MONASH University

FIT1043 Introduction to Data Science

Week 4

Ian K T Tan

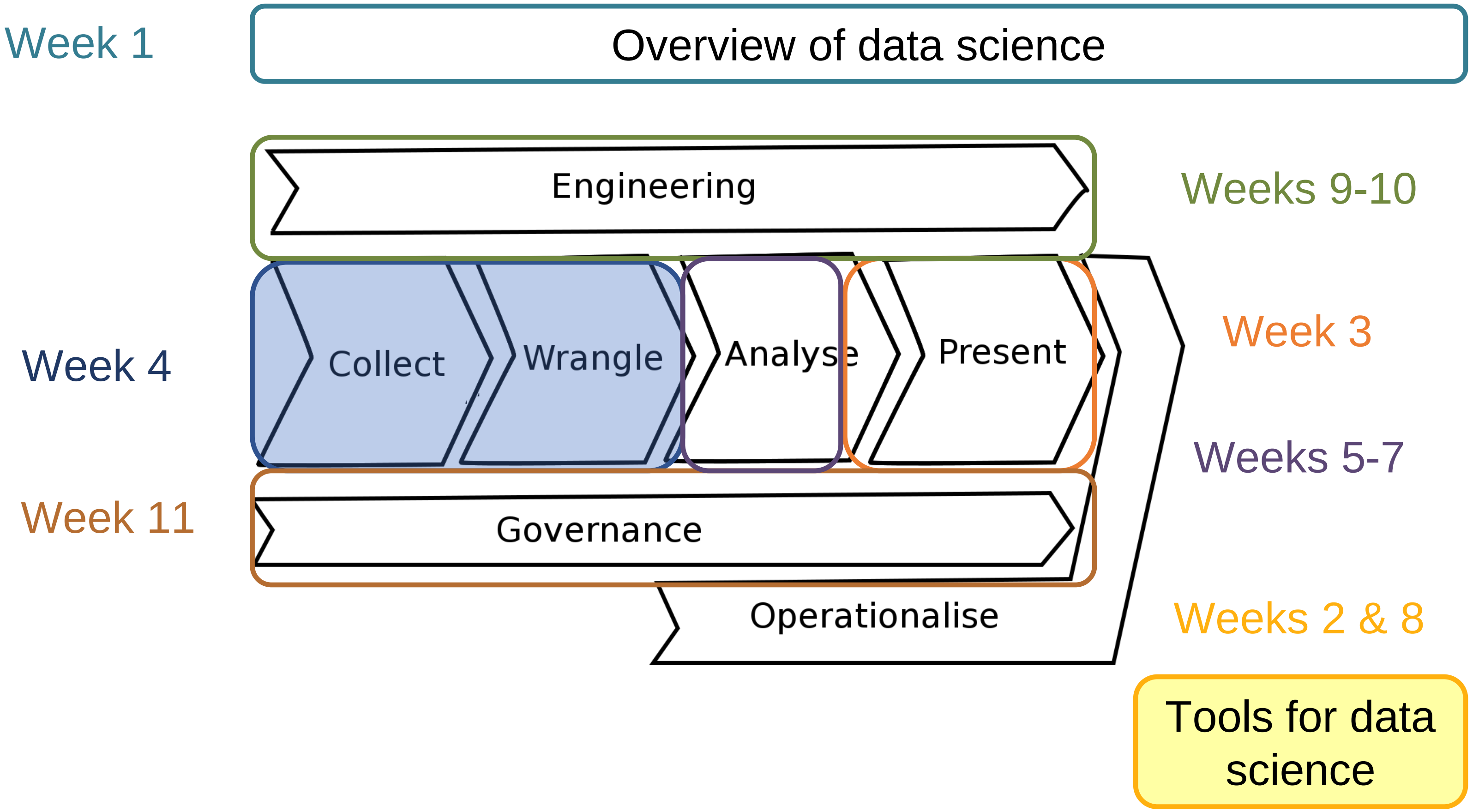
School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 3 Coverage
Visualisation in Python
Descriptive Statistics

MONASH
University

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	Assignment 2
9	Characterising data and "big" data	
10	Big data processing	
11	Issues in data management	Assignment 3
12	Industry Guest Lecture	


