Εξόρυξη δεδομένων

Α. Κολοβού

Εξόρυξη δεδομένων

Είναι η διαδικασία της αυτόματης ανακάλυψης χρήσιμων πληροφοριών μέσα από μεγάλα σύνολα δεδομένων

Μέσα από την εξόρυξη δεδομένων αντλούμε χρήσιμες πληροφορίες οι οποίες σε οποιαδήποτε άλλη περίπτωση θα έμεναν αγνωστες

Επίσης παρέχουν δυνατότητες πρόβλεψης του αποτελεσματος για μια μελλοντική παρατήρηση, πχ πόσα θα ξοδέψει ένας πελάτης στην επόμενη αγορά του από ένα κατάστημα

1 BUSINESS PROBLEM

WHY?....WHY?....WHY?....

ONE OF THE MANY TRAITS OF

A GOOD DATA SCIENTIST!

Η διαμόρφωση του προβλήματος

Η διαμόρφωση προβλήματος είναι η διαδικασία ανάλυσης ενός προβλήματος για την απομόνωση των επιμέρους στοιχείων που πρέπει να αντιμετωπιστούν για την επίλυσή του.

Η διαμόρφωση του προβλήματος βοηθά στον προσδιορισμό της τεχνικής προσέγγισης του έργου σας και παρέχει ένα σαφές σύνολο στόχων και κριτηρίων επιτυχίας.

Κατά την εξέταση μιας λύσης εξόρυξης/ η και μηχανικής μάθησης, η αποτελεσματική διαμόρφωση του προβλήματος μπορεί να καθορίσει αν το έργο σας θα επιτύχει τελικά ή όχι.

Η κατανόηση του προβλήματος

Για να κατανοήσετε το πρόβλημα, εκτελέστε τις ακόλουθες εργασίες:

- Δηλώστε τον στόχο του προϊόντος που αναπτύσσετε ή αναδιαμορφώνετε.
- Καθορίστε αν ο στόχος επιλύεται καλύτερα με τη χρήση εξόρυξης δεδομένων.
- Βεβαιωθείτε ότι διαθέτετε τα δεδομένα που απαιτούνται για την εκπαίδευση ενός μοντέλου.

Ο στόχος είναι η απάντηση στην ερώτηση "**Τι προσπαθώ να πετύχω;**".

Παραδείγματα

Application	Goal
Weather app	Υπολογίστε τη βροχόπτωση σε εξάωρα διαστήματα για μια γεωγραφική περιοχή.
Video app	??
Mail app	??
Map app	??
Banking app	??
Dining app	??

Παραδείγματα

Application	Goal
Weather app	Υπολογίστε τη βροχόπτωση σε εξάωρα διαστήματα για μια γεωγραφική περιοχή.
Video app	Προτείνει χρήσιμα βίντεο.
Mail app	Ανίχνευση ανεπιθύμητης αλληλογραφίας.
Map app	Υπολογισμός του χρόνο ταξιδιού.
Banking app	Εντοπισμός δόλιων συναλλαγών.
Dining app	Προσδιορίστε την κουζίνα από το μενού ενός εστιατορίου.

2 DATA ACQUISITION

Data is the driving force ...

Τι προδιαγραφές πρέπει να έχουν τα δεδομένα;

Άφθονα!! Όσο περισσότερα σχετικά και χρήσιμα παραδείγματα υπάρχουν στο σύνολο των δεδομένων σας, τόσο καλύτερο θα είναι το μοντέλο σας.

Συνεπή. Η ύπαρξη δεδομένων που συλλέγονται με συνέπεια και αξιοπιστία θα παράγει ένα καλύτερο μοντέλο. Για παράδειγμα, ένα μοντέλο καιρού με βάση το ML θα επωφεληθεί από δεδομένα που συλλέγονται επί πολλά χρόνια από τα ίδια αξιόπιστα όργανα.

Αξιόπιστη πηγή. Κατανοήστε από πού θα προέρχονται τα δεδομένα σας. Τα δεδομένα θα προέρχονται από αξιόπιστες πηγές που εσείς ελέγχετε, όπως τα αρχεία καταγραφής από το προϊόν σας, ή θα προέρχονται από πηγές στις οποίες δεν έχετε μεγάλη εικόνα, όπως η έξοδος από ένα σύστημα μηχανικής μάθησης;

Τι προδιαγραφές πρέπει να έχουν τα δεδομένα;

Ορθότητα. Σε μεγάλα σύνολα δεδομένων, είναι αναπόφευκτο ότι ορισμένες ετικέτες θα έχουν εσφαλμένες τιμές, αλλά αν περισσότερο από ένα μικρό ποσοστό των ετικετών είναι εσφαλμένες, το μοντέλο θα παράγει κακές προβλέψεις.

Αντιπροσωπευτικά. Τα σύνολα δεδομένων θα πρέπει να είναι όσο το δυνατόν πιο αντιπροσωπευτικά του πραγματικού κόσμου. Με άλλα λόγια, τα σύνολα δεδομένων θα πρέπει να αντικατοπτρίζουν με ακρίβεια τα γεγονότα, τις συμπεριφορές των χρηστών ή/και τα φαινόμενα του πραγματικού κόσμου που μοντελοποιείται. Η εκπαίδευση σε μη αντιπροσωπευτικά σύνολα δεδομένων μπορεί να προκαλέσει κακή απόδοση όταν το μοντέλο καλείται να κάνει προβλέψεις στον πραγματικό κόσμο.

What counts as "a lot" of data?

Data set	Size (number of examples)
Iris flower data set	150 (total set)
MovieLens (the 20M data set)	20,000,263 (total set)
Google Gmail SmartReply	238,000,000 (training set)
Google Books Ngram	468,000,000,000 (total set)
Google Translate	trillions

2 DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- API'S
- ONLINE REPOSITORIES

Web Scraping

Η μέθοδος που ονομάζεται web scraping είναι κάπως η μέση λύση μεταξύ της συλλογής των δικών σας δεδομένων και της χρήσης των δεδομένων κάποιου άλλου.

Μπορείτε να αποκτήσετε πρόσβαση στα δεδομένα άλλων ανθρώπων πηγαίνοντας στους ιστοτόπους τους και επιλέγοντας ακριβώς ποια μέρη θέλετε να συλλέξετε.

Σαν πρώτη ανάγνωση, αυτό μοιάζει καλό, αλλά το web scraping έχει τις δικές του επιφυλάξεις.

Αρχικά, η εξαγωγή δεδομένων από ιστότοπους μέσω της scraping μπορεί να είναι δύσκολη. Τα εργαλεία που χρησιμοποιούνται, όπως το Selenium, απαιτούν καλή γνώση της HTML και της γλώσσας XML Path Language (XPath). Επιπλέον, τα σενάρια που απαιτούνται για την πλοήγηση σε ιστότοπους και την απόκτηση των απαιτούμενων πληροφοριών μπορεί να είναι μακροσκελή και να απαιτούν πολύ χρόνο για τη συγγραφή τους.

Επιπλέον, η απόσπαση ιστοσελίδων μπορεί κατά καιρούς να είναι ανήθικη ή ακόμη και παράνομη. Ενώ ορισμένοι ιστότοποι δεν έχουν ενδοιασμούς, άλλοι μπορεί να είναι λιγότερο ανεκτικοί. Δεν είναι ασυνήθιστο για τους ιστότοπους να ανεβάζουν δεδομένα που προστατεύονται από πνευματικά δικαιώματα ή να θέτουν όρους που καθορίζουν τις προϋποθέσεις για το scraping.

