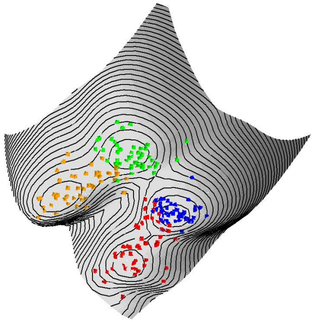
IIC2433 Minería de Datos

Profesor: Karim Pichara
Departamento de Ciencia de la Computación

Who is the professor?



Data Mining

- Research area that belongs to Computer Science and Statistics (and may be Physics, and Math)
- It studies how to develop automatic tools to analyze data
- Solutions have to be:
  - fast enough (big data)
  - accurate enough (flexible & adaptive models)

Mining of gold from rocks ⇨ Gold Mining

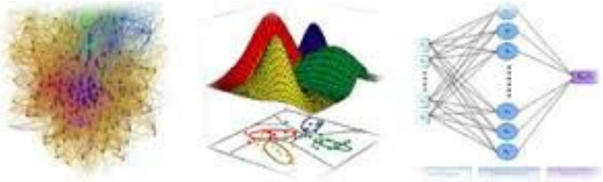
Mining of Data from Information sources ⇨ Data Mining

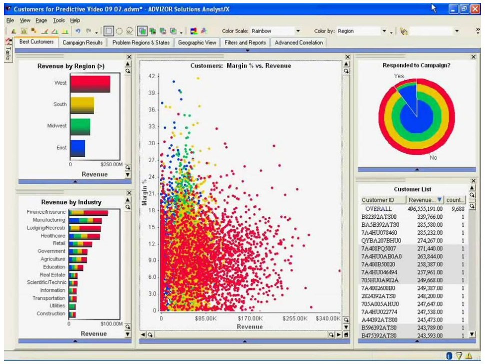
What name to use?

- Data Science: If you want to find a job
- Machine Learning: If you want to hire people
- Artificial Intelligence: If you want to raise money
- Data Mining: Do not use it anymore

Importance of Data Mining

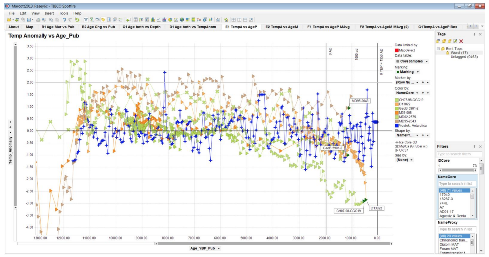
- Current databases can not be analyzed manually
- Automatic tools are needed to perform different tasks on data



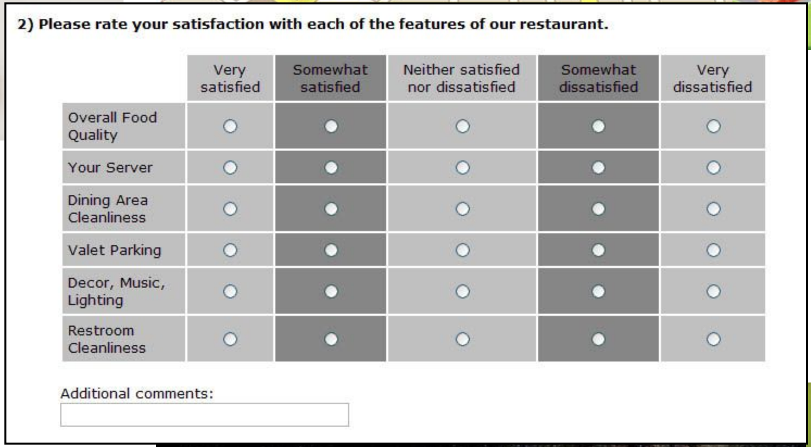


Customer List

Customer ID	Revenue...	count...
OVERALL	496,555,191.00	9,688
B82392ATS00	339,766.00	1
BA5B392ATS0	285,580.00	1
7A4HU078460	285,232.00	1
QYBA207BHU0	274,267.00	1
7A408PQ5007	271,440.00	1
7A4HU0AB0A0	263,844.00	1
7A400B50020	258,387.00	1
7A4HU046494	257,961.00	1
705HU0A902A	249,668.00	1
7A4002600B0	249,387.00	1
2824392ATS0	248,200.00	1
705A005AHU0	247,647.00	1
7A4HU022774	247,538.00	1
A44392ATS00	245,473.00	1
B596392ATS0	243,789.00	1
B475392ATS0	243,593.00	1



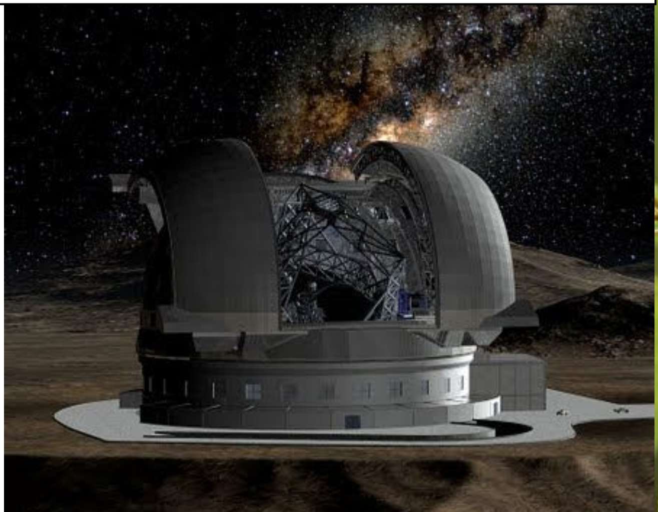
Data Everywhere!!



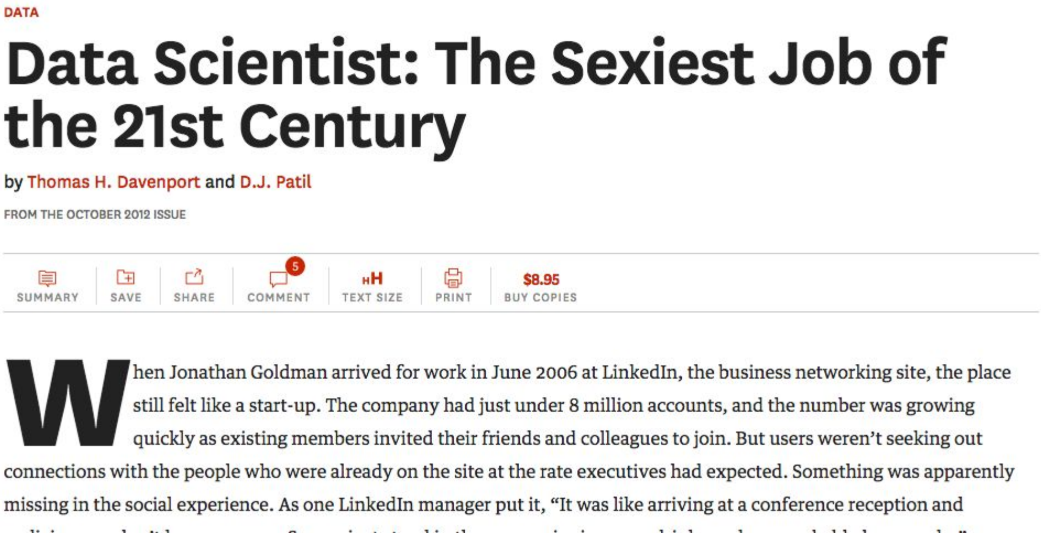
2) Please rate your satisfaction with each of the features of our restaurant.