Week 1
Overview of data science
Engineering
Weeks 9-10
Week 4
Collect
Wrangle
Analyse
Present
Week 3
Weeks 5-7
Week 11
Governance
Operationalise
Weeks 2 & 8
Tools for data science

Week 4 Outline

Data Sources

- Open data
- API

Data Wrangling

- Motivation
- Data quality problems
- Data auditing in Python
- Techniques to handle the data quality problems

Learning Outcomes

Week 4

By the end of this week you should be able to:

- Explain open data and linked open data
- Explain how to access to new data sources through APIs
- Identify how different APIs work
- *Inspect data quality problems in datasets and recommend solutions to fix them*
- *Use data wrangling operations in Python*

Introduction to Resources

Introduction to Resources

Finding and Using Data

Important to a Data Science Project

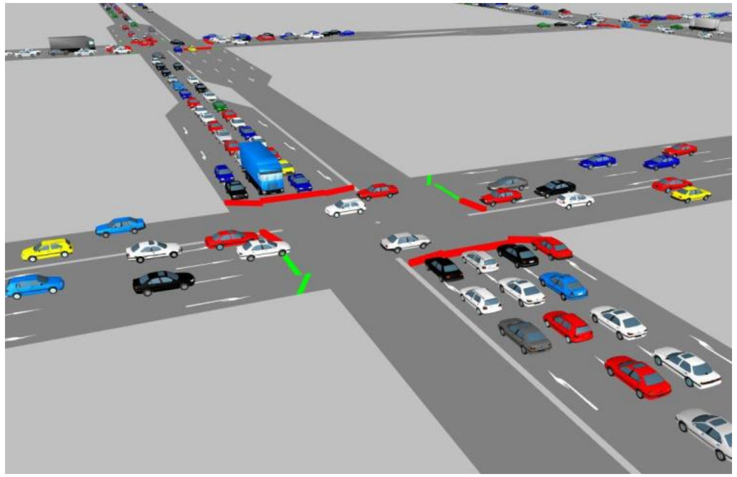
- Existing data sources
- Clever and creative use of multiple existing data sources
- Access to new data sources

Where to Find and How to Use

Data Sources

Traffic Forecasting

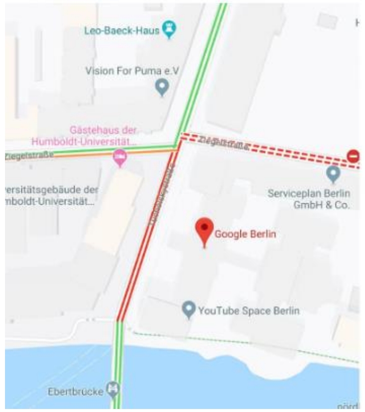
- Blockages (congestion)
- Alternate routes
- Unusual situations
- Clear

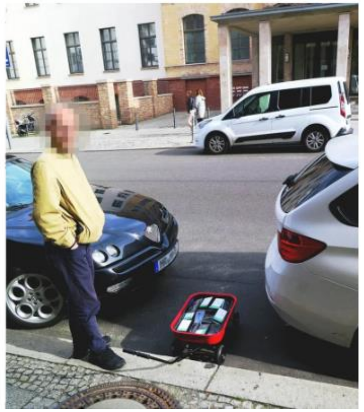


source: math.tu-berlin.de

Critical Data

- Maps
- Individual GPS data to monitor traffic flow
- Incidences (and events)
- Weather


Leo-Baeck-Haus
Vision For Puma e.V
Gästehaus der Humboldt-Universität...
Ziegelstraße
...ersitätsgebäude der ...mboldt-Universität...
Serviceplan Berlin GmbH & Co.
Google Berlin
YouTube Space Berlin
Ebertbrücke



Introduction to Open Data

Organizations that provide machine readable data sources to support data science

MONASH University

Worldwide Open Data Initiatives

From The World Bank

- Open Data involves **significant cost for public resources** as well as human resource effort, especially that of the civil service.
- However, there are **significant potential benefits** that come with it. Data, much like oil, has intrinsic value in that they need to be “refined” to realise their potential.
- When government data are made accessible and re-usable, they enable individuals, organizations and even governments themselves to innovate and collaborate.