2 DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- API'S
- ONLINE REPOSITORIES

Πηγές/ Αποθετήρια δεδομένων

Παραδείγματα από “Ανοιχτά δεδομένα” (open data sources)

https://data.worldbank.org/
the world’s most comprehensive data regarding what’s happening in different countries across the world

https://www.who.int/gho/database/en/
WHO keeps track of health-specific statistics of its 194 Member States.

https://www.google.com/publicdata/directory
This directory can help you explore vast amounts of public-interest datasets.

http://open-data.europa.eu/en/data/
Access whatever open data EU institutions, agencies and other organizations publish on a single platform

τα ανοικτά δεδομένα μπορούν να δώσουν μεγάλη ώθηση για τη μηχανική μάθηση με σκοπό να βοηθήσουν στην καταπολέμηση παγκόσμιων προβλημάτων όπως οι ασθένειες, το έγκλημα ή η πείνα. Μπορούν να συμβάλουν στον μετασχηματισμό του τρόπου με τον οποίο κατανοούμε και συνεργαζόμαστε με τον κόσμο.

Πηγές/ Αποθετήρια δεδομένων

These datasets remove barriers and provide access to critical information quickly, safely, and easily, eliminating the need to search for and onboard large data files.

Παραδείγματα από “Ανοιχτά δεδομένα” (open data sources)

https://data.fivethirtyeight.com/
a great site for data-driven journalism and story-telling

http://www.census.gov/data.html
the biggest statistical agency of the federal government. It stores and provides reliable facts and data regarding people, places, and economy of America.

https://www.data.gov/
It was only recently that the decision was made to make all government data available for free.

https://wiki.dbpedia.org/
DBpedia aims at getting structured content from the valuable information that Wikipedia created.

Πηγές/ Αποθετήρια δεδομένων

Παραδείγματα από “Ανοιχτά δεδομένα” (open data sources)

https://github.com/freeCodeCamp/open-data
You can find datasets, analysis of the same and even demos of projects based on the freeCodeCamp data.

https://www.yelp.com/dataset
a subset of nothing but our own businesses, reviews and user data for use in personal, educational and academic pursuits..

https://data.unicef.org/
it has compiled relevant data on education, child labor, child disability, child mortality, maternal mortality, water and sanitation, etc.

https://www.kaggle.com/datasets
The platform supports open and accessible data formats.it is not just a data repository. Each dataset stands for a community that enables you to discuss data, find out public codes and techniques, and conceptualize your own projects in Kernels.

Ελληνικά Αποθετήρια δεδομένων

Υποδομή για “ανοιχτά δεδομένα” (open data sources)

https://inventory.clarin.gr/
είναι η εθνική Υποδομή Γλωσσικών πόρων και Τεχνολογιών στην Ελλάδα. Στόχοι του ειναι η συλλογή, η τεκμηρίωση, η συντήρηση και ο διαμοιρασμός ψηφιακών γλωσσικών πόρων, εργαλείων γλωσσικής τεχνολογίας καθώς και πιστοποιημένων διαδικτυακών υπηρεσιών γλωσσικής επεξεργασίας

https://data.gov.gr/
δεδομένα δημοσιευμένα από την κεντρική διοίκηση, οργανισμούς τοπικής αυτοδιοίκησης και άλλες υπηρεσίες

https://opendata.bankofgreece.gr/el/home

https://catalog.hcapdata.gr/dataset/

Το O2hub είναι η ψηφιακή πλατφόρμα διάθεσης ανοικτών δεδομένων (Open Data) και ανοικτών προγραμματιστικών διεπαφών (Open APIs) από τις θυγατρικές εταιρείες του Υπερταμείου (ΕΕΣΥΠ).

2 DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- API'S
- ONLINE REPOSITORIES

APIs

Όταν ένας πελάτης επιθυμεί πρόσβαση σε ορισμένα δεδομένα από έναν ξένο διακομιστή, υποβάλλει αίτηση στον εν λόγω διακομιστή.

Όταν ο διακομιστής λάβει την αίτηση, παράγει μια απάντηση που στέλνει πίσω στον πελάτη.

Ένα API παίζει το ρόλο του μεσάζοντα σε αυτή την ανταλλαγή.

Τα API των διαφόρων πηγών διαφέρουν ως προς την προσβασιμότητα και την εφαρμογή.

Οι περισσότεροι ιστότοποι έχουν τις δικές τους μοναδικές απαιτήσεις για τα αιτήματα και ορίζουν μια μοναδική μορφή για τις απαντήσεις τους. Διαφέρουν επίσης όσον αφορά τους περιορισμούς τους. Για παράδειγμα, ορισμένα API περιορίζουν τον αριθμό των αιτημάτων που μπορείτε να κάνετε σε μια ημέρα.

Μπορείτε να αποκτήσετε API_KEY υποβάλλοντας αίτημα στην αντίστοιχη ενότητα για προγραμματιστές στους New York Times.

APIs (παράδειγμα)

```
# obtain the URI for access to articles for a given query, page number, sorting order
def get_URI(query:str, page_num:int, sort_order:str, API_KEY:str) -> str:

   # create a URI string
   URI = 'https://api.nytimes.com/svc/search/v2/articlesearch.json?q='+query

   # if a page number is mentioned, add it to the URI
   if page_num:
       add_to_URI = '&page='+str(page_num)
       URI+= add_to_URI

   # if the sorting order is mentioned, add it to the URI
   if sort_order:
       add_to_URI = '&sort='+sort_order
       URI += add_to_URI

   # add the given API key to the URI
   add_to_URI = '&api-key='+API_KEY
   URI += add_to_URI

   # return the new URI
   return URI
```

Μπορείτε να αποκτήσετε API_KEY υποβάλλοντας αίτημα στην αντίστοιχη ενότητα για προγραμματιστές στους New York Times.

APIs (παράδειγμα)

Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να λάβουμε τα τελευταία 1000 άρθρα που σχετίζονται με τους Χειμερινούς Ολυμπιακούς Αγώνες. Μπορούμε να λάβουμε το URI για το αίτημα χρησιμοποιώντας τη συνάρτηση.

```
# get the URI needed for the articles related to Winter Olympics in page 1 from newest to oldest
URI = get_URI(query='Winter Olympics', page_num=1, sort_order='newest', API_KEY=API_KEY)
```

Η δημιουργία αιτήσεων στην Python είναι μια απλή εργασία με το πακέτο "requests". Μόλις κάνετε μια αίτηση και λάβετε μια απάντηση από τον διακομιστή, πρέπει να αναλύσετε την απάντηση, η οποία είναι σε μορφή JSON (για την περίπτωση αυτή αλλά είναι και το πιο συνηθισμένο).

```
import requests

# make a request to the server
response = requests.get(URI)

# parse the response (only works for JSON format)
data = response.json()

# print response
print(data)
```

{'status': 'OK', 'copyright': 'Copyright (c) 2022 The New York Times Company. All Rights Reserved.', 'response': {'docs': [{'abstract': 'The mountains are too crowded. The sport is too expensive. Several resorts are trying to fix a number of problems. How are they doing?', 'web_url': 'https://www.nytimes.com/2022/02/05/style/vail-ski-resorts-crowds.html', 'snippet': 'The mountains are too crowded. The sport is too expensive. Several resorts are trying to fix a number of problems. How are they doing?', 'lead_paragraph': 'When Tim Pham learned to ski in the 1980s, the sport seemed simpler. He would go to quiet resorts in Northern California like Sugar Bowl, where he would show up any time of day, buy a $35 lift pass, and ski without facing lines or crowds.', 'print_section': 'ST', 'print_page': '1', 'source': 'The New York Times', 'multimedia': [{'rank': 0, 'subtype': 'xlarge', 'caption': None, 'credit': None, 'type': 'image', 'url': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-articleLarge.jpg', 'height': 750, 'width': 600, 'legacy': {'xlarge': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-articleLarge.jpg', 'xlargewidth': 600, 'xlargeheight': 750}, 'subType': 'xlarge', 'crop_name': 'articleLarge'}, {'rank': 0, 'subtype': 'popup', 'caption': None, 'credit': None, 'type': 'image', 'url': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-popup.jpg', 'height': 500, 'width': 400, 'legacy': {}, 'subType': 'popup', 'crop_name': 'popup'}, {'rank': 0, 'subtype': 'blog480', 'caption': None, 'credit': None, 'type': 'image', 'url': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-blog480.jpg', 'height': 600, 'width': 480, 'legacy': {}, 'subType': 'blog480', 'crop_name': 'blog480'}, {'rank': 0, 'subtype': 'blog533', 'caption': None, 'credit': None, 'type': 'image', 'url': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-blog533.jpg', 'height': 666, 'width': 533, 'legacy': {}, 'subType': 'blog533', 'crop_name': 'blog533'}, {'rank': 0, 'subtype': 'blog427', 'caption': None, 'credit': None, 'type': 'image', 'url': 'images/2022/01/31/fashion/SKIING-3/merlin_200529699_6287fd74-46b1-4d56-8877-e432d4941797-blog427.jpg', 'height': 534, 'width': 4

