

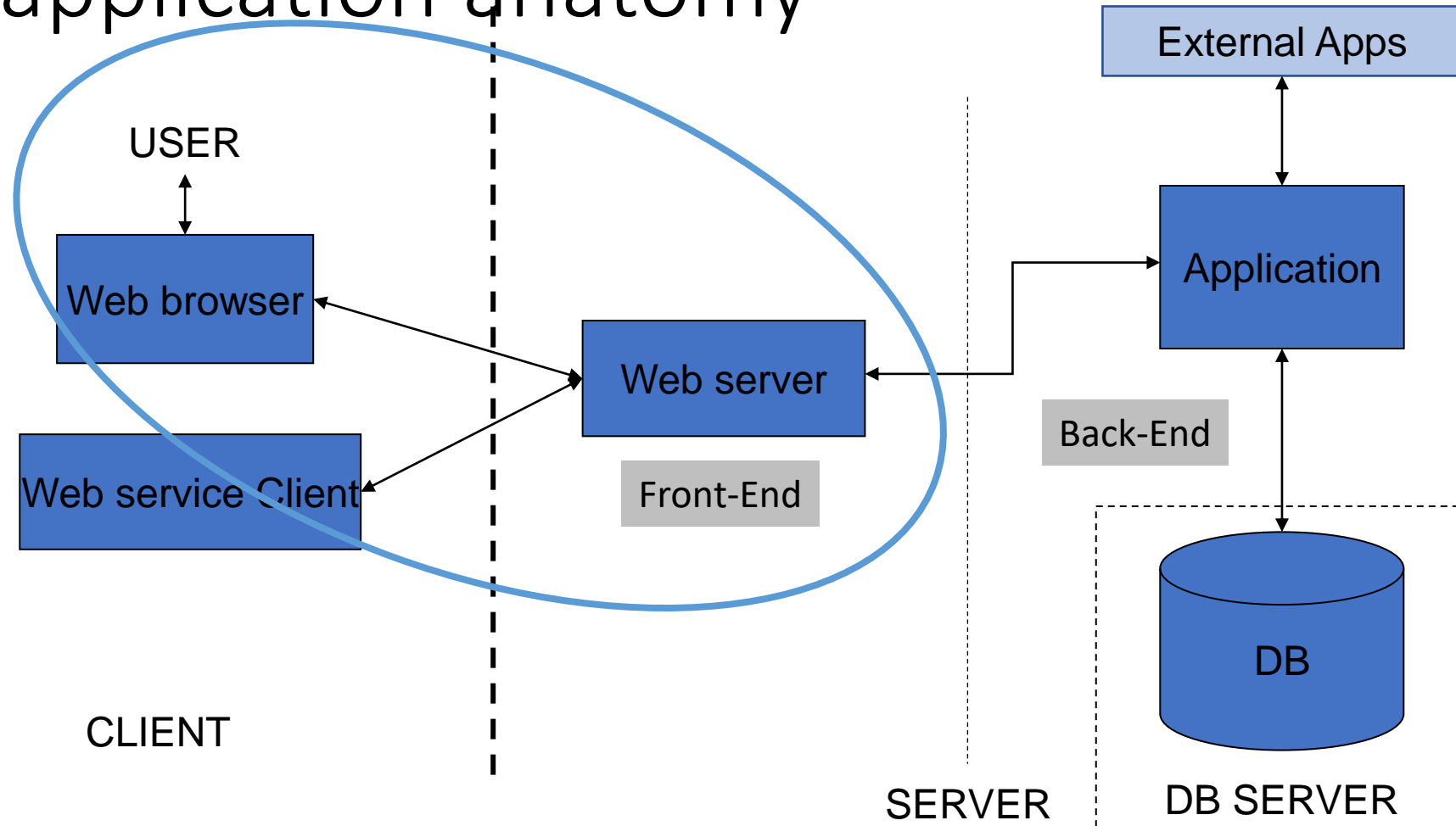
# Web, Web applications & APIs

DBW

# Outline

- Web basics
  - HTTP servers and browsers
  - Languages
  - Software
- Concept and types of Web applications. Web services
- Languages involved
- The Model-View-Controller paradigm
- Transmission details. CGI Protocol
- Common architectures
- Time issues
- Authentication/authorization. Cookies & session persistence
- Hints
- Web services & APIs
  - REST
  - Languages
  - Examples

# Web application anatomy



# Web (HTTP) Servers & Clients

## *Servers*

- **Listen** at a TCP port (80/443), and **understand HTTP**.
- Data can be text or binary files (*resources*) stored locally in the server or generated dynamically by **server-side applications**
- Example SimpleHTTPServer ([Perl](#), [Python](#))

## *Clients*

- **Make requests** to a server **using HTTP**.
- **Resources are identified by URLs**
- Normal browsers “understand” **combine information** interpreting HTML/CSS to provide a **graphical output**.
- Most **browsers can execute client-side applications** (code is obtained from the server)

# Languages involved (web interfaces)

- **HTML: Contents management language**

- Defines contents and structure of the page, includes the necessary links to all elements
- Tag formatted language (...<p>Some text</p>...)

- **CSS: Formatting language**

- Defines how the contents is represented in the user browsers
- `P {font-family:Times; font-family: 10pt; display:block; background-color:black}`

- **Javascript family: Task execution languages**

- Used for client-side applications
- Plain or in richer variants like Typescript, and frameworks like Angular, React, Vue, ...

# Languages involved (web services)

## XML

Most traditionally used by web applications

- Same structure as HTML, but with no fixed tags
- Requires XML-schema to specify tags and check coherence

```
<Course id="DBW">
  <Acronym>DBW</Acronym>
  <Title>Databases and Web applications</Title>
  <Students>
    <Student id="1">
      <name>Josep</name>
      ...
    </Student>
    ...
  </Students>
</Course>
```



## JSON

Increasing replacing XML

- Natively understood by Javascript
- Can be validated using JSON-Schemas (not mandatory)

```
Course: {
  "id": "DBW",
  "Acronym": "DBW",
  "Title": "Databases and Web applications",
  "Students": [
    {"id": 1, "name": "Josep", ...}, ...
  ]
}
```

