

Big Data Week is one of the most unique global platforms of interconnected community events focusing on the social, political, technological and commercial impacts of Big Data

Follow all the events at

**bigdataweek.com**

Official Event Hashtag **#bdw13**

Big Data Week brand and concept copyright © 2013 Big Data Week - produced by media140

# BDW13: Introducción al Data Mining

BIG\_DATA\_WEEK\_2013

Óscar Marín Miró  
@oscarmarinmiro  
@outliers\_es  
oscar@outliers.es



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License



**Outliers**  
Because differences matter.

# CONTENIDOS

INTRODUCCIÓN A LA MINERÍA DE DATOS

ADQUISICIÓN

ANÁLISIS

GEPHI

REFERENCIAS

Material del curso en <http://assets.outliers.es/bdw13/datamining>



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License

# INTRODUCCIÓN A LA MINERÍA DE DATOS



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License

# DE LOS DATOS A LA SABIDURÍA

“Los **datos**, organizados y empleados debidamente, pueden convertirse en información.

La **información**, absorbida, comprendida y aplicada por las personas, puede convertirse en conocimientos.

Los **conocimientos** aplicados frecuentemente en un campo pueden convertirse en sabiduría, y la **sabiduría** es la base de la acción positiva”

Michael Cooley: “Architect or Bee?” Hogarth Press, London, UK, 1987.

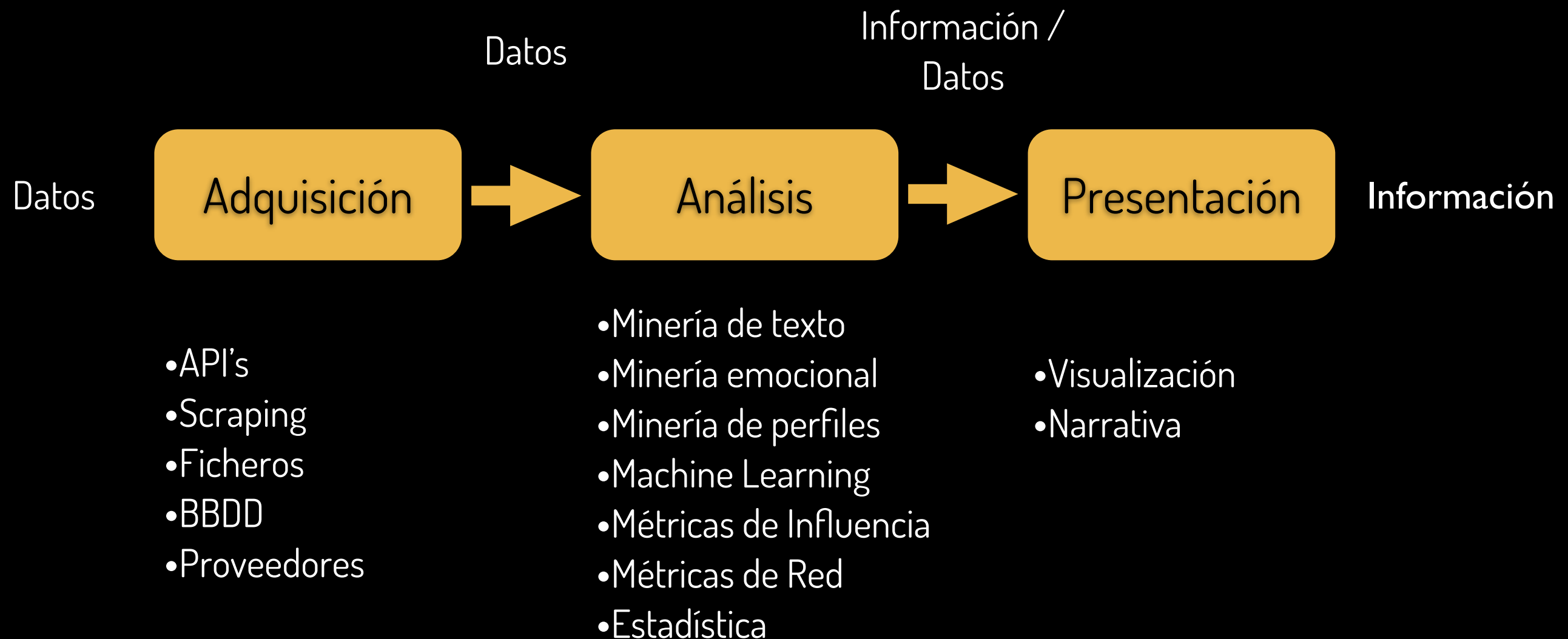
# DE LOS DATOS A LA SABIDURÍA



## ¿QUÉ ES EL ANÁLISIS DE DATOS?

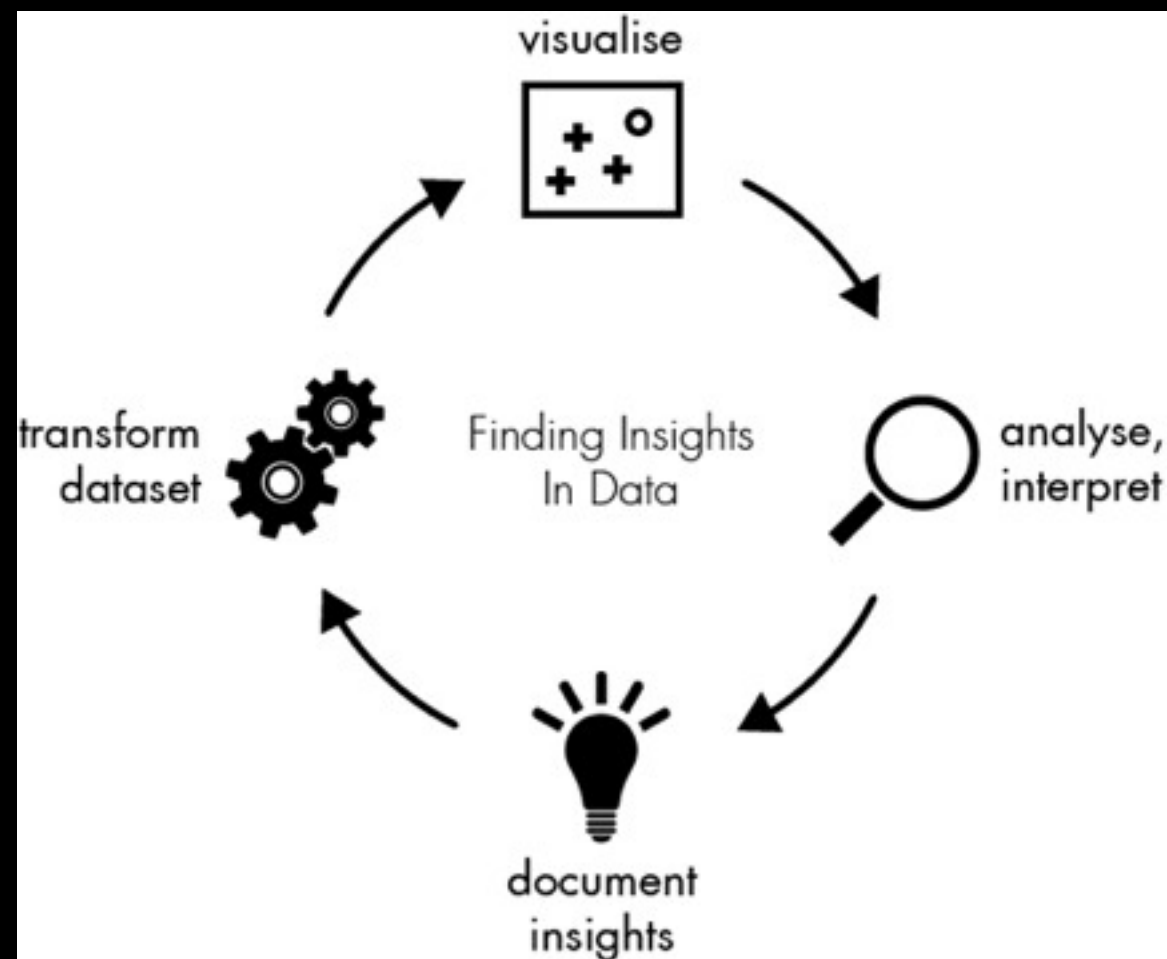
- ▶ Proceso multidisciplinar
- ▶ Pasar de los datos a la información
- ▶ Nuevas maneras de estructurar datos desestructurados
- ▶ Cuantitativo o cualitativo
- ▶ Automático, semi-automático, manual
- ▶ El dataset puede ser el punto de partida y/o de llegada

# PIPELINE DE DATOS





# FASES DE UN TRABAJO (VISUALIZACIÓN)



## MÉTODOS: ADQUISICIÓN

- ▶ Adquisición
  - ▶ API's
  - ▶ Scraping
  - ▶ Crawling
  - ▶ Open Data
  - ▶ Ficheros (excel, tsv, csv)

## MÉTODOS: ANÁLISIS

- ▶ Análisis
  - ▶ Numérico / Estadístico (p.ej: histograma de followers)
  - ▶ Contenido (p.ej: análisis semántico)
  - ▶ Relaciones: Grafos entre cualquiera de los anteriores

## MÉTODOS: PRESENTACIÓN

- ▶ Presentación
  - ▶ Visualización de Redes (grafos)
  - ▶ Nubes de palabras (contenido)
  - ▶ Gráficos estadísticos (numérico)

## HERRAMIENTAS

- ▶ Programación:
  - ▶ Python (multipropósito)
  - ▶ R (análisis estadístico)
  - ▶ Librería Pattern para Python
- ▶ Sin programación
  - ▶ Gephi (análisis de redes)
  - ▶ Wordle (nubes de palabras)

# ADQUISICIÓN DE DATOS

# ADQUISICIÓN DE DATOS

- Traer a memoria datos que existen fuera de ella
- ¿Dónde?
  - Fichero local (CSV, JSON, XML)
  - BBDD
  - Internet:
    - API
    - Scraping
    - Crawling
    - Ficheros remotos (CSV, JSON, XML)

