Measured Futures

Blanquerna 2018 Futures Summer School

Blanquerna
UNIVERSITAT RAMON LLULL

Enric Escorsa O'Callaghan

Barcelona, 3th of July 2018

IALE Tecnología

Specialists in cross-domain information services

Course Program

Part 1. ANALYSING AND ANTICIPATING TODAY

1.1. Traditional Intelligence & Foresight Tasks
1.2. Search, analysis (data & text mining), visualization…

Part 2. CURRENT CHALLENGES

2.1. A changing environment
2.2. The era of automatization: AI and Machine Learning
2.3. The changing role of the analyst-data scientist-futurist?

Traditional Intelligence & analytics

All organizations need to stay updated on what is going on

Source: IALE Tecnología

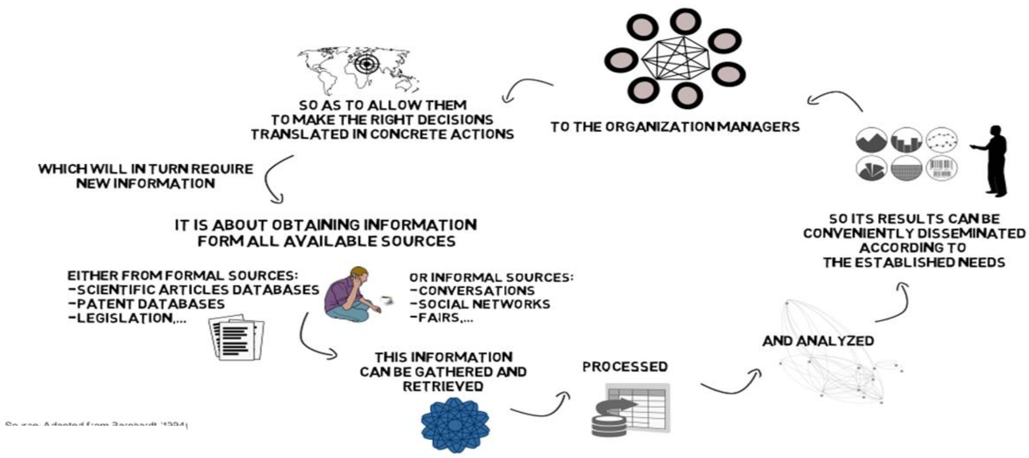
"Competitive Intelligence is the process of obtaining, analyzing, interpreting and diffusing the strategic value information on the industry and competitors, that is transferred to the decision makers at the right moment."

(Gibbons & Prescott, 1996)

"The Information Watch consists in observing and analyzing the scientific, technical, technological, and economical environment of a company to counteract potential threats and seize growth opportunities."

(Jakobiak, 1992)

CI lifecycle



Source: IALE adaptat de Bernhardt (1994)

Added value of Intelligence analytics

Keeping us up-to-date

- What are the main research lines in an area or domain?
- What technologies are emerging?
- What are the main players to be aware of?
  (Countries, reference centers, teams, people…)
- What is the market/impact/social context…

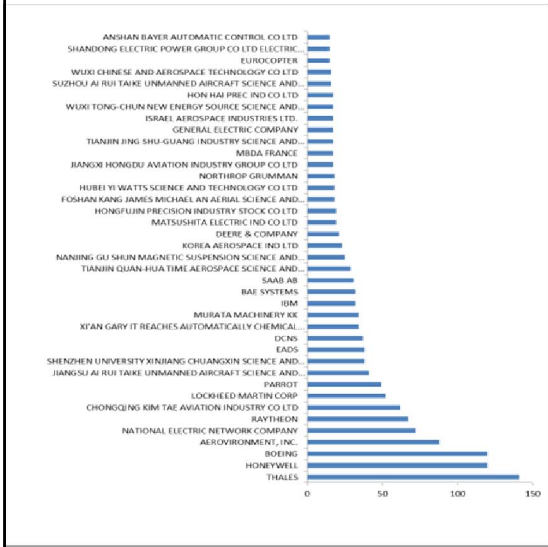
✓ Avoid carrying out R&D on what is already invented!

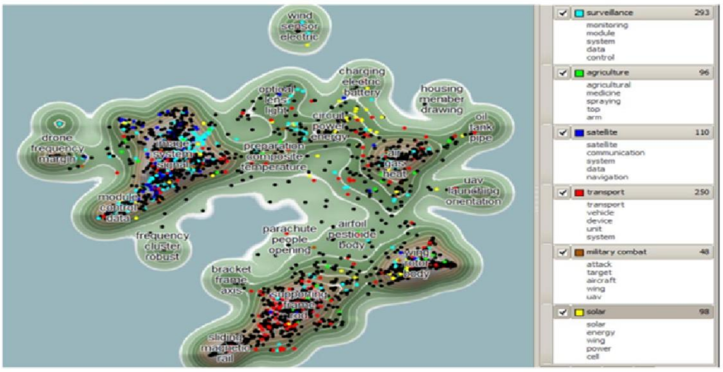
✓ Profiting from what already exists!

  - Realise of current opportunities for innovation/diversification
  - Buying/licensing an interesting patent,
  - Go for an strategic alliance / Joint venture
  - Hire talent,….

Source: Scott Simmerman

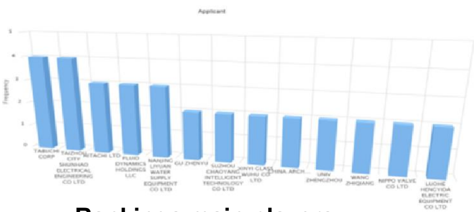
Perceiving the impact of new technologies



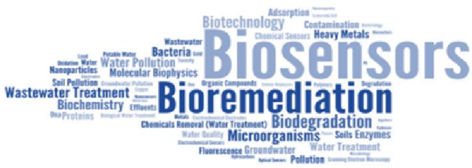


Source: IALE Tecnologia & IFICLAIMS "Hovering over the drone landscape"

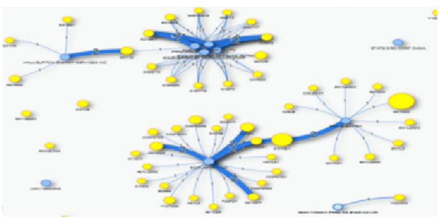
Domain characterization



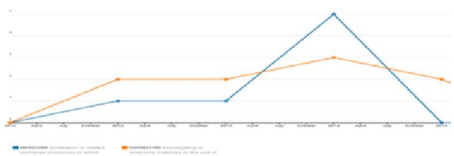
Rankings main players



Topics



Collaboration networks

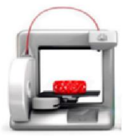


Trends

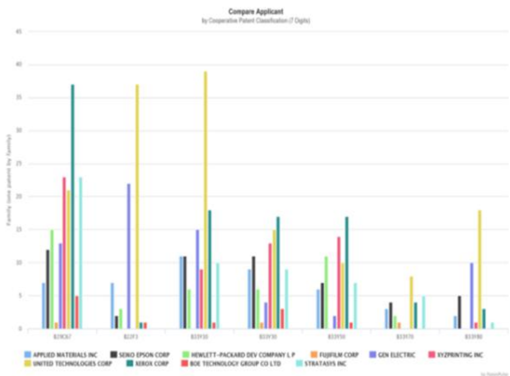
Source: PatentPulse, Patentinspiration,…

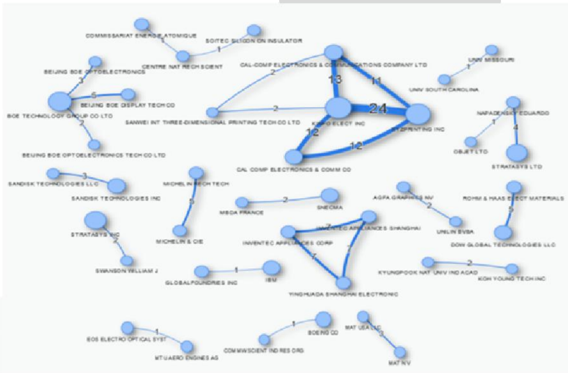
State of the Art of a domain: e.g. 3D Printing

Latest developments, evolution, emerging areas, Key players, collaboration networks



Source:
3DSystems

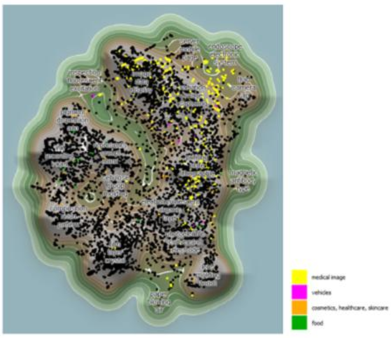




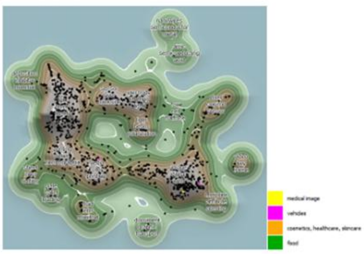
Source: PatentPulse – Matheo Software www.matheo-software.com

Competitive environment

E.g. Patent portfolio analysis & benchmarking (The FUJI vs KODAK case)



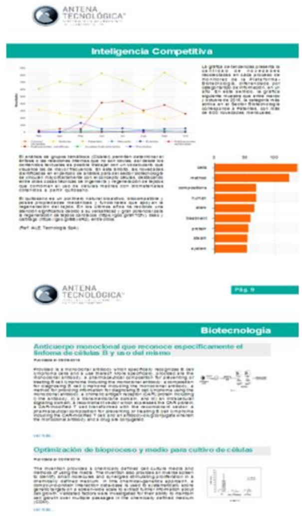
FUJI's portfolio



KODAK's portfolio

Read more...

