Analyzing a CRIS: From data to insight in university research
“The amount of data we produce every day is truly mind boggling”.

Data is growing faster than even before and by the year 2020, about 1.7 Mb of new information will be created every second for every human being on the planet.

Bernard Marr (Forbes 2018)
The famous Internet minute...

2018 What happens in an Internet minute
Is so relevant, that we can say:

Data is the new oil

https://scientistcafe.com
The human brain cannot process so much data...

And that’s why we start talking about data analytics....

Don't know what to do!!

So much data....
Let’s talk about **data analytics**... an interview:

Is something new?

No! Analytics is all around us! We use it in our day to day. Every time we choose something: our professional career, our next holidays... we choose options based on our experiences, or related on what happens around us... It’s a natural process.
Let’s talk about **data analytics**... an interview:

Where is analytics used?

Almost everywhere, all the big corporations collects lots of data and predict, how to purchase or the client preferences. Political parties, use it to send personalized messages ...
So we can define:

“Data analytics is the ability to collect and use data to generate insights that inform fact-based decision making.”
We can classify the data analytics process in four stages:

**Descriptive**
Explains what happened.

**Diagnostic**
Explains why it happened.

**Predictive**
Forecasts what might happen.

**Prescriptive**
Recommended an action based on the forecast.

It's easy to conclude that more value is added as you advance in the list below!
Data analytics lifecycle is an iterative process:

- Explore available data
- Formulate the question
- Formulate the hypothesis

- Determine necessary data
- Collect and clean data
- Analyze data consistency

- Implement the models, using the best in each case
- Validate the model

- Find which variables tells and which predicts
- Select possible models and algorithms to use
- Define performance metrics

- Decision-making based on the results
- Define standards
- Goals to achieve standards
- Plan and allocate resources

- Perform results
- Generate the correct visualizations to communicate results
- Make improvement recommendations

- Use
- Prepare data
- Plan model
- Develop model
- Communicate
So it’s not trivial to find results through data, it is necessary to draw an elaborate plan and take some actions. New kind of experts appeared with new skills...

**Data Scientist**
also known as Data Managers, statisticians.

- A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

- **Skills:** Mathematics, Programming, Communication

- Will use programmes such as: SQL, Python, R

**Data Engineers**
also known as database administrators and data architects.

- They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

- **Skills:** Programming, Mathematics, Big data

- Will use programmes such as: Hadoop, NoSQL, and Python

**Data Analysts**
also known as business Analysts.

- They typically help people from across the company understand specific queries with charts.

- **Skills:** Statistics, Communication, Business knowledge

- Will use programmes such as: Excel, Tableau, SQL

Having all this in mind, at SIGMA, since 2011, we are developing analytical environments to give universities tools to easy their information analysis both in research and academic areas.

In this way, we have experienced data scientists & data engineers that create analytical environments for Teaching, Students, Economic management, Research and so on.
Thanks to our work close to the universities, we also have **data analysts** profiles experienced in higher education processes (both academic and research)
Focusing in research area, it is easy to do when we have the enough maturity in the CRIS system that incorporates the largest number of scientific information of the university with quality.
To develop a project of data analytics and focusing on the Research area, we followed the next 6 steps:

1. Classify the Research scope in distinctive knowledge areas
2. Translate the CRIS (CERIF compliant) model to an analytical data model
3. Define main indicators in each knowledge area with university experts collaboration
4. Implement (calculate if necessary) the indicators and main analytical outputs
5. Validate the data quality (data cleansing)
6. Open product on the university, improve process (iterative)
1. Classify the Research scope in distinctive knowledge areas, called analytical cubes:

- Awards
- Scientific Production
- Projects & contracts
- Research groups
- Bibliometrics / Altmetrics
- Entrepreneurship
- Patents
2. Translate de CRIS (CERIF) model into a star model, unlinking the analytical environment from the transactional environment.
3) Define the main indicators in each knowledge area or cubes with the university experts collaboration. i.e:

Scientific Production

We can analyze many facts:
- # of citations at different levels
- # of publications of all kinds
- Reputation (author&publication)
- Rankings of different kinds
- ....

Through many dimensions:
- Author data
- Year of publication
- Impact (wos, scopus, google scholar...)
- Citations (wos, scopus, google scholar...)
- SJR
- ....
Projects & Contracts:

We can analyze many facts:
- # of projects
- # of active projects
- Expected income
- Real income
- Total income
- % of public financing
- % of private financing
- ....

Through many dimensions:
- Award
- Grant
- Associates
- Country
- ....
4. Implement (calculate if necessary) the indicators and main analytical outputs with an analytical tool.
5. Validate the data quality (avoiding redundancies, referential integrity, range of values...)

