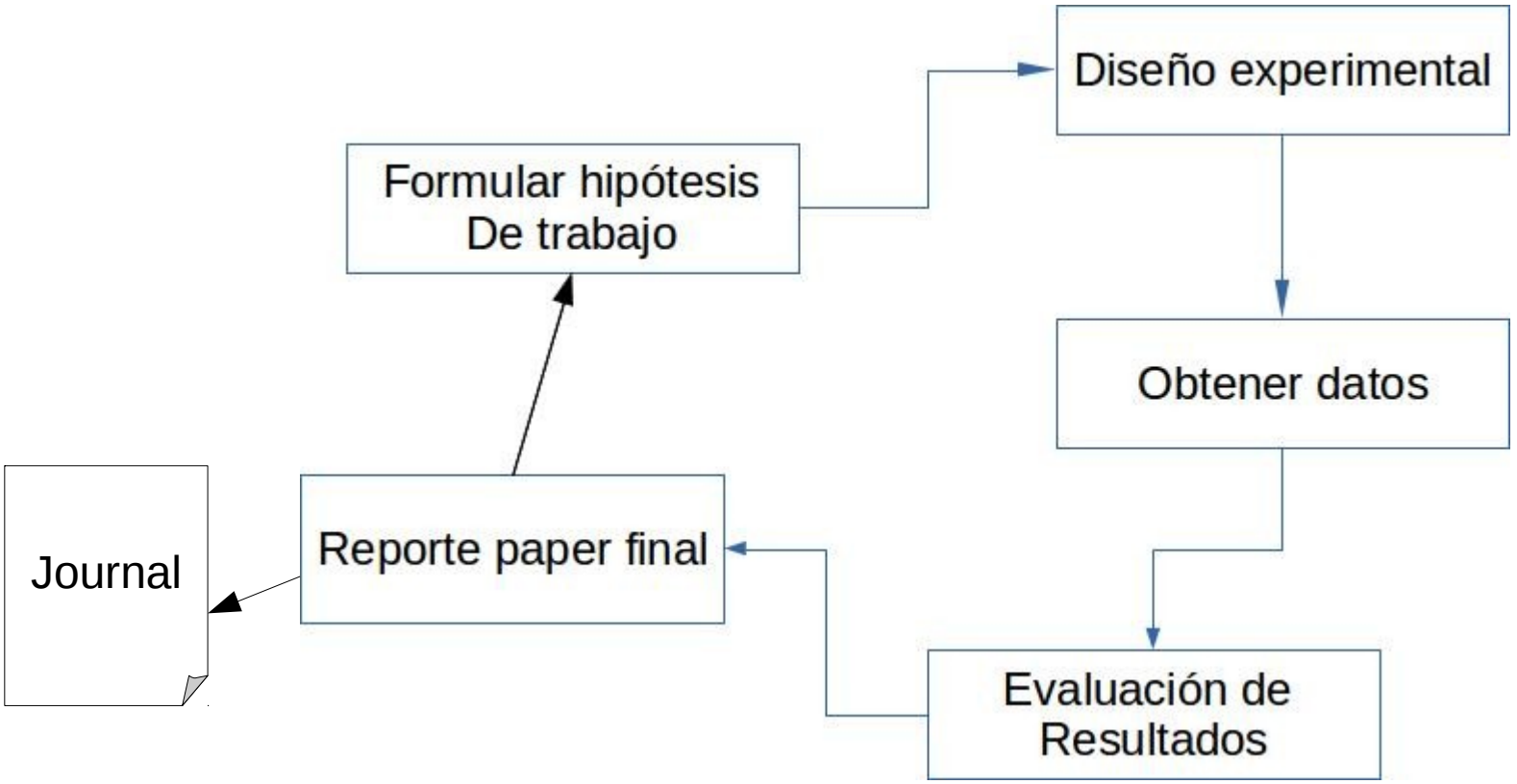
Open Research

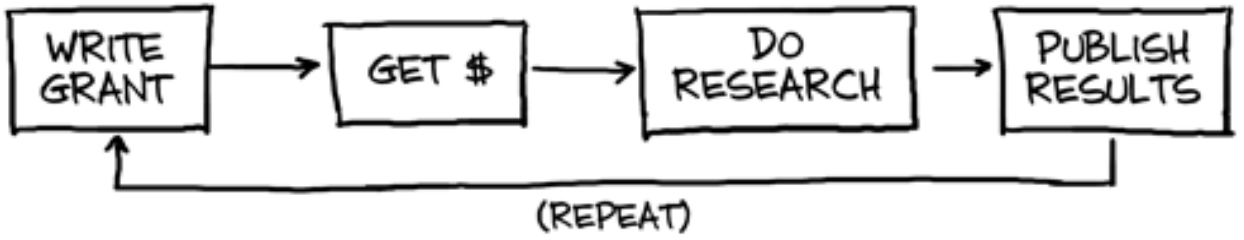
Ana Laura Diedrichs
Lab. DHARMa
Grupo GRIDTICs

Ciclo de investigación (simplificado)



