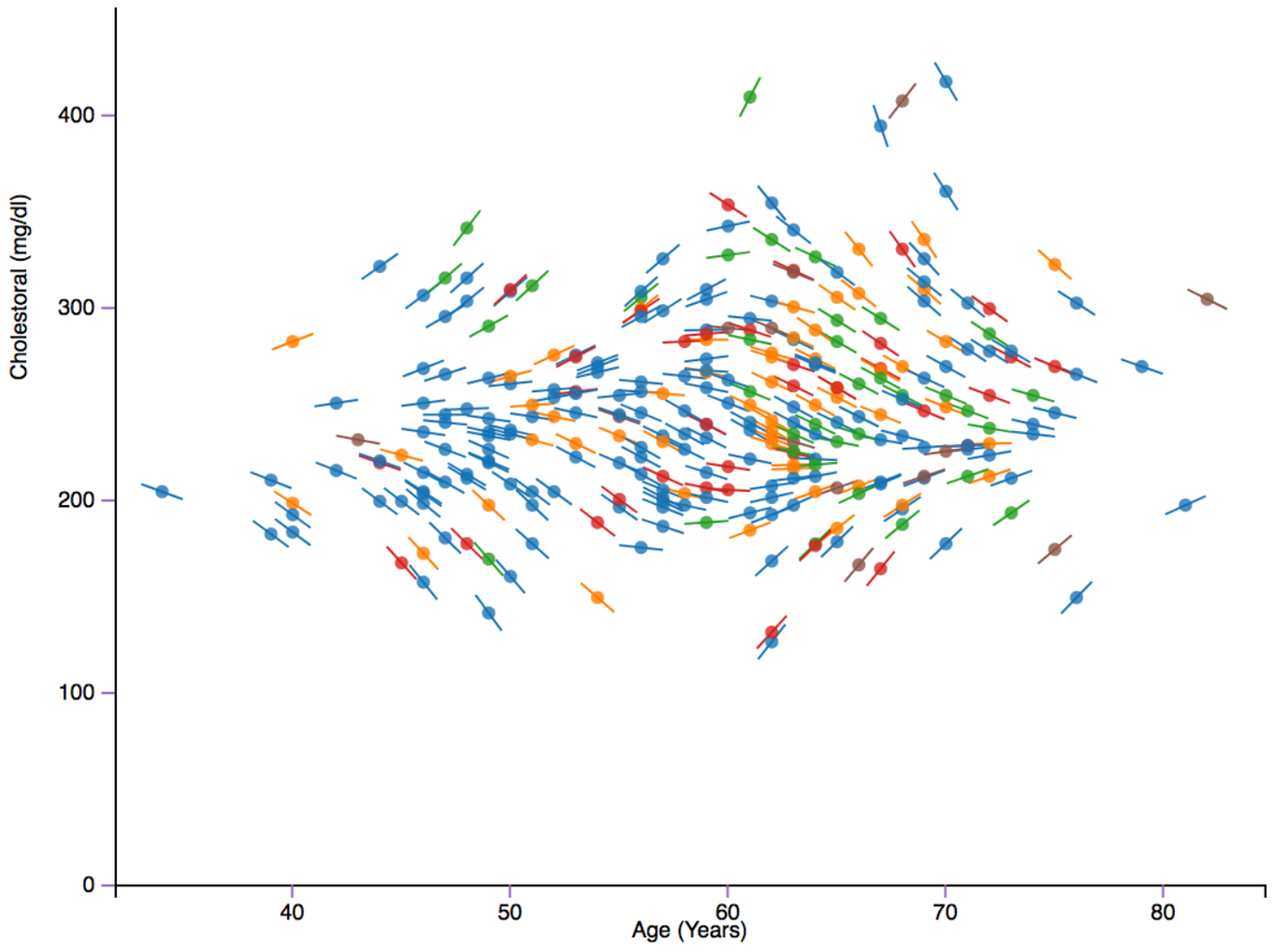
Flow-based Scatterplot



Sebastián Ricke
Jose Pablo Domínguez
Guillermo Espinoza

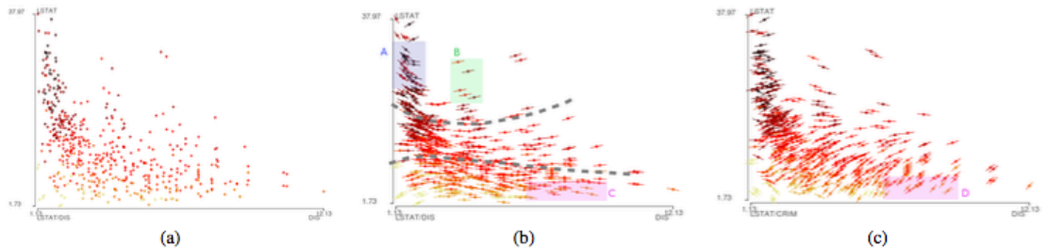


Figure 1: (a) Traditional scatter plot between two variables (b) Sensitivity visualization of the same two variables, where data points are augmented with derivatives. (c) Sensitivity visualization for a third variable, useful for analyzing tri-variate correlations.

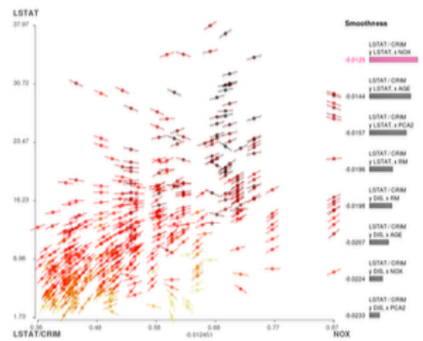


Figure 2: Smoothness ranking view. Next to each augmented scatterplot, we show the smoothness ranking of other variables.

bound by the inherent loss of information that occurs when projecting high dimensional data into a 2D space. In our case, we can plot the sensitivity of another variable with respect to one of the scatterplot axes. If we show the derivatives of a variable with respect to another variable (different from the ones used in projection), then we can begin making queries and formulating hypotheses about tri-variate correlations, instead of bi-variate queries that are typical of 2D scatterplots. An example is shown in Figure 1(c), where we show the same data points as before, using two variables named LSTAT and DIS, but we plot the sensitivity of the variable in Y (LSTAT) with respect to another variable named CRIM. Note that, although the data points have the same location in the X-Y plane, the sensitivities differ. We immediately have a different sense of flow, which changes the way we begin to formulate hypotheses about the three variables. For example, we see that, for points in region D in Figure 1(c), variable CRIM increases as DIS increases, but the same cannot be said about LSTAT, which only seems to increase when LSTAT is larger and decreases when LSTAT is low. Therefore, we may regard sensitivity derivatives as another attribute of nodes that represents relationships between two particular variables. Sensitivity