

XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

19-21 de Noviembre de 2014

Auditorio y Centro de Congresos

Víctor Villegas. Murcia





XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

SIMULACIÓN Y BIG DATA: UNA MANERA DE DESCUBRIR CONOCIMIENTO EN LOS SERVICIOS DE URGENCIAS

Emilio Luque

emilio.luque@uab.es

**High Performance Computing for
Efficient Applications and Simulation (HPC4EAS)**

**Computer Architecture and Operating Systems (CAOS) Department
Universitat Autònoma de Barcelona (UAB)**



High Performance Computing FOR
Efficient Applications and Simulation



UAB
Universitat Autònoma
de Barcelona



XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

High Performance Computing for Efficient Applications and Simulation



Staff Members (UAB)

Dr. Emilio Luque (Professor)
Dr. Dolores Rexáchs (Ass. Prof)
Dr. Remo Suppi (Ass. Prof)
Dr. Daniel Franco (Ass. Prof)
Dr. Elisa Heymann (Ass. Prof)
Dr. Francisco Epelde (MD-Tauli Hospital)

Postdoc Researchers (UAB)

Dr. Álvaro Wong
Dr. Manel Taboada
Dr. Marcela Castro

Postdoc researchers (External Collaborations)

Dr. Gonzalo Zarza
Dr. Ronal Muresano
Dr. Roberto Solar
Dr. Carlos Núñez
Dr. Sandra Méndez
Dr. Leonardo Fialho
Dr. Eduardo C. Cabrera
Dr. Cristian Tissera (UNSL)

PhD Students (UAB):

- Javier Panadero
- Francisco Borges
- Albert Gutierrez
- Eva Bruballa
- Cecilia Jaramillo
- Joe Carrión
- Zhengchun Liu
- Jorge Villamayor
- Laura Espínola

PhD Students (Foreing Institutions Collaborations):

- Diego Montezanti (UNLP)
- Silvana Lis Gallo (UNLP)
- Diego Encinas (UNLP)
- Adriana Gaudiani (UNGS)



Universitat
Autònoma
de Barcelona

Corporació
Parc Taulí



<http://grupsderecerca.uab.cat/hpc4eas/>



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

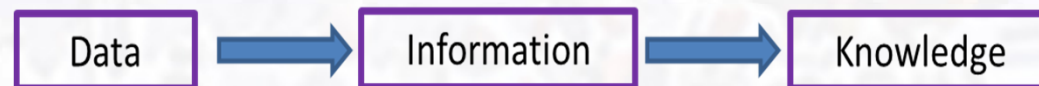


19-21 Noviembre 2014

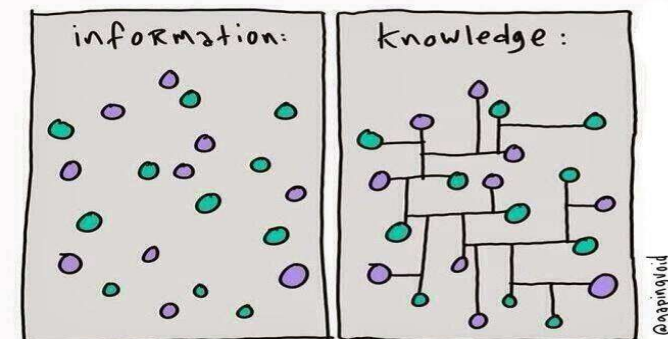
Auditorio y Centro de Congresos Víctor Villegas
Murcia

LA EVOLUCIÓN DE LOS PARADIGMAS CIENTÍFICOS

- **Ciencia Empírica:** Descripción. Sin capacidad de predecir o muy limitada. Sin posibilidad de generalizar
- **La Teoría en la Ciencia:** Elaboración y verificación de Modelos Capacidad de generalización y Predicción.
- **Ciencia Computacional (Large-Scale Computational Simulation):** Más allá del mundo “físico” con la “colaboración” del computador.
- **Data management and Exploitation for Knowledge Extraction (e Science):** Work together with Theory, Experiments and Simulation



Data-Intensive Scientific Discovery





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Source: The Fourth Paradigm: Data-Intensive Scientific Discovery.

Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Research Redmond, Washington



XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

La Simulación



XXXV

Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

La Simulación

Definición de **R. E. Shannon**:

"La simulación es el proceso de diseñar un **modelo de un sistema real** y llevar a cabo experimentos con él, a fin de **comprender el comportamiento del sistema** y evaluar nuevas estrategias - dentro de los límites impuestos **por un criterio o conjunto de ellos** - **para operar el sistema** "

La **simulación** es la realización de experimentos utilizando un modelo que representa la realidad y donde se aplica una hipótesis o un conjunto de hipótesis de trabajo.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

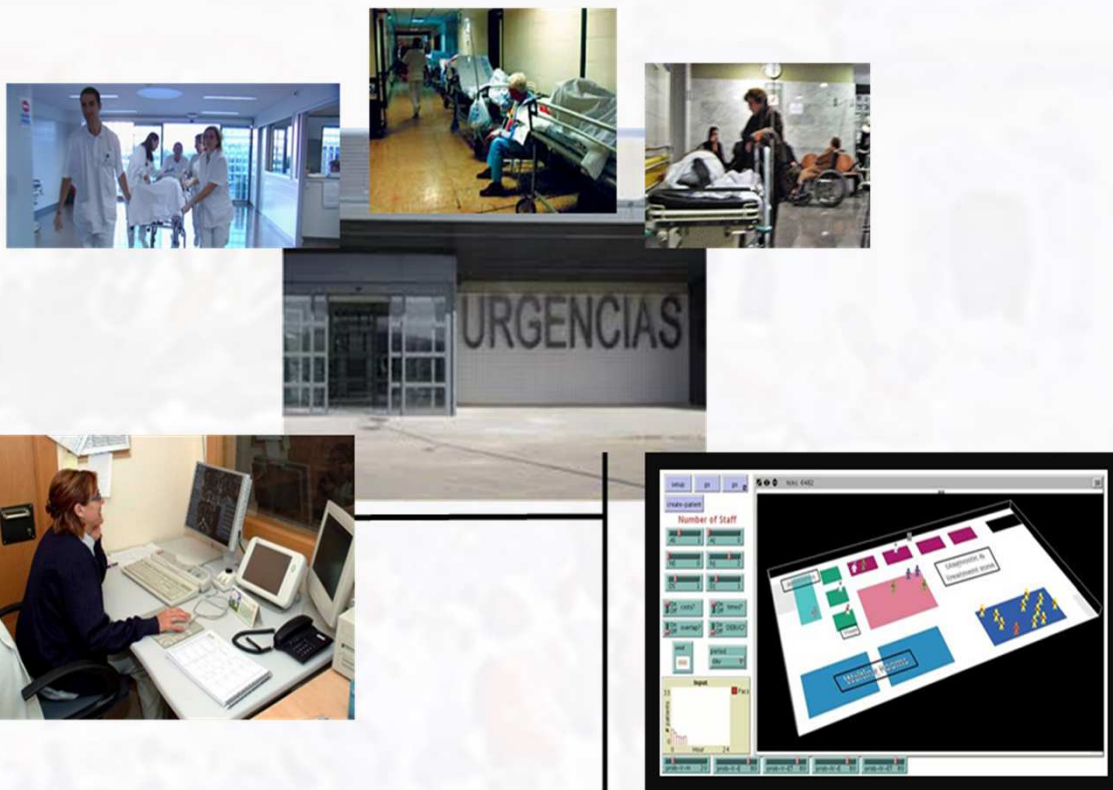
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

SIMULACIÓN: ¿QUÉ PASARÍA SI...?



Con la simulación *confiamos en que podemos ir más allá. Siempre la esperanza que se tiene es que un simulador, una vez sintonizado es capaz de ir más allá de puramente hacer un mímico de la realidad, porque si sólo hace el mímico de la realidad, sólo la repite, significa que la comprendemos pero que no podemos afirmar nada sobre ella que no podamos ver antes.*



XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

Simulation of Healthcare Emergency Departments and Nosocomial Infection

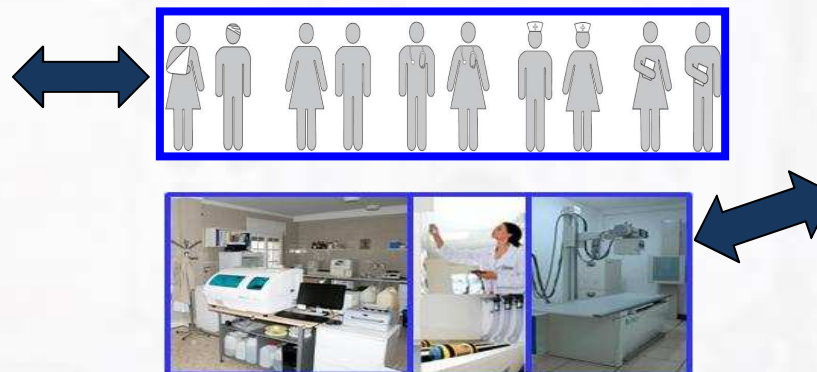


CARACTERÍSTICAS DEL MODELO CONCEPTUAL

1) Agentes Activos

Pacientes
 Acompañantes de pacientes
 Personal de Admisión
 Técnicos Sanitarios
 Enfermeras (Triage, Asistenc.)
 Médicos (Emergencias, especialistas)

Tipos de Agentes



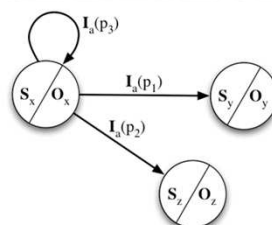
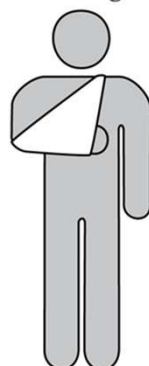
2) Agentes Pasivos

Laboratorios
 Hospital
 Sistema informático
 Sistema de megafonía
 Tubos neumáticos

Máquinas de Estado de Moore



Patient agent



Current state / Output	Input	Next state / Output
....		
....		
Sx / Ox	I ₁ (p1)	Sy / Oy
Sx / Ox	I ₂ (p2)	Sz / Oz
Sx / Ox	I ₃ (p3)	Sx / Ox
....		
....		

State variables

STATE Variables	Values	Observability
Name/identific <id>	Unique per agent	I
Personal details	Gender, Medical history (cardiology, pulmonology, neurological...); Allergies (yes-no); Treatments that received (classified into therapeutic groups: bronchodilators, vasodilators, etc.); Origin (national or immigrant)	I
Location	Entrance, Admissions, Waiting Room, Triage, Treatment Zone	E
Action	Idle, Requesting information from <id>, Giving information to <id>, Searching, Moving to <location>, Waiting for ambulance	E
Physical condition	Healthy; Hemodynamic-Constant; Barthel Index (degree of dependence)	E/I/N
Symptoms (patients)	Healthy, Cardiac/respiratory arrest, severe/moderate trauma, headache, vomiting, diarrhea	E/I
Communication skills	Low, Medium, High	E
Level of experience (doctors)	Resident (1 to 5); Junior (5-10); Senior (10- 15) and Consultant (over 15 years)	E/I
Level of experience (triage nurses)	Low, Medium, High	E/I
Level of experience (emergency nurses)	Low, Medium, High	E/I
Level of experience (admissions)	Low, Medium, High	E/I



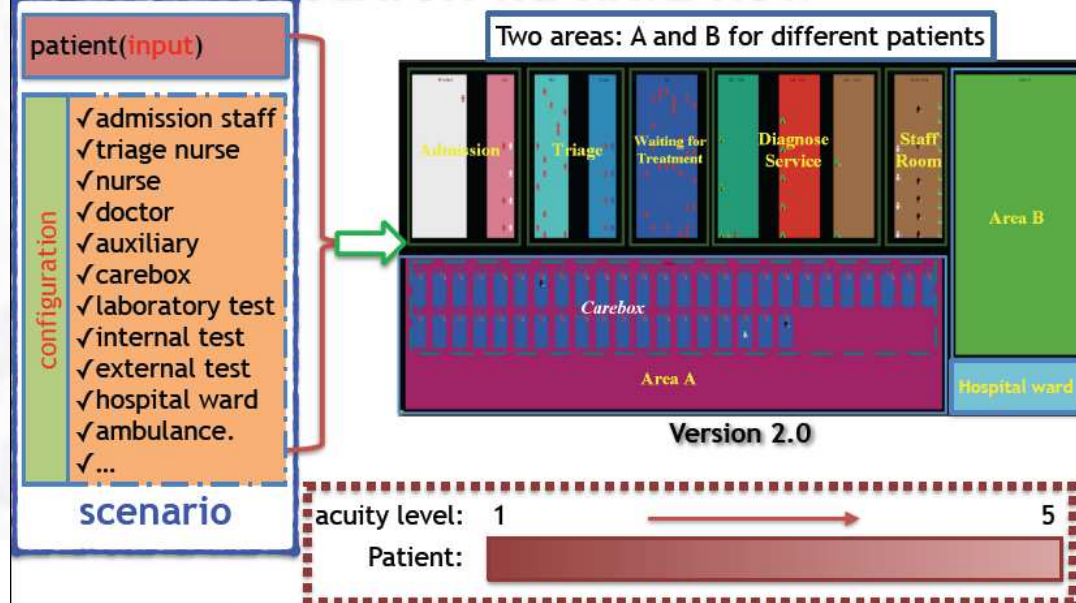
XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

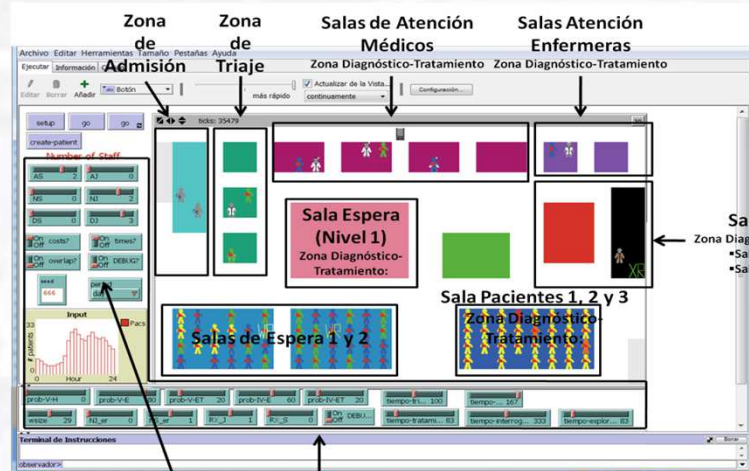


19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

THE ED SIMULATOR WE HAVE NOW



The simulator is not dedicated for any specific ED, that can simulate other EDs who works under the similar healthcare policy by tuning process (adaptable).



