Background: Massive and complex data

1. Massive monitoring and sensor technology - everywhere
2. Storage
3. Data distribution
4. Good understanding of main effects: now analyse rare effects, extremes, interactions, weak signals. For this we need much data!
5. Industry is more and more exploratory: collect ALL data “in hope” ...

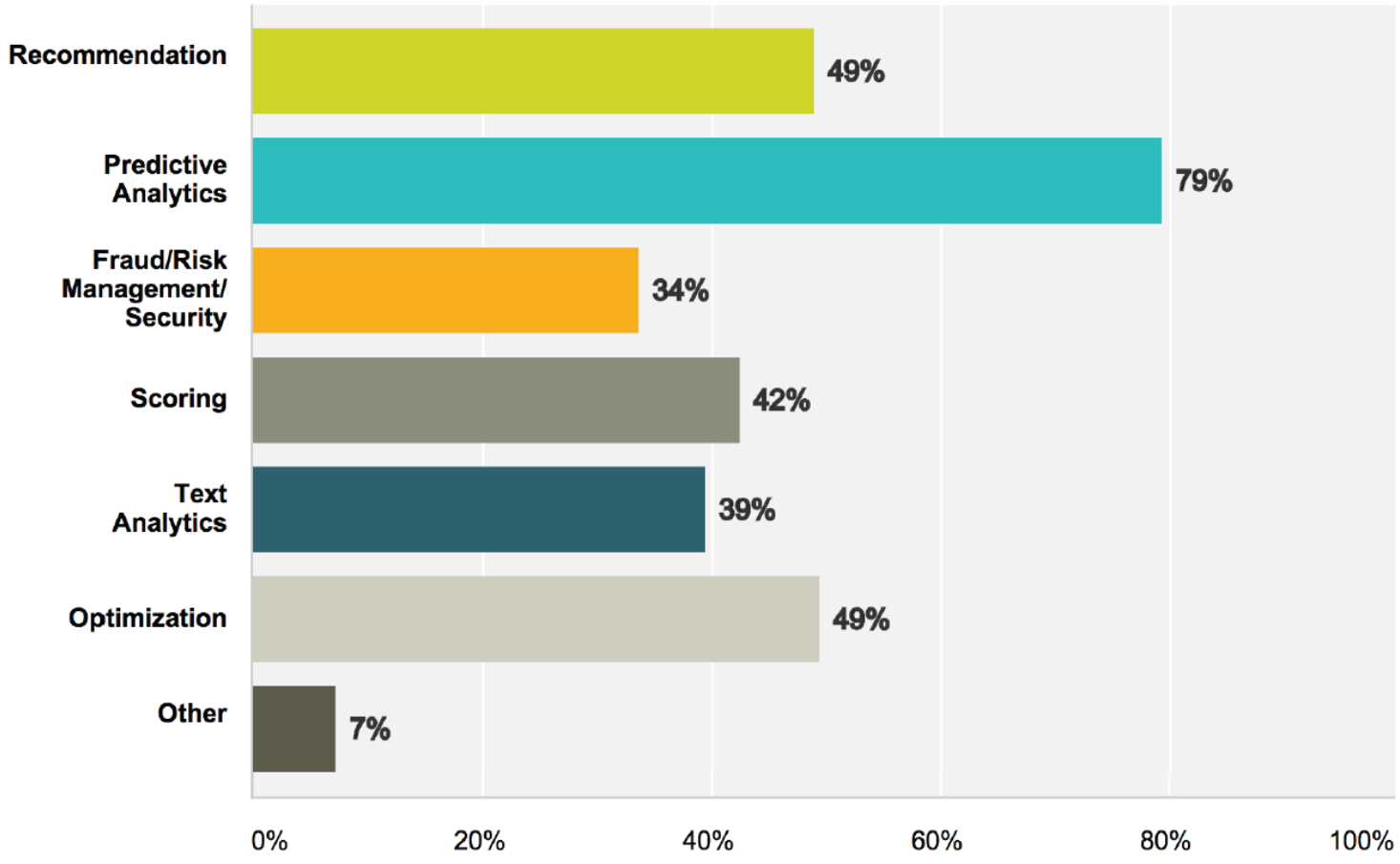
McKinsey Global Institute




June 2011

Big data: The next frontier for innovation, competition, and productivity

What does the industry do with big data?



