UNIVERSIDAD ADOLFO IBÁÑEZ

Spatial Analytics Toolbox
Satellite Imagery, Web Scraping, Hedonic Modelling

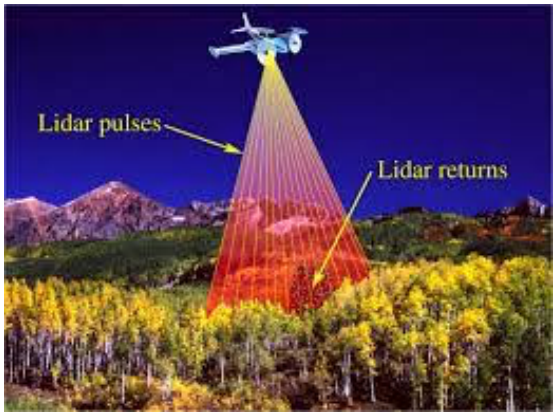
Esteban Lopez Ochoa
Master in Business Analytics

11/05/2020

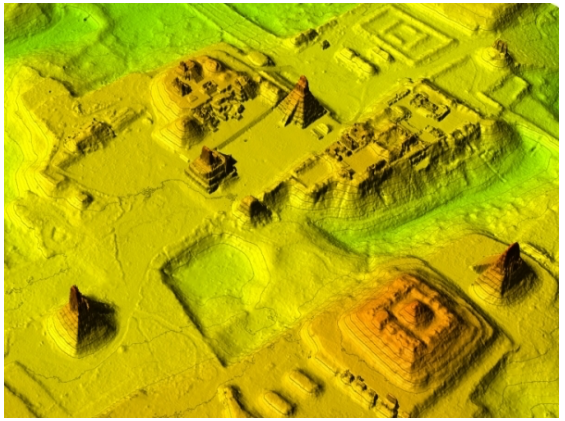
UNIVERSIDAD ADOLFO IBÁÑEZ

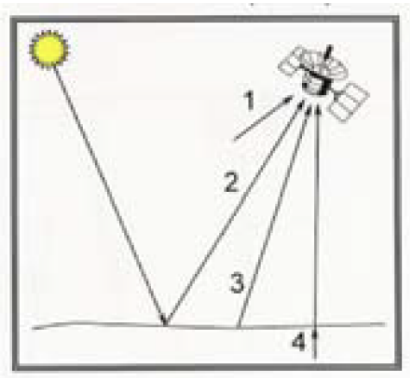
SATELLITE IMAGERY

What's remote sensing?



- "… gathering information from a distance"
- "… detec and analyze features and phenomena on Earth's surface"
- Types
  - Active Sensors – propagate energy, record response
    - Flash photos, LiDAR, Radar, Sonar
  - Passive sensors – record reflected or emitted energy
    - Aerial cameras, most common satellites





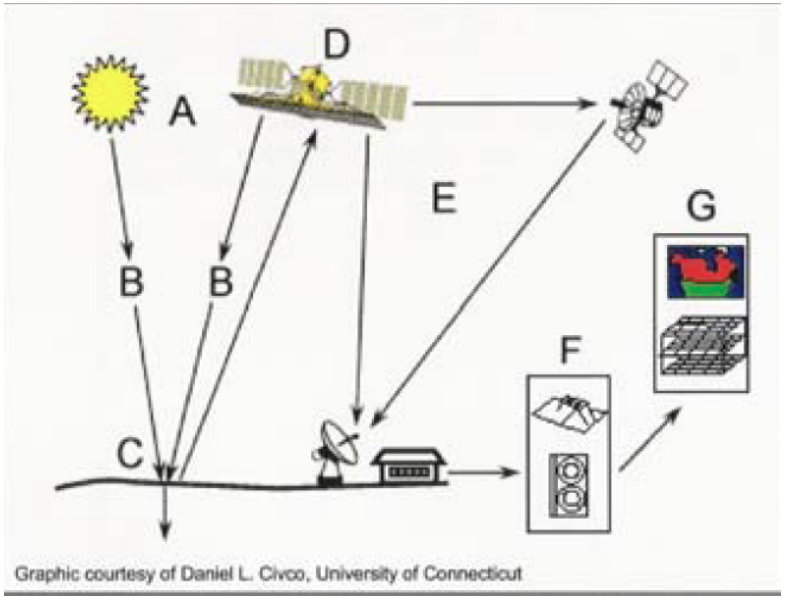
UNIVERSIDAD ADOLFO IBÁÑEZ

Active vs. Passive RS

Type	Active	Passive
Frequency	At will	Depending on instrument
Level of Detail	High	Depends on Instrument, but generally medium-low
Accuracy	Needs rectification - Earth surface is uneven - Elevation differences - Tilt of the camera in the plane and the plane itself	Earth centered datum, doesn't require much more work
Costs	High	Low-Medium
Availability	Low	High

