

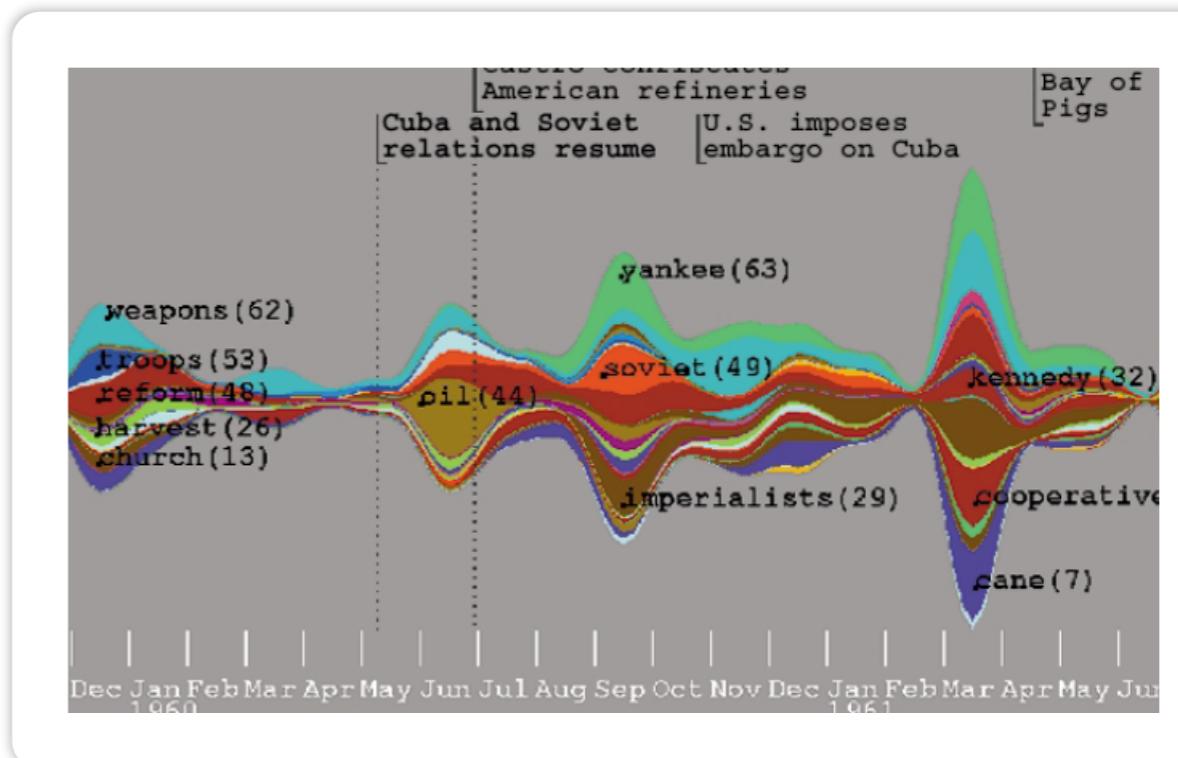
8

SÉRIES TEMPORAIS – PARTE II

Profa. Raquel C. de Melo Minardi

THEMERIVER: VISUALIZING THEME CHANGES OVER TIME

S. Havre, B. Hetzler e L. Nowell
IEEE Symposium on Information Visualization
2000



- Como os temas tratados nos textos de um determinado autor mudam com o tempo?
- *Themeriver* provê uma visão macro das mudanças temáticas em corpos de documentos na dimensão tempo

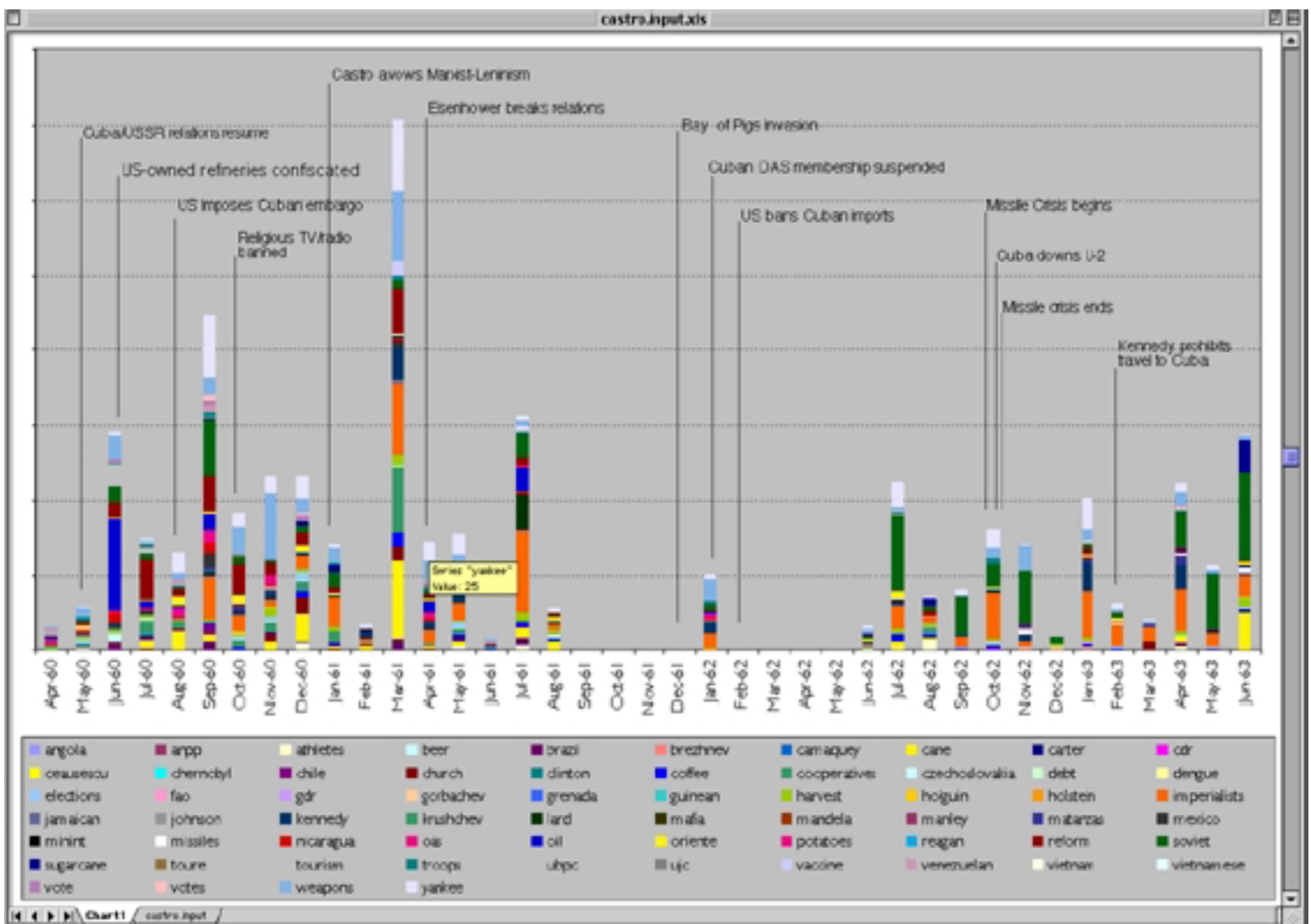
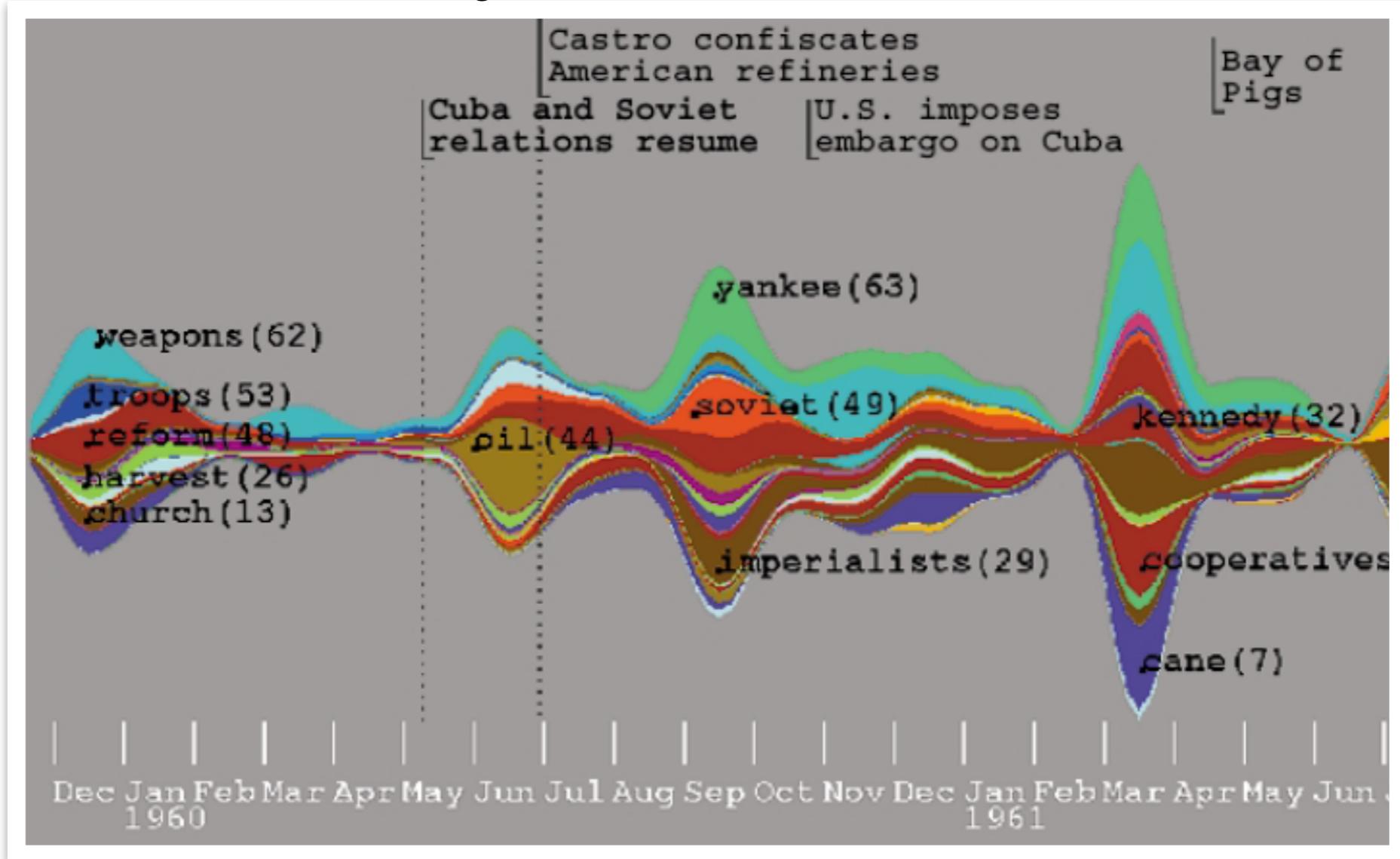


Figure 2: Like ThemeRiver™ in Figure 1, this histogram uses the Castro collection data and depicts changes in thematic content over time.

- Para que se possa responder à pergunta, é preciso associar os histogramas ao longo do tempo
- Como as barras estão ancoradas na sua base, a posição de cada tema pode variar consideravelmente ao longo do tempo, dificultando a identificação e associação dos objetos visuais

Falas, entrevistas, artigos e outros textos referentes a Fidel Castro



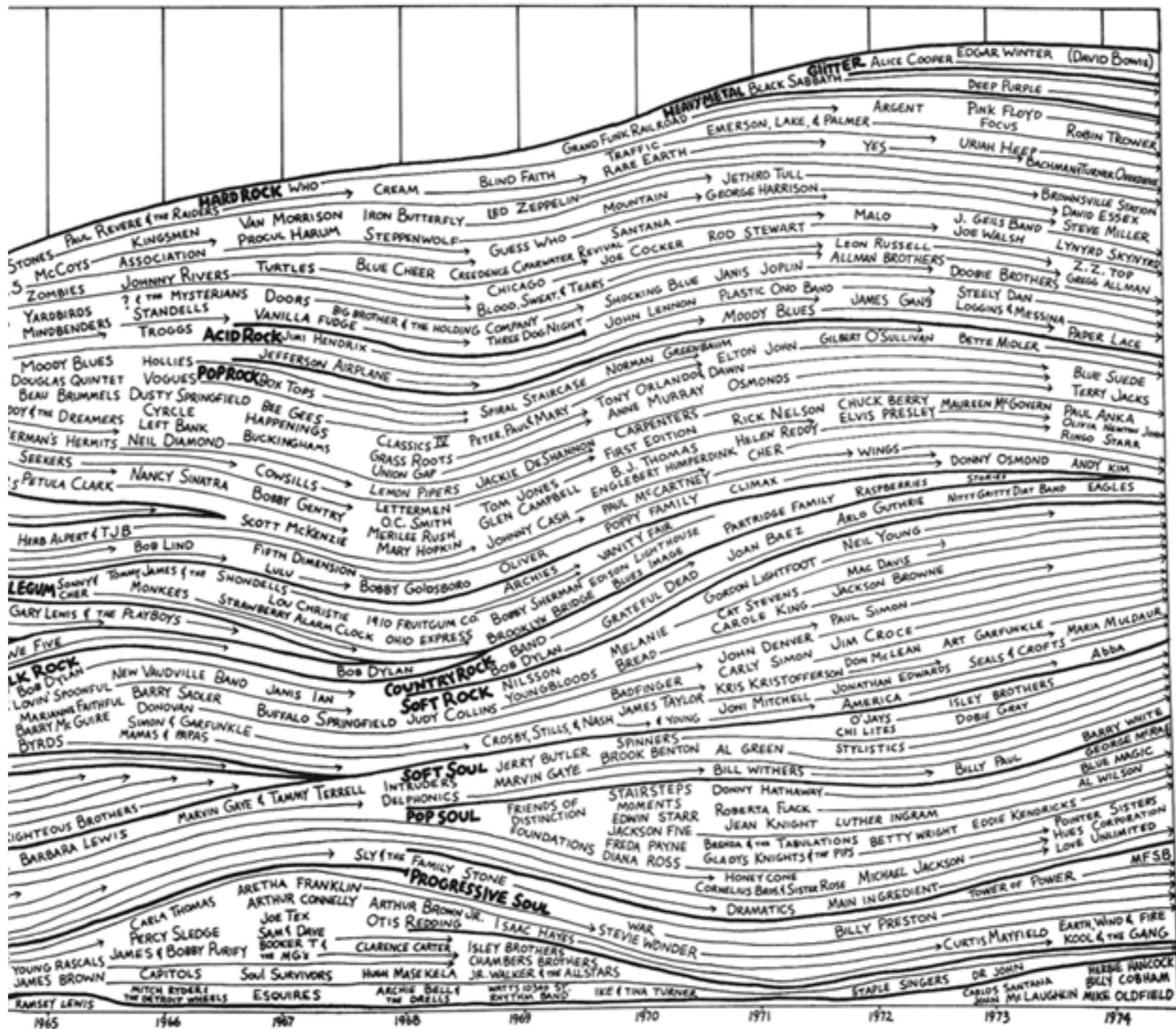
- Uso do eixo x para delimitação dos intervalos de tempo
- Uso de texto complementar como legenda dos tópicos mais relevantes

REPRESENTAÇÃO VISUAL

- Assim como no histograma, o *Themeriver* utiliza as espessuras das camadas para representar a frequência dos temas
- Porém, os temas permanecem conectados por curvas contínuas e suaves
- Metáfora:
 - O fluxo do rio representa o fluxo do tempo

are repeated, as they resurface in fresh currents. The multiple, parallel flows locate music-makers in two dimensions—*linking* musical parents and offspring from 1955 to 1974, and *listing* contemporaries for each year.¹⁰ With an intense richness of detail (measuring in at 20% of the typographic density of a telephone book), this nostalgic and engaging chart fascinates many viewers—at least those of a certain age. Also the illustration presents a somewhat divergent perspective on popular music: songs are not merely singles—unique, one-time, *de novo* happenings—rather, music and music-makers share a pattern, a context, a history.