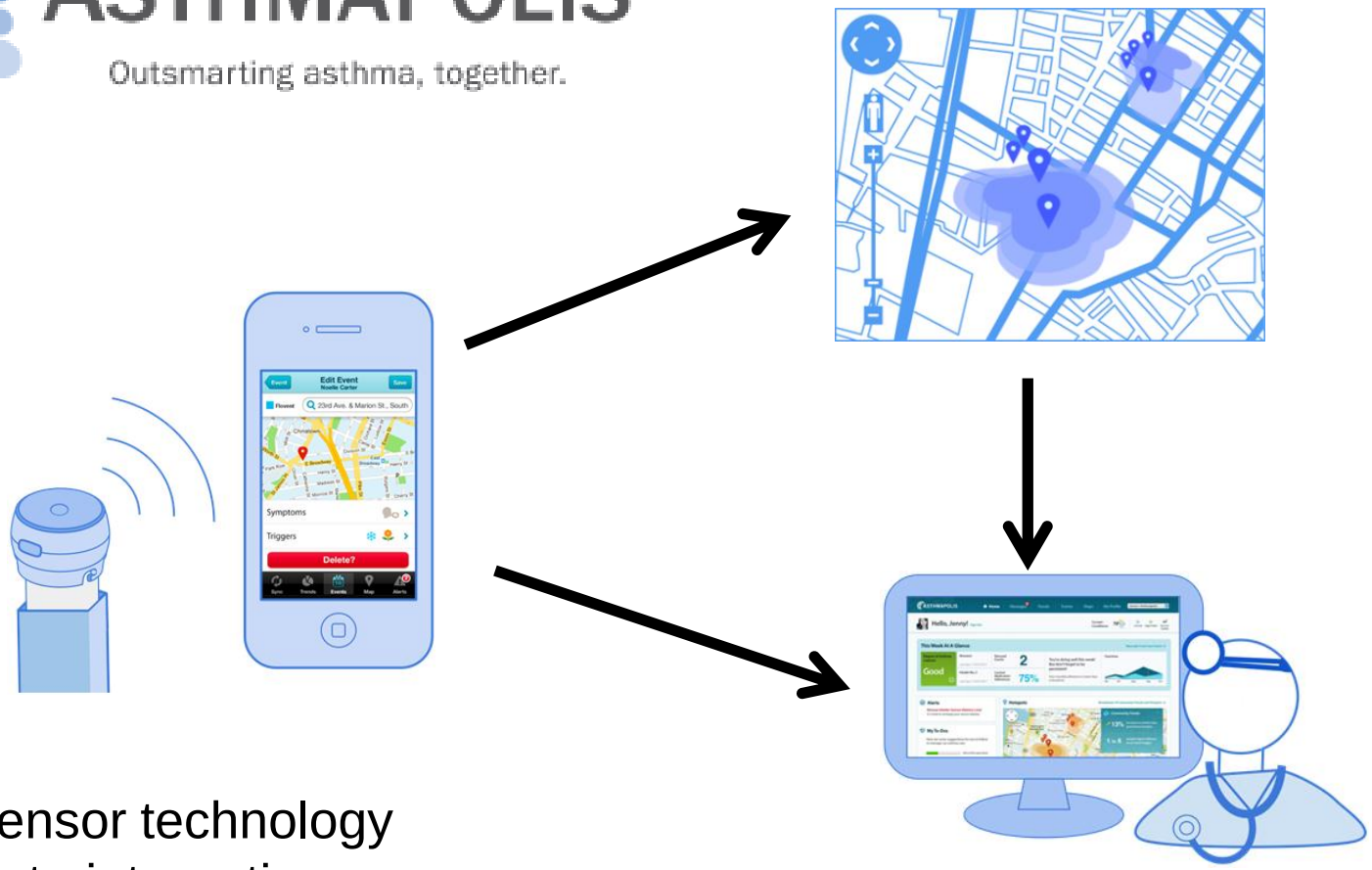
Big data versus lots of data

Before	Now
Instead of relying on a sample of data to predict outcomes ...	... we can now use **all** the data.
Instead of analysing data with no time constraints ...	... we need now to do **real time** analysis
Data were homogeneous, uni-scale, uni-type, structured data bases ...	... now we have to analyse **unstructured** data, emails, twitter, video, images, networks, sensor data.

Lot's of data - all data! - does not mean problem solved:
If we are investigating the mechanisms of a drug, having all breast cancer patients does not mean that we know the answer.

ASTHMAPOLIS

Outsmarting asthma, together.



- sensor technology
- data integration
- real time & post analysis
- personalised (real time) prevention and therapy

A remote monitor embedded in insulin pen caps could help personalize diabetes treatment

June 3, 2013 9:41 am by Stephanie Baum | 0 Comments

