



*Architecture, Development Model and Future
Trends of Web Search Engines*



Marcelo De Barros

Bing UX Features and Shared Tools Team

Plan for the next hour

- I'll present you a simplified view of Search Engines architecture.
- I'll try **not** to use jargon without explaining it. Stop me if I forget.
- I'll talk about the future trends around search engines (my own opinion).
- You ask questions if you have them.
- If I can't answer, I'll follow up with someone who can.

This deck represents an overview of Search Engines.

Some technical implementation details will be omitted on purpose 😊



**WHAT
IF I
TOLD
YOU**

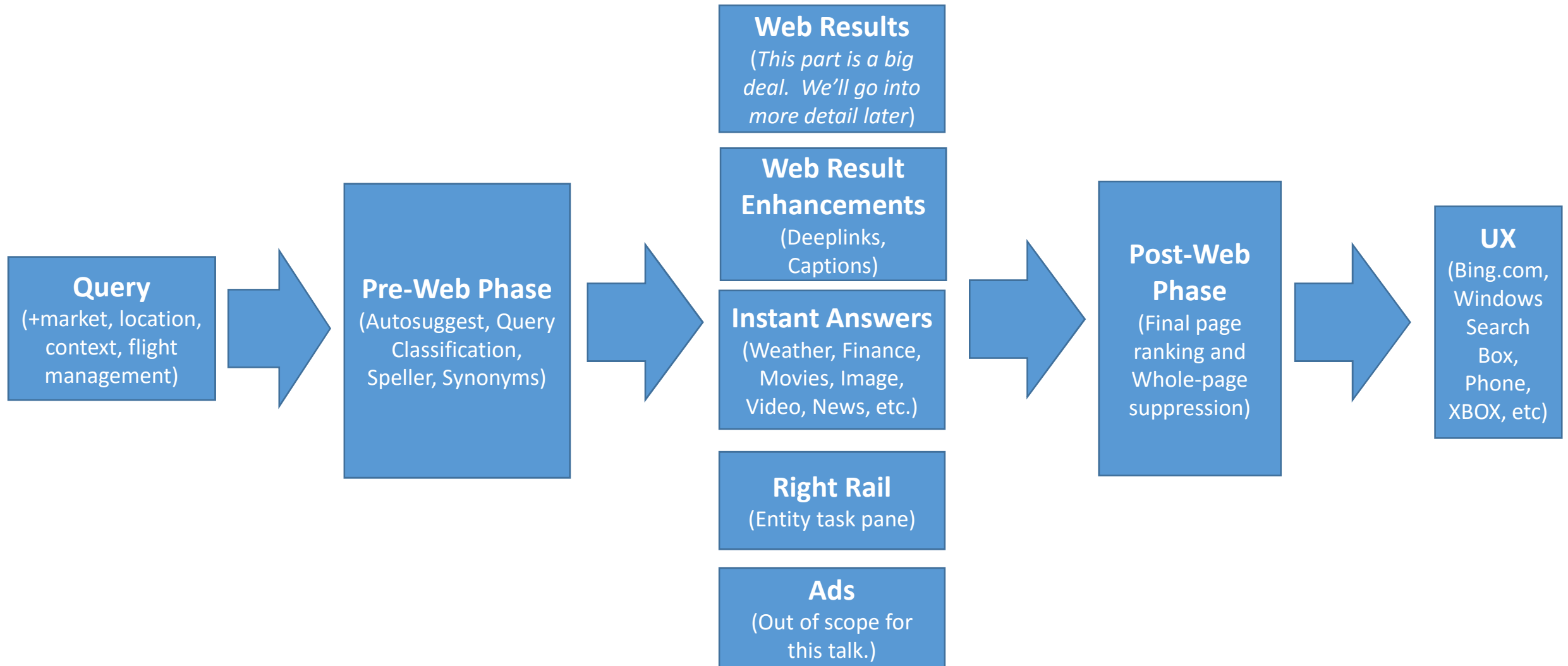
**THAT
READING A
POWERPOINT
ALoud IS NOT
THE SAME AS
TEACHING**

The Anatomy of a Bing SERP – Search Engine Results Page

The image shows a Bing search results page for the query "katy perry". The search bar at the top contains the text "katy perry" and shows "49,200,000 RESULTS" and "Any time" filter. The page is annotated with several callouts:

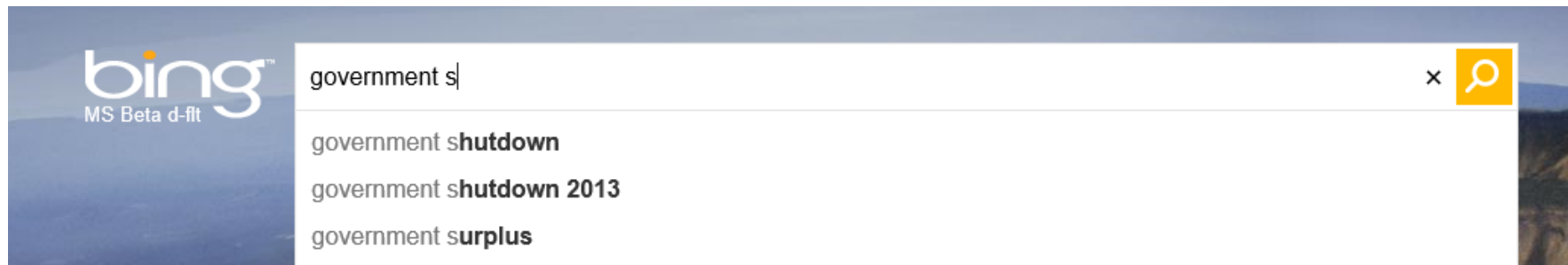
- Search Box, Navigation, Settings:** Points to the top navigation bar and search bar area.
- Instant Answers (News, Images):** Points to the "News about Katy Perry" section, which includes a news snippet from ABC News and a grid of images.
- Web Results:** Points to the "KATY PERRY // Official Website" and "Katy Perry - Wikipedia" search results.
- Deeplinks:** Points to the "Early life · Career · Art" links at the bottom of the Wikipedia result.
- Task Pane:** Points to the right-hand sidebar containing a biography, social media links, and a list of songs and albums.

