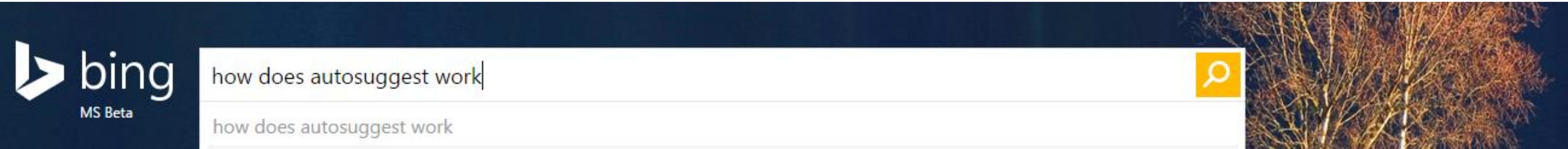


3. Provide more relevant search results
   - Search Result Pages (SERPs) tend to be optimized for popular queries

4. Provide direct answers

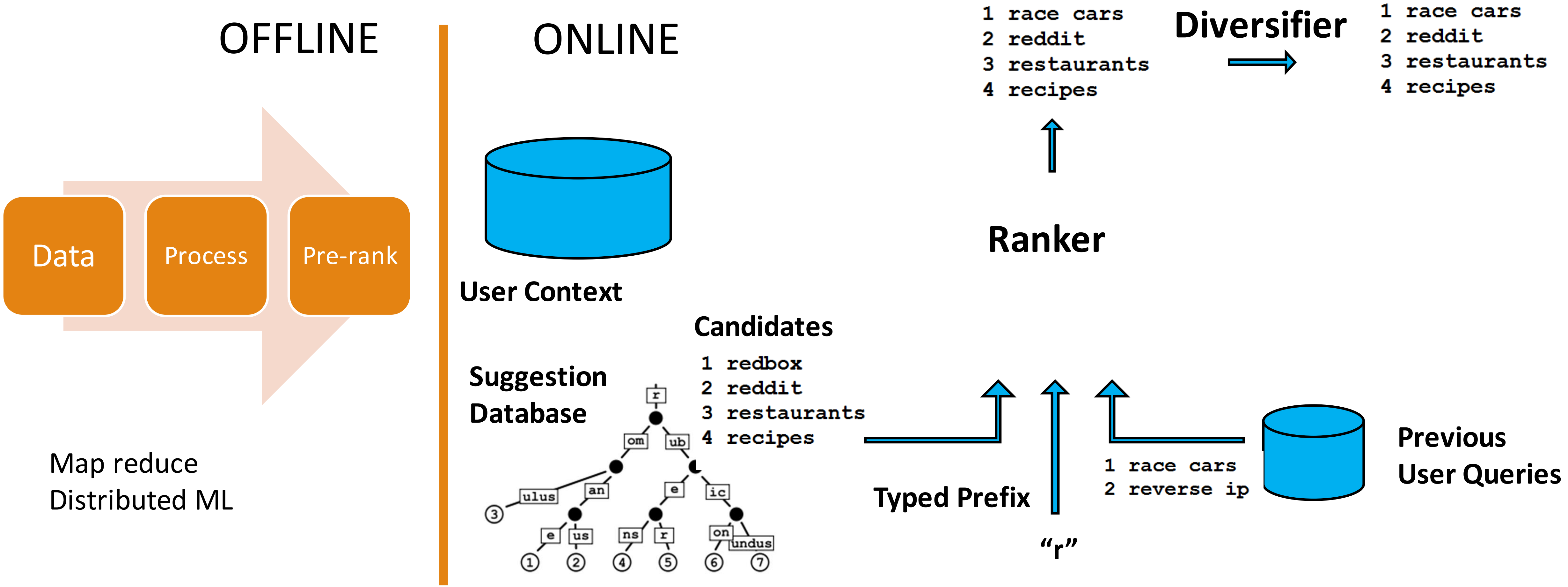


AUTOSUGGEST - BEHIND THE SCENES



Overall architecture

(very simplified)

