Intro to Spatial Data Science with R

Alí Santacruz

amsantac.co

JULY 2016

About me

- Expert in geomatics with a background in environmental sciences
- R geek
- PhD candidate in Geography
- Interested in *Spatial Data Science*
- Author of several R packages (available on CRAN)

Purpose of this talk

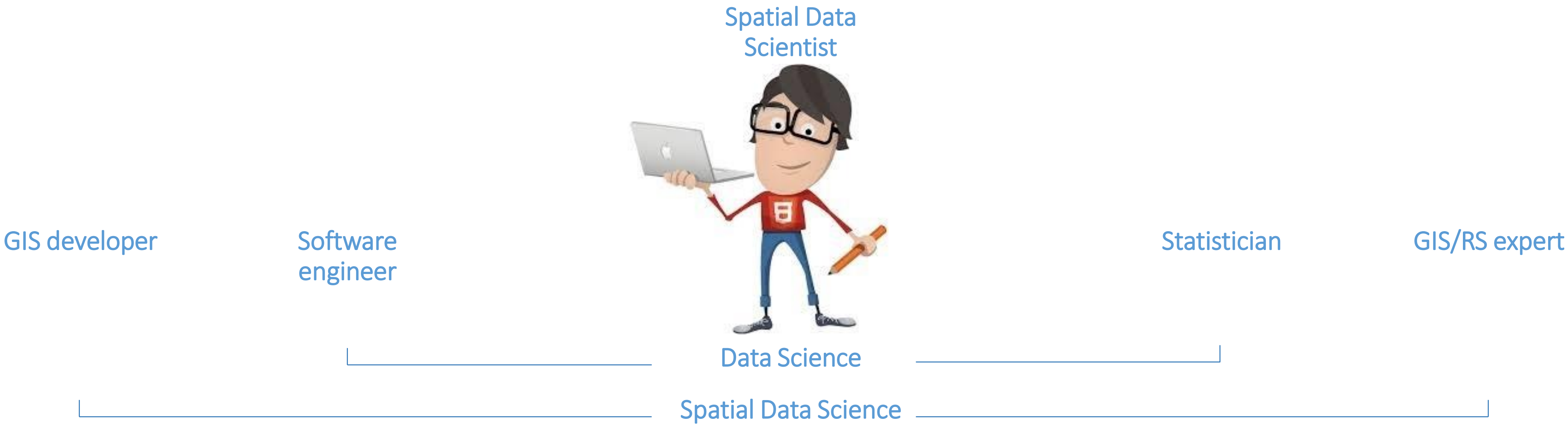
- Discuss what *Spatial Data Science* is

- Give an introductory explanation about how to conduct *Spatial Data Science* with R

What is Spatial Data Science?

Spatial Data Scientist (n.):

Person that is better in spatial data analysis than a GIS developer and better in software engineering than a GIS/RS expert



