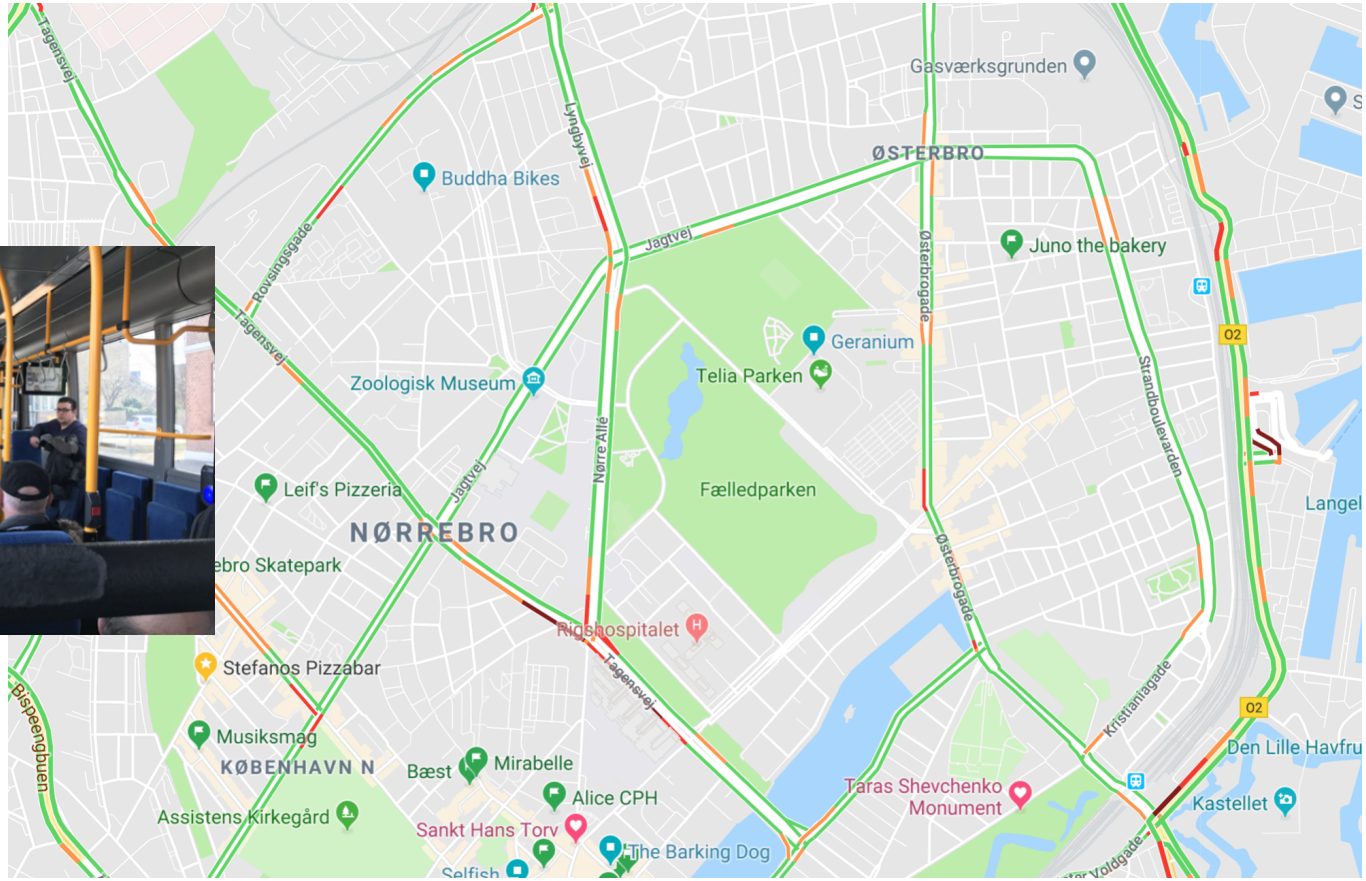
42578 Advanced Business Analytics

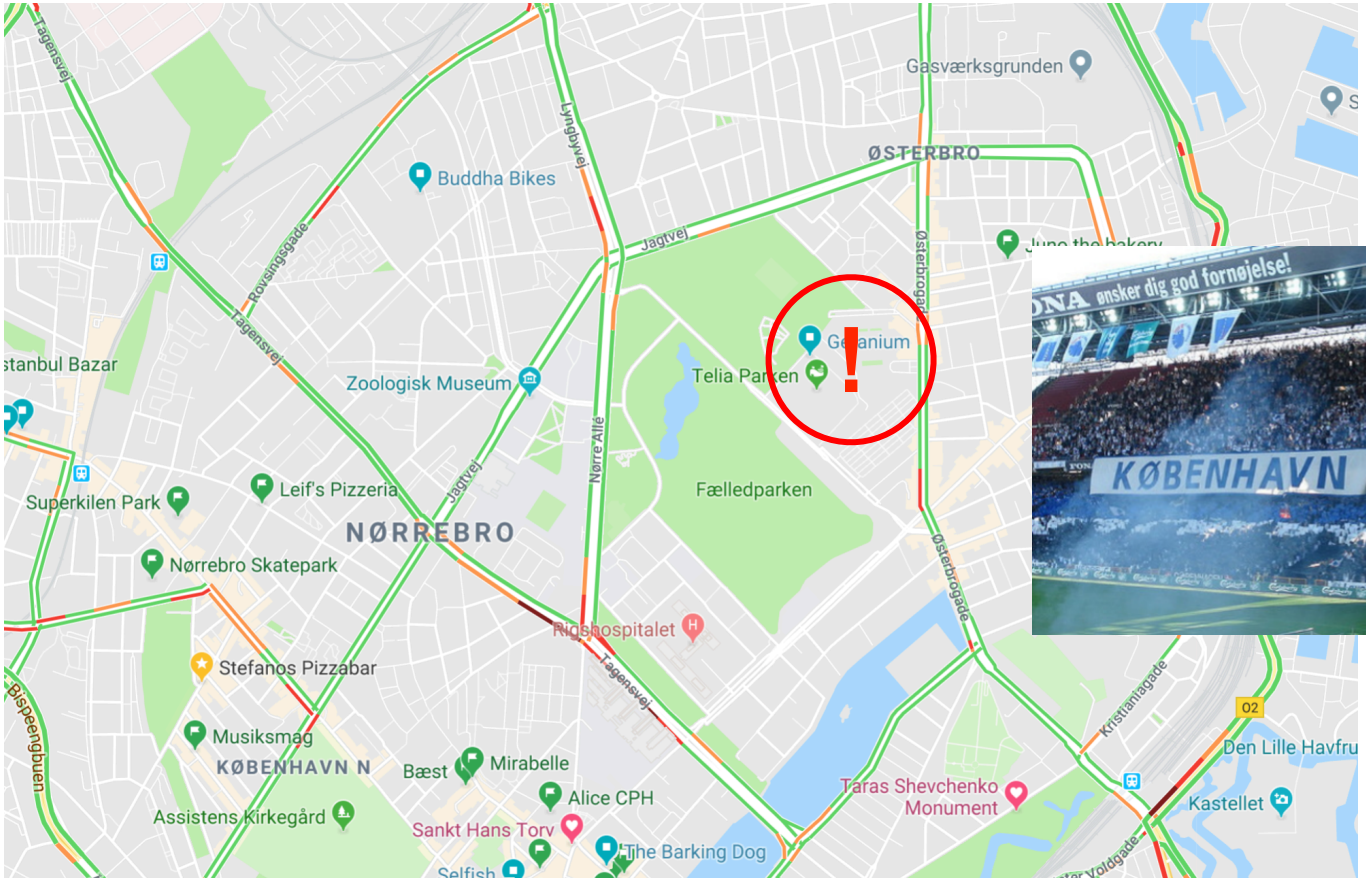
Web Data Mining

Stanislav Borysov [stabo@dtu.dk]

