Introducción a APIs

Delta Analytics construye capacidad técnica alrededor del mundo.

El contenido de este curso está siendo desarrollado activamente por Delta Analytics, una organización sin fines de lucro 501(c)3 del Área de la Bahía que apunta a capacitar a las comunidades para aprovechar sus datos.

Por favor comuníquese con cualquier pregunta o comentario a inquiry@deltanalytics.org.

Descubre más sobre nuestra misión aquí.

Modulo 3.2:
APIs

APIs

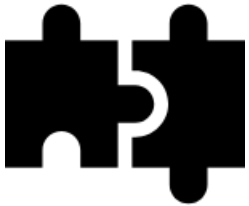
El tópico de este módulo

¿Cómo obtenemos datos de un tercero, como Kiva, Twitter, o Google?

Estos servicios ofrecen APIs (Application Programmable Interface), o intercambios/contratos entre programas.

APIs establecen como te debes comunicar con un tercero para obtener los datos que te interesan.

Luego, una vez que has aprendido a obtener datos,
¿Cuáles son las consideraciones éticas para acceder a ellos?



Module Checklist:

- ❑ Interfaces
- ❑ HTTP
- ❑ Formatos de datos (JSON)
- ❑ Consumir APIs en Python
- ❑ Ética de acceso de datos

HTTP

¿Cómo se relacionan estos términos entre sí?

Para entender mejor los términos de la diapositiva anterior, considera la siguiente analogía:

Tu quieres que tu correo sea recibido por la persona x y que esta te responda por correo.

HTTP

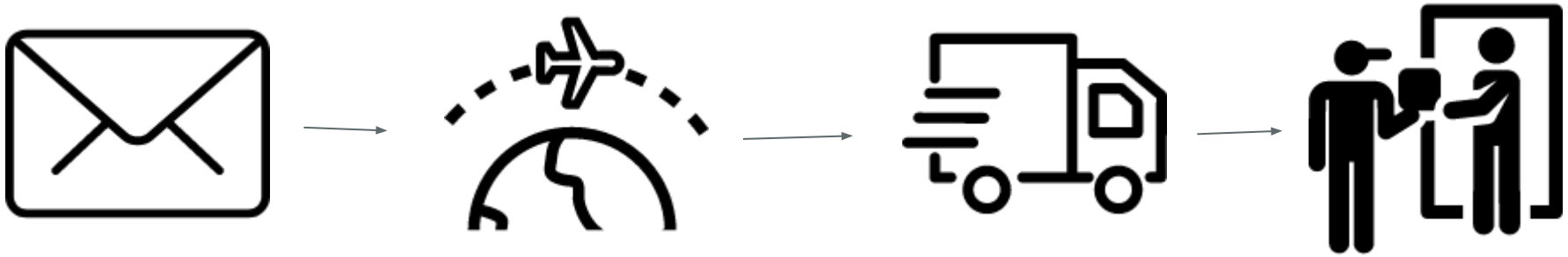
¿Cómo se relacionan estos términos entre sí?

- La **API or interfaz** provee un contrato entre tu y el sistema de entrega de correo, y entre tu y la persona x:
  - Si le das un sobre con una dirección, le agregas una estampilla, y se lo pasas a un funcionario de correos, tu sobre llegará a destino.
  - Tú y la persona x han acordado como comunicarse de antemano.

HTTP

¿Cómo se relacionan estos términos entre sí?

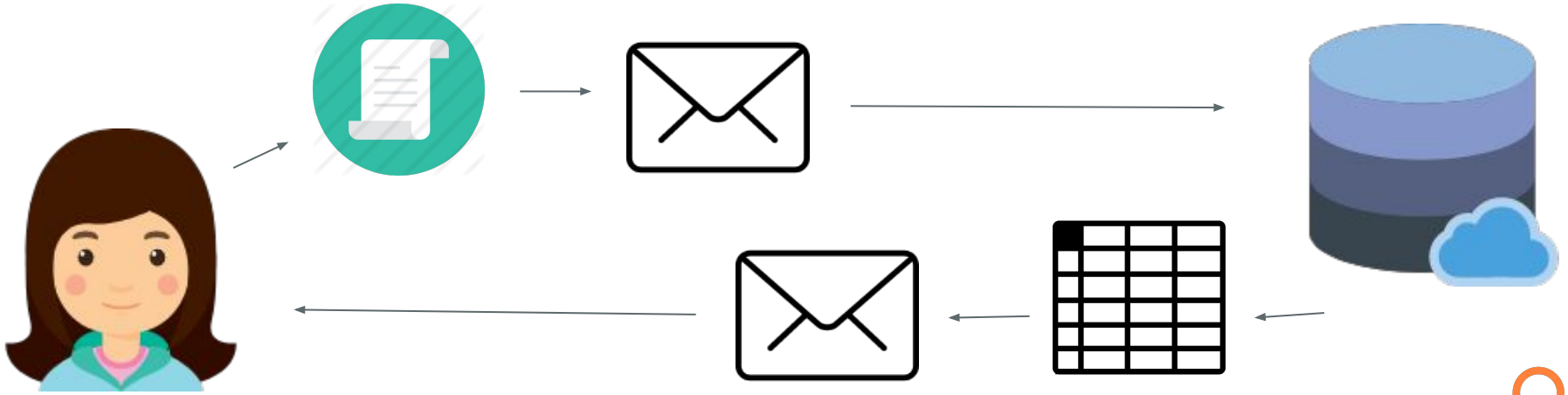
- **HTTP** es el medio por el cual es entregado. Tu correo podría ser entregado por avión, camión, o a pie – no te interesa, en tanto llegue a su destino.