¹⁰ Among the missing are The Weavers, Pete Seeger, Bonnie Raitt, and Lou Reed and The Velvet Underground.



PONTOS POSITIVOS

- Um tema pode aparecer, desaparecer e reaparecer e permanece facilmente reconhecível devido a manutenção da cor e posicionamento
- Exibição de informação de contexto: tempo e eventos importantes
- Exibição dos dados brutos
- Possibilidade de navegação pelo rio

German economic
and monetary union

OPEC agrees to
raise oil price

NATO to redefine
military strategy

Iraq invades
Kuwait

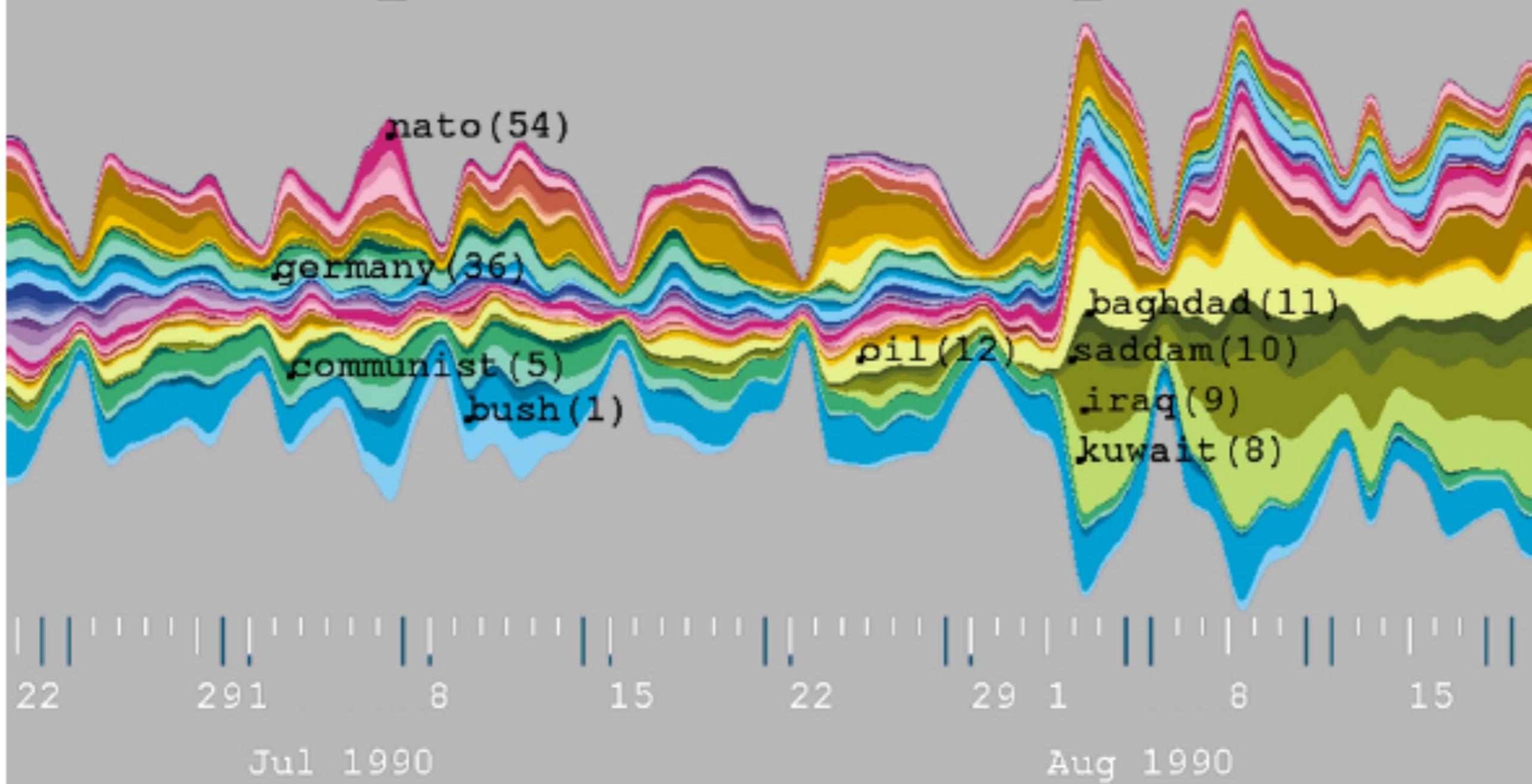


Figure 3: AP data from July - August 1990. A wide current in the river indicates heavy use of a topic, while changes in color distribution correlate to changes in themes.

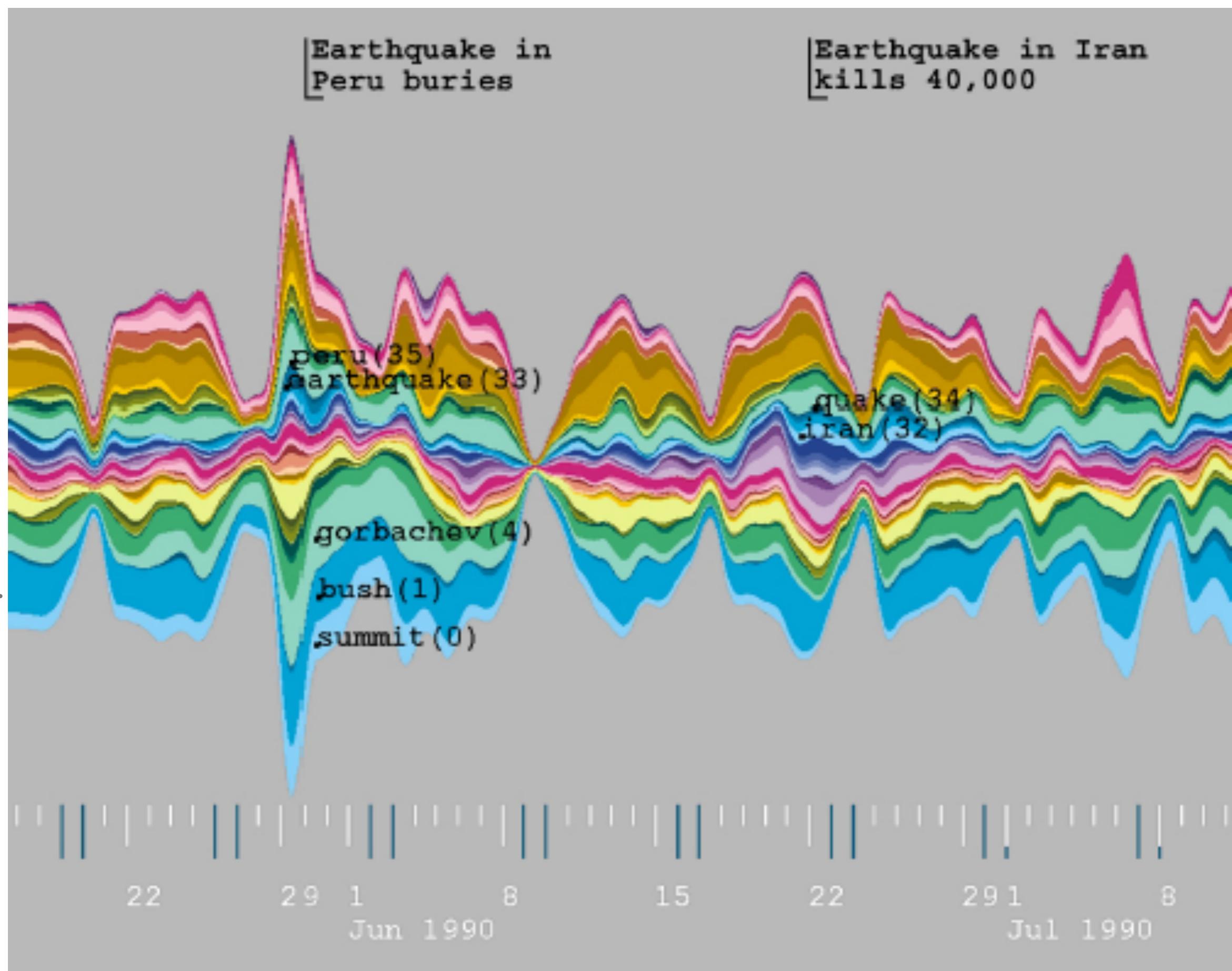


Figure 4: ThemeRiver™ of AP data from June - July 1990 identifies very different events from those revealed immediately afterwards (Figure 3).

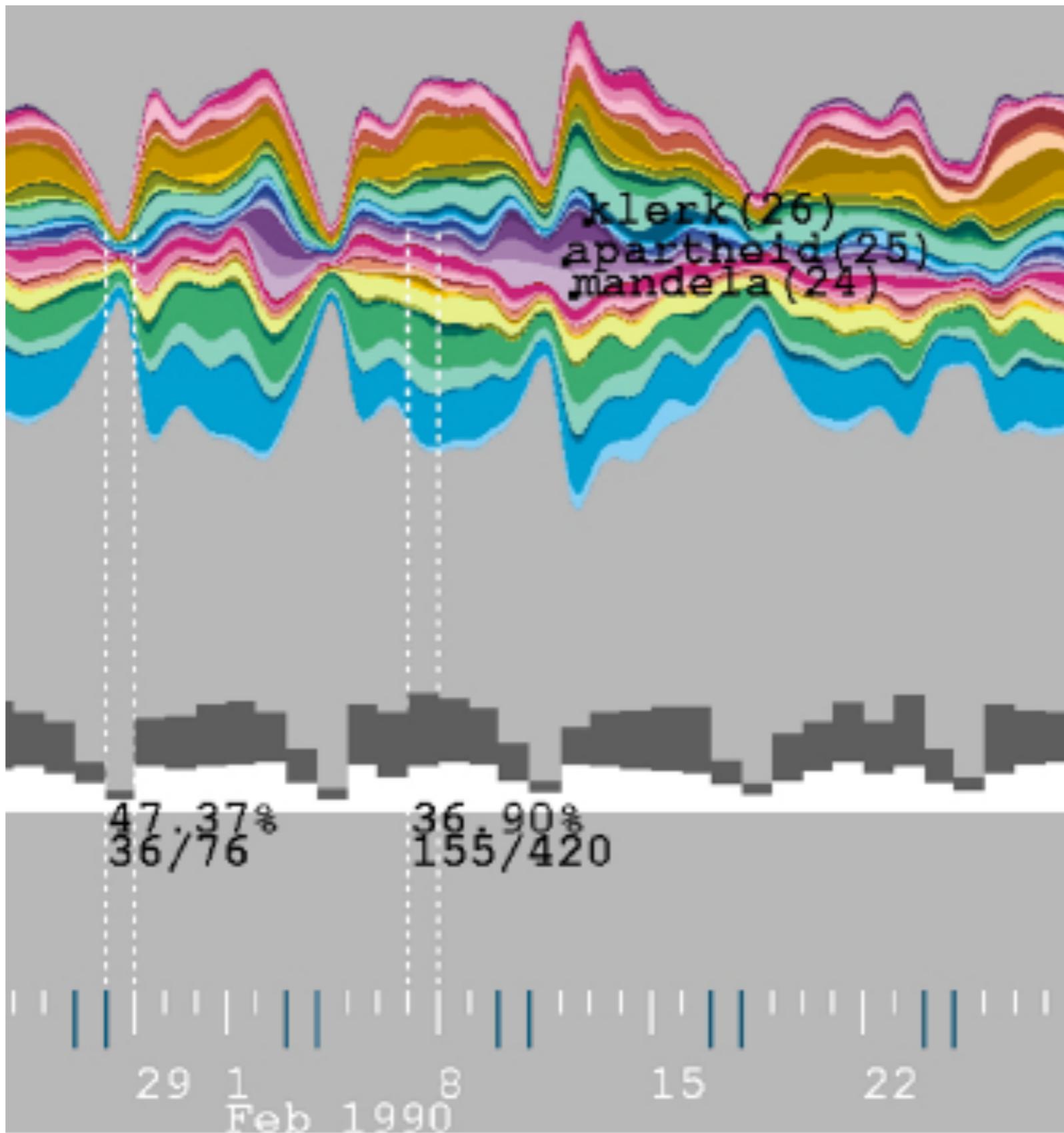


Figure 5: The addition of a histogram to Theme-River™ reveals that news is light on Sundays, not that themes shift.

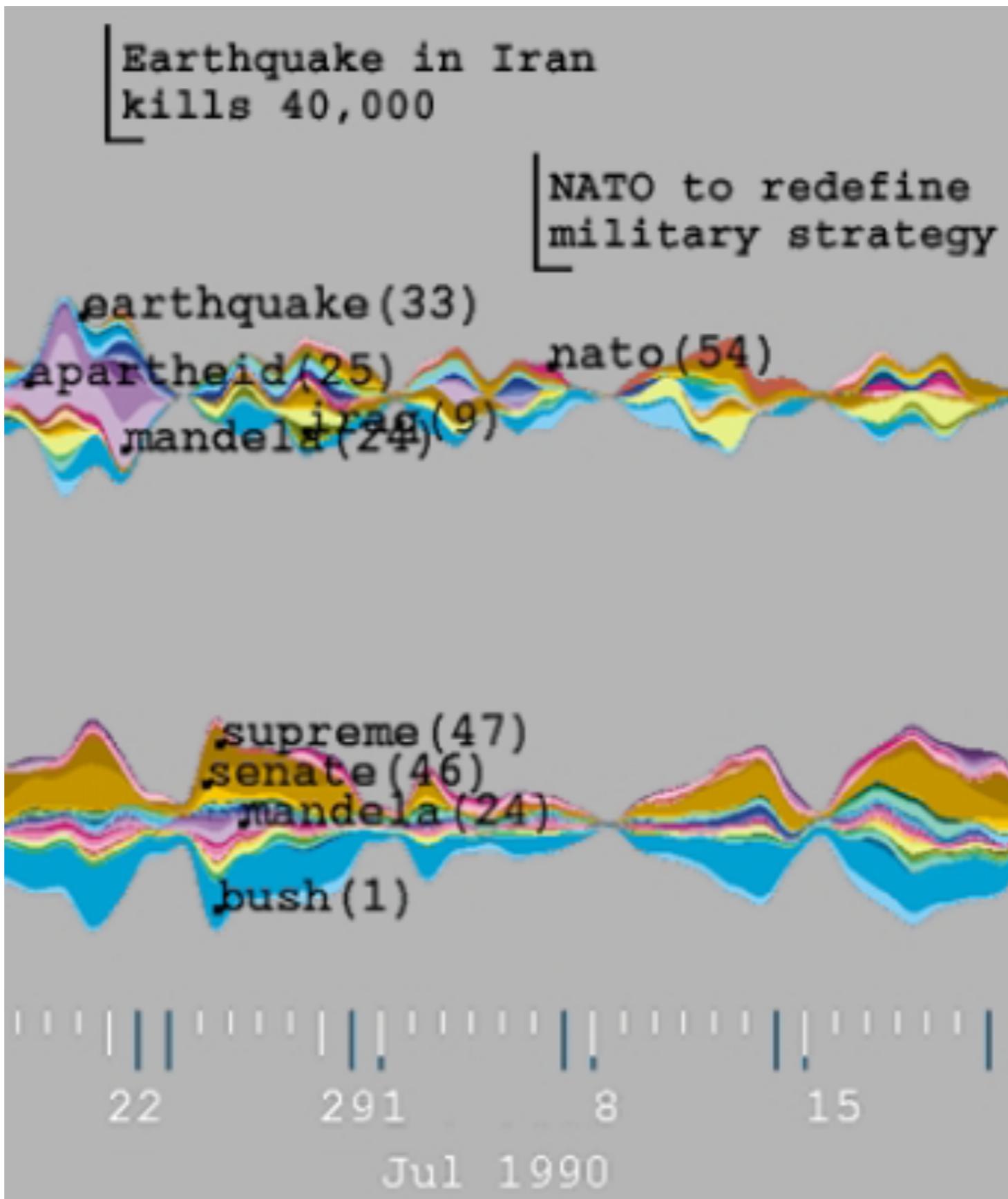


Figure 6: Parallel rivers let users compare AP data from Washington, D.C. and New York from the same time period.