Spatial Data Scientist

GIS developer

Software engineer

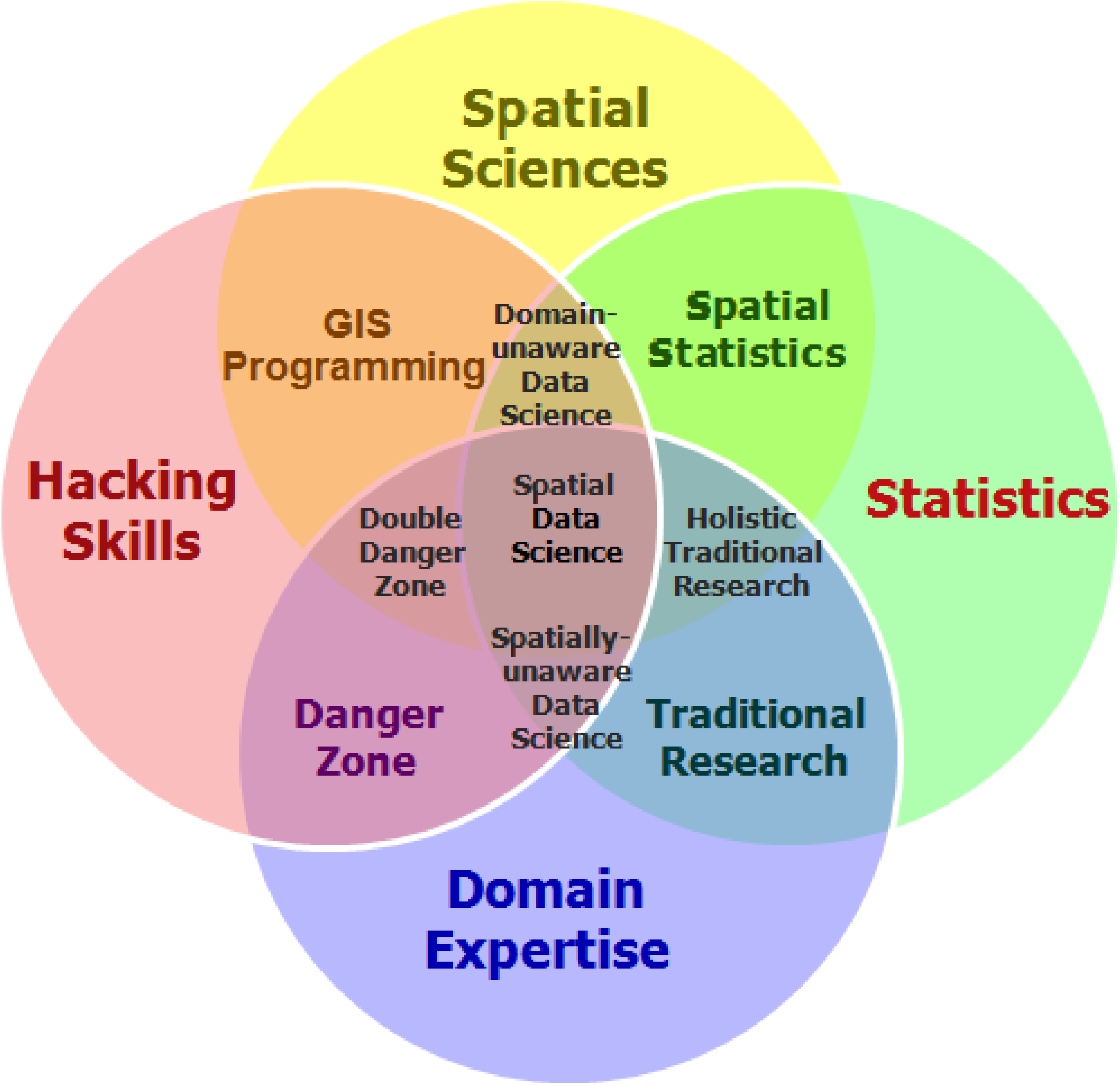
Statistician

GIS/RS expert

Data Science

Spatial Data Science

Spatial Data Science



All they are combined for **data analysis** in order to …

Support a better decision making

"The key word in data science is not data; it is science"
Jeff Leek. Data Science Specialization. Coursera.

Spatial Data Scientist

MATH & STATISTICS

- Machine learning
- Spatial statistics
- Statistical modeling
- Experiment design
- Statistical inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

SPATIAL SKILLS

- Spatial data structures
- Geodesy
- Spatial analysis
- Spatial data infrastructures and standards
- GIS & Remote sensing procedures and technologies
- Cartography
- Photogrammetry

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate and knowledgeable about the business
- Curious about data
- Problem solver
- Strategic, proactive, creative, innovative and collaborative
- Story telling skills
- Translate data-driven insights into decisions and actions
- Able to engage with senior management