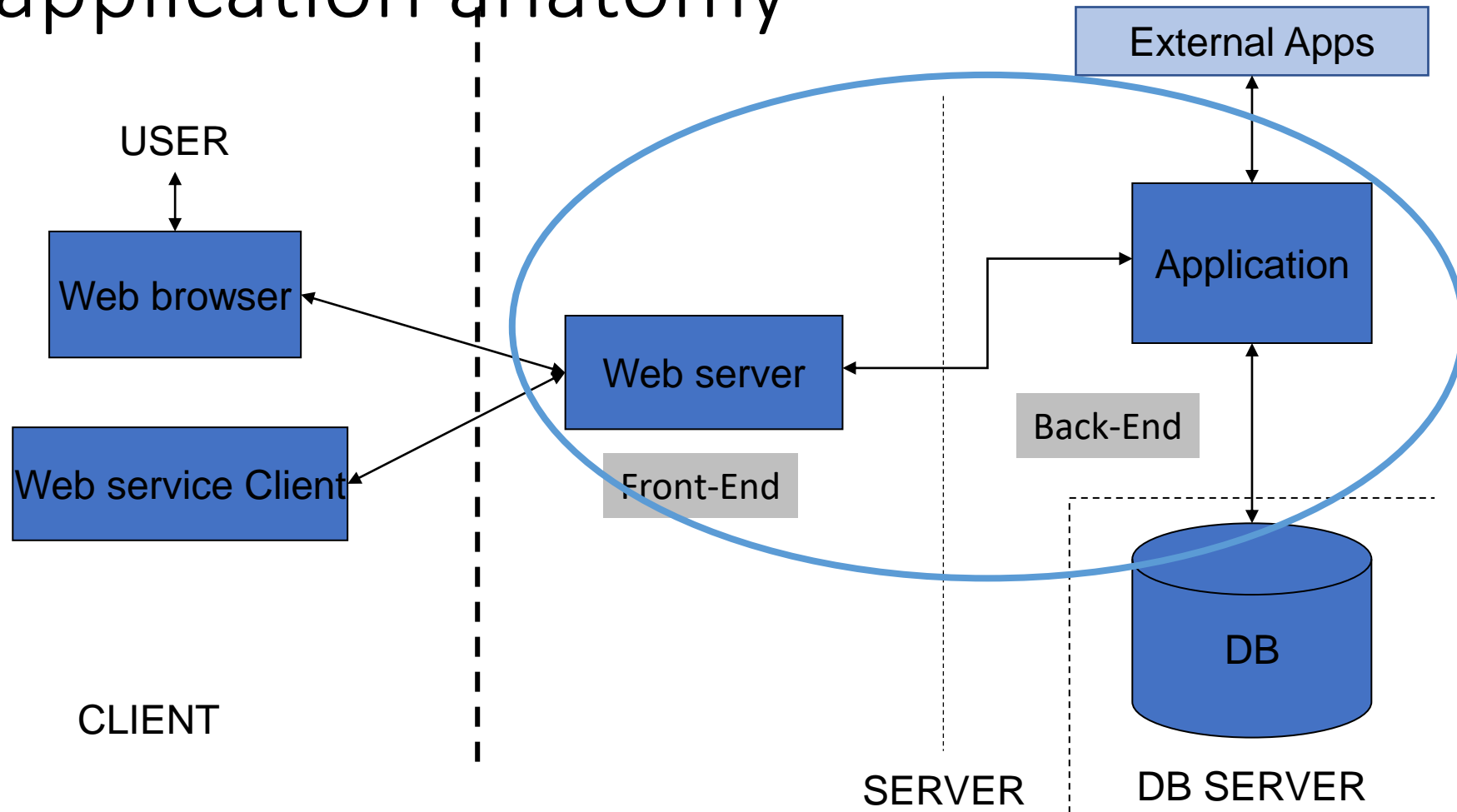
## YAML (kind of friendly JSON)

Used mainly for configuration files.

```
Course:
  id: DBW
  Acronym: DBW
  Title: Databases and Web applications
  Students:
    - 1 :
      name: Josep
    - ...
```



# Web application anatomy



# Definition & types

- A Web application is a **dynamic extension** of a Web server.
  - Adapts to user input
  - Can serve non-static information (generated in real-time)
  - Uses standard protocols (HTTP(S), SMTP)
  - Users interact with the application mainly using Web browsers
- Presentation-oriented
  - Generates dynamic Web pages (HTML/CSS/JS) responding to user queries
  - Usual way to provide bioinformatics results
- Service-oriented
  - Interacts with other applications (XML/SOAP, REST)
  - Allows to build automatic workflows for complex analyses



# Frontend (interacts with clients)

- Application must be compatible with standard web browsers
  - HTTP protocol: GET, POST, (PUT)
- User input comes from URL's or HTML forms
- Output must be in known languages: HTML, CSS, Javascript, XML, JSON
- Output may invoke other programs (plug-ins) though MIME types
  - Almost obsolete, fully replaced by HTML v5
- HTML v5 include a variety of native functionalities
  - Audio/video, SVG graphics, MathML, GeoLocalization, parallel processing, ...
- Modern browsers are able to run **client-side apps** (Java applets, and Javascript)
  - **Client-side applications** are fully qualified applications, interpreted and run by the browser
  - Java applets are almost obsolete (still seen in bioinformatics) due to security issues
  - **Javascript** is behind dynamic behaviour of modern web sites.
    - Asynchronous interaction with server (new requests do not require reload)
    - JsMol, NGLview (molecular structures), Jbrowse (genomic data)
    - Component libraries for bioinformatics start to be available (<https://ebi-webcomponents.github.io/nightingale/#/>)
  - **Single Page Applications.** Full self-contained applications that run entirely on the browser.
- Client-side apps are generated on the server and sent to the browser as part (or all) of the output.

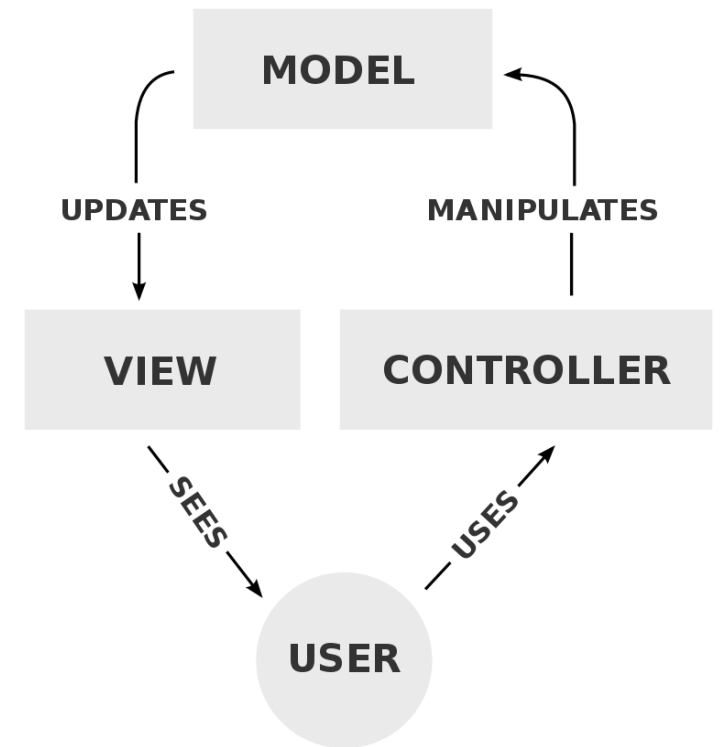
# Backend (the part that do the work)

- An application is invoked by web server on receiving the request
  - **External application (“CGI”)**
    - Executable running in the server machine.  
Can be written in **any** language.
    - Get input from **standard input** and writes in the **standard output**. Web server redirects both.
  - **Server embedded.**
    - Web server is able to execute the application as a child process (may require a driver)
    - Usual languages: **Python**, **NodeJS**, Perl, Java
    - “Designed for Web” languages: **PHP**, ASP, .NET, JSP
- Java, Python, JS server-side applications require special servers
  - Installed normally as a secondary web server (port != 80)
  - The main server acts as a “proxy”