Consola de Configuración de Parámetros



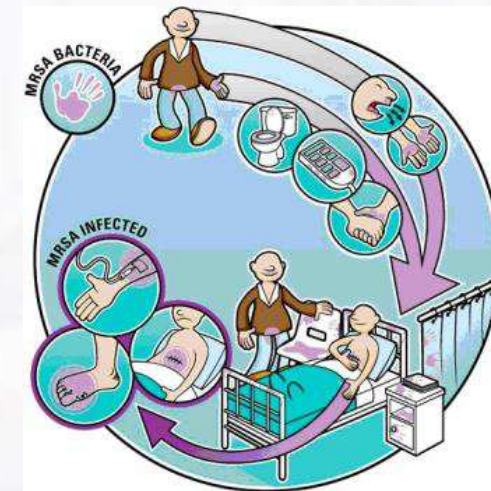
Nosocomial Infection Propagation

- It is a kind of infection caused by microorganisms and occur in healthcare environments.

What are the consequences?

- Increase the length of stay (LoS).
- Need specific treatment (MRSA is resistant to the common antibiotics).
- Increase mortality and morbidity.

Why in the Hospital Emergency Service?



Because it's the majority entrance door for patients to the Hospital.
It is one of the most complex units.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

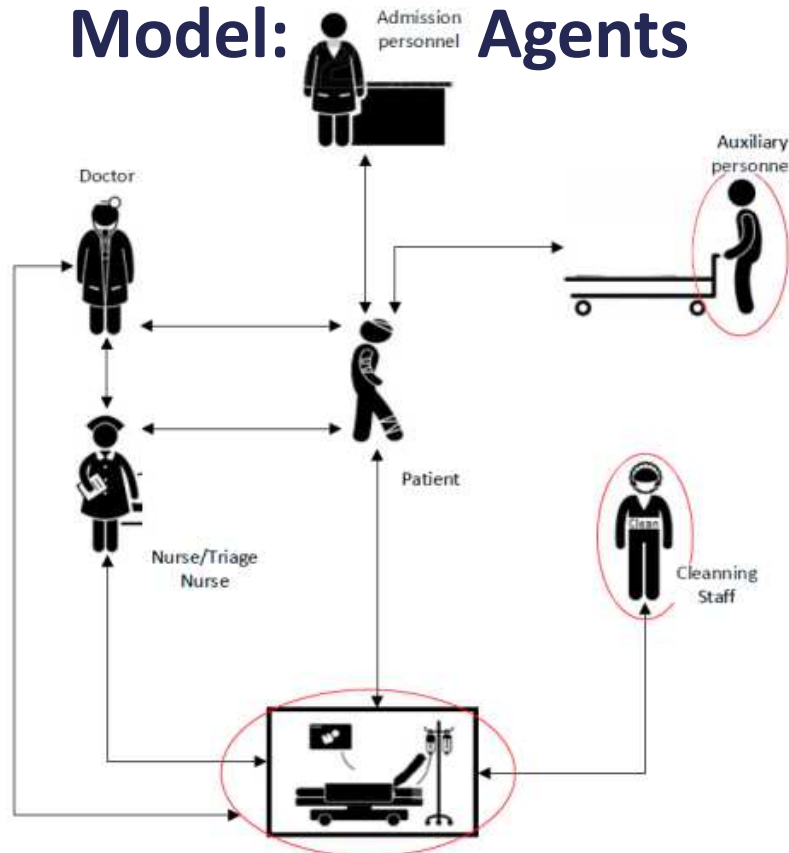


19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Objetives

Contact Propagation Model: Agents



Create a contact propagation model of Methicillin-Resistant Staphylococcus Aureus (MRSA) in a hospital emergency department by using agent-based modeling and simulation technique (ABMS).

ED-Simulator

Admission personnel

Patient

Doctors (S, J)

Nurses(S, J)

Auxiliary Personnel

Cleaning Staff

AA

Information System

CareBox

PA



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

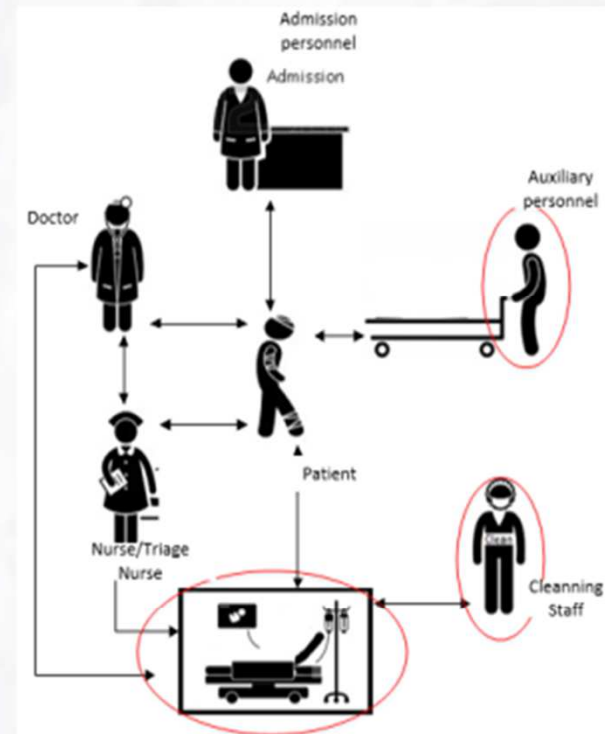
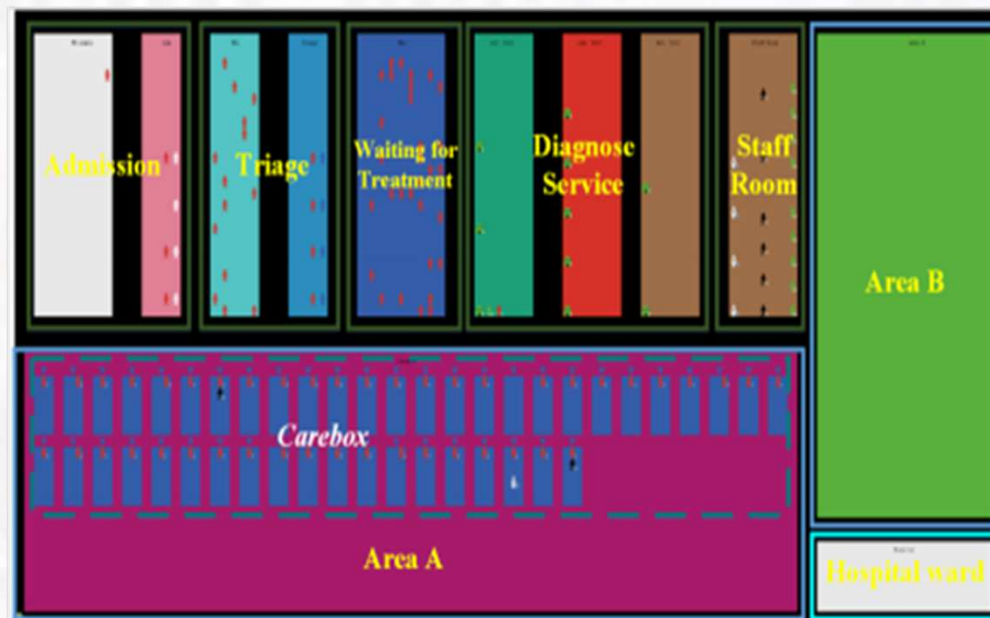
IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia



Five acuity levels divided in 2 areas: A and B.

Physical contact model.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

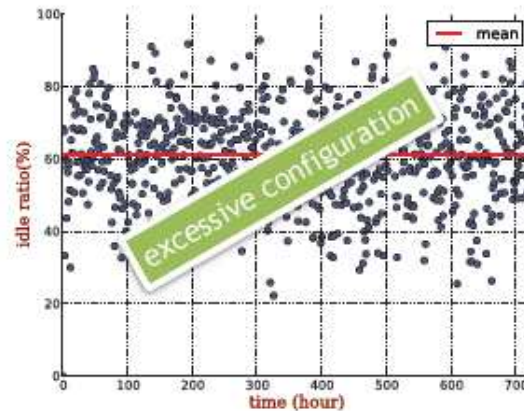


19-21 Noviembre 2014

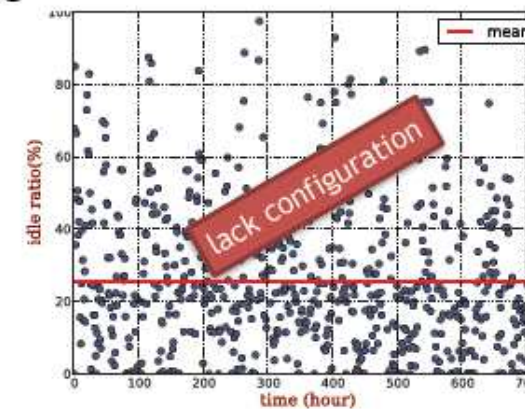
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Workload of physical resource

Laboratory test

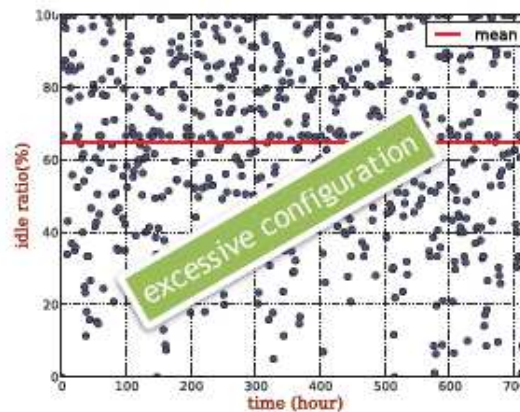


Internal test

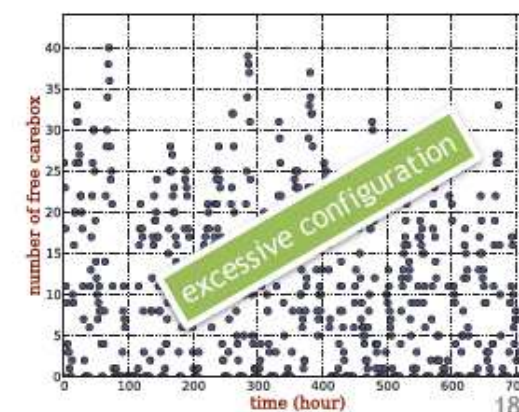


$$idle_rate = \frac{\sum_{i=0}^n length_of_idle_time}{n \times length_of_working_time} \times 100\%$$

External test



Carebox



Efficiency



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

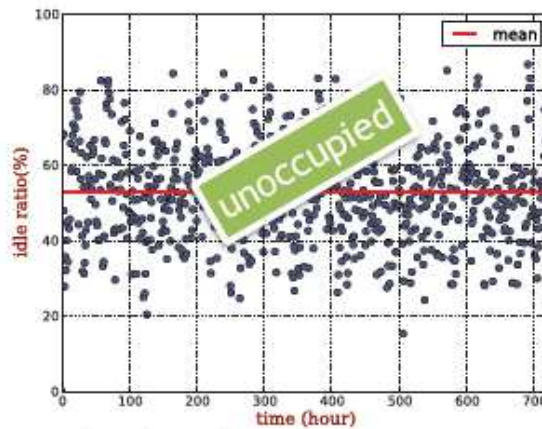
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

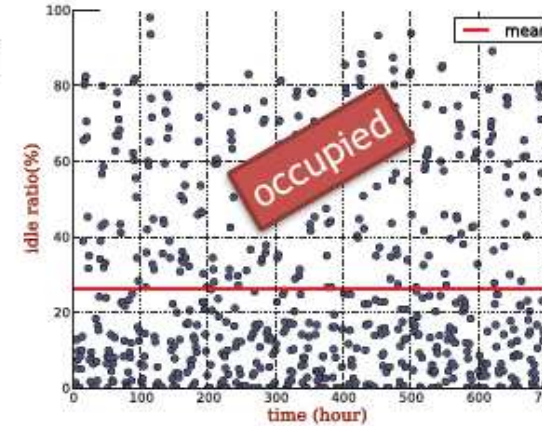
Nurse in area A



n = number of nurse in area A

Workload of sanitary staff

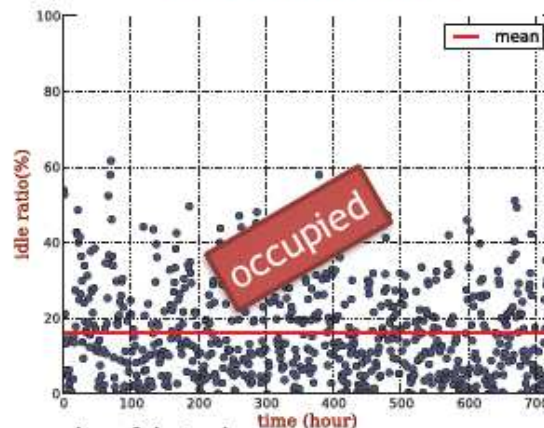
Nurse in area B



n = number of nurse in area B

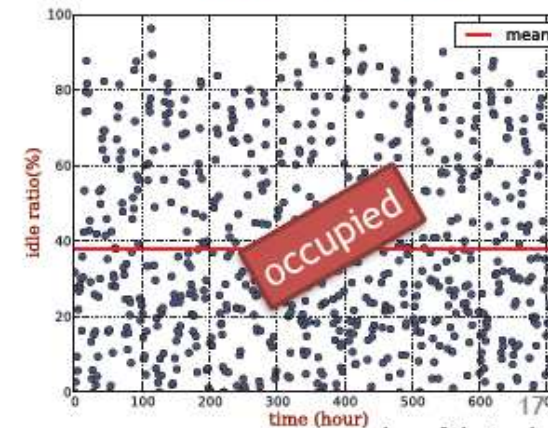
$$idle_rate = \frac{\sum_{i=0}^n length_of_idle_time}{n \times length_of_working_time} \times 100\%$$

Doctor in area A



n = number of doctor in area A

Doctor in area B



n = number of doctor in area B

Efficiency



XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

Big Data



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Introduction to Big Data

What is Big Data?

What makes data, "Big" Data?

understanding the data deluge: comparison of scale with physical objects

1 megabyte

(A large novel)



x 1000

1 gigabyte

(Information in the human genome)



Height of a short person

x 1000

1 terabyte

(Annual world literature production)



Length of the Auckland Harbour Bridge

x 1000

1 petabyte

(All US academic research libraries)



Length of New Zealand

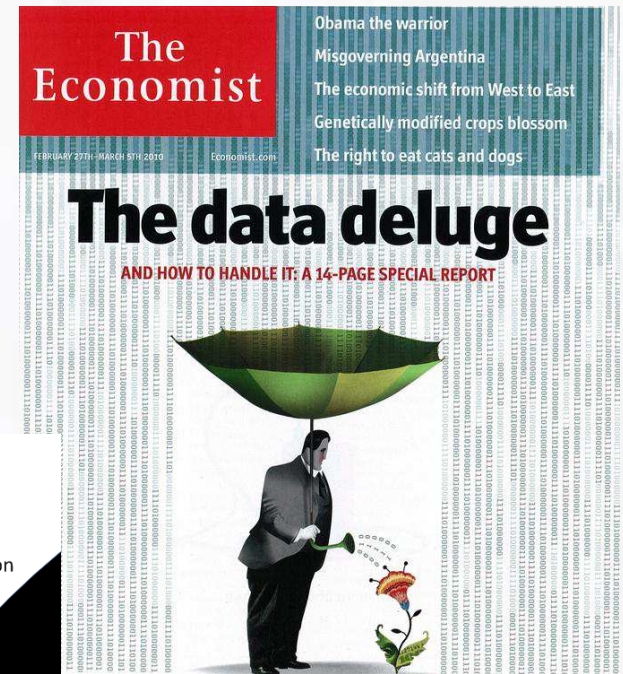
x 1000

1 exabyte

(Two thirds of annual production of information)



Diameter of the Sun





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Big Data Definition

No single standard definition...

