Analyzing Time Structured Corpora

Corpus Statistics Research Group launch event

Birmingham, 11th Feb 2016

Tony Hennessey (University of Nottingham)

joint work with R. Carrington, Y. van Gennip, M. Mahlberg, S. Preston, K. Severn, V. Wiegand

Overview

Overview

How to look at the time dependency in the properties of a corpus.

- Recap terminology and describe the main example used throughout the presentation.
- Binning data and how to think about binning mathematically.
- Using kernels which are better than bins.

Setting the scene (and a bit of a recap)

X - some matrix representation of the corpus

$$\begin{pmatrix} 0 & 2 & 2 & 1 & 0 & \dots \\ 0 & 0 & 2 & 1 & 1 & \dots \\ 1 & 0 & 0 & 1 & 1 & \dots \\ 1 & 1 & 0 & 0 & 1 & \dots \\ 1 & 1 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Setting the scene (and a bit of a recap)

X - some matrix representation of the corpus

$$
\begin{array}{c}
 \\
\texttt{doc 01} \\
\texttt{doc 02} \\
\texttt{doc 03} \\
\texttt{doc 04} \\
\texttt{doc 05} \\
 \\
 \\
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
\texttt{aardvark} & \texttt{abacus} & \texttt{badger} & \texttt{bandicoot} & \texttt{bonsai} & \\
\end{array} \\
\begin{pmatrix}
0 & 2 & 2 & 1 & 0 & \dots \\
0 & 0 & 2 & 1 & 1 & \dots \\
1 & 0 & 0 & 1 & 1 & \dots \\
1 & 1 & 0 & 0 & 1 & \dots \\
1 & 1 & 1 & 0 & 0 & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\end{array}
$$

document-term matrix

Setting the scene (and a bit of a recap)

$f(\mathbf{X})$ - some function that we apply to the corpus

Setting the scene (and a bit of a recap)

f($\mathbf{X}$) - some function that we apply to the corpus

The cosine of the angle between words in a vector space which was derived using a matrix factorization.
($\mathbf{X} = \mathbf{USV}^{\mathsf{T}}$ singular value decomposition)

This measure quantifies the degree of association between words i.e. a bigger value implies closer association.

Setting the scene (and a bit of a recap)

THE TIMES
Digital Archive 1785-2008

X (document-term matrix)

- 11,543,110 documents
- 472,331 terms

Setting the scene (and a bit of a recap)

THE TIMES
Digital Archive 1785-2008

X (document-term matrix)

- 11,543,110 documents
- 472,331 terms

Meta-data for each document includes a date

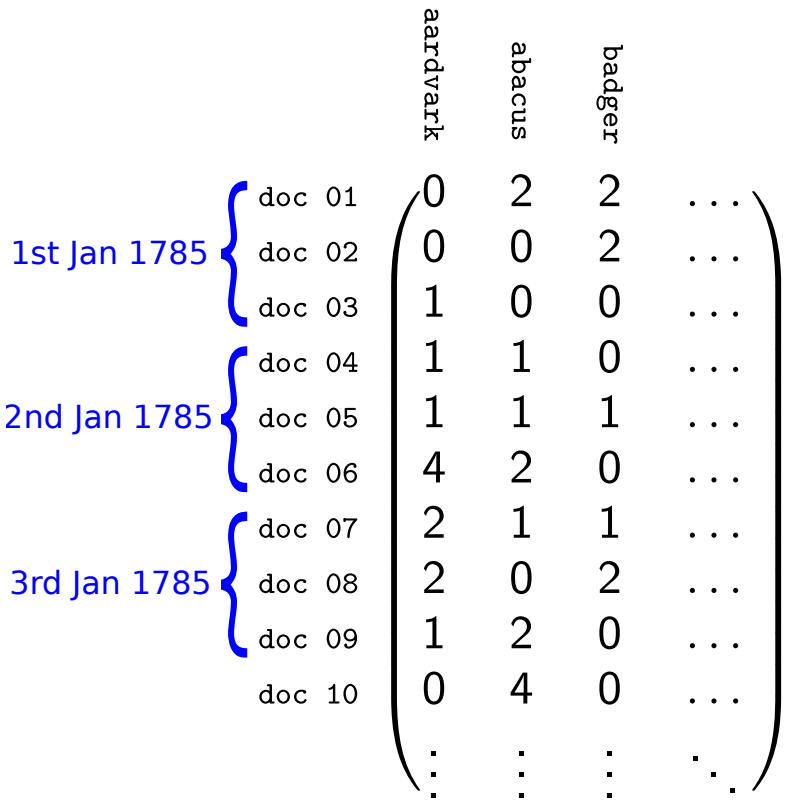
How does the corpus change with time?

Let us try binning the data using dates.

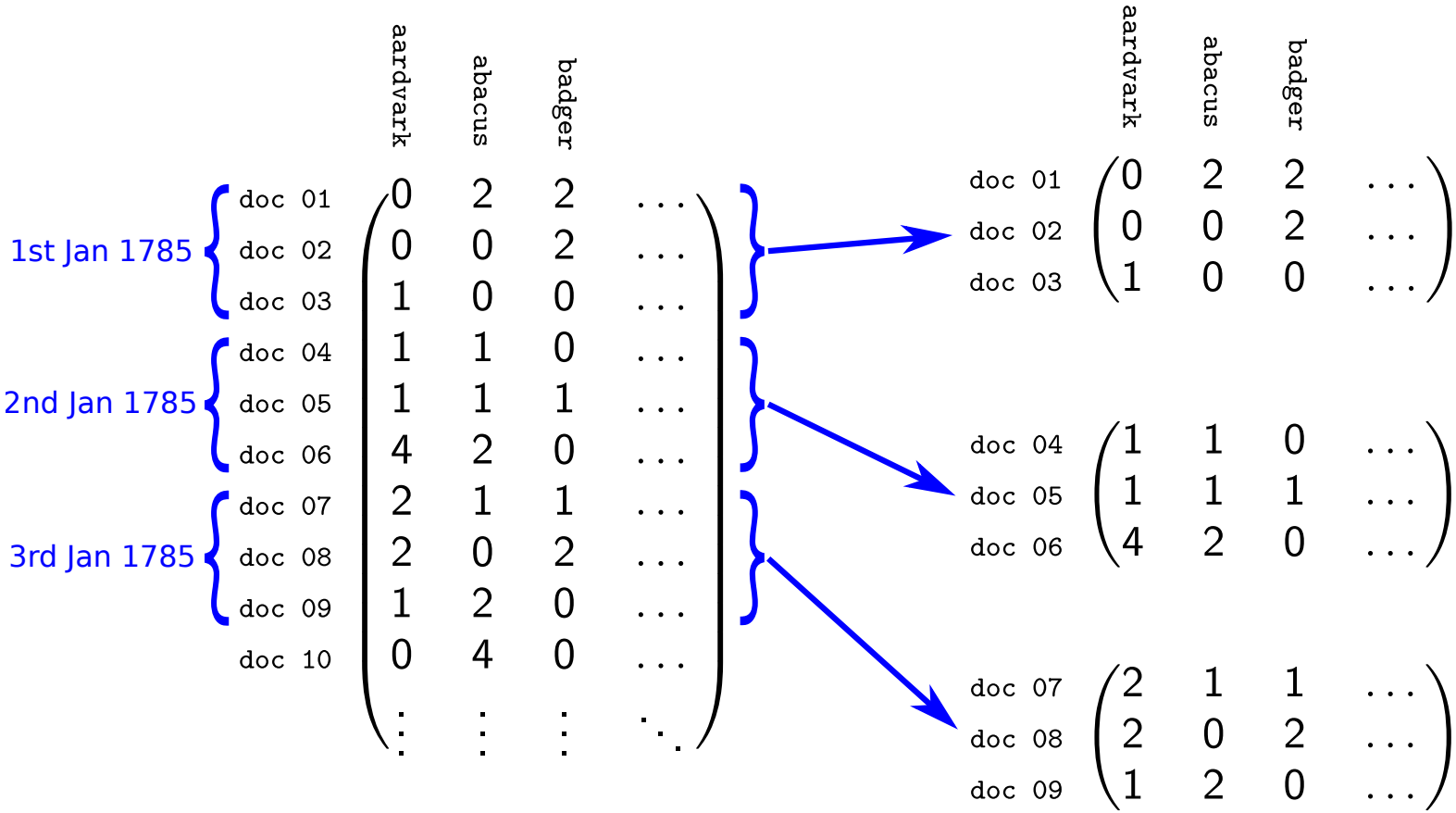
Binning by date

$$X = \begin{array}{c} \\ \texttt{doc 01} \\ \texttt{doc 02} \\ \texttt{doc 03} \\ \texttt{doc 04} \\ \texttt{doc 05} \\ \texttt{doc 06} \\ \texttt{doc 07} \\ \texttt{doc 08} \\ \texttt{doc 09} \\ \texttt{doc 10} \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{aardvark} & \texttt{abacus} & \texttt{badger} & \\ \end{array} \\ \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \\ 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 4 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Binning by date



Binning by date



Binning by date

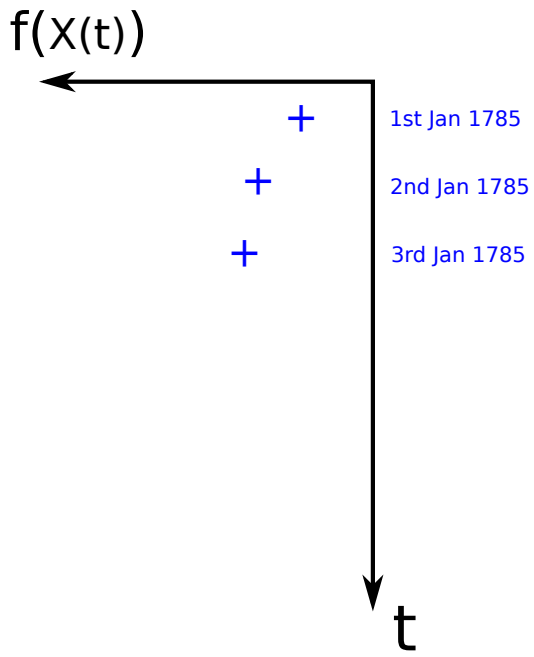
$$
X(t) \qquad \begin{matrix} & \texttt{aardvark} & \texttt{abacus} & \texttt{badger} & \\ \end{matrix}
$$

$$
X(\text{t = 1st Jan 1785}) = \begin{matrix} \texttt{doc 01} \\ \texttt{doc 02} \\ \texttt{doc 03} \end{matrix} \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \end{pmatrix}
$$

$$
X(\text{t = 2nd Jan 1785}) = \begin{matrix} \texttt{doc 04} \\ \texttt{doc 05} \\ \texttt{doc 06} \end{matrix} \begin{pmatrix} 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \end{pmatrix}
$$

$$
X(\text{t = 3rd Jan 1785}) = \begin{matrix} \texttt{doc 07} \\ \texttt{doc 08} \\ \texttt{doc 09} \end{matrix} \begin{pmatrix} 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \end{pmatrix}
$$

Binning by date



$$X(t) = \begin{array}{c} \\ \texttt{doc 01} \\ \texttt{doc 02} \\ \texttt{doc 03} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{aardvark} & \texttt{abacus} & \texttt{badger} & \end{array} \\ \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \end{pmatrix} \end{array}$$

$$X(t = \text{1st Jan 1785}) = \begin{array}{c} \\ \texttt{doc 04} \\ \texttt{doc 05} \\ \texttt{doc 06} \end{array} \begin{pmatrix} 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \end{pmatrix}$$

$$X(t = \text{2nd Jan 1785}) = \begin{array}{c} \\ \texttt{doc 07} \\ \texttt{doc 08} \\ \texttt{doc 09} \end{array} \begin{pmatrix} 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \end{pmatrix}$$

$$X(t = \text{3rd Jan 1785}) =$$

Binning by date

Identity matrix

$$\mathbf{X} = \mathbf{I}\,\mathbf{X}$$

$$\begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \\ 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 4 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix} \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \\ 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 4 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Binning by date

Filter by date

$$\mathbf{X}(t) = \mathbf{b}(t)\,\mathbf{X}$$

$$\begin{pmatrix} \mathbf{0} & \mathbf{2} & \mathbf{2} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{2} & \dots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix} \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \\ 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 4 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where $t =$ '1st Jan 1785'