A Runtime Stack in One Slide



The User Query

- Things that a search engine might know even *before* we get to the web search:
 - Your query
 - Your entry point (Windows Search Box, Bing.com, Phone, XBOX, etc.)
 - Your market (country + language)
 - Your location (sometimes...)
 - Your past queries (sometimes...)
 - Your identity from logged-in experiences (sometimes)
 - Which flights (experiments) you are in
- Query Formulation via Autosuggest (traditional *trie* data structure)



Understanding Popular Pages

- Search engines know a number of statistics about the pages:

[KATY PERRY // Official Website // \(1.25%\)](#)

www.katyperry.com ▾ [Translate this page \(0%\)](#)

The official website of Katy Perry. New Album 'Prism' coming October 22, 2013!
Featuring the single ROAR!

[Videos; \(0.38%\)](#)

[Movie \(0%\)](#)

[Tours; \(0.2%\)](#)

[One of The Boys \(0%\)](#)

[Discography; \(0.01%\)](#)

[North America \(0%\)](#)

[Bio; \(0.1%\)](#)

[Tour-Diary \(0%\)](#)

[News; \(0.01%\)](#)

[See more results \(0.02%\)](#)

[Katy Perry - Wikipedia, the free encyclopedia \(1.64%\)](#)

en.wikipedia.org/wiki/Katy_Perry ▾ [Translate this page \(0%\)](#)

Sample Data

(not real numbers)

- Such information helps in ranking decisions, as well as caching decisions and placement decisions

Query Rewriting: Spelling and Synonyms (*pre-web*)

- Spelling:
 - Dictionary (per language)
 - Logs (words proximity, clustering techniques, ranking within clusters)

The screenshot shows a Bing search interface with the query 'ktay perry'. The search bar is at the top, and the results are listed below. The first result is 'katy perry' with a rank of 27.379032, which is highlighted with a green box. Other results include 'ktay perry', 'katy parry', 'katy perry'', 'katyperry', 'katey perry', 'kathy perry', 'katie perry', and 'kay perry', each with a rank and a link to 'Features - Local NGrams - Odyssey'.

WEB IMAGES VIDEOS MAPS NEWS MORE

bing MS Beta

ktay perry

49,700,000 RESULTS Any time ▾

Including results for *katy perry*.
Do you want results only for *ktay perry*?

- [katy perry](#) Ranker: 27.379032
Features - Local NGrams - Odyssey
- [ktay perry](#) Ranker: 13.263762 Original
Features - Local NGrams - Odyssey
- [katy parry](#) Ranker: 4.836921
Features - Local NGrams - Odyssey
- [katy perry'](#) Ranker: 3.503151
Features - Local NGrams - Odyssey
- [katyperry](#) Ranker: 3.23323
Features - Local NGrams - Odyssey
- [katey perry](#) Ranker: 2.711165
Features - Local NGrams - Odyssey
- [kathy perry](#) Ranker: 2.006864
Features - Local NGrams - Odyssey
- [katie perry](#) Ranker: 1.353943
Features - Local NGrams - Odyssey
- [kay perry](#) Ranker: -1.88421
Features - Local NGrams - Odyssey

The screenshot shows a Bing search interface with the query 'perfecct'. The search bar is at the top, and the results are listed below. The search bar is highlighted with a blue box. The results are listed below, and the first result is 'perfect' with a rank of 90,500,000. The search bar is highlighted with a blue box.

perfecct

Web Images Videos Maps News Explore

90,500,000 RESULTS Any time ▾

Including results for *perfect*.
Do you want results only for *perfecct*?

Query Rewriting: Spelling and Synonyms (*pre-web*)