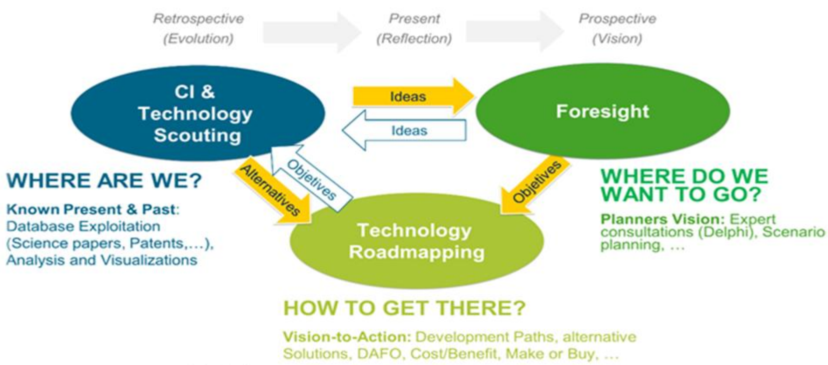
Monitoring. Alert systems

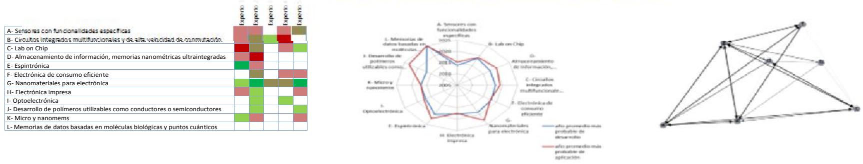


Source: Newsletters generated with VIGIALE platform by IALE Tecnología for the Ministry of C&T of Argentina
http://antenatecnologica.mincyt.gob.ar/

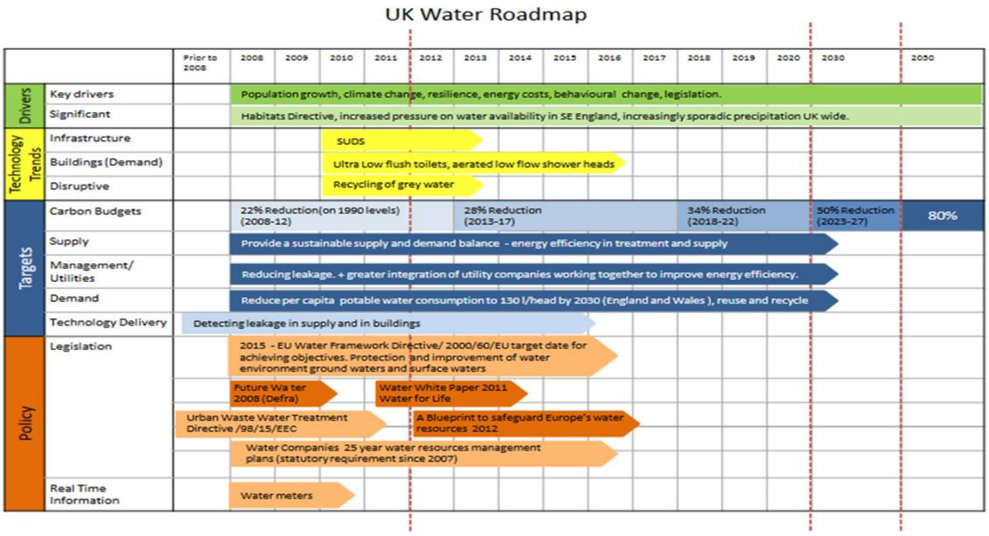
Prospective Studies

For orientation and industrial sectors policy



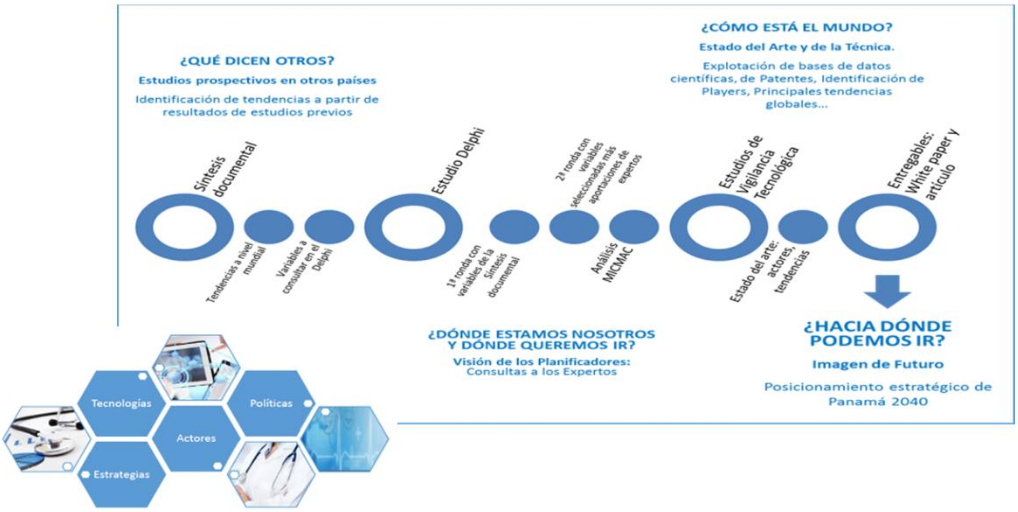


UK Energy, Water and Waste Roadmaps to 2050 (May 2012)



Source: http://www.retrofit2050.org.uk/sites/default/files/resources/2050_Roadmaps.pdf

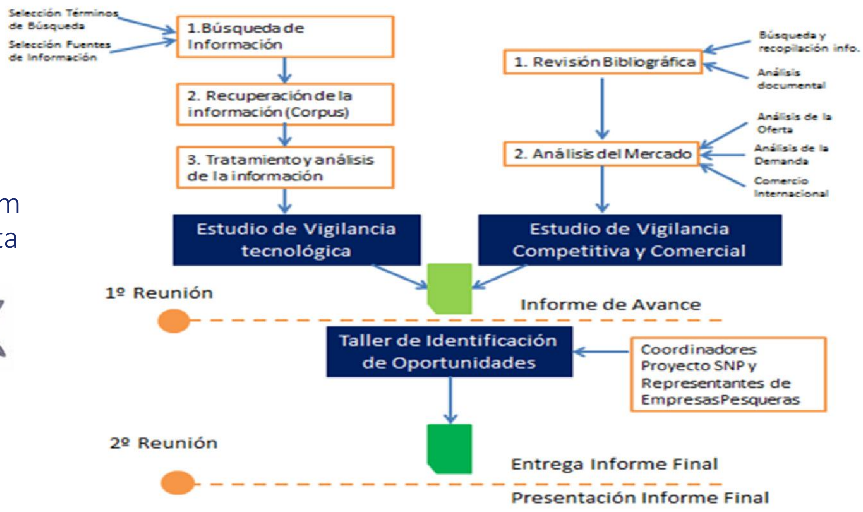
Examples



Source: IALE. Estudios Prospectivos sectores Agua /Salud /Educación para Panamá al año 2040

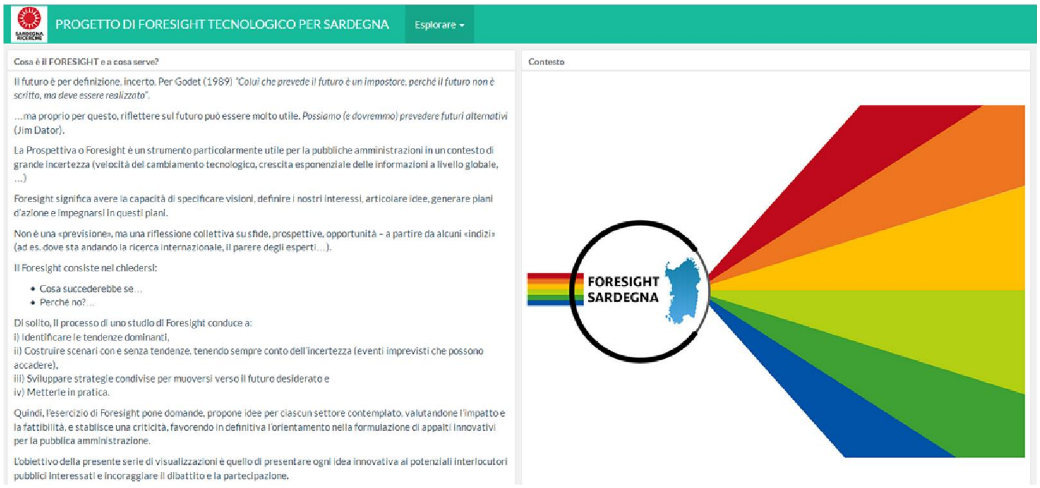
Examples

Study on
Enriched foods from
Peruvian Anchoveta



Source: IALE Tecnologia

Examples

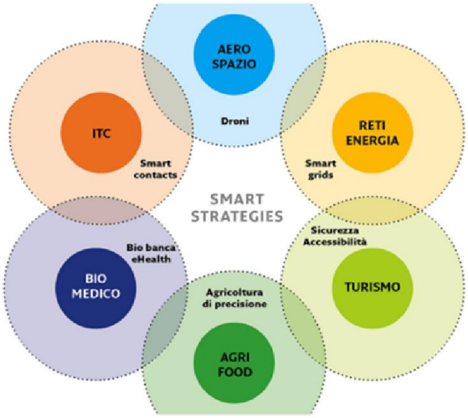


UNIONE EUROPEA
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR

PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA

Esplorare

Nuove idee su cui riflettere



Le «smart specialisation strategies» (S3), 2014-2020

Le «smart specialisation strategies» (S3) definiscono una generazione avanzata di politiche regionali di promozione e gestione dell'innovazione.