	Very satisfied	Somewhat satisfied	Neither satisfied nor dissatisfied	Somewhat dissatisfied	Very dissatisfied
Overall Food Quality	○	○	○	○	○
Your Server	○	○	○	○	○
Dining Area Cleanliness	○	○	○	○	○
Valet Parking	○	○	○	○	○
Decor, Music, Lighting	○	○	○	○	○
Restroom Cleanliness	○	○	○	○	○

Additional comments:



Harvard Business Review



DATA

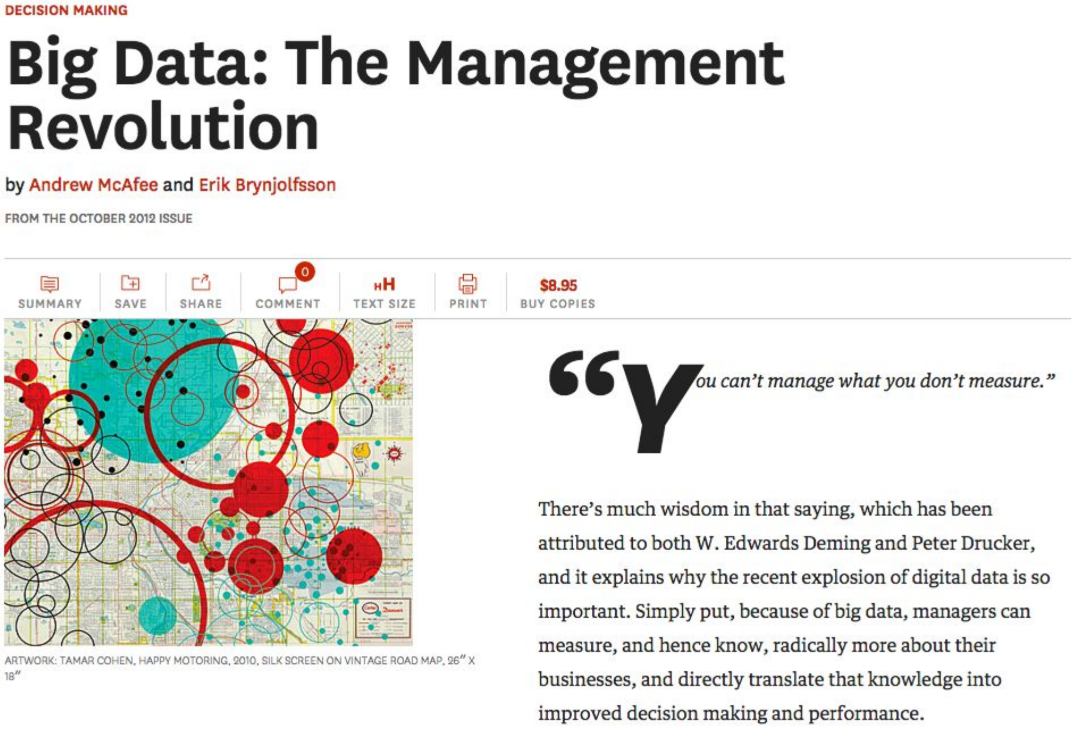
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/



DECISION MAKING

Big Data: The Management Revolution

by Andrew McAfee and Erik Brynjolfsson

FROM THE OCTOBER 2012 ISSUE

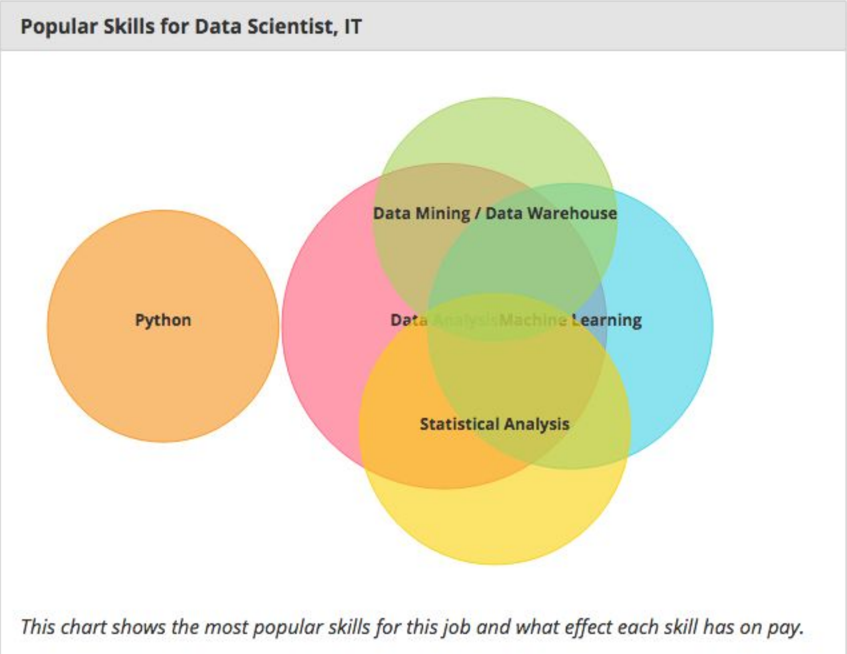
ARTWORK: TAMAR COHEN, HAPPY MOTORING, 2010, SILK SCREEN ON VINTAGE ROAD MAP, 26" X 18"

"You can't manage what you don't measure."

There's much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker, and it explains why the recent explosion of digital data is so important. Simply put, because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance.

https://hbr.org/2012/10/big-data-the-management-revolution&cm_sp=Article-_-Links-_-Top%20of%20Page%20Recirculation