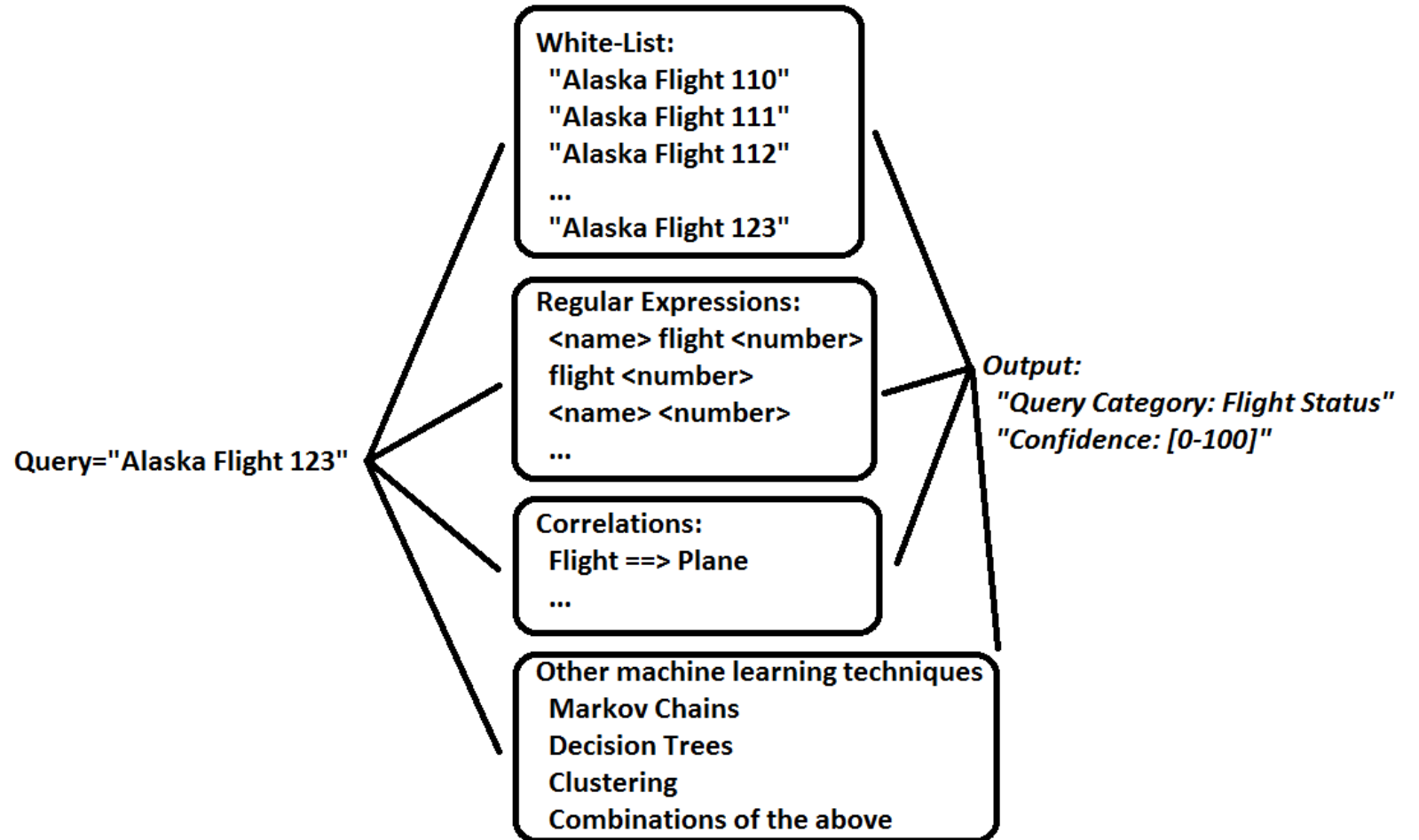
- Synonyms:
 - Clustering techniques

Term	Candidate	Tier1 Score
[-] car	car	N/A
	auto	999966
	automobile	999885
	automotive	999431
	carr	996370
	cars	999991
	motor vehicle	918356
	truck	999848
	van	998102
	vehicle	999862
	vehicles	999752
[-] repair	repair	N/A
	fix	999851
	fixing	1000000
	maintenance	999855
	owners	999849

The screenshot shows the Bing search engine interface. At the top, there are navigation links for WEB, IMAGES, VIDEOS, MAPS, NEWS, and MORE. The search bar contains the text "car repair" and a magnifying glass icon. Below the search bar, it displays "110,000,000 RESULTS" and "Any time" with a dropdown arrow. The first search result is titled "DIY Auto Repair Help - Car Maintenance, Troubleshooting, ..." from the website "autorepair.about.com". The snippet below the title reads: "You can do your own auto repairs by following out easy step-by-step do-it-yourself tutorials which show you how to diagnose, troubleshoot, repair, fix, modify and ...". Below the snippet are links for "Fix It Yourself", "Troubleshooting", "DIY Repairs", and "My Car Won't Start".

Query Rewriting: Query Classification (*pre-web*)


- Query Classification:
 - Fast Classifiers
 - White-List
 - Regular Expressions
 - Correlations
 - Other techniques



Instant Answers: a Federated Model

- After pre-web components are run, the query is federated out (dispatched) to dozens of *Answer Services*
- *Anybody* can ship an Answer service, and *any* answer can trigger for *any* query
- Answers vary widely in complexity. Some (like Flights/News/Stocks) have up-to-the-minute data requirements. Some (like Image/Video) have full indices and relevance stacks.

[Flight status for American 10](#)
flightstats.com · 1 minute ago






 Departing on time at 8:45 PM from LAX

FROM	LAX Los Angeles	8:45 PM 10/14/2013
TO	JFK New York	5:05 AM 10/15/2013

the post · Bruno Mars height
5 feet 5 inches (1.65 meters)
Find out more on: [Freebase](#)

Current time in Dubai, Dubai, United Arab Emirates
07:23:41 PM
10/14/2013 · Arabian Standard Time

age ra [Weather in Hyderabad, India](#)
bing.com/weather · Data from AccuWeather





Now	Mon	Tue	Wed	Thu	Fri
80° Clear					
°F °C	88° / 67°	88° / 66°	86° / 68°	84° / 68°	83° / 67°

Change providers: [Foreca](#) · [Weather Underground](#) · [Compare all](#)




[Cooking School near Seattle, Washington](#)
bing.com/local

- 1 Blue Ribbon Cooking LLC
2501 Fairview Ave E, Seattle · (206) 328-2442
[Directions](#)
- 2 Nu Culinary
6523 California Ave Sw, Seattle · (206) 932-3855
[Directions](#) · ★★★★★ 1 review
- 3 Bon Vivant School of Cooking
4925 Ne 86th St, Seattle · (206) 525-7537
[Directions](#)

[Videos of video how to install a stereo](#)
bing.com/videos

 6:41	 6:01	 23:14	 10:30
How to wire / Install a Car St... YouTube	How to Install Your Own Car ... YouTube	How-To Install a Car Stereo Sys... YouTube	How to install an aftermarket car ... YouTube

[Movies near Topeka, KS](#)
msn.com/movies

 Captain Phillips ★★★★★ · Drama · PG-13 · 2hr 14min	 Rush ★★★★★ · Action · R · 2hr 03min
 Insidious: Chapter 2 ★★★★★ · Horror · PG-13 · 1hr 45min	 Runner Runner ★★★★★ · Suspe · R · 1hr 31min

Instant Answers: Targeted Experiences

- Instant answers are a great way to meet users' demands
- Users no longer have patience for the traditional blue links 😊

Web Images Videos Maps News

17,400,000 RESULTS Any time ▾

Polynomial equation solver

Standard form:
 $x^{10} - 3x^3 + 9 = 0$

Solutions based on [Jenkins–Traub algorithm](#):

$x_1 = -1.213921 - 0.454848i$
 $x_2 = -1.213921 + 0.454848i$
 $x_3 = -0.660282 - 0.958082i$
 $x_4 = -0.660282 + 0.958082i$

[Show all ▾](#)

🔍

Web Images Videos Maps News Explore

12,700,000 RESULTS Any time ▾

Periodic table

Chemical group Physical state Discovered Found on Earth Density

Legend:

- Hydrogen (Green)
- Alkali metals (Yellow)
- Alkali-earth metals (Light Blue)
- Transition metals (Orange)
- Post-transition metal (Dark Blue)
- Metalloid (Light Green)
- Polyatomic nonmetal (Light Yellow)
- Diatomic nonmetal (Light Orange)
- Noble gas (Red)
- Lanthanide series (Light Grey)
- Actinide series (Dark Grey)

🔍

Web Images Videos Maps News Explore

Also try: [Fibonacci Sequence Python Code](#) · [Python Fibonacci Recursion](#) · [Fibon...](#)

169,000 RESULTS Any time ▾

Fibonacci

Python ▾

```
1 def calculate_fib(n):
2     fib, tmp = 0, 1
3     for i in range(n):
4         fib, tmp = tmp, fib + tmp
5     return fib
6
7 if __name__ == "__main__":
8     print calculate_fib(5),
9     print calculate_fib(10),
10    print calculate_fib(15),
```

5 55 610

Powered by [HackerRank.com](#)