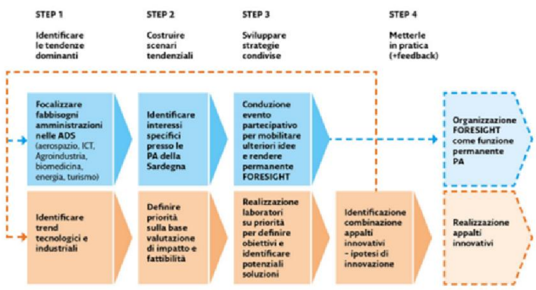
Quando una specializzazione è veramente smart?

• Quando si fonda su una visione «distintiva» • Quando sceglie in modo strategico • Quando sperimenta e innova • Quando le scelte sono frutto di un processo collettivo di «scoperta imprenditoriale» • Quando le scelte sono «outward looking», cioè guardano al contesto externo

PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA

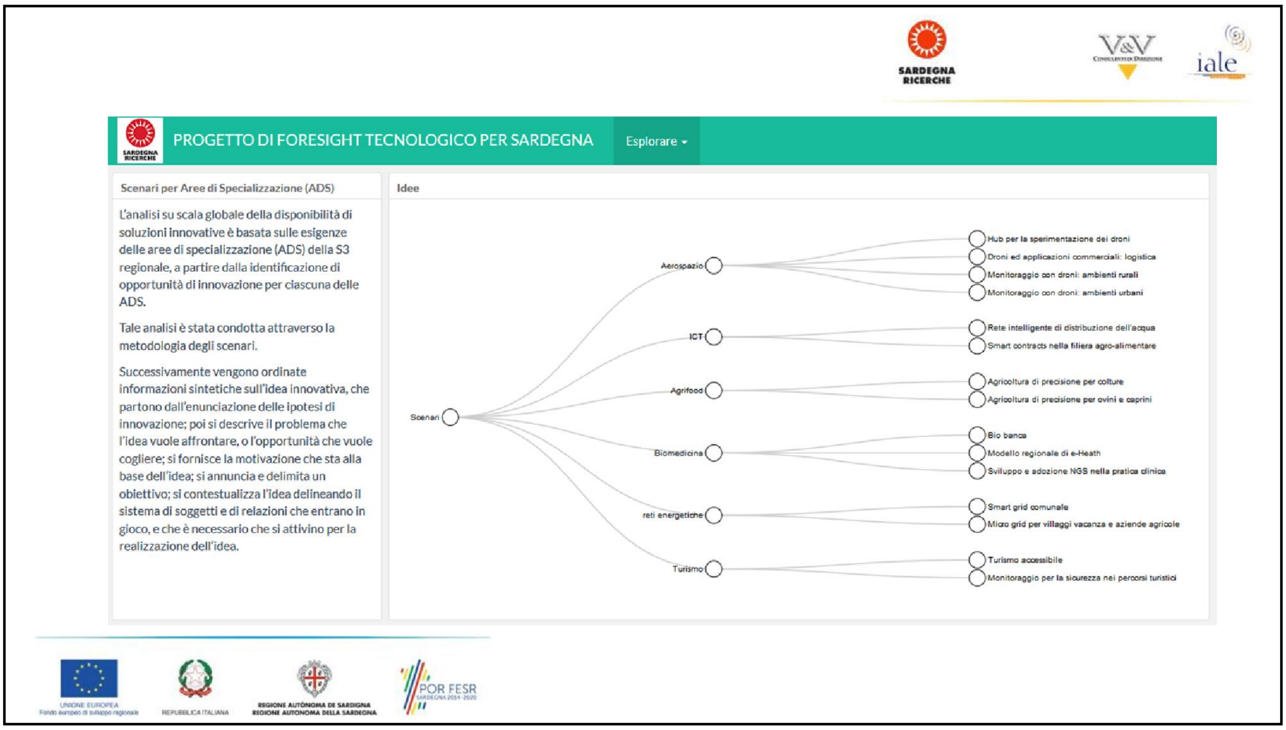
Esplorare

Il processo di costruzione di Scenari



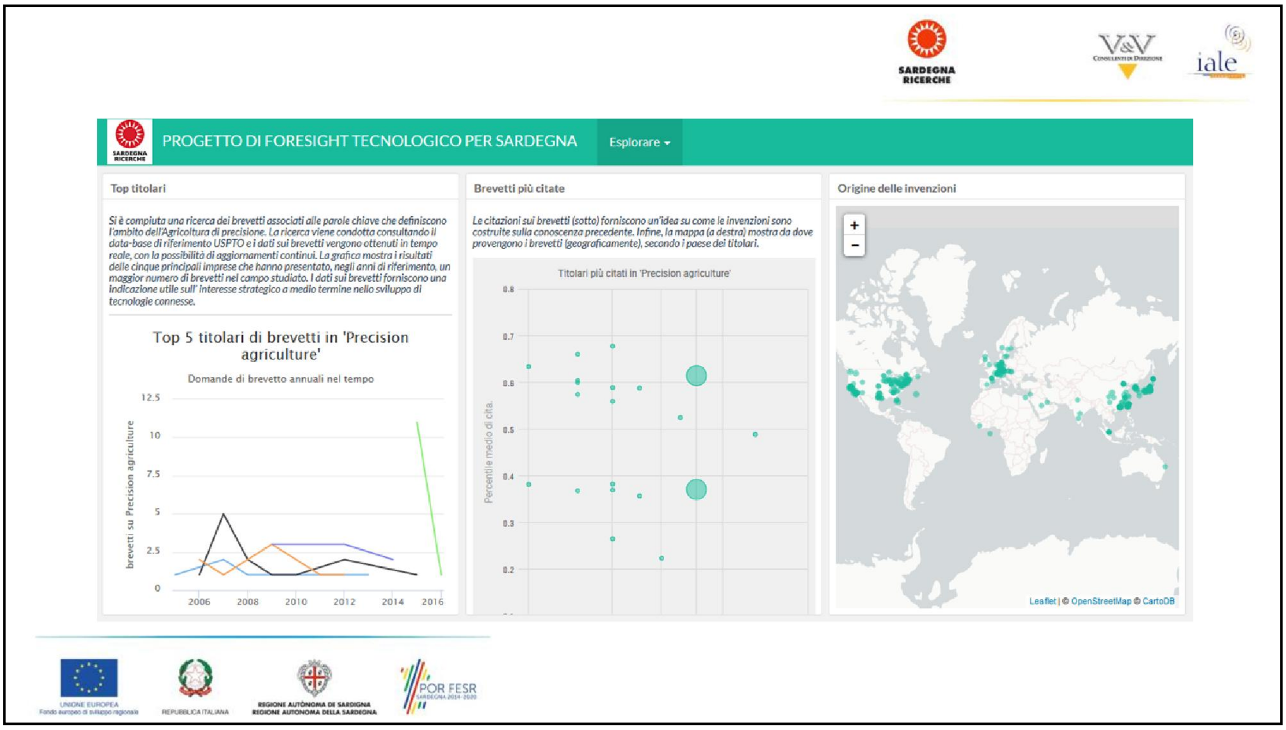
Steps

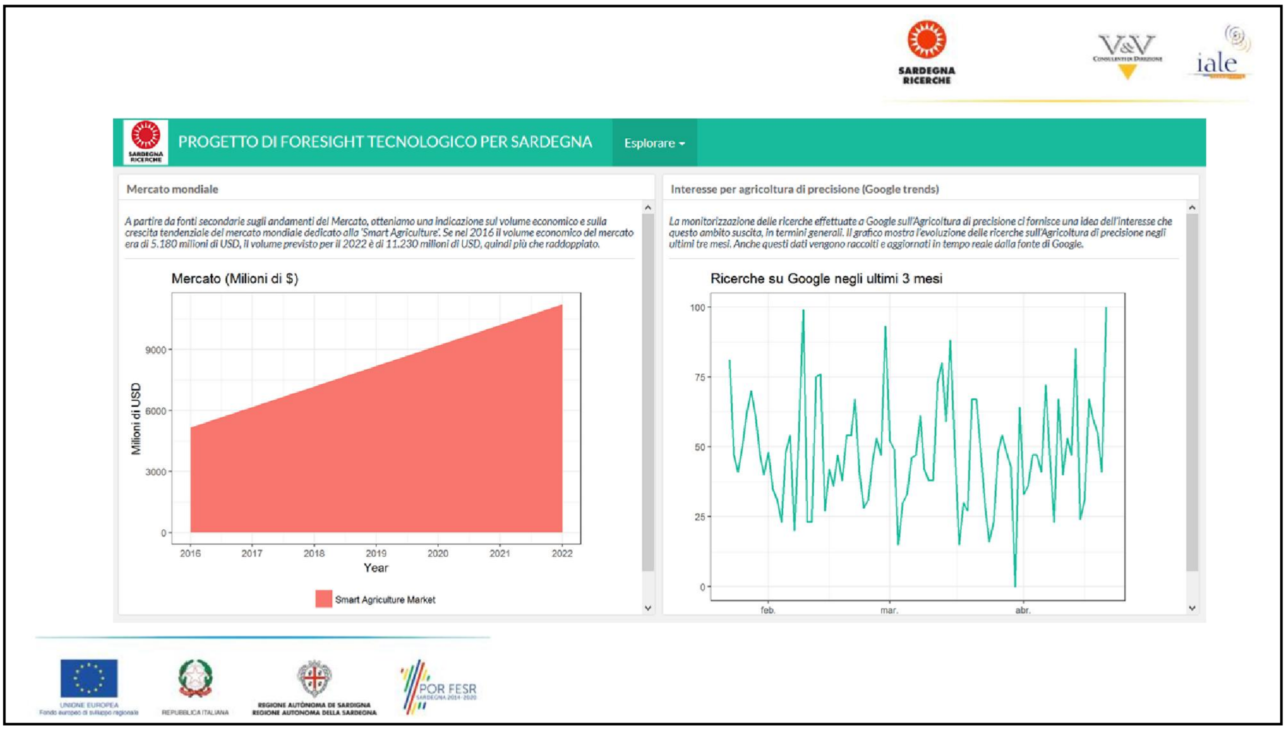
Step 1: Identificare le tendenze dominanti,
Step 2: Costruire scenari con e senza tendenze, tenendo sempre conto dell'incertezza (eventi imprevisti che possono accadere),
Step 3: Sviluppare strategie condivise per muoversi verso il futuro desiderato,
Step 4: Metterle in pratica.