# Model – View - Controller (MVC) concept

- Helps to “rationalize” the structure of a web Applications
- Specially suited for data retrieval, but is generally aplicable
- **“Model”**: The actual data that is required to anwser the request (“data model” structure)
- **“Controller”**: Part of the application that understands the request and perform the necessary logic
- **“View”**: Translates formats the data model into the appropriate output (either HTML, JSON, etc)
- Controllers and views should (ideally) be **kept separated** in the code (even in diferent scripts/computers).
  - In Single-Page Applications, the “controller” in executed on the server and the “view” on the client



# Transmission details. CGI Protocol

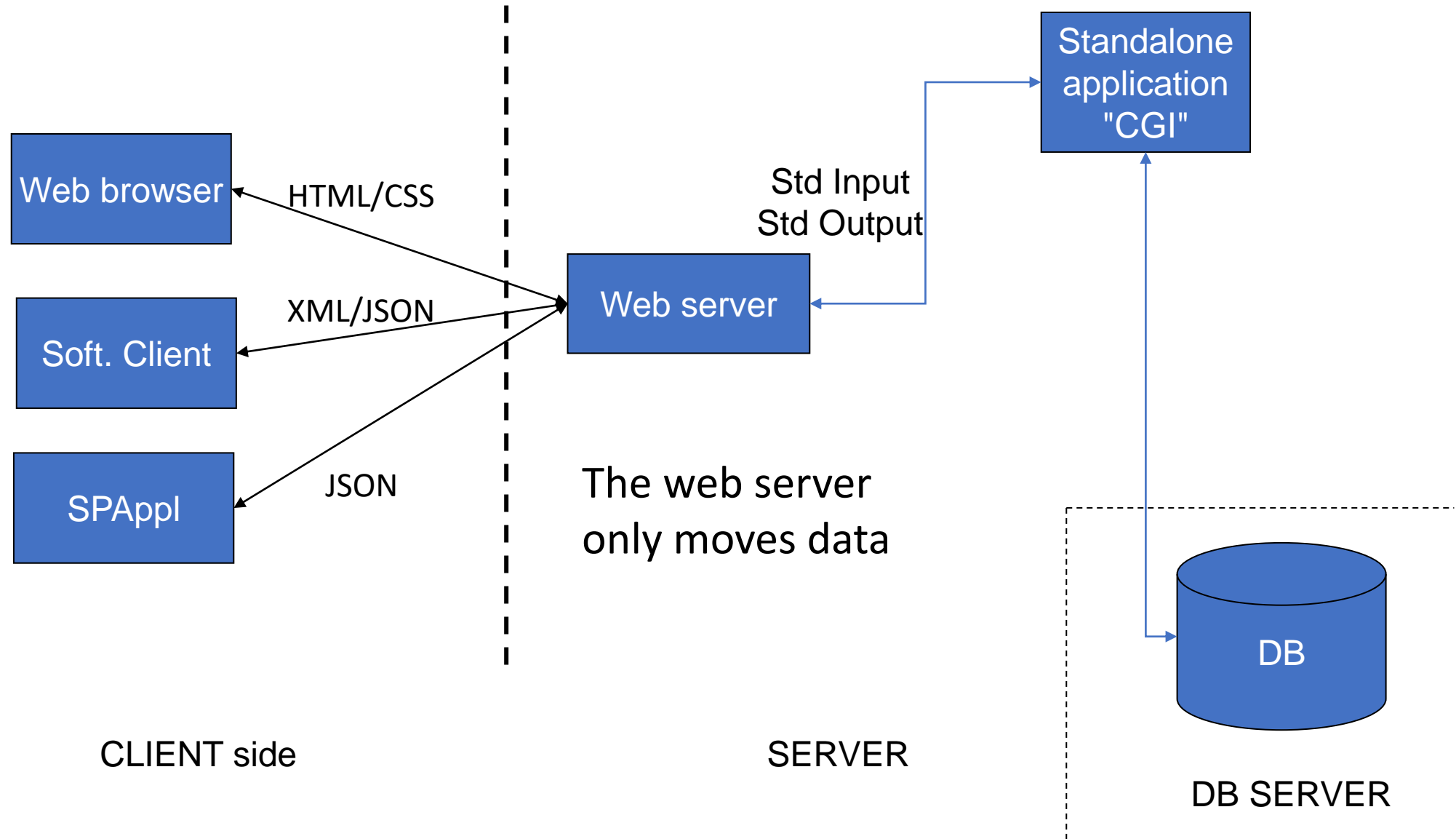
- **Common Gateway Interface (CGI)**

- **Formal interface** between Web server and external applications
- CGI interface provides
  - **Environment variables** including information from the browser-server conversation
  - **POST input data**, as standard input
  - **Redirection** of application standard output & error to Web stream.