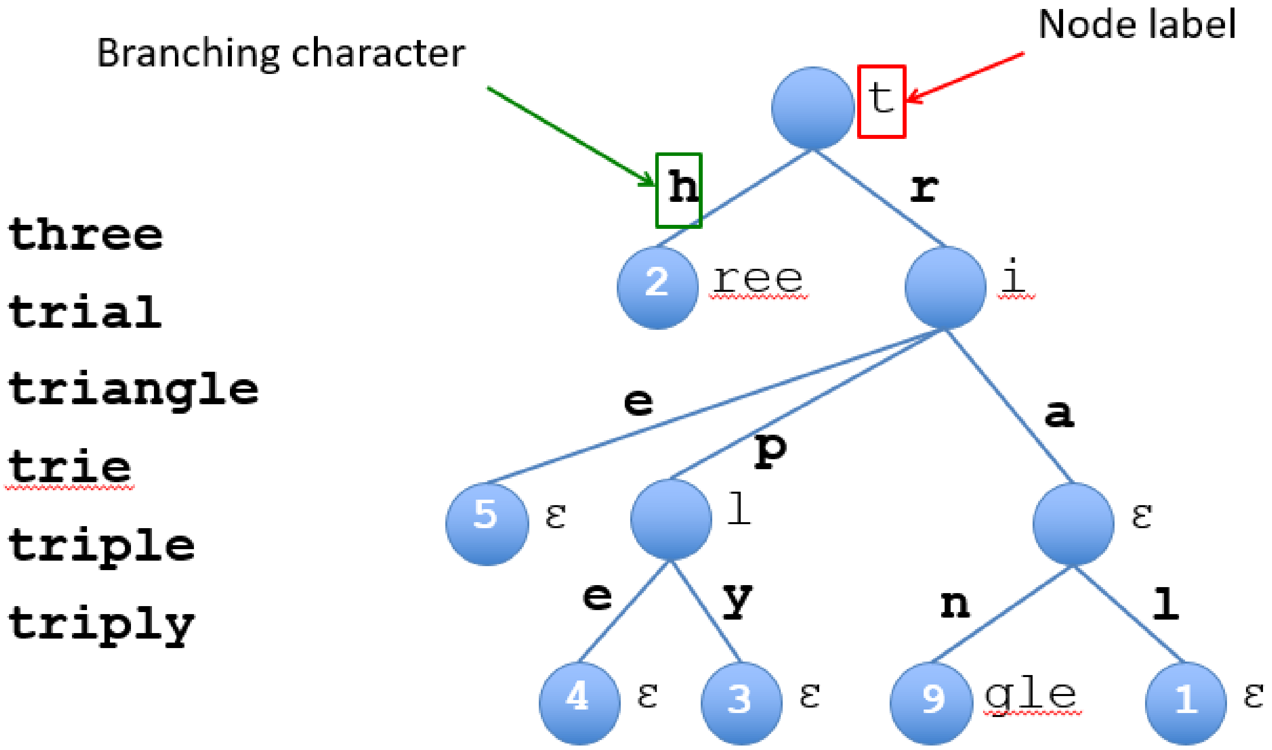


Pre-ranking

- Each query **q** is associated with (an estimate of) the probability **P(q)** that a user will use it.
- The estimate is based on:
  - how many times **q** was typed in the past
  - how recently
  - … and other factors

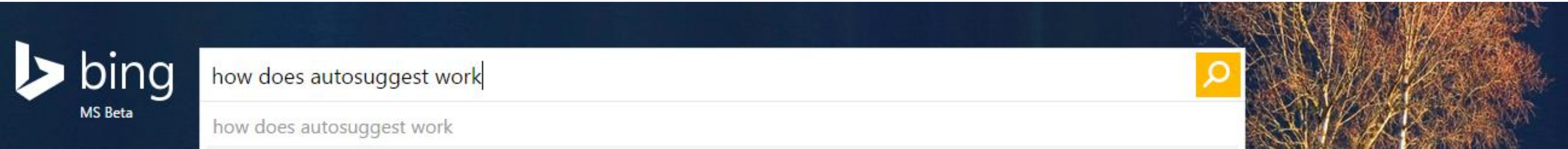
Suggestion Database - Candidate Generation

(Scored) compacted tries



- Compressed data structure
  - 1.1B suggestions and their metadata fit in < 30 GB
- Very efficient retrieval of top-k completions
  [Hsu and Ottoviano, WWW 2013]
- Inverted Index over queries for non prefix match suggestions

AUTOSUGGEST – CHALLENGES



Context Matters

What should we suggest?