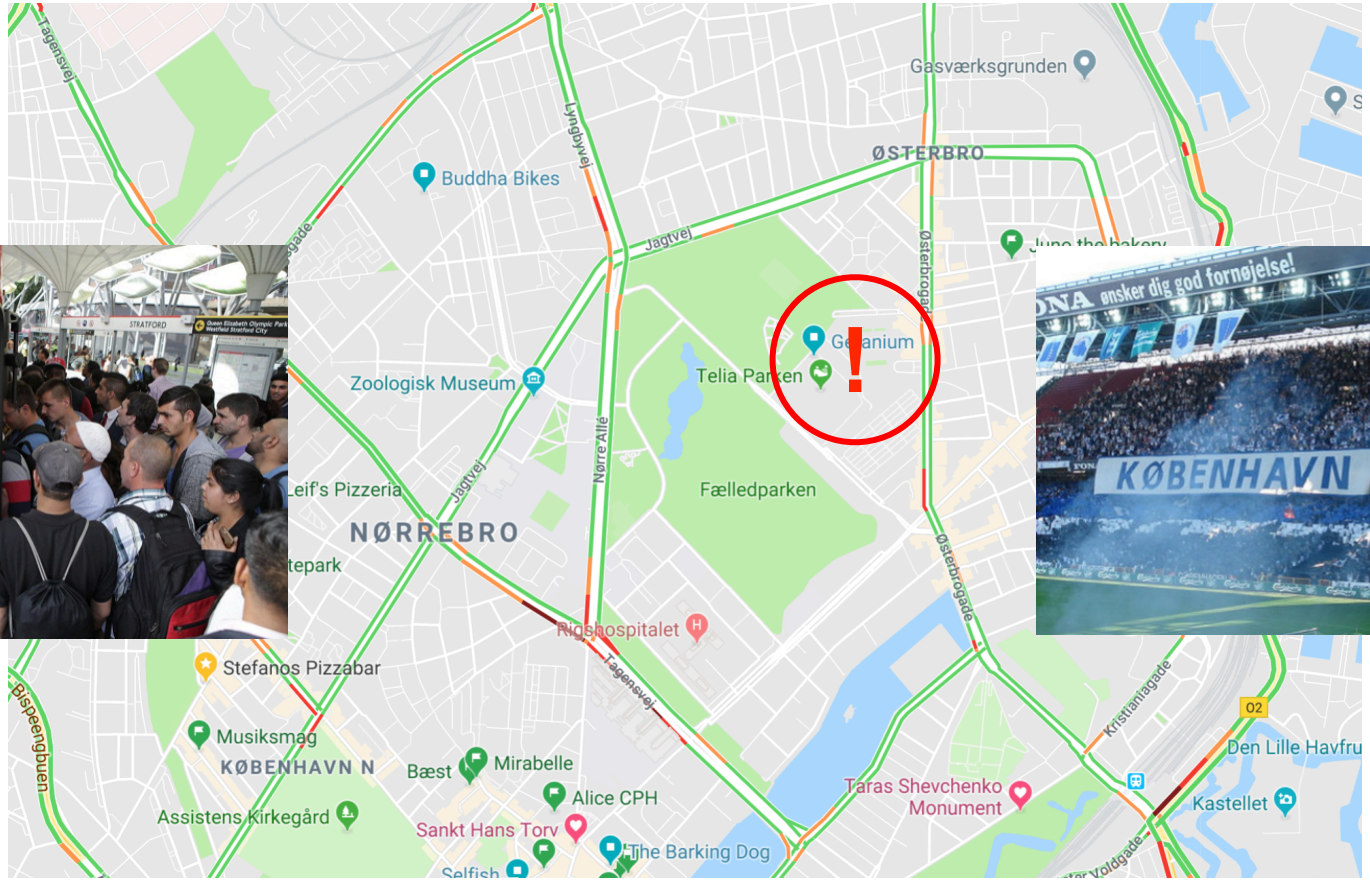
Transport demand problem: Average day



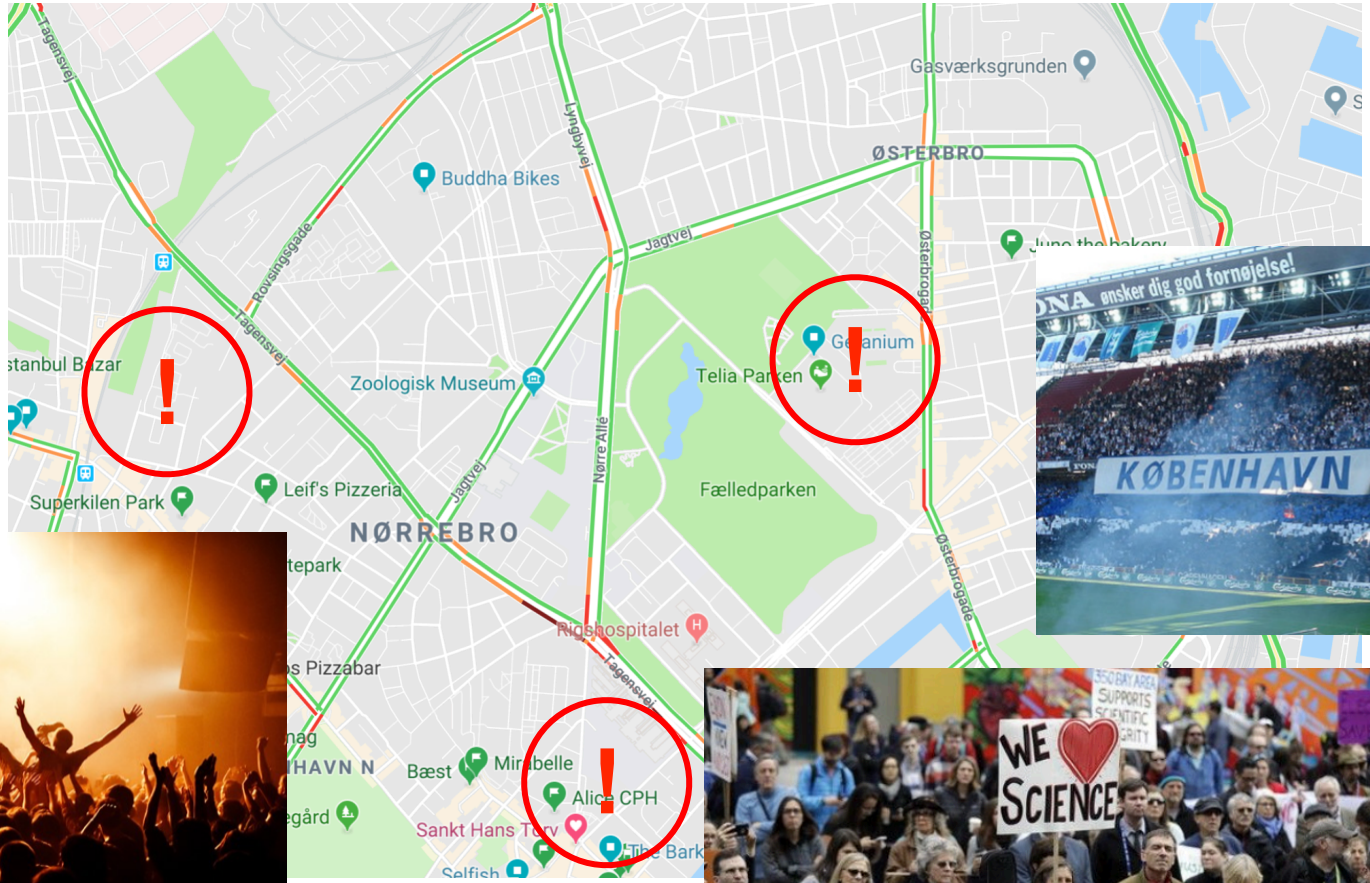
Transport demand problem: Large event



Transport demand problem: Large event



Transport demand problem: A lot of stuff…

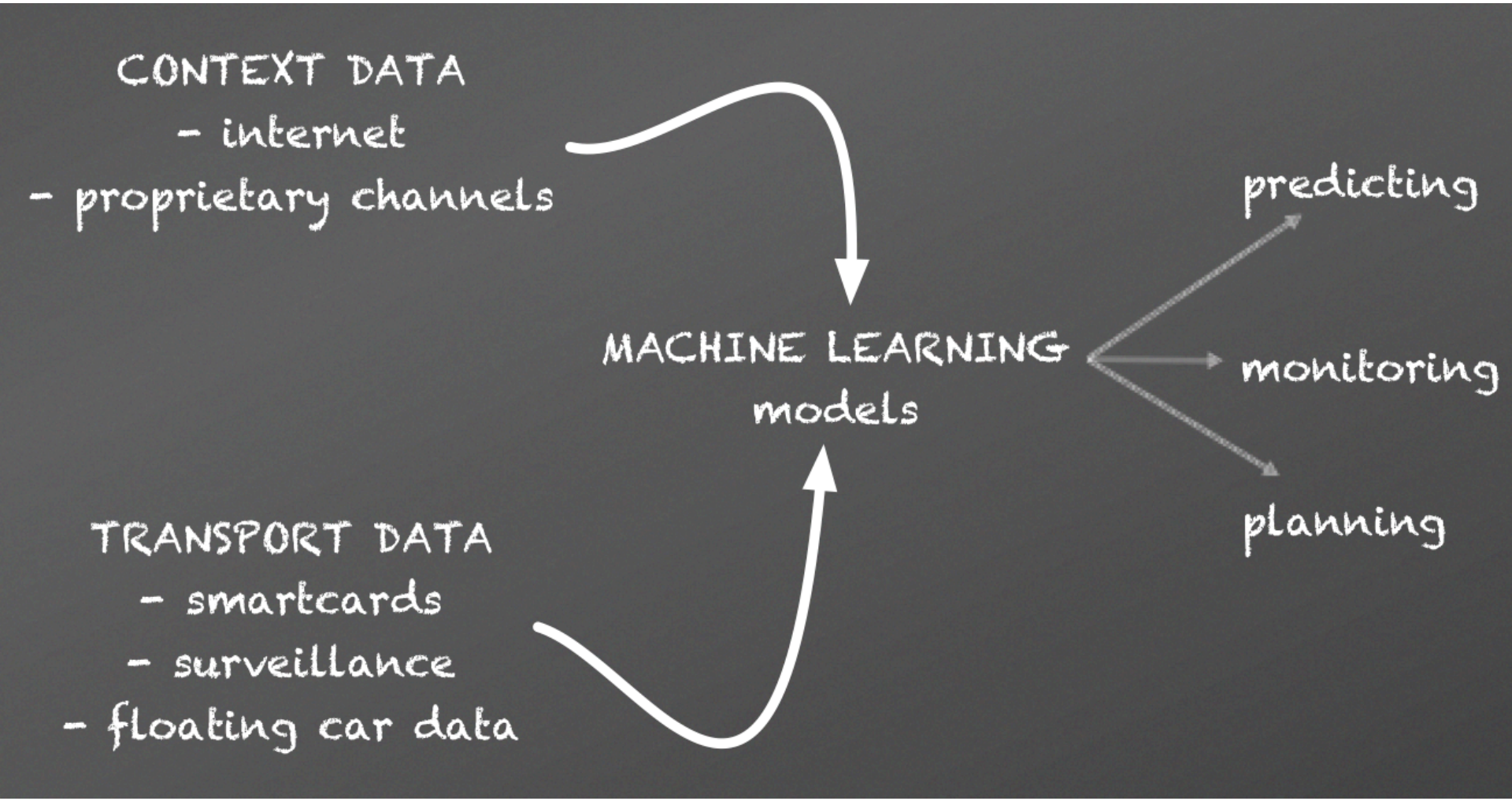


© Daniel Leal-Olivas / i-images
STRATFORD
Queen Elizabeth Olympic Park
Westfield Stratford City

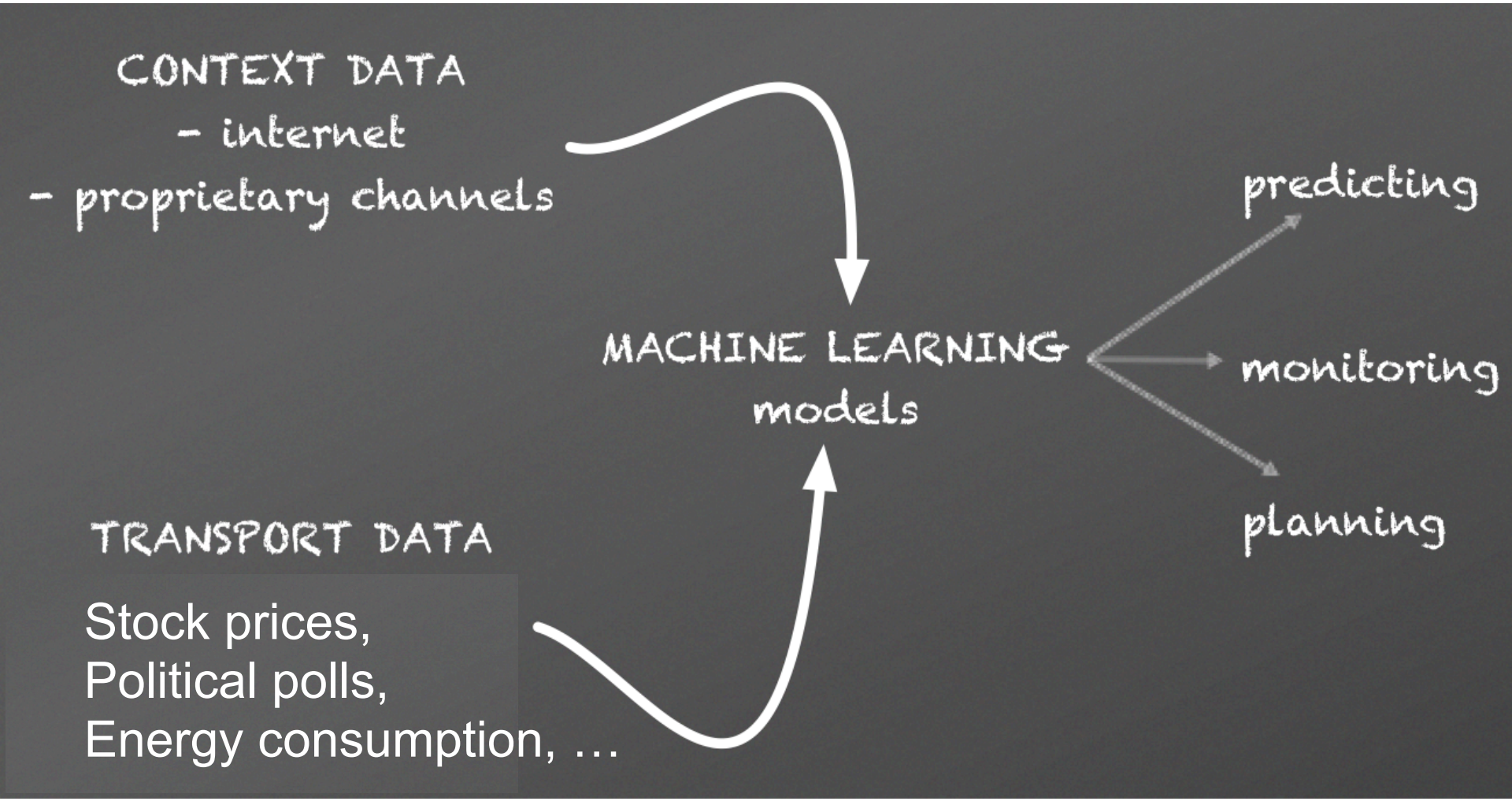
Where to get the information?