derivatives of U with respect to V shows the relationship between U and V for each data point, and the projection variables (X, Y) decide where to locate these nodes of such derivative attribute. Some particular projections might place these nodes in a way that show global trends and correlations between variables U and V, which helps us understand the relationship between both U and V, and X and Y. In this paper, we show a number of operations, based on flow analysis, to help us identify these relationships.

3.1 Computing Sensitivities

As described before, there are different ways to compute the sensitivity of one variable with respect to another. In this paper, we follow a variational approach, where the sensitivity can be approximated by the partial derivative of one variable with respect to another. Since we do not know the analytic closed form of the function between two variables in the general case, we approximate the partial derivatives using linear regression. Because we do this in different neighborhoods around each point, we employ the method of moving least squares. We obtain the partial derivatives of a variable y with respect to x considering the Taylor approximation of y around a given point (x_0, y_0):

$$y_i = y_0 + \frac{\partial y}{\partial x}(x_i - x_0) \tag{1}$$

Then, we approximate the partial derivatives for point (x_0, y_0) in a neighborhood of N points, as:

$$\frac{\partial y}{\partial x} \approx \frac{\sum_{i=0}^{N}(y_i - y_0)(x_i - x_0)}{\sum_{i=0}^{N}(x_i - x_0)^2} \tag{2}$$

With this information, we augment the scatterplot using tangent line segments on each data point. Each tangent line is computed as follows. For a given point (x_0, y_0), we trace a line between points $(x_0 - \delta vx, y_0 - \delta vy)$ and $(x_0 + \delta vx, y_0 + \delta vy)$, where $(vx, vy) = normalize(1, \frac{\partial y}{\partial x})$ and δ is a parameter that controls the length of the tangent lines.