HTTP

¿Cómo se relacionan estos términos entre sí?

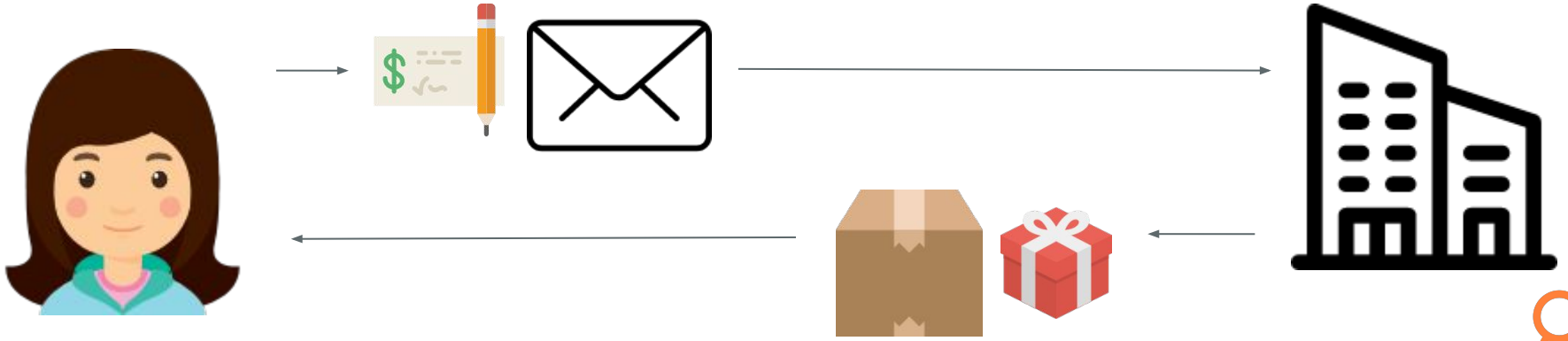
- Para recapitular:
  Puedes usar **Python** para enviar mensajes **HTTP** a un servicio tercero que provee su **API**, la cual **formateará los datos**, y te los enviará.



HTTP

¿Cómo se relacionan estos términos entre sí?

- El **formato de datos** es una forma acordada de cómo empaquetar el mensaje.
    - Si la persona X es una persona en una compañía que te podría enviar un producto, tu formato de datos podría ser un formulario de orden con un cheque, y como respuesta, la persona x te enviará de vuelta un producto envuelto en una caja de cartón.



Interfaces

Interfaces

Analogía de televisión

Una TV podría estar conectada a otros dispositivos, como

Un dvd

Un control remoto

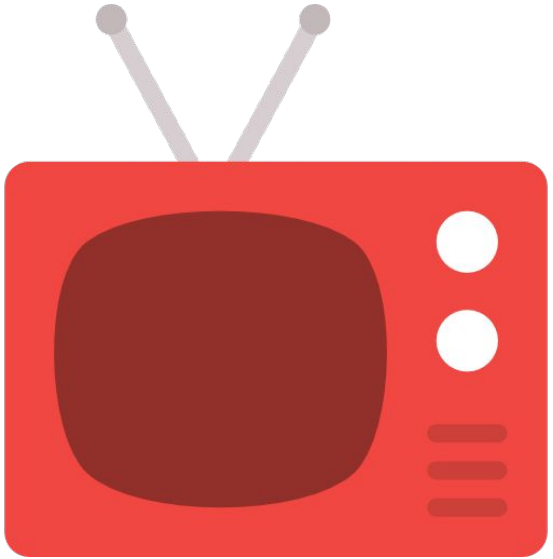
Parlantes

Analogía de televisión

El DVD podría ser de la Compañía A.