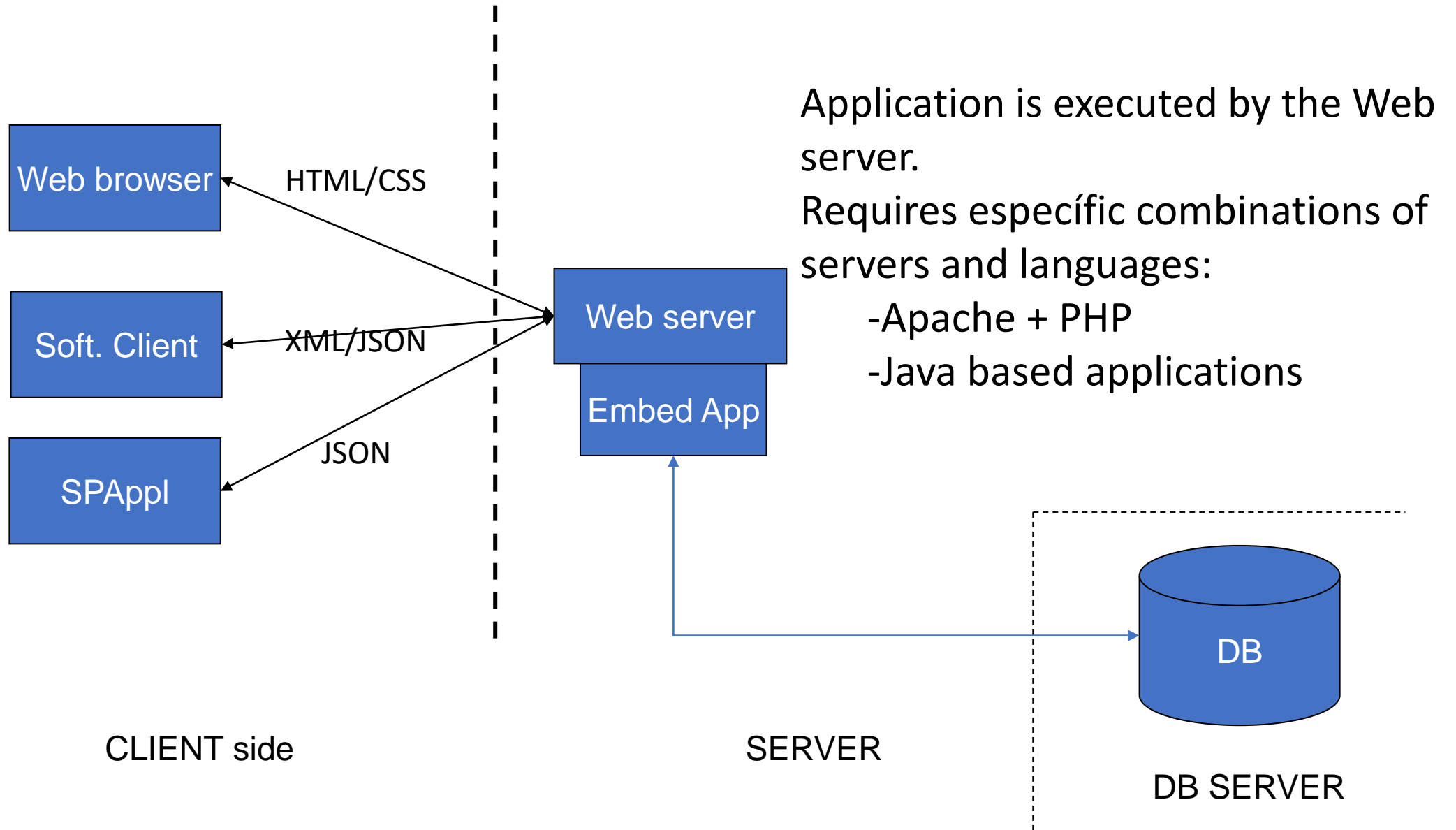
- **Practical CGI use**

- Languages have specific extensions to deal (Hide) CGI from the development process
  - Input GET and POST, and CGI variables available at predefined variables/objects
  - No need to process HTTP
    - Cookies, authentication, etc.
- Web programming frameworks
  - **Slim**, Symphony (PHP), **Flask**, Django (Python)
  - Programming helpers acting as an interface between CGI and the programming language
  - Provide “easy” web applications

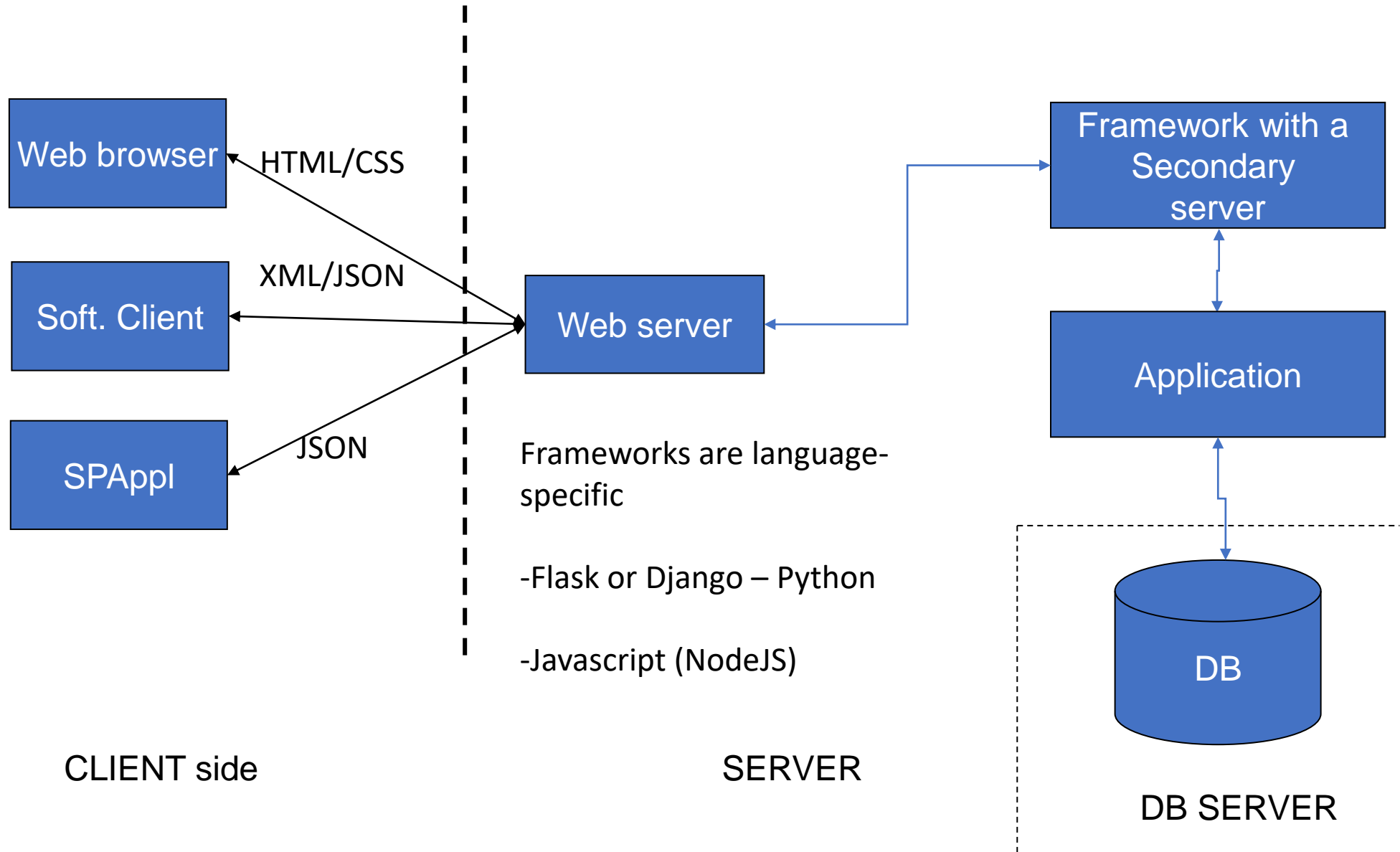
# Architectures (I) The classical model (obsolete)