Binning by date

Filter by date

$$\mathbf{X}(t) = \mathbf{b}(t)\,\mathbf{X}$$

$$\begin{pmatrix} 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \dots \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \dots \\ \mathbf{4} & \mathbf{2} & \mathbf{0} & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix} \begin{pmatrix} 0 & 2 & 2 & \dots \\ 0 & 0 & 2 & \dots \\ 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ 4 & 2 & 0 & \dots \\ 2 & 1 & 1 & \dots \\ 2 & 0 & 2 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 4 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where $t =$ '2nd Jan 1785'

Binning by date

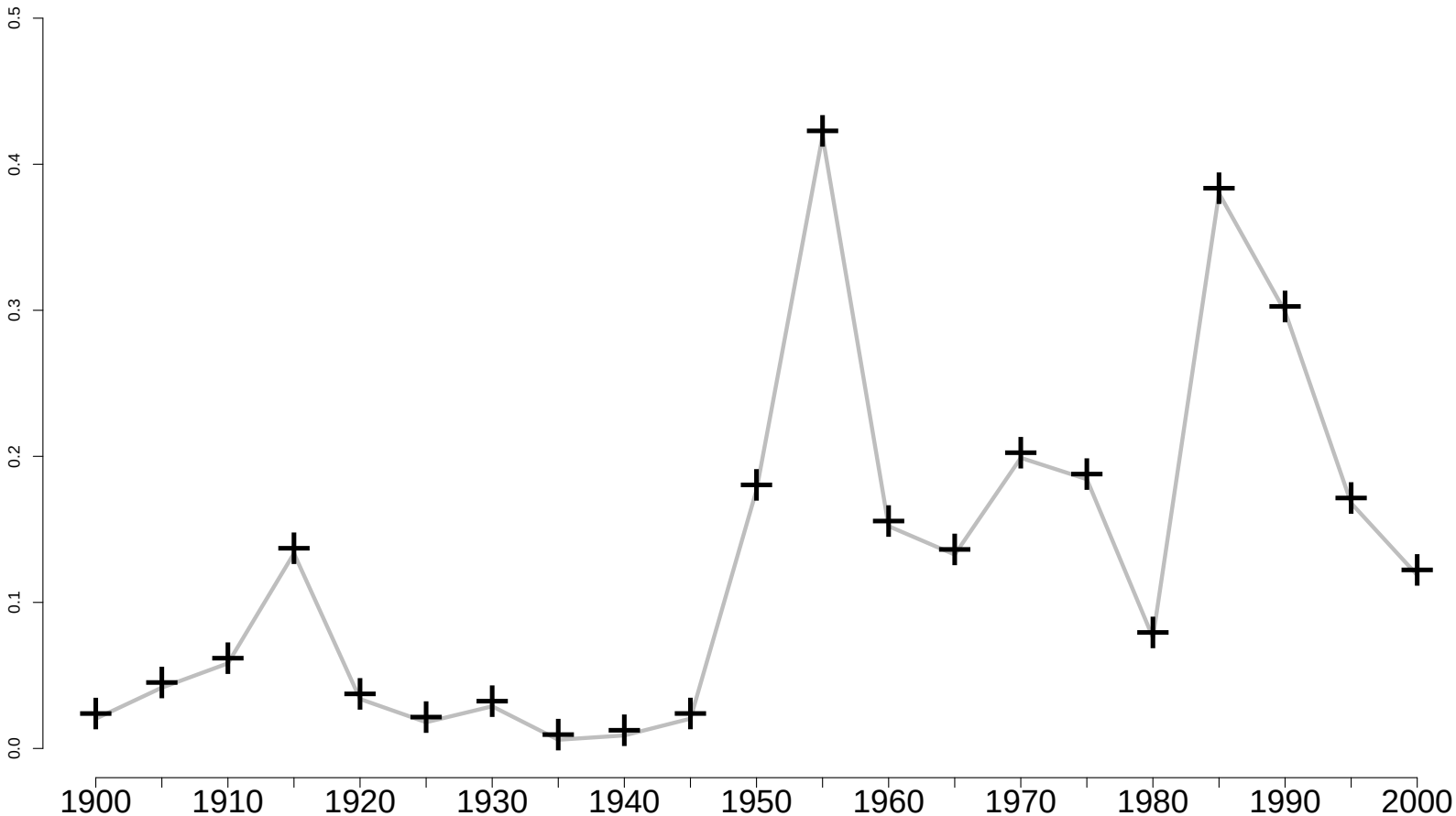
How wide should the bins be?

- depends on your research question
  - e.g. over what time scale are you interested in examining change?
- depends on your data
  - e.g. how sparsely distributed are the traits you are looking at likely to be?

Binning by date

An example of binning using the TDA

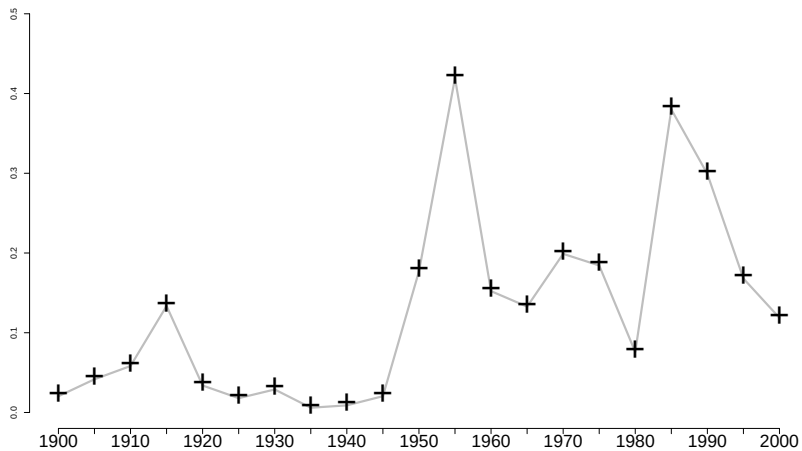
just showing $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Binning by date

An example of binning using the TDA

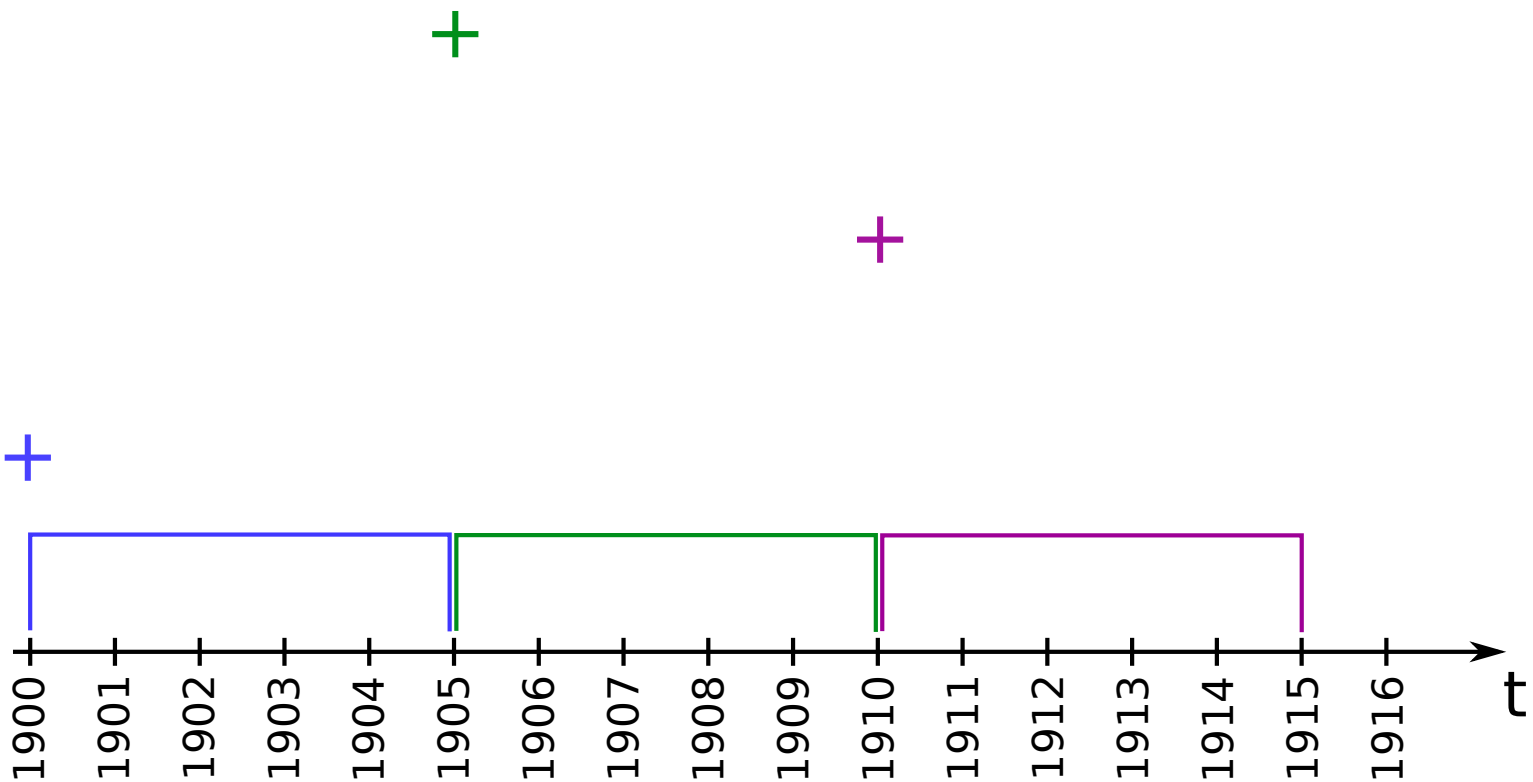
just showing $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



We used 5 year bins because

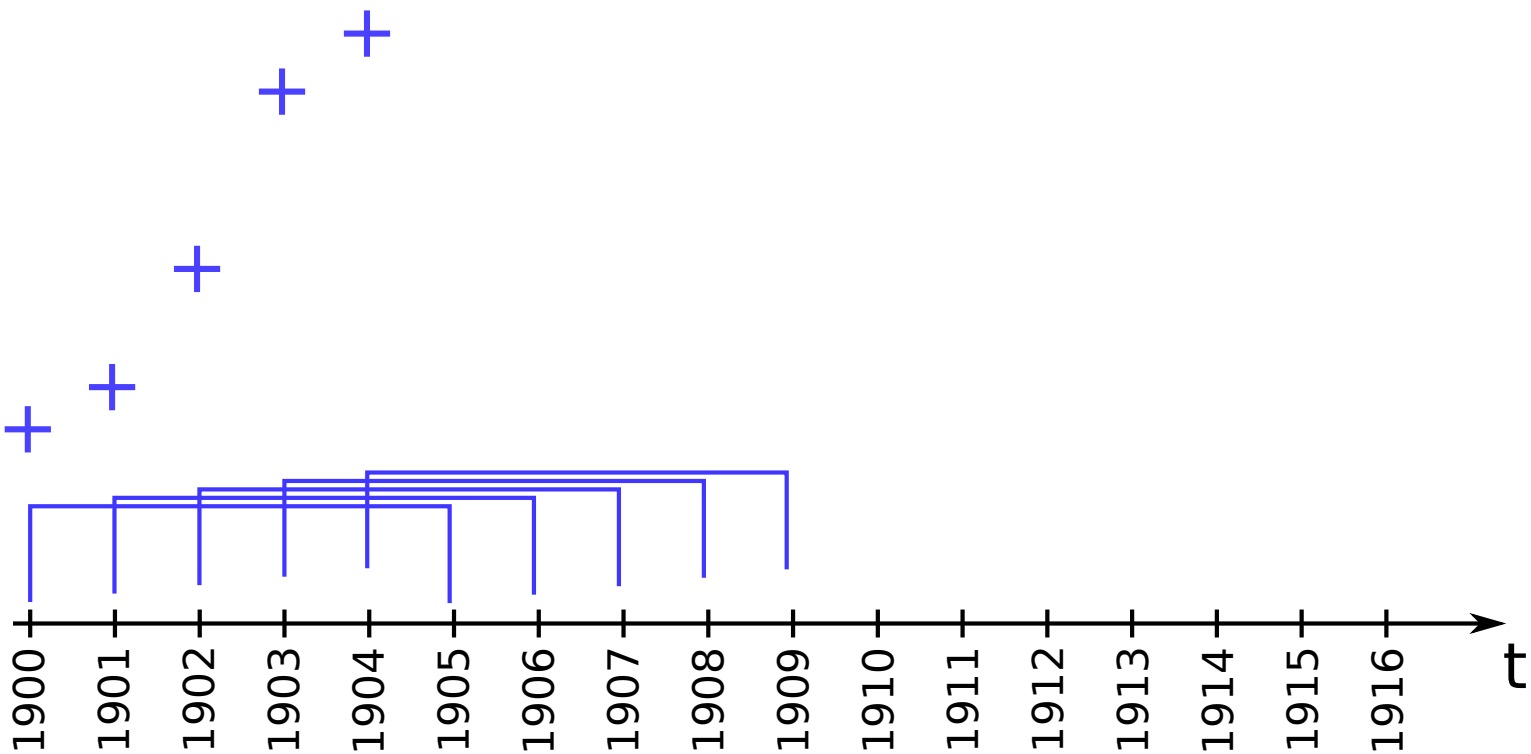
- the number of articles about smoking are quite sparsely distributed
- we are mainly interested in long term trends