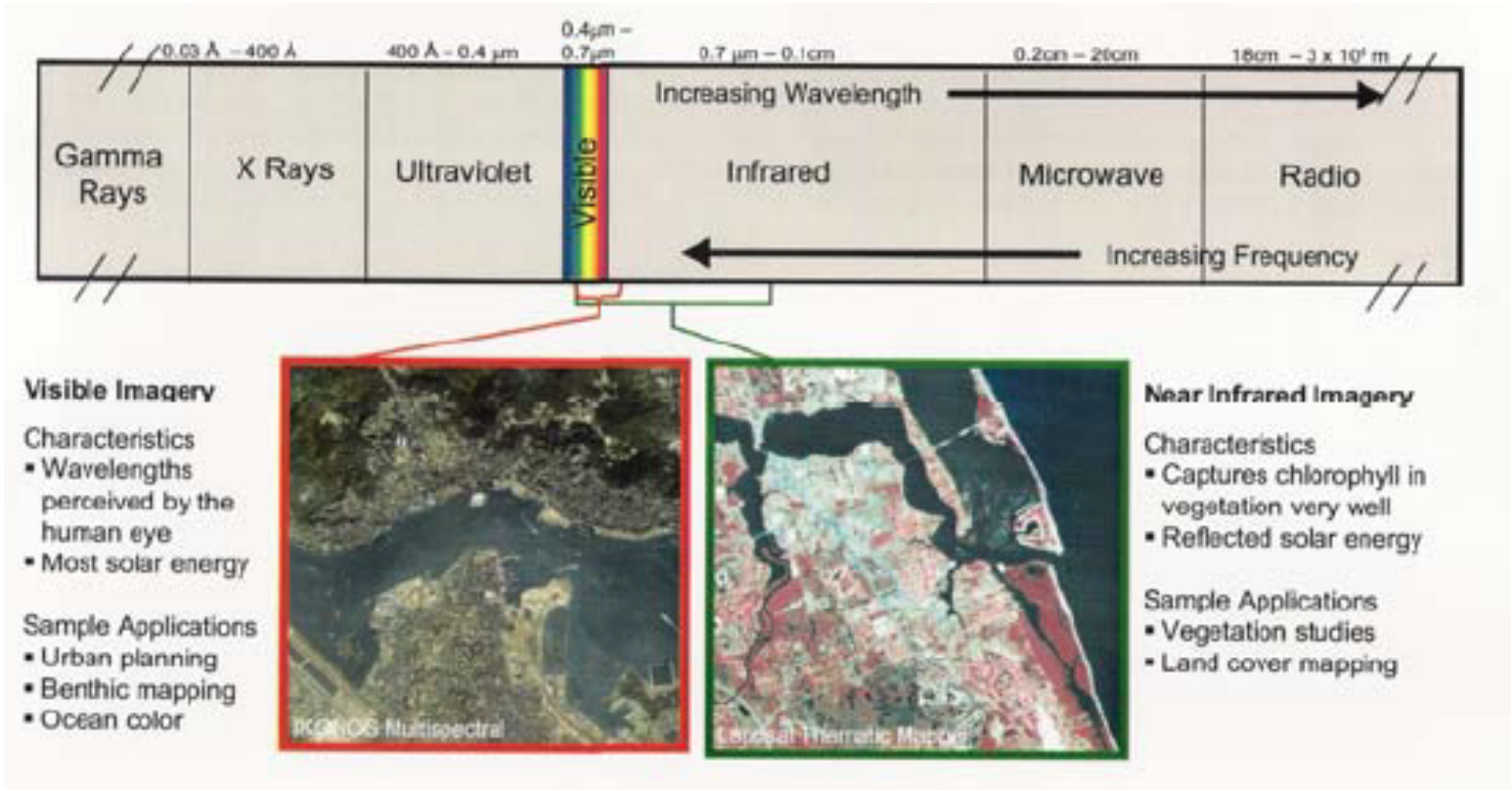
RS: Passive sensors: Satellites

- Light has both wave and particle characteristics
- Sensors detect variations in electromagnetic energy
  - A: Energy source / illumination
  - B: Radiation / atmospheric surface
  - C: Interaction with surface
  - D: Sensor records energy
  - E: Transmission, reception, processing
  - F: Interpretation/analysis
  - G: Application



