

# Lecture 9: Data Mining, Data Analytics and Big Data

Maaïke Limper, Antonio Romero, Manuel Martin



# Introduction

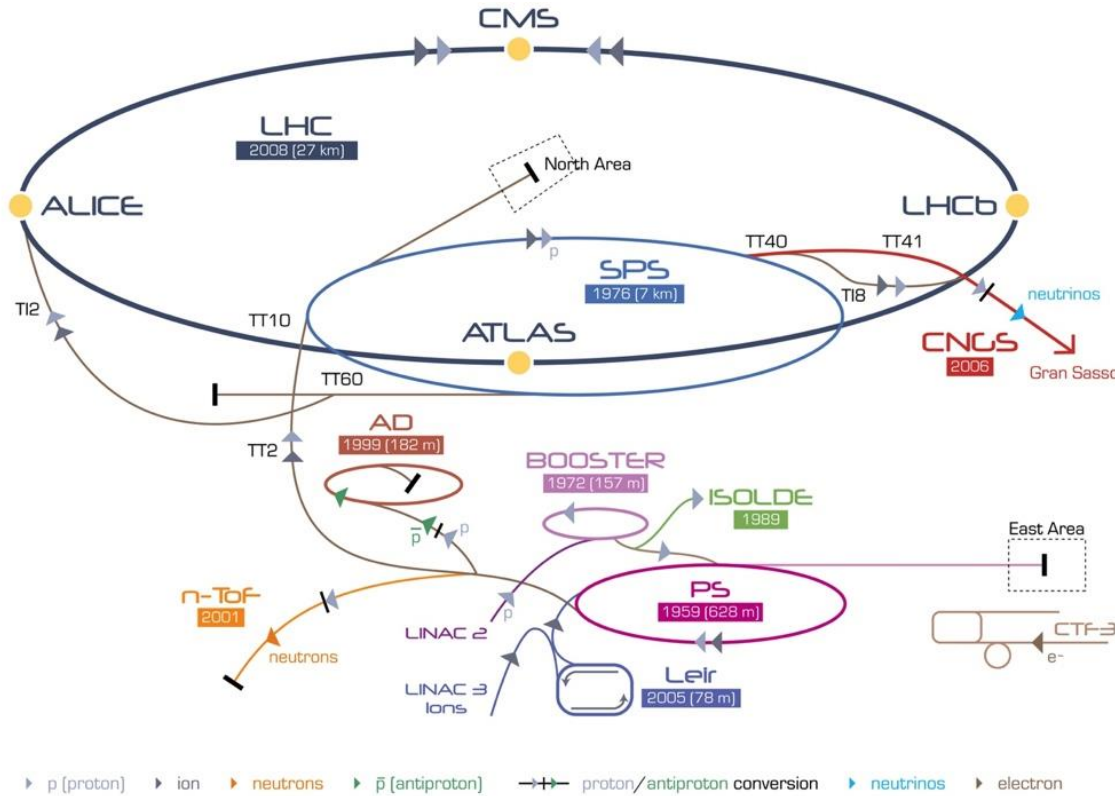
- Two openlab Projects in IT-DB
  - Data Analytics
  - In-Database Physics Analysis
- Both using data analytics
- Projects have different scopes



# Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

# CERN is a extreme data environment



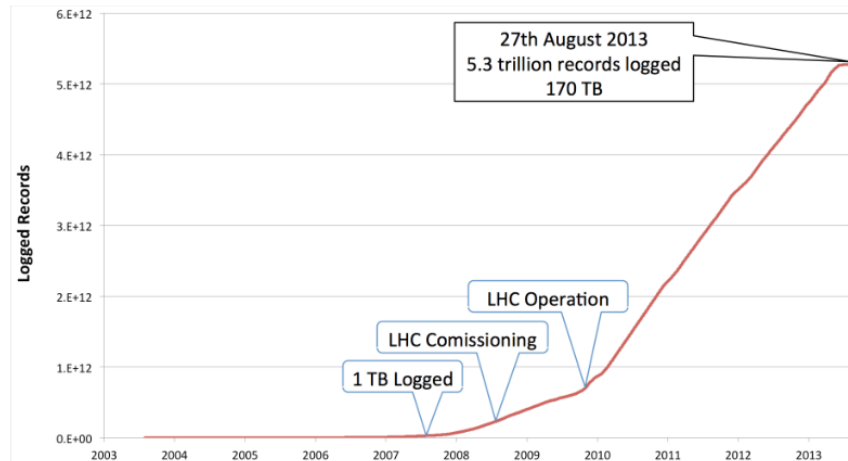
- Control and operations
  - Million of sensors, signals
  - Large number of control devices
  - Equipment
- Monitoring and logging
- Supporting IT infrastructure
  - Databases
  - Network
  - Services

LHC Large Hadron Collider SPS Super Proton Synchrotron PS Proton Synchrotron

AD Antiproton Decelerator CTF-3 Clic Test Facility CNCS Cern Neutrinos to Gran Sasso ISOLDE Isotope Separator OnLine DEvice  
LEIR Low Energy Ion Ring LINAC LINear ACcelerator n-ToF Neutrons Time Of Flight

# CERN Data Investment

- CERN generates huge amount of data everyday
  - Accelerator Logging Service around 275 GB/day
- CERN has great monitoring and logging systems
  - Large amount of data has been stored over years