Binning by date



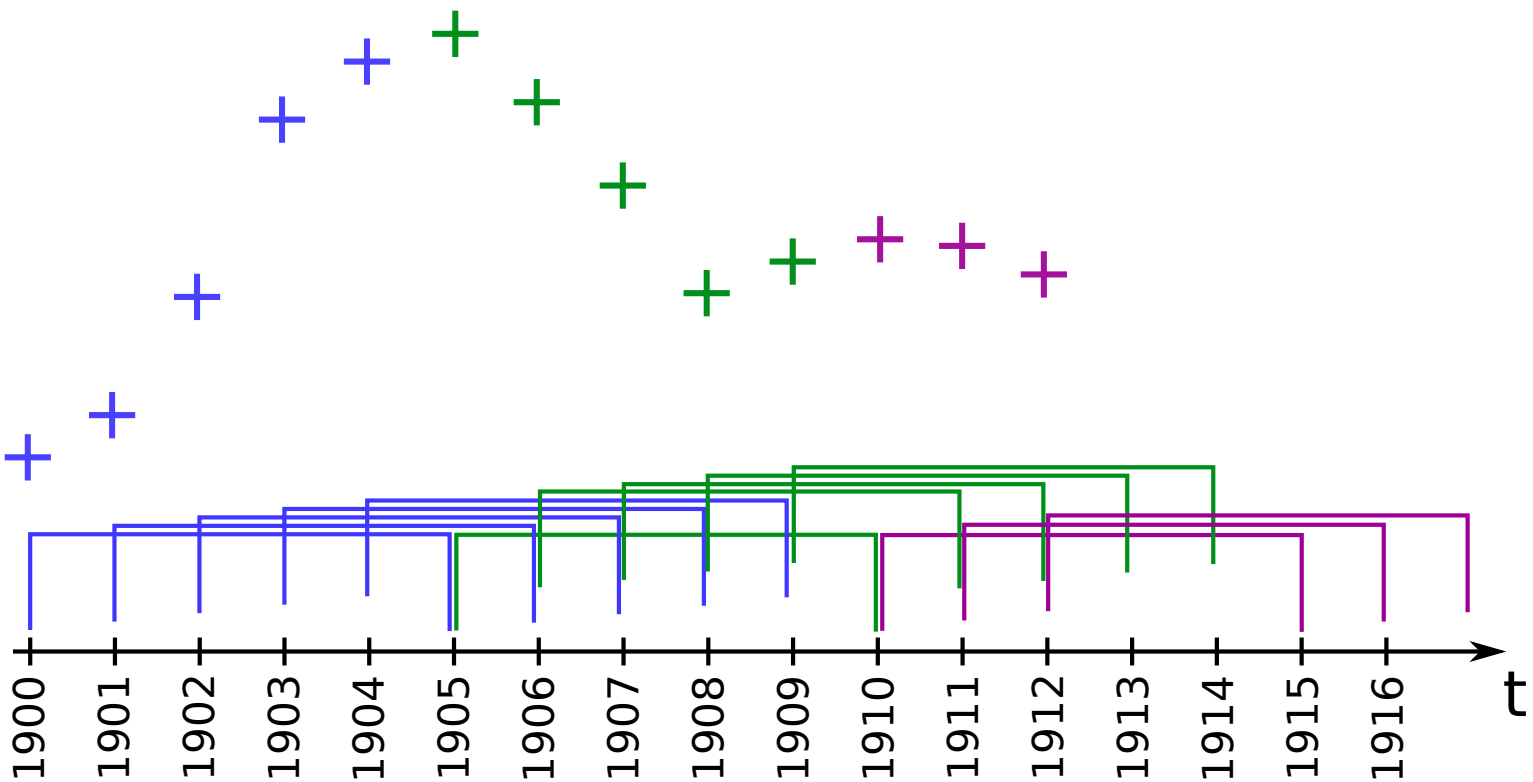
Binning by date

Sliding the bins



Binning by date

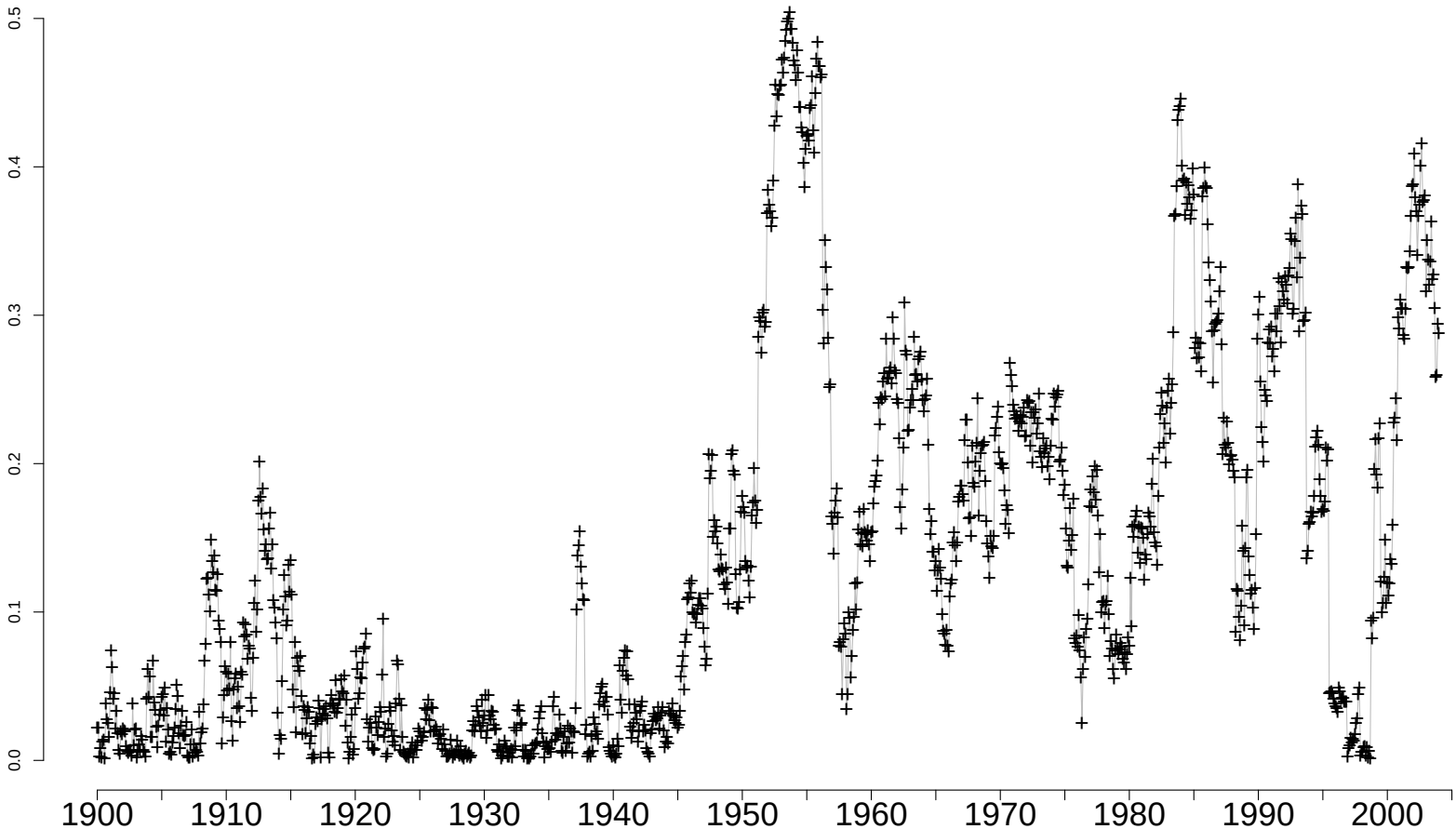
Sliding the bins



Binning by date

Sliding the bins for the TDA example

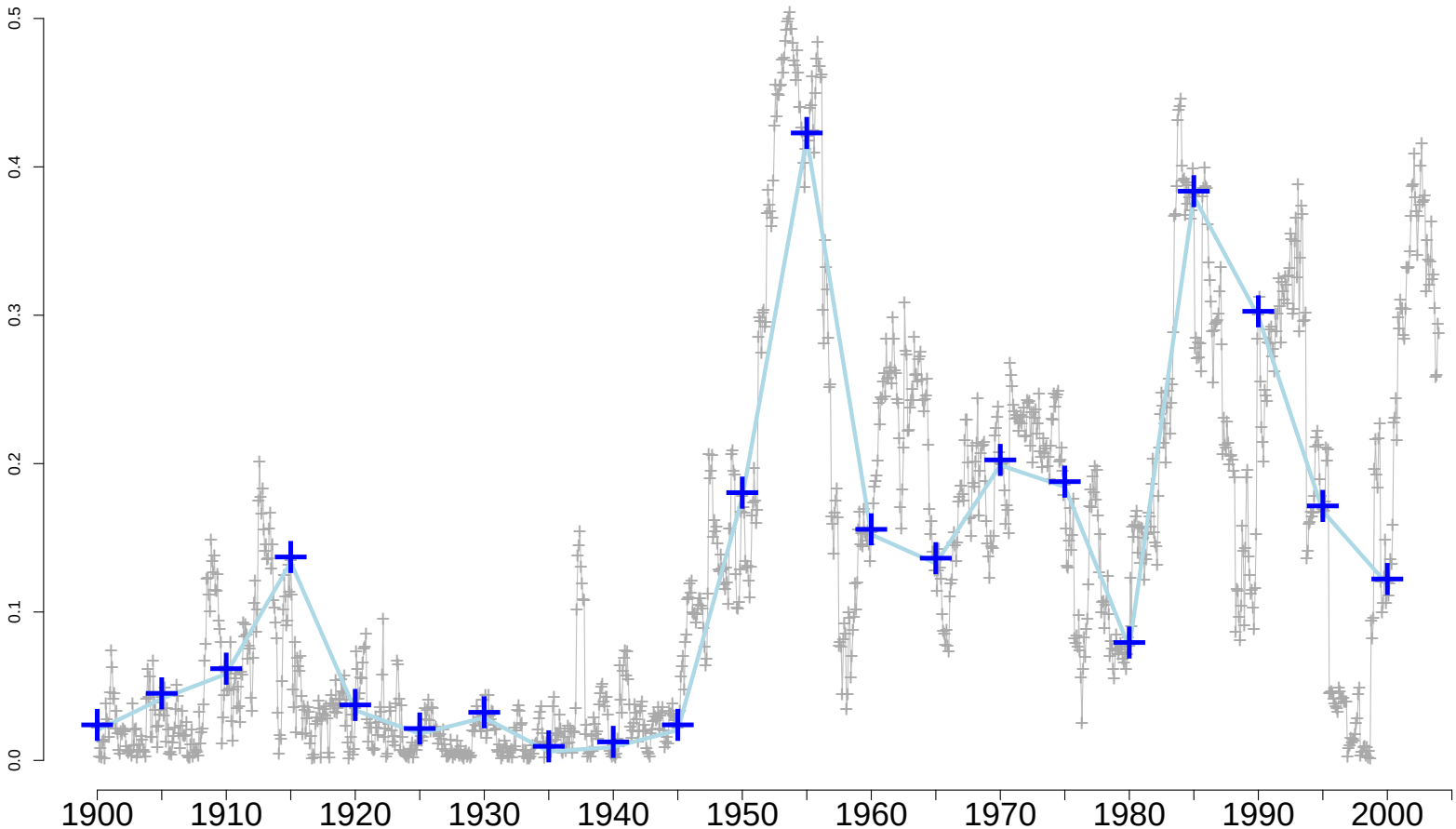
revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Binning by date