Mientras que la TV de la Compañía B.

Trabajadores de la Compañía A podrían nunca haber hablado con los de la Compañía B, y aun así, estos dispositivos son capaces de comunicarse entre sí!



Interfaces

Analogía de televisión

Piensa del HDMI como una interfaz común que dos dispositivos usan para comunicarse.

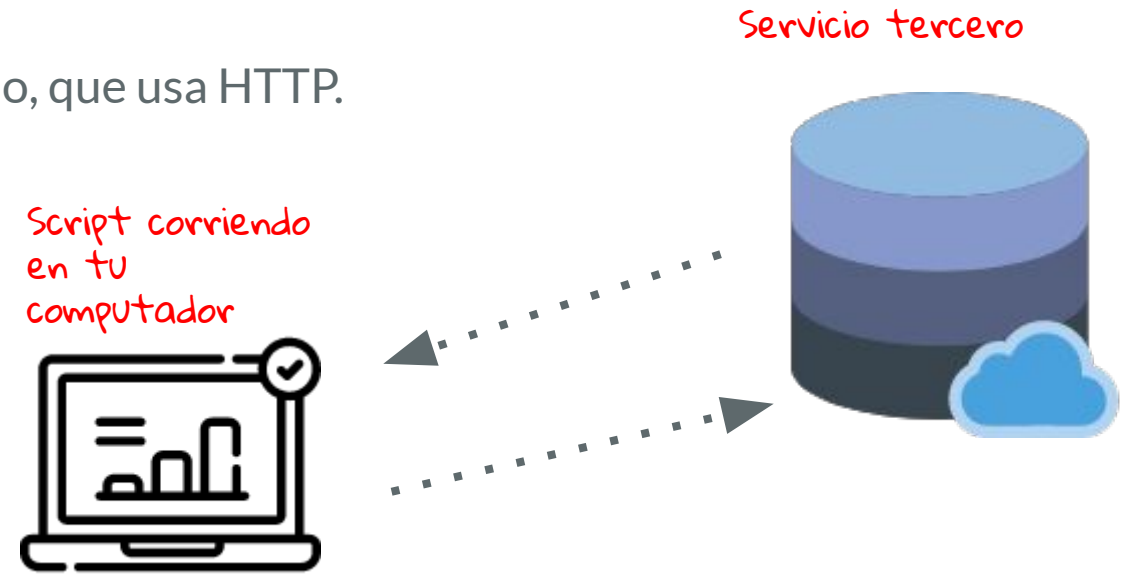
El DVD tiene un puerto HDMI físico para enviar video.

El TV tiene un puerto HDMI físico para recibir video.

Analogía de televisión

Análogamente, puedes enviar un mensaje a un servicio en la web, y ellos te enviarán un mensaje con los datos que solicitaste.

Discutiremos este método, que usa HTTP.



HTTP

HTTP

HTTP es corto de Hypertext Transfer Protocol.

HTTP abarca un conjunto de reglas y procesos para enviar y recibir mensajes en la world wide web.

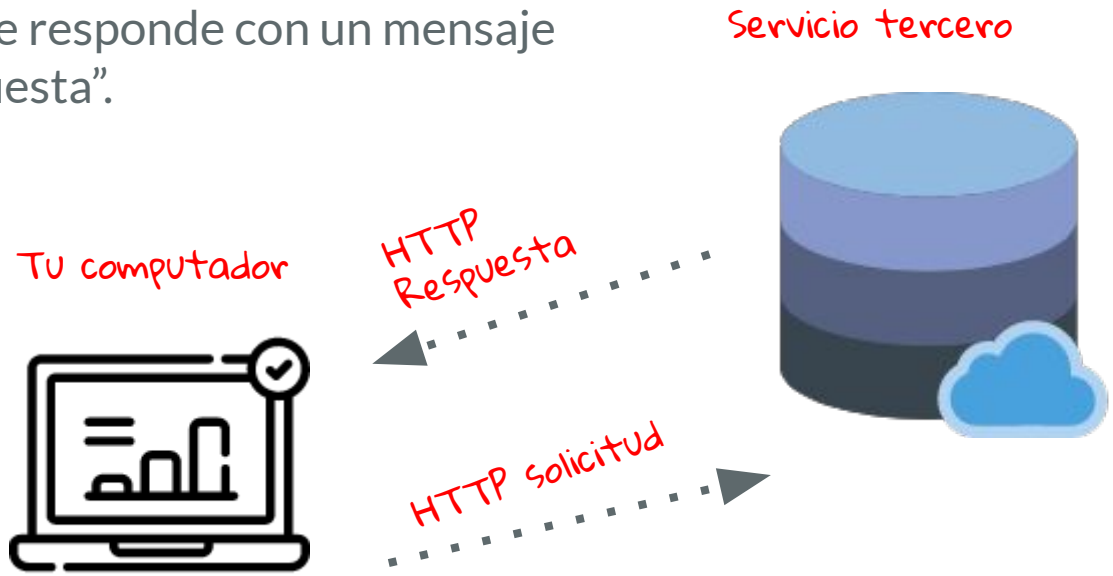
Típicamente tú no compones estos mensajes – tu navegador los produce, o escribes un programa que los produce.