RS: Passive sensors: Satellites

- Satellites detect different wave lengths of the electromagnetic spectrum



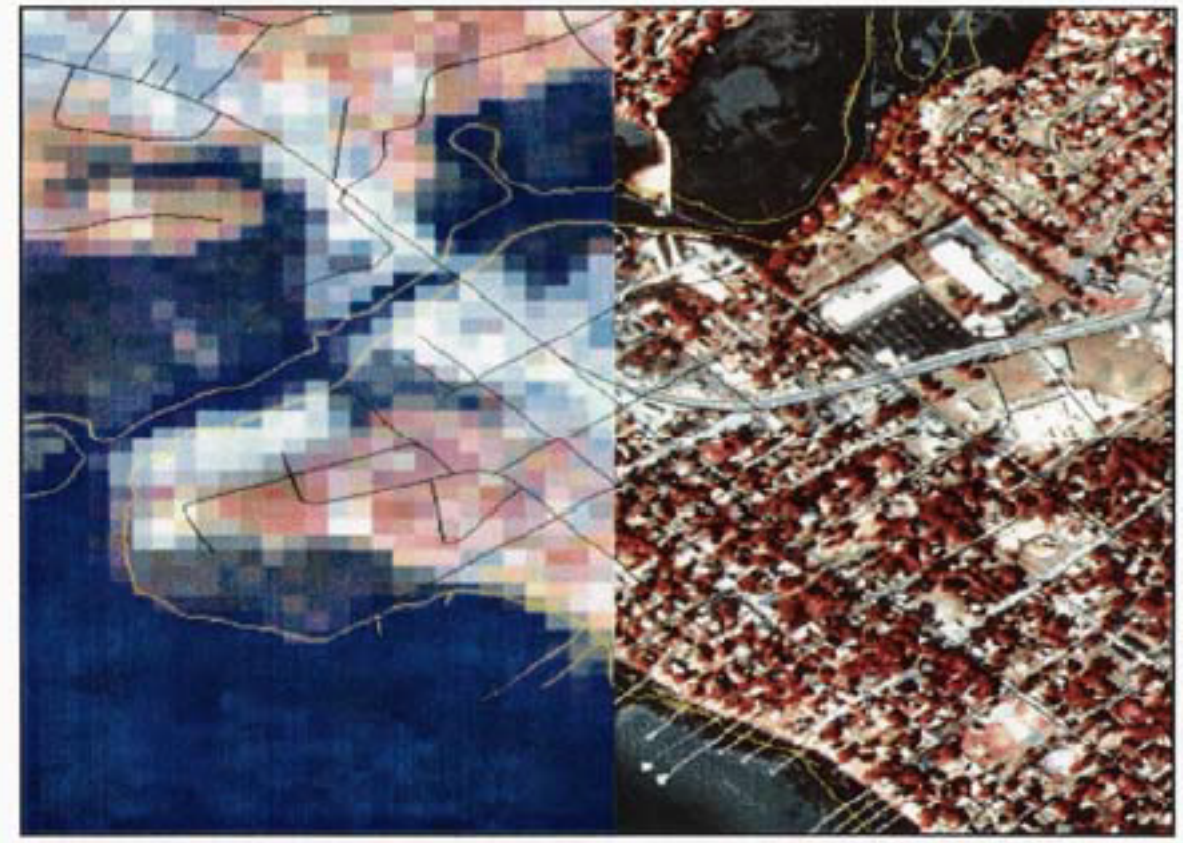
UNIVERSIDAD ADOLFO IBÁÑEZ

Platforms and Sensors: Space

Satellite Name	Sensor	Sample Applications
★ SPOT	HRV, panchromatic	shallow water mapping, vegetation mapping
★ IKONOS 1	IKONOS	land cover, higher resolution feature extraction
RADARSAT-1	synthetic aperture radar (SAR)	land cover, sea ice, oil spill detection
★ Landsat-7	enhanced thematic mapper (ETM+)	land cover, suspended sediments, surface temperature
Seastar	SeaWiFs	ocean color
GOES-east	sounder, imager	sea surface temperature, atmospheric properties
NOAA TIROS 16	AVHRR	sea surface temperature, normalized difference vegetation index (NDVI)
TERRA	MODIS, ASTER	ocean color, water vapor, sea surface temperature
★ QuickBird	multispectral, panchromatic	land cover, higher resolution feature extraction

UNIVERSIDAD ADOLFO IBÁÑEZ

Platforms and Sensors: Space



Landsat image with 30-meter resolution.

Digital orthophoto with 1-meter resolution.