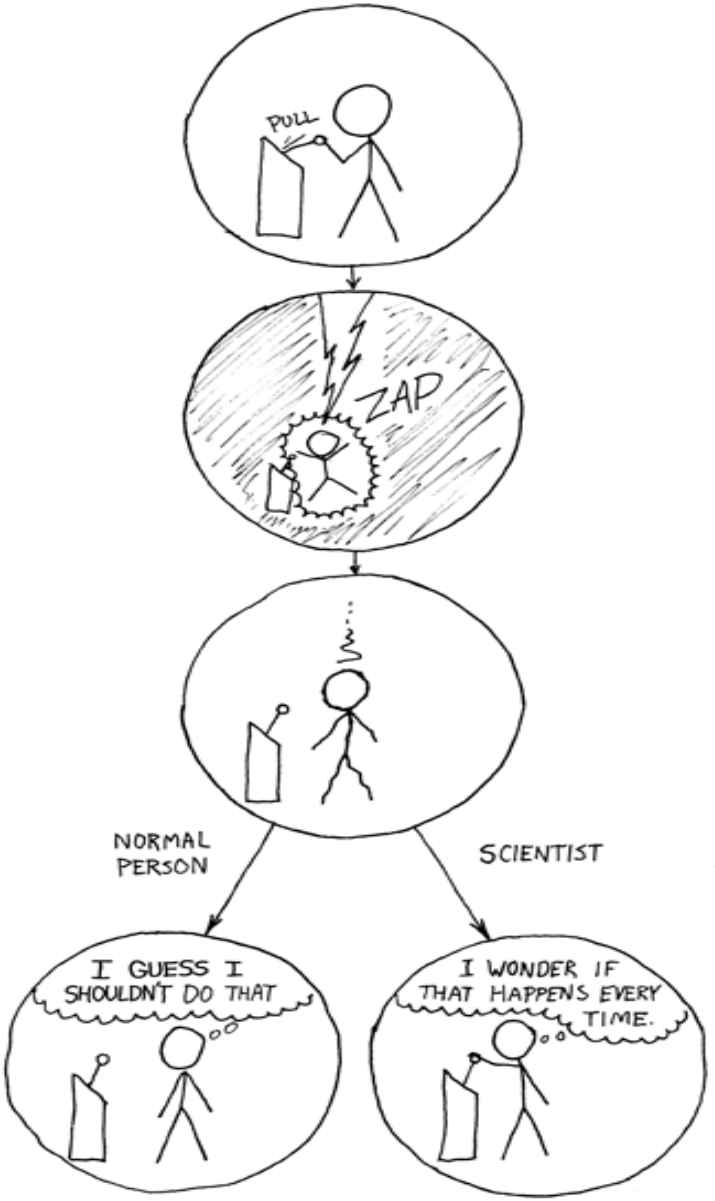
THE GRANT CYCLE

HOW IT'S SUPPOSED TO WORK:



HOW IT REALLY WORKS:

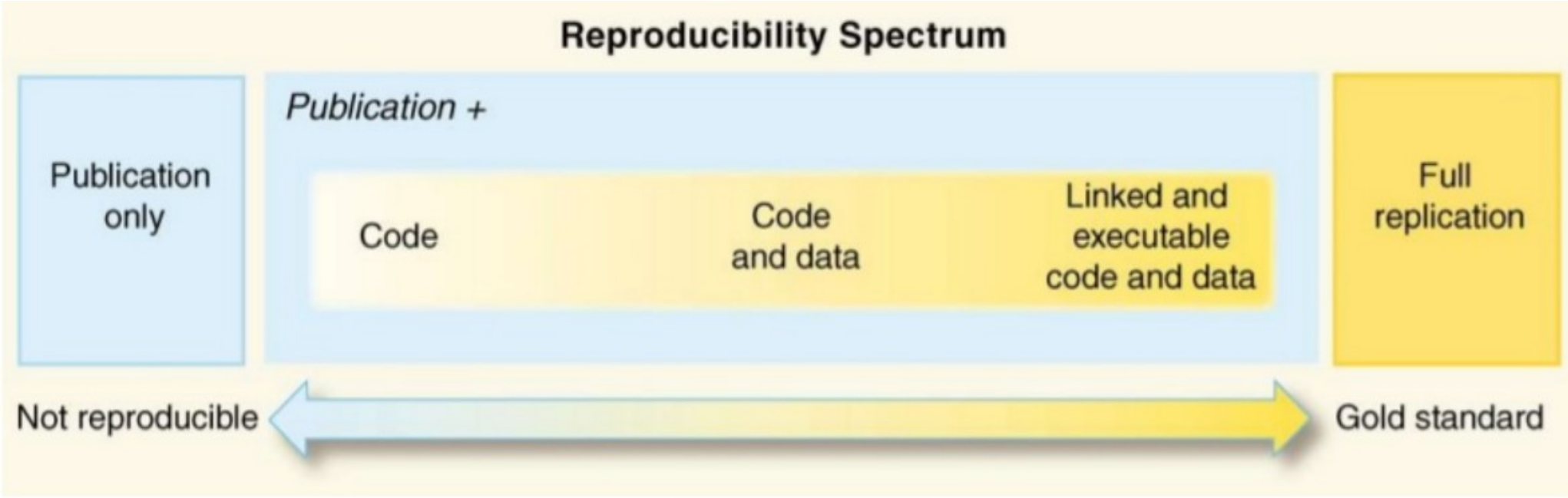
Investigación reproducible



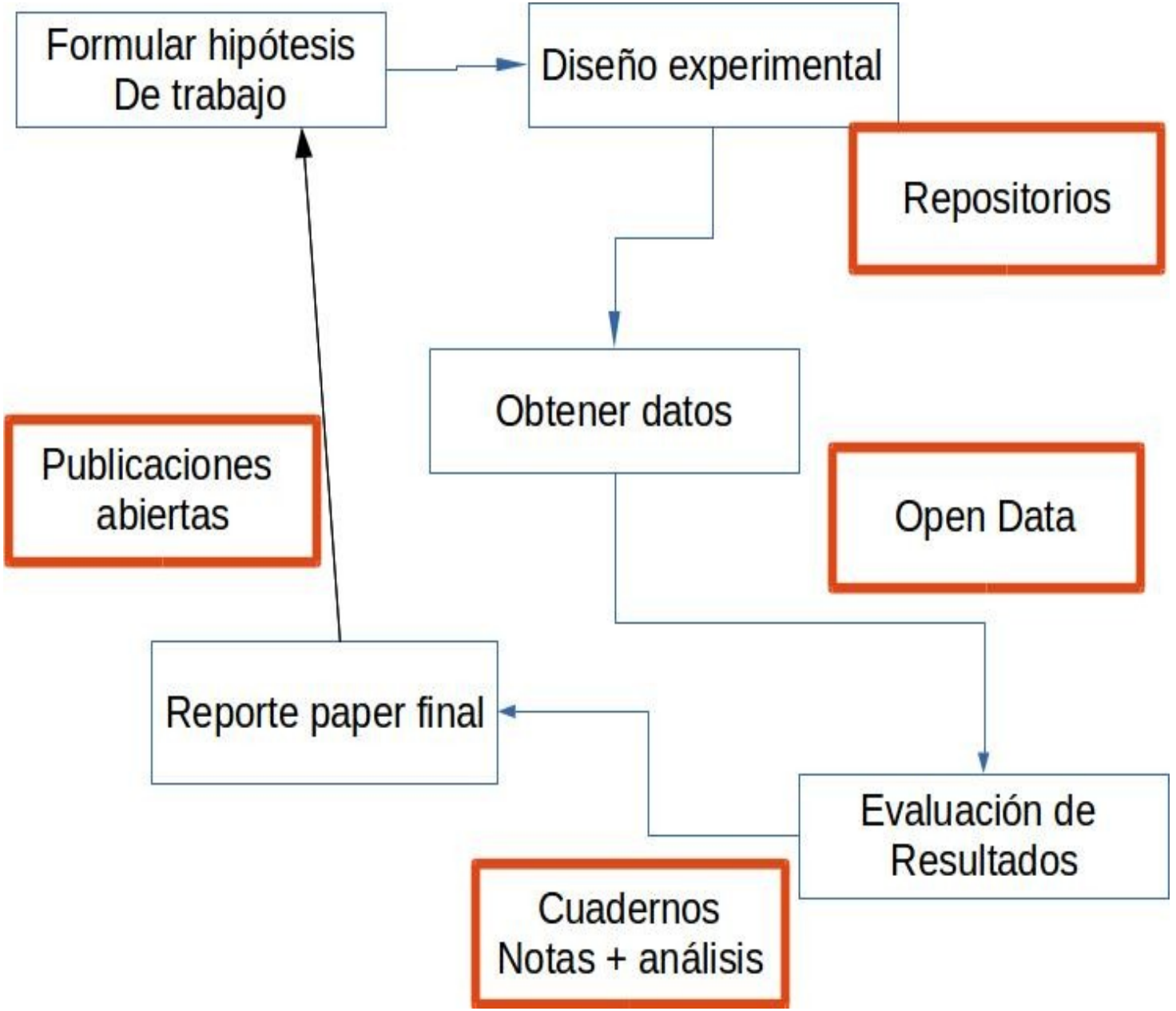
- Métodos (scripts) + datos = resultados
- Bajo ciertas condiciones o áreas,es costoso o imposible reproducir al 100%

http://xkcd.com/242/

Investigación reproducible



"Reproducible Research in Computational Science". **RD Peng** Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847


Formular hipótesis
De trabajo
Diseño experimental
Repositorios
Obtener datos
Publicaciones
abiertas
Open Data
Reporte paper final
Evaluación de
Resultados
Cuadernos
Notas + análisis

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

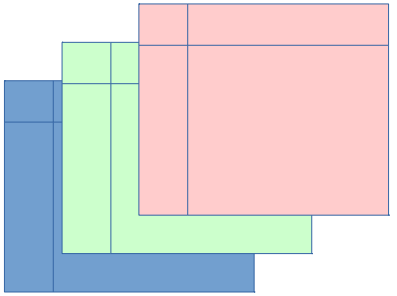
Datos

Datos crudos

* encuestas
* información georeferenciada
* datos de sensores
* imágenes satelitales