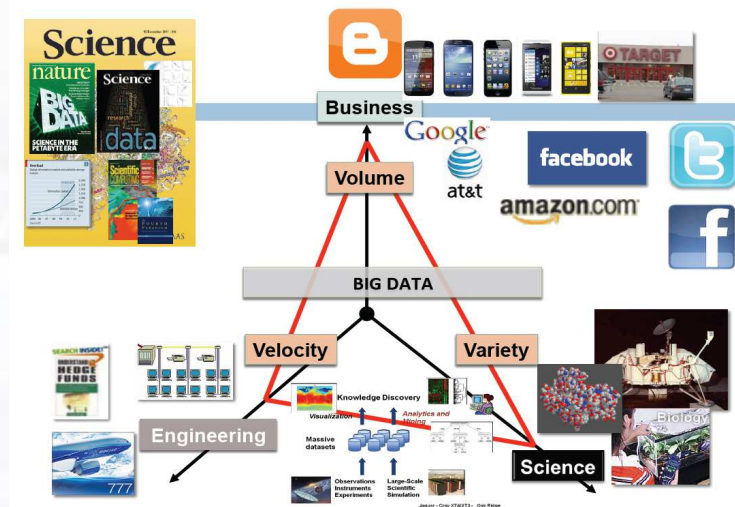
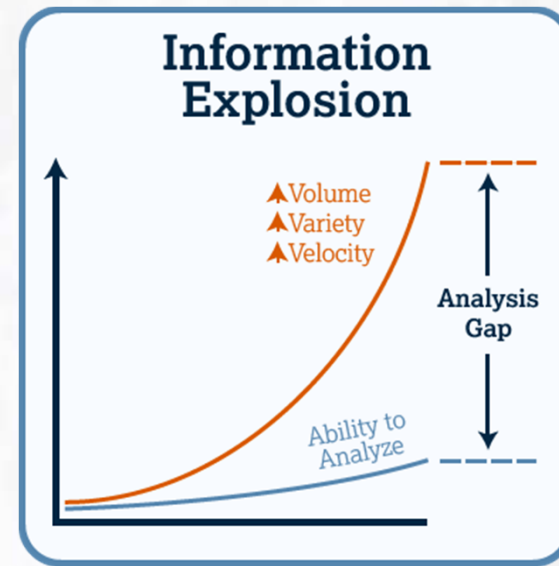
“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...

Big Data is any thing which is crash Excel.

Small Data is when is fit in RAM. Big Data is when is crash because is not fit in RAM.



Or, in other words, Big Data is data in volumes too great to process by traditional methods.





XXXV

Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

The promise of Big Data

- Data contains information of great business value
- If you can extract those insights you can make far better decisions
- ...but is data really that valuable?



XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



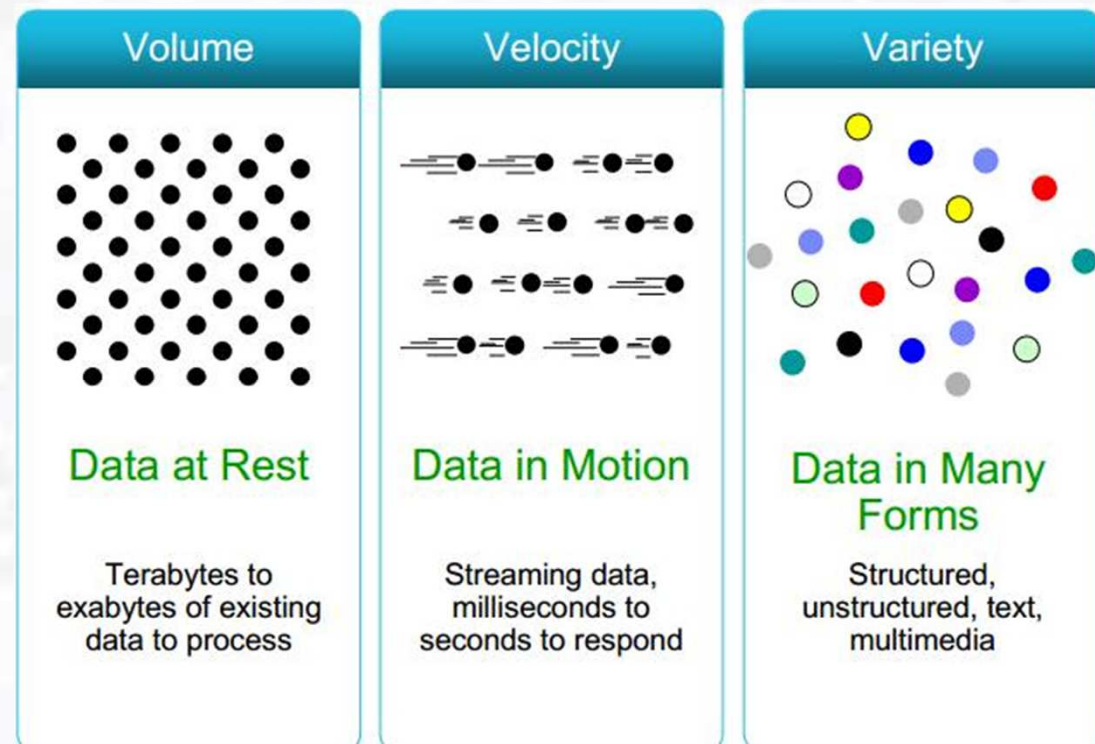
19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

From WWW to VVV (3V's)

- **Volume**
 - data volumes are becoming unmanageable

- **Variety**
 - data complexity is growing
 - more types of data captured than previously

- **Velocity**
 - some data is arriving so rapidly that it must either be processed instantly, or lost
 - this is a whole subfield called "stream processing"





XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Some Make it 4V's

*"The **greatest** enemy of knowledge is not ignorance, it is the illusion of knowledge."*

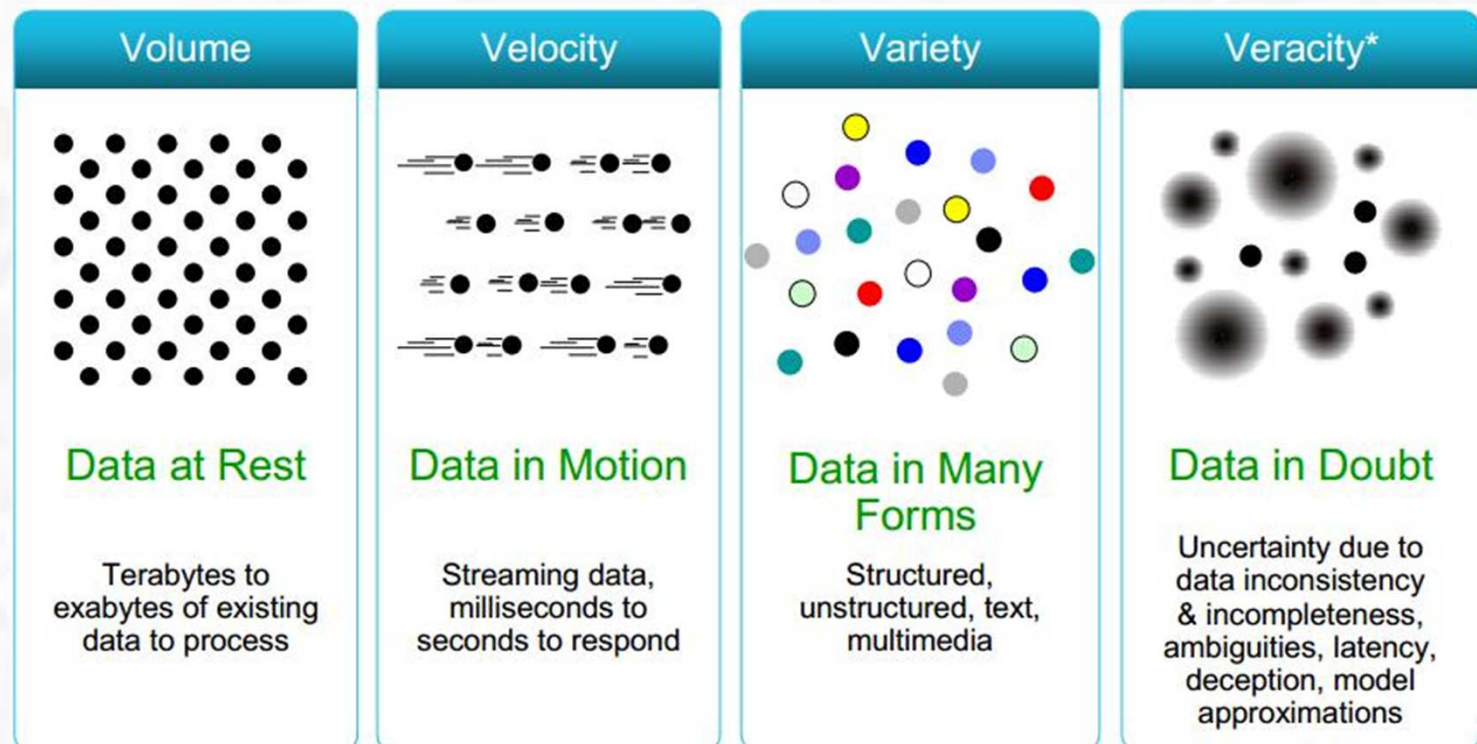
- A huge problem in practice
 - any manually entered data is suspect
 - most data sets are in practice deeply problematic

- Even automatically gathered data can be a problem

- systematic problems with sensors
- errors causing data loss
- incorrect metadata about the sensor

- **Never, never, never** trust the data without checking it!

- garbage in, garbage out, etc





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

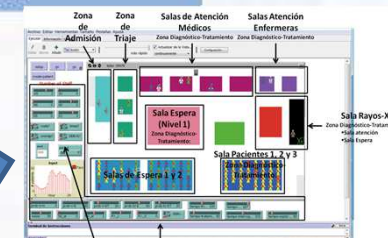


19-21 Noviembre 2014

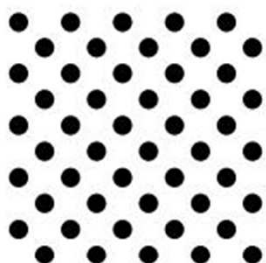
Auditorio y Centro de Congresos Víctor Villegas Murcia

Can we include 5V's?

Inclusion of the 5th V: Virtual (simulation generated) data



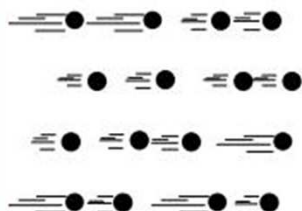
Volume



Data at Rest

Terabytes to exabytes of existing data to process

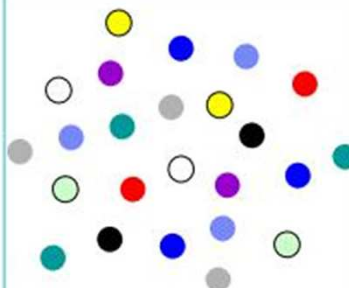
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

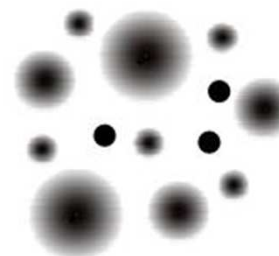
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

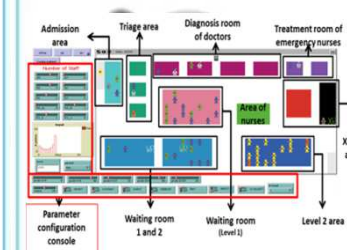
Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Virtual



Data from "Virtual Worlds"

Simulation as a sensor of the real world



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

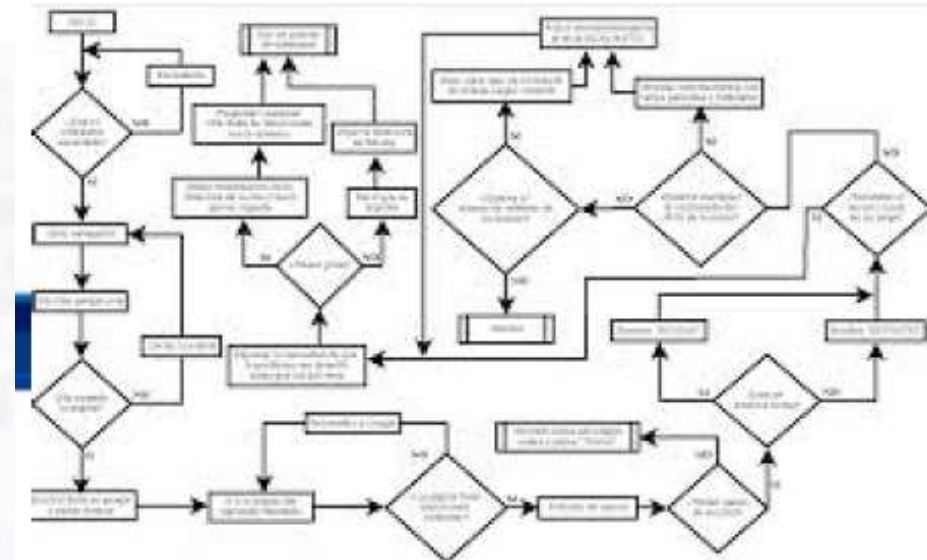
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Algunos algoritmos de minería de datos





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

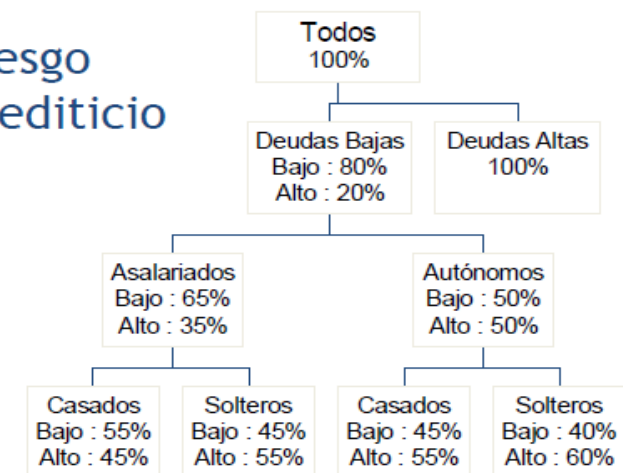
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Arboles de decisión

Arboles de Decisión (modelos predictivos)

- Son útiles para predecir el valor de ciertos atributos a partir de otros dados
- Se realiza una clasificación, empezando por los atributos más discriminantes hasta terminar por los menos discriminantes.
- De esta manera es posible establecer una serie de reglas sobre qué factores (atributos) afectan a cada caso del modelo, y tener una base a partir de la cual estimar como responderán otros casos en el futuro (predicción)

Riesgo Crediticio





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas

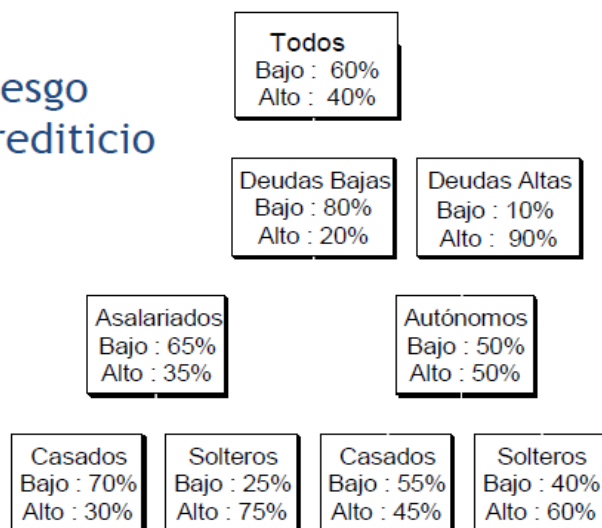
Murcia

Arboles de decisión

Árboles de Decisión

También se utilizan cuando se quiere obtener un modelo para predecir una variable a partir de otras variables, pero a diferencia de las regresiones, con el árbol obtenemos una clasificación de grupos, obteniendo una serie de reglas.