The Sanofi US competition to encourage companies to devise new ways to improve outcomes and lower healthcare costs associated with diabetes has its demo day today. In a contest that generally revolves around various health IT approaches, one company has developed a remote monitoring device to generate big data using the cap of an insulin pen.

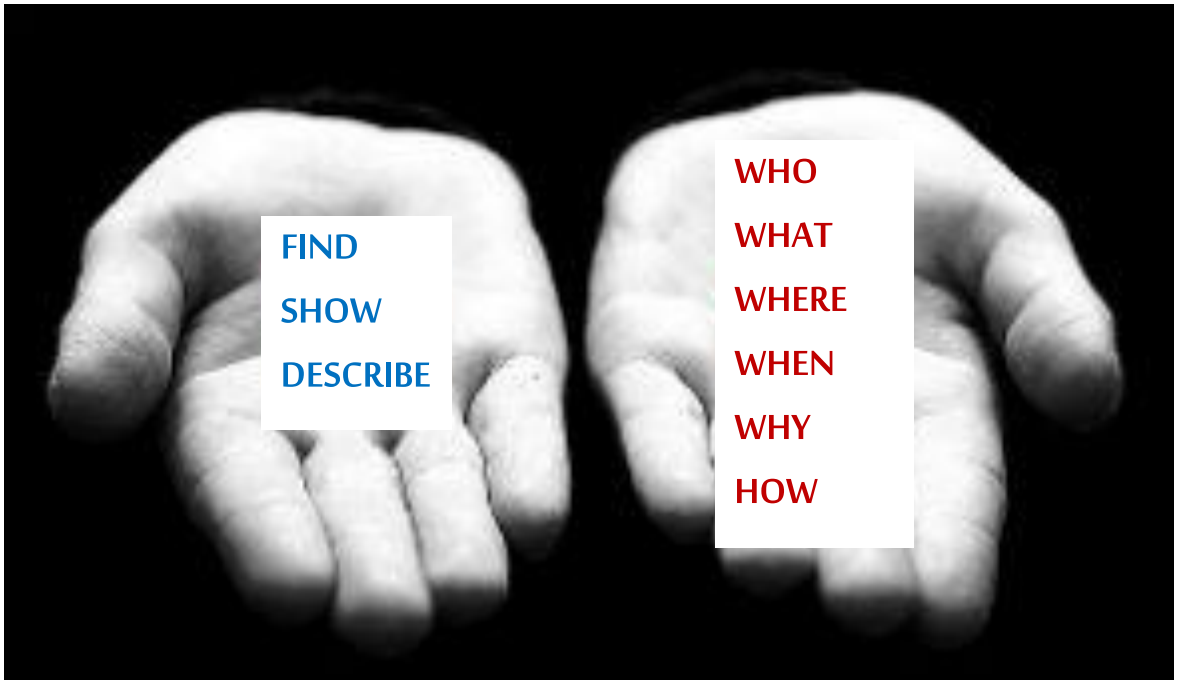
Cambridge, Massachusetts-based Common Sensing produced the GoCap — a replacement cap for prefilled insulin pens that records the level of insulin administered daily and the times it was taken. It transmits that information using Bluetooth to a mobile phone or connected glucometer. The idea is to provide a steady stream of relevant information transmitted in an easily digestable format to alert healthcare professionals to potential problems early enough before they require hospitalization and ramp up healthcare costs.