DESAFIOS

- O número de termos que podem ser exibidos é limitado.
Como exibir um grande número de padrões e ainda assim manter o poder de discriminação entre eles?
- Nossa percepção de cores depende do contexto e como os temas aparecem e desaparecem, é difícil saber que cores estarão adjacentes
- Uma solução é agrupar os temas em famílias e exibi-los em cores análogas

AVALIAÇÃO

- Os usuários compreendem a metáfora do rio?
- Conseguem identificar temas mais frequentemente discutidos?
- A visualização é útil no levantamento de novas questões sobre os dados?
- Eles interpretam os dados da forma esperada?
- Como a interpretação difere da interpretação do histograma que exibe os mesmos dados?

AVALIAÇÃO

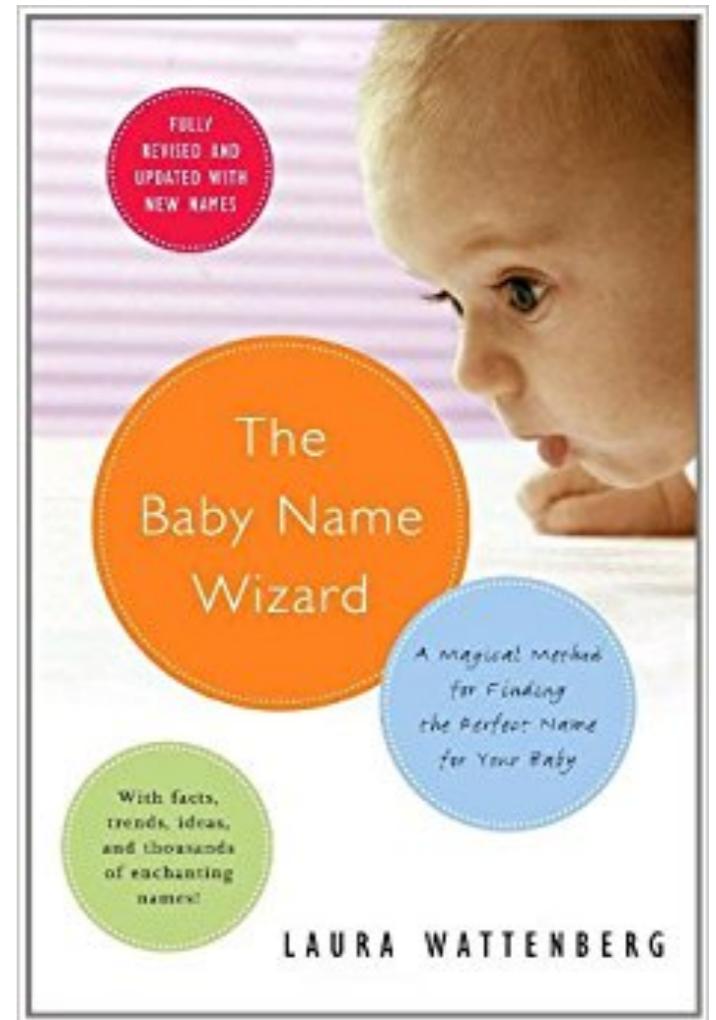
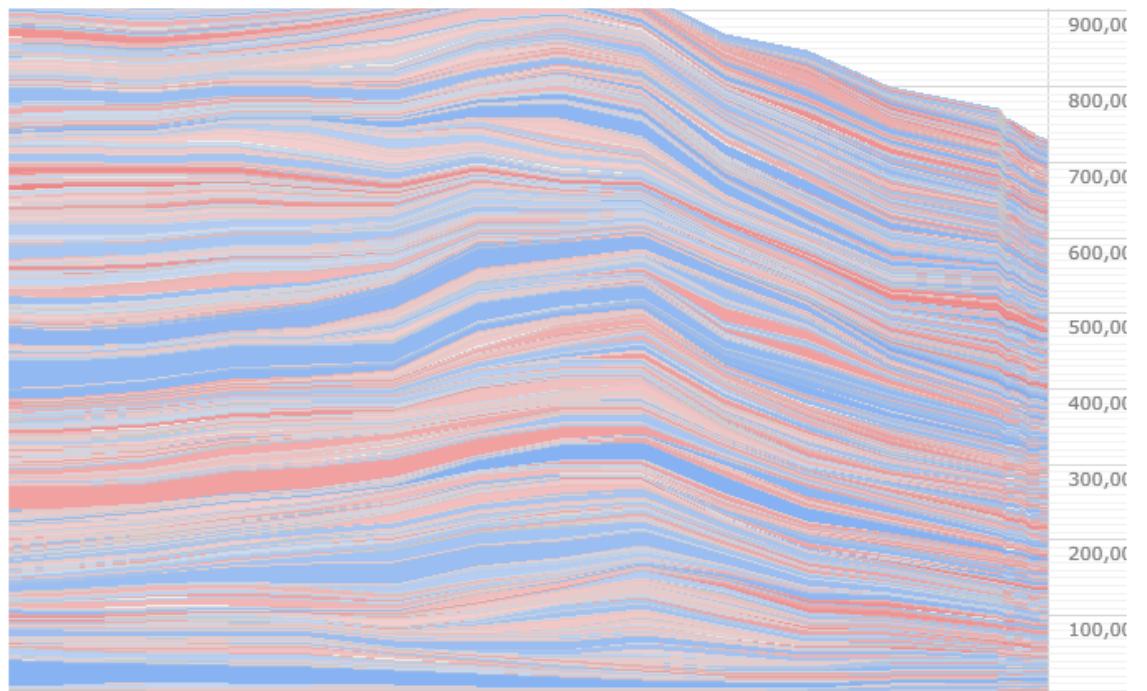
- Público
 - 2 usuários
- Questões específicas
 - Em julho de 1962, quais eram os temas mais discutidos?
 - Onde um novo tema é introduzido?
- Questões gerais
 - O que te parece interessante? O que você gostaria de explorar?
 - Como você gostaria de modificar ou manipular a visualização?

RESULTADOS DA AVALIAÇÃO

- Usuários não tiveram dificuldades em entender a metáfora
- Foram capazes de identificar temas fortemente representados e de entender o relacionamento entre a espessura da faixa e a representação do tema
- A exploração levantou questões sobre as razões da predominância de certos temas
- Acharam a visualização útil
- Acharam difícil analisar temas menos frequentes
- Sugeriram possibilitar a visão dos valores dos dados brutos (número e documentos relacionados) através da interação
- Sugeriram a possibilidade de reordenar os temas

BABY NAMES, VISUALIZATION, AND SOCIAL DATA ANALYSIS

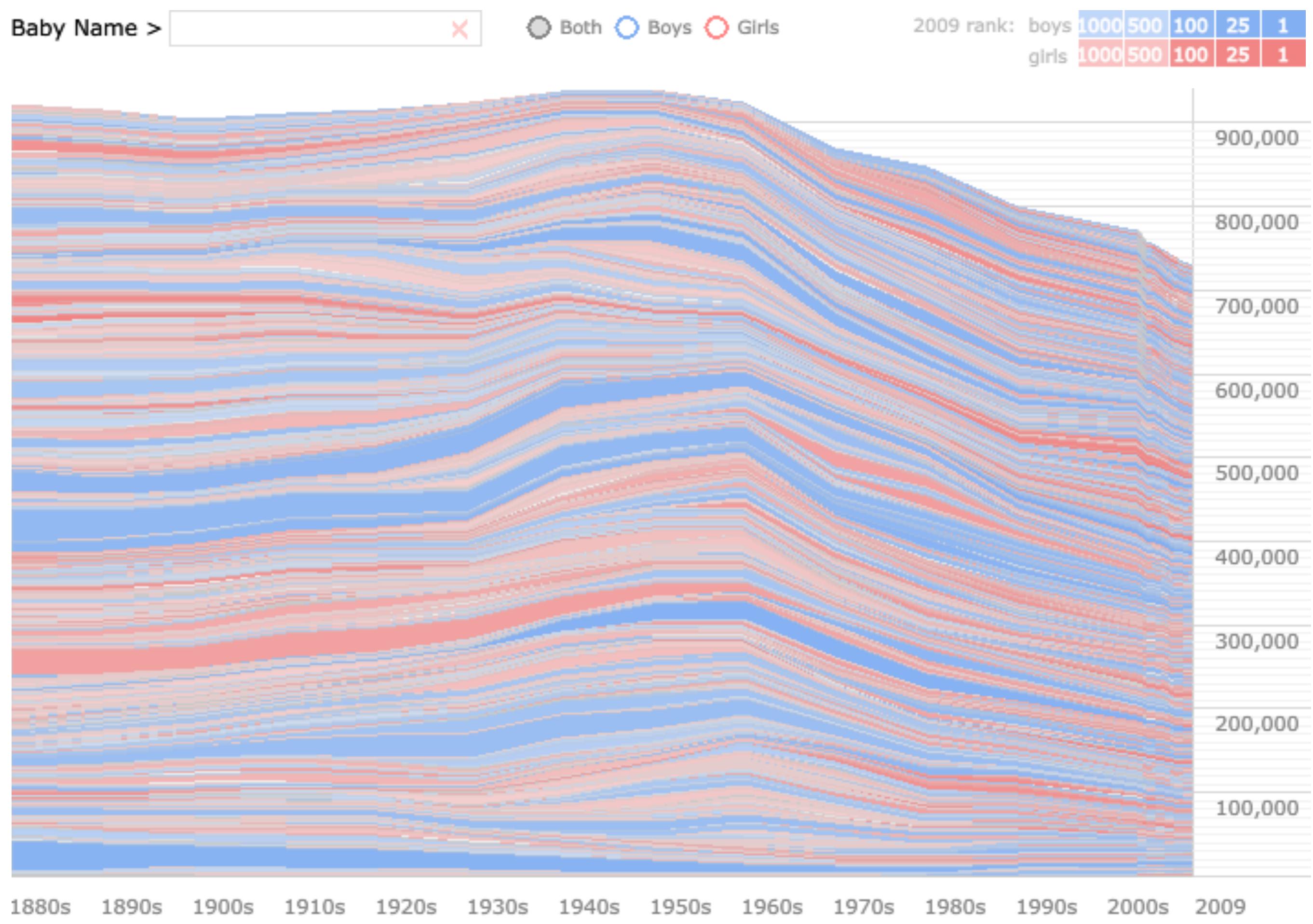
M. Wattenberg
IEEE Symposium on Information Visualization
2005



- O NameVoyager foi criado para chamar a atenção para o livro *The Baby Name Wizard*
- Ele permite a exploração de dados sobre nomes, especialmente, a popularidade dos nomes ao longo do tempo
- 500.000 acessos nas primeiras duas semanas e 10.000 acessos por dia após este período
- Pessoas engajadas com a visualização, gastando tempo considerável na exploração visual dos dados e na descoberta de fatos sobre os nomes
- 6.000 séries temporais

- Dados provenientes do Social Security Administration (SSA) dos Estados Unidos
- Para cada década desde 1.900, a lista dos 1.000 nomes de meninos e meninas mais comuns é publicada
- 6.000 nomes no total

- Wattenberg acredita que o uso do *NameVoyager* se assemelha mais ao uso de um jogo que ao uso de uma ferramenta estatística
- Contraste com a visão tradicional de que a visualização da informação é uma tarefa orientada a resolução de problemas



REPRESENTAÇÃO VISUAL

- Gráfico de área empilhada
- Eixo x representa as datas
- Eixo y representa a frequência para os nomes em termos da sua ocorrência a cada um milhão de bebês
- Cada faixa corresponde a um nome e sua largura é proporcional a frequência do nome em um dado instante do tempo

REPRESENTAÇÃO VISUAL

- As cores são baseadas em azul para meninos e rosa para meninas
- O brilho também contribui na indicação da popularidade de cada nome sendo as faixas mais escuras referentes a nomes mais populares atualmente
- Nomes populares tendem a ser mais procurados
- O uso do brilho faz que não seja necessário usar bordas nas faixas
- Desenham apenas as faixas com mais de 2 pixels de largura
- Na prática, cerca de 200 faixas

MANTRA DA VISUALIZAÇÃO DE DADOS

“

Overview first, zoom and filter, details
on demand

-B. Schneiderman

INTERATIVIDADE

- No início o usuário vê os 6.000 nomes
- Pode selecionar por sexo
- Por prefixo (applet reage a cada nova tecla pressionada)
- Ou mesmo buscar por nomes específicos