Riesgo
Crediticio



Ejemplo de lectura:

Alto riesgo crediticio, clientes con:

- Deudas altas
- Deudas bajas, asalariados y solteros
- Deudas bajas, autónomos y solteros

Bajo riesgo crediticio, clientes con:

- Deudas bajas, asalariados y casados
- Deudas bajas, autónomos y solteros



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas

Murcia

Regresiones

Regresiones

Se utilizan cuando se quiere obtener un modelo para predecir una variable (variable dependiente) a partir de otras variables (variables independientes).

Como resultado del modelo tenemos una ponderación o coeficientes para cada una de las variables y se consigue dar una puntuación a cada uno de los registros de la base de datos.

FACTOR CLAVE	Coeficiente
Edad	
<= 30 años	150
31 - 40 años	100
41 - 50 años	50
> 50 años	0
Antigüedad	
< 1 año	120
1-2 años	100
2-3 años	40
> 3 años	0
Consumo mes anterior a la fecha corriente	
Bajo	200
Medio	100
Alto	0
Número de incidentes de pago	
2 o más	100
1	60
0	0

Ejemplo de lectura: una persona de 45 años con 4 años de antigüedad, un consumo medio durante el mes anterior y sin incidentes de pago obtendrá la puntuación global de:

$$50 + 0 + 100 + 0 = 150$$



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

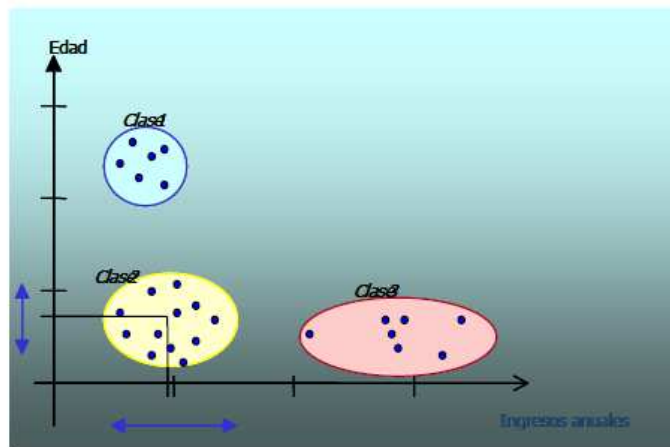
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Clustering

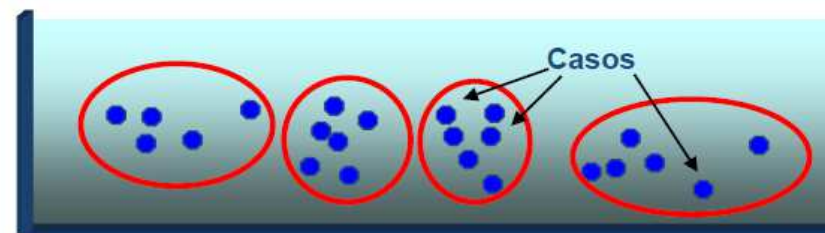
Clustering

Se utiliza para hallar agrupaciones de registros de acuerdo a características comunes, obteniendo grupos de registros con comportamiento o características parecidas entre ellos y comportamiento o características diferenciadas con los otros grupos..

El *clustering* es útil especialmente para realizar tareas de Marketing tales como segmentación de clientes.



Clustering





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

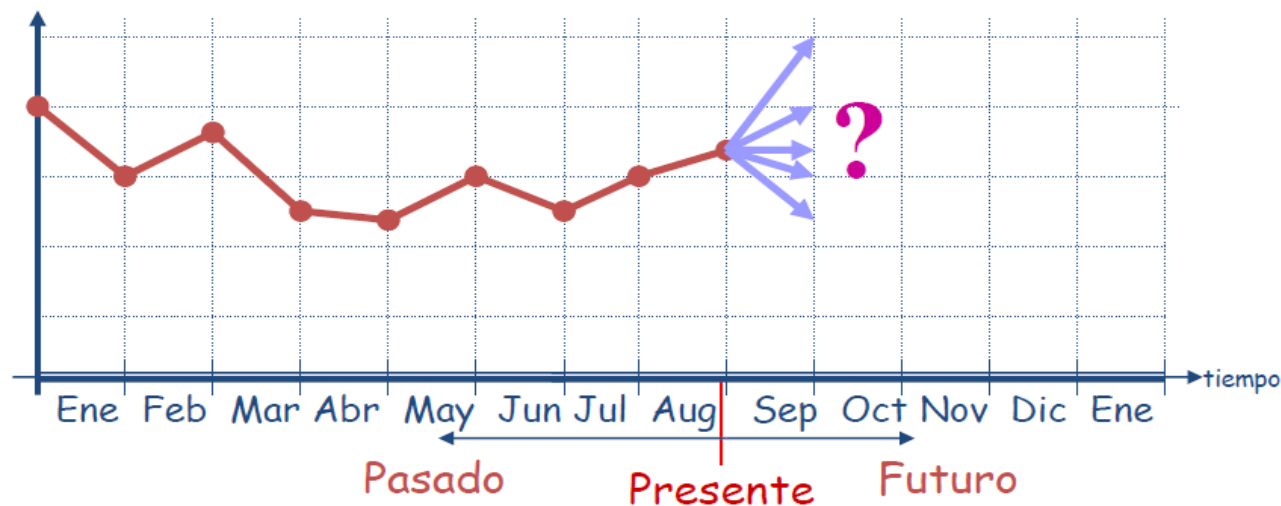
Series temporales

Análisis de Series Temporales

Se utilizan para predecir el comportamiento de variables que dependen del tiempo.

Se necesita una base de datos histórica.

Se utilizan para la previsión de demanda o stocks en un almacén o para la previsión del número de periódicos a enviar a un quiosco, etc.





XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

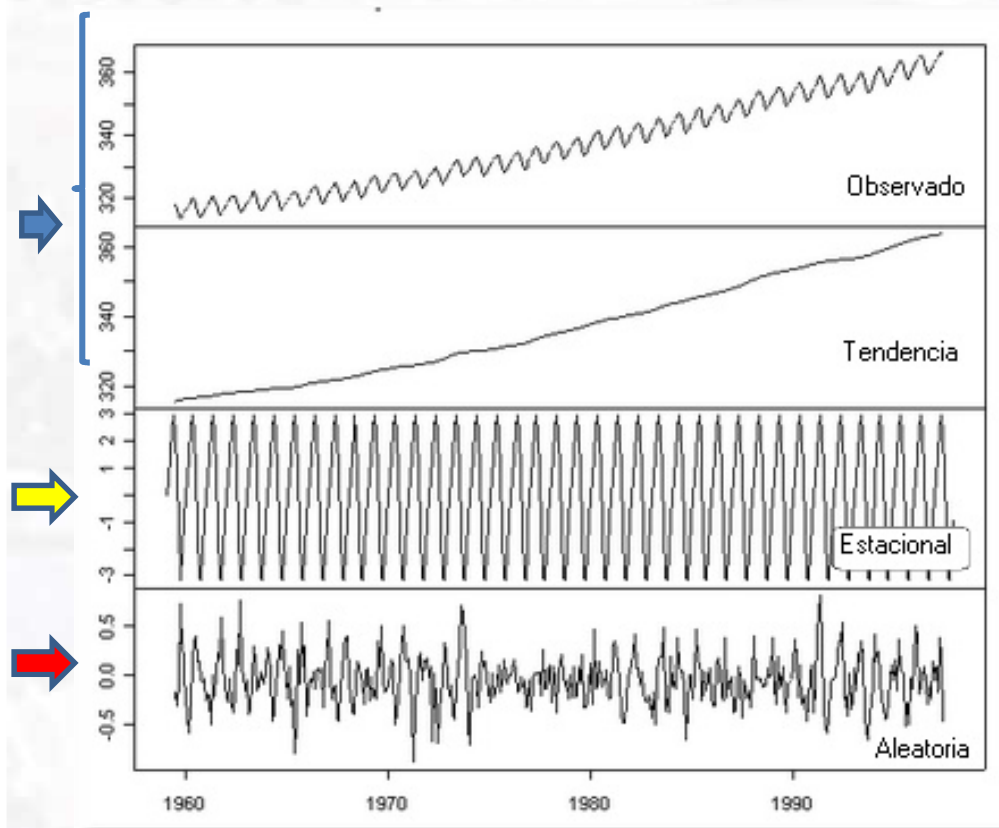
Series temporales

El análisis clásico de las series temporales se basa en la suposición de que los valores que toma la variable de observación es la consecuencia de cuatro componentes:

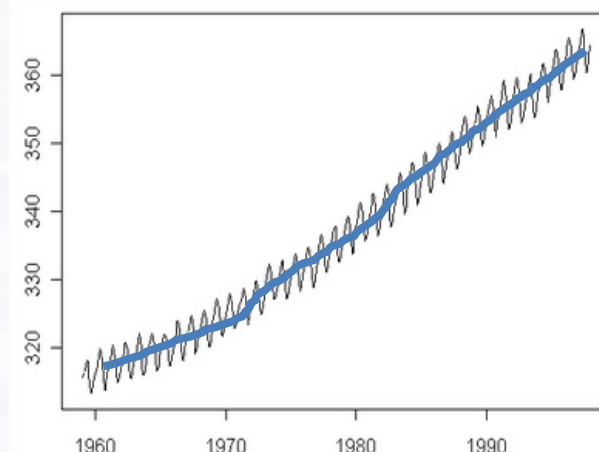
- 1. Tendencia.** Indica la **marcha general** y **persistente** del fenómeno observado, es una componente de la serie que refleja la evolución a largo plazo. Es la dirección general de la variable en el periodo de observación, es decir el cambio a largo plazo de la media de la serie.
 - a) Por ejemplo, el uso creciente de Internet en la sociedad, independientemente de que en un mes concreto en un país, por determinadas causas, haya una baja en la utilización de Internet.
- 2. Variación estacional (Estacionalidad).** Corresponde a **fluctuaciones periódicas** de la variable, en periodos relativamente cortos de tiempo. Es el **movimiento periódico de corto plazo**. Se trata de una componente causal debida a la influencia de ciertos fenómenos que se repiten de **manera periódica** en un **año** (las estaciones), una **semana** (los fines de semana) o un **día** (las horas puntas) o cualquier otro periodo.
- 3. Variación cíclica.** Es el componente de la serie que recoge las **oscilaciones periódicas** de **amplitud superior a un año**. Movimientos normalmente irregulares alrededor de la tendencia, en las que a diferencia de las variaciones estacionales, tiene un **período y amplitud variables**, pudiendo clasificarse como cíclicos, cuasicíclicos o recurrentes.
- 4. Variación aleatoria.** Accidental, de **carácter errático**, también denominada **residuo**, no muestran ninguna regularidad, debidos a fenómenos de carácter ocasional como pueden ser tormentas, terremotos, inundaciones, huelgas, guerras, avances tecnológicos, etc.



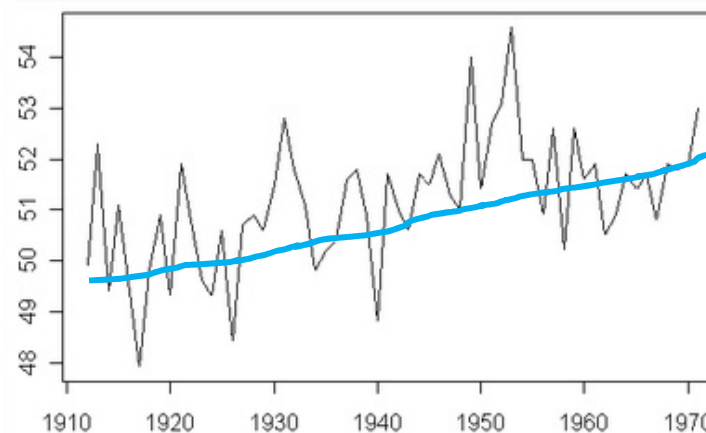
Ejemplos de series temporales



Descomposición de una serie temporal en sus componentes



Serie temporal con **tendencia**



Serie temporal con **tendencia menos pronunciada**



XXXV

Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Ojo con Big Data y su capacidad de predicción

“Detecting influenza epidemics using search engine query data”

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention (CDC).
Nature Vol 457, 19 February 2009.

Desde su publicación, *“Flu Trends”* ha sobreestimado las visitas a centros médicos relacionadas con la gripe, siendo especialmente inexacto en los picos. En el 2012-13 predijo el doble y en el 2011-12 sobre estimó más del 50%. Estos datos se publicaron en el paper:

“La parábola de la “gripe” de Google: Trampas en el análisis Big Data”. David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani.
SCIENCE Vol. 343 14 March 2014.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas

Murcia

Roberto Rigobon profesor en el Massachusetts Institute of Technology

Lo que hay que entender es que existe una gran distancia entre los **datos** y la **información**.

Producimos una inmensa cantidad de datos, pero eso no significa que estemos haciendo más o mejor información.

- Por ejemplo: gracias al procesador de texto, hoy en día escribir es más fácil que nunca, si lo comparamos con el esfuerzo que tenía que hacer Cervantes para escribir. Pero no estamos haciendo más literatura de calidad que en tiempos de Cervantes.
- Ni **Newton**, ni **Einstein**, tenían Big Data, ni siquiera computadoras. Kepler descubrió las leyes del movimiento de los planetas, estudiando miles de observaciones que había hecho **Tycho Brahe** a simple vista

Podemos aprender mucho de los datos, pero se ha sobrevalorado su potencial de tener un contenido informativo.

Hay mucha distancia entre **datos e información** pero más aún entre **información y conocimiento**.



XXXV

Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Roberto Rigobon profesor en el Massachusetts Institute of Technology

Un ejemplo:

Pongamos el caso de Amazon, que vende cerca de 53 millones de productos.