Sliding the bins for the TDA example

revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Using a kernel

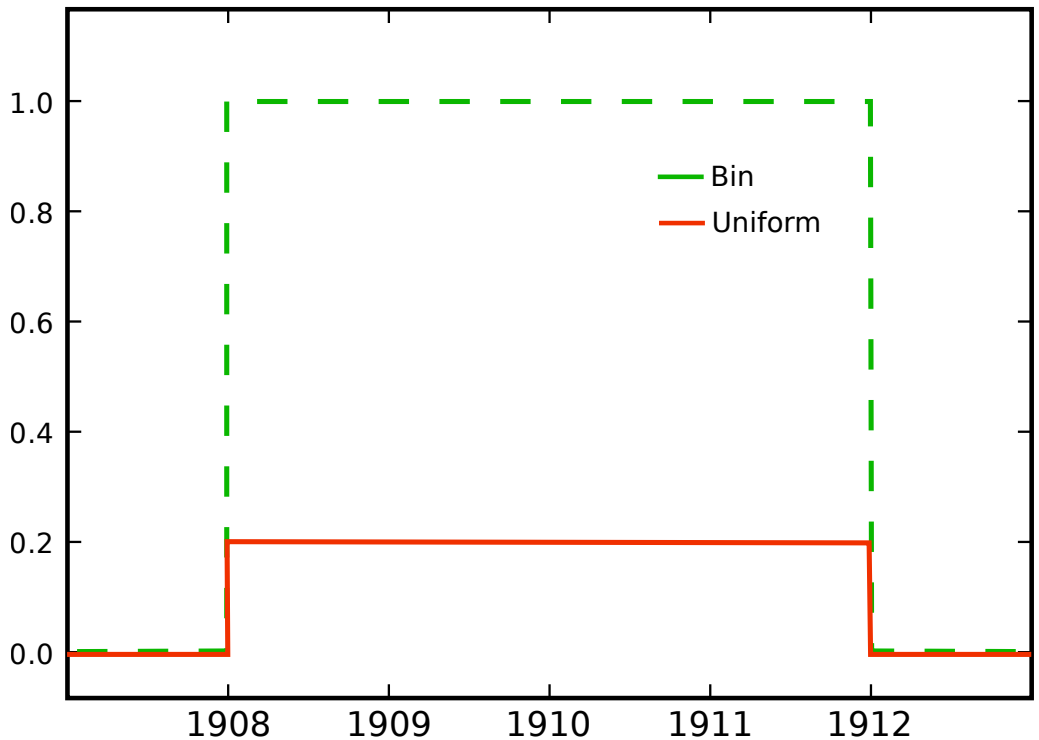
Can we do better?

Yes. Use a kernel.

Using a kernel

Why use a kernel? Why not just bin?

- A kernel takes account of the width of your data collection window i.e. if you bin, as your bins get wider your effect will get bigger; with a kernel it will not.



Using a kernel

Why use a kernel? Why not just bin?

- A kernel takes account of the width of your data collection window i.e. if you bin, as your bins get wider your effect will get bigger; with a kernel it will not.

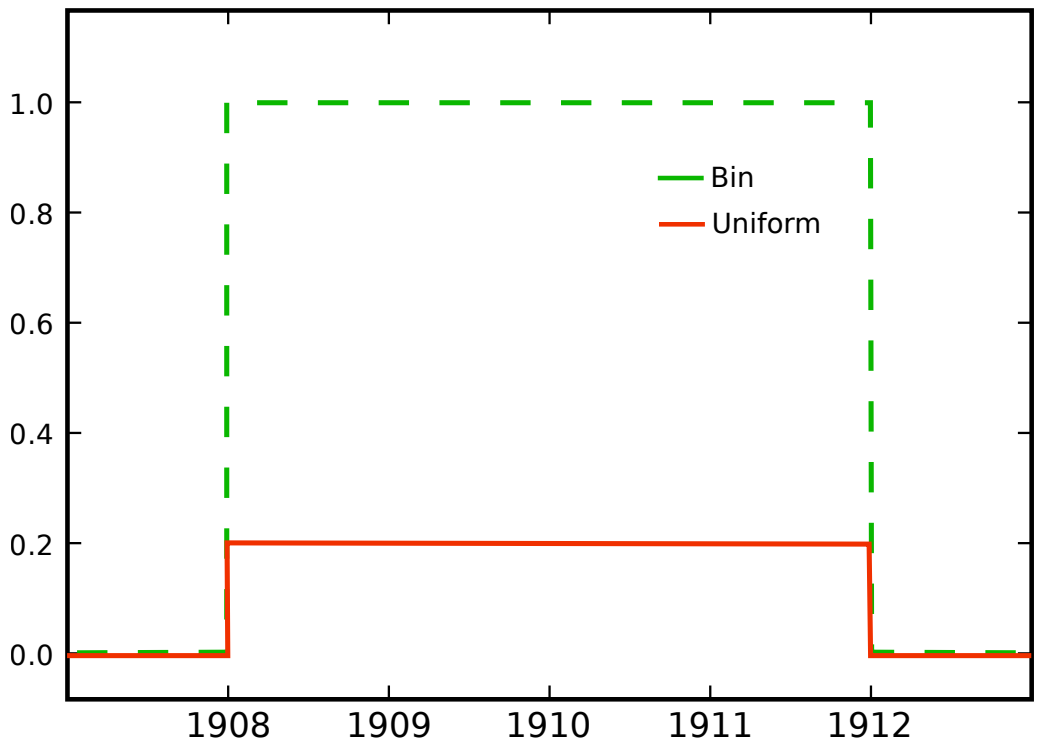
$$\mathbf{k}(t) = \frac{1}{w}\,\mathbf{b}(t)$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \frac{1}{w} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \frac{1}{w} & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{w} & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \frac{1}{w} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Using a kernel

Why use a kernel? Why not just bin?

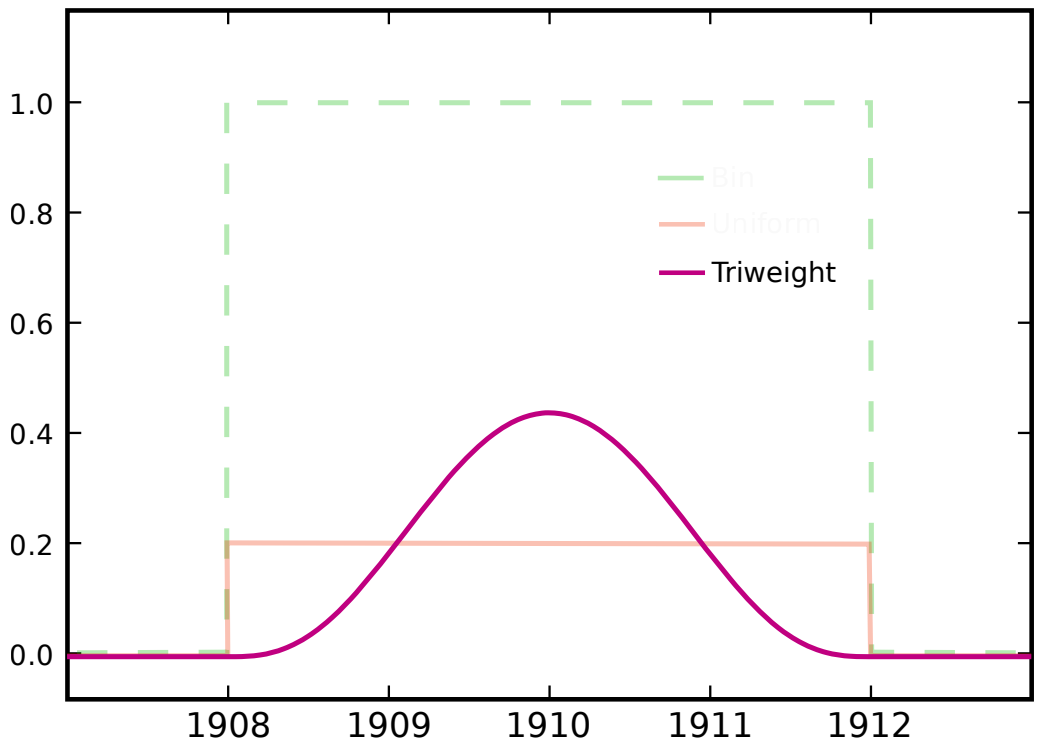
- A kernel takes account of the width of your data collection window i.e. if you bin, as your bins get wider your effect will get bigger; with a kernel it will not.
- With a kernel we can control smoothing.



Using a kernel

Why use a kernel? Why not just bin?

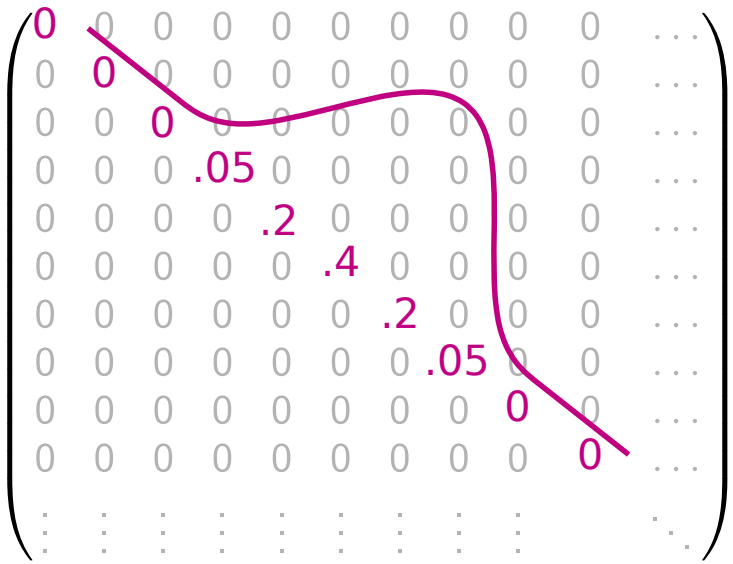
- A kernel takes account of the width of your data collection window i.e. if you bin, as your bins get wider your effect will get bigger; with a kernel it will not.
- With a kernel we can control smoothing.



Using a kernel

Why use a kernel? Why not just bin?

- A kernel takes account of the width of your data collection window i.e. if you bin, as your bins get wider your effect will get bigger; with a kernel it will not.
- With a kernel we can control smoothing.