# ADQUISICIÓN DE DATOS: FORMATOS

TSV

```
"valor1\tvalor2\tvalor3\t\n"
```

JSON

```
{ [
```

```
{ "campo1": "valor1", "campo2": "valor21",
```

```
{ "campo1": "valor1N", "campo2": "valor2N"
```

```
}]
```

XML

```
<items>
```

```
<item campo1="valor1" campo2="valor21" />
```

```
<item campo1="valor1N" campo2="valor2N" /> </items>
```



# ADQUISICIÓN DE DATOS: FUENTES



<http://shop.oreilly.com/product/0636920018254.do>

<http://www.quora.com/Data/Where-can-I-get-large-datasets-open-to-the-public>

<http://www.google.com/publicdata/directory>

<http://data.worldbank.org>

## Application Programming Interface

- ▶ API's
  - ▶ Facebook Open Graph API
  - ▶ YouTube, API
  - ▶ Google :-[
  - ▶ Foursquare, Venues
  - ▶ Twitter

# ADQUISICIÓN DE DATOS: TWITTER API

- ▶ GET API:
  - ▶ Parámetros: consulta booleana, lenguaje, geocode
  - ▶ Sin autenticación, pero en proceso de cambio, ojo!
  - ▶ Tweets 'incompletos': sin info de usuario ni de RT
  - ▶ 150 peticiones por hora
  - ▶ No documentado, pero puedes buscar 'places'

# ADQUISICIÓN DE DATOS: TWITTER API

- ▶ Streaming API:
  - ▶ Parámetros:
    - ▶ Follow: hasta 5000 usuarios
    - ▶ Keywords (track): hasta 400
    - ▶ Location: hasta 25 cajas
  - ▶ Requiere una cuenta twitter
  - ▶ El follow te da tweets, RT's origen y destino; replies
  - ▶ Empieza a limitar si el volumen que te dan > 1%

# ADQUISICIÓN DE DATOS: TWITTER API

- ▶ El API REST completa
- ▶ Anatomía de un Tweet
- ▶ Herramientas para acceder

# ADQUISICIÓN DE DATOS: STREAMING DE TWITTER

- ▶ Vamos a 'enchufarnos' al streaming de twitter
- ▶ La librería que vamos a usar es tweetstream
- ▶ Seguiremos un hashtag que sea trending topic ahora mismo
- ▶ Cogemos 100 tweets y los volcaremos en un fichero json
- ▶ El ejercicio está en `acq/twitterStreamGet.py`

# ADQUISICIÓN: PATTERN.WEB

- ▶ <http://www.clips.ua.ac.be/pages/pattern-web>
- ▶ Descarga de urls genéricas
- ▶ Web crawling (clase [web.Spider](#))
- ▶ Conversión de HTML a texto
- ▶ Acceso a través de web.SearchEngine: Google, Bing, Yahoo, Twitter, Facebook, Wikipedia, Flickr...
- ▶ Abrir `acq/web.py`

# ANÁLISIS



# ANÁLISIS DE DATOS

- ▶ Análisis cuantitativos (métricas)
- ▶ Análisis de contenido (cuantitativo y cualitativo: sentiment, entidades, NLP)
- ▶ Análisis de relaciones

# ANÁLISIS CUANTITATIVO: HISTOGRAMA FOLLOWERS

- ▶ ¿Todos los HT siguen el mismo perfil de followers?
- ▶ ¿Cuál es el número medio de followers por HT?

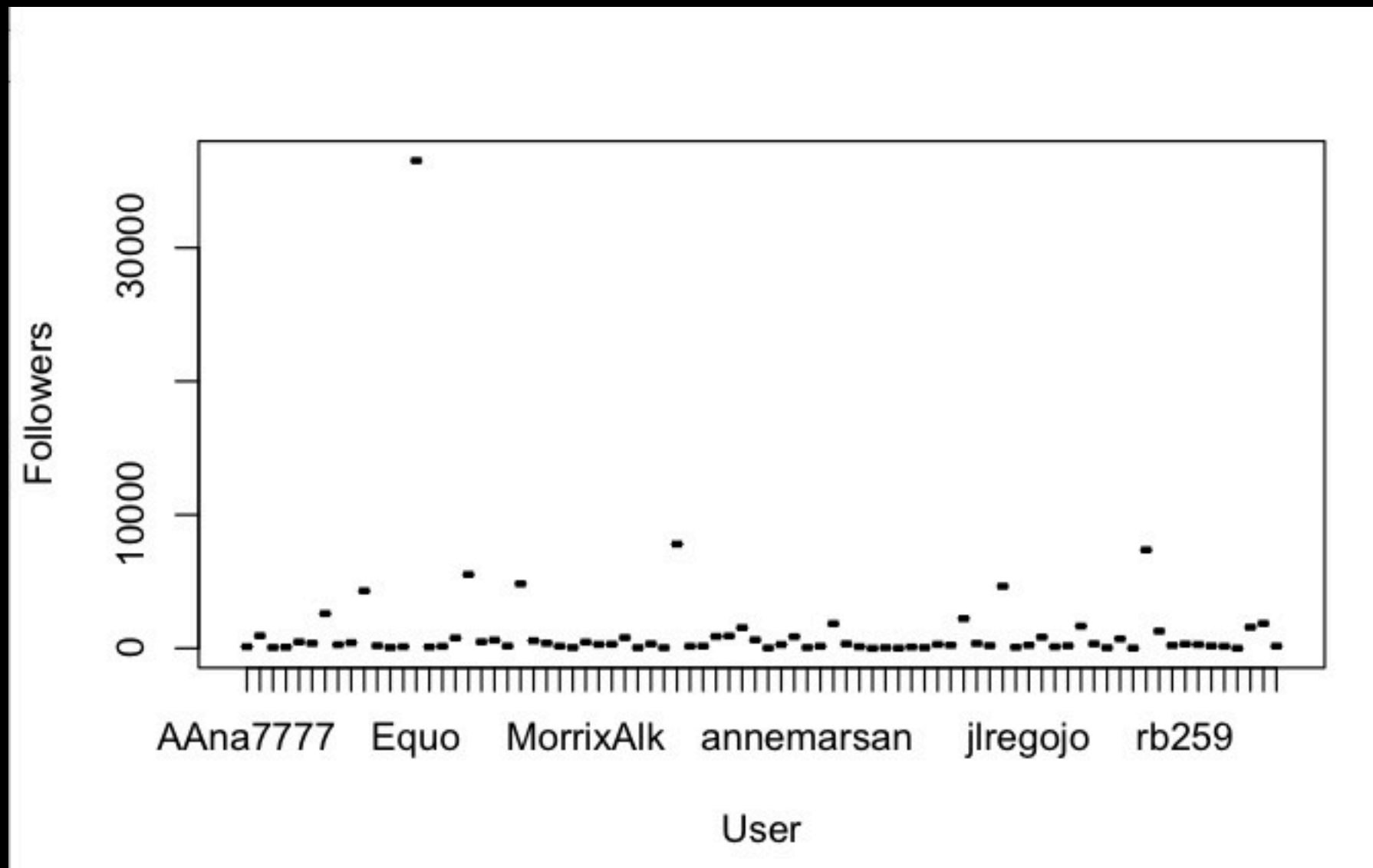
# ANÁLISIS CUANTITATIVO: R

- ▶ R: Lenguaje de programación para análisis estadístico y gráfico
- ▶ Instalación