Imaginemos que queremos medir cuánto han cambiado los precios en Amazon.

- Tenemos la capacidad de bajar toda esa información; no supondría ningún problema.
- Pero algo más de 22 millones de esos productos son libros o música.
- Y solo 500 de esos libros representan más del 90% de las ventas. Solo se necesitarían esos 500 libros para obtener un dato relevante.
- ¿Para qué bajarse el resto?

Es totalmente inútil. Es un ejemplo de que no se trata de fuerza bruta.



XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

Simulation and “Big Data”



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

The simulator: source of Big Data

The simulator as a sensor and main data source
can simulate any possible situation :

Different values assigned to the random number generation seed, will affect the way incoming patients come into service in each hour.



Different iterations for each scenario.



The number of generated data will increase even more.

Generated data in each simulation

LoS (Length of Stay) in each location

Number of patients

Staff Configuration

Patients/hour/
location

WRO

AD

WR1

TR

WR2

DC

WR3

RX

Typically 400

5 (types) x 2 (levels) =
10 data

24 hours x 8 locations

$(8 \times 400) + 10 + (24 \times 8) = 3,402$ generated data in each execution

TOTAL GENERATED DATA = $5,670,000 \times 3,402 = 1,928,934 \times 10^{10}$



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

¿Big Data?.... seguro “Data intensive”

1 escenario = 1 simulación

Configuración “staff”

Llegada pacientes al servicio

Tipo /número (min-máx)

Nivel
experiencia

INPUT DATA

AS (1-3)	TN (1-3)	D (1-4)	NER (1-2)	RX (1-2)	Nivel 2 ?	Junior	Senior	100%	80%	50%	20%	...
-------------	-------------	------------	--------------	-------------	--------------	--------	--------	------	-----	-----	-----	-----

Configuraciones posibles (nivel 1) ID

$9_{AS} \times 9_{TN} \times 14_D \times 5_{NER} \times 5_{RX} = 28350 \times 20_{(p.e.)} = 567000$ escenarios posibles

Datos generados por simulación

LoS (Length of Stay) en cada servicio

Entrada Pacientes

Configuración “staff”

Pac/hora/servicio

WR0	AD	WR1	Tr	WR2	DC	WR3	NER / RX	<400	10	24x8
-----	----	-----	----	-----	----	-----	----------	------	----	------

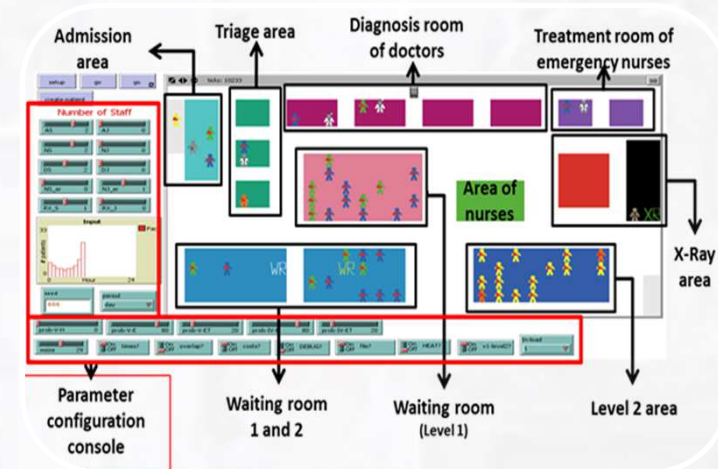
$(8 \times 400) + 10 + (24 \times 8) = 3402$ datos generados en cada ejecución

TOTAL DATOS GENERADOS = $567000 \times 3402 = 1,928934 \times 10^9$



Modelo de trabajo experimental

El simulador como fuente de datos



1 escenario = 1 simulación

Configuración "staff"

Llegada pacientes al servicio

Tipo /número (min-máx)					Nivel experiencia		INPUT DATA				
AS (1-3)	TN (1-3)	D (1-4)	NER (1-2)	RX (1-2)	Nuevo staff	Junior	Senior	100%	80%	50%	20% ...

Configuraciones posibles (Área B) ID

$$9_{AS} \times 9_{TN} \times 14_D \times 5_{NER} \times 5_{RX} = 28350 \times 20_{(p.e.)} = 567000 \text{ escenarios posibles}$$



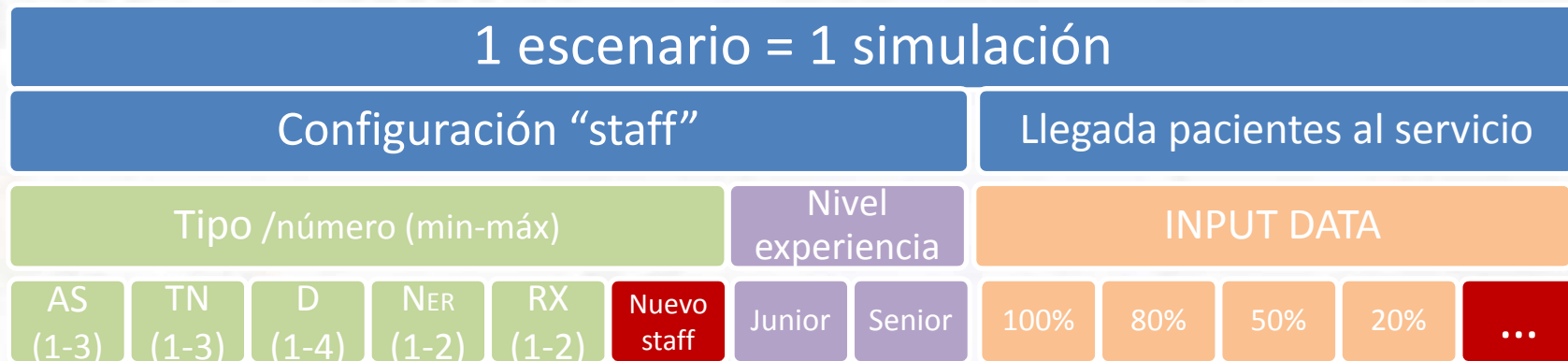
XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



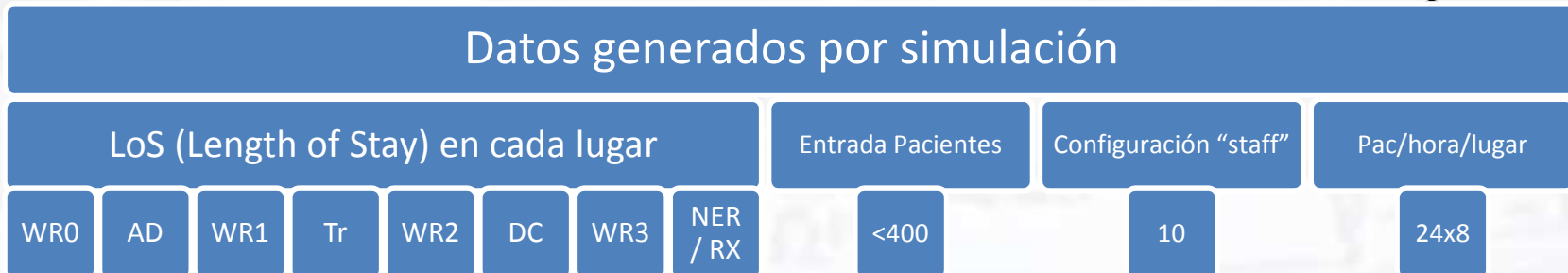
19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

¿Qué registramos?



Configuraciones posibles (Área B) ID

$$9_{AS} \times 9_{TN} \times 14_D \times 5_{NER} \times 5_{RX} = 28350 \times 20_{(p.e.)} = 567000 \text{ escenarios posibles}$$



$$(8 \times 400) + 10 + (24 \times 8) = 3402 \text{ datos generados en cada ejecución}$$

$$\text{TOTAL DATOS GENERADOS} = 567000 \times 3402 = 1,928934 \times 10^9$$



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

SPECIFIC OBJECTIVES

To gain knowledge about the system behavior for **any possible situation** in the Emergency Department.

- **Unexpected** increase in the demand for service, mass accidents, epidemics, etc.

To anticipate solutions for **unusual situations**.

- Provide **prediction models** of staff demand and other resources in the service, as a reference for making decisions in any possible reality.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

POSSIBILITY OF SIMULATOR

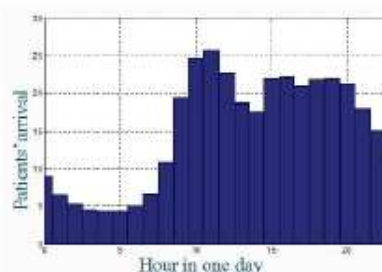
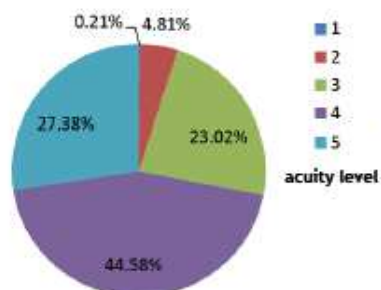
Simulation condition:

Simulation Time: 720 hours(30 days)

Execution Time: 3 minutes per one scenario

Patient arrival (input)

Real Data from Hospital of Sabadell



397 patients / day

Virtual ED

Scenario configuration

name	value	meaning
<i>n-of-adm</i>	4	admission
<i>n-of-tn</i>	4	triage nurse
<i>n-of-jnA</i>	1	junior nurse in A
<i>n-of-snA</i>	7	senior nurse in A
<i>n-of-jnB</i>	1	junior nurse in B
<i>n-of-snB</i>	2	senior nurse in B
<i>n-of-jdA</i>	3	junior doctor in A
<i>n-of-sdA</i>	3	senior doctor in A
<i>n-of-jdB</i>	2	junior doctor in B
<i>n-of-sdB</i>	2	senior doctor in B
<i>n-of-auxi</i>	10	auxiliary
<i>n-of-cb</i>	60	carebox
<i>n-of-int-tr</i>	4	internal test room
<i>n-of-lab-td</i>	9	external test room
<i>n-of-ext-td</i>	3	lab. test room
<i>n-of-amb</i>	8	ambulance
<i>area-B-cap</i>	70	capacity of B



XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

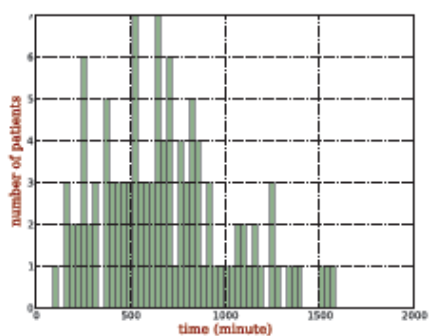
IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



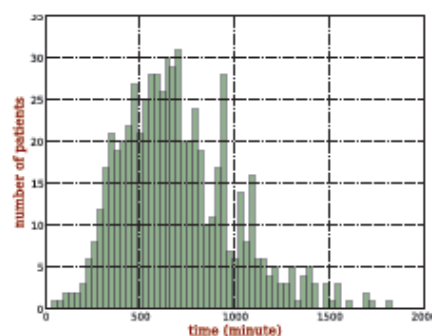
19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Area A

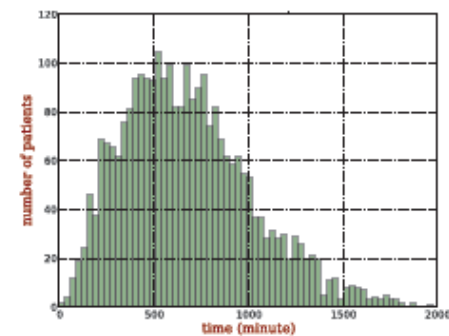
Patients with acuity level 1



Patients with acuity level 2



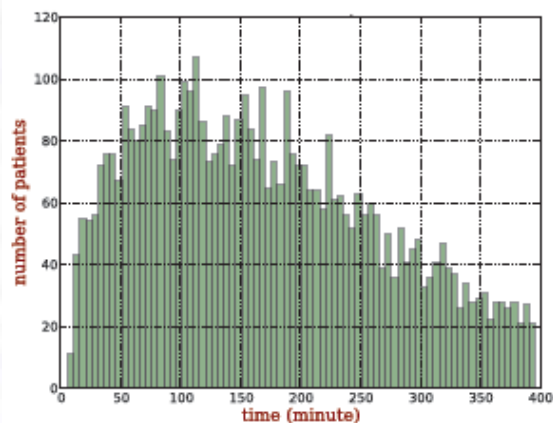
Patients with acuity level 3



LoS distribution

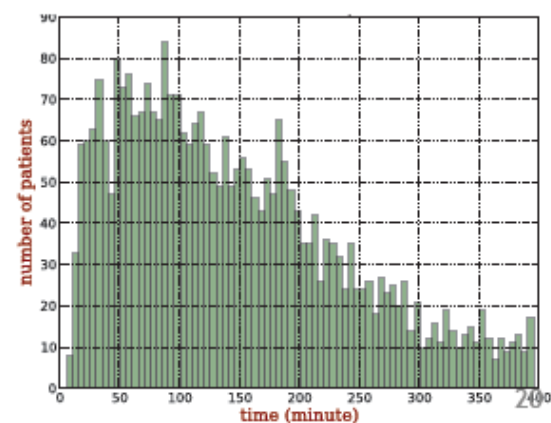
Area B

Patients with acuity level 4



QoS ?