# Architectures (II) Embedded Application Server



# Architectures (III) Application frameworks



# Known issues in web applications

- **Time issues**

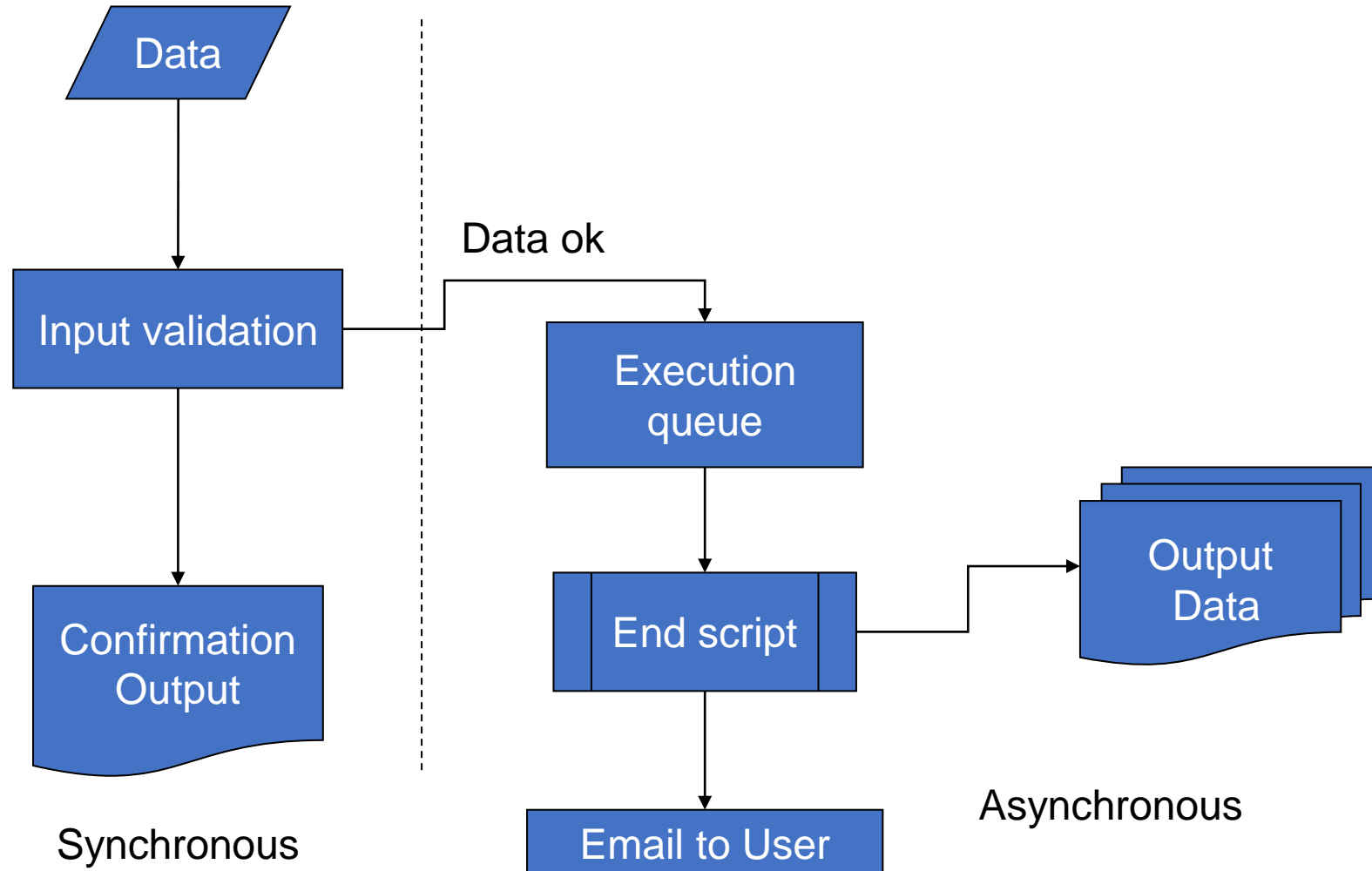
- Web users require “instant” responses
- Most web browsers (and network helpers) may have a short “timeout”
- Application that lasts more than 1-2 mins must be asynchronous

- **Persistence, and User recognition**

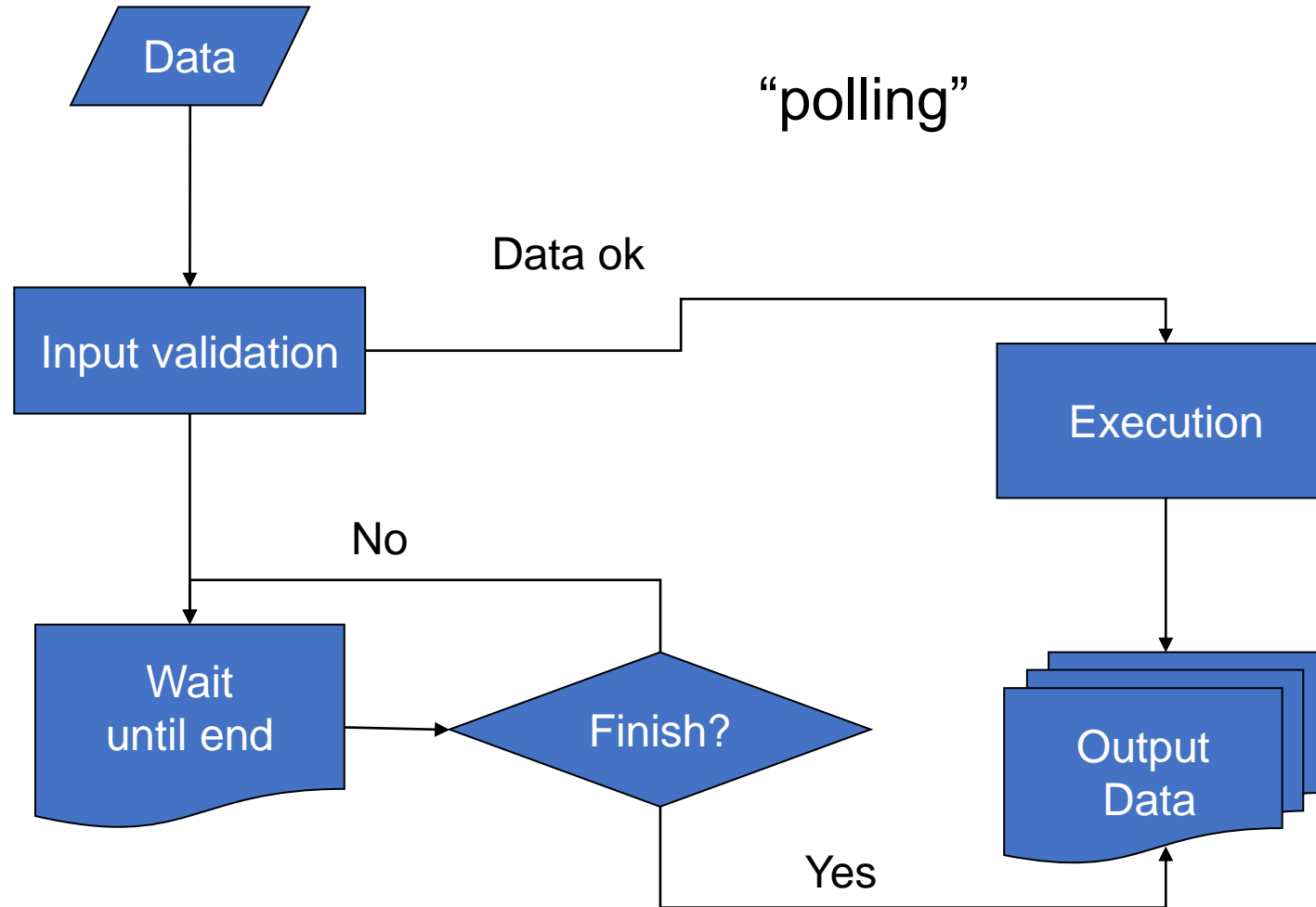
- HTTP protocol is **not persistent**: Connection closes short time (~20s) after the server answers
- Applications need to **recognize returning users**
  - Authentication (user only must write the login/password once)
  - Keep personal preferences, and private data
  - Grant access to given resources according to previous requests
  - Avoid requesting known data more than once
  - Avoid “reloads”