Using a kernel

Some examples of kernels

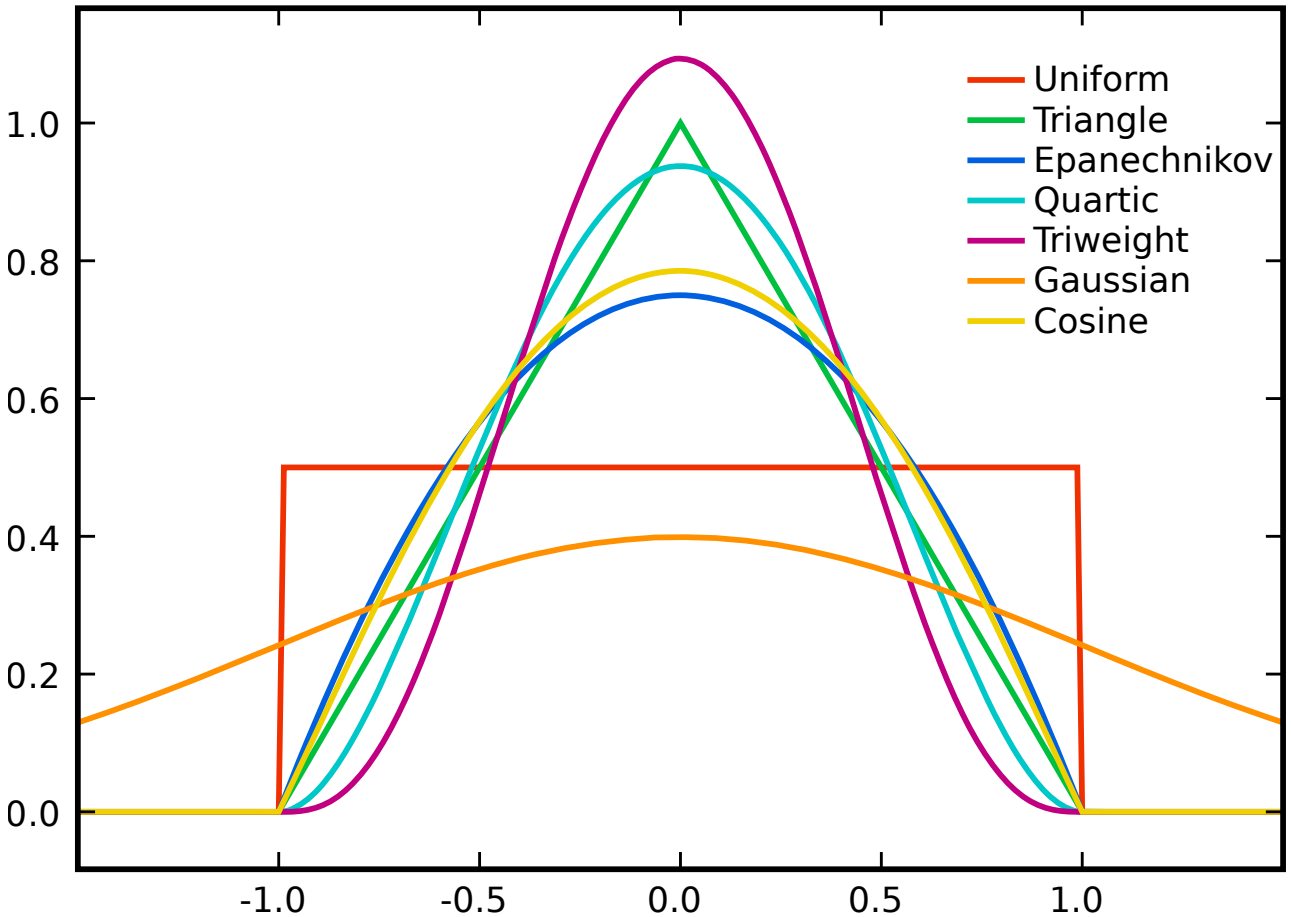
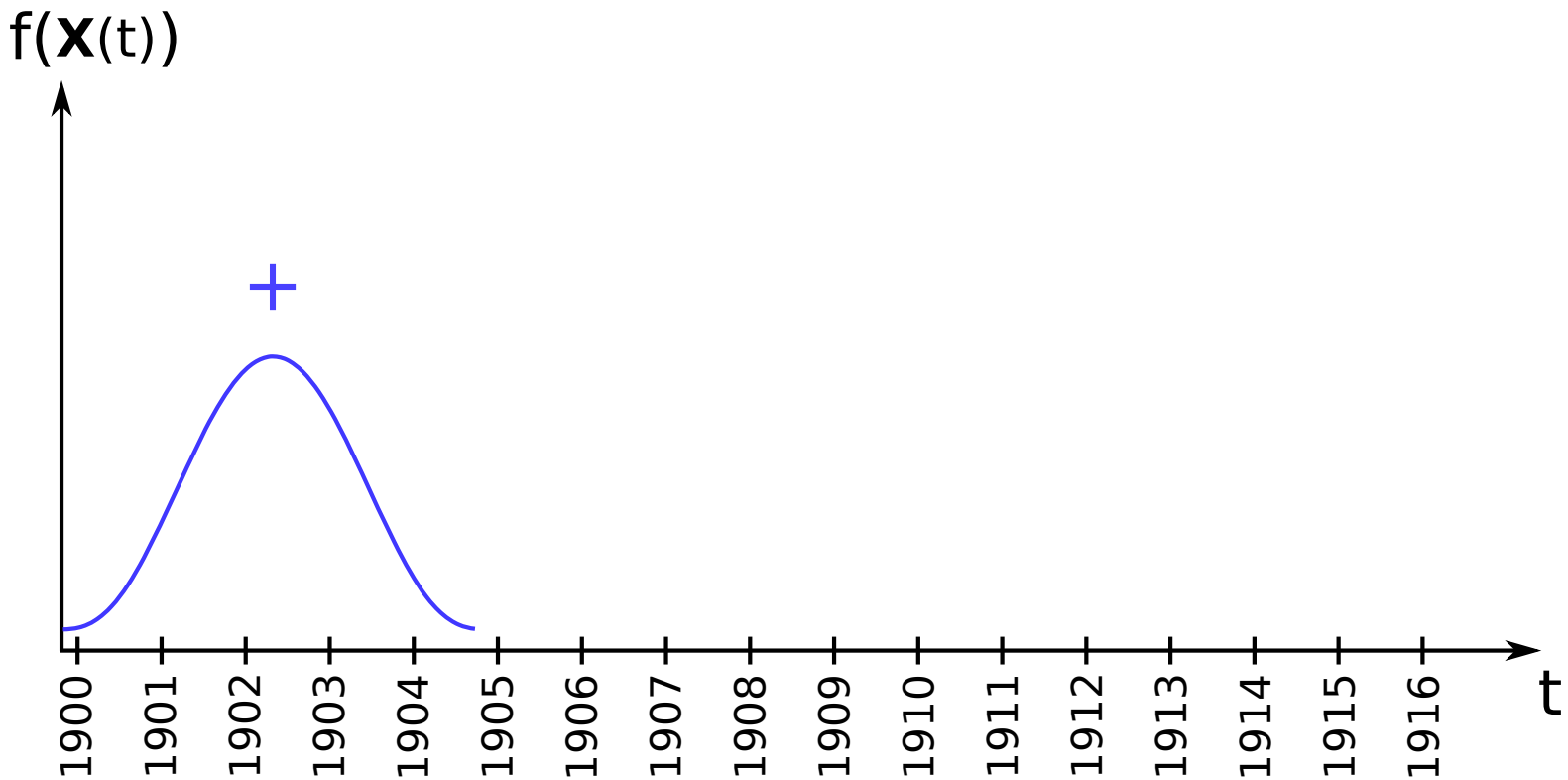


image by Brian Amberg (wikicommons)