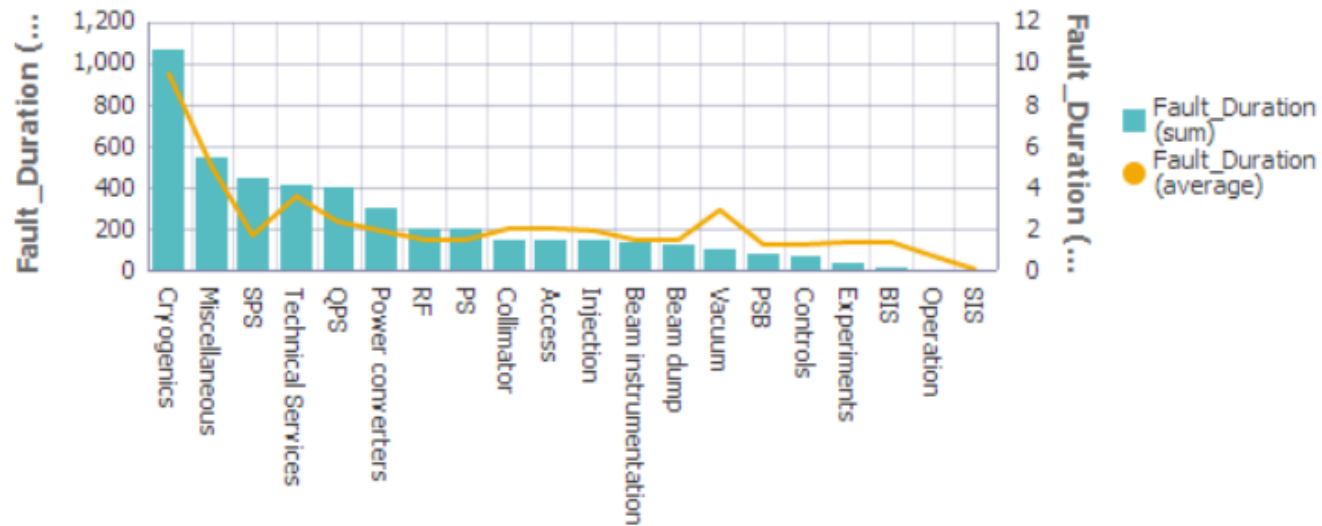
DIAMON



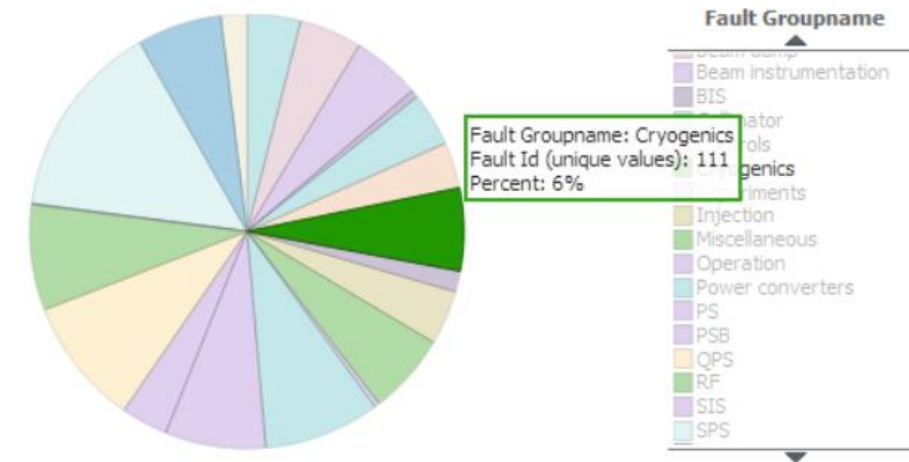
# Faults

- Some faults cannot be avoid
- Decrease the availability for running physics

Fault\_Duration (sum), Fault\_Duration (average) by Fault\_Groupname



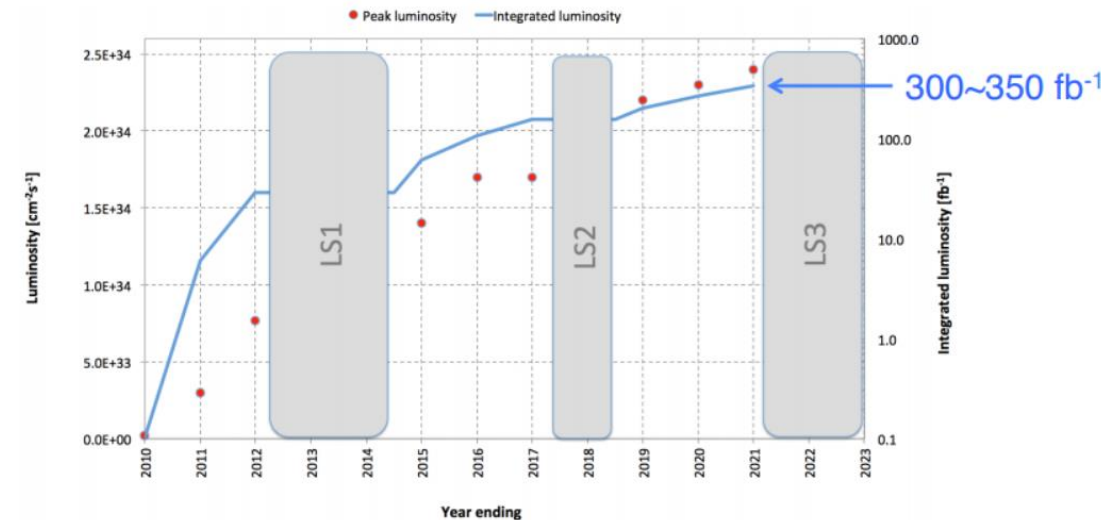
Fault Id (unique values) by Fault\_Groupname