UNIVERSIDAD ADOLFO IBÁÑEZ

Platforms and Sensors: Space

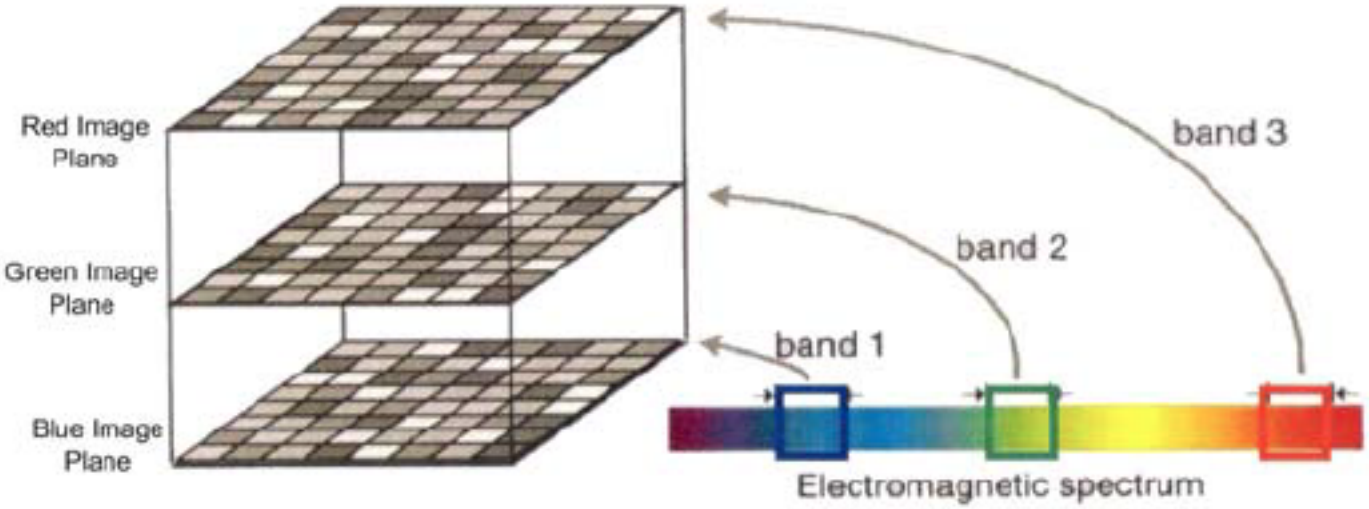
Small scale, 1:100,000. Greater area covered, less detail; features appear smaller.

Large scale 1:12,000. Smaller area covered, more detail; features appear larger.

UNIVERSIDAD ADOLFO IBÁÑEZ

Spectral Resolution

- Number of bands or channels detected by a sensor
- Computers display multi-band images by assigning the bands to 1 of 3 image planes: Red, Green Blue