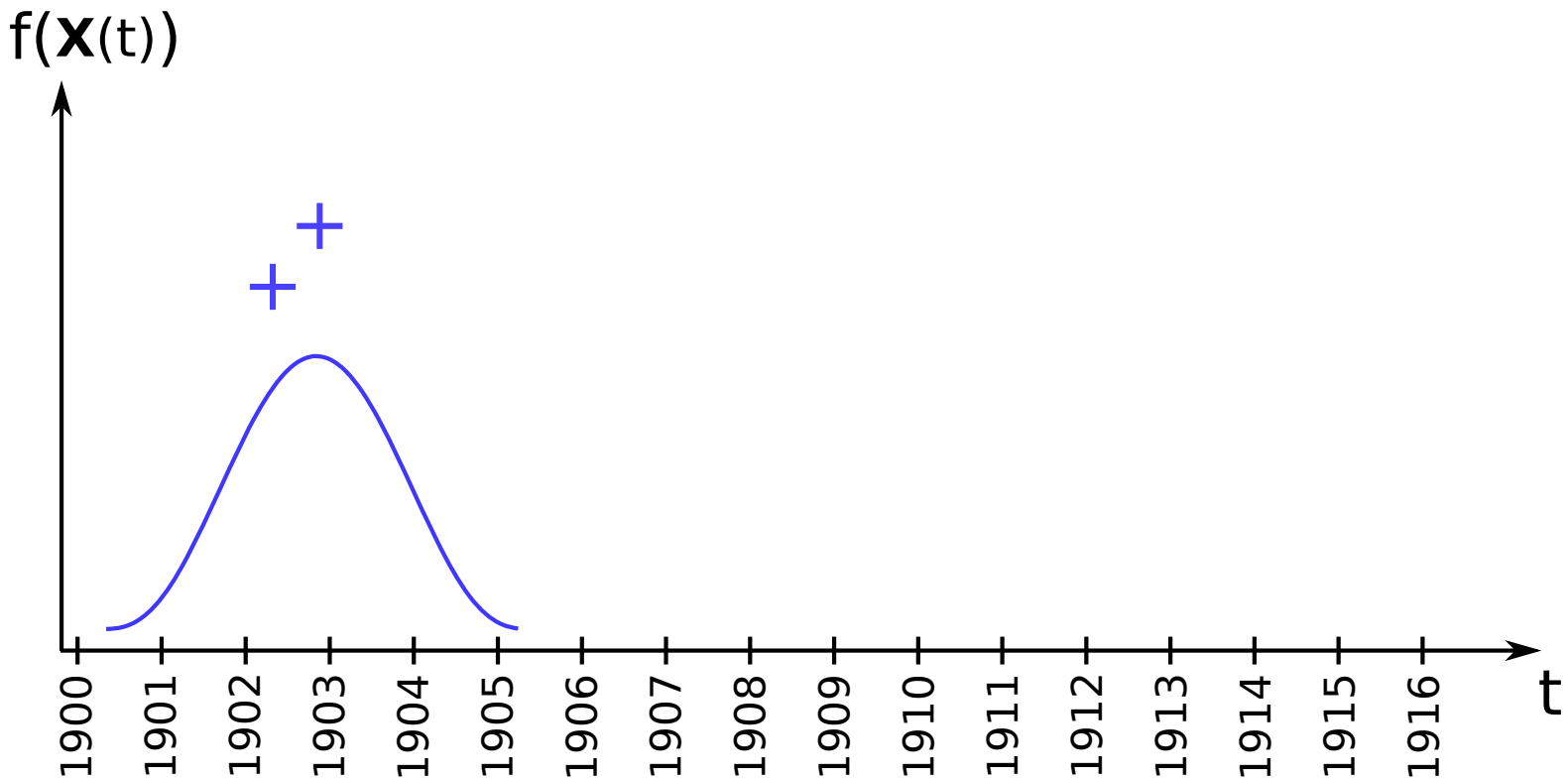
Using a kernel

Sliding the kernel



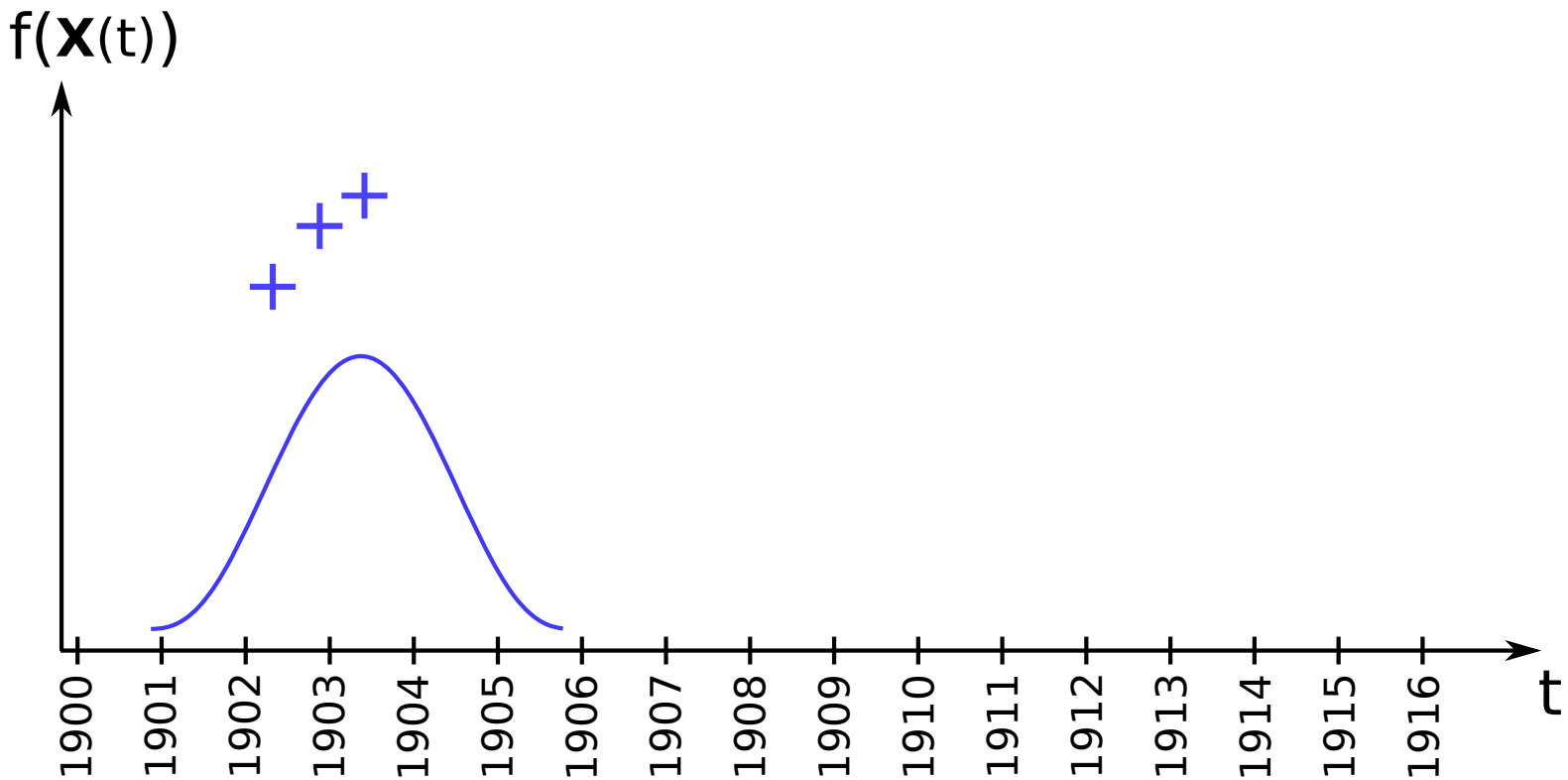
Using a kernel

Sliding the kernel



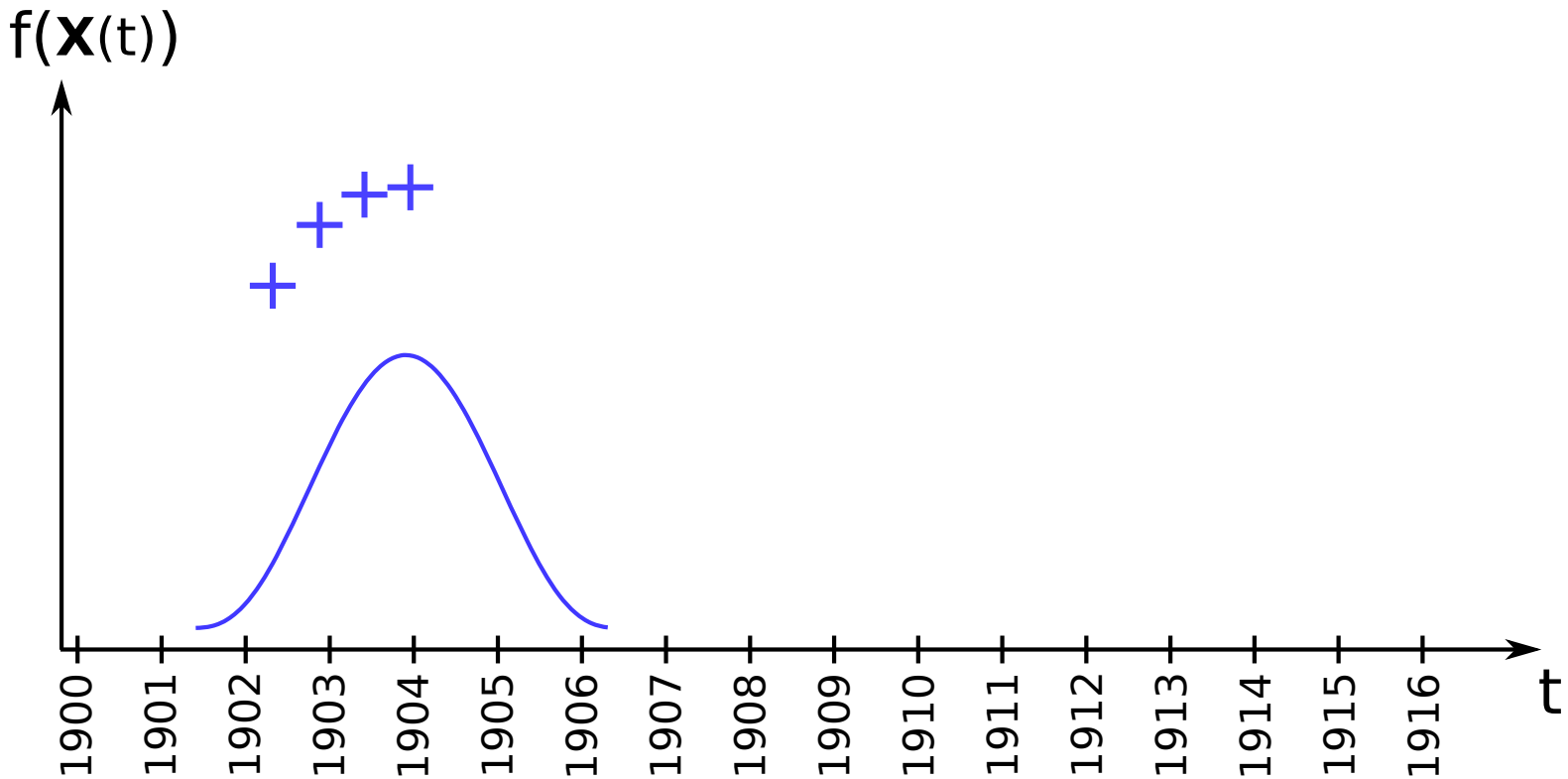
Using a kernel

Sliding the kernel



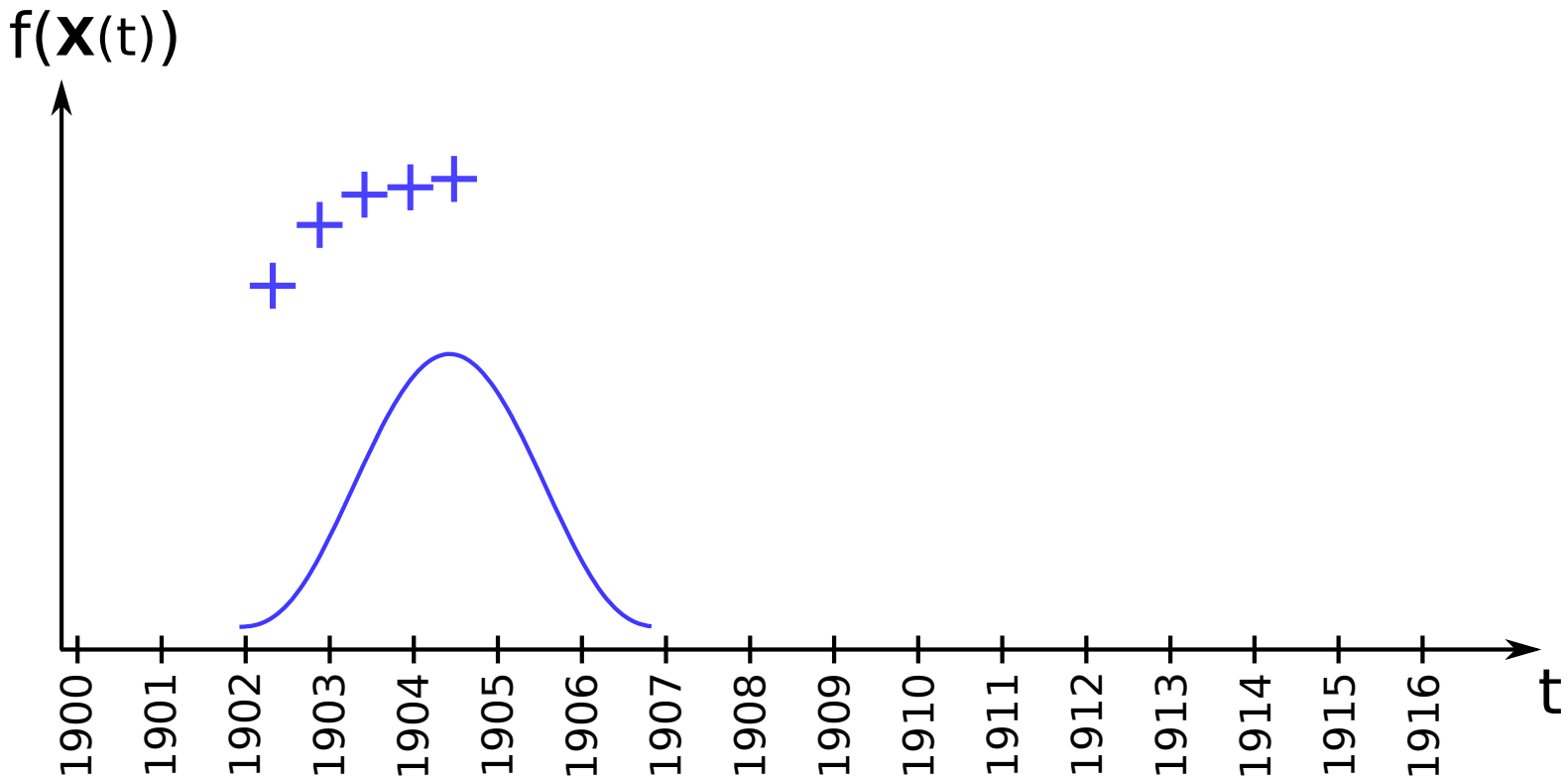
Using a kernel

Sliding the kernel



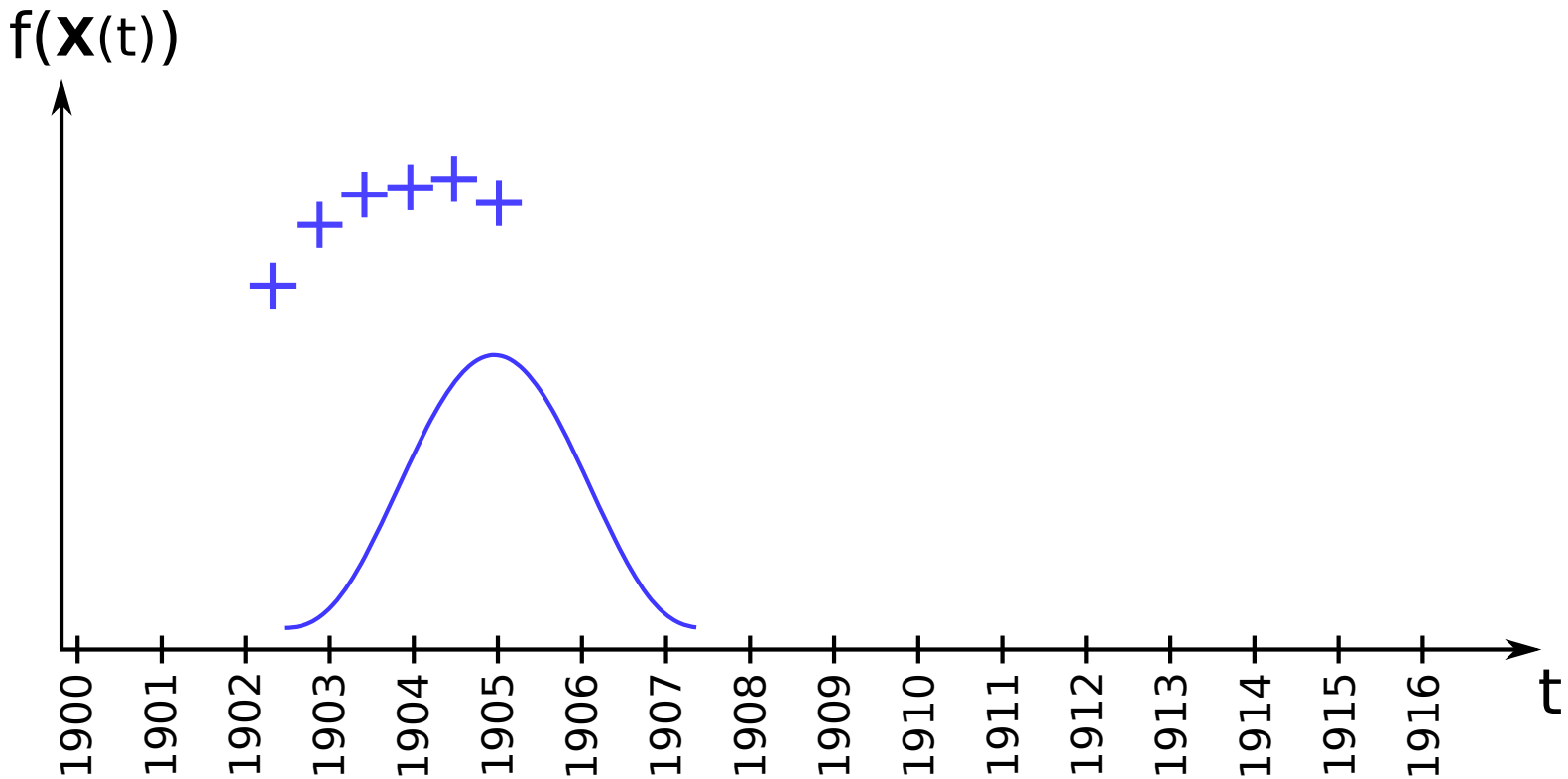
Using a kernel

Sliding the kernel