James White co-founded Common Sensing with Richard Whalley, both MIT graduates. In a phone interview with MedCity News, White — who is also the chief technology officer — said one reason for starting the company is that the data to gauge risk in diabetes patients just isn't there.

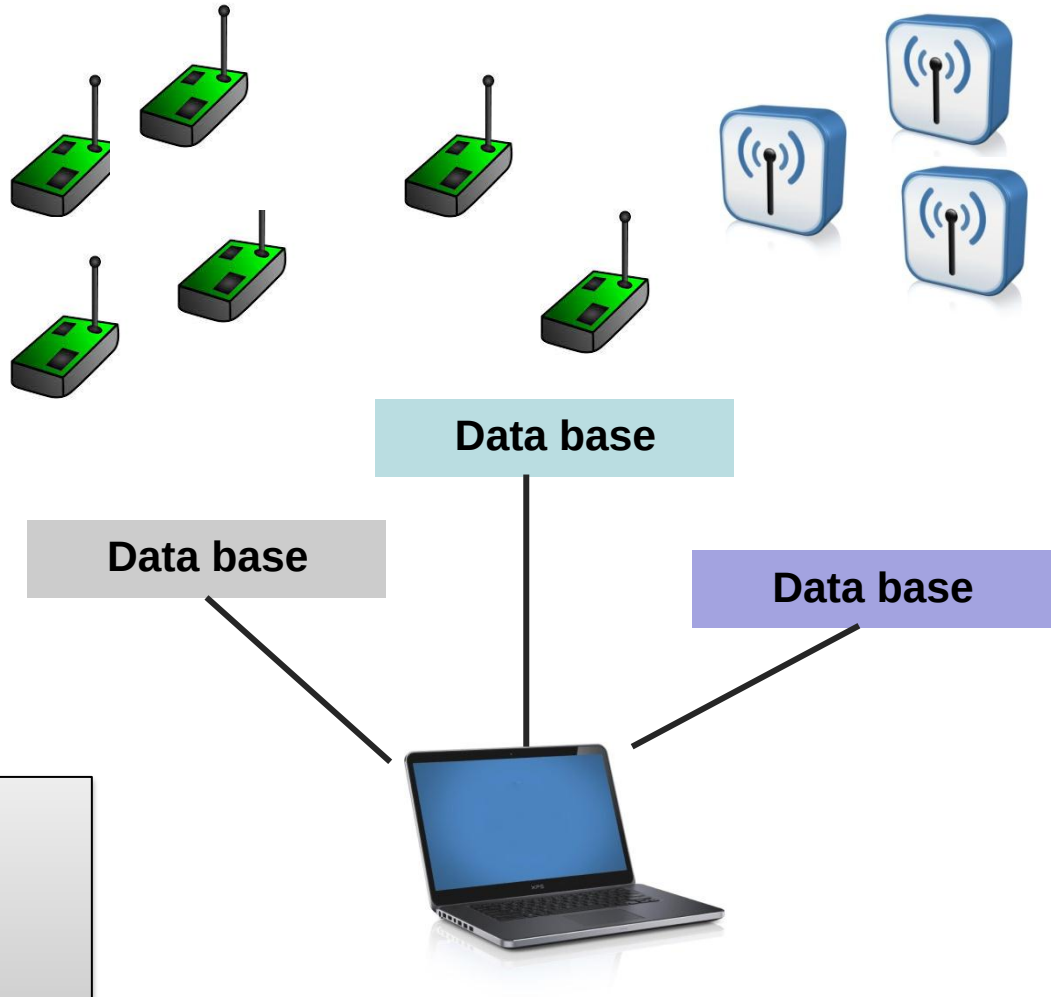
TransCelerate BioPharma will develop shared industry research and development solutions to simplify and accelerate the delivery of innovative products to patients, … enabled by broad participation and collaboration across the global research and development community."

