BIG DATA AS A GAME-CHANGER OF CLINICAL RESEARCH STRATEGIES

RAFAEL SAN MIGUEL

DATA SCIENTIST UNIR

DR. JAVIER GÓMEZ PAVÓN

DATA SCIENTIST HOSPITAL CENTRAL DE LA CRUZ ROJA

ORGANIZER

{ paradigma

Rafael San Miguel Carrasco
PhD Javier Gómez Pavón
PhD Beatriz Ares Castro-Conde

BIG DATA AS A GAME-CHANGER OF CLINICAL RESEARCH STRATEGIES


In the backpack:
• Having read a few books about R, SAS, Hadoop and statistics
• The feeling that healthcare would be a good place to start

How it all started

"I am not sure what you are trying to achieve."

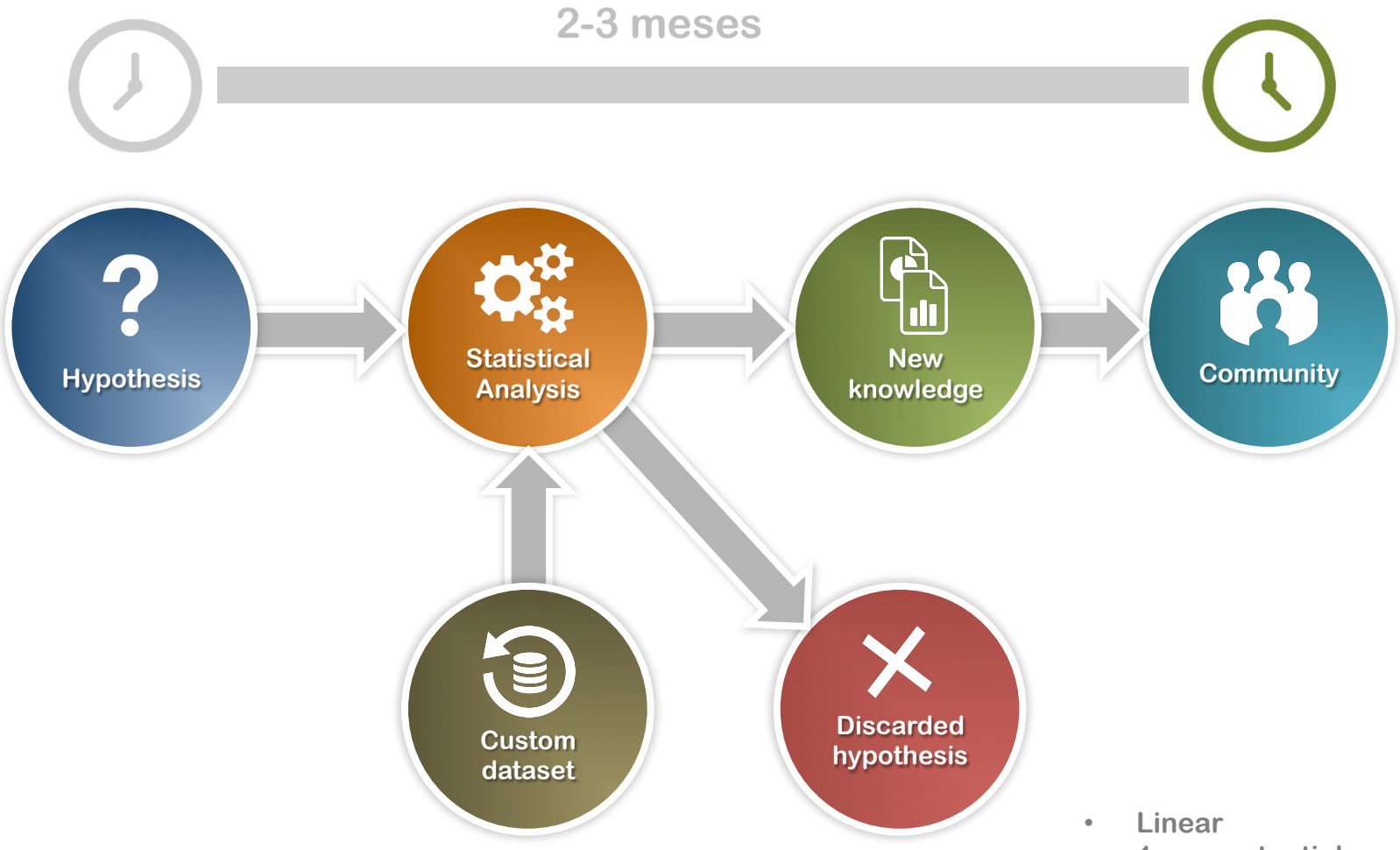
Starbucks, February 2015

How it all ended

"We should create a company to sell this as a product."