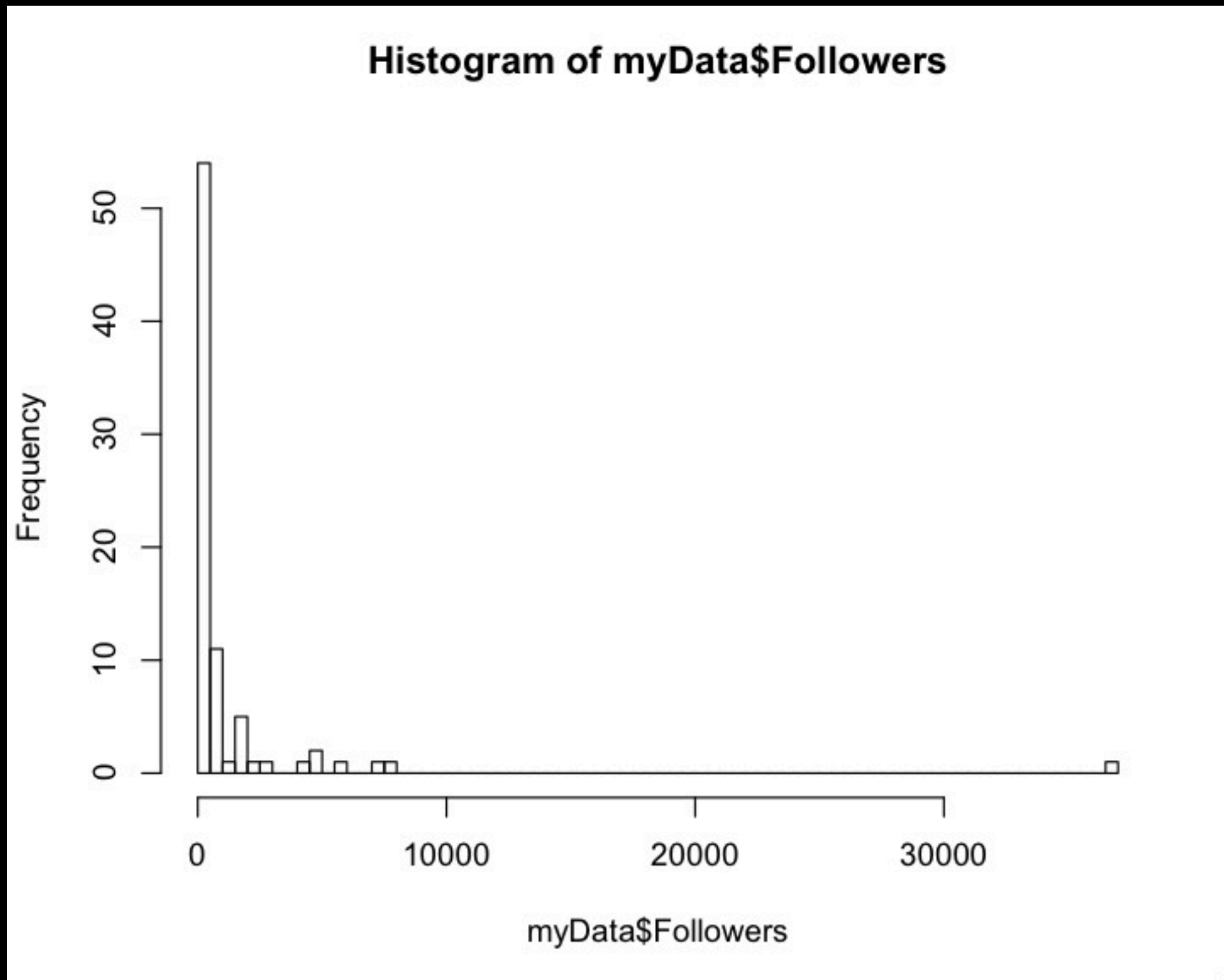
# ANÁLISIS CUANTITATIVO: IMPORTACIÓN Y ANÁLISIS

```
> myData = read.csv("/Users/oscarmarinmiro/PycharmProjects/Cursos/BigDataWeek/DataMining/analysis/histog
> head(myData)
      User Followers
1 Pippilotta_here    333
2   manolonight    189
3     dante_w     233
4       Equo    36531
5 feminismos_sol    4648
6 MARIANSANTIAG09    382
> myData$User
 [1] Pippilotta_here manolonight    dante_w       Equo       feminismos_sol MARIANSANTIAG09 esp
[11] marcosac3       de1969       ChaoticLibby  AfryLuNa     AnalisisSol15M GallaeciaReyno Mar
[21] rocidecastro    AsambleaVirtuaI ASantG       AranCaLe     marcdosan     aresmg1369     Alv
[31] SanidadEnLucha LaMuertedelPop NewellOsterberg Julioqc     angelessmig   JulianDiazgzg AA
[41] AsambleaAgra    MorrixAlk    yarr969      violetacela   mueve_tu_dinero peritoscostaluz Jos
[51] natimbs        belendemusica paquiron     AteneoAlkorcon UJCEArganda   JANietoPangea aul
[61] XJPeake        AnaMG12     beni_man     annemarsan   CiroOlivares5 HsalasteleSUR  lpd
[71] ane85649714    jlregojo    anamariacruzado aladeltas    Sarroyoscat   caboixo       Hos
80 Levels: AAAna7777 ASantG AfryLuNa AlexNyaklus Alvaro_ESHabbo AnaMG12 AnalisisSol15M AranCaLe AsambleaA
> myData$Followers
 [1] 333 189 233 36531 4648 382 199 1546 309 294 339 2228 60 61 2604 100 1
[30] 118 7807 567 802 4831 870 172 121 922 177 295 418 462 166 1580 43 73
[59] 334 1861 886 374 18 57 123 5531 836 53 151 47 281 234 27 633 1
> mean(myData$Followers)
[1] 1304.013
> median(myData$Followers)
[1] 289.5
> quantile(myData$Followers)
      0%      25%      50%      75%     100%
 0.0    121.0    289.5    810.5 36531.0
> plot(myData)
> plot(myData)
> hist(myData$Followers)
> hist(myData$Followers,breaks=20)
```

# ANÁLISIS CUANTITATIVO: IMPORTACIÓN Y ANÁLISIS



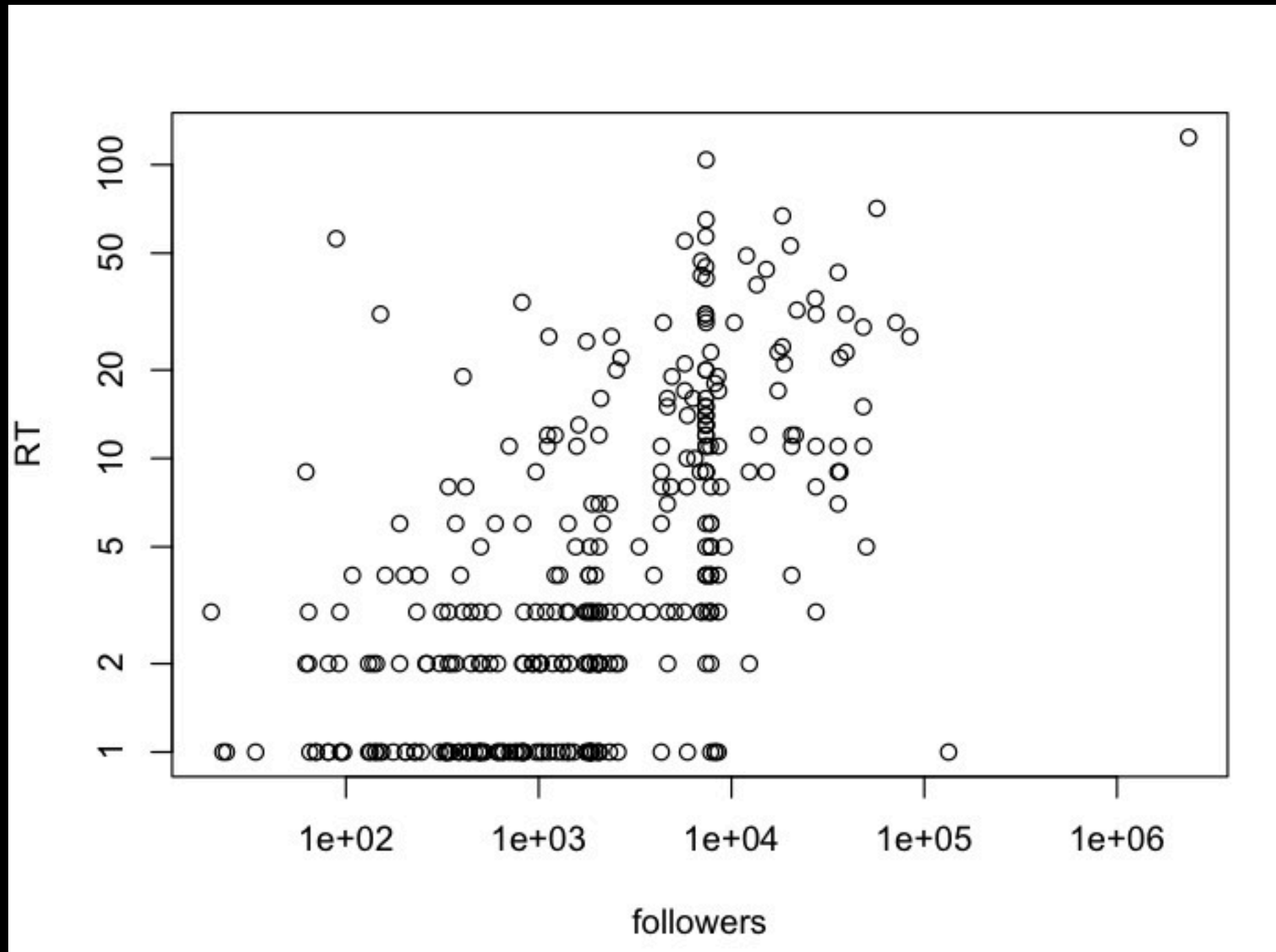
# ANÁLISIS CUANTITATIVO: HISTOGRAMAS



# ANÁLISIS CUANTITATIVO: DISPERSIÓN RT/FOLLOWERS

```
>  
> myData = read.csv("/Users/oscarmarinmiro/PycharmProjects/Cursos/BigDataWeek/DataMining/analysis/dispersionTW.23f.csv")  
> head(myData)  
  followers RT  
1     4652 16  
2     7392 41  
3     7389 29  
4     3838  3  
5     1846  1  
6       131  1  
> plot(myData)  
> plot(myData, log="xy")  
> |
```

# ANÁLISIS CUANTITATIVO: DISPERSIÓN RT/FOLLOWERS





# ANÁLISIS: LIBRERÍA 'PATTERN'

- ▶ Instalar la librería desde PyCharm (Preferences/ Python Interpreters/ (Ventana Derecha) / Install)
- ▶ En github: <https://github.com/clips/pattern>
- ▶ Ejemplos en <https://github.com/clips/pattern/tree/master/examples>
- ▶ Documentación en <http://www.clips.ua.ac.be/pages/pattern>

# ANÁLISIS NLP: PATTERN.TEXT

- ▶ <http://www.clips.ua.ac.be/pages/pattern-en>
- ▶ Lematización y Conjugación
- ▶ POS-Tagging
- ▶ Chunking
- ▶ Abrir Pattern/text.py

# ANÁLISIS DE PATRONES: PATTERN.SEARCH

- ▶ <http://www.clips.ua.ac.be/pages/pattern-search>
- ▶ Matching sintáctico, semántico y de raíces
- ▶ Taxonomías
- ▶ Abrir Pattern/match.py

# SENTIMENT ANALYSIS

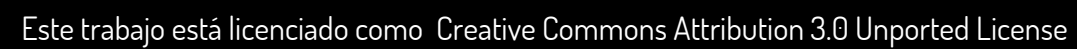
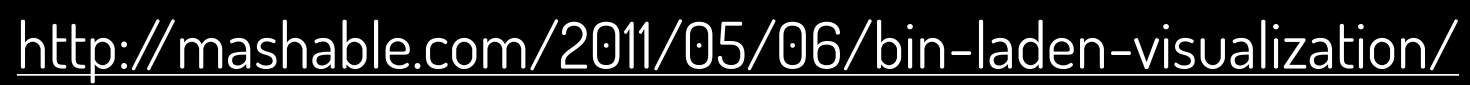
- ▶ Con apoyo de la librería Pattern
- ▶ Tipo I: 'Las sucias calles de la ciudad'
- ▶ Tipo II: 'Barcelona tiene buen clima'
- ▶ Tipo III: 'Odio el servicio técnico de Movistar'
- ▶ Tipo IV: 'Vodafone apesta'
- ▶ Abrir `Pattern/basicSentiment.py`