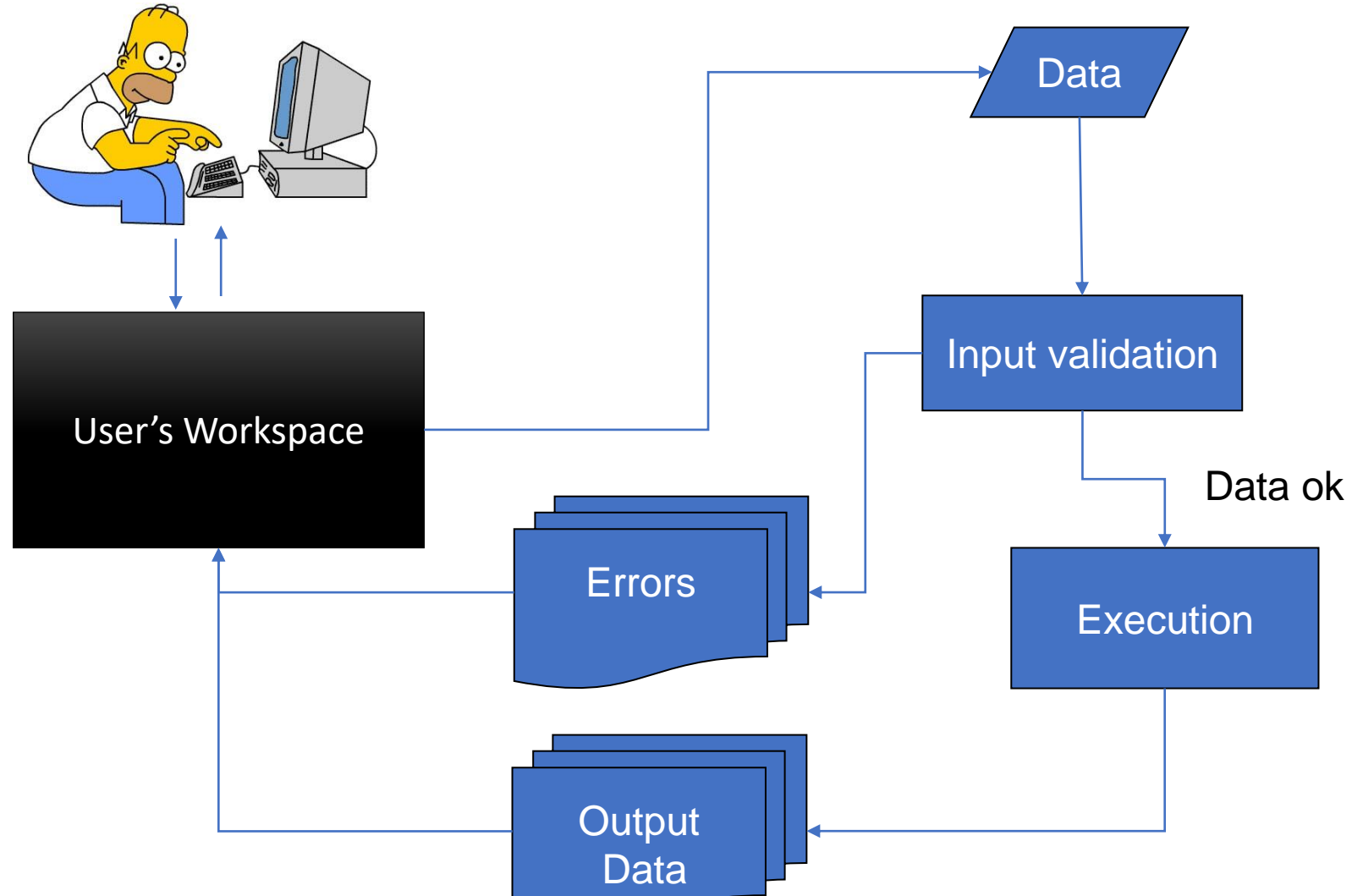
# Time issues. Usual strategies (1)



# Time issues. Usual strategies (2)

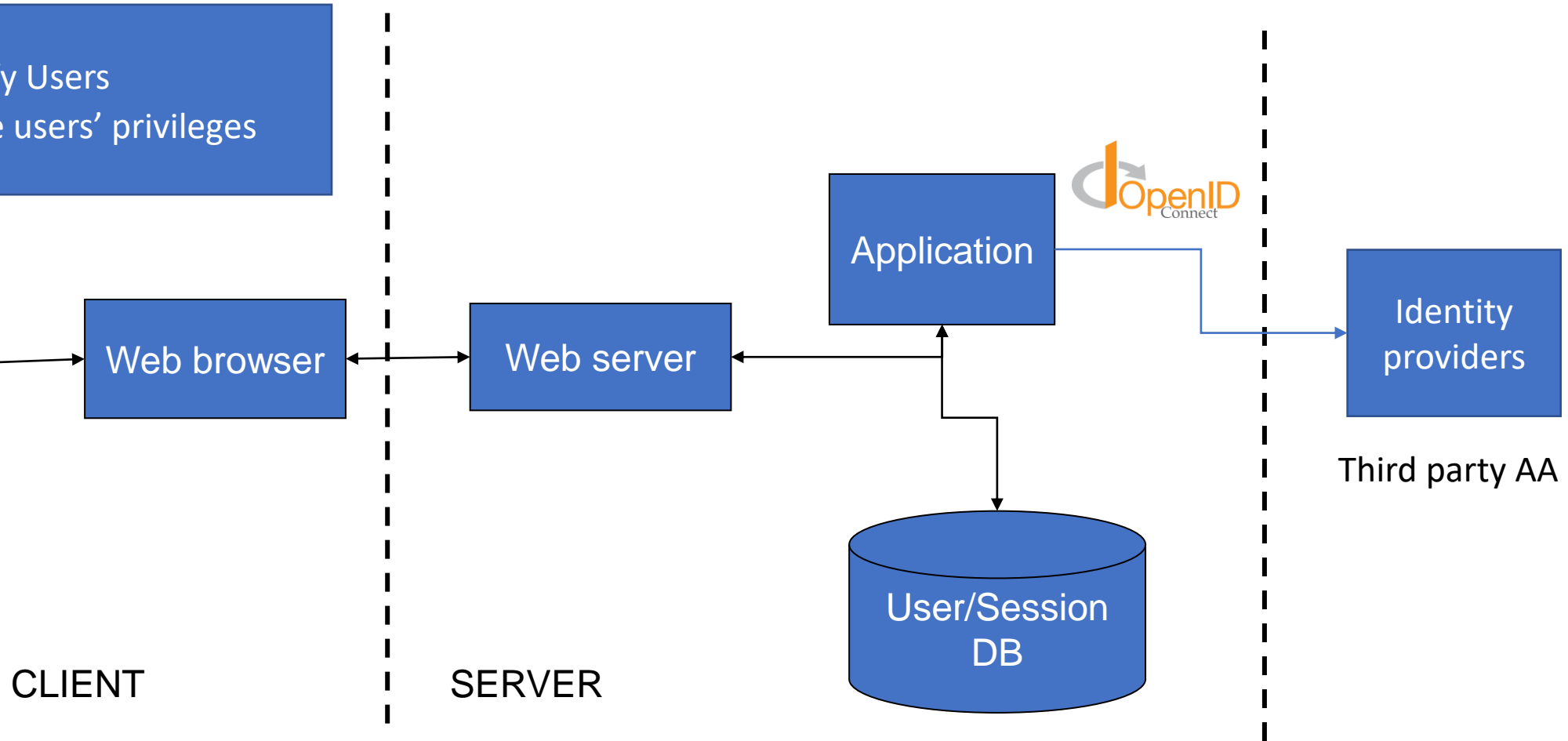
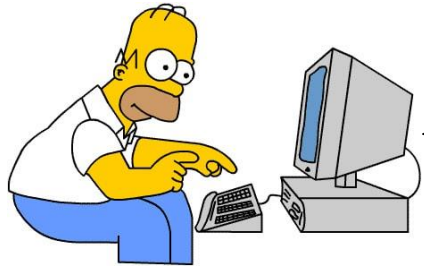