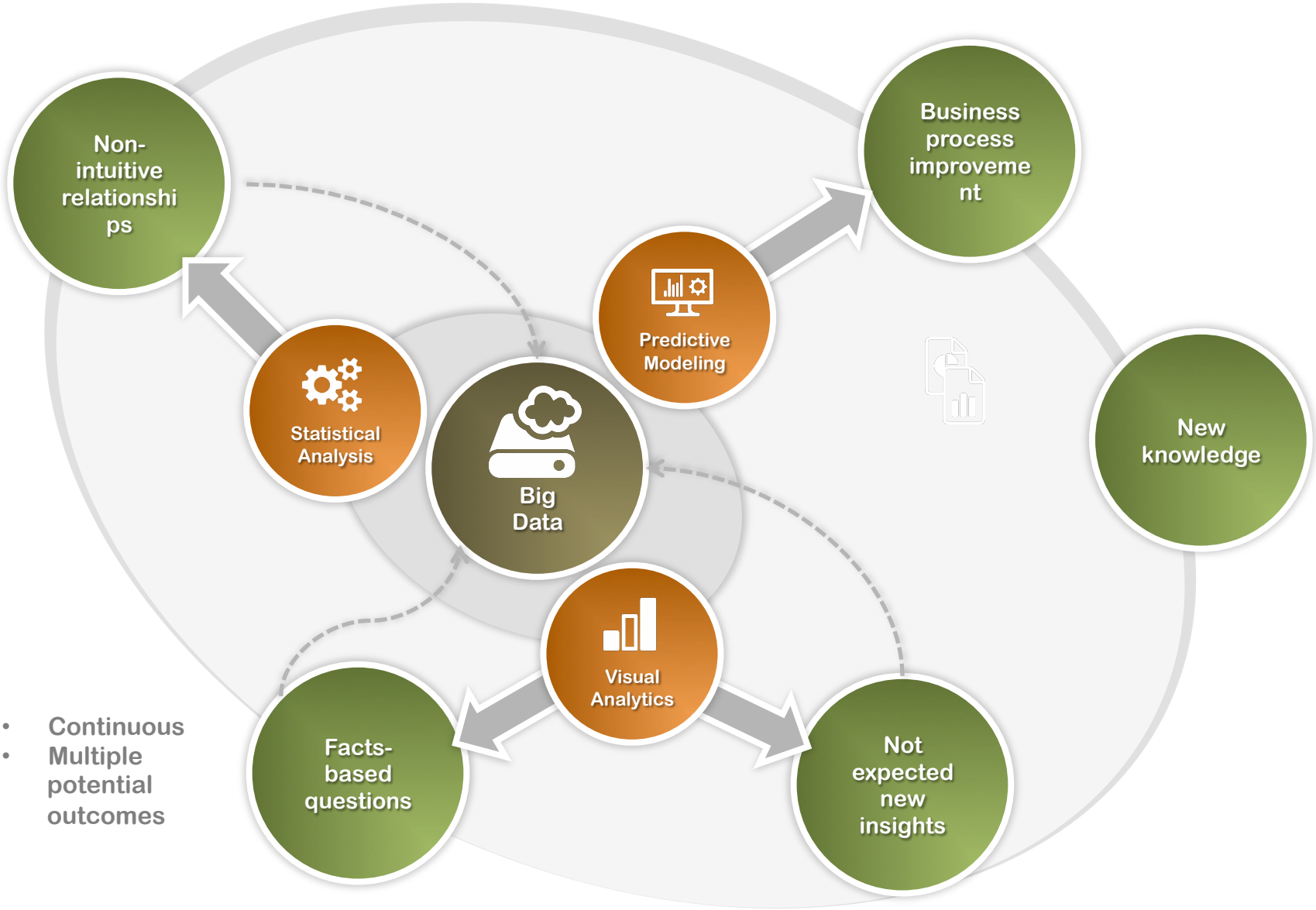
Irish Pub, September 2015

Traditional clinical research



- Linear
- 1 max potential outcome
- High prob of failure
- Long play

Big data as a game-changer



Goals

- How can big data enhance procedures used to build predictive models over traditional approaches, like hypothesis-based clinical research?
- Can big data help to measure ROI from research initiatives and programs, in terms of patients' quality of life or cost?
- Can big data produce net new knowledge for the medical community? If so, is it useful to optimize limited resources and enhance planning and forecasting processes?
- Can big data help improve prediction accuracy of clinical performance over traditional inference techniques from small samples?
- How can exploratory analysis be made available through ecosystems like Hadoop?

3 key areas

STATISTICAL ANALYSIS

Efficiency analysis of programs geared towards providing assistance to nursing homes.