# MOOD ANALYSIS

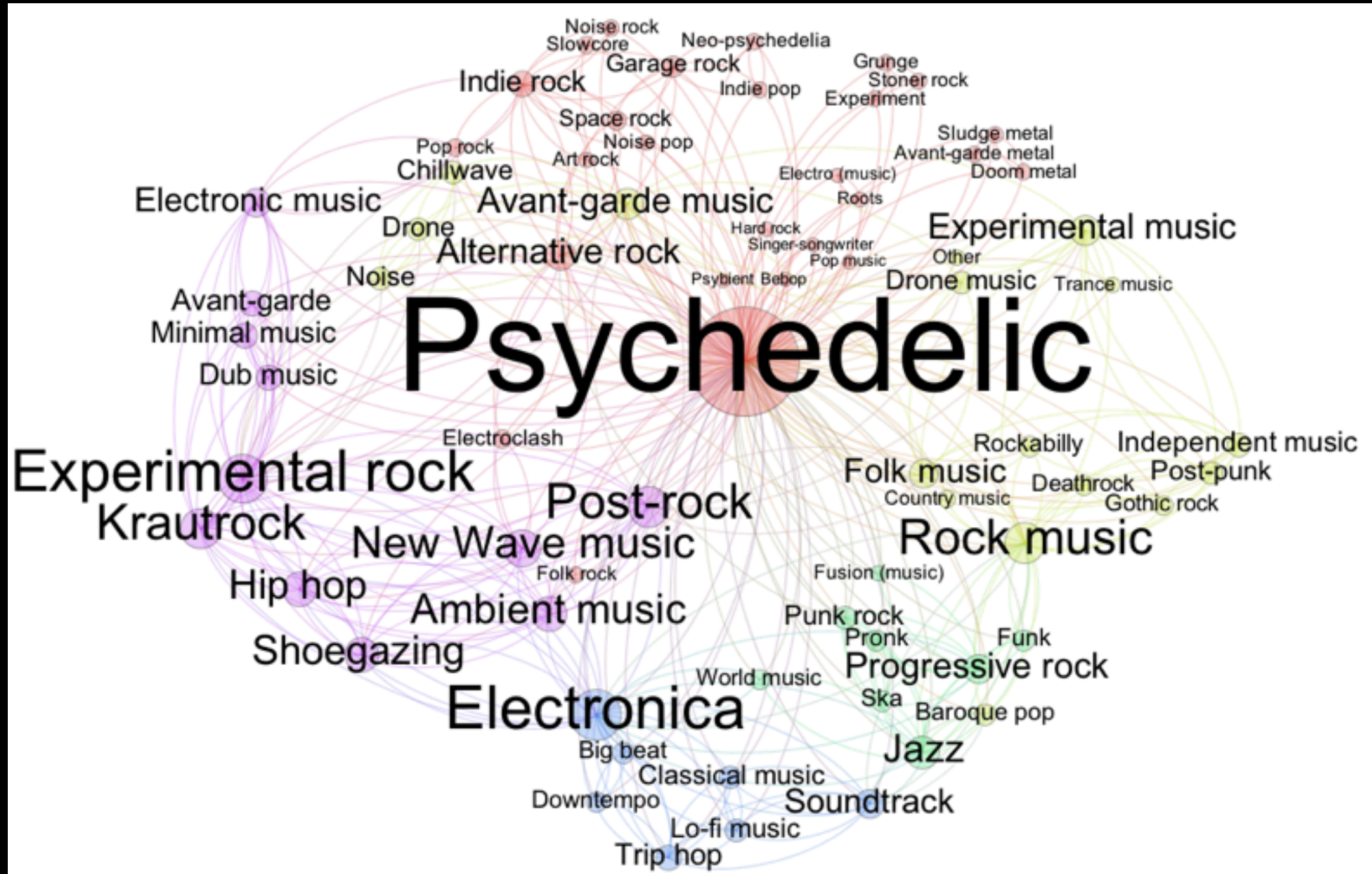
- ▶ Con apoyo de la librería Pattern
- ▶ Es una categorización
- ▶ No haremos caso de las negaciones
- ▶ Dos mood: Tristeza y Alegría
- ▶ Abrir Pattern/basicMood.py

GEPHI







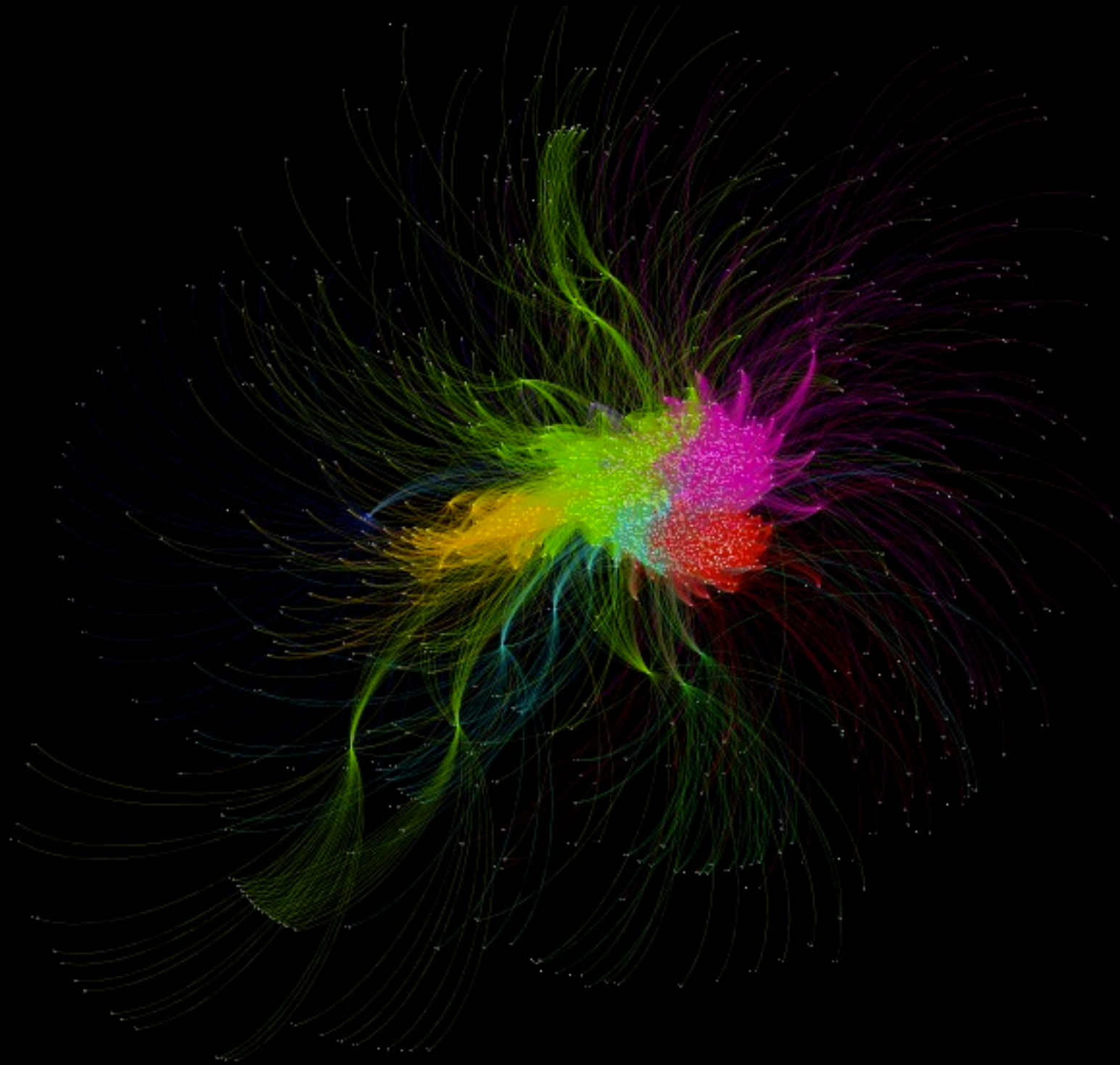


<http://blog.ouseful.info/2012/07/03/visualising-related-entries-in-wikipedia-using-gephi/>