In our experiments, we compute the neighborhood of N points as an isotropic region around each point of a radius W. This radius controls how local or global is the flow. When W is small, the derivatives capture the local variability of data and reveal localized trends. On the other hand, when W is large, the flow represents the global trend in the data. An example is shown in Section 5.2. The variable width helps us reveal local trends where the global correlation is low. Instead of making an automatic decision in terms of correlation, flow-based scatterplots offer the option to the analyst to explore the spectrum of trends and correlations interactively.

3.2 The Smoothness of a Flow Scatterplot

As can be seen from Figures 1(b-c), sensitivities provide a sense of *flow* of the data points in the projection space. This flow helps reveal overall trends. For some projections, these sensitivities show certain critical regions, where linear trends coincide at some point but then diverge. This means that data points in that region can either go up or down, possibly depending on other variables. This suggests that this particular projection is hiding a lot of complexity that may be identified through a different projection. To measure the complexity of a flow-based scatterplot, we turn to second derivatives, which tell us how fast the tangent lines change in a

Idiom basado en el paper *Flow-based Scatterplots for Sensitivity Analysis* publicado por los autores Yu-Hsuan Chan, Carlos D. Correa y Kwan-Liu Ma de la universidad de California en Davis

Vista del paper

Problema

- Entender correlaciones y tendencias en un un dataset complejo que implica múltiples dimensionalidades

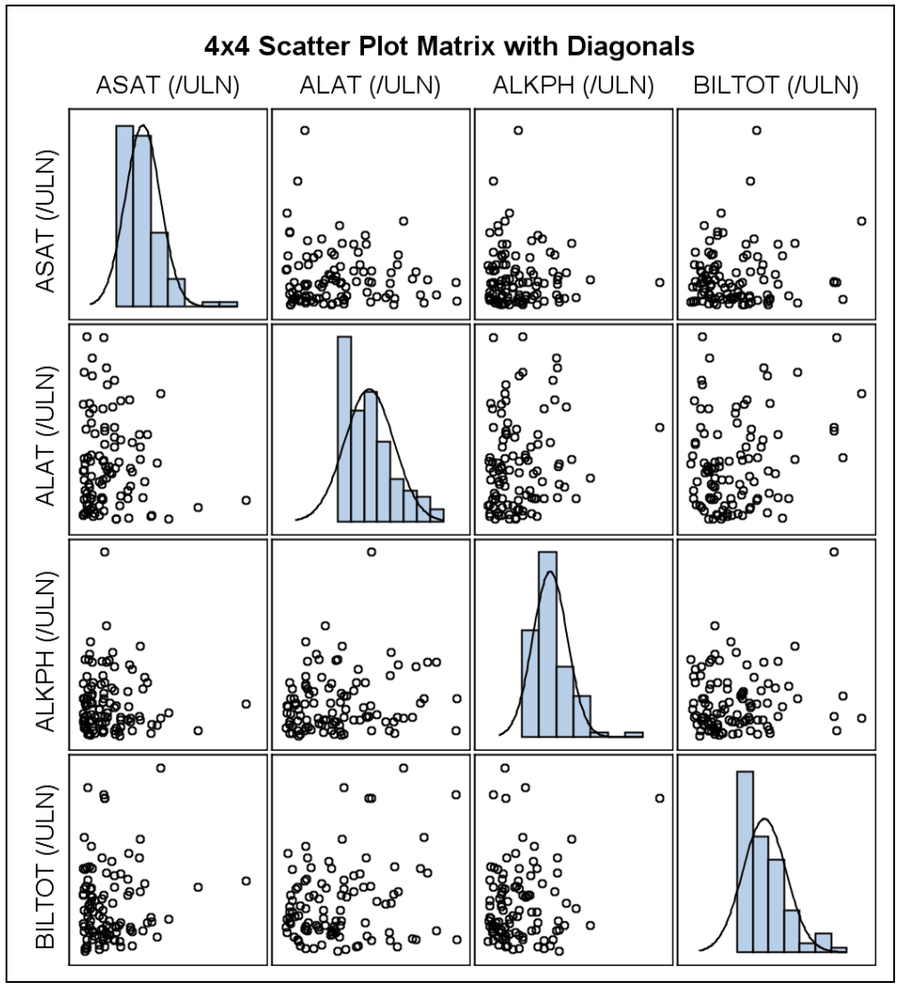
- La sensibilidad de un dataset, es decir, la tasa de cambio entre una variable y otra, puede ayudar al análisis