Entity (or Side) Pane

- The Entity Pane is a special kind of Instant Answer
- It pulls in content from various answers and displays it all together in one place
- Search engines keep a graph of entities on the Web

Gravity (2013)



Dr. Ryan Stone is a brilliant medical engineer on her first shuttle mission, with veteran astronaut Matt Kowalsky. But on a seemingly routine spacewalk, disaster strikes. The shuttle is destroyed, leaving Stone and Kowalsky completely alone - tethered to nothing but each other and spiraling out into the blackness. The deafening silence tells them th... +

Watch trailer: [Moviefone](#)

Summary: PG-13 · 1hr 31min · SciFi/Fantasy

Director: [Alfonso Cuarón](#)

Reviews

★★★★★ 90,261 ^

[IMDB](#)

8.8/10

[MSN](#)

4.5/5 stars

[Flixster](#)

97% positive

Cast



[Sandra Bullock](#)
Dr. Ryan Stone



[George Clooney](#)
Matt Kowalsky

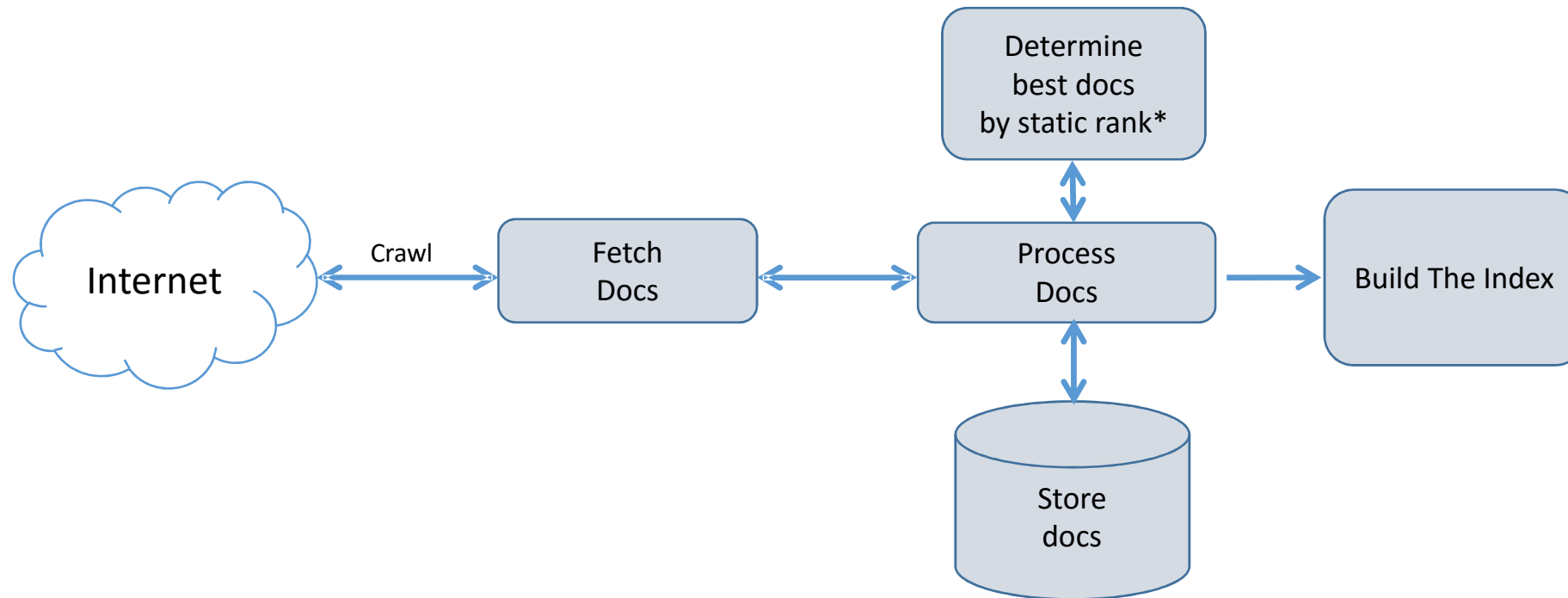
An Aside on Web Relevance

Web Relevance

- Objective: find the 10 (sometimes more, sometimes less) most relevant blue links for the query and put them in the right order on the page
- How this happens in 6 oversimplified steps:
 1. Acquire billions of web documents and index them
“this is a hard problem from many angles, mainly from a scalability and storage standpoint”
 2. Match each user query to some possibly relevant web docs
 3. Use machine learning to rank the candidate web docs
 4. Return the top ten (give or take) to the user
 5. Do this globally
 6. Do this in a blink of an eye!!!

Where do the documents come from?

Generation of the Index = the process of crawling/storing docs and building the index



*Static Rank = the query-independent importance score that we assign to every document on the web