Patients with acuity level 5





XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



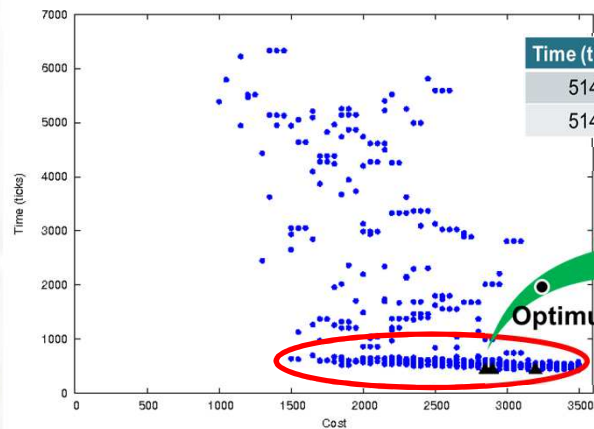
19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

Example: from “Regular” to “Exceptional” situations

Average patient “LoS” for all possible staff configurations

Cost constraint ≤ 3500 €

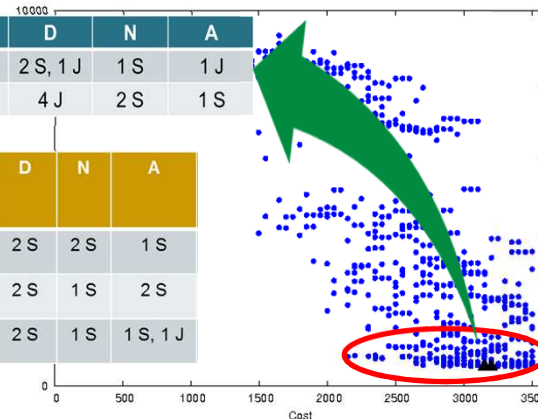
4 p/hr
Regular



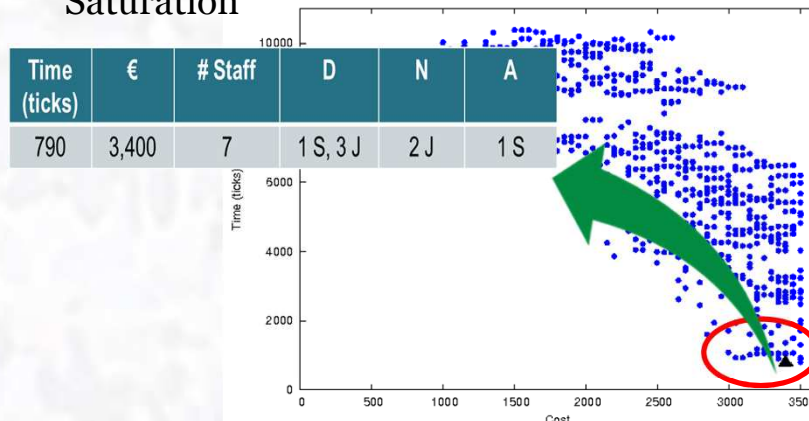
Time (ticks)	€	# Staff	D	N	A
514	3,150	5	2 S, 1 J	1 S	1 J
514	3,200	7	4 J	2 S	1 S

Time (ticks)	€	# Staff	D	N	A
428	3,200	5	2 S	2 S	1 S
428	2,900	5	2 S	1 S	2 S
428	2,850	5	2 S	1 S	1 S, 1 J

9 p/hr
Complex



13 p/hr
Saturation

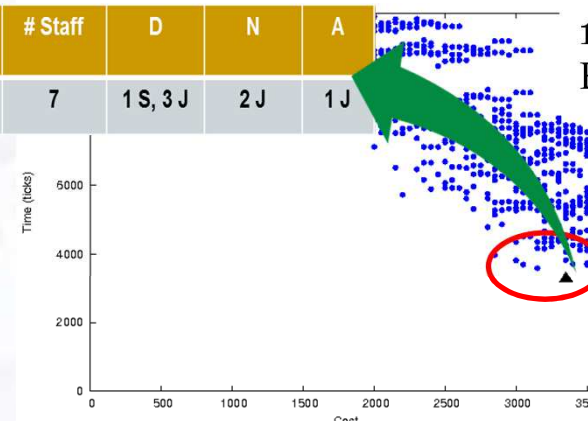


Time (ticks)	€	# Staff	D	N	A
790	3,400	7	1 S, 3 J	2 J	1 S



Time (ticks)	€	# Staff	D	N	A
3,266	3,350	7	1 S, 3 J	2 J	1 J

17 p/hr
Excepcional!





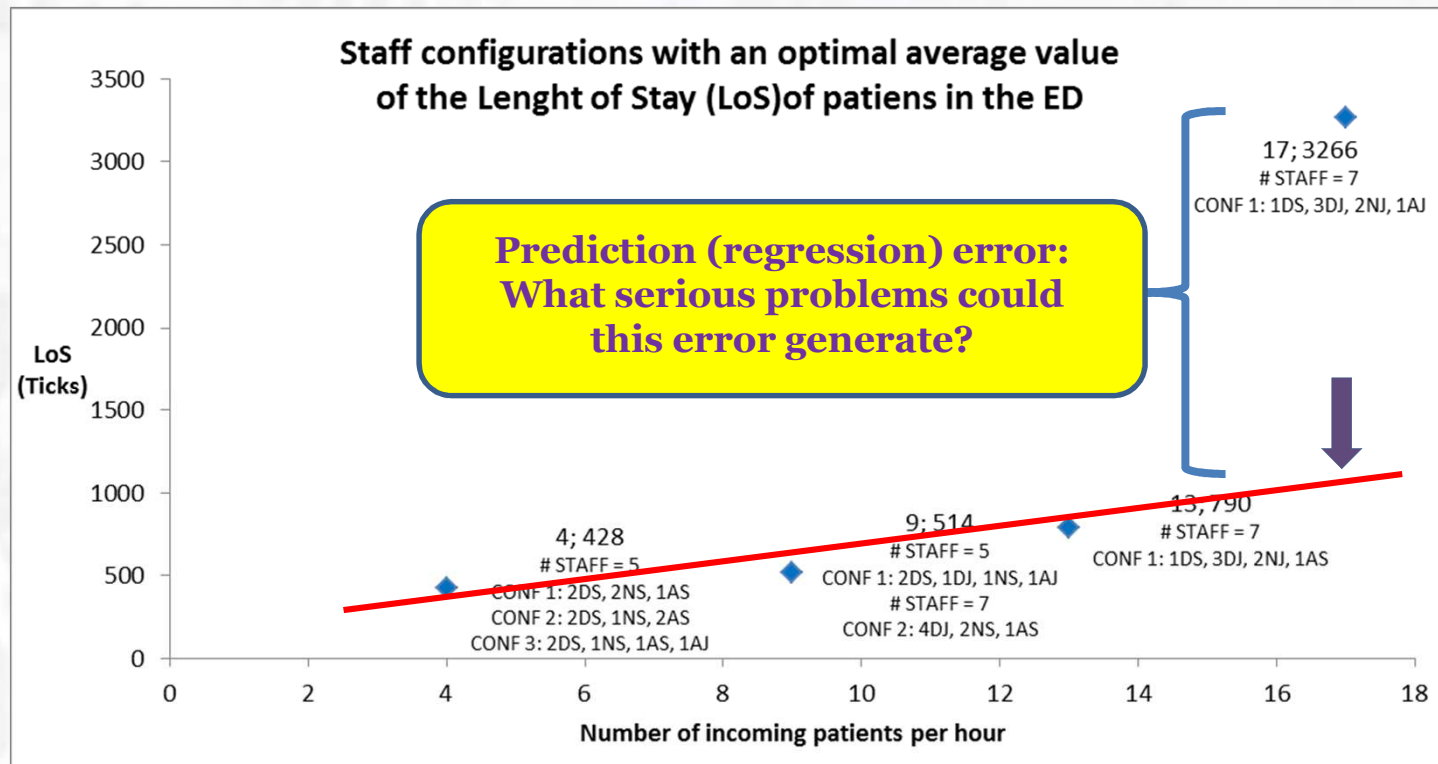
XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia

JUSTIFICATION



Data generated by simulation can be a more reliable source for predicting the behavior of the real system.

Incoming Patients	4 pat/h 9 pat/h 13 pat/h 17 pat/h
-------------------	--



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia

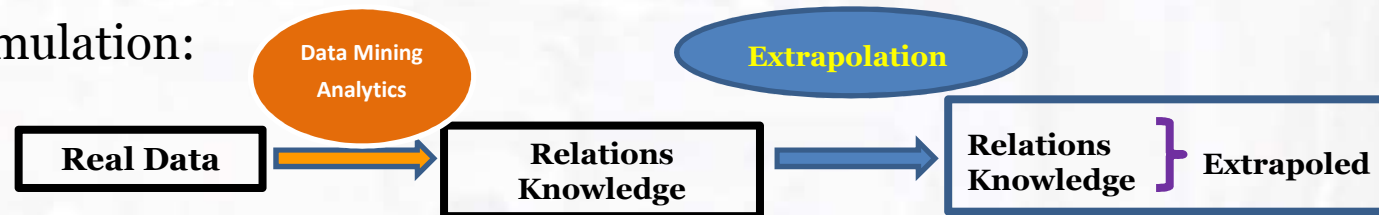


19-21 Noviembre 2014

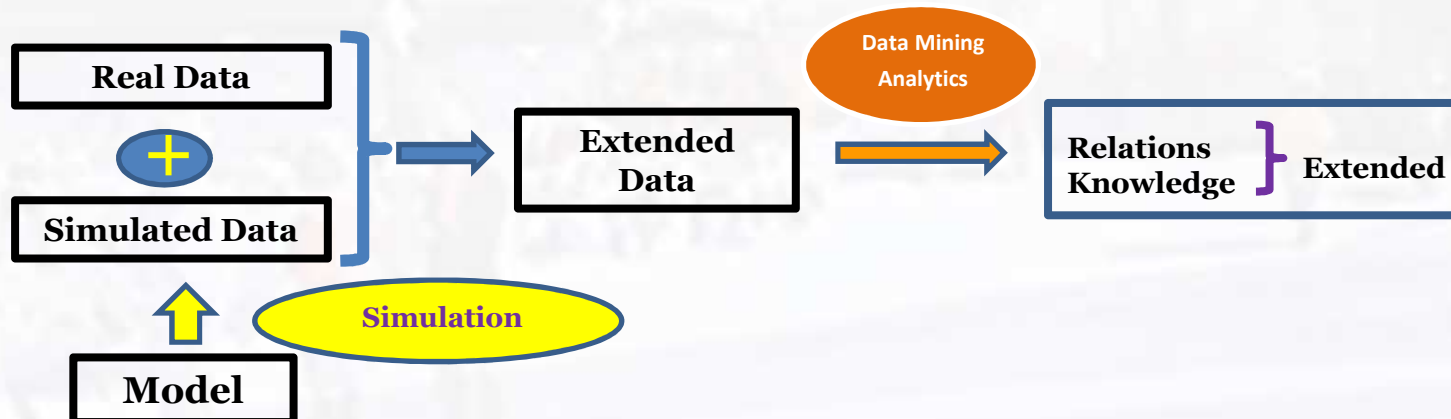
Auditorio y Centro de Congresos Víctor Villegas
Murcia

HYPOTHESIS

Without simulation:



Data obtained directly from the real system will be complemented with data **generated by the simulator**. This will allow us to obtain much more refined behavior models of the ED, an extra knowledge which would not be possible to obtain without the simulation:





XXXV

**Congreso Nacional de la Sociedad
Española de Medicina Interna (SEMI)**

IV Congreso Ibérico de Medicina Interna

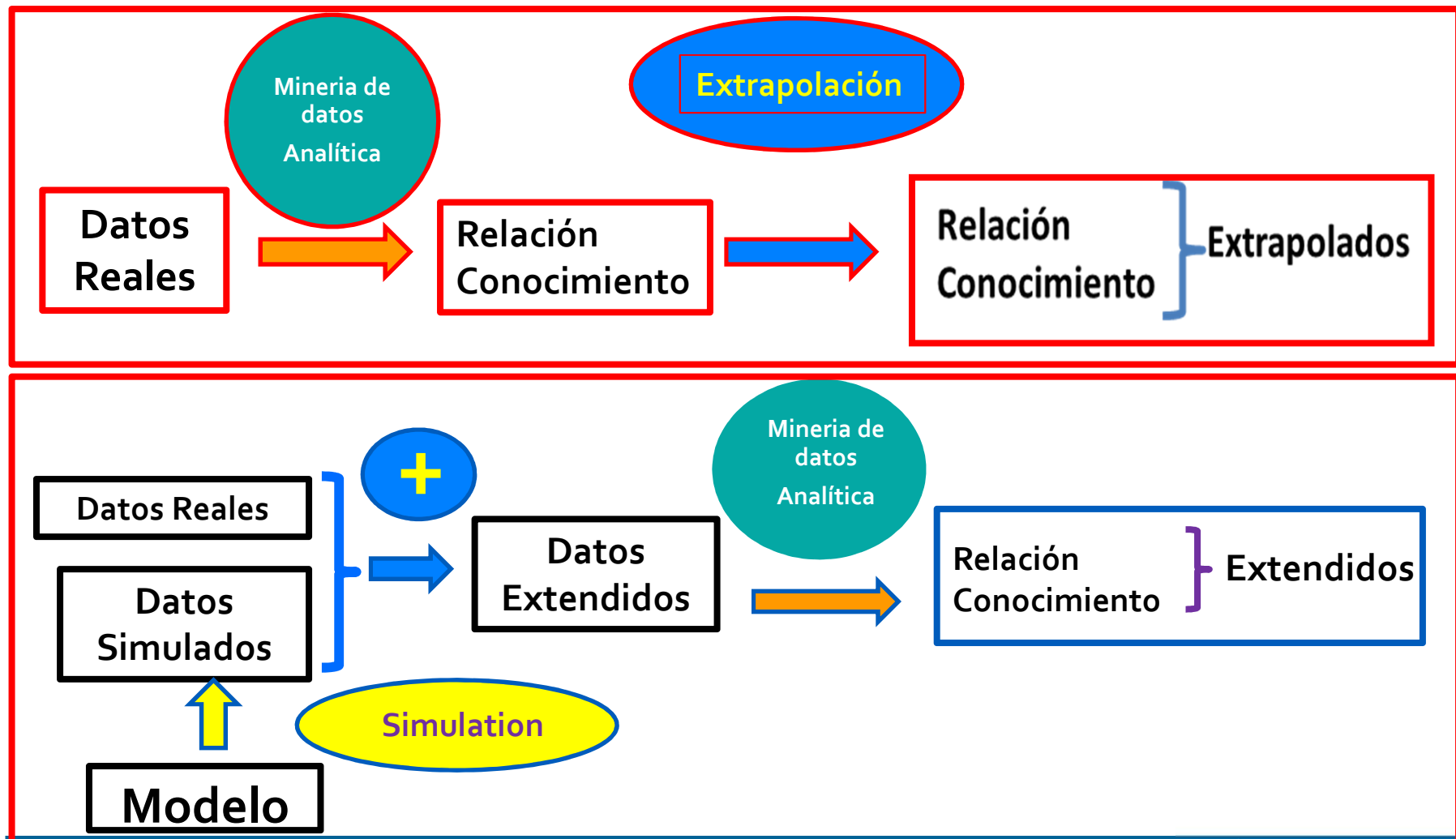
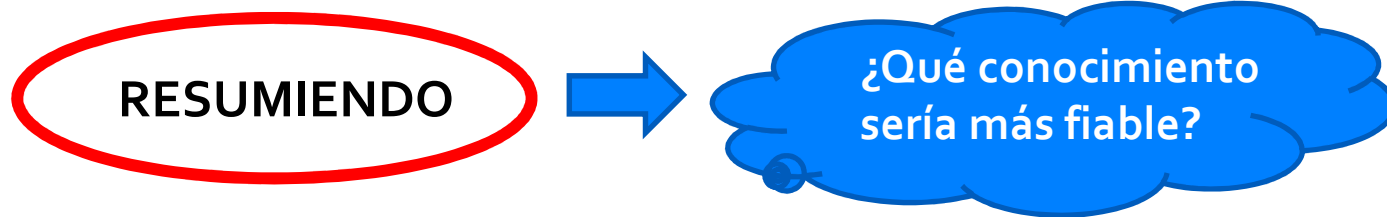
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