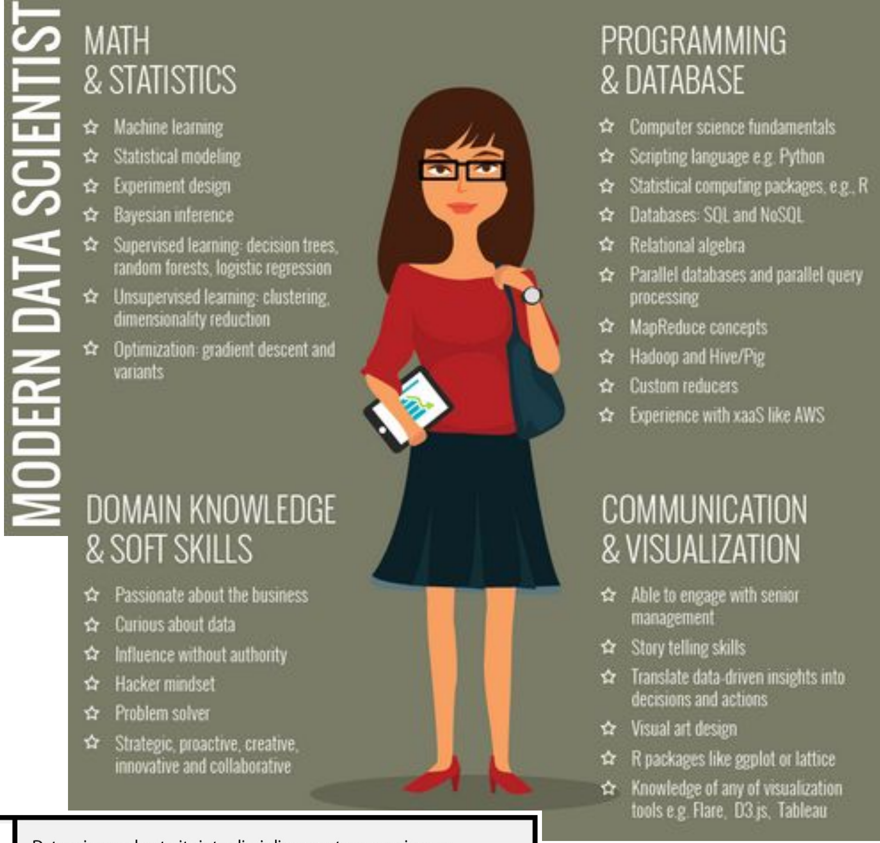
Skills you will obtain from IIC2433



Popular Skills for Data Scientist, IT

This chart shows the most popular skills for this job and what effect each skill has on pay.

http://www.payscale.com/research/US/Job=Data_Scientist,_IT/Salary



MODERN DATA SCIENTIST

MATH & STATISTICS

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

PROGRAMMING & DATABASE

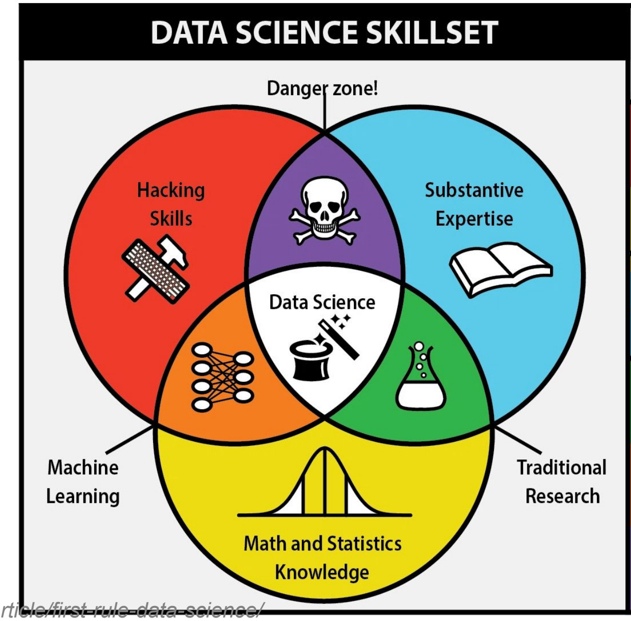
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



DATA SCIENCE SKILLSET

Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills, math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

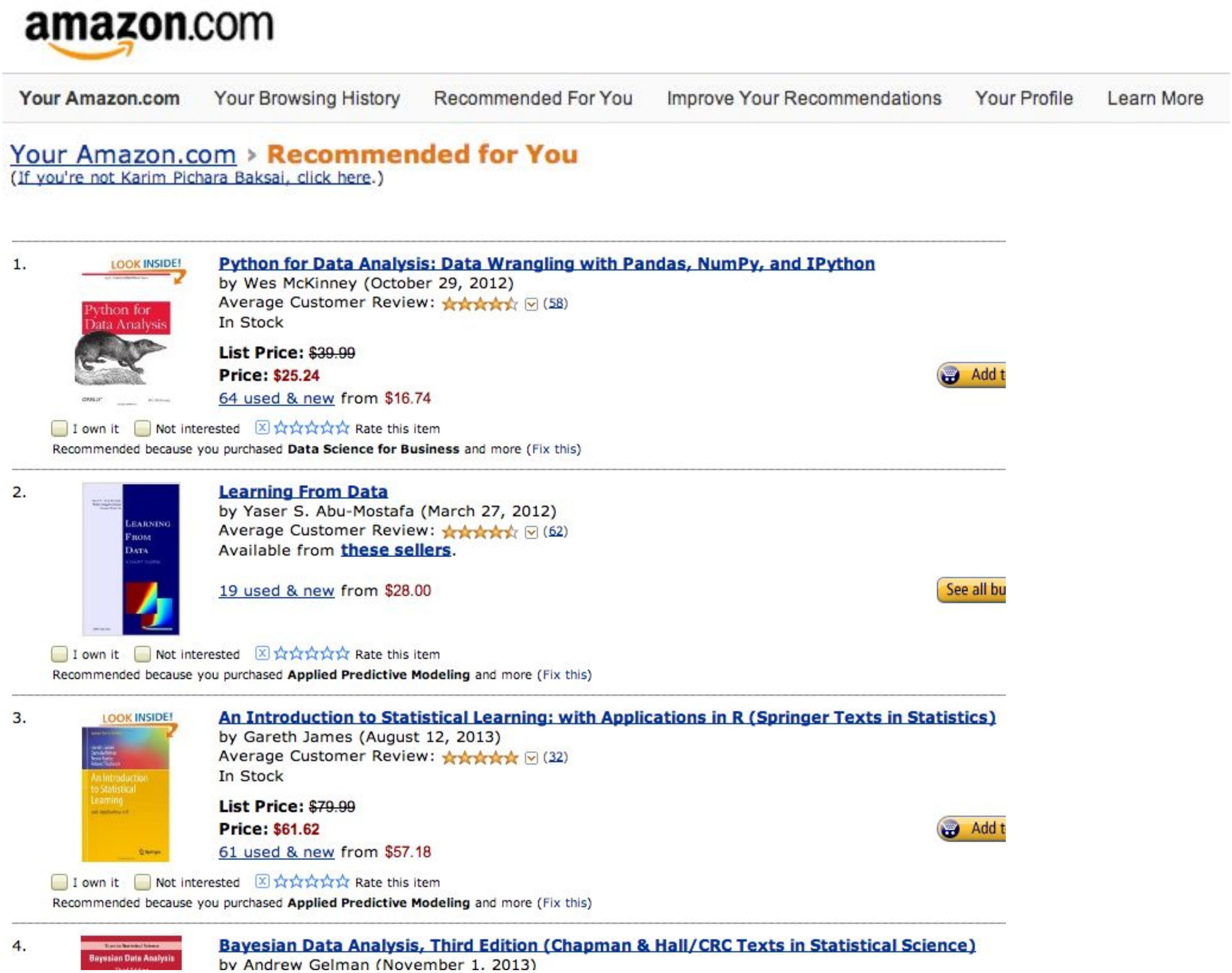
Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

http://berkeleysciencereview.com/article/first-rule-data-science/

Recommender Systems



Your Amazon.com › **Recommended for You**
(If you're not Karim Pichara Baksai, click here.)