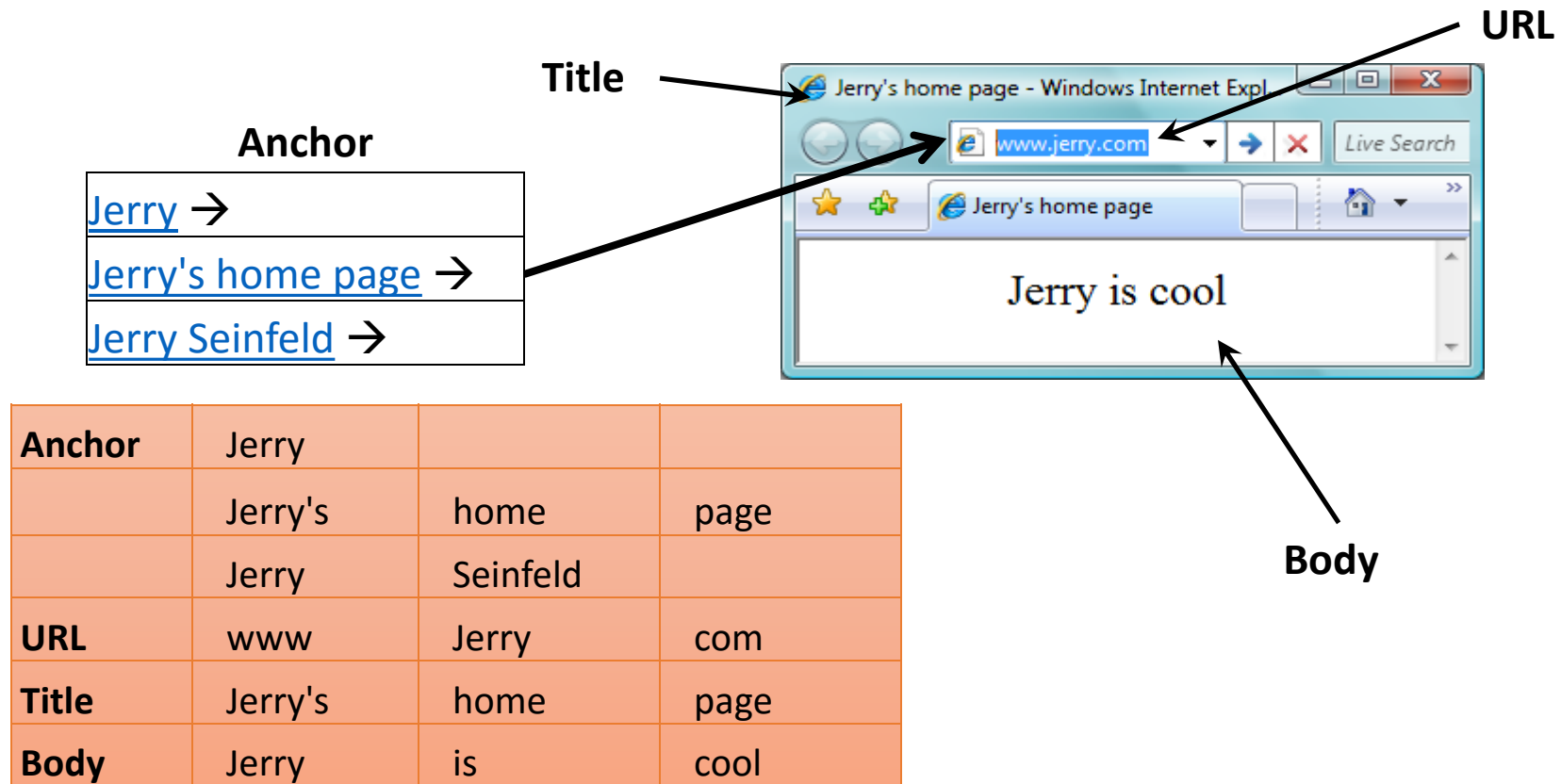
How are documents served for a query?

Index-Serve = the process of hosting docs and returning them for incoming queries

- Search Engines have multiple-tiers platform to balance for freshness, relevance, index depth, and cost. Bing for example:
 - Fresh tier
 - **Millions** of documents
 - Doc discovery to hosting takes <1min
 - Main tier
 - **Billions** of documents
 - Updates in <1 day
 - Depth tier
 - **Many Billions** documents
 - Updates in <7 days
- Includes both En-us docs as well as global docs

How is a query matched to documents?

There are four basic streams (text sources): **Anchor**, **URL**, **Title**, and **Body (or AUTB)**



- AUTB is just the basics that Bing uses. Other engines might use other streams. We also rely heavily on Speller and Synonym expansion.

So Far, We Have a Big Pile of Documents

- We've matched a few thousand (or more) documents to your query
- Now we just need to get them in the right order
- How do we do that? Machine learning!

Machine Learning is like Guess Who



The Steps of Machine Learning

Machine learning helps a machine answer human questions (e.g. what are the best docs for this query?) by quantifying human questions into scores.

1. First, create some examples *where you know the right answer*. This is called **training data**.
2. Figure out some important *easy* and *quantifiable* questions to ask of those examples. These questions are called **features**.
3. Use the training data to “learn” how to get the known examples right by adjusting weights until the numbers work out. This is called the **training process**.
4. Then, for *new* examples, the system can take an educated guess at the right answer. This is called **generalization**.
5. **Measure** how well you do. Do this early and often.
6. Then go back and fix the problems. This is called **tuning**. Rinse and repeat.

Web Ranking Features

Features can be for the query, the doc, or both. Here are just a few examples of many that are used by different search engines:

- Do the query and doc belong to the same **category**? (sports, movies, etc.)
- Do the query and doc come from the same **geographical origin**?
- How many times does the query term appear in the doc? (**frequency**)
- Does the query have any **known phrases**? e.g. {*star wars* trailer}
- How important is the doc? (Remember **static rank**?)
- We also look at **doc clicks**.
- Has the doc been classified as **junk/spam/adult**?
- What **query terms** have people used in the past to get to (click on) this doc? (queries association technique)
- And many, many more!

*In the end, each query/doc pair gets a **dynamic rank** score. The docs are ordered by this score.*

How do we gather training data?

Relevance Measurement: judges assess query/doc pairs on a five-point scale. This is used for both training and testing.

The process of pulling in the top N docs for a query and storing them is called **scraping**.

We use these judgments to train, test and measure our rankers (machine learning models).

The image shows a screenshot of a Windows Internet Explorer browser window. The main window displays the SHOP.COM website. At the top, there is a banner for '20% OFF GIFTS for a limited time only' with a coupon code '20OFFGIFTS'. Below the banner, the SHOP.COM logo is visible, along with navigation links for 'stores', 'clothing', 'shoes', 'beauty', 'home & housewares', 'electronics', 'all departments', 'gifts & registry', and 'sale'. A search bar is present with the text 'I'm shopping for' and a 'find now' button. Below the search bar, there are several sponsored links and a product listing for 'The Streetcar Named Desire - The Music Of Max Steiner'. The product listing includes the title, format (CD), performer (Original Soundtrack), and price information: 'was \$13.08 now: \$11.99 - \$13.49'. A 'Buy it here' button is visible below the product listing.

On the left side of the browser window, there is a smaller window titled 'HRS Web Judging - W...'. This window displays a judging interface for the query 'A Street Car Named DEsire'. The interface includes a rating scale with radio buttons for 'Perfect [t]', 'Excellent [r]', 'Good [e]', 'Fair [w]', 'Bad [q]', 'Page Did Not Load [2]', and 'Detrimental [1]'. There are also buttons for '<< [c]', 'Review [v]', and '>> [b]'. The URL in the address bar of this window is 'http://www.shop.com/+p94973390-st.shtml'.

Bringing all together – recap!