Model P (Query | UserContext, Time)

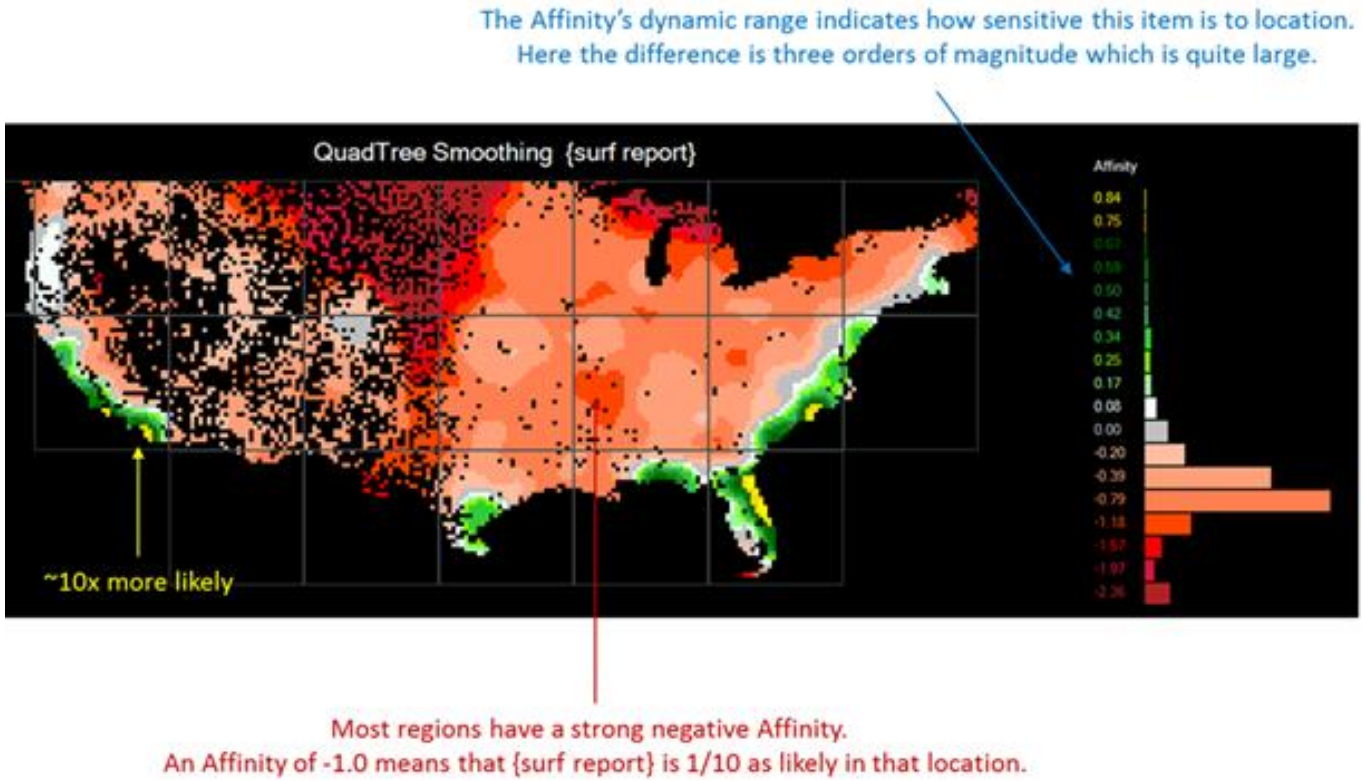
Alice	Bob	Charlie
User previous queries: - movie streaming - imdb ranking	User location: 	Day of week: Sunday Time of day: 12h30 Device: mobile
netflix	**new york times**	**nearby restaurants**