- Data deduplication
- Quality check
- Data normalization
- Data standardization
6. Open product on the university, improve the process (if necessary) → start the iterative process
When we have knowledge, we have the first insights and we can act based on it...
Examples: we can analyze all the authors of the institution...
(dashboards)

Publications & Bibliometrics

<table>
<thead>
<tr>
<th>Year</th>
<th>Num. of Articles</th>
<th>Num. of Authors</th>
<th>Num. of Journals</th>
<th>Scopus Citations</th>
<th>WOS Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5,5k</td>
<td>1,49k</td>
<td>2,28k</td>
<td>61,76k</td>
<td>57,11k</td>
</tr>
</tbody>
</table>

Top 10 Journals by Articles

- PLoS ONE [1832-5283] 183
- Gazette sanitaria [0213-9111] 57
- La Vanguardia... 46
- InDestr: revista p... 37
- Archivos de Bronconeumol... 47
- Lecture Notes in Computer... 46

Articles by Language

- English: 61
- Spanish: 25
- Other: 1

Authors & Articles by Genre

- Articles: 889
- 8.53k
- 2.25k

Top Citations Articles

- "If you are good, I get better": the role of social hierarchy in perceptual decision-making
- Total Scopus Citations: 15

Top H-Index Authors

<table>
<thead>
<tr>
<th>Authors</th>
<th>Max H-Index Scopus</th>
<th>Max H-Index WOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SERRA GUIGO, RODERIC</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>MARTINEZ COMAS, DAVID</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>DEU SUNYER, JORDI</td>
<td>72</td>
<td>45</td>
</tr>
<tr>
<td>LOPEZ MALDONADO, RAFAEL</td>
<td>53</td>
<td>44</td>
</tr>
<tr>
<td>BOQUE ANTó, JOSE MARIA</td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>BUSTOS BERTRANPETIT, JAUME</td>
<td>52</td>
<td>40</td>
</tr>
<tr>
<td>MAJJO CALAFELL, FRANCESC D'ASSIS</td>
<td>48</td>
<td>39</td>
</tr>
<tr>
<td>GARRIGA POSAS, FRANCESC</td>
<td>36</td>
<td>38</td>
</tr>
</tbody>
</table>
Examples: ...or only one author

Publications & Bibliometrics

Top 18 Journals by Articles

- Genome Research [1886-9051]
- Nature [0028-0836]
- Nucleic Acids Research [0028-0836]
- Bioinformatics [1367-4884]
- PLoS ONE [1932-6203]
- Nature Methods [1548-0338]

Top Citations Articles

1. Absence of canonical marks of active chromatin in developmentally regulated genes
2. An encyclopedia of mouse DNA elements (Mouse ENCODE)
3. An integrated encyclopedia of DNA elements in the human genome
4. ASPicGeneID: a lightweight pipeline for gene prediction and alternative isoforms detection
5. Assessment of transcript reconstruction methods for RNA-seq
6. BLUEPRINT to decode the epigenetic signature written in blood

Authors & Articles by Genres

- 1 article by genre

Top H-Index Authors

- SERRA GUIGO, RODERIC
  - Max H-Index Scopus: 65
  - Max H-Index WOS: 49
Examples: ...or the authors from one article

Publications & Bibliometrics

<table>
<thead>
<tr>
<th>Year</th>
<th>Num. of Articles</th>
<th>Num. of Authors</th>
<th>Num. of Journals</th>
<th>Scopus Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

Top 10 Journals by Articles

Social Cognitive and Affective Neuroscience [1749-5924]
1

Articles by Language

Authors & Articles by Genre

<table>
<thead>
<tr>
<th>Article</th>
<th>Total Scopus Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>'If you are good, I get better': the role of social hierarchy in perceptual decision-making</td>
<td>15</td>
</tr>
</tbody>
</table>

Top H-Index Authors

<table>
<thead>
<tr>
<th>Authors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GALLES SEBASTIAN, NURIA</td>
<td></td>
</tr>
<tr>
<td>DECO, GUSTAVO</td>
<td></td>
</tr>
<tr>
<td>PANNUNZI, MARIO</td>
<td></td>
</tr>
<tr>
<td>GARCIA SANTAMARIA, HERNANDO</td>
<td></td>
</tr>
<tr>
<td>GIMENO AYNETO, ALBA</td>
<td></td>
</tr>
</tbody>
</table>
Examples: ...we can elaborate university author rankings based on the main impact indicators that we have stored in the CRIS through the years