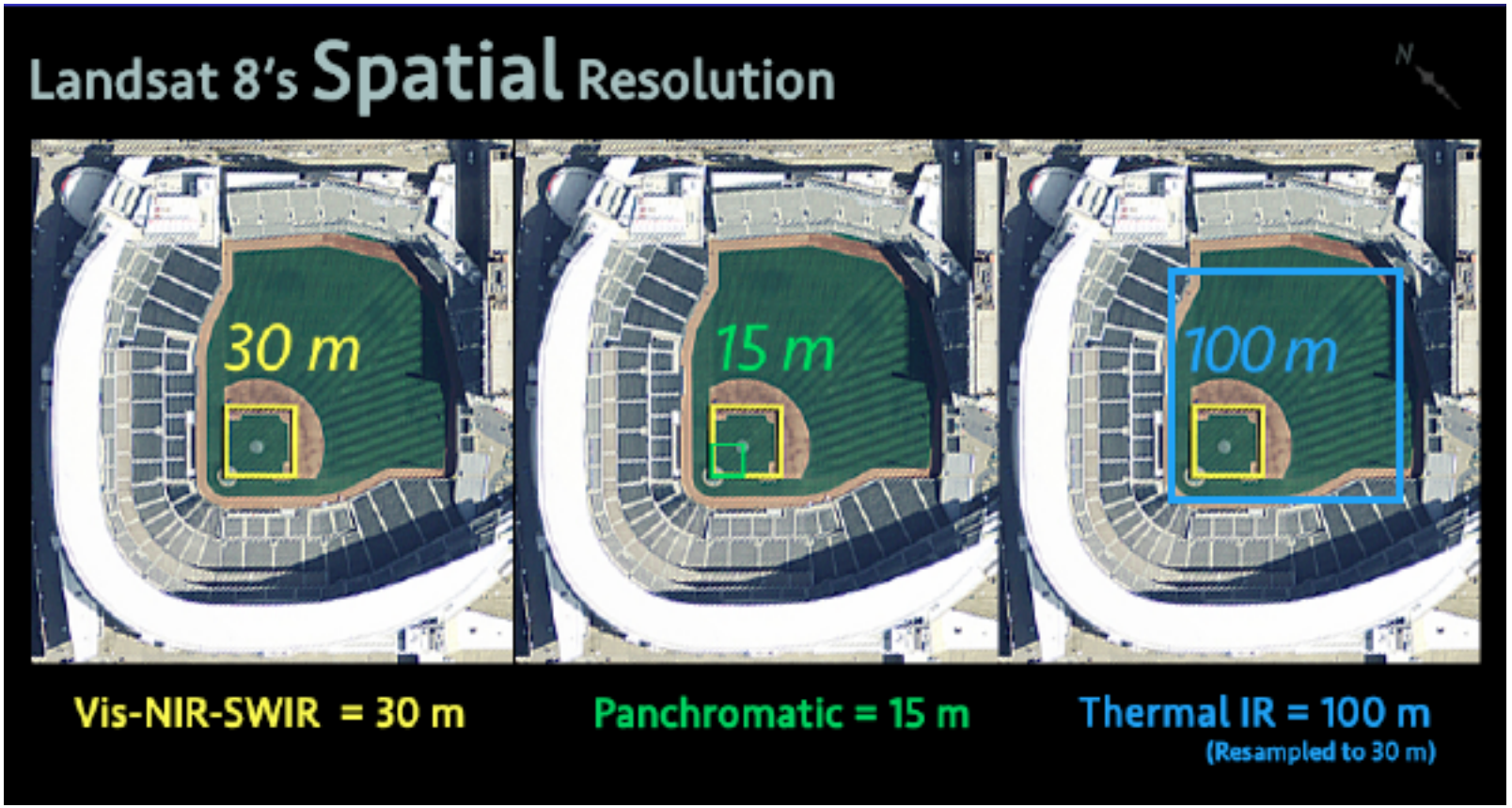


UNIVERSIDAD ADOLFO IBÁÑEZ

Landsat7 vs Landsat8

Landsat-7 ETM+ Bands (μm)			Landsat-8 OLI and *TIRS* Bands (μm)		
			30 m Coastal/Aerosol	0.435 - 0.451	Band 1
Band 1	30 m Blue	0.441 - 0.514	30 m Blue	0.452 - 0.512	Band 2
Band 2	30 m Green	0.519 - 0.601	30 m Green	0.533 - 0.590	Band 3
Band 3	30 m Red	0.631 - 0.692	30 m Red	0.636 - 0.673	Band 4
Band 4	30 m NIR	0.772 - 0.898	30 m NIR	0.851 - 0.879	Band 5
Band 5	30 m SWIR-1	1.547 - 1.749	30 m SWIR-1	1.566 - 1.651	Band 6
Band 6	60 m TIR	10.31 - 12.36	*100 m TIR-1*	*10.60 – 11.19*	Band 10
			100 m TIR-2	*11.50 – 12.51*	Band 11
Band 7	30 m SWIR-2	2.064 - 2.345	30 m SWIR-2	2.107 - 2.294	Band 7
Band 8	15 m Pan	0.515 - 0.896	15 m Pan	0.503 - 0.676	Band 8
			30 m Cirrus	1.363 - 1.384	Band 9