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License



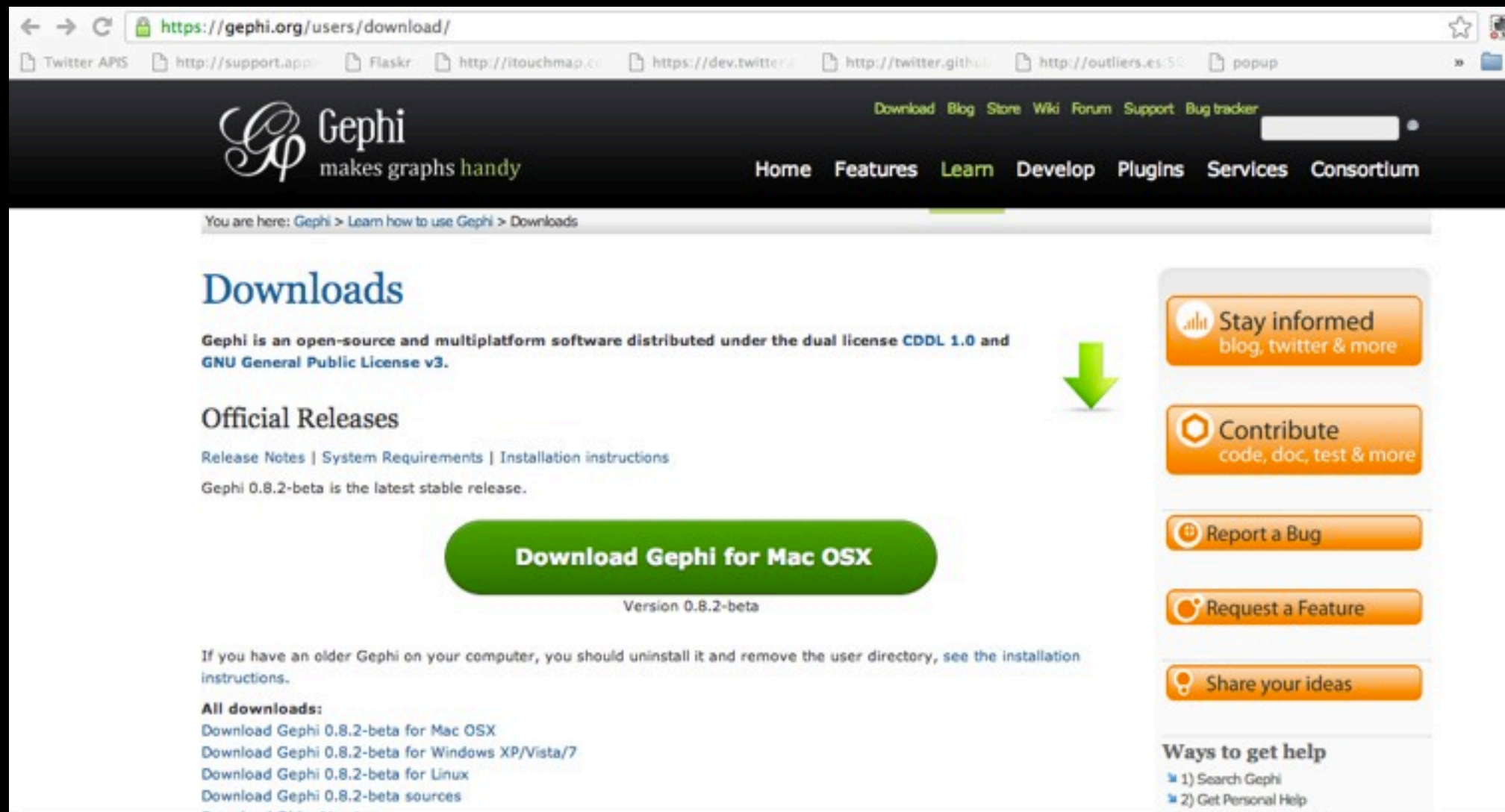


Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License

# DOWNLOAD

<https://gephi.org/users/download/>

<https://launchpad.net/gephi/+download>



The screenshot shows the Gephi Downloads page in a web browser. The browser's address bar displays the URL <https://gephi.org/users/download/>. The page features the Gephi logo and tagline "makes graphs handy" at the top left. A navigation menu at the top right includes links for Download, Blog, Store, Wiki, Forum, Support, and Bug tracker. Below this, a secondary menu lists Home, Features, Learn, Develop, Plugins, Services, and Consortium. The main content area is titled "Downloads" and includes the text: "Gephi is an open-source and multiplatform software distributed under the dual license CDDL 1.0 and GNU General Public License v3." Under the "Official Releases" section, there are links for Release Notes, System Requirements, and Installation instructions, followed by the statement: "Gephi 0.8.2-beta is the latest stable release." A large green button labeled "Download Gephi for Mac OSX" is prominently displayed, with "Version 0.8.2-beta" written below it. To the right of the main content, a vertical sidebar contains five orange buttons: "Stay informed" (with a RSS icon), "Contribute" (with a gear icon), "Report a Bug" (with a bug icon), "Request a Feature" (with a plus icon), and "Share your ideas" (with a lightbulb icon). At the bottom of the sidebar, a "Ways to get help" section lists "1) Search Gephi" and "2) Get Personal Help".



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License

# INSTALAR

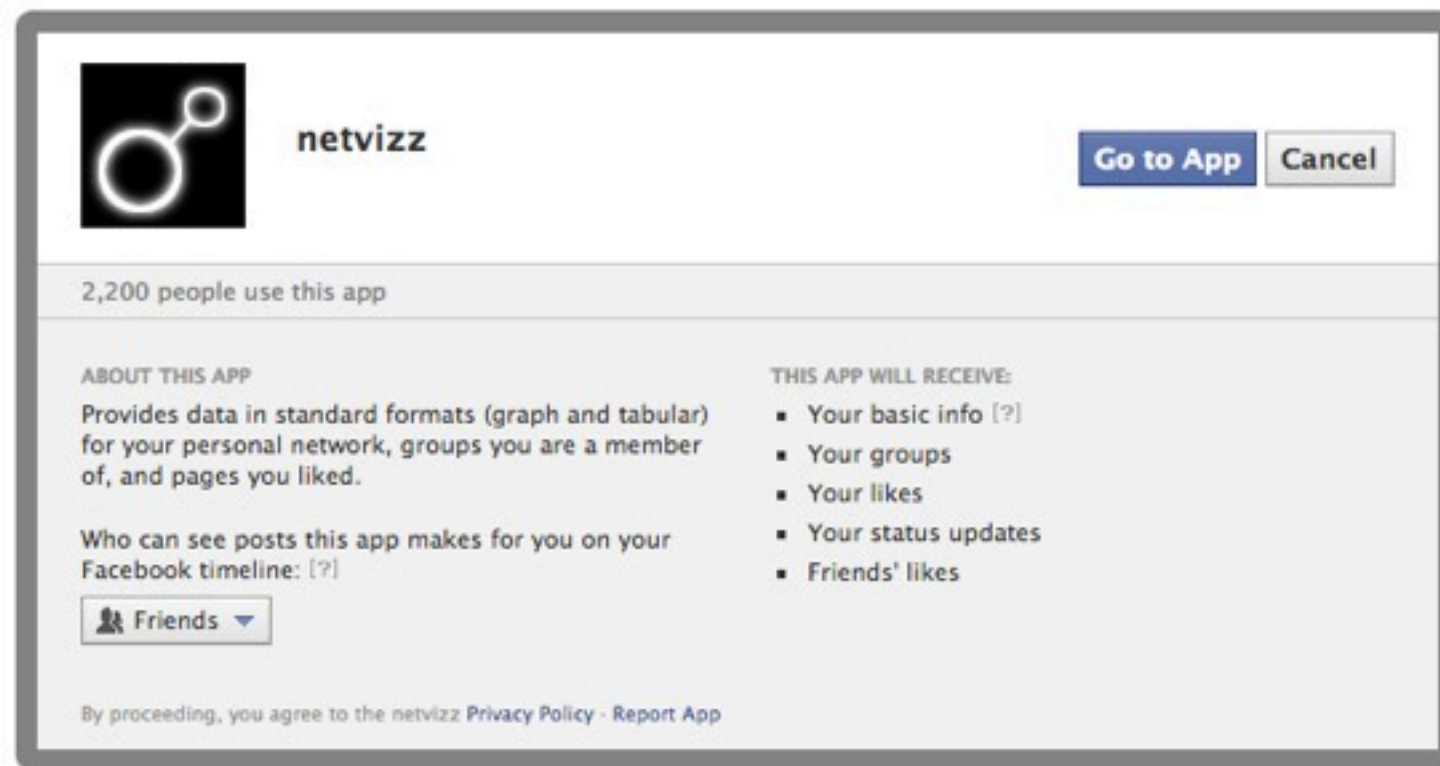
<https://gephi.org/users/install/>



The screenshot shows the Gephi website's installation page. The header features the Gephi logo and navigation links: Download, Blog, Store, Wiki, Forum, and Support. The main content area is titled "Installing Gephi" and includes a subheading "Installing the software". Under the "Windows" section, two steps are listed: 1. Be sure you have a recent Java JRE installed on your system. Download Free Java [here](#). 2. After the download completes, run the installer and follow the steps. Below the text is a screenshot of the "Setup - Gephi" window. The window has a blue sidebar with a computer icon and a main white area with the title "Welcome to the Gephi Setup Wizard". The text in the main area reads: "This will install Gephi 0.7 on your computer. It is recommended that you close all other applications before continuing. Click Next to continue, or Cancel to exit Setup."

# APLICACIONES: NETVIZZ

<https://apps.facebook.com/netvizz/>





# APLICACIONES: NETVIZZ

## netvizz v0.8

This application allows you to extract data from different sections of the Facebook platform for research purposes. It creates network files in the [gdf format](#) (a simple text format that specifies a graph) as well as statistical files using a [tab-separated format](#).

These files can then be analyzed and visualized using graph visualization software such as the powerful and very easy to use [gephi](#) platform or statistical tools such the interactive visualization software [Mondrian](#).

Big networks may take some time to process. **Be patient!**

Privacy policy and credits are [here](#). Non-commercial use only.

Developing and hosting netvizz costs time and money. If the tool is useful for you, please consider to

[Donate](#)

## your personal friend network:

Creates a network file with all the friendship connetions in your personal network.