- Internet (obviously)
- …

Predictions in Context

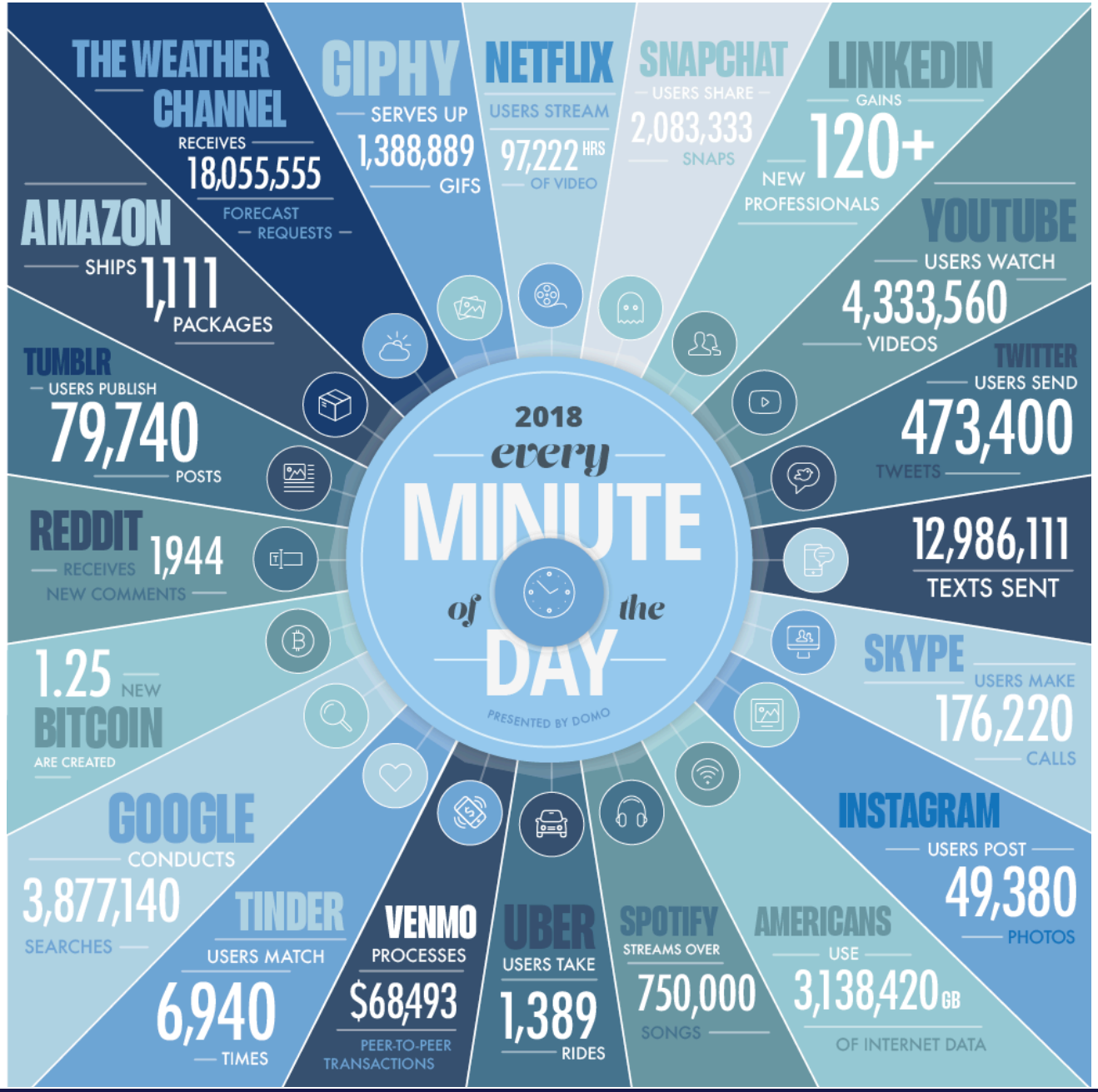


Predictions in Context

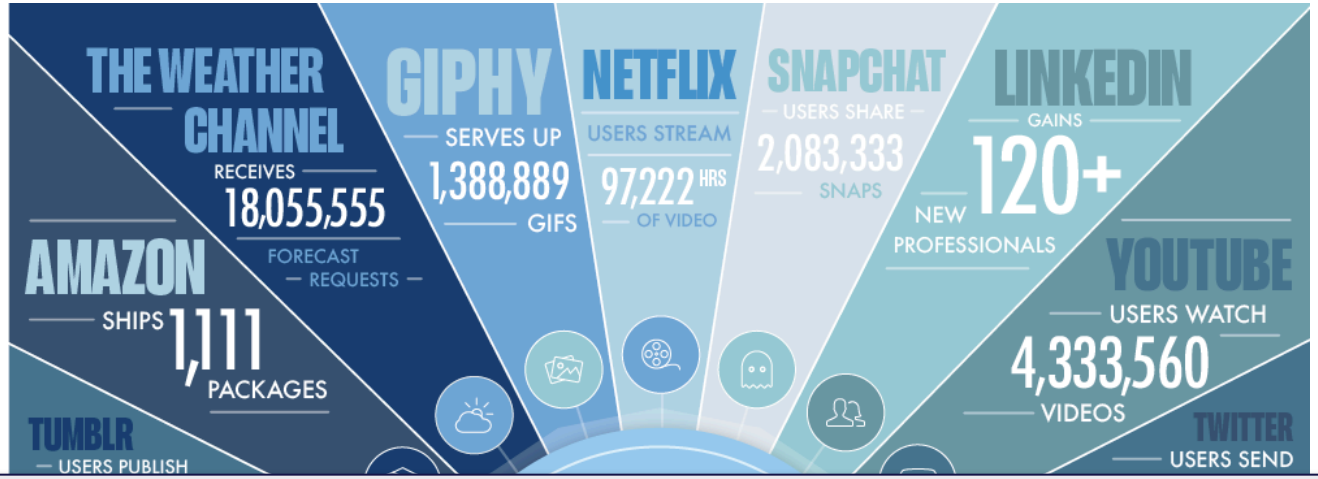


Let’s focus on the Internet

https://www.domo.com/learn/data-never-sleeps-6

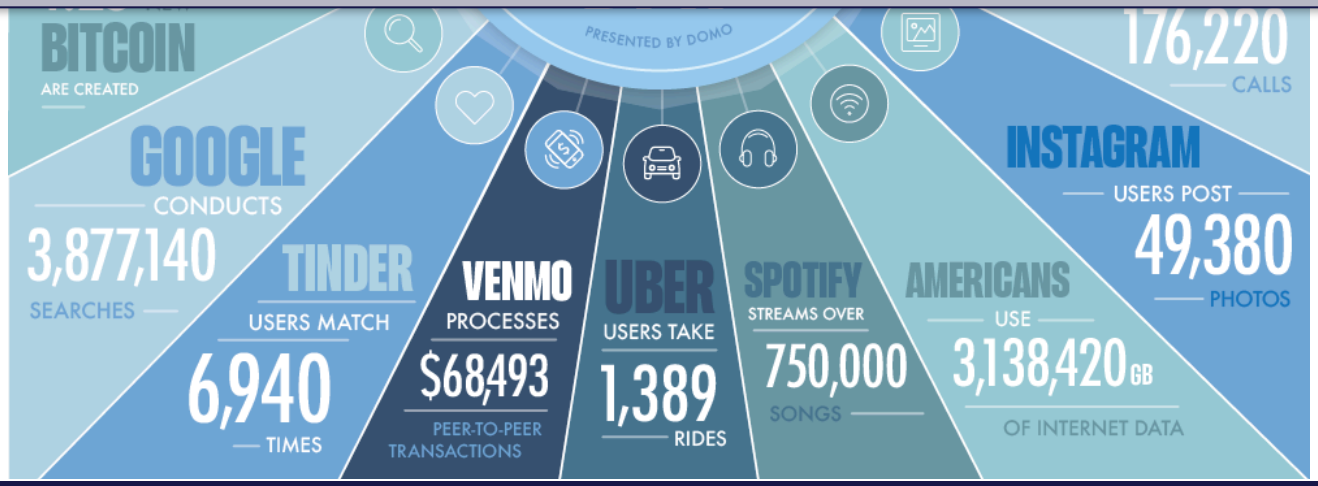


https://www.domo.com/learn/data-never-sleeps-6



A digital footprint of the humanity

(past, present, future,
facts, emotions, interpretations,
expectations, attitudes, plans, …)



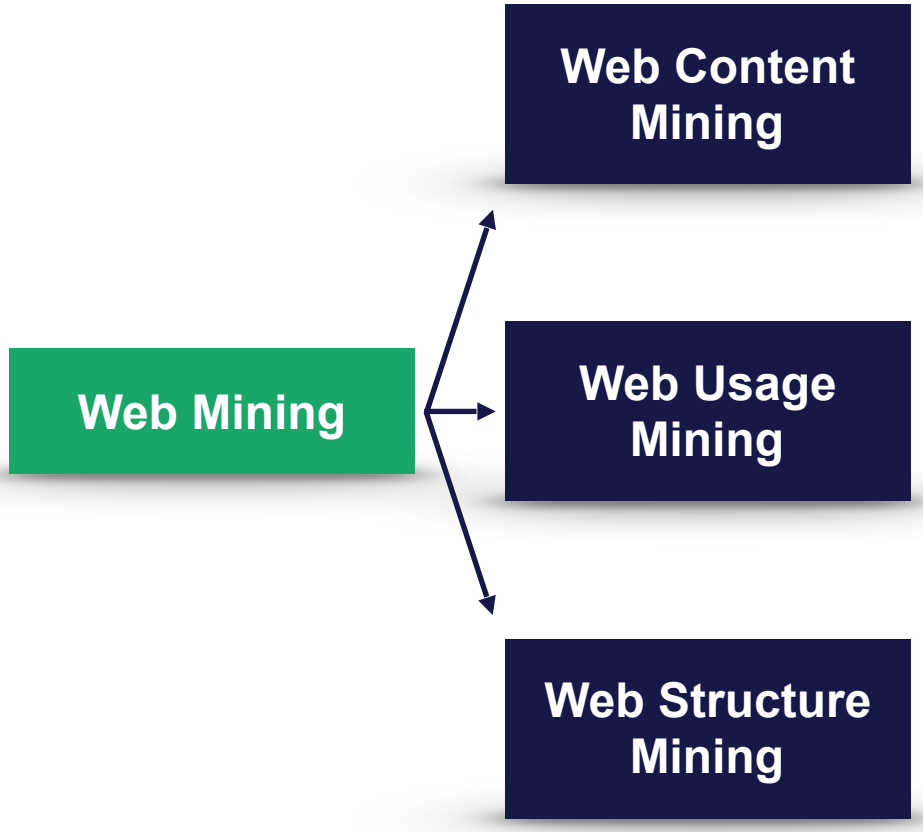
What is Web mining

- The potential of extracting valuable knowledge from the Web is quite evident
- Web mining is the collection of methods to fulfil this potential

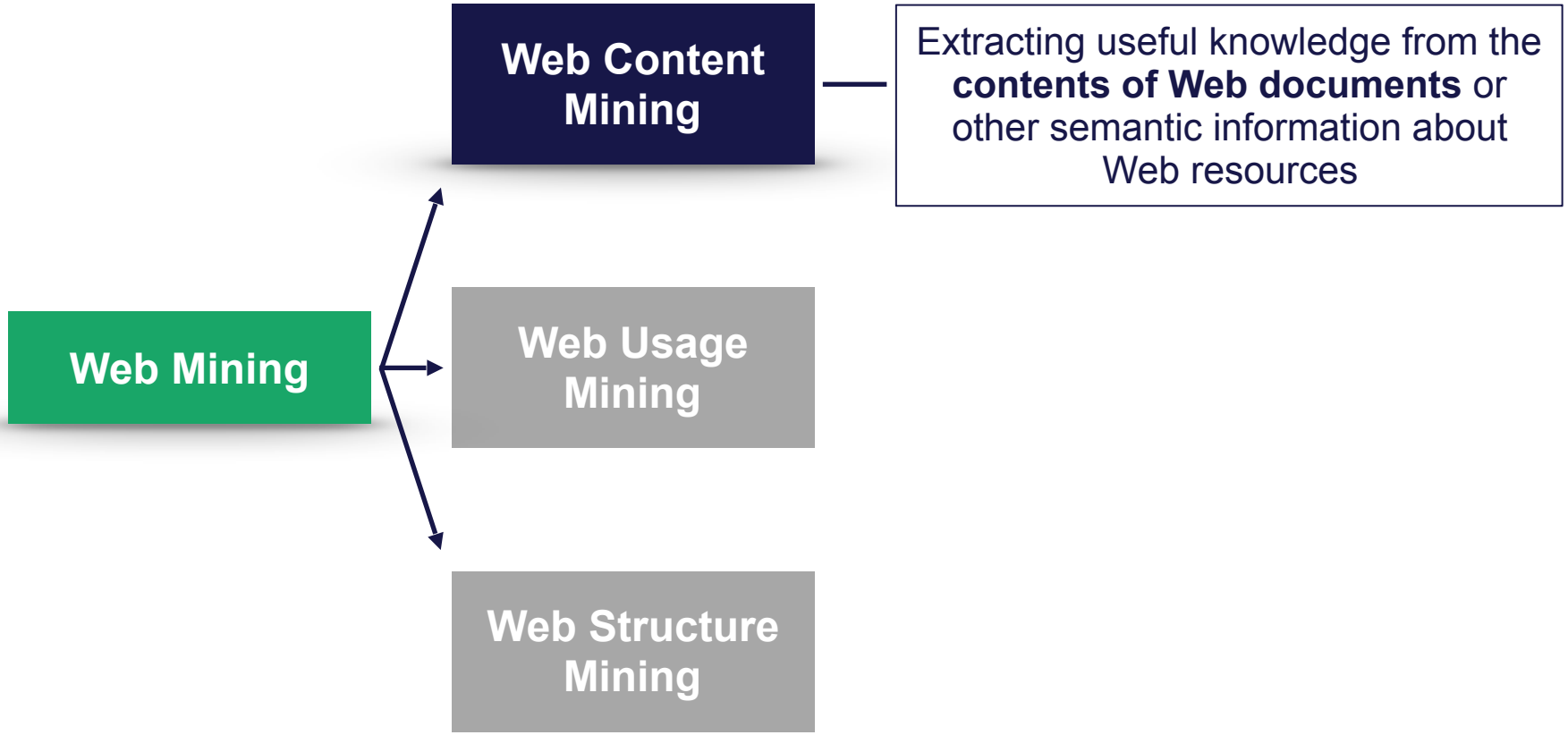
Web mining is application of *data mining* and *machine learning* techniques to extract useful knowledge from the content, structure, and usage of Web resources

- Nowadays, it is more relevant than at any other time during the history of the Web