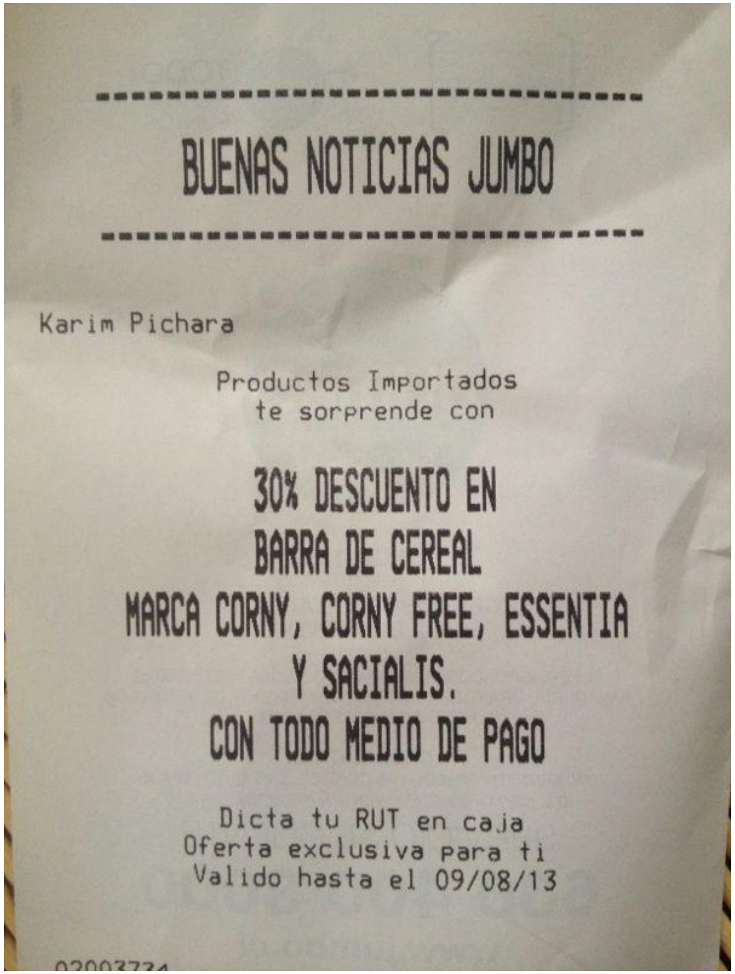
1. **Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython**
by Wes McKinney (October 29, 2012)
Average Customer Review: ★★★★☆ (58)
In Stock
List Price: ~~$39.99~~
Price: $25.24
64 used & new from $16.74
Add t
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **Data Science for Business** and more (Fix this)

2. **Learning From Data**
by Yaser S. Abu-Mostafa (March 27, 2012)
Average Customer Review: ★★★★☆ (62)
Available from **these sellers**.
19 used & new from $28.00
See all bu
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **Applied Predictive Modeling** and more (Fix this)

3. **An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)**
by Gareth James (August 12, 2013)
Average Customer Review: ★★★★★ (32)
In Stock
List Price: ~~$79.99~~
Price: $61.62
61 used & new from $57.18
Add t
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **Applied Predictive Modeling** and more (Fix this)

4. **Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)**
by Andrew Gelman (November 1, 2013)

Personalized discounts



BUENAS NOTICIAS JUMBO

Karim Pichara

Productos Importados
te sorprende con

**30% DESCUENTO EN
BARRA DE CEREAL
MARCA CORNY, CORNY FREE, ESSENTIA
Y SACIALIS.
CON TODO MEDIO DE PAGO**

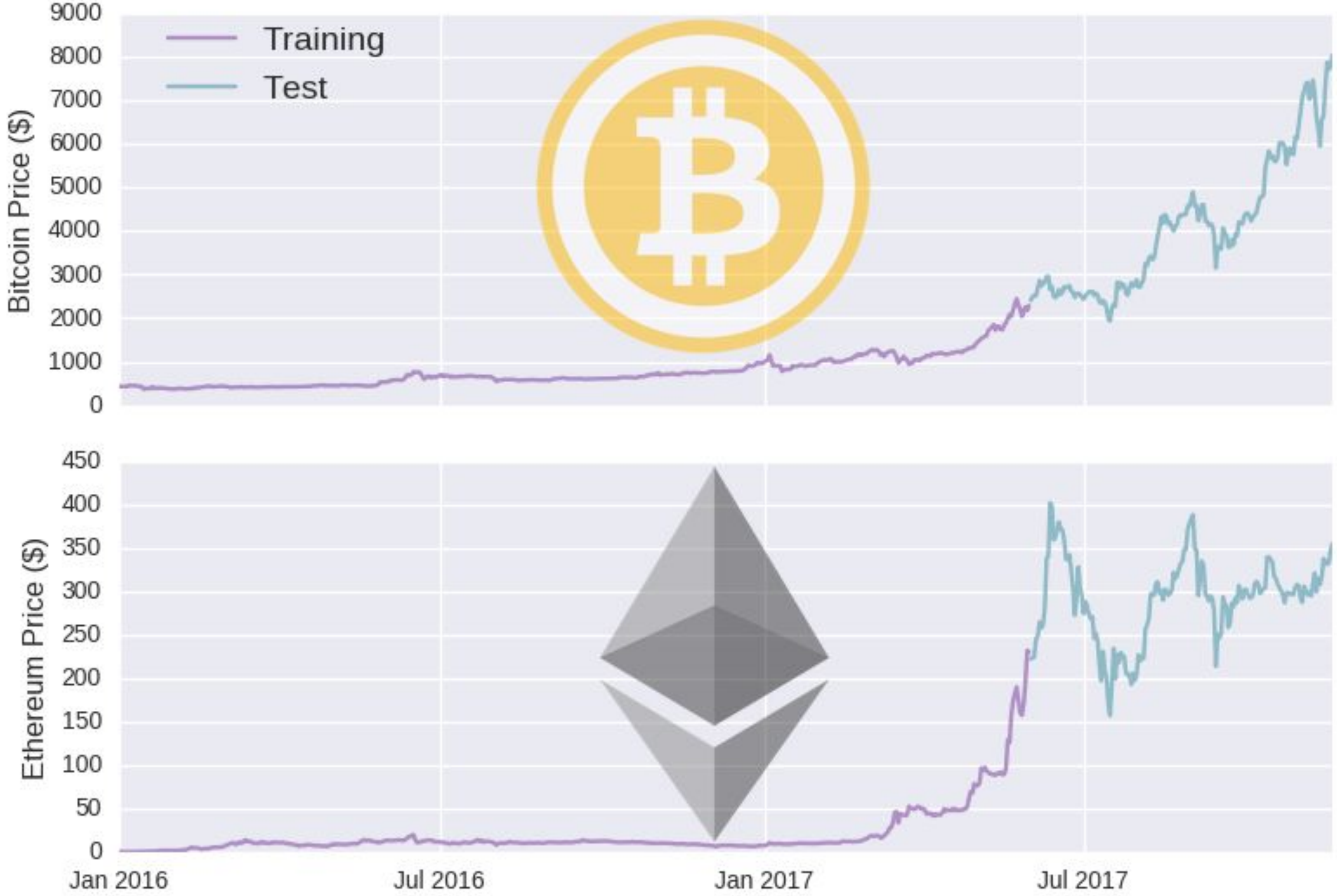
Dicta tu RUT en caja
Oferta exclusiva para ti
Valido hasta el 09/08/13