procesamiento

Datos procesados

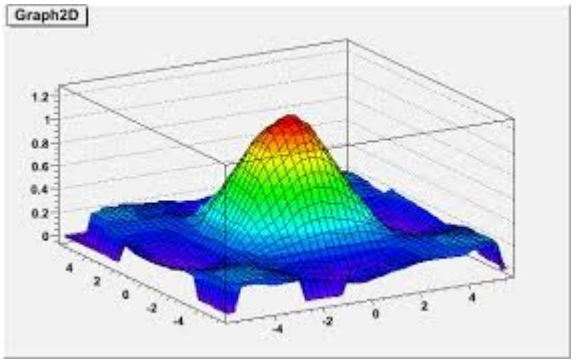


Formatos archivos abiertos perduran en el tiempo

.tiff, geojson, json, csv

análisis

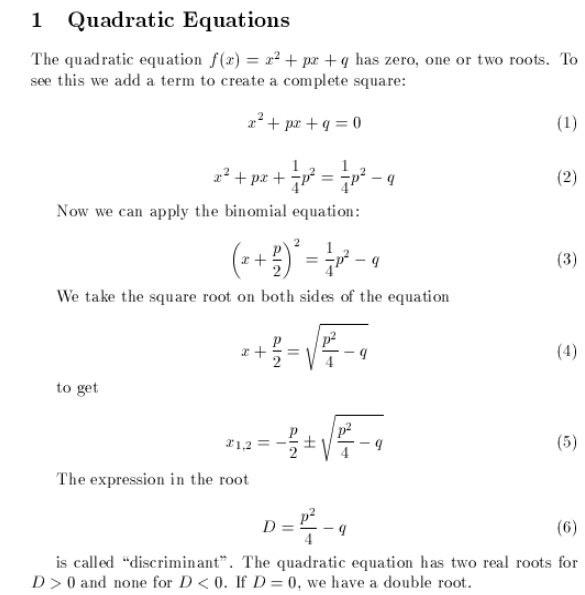
Resultados



Herramientas

LaTeX

documentos



1 Quadratic Equations

The quadratic equation $f(x) = x^2 + px + q$ has zero, one or two roots. To see this we add a term to create a complete square:

$$x^2 + px + q = 0 \quad (1)$$

$$x^2 + px + \frac{1}{4}p^2 = \frac{1}{4}p^2 - q \quad (2)$$

Now we can apply the binomial equation:

$$\left(x + \frac{p}{2}\right)^2 = \frac{1}{4}p^2 - q \quad (3)$$

We take the square root on both sides of the equation

$$x + \frac{p}{2} = \sqrt{\frac{p^2}{4} - q} \quad (4)$$

to get

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \quad (5)$$

The expression in the root

$$D = \frac{p^2}{4} - q \quad (6)$$

is called "discriminant". The quadratic equation has two real roots for $D > 0$ and none for $D < 0$. If $D = 0$, we have a double root.

git

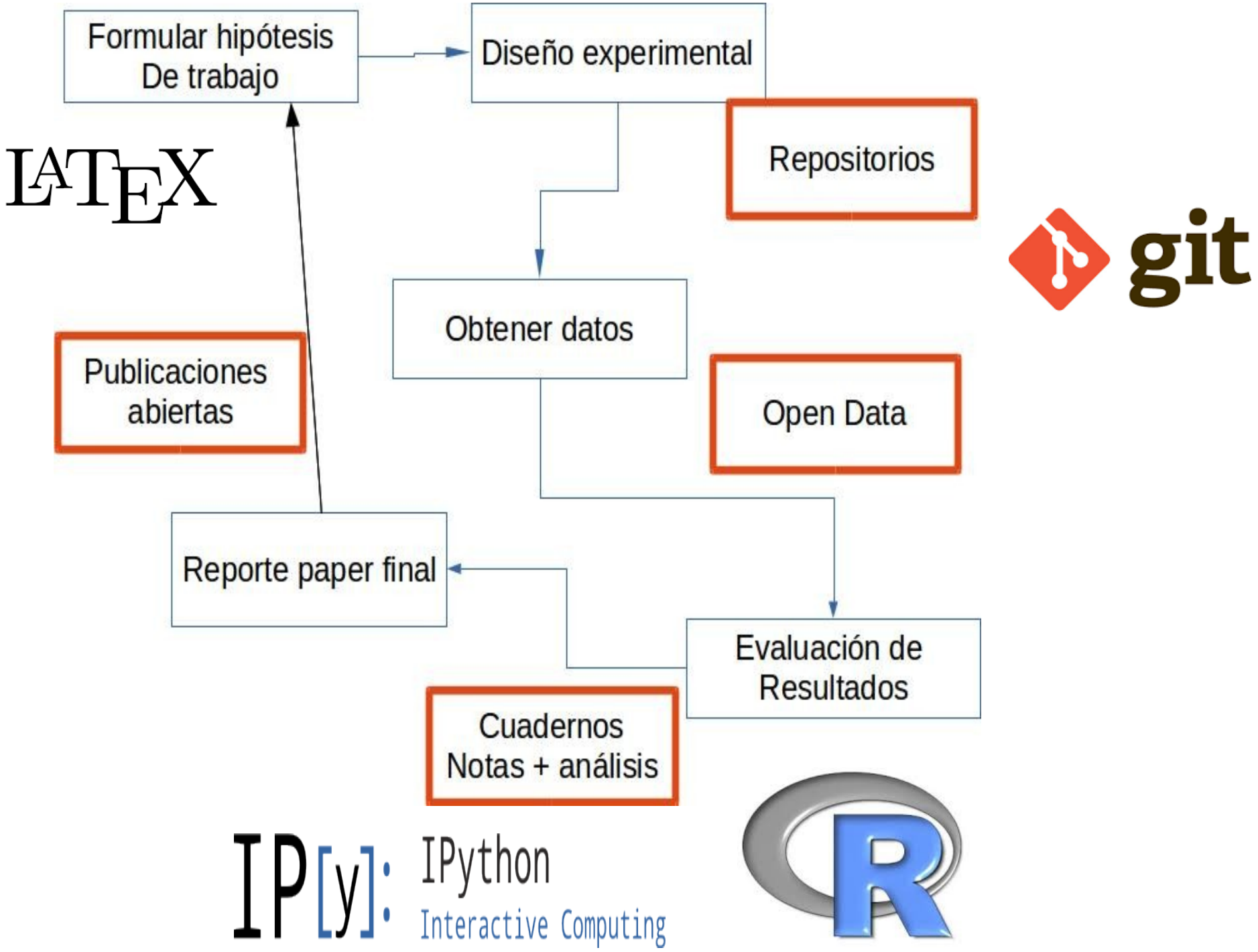
Versionar código y documentos + trabajo colaborativo