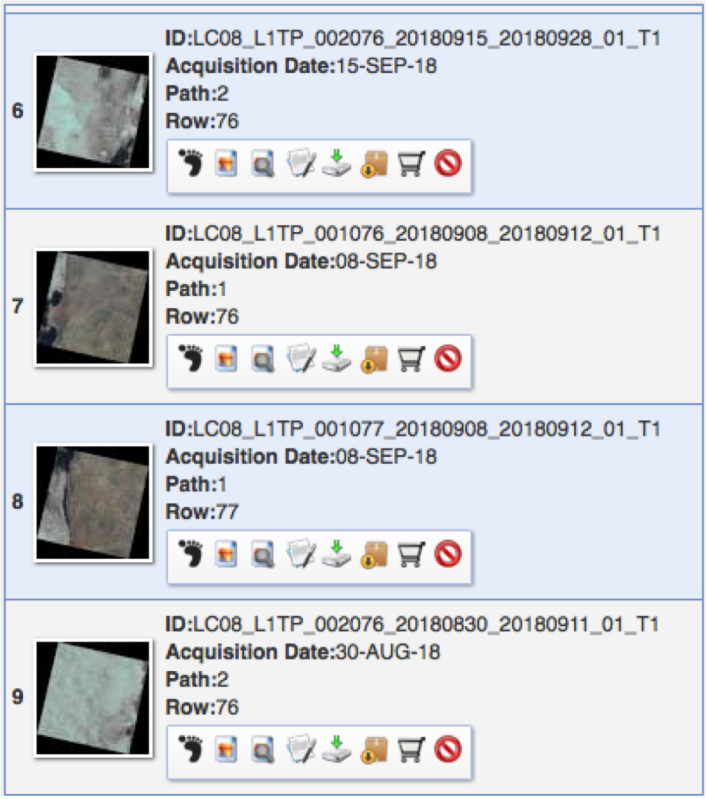
Landsat8

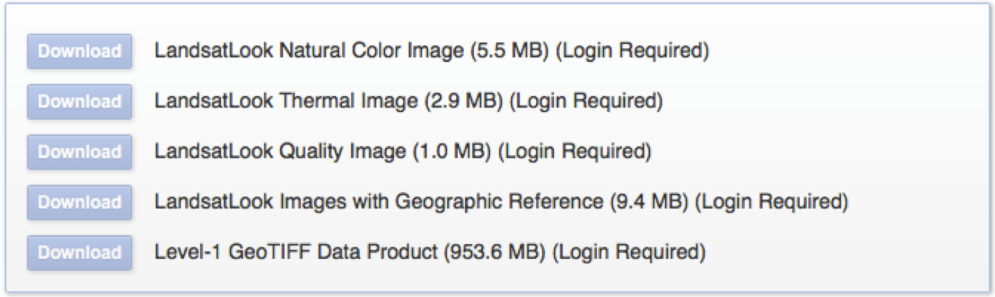


Source: NASA.2016. Landsat 8 Overview

Getting Landsat8 Imagery

- Step 1: Register at: https://ers.cr.usgs.gov/login
- Step 2: Login to: https://earthexplorer.usgs.gov
- Step 3: Choose method of selection of area
- Step 4: Browse specific date of Image you want to download
  - Watch for the presence of clouds
- Step 5: Download (usually about 1GB)





Basic Processing

- Steps:
  1. Load the data
     1. Create a list of filenames
  2. Stack the data
  3. Crop the data
  4. Visual analysis
     1. Image RGB plotting of different bands
  5. Index Calculation
  6. Polygon creation

Common Possibilities of Band Combinations

Natural Color	4 3 2
False Color (urban)	7 6 4
Color Infrared (vegetation)	5 4 3
Agriculture	6 5 2
Atmospheric Penetration	7 6 5
Healthy Vegetation	5 6 2
Land/Water	5 6 4
Natural With Atmospheric Removal	7 5 3
Shortwave Infrared	7 5 4
Vegetation Analysis	6 5 4