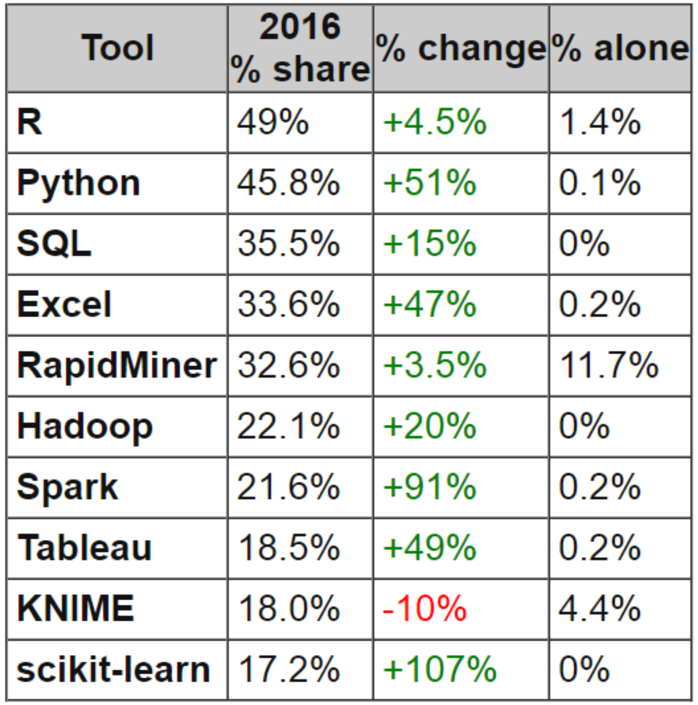
HACKING SKILLS

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Spatial databases
- Parallel databases and parallel query processing
- MapReduce concepts
- Distributed storage and processing frameworks e.g. Hadoop, Spark
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

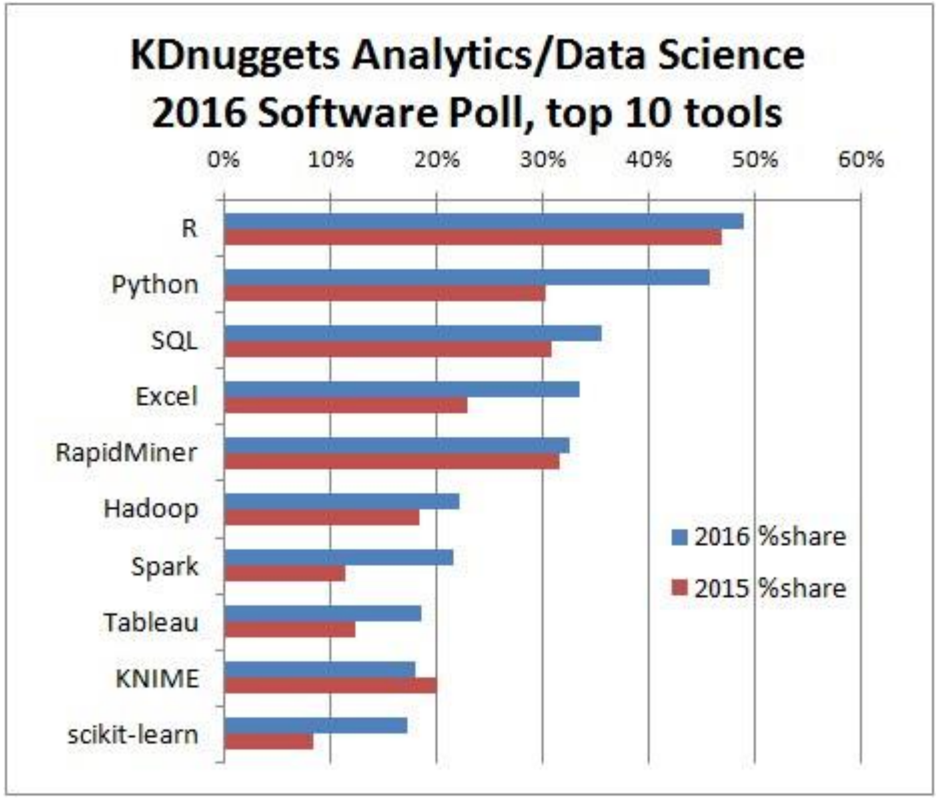
Modified from gettingsmart.com

Hacking skills

- **Programming languages:** Python and R (and others)



Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%



http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Why should we use R

- Free and open-source
- A large and comprehensive set of packages (> 8600)
  - Data access
  - Data cleaning
  - Analysis
  - Visualization and report generation
- Excellent development environments – RStudio IDE
- An active and friendly developers community
- A huge users community: > 2 million

Why R for spatial analysis

- 160+ packages in CRAN Task View: Analysis of Spatial Data
    - Classes for spatial (and spatio-temporal) data
    - Spatial data import/export
    - Exploratory spatial data analysis
    - Support for vector and raster operations
    - Spatial statistics
    - Data visualization through static and dynamic (web) graphics
    - Integration with GIS software
    - Easy integration with techniques from non-spatial packages

R classes for spatial data

- Before 2003:
    - Several packages with different assumptions on how spatial data was structured