# A look into the Future

- Future LHC upgrades will increase luminosity
  - Computing resources needs will be higher
  - Data generated will increase drastically

Parameter	2010	2011	2012	design value
Beam energy	3.5	3.5	4	7
$\beta^*$ in IP 1 and 5 (m)	2.0/3.5	1.5/1.0	0.6	0.55
Bunch spacing (ns)	150	75/50	50	25
Max. number of bunches	368	1380	1380	2808
Max. bunch intensity (protons per bunch)	$1.2 \times 10^{11}$	$1.45 \times 10^{11}$	$1.7 \times 10^{11}$	$1.15 \times 10^{11}$
Normalized emittance at start of fill (mm mrad)	$\approx 2.0$	$\approx 2.4$	$\approx 2.5$	3.75
Peak luminosity ( $\text{cm}^{-2}\text{s}^{-1}$ )	$2.1 \times 10^{32}$	$3.7 \times 10^{33}$	$7.7 \times 10^{33}$	$1 \times 10^{34}$
Max. mean number of events per bunch crossing	4	17	37	19
Stored beam energy (MJ)	$\approx 28$	$\approx 110$	$\approx 140$	362



- Next accelerators
  - Future Circular Collider (80-100 km)



# The objective – Improve our systems

Monitoring and Diagnostics Systems

Data Analytics

Predictive and Proactive systems



# Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

# openlab Data Analytics Project

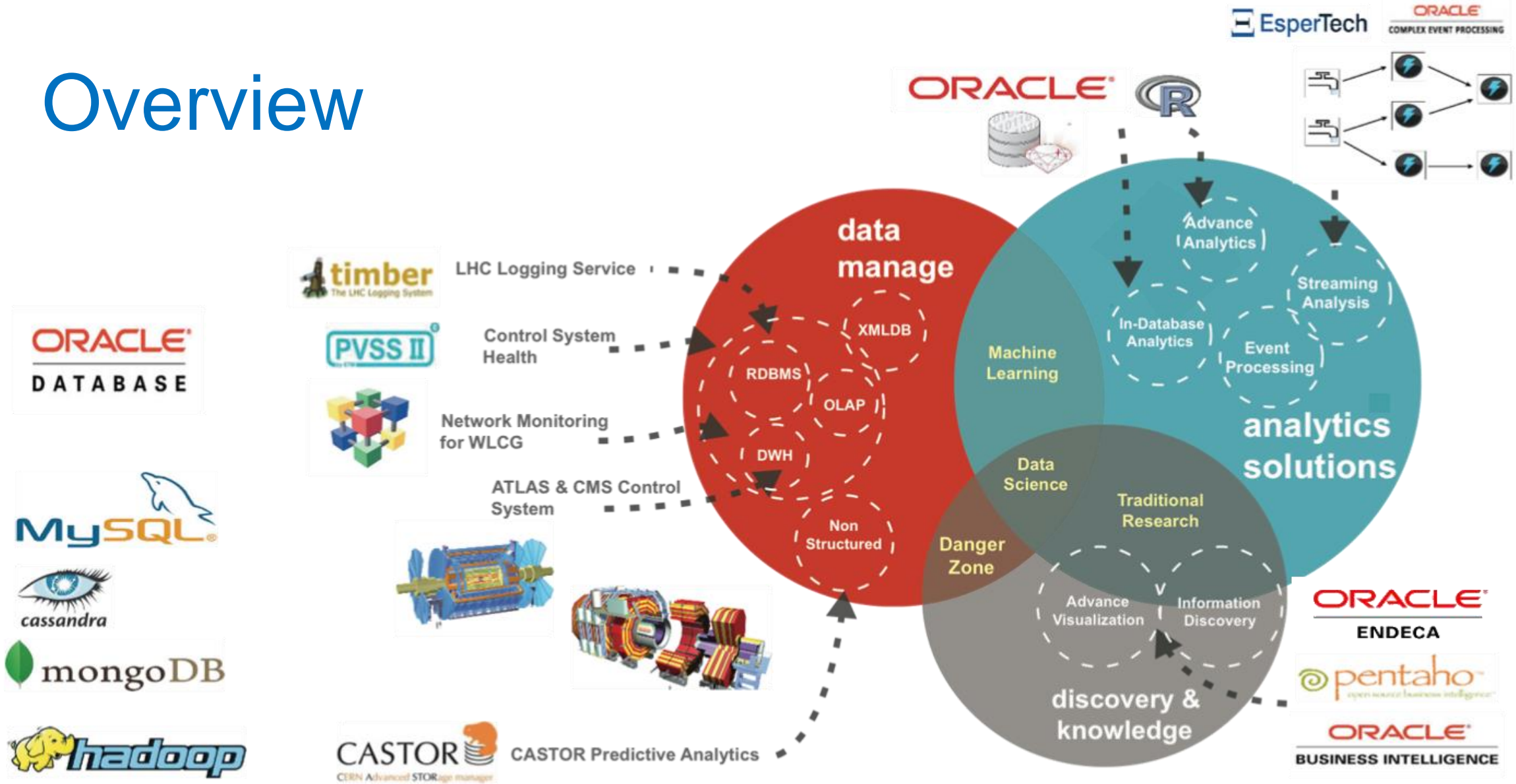
- Optimize our systems
  - Reducing and predicting faults and corrective interventions
  - Increase the availability and operations efficiency
- Profit from CERN data investment by using data analytics
  - Extract knowledge
  - Discover useful information
  - Suggest conclusions
  - Support decision making
- Control and Monitoring Systems
  - Proactive
  - Predictive
  - Intelligent