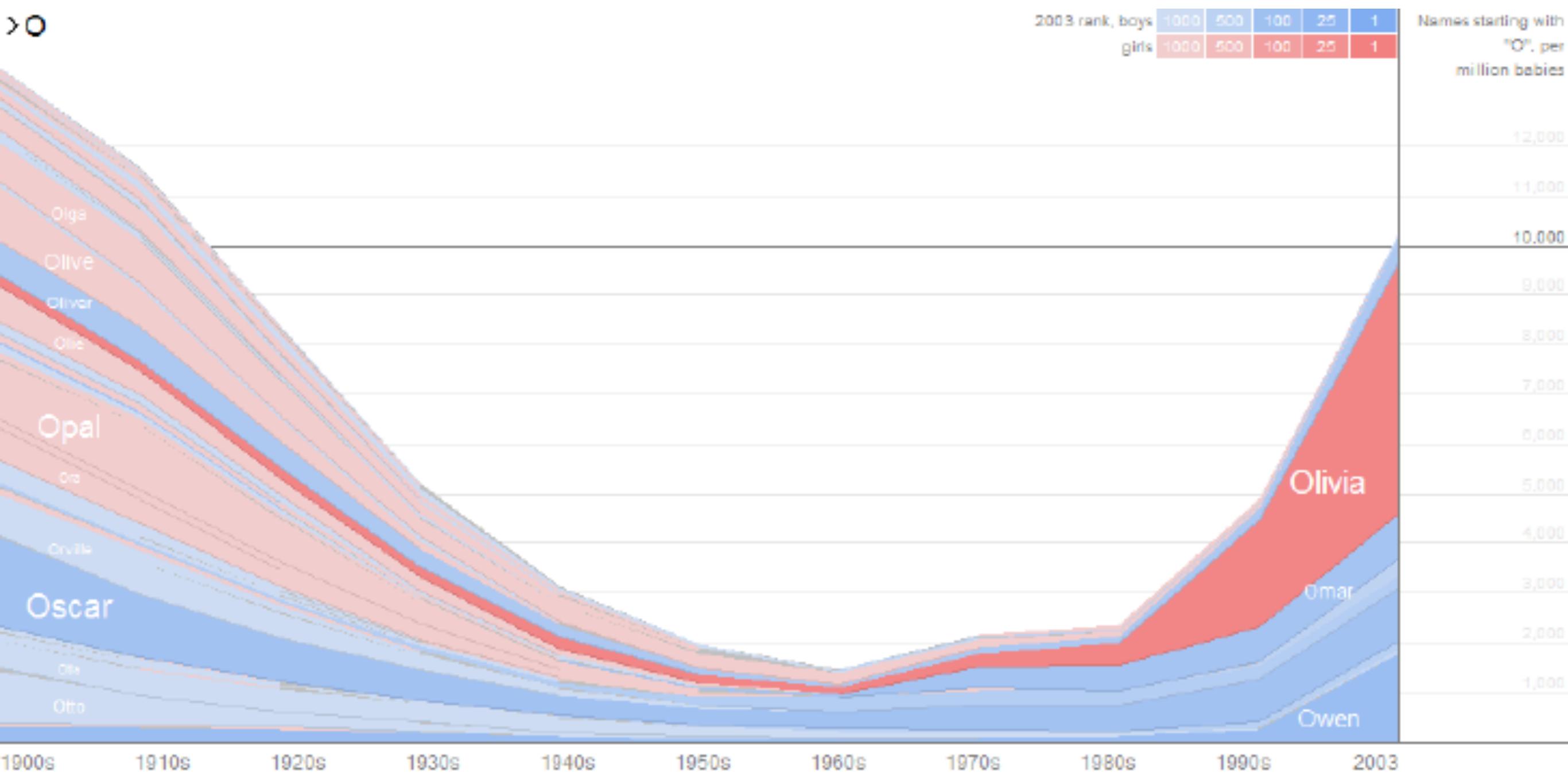


Figure 2. Names beginning with O

► LAT

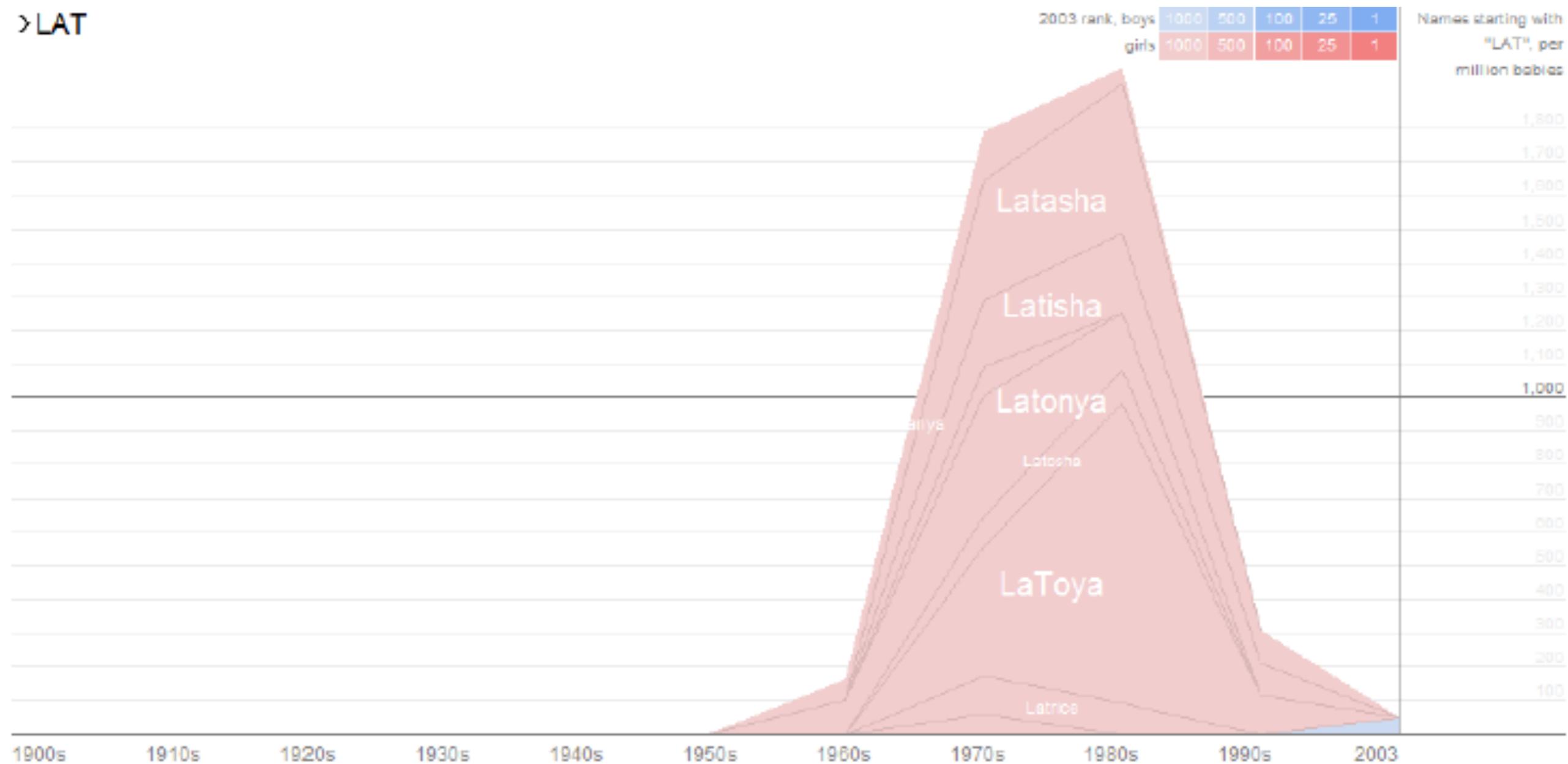
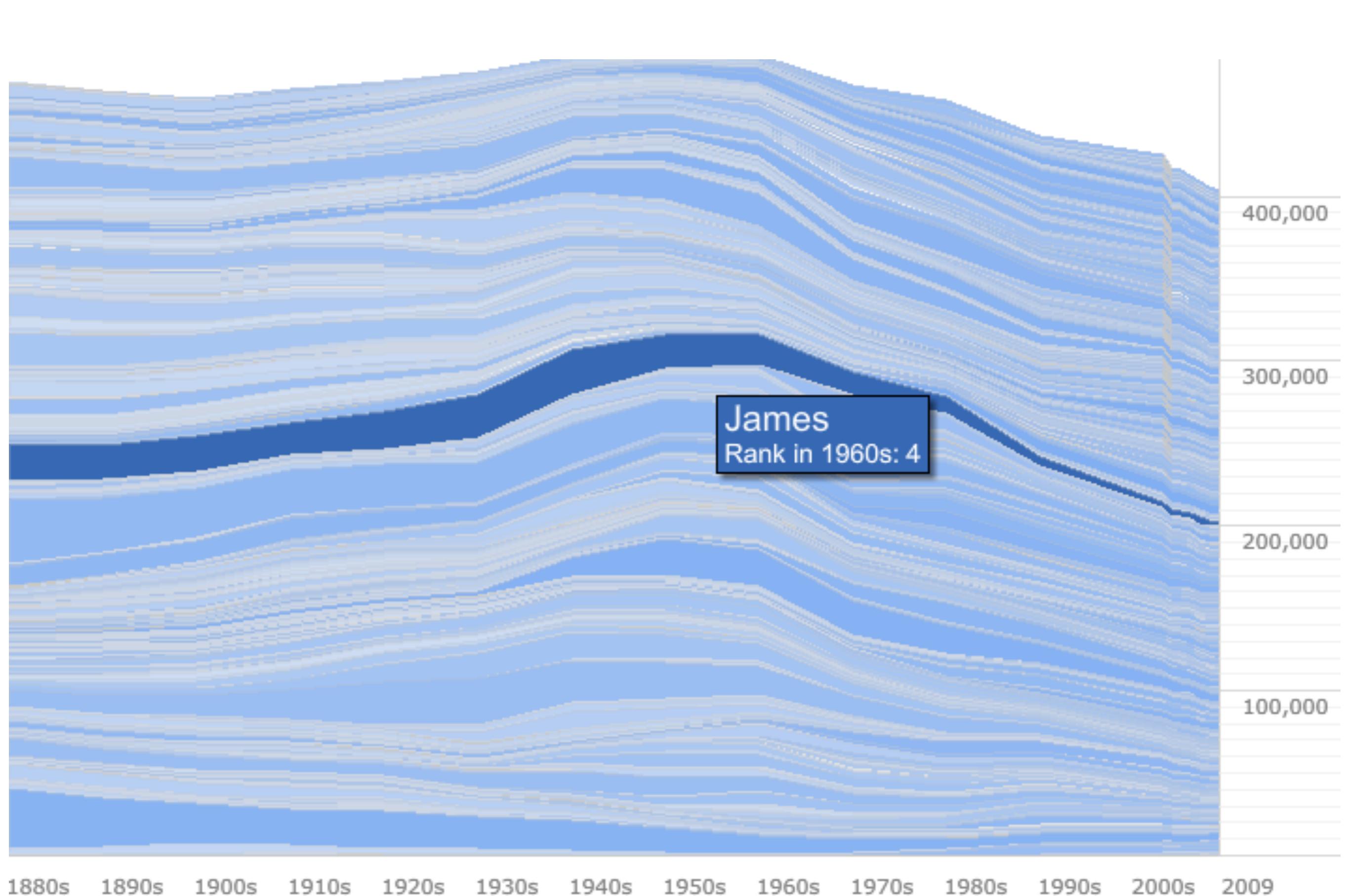
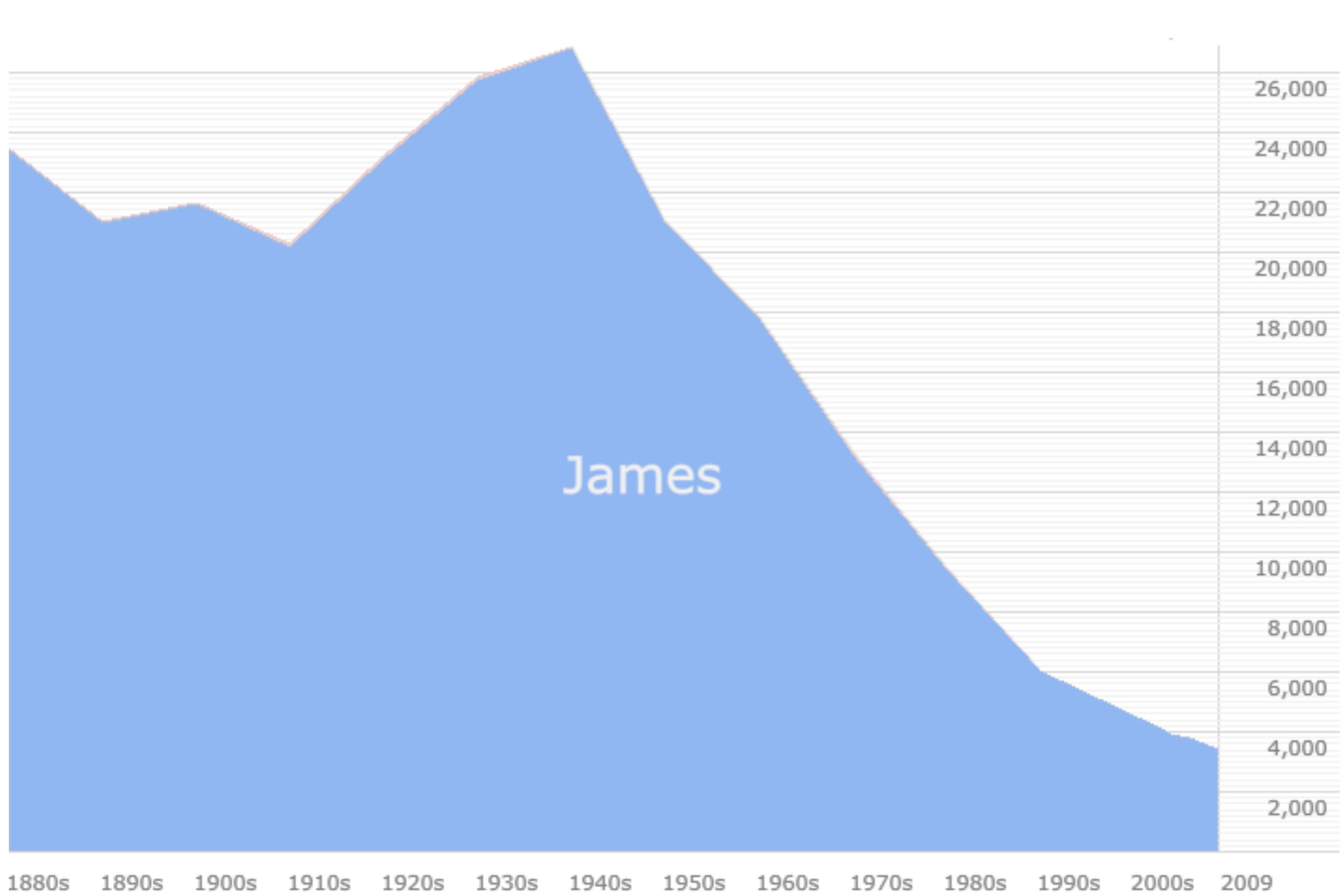


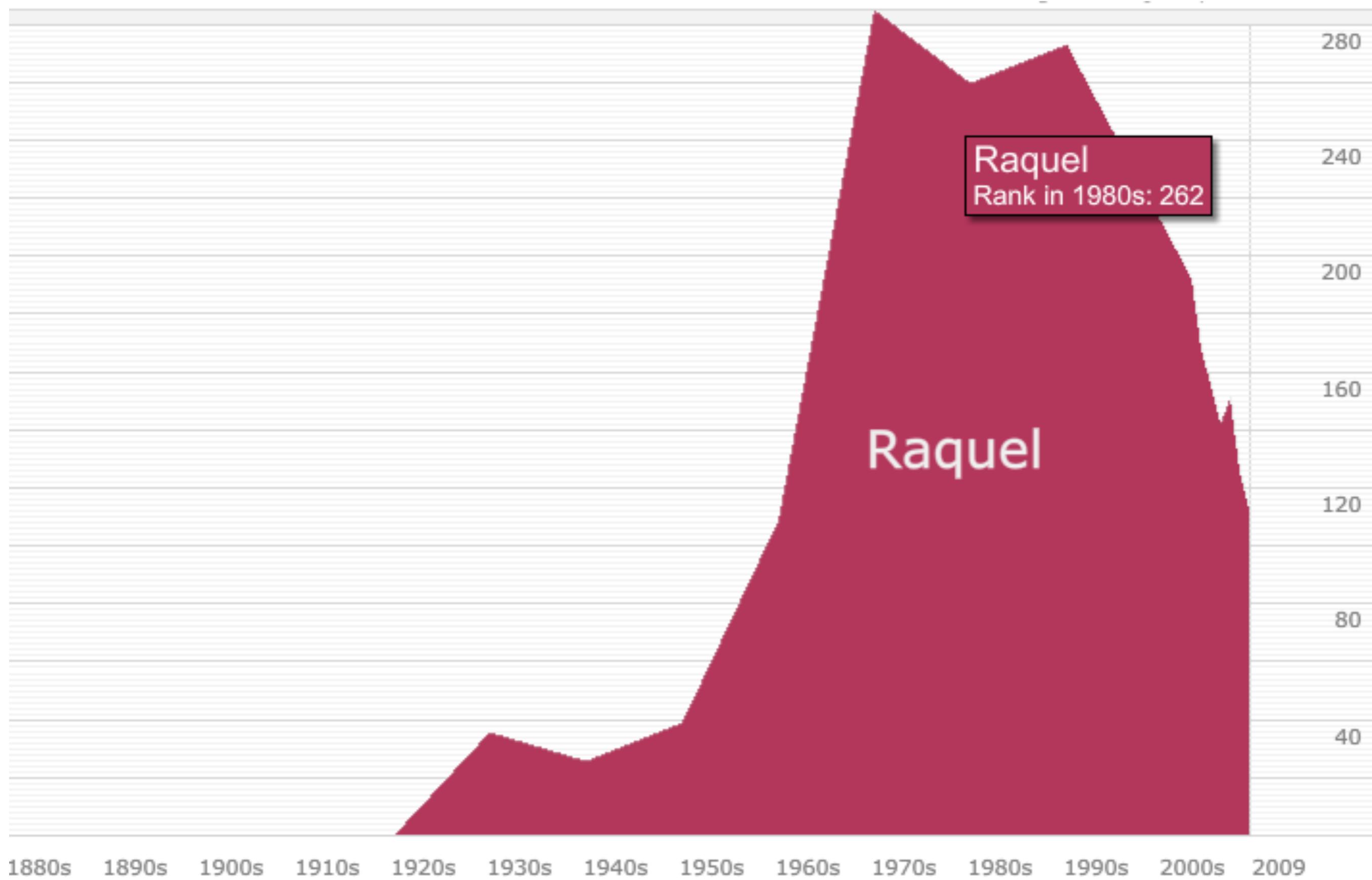
Figure 3. Names Beginning with LAT



1880s 1890s 1900s 1910s 1920s 1930s 1940s 1950s 1960s 1970s 1980s 1990s 2000s 2009