User Location

Some queries are much more popular in some places than in others

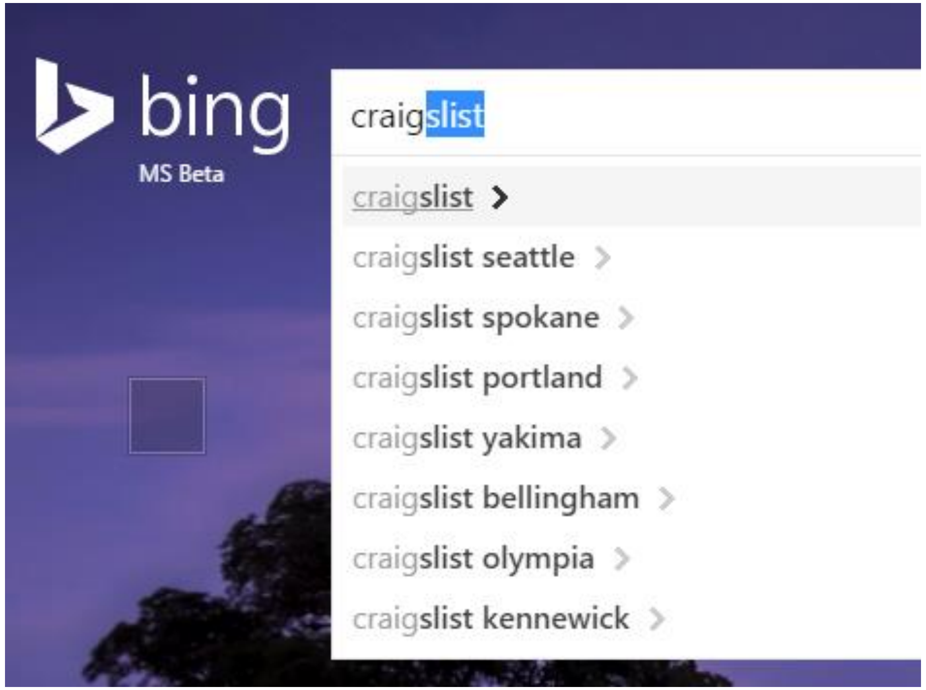
Modify Query Prior Prob by

Affinity (Query, Location) learnt from data

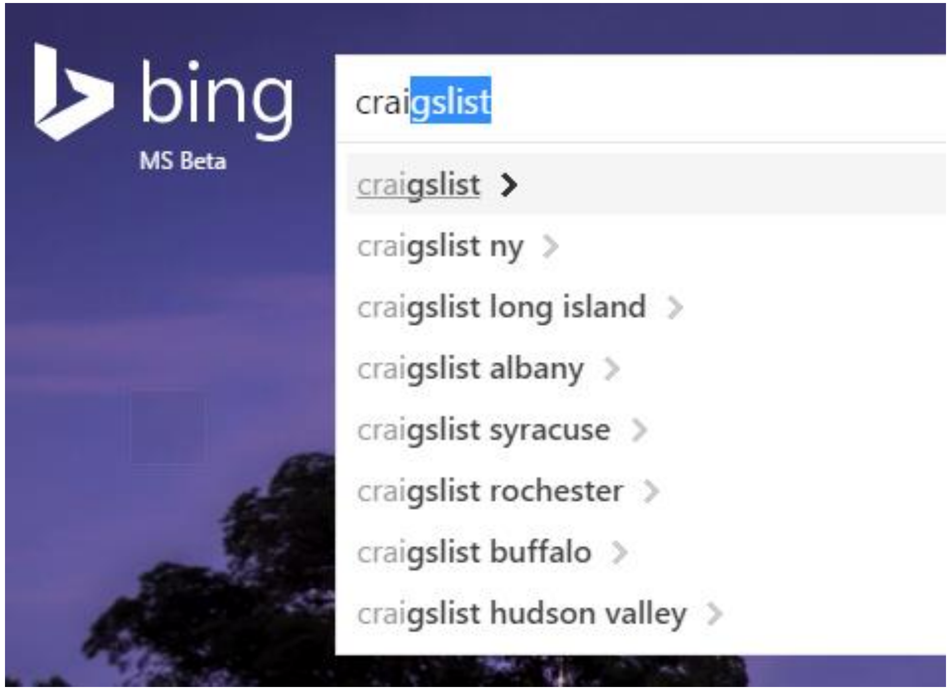


Localized Suggestions in Action

User Location: Redmond, WA