Benefits of Open Data

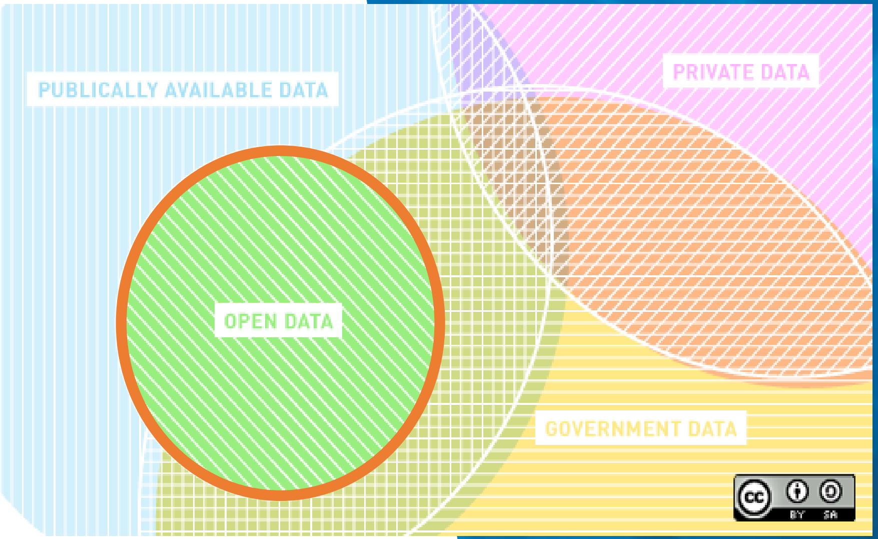
From The World Bank

Transparency

- Open Data can potentially help in oversight by government and hence potentially help reduce wastage and mitigate potential corruption.
- Makes it easier to track public budget expenditures and relate them to the impacts, e.g. police resources expenditure.

Public Service Improvement

- Open Data gives citizens the data they need to engage their governments and contribute to the improvement of public services.
- For example, citizens can use Open Data to contribute to road safety planning.
- Note that it is also useful for foreign countries and their citizens in comparing best practises or to learn from each other.