HTTP

Mensajes HTTP

Cuando envías un mensaje HTTP al servidor del tercero, se llama una “solicitud”.

El servidor del tercero que responde con un mensaje HTTP, que se llama “respuesta”.



HTTP

Mensajes HTTP son como cartas y sobres

Carta en un sobre

Tu dirección Estampilla

Dirección de destino

Querido ___,

…
…
…

Mensaje HTTP

Línea de inicio (dirección de destino,
método)

Cabeceras

Cuerpo del mensaje

Solicitud HTTP

HTTP

HTTP Solicitud

Anatomía de una solicitud HTTP

Línea de inicio

Request URL: https://api.kivaws.org/v1/loans/search.json?country_code=KE&per_page=500
Request Method: GET

"GET" es el método HTTP que le dice al servidor que quieres obtener datos.

La URL le dice al servidor que quieres. Veremos esto en más detalle a continuación.

Cabeceras

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8
Accept-Encoding: gzip, deflate, br
Accept-Language: en-US,en;q=0.9
Connection: keep-alive
Host: api.kivaws.org
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/65.0.3325.181 Safari/537.36

Solo necesitas saber que existen las cabeceras HTTP. En general, no te debería preocupar.

Sin cuerpo de solicitud. Solicitudes GET no tienen cuerpo

HTTP

HTTP Request

Anatomía de una url

https://api.kivaws.org/v1/loans/search.json?country_code=KE&per_page=500

Protocolo	https
Nombre de dominio	api.kivaaws.org
Ruta	/v1/loans/search.json
Parámetros de solicitud	?country_code=KE&per_page=500

Url como función

https://api.kivaws.org/v1/loans/search.json?country_code=KE&per_page=500

Ahora que entiendes todas las partes que conforman una url, re-ordenemos las partes como otro concepto que probablemente ya conoces: la función.

Piensa en la url como una función que toma parámetros:

```
from api.kivaws.org import search_loans

search_loans(country_code='KE', per_page=500)
```

Parámetros de solicitud

Los parámetros de solicitud contienen información extra para esta. A veces no son necesarios – revisa la documentación de la api de la url para ver cuáles son necesarios.

Especifica pares de llave valor.

Por ejemplo:
?country_code=KE&per_page=500

- Parámetros de solicitud empiezan con “?”
- Seguidos de una llave (country_code)
- “=” separa la llave del valor (KE)
- Luego un “&” separa un par valor/llave de otro

HTTP

HTTP Request

¿Cómo encontrar las URLs de una API?

Request URL: `https://api.kivaws.org/v1/loans/search.json?country_code=KE&per_page=500`
Request Method: `GET`

API Reference

This is reference documentation for all methods exposed by the Kiva API.

Lenders

```
GET /lenders/:lender_ids
GET /lenders/:lender_id/loans
GET /lenders/:lender_id/teams
GET /lenders/newest
GET /lenders/search
```

LendingActions

```
GET /lending_actions/recent
```

Loans

```
GET /loans/:ids
GET /loans/:id/journal_entries
```

Servicios típicamente publican urls para acceder a datos.

Busca por “{nombre Compañía} API documentation”.

URLs están típicamente organizadas por recursos. E.g. prestamistas, acciones de préstamo, y préstamos son recursos disponibles en la tabla a la izquierda.

Respuesta HTTP

HTTP

HTTP Response

Anatomía de una respuesta HTTP

Línea de inicio

Status Code: 200 OK

El código de status te dice si el servidor te ha respondido de forma exitosa. Revisaremos esto a continuación.

Cabeceras

```
Access-Control-Allow-Origin: *
Cache-Control: private, no-store, no-cache, mus
date, max-age=0, post-check=0, pre-check=0, pr
lidate, no-transform
Content-Encoding: gzip
Content-Length: 32385
Content-Type: application/json; charset=UTF-8
Date: Fri, 06 Apr 2018 05:18:22 GMT
Expires: Tue, 03 Jul 2001 06:00:00 GMT
Last-Modified: Fri, 06 Apr 2018 05:18:22 GMT
Pragma: no-cache
Server: Apache/2.4.18 (Ubuntu)
Vary: Accept-Encoding
X-RateLimit-Overall-Limit: 60
X-RateLimit-Overall-Remaining: 60
```

El content-type, o formato en el que el cuerpo es formateado es JSON. Examinaremos esto en más detalle en las siguientes diapositivas.