PREDICTIVE MODELING

Generation of models that connect admission-related data with key target variables as length of stay, admission rate or mortality rate.

DATA VISUALIZATION

Hadoop-based visual analytics platform for domain experts to perform exploratory analysis and on-the-go clinical research.

Database

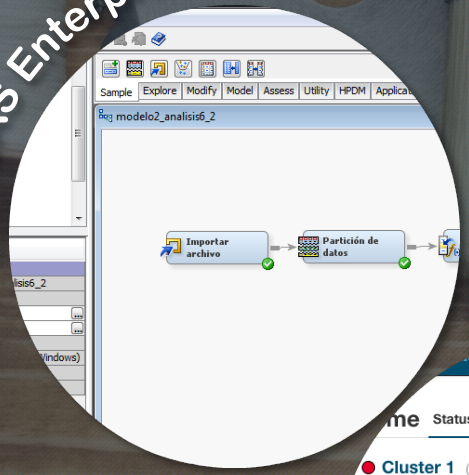
Not a sample: full database

12.000 clinical records from elderly patients admitted in Acute Unit from 2006 to today

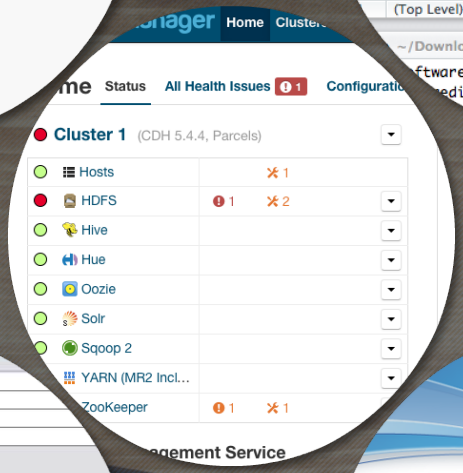
Demographics
Diagnosis
Treatment
Functional and mental status
Admission-related complications
Visits to emergency units
Lab tests
Administered drugs

Well-known tools

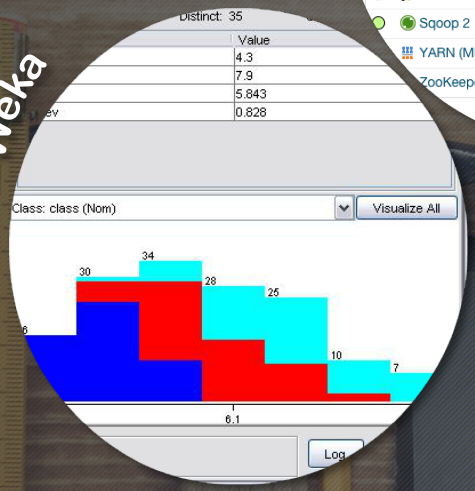
SAS Enterprise Miner



RStudio



Weka



IBM SPSS

Statistical analysis

Evaluating ROI of specialized assistance programs

Background

Programs geared towards providing assistance from a geriatrics doctor to nursing homes

- Give better care to elderly patients
- Fewer support from hospitals
- Lowering the cost to deliver healthcare

Goal: validate that these programs can provide the expected ROI

Target parameters

- CRF (functional status)
- CRM (mental status)
- Barthel index
- Number of complications (lower)
- Admissions
- Readmissions
- LOS
- Survival
- Lab tests
- Number of administered drugs

Methodology

Chi-Square test (X^2 statistic) was used to check for an statistically significant difference in key clinical variables between patients receiving specialized assistance and the control group.

```
aggregate(I_Index_CRF_previo~B_AGR, data=datos, var)
# B_AGR I_Index_CRF_previo
# 1    NO          1.374176
# 2    SI          2.086957
shapiro.test(datos$I_Index_CRF_previo)
# Shapiro-Wilk normality test
#
# data:  datos$I_Index_CRF_previo
# W = 0.8604, p-value = 1.11e-09
ggplot(datos, aes(x=I_Index_CRF_previo, fill=B_AGR)) + geom_histogram(binwidth=1, alpha=1/2)

ansari.test(I_Index_CRF_previo~B_AGR, datos)
# Ansari-Bradley test
#
# data:  I_Index_CRF_previo by B_AGR
# AB = 3519, p-value = 0.282
# alternative hypothesis: true ratio of scales is not equal to 1

t.test(I_Index_CRF_previo~B_AGR, data=datos, var.equal=TRUE)
# Two Sample t-test
```

Results

These clinical parameters were found to improve:

- CRF (functional status)
- CRM (mental status)
- Barthel index
- Number of complications (lower)

New knowledge

Elderly patients receiving specialized assistance exhibit better functional and mental status, less disability and fewer complications during admission.