# Data analytic challenges at CERN

- Very dynamic and heterogeneous environment
  - Projects (number of projects)
  - Requirements (data analytic needs)
  - Technologies (problem-driven)
  - Data sources (raw, structured and unstructured)
- Large amount of data
- Education and Training
  - Users know the data and the questions
  - Help them how to connect both

# Overview



# Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

# What is R

- Language for statistical computing and graphics
- Standard and advanced statistical techniques
- Integrated suite of software facilities
  - Data manipulation
  - Calculation
  - Graphical display
- Free and Open Source



# Why R is good for data analysis

- Powerful
  - Advanced statistics
  - Plotting
- Extensible
  - Over 5800 CRAN packages extending base functionality
- Standard de facto for data analytics
- Great user community
  - Over 2 million R users worldwide
- Active development, frequently updated

# IDE for R

- Multiple IDEs for R
- RStudio
  - Open source version
  - Windows, Mac, Linux
  - Web (RStudio server)

A screenshot of the RStudio IDE interface. The main editor window displays R code for downloading and analyzing household power consumption data. The console window shows the execution of the code. The Environment pane on the right shows the 'data' object with 2880 observations and 10 variables. The Plots pane at the bottom right shows a histogram titled 'Global Active Power' with a red color scheme, plotting Frequency (0 to 1200) against Global Active Power (kilowatts) (0 to 6).

```
1 ##download and unzip the data
2 if (!file.exists("power_consumption.zip"))
3 {
4   download.file("https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2Fhousehold_power_consumption.zip",
5                 "power_consumption.zip")
6 }
7
8
9 unzip(zipfile = "power_consumption.zip", exdir = "data")
10
11 ##read data
12 data <- read.table(file = "data/household_power_consumption.txt", header = TRUE, sep = ";", na.strings = "?",
13                   colClasses=c("character", "character", "numeric", "numeric",
14                                 "numeric", "numeric", "numeric", "numeric", "numeric"))
15
16 ##convert to datetime and filter
17 data$Date <- as.Date(data$Date, "%d/%m/%Y")
18 data <- data[data$Date >= as.Date("2007-02-01") & data$Date <= as.Date("2007-02-02"),]
19
20 ##create new column with date and time
21 data$DateTime <- strptime(paste(data$Date, data$Time), "%Y-%m-%d %H:%M:%S")
22
23
24 ##open png and create the plot
25 ##png(filename = "figures/plot1.png", width = 480, height = 480)
26
27 hist(data$Global_active_power, col = "red", xlab = "Global Active Power (kilowatts)",
28      main="Global Active Power", ylim = c(0,1200))
```



# R example

```
> head(airquality)
```

```
  Ozone Solar.R Wind Temp Month Day
1    41    190  7.4   67     5   1
2    36    118  8.0   72     5   2
3    12    149 12.6   74     5   3
4    18    313 11.5   62     5   4
5     NA     NA 14.3   56     5   5
6    28     NA 14.9   66     5   6
```

## New York Air Quality Measurements

```
> summary(airquality)
```

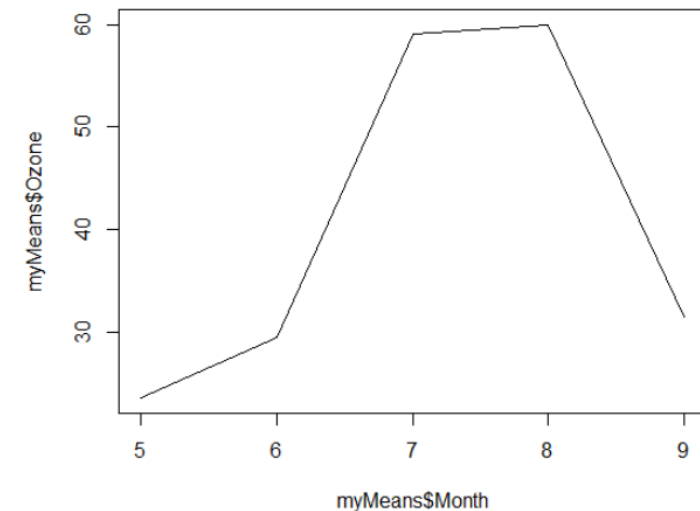
Ozone	Solar.R	Wind	Temp	Month	Day
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. : 56.00	Min. : 5.000	Min. : 1.0
1st Qu.: 18.00	1st Qu.: 115.8	1st Qu.: 7.400	1st Qu.: 72.00	1st Qu.: 6.000	1st Qu.: 8.0
Median : 31.50	Median : 205.0	Median : 9.700	Median : 79.00	Median : 7.000	Median : 16.0
Mean : 42.13	Mean : 185.9	Mean : 9.958	Mean : 77.88	Mean : 6.993	Mean : 15.8
3rd Qu.: 63.25	3rd Qu.: 258.8	3rd Qu.: 11.500	3rd Qu.: 85.00	3rd Qu.: 8.000	3rd Qu.: 23.0
Max. : 168.00	Max. : 334.0	Max. : 20.700	Max. : 97.00	Max. : 9.000	Max. : 31.0
NA's : 37	NA's : 7				