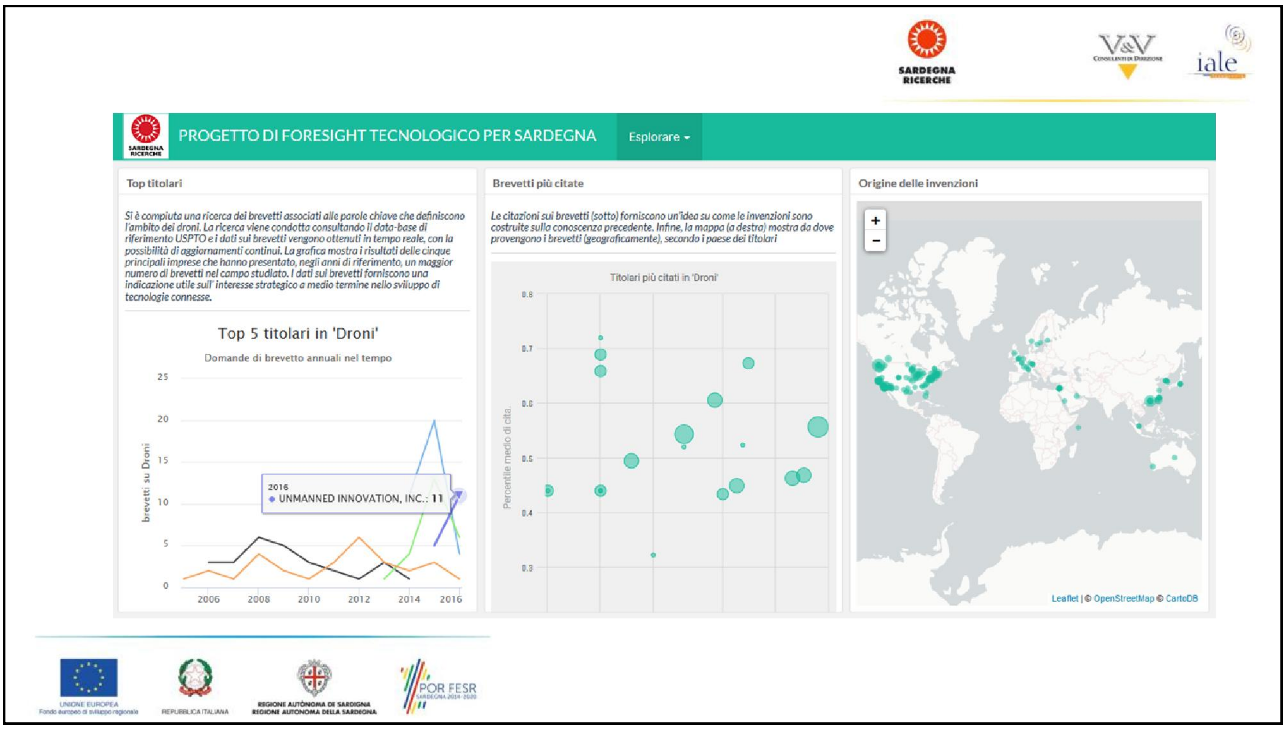

SARDEGNA RICERCHE
V&V
iale
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
Scenari per Aree di Specializzazione (ADS)
L'analisi su scala globale della disponibilità di soluzioni innovative è basata sulle esigenze delle aree di specializzazione (ADS) della S3 regionale, a partire dalla identificazione di opportunità di innovazione per ciascuna delle ADS.
Tale analisi è stata condotta attraverso la metodologia degli scenari.
Successivamente vengono ordinate informazioni sintetiche sull'idea innovativa, che partono dall'enunciazione delle ipotesi di innovazione; poi si descrive il problema che l'idea vuole affrontare, o l'opportunità che vuole cogliere; si fornisce la motivazione che sta alla base dell'idea; si annuncia e delimita un obiettivo; si contestualizza l'idea delineando il sistema di soggetti e di relazioni che entrano in gioco, e che è necessario che si attivino per la realizzazione dell'idea.
Idee
Scenari
Aerospazio
Hub per la sperimentazione dei droni
Droni ed applicazioni commerciali: logistica
Monitoraggio con droni: ambienti rurali
Monitoraggio con droni: ambienti urbani
ICT
Rete intelligente di distribuzione dell'acqua
Smart contracts nella filiera agro-alimentare
Agrifood
Agricoltura di precisione per colture
Agricoltura di precisione per ovini e caprini
Biomedicina
Bio banca
Modello regionale di e-Heath
Sviluppo e adozione NGS nella pratica clinica
reti energetiche
Smart grid comunale
Micro grid per villaggi vacanza e aziende agricole
Turismo
Turismo accessibile
Monitoraggio per la sicurezza nei percorsi turistici
UNIONE EUROPEA
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR


SARDEGNA RICERCHE
V&V
iale
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
IDEA 1: Agricoltura di precisione per colture (Agrifood)
Motivazione generale
L'applicazione delle innovazioni dell'Agricoltura di precisione alle singole colture risponde all'esigenza di migliorarne la redditività e sostenibilità, per favorirne l'adozione da parte delle singole aziende e contribuire così alla valorizzazione delle specializzazioni produttive della regione. Le colture a cui può essere applicata l'Agricoltura di precisione riguardano i più importanti comparti dell'agricoltura regionale, cominciando dai seminativi per finire con colture specifiche come la vite e l'olivo.
Obiettivo
L'obiettivo principale della "idea innovativa" sull'Agricoltura di precisione per le singole colture è quello di aumentarne le capacità competitive e la sostenibilità (ambientale, economica, sociale); si vuole inoltre contribuire al miglioramento della singola filiera di appartenenza (prodotti da forno e vitivinicola, nei casi esaminati).
Impatto
6
0
Impatto
10
Medio. Le filiere a cui ciò si applica sono numerose ma non è del tutto evidente quanto le filiere presentano un vantaggio competitivo potenziale.
Fattibilità tecnica ed amministrativa
8
0
Fattibilità
10
Rosa ampia di soggetti pubblici da coinvolgere: organismo certificatore regionale delle produzioni agricole, Provincia/Comune, Consorzi di bonifica, Agenzie regionali dedicate ai controlli e a sviluppo in agricoltura, Sezioni di ispezione degli alimenti di origine animale. Ipotesi di costo: aprox. 1,5 M€. In sinergia con altre idee innovative. Adatto a PCP/PPI
Benefici
*Riduzione costi di produzione e impiego di concimi e fitofarmaci. *Migliorare uso di risorse scarse (suolo, acqua). *Limitare le perdite alla raccolta. *Migliorare qualità della produzione dei cereali.'
Benefici PA
*Riduzione costi di produzione e impiego di concimi e fitofarmaci. *Migliorare uso di risorse scarse (suolo, acqua). *Sostegno alle materie prime regionali per filiera dei prodotti da forno. *Favorire la modernizzazione dellagricoltura
Criticità
*Dimensioni piccole delle aziende agricole e difficoltà di aggregazione degli interventi.
UNIONE EUROPEA
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR


SARDEGNA RICERCHE
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
Contesto
L'agricoltura di precisione è una strategia gestionale dell'agricoltura che si avvale di moderne strumentazioni -un insieme di tecnologie formate dal Global Navigation Satellite System (GNSS), sensori e immagini satellitari e aeree, insieme a Geographic Information Systems (GIS)- ed è mirata all'esecuzione di interventi agronomici tenendo conto delle effettive esigenze colturali e delle caratteristiche biochimiche e fisiche del suolo.
Satellites monitoring crops
Robots testing the ripeness of grapes
Drones destroying weeds
Concetti associati all'idea
La wordcloud visualizza con un colpo d'occhio le parole chiave più ricorrenti nelle riviste specializzate che si occupano di agricoltura di precisione.
tracciabilita mappatura sensore precisione satellite gestione decisioni sistemi agroalimentare gps agricoltura raccolta internet gnss uav smart coltivazione meteo dss gis fertilizzazione irrigazione remoto georeferenziazione rilevamento supporto
Esempi di implementazione (recenti sviluppi)
Gli esempi qui di seguito sono il frutto di una esplorazione di data-base internazionali (Technology scouting) e elencano informazioni utili sulle attività dei principali protagonisti del settore e su recenti esperienze internazionali nel settore dell'agricoltura di precisione:
CONSIGLIO PER LA RICERCA E LA SPERIMENTAZIONE IN AGRICOLTURA CRA Italia Device and a method for reducing the dispersion into the environment of dust from abrasion of dressed seeds by pneumatic precision sewing machines EP2611276A1
MASCHIO GASPARDO Seed distribution element for precision seed drills, seed drill including said element - WO2012062876A1; Sowing element for pneumatic precision seed drills - WO2015158900a1
MATERMACC Improved seed dispenser for a precision automatic sower.
SFOGGIA Seeder combined with fertilising apparatus
AgriXP Piattaforma progettata dagli agricoltori e dai loro consulenti agricoli per tracciare i dati sul campo
DJI Droni per agricoltura
DIGI Droni per manodopera agricola -W02017053135A1
AEROBOTS Produttore e operatore di UAV per l'agricoltura
ITERIS Supporto decisionale per l'irrigazione e la raccolta delle coltivazioni.
AGCO Tecnologie di precisione: FUSE "connecting your farm".
TRIMBLE NAVIGATION GPS and positioning solutions
TOPCON Sistemi di misura laser, alle antenne della stazione base gnss, ai metodi di sincronizzazione dei macchinari)
POR FESR