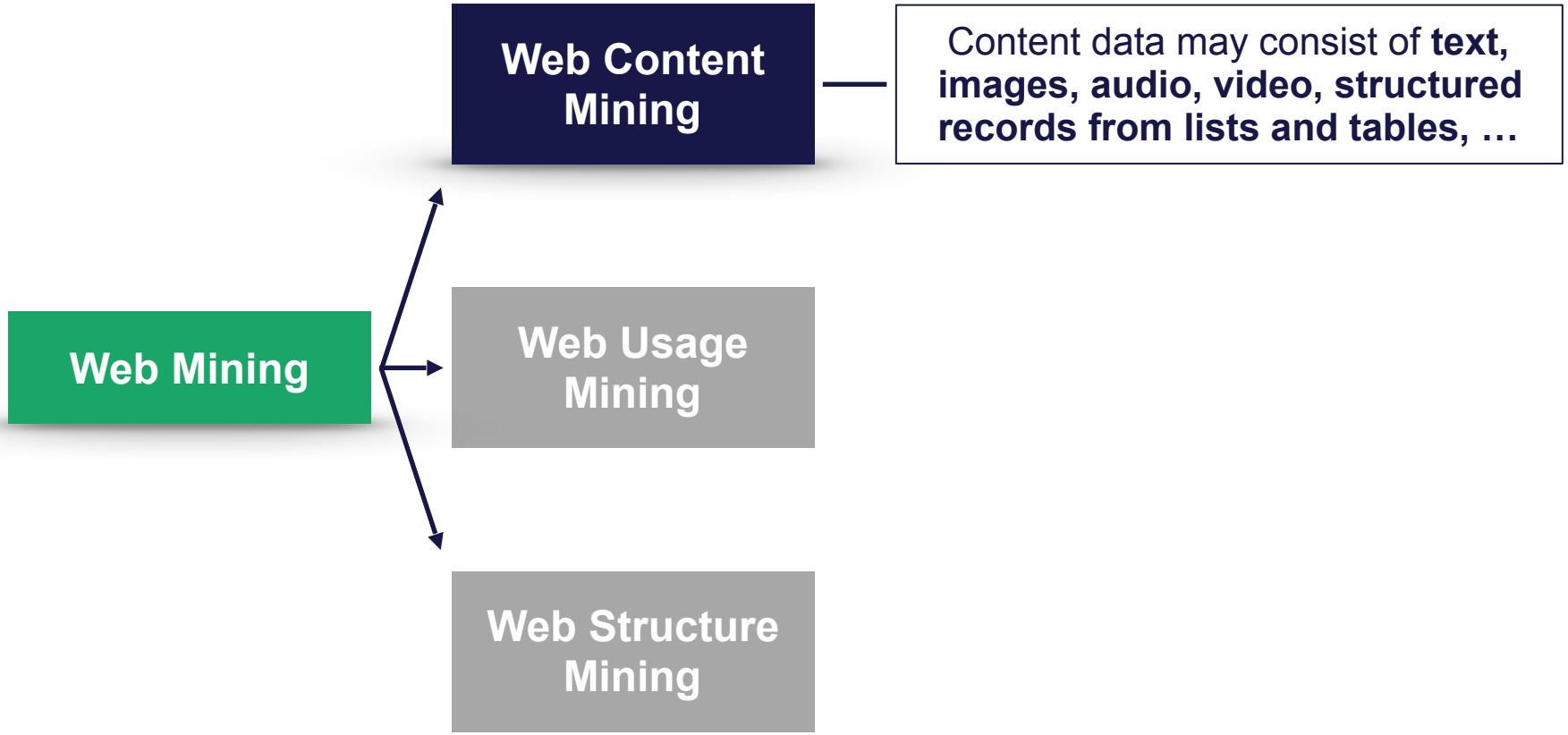
Types of Web Mining



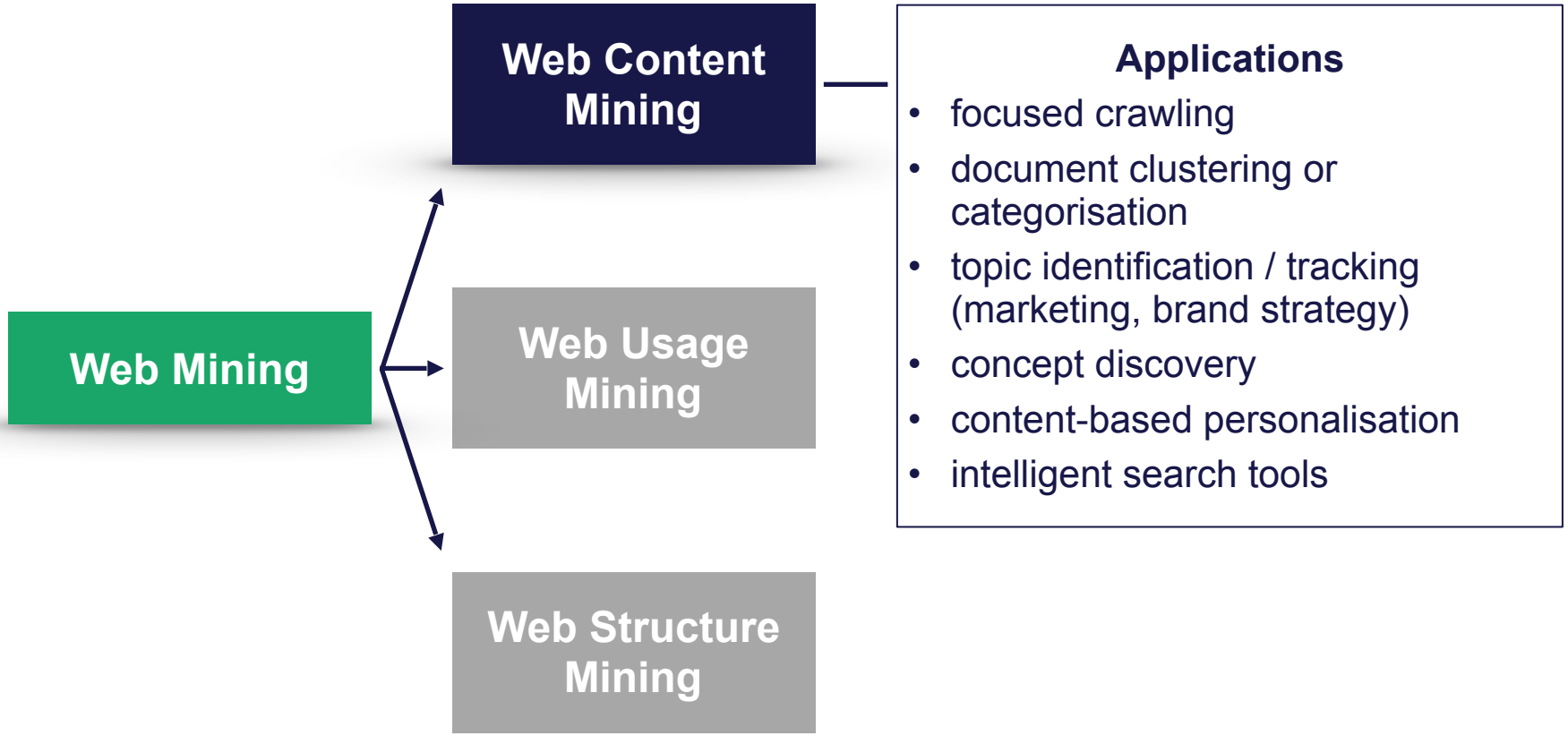
Types of Web Mining



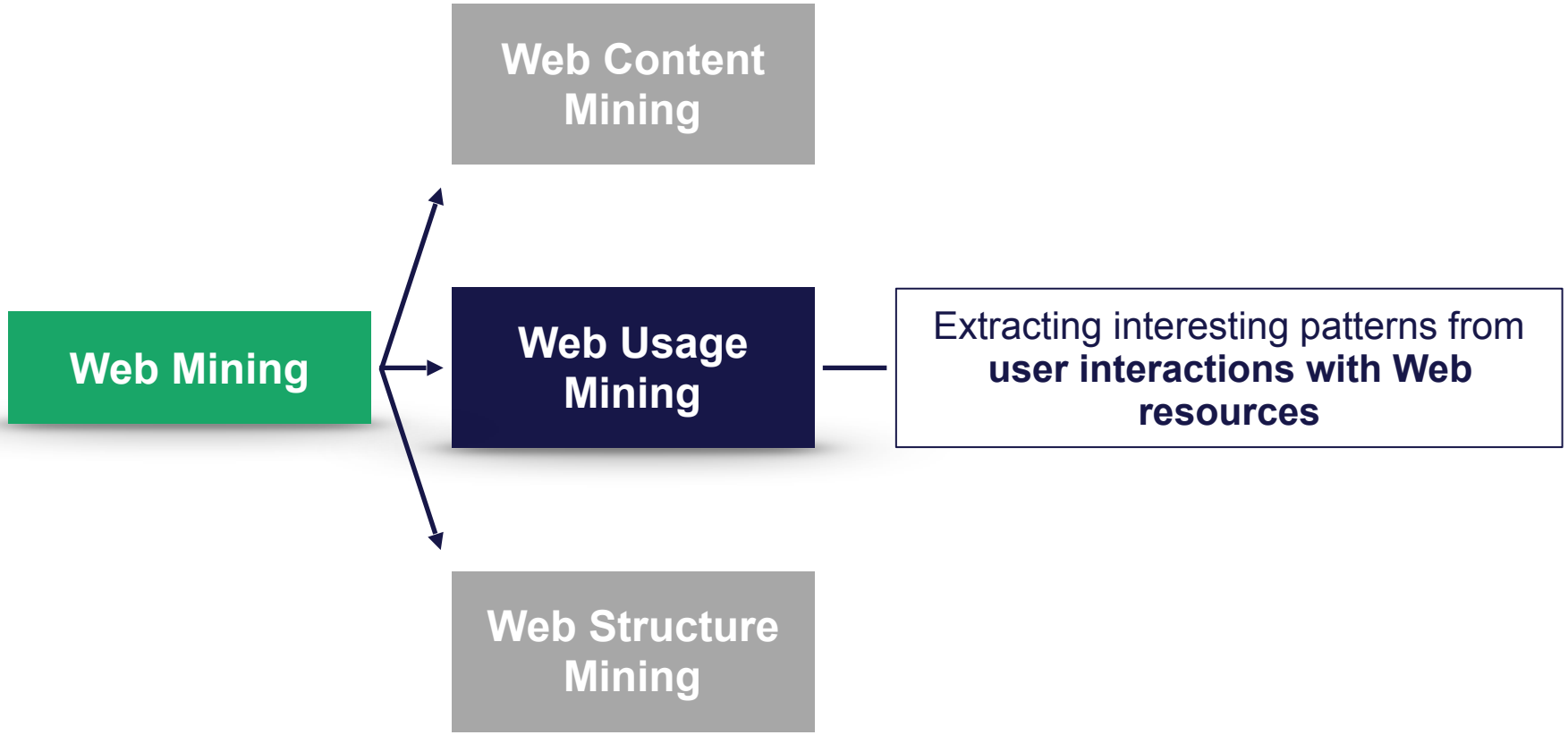
Types of Web Mining



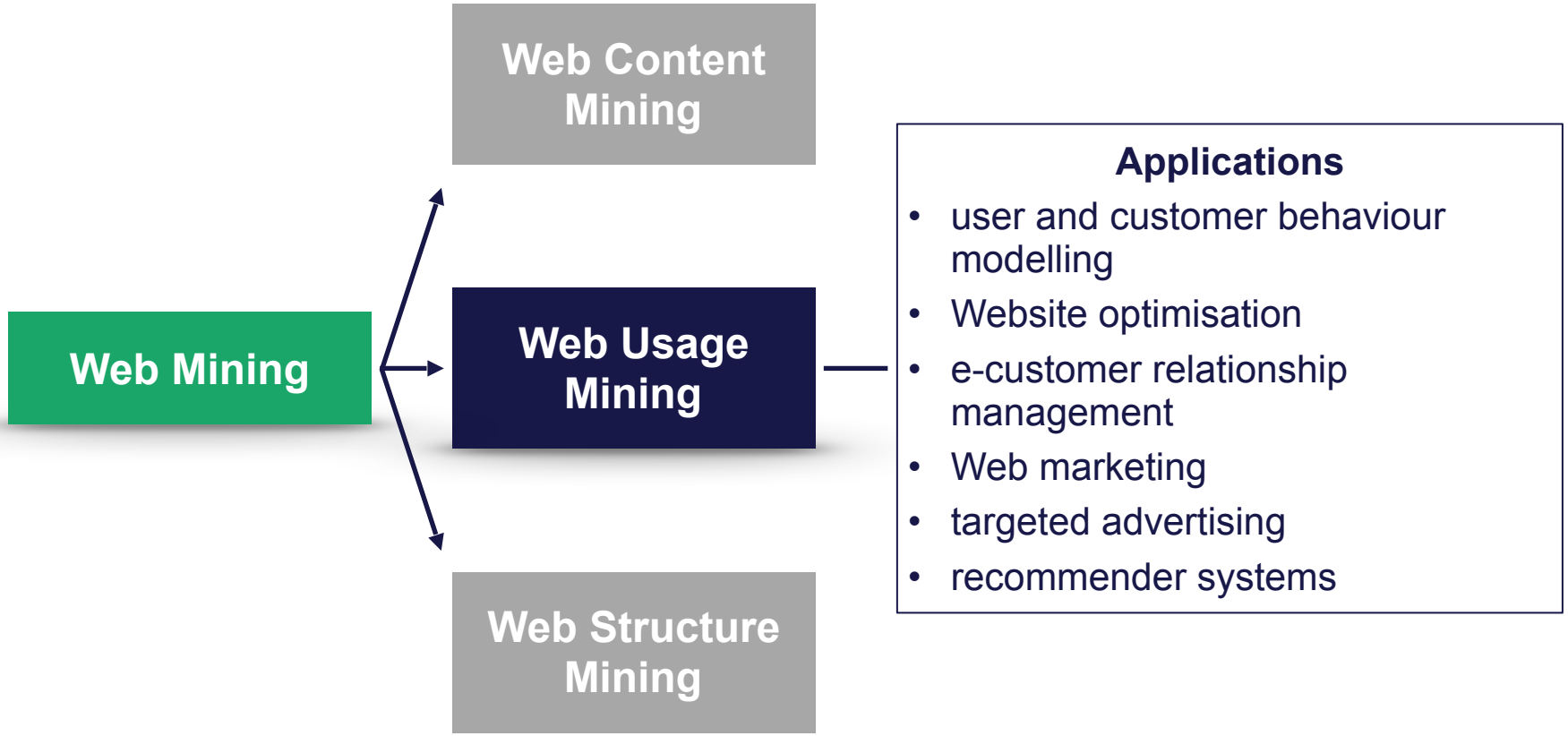
Types of Web Mining



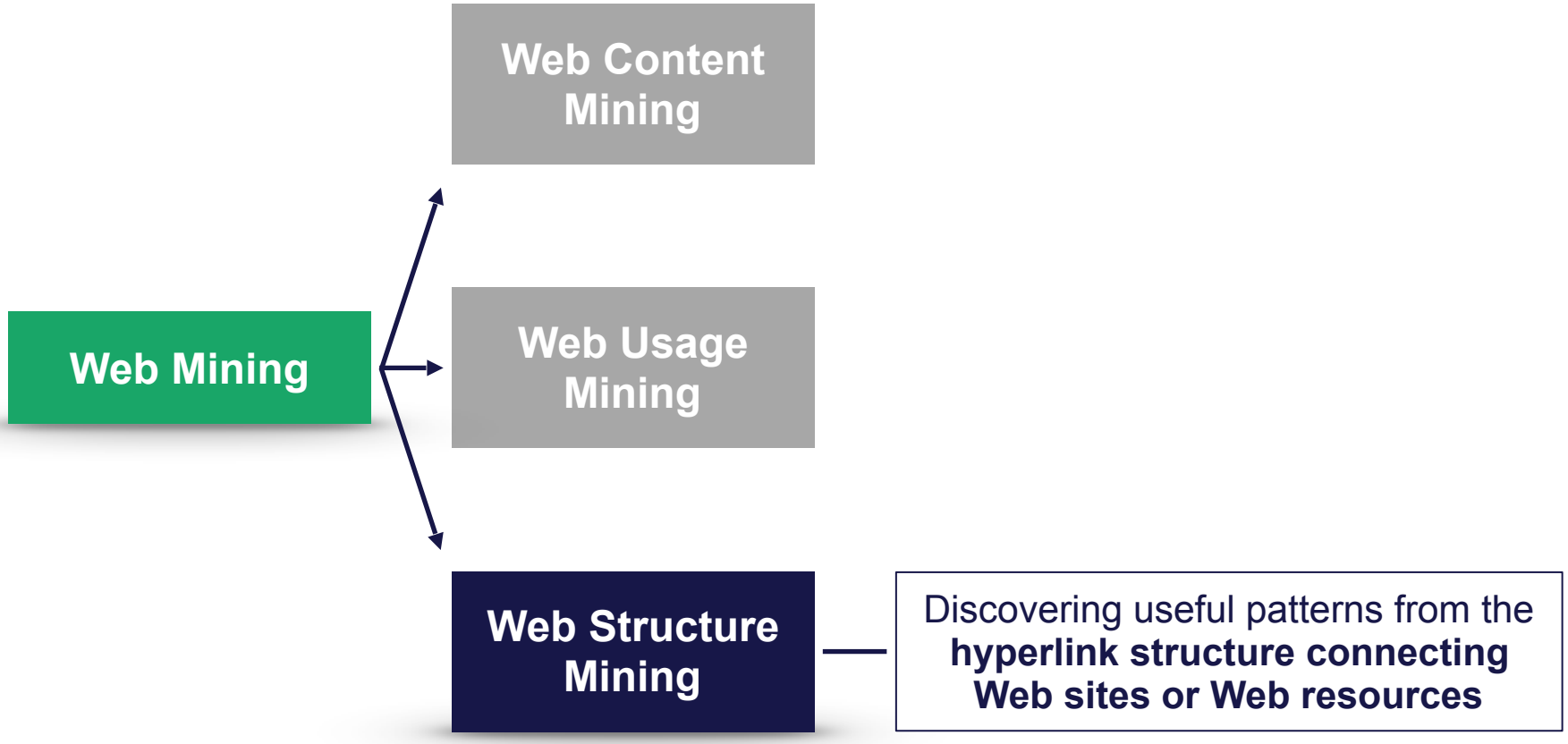
Types of Web Mining