- combine resources
- standardise
- integrated analysis – borrowing strength across similar cohorts and compounds, based on chemistry, biology, epidemiology

Big Data's two faces



Data engineering



- database management
- information engineering
- information retrieval
- ontologies & semantics
- structured and unstructured

Find, visualise, summarise

Big Data: exploratory, descriptive

Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips

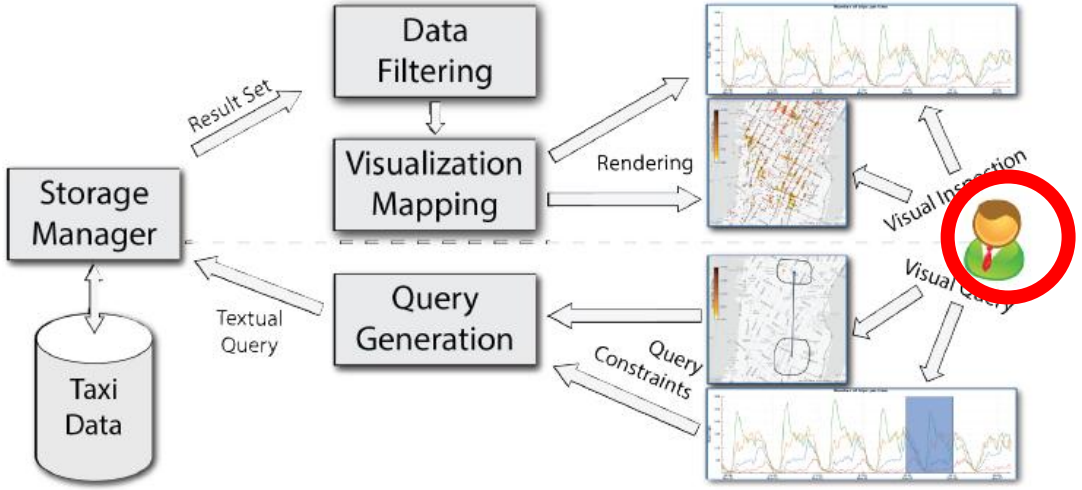
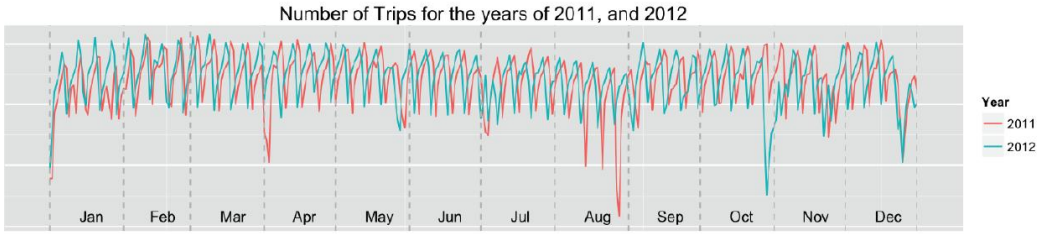