02003734

Plant based food

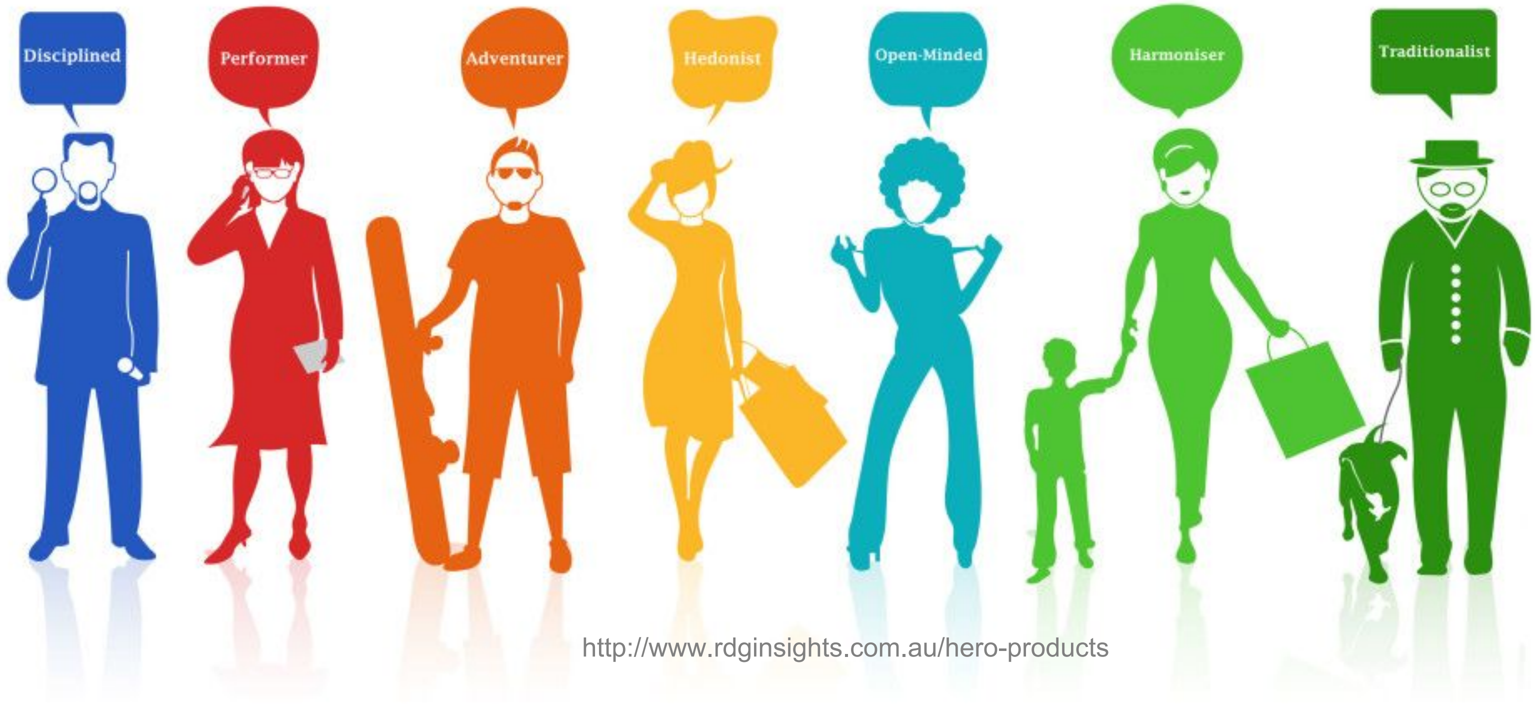
GIUSEPPE
NotCo
NOT CHEESE
Hecha con ingredientes vegetales
Queso CREMA
NOT MAYO
Sabor ORIGINAL
Hecha con ingredientes vegetales
NOT MILK
Made with plant-based ingredients only
Almonds
Peas
Rice
Nuts
Linseed
Coconut
Vanilla
NOT YOGURT
Hecha con ingredientes vegetales

Prediction of Crypto-Currency price



https://dashee87.github.io/deep%20learning/python/predicting-cryptocurrency-prices-with-deep-learning/

Customer profiling

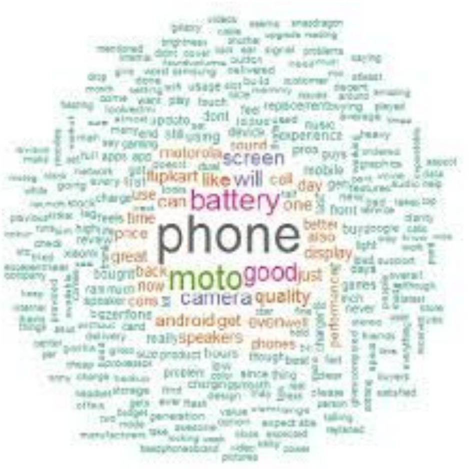


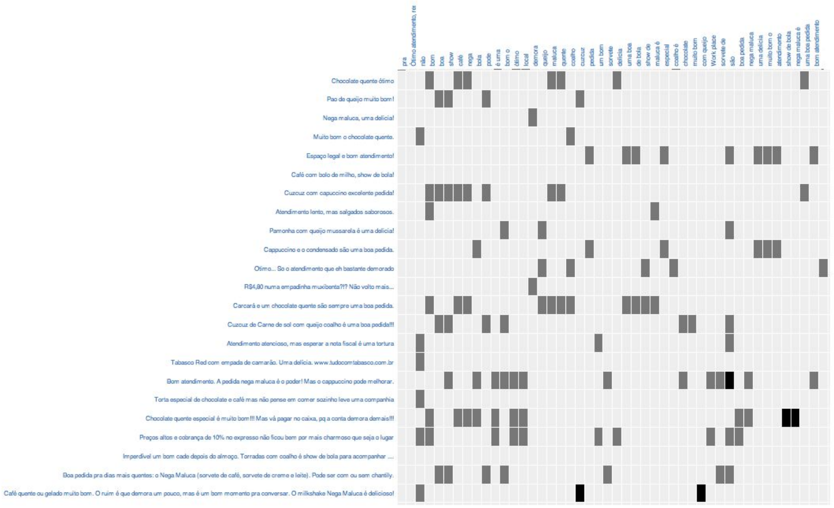
http://www.rdginsights.com.au/hero-products