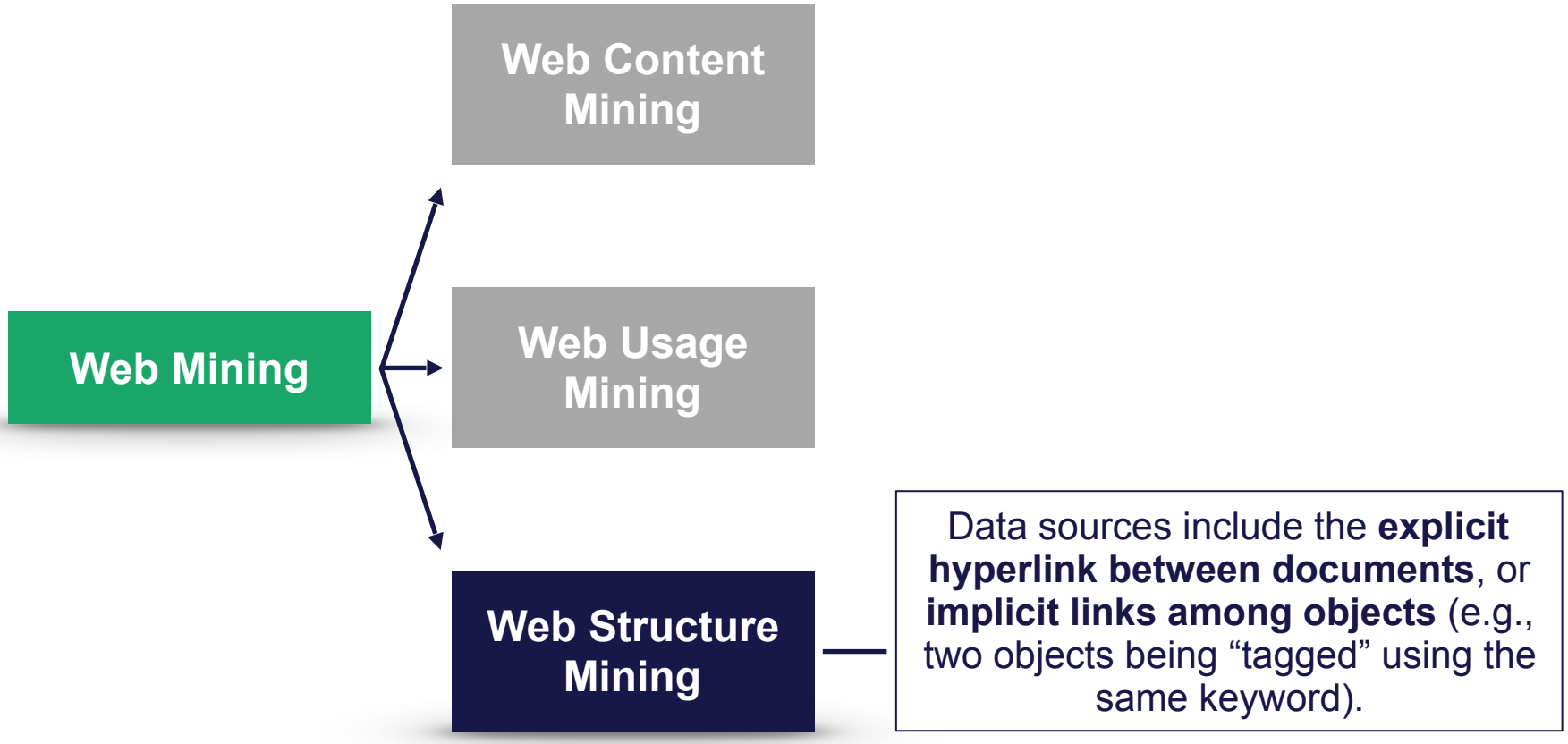
Types of Web Mining



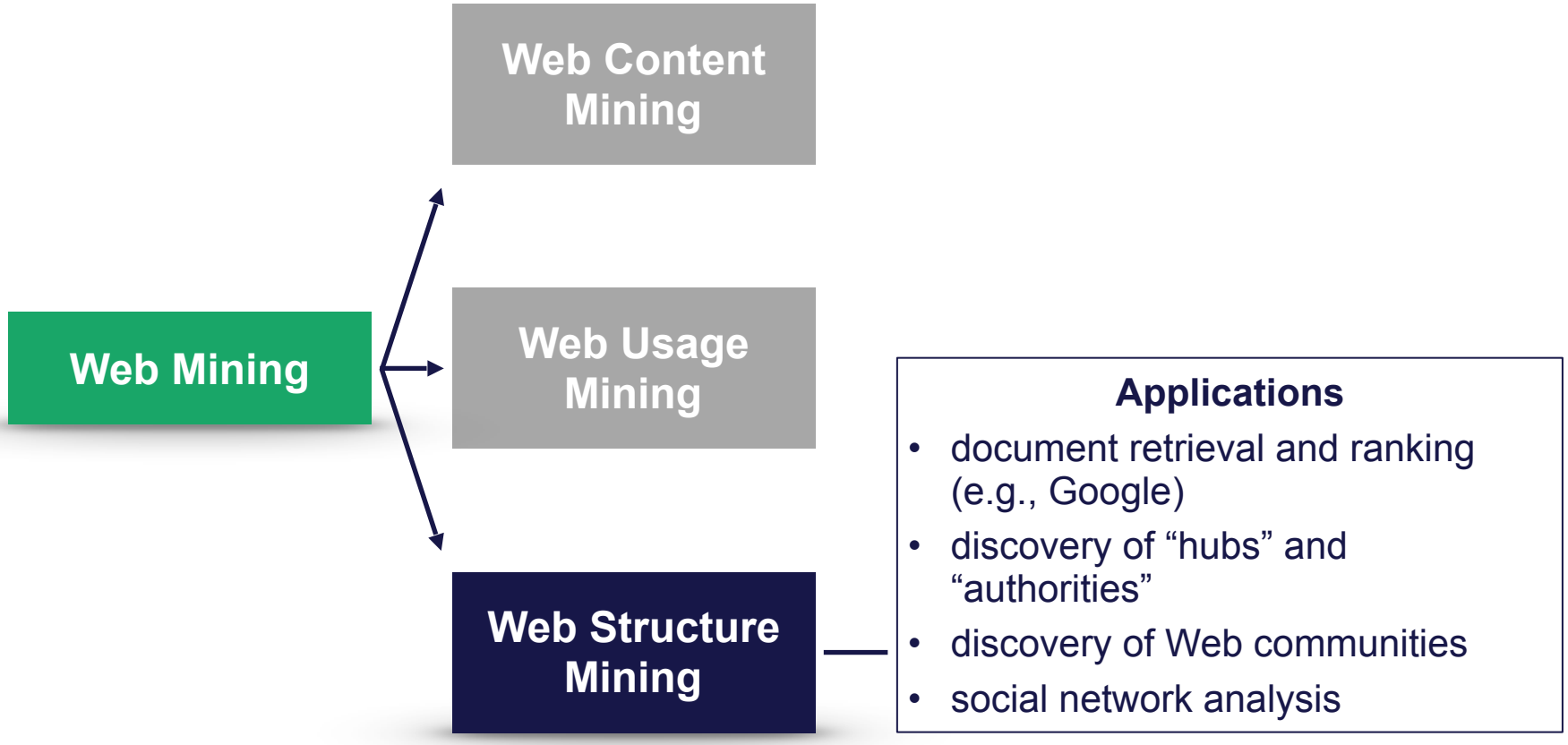
Types of Web Mining



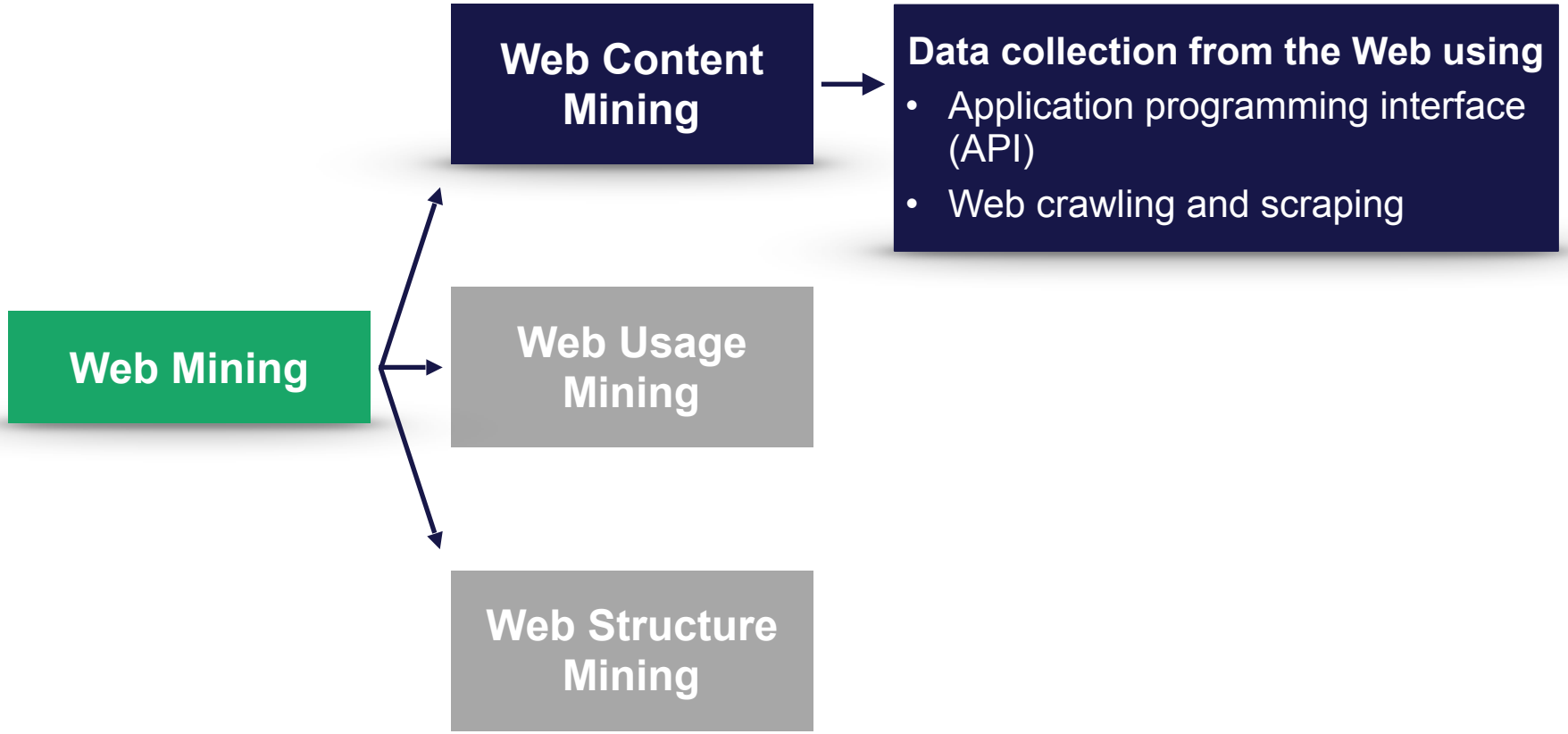
Types of Web Mining



Types of Web Mining



Today's focus: Web content mining



Transport demand problem: Context data

- **Examples of the useful online sources:**
  - ?