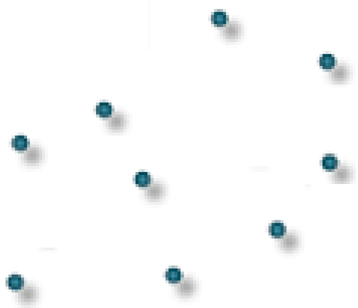
- From 2003:
    - ‘sp’ package: extends R classes and methods for spatial data (vector and raster)

- From 2010:
    - ‘raster’ package: deals with raster files stored in disk that are too large to be loaded on memory (RAM)

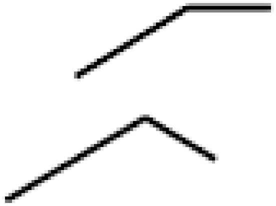
R classes for spatial data

sp package

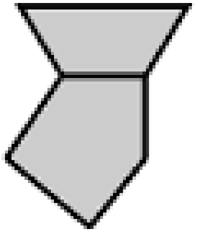
SpatialPointsDataFrame



SpatialLinesDataFrame

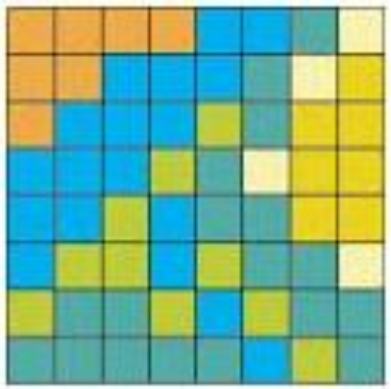


SpatialPolygonsDataFrame



SpatialPixelsDataFrame

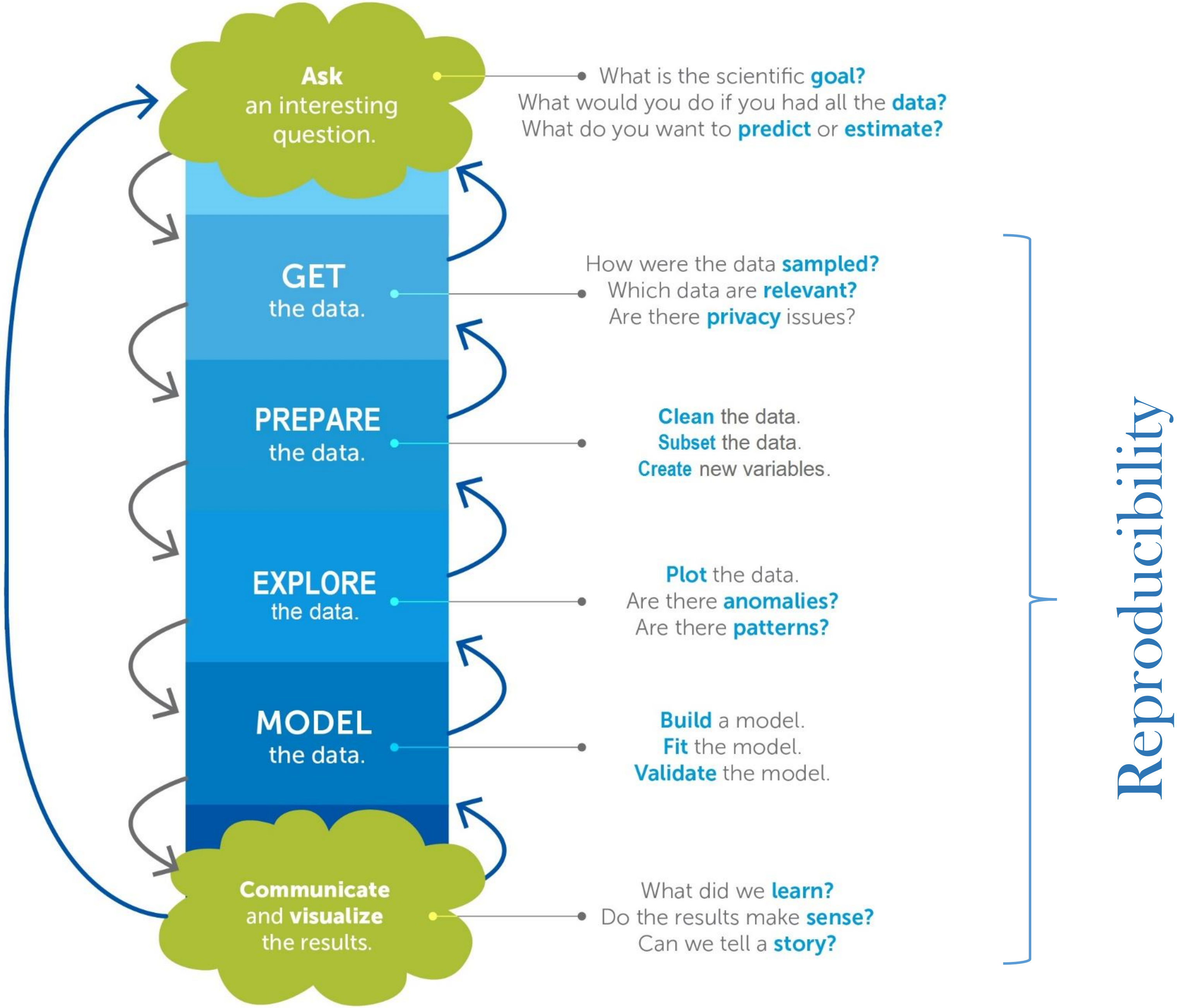
SpatialGridDataFrame



raster package
(recommended)

RasterLayer
RasterStack
RasterBrick

The **Data Science** Process



Ask an interesting question.

- What is the scientific **goal?**
- What would you do if you had all the **data?**
- What do you want to **predict** or **estimate?**

GET the data.

- How were the data **sampled?**
- Which data are **relevant?**
- Are there **privacy** issues?

PREPARE the data.

- **Clean** the data.
- **Subset** the data.
- **Create** new variables.

EXPLORE the data.

- **Plot** the data.
- Are there **anomalies?**
- Are there **patterns?**

MODEL the data.

- **Build** a model.
- **Fit** the model.
- **Validate** the model.

Communicate and **visualize** the results.

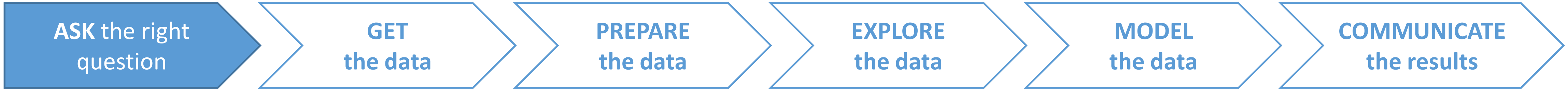
- What did we **learn?**
- Do the results make **sense?**
- Can we tell a **story?**

Modified from science2knowledge

Domain expertise