- Detectar más fácilmente los factores más importantes que influyen en la correlación y sensibilidad de los datos

Contexto

No se ha logrado encontrar un análisis multidimensional que permita simultáneamente:

- Evitar en la mayor medida posible la pérdida de información al proyectar los datos
- Extender el número limitado de variables que se pueden visualizar en un scatterplot
- Una visualización interactiva que facilite el aprendizaje y los objetivos del usuario.

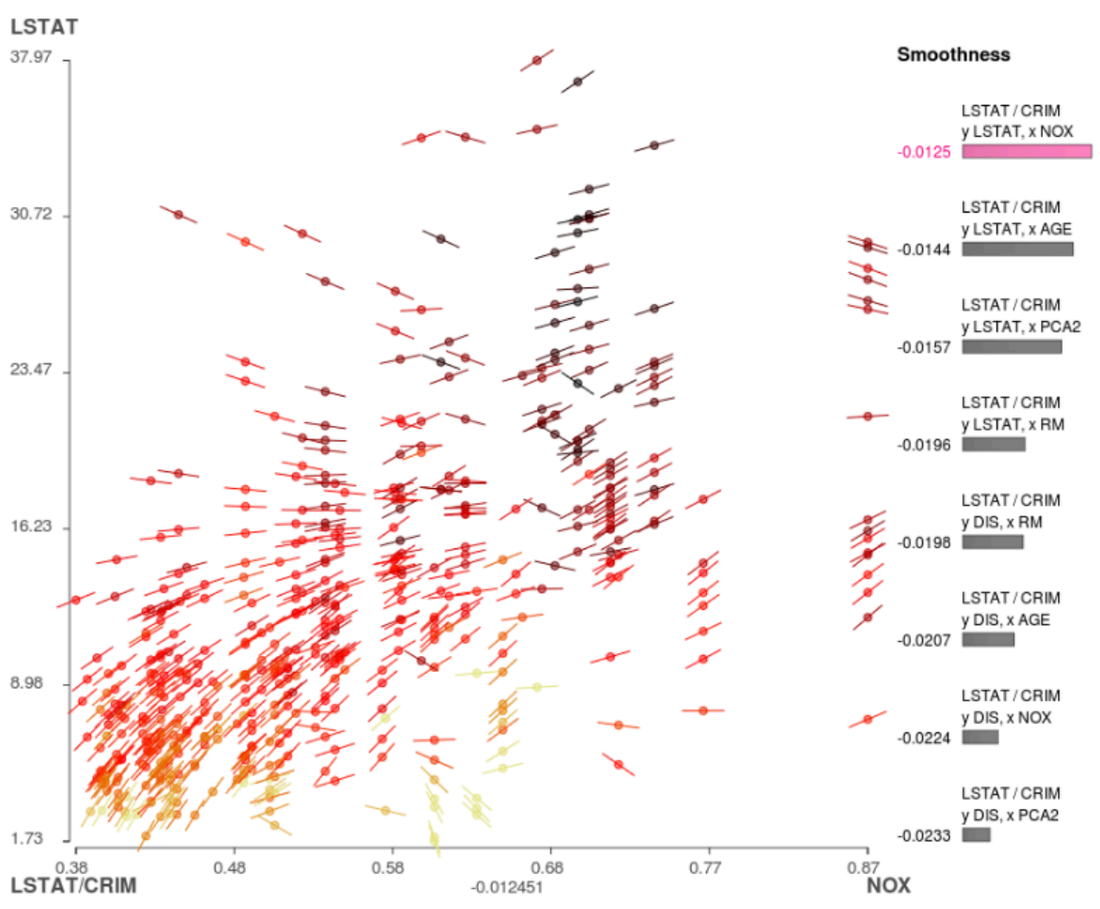


Beneficiados

- Investigadores que no necesariamente tienen conocimientos que le permitan realizar un análisis de correlación y sensibilidad en un dataset multivariable.
- Estudiantes que buscan profundizar sus conocimientos en visualización analítica