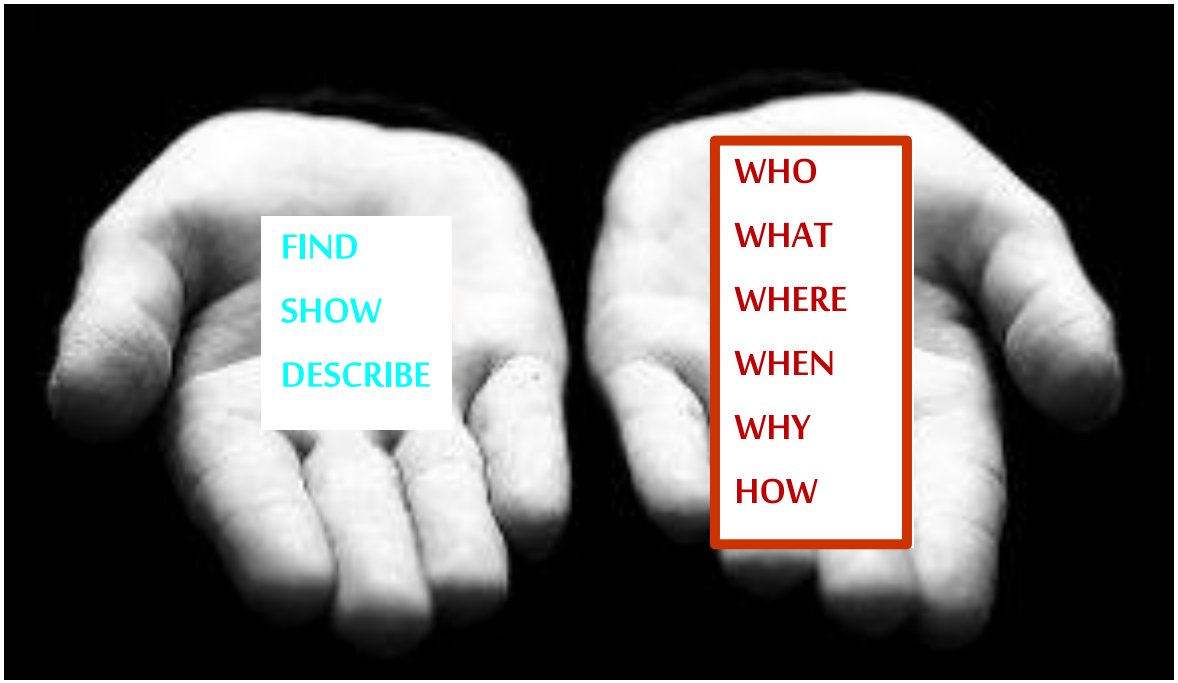
Fig. 3. High-level architecture of TaxiVis.