- **Is this A or B or C?** :: classification
- **Is this weird?** :: anomaly detection
- **How much/how many?** :: regression
- **How is it organized?** :: clustering
- **How will it change?** :: prediction

"The key word in data science is not data; it is science"
Jeff Leek. Data Science Specialization. Coursera.



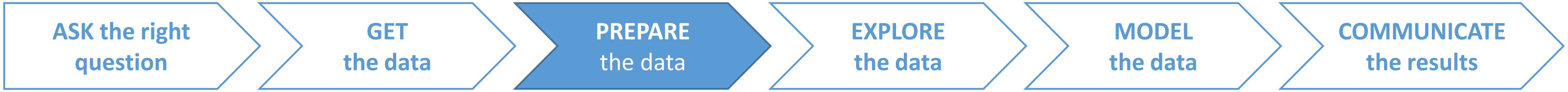
ASK the right question → **GET** the data → **PREPARE** the data → **EXPLORE** the data → **MODEL** the data → **COMMUNICATE** the results

- **Import vector layers:** rgdal, raster packages
- **Import raster layers:** raster package
- **Get geocoded data from APIs:** twitteR package, see example
- **Download satellite images/geographic data:** raster, modis, MODISTools packages

For this slide and following ones see code and examples in in this webpage

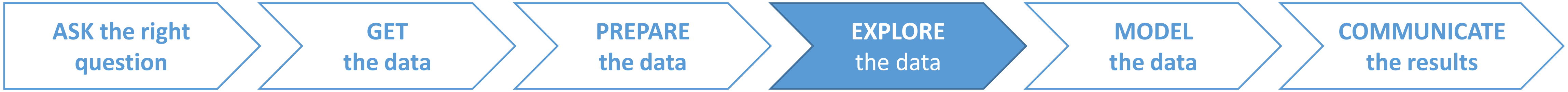
- **Data cleaning, subset, etc.**
    - Manipulate data with “verbs” from dplyr and other Hadley-verse packages
    - Spatial subset (sp, raster packages)
- **Vector operations:**
    - Operations on the attribute table (sp package)
    - Overlay: union, intersection, clip, extract values from raster data using points/polygons (raster, rgeos packages)
    - Dissolve (sp, rgeos packages), buffer (rgeos package)
    - Rasterize vector data (raster package)
- **Raster operations:**
    - Map algebra, spatial filters, resampling, … (raster package)
    - Vectorize raster data (rgdal, raster packages)