MONASH University

Benefits of Open Data

From The World Bank

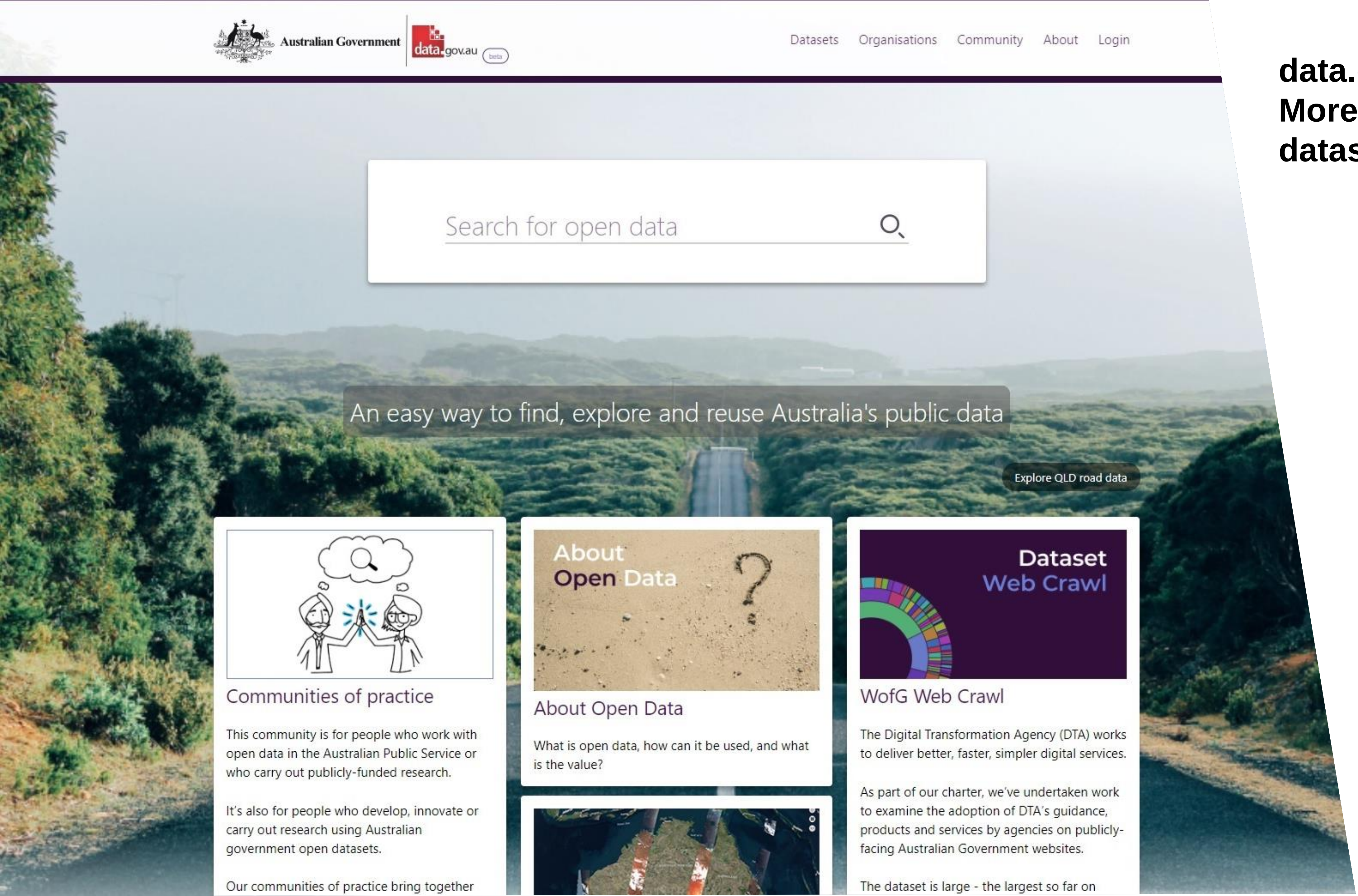
Innovation and Economic Value

- Public data are key resources for social innovation and even economic growth.
- Open Data provides new opportunities for governments to collaborate with citizens by giving citizens access to data.
- Businesses are able to use Open Data to better understand potential markets and build new data-driven products.

Efficiency

- Easier and less costly for government ministries to discover and access their own data or data from other ministries, which reduces acquisition costs, redundancy and overhead.
- Empower citizens with the ability to alert governments to gaps in public datasets and to provide more accurate information.

data.gov.au: More than 30k datasets.



An easy way to find, explore and reuse Australia's public data

Communities of practice

This community is for people who work with open data in the Australian Public Service or who carry out publicly-funded research.

It's also for people who develop, innovate or carry out research using Australian government open datasets.

Our communities of practice bring together

About Open Data

What is open data, how can it be used, and what is the value?

WofG Web Crawl

The Digital Transformation Agency (DTA) works to deliver better, faster, simpler digital services.

As part of our charter, we've undertaken work to examine the adoption of DTA's guidance, products and services by agencies on publicly-facing Australian Government websites.