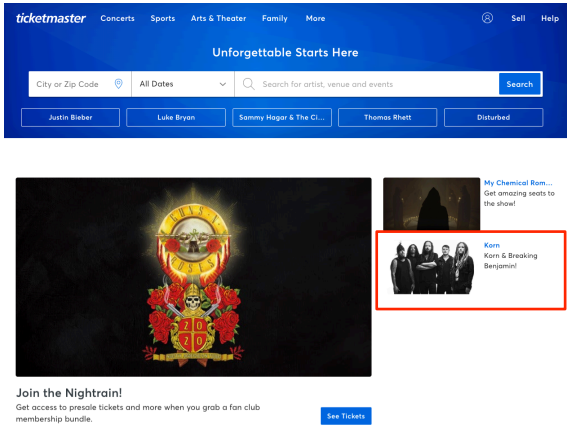
Transport demand problem: Context data

- **Examples of the useful online sources:**
  - Events announcements
  - Social networks
  - News
  - Waze, Google, Inrix
  - Weather
  - …

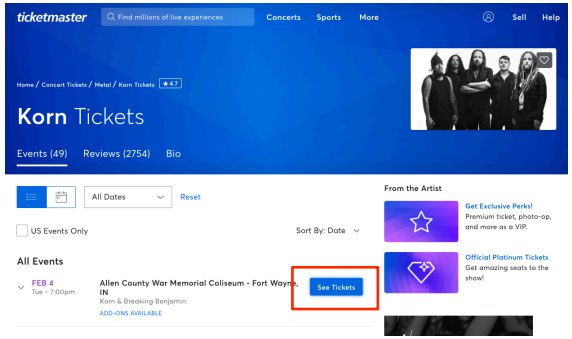
Example: ticketmaster.com

- Go to https://www.ticketmaster.com/
- Click on one of the events listed
- Click “see tickets”
- Click “more info” next to the event title in the top left corner

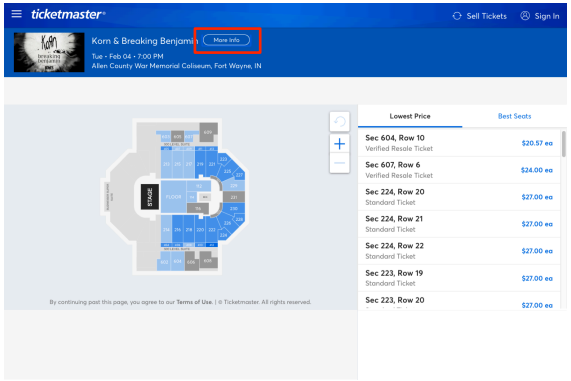
1



2

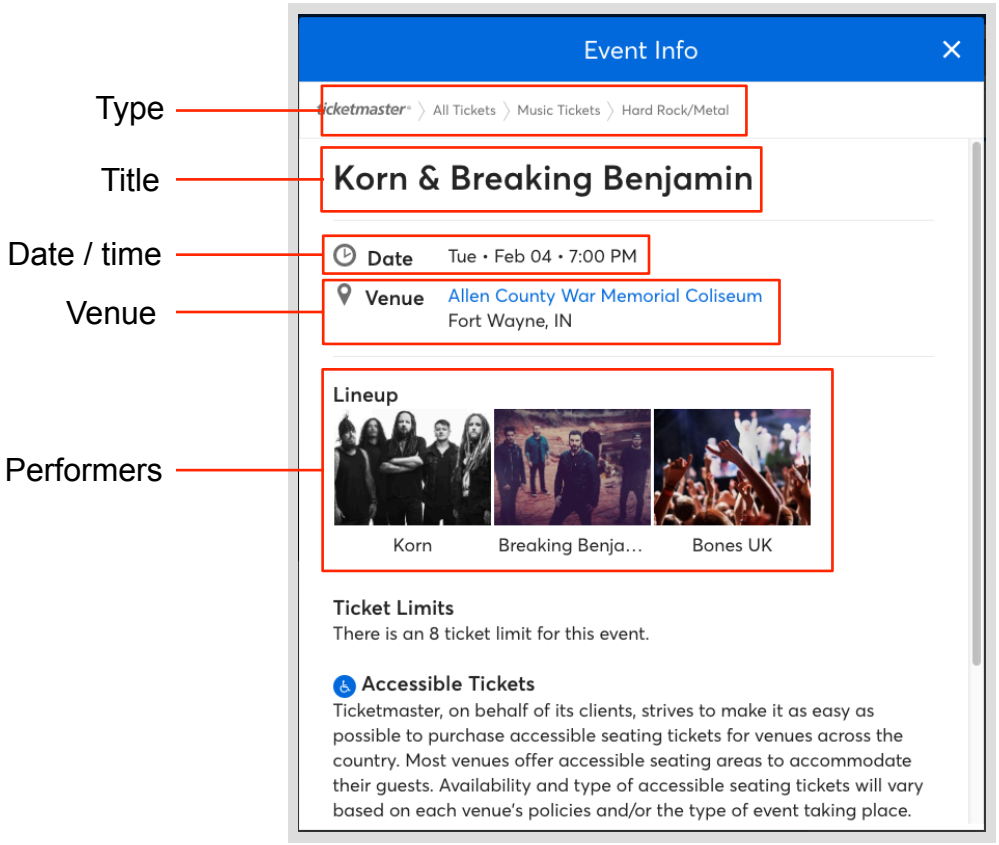


3



Example: ticketmaster.com

- A lot of useful information… How to get (“extract”) it?

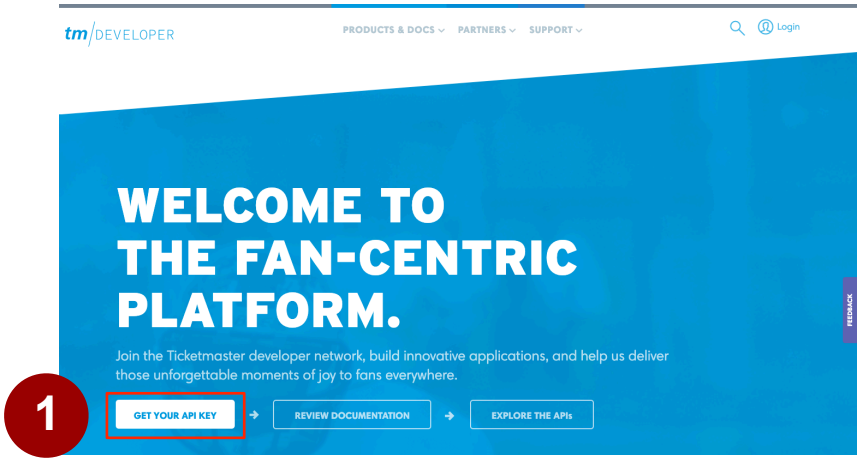


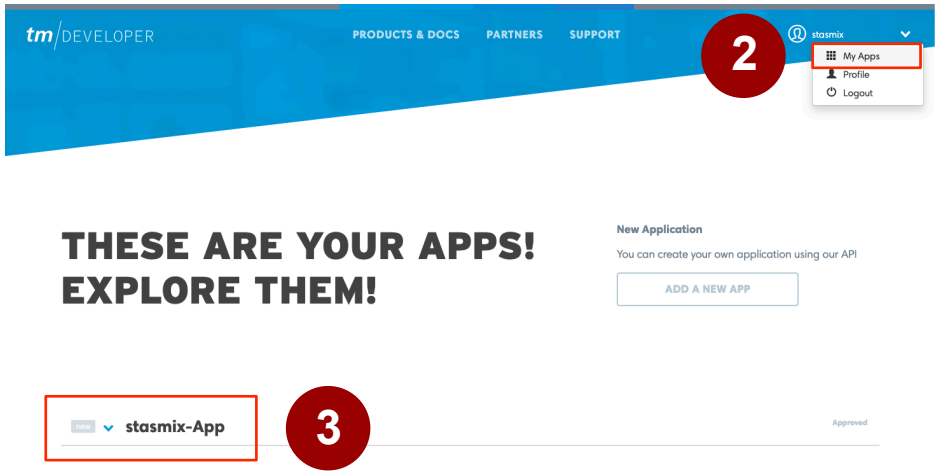
Application programming interface (API)

- The easiest way to get data from the Internet 🙂
- Data is structured or semi-structured
- **Steps:**
    1. Make sure the data source provide APIs for data collection.
    2. Obtain API key or other forms of authorisation.
    3. Read documentation
    4. Coding

API Example: ticketmaster.com

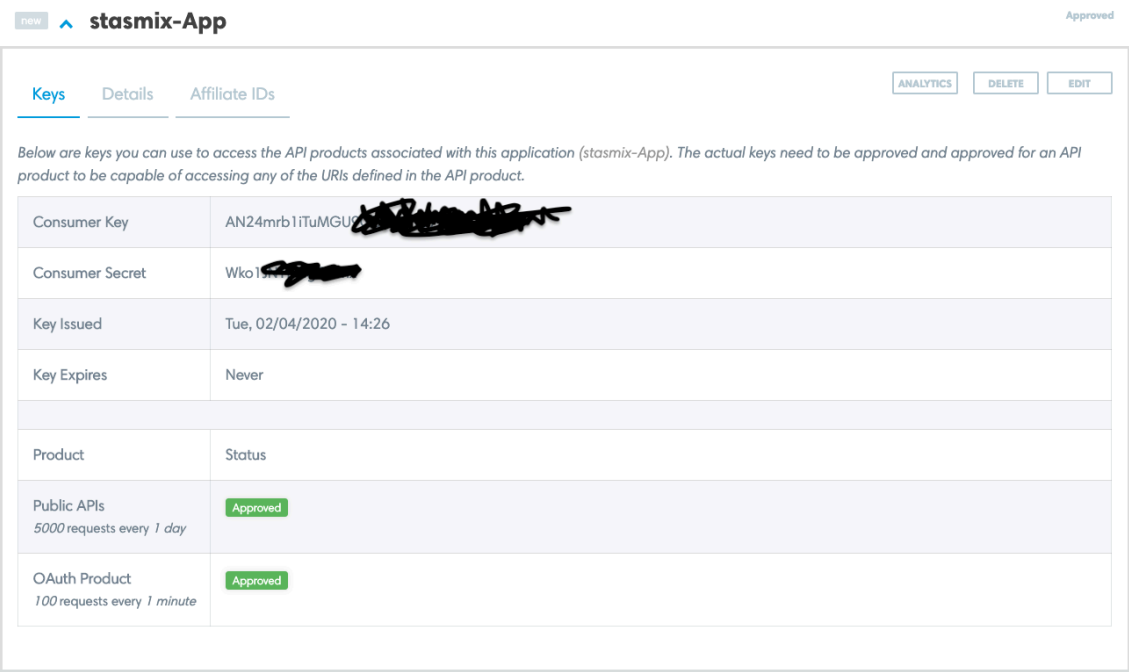
- Go to https://developer.ticketmaster.com/
- Click on “GET YOUR API KEY”
- Sign up
  (create a new email address if you don’t want to use your own one
- Log in, go to “My Apps” and click on your first automatically created app





API Example: ticketmaster.com

- Here you can see your credentials to make API requests and other information such as usage quotas
- **Always keep your credentials private!**



API Example: ticketmaster.com

- Go to the documentation page (we will use **Discovery API**) https://developer.ticketmaster.com/products-and-docs/apis/discovery-api/v2/

- There are many ways to authorise yourself and use API
- We will use the simplest "GET" method which responds to a plain URL
- A GET request URL has the simple structure:

  url_base**?**parameter1=value1**&**parameter2=value2 …

- Scroll down to "Examples" below the map and copy-paste it to a new tab (your API key is automatically appended in the examples)

url_base

Get a list of all events in the United States https://app.ticketmaster.com/discovery/v2/events.json?