James



Raquel
Rank in 1980s: 262

Raquel

AVALIAÇÃO

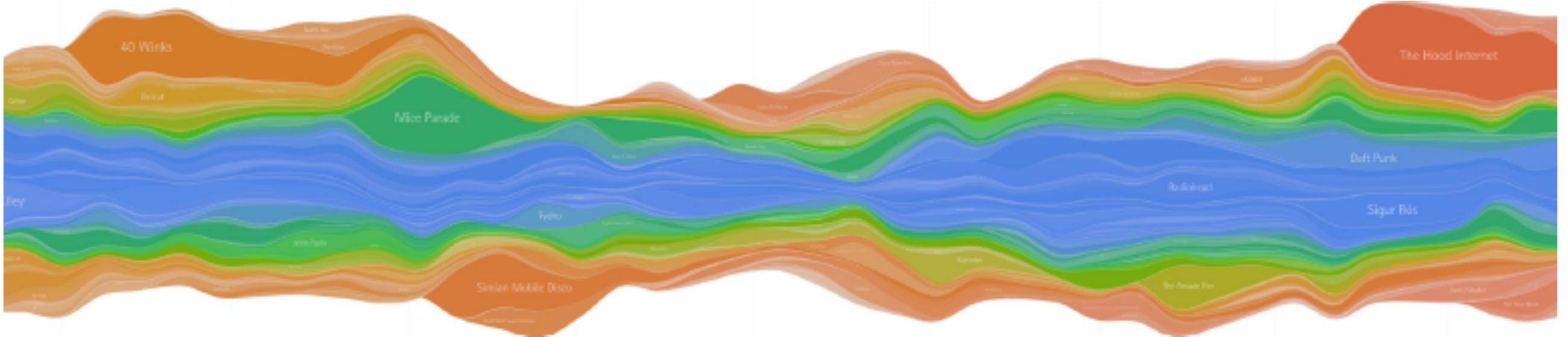
- Os autores avaliam os comentários deixados pelos usuários
- Enfatizam o caráter de atividade social da visualização, da visualização colaborativa, do diálogo na “mineração” dos dados
- Fator surpresa
- Tentam entender o motivo da popularidade da visualização

AVALIAÇÃO

- Background comum
- O uso de animações produz o efeito espectador, ou seja, usa uma interface expressiva
 - S. Reever et al., *Designing the spectator experience*, Conference for Human-Computer interaction, 2005
- Compartilhamento de descobertas

STACKED GRAPHS – GEOMETRY & AESTHETICS

*L. Byron e M. Wattenberg
IEEE Symposium on Information Visualization
2008*



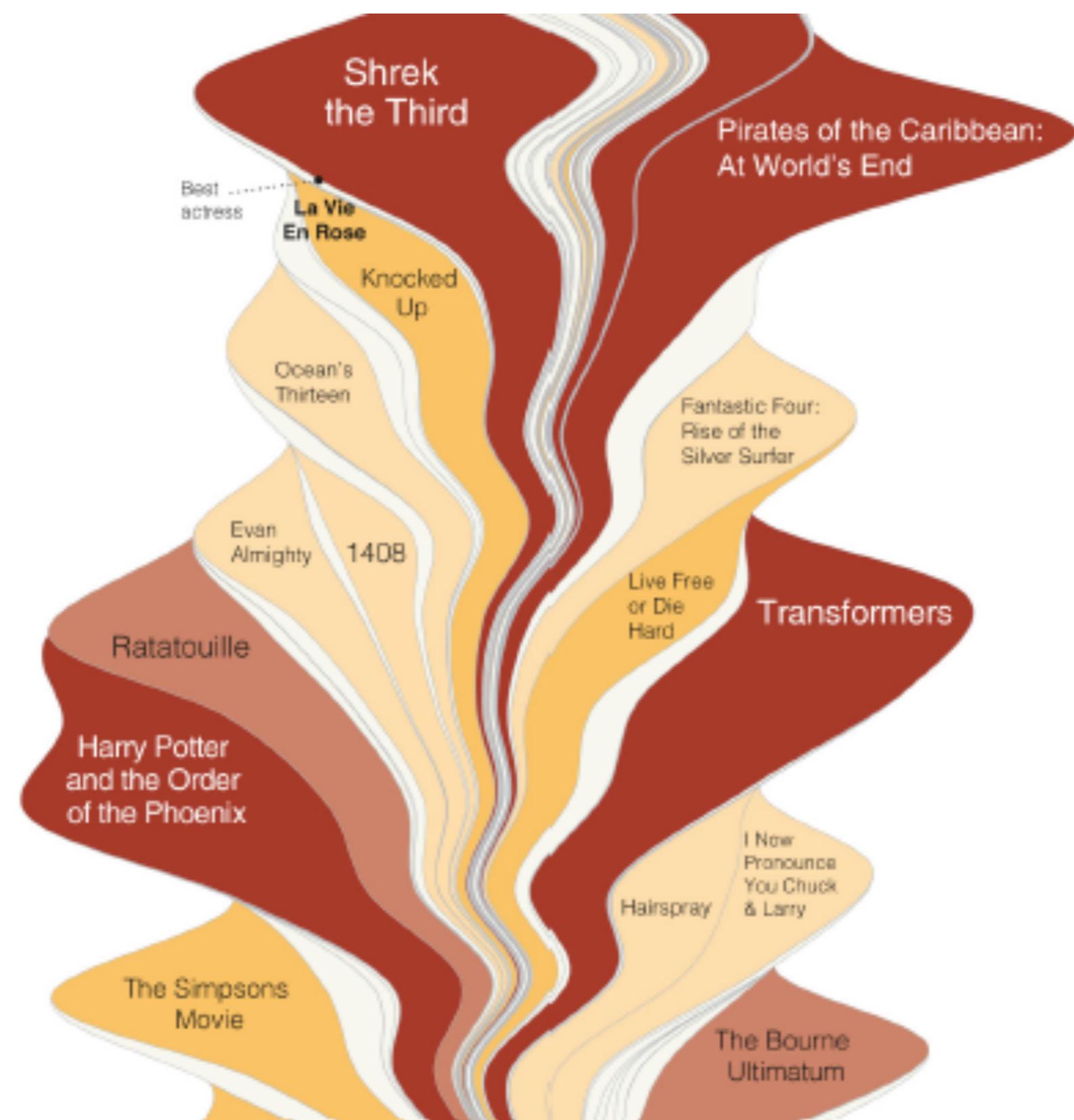


fig 2 – films from the summer of 2007

REPRESENTAÇÃO VISUAL

- Baseada em gráfico de áreas empilhadas, com ênfase nos aspectos de estética
- Eixo x representa as datas (semanas)
- Cada faixa corresponde a uma série temporal (um artista)
- A espessura da faixa representa o valor quantitativo a ser representado (número de vezes que se ouve uma música do artista)
- A cor representa a data em que um determinado artista foi descoberto pelo usuário
- A saturação indica a popularidade do artista

OBJETIVO

- Criar uma representação que não tivesse aspecto matemático ou científico, mas emocionalmente prazeroso
- Pessoas associam eventos vividos nos seus hábitos musicais

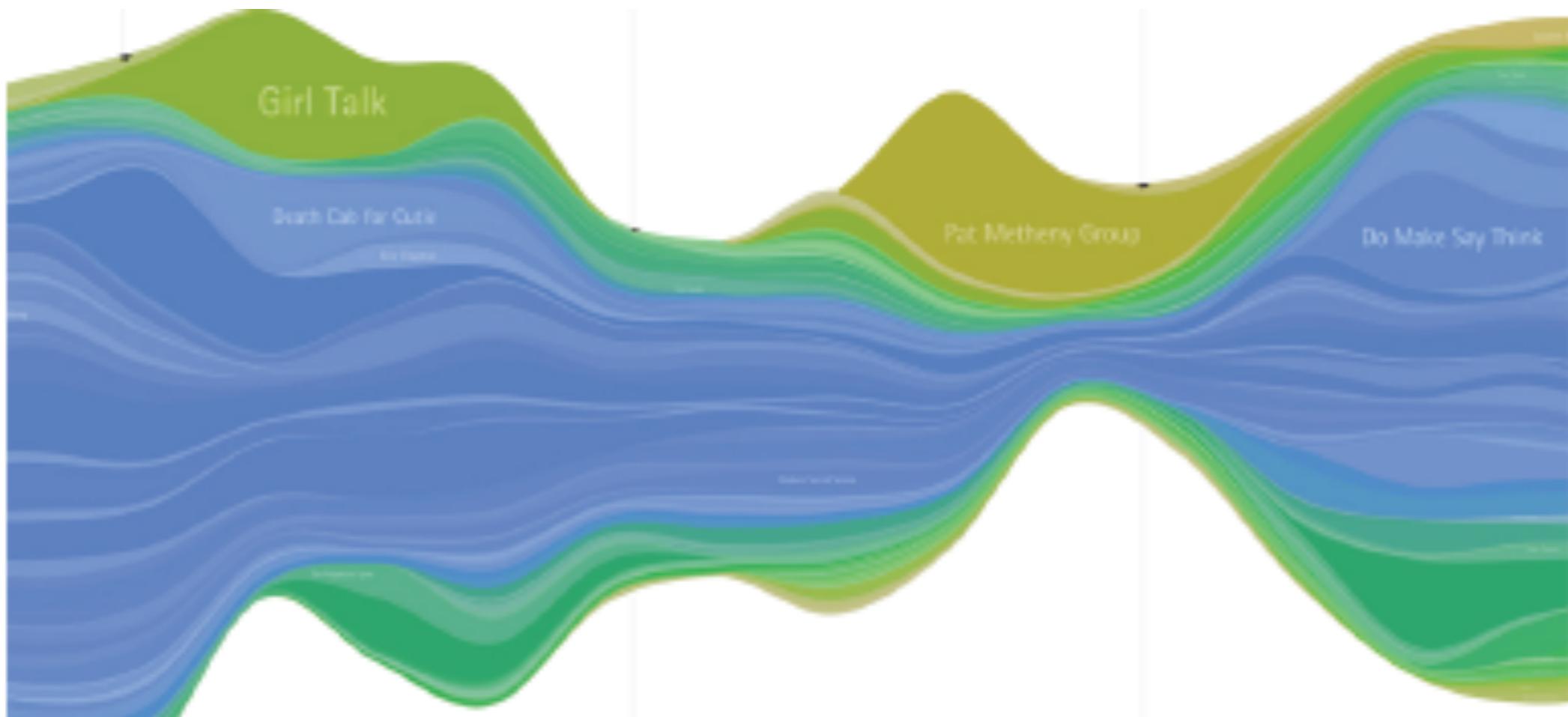
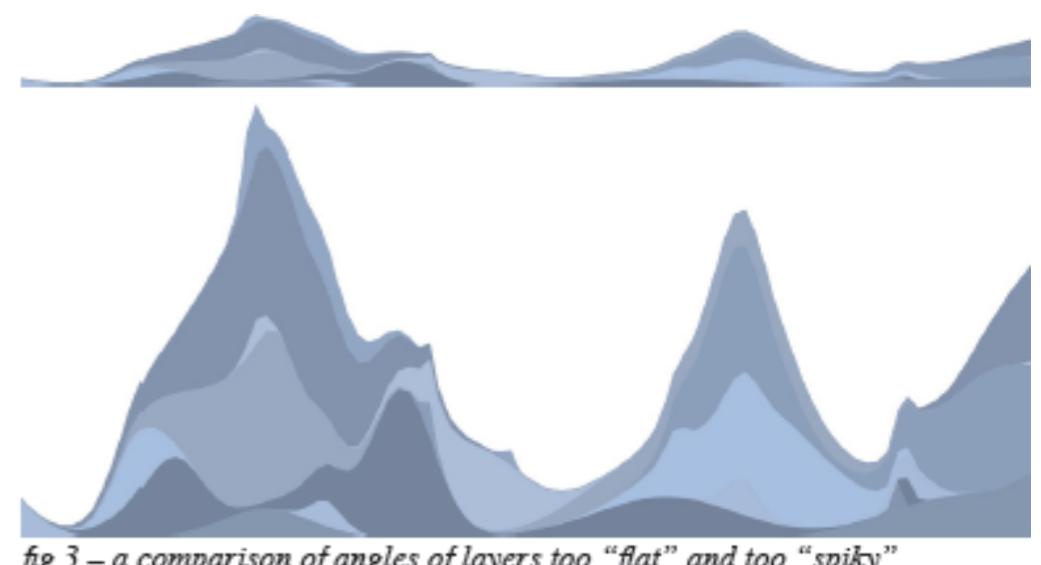


fig 1 – section from Listening History of primary author

CONSIDERAÇÕES ESTÉTICAS: LEGIBILIDADE

- A. Mudanças em camadas intermediárias provocam distorções em outras camadas ao redor
- B. Duas camadas da mesma espessura mas de inclinações diferentes podem parecer ter espessuras diferentes
- C. É preciso haver um compromisso entre o uso de gráficos muito planos e os cheios de picos (princípios da média de 45° de Cleveland)
- D. É preciso garantir que as camadas possam ser facilmente diferenciáveis
- E. Considerar o aspecto da estética



ALGORITMO

- É preciso considerar propriedades da silhueta do gráfico uma vez que ela implica nas inclinações e nas curvaturas de cada camada
- A ordenação das camadas é outro aspecto importante podendo levar a visualizações totalmente diferentes
- Os rótulos das camadas são essenciais para facilitar a identificação dos dados através dos objetos visuais
- As cores distinguem as camadas e podem ainda representar outras dimensões dos dados

g_0 é chamada base line

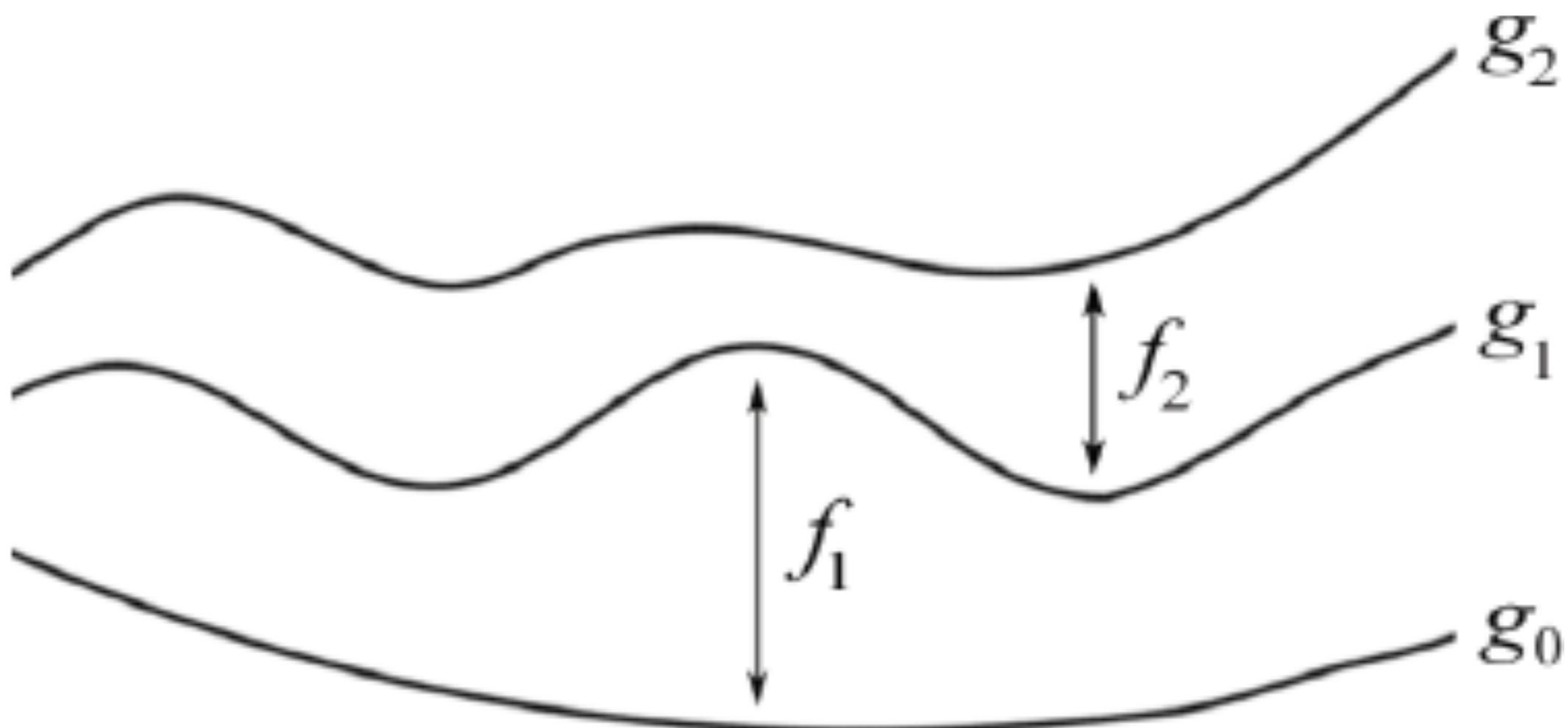


fig 4 – a visual description of stacked graph functions f_i and g_i for $n=2$ as used in this section