The dataset is large - the largest so far on

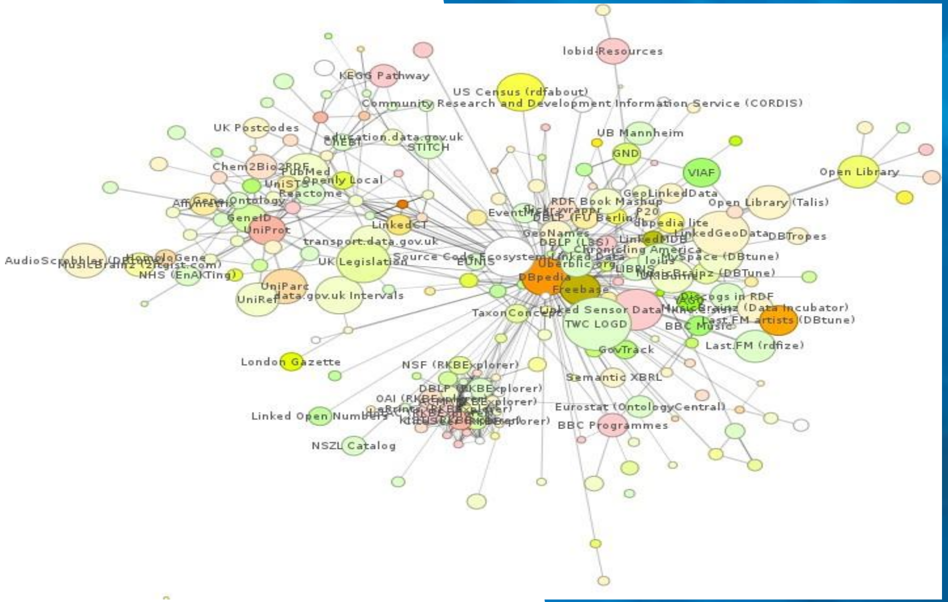
Common Open Data Format

Comma Separated Value (CSV) Files

- Note that the separator can be a semi-colon, colon, comma or even a tab. The CSV name is used to describe them all.

Linked Open Data (LOD)

- Triples: subject, verb and object
- E.g. DBpedia – a dataset containing extracted data from Wikipedia; it contains about 3.4 million concepts described by 1 billion triples.
- Enables data from different sources to be connected and queried.



API (Application Programmer Interface)