SARDEGNA RICERCHE
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
Top titolari
Si è compiuta una ricerca dei brevetti associati alle parole chiave che definiscono l'ambito dell'Agricoltura di precisione. La ricerca viene condotta consultando il data-base di riferimento USPTO e i dati sui brevetti vengono ottenuti in tempo reale, con la possibilità di aggiornamenti continui. La grafica mostra i risultati delle cinque principali imprese che hanno presentato, negli anni di riferimento, un maggior numero di brevetti nel campo studiato. I dati sui brevetti forniscono una indicazione utile sull' interesse strategico a medio termine nello sviluppo di tecnologie connesse.
Top 5 titolari di brevetti in 'Precision agriculture'
Domande di brevetto annuali nel tempo
brevetti su Precision agriculture
12.5
10
7.5
5
2.5
0
2006
2008
2010
2012
2014
2016
Brevetti più citate
Le citazioni sui brevetti (sotto) forniscono un'idea su come le invenzioni sono costruite sulla conoscenza precedente. Infine, la mappa (a destra) mostra da dove provengono i brevetti (geograficamente), secondo i paese dei titolari.
Titolari più citati in 'Precision agriculture'
Percentile medio di cita.
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Origine delle invenzioni
Leaflet | © OpenStreetMap © CartoDB
POR FESR


SARDEGNA RICERCHE
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
Mercato mondiale
A partire da fonti secondarie sugli andamenti del Mercato, otteniamo una indicazione sul volume economico e sulla crescita tendenziale del mercato mondiale dedicato alla 'Smart Agriculture'. Se nel 2016 il volume economico del mercato era di 5.180 milioni di USD, il volume previsto per il 2022 è di 11.230 milioni di USD, quindi più che raddoppiato.
Mercato (Milioni di $)
Milioni di USD
Year
Smart Agriculture Market
Interesse per agricoltura di precisione (Google trends)
La monitorizzazione delle ricerche effettuate a Google sull'Agricoltura di precisione ci fornisce una idea dell'interesse che questo ambito suscita, in termini generali. Il grafico mostra l'evoluzione delle ricerche sull'Agricoltura di precisione negli ultimi tre mesi. Anche questi dati vengono raccolti e aggiornati in tempo reale dalla fonte di Google.
Ricerche su Google negli ultimi 3 mesi
feb.
mar.
abr.
POR FESR


SARDEGNA RICERCHE
PROGETTO DI FORESIGHT TECNOLOGICO PER SARDEGNA
Esplorare
Top titolari
Si è compiuta una ricerca dei brevetti associati alle parole chiave che definiscono l'ambito dei droni. La ricerca viene condotta consultando il data-base di riferimento USPTO e i dati sui brevetti vengono ottenuti in tempo reale, con la possibilità di aggiornamenti continui. La grafica mostra i risultati delle cinque principali imprese che hanno presentato, negli anni di riferimento, un maggior numero di brevetti nel campo studiato. I dati sui brevetti forniscono una indicazione utile sull' interesse strategico a medio termine nello sviluppo di tecnologie connesse.
Top 5 titolari in 'Droni'
Domande di brevetto annuali nel tempo
brevetti su Droni
2016
UNMANNED INNOVATION, INC.: 11
Brevetti più citate
Le citazioni sui brevetti (sotto) forniscono un'idea su come le invenzioni sono costruite sulla conoscenza precedente. Infine, la mappa (a destra) mostra da dove provengono i brevetti (geograficamente), secondo i paese dei titolari
Titolari più citati in 'Droni'
Origine delle invenzioni
Leaflet | © OpenStreetMap © CartoDB
POR FESR

Current Challenges

Information explosion ("Infoxication")

Exponential growth

Globalized

Heterogeneous

Scopus (>57M registers)
Isi Web of Knowledge (>90M)
Medline Pubmed (>24M)
Ei Compendex (>20M)
ACM DL (>400K)
Espacenet/USPTO/Patentscope (>58M Patents)
...

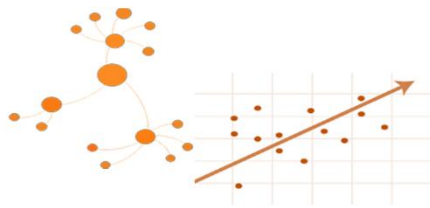
- ✓ Databases: Publications, patents, market...
- ✓ Open Access repositories
- ✓ OpenData from Governments
- ✓ Internet (free/invisible),
- ✓ Social networks
- ✓ Sensors data, IoT
- ✓ GIS data,
- ✓ Multimedia (video, audio)...

Advanced CI

Data mining tools

- ✓ Descriptive analytics
- ✓ Predictive analytics and Machine Learning

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa



Datamining or Knowledge Discovery in Databases KDD

Data mining is the process of extraction of implicit, previously unknown, and potentially useful information from data.

It uses powerful analytical tools and statistical metrics to process large volumes of data to reveal patterns and associations.

Advanced CI

Text mining and Natural Language Processing Techniques (NLP)

- ✓ extracting topics o theme clusters,
- ✓ identifying entities such as names, substances, trademarks, or companies
- ✓ detecting relationships between facts or concepts, etc.
- ✓ understanding people's sentiments, preferences

Basic process

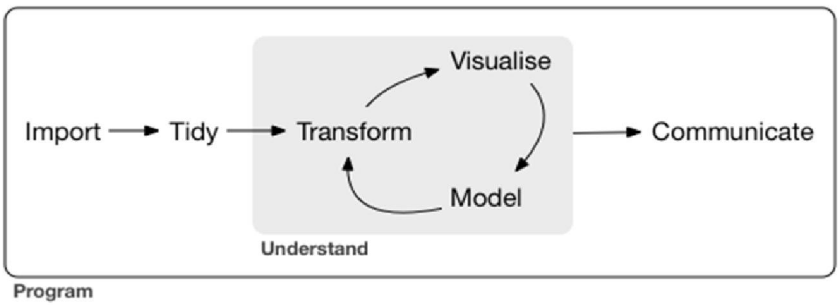
Pre-treatament
- Obtaining Data
- Creation of descriptors, dictionaries…
- Tyding (eliminating duplicates, missing values,….)

Treatment
- Simple counts of information fields
- Co-ocurrence analysis (co-word, citations,..)
- Creation of Analytical Models and its automatization (ML)

Visualization
- Histograms, Matrices, Graphs, Maps,…

Communication

Basic process (iterative understanding)



Source: Hadley Wickham R for Data Science http://r4ds.had.co.nz

Reproducible research

In the current context of global science trends (open access, open data, and open source tools), Reproducibility becomes crucial.

Writing code for our analysis allows:

1. Transparency throughout all the analytical process (not only in the results)
2. Sharing and collaboration
3. Re-executing this analysis either with same data or with new data (Reproducibility)

```
mydata <- 15 #I define a variable, my data
sum(mydata, 10) #I apply operations and functions,…

## 25 #I obtain results
```

…then I can save all that in a file: **"myanalisis"** and share it!

Platforms for collaboration & version control

A form of resistance!:

- ✓ Know the data sources
- ✓ Create your own tools
- ✓ Be accountable, be transparent
- ✓ Keep reproducibility
- ✓ Collaborate / Share / Learn…

Program or be programmed

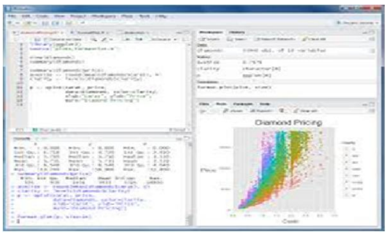
Microsoft is acquiring GitHub for $7.5

As of June 2018, there were 28 million developers in the GitHub community, as well as 85 million code repositories, making it the world's largest host of source code.