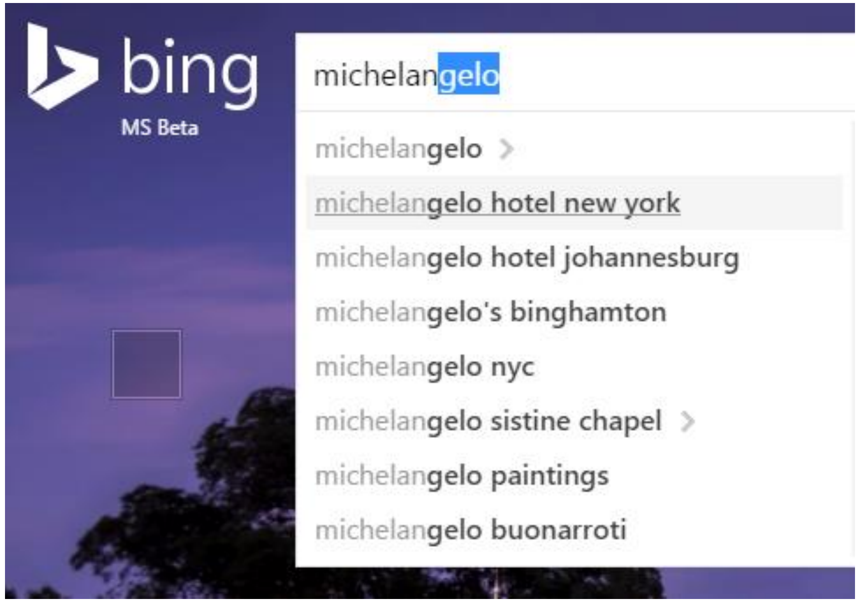
User Location: New York, NY



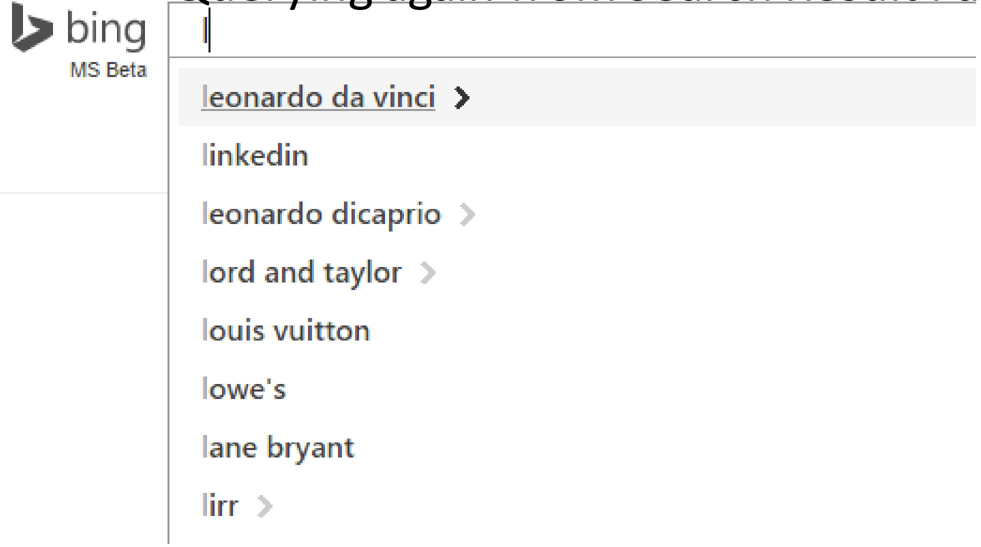
Previous Query

Modify Query Prior by *Affinity (Query, Previous Query)*

User Previous Query:

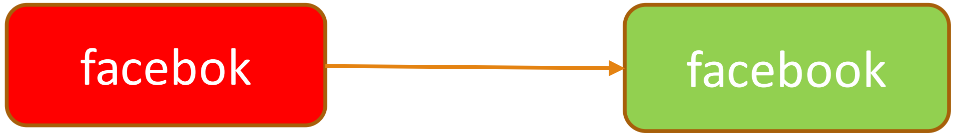


Querying again from Search Result Page:

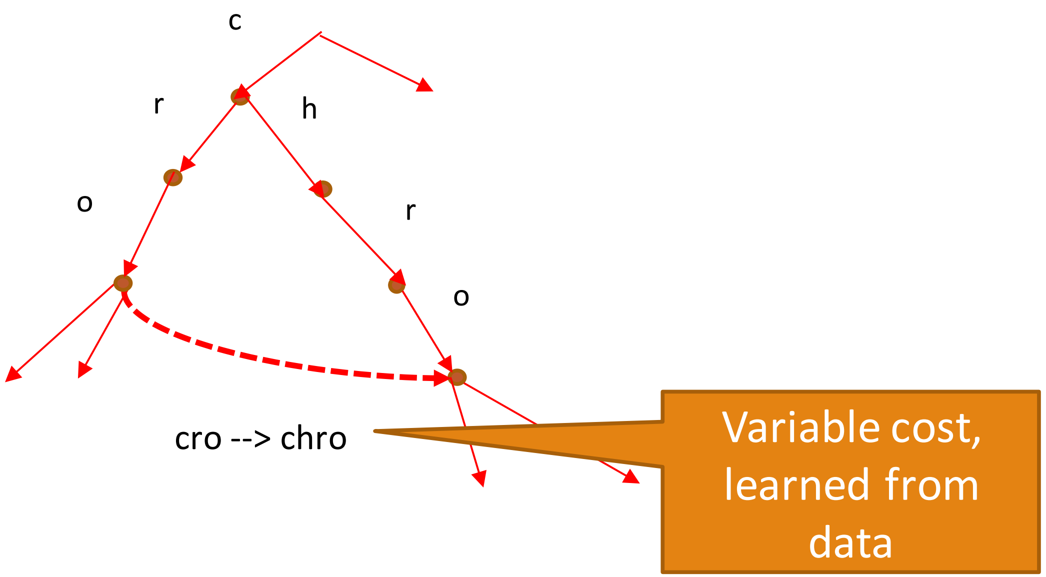


Spelling Corrections

5-15% of submitted queries contain spelling errors



- Offline spell corrections
  - Popular misspellings are stored in the trie with a pointer to their correction
- Online spell corrections
  - Exploration of the trie using an error model learned from the data (frequent typos have low penalty) [Duan and Hsu, WWW 2011]