Inspection tool

Ferreira et al IEEE Trans Visual Comp Graph 2013

The statistical science side

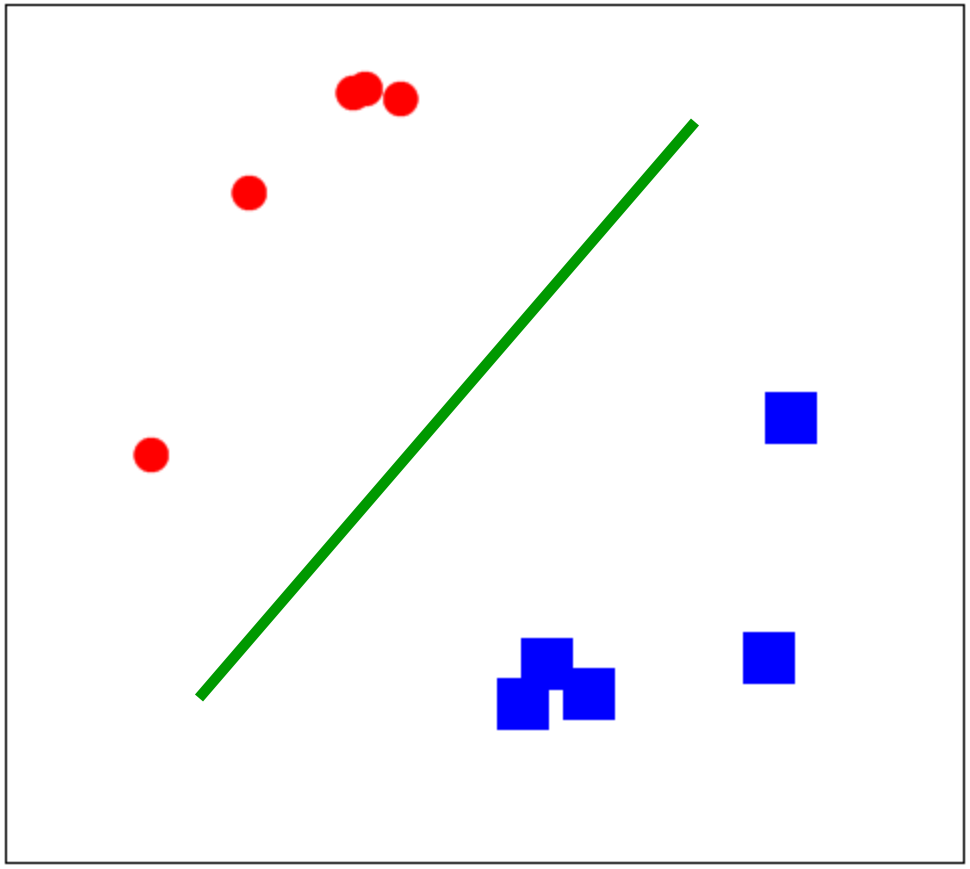


Inference for big data

Exploiting knowledge

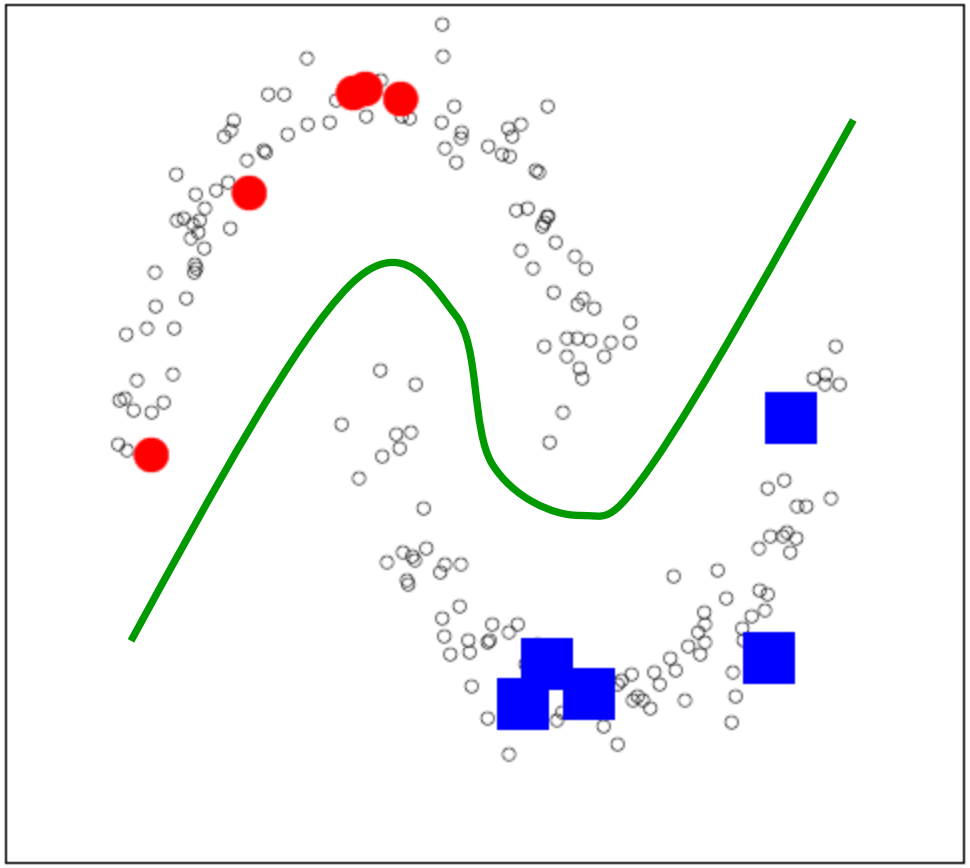
horizontal, generic, all purpose "SAS-type" methods	model based methods that exploit knowledge and structure in new data contexts
raw data exploration passive data collection	controlled data collection modelling to extract knowledge