**Auditorio y Centro de Congresos Víctor Villegas
Murcia**

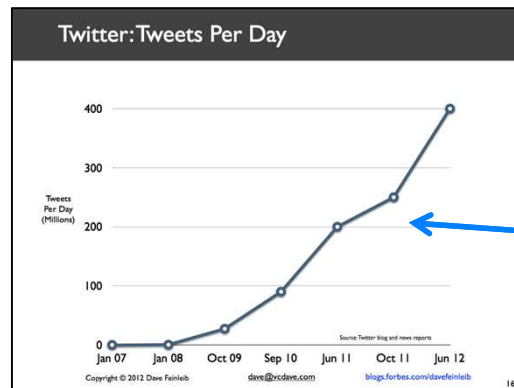
**Gracias por su
atención**



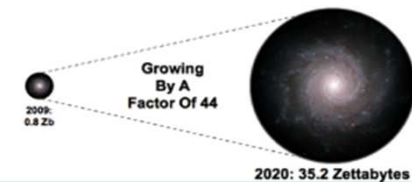
Characteristics of Big Data:

1-Scale (Volume)

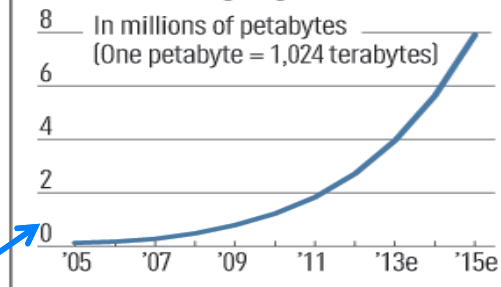
- **Data Volume**
 - 44x increase from 2009 2020
 - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially



The Digital Universe 2009-2020



Data storage growth



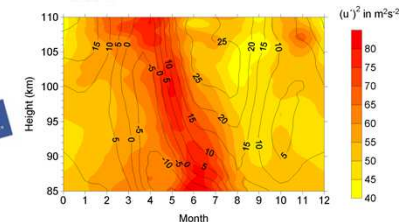
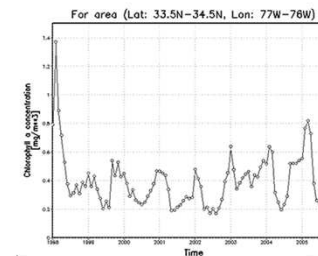
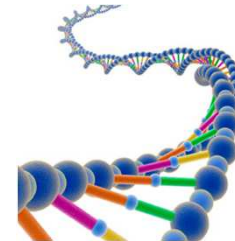
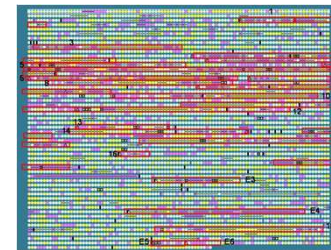
Exponential increase in collected/generated data

Characteristics of Big Data:

2-Complexity (Variety)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc...
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

To extract knowledge → all these types of data need to be linked together



Characteristics of Big Data:

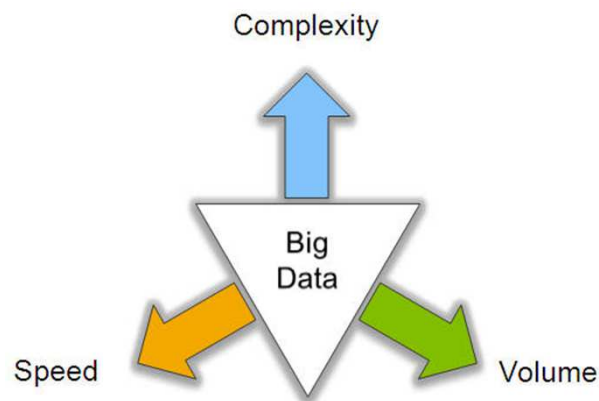
3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- **Examples:**

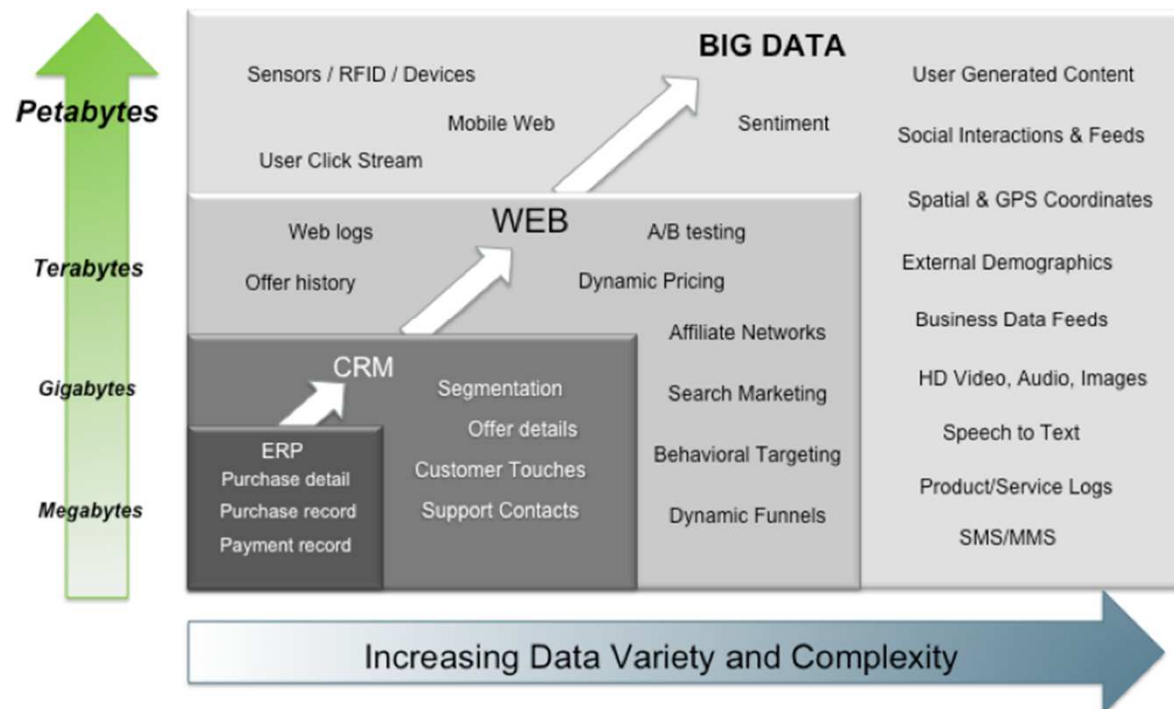


- **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions right now for store next to you
- **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction

Big Data: 3V's



Big Data = Transactions + Interactions + Observations



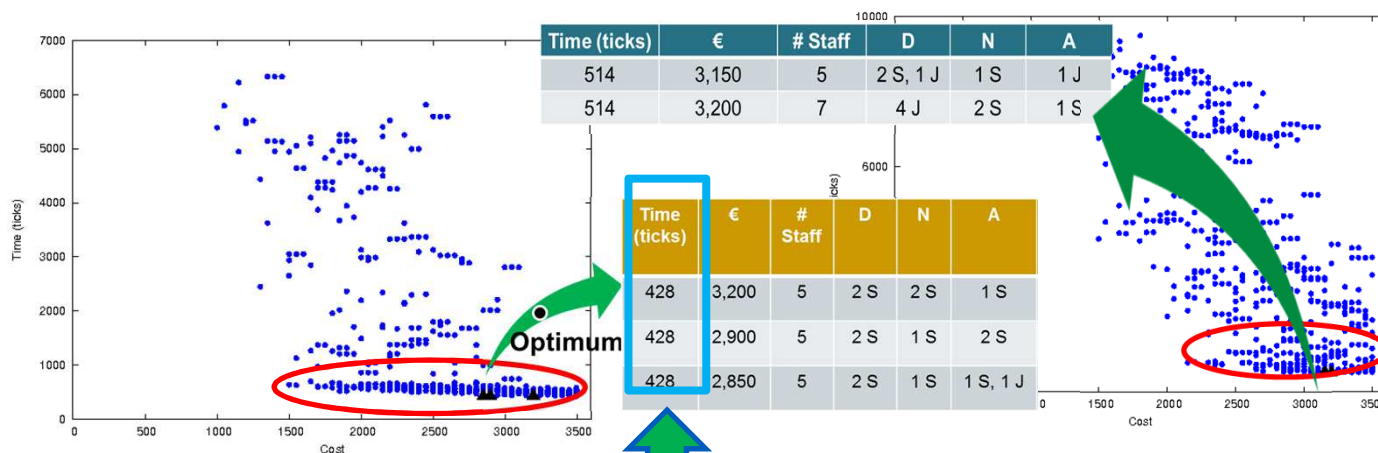
Source: Contents of above graphic created in partnership with Teradata, Inc.

Example: from “Regular” to “Exceptional” situations

Average patient “LoS” for all possible staff configurations

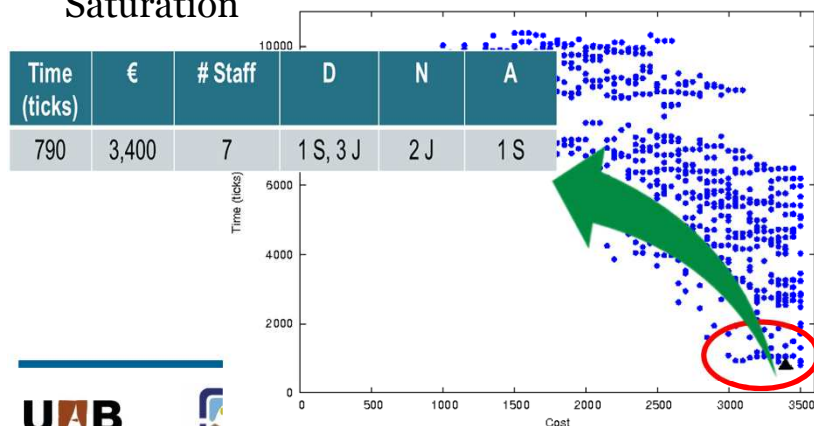
Cost constraint ≤ 3500 €

4 p/hr
Regular

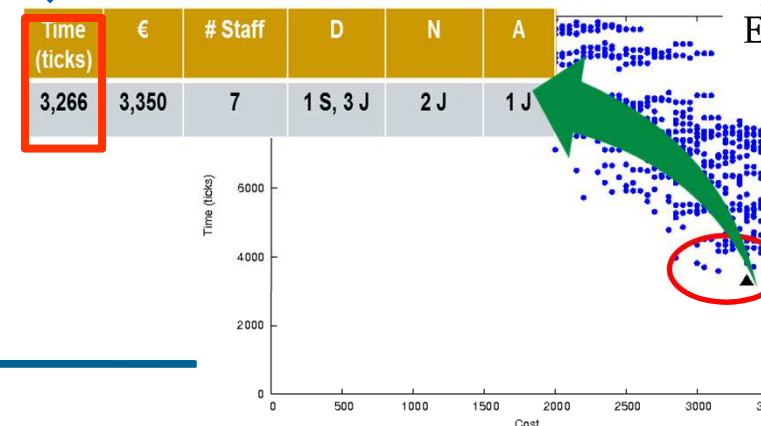


9 p/hr
Complex

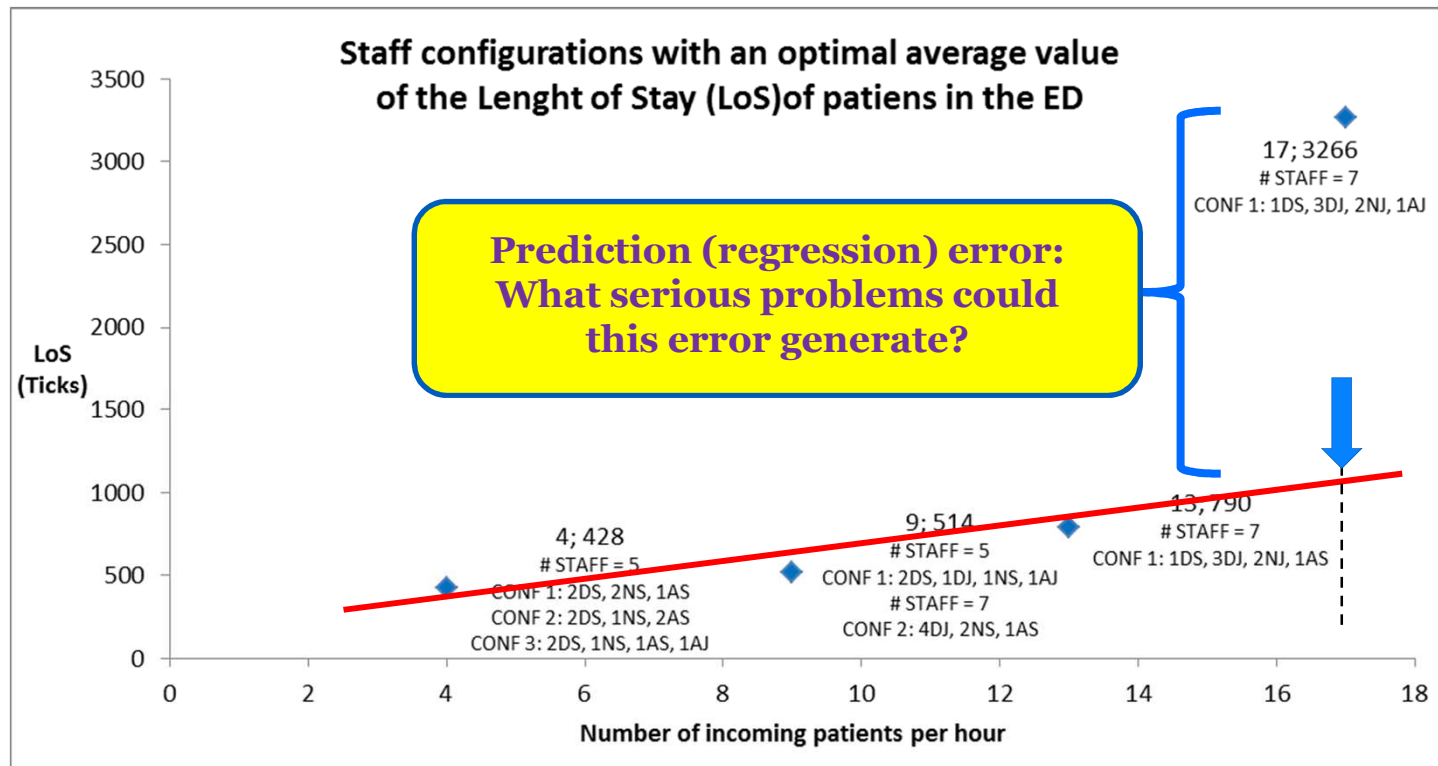
13 p/hr
Saturation



17 p/hr
Excepcional!



Justification

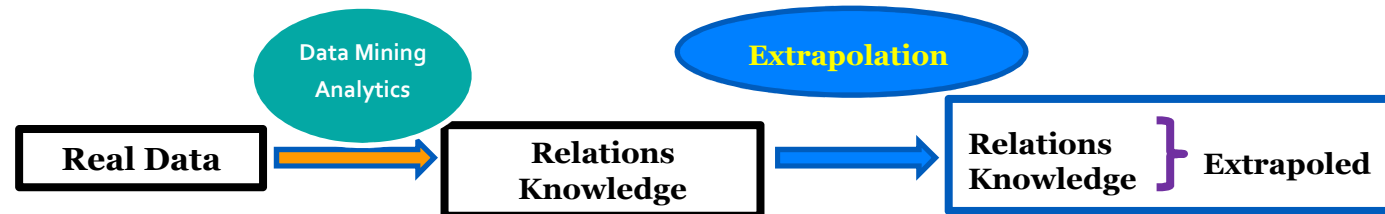


Data generated by simulation can be a more reliable source for predicting the behavior of the real system.