Assim o topo do gráfico é dado por

$$g_i = g_0 + \sum_{j=1}^i f_j$$

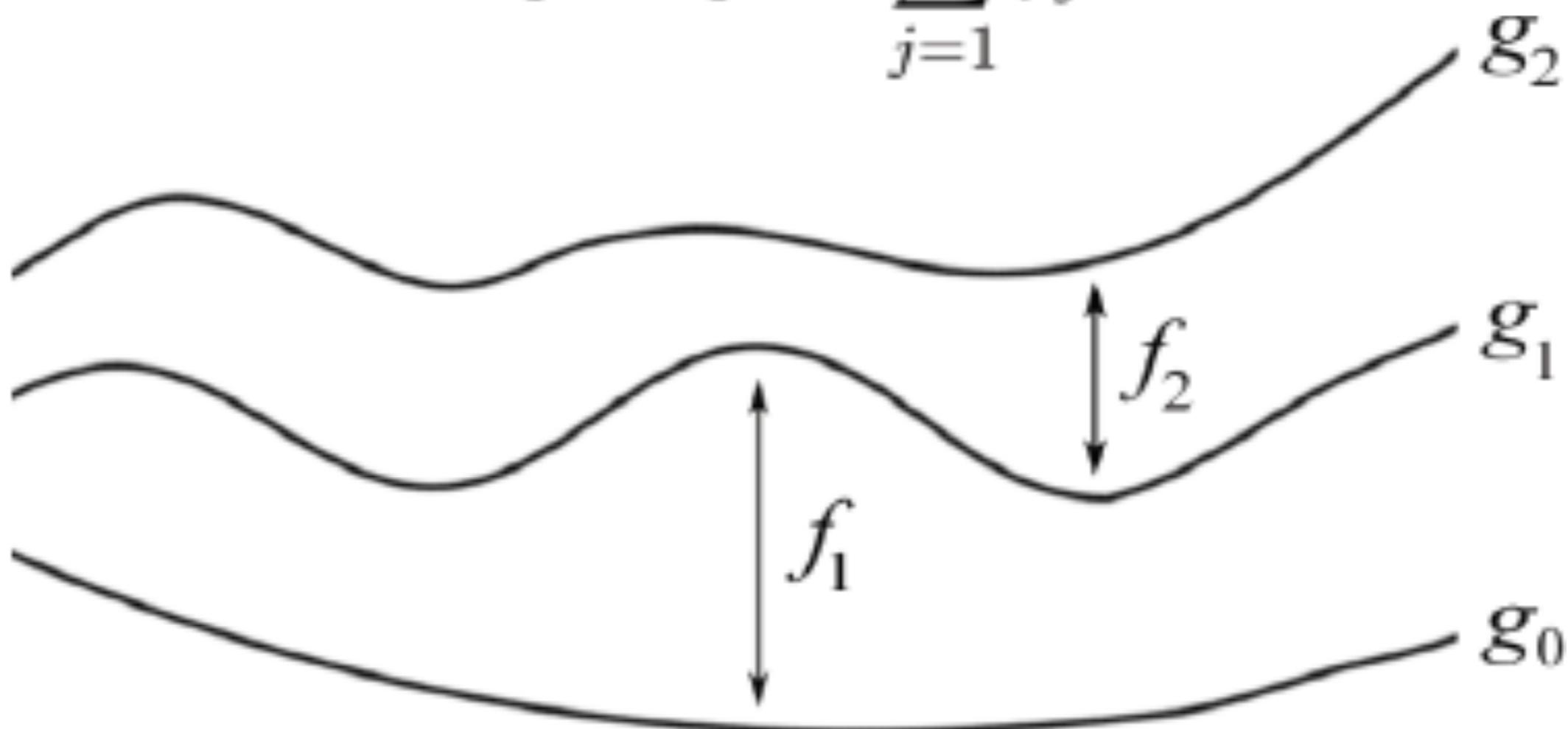


fig 4 – a visual description of stacked graph functions f_i and g_i for $n=2$ as used in this section

Quando $g_0 = 0$

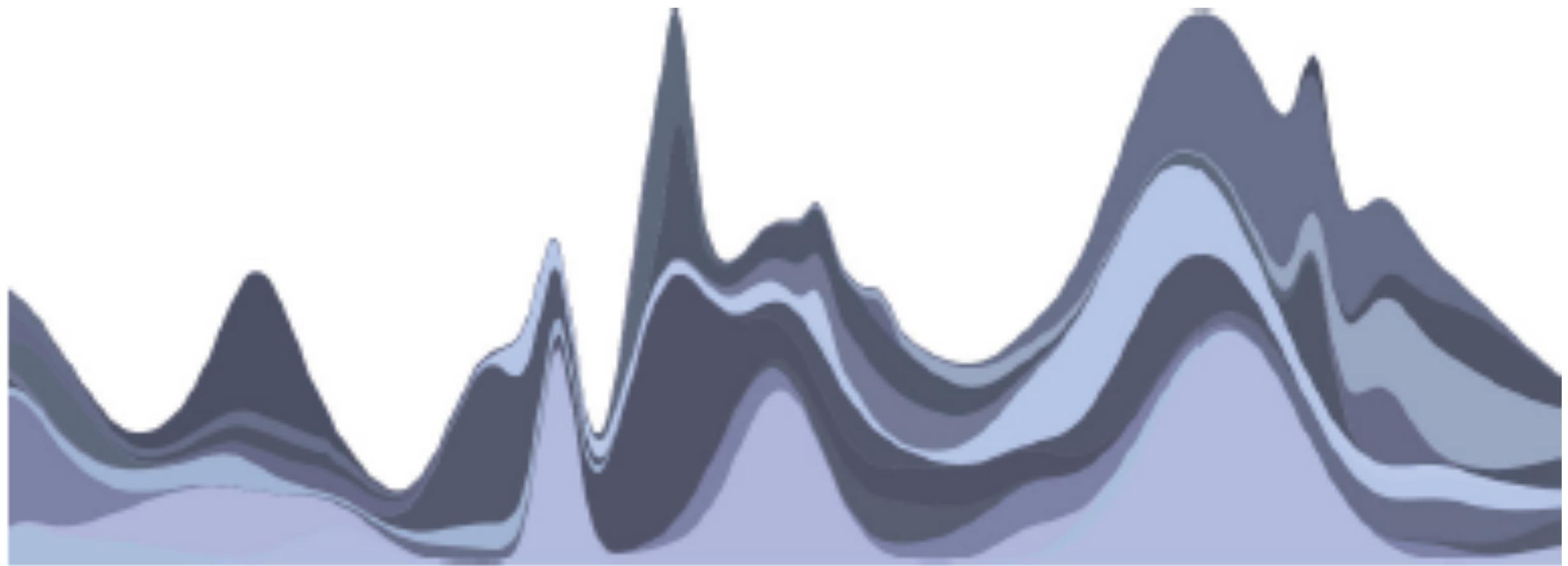


fig 5 – A traditional stacked graph with a baseline $g_0 = 0$

A soma das séries é legível mas a legibilidade de séries individuais fica prejudicada

A solução proposta pelo autor do *Themeriver* para este problema é o uso de um layout simétrico em relação a x

$$g_0 + g_n = 0$$

Como

$$g_i = g_0 + \sum_{j=1}^i f_j$$

então

e

$$2g_0 + \sum_{i=1}^n f_i = 0 \qquad g_0 = -\frac{1}{2} \sum_{i=1}^n f_i$$

É possível mostrar que este procedimento diminui a silhueta da curva deixando a curva mais suave

Os autores propõem dois novos procedimentos para melhorar a estética das curvas que consistem em obter g_0 visando:

1. Minimizar a inclinação em cada valor de X

$$wiggle(g_0) = \sum_{i=0}^n g_i'^2 = \sum_{i=0}^n (g_0' + \sum_{j=1}^i f_j')^2$$

Como o valor de X que minimiza

$$\sum_{i=1}^n (x + a_i)^2$$

is

$$x = -\frac{1}{n} \sum_{i=1}^n a_i$$

$$g'_0 = -\frac{1}{n+1} \sum_{i=0}^n \sum_{j=1}^i f'_j$$

2. Minimizar as médias dos quadrados das inclinações em pontos médios de cada camada ponderando pela espessura da camada

$$\text{weighted_wiggle}(g_0) = \sum_{i=1}^n \left(\frac{1}{2}(g'_i + g'_{i-1}) \right)^2 f_i = \sum_{i=1}^n \left(g'_0 + \frac{1}{2}f'_i + \sum_{j=1}^{i-1} f'_j \right)^2 f_i$$

$$g'_0 = -\frac{1}{\sum f_i} \sum_{i=0}^n \left(\frac{1}{2}f'_i + \sum_{j=1}^{i-1} f'_j \right) f_i$$

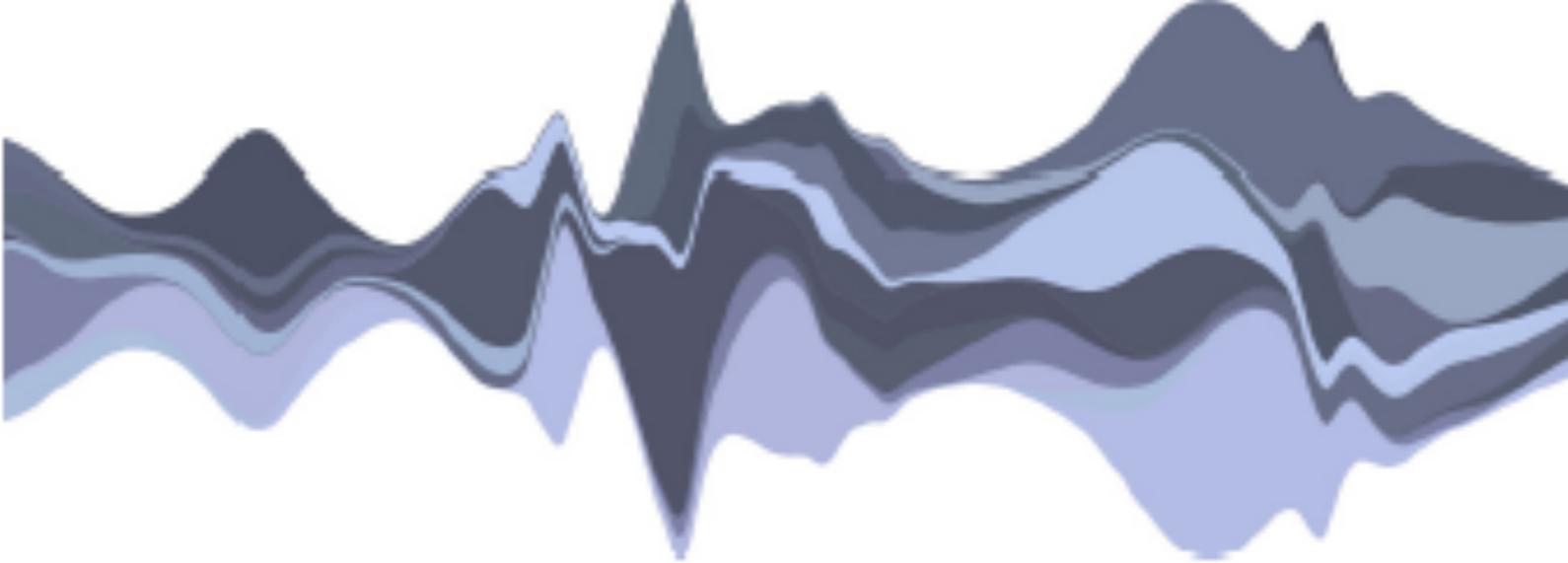


fig 6 – the same data set using the ThemeRiver layout algorithm

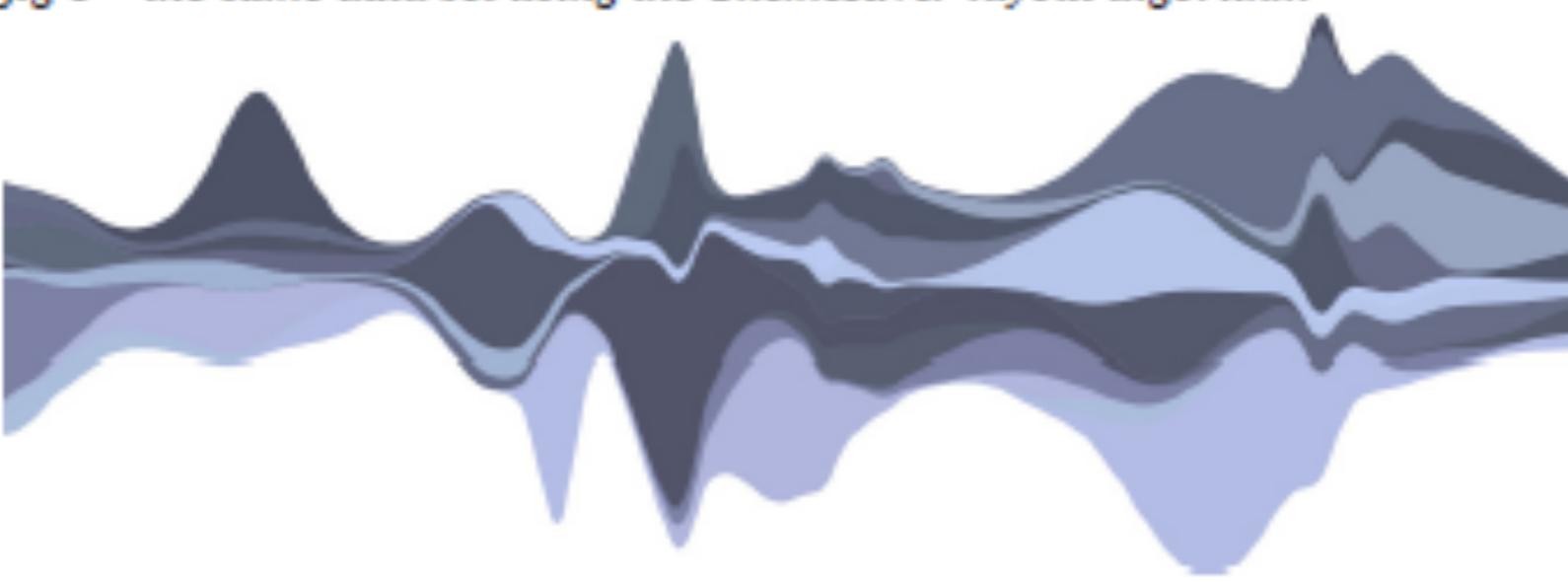


fig 7 – the same data set optimized to reduce the "wiggle" function, or overall variation in slope