1. A query comes in via one of several entry points
2. Some contextual information comes with the query
3. A few core services (e.g. Speller, Alteration) process the query
4. The query is “federated out” to Web, Answers, Task Pane, etc.
5. A subset of answers trigger for the query
6. The web ranker matches *many* documents and returns the top 10
7. All of this takes few milliseconds...
8. Now, we have a big pile of stuff waiting to be rendered on the page

Page Coherence

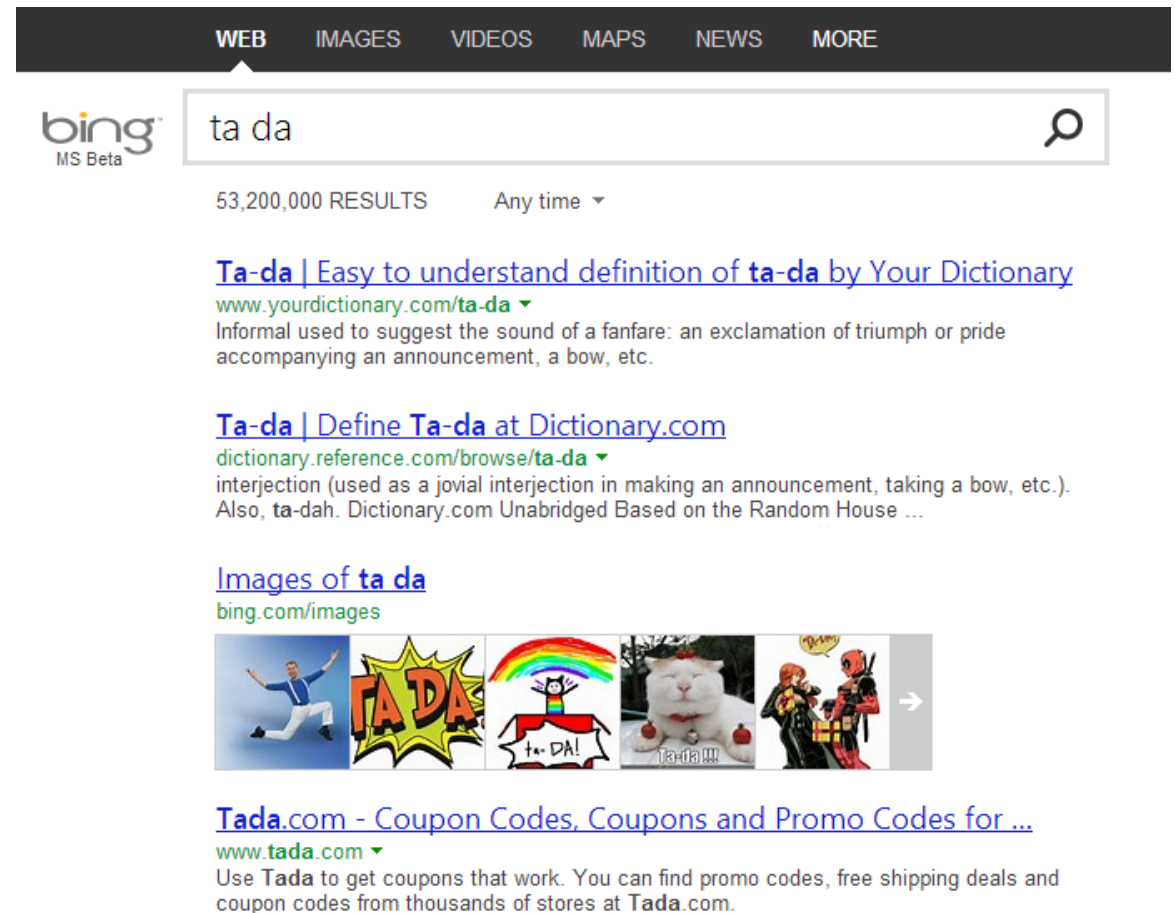
- It doesn't look good to show apples and oranges intertwined...
 - Jaguar: The Car? The Animal? The City?
- Need to apply **suppression**, and then
- Need to apply **final ranking**
- **Coherence** between web docs and answers is a key component
- **Past data** (user clicks) is also important
- The job of suppression is to **minimize defects**
 - A defect = irrelevant or otherwise bad content for a query
 - Components that perform poorly lose credibility
- The job of the final page ranking is to **push the best stuff to the top and the less-good stuff toward the bottom**
 - This is done via a metrics derived from click-info

The screenshot shows a Bing search for "grizzly bear". The search bar contains "grizzly bear" and shows "8,000,000 RESULTS". Below the search bar is a grid of eight images of grizzly bears. To the right is a knowledge panel titled "Grizzly bear" with a small image of a grizzly bear. The panel includes a Wikipedia snippet: "The grizzly bear, is any North American subspecies of the brown bear, such as the mainland grizzly, the Kodiak, the peninsular grizzly and the recently extinct California grizzly. Specialists sometimes call the grizzly the No...". Below this is the scientific name "Ursus arctos" and biological classification "Subspecies". It also lists "Belongs to: Brown bear" and "Notables: Old Ephraim". At the bottom of the panel is a section "People also search for" with small images and labels for "American", "Kodiak bear", "Polar bear", "Gray wolf", and "Moose".

The screenshot shows a Bing search for "grizzly bear band". The search bar contains "grizzly bear band" and shows "1,080,000 RESULTS". Below the search bar are search results for "Grizzly Bear" from "grizzly-bear.net", described as an "Indie-rock band from Brooklyn". There are links for "Live", "Videos", "Music", "About", and "Contact". To the right is a knowledge panel titled "Grizzly Bear" with a small image of the band. The panel includes a Wikipedia snippet: "Grizzly Bear is an American rock band from Brooklyn, New York, formed in 2002. The band consists of Edward Droste, Daniel Rossen, Chris Taylor and Christopher Bear." Below this is a "KLOUT" score of 84 and a list of "Songs" including "Slow Life (With Victoria Legrand)", "Service Bell", "Deep Blue Sea", and "This Song".