GitHub

Open Source Statistical Programming Languages & Reproducible research

E.g. Python, R,…



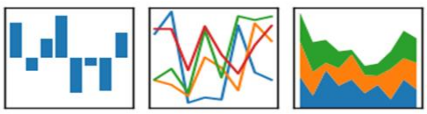
- ✓ Free and open source
- ✓ Dynamic supporting communities
- ✓ Allow for Reproducible research

Packages/libraries for **Data Manipulation**

Tidyverse, ensemble of packages including:
- ***Dplyr*** (for data manipulation)
- ***Tidyr*** (for data tidying)

https://www.tidyverse.org/

- **Pandas** (working with tables)
- **Scipy** (numerical computation),…



http://pandas.pydata.org/

Packages/libraries for **Text Mining**

Many packages:

- ***tm***
- ***OpenNLP***
- ***Tidytext***
- ***quanteda***
- ***stringr…***

https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

python™

spaCy

https://alpha.spacy.io/

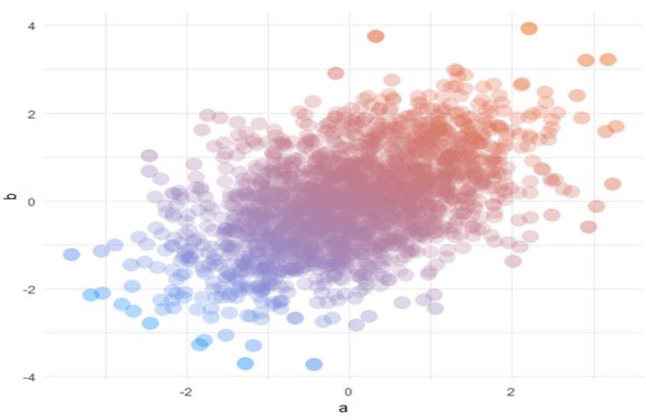
gensim
topic modelling for humans

https://radimrehurek.com/gensim/

Packages/libraries for **Visualization** in R

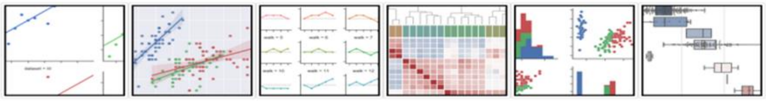
ggplot2
http://ggplot2.org/

ggplot2

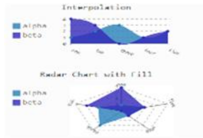


Hadley Wickham

Packages/libraries for **Visualization** in Python

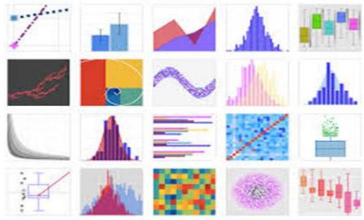


Seaborn http://seaborn.pydata.org/

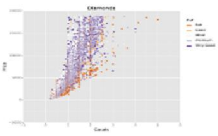


Pygal (interactiu)
http://pygal.org

Matplotlib http://matplotlib.org/



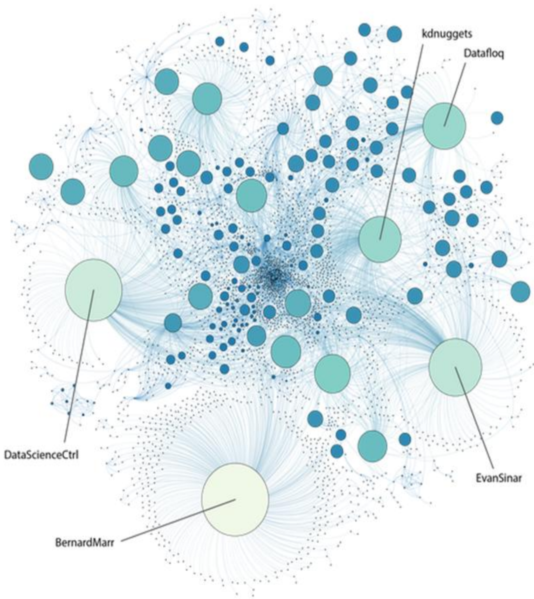
Plotly (interactiu)
https://plot.ly/



gglot http://ggplot.yhathq.com/
Wrapper de la llibreria ggplot de R per a Python

Bokeh (interactiu) https://bokeh.pydata.org

Example: Influencers in a domain

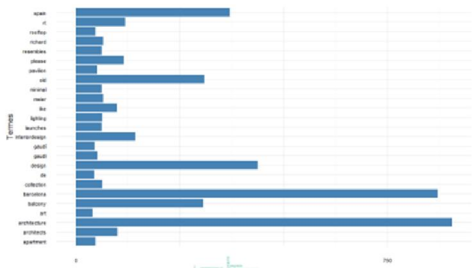


Source: nbviewer.jupyter.org

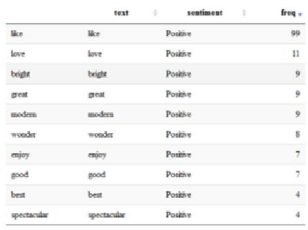
Example: Social network analysis for an event

Active users

Topics in conversations



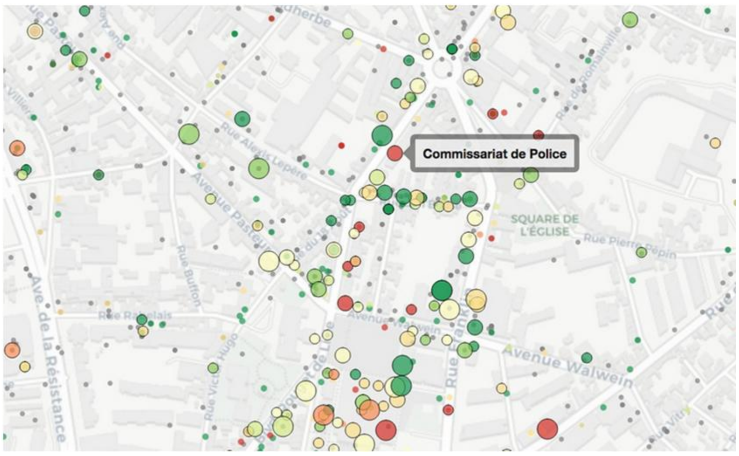
Sentiments



	text	sentiment	freq
like	like	Positive	99
love	love	Positive	11
bright	bright	Positive	9
great	great	Positive	9
modern	modern	Positive	9
wonder	wonder	Positive	8
enjoy	enjoy	Positive	7
good	good	Positive	7
best	best	Positive	4
spectacular	spectacular	Positive	4

Source: IALE

Example: maps



Subjective Cartography form Google Places Opinions data. Source: https://twitter.com/datagistips

Patents and technology areas

1 Food, beverages
2 Tobacco products
3 Textiles
4 Wearing apparel
5 Leather articles
6 Wood products
7 Paper
8 Petroleum products, nuclear fuel
9 Basic chemicals
10 Pesticides, agrochemical products
11 Paints, varnishes
12 Pharmaceuticals
13 Soaps, detergents, toilet preparations
14 Other chemicals
15 Man-made fibres
16 Rubber and plastic products
17 Non-metallic mineral products
18 Basic metals
19 Fabricated metal products
20 Energy machinery
21 Non-special-purpose machinery
22 Agricultural and forestry machinery
23 Machine tools
24 Special-purpose machinery
25 Weapons and ammunition
26 Domestic appliances
27 Office machinery and computers
28 Electric motors, generators, transformers
29 Electric distribution, control, wire, cable
30 Accumulators, batteries
31 Lighting equipment
32 Other electrical equipment
33 Electronic components
34 Signal transmission, telecommunications
35 Television and radio receivers, audiovisual electronics
36 Medical equipment
37 Measuring instruments
38 Industrial process control equipment
39 Optical instruments
40 Watches, clocks
41 Motor vehicles
42 Other transport equipment
43 Furniture, consumer goods

Foods & Beverages
Wearables

Pharmaceutical

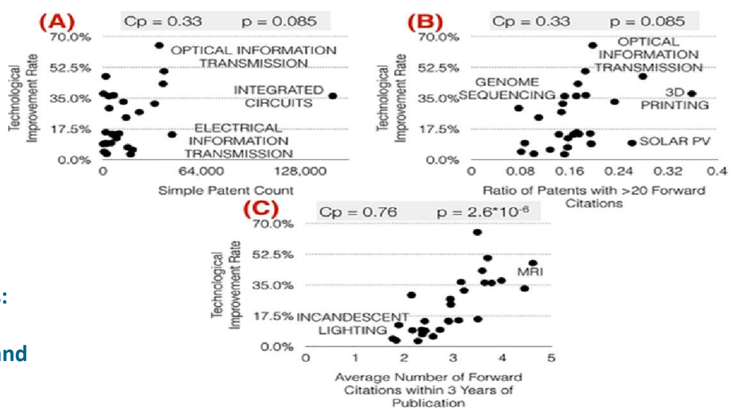
Computers
Batteries
Lighting equipment

Medical devices

Source: http://www.epo.org

Technology Area Assessment

Fig 1. Technological Improvement Rates vs Simple Patent Count (A), ratio of patents with greater than 20 citations (B), and average number of forward citations within 3 years of publication (C); the Pearson correlation coefficient (cp), the null hypothesis acceptance (cutoff at p = 0.05) and the values of the independent variable for the domains having maximum and minimum values are shown in the upper right corner.



Fastest-developing technologies: optical and wireless communications, 3-D printing, and MRI technology

Source: Benson CL, Magee CL (2015) Quantitative Determination of Technological Improvement from Patent Data. PLoS ONE 10(4): e0121635. doi:10.1371/journal.pone.0121635
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0121635

Technology Area Assessment

Machine Learning & AI

Basic statistical terminology

Observation (Normally Rows in a table)

=Observation = Sample = Example = Instance = Record = Mostra

Feature (Normally Columns in a Table)

=Characteristic = Predictor = Attribute = Input = Regressor

=Independent variable

?

Response (Each value we want to predict)

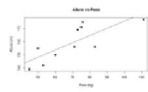
= Response = Target = outcome = Output = Label = Dependent variable

	Colour	texture	size
Sample 1	Golden yellow	Smooth	6,5 cm
Sample 2	...	...	...
..	...	...	...

→ Golden apple?

Main machine Learning Methods

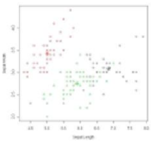
✓ ***Supervised Learning*** *Classification and Regression*



We can predict a response from several atributs (or independent variables). We generate a function to map inputs to desired outputs. **(p.ex. Filtering spam mails)**. We train this model until a desired level of precision over the training data.