# R example

```
2
3 library(plyr)
4
5 ## subset needed data
6 myData <- airquality[,c("Month","Ozone","Solar.R","Wind")]
7
8 ##group by Month and calculate means
9 myMeans <- ddp1y(.data = myData,
10                .variables= .(Month),
11                .fun= numcolwise(mean, na.rm = TRUE))
12
13 ## show calculated dataset
14 myMeans
15
16
17 plot(x = myMeans$Month, y = myMeans$Ozone, type = "l")
18
```

myMeans

Month	Ozone	Solar.R	Wind
5	23.61538	181.2963	11.622581
6	29.44444	190.1667	10.266667
7	59.11538	216.4839	8.941935
8	59.96154	171.8571	8.793548
9	31.44828	167.4333	10.180000



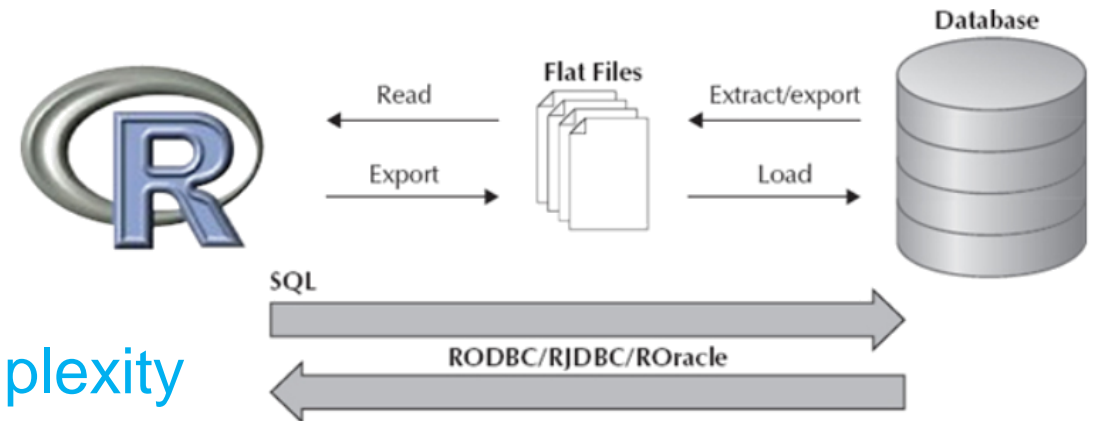
# Some R Resources

- CRAN (<http://cran.r-project.org>)
- R Tutorial (<http://www.r-tutor.com/>)
- R-bloggers (<http://www.r-bloggers.com/>)
  
- Free courses
  - Coursera - “Data Science” (Johns Hopkins University)
  - Lynda - “Up and Running with R”
  - O’Really - “Try R” (<http://tryr.codeschool.com>)
  
- Learn R, in R
  - Swirl (<http://swirlstats.com>)



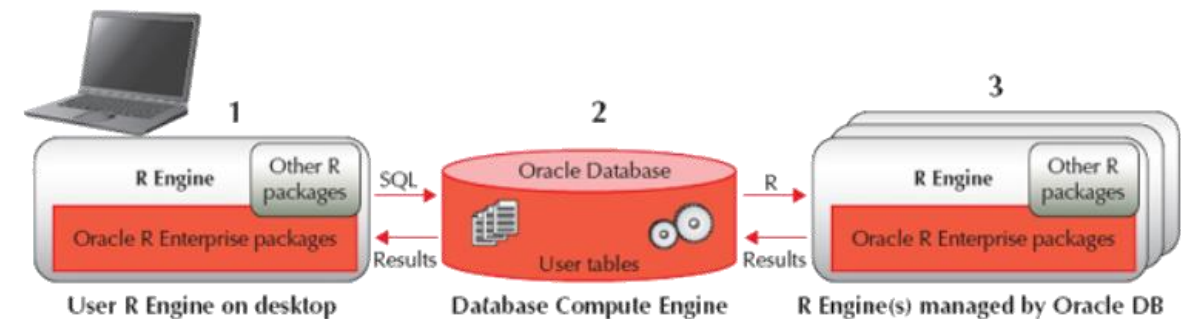
# Traditional R and Database Interaction

- Data has to be moved from database to client
  - Client need to store it in HDD
- R user needs to know SQL
  - Mixing R and SQL
- Not multithreaded or parallel
  - Use special packages increasing complexity
- R client has to load everything in memory



# Oracle R Enterprise

- A database-centric environment for analytical processes in R
  - Allows to use the database server to run R scripts
  - R working on data directly in the database
  - Integration with the SQL language
    - Run R scripts from SQL