For slides 14 - 18 see code and examples in this webpage



- **Descriptive statistics:** central tendency and spread measures
- **Exploratory graphics (2D, 3D):** scatter plot, box plot, histogram, …
- **Spatial autocorrelation:**
    - Global spatial autocorrelation statistics: Moran’s I, Geary’s C, Getis and Ord’s G(d) (spdep package)
    - Local spatial autocorrelation statistics: Moran’s Ii, Getis and Ord’s Gi y Gi*(d) (spdep package)

For slides 14 - 18 see code and examples [in this webpage]



ASK the right question

GET the data

PREPARE the data

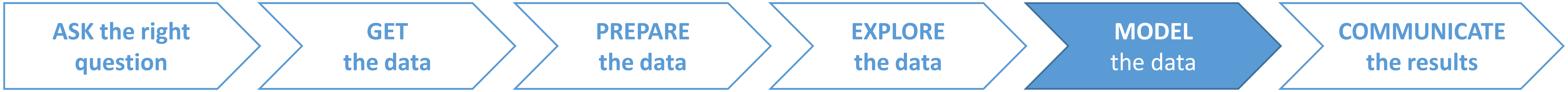
EXPLORE the data

MODEL the data

COMMUNICATE the results

- **Regression:**
  - Spatial autoregressive models (spdep package)
  - Geographically weighted regression (spgwr package)
- **Classification (Machine Learning):**
  - Supervised: RandomForests, SVM, boosting, … (caret package)
  - Non-supervised: k-means clustering (stats package)
- **Spatial statistics:**
  - Geostatistics (gstat, geoR, geospt packages and others)
  - Spatial point patterns (spatstat package)

For slides 14 - 18 see code and examples in this webpage



ASK the right question

GET the data

PREPARE the data

EXPLORE the data

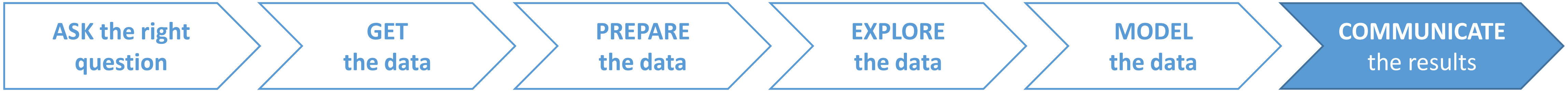
MODEL the data

COMMUNICATE the results

- **Static or interactive maps:** tmap, leaflet, mapview packages

- **Interactive graphics, web apps and dashboards:**
  - plotly (example), rcharts, googleVis (example) packages
  - shiny, see example
  - flexdashboard, see example

For slides 14 - 18 see code and examples in this webpage



ASK the right question

GET the data

PREPARE the data

EXPLORE the data

MODEL the data

COMMUNICATE the results

Don’t forget: **Reproducibility!**

- R code and output for examples shown in this webinar (slides 17-21) can be **reproduced** with this .Rmd document using RMarkdown
- See this example about reproducible spatial analysis using interactive notebooks
- Learn more about reproducible geoscientific research

Integrating R with GIS software

- **QGIS:** see example in this post
- **ArcGIS:** arcgisbinding package, see example in this post
- **GRASS GIS:** version 6, spgrass6 package; version 7, rgrass7 package
- **gvSIG:** more info in this post
- **SAGA:** RSAGA package
- **GME (Geospatial Modelling Environment):** more info in this webpage

References / Online resources

- Bivand, R., Pebesma, E., Gómez-Rubio, V. 2013. Applied Spatial Data Analysis with R. New York: Springer. 2nd ed.
- R-SIG-Geo mailing list
- CRAN Task View: Analysis of Spatial Data
- Facebook groups: GIS with R, R project en Español
- Google+ groups: Statistics and R, R Programming for Data Analysis
- My blog: amsantac.co/blog.html

Thanks!

If you have any question feel free to contact me:

amsantac.co/contact.html