APIs (παράδειγμα)

Ευτυχώς, η συνάρτηση json_normalize του Pandas μας επιτρέπει να χειριζόμαστε με ευκολία εμφωλευμένα αντικείμενα JSON.

```
from pandas.io.json import json_normalize

# convert JSON response to a data frame
df = json_normalize(data['response'], record_path=['docs'])
```

Ακολουθούν οι τίτλοι και οι ημερομηνίες δημοσίευσης των 5 πρώτων άρθρων:

```
# show the headline and publication date
df[['headline.main', 'pub_date']].head(5)
```

	headline.main	pub_date
0	Who Gets to Ski?	2022-02-05T10:00:10+0000
1	Long-Track Speedskating: How It Works and Who'...	2022-02-05T09:54:24+0000
2	Race, Crash, Surgery, Rehab. Then Repeat.	2022-02-05T08:00:11+0000
3	U.S. broadcast coverage continues with freesty...	2022-02-05T05:05:35+0000
4	Planning for the 2026 Winter Games is already ...	2022-02-05T04:51:49+0000

APIs (παράδειγμα)

Όπως μπορείτε να παρατηρήσετε, τα δεδομένα που πήραμε σε αυτό το request περιέχουν μόνο 10 εγγραφές, οι οποίες υπολείπονται των 1000 άρθρων που επιθυμούμε να συλλέξουμε. Τι συμβαίνει λοιπόν;

Το API των New York Times διατηρεί ένα **όριο** 10 άρθρων για κάθε αίτηση.

Όλα τα APIs έχουν περιορισμούς

Για τις περισσότερες περιπτώσεις, η συλλογή των δικών σας δεδομένων είναι μια παράλογη ιδέα.

Η προμήθεια πληροφοριών της απαιτούμενης ποσότητας και ποιότητας απαιτεί σημαντικό χρόνο, χρήμα, ανθρώπινο δυναμικό και πόρους.

APIs (social media)

Τα δεδομένα των μέσων κοινωνικής δικτύωσης είναι κάθε πληροφορία που συλλέγεται από τις πλατφόρμες μέσων κοινωνικής δικτύωσης και παρέχει πληροφορίες σχετικά με τις δραστηριότητες των χρηστών στην πλατφόρμα.

Ορισμένα από τα δεδομένα περιλαμβάνουν μετρήσεις όπως ο αριθμός των likes, η αύξηση των followers, ο αριθμός των shares, η διάρκεια engagement και άλλα.

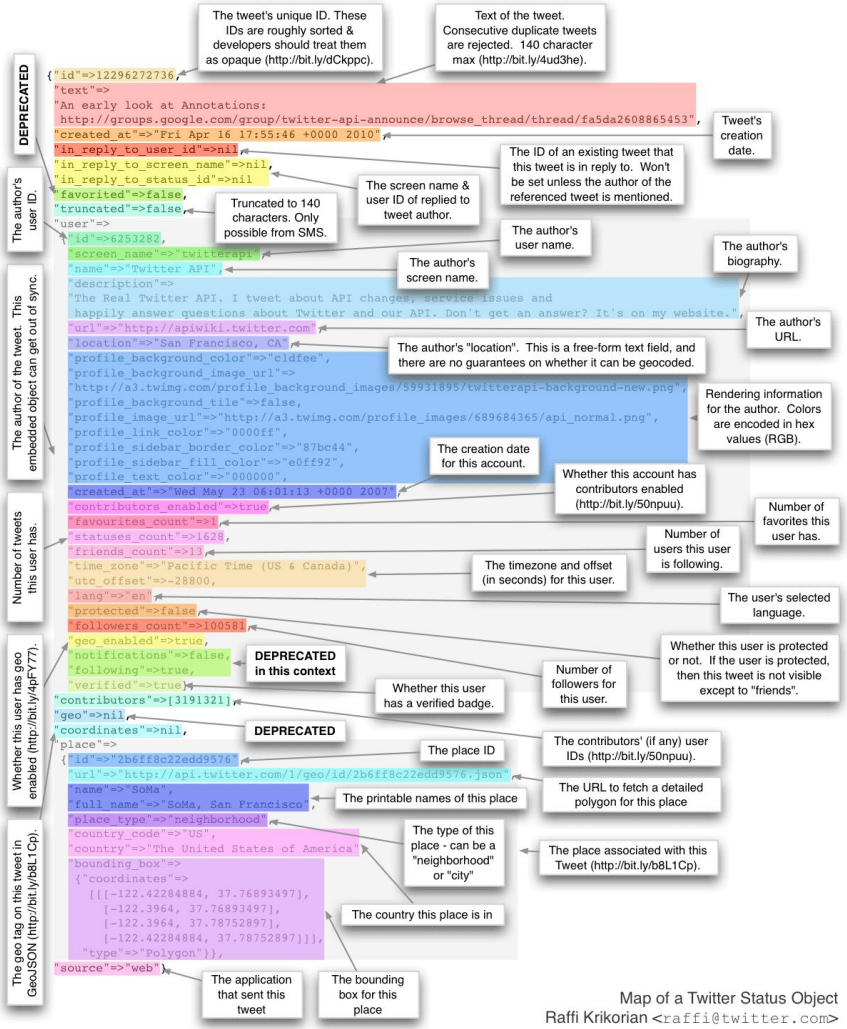
Αξίζει επίσης να αναφερθεί η εξόρυξη δεδομένων από τα μέσα κοινωνικής δικτύωσης είναι ένας πολύ ενεργός χώρος με πολλές εφαρμογές και προεκτάσεις.

Η συλλογή προσωπικών δεδομένων είναι ένα ευαίσθητο θέμα μεταξύ των διαδικτυακών κοινοτήτων και της νομοθετικής κοινότητας στο σύνολό της. Αυτό οφείλεται στο γεγονός ότι, ενώ η συλλογή δεδομένων σχετικά με τις δραστηριότητες των χρηστών στα διάφορα κοινωνικά δίκτυα θεωρείται κάπως δεοντολογική, η συλλογή προσωπικών δεδομένων δεν είναι.

Η συλλογή δεδομένων από τα μέσα κοινωνικής δικτύωσης δεν είναι τόσο απλή όσο ακούγεται, για πολλούς λόγους. Ο τύπος των δεδομένων που συλλέγονται εξαρτάται από την πλατφόρμα και τη σημασία των δεδομένων για έναν οργανισμό.

Twitter's JSON response

The streaming API returns tweets, as well as several other types of messages (e.g. a tweet deletion notice, user update profile notice, etc), all in JSON format.



Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Άσκηση

- Επιλέξτε ένα κοινωνικό δίκτυο της αρεσκείας σας
- Μελετήστε αν παρέχει API
- Τι δεδομένα και τι μεταδεδομένα επιστρέφει το API (αν υπάρχει)
- Τι περιορισμούς έχει ;
- Πόσο κοστίζει ;

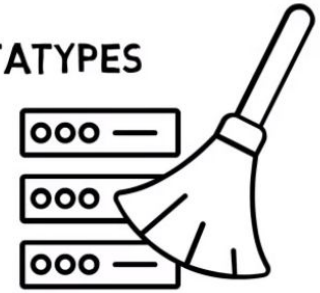
3 DATA PREPARATION

DATA CLEANING

INCONSISTENT DATATYPES

MISSPELLED ATTRIBUTES

MISSING AND DUPLICATE VALUES



TRANSFORMATION

(4) EXPLORATORY DATA ANALYSIS

DEFINES AND REFINES

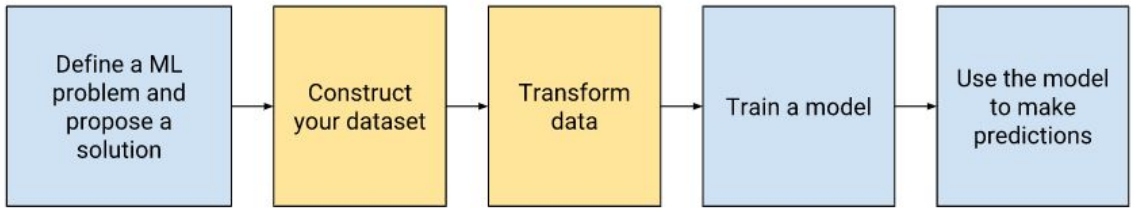
THE SELECTION OF FEATURE

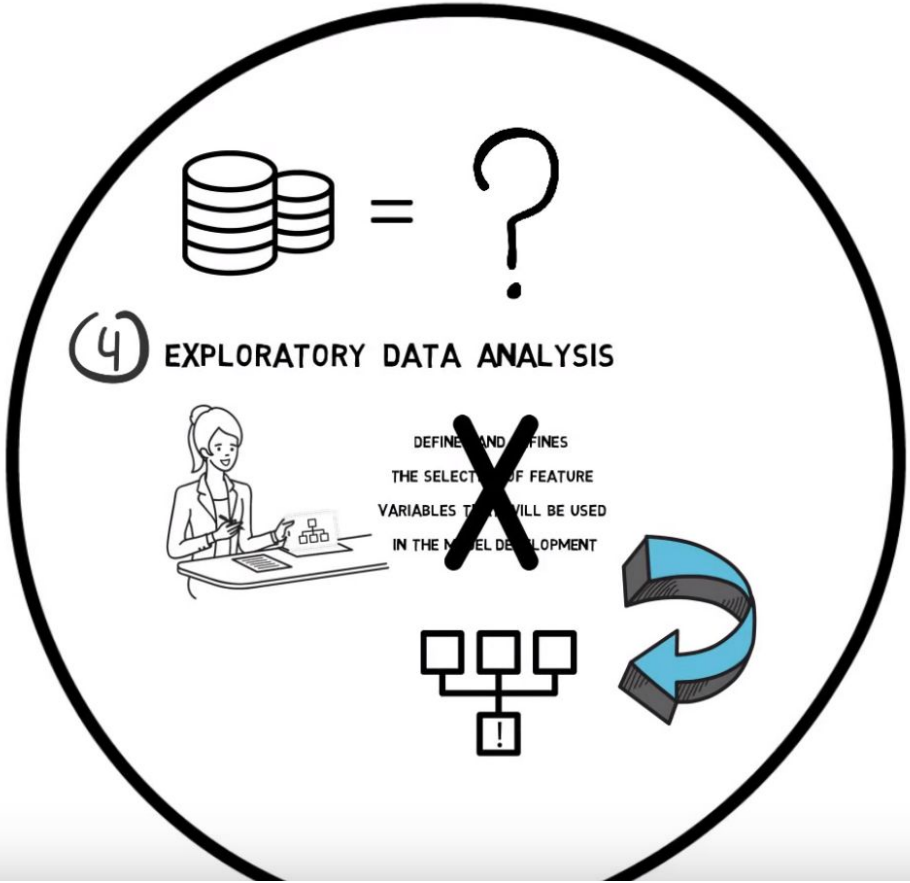
VARIABLES THAT WILL BE USED

IN THE MODEL DEVELOPMENT

EDA different from ML

Ανάλυση δεδομένων - Πρόβλεψη βασισμένη σε δεδομένα ... είναι δύο εντελώς διαφορετικά πράγματα