This means that these patients have better quality of life, which proves that these programs provide a true return on investment.

Predictive modeling

What key clinical variables can be predicted?

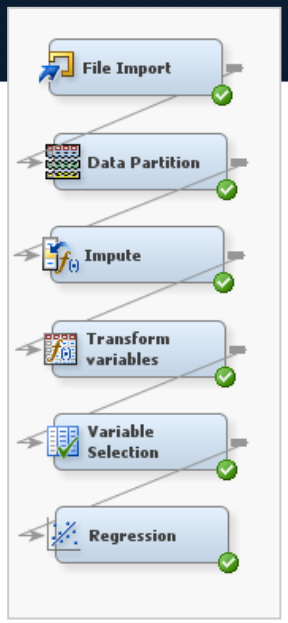
Background

Understanding what features of a nursing home lead to higher performance constitute a desirable goal for the medical community, because …

Length of stay becomes a key clinical variable after an elderly patient is admitted., because …

Patients mortality is a key clinical variable. I guess this raises no discussion …

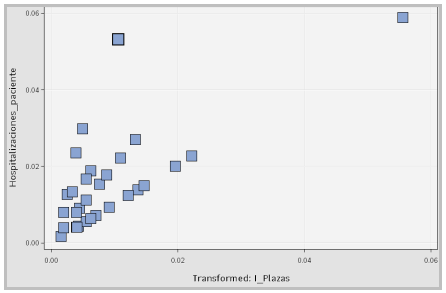
Methodology



Diagrams

Entrenamiento	
Variables	
Fórmulas	
Interacciones	
Código SAS	
Métodos predeterminados	
Inputs intervalo	Mejor
Variables objetivo tipo inter	Ninguno
Inputs de clase	Indicadores "Dummy"
Variables objetivo tipo clase	Ninguno
Tratar ausentes como Nivel	No
Propiedades de la muestra	
Método	Primer N
Tamaño	Predeterminado
Semilla aleatoria	12345
Agrupamiento óptimo	
Número de clases	4
Valores ausentes	Utilizar en Búsqueda
Método de agrupación	
Valor de corte	0.1
Agrupar ausentes	No
Número de clases	Variables
Añade un valor mínimo al va	Sí
Valor de compensación	1
Puntuación	
Utilizar Meta transformació	Sí
Ocultar	Sí
Rechazar	Sí
Informe	
Estadísticos de sumarizació	Sí

Transformation



Visual inspection

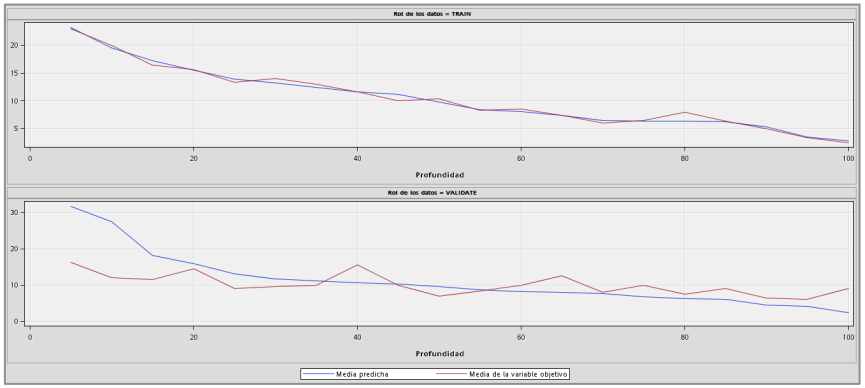
Model Fit Statistics			
R-Square	*0.9706*	*Adj R-Sq*	*0.9523*
AIC	*35.6839*	*BIC*	*1.5073*
SBC	*77.0600*	*C(p)*	*298.3080*

Model fit indicators

Data role=VALIDATE Target variable=I_Length_of_stay Target label=I_Length_of_stay

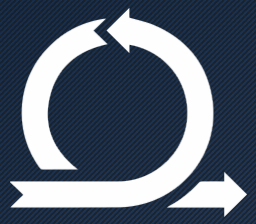
Depth	Observations	Mean of target variable	Mean of predicted value
5	3	16.3333	31.8414
10	2	12.0000	27.5875
15	2	11.5000	18.1360
20	2	14.5000	15.8958
25	2	9.0000	13.0843
30	2	9.5000	11.7780
35	2	10.0000	11.1453
40	2	15.5000	10.6869

Quantile-based comparison



Agnostic approach

No prior questions or hypothesis for the analysis
No focus on particular input variables



… and multiple iterations/configurations to come up with as many results as possible

ITERATION 1

Mortality can be fully predicted through another variable:

Place of Exitus

ITERATION 2

Mortality can be predicted through another variable:

Morphine

ITERATION 3

Mortality can be predicted through several variables:

Digoxin

Number of lab tests

Occurrence of pressure ulcers

Non-intuitive relationships

Event classification table

Data role=TRAIN Target=B_Exitus_Horus Target label=B_Exitus Horus

False negative	*True negative*	*False positive*	*True positive*
8	35	6	24

Data role=VALIDATE Target=B_Exitus_Horus Target Label=B_Exitus Horus

False negative	*True negative*	*False positive*	*True positive*
3	11	.	6

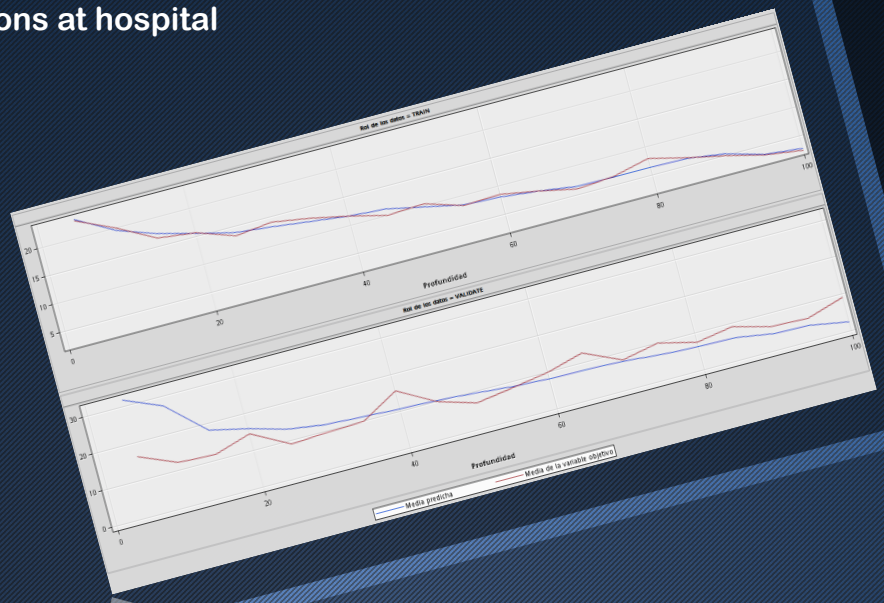
Adverse reactions to drugs leading to higher mortality rate

We can predict length of stay

The resulting model was found to be statistically signficant, accounting for 95,23% (R-Square) of the target variable's variance.

LOS can be predicted through:

Previous number of admissions
Gender
Diagnoses as acute kidney failure, respiratory infection and acute bronchitis
Barthel index prior to admission
Falls
Total amount of administered drugs
Need for urinary catheter
Infections at hospital



Business process improve ment

Business process improve ment

$$
\begin{aligned}
Admission\_rate \\
&= 0.0419 + (INV\_I\_PBeds \times 0.9481) \\
&+ (TI\_N\_Nursinghome\_city1 \times -0.0104) \\
&+ (TI\_N\_Nursinghome\_city15 \times -0.0194) \\
&+ (TI\_N\_Nursinghome\_city3 \times -0.00771)
\end{aligned}
$$

The transformed variable that represents the inverse of the number of beds in the nursing home can accurately predict the admission rate from that nursing home.

The model can explain up to 86% of the target's variance.

A Geriatric Unit can accurately forecast the number of admissions from currently served nursing homes, and make better choices with regards to new nursing homes to be served in the future.

We can predict admission rate

Data visualization

Playing with datasets and finding new insights on-the-go

Background

Providing domain experts with an effective tool to discover patterns or relationships among data variables through exploratory analysis and visual inspection.

A tool to go beyond reported findings and discover new insights in the datasets.

Using Hadoop to ensure that it could scale out to process millions of clinical records with literally no changes to the current architecture.

Architecture

CDH Cloudera Hadoop

HUE Hadoop User Experience

Apache Solr

So, how does it look like?

What are the defining characteristics of a patient with a cardiac insufficiency?

- CIE-9 value 428 is selected from the diagnosis list
- Charts and indicators related to gender, age, year of admission, CRF, CRM and Barthel index are updated and displayed.