Source: ESRI

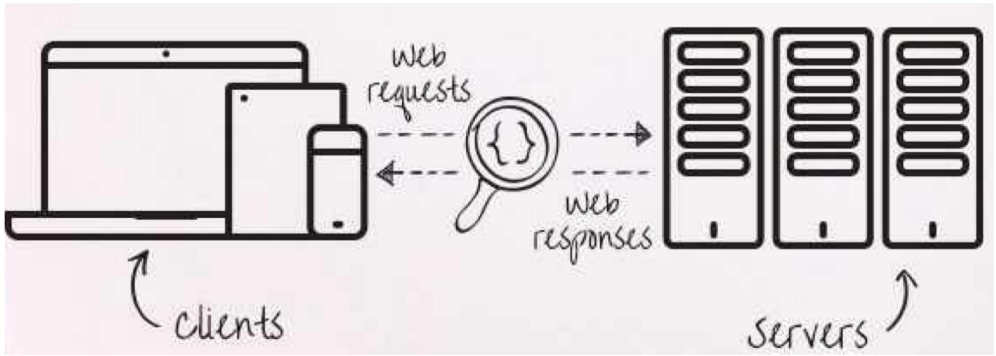
UNIVERSIDAD ADOLFO IBÁÑEZ

WEB SCRAPING

Introduction

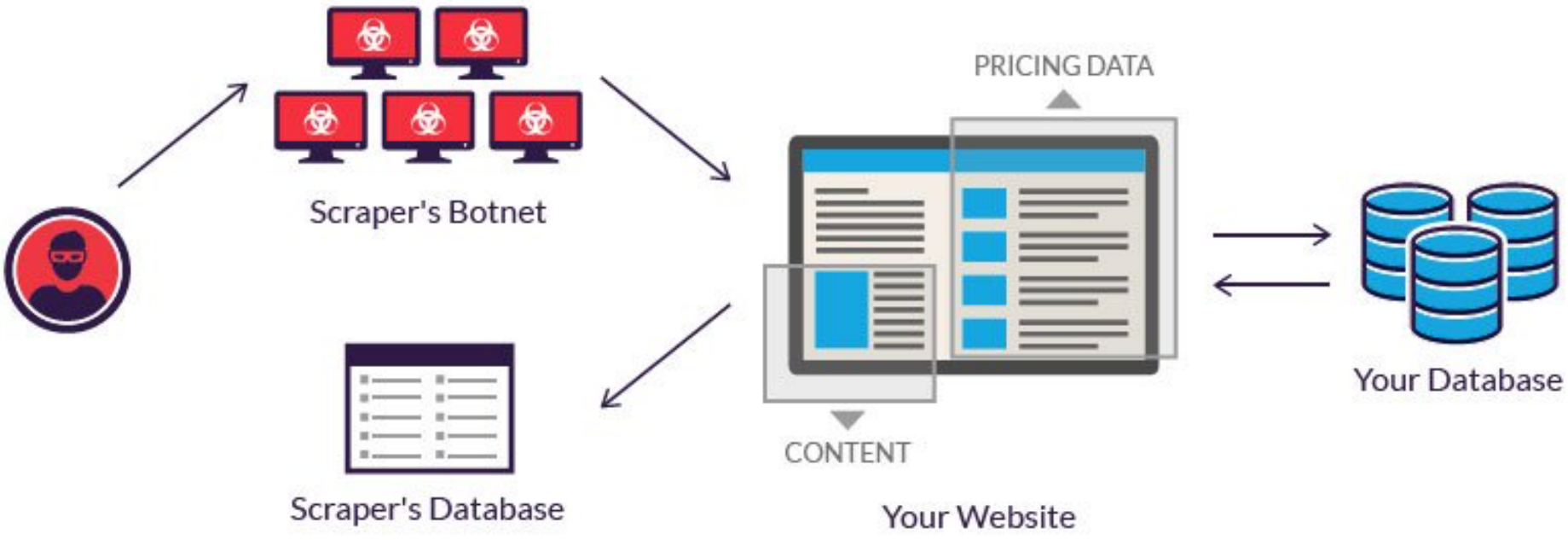
- Web-scraping is an ‘artisan’ method to scrape data from websites in a automated setting

- Basic principle:
  - Free access to information by point-n-click
  - Automatization of point-n-click
- Ways to scrape data
  - Human copy-paste
  - Text pattern matching
  - API interface
- Basic knowlegde:
  - HTML or PHP web coding
- WARNING
  - Not all websites are scrape-friendly



UNIVERSIDAD ADOLFO IBÁÑEZ

General process and how companies view it



UNIVERSIDAD ADOLFO IBÁÑEZ

Web-Scraping Process

- Step 1: Identify website to scrape from
  - Main characteristic: well-structured URL address
- Step 2: Make a list of the elements you wish to scrape
- Step 3: Design what how would you like your data to look like based on Step 2
- Step 4: Take one-webpage example to start creating your code
  - Identify elements in the HTLM context → CSS selector
  - Parse the HTML element to text format
  - Organize elements in your data object
- Step 5: Construct a loop
  - Depends on how the website is organized
- Step 6: Create safety measures and save the data
  - Add time in-between scrapings
  - Scrape at times that are unlikely to become a burden

UNIVERSIDAD ADOLFO IBÁÑEZ

Application in R

- Objectives:
  - Lear how to scrape information from a scrape-friendly website
  - Store some data in order to conduct hedonic analysis
- Steps
  - Browse and familiarize yourself with the website
  - Create URLs and download the site to R
  - Create scraping code for a single example
  - Doing a loop to scrape the whole website
  - Assessing scraping and data quality

UNIVERSIDAD ADOLFO IBÁÑEZ

Bringing it all together

HEDONIC MODELLING

UNIVERSIDAD ADOLFO IBÁÑEZ

Introducción

- Hedonic Modelling is one of the most common techniques to analyze urban issues in the WTP context
- Basic Idea: Decompose the price of a good in the different parts that compose it.
  - Recover a shadow price for each variable that compose the good
  - Where: $y = f(X, S, Z)$
    - y: price of the good
    - X: intrinsic characteristics of the good
    - S: Spatial characteristics of the good
    - Z: Unobservable characteristics of the good

UNIVERSIDAD ADOLFO IBÁÑEZ

Issues in Spatial Data Analysis

McMillen (2010, JRS)

- Main Challenges
  - NEVER a data set is good enough that contains all independent variables affecting the dependent variables
  - Variables also need to be measured properly
  - Functional form needs to be correctly specified.
    - Any or more of these always fail to happen
- Additional Challenges
  - Spatially correlated missing variables
- Common approach
  - Begin with simple functional form
  - Test for spatial autocorrelation
  - Estimate a spatial parametric model
- Main Problem:
  - We are likely going to fail to identify the source of the spatial autocorrelation

UNIVERSIDAD ADOLFO IBÁÑEZ

Issues in Spatial Data Analysis

McMillen (2010, JRS)