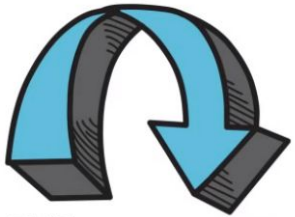


THE MOST

IMPORTANT STEP

IDENTIFY THE MODEL THAT BEST FITS THE BUSINESS REQUIREMENT

TRAINS THE MODELS ON THE TRAINING DATASET AND TEST



SELECT THE BEST PERFORMING MODEL

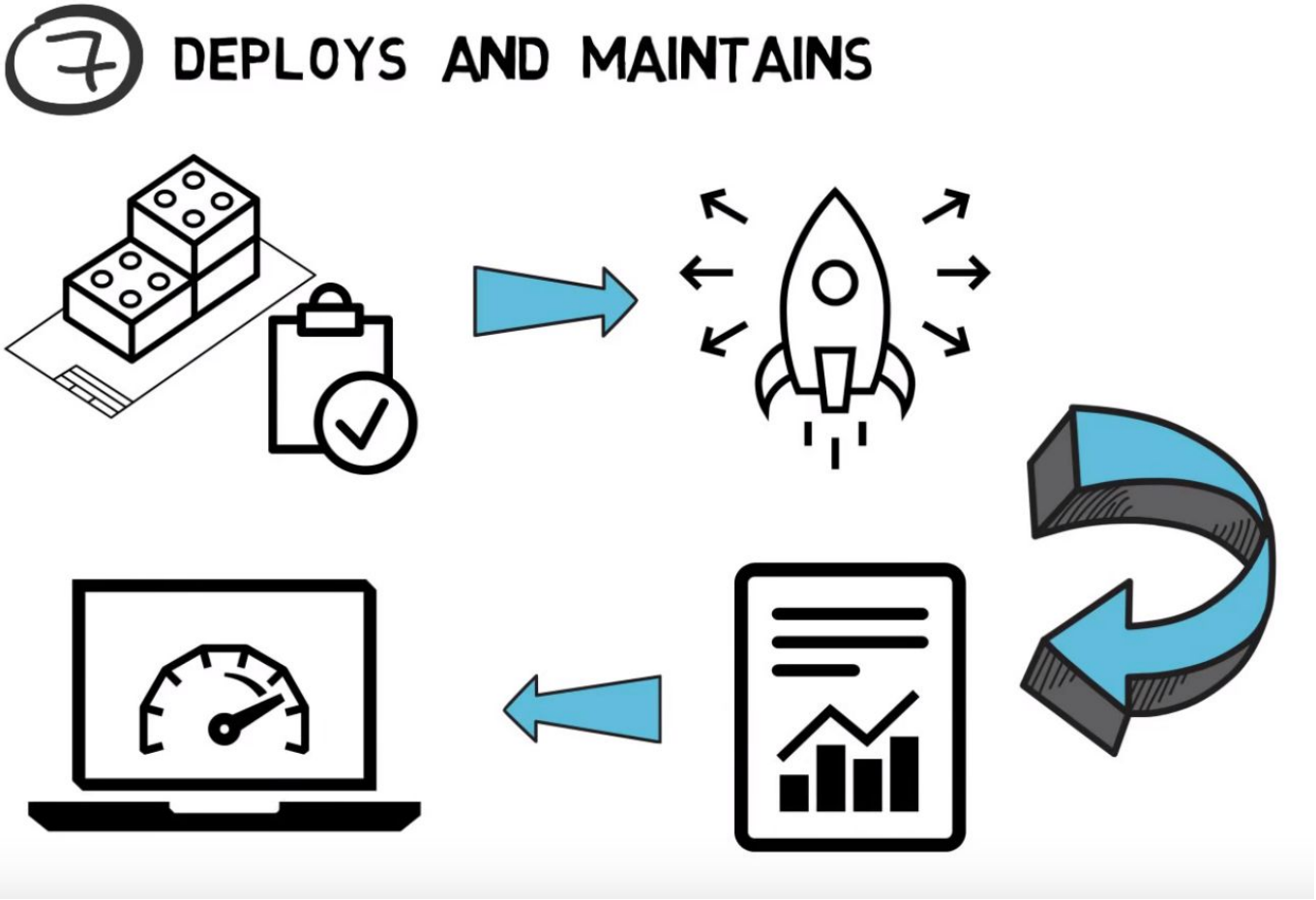
python™

R

sas

6 VISUALIZATION AND COMMUNICATION

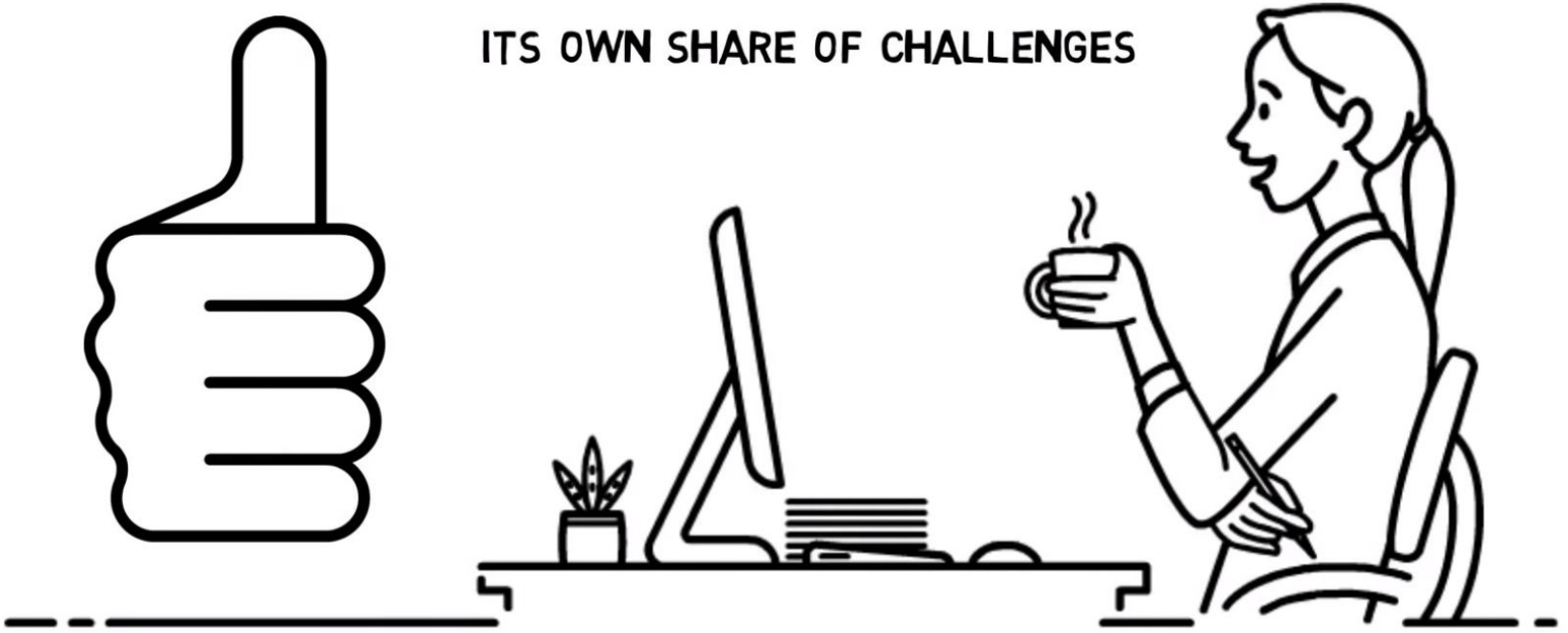
tableau Power BI QlikView


7
DEPLOYS AND MAINTAINS

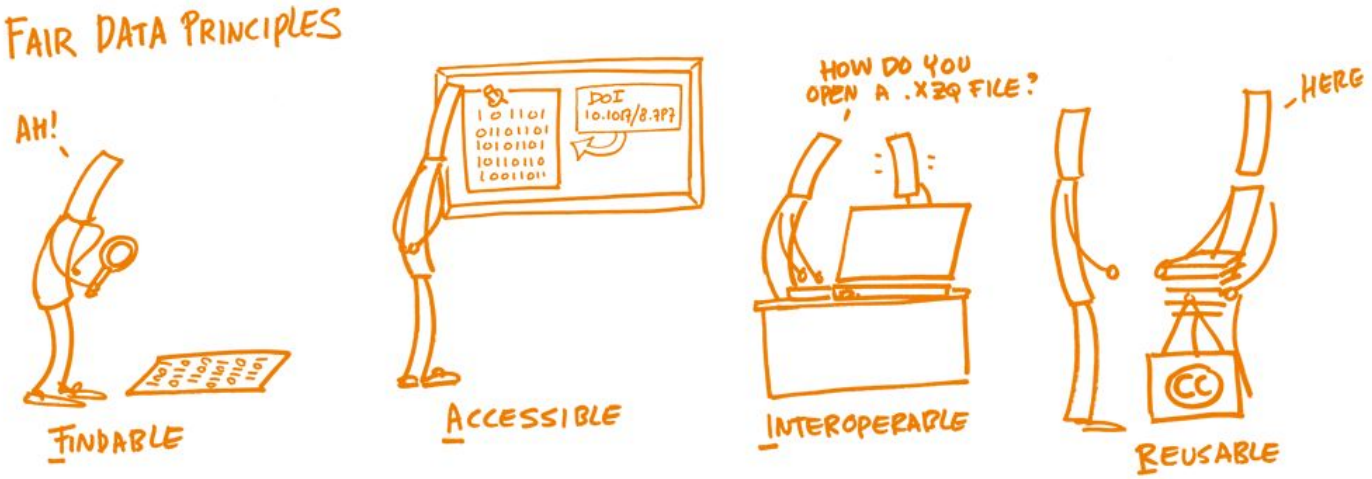
THE DAILY ROUTINE OF A DATA SCIENTIST IS A WHOLE LOT OF FUN,

HAS A LOT OF INTERESTING ASPECTS AND COMES WITH

ITS OWN SHARE OF CHALLENGES



FAIRness



https://www.fosteropenscience.eu/learning/assessing-the-fairness-of-data/#/id/5c52e8cf0d3def29462d8cb5

Fairness (2)

Fairness in data analysis is to use data in a way that doesn't create or reinforced bias.

Δικαιοσύνη στην ανάλυση δεδομένων είναι η χρήση των δεδομένων με τρόπο που δεν δημιουργεί ή ενισχύει την προκατάληψη.

Fairness - types of bias

automation bias *is a tendency to favor results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each.*

confirmation bias *occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution*

coverage bias *Data is not selected in a representative fashion*

experimenter's bias *a model builder may actually keep training a model until it produces a result that aligns with their original hypothesis*

group attribution bias *is a tendency to generalize what is true of individuals to an entire group to which they belong*

implicit bias *occurs when assumptions are made based on one's own mental models and personal experiences that do not necessarily apply more generally*

non-response bias *Users from certain groups opt-out of surveys at different rates than users from other groups*.

reporting bias *The fact that the frequency with which people write about actions, outcomes, or properties is not a reflection of their real-world frequencies or the degree to which a property is characteristic of a class of individuals.*

selection bias *Errors in conclusions drawn from sampled data due to a selection process that generates systematic differences between samples observed in the data and those not observed.*

https://pair.withgoogle.com/explorables/measuring-fairness/

How data is changing the world

Language

Οι πολιτισμοί σήμερα επηρεάζουν ο ένας τον άλλον μέσω των μέσων κοινωνικής δικτύωσης, των επιχειρήσεων και των ταξιδιών. Παρακολουθούμε ταινίες που γυρίστηκαν στο εξωτερικό και μεταφράζονται στα αγγλικά μέσω υποτίτλων. Μεταφράζουμε μηνύματα ηλεκτρονικού ταχυδρομείου σε επιχειρηματικούς εταίρους στις χώρες όπου ελπίζουμε να επεκταθούμε. Γνωρίζουμε άλλους ανθρώπους, μορφωνόμαστε και εξερευνούμε σε έναν κόσμο όπου μιλιούνται χιλιάδες γλώσσες και διάλεκτοι.