Examples of Algorithms) ***Regression, Decision trees, Random Forest, kNN, Logístic regression,…***

✓ ***Unsupervised Learning*** *Clusterizing*



Here we do not have a Response or variable to predict or estimate. We use it to cluster a population in different groups. **(p.ex. For segmenting grups of clients.)**

Examples of Algorithms) ***Apriori, K-means, hierarchic clusterization***

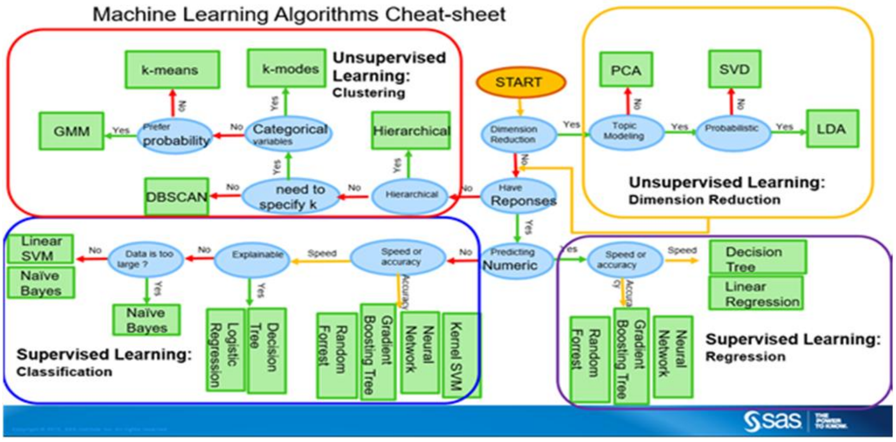
✓ *Reinforced Learning*

Automatism exposed to an environment where it can continuously self-train through trial and error and reward. It learns from experience trying to capture best knowledge to make precise decisions.

Ex) ***Markov Decision Processes***

The learning agent can be a neural net that maps states and actions (Deep learning)… **(p.e. artificial vision),….**

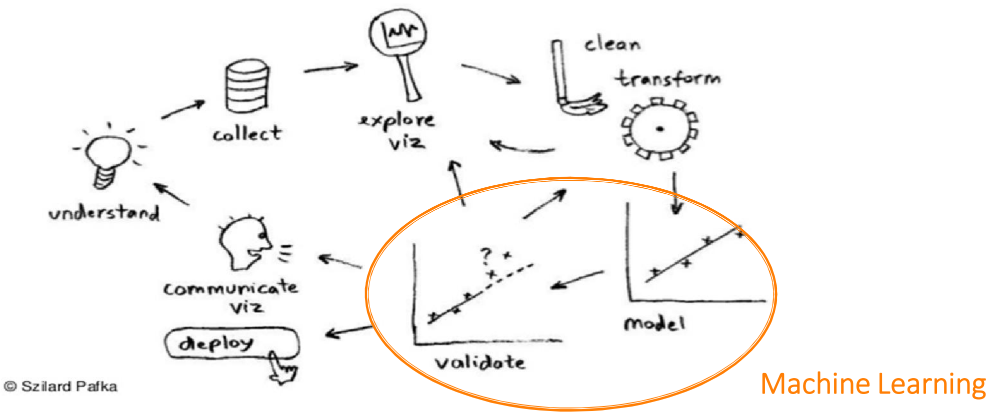
Main machine Learning Methods: Which Algorithm should I use



Sources:
Dousset, Bernard https://www.irit.fr/-Publications-?code=114&nom=Dousset+Bernard&lang=fr

Classification of main ML Algorithms. Source: Hui Lui, SAS
http://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use

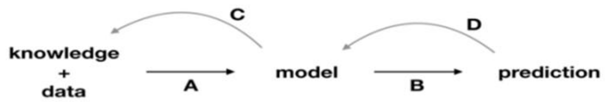
Data Science cycle



Source: "Data Science Methods and Tools" by Szilard Pafka via Angela Bassa

Data Science cycle

Typical workflow in ML



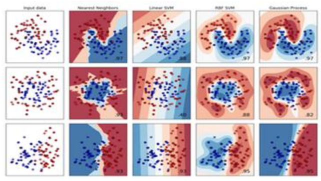
A. Modelling is a process in which domain knowledge and data are turned into models.
B. Models are used to generate predictions.
C. Understanding of model structure may increase our knowledge and in consequence leads to a better model. *DALEX helps here.*
D. Understanding of drivers behind particular model predictions may help to correct wrong decisions and in consequence leads to a better model. *DALEX helps here.*

Source: https://github.com/pbiecek/DALEX_docs/blob/master/workshops/eRum2018/Workshop_eRum_2018_part1.pdf

Packages/libraries for **Machine Learning**

Many packages:
- ***e1071***
- ***randomForest***
- ***Caret,***
- ***…***

https://cran.r-project.org/web/packages/caret/caret.pdf



http://scikit-learn.org/stable/

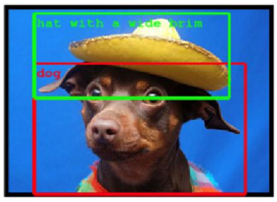
Environments and Packages for Deep Learning (many are Open Source)

- Torch i Pytorch (Facebook)
- TensorFlow (Google)
- CNTK (deep learning) i DMTK (machine learning) from Microsoft
- Deepmask (deep learning per a visió artificial) from Facebook
- Caffe2 (deep learning for Artificial vision) Yangqing Jia, Facebook
- DSSTNE (deep learning) d'Amazon
- Mahout (Machine learning) i BigDL (Deep Learning) from Apache
- SystemML from IBM (also Watson & Bluemix)
- PaddlePaddle from Baidu (Andrew Ng)
- MxNet (Pedro Domingos, U.Washington, Amazon)
- DL4J (Deeplearning for Java, Open source)

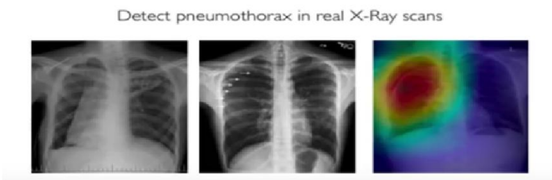
...

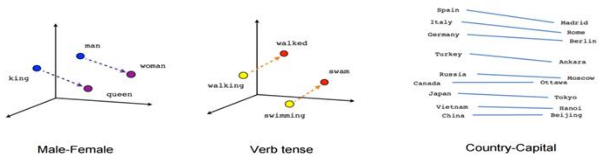
- Keras (François Chollet, Google)

Python API for Tensorflow



Object detection in Pictures, Source:Google Blog.



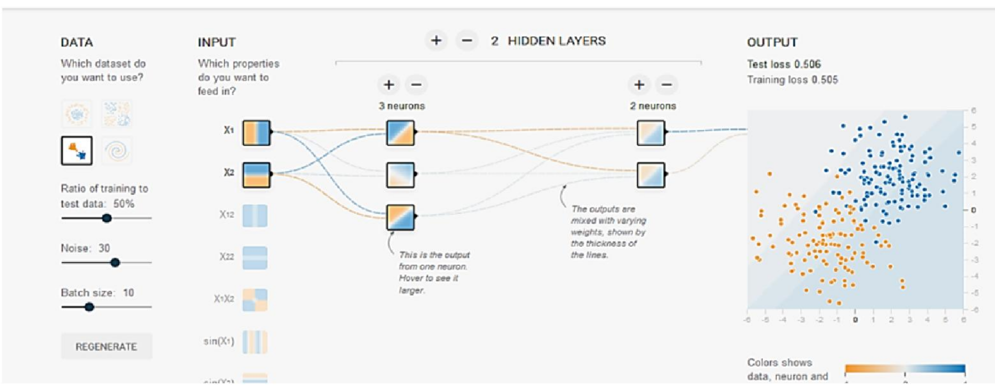


Associacions Word2Vec Font: Tensorflow

Font: MIT Introduction to Deep Learning 6.S191: Lecture 1 Foundations of Deep Learning Lecturer: Alexander Amini

Neural Networks and Deep Learning

TensorFlow from Google is Opensource and allows you to experiment with Neural Nets (http://playground.tensorflow.org)

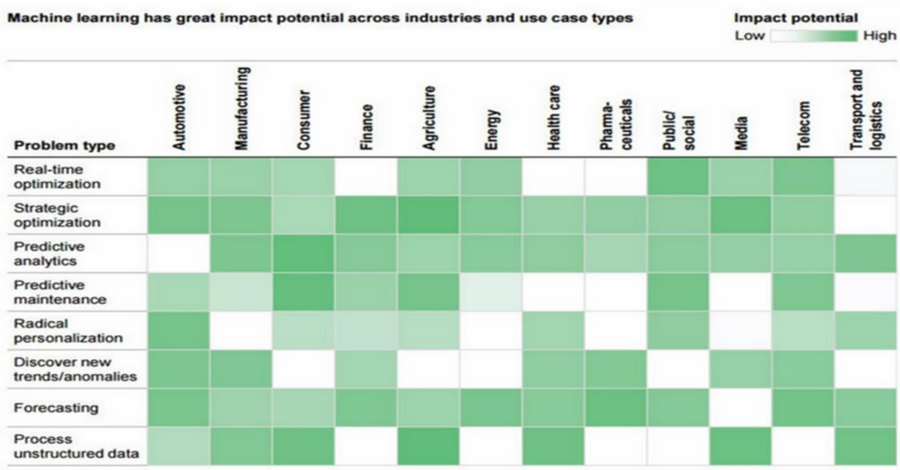


Source: Neural Networks and Deep Learning by Michael Nielsen & Deep Learning by Ian Goodfellow, Yoshua Bengio y Aaron Courville.