Using a kernel

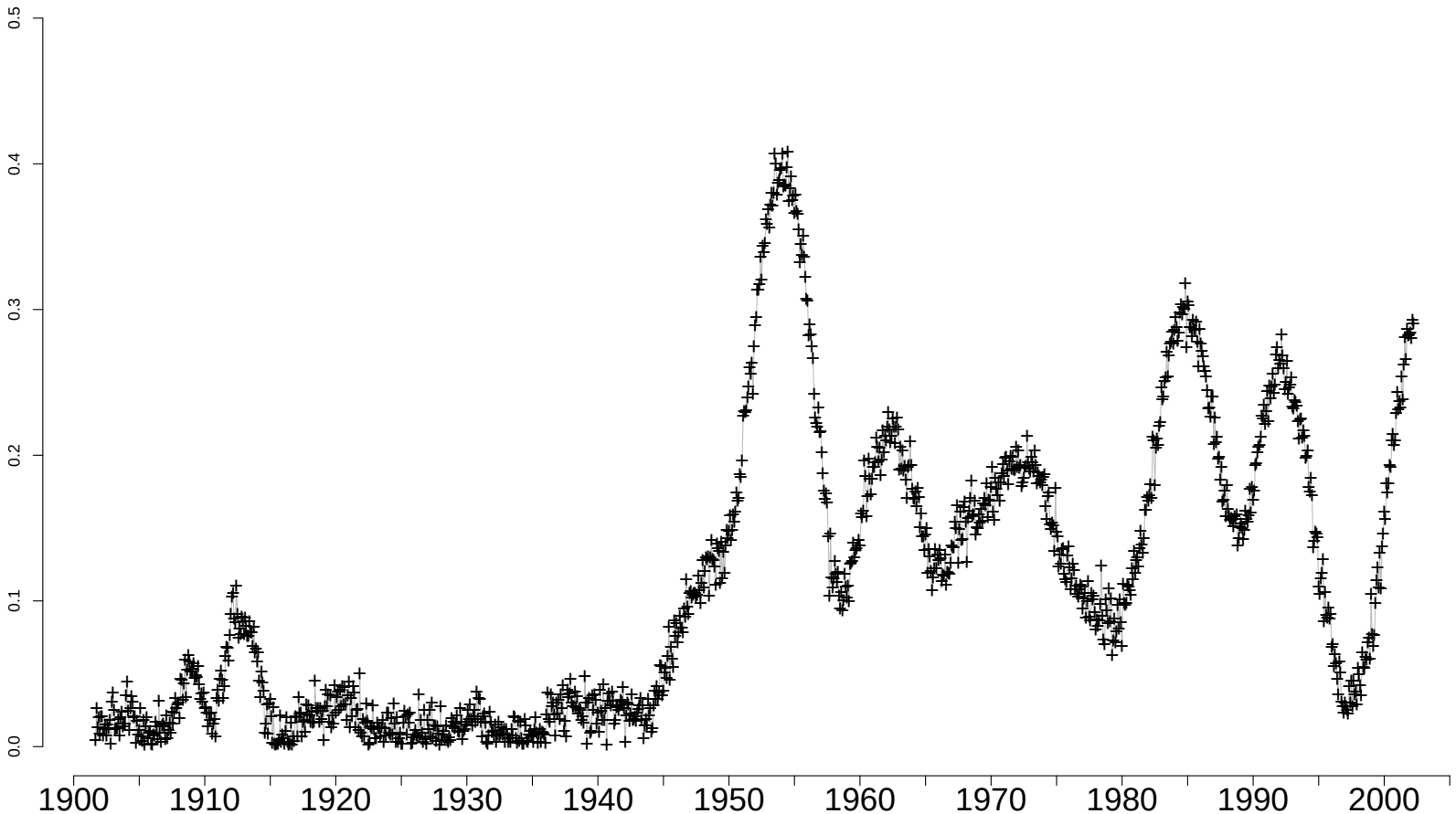
Sliding the kernel



Using a kernel

Triweight kernel for the TDA example

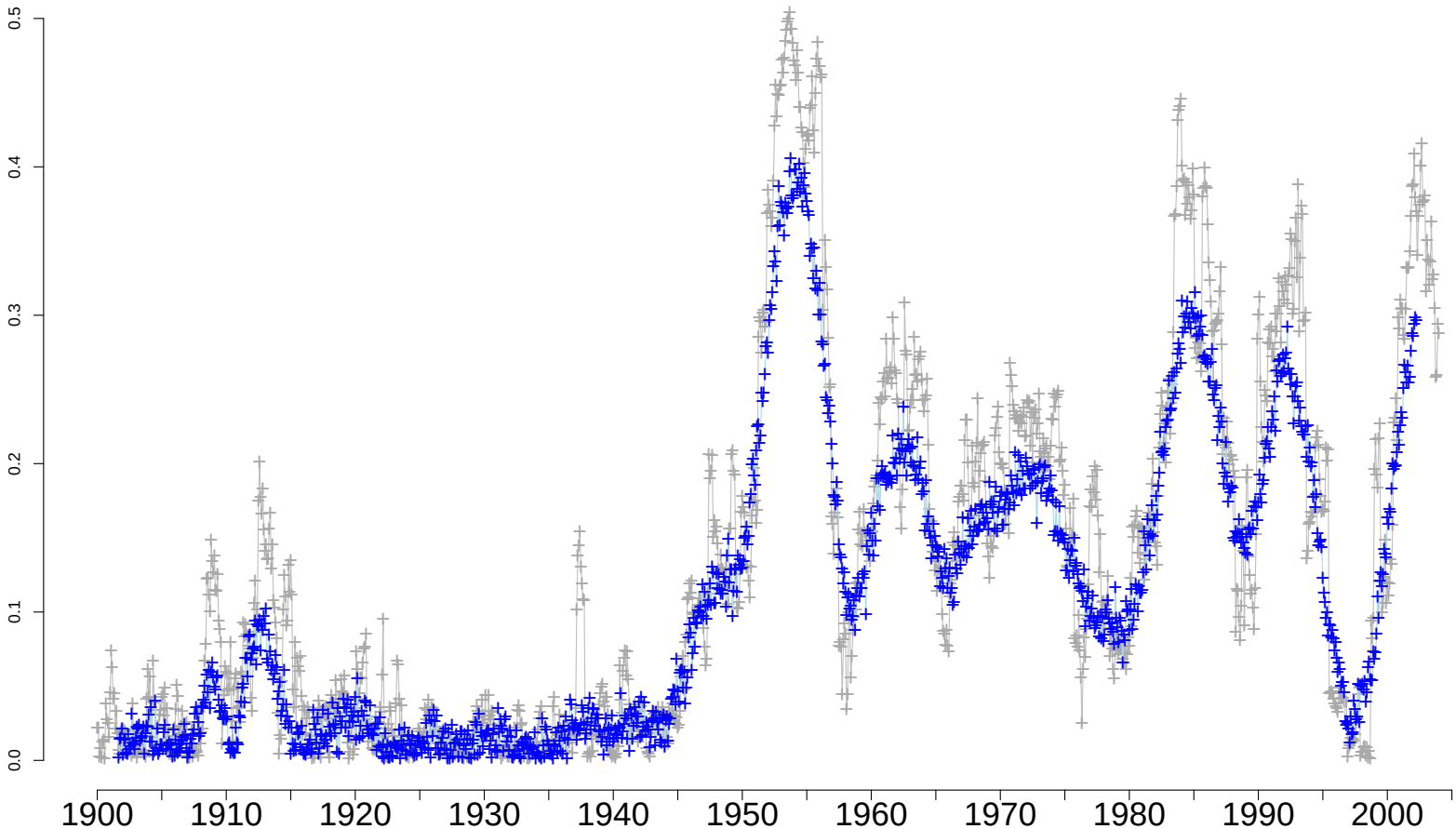
revisit $f(\,\mathbf{X}(t)\,)$ for 'smoking' and 'cancer'



Using a kernel

Triweight kernel for the TDA example

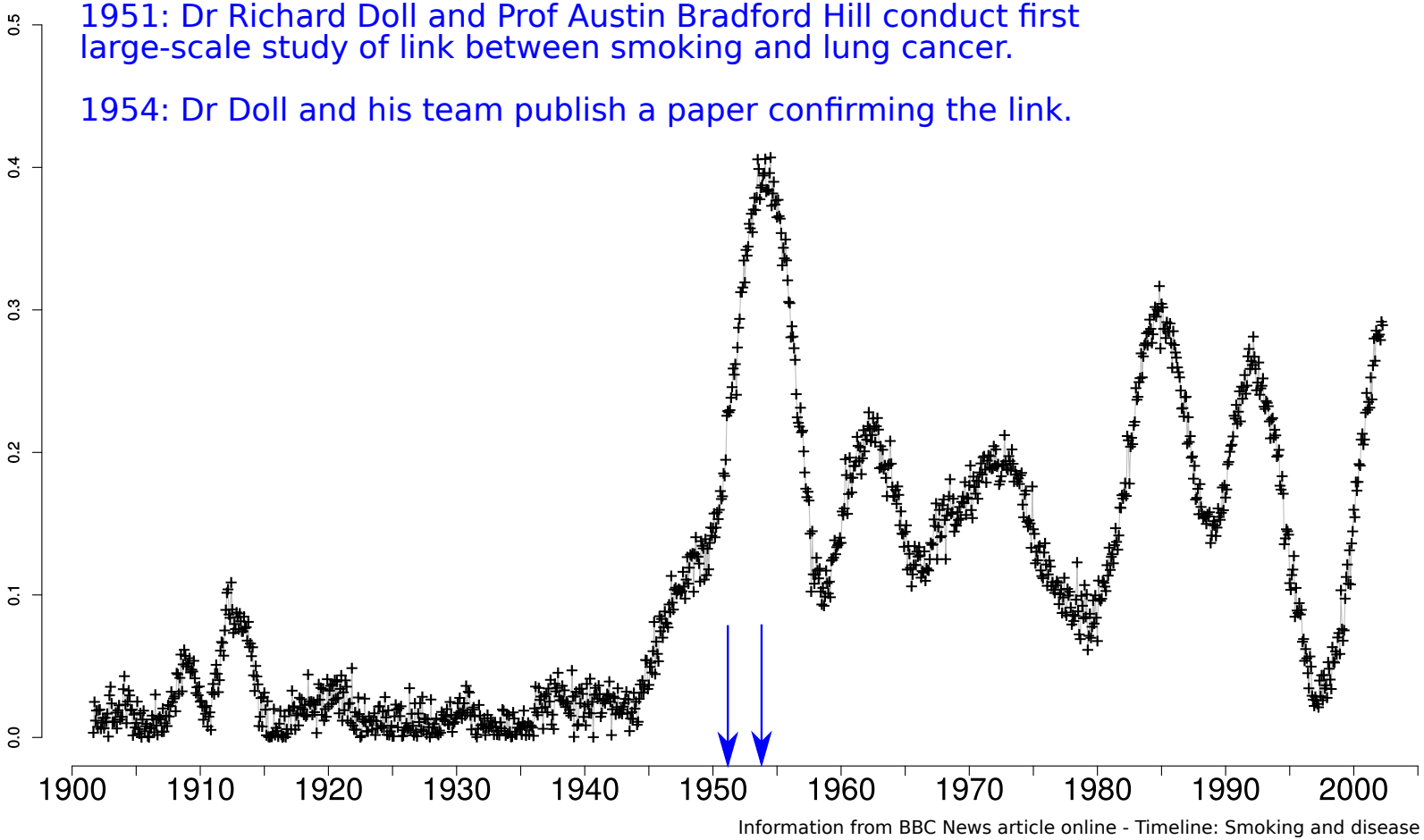
revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Using a kernel