Τα Μεγάλα Γλωσσικά Μοντέλα, όπως το GPT-3, εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων κειμένου από το διαδίκτυο και είναι ικανά να παράγουν κείμενο που μοιάζει με ανθρώπινο κείμενο, αλλά μπορεί να μην παράγουν πάντα αποτελέσματα που να συνάδουν με τις ανθρώπινες προσδοκίες ή τις επιθυμητές τιμές.

Το ChatGPT βασίζεται στο αρχικό μοντέλο GPT-3, αλλά έχει εκπαιδευτεί περαιτέρω με τη χρήση ανθρώπινης ανατροφοδότησης για την καθοδήγηση της διαδικασίας εκμάθησης, με συγκεκριμένο στόχο την άμβλυνση των προβλημάτων κακής ευθυγράμμισης του αρχικού μοντέλου. Η συγκεκριμένη τεχνική που χρησιμοποιήθηκε, η οποία ονομάζεται *Reinforcement Learning from Human Feedback*.

Food

Σύμφωνα με εκθέσεις του Οργανισμού Τροφίμων και Γεωργίας, απέχουμε πολύ από τους στόχους μας για την εξάλειψη της παγκόσμιας πείνας έως το 2030.

Για την επίλυση αυτού του προβλήματος, στρεφόμαστε σε μια σύνδεση μεταξύ των μεγάλων δεδομένων και της γεωργίας. Η γεωργία ακριβείας, όπως αποκαλείται, θα λαμβάνει υπόψη τις καιρικές συνθήκες, την ποιότητα του αέρα, το νερό, το άζωτο και άλλες συνθήκες που σχετίζονται με κάθε επιμέρους τετραγωνικό καλλιεργήσιμης γης.

Σήμερα ο γεωργικός εξοπλισμός αποτελείται από μεγαμηχανές με διεπαφές tablet και ένα λογισμικό που κατευθύνει τους ελκυστήρες προς τα πού να πάνε. Τα μηχανήματα δεν οδηγούνται από τους αγρότες - οδηγούνται από πληροφορίες από το έδαφος και τον ουρανό.

Ήδη οι μεγαλύτερες γεωργικές εταιρείες του κόσμου επενδύουν δισεκατομμύρια δολάρια σε αναλύσεις γεωργικών εκμεταλλεύσεων, ελπίζοντας να αυξήσουν την παραγωγή καλλιεργειών κατά ένα σημαντικό ποσοστό.

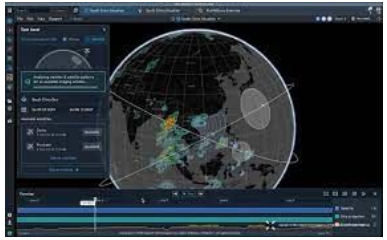
Finance

Η συντριπτική πλειονότητα των μετοχών που διαπραγματεύονται καθημερινά γίνεται μέσω της μεθόδου black-box. Αυτό περιλαμβάνει τη χρήση αλγορίθμων υπολογιστών για τη διαπραγμάτευση μετοχών με τη χρήση πληροφοριών σχετικά με τις τιμές των μετοχών, την ποσότητα και το χρονοδιάγραμμα. Στον κόσμο της λιανικής τραπεζικής, εν τω μεταξύ, τα μεγάλα δεδομένα αποσκοπούν στην παραγωγή πιο αποτελεσματικών υπηρεσιών διανομής, προμήθειας και χρηματοοικονομικών υπηρεσιών για τις εταιρείες και τους πελάτες τους. Αυτός ο συνδυασμός μεταξύ χρηματοοικονομικών και τεχνολογίας ονομάζεται κατάλληλα "fintech".

Τα νέα συστήματα fintech θα αντικαταστήσουν εκείνα που είναι σήμερα υπεύθυνα για πράγματα όπως η επεξεργασία δανείων και η παρακολούθηση λογαριασμών. Και φαίνεται ότι τα ιδρύματα ενδιαφέρονται ιδιαίτερα να κάνουν τον μετασχηματισμό. Το 2008, πραγματοποιήθηκαν 930 εκατομμύρια δολάρια σε παγκόσμιες επενδύσεις σε fintech.

Οι νεοσύστατες επιχειρήσεις Fintech ισχυρίζονται ότι είναι σε θέση να κρίνουν καλύτερα εάν ένας πελάτης είναι επιλέξιμος για δάνειο, συγκεντρώνοντας δεδομένα σχετικά με τις συναλλαγές του πελάτη τους. Αυτό περιλαμβάνει συναλλαγές που πραγματοποιούνται μέσω κάρτας, επιταγής, μετρητών, καθώς και σχετικά δεδομένα όπως η τοποθεσία της επιχείρησής τους και οι οικονομικές τους τάσεις. Αυτά μπορούν να υπαγορεύσουν εάν ένας πελάτης μπορεί να είναι σε θέση να αποπληρώσει το δάνειό του.

War



Ίσως το μεγαλύτερο πρόβλημα με τη χρήση μεγάλων δεδομένων στον τομέα του πολέμου είναι ότι η τεχνολογία δεν παραμένει ποτέ αποκλειστική για πολύ. Όπως η “συνταγή” για μια πυρηνική βόμβα, τα έξυπνα εργαλεία που δημιουργούμε για να μας δώσουν πλεονέκτημα έναντι των εχθρών μας μπορούν να αντιγραφούν και να αναδημιουργηθούν για να λειτουργήσουν εναντίον μας.

Τα μεγάλα δεδομένα χρησιμοποιούνται για να σκοτώσουν μεγαλύτερο αριθμό εχθρών και να κρατήσουν περισσότερους στρατιώτες ασφαλείς, εντοπίζοντας ποιες τοποθεσίες είναι πιθανότερο να αποτελέσουν σημείο ενέδρας του εχθρού. Ο στόχος είναι το στρατιωτικό προσωπικό να βλέπει τα δεδομένα και να μπορεί να τα κατανοήσει γρήγορα, ώστε να μπορεί να καταλήξει σε ένα αποτελεσματικό σχέδιο δράσης

Μια από τις εταιρείες που ειδικεύεται σε αυτό ονομάζεται Palantir. Αν και η ίδια η εταιρεία δεν προορίζεται αποκλειστικά για στρατιωτική χρήση, ένα μεγάλο ποσοστό των εσόδων της προέρχεται από τον τομέα της άμυνας. Ένα ιστορικό των πελατών της περιλαμβάνει το FBI, την NSA και τον αμερικανικό στρατό. Η ιστοσελίδα τους υπόσχεται ότι το λογισμικό τους μπορεί να βοηθήσει τους πελάτες τους να αποκτήσουν τακτικό πλεονέκτημα, "μετατρέποντας γρήγορα μεγάλα δεδομένων σε σχέδια δράσης".

Medicine

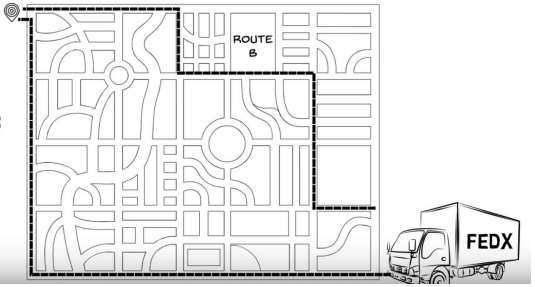
Η γονιδιωματική, που σήμερα θεωρείται τομέας των μεγάλων δεδομένων λόγω του εκπληκτικού όγκου των πληροφοριών που παράγονται, ασχολείται με τη χαρτογράφηση και τη μελέτη του πολύπλοκου κόσμου των ανθρώπινων γονιδιωμάτων.