Dificultades

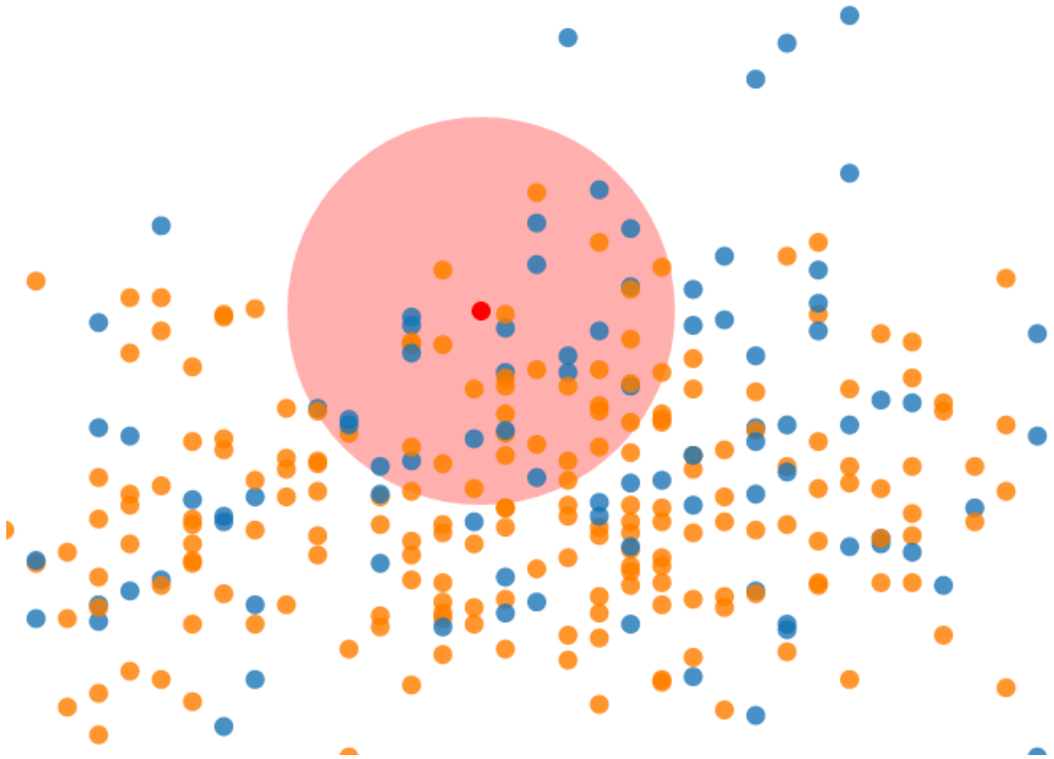
- Funciones no nativas en d3
- Procesamiento computacional



Vista presentada por los autores en el paper

Funciones

- Cálculo de derivadas
- Cálculo smoothness
- Selección dinámica del radio de la derivada



Radio de la derivada

$$\frac{\partial y}{\partial x} \approx \frac{\sum_{i=0}^{N}(y_i - y_0)(x_i - x_0)}{\sum_{i=0}^{N}(x_i - x_0)^2}$$