IPython Interactive Computing

código

python™

R


Formular hipótesis
De trabajo
Diseño experimental
Repositorios
LaTeX
git
Obtener datos
Publicaciones
abiertas
Open Data
Reporte paper final
Evaluación de
Resultados
Cuadernos
Notas + análisis
IP[y]: IPython
Interactive Computing
R

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Addins | Project: (None)

entropy.R | tmin-experiment.R | sensores.R | utils-plotting.R | Makefile

Source on Save | Run | Source

```
#'
#' Para trabajar con este problema se abordó un enfoque de metaheurística con un algoritmo poblacional como el algoritmo genético.
#' Considero un escenario discreto, es decir una cantidad fija de posiciones.
#' Diseñamos una función $g(x)$ de evaluación
#' que ayude a determinar la distancia entre los sensores, por ejemplo,
#' una medida de captura entre los eventos anteriores, o una medida de cobertura.
#' La función objetivo o puntaje (o rendimiento a evaluar) podría tener la siguiente forma:
#'  $$f(x)=w_1*numberOfSensorsCost(x)+w_2*G(x)$$
#' Donde $number_of_sensors$ es una función lineal decreciente que otorga mayor puntaje mientras menos sensores sean, y $G$
#' es la media de la desviación estandar entre los sensores.
#' Para correr los experimentos utilizo R y las siguientes librerias.
#'
library(ggplot2)
library(GA)
#' Configuro una semilla, para reproducibilidad de los experimentos.
SEED <- 3.8
#' Cargo el archivo con la matriz de distancia de los sensores
distancias.sensores <- read.csv("distancias.csv", row.names=1)
#' La siguiente es la variable que almacena los nombres de los sensores
NAMES <- names(distancias.sensores)
print(NAMES)
#' Constante para almacenar cantidad total de sensores
TOTAL <- length(NAMES)
#' Coeficiente del primer término
W_1 <- 0.004
#' Coeficiente del segundo término
W_2 <- 0.6
#' Matriz que almacena las posiciones relativas de los sensores, para mostrar gráficamente la solución
position.table <- data.frame( H1 = c('S4','S20','S11','S15','S19'),
                  H2 = c('S3','S7','S10','S14','S18'),
                  H3 = c('S2','S6','S9','S13','S17'),
                  H4 = c('S1','S5','S8','S12','S16'))
#' Función lineal que otorga mayor puntaje mientras menos sensores sean
#
numberOfSensorsCost <- function(x){ return( ((-1)*x) + TOTAL )}
```

92:43 (Top Level) | R Script

Environment | History

Import Dataset | List

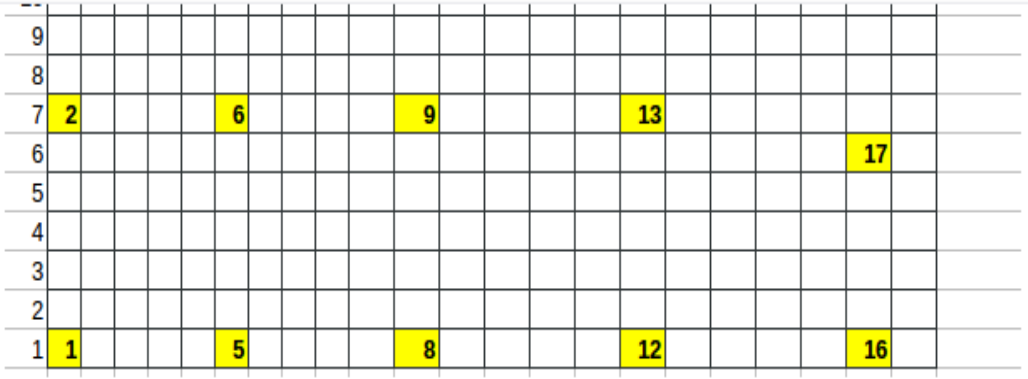
Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