New knowledge

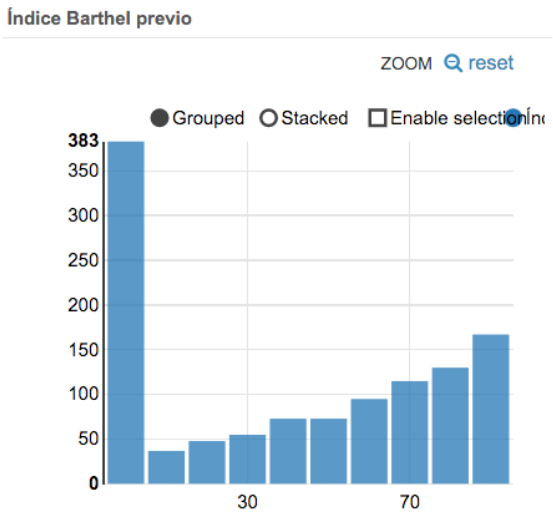
Sexo

Mujer (978)
Varón (300)
No identificado (1)

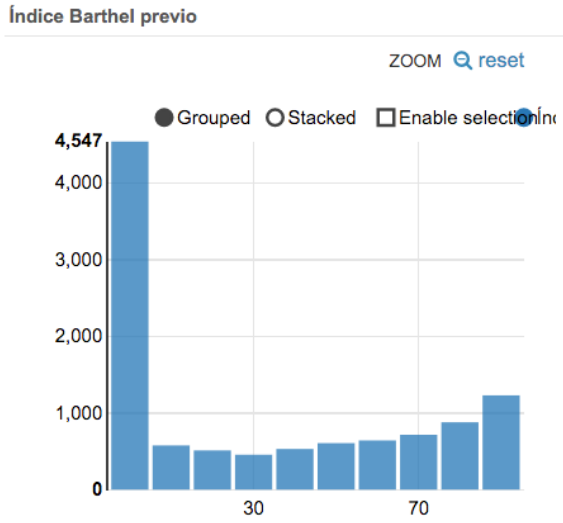
Edad

50 - 55 (0)
55 - 60 (0)
60 - 65 (1)
65 - 70 (0)
70 - 75 (17)
75 - 80 (83)
80 - 85 (275)
85 - 90 (439)
90 - 95 (320)
95 - 100 (128)
100 - 105 (11)
105 - 110 (4)
110 - 115 (1)

Diagnosis 428



All diagnoses

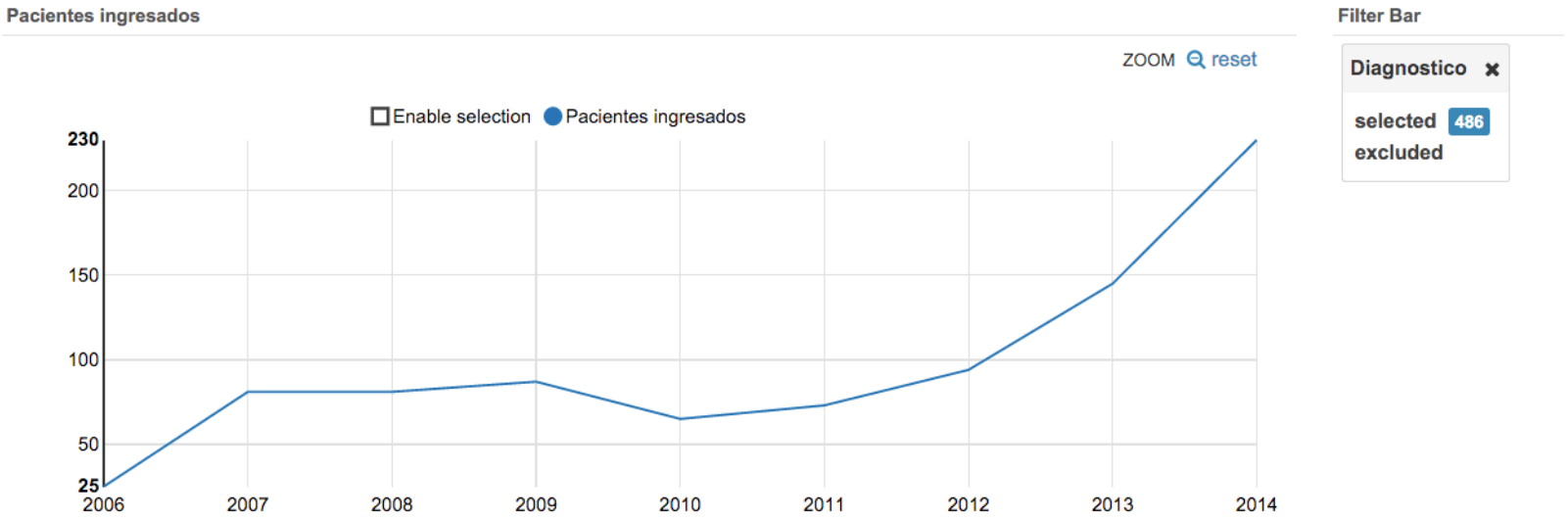


Most frequent profile is a 85-95 years old patient, woman, with Barthel index higher than average.