- Solutions:
  - Direct: Include the omitted variable in the hedonic model
  - Indirect: Assess the robustness of the results to alternative model specifications
    - Main idea: If results hold, we can trust them better than if not
    - Spatial Econometrics can be viewed as an alternative model specification
- BUT!
  - Spatial Econometric Models are mostly estimated using Maximum Likelihood assuming that the 1) functional form and 2) error distributions ***are know in advance***
    - Source of spatial autocorrelation is often unknown
    - Large samples are a problem because of dimensionality, but ML methods require large samples to produce accurate results

UNIVERSIDAD ADOLFO IBÁÑEZ

Issues in Spatial Data Analysis

McMillen (2010, JRS)

- Shifting the focus
  - Accepting that obtaining efficient and consistent estimates is virtually impossible
  - Spatial Econometric models become just another tool to guide model specification and robustness of results
- Nonparametric and Semiparamentric Models
  - A great alternative when the objective is to guide policy making → model specification becomes the central issue
    - Starting point: Model structure is unknown → relax assumptions
    - Even more accurate causal estimations can be obtain if a well-defined area (zoning, school boundaries, etc.) is analyzed

UNIVERSIDAD ADOLFO IBÁÑEZ

Parametric vs. Nonparametric

- Parametric Functional forms
  - SAR $\boldsymbol{Y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X}\beta + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{u}$
  - SEM $\boldsymbol{Y} = \boldsymbol{X}\beta + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{e}$

- Non Parametric
  - Kernel Regression:
    - Identical to WLS regression

$$\min(\alpha) = \sum_{i=1}^{n} (y - \alpha)^2 K(\psi_i) y_i$$

$$, \psi_i = \frac{x_i - x}{h}$$

UNIVERSIDAD ADOLFO IBÁÑEZ

Parametric vs. Nonparametric

- Non Parametric
  - Locally Weighted Regression (LWR)

$$\sum_{i=1}^{n}(y_i-\alpha-\beta'(x_i-x))^2K(\psi_i)$$

$$Z_i=\begin{pmatrix}1\\x_i\end{pmatrix}\qquad\theta=\begin{pmatrix}\alpha\\\beta\end{pmatrix}$$

$$\hat{\theta}(x)=(\textstyle\sum_{i=1}^{n}K(\psi_i)Z_iZ_i')^{-1}\sum_{i=1}^{n}K(\psi_i)Z_i'y_i$$

A regression of wy on wz, where

$$w_i=K\left(\frac{x_i-x}{h}\right)^{1/2}$$

UNIVERSIDAD ADOLFO IBÁÑEZ

Parametric vs. Nonparametric

- Non Parametric
  - LWR with two variables

$$K(\psi_i) = k\left(\frac{x_{1i}-x_1}{h_1}\right) k\left(\frac{x_{2i}-x_2}{h_2}\right)$$

$$Z_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \end{pmatrix} \qquad \theta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\hat{\theta}(x) = \left(\sum_{i=1}^{n} K(\psi_i) Z_i Z_i'\right)^{-1} \sum_{i=1}^{n} K(\psi_i) Z_i' y_i$$

A WLS regression of y on 1, x1, x2

UNIVERSIDAD ADOLFO IBÁÑEZ

Parametric vs. Nonparametric

- Non Parametric
  - Conditionally Parametric Model (CPAR)
  - Same as LWR but only some variables enter the Kernel function

    - LWR $K(\psi_i) = k\left(\frac{x_{1i}-x_1}{h_1}\right) k\left(\frac{x_{2i}-x_2}{h_2}\right)$

    - CPAR $K(\psi_i) = k\left(\frac{x_{1i}-x_1}{h_1}\right)$

UNIVERSIDAD ADOLFO IBÁÑEZ

Parametric vs. Nonparametric

- Non Parametric
  - Geographically Weighted Regression (GWR)
    - SPECIAL case of CPAR

Kernel weights are a function of straight-line distance between observations

Let z_1= longitude, z_2 = latitude

CPAR: Each coefficient varies by z_1 and z_2

$$\sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \, k\left(\frac{z_{1i}-z_1}{h_1}\right) k\left(\frac{z_{2i}-z_2}{h_2}\right)$$

GWR: Coefficients by straight-line distance between each observation and the target point

$$\sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \, k\left(\frac{d_i}{h}\right)$$

UNIVERSIDAD ADOLFO IBÁÑEZ

Application in R

- Objective:
  - Evaluate model specification in a hedonic setting of housing prices on housing characteristics
  - Compare functionality between OLS and LWR and CPAR

- Steps:
  - Calculate distances and arrange data
  - OLS
  - CPAR
  - Comparison
  - Visualization