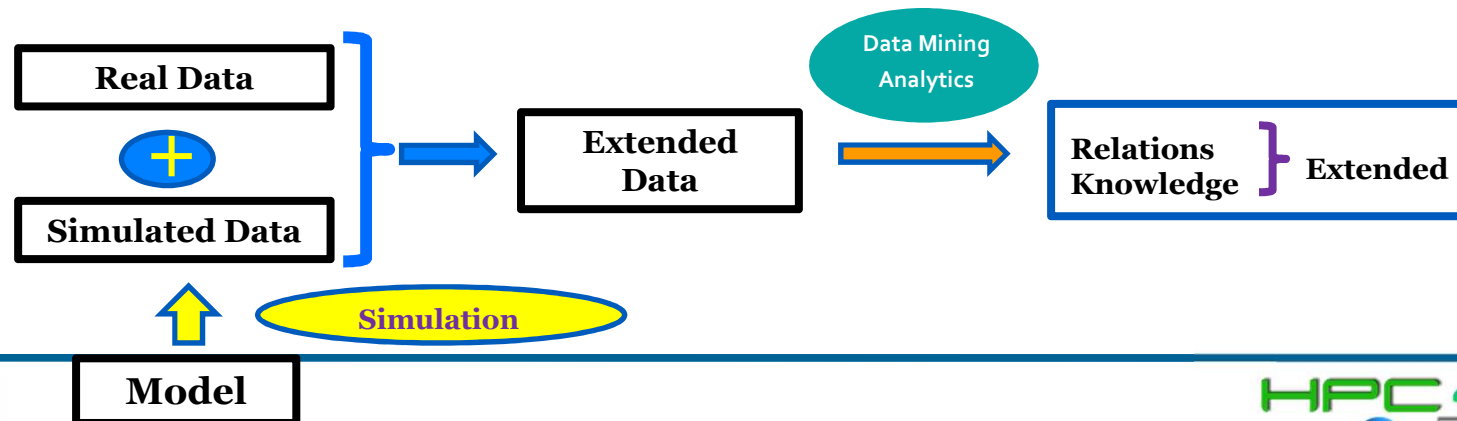
Incoming Patients	4 pat/h 9 pat/h 13 pat/h 17 pat/h
-------------------	--

Hypothesis

Without simulation:



Data obtained directly from the real system will be complemented with data **generated by the simulator**. This will allow us to obtain much more refined behavior models of the ED, an extra knowledge which would not be possible to obtain without the simulation:

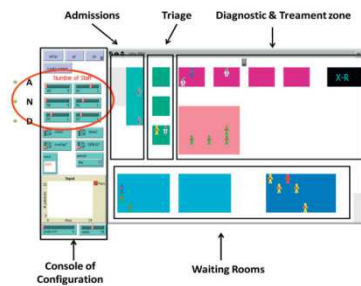


The promise of Big Data

- Data contains information of great business value
- If you can extract those insights you can make far better decisions
- ...but is data really that valuable?



Planteamiento del problema

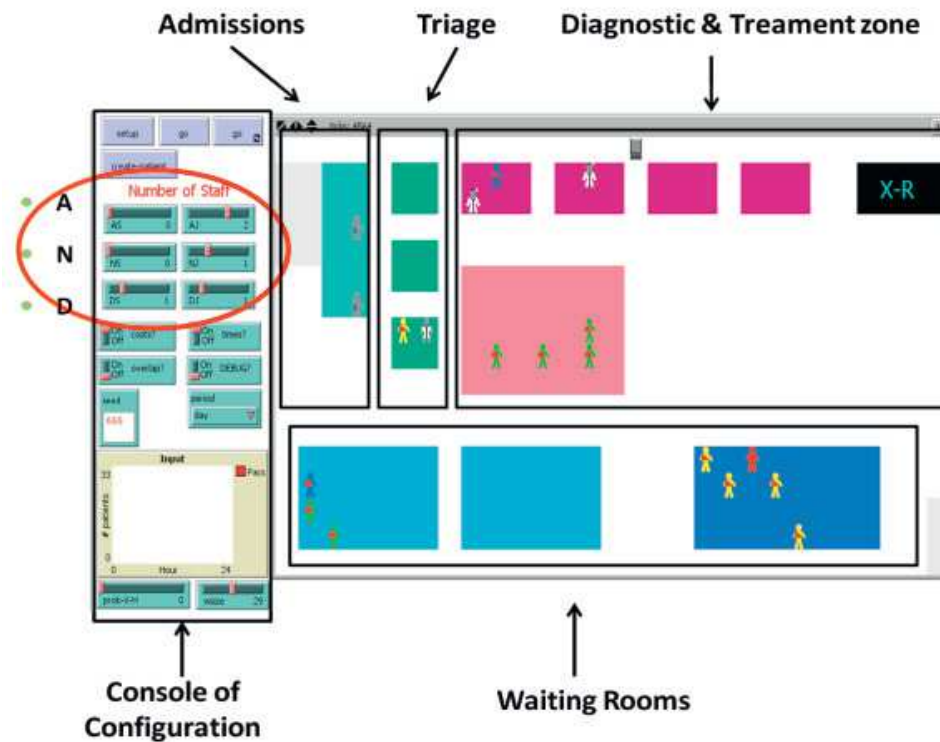


Contamos con un simulador del SUH validado y verificado.

Podemos ejecutar la simulación para diferentes escenarios según la carga de entrada (número de pacientes) y configuraciones del personal sanitario.

La cantidad de datos generados únicamente depende del número de ejecuciones realizadas. (HPC)

Planteamiento del problema



**“SENSOR” DEL
SISTEMA REAL**

Generación de
datos masivos
(Big Data)

Extracción de
información
(Data mining)

Generación de
conocimiento



XXXV Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014
Auditorio y Centro de Congresos Víctor Villegas
Murcia



UAB
Universitat Autònoma
de Barcelona

**A comprehensive
University**



**"Vila Universitaria"
University Residence**



Faculties

**UAB
Campus**



**School of
Engineering**





XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas

Murcia

¿Qué nos puede aportar la simulación en procesos de Big Data?

- Tenemos el mundo real de salud, en nuestro caso el Servicio de Urgencias (ED).
- Además nosotros tenemos nuestro mundo simulado del ED.
- A partir de aquí uno puede sacar un montón de información. Con toda esta información, a través de Big Data (o Business Intelligence), sacamos un cierto conocimiento sobre lo que pasa aquí.
- Hasta aquí sería lo normal (real), es decir, por aquí sacamos el conocimiento que corresponde a los datos reales que podemos obtener, pero no más allá de los datos reales que podemos obtener, de las situaciones más o menos estándar que se puedan dar.
- **La simulación nos permitiría poder llegar a generar un conocimiento que sin la simulación no podríamos tener, porque somos capaces de aportar una serie de datos que el mundo real no nos los puede dar porque esas situaciones no son situaciones de las que disponemos normalmente, llamémosle, información de situaciones excepcionales, ideales o especiales, por una u otra razón: porque cambiemos los parámetros en la forma en que nosotros estimemos que podría ser interesante de cara al futuro, a problemas que ocurran, etc.**
- **¿Qué conocimiento extra estaríamos obteniendo?, es decir, ¿qué más somos capaces de obtener?. Por aquí (real) sabemos el conocimiento, pero este (simulación) nos permitiría tener un conocimiento mayor que probablemente nos podría permitir actuar en situaciones extremas, diferentes, en situaciones que no se dan normalmente. Encontrar ese tipo de comportamiento.**
- En un futuro podemos pensar en una realimentación hacia el simulador (modelo y simulación) en función del tipo de información que hemos obtenido, pero que viésemos, o que nos falta información, o que podemos obtener más conocimiento, pero no lo tenemos en ese caso disponible y pudiésemos “perdírselo” al simulador, mejorando los modelos o generando nuevas simulaciones.
- Es decir, podríamos actuar sobre esa “realidad virtual” que nos proporciona el simulador para que nos generase más información y pudiésemos sacar más conocimiento.



XXXV

Congreso Nacional de la Sociedad Española de Medicina Interna (SEMI)

IV Congreso Ibérico de Medicina Interna

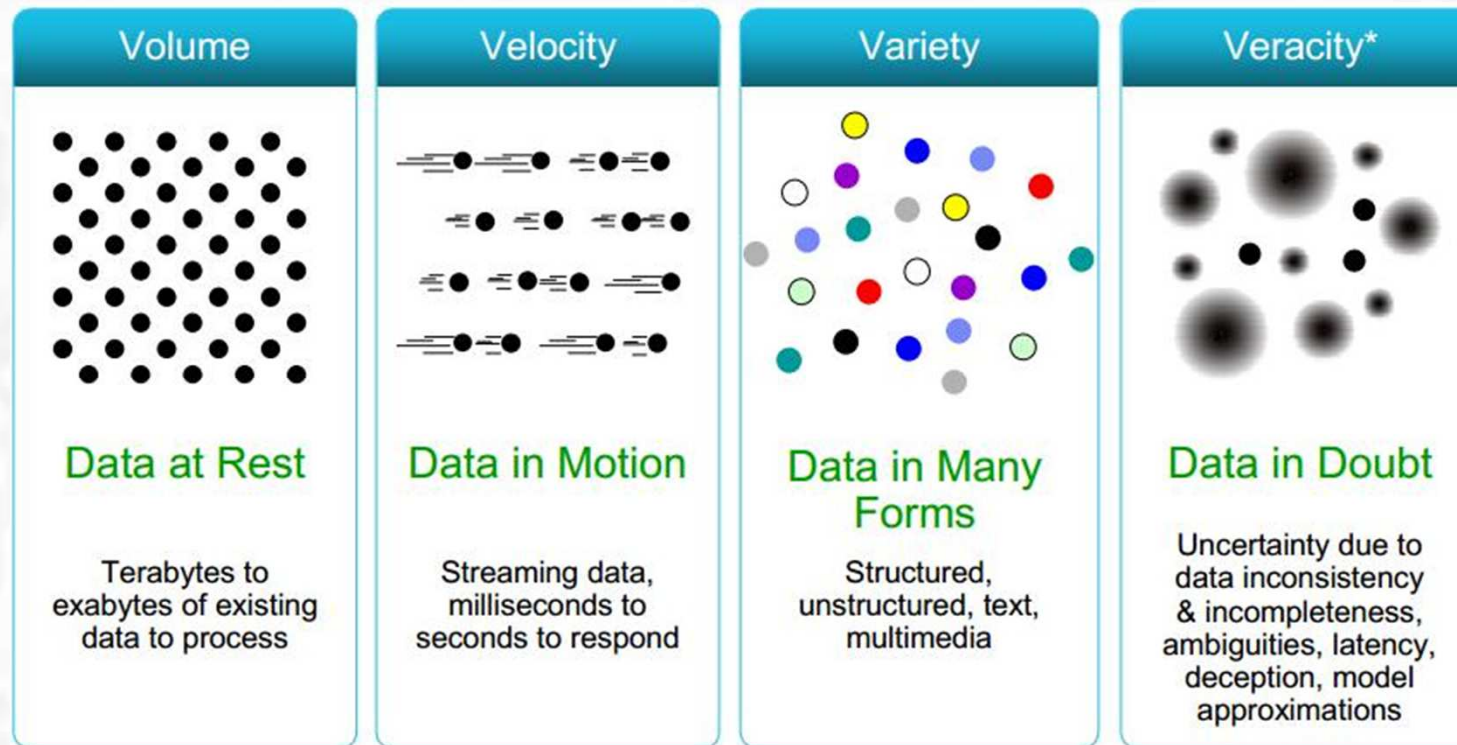
II Congreso de la Sociedad de Medicina Interna de la Región de Murcia



19-21 Noviembre 2014

Auditorio y Centro de Congresos Víctor Villegas
Murcia

Some Make it 4V's



Objetives

Create a contact propagation model of Methicillin-Resistant Staphylococcus Aureus (MRSA) in a hospital emergency department by using agent-based modeling and simulation technique (ABMS).

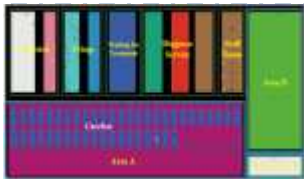
Context



Emergency Departments



Agent Based Modeling and Simulations (ABMS)

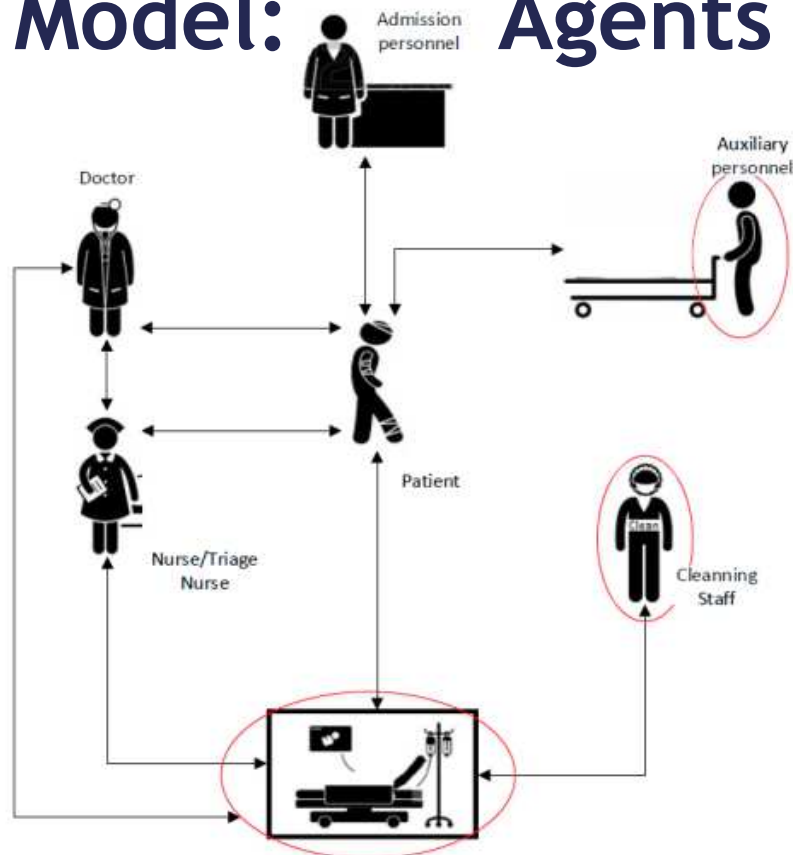


ED-Simulator

Objective

Create a contact propagation model of Methicillin-Resistant Staphylococcus Aureus (MRSA) in a hospital emergency department by using agent-based modeling and simulation technique (ABMS).

Contact Propagation Model: Agents



ED-Simulator

Admission personnel

Patient

Doctors (S, J)

Nurses(S, J)

Auxiliary Personnel

Cleaning Staff

AA

Information System

CareBox

PA