Derivada

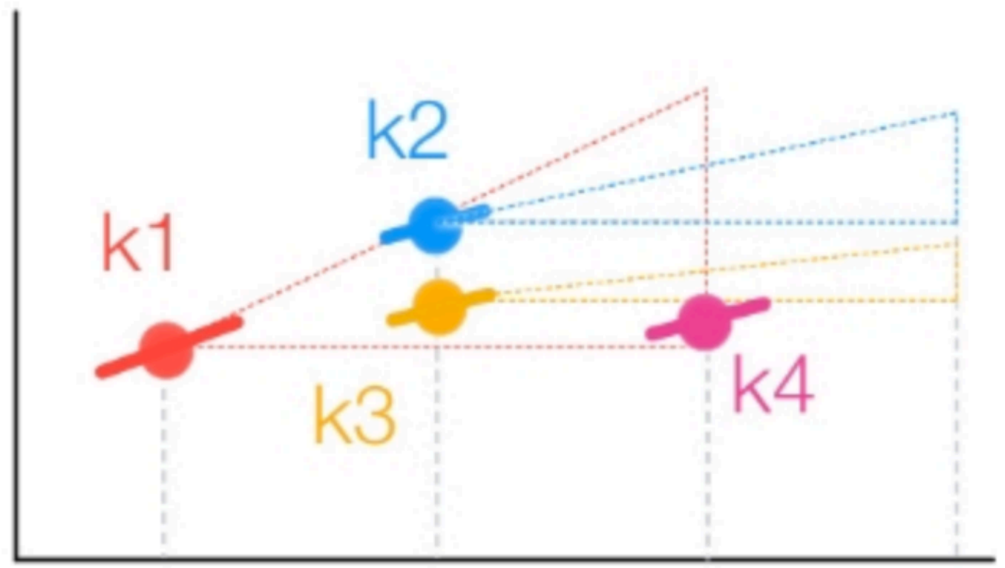
$$C_i \approx \frac{\sum_{j=0}^{Neighborhood}(\frac{\partial y}{\partial x}|_{x_j,y_j} - \frac{\partial y}{\partial x}|_{x_i,y_i})(x_j - x_i)}{\sum_{j=0}^{Neighborhood}(x_j - x_i)^2}$$

Smoothness

Streamlines

- Integración de las derivadas a lo largo de una streamline utilizando el método de Runge-Kutta
- Consideración: Resultado final susceptible al lugar dónde se comienza a integrar

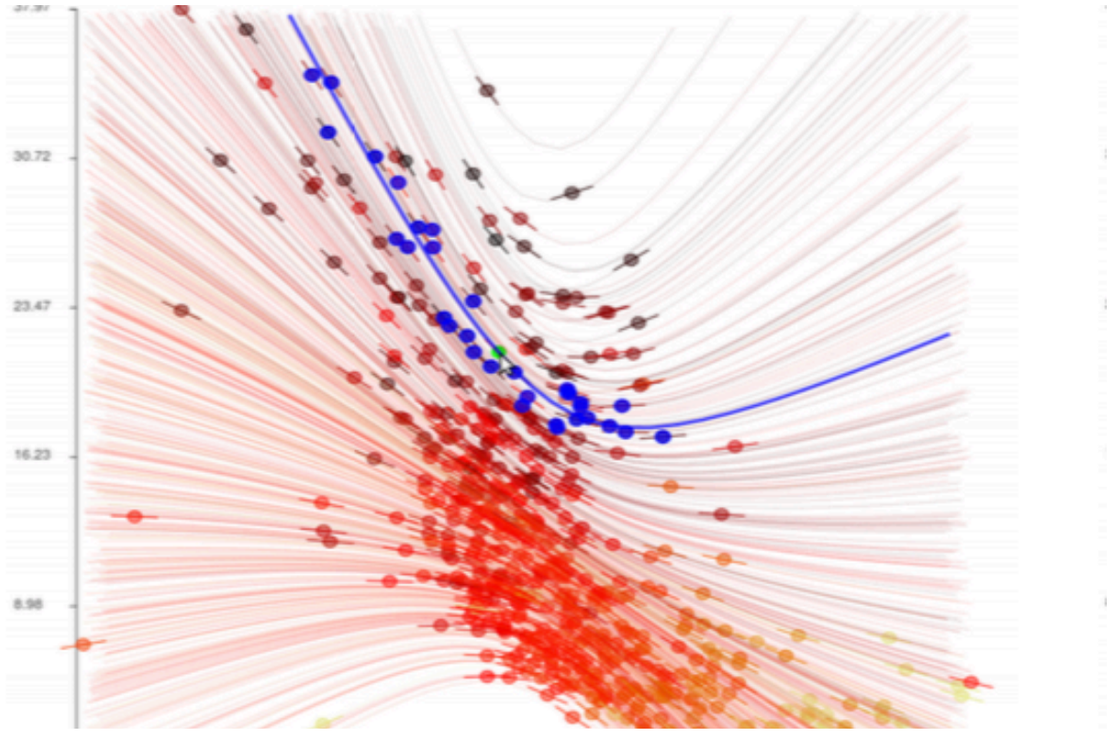
$$p'_k = p_k + hv(p_k)$$
$$p_{k+1} = p_k + hv(p'_k)$$