Install | Update

Name	Description	V...
User Library		
akima	Interpolation of Irregularly and Regularly Spaced Data	0.5-12
assertth	Easy pre and post assertions.	0.1
backpor	Reimplementations of Functions Introduced Since R-3.0.0	1.0.4
base64e	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.62.1
BiocGen	S4 generic functions for Bioconductor	0.20
BiocInst	Install/Update	1.24



Parcela de interés de 0.7 ha remarcada en azul. Los puntos azules son ubicaciones aproximadas

Para trabajar con este problema se abordó un enfoque de metaheurística con un algoritmo poblacional como el algoritmo genético. Considero un escenario discreto, es decir una cantidad fija de posiciones. Diseñamos una función $g(x)$ de evaluación que ayude a determinar la distancia entre los sensores, por ejemplo, una medida de captura entre los eventos anteriores, o una medida de cobertura. La función objetivo o puntaje (o rendimiento a evaluar) podría tener la siguiente forma:

$$f(x) = w_1 * numberOfSensorsCost(x) + w_2 * G(x)$$

Donde $number_of_sensors$ es una función lineal decreciente que otorga mayor puntaje mientras menos sensores sean, y G es la media de la desviación estandar entre los sensores. Para correr los experimentos utilizo R y las siguientes librerías.

```
library(ggplot2)
library(GA)
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

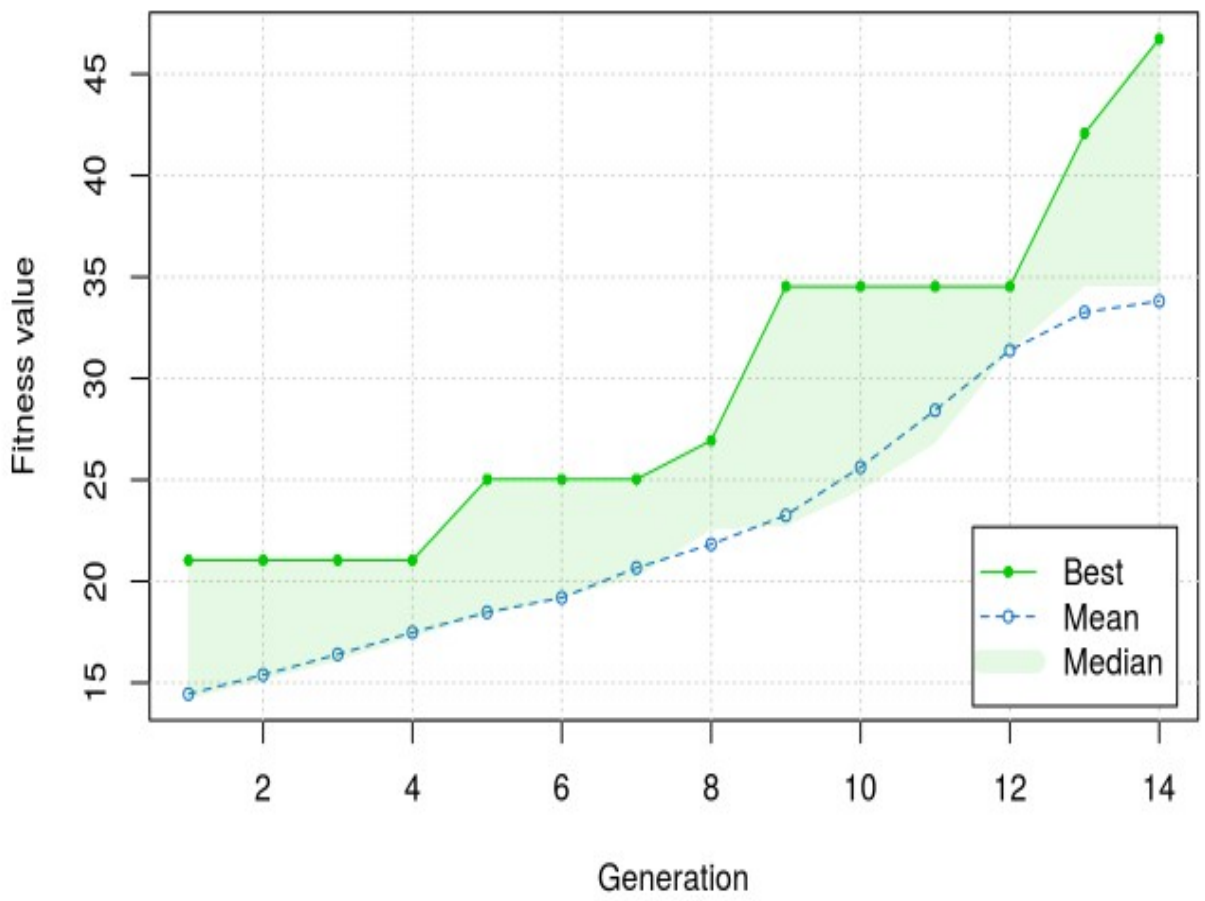
```
## Package 'GA' version 3.0.2
```

```
## Type 'citation("GA")' for citing this R package in publications.
```

Configuro una semilla, para reproducibilidad de los experimentos.

A continuación observamos como fue aumentando el fitness a medida que transcurren las generaciones