countryCode=US&apikey=AN24mr[redacted]N6

parameter1 parameter2 (your API customer secret)

API Example: ticketmaster.com

- Congratulations! You will get a structured response in the JSON format


```
{"_embedded":{"events":[{"name":"T3 ENT Birthday Bash","type":"event","id":"G5eVZpdR-BGMM","test":false,"url":"https://www.ticketmaster.com/t3-ent-birthday-bash-
columbia-south-carolina-03-28-2020/event/2D00583BE1E8CDF7","locale":"en-us","images":[{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-
e35c8d587497_495231_RETINA_PORTRAIT_16_9.jpg","width":640,"height":360,"fallback":false},{"ratio":"4_3","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
a6fe-e35c8d587497_495231_CUSTOM.jpg","width":305,"height":225,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-
e35c8d587497_495231_RETINA_LANDSCAPE_16_9.jpg","width":1136,"height":639,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
a6fe-e35c8d587497_495231_RECOMENDATION_16_9.jpg","width":100,"height":56,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
a6fe-e35c8d587497_495231_TABLET_LANDSCAPE_16_9.jpg","width":1024,"height":576,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-
437f-a6fe-e35c8d587497_495231_EVENT_DETAIL_PAGE_16_9.jpg","width":205,"height":115,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-
f05d-437f-a6fe-e35c8d587497_495231_TABLET_LANDSCAPE_LARGE_16_9.jpg","width":2048,"height":1152,"fallback":false},
{"ratio":"3_2","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-e35c8d587497_495231_ARTIST_PAGE_3_2.jpg","width":305,"height":203,"fallback":false},
{"ratio":"3_2","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-
e35c8d587497_495231_RETINA_PORTRAIT_3_2.jpg","width":640,"height":427,"fallback":false},{"ratio":"3_2","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
a6fe-e35c8d587497_495231_TABLET_LANDSCAPE_3_2.jpg","width":1024,"height":683,"fallback":false}],"sales":{"public":{"startDateTime":"2020-01-
31T15:00:00Z","startTBD":false,"endDateTime":"2020-03-29T01:00:00Z"}},"dates":{"start":{"localDate":"2020-03-28","localTime":"19:00:00","dateTime":"2020-03-
28T23:00:00Z","dateTBD":false,"dateTBA":false,"timeTBA":false,"noSpecificTime":false},"timezone":"America/New_York","status":
{"code":"onsale"},"spanMultipleDays":false},"classifications":[{"primary":true,"segment":{"id":"KZFzniwnSyfZ7v7nJ","name":"Music"},"genre":
{"id":"KnvZfZ7vAv1","name":"Hip-Hop/Rap"},"subGenre":{"id":"KZazBEonSMnZfZ7vaa1","name":"French Rap"},"type":{"id":"KZAyXgnZfZ7v7nI","name":"Undefined"},"subType":
{"id":"KZFzBErXgnZfZ7v7lJ","name":"Undefined"},"family":false}],"promoter":{"id":"494","name":"PROMOTED BY VENUE","description":"PROMOTED BY VENUE / NTL /
USA"},"promoters":[{"id":"494","name":"PROMOTED BY VENUE","description":"PROMOTED BY VENUE / NTL / USA"}],"info":"T3 Entertainment & Hot 103.9 present Birthday Bash
with Youngboy & Special Guests","pleaseNote":"Please Note: Please Note: This Venue Has a Clear Bag Policy. Click for details. Age Limit: Everyone requires a ticket
regardless of age. Event Information & Details on Colonial Life Arena Website","priceRanges":[{"type":"standard","currency":"USD","min":53.0,"max":163.0}],"seatmap":
{"staticUrl":"https://maps.ticketmaster.com/maps/geometry/3/event/2D00583BE1E8CDF7/staticImage?type=png&systemId=HOST"},"_links":{"self":
{"href":"/discovery/v2/events/G5eVZpdR-BGMM?locale=en-us"},"attractions":[{"href":"/discovery/v2/attractions/K8vZ917pkBV?locale=en-us"},
{"href":"/discovery/v2/attractions/K8vZ917b8z7?locale=en-us"},{"href":"/discovery/v2/attractions/K8vZ91714zf?locale=en-us"},
{"href":"/discovery/v2/attractions/K8vZ917bJa0?locale=en-us"}],"venues":[{"href":"/discovery/v2/venues/KovZpZAE7dtA?locale=en-us"}]},"_embedded":{"venues":
[{"name":"Colonial Life Arena","type":"venue","id":"KovZpZAE7dtA","test":false,"url":"https://www.ticketmaster.com/colonial-life-arena-tickets-
columbia/venue/369551","locale":"en-us","aliases":["carolina center","colonial center"],"images":
[{"ratio":"16_9","url":"https://s1.ticketm.net/dbimages/17907v.jpg","width":205,"height":115,"fallback":false},
{"ratio":"4_3","url":"https://s1.ticketm.net/dam/v/668/18c8f001-6426-484a-84be-
937112ef2668_404381_SOURCE.jpg","width":2304,"height":1728,"fallback":false}],"postalCode":"29208","timezone":"America/New_York","city":{"name":"Columbia"},"state":
{"name":"South Carolina","stateCode":"SC"},"country":{"name":"United States Of America","countryCode":"US"},"address":{"line1":"801 Lincoln St."},"location":
{"longitude":"-81.036104","latitude":"33.9942619"},"markets":[{"name":"South Carolina","id":"121"}],"dmas":[{"id":243},{"id":256},{"id":278},{"id":291},
{"id":292}],"social":{"twitter":{"handle":"@CLAmktg"}},"boxOfficeInfo":{"phoneNumberDetail":"Colonial Life Arena: (803)576-9200 Gamecock Athletics Box Office:
(803)777-4274","openHoursDetail":"Monday-Friday: 9 AM to 5 PM. Weekends: Open for events only and hours will vary by that event. For most accurate information please
visit the event page.","acceptedPaymentDetail":"Visa, Mastercard, Discover, American Express, Cash and Carolina Card (only for non-USC
Athletics).","willCallDetail":"Will call tickets will only be given out if the customer provides the credit card used to purchase and an ID that matches the name on
the account. If the original card used to purchase cannot be provided, we will need to see a bank/credit card statement with the purchase line on
it.},"parkingDetail":"There are parking lots available within walking distance of Colonial Life Arena. Accessible parking is available in all of these lots. In order
to park in the spaces reserved for disabled guests, the vehicle must have a state issued handicapped window pass, sticker, or license plate. Any vehicle may drop off
a disabled patron near a Colonial Life Arena entrance and park their car in one of the lots. A parking fee will be charged to every patron who parks in any of the
lots. The charge varies by show. The lots are CASH ONLY. Shuttles For some events, free parking is available in the Pendleton Street Garage on the USC campus and a
shuttle will run beginning two hours prior to event start and will run for one hour following the conclusion of the event. The round-trip shuttle ride costs $3 cash
per person. Check individual event pages on www.ColonialLifeArena.com for specific information and availability or call (803)576-
9200.","accessibleSeatingDetail":"Accessible seating is available at various locations throughout the Colonial Life Arena. Handicap seating is available to the patron
with the disability and all guests in the party, up to published ticket limits, if available. To purchase handicap accessible seating, please visit the box office
during business hours, call Ticketmaster at 800-745-3000 or go online at www.Ticketmaster.com We reserve the right to re-seat or remove any guests from the property
that violate the Federal ADA laws. If you purchased seats in a designated ADA seat location in error, please advise Ticketmaster so they can try to relocate your
seats. This is best to be completed before the show but can also be handled at the box office on the day of the show.","generalInfo":{"generalRule":"For Drop-Offs —
There is often a drop off area at the corner of Lincoln and College streets. This may vary depending on traffic flow for each event but this is usually the case.
Colonial Life Arena Policy: All persons and items entering Colonial Life Arena are subject to inspection. Please visit www.ColonialLifeArena.com to see a full list of
prohibited items. Please note that these can vary by show and change without notice. There is no re-entry. Alcohol sales: Beer, wine and soft drinks are all available
at specific events subject to crowd management policies and regulations. Patrons are required to have a valid driver's license from any state, military ID, passport,
or valid government issued ID.","childRule":"The age for requiring a ticket varies per show. Please visit the individual event page for more information that will be
more accurate for each show. If a child is going to sit in a seat, they will need a ticket regardless of age. If the child is going to sit in the parents lap they
must still adhere to the age policy for purchasing a ticket. For example: If the show requires a ticket for children 2 and older but a 3 year old is going to sit in
the parents lap, the child will require a ticket."},"upcomingEvents":{"_total":43,"ticketmaster":43},"_links":{"self":{"href":"/discovery/v2/venues/KovZpZAE7dtA?
locale=en-us"}}}},"attractions":[{"name":"Youngboy Never Broke
Again","type":"attraction","id":"K8vZ917pkBV","test":false,"url":"https://www.ticketmaster.com/youngboy-never-broke-again-tickets/artist/2385776","locale":"en-
us","externalLinks":{"twitter":[{"url":"https://twitter.com/ggyoungboy"}],"musicbrainz":[{"id":"127d8614-0876-4408-886c-6f110b047a78"}],"homepage":
[{"url":"http://youngboynba.com/"}]},"aliases":["NBA Youngboy"],"images":[{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-
e35c8d587497_495231_RETINA_PORTRAIT_16_9.jpg","width":640,"height":360,"fallback":false},{"ratio":"4_3","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
a6fe-e35c8d587497_495231_CUSTOM.jpg","width":305,"height":225,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-a6fe-
e35c8d587497_495231_RETINA_LANDSCAPE_16_9.jpg","width":1136,"height":639,"fallback":false},{"ratio":"16_9","url":"https://s1.ticketm.net/dam/a/497/bdbfdc11-f05d-437f-
```


- To make it more human readable, you can either use JSON formatter such as https://jsonformatter.curiousconcept.com/ or directly use Python (next slide)

API Example: ticketmaster.com

In [1]:
```
import requests

api_key = "AN24[redacted]gN6"
# It's a very bad programming practice
# to put passwords or secret keys into the code…
url_base = "https://app.ticketmaster.com/discovery/v2/events.json"
params = ["countryCode=US", "apikey="+api_key]

request_url = url_base + "?" + "&".join(params)

response = requests.get(request_url)
data = response.json()
data
```

Out[1]:
```
{'_embedded': {'events': [{'_embedded': {'attractions': [{'_links': {'self':
{'href': '/discovery/v2/attractions/K8vZ917pkBV?locale=en-us'}},
       'aliases': ['NBA Youngboy'],
       'classifications': [{'family': False,
         'genre': {'id': 'KnvZfZ7vAv1', 'name': 'Hip-Hop/Rap'},
         'primary': True,
         'segment': {'id': 'KZFzniwnSyZfZ7v7nJ', 'name': 'Music'},
         'subGenre': {'id': 'KZazBEonSMnZfZ7vkvl', 'name': 'Hip-Hop/Rap'},
         'subType': {'id': 'KZFzBErXgnZfZ7vAd7', 'name': 'Musician'},
         'type': {'id': 'KZAyXgnZfZ7v7la', 'name': 'Individual'}}],
       'externalLinks': {'homepage': [{'url': 'http://youngboynba.com/'}],
        'musicbrainz': [{'id': '127d8614-0876-4408-886c-6f110b047a78'}],
        'twitter': [{'url': 'https://twitter.com/ggyoungboy'}]},
       'id': 'K8vZ917pkBV',
       'images': [{'fallback': False,
         'height': 360,
         'ratio': '16_9',

                                   ...

      'venues': [{'_links': {'self': {'href': '/discovery/v2/venues/KovZpZAE7d
  tA?locale=en-us'}},

                                   ...

              'address': {'line1': '801 Lincoln St.'},
              'aliases': ['carolina center', 'colonial center'],

                                   ...
```

JSON format

One of the most popular web data formats

```
{
   “var_1”: “string_val”,
   “var_2”: 1.56,
   “var_3”: true,
   “var_4”: [1, 2, 3],
   “var_5”: {
      “var_6”: [”a”, “b”, “c”]
   }
}
```

Can be parsed to a Python dictionary using json library

```
import json
```

APIs: Summary

- parameters
- authorization

- JSON
- XML
- …

APIs: Problems

- No API is provided
- Free access is limited
- Response does not contain all the information

API Problems Example: eventbrite.com

- Last semester, we used eventbrite.com API as an example
- However, when I checked it again last week, I found that the event search API was shut down on Thursday, December 12, 2019.
- There are still some API functionality left but maybe it is not enough for our purposes.
- Also, maybe eventbrite.com has more events and provides better event description, so we would like to use it instead of ticketmaster.com

- **What to do?**

API Problems Example: eventbrite.com

- Last semester, we used eventbrite.com API as an example
- However, when I checked it again last week, I found that the event search API was shut down on Thursday, December 12, 2019.
- There are still some API functionality left but maybe it is not enough for our purposes.
- Also, maybe eventbrite.com has more events and provides better event description, so we would like to use it instead of ticketmaster.com

- **What to do?**

Generic web crawler / scraper!