Visualización. Fuente: haroldserrano.com

Selección Streamline

- Permite hacer énfasis sobre una streamline y sus puntos más cercanos.
- Guía el análisis de sensibilidad al localizar la atención sobre los datos más relevantes



Selección de una streamline

Dataset

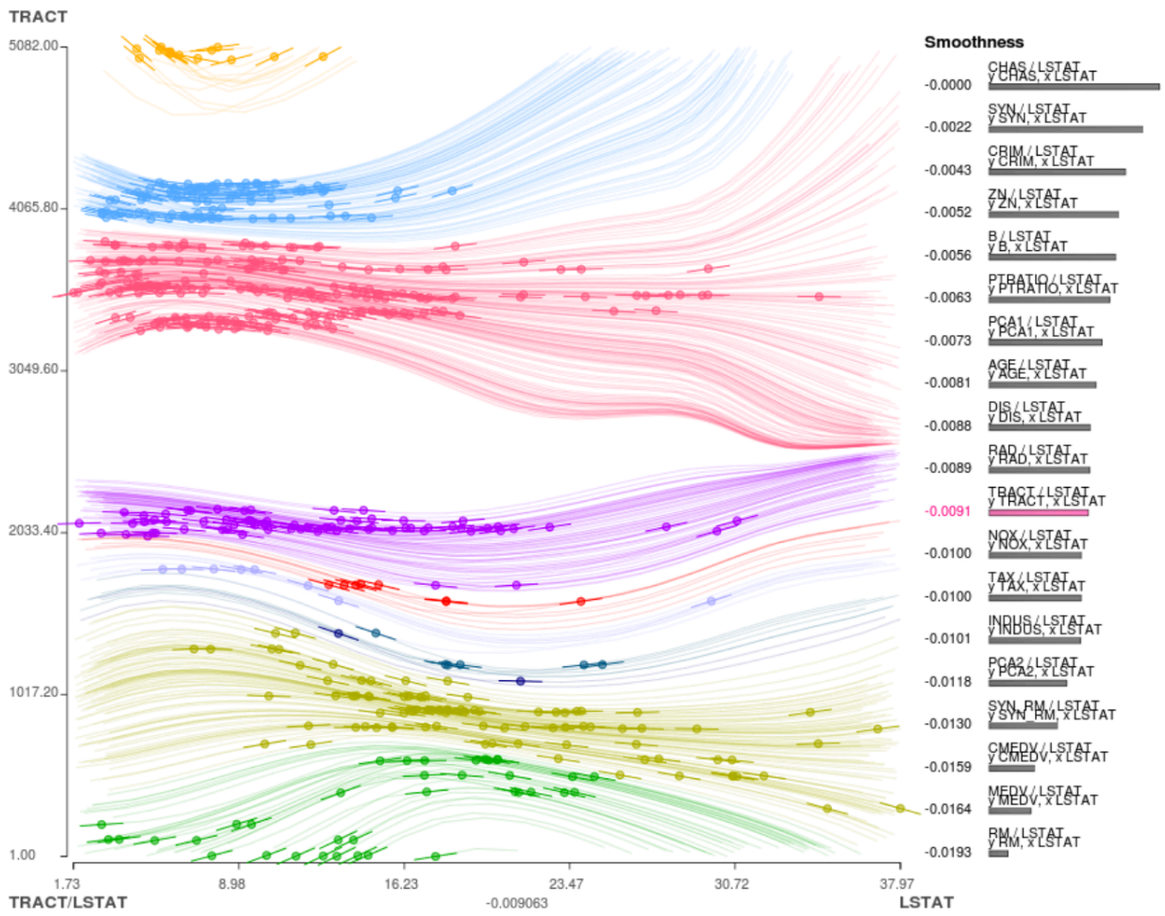
- Enfermedades cardíacas registradas en el hospital de Cleveland el año 1988
- 303 instancias, con 75 atributos de las cuales los investigadores seleccionaron los 14 más importantes

Algunos atributos

- Edad
- Sexo
- Presión
- Colesterol
- Latido del corazón
- Tipo de dolor
- Resultado del diagnóstico
- entre otros…

Futuros desafíos

- Separación por clusters



Ejemplo clusters