**Step 1** – Select user data to include in the file (sex, interface language, and account age ranking are standard):

☒ friends' like and post count (public and visible to logged user), includes counts for received likes and comments on posts, adds an additional  $\pm 4$  seconds of waiting time per friend

**Step 2** – create a gdf file from your personal network by clicking [here](#)

file fields: sex: user specified sex, locale: user selected interface language, agerank: accounts ranked by creation date where 1 is youngest, like\_count: number of user likes, post\_count: number of user posts, post\_like\_count: number of likes on user's posts, post\_comment\_count: number of comments on user's posts, post\_engagement\_count: post\_comment\_count + post\_like\_count

**Attention:** data depends on your friends' privacy settings and the filtering choices you made for your newsfeed.



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License

# APLICACIONES: NETVIZZ

## netvizz v0.8

getting connections (181):  
0 35 70 105 140 175

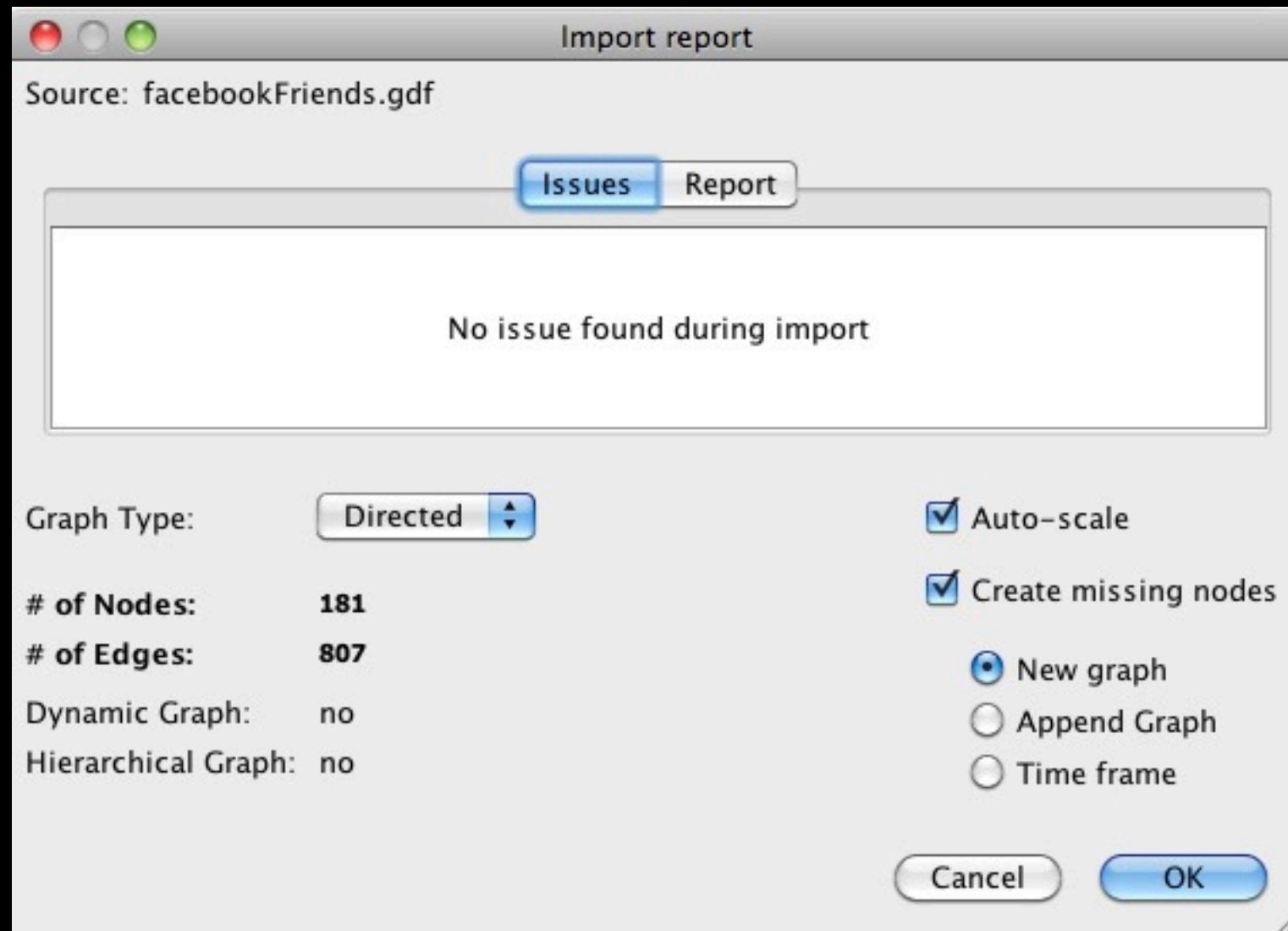
## download

181 nodes, 807 edges

Your **gdf file** (right click, save as...).

Your **tab file** (right click, save as...).

# APLICACIONES: NETVIZZ



# MINERÍA DE RELACIONES

- ▶ Algunas métricas de red
  - ▶ Centralidad: Closeness, Betweenness, PageRank
  - ▶ Densidad
  - ▶ Diámetro
  - ▶ Separación media
  - ▶ Coeficiente de clustering
  - ▶ Componentes conectadas



# APLICACIONES: NETVIZZ

- ▶ Cargar el fichero en Gephi:
  - ▶ Aplicar el layout Force Atlas con Attraction Strength=0.1
  - ▶ Rankear nodos por Degree. ¿Qué significa el Degree en esta red? Rankear las etiquetas por Degree también
  - ▶ Particionar por sexo, mirar la tarta de reparto
  - ▶ Particionar por locale, mirar la tarta de reparto



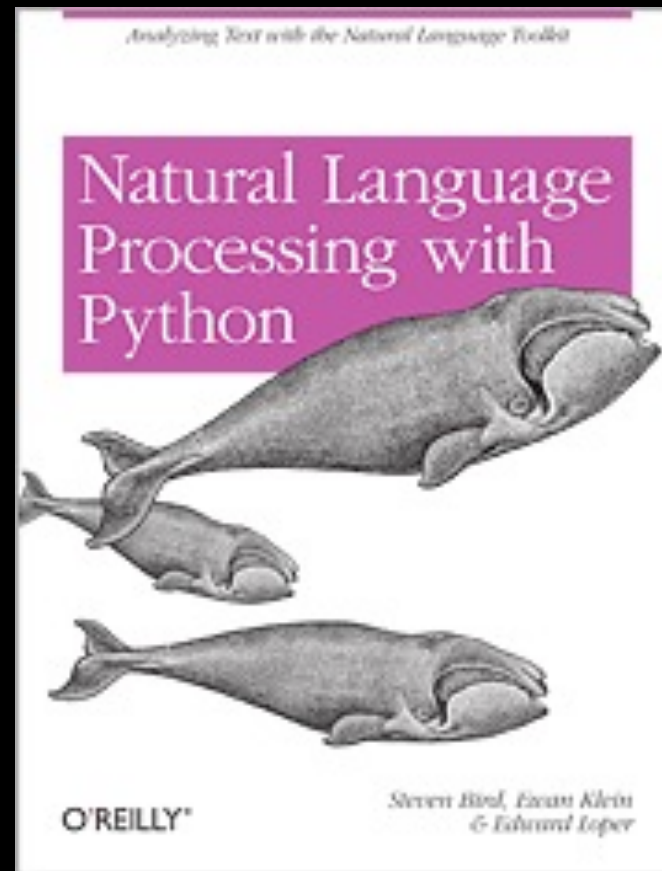
# APLICACIONES: NETVIZZ

## ► Métricas:

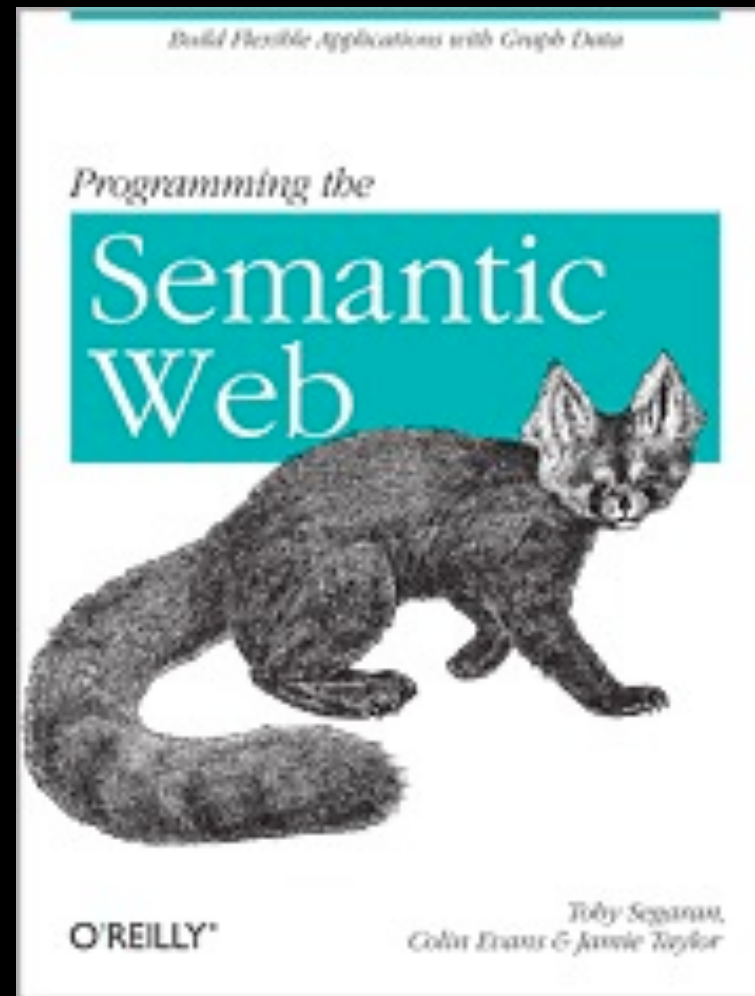
- ¿Cuál es el degree medio? ¿Cuál es el amigo con más degree? ¿Hay alguno con degree 'cero'? ¿¿QUÉ TIPO DE AMIGO ES ESE??
- ¿Cuál es la densidad del grafo? ¿Qué dice eso de mis grupos sociales?
- ¿Cuál es el camino medio? ¿Y el diámetro de red?
- ¿Cuál es el coeficiente de clustering?
- ¿Cuál es el amigo con mayor centralidad?
- Calcular modularidad y particionar por comunidad. ¿Cuántos grupos tengo? ¿Coinciden con familia, colegio/universidad y trabajo(s)?
- Configurar Preview y salvar en PDF al gusto