```
plot(GA.MODEL)
```



DISCUSION

Para este modelo de optimización, tan sólo fueron utilizadas dos medidas en la función de evaluación: una basada en el costo por cantidad de sensores y otra sobre la dispersión de las distancias. Dejo abierta la discusión sobre que medidas utilizar para determinar la diferencia entre las

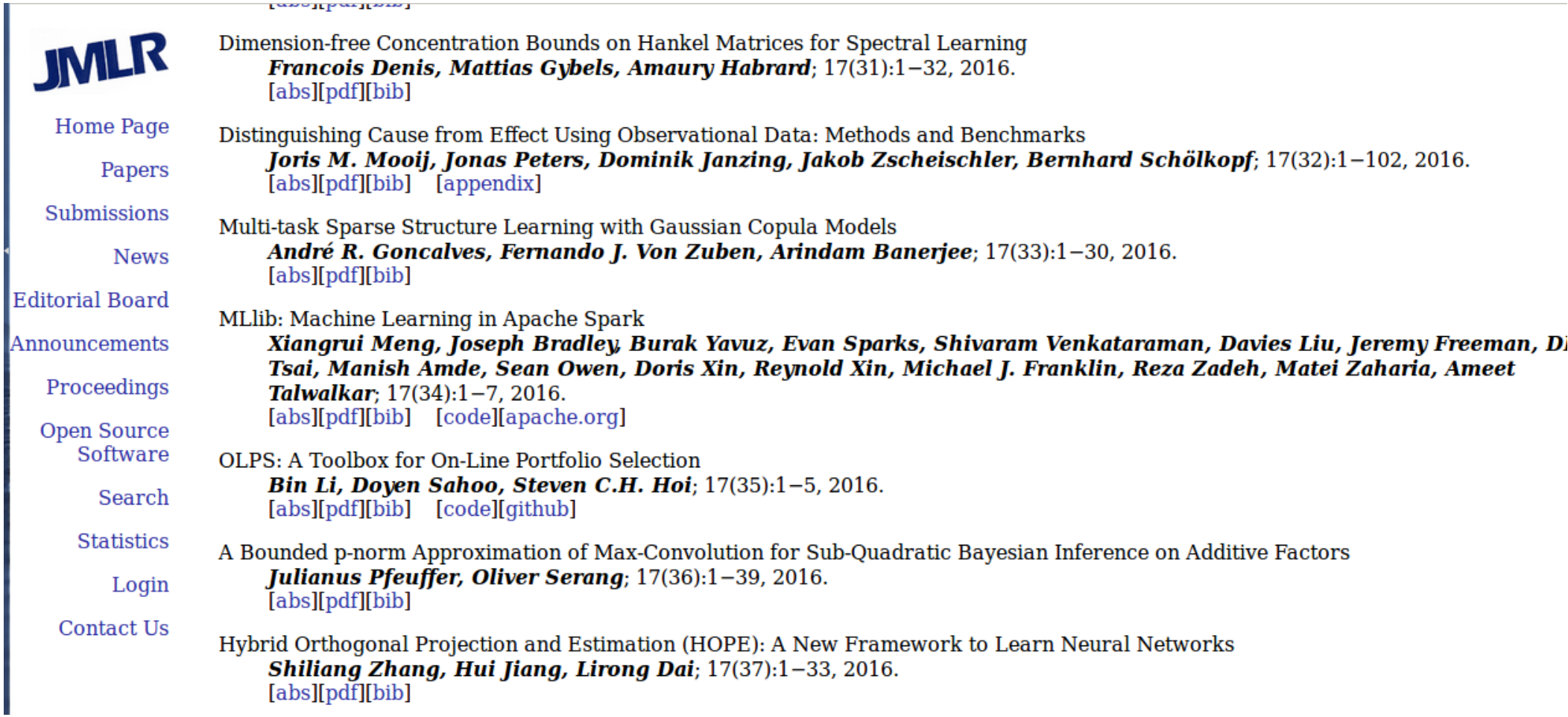
#opendata

- Lista de repositorios variados de datos

  http://www.nature.com/sdata/policies/repositories
- UCI Machine learning data repository http://archive.ics.uci.edu/ml/
- CRAWDAD A community resource for archiving wireless data at Dartmouth http://crawdad.org/
-

#openaccess

- Permitir el acceso a libre a las publicaciones científicas



JMLR

Home Page
Papers
Submissions
News
Editorial Board
Announcements
Proceedings
Open Source Software
Search
Statistics
Login
Contact Us

Dimension-free Concentration Bounds on Hankel Matrices for Spectral Learning
Francois Denis, Mattias Gybels, Amaury Habrard; 17(31):1−32, 2016.
[abs][pdf][bib]

Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, Bernhard Schölkopf; 17(32):1−102, 2016.
[abs][pdf][bib] [appendix]

Multi-task Sparse Structure Learning with Gaussian Copula Models
André R. Goncalves, Fernando J. Von Zuben, Arindam Banerjee; 17(33):1−30, 2016.
[abs][pdf][bib]

MLlib: Machine Learning in Apache Spark
Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, D
Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, Ameet Talwalkar; 17(34):1−7, 2016.
[abs][pdf][bib] [code][apache.org]

OLPS: A Toolbox for On-Line Portfolio Selection
Bin Li, Doyen Sahoo, Steven C.H. Hoi; 17(35):1−5, 2016.
[abs][pdf][bib] [code][github]

A Bounded p-norm Approximation of Max-Convolution for Sub-Quadratic Bayesian Inference on Additive Factors
Julianus Pfeuffer, Oliver Serang; 17(36):1−39, 2016.
[abs][pdf][bib]

Hybrid Orthogonal Projection and Estimation (HOPE): A New Framework to Learn Neural Networks
Shiliang Zhang, Hui Jiang, Lirong Dai; 17(37):1−33, 2016.
[abs][pdf][bib]

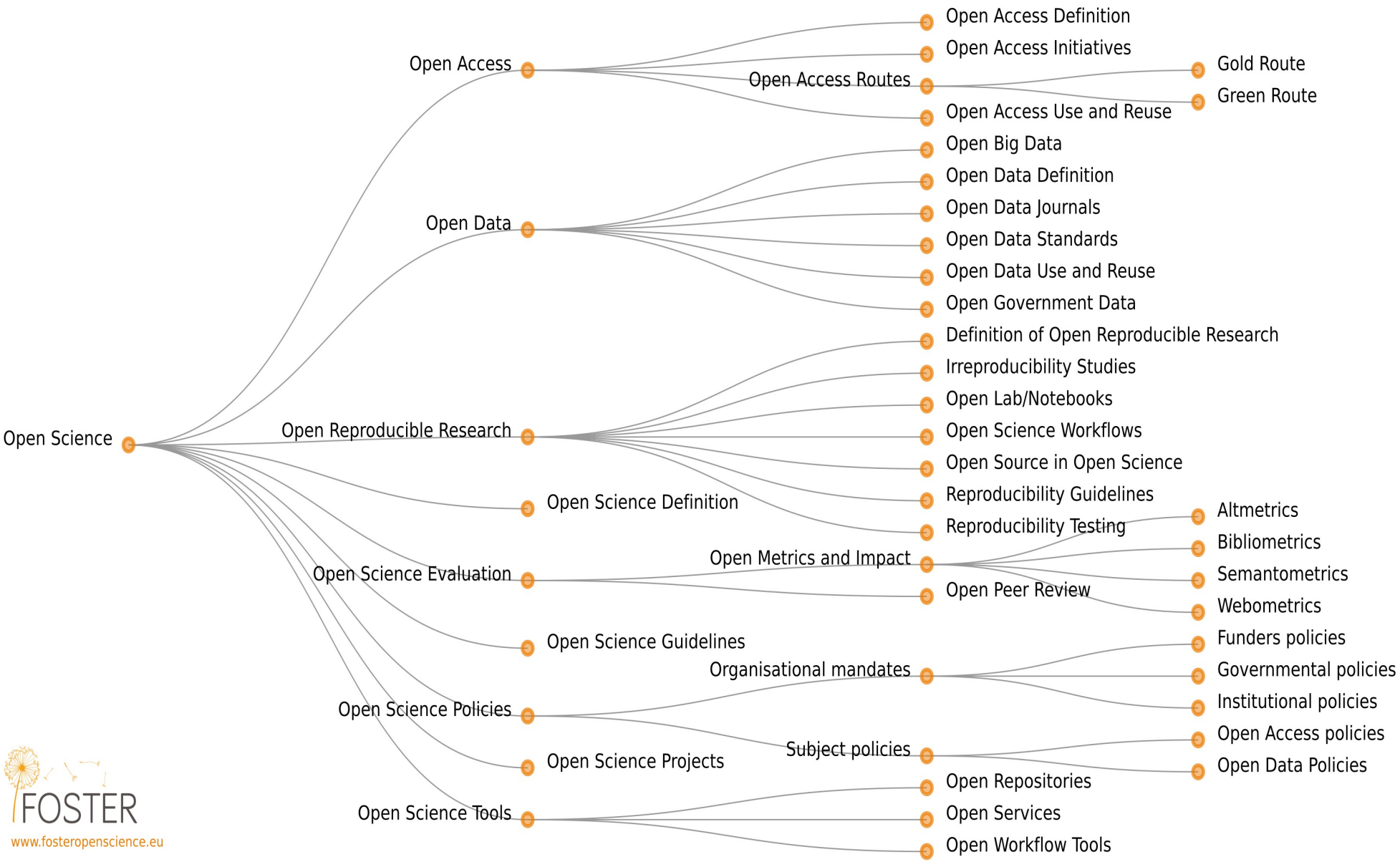
#openscience claras ventajas

- Mayor visibilidad
- Mayor citas
- Enriquecimiento en las contribuciones
- +

#openscience desafíos

- Cambio de paradigma y cultural
- Protección de derechos intelectuales
- Adecuación a nuevas herramientas
- -

Open Science Taxonomy



Nancy Pontika; Petr Knoth; Matteo Cancellieri; Samuel Pearce (2015). "Fostering Open Science to Research using a Taxonomy and an eLearning Portal". Retrieved 12 August 2015

Más en:

- http://openaccessweek.org/
- http://whyopenresearch.org/
- https://cran.r-project.org/web/views/ReproducibleResearch.html

Gracias!

#openscience #opendata

Ana.diedrichs@frm.utn.edu.ar
@anadiedrichs

Lab. DHARMa
Grupo GRIDTICs

UTN
FACULTAD REGIONAL MENDOZA

CONICET
Consejo Nacional de Investigaciones Científicas y Técnicas