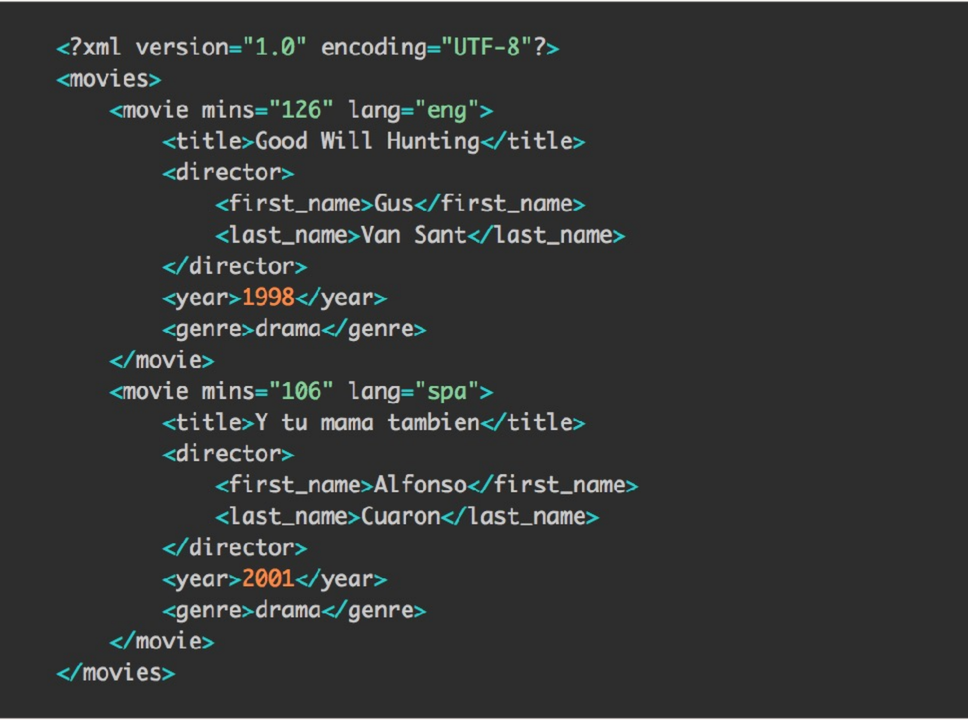
Basics of XML

Sources:
G.Sanchez, H.
Wickham



```xml
<?xml version="1.0" encoding="UTF-8"?>
<movies>
    <movie mins="126" lang="eng">
        <title>Good Will Hunting</title>
        <director>
            <first_name>Gus</first_name>
            <last_name>Van Sant</last_name>
        </director>
        <year>1998</year>
        <genre>drama</genre>
    </movie>
    <movie mins="106" lang="spa">
        <title>Y tu mama tambien</title>
        <director>
            <first_name>Alfonso</first_name>
            <last_name>Cuaron</last_name>
        </director>
        <year>2001</year>
        <genre>drama</genre>
    </movie>
</movies>
```

Basics of XML and HTML

Goal

XML & HTML

The goal of these slides is to give you a **crash introduction to XML and HTML** so you can get a good grasp of those formats for the rest of the lectures

Synopsis

In a nutshell

We'll cover a the following concepts:

- Importance of XML and HTML
- Hierarchical Structure

XML and HTML

Why you should care about XML and HTML?

- Large amounts of data and information are stored, shared and distributed using HTML and XML-dialects
- They are widely adopted and used in many applications
- Working with data from the Web means dealing with HTML

XML

eXtensible Markup Language

```xml
<?xml version="1.0" encoding="ISO8859-1" ?>
<CATALOG>
  <PLANT>
    <COMMON>Bloodroot</COMMON>
    <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$2.44</PRICE>
    <AVAILABILITY>031599</AVAILABILITY>
  </PLANT>

  <PLANT>
    <COMMON>Columbine</COMMON>
    <BOTANICAL>Aquilegia canadensis</BOTANICAL>
    <ZONE>3</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$9.37</PRICE>
    <AVAILABILITY>030699</AVAILABILITY>
  </PLANT>

  <PLANT>
    <COMMON>Marsh Marigold</COMMON>
    <BOTANICAL>Caltha palustris</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT> Mostly  Sunny </LIGHT>
    <PRICE> $6.81 </PRICE>
    <AVAILABILITY>   051799 </AVAILABILITY>
  </PLANT>

  <PLANT>
```

Some Definitions

> *"XML is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable"*

http://en.wikipedia.org/wiki/XML

> *"XML is a data description language used for describing data"*

Paul Murrell

Introduction to Data Technologies

Some Definitions

> *“XML is a very general structure with which we can define any number of new formats to represent arbitrary data”*

> *“XML is a standard for the semantic, hierarchical representation of data”*

Deb Nolan & Duncan Temple Lang

XML and Web Technologies for Data Sciences with R

About XML

XML

XML stands for **eXtensible Markup Language**

Broadly speaking ...

XML provides a flexible framework to create formats for describing and representing data

Markups

Markup

A **markup** is a sequence of characters or other symbols inserted at certain places in a document to indicate either:

- how the content should be displayed when printed or in screen
- describe the document's structure

Markup Language

A markup language is a system for **annotating** (i.e. *marking*) a document in a way that the content is distinguished from its representation (eg LaTeX, PostScript, HTML, SVG)

Markups

XML Markups

In XML (as well as in HTML) the marks (aka *tags*) are defined using angle brackets: < >

```
<mark>Text marked with special tag</mark>
```

Extensible

Extensible?

The concept of *extensibility* means that we can define our own marks, the order in which they occur, and how they should be processed. For example:

- <my_mark>
- <awesome>
- <boring>
- <pathetic>

About XML

XML is NOT

- a programming language
- a network transfer protocol
- a database

XML is

- more than a markup language
- a generic language that provides structure and syntax for representing any type of information
- a meta-language: it allows us to create or define other languages

Minimalist Example

XML Example

Ultra Simple XML

```
<movie>
  Good Will Hunting
</movie>
```

- one single element *movie*
- start-tag: <movie>
- end-tag: </movie>
- content: Good Will Hunting

XML Example

Ultra Simple XML

```
<movie mins="126" lang="en">
   Good Will Hunting
</movie>
```

- xml elements can have **attributes**
- attributes: mins (minutes) and lang (language)
- attributes are *attached* to the element's start tag
- attribute values **must be quoted!**

XML Example

Minimalist XML

```
<movie mins="126" lang="en">
   <title>Good Will Hunting</title>
   <director>Gus Van Sant</director>
   <year>1998</year>
   <genre>drama</genre>
</movie>
```

- an xml element may contain other elements
- *movie* contains several elements: *title, director, year, genre*

XML Example

Simple XML

```
<movie mins="126" lang="en">
  <title>Good Will Hunting</title>
  <director>
    <first_name>Gus</first_name>
    <last_name>Van Sant</last_name>
  </director>
  <year>1998</year>
  <genre>drama</genre>
</movie>
```

- Now *director* has two child elements: *first_name* and *last_name*

XML Hierarchy Structure

Conceptual XML

```
<Root>
  <child_1>...</child_1>
  <child_2>...</child_2>
    <subchild>...</subchild>
  <child_3>...</child_3>
</Root>
```

- An XML document can be represented with a **tree structure**
- An XML document must have **one single Root** element
- The Root may contain child elements
- A child element may contain subchild elements