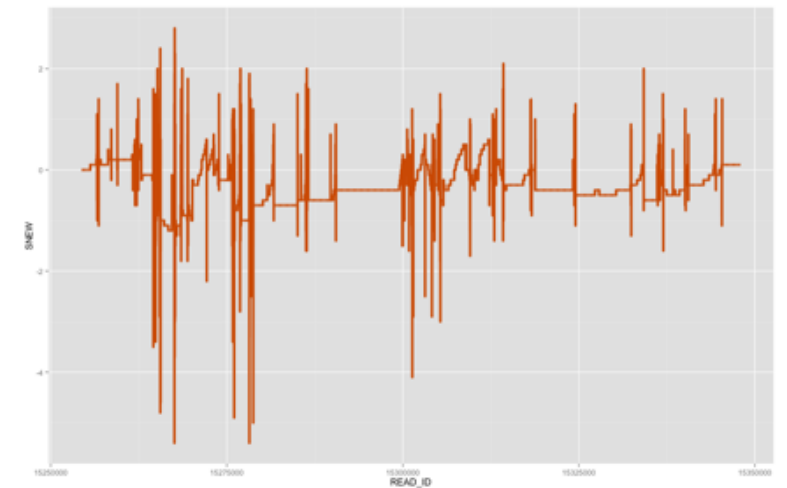
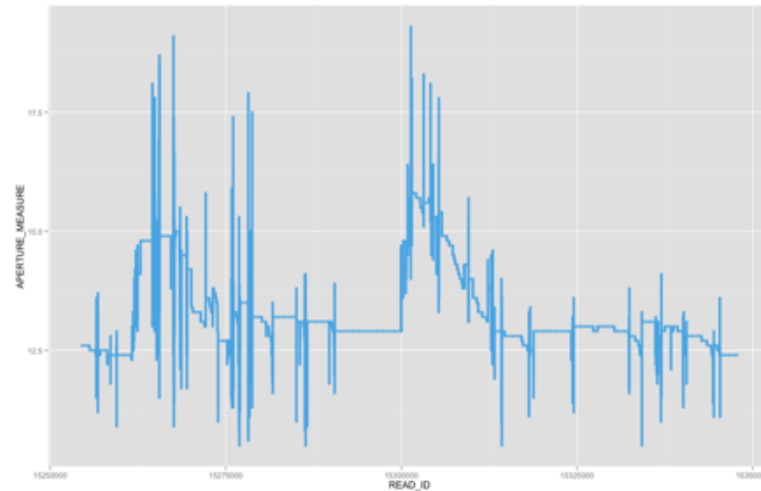
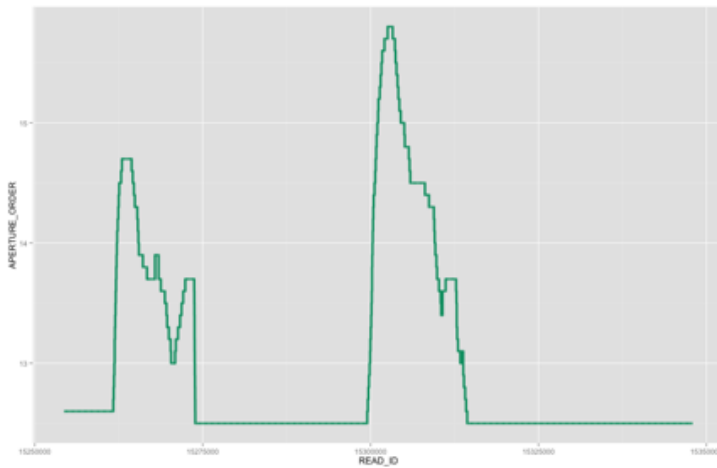
- Transparency Layer
  - Transparently interact with the database in R

# Benefits of using ORE

- Allows in-database data analysis
- Provide data parallelism and resource management
- Execute R scripts in database server machine
  - Scalability and performance
  - Eliminate memory constraint
- Oracle databases are widely used at CERN

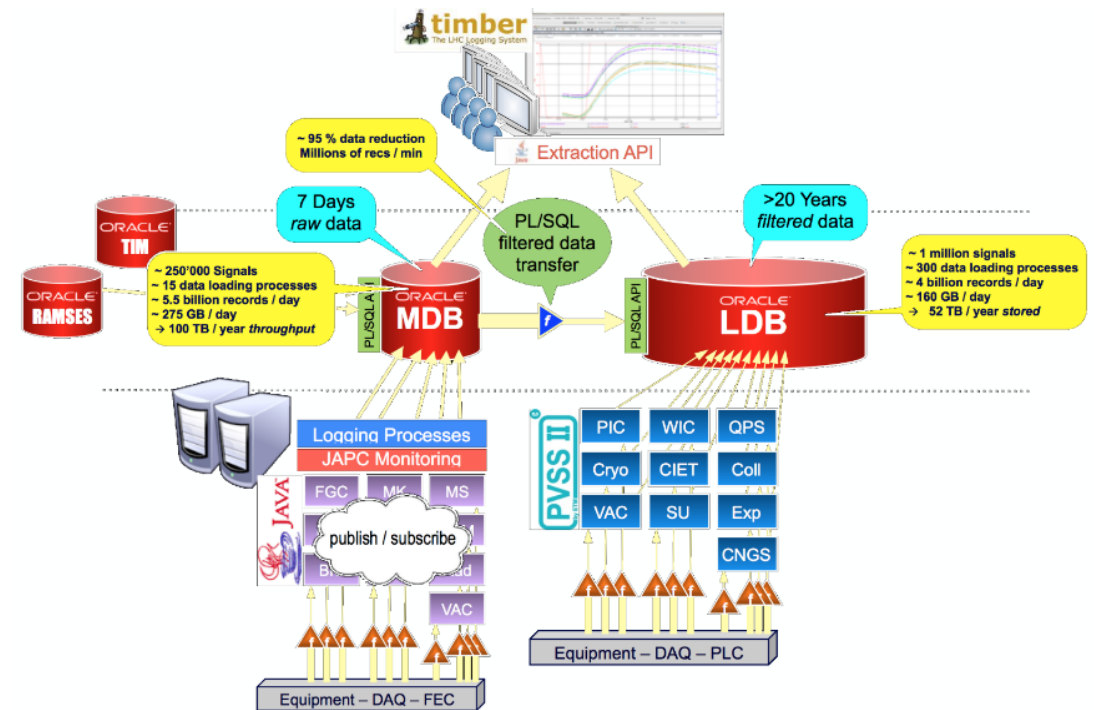
# Cryogenics Use Case: Faulty Valves Detection