Routines that provide programmatic access to an application or data

Application Programmer Interface

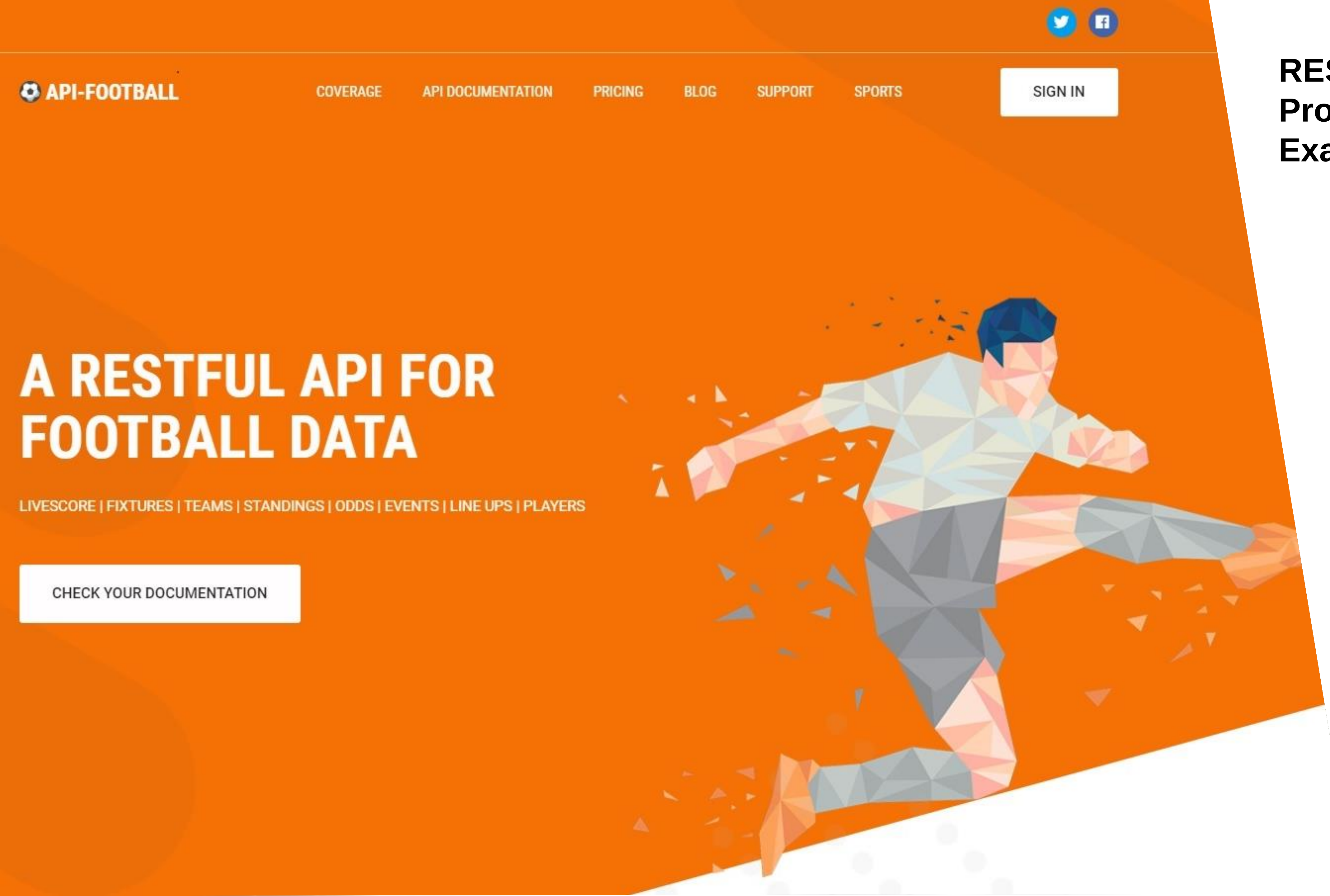
API

- API is like a user interface but it is designed for computers to access to the functionality of a software (google maps)
- Computers talk to each other
- API consumers vs API providers

Supplementary Materials (on Moodle)

- What is an API?
- What is a RESTful API (REST API)?

