Obtaining data

- Typically obtained from the internet automatically using UNIX scripts
  - Commonly used languages
    - Shells (bash, tcsh, etc.)
    - Perl
    - Python
    - Ruby
- Services often provide protocols for data access
  - e.g. Facebook, Twitter
- Standard parsers for standard data formats
  - Apache Tika: extracts metadata and text from >100 different file types
  - e.g. PPT, XLS, PDF, XML, ODF, EPUB, ZIP, MP4, FLV, …

Obtaining data *(cont.)*

- Web-scraping
  - complicated due to the need for user-interaction in retrieving online documents, especially from DHTML-rich websites
    - e.g. Some pages load only when you scroll to the end of the page
  - Many tools allow us to mimic user-interaction as if accessed through a web-browser
  - **Scripting environment**
    - Selenium Webdriver (Java, Python, C#, Groovy, Perl, PHP, Ruby)
    - PhantomJS (Javascript)
    - Capybara (Ruby)
    - Zombie.js (Javascript)
    - WebHarvest (Java)
    - *etc*
  - **Learn from user-interaction**
    - Selenium IDE
    - Commercial ($$$): Mozenda, Import.io
    - *etc*

Obtaining data *(cont.)*

- Online image repositories

- Stalk your friends