Web scraping: Overall idea

- After browsing the eventbrite.com website, you find out that each event's information can be found at https://www.eventbrite.dk/e/event_id where each event has its own unique id *event_id*

- **Pseudo code:**

Web crawling

```
get a list of event ids {ids}, for example, from the main page eventbrite.com
for each event_id in {ids}
    access https://www.eventbrite.dk/e/event_id, store page content in C
    extract event name n, date d, time t, etc from C
    store (event_id, n, d, t) in database
```

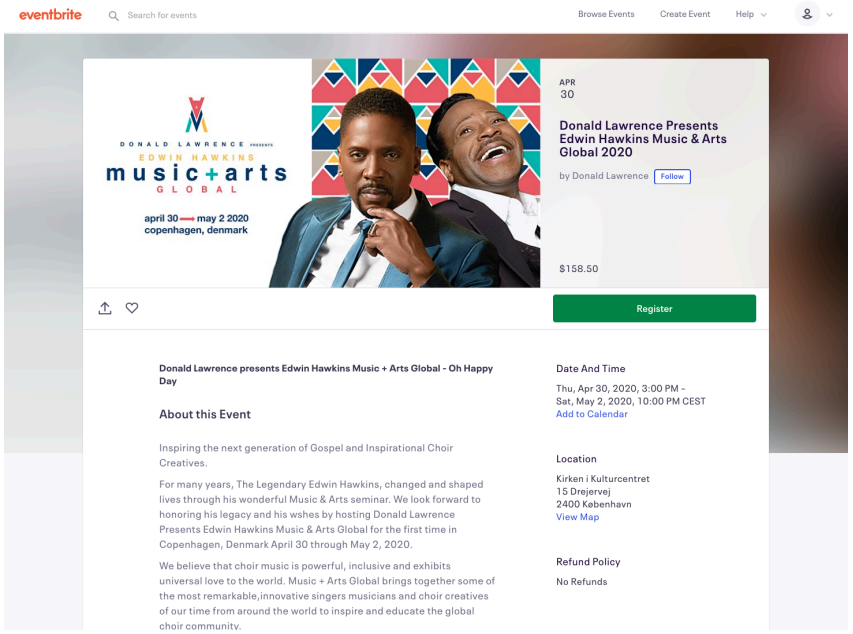
Web scraping

Example: Eventbrite.com

- Go to eventbrite.com
- Click on one of the events listed
- View the webpage HTML source
  - **Chrome**
    - *Right click → View page source*
    - *View → Developer → View source*
  - **Safari**
    1. *Preferences → Advanced → Show develop menu*
    2. *Develop → Show page source*
  - **Firefox**
    Tools → Web developer → Page source
  - **Edge**
    Right click → View source

Example: Eventbrite.com

What you see



What your browser sees

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" xmlns:fb="http://ogp.me/ns/fb#" lang="en-us">
    <head>
<meta charset="utf-8" />
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<meta name="viewport" content="initial-scale=1, width=device-width">
<meta name="description" content="Eventbrite - Donald Lawrence presents Donald Lawrence Presents Edwin Hawkins Music &amp; Arts Global 2020 - Thursday, April 30,
2020 | Saturday, May 2, 2020 at Kirken i Kulturcentret, København, denmark. Find event and registration information."/>
<meta name="keywords" content="Eventbrite, Donald Lawrence Presents Edwin Hawkins Music &amp; Arts Global 2020, Donald Lawrence Presents Edwin Hawkins Music &amp;
Arts Global 2020 tickets, Donald Lawrence Presents Edwin Hawkins Music &amp; Arts Global 2020 København, Donald Lawrence Presents Edwin Hawkins Music &amp; Arts
Global 2020 København tickets"/>
<meta http-equiv="Content-Language" content="en" />
<link rel="canonical" href="https://www.eventbrite.com/e/donald-lawrence-presents-edwin-hawkins-music-arts-global-2020-registration-79985461669" />
<link rel="alternate" href="android-app://com.eventbrite.attendee/http/www.eventbrite.com/e/79985461669" />
<meta name="y_key" content="d92e23811007b438">
<meta name="msvalidate.01" content="A9AB07B7E430E4608E0BC57AFA5004AA" />
<!--
        ++++++++++++++++++++++++++++++++++
        ++++++++++++++++++++++++++++++++++
        +++++++++.                .+++++++++
        +++++++.                     :+++++++
        +++++      ++++++++++      .+++++
        ++++      ++++++++++++       .++++
        ++++     ++++++++++++        ++++
        +++;     +++                 +++++++
        +++:            ++++++++++++++++++
        ++++.      ,++++++++++++++++++++
        +++++      ++++++++++++ +++++++++
        +++++++      +++++++      +++++++
        ++++++++            ++++++++++
        ++++++++++.           ++++++++++++
        ++++++++++++++++++++++++++++++++++
        ++++++++++++++++++++++++++++++++++

Oh hi there, whatcha looking for? Wanna chat about joining our team?
Check us out at https://www.eventbrite.com/careers
-->
<meta name="robots" content="index, follow, " />
    <link rel="icon" type="image/png" href="https://cdn.evbstatic.com/s3-build/perm_001/8a4e23/django/images/favicons/favicon-32x32.png" sizes="32x32">
    <link rel="icon" type="image/png" href="https://cdn.evbstatic.com/s3-build/perm_001/477279/django/images/favicons/favicon-194x194.png" sizes="194x194">
    <link rel="icon" type="image/png" href="https://cdn.evbstatic.com/s3-build/perm_001/70b79b/django/images/favicons/favicon-96x96.png" sizes="96x96">
    <link rel="icon" type="image/png" href="https://cdn.evbstatic.com/s3-build/perm_001/67b5d4/django/images/favicons/android-chrome-192x192.png" sizes="192x192">
    <link rel="icon" type="image/png" href="https://cdn.evbstatic.com/s3-build/perm_001/08b9c2/django/images/favicons/favicon-16x16.png" sizes="16x16">
    <link rel="shortcut icon" href="https://cdn.evbstatic.com/s3-build/perm_001/471d12/django/images/favicons/favicon.ico">
<link rel="mask-icon" href="https://cdn.evbstatic.com/s3-build/perm_001/a24d6e/django/images/favicons/safari-pinned-tab.svg" color="#f6682f">
<meta name="apple-mobile-web-app-title" content="Eventbrite">
<meta name="application-name" content="Eventbrite">
<meta name="msapplication-TileColor" content="#f6682f">
<meta name="msapplication-TileImage" content="https://cdn.evbstatic.com/s3-build/perm_001/8ad267/django/images/favicons/mstile-144x144.png">
<meta name="theme-color" content="#f6682f">
    <link rel="manifest" href="https://cdn.evbstatic.com/s3-build/perm_001/fe7929/django/images/favicons/manifest.webmanifest">
<link rel="apple-touch-icon" href="https://cdn.evbstatic.com/s3-build/perm_001/eba891/django/images/touch_icons/apple-touch-icon-180x180.png">
    <!-- twitter -->
    <meta name="twitter:card" content="summary_large_image" />
    <meta name="twitter:site" content="@eventbrite" />
    <meta name="twitter:title" content="Donald Lawrence Presents Edwin Hawkins Music &amp; Arts Global 2020" />
        <meta name="twitter:description" content="Donald Lawrence presents Edwin Hawkins Music + Arts Global  -  Oh Happy Day" />
        <meta name="twitter:image" content="https://img.evbuc.com/https%3A%2F%2Fcdn.evbuc.com%2Fimages%2F80716413962F36736482100182F142Foriginal.20191112-211640?
w=1000&amp;auto=format%2Ccompress&amp;q=75&amp;sharp=10&amp;rect=0%2C0%2C2160%2C1080&amp;s=94f0f314e58519a87546d2d1c79a585a" />
        <meta name="twitter:app:name:iphone" content="Eventbrite" />
        <meta name="twitter:app:url:iphone" content="com-eventbrite-attendee://event/79985461669/?referrer=eiosurlxtcar&amp;" />
        <meta name="twitter:app:id:iphone" content="487922291" />
```

XML/HTML format

```html
<!DOCTYPE html >
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
</head>
<body>
<div id="header">
<ul id="A">
    <li><a href="#">A link</a></li>
</ul>
<div id="B"><span class="about">About something</span></div>
</div>

<div id="main">Main container...</div>

<p id="footer">This is the footer. <a href="#">Yet another link</a></p>
</body>
</html>
```

Web crawlers: Selected useful Python methods

- Read html files:

```
import urllib.request
contents = urllib.request.urlopen("http://example.com/foo/bar").read()
#returns a string containing the source html content for the url
```

- Regex that finds specific patterns in a text:

```
import re
results = re.findall(pattern, contents)
for result in results:
    # do something
#from in contents, find string(s) that matches the pattern specified by regex
```

- Wait for several seconds to reduce the risks of being detected and banned

```
import time
time.sleep(5)
#wait for 5 seconds.
```

Regular Expression

- Regex - An advanced search.
  - “Normal search” only deals with finding fixed character sequences.
  - Regex can handle various patterns.

- An interactive tutorial:
  - http://regexone.com/
- A place to quickly test a written regex against a source text:
  - http://regexpal.com/

Regular Expression

Regular- Expression Patterns

^ Matches beginning of line.
$ Matches end of line.
. Matches any single char except newline.
[...] Matches any single char in brackets.
[^...] Matches any single char not in brackets.
\w Matches word characters.
\W Matches nonword characters.
\s Matches whitespace.
\S Matches nonwhitespace.
\d Matches digits.
\D Matches nondigits.
\A Matches beginning of string.
\Z Matches end of string.
\z Matches end of string.
\G Matches point where last match finished.
x| y Matches either x or y.

[0-9] Match any digit; same as [0123456789]
[a-z] Match any lowercase ASCII letter
[A-Z] Match any uppercase ASCII letter
[a-zA-Z0-9] Match any of the above
[^aeiou] Match any other than a lowercase vowel
[^0-9] Match anything other than a digit.

The most useful one for web crawlers

<tag>(.*?)</tag>

match everything surrounded by <tag><tags>

Regular Expression: Example

HTML content:

```
<div class="txt-block" itemprop="actors" itemscope itemtype="http://schema.org/Person">
   <h4 class="inline">Stars:</h4>
   <name size=3>Ben Ziegler</name>,
   <name size=5>Glenna Hill</name>,
   <name size=4>Jason Woolfolk</name>
   <span class="ghost">|</span>
   <span class="see-more inline nobr">
      <a href="fullcredits?ref_=tt_ov_st_sm" itemprop='url'> See full cast and crew</a> &raquo;
   </span>
</div>
```

Regular Expression: Example

- Match the three names surrounded by <name> tags:
  - **<name size=\d>(.*?)</name>**

☑ Case insensitive (i) ☑ ^$ match at line breaks (m) ☑ Dot matches all (s; via XRegExp)

```
<name size=\d>(.*?)</name>
```

```
 <div class="txt-block" itemprop="actors" itemscope itemtype="http://schema.org/Person">
                            <h4 class="inline">Stars:</h4>
<name size=3>Ben Ziegler</name>,
<name size=5>Glenna Hill</name>,
<name size=4>Jason Woolfolk</name>                          <span class="ghost">|</span>
                          <span class="see-more inline nobr">
<a href="fullcredits?ref_=tt_ov_st_sm" itemprop='url'> See full cast and crew</a> &raquo;
                          </span>
                        </div>
```

Regular Expression: Example

```python
import re

html_content = """
 <div class="txt-block" itemprop="actors" itemscope>
     <h4 class="inline">Stars:</h4>
     <name size=3>Ben Ziegler</name>,
     <name size=5>Glenna Hill</name>,
     <name size=4>Jason Woolfolk</name>
     <span class="ghost">|</span>
     <span class="see-more inline nobr">
         <a href="fullcredits?ref_=tt_ov_st_sm" itemprop='url'>
             See full cast and crew
         </a> &raquo;
     </span>
 </div>
"""
names = re.findall(r'<name size=\d>(.*?)</name>', html_content)
for name in names:
    print(name)
```

```
Ben Ziegler
Glenna Hill
Jason Woolfolk
```

Web crawlers / scrapers: Summary

Get a list of URLs → Get content → Extract data

Web crawlers / scrapers: Summary

- Try not to (almost never) implement yourself
- A lot of frameworks are available, e.g., Scrapy (Python)
  - Tricks to imitate a real user not to be blocked
- A lot of HTML parsers are available (PyQuery, BeautifulSoup)
  Easy parsing functionality (e.g., *"get all blocks where class=event"*)
- Read Terms & Conditions carefully, obey "robots.txt" rules

Web crawlers / scrapers: Summary

- Try not to (almost never) implement yourself
- A lot of frameworks are available, e.g., Scrapy (Python)
  - Tricks to imitate a real user not to be blocked
- A lot of HTML parsers are available (PyQuery, BeautifulSoup)
  Easy parsing functionality (e.g., *"get all blocks where class=event"*)
- **Read Terms & Conditions carefully, obey "robots.txt" rules!**

APIs vs Web crawlers

	Pros	Cons
Third party APIs	• Convenient, easy to use • Safe, won't be blocked • Fast	• Need to manage API keys • Inflexible • Limit on access
Web crawlers / scrapers	• Very flexible. Theoretically, you can collect anything you find.	• Data structure can change • May be blocked • Terms & Conditions

Exercises

- Part 1 - Twitter.com API + a bit of NLP
- Part 2 - Weatherbit.io API + a bit of data visualisation
- Parts 3&4 - Web scraping (optional)