# Time issues. Usual strategies (3)



# Personalization (Authentication/Authorization)

**Authentication:** Identify Users  
**Authorization:** Manage users' privileges



A DB stores history of user connections and activities

# Authentication/Authorization schemes

- **Server based authentication**

- Based on unix-like passwd files (login / passwd)
- Protects folders and sub-folders ([.htaccess](#) files)
- Identity via CGI variable ([REMOTE\\_USER](#))
- May **require access to server configuration**

- **Application based (both authentication and authorization)**

- Do not require access to server configuration
- Authentication and environment managed by the application itself (via local DB)
- Full control from the application (login / passwd, SSL Keys, ...)
- Persistence via Cookies or language specific constructs

- **Third party authentication (single sign-on, SSO)**

- Authentication is done by identity providers (Google, openID, ELIXIR, ....), or other apps (eGroupWare, Drupal, ...)
- **Users can have a single point for authentication (SSO)**
- Protocols
  - OAuth, OpenID Connect
- Bioinformatics world
  - *Bona fide* Researchers
  - European Life Sciences Id / ELIXIR



# User identification from the application:

## cookies

- Small amount of text information stored by the server in the users' web browser as key /value pairs
- Do not require user/password (user do not need to be aware of)
- Limited to 4Kb
- Retrieved automatically as part of CGI handshake
- Cookies do not identify persons but browsers!!
- **ID:** a unique ID generated by the server
- **Origin:** server URL. Browsers send back cookies to the servers that created them (no other servers can get the data)
- **Expiration date.** Cookies can last for a single session or till a specified date

# Web application layout hints

- Static contents (text, images, etc. ) stored as normal web resources
  - For optimization, some servers keep them separated from scripts
- Dynamic pages managed by server scripts
  - No general rule, depends on language and programming style
  - The easy way: Each different screen is managed by a specific script.
  - Web frameworks use normally a “routing” mechanism to associate code to incoming URLs
  - **Single-Page-Applications** ([http://itsnat.sourceforge.net/php/spim/spi\\_manifesto\\_en.php](http://itsnat.sourceforge.net/php/spim/spi_manifesto_en.php)). Manage everything in a single client-side applications with asynchronous communication with the server.
- Global variables
  - Each script acts in a separated HTTP transaction!
  - All scripts should load the **same global environment**, usually included from a single file
- Protected/public data
  - Protected data should be stored **outside of the web directory tree**, and be accessed only programmatically
  - Output data may be placed inside the web tree if it is already HTML/CSS

# Web application layout hints

- Temporary data
  - Can be stored anywhere
  - Most languages provide automatic temporary directories and file names.
  - Be aware that applications can be executed by the “web” user, check permissions.
  - Should be deleted after use!!
- Beware of multiple concurrent users
  - Use request-specific file names for temporary data and results
  - Use user-based or process-based directories
  - Think in a [queueing system](#) for lengthy operations
- Collect statistics of use!!

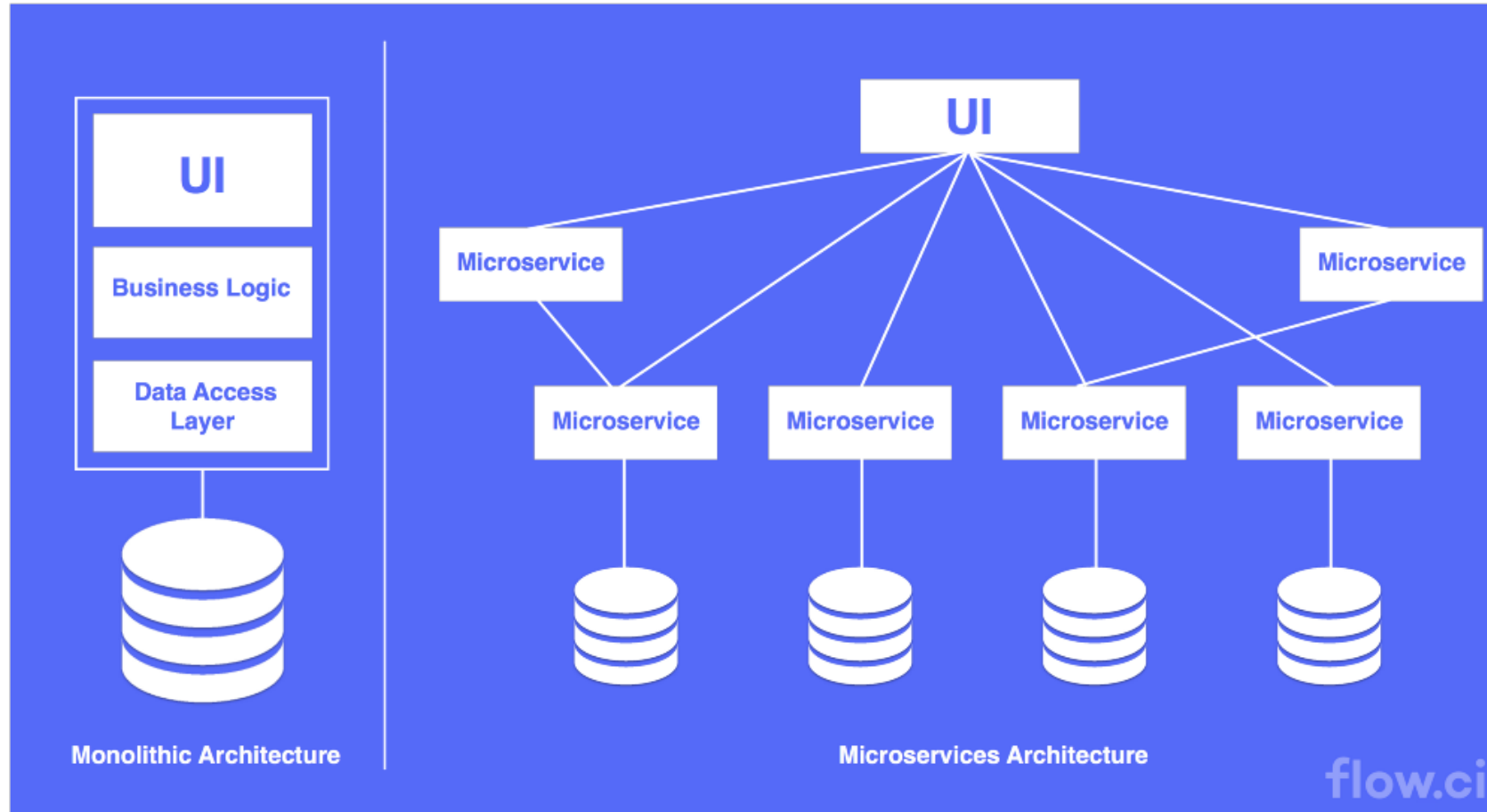


DBW - APIs

# APIs – Programmatic access

- API = “Application programming interface”
  1. Set of routines, functions, or procedures/methods, offered as a library to be used inside other software (API = “software library”)
  2. Any web service providing remote functionalities to be used inside other software or most usually as data provider (API = “web service”)
- In bioinformatics the approach is generally used to access data or to allow communication among application components (“microservices architecture”)
- Strategies: SOAP/XML, XML-RPC (remote executions), **REST**
- Authentication implicit via OAuth2 / openIDConnect