- What is the objective?
  - Predict faulty valves before they actually fail
- How?
  - Valve receive an aperture order value (**aperture order**)
  - Effective aperture realized by the valve (**aperture measured**)
  - Analyzing the difference between both (**S = aperture order - aperture measured**)



# Cryogenics Use Case: Faulty Valves Detection

- Signals used
  - **S** = aperture order - aperture measured
- Features extractions based on **S**
  - Variance
  - Percentile 99.9
  - Rope distance
  - Noise Band
- **Automatic Faulty Valves Detection System**
  - SVM - Support Vector Machine
- The learning set 44 valves
- Three different status
  - Faulty,
  - Not faulty
  - Unknown
- Data comes from Accelerator Logging Service





# Summary

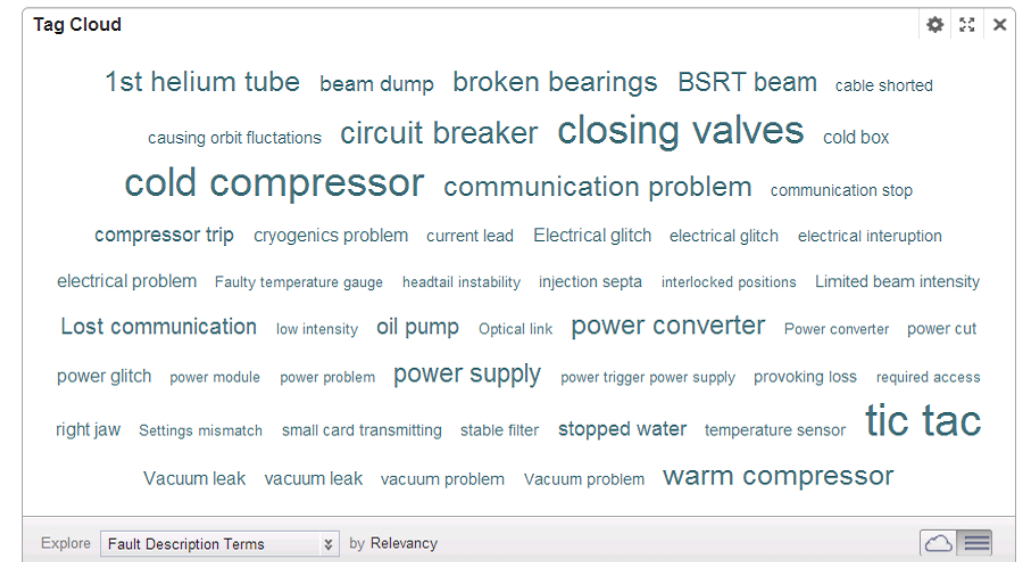
- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

# Data Discovery

- Interactive and visual analytics
- Intended to be used by the end users
  - Enabling them to use their intuition and knowledge of the data
- Powerful customization of dashboards and visualizations
  - Without intervention of IT
- Structure and unstructured data
- Mainstream field in Data Analytics

# Endeca Information Discovery

- Data discovery platform
  - Analyze information of any type and any source
- Flexible and user friendly
- Professional ETL tool
- Powerful text analysis
  - Sentiment
  - Text tagging, entities extraction
  - Multi language features



# Endeca Use Case

- Data Discovery for **improving** the Accelerator Complex Operations
- Electronic Logbook
  - Log of events in the accelerator complex

Statistics for the eLogbook: PS From: 20120815 Period: Morning To: 20120817 Period: Morning

Start Statistics ...

Availabilities			
Lines	In Super Cycle	In Fault	Availabilities (%)
AD	55 [h]	7 [h] 57 [min] 37[s]	85%
EASTA	55 [h]	7 [h] 57 [min] 37[s]	85%
EASTB	55 [h]	7 [h] 57 [min] 37[s]	85%
SFTPRO	6 [h] 54 [min] 29[s]	5 [h] 45 [min] 18[s]	16%
EASTC	55 [h]	7 [h] 57 [min] 37[s]	85%
CNGS	55 [h]	10 [h] 22 [min] 58[s]	81%
LHC PROBE	55 [h]	7 [h] 57 [min] 37[s]	85%
I_LHC	31 [h] 31 [min] 33[s]	4 [h] 37 [min] 40[s]	85%
LHC	55 [h]	7 [h] 57 [min] 37[s]	85%
TOTAL	423 [h] 26 [min] 02[s]	68 [h] 31 [min] 38[s]	83%