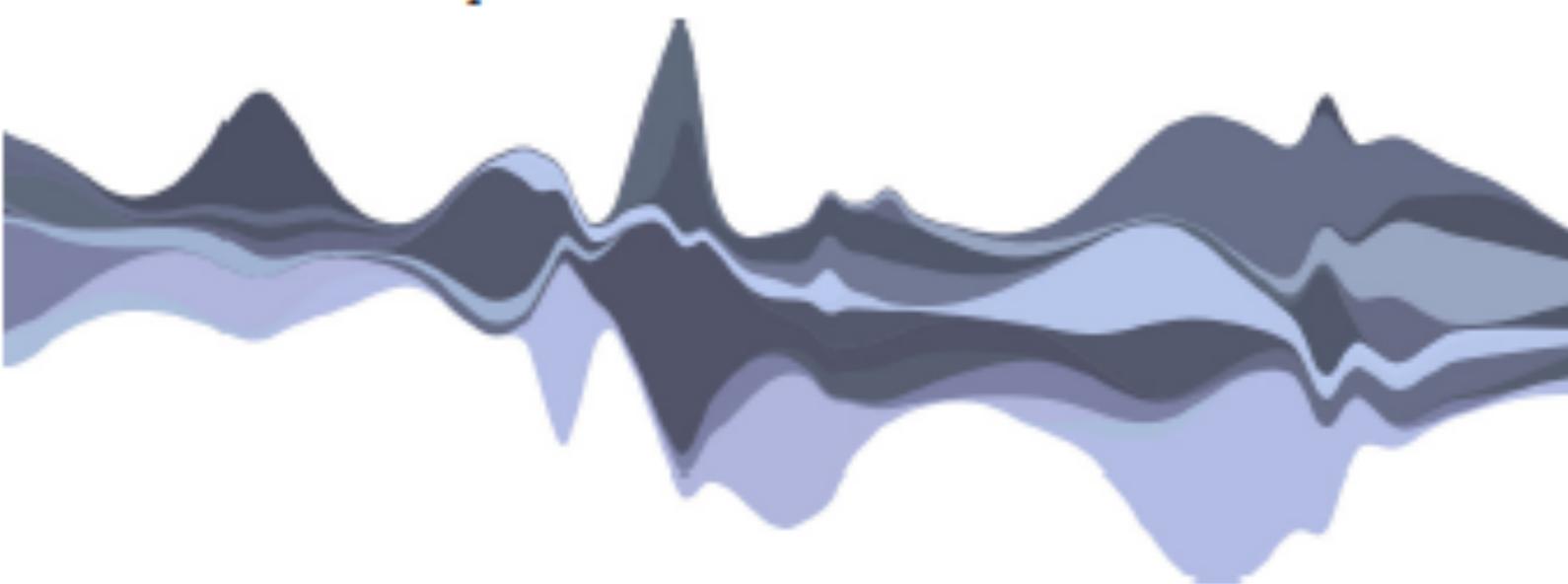


fig 8 – the same data set optimized to reduce the "weighted_wiggle," the algorithm used in Streamgraph

Paleta de cores gerada com o uso do Photoshop através de **imagens altamente saturadas da natureza**

Não usam todos os matizes existentes mas selecionam cores altamente contrastantes para revelar os diferentes comportamentos das séries

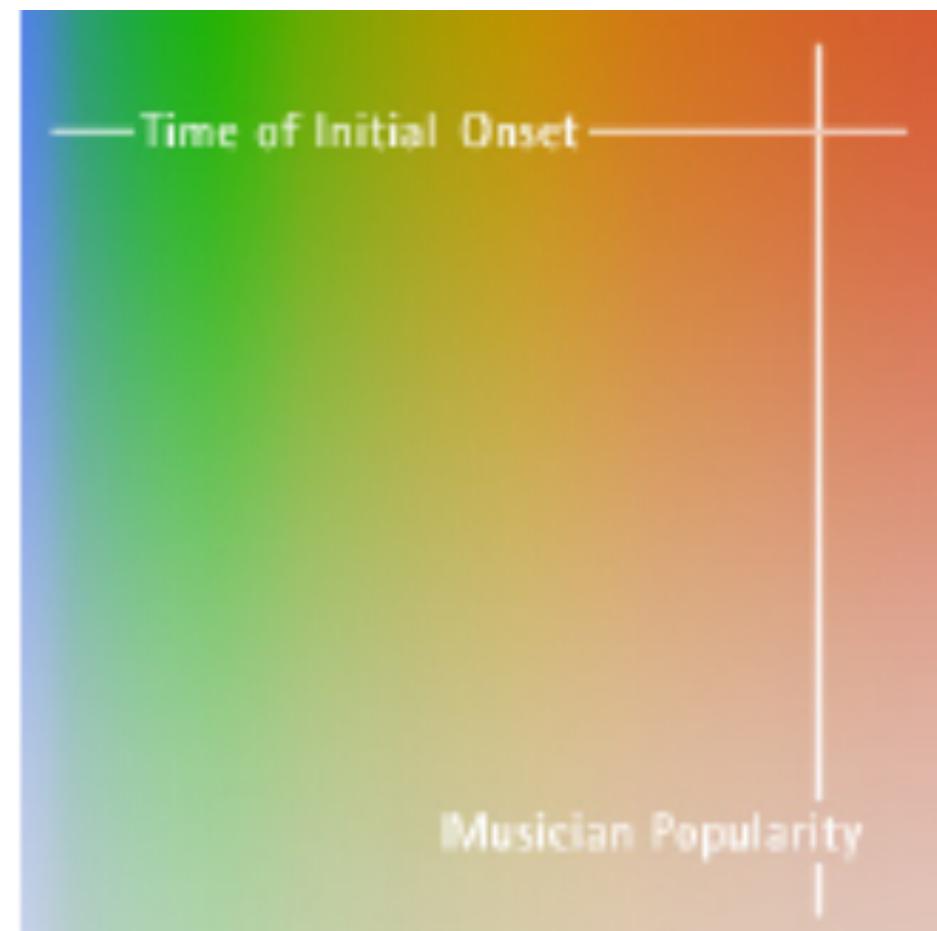


fig 9 – the 2d color palette used in Listening History

A variação horizontal varia de cores frias a quentes e reflete a data em que o artista foi descoberto pelo usuário

“**Núcleo frio**” de artistas conhecidos e as camadas “**quentes e novas**” das descobertas recentes

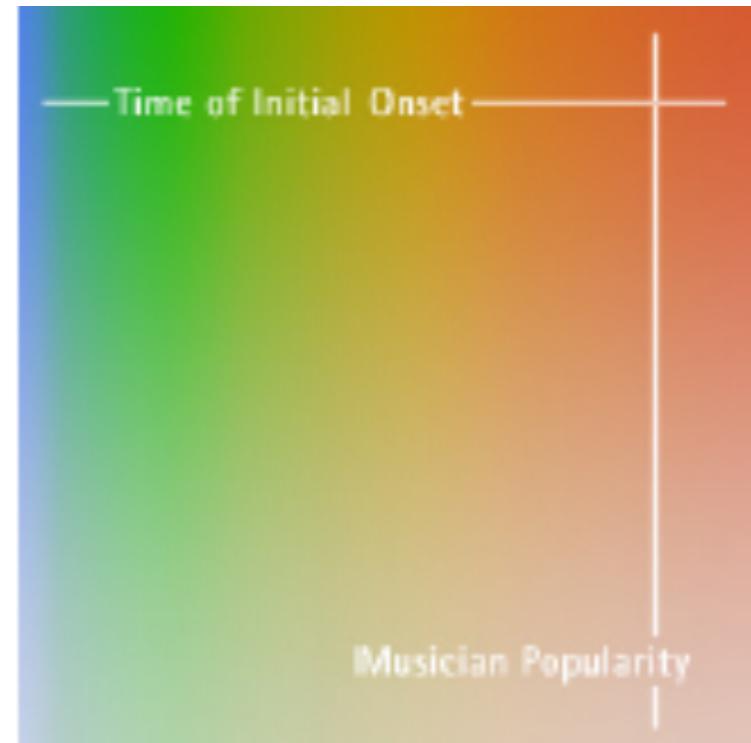
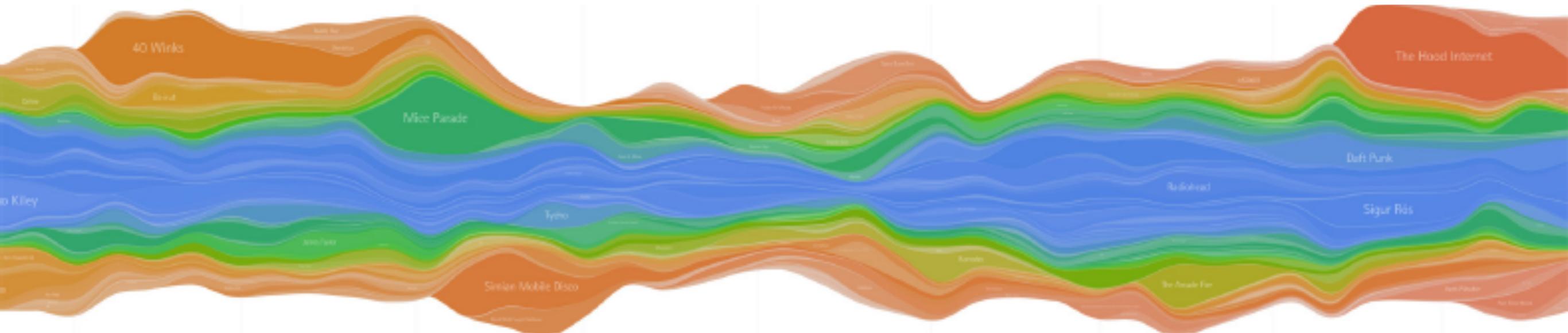


fig 9 – the 2d color palate used in Listening History

A variação vertical indica frequência

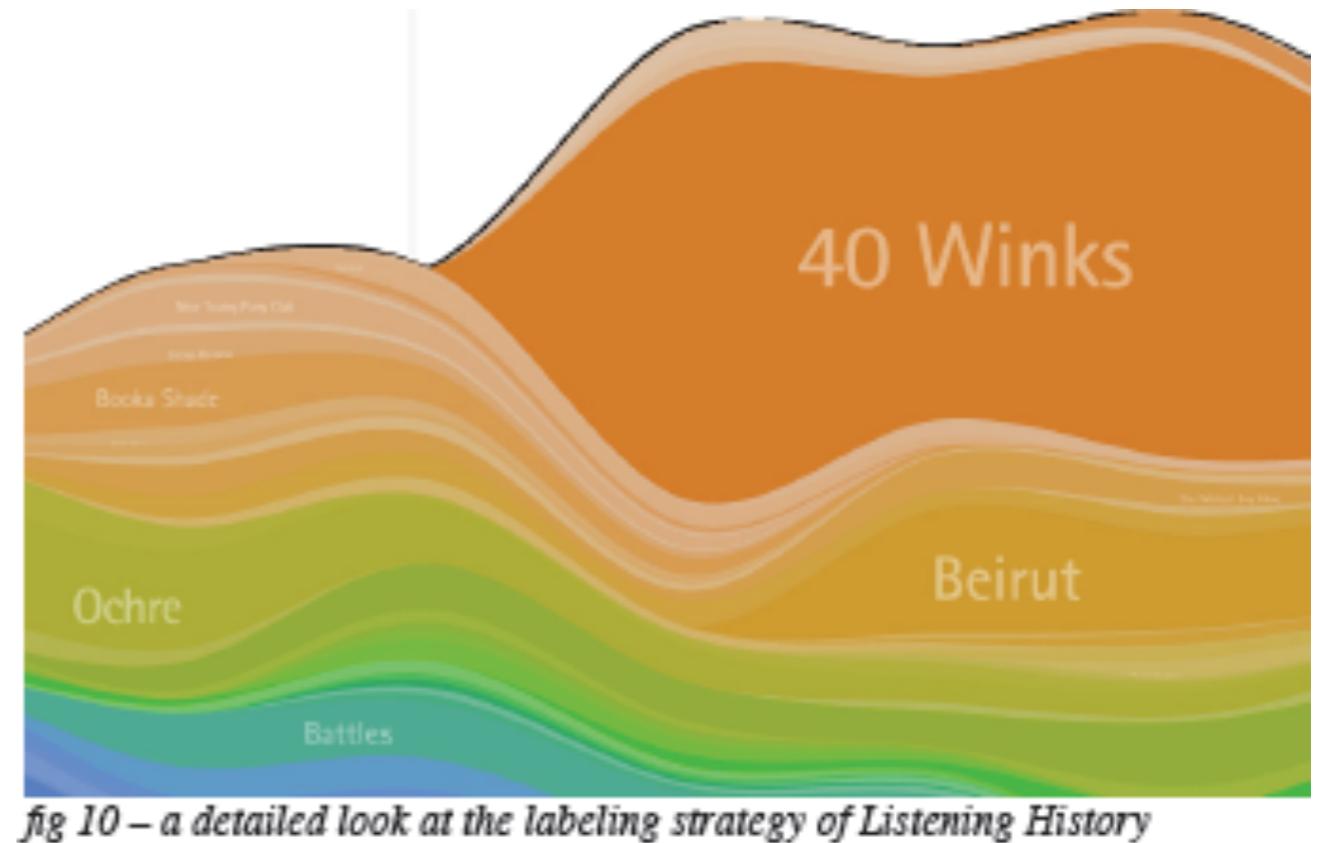


Muitas cores = dificuldade no uso de legendas

Escolha da posição da camada onde poderia ser usada a maior fonte possível usando algoritmo de força bruta (off-line)

Na versão on-line, rótulos nos eventos do mouse

Fonte de cor branca com transparência para conectar o rótulo a camada



EVOLUTIONARY COMPUTATION FOR LABEL LAYOUT ON UNUSED SPACE OF STACKED GRAPHS

*A. Toledo, K. Sookhanaphibarn, R. Thawonmas e F.
Rinando*

*ISRN Artificial Intelligence
2012*

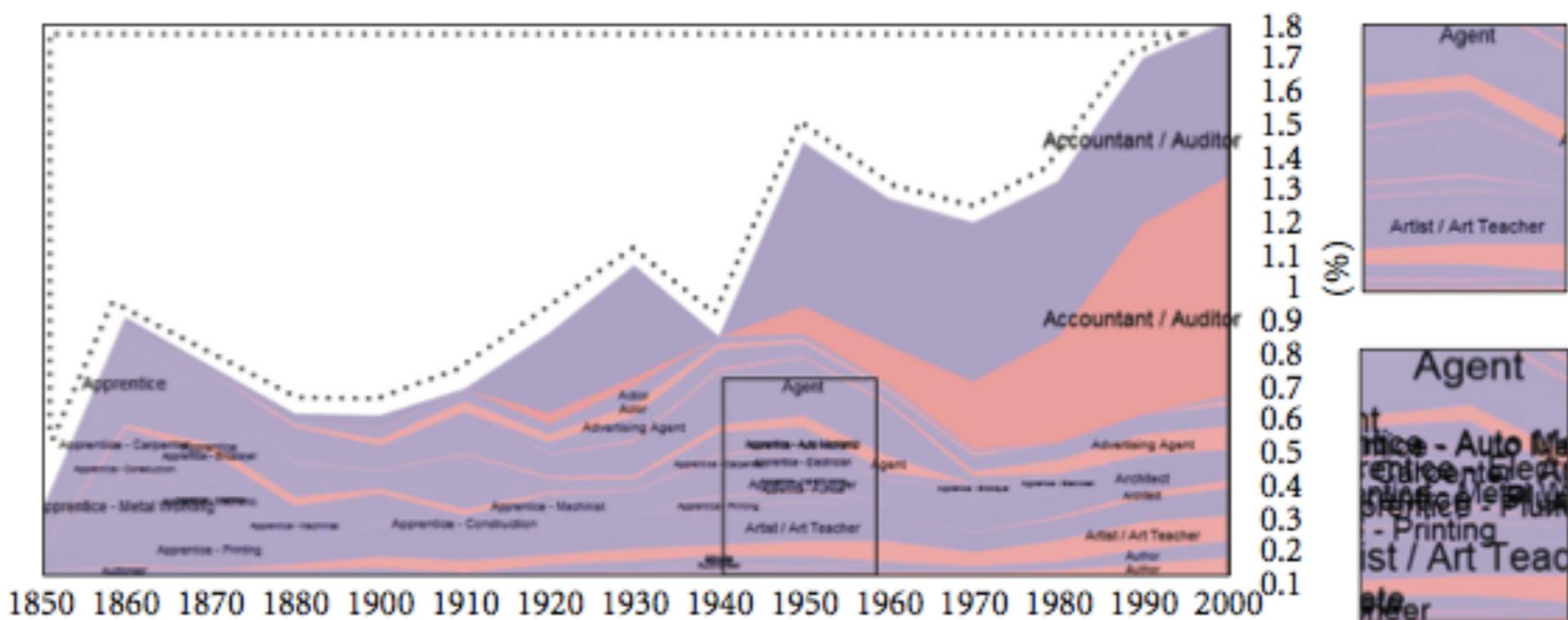


FIGURE 1: Stacked graph view (left) showing 40 job occupation names starting with the prefix “A”, and $TLV = 0$ pixels. A number of labels are illegible due to either overlapping or small size. The solid line encloses the visualization area, and the dotted line the unused space. At the right, an increased section of the same view (top) with $TLV = 12$ pixels (some labels are hidden). The same section (bottom) can be seen with $TLV = 0$ (some labels are overlapped) and with their minimum font size increased three times.