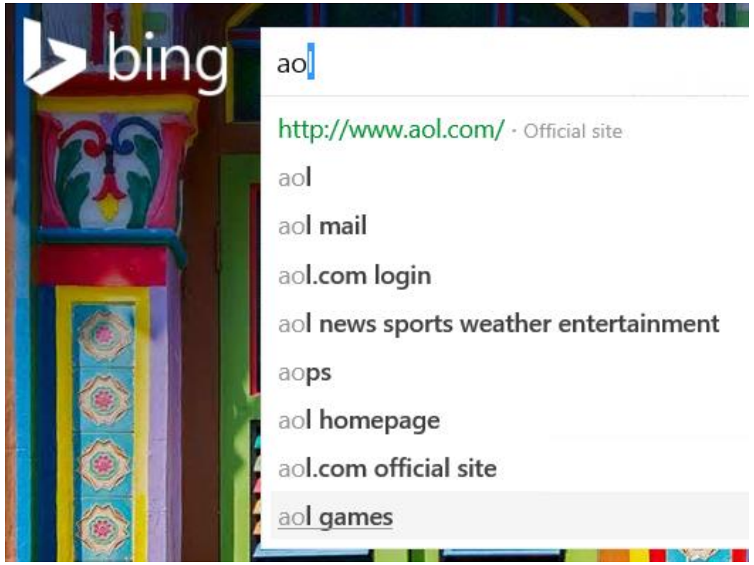


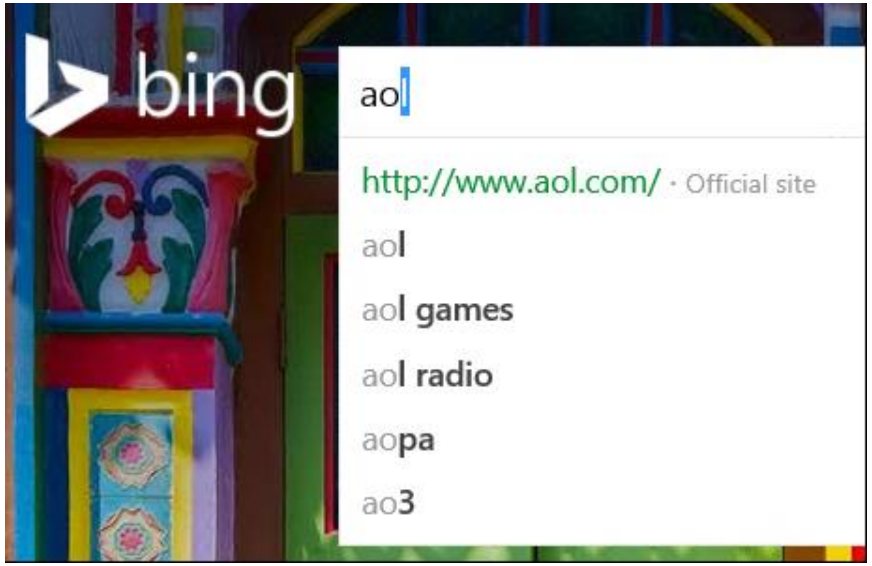
Diversity

Avoid showing “duplicate” suggestions e.g.:

- “aol”, “aol homepage”, “aol.com official site”

➔ Represent a diverse set of intents when input is underspecified





More Challenges

- Filtering queries with explicit sexual intent, offensive queries and queries inciting to commit crimes
- Filtering Spam queries
- Ensuring high availability and low latency
- …

Microsoft

Trade-offs – what to build and how to build it

This is were understanding the fundamentals of computing at scale kicks in

There are no easy answers:

Each potential solution has a cost in terms of complexity, storage, compute usage, serving cost

These costs need to be weighed up against product impact (which we won’t cover here)

Everything needs to be measured:

- Instrument production systems
- Profile during coding, don’t assume
- Gate on performance during builds and deployments
- Continuously evaluate and re-evaluate as systems change

Systems vs fundamentals: Lesson 1

Core trie data structure optimized to do micro second lookups

Custom data structure with per processor optimization to reduce cost online

Extreme computing win, yes?

Partly:

Code so complex it's almost impossible to maintain

Extreme computing requires extreme engineering: turns out the serialization code was 100x slower than the data structure

Systems vs fundamentals: Lesson 2

Let’s go back to basics. We have a middle tier workflow engine written in C#. You’ve been asked to check at runtime whether a string is a duplicate in a list of strings you’ve already. How do you do this?

A hash table/map? Right?

Depends.

It turns out for sufficiently small collections, iterating every time is not much slower.

Also, because every machine is serving multiple requests, it also reduces the impact on other requests by reducing memory overhead and garbage collection pressure.

Sometimes, simple data structures and algorithms are more efficient overall

Systems vs fundamentals: Lesson 3

Map-reduce is awesome, it allows us to process petabytes of data

However, at some point even map-reduce doesn't scale

Issues:

Data skew

Logs, even split across partitions, larger than can be read or processed

Sampling as a solution

Systems vs fundamentals: Lesson 4

The real complexity is not in the code or the individual components though

The system overall is larger than most engineers or scientists can reason over

A lot of extreme engineering is in place to allow extreme computing flourish

This abstraction needs to be balanced against needing to understand the complexity to produce elegant and efficient solutions

There is no easy answer: it’s part science, part art, part experience

Conclusion

The techniques you are learning in this course provide a strong foundation to work at scale

We focus on these fundamentals

- When we design systems

- When we interview candidates

However extreme the computing, it fails if the systems are in place to facilitate it

Personal learning: the optimisations that mattered a generation or so ago matter again

Paul Baecke

Add me on LinkedIn if you want to stay in touch:

https://uk.linkedin.com/in/paulbaecke