### Rankings

<table>
<thead>
<tr>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

### Top H-Index Authors

<table>
<thead>
<tr>
<th>Authors</th>
<th>Q</th>
<th>Total Scopus Citations</th>
<th>Total WOS Citations</th>
<th>Max H-Index Scopus</th>
<th>Max H-Index WOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totales</td>
<td></td>
<td>61762</td>
<td>57113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERRA QUIGO, RODERIC</td>
<td></td>
<td>7500</td>
<td>7314</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>MARTINEZ COMAS, DAVID</td>
<td></td>
<td>523</td>
<td>515</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>DEU SUNYER, JORDI</td>
<td></td>
<td>761</td>
<td>755</td>
<td>72</td>
<td>45</td>
</tr>
<tr>
<td>LOPEZ MALDONADO, RAFAEL</td>
<td></td>
<td>729</td>
<td>663</td>
<td>53</td>
<td>44</td>
</tr>
<tr>
<td>BOQUÉ ANTÓ, JOSE MARIA</td>
<td></td>
<td>424</td>
<td>425</td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>BUSQUETS BERTRANPETIT, JAUME</td>
<td></td>
<td>588</td>
<td>468</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td>MAJO CALAFELL, FRANCESC D'ASSIS</td>
<td></td>
<td>49</td>
<td>41</td>
<td>48</td>
<td>39</td>
</tr>
<tr>
<td>GARRIGA POSAS, FRANCESC</td>
<td></td>
<td>191</td>
<td>101</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>MADUEÑO GARCIA DE HERREROS, ANTONIO</td>
<td></td>
<td>349</td>
<td>348</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>GRANADOS VALVERDE, OLGA</td>
<td></td>
<td>199</td>
<td>181</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>PUJADAS SERRA, CONSUELO</td>
<td></td>
<td>306</td>
<td>306</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>GARRETA GUAL, JORDI</td>
<td></td>
<td>21</td>
<td>15</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>DE ROVIRA BENACH, JUAN</td>
<td></td>
<td>164</td>
<td>143</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>TRULL LLORETA, JOSÉ</td>
<td></td>
<td>486</td>
<td>380</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>AYMERICH GARCIA, JUDITH</td>
<td></td>
<td>716</td>
<td>649</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td>BIGAS LOPEZ, NURIA</td>
<td></td>
<td>635</td>
<td>633</td>
<td>-</td>
<td>24</td>
</tr>
</tbody>
</table>
Examples: ... we can geolocate the different congresses where the university researches take part the most.

<table>
<thead>
<tr>
<th>Congress country</th>
<th>Capa de area</th>
<th>Total Congress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshop on Collateral Sanctions</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Workshop on Criminal Records</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Workshop on Databases and Corpora in Linguistics</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Workshop on Forensic Authorship Identification</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>XVIII Biennial International Conference on Infant Studies</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>“Change in political attitudes: Panels and experiments’ Workshop</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>“The Friars and their Influence in Medieval Spain”</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>+ Democracia: II Congreso de Análisis Político Crítico</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>1a Jornadas transfronterera de la recerca Empordà/Rosselló (s. VI-I a.n.e.)</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>1a Jornades Transfrontereres de la recerca arqueològica Empordà/Rosselló (s. VI-I a.n.e.)</td>
<td>0</td>
<td>200</td>
</tr>
</tbody>
</table>
Examples: ... We can show the co-authoring evolution through the years and thus we could discover if the internalization degree of the researchers of the institution advance correctly...
Examples: ...we can know the evolution of authors of my institution that publish in first quartile journals

![Bar chart showing the evolution of authors publishing in first quartile journals over the years 2014 to 2017. The chart compares the number of authors in the first quartile versus the rest of the quartiles.]
Examples: ...and as a researcher, the state of my research group. For example, the new projects of the group through the years, to determine the group evolution.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>214 - Propiedades Nanométricas de la Materia</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>321 - G-FOR (Grupo de Fotónica, Información Cuántica y Radiación y Dispersión de Ondas)</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Examples... we can track the projects discovering which researchers have more funding and in which areas, ordering by the project results impact, determining in which areas is the university, top. This can provide an evidence of where is best the university and could focus on it.
Examples… Finally, we can discover the collaboration degree between the researchers of our institution, through the researchers interactions.

We can discard the weak interactions and which researchers have more influence.
To have into account:

1) We need the **time** variable that shows the evolution, to explain what had happened

2) The information should be from **more than one institution**, providing comparatives between institutions

With all of this information we can know exactly which are the impact authors of our institution and if they are always the same every year or not. Also a researcher could track the impact evolution of his research group
Conclusions:

We have the **descriptive and diagnostic states** complete

- The research information is now FAIR in order to ease the reporting, analyses and access to the information
- Users can see what happens in the past and why it happens
- The process is continuously improved
- This is more a **service** than a product

We are now boarding the **predictive state** in order to find the facts (using data mining and machine learning technologies) to see what could happen in the future, based on what had happened
Conclusions:

We should not be afraid to face a project like this if we have the enough data maturity in the CRIS (CERIF compliant)

(data with quality, accessible and reusable enough)
Thank you so much for your attention.

Anna Guillaumet
@annaguillaumet
anna.guillaumet@sigmaaie.org