Customer Abandonment

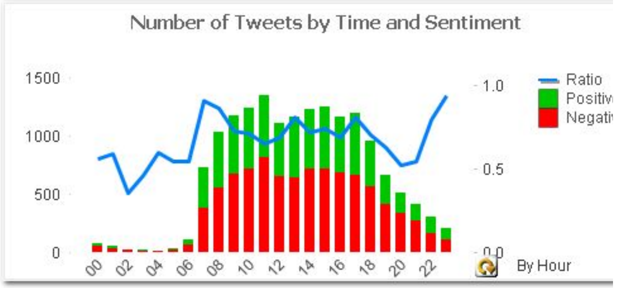
Employee Evaluation

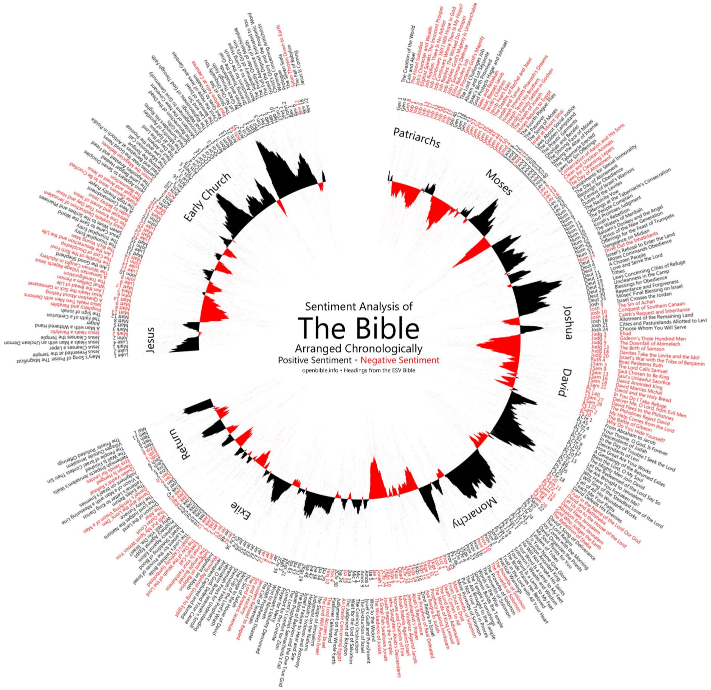
Text mining


phone
battery
moto
good
camera
quality
screen
motorola
android
get
great
flipkart
like
will
speakers

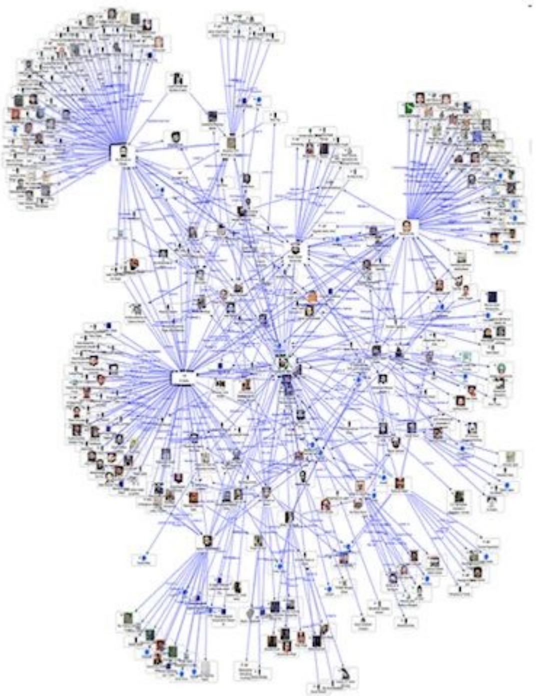

pra
Ótimo atendimento, rer
não
bom
boa
show
café
nega
bola
pode
é uma
bom o
ótimo
local
demora
queijo
maluca
quente
coalho
cuzcuz
pedida
um bom
sorvete
delícia
uma boa
de bola
show de
maluca é
especial
coalho é
chocolate
muito bom
com queijo
Work place
sorvete de
são
boa pedida
nega maluca
uma delícia
muito bom o
atendimento
show de bola
nega maluca é
uma boa pedida
bom atendimento
Chocolate quente ótimo
Pao de queijo muito bom!
Nega maluca, uma delicia!
Muito bom o chocolate quente.
Espaço legal e bom atendimento!
Café com bolo de milho, show de bola!
Cuzcuz com capuccino excelente pedida!
Atendimento lento, mas salgados saborosos.
Pamonha com queijo mussarela é uma delicia!
Cappuccino e o condensado são uma boa pedida.
Otimo... So o atendimento que eh bastante demorado
R$4,80 numa empadinha muxibenta?!? Não volto mais...
Carcará e um chocolate quente são sempre uma boa pedida.
Cuzcuz de Carne de sol com queijo coalho é uma boa pedida!!!
Atendimento atencioso, mas esperar a nota fiscal é uma tortura
Tabasco Red com empada de camarão. Uma delicia. www.tudocomtabasco.com.br
Bom atendimento. A pedida nega maluca é o poder! Mas o cappuccino pode melhorar.
Torta especial de chocolate e café mas não pense em comer sozinho leve uma companhia
Chocolate quente especial é muito bom!!! Mas vá pagar no caixa, pq a conta demora demais!!!
Preços altos e cobrança de 10% no expresso não ficou bem por mais charmoso que seja o lugar
Imperdivel um bom cade depois do almoço. Torradas com coalho é show de bola para acompanhar ....
Boa pedida pra dias mais quentes: o Nega Maluca (sorvete de café, sorvete de creme e leite). Pode ser com ou sem chantily.
Café quente ou gelado muito bom. O ruim é que demora um pouco, mas é um bom momento pra conversar. O milkshake Nega Maluca é delicioso!

fail
worst
sucks
delay
waiting
wtf
delays
hate
shit
damn
wait
awful
nightmare
4p07z
ebay
hates

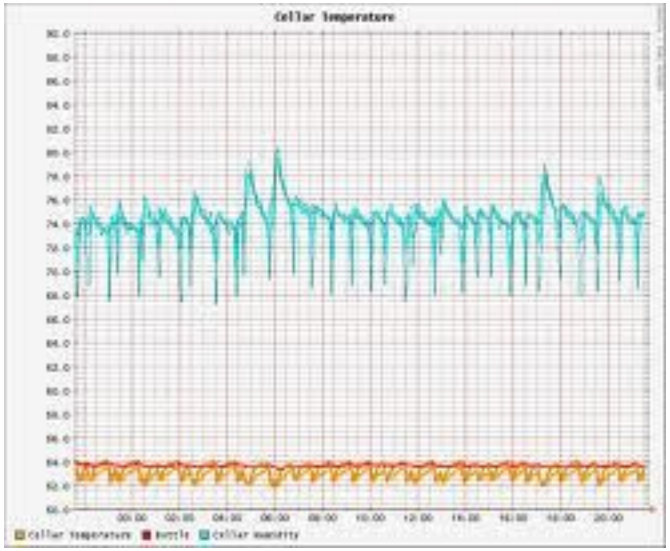

Number of Tweets by Time and Sentiment
1500
1000
500
0
1.0
0.5
0.0
00
02
04
06
08
10
12
14
16
18
20
22
By Hour
Ratio
Positiv
Negati


Sentiment Analysis of
The Bible
Arranged Chronologically
Positive Sentiment • Negative Sentiment
openbible.info • Headings from the ESV Bible
Patriarchs
Moses
Joshua
David
Monarchy
Exile
Return
Jesus
Early Church