Cuerpo del mensaje


```
{
  "paging": {
    "page": 1,
    "total": 150934,
    "page_size": 500,
    {
      "paging": {
        "page": 1,
        "total": 150934,
        "page_size": 500,
        "pages": 302
      },
      "loans": [{
        "id": 1501869,
        "name": "George",
        "description": {
          "languages": ["en"]
        },
...
```

Códigos de estado HTTP

2xx	Exito!
3xx	Redirección
4xx	Hiciste algo mal
5xx	El servidor hizo algo mal

Códigos de estado HTTP comunes

200	Exito!
301	El servidor te está redirigiendo. No debiese ser un problema para ti.
400	El servidor piensa que estás haciendo una mala solicitud. Revisa tu url y parámetros de solicitud.
401/403	No tienes la autorización adecuada para el recurso que solicitas. Revisa que la estás proporcionando correctamente.
404	El recurso no existe. Revisa tu url.
500	Algo falló en el servidor, así que no te puede proporcionar lo que quieres.

HTTP

En las diapositivas anteriores, aprendiste sobre urls y cabeceras.

Le dicen a un mensaje HTTP
1. Donde llegar, e
2. Información sobre el recurso

A continuación nos enfocaremos en el cuerpo del mensaje, donde viven los datos que nos interesan!

Formatos de Datos (JSON)

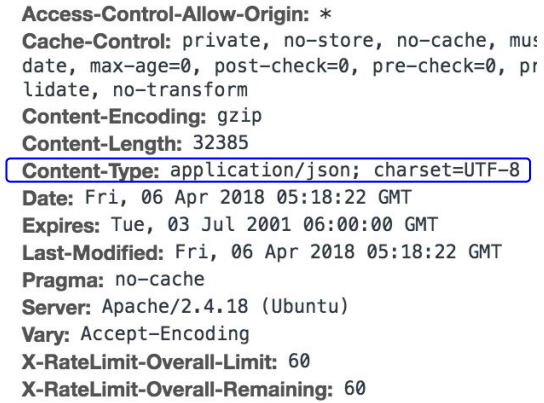
Formato de datos

Un formato de datos especifica como un modelo de datos es representado en la memoria.

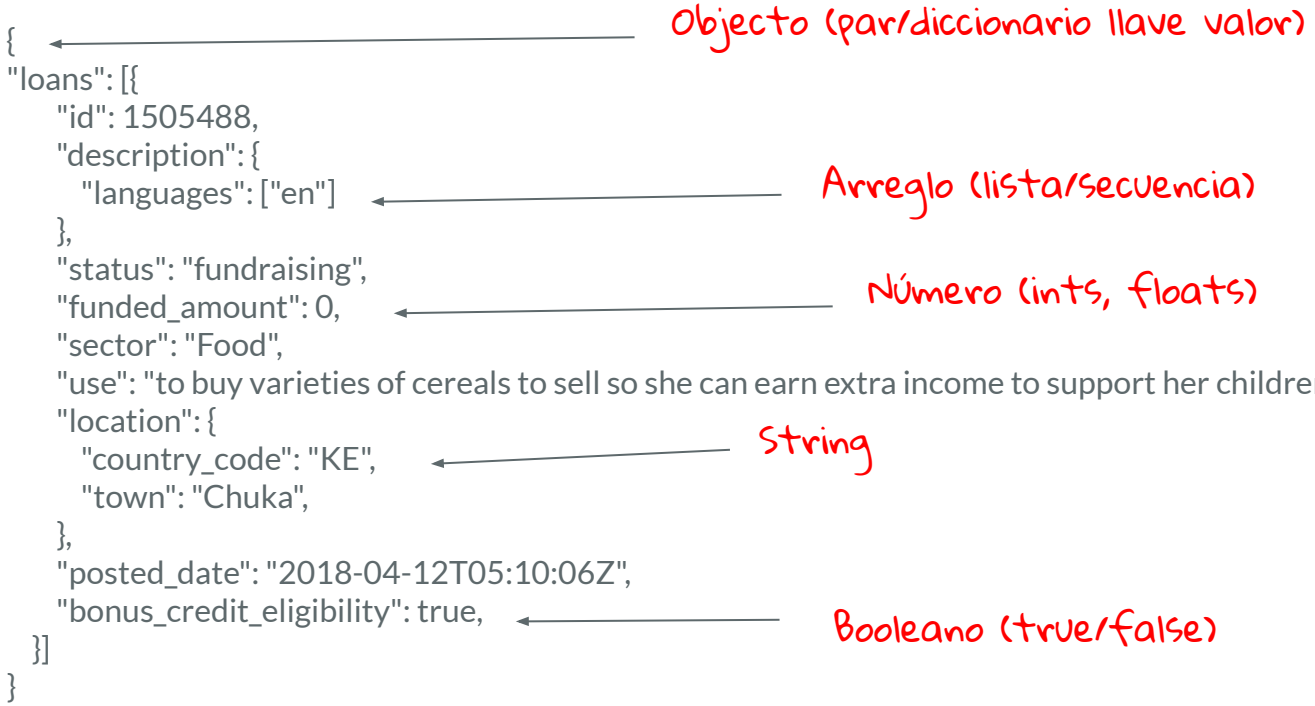
Veremos un ejemplo concreto de esto en la siguiente diapositiva.