Εταιρείες όπως η Personal Genome Diagnostics (PGDx) προσφέρουν εξατομικευμένη και εξειδικευμένη φροντίδα για άτομα που έχουν διαγνωστεί με καρκίνο. Ένα δείγμα από το σάλιο και τον όγκο αποστέλλεται από το γραφείο του ογκολόγου, όπου μια ομάδα της PGDx καθαρίζει και προετοιμάζει τα δείγματα για να εισέλθουν σε ένα μηχάνημα αλληλούχισης. Μέσα σε αυτά τα μηχανήματα τα κύτταρα γίνονται δεδομένα, το γονιδίωμα μειώνεται σε gigabytes πληροφοριών που αποκαλύπτουν σημαντικές λεπτομέρειες σχετικά με την κακόβουλη ζωή του όγκου - πχ αν μεταλλάσσονται οι πρωτεΐνες και αν συνεχίζει να εξαπλώνεται. Από εκεί και πέρα, η εταιρεία μπορεί επίσης να προτείνει τα φάρμακα που είναι καταλληλότερα για να κρατήσουν τις μεταλλάξεις υπό έλεγχο. Πολλές φορές, ωστόσο, το τέλειο φάρμακο για έναν συγκεκριμένο ασθενή απλώς δεν υπάρχει. Η ανάπτυξη φαρμάκων κινείται με πολύ αργούς ρυθμούς για να συμβαδίσει με τη γονιδιωματική.

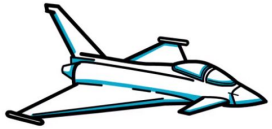
Transportation

LOGISTICS COMPANIES LIKE DHL, FEDEX HAVE DISCOVERED THE BEST TIME AND ROUTES TO SHIP

BEST TIME:
12:00PM



AIRLINE COMPANIES CAN NOW EASILY PREDICT FLIGHT DELAY AND NOTIFY THE PASSENGERS

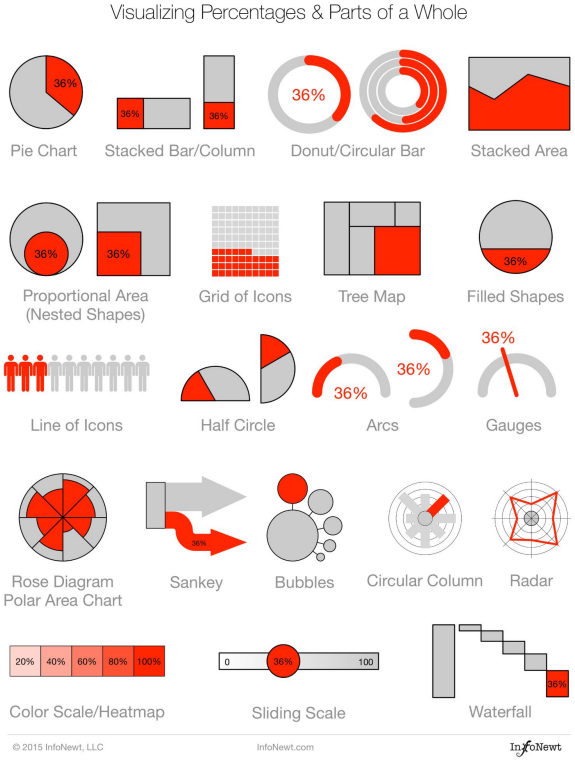


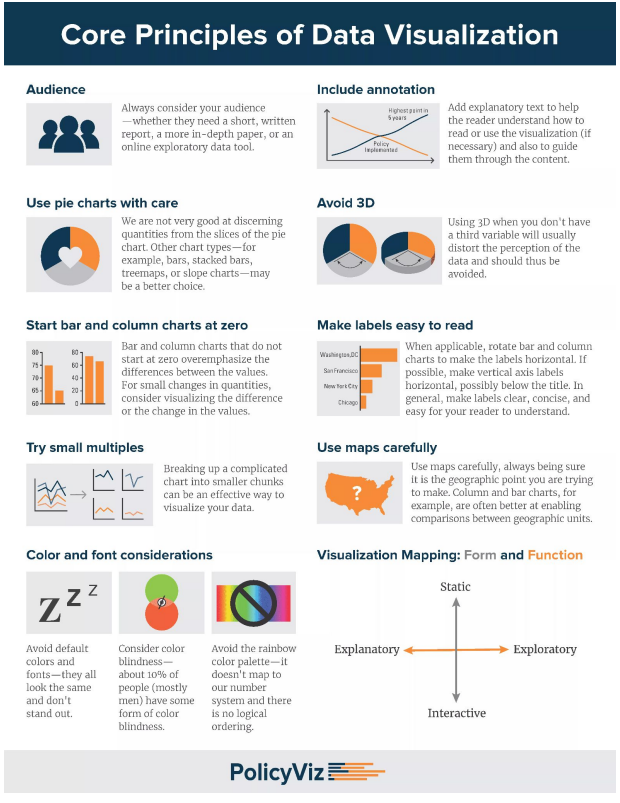
H487	CANCELLED
PP87	DELAYED

President Barack Obama's Big Data Keynote -- Strata + Hadoop World 2015

The Role of Big Data Visualization in an Era of Pandemics

Data visualization is equally important





Case Studies

Twitter Accurately Predicts Politician's Victory at New Hampshire Primary (2016)

The research firm (Globalpoint Research), predicts that this based on Tweets:

> *Romney would not only have the edge, second place would be a tight call between Ron Paul, Jon Huntsman and Rick Santorum.*

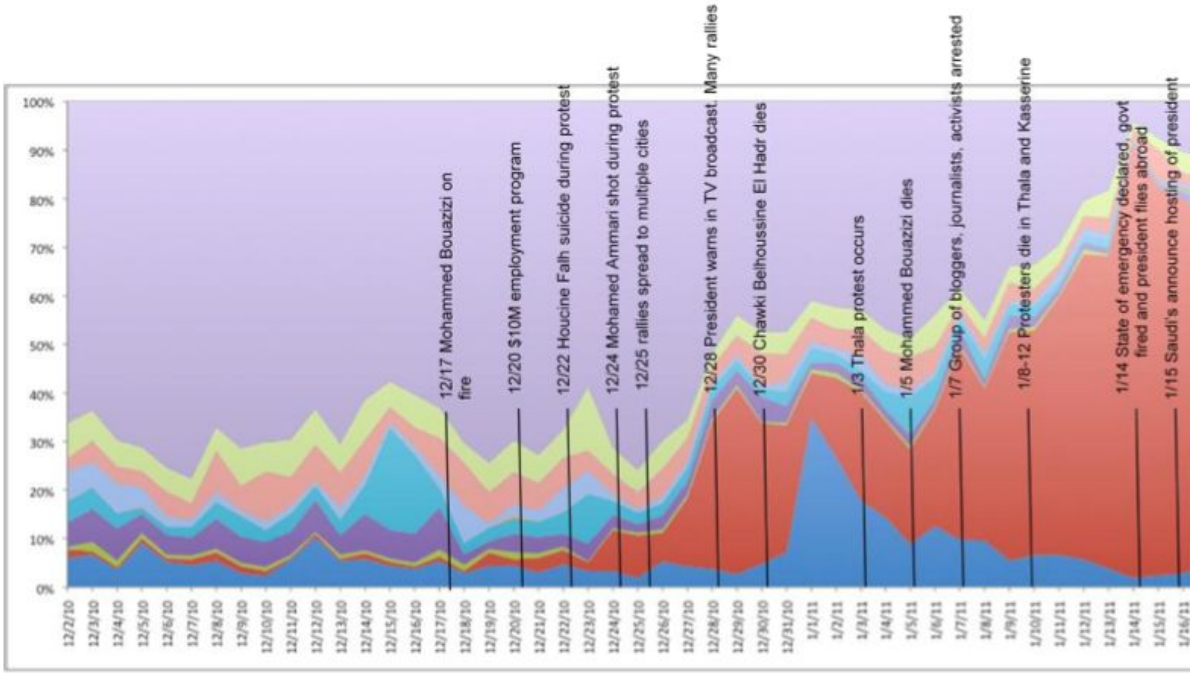
The actual results are as follows:

> *Romney came out on top with 37% of the vote, followed by Paul (23%), Huntsman (17%), Gingrich (10%), Santorum (10%) and Texas Gov. Rick Perry (1%).*

Did Twitter predict the revolution in Egypt?