# Microservices architecture



# REST (REpresentational State Transfer)

- REST-ful web services
  - Used primarily to serve data
  - Data can be pre-processed at the server-side (so becoming a kind of RPC)
  - Controlled through HTTP and called using standard URLs
    - `/api/{store}/{id}/option.format?option`
  - HTTP interfaces to Data repositories
- HTTP based (GET, PUT, POST, DELETE)
  - Allow to GET, PUT (update), POST (insert), and DELETE a resource in the data portal (i.e. a DB)
- Using REST APIs, application components can be independent (and distributed) as long they communicate using HTTP and a known format

# Data exchange languages

- Data exchange formats
  - XML: Most traditionally used by web services (SOAP, RPC)
    - Same structure as HTML, but with no fixed tags
    - Requires XML-schema to specify tags and check coherence
  - JSON: Data interchange format replacing XML (most popular)
    - Natively understood by Javascript
- Both require “schemas” to validate data model

The logo for XML, featuring the letters 'xml' in red lowercase font, flanked by blue angle brackets '<' on the left and '>' on the right.The logo for JSON, featuring the letters 'JSON' in blue uppercase font, enclosed in blue curly braces '{' on the left and '}' on the right. The letter 'O' is stylized with a black and white gradient.

# Programmatic Access (client side)

- Perl

```
use LWP:Simple;  
use JSON;  
my $content = decode_json(get('http://...'));
```

- PHP

```
$data = json_decode(file_get_contents("http://...."));
```

- Python

```
import requests  
data = requests.get('http://...').json()
```

# Web service : server side

- Usual web applications but...
  - Output is not meant to be shown in browsers (no HTML, CSS, JS)
  - Headers required
    - `Content-type: text/xml | text/plain | application/json | application/x-gzip | image/png`
      - Define the type of data being sent
    - `Content-Disposition: attachment; filename=file_name`
      - Force download (when seen from a browser)
    - `Access-Control-Allow-Origin: *`
      - Allow access from any client (to avoid security checks on JS/AJAX)
  - Formats can change
    - In theory should be requested via HTTP (Accept...) but normally are included in the URL
  - Error handled via HTTP codes
    - 200 ok, 404 not found, ...
  - Prevent caching
    - Header: `Cache-Control: no-cache`
- Programming frameworks are very useful here due to the complex routing
- A “quite complex” backend : <http://mmb.pcb.ub.es/gitlab/MMBData/MMBApi>

# RESTful URLs

- No standard
  - A typical schema is  
`/api/{store}/{id}/option.format?options`

ex. `http://mmb.pcb.ub.es/api/pdb/2ki5/entry.json`

- Documented via
  - *Ad-hoc* help pages
  - OpenAPI (a.k.a. Swagger) (recommended)
    - [ICGC example](#)





## MMB Data repository API Interface

### Statistics

`/info[.xml|.json]` PDB/Uniprot data repository information (default json)

### Protein Data Bank

`/pdb/{id}/entry[.xml|.json]` PDB full entry metadata: Ascession Date, Experiment type, res chain ids, ligand data, remarks, chain sequences, sequence c (json). Individual fields can be recovered, completing the URI

#### Parameters

#### Usage example

```
/pdb/2ki5/entry? [.xml | .json]
```

SwissProt Hit on chain 0 (A) sequence.

```
/pdb/2ki5/entry/chains/0/swpHit/idHit
```

`/pdb/{id}[_bn{n}]` Headers from PDB file

`/headers[.gz]`

**Options** `{id}_bn{n}`: Biounit {n} instead of assymmetric unit



## API ENDPOINTS

### Overview

ICGC API is a set of RESTful endpoints -- programmable interfaces over the Web -- that allows third-party developers to build automation scripts and apps. This documentation describes ICGC API in details, including data model information of both inputs (parameters) and outputs (response records). It also allows you to interact with and test out each API directly on this page, which shall provide clear insights into how the API responds to different parameters.

### Endpoints

<https://dcc.icgc.org>

#### browser

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

#### PQL

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

#### analysis

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

POST	<code>/v1/analysis/enrichment</code>	Submits an asynchronous enrichment analysis request. Users must poll the status using the GET resource
GET	<code>/v1/analysis/enrichment/{analysisId}</code>	Retrieves an enrichment analysis by id
POST	<code>/v1/analysis/phenotype</code>	Creates a new Phenotype analysis by providing IDs of Donor entity sets.
GET	<code>/v1/analysis/phenotype/{analysisId}</code>	Retrieves the result of a phenotype analysis by its ID.
POST	<code>/v1/analysis/survival</code>	Creates a new Survival Plot analysis by providing IDs of Donor entity sets.
GET	<code>/v1/analysis/survival/{analysisId}</code>	Retrieves the result of a phenotype analysis by its ID.
POST	<code>/v1/analysis/union</code>	Creates a set analysis asynchronously. Status can be retrieved by polling the /{id} GET endpoint.
POST	<code>/v1/analysis/union/preview</code>	Retrieves a sample data of a set analysis as preview.

# Full entries and sequences

- `/api/pdb/{id}/entry/` `/api/pdbMonomer/{id}/entry/`
  - Full data in XML or JSON
- `/api/pdb/{id}.fasta`
- `/api/uniprot/{id}/entry`
- `/api/uniprot/{id}.fasta`

# PDB search options

- `/api/pdb/` Search on PDB
  - **resmin=value, resmax=value** Min Max for resolution (XRAY only)
  - **qcompType=(prot, nuc, prot-nuc, carb, other)** Compound types.
  - **qexpType=(ELECTRON\_CRYSTALLOGRAPHY, ELECTRON\_MICROSCOPY, FLUORESCENCE\_TRANSFER, INFRARED, NEUTRON\_DIFRACTION, NMR, SOLID-STATE\_NMR, X-RAY)** Type of Experiment.
  - **query=txt** Text query
    - **queryOn=(header, compound, sources, authors)**
  - **sequence=seq** Sequence match
    - **molTy=(protein | na)** Sequence type
    - **seqType=(exact | regex)** Type of sequence match (exact | regular expression )

# PDB options for structures

- `/api/pdb/{id}/ /api/pdbMonomers/{id}/`
  - Default: standard PDB coordinates(possible .gz)
  - Available filters
    - **bunit={n}** Show Biounit n instead of the Assimetric Unit
    - **noheaders=1** Skip PDB headers (implicit in the following filters)
    - **group=(ATOM | HETATM)** PDB label selection. (HETATM includes CONECT)
    - **groupRes=[!](POLAR | APOLAR | NUC | PROT)** Residue type selection. "!" negates
    - **groupAt=[!](POLAR | APOLAR | NOH | BACK | NABACK)** Atom type selection, "!" negates
    - **filter=[!][RES]nres:chain.atom/model** Atom filter using J(s)Mol format ("!" negates selection)