Por lejos, el formato más popular para enviar información por la web es JSON.

Cabeceras

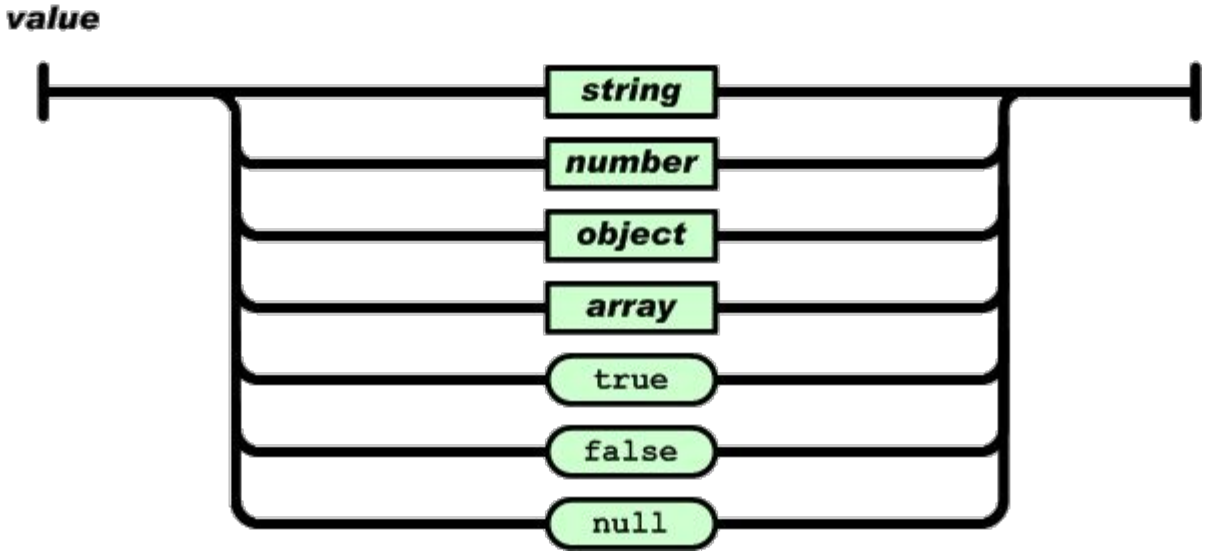


Anatomía de un objeto JSON



Valores

Un valor puede ser cualquiera de estos tipos:



Objectos

Un objeto es un par llave valor. Representado por:
- Dos llaves cursivas, de abertura y cierre { }
- Una llave, que debe ser un string
- Un : separa una llave de
- un valor

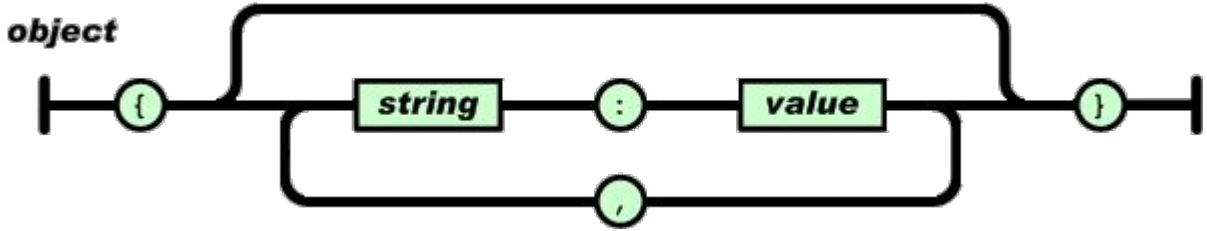
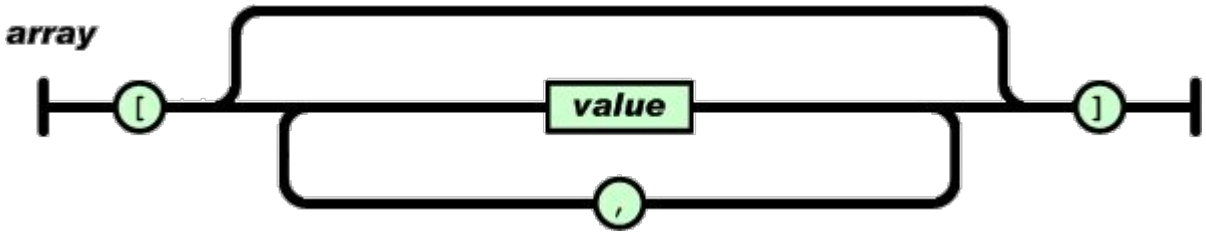