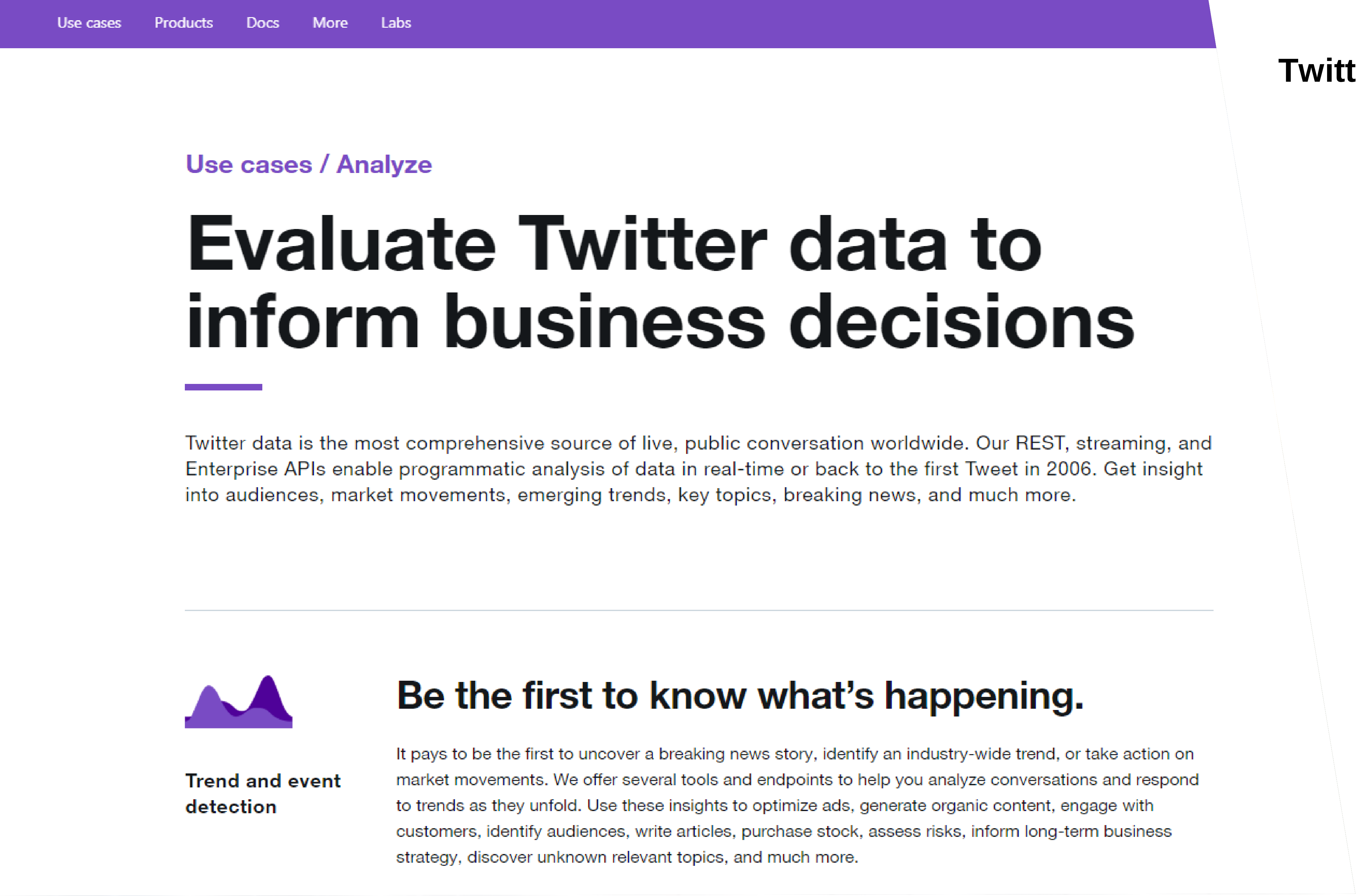
REST API Provider Example



A RESTFUL API FOR FOOTBALL DATA

LIVESCORE | FIXTURES | TEAMS | STANDINGS | ODDS | EVENTS | LINE UPS | PLAYERS

Twitter



Use cases / Analyze

Evaluate Twitter data to inform business decisions

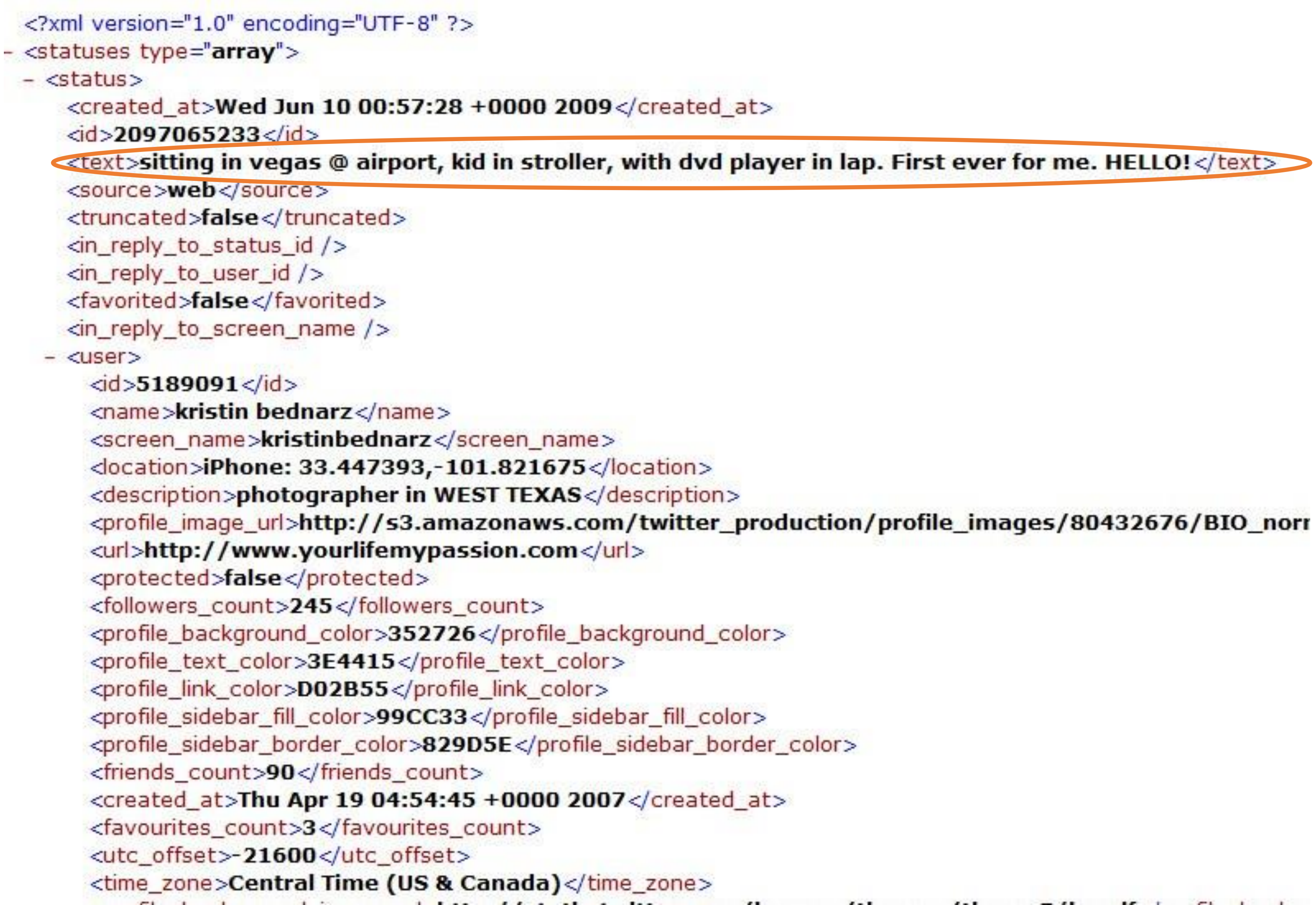
Twitter data is the most comprehensive source of live, public conversation worldwide. Our REST, streaming, and Enterprise APIs enable programmatic analysis of data in real-time or back to the first Tweet in 2006. Get insight into audiences, market movements, emerging trends, key topics, breaking news, and much more.