Which trends are observed in admissions related to pneumonia?

- CIE-9 value 486 is selected from the diagnosis list. The chart displaying yearly number of admissions is updated.

Not expected new insights

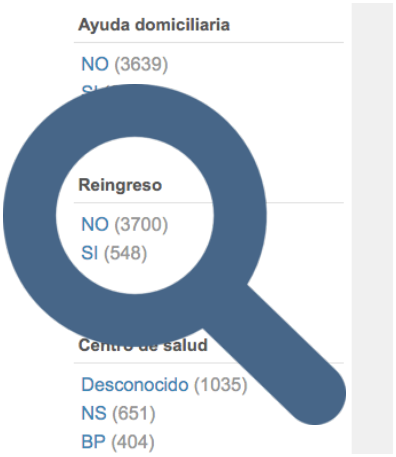


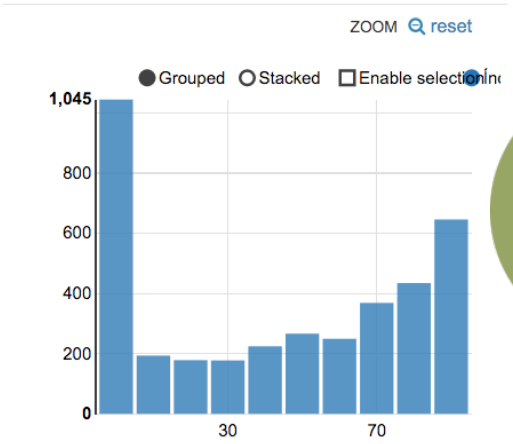
The number of admissions related to pneumonia was 81 in 2007 and 230 in 2014, which means that it has increased by 4 times in the last 7 years.

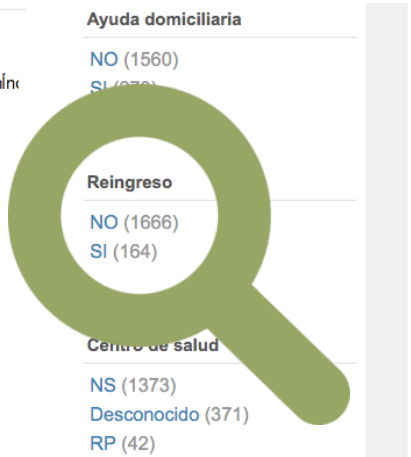
Are there variations in readmission rates for patients from Hospital and Hospital B?

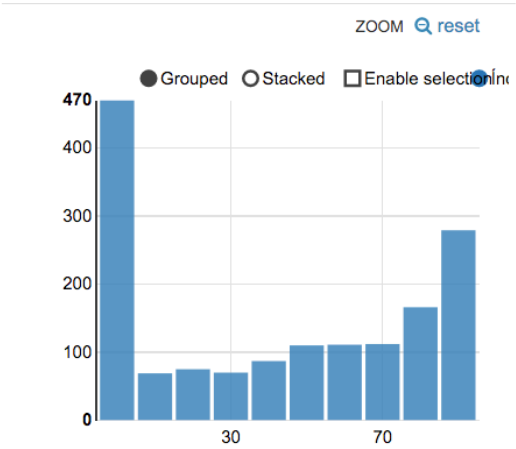
- Hospital A is selected from the Source of Admission list. The number of readmissions is 548 from a total of 4.248; therefore, the readmission rate is 12%
- For Hospital B, the number of readmissions is 164 from a total of 1.830, so the readmission rate is only 8%.

New knowledge







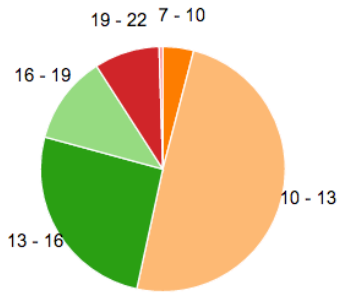


Therefore, readmission rate for patients from Hospital A seems to be higher than from Hospital B.

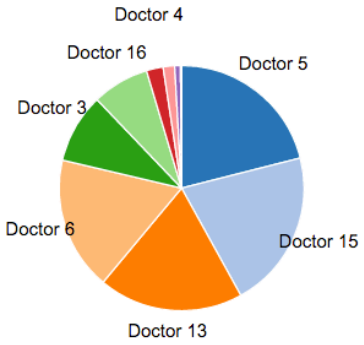
How was workload distributed among doctors in the Acute Unit in 2014?