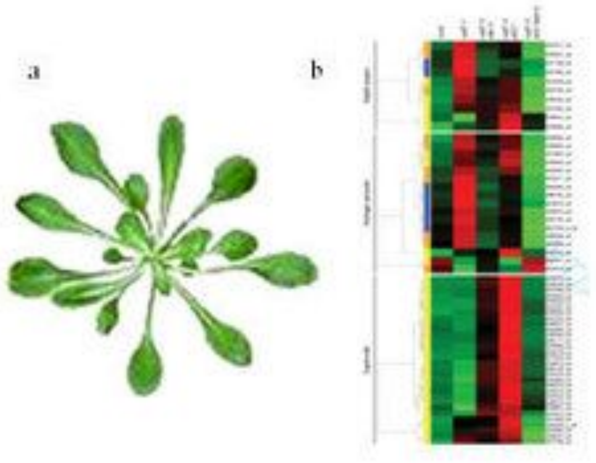
Social Networks data mining

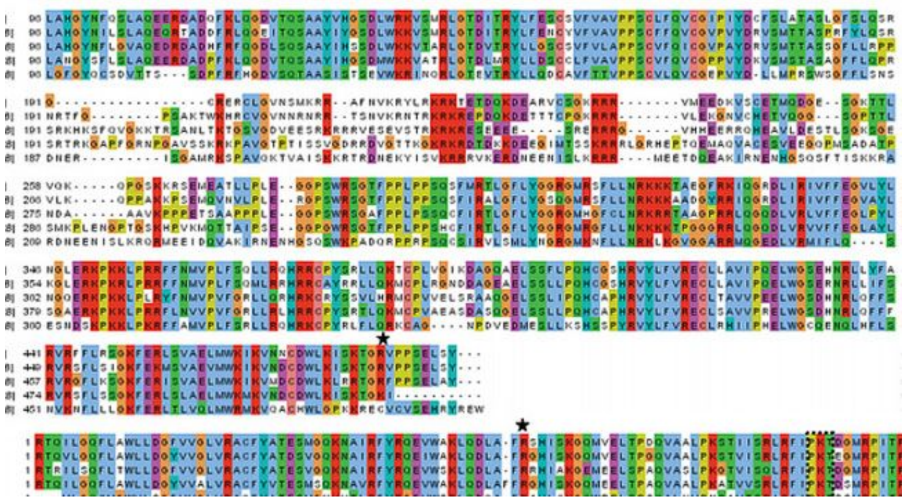


From offline to online world: Reporting and analysis

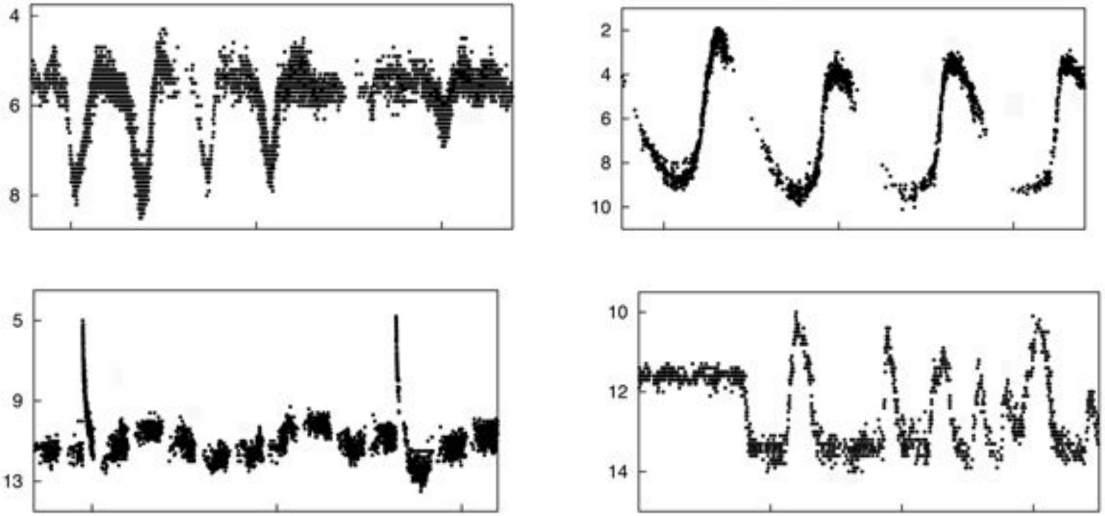


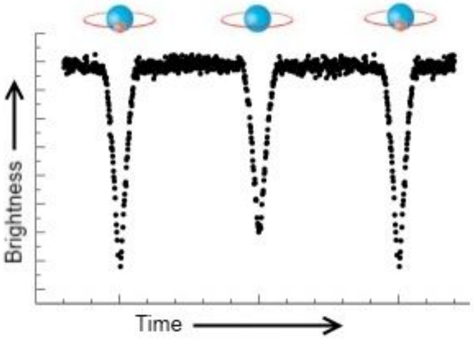
Biotechnology

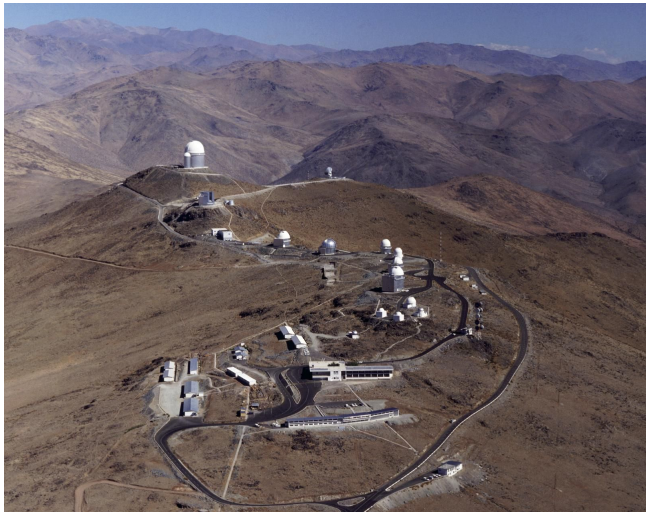




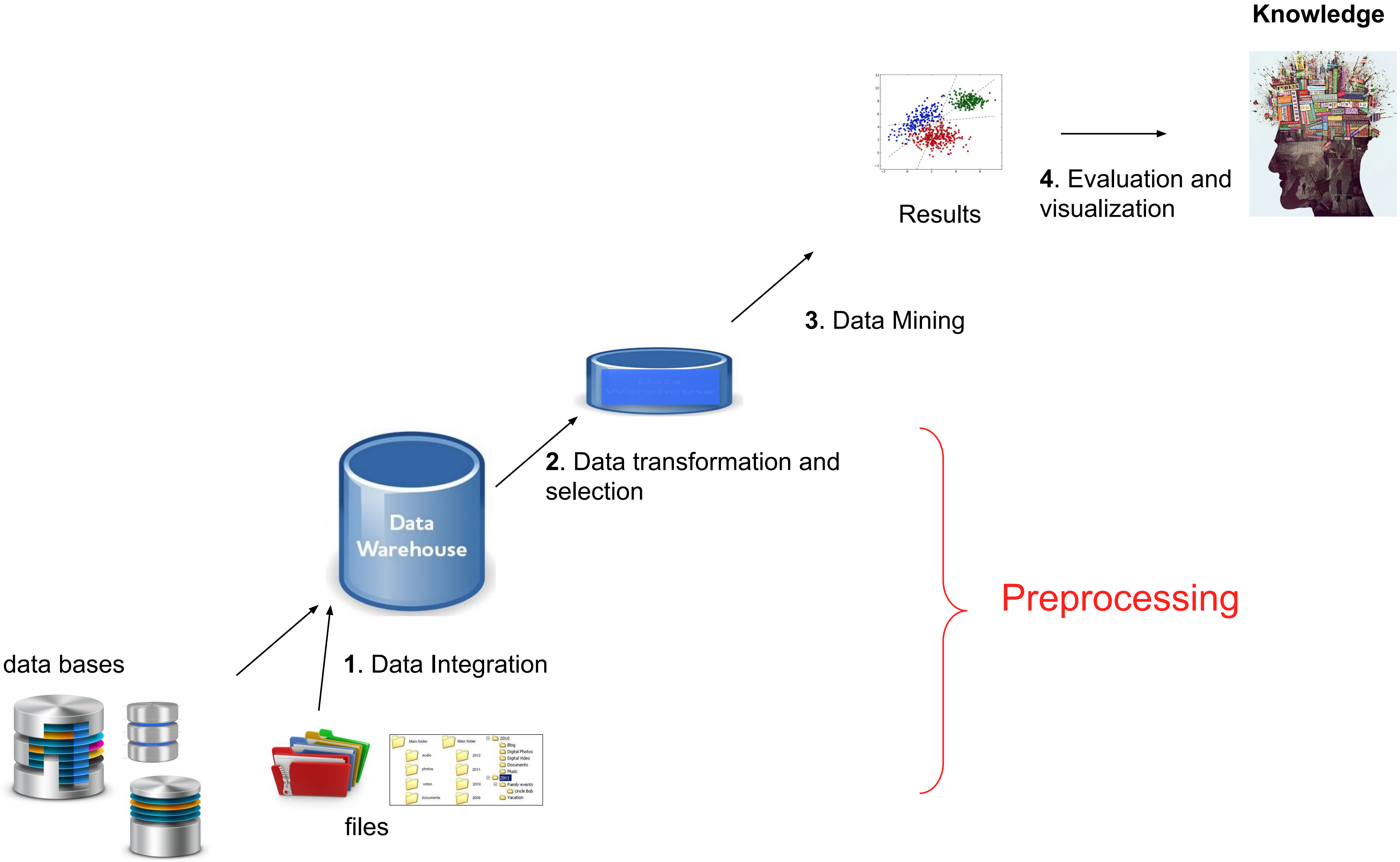
Automatic classification of Variable stars







Contenidos del Curso



Knowledge