Some companies using Open Source Statistical Programming Languages (e.g. R & Python) today

Source: http://blog.revolutionanalytics.com/2017/06/applications-earl-sf-2017.html; https://www.quora.com/What-top-tier-companies-use-Python

AI and the age of automatization



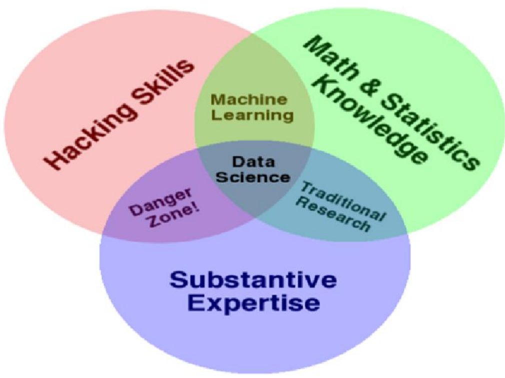
"Heat map" of automation

Source: Potential impact of ML in several industries. Source: McKinsey Global Institute (MGI), December 2016 "The age of analytics: competing in a data-driven world"
http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of analytics-competing-in-a-data-driven-world

Intelligence Professionals today

CI Professionals today

"Statisticians will be the sexiest profession in the next decade"
Hal Varian, Chief economist at Google (2009)



Source: Drew Conway (author of Machine Learning for Hackers.) Data Consulting 2015
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

CI Professionals today

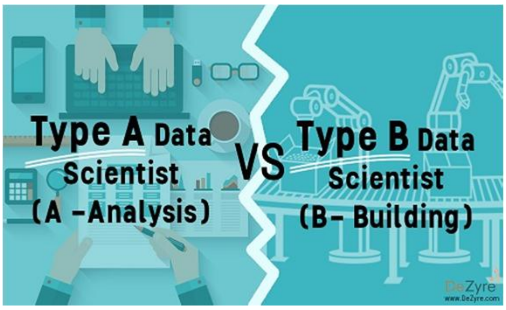
- ✓ Domain Expertise
  - ✓ Business and industry Knowledge, context, social, cultural, normative, IP, analytical mind, strategic thinking, etc.
- ✓ Mathematics & Statistics
  - ✓ Some key issues: control overfitting, detect/handle outliers, differenciate correlation/causation, Inference algorithms, etc.
- ✓ Data management
  - ✓ Collect & Wrangling, Data Visualization, Tools, APIs, mashups, dashboards, DDBB querying, programming languages (SQL, Python, R,...)...
- ✓ Communication
  - ✓ Storytelling, report drafting, configuring tailored alert systems, newsletters, social media, etc

Sources:
Roger Huang https://www.springboard.com/blog/how-to-become-a-data-scientist/
Thomas C. Redman "Data Scientists get out and talk to people" https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

CI Professionals today

Some distinguish from **type "A" data Scientist** vs **Type "B" data Scientist**

Type "A" (Analysis):
have strong statistics skill and the ability to work with messy data and communicate results.

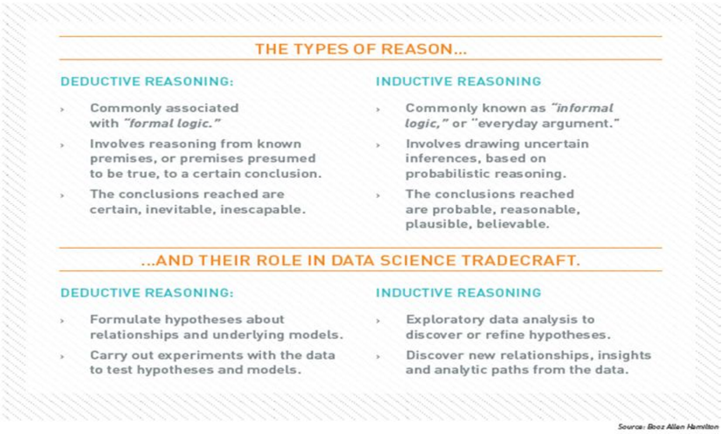


Type "B" (build):
have very strong coding skills, maybe have a background in software engineering, and focus on putting machine learning models, such as recommendation systems, into production.

Sources:
Michael Hochser https://www.quora.com/What-is-data-science/answer/Michael-Hochster
Emily robinson http://hookedondata.org/Advice-for-Applying-to-Data-Science-Jobs/
Dzyre https://www.dezyre.com/article/type-a-data-scientist-vs-type-b-data-scientist/194

Theory driven (deductive) vs Data driven (inductive)

Traditional science (deductive)
Observation->
Hipotesis->
Experiment->
validation



The Types of Reason and Their Role in Data Science Tradecraft

Machine Learning (inductive)
Data&knowledge->
Ass./patterns
(modeling)->
Insights/Predictions

Sources:

Kuonen, Diego Keynote NTTS 2017, Brussels, Belgium March 14, 2017

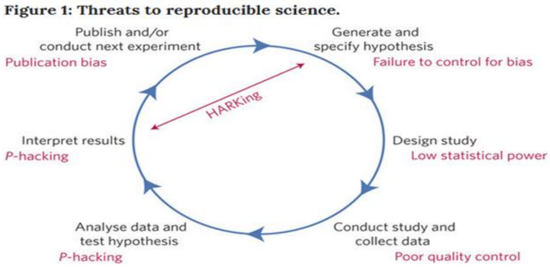
Booz Allen Hamilton. “The Field Guide to Data Science” 2015. Web Accessed 1 February 2017. SSRN

Theory driven (deductive) vs Data driven (inductive)

Dangers of deductive reasoning:

- ✓ P-hacking, HARKing (hypothezing after the results are known)
- ✓ Need for constantly contrasting and leveraging results with the real world

Figure 1: Threats to reproducible science.



An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], P-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Source: Mainfesto Reproducibility: http://www.nature.com/articles/s41562-016-0021

Theory driven (deductive) vs Data driven (inductive)

Dangers of inductive reasoning:

- ✓ Today certain automatisms make decisions that affect business and sometimes our culture and our lives (**biased** decisions based in race, location or gender)
- ✓ Algorithms are **Black boxes** for the common of mortals. Are algorithms fair?

(at least they need to be more transparent)

> *"All models are wrong but some of them are useful"*
>
> George Box (1919-2013)

Replacing one type of reasoning for the other (both need to go together; we actually need to mix them)

We must always go back to the real world and keep humans in the loop

Sources:

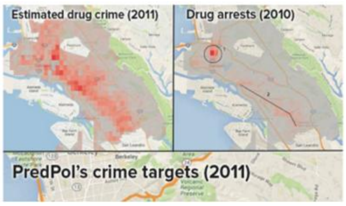
Peter Norvig, Director Research Google "the unreasonable effectiveness of data" Stanford https://www.youtube.com/watch?v=yvDCzhbjYWs

Katharine Jarmul PyData, Amsterdamm 2017 https://www.youtube.com/watch?v=hDgXIUM3Rmw

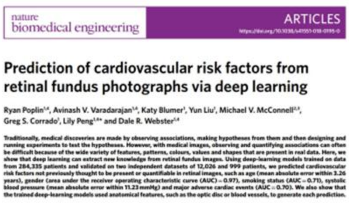
Ethical aspects: ALGORITHMIC IMPACT ASSESSMENT (AIA)

iale TECNOLOGIA

Automated crime prediction



Early diagnosis of deseases


nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0195-0

Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Ryan Poplin[1,4], Avinash V. Varadarajan[1,4], Katy Blumer[1], Yun Liu[1], Michael V. McConnell[2,3], Greg S. Corrado[1], Lily Peng[1,4*] and Dale R. Webster[1,4]

Traditionally, medical discoveries are made by observing associations, making hypotheses from them and then designing and running experiments to test the hypotheses. However, with medical images, observing and quantifying associations can often be difficult because of the wide variety of features, patterns, colours, values and shapes that are present in real data. Here, we show that deep learning can extract new knowledge from retinal fundus images. Using deep-learning models trained on data from 284,335 patients and validated on two independent datasets of 12,026 and 999 patients, we predicted cardiovascular risk factors not previously thought to be present or quantifiable in retinal images, such as age (mean absolute error within 3.26 years), gender (area under the receiver operating characteristic curve (AUC) = 0.97), smoking status (AUC = 0.71), systolic blood pressure (mean absolute error within 11.23 mmHg) and major adverse cardiac events (AUC = 0.70). We also show that the trained deep-learning models used anatomical features, such as the optic disc or blood vessels, to generate each prediction.

1) Respect the public's right to know which systems impact their lives and how they do so by **publicly listing and describing algorithmic systems used to make significant decisions affecting identifiable individuals or groups,** including their purpose, reach, and potential public impact;
2) Ensure greater accountability of algorithmic systems by **providing a meaningful and ongoing opportunity for external researchers to review, audit, and assess these systems using methods that allow them to identify and detect problems;**
3) **Increase public agencies' internal expertise and capacity** to evaluate the systems they procure
4) **Ensure that the public has a meaningful opportunity to respond to and, if necessary, dispute an agency's approach to algorithmic accountability.**

Source: https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde; https://ainowinstitute.org/aiareport2018.pdf

In summary: current roles of the Intelligence analyst

- ✓ Human resource for Intelligence/Data Science/Analytical /Futurist tasks is its main asset. An it is an scarce one.
- ✓ Need for a professional figure dedicated to understanding data, its origin and nature, the involved adquisition methods, its quality and context
- ✓ That is able to propose/derive possible uses according to specific motivations or social, organizational or industrial problems, acting as the bridge between human and/or organizational needs and algorithms potentially augmenting our capabilities.
- ✓ That ensures the dialog between deductive and inductive reasoning
- ✓ That is able to work cooperatively in networks
- ✓ That is able to comunicate and share analytical processes and results

Supporting strategic decisions, leveraging -based on them- the necessary data, contrasting the value of information and distinguishing from noise, contrasting the veracity and usefulness of models with the real world, and proposing creative ways of identifying problems and deploying meaningful shared narratives.

A brilliant, challenging future, awaits for the intelligence analyst!

Thank you very much!

enric@ialetecnologia.com
https://twitter.com/ialeT
www.ialetecnologia.com