- 2014 is selected from the list of years

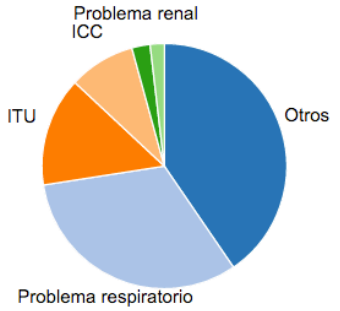
Hora



Doctor



Motivo ingreso



Most patients are admitted by noon

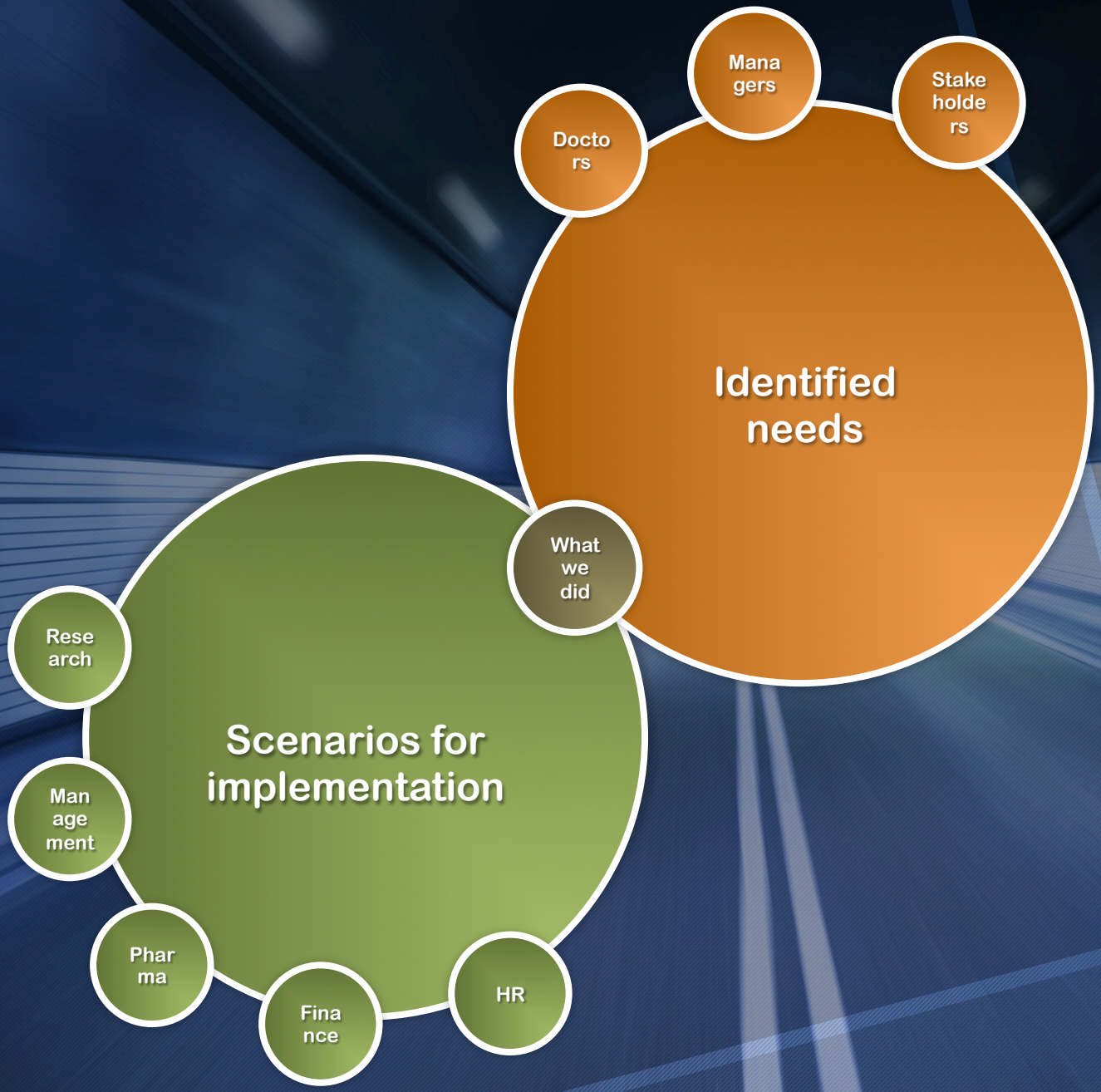
New knowledge

Doctors' workload requires further review

Not expected new insights

Doctors should specialize in ITU and respiratory diseases

Business process improvement



Looking into the future


Takeaways
1
Big data analytics can be more efficient than hypothesis-based research
2
Exploratory analysis is a must when it comes to discover net new knowledge
3
Predictive modeling is geared towards business process improvement
4
Extracting insights shouldn't be a one-shot activity, but a continuous process

Thanks!

Keep calm … and send us feedback

rsanmcar@gmail.com