Results

4. Evaluation and visualization

3. Data Mining

2. Data transformation and selection

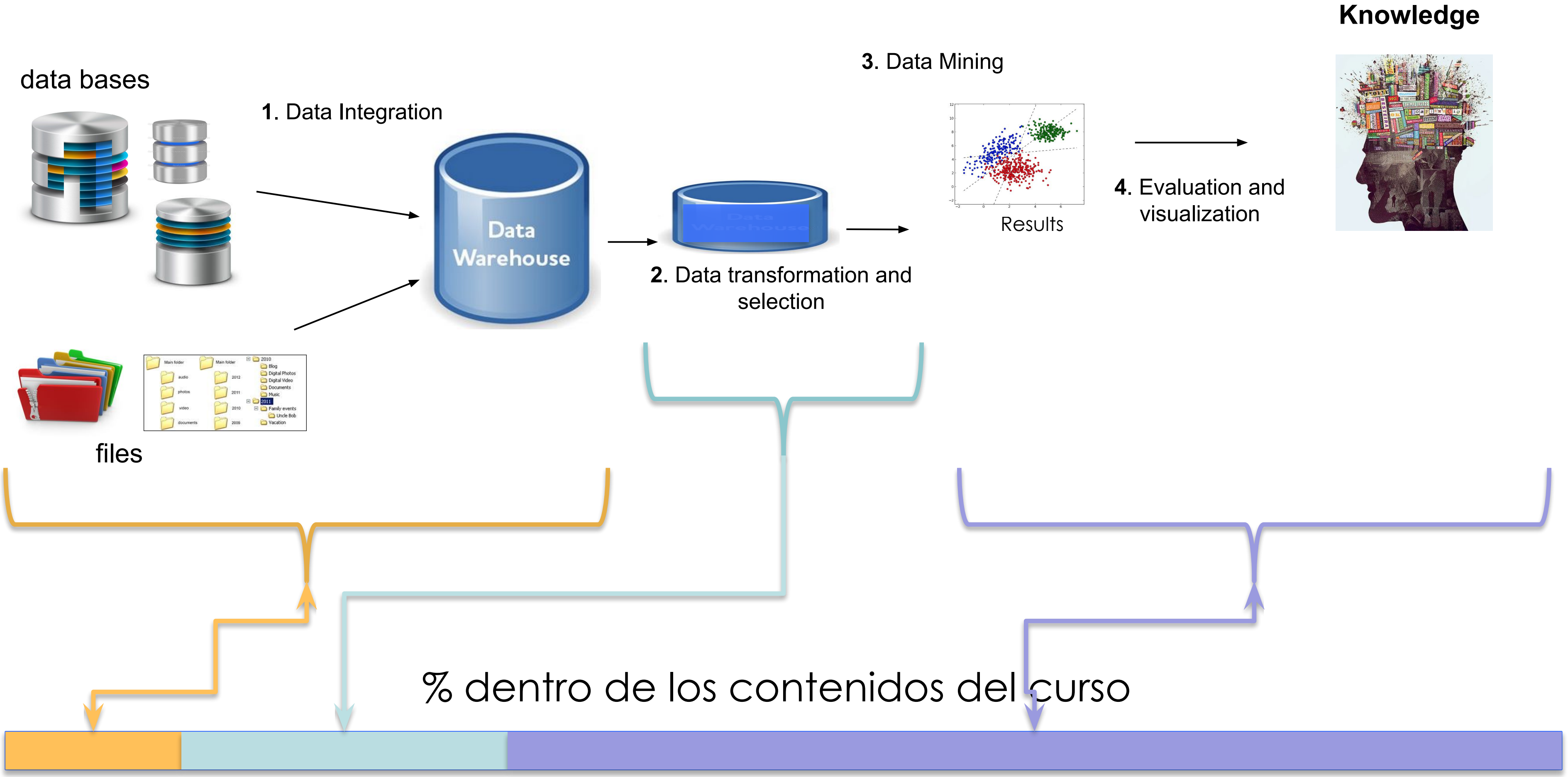
Preprocessing

data bases

1. Data Integration

files

Contenidos del Curso



Knowledge

3. Data Mining

data bases

1. Data Integration

2. Data transformation and selection

4. Evaluation and visualization

files

% dentro de los contenidos del curso

I WANT ANSWERS
CALL THE DATA SCIENTIST. RIGHT NOW!