Image credit: https://www.json.org

Arreglo

Un arreglo, o lista de valores, es representado por:

- Un paréntesis de apertura y cierre
- Un valor
- , separan valores en la lista



Consumiendo APIs en Python

APIs en Python

Paquete Requests

El paquete “Requests” es el recomendado para uso general por la documentación de Python.

Su documentación puede ser encontrada aquí.

Instrucciones de instalación serán incluidas en el Jupyter Notebook.

s

APIs in Python

¿Cómo usar Requests?

```
1 import requests
2 >>> r = requests.get('https://api.github.com/user', auth=('user', 'pass'))
```

Envía un mensaje HTTP y guarda la respuesta como r

```
3 >>> r.status_code
200
```

Revisa que la respuesta fue exitosa

```
4 >>> r.headers['content-type']
'application/json; charset=utf8'
```

Revisa el formato del cuerpo de respuesta

```
5 >>> r.json()
{u'private_gists': 419, u'total_private_repos': 77, ...}
```

Examina el cuerpo de respuesta

En las diapositivas previas, aprendimos sobre interfaces, http, json, y usar el paquete requests en python.

Has ganado un superpoder! Con un gran poder viene una gran responsabilidad.

Discutamos ética de acceso de datos.

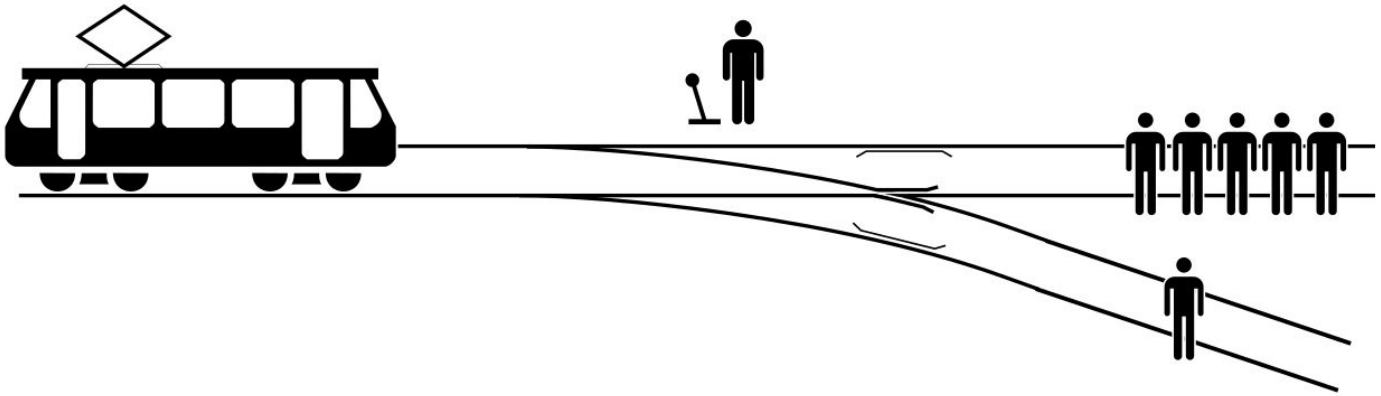
Ética de Acceso de Datos

Ética de los datos

Ética

¿Qué es la ética?

Ética son reglas sociales compartidas en una sociedad que la gente sigue para su beneficio.



Ética de ciencia de datos

Tópicos en ética de ciencia de datos pueden incluir

- Consentimiento informado
- Data ownership
- Privacidad
- Justicia algorítmica

No discutiremos todos estos, ya que este módulo se enfoca en acceso a datos. Con ese fin, nos enfocaremos en consentimiento informado y privacidad respecto a métodos de recolección de datos.

(Fuente https://www.coursera.org/learn/data-science-ethics)

Caso de estudio real

- La pregunta planteada por el caso de estudio es: Cuando la data que quieres no está disponible públicamente, pero es disponibilizada por hackers, ¿puedes usar esos datos éticamente?