We need to turn traditional and new methods of statistical modeling and inference, into powerful discovery tools, able to handle large quantities of complex data, often in real time.



- **Data (2 dimensions) with classification known.**
- **Supervised**
- **Classification rule.**

We can use the unlabeled data to estimate the decision boundary accurately.



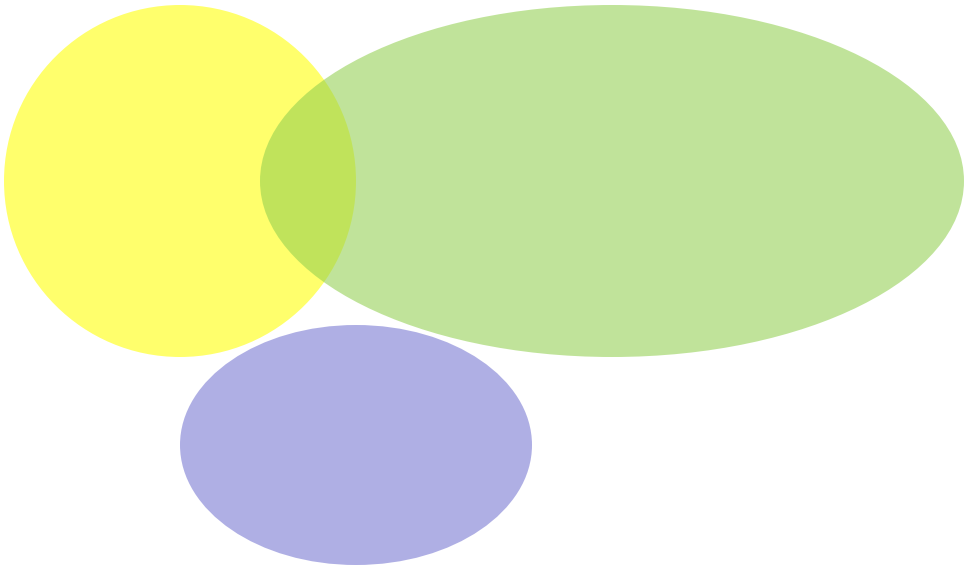
- **Classified and not classified data**
- **Semi-supervised**
- **Makes a much better rule (less false positives/negatives)**
- **Assumption: P(Y = 1 | X = x) is smooth relative to the clusters → Model!**

Run ten million regressions, with very few repetitions each:

$$i = 1, \ldots, p = 1000000 \quad \text{individuals/units}$$

$$\text{j}=1,\ldots,\text{n}_i = 5, \quad \text{observations per unit}$$

$$\text{y}_{i,j} = \alpha_i + \beta_i \text{X}_{i,j} + \varepsilon_{i,j}$$



- clustering of units
- with overlap
- changing in time

- supervised borrowing information

Sequential, real time multi-sensor change-point detection

$i = 1, \ldots, p = 1000$ - 1000000 sensors, patients, web-based indexes

$y_i(t)$ data streams, observed in time $t = 1, 2, \ldots$

At an unknown time τ, there are changes in a subset A of sensors.
A is small but unknown and the size of the change also unknown.

Detect the change, keeping probability of false allarms as low as possible.

Structured or un-structured modelling of the sensors.

- *There are new possibilities in using all data*
- Inferential uncertainty is fundamental in decision making
- Time series data allow intervention in real time.
- Causal effects enable effective actions.
- Rapid computations of solutions are required.