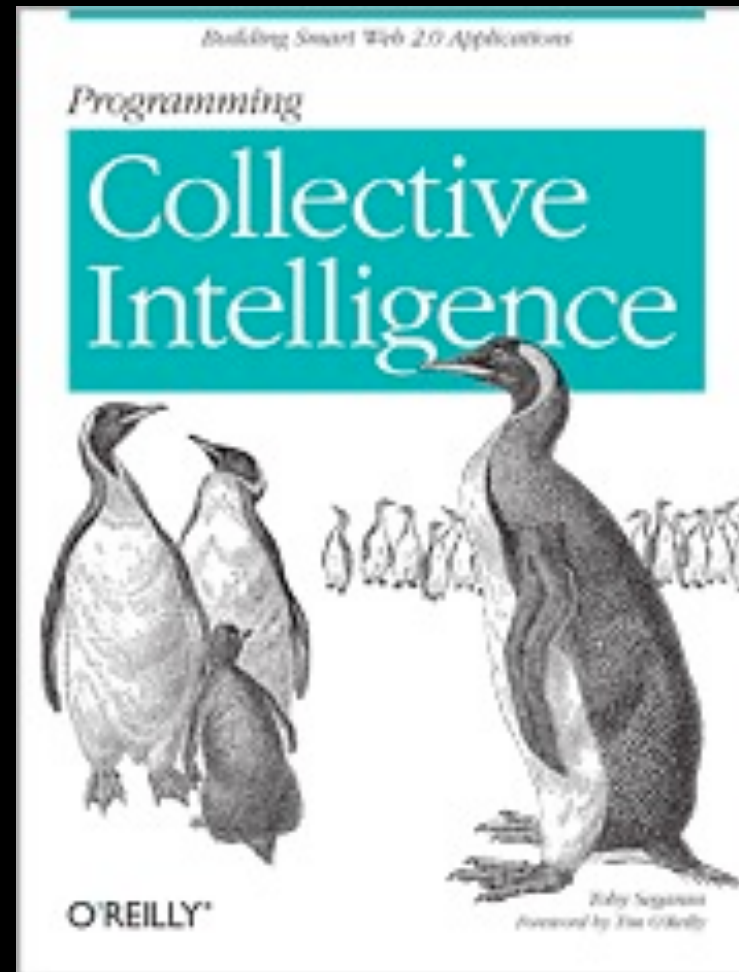
# REFERENCIAS



<http://shop.oreilly.com/product/9780596516499.do>



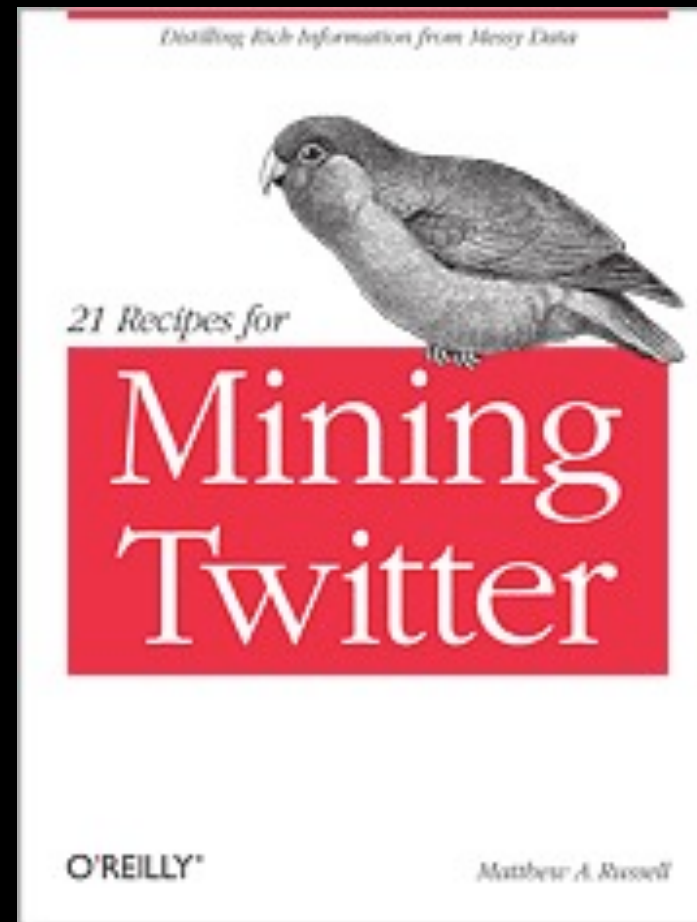
<http://shop.oreilly.com/product/9780596153823.do>