Systems for AD		
GROUP NAME	FAULT NAME	DURATION
PS	Power supply	3 [h] 47 [min] 34[s]
PS	RF	3 [h] 13 [min] 33[s]

4	23:42	SUP	Global Post Mortem Event Event Timestamp: 10/06/12 23:42:39.163 Fill Number: 2718 Accelerator / beam mode: PROTON PHYSICS / STABLE BEAMS Energy: 4000080 [MeV] Intensity B1/B2: 15509 / 14217 [e^10 charges] Event Category / Classification: PROGRAMMED_DUMP / MULTIPLE_SYSTEM_DUMP First BIC input Triggered: First_USR_PERMIT change: Ch 1-Programable Dump bi: A T -> F on CIB.CCR.LHC.B1
5	23:42	SUP	Global Post Mortem Event Confirmation Dump Classification: Programmed Dump Operator / Comment: papotti / End of physics fill, clean dump.
6	23:42	SUP	<b>BEAM MODE &gt; BEAM DUMP</b> LHC RUN CTRL: BEAM MODE changed to BEAM DUMP
7	23:42	SUP	<b>BEAM MODE &gt; BEAM DUMP</b> LHC RUN CTRL: BEAM MODE changed to BEAM DUMP
8	23:42	SUP	ELOGBOOK: STARTING B1 MKISS
9	23:43	SUP	ELOGBOOK: STARTING B2 MKISS
10	23:44	SUP	LHC SEQ: beam dump handshake closed; LHC=STANBY, EXP=VETO
11	23:44	SUP	LHC SEQ: MCS checks finished
12	23:45	SUP	LHC SEQ: SMP pre-operational checks finished
13	23:45	SUP	LHC SEQ: BIS pre-operational checks finished
14	23:48	SUP	<b>BEAM MODE &gt; RAMP DOWN</b> LHC RUN CTRL: BEAM MODE changed to RAMP DOWN
15	23:48	SUP	LHC SEQ: BPMLHC calibration finished. Overall result: SUCCESS Chosen bunch spacing: (B1 & B2) BUNCH_50NSEC (manually chosen) (For more details see BI-LHC ELogBook)

# Endeca Use Case

- Electronic Logbook
- Endeca PoC

# DEMO

The screenshot displays the Endeca Information Discovery LogBook EID interface. At the top, navigation tabs include Events, Faults, Operational Issues, FEC & Devices, and Operational Modes. A search box and a 'Selected Refinements' section are visible on the left. The main area shows a 'Summarization Bar' with the following statistics:

- Total Events: 493,478
- LHC Events: 38.84%
- PS Complex Events: 21.62%
- SPS Events: 12.54%
- History Events: 3.57%
- ISOLDE Events: 4.25%

Below the summarization bar is a 'Tag Cloud' for 'Faulty SIS nodes' with various tags such as 'Analysis Context', 'beam control SPS', 'BEAM MODE', 'beam mode', 'BEAM SETUP', 'bunch spacing', 'check result summary', 'collimators warnings', 'converter fault', 'dump protection', 'energy thresholds', 'Event category', 'Failed channels', 'Failed devices', 'fast extraction', 'filling period', 'flat top', 'frequency checks', 'inj handshake status', 'injection handshake', 'injection handshake starting', 'injection protection coll', 'injection settings', 'interlocks reset', 'issue report', 'main circuit', 'Missing devices', 'NONE Missing devices', 'Operator Buttons', 'Overall analysis result', 'Overall result', 'parking starting', 'power supply', 'POWERING FAILURE', 'pre-operational checks', 're-arm circuit', 'Reset details', 'ring cleaning coll', 'stable beams', 'standby service', 'successful non-interactive sequence', 'Superlock circuit', 'Test PIC2', 'timing user', 'transfer line coll', 'unsuccessful contact', 'valid information', 'XPOC error', and 'XPOC PRO'.

A 'Chart' section displays a stacked bar chart titled 'Event Id (unique values) by Complex, Operational Mode'. The x-axis lists complexes: LHC, TESTS, PS Complex, SPS, ISOLDE, HISTORY, BE-BI, UA9, TE-ABT, TE, RF, SHUTDOWN, TE-VSC, EN-ICE, R2E, and AT-MCS. The y-axis represents 'Event Id (unique values)' from 0K to 240K. The legend includes operational modes: BEAM IN, Commissioning, LHC\_MASTERSHIP, NO BEAMS, RUN, SETTING\_UP, SETUP, SPS\_MASTERSHIP, STABLE BEAMS, Stand-by, Unselected line, and UNSTABLE BEAMS.

At the bottom, a 'Results Table' shows 0 records selected. The table has columns: Complex, Logbook, Operational Mode, Line Name, Event Date (Year-Mo...), Event Comment, Fault Groupname, Fault Name, and Shift Start Date (Year...). The visible rows are:

Complex	Logbook	Operational Mode	Line Name	Event Date (Year-Mo...)	Event Comment	Fault Groupname	Fault Name	Shift Start Date (Year...)
AT-MCS	Test Facility	Unselected line		2/9/10	Vaccum leak IA3 since t...			2/8/10 1
AT-MCS	Test Facility	Unselected line		2/10/10	Mesurements on IP SOL...			2/9/10 C
AT-MCS	Test Facility	Unselected line		2/10/10	Solenoid Central. 400A ...			2/9/10 C
AT-MCS	Test Facility	Unselected line		2/10/10	Sol Cen. 400A. 10.02V ...			2/9/10