Topsy Labs (*owned by Apple, now closed*) did a correlation analysis of Tweets based on the hashtags #yemen, #iran and #egypt.

What's most interesting is that the graph strongly correlates with actual events marked on the timeline:



Twitter is surprisingly accurate at predicting unemployment

The blue line tracks initial seasonally adjusted claims for unemployment insurance in official Department of Labor statistics. The red line is from a model, now updated weekly, that predicts unemployment claims based solely on the ebb and flow of Twitter missives like "I just lost my job. Who's buying my drinks tonight?"

Sources: Initial Claims for Unemployment Insurance (seasonally adjusted), U.S. Department of Labor; Prediction, University of Michigan Social Media Job Loss Index.

Twitter knows how you will be feeling this Friday

Ο Scott Golder (@redlog), από το Πανεπιστήμιο του Cornell, έριξε μια “ματιά” σε σχεδόν μισό δισεκατομμύριο Tweets. Αυτό που ήθελαν να διαπιστώσουν οι ερευνητές ήταν το εξής:

How do our moods and feelings change throughout the day, week and year?

Το ενδιαφέρον εδώ είναι ότι ανέλυσαν μηνύματα από 2,4 εκατομμύρια ανθρώπους σε 84 διαφορετικές χώρες. Αυτό έχει πολύ νόημα, καθώς θα μπορούσαν επίσης να συγκρίνουν διαφορετικές κουλτούρες, με διαφορετικές εβδομαδιαίες ρουτίνες και εποχές δίπλα-δίπλα.

Twitter knows how you will be feeling this Friday

Ακολουθούν μερικά από τα πιο ενδιαφέροντα ευρήματα, ορισμένα πιο προφανή από άλλα:

Καθημερινά: Η μεγαλύτερη “αρνητικότητα” κατά τη διάρκεια της εβδομάδας συμβαίνει την Παρασκευή και (ξαφνικά) εξαφανίζεται αργά το απόγευμα.

Εβδομαδιαία: Σε καθημερινή βάση, είμαστε πιο ευτυχισμένοι το πρωί και στη συνέχεια η διάθεσή μας παίρνει την κατιούσα. Αργά το απόγευμα ανακάμπτει με άλλη μια αιχμή θετικότητας.

Εποχές: Μέσα από τη σύγκριση διαφορετικών χωρών, οι ερευνητές θέλησαν να μάθουν αν το φως του ήλιου σε διαφορετικές εποχές επηρεάζει τη διάθεσή μας. Το συμπέρασμα ήταν το εξής: "Δεν είναι το πόσο φως της ημέρας λαμβάνετε, αλλά το σχετικό φως της ημέρας - αν οι μέρες γίνονται μεγαλύτερες ή μικρότερες - που κάνει τη διαφορά στη θετική διάθεση".

https://news.cornell.edu/stories/2011/09/study-tweets-reveals-mood-patterns-worldwide

Καλειδοσκόπιο Κοινωνικών δεδομένων

http://www.socioscope.gr/

Το Καλειδοσκόπιο Κοινωνικών Δεδομένων αναπτύχθηκε στα πλαίσια του έργου «Δυναμική Διαχείριση Βάσεων Κοινωνικών Δεδομένων και Χαρτογραφικών Αναπαραστάσεων» της πρόσκλησης «ΚΡΗΠΙΣ» της ΓΓΕΤ (2013-2015) μεταξύ του Εθνικού Κέντρου Κοινωνικών Ερευνών και του Ινστιτούτου Πληροφοριακών Συστημάτων του Ερευνητικού Κέντρου «ΑΘΗΝΑ». Αποτελεί πλατφόρμα οπτικής ανάλυσης και χαρτογραφικής αναπαράστασης κοινωνικών και πολιτικών δεδομένων με στόχο την υποστήριξη και ενίσχυση της κοινωνικής έρευνας και τη διάθεση ανοιχτών κοινωνικών δεδομένων στο ευρύ κοινό.

Netflix: Using Big Data to Drive Big Engagement

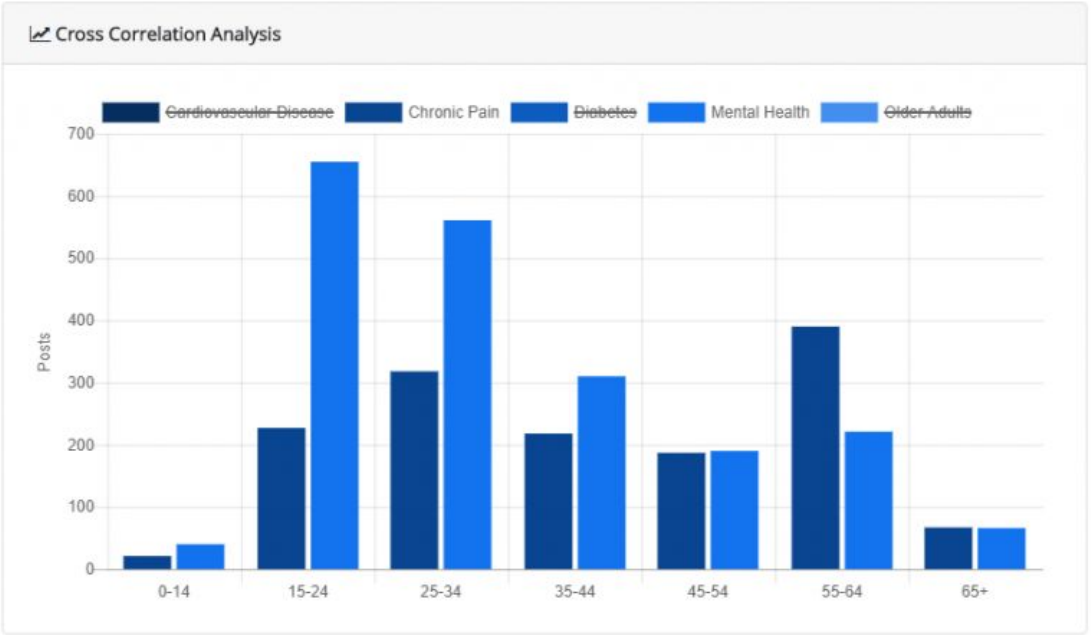
Yes, Netflix is the largest internet-television network in the world. But what most people don't realize is that, at its core, Netflix is a customer-focused, data-driven business. Founded in 1997 as a mail-order DVD company, it now boasts more than 53 million members in approximately 50 countries.

If you watch The Fast and The Furious on Friday night, Netflix will likely serve up a Mark Wahlberg movie among your personalized recommendations for Saturday night. This is due to data science. But did you know that the company also uses its data insights to inform the way it buys, licenses, and creates new content? House of Cards and Orange is the New Black are two examples of how the company leveraged big data to understand its subscribers and cater to their needs.

Social Media Text Analytics for Public Health

Mosaic Data Science designed and prototyped a web platform that facilitates interactive analysis of social media data related to chronic health conditions and disease management.

The results show that while mental health is a primary concern among young adults, social media users in their 50's and early 60's are at least as concerned with chronic pain. This type of insight, if supported by further research, could help to direct health outreach to the particular conditions most problematic among different age groups.



Assignment 1: What do you think data mining is for?

Identify a problem *from your own experience* that you think would be amenable to data mining. Describe:

- What the data is.
- What type of benefit you might hope to get from data mining.
- What type of data mining (classification, clustering, etc.) you think would be relevant.
- Name one type of data mining that you think would not be relevant, and describe briefly why not.
- For each, illustrate with an example, e.g., if you think clustering is relevant, describe what you think a likely cluster might contain and what the real-world meaning would be.

Read more

https://towardsdatascience.com/

https://www.mosaicdatascience.com/

https://www.reddit.com/r/dataisbeautiful/

https://www.reddit.com/r/datascience/

https://dssg.uchicago.edu/blog/

https://knowyourdata.withgoogle.com/

https://coolinfographics.com/