<http://shop.oreilly.com/product/9780596529321.do>



<http://shop.oreilly.com/product/0636920010203.do>

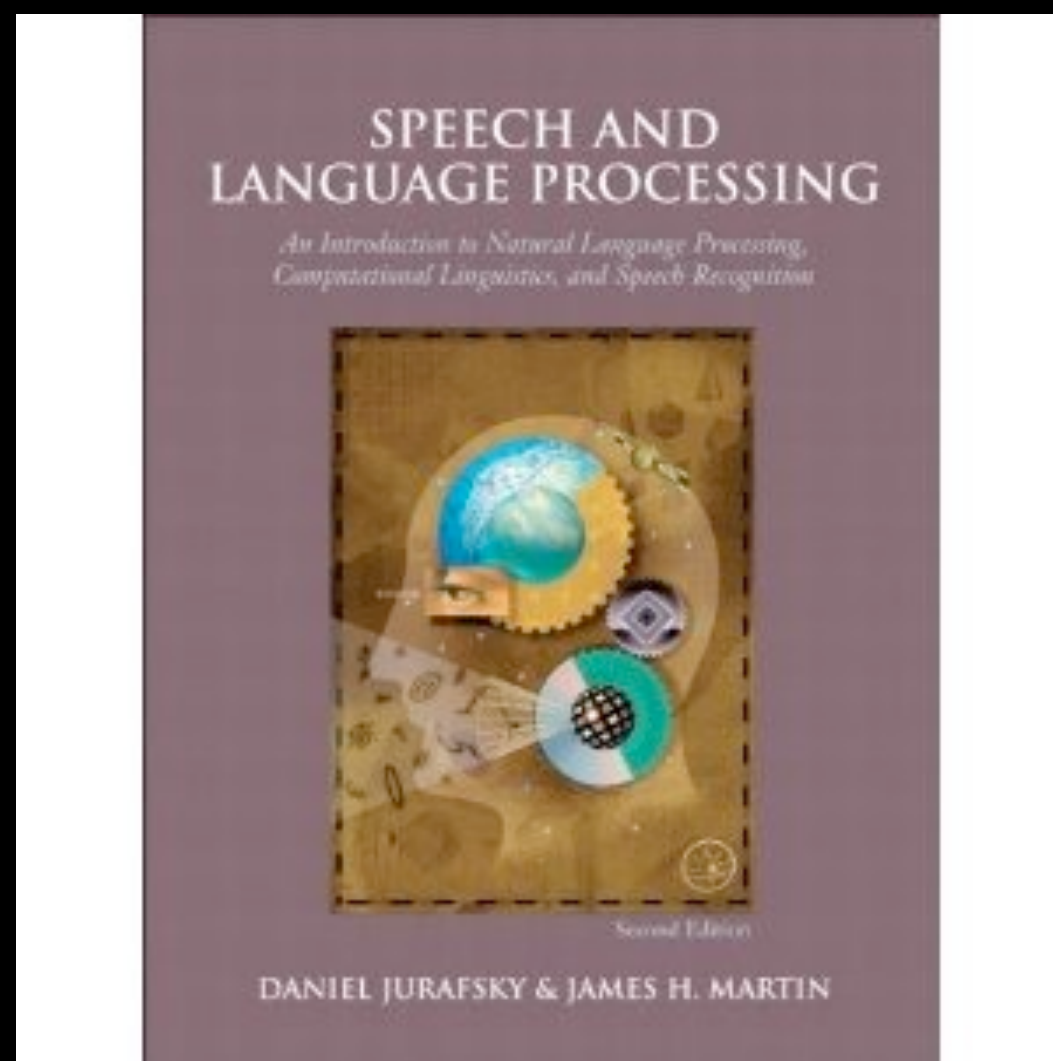


<http://shop.oreilly.com/product/0636920018261.do>

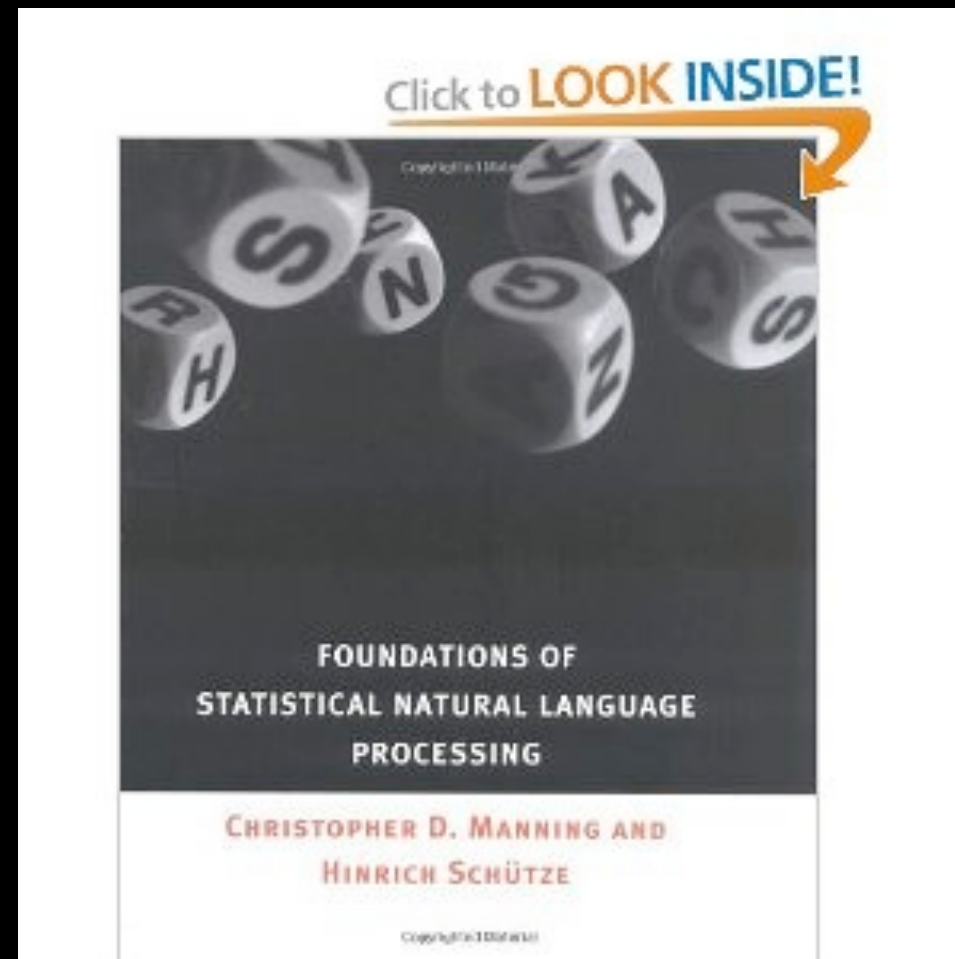




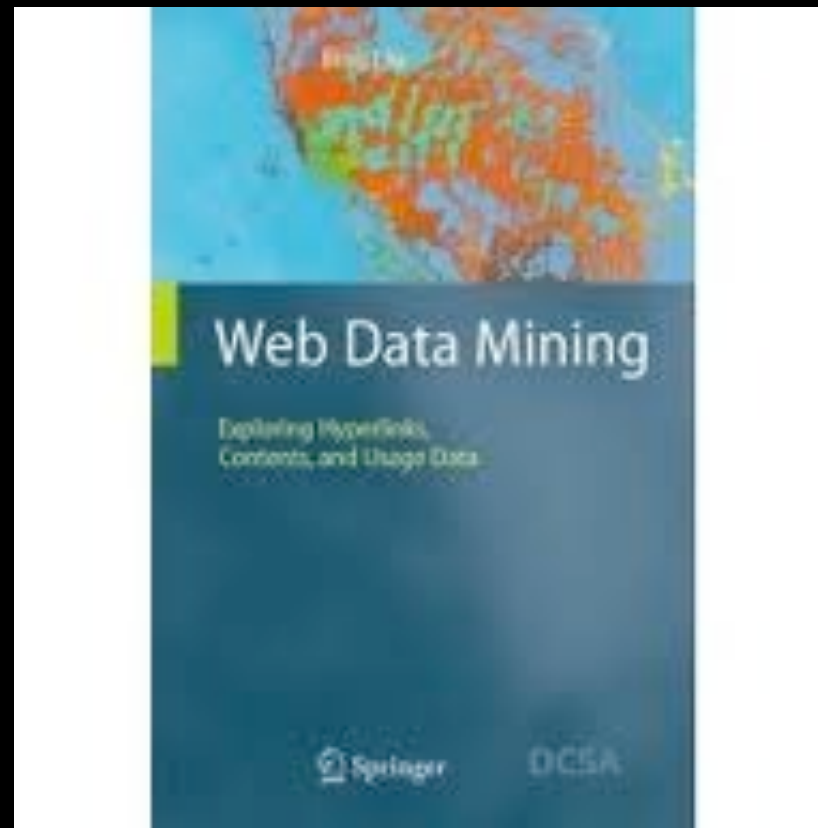
<http://shop.oreilly.com/product/0636920020424.do>



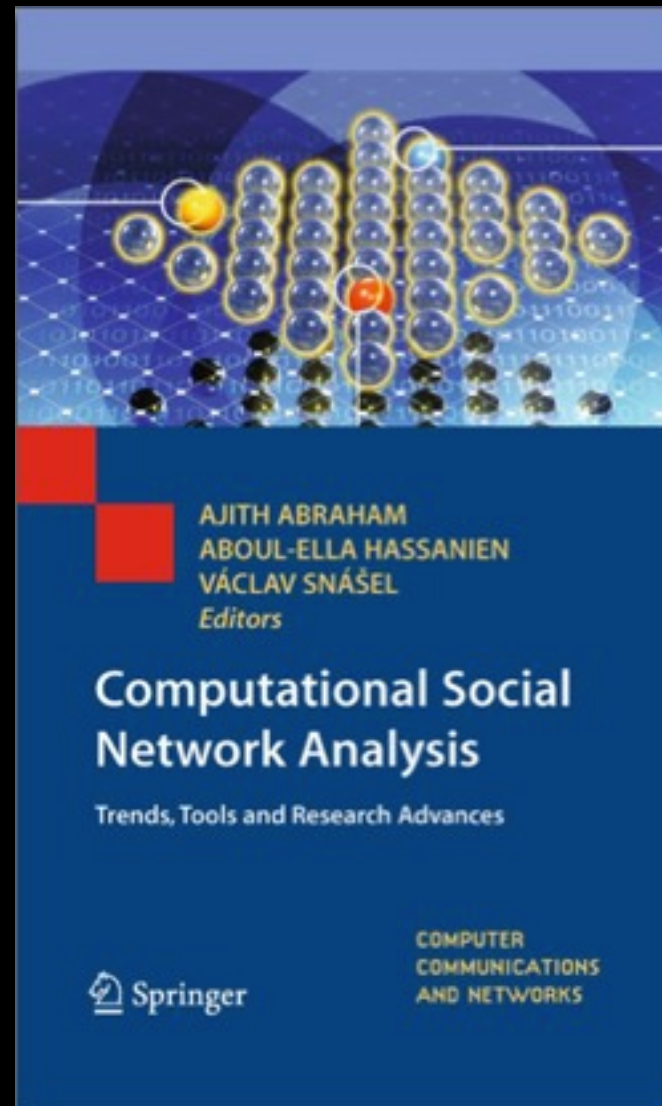
<http://www.amazon.com/Speech-Language-Processing-Daniel-Jurafsky/dp/0131873210>



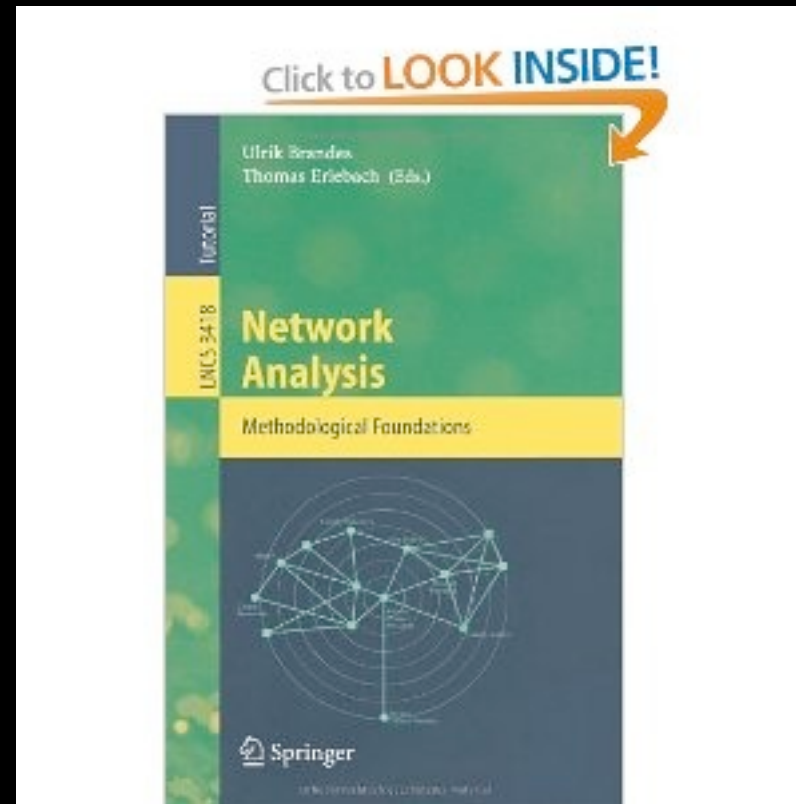
<http://www.amazon.com/Foundations-Statistical-Natural-Language-Processing/dp/0262133601>



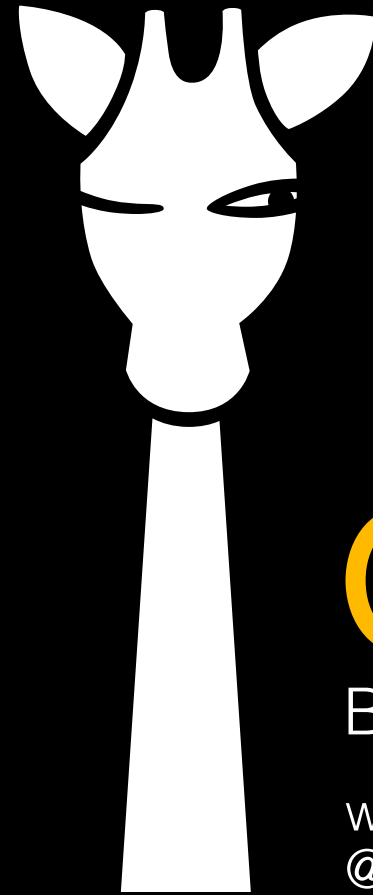
<http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3540378812>



<http://www.springer.com/computer/communication+networks/book/978-1-84882-228-3>



<http://www.amazon.com/Network-Analysis-Methodological-Foundations-Theoretical/dp/3540249796>



# Outliers

Because differences matter.

[www.outliers.es](http://www.outliers.es)  
[@outliers\\_es](https://twitter.com/outliers_es)



Este trabajo está licenciado como Creative Commons Attribution 3.0 Unported License