UX (User Experience, or UI)

- After Whole-Page Relevance decides what to show, it passes the final content to the UX layer
- The content is rendered beautifully on the page
- The layout is customized by entrypoint, but the content is (mostly) the same
- UX Server: ASP.Net
- UX Client: Java Script (Libraries) + HTML + CSS3



The screenshot shows a Bing search results page for the query "ta da". At the top, there is a navigation bar with links for WEB, IMAGES, VIDEOS, MAPS, NEWS, and MORE. The Bing logo and "MS Beta" are visible on the left. The search bar contains the text "ta da" and a magnifying glass icon. Below the search bar, it indicates "53,200,000 RESULTS" and "Any time". The first result is from "Your Dictionary" with the title "Ta-da | Easy to understand definition of ta-da by Your Dictionary" and the URL "www.yourdictionary.com/ta-da". The description reads: "Informal used to suggest the sound of a fanfare: an exclamation of triumph or pride accompanying an announcement, a bow, etc." The second result is from "Dictionary.com" with the title "Ta-da | Define Ta-da at Dictionary.com" and the URL "dictionary.reference.com/browse/ta-da". The description reads: "interjection (used as a jovial interjection in making an announcement, taking a bow, etc.). Also, ta-dah. Dictionary.com Unabridged Based on the Random House ...". Below the text results, there is a section titled "Images of ta da" with the URL "bing.com/images". This section displays a horizontal carousel of five images: a man in a blue suit performing a bow, a yellow comic book-style explosion with the word "TADA" in red, a cartoon character with a rainbow above their head and a sign that says "ta DA!", a white cat with a red bow around its neck and the text "Tada!!!", and a Deadpool character holding a gift box. A right-pointing arrow is visible at the end of the carousel. The final result is from "Tada.com" with the title "Tada.com - Coupon Codes, Coupons and Promo Codes for ..." and the URL "www.tada.com". The description reads: "Use Tada to get coupons that work. You can find promo codes, free shipping deals and coupon codes from thousands of stores at Tada.com."

Engineering Development Rhythm

Inner Dev Loop

- Feature development
- Concludes at checkin

Outer Dev Loop

- Build validation
- Concludes at PROD deployment

Monitoring

- Live Site quality
- Continuous

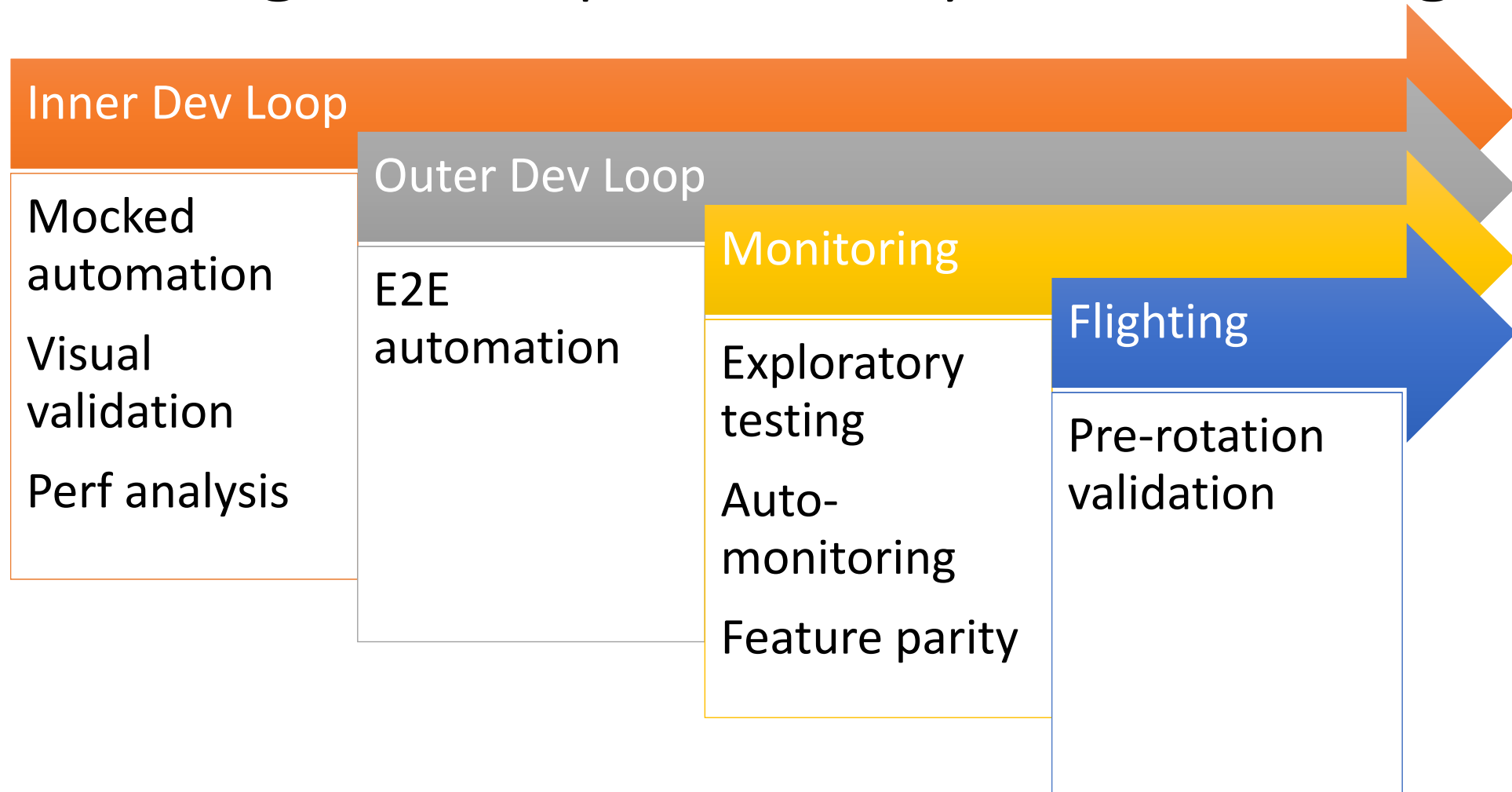
Flighting

- Controlled exposure of features

Development is composed of discrete states

<https://www.youtube.com/watch?v=SiPtRjiCe4U>


Engineering Development Rhythm - Testing



Testing is composed of overlapping states

Engineering Development Rhythm

- Hundreds of engineers across many continents!
- Shipping multiple times a day (millions of lines of code):
 - *Continuous Delivery* → “your check-in will go to production soon!!!”
- Tens of thousands of automated tests
 - If any fails → don’t ship
 - Don’t write tests? Well, good luck shipping to hundreds of millions of users!!!
- Flight everything → Analyze the data → Ship or fail fast!!!

Flowers at 1-800-FLOWERS - Same Day Delivery Available. 

 1800flowers.com

★★★★★ (183920 reviews) · 37,000+ followers on Twitter

Same Day Delivery Available. 100% Satisfaction at 1-800-FLOWERS.