- Miraremos a la historia detrás, el debate, y las decisiones reales tomadas por investigadores. Luego te presentaremos una serie de lineamientos que te ayudarán a tomar una decisión si te encuentras esta situación en el futuro.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Ética de los datos

Historia detrás y hechos

- Un grupo de cientistas de datos querían investigar crowdfunding .
- Estaban interesados en datos de Patreon, un website de crowdfunding.
- Patreon no otorgaba una API. Los investigadores consideraron escrapear páginas de Patrón, porque habría decenas de miles de páginas.
- Sin embargo, no podían asegurar que llegarían a todas las páginas de proyectos. Esto significaba que los datos serían una muestra de conveniencia y no una muestra representativa para un estudio.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Historia detrás (continuación)

- Mientras consideraban esto, Patreon fue hackeado en Octubre 2015. Toda la información fue hecha pública, incluyendo datos públicos de los proyectos (de interés para los investigadores), mensajes privados, emails, y claves.

- Como se accedió a los datos no es legal en US.

- Sin embargo, los datos de interés a los investigadores ya era pública previo al hack, ahora fácilmente accequible.

- Algunos investigadores opinaban que como la información ya era pública, podrán usarla éticamente, mientras otros estaban en desacuerdo.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Ética de los datos

Casos previos

- Al no llegar a acuerdo, los investigadores miraron a casos pasados en periodismo.

- Algunos ejemplos:
  Un diario del estado de New York publicaron nombres y direcciones de dueños de armas viviendo en el área, tomando los datos de fuentes públicas. Esta gente atacó al diario, así que hacer datos públicos más públicos no siempre visto como apropiado.

  - Mark Burnett, un investigador de seguridad, recolectó usuarios y claves de fuentes ilegales. Publicó su colección para facilitar la investigación. Dijo que no se podría causar un mayor daño ya que estos datos ya eran públicos, y ahora podían ser usados más fácilmente para propósitos investigativos.

- Luego de examinar los casos pasados, los investigadores debatieron.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Ética de los datos

Debate

Argumentos en favor del uso

1. Los datos son públicos, como un diario.
2. Esperamos servir el bien común con nuestra investigación.
3. Es la data que queremos, pero no podemos obtener por otros métodos.

Argumentos en contra

1. Investigadores tienen una capacidad limitada para distinguir entre información pública y privada en el área de datos hackeados.
2. Podrían ver datos privados al limpiar los datos.
3. Quizás legitiman actividad legal.
4. Violando la esperanza de privacidad de los usuarios.
5. Usando los datos de los usuarios sin su consentimiento.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Ética de los datos

Decisión

- Al final, los investigadores decidieron no usar los datos habilitados por los hackers.

- Sintieron que los argumentos en contra eran más fuertes.

- Inclusive cuando utilizarían los datos para bien y datos que ya eran públicos, decidieron que los negativos sobrepasaban a los positivos.

(Fuente: https://bdes.datasociety.net/wp-content/uploads/2016/10/Patreon-Case-Study.pdf)

Ética de los datos

Ética de ciencia de datos leyes/guias

- Las leyes empiezan a alcanzar la práctica, al ponerse en efecto en Mayo del 2018 las leyes EU's General Data Protection Regulation (GDPR).
  - Estas leyes tienen como propósito proteger la privacidad de todos los miembros de la UE. Por ejemplo, entidades colectando datos deben especificar un propósito y usarlos para ese propósito únicamente.

- Ciencia de datos como industria carece de un código de ética universal. Hay compañías individuales que están creando "manifiestos de ética de datos". La siguiente diapositiva presentará una colección de principios.

Ética de los datos

Código de conducta de dos puntos

- El Profesor H.V. Jagadish, de la Universidad de Michigan, propuso un código de ética para ciencia de datos de dos puntos:

    1. No sorprendas al sujeto en colección y uso de datos.

    2. “Adueñate” de las consecuencias. Si el proceso da pie a eventos no deseados, arréglalo.

(Fuente: https://www.coursera.org/learn/data-science-ethics)

Checklist:

- ✓ Interfaces
- ✓ HTTP
- ✓ Formato de Datos (JSON)
- ✓ Consumiendo APIs en Python
- ✓ Ética de acceso de datos

Recursos Avanzados

Quieres ir más allá? Aquí hay algunos recursos que recomendamos:

- HTTP Overview from Mozilla
- JSON overview from Mozilla
- Requests package
- Data Science Ethics course

Felicidades! Terminaste el módulo!

Obtén más información sobre el machine learning de Delta para una buena misión aquí.