Trend and event detection

Be the first to know what's happening.

It pays to be the first to uncover a breaking news story, identify an industry-wide trend, or take action on market movements. We offer several tools and endpoints to help you analyze conversations and respond to trends as they unfold. Use these insights to optimize ads, generate organic content, engage with customers, identify audiences, write articles, purchase stock, assess risks, inform long-term business strategy, discover unknown relevant topics, and much more.

Sample Twitter XML Data



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <statuses type="array">
  - <status>
      <created_at>Wed Jun 10 00:57:28 +0000 2009</created_at>
      <id>2097065233</id>
      <text>sitting in vegas @ airport, kid in stroller, with dvd player in lap. First ever for me. HELLO!</text>
      <source>web</source>
      <truncated>false</truncated>
      <in_reply_to_status_id />
      <in_reply_to_user_id />
      <favorited>false</favorited>
      <in_reply_to_screen_name />
    - <user>
        <id>5189091</id>
        <name>kristin bednarz</name>
        <screen_name>kristinbednarz</screen_name>
        <location>iPhone: 33.447393,-101.821675</location>
        <description>photographer in WEST TEXAS</description>
        <profile_image_url>http://s3.amazonaws.com/twitter_production/profile_images/80432676/BIO_nor
        <url>http://www.yourlifemypassion.com</url>
        <protected>false</protected>
        <followers_count>245</followers_count>
        <profile_background_color>352726</profile_background_color>
        <profile_text_color>3E4415</profile_text_color>
        <profile_link_color>D02B55</profile_link_color>
        <profile_sidebar_fill_color>99CC33</profile_sidebar_fill_color>
        <profile_sidebar_border_color>829D5E</profile_sidebar_border_color>
        <friends_count>90</friends_count>
        <created_at>Thu Apr 19 04:54:45 +0000 2007</created_at>
        <favourites_count>3</favourites_count>
        <utc_offset>-21600</utc_offset>
        <time_zone>Central Time (US & Canada)</time_zone>
```

Twitter Developer API

Comprehensive

- Library interfaces for Java, C++, JavaScript, **Python**, Perl, PHP, and others
- Allows other applications to manage Twitter data for users
- Extensive developer policy

What can you do?

- Filter real-time tweets: Get only the Tweets you need by using advanced filtering tools with the real-time streaming API.
- Embed Tweets, Timelines, and more within your website.

Other APIs

- Many companies are exposing their data and their website functionality as APIs for others to make use of:
  - Facebook API
  - Twitter API
  - LinkedIn API
  - Google Maps API
  - YouTube API
  - Amazon Advertising API
  - TripAdvisor API

Data Wrangling

MONASH University