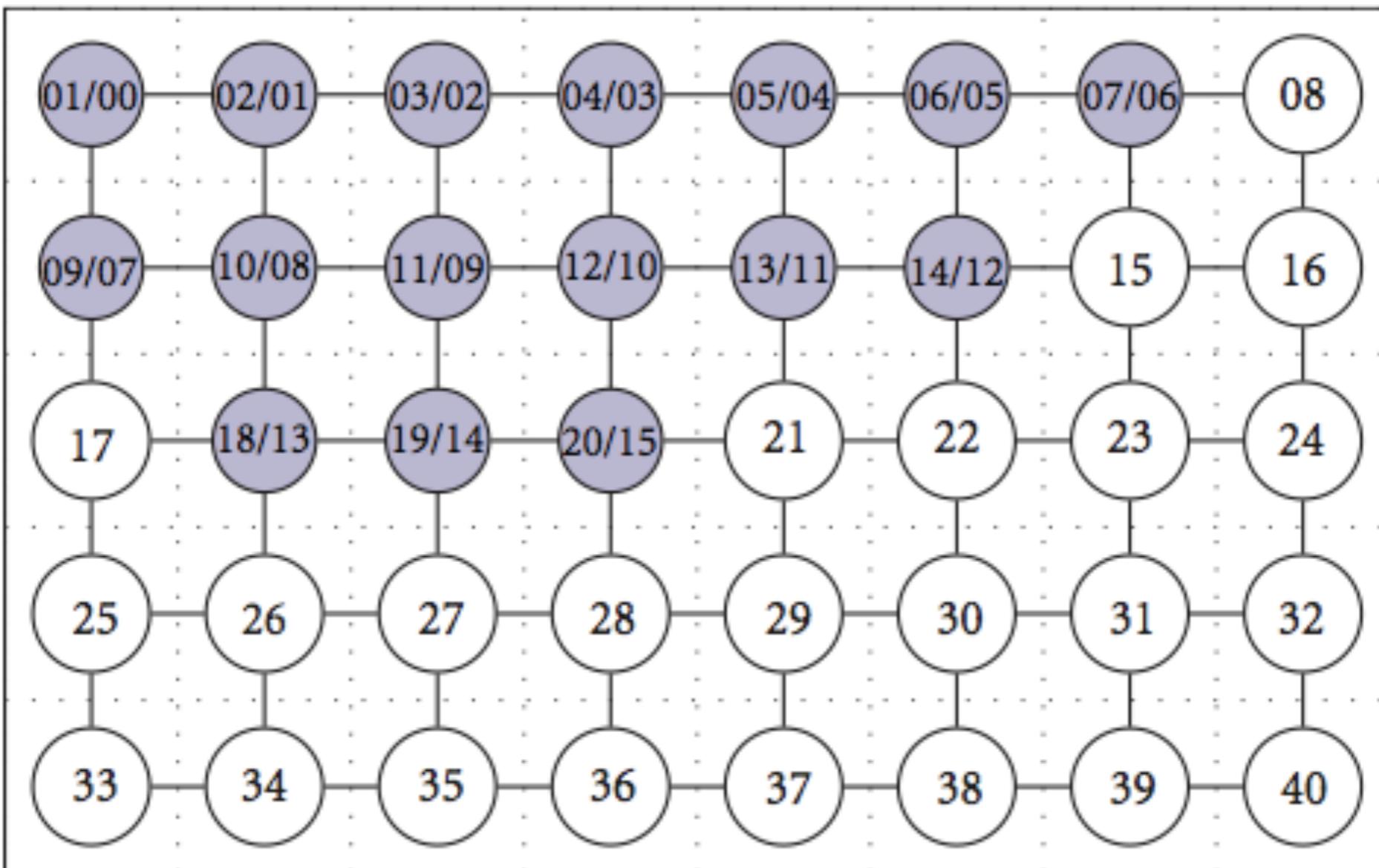


FIGURE 2: Partition of the visualization area of the stacked graph of Figure 1. Partitions in blue color correspond to the candidate partitions for label layout.

Chromosome A: 0000 0110 0010 1000 0111 1001

Chromosome B: 1001 0101 0110 0011 0001 0010

Ordenação das camadas

Em alguns casos a ordenação natural dos dados é a mais apropriada

Neste caso, a ordenação influencia na aparência do gráfico e pode ser escolhida com objetivos estéticos



fig 12 – an unsorted data set, exhibiting the type of “burstiness” apparent in last.fm and box office data sets

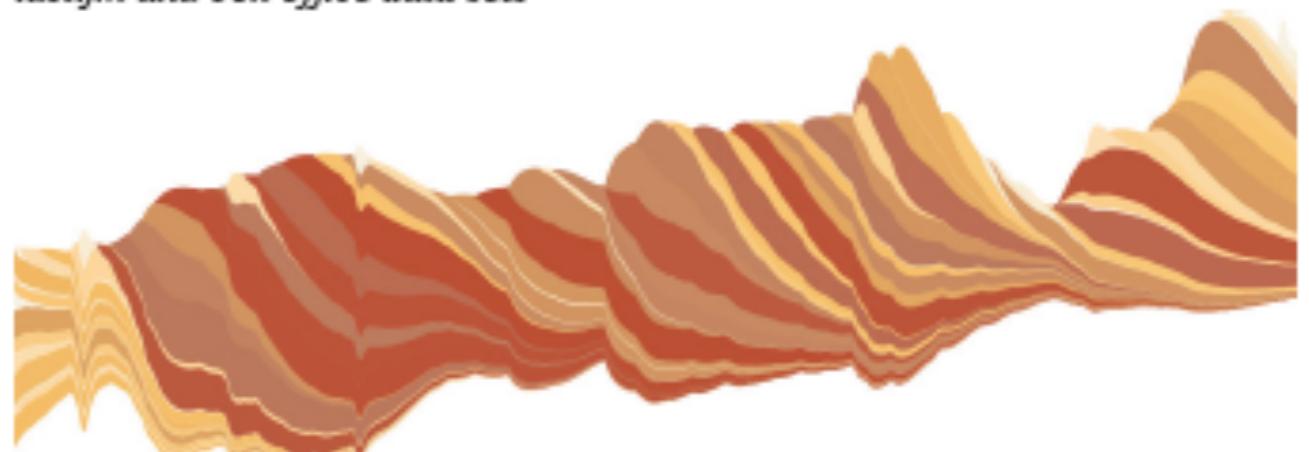


fig 13 – the same data set, naively sorted in order of “onset time” exhibiting the distracting diagonal striping effect



fig 14 – the same data set sorted using the weighted “inside out” strategy to highlight the initial onset of each time series

Ordenação das camadas

Camadas **desordenadas**



fig 12 – an unsorted data set, exhibiting the type of “burstiness” apparent in last.fm and box office data sets

Camadas ordenadas
por ***onset time***

(descoberta do artista)

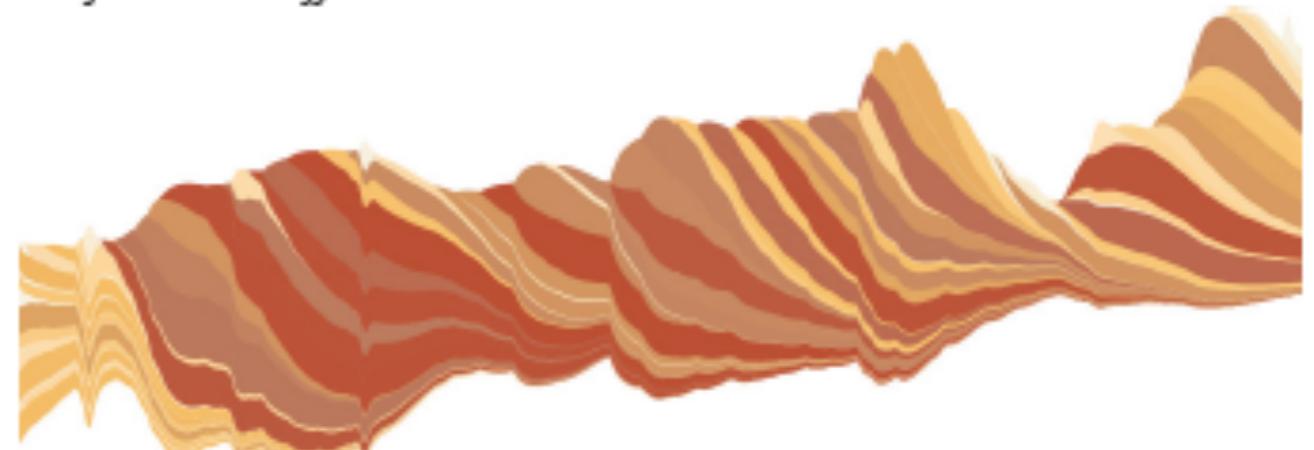


fig 13 – the same data set, naively sorted in order of “onset time” exhibiting the distracting diagonal striping effect

Camadas ordenadas
por ***inside-out***

(onset antigo no meio e onset recente
em baixo e em cima)

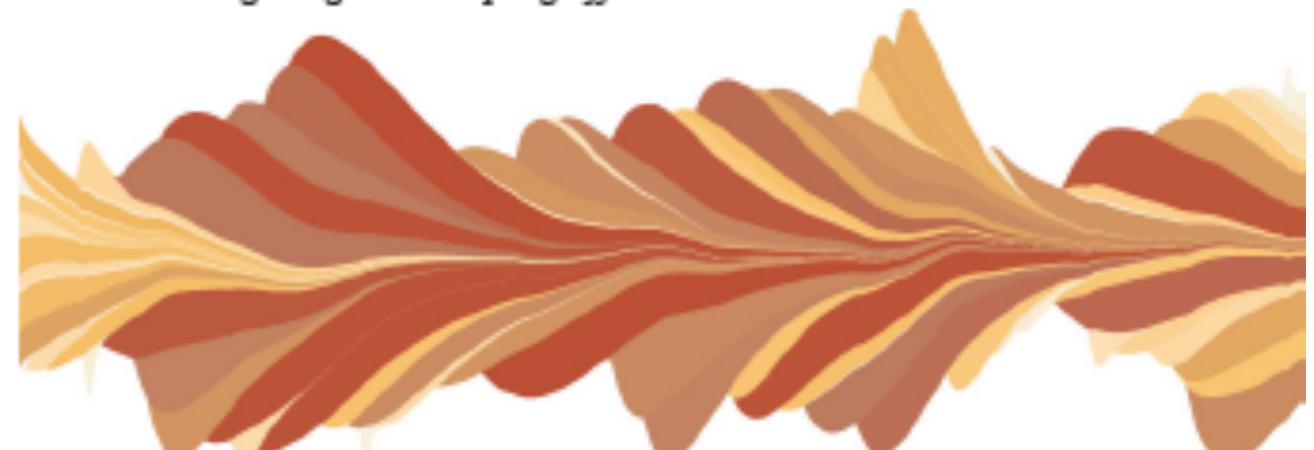


fig 14 – the same data set sorted using the weighted “inside out” strategy to highlight the initial onset of each time series

Ordenação das camadas

Camadas ordenadas
por *inside-out*

Solução:

1. Ordenar as camadas por *onset time*
2. Para cada camada, calcular a soma dos valores da série (peso)
3. Insere sequencialmente na lista testando se a soma do peso da primeira metade da lista é maior que a metade do peso total, insere no fim. Caso contrário insere no começo.

Ordenação das camadas

Outra possibilidade é através de uma métrica de **volatilidade**

Séries com menos modificações no centro e séries com mais modificações no exterior

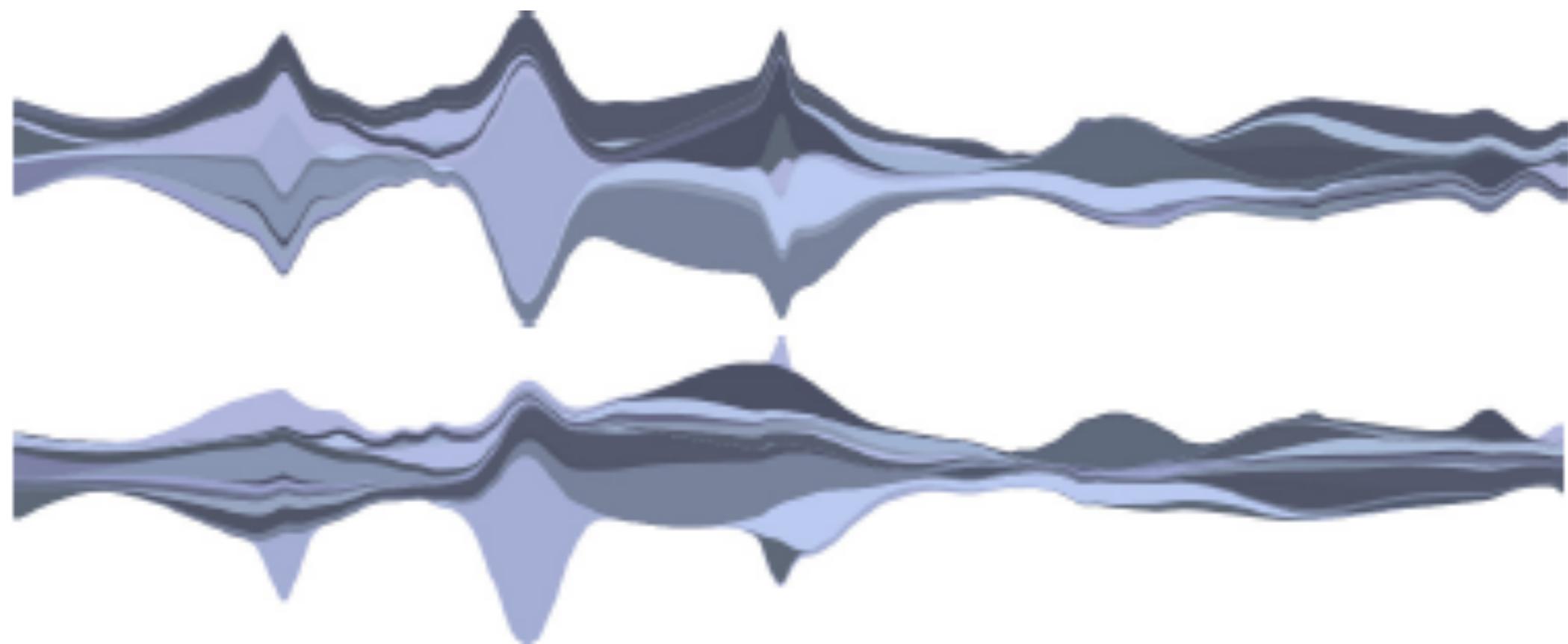


fig 15 – before and after applying an “inside out” sort using a “measure of volatility” in place of “onset time”