Triweight kernel for the TDA example

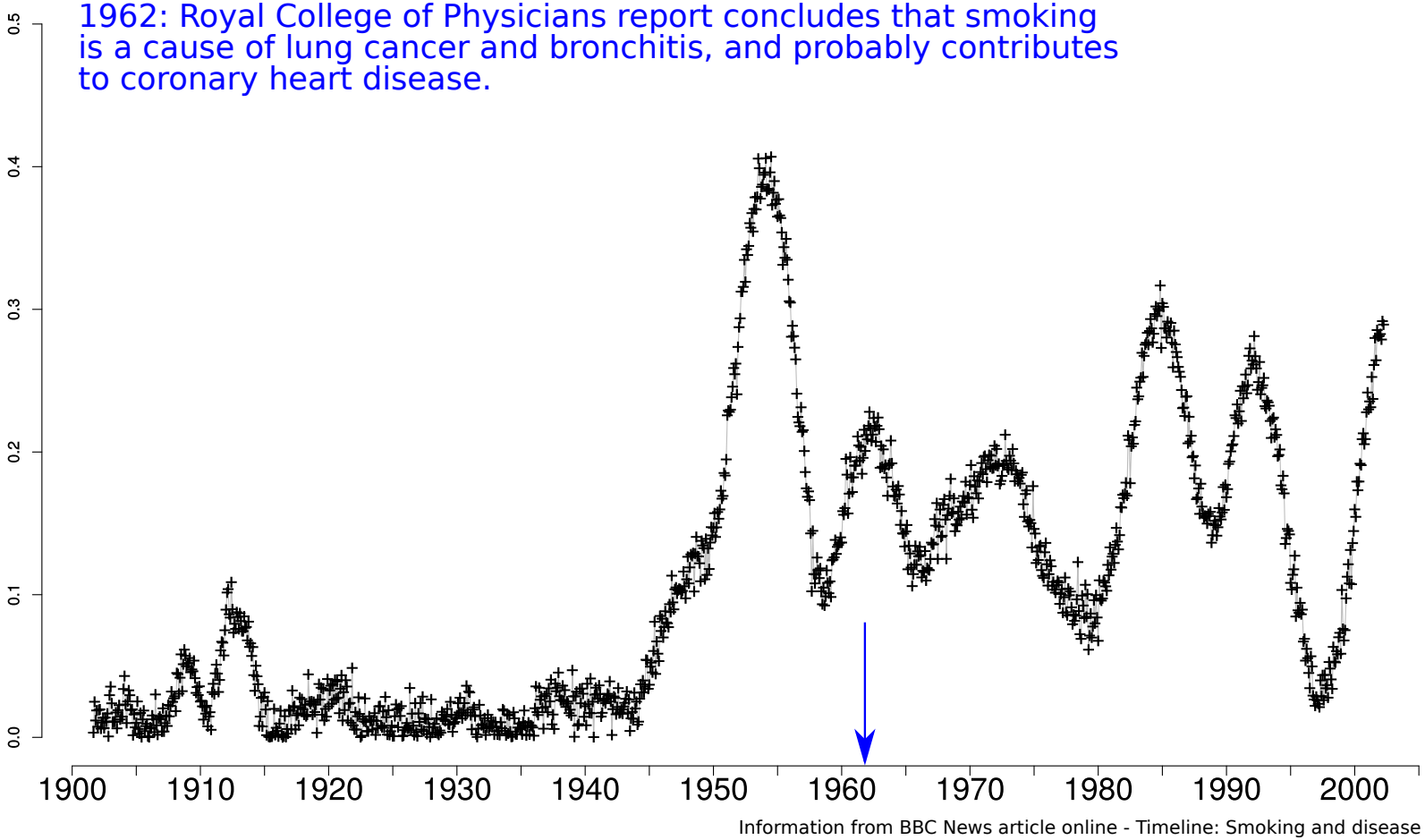
revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Using a kernel

Triweight kernel for the TDA example

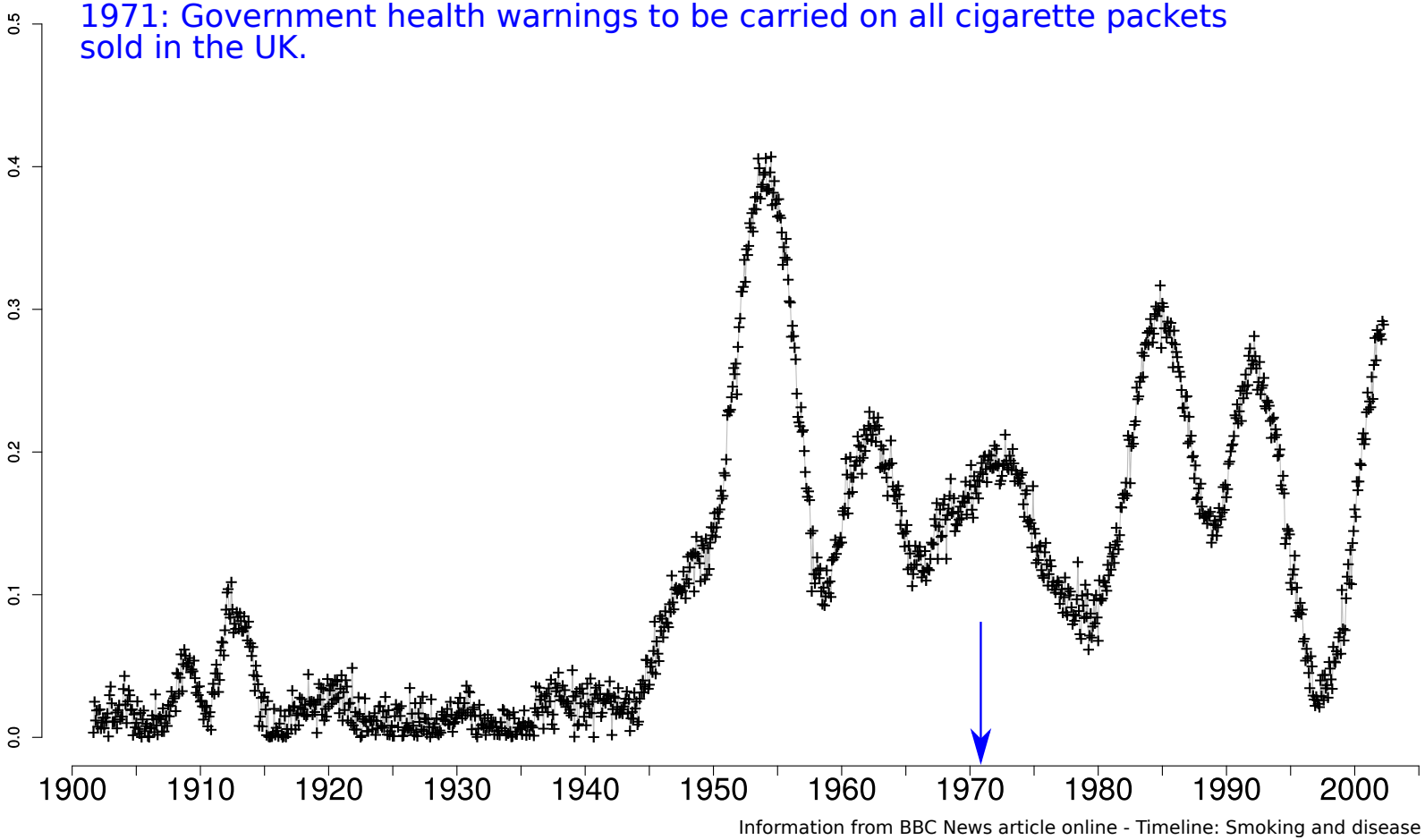
revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Using a kernel

Triweight kernel for the TDA example

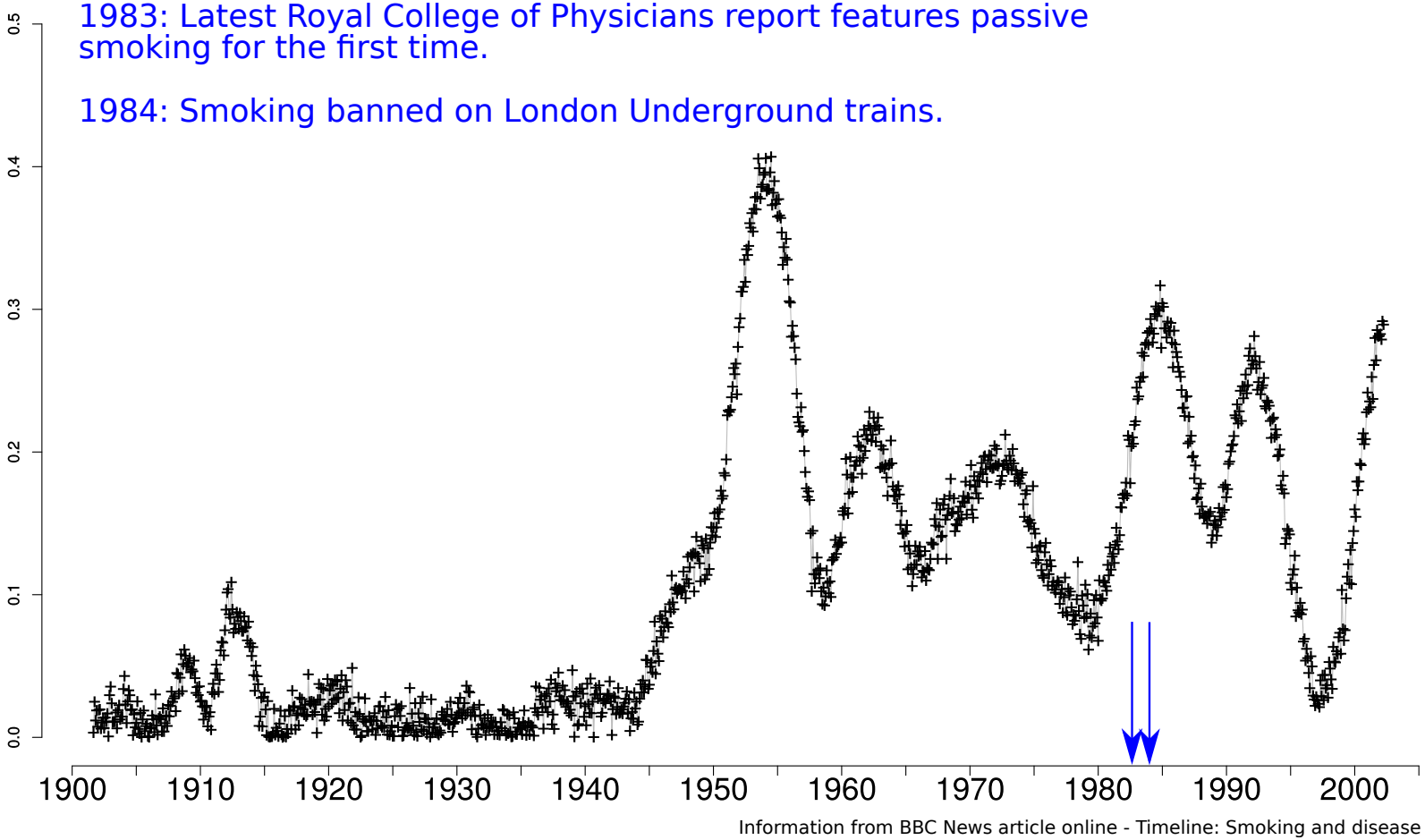
revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



Using a kernel

Triweight kernel for the TDA example

revisit $f(\mathbf{X}(t))$ for 'smoking' and 'cancer'



The End

The End.