[Anniversary Flowers.](#)

[Best Selling Flowers.](#)

[Rose Spectacular.](#)

[Birthday Flowers & Gifts.](#)

[Fresh Cuts.](#)

[Gift Baskets.](#)



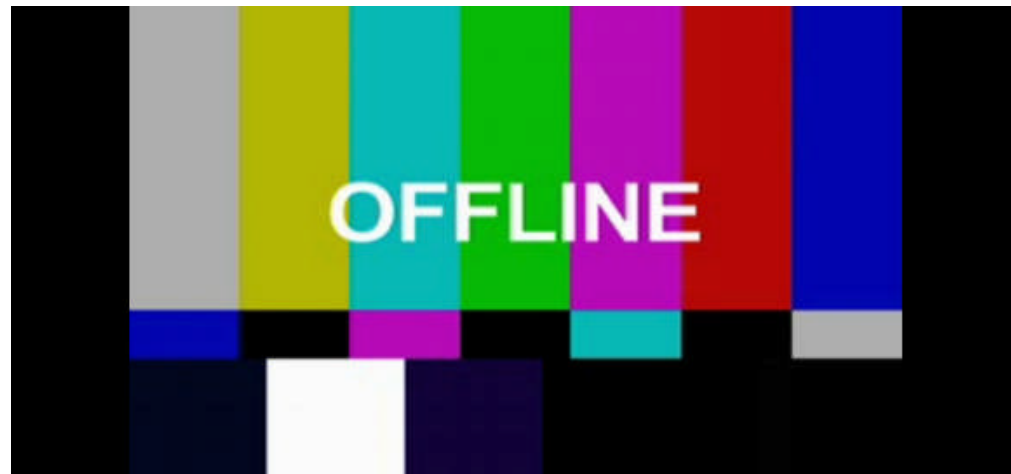
Guardrail Metrics	Treatment	Control	Delta [%]	Pval
Quick Back 20	0.2295	0.2281	0.0014 [0.60%]	< 0.001
Algo Pane Load Time(Overall PLT)	1212	1208	4.055 [0.34%]	< 0.001
Revenue /UU	1.088	1.075	0.0130 [1.21%]	< 0.001
Truncated Revenue / UU	0.8571	0.8504	0.0067 [0.79%]	< 0.001
Distinct Queries / UU	14.67	14.67	-	1.001
Average Log Record Size (in KB)	111.4	111.1	0.2545 [0.23%]	< 0.001

Search Engines Future Trends and Challenges



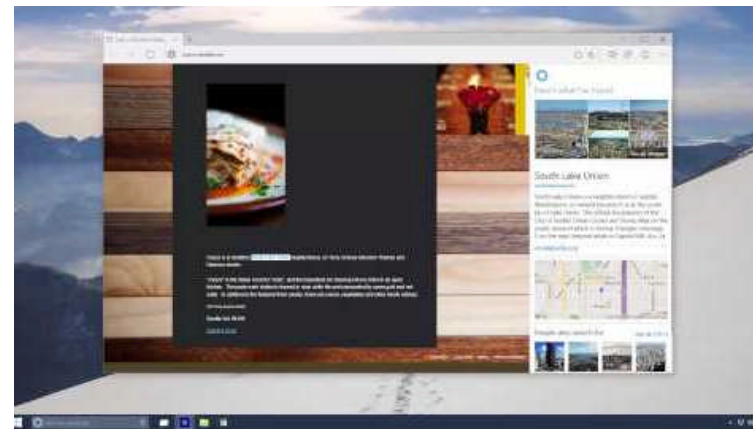
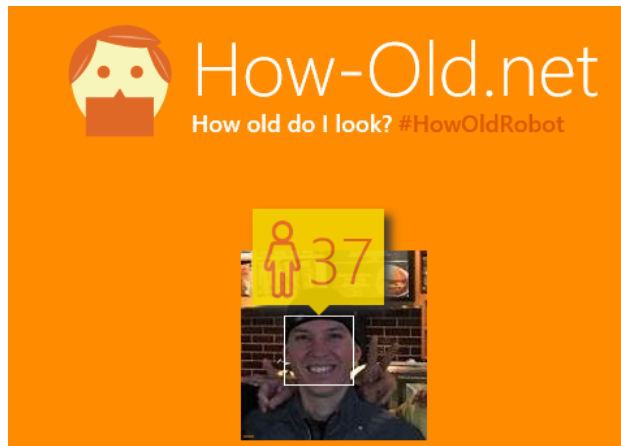
Search Engines Future Trends and Challenges

- **Data:** not every data is in the index...
 - Offline data – other formats
 - Live data - happening now, I mean, really, NOW!!!



Search Engines Future Trends and Challenges

- **AI:** it is only in its infancies
 - Image and Video Understanding
 - Media → Features → Syntax → Semantics
 - Personal Assistant (Cortana, Siri, Google Now, Alexa)

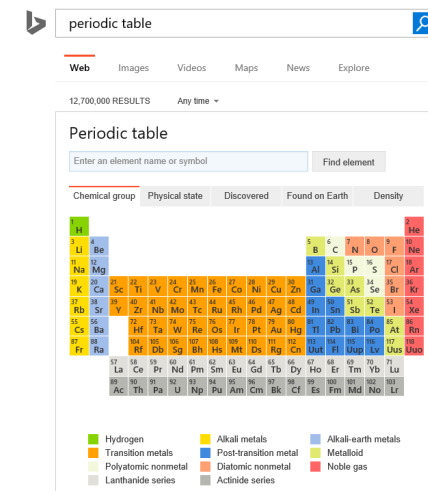


Search Engines Future Trends and Challenges

- **Fundamentals:** more connections, ~~less~~ no patience
 - Internet of Things (IoT)
 - Availability across devices (phones, wearables, cars, things)
 - Poses unique User Interface challenges
 - Poses unique privacy concerns
 - Performance:
 - Not fast... but NOW!!!!
 - Fun experiment: slowdown flight → revenue hit!
 - Pushing the limits of techniques
 - Algorithms, distributed computation, hardware, networks, caching, programming languages, etc.
 - Faster data analysis
 - Data is becoming cheaper...
 - However, useful information from the massive data sets is hard!

Search Engines Future Trends and Challenges

- **Collaboration:** search is also about connecting services
 - No more blue-links: the answer must be right there!
 - Many specialized companies
 - *Servicefication* of platforms



Q&A