Understanding Document Collections

Michael DiBernardo
November 7th, 2005

Purpose

- General description of the field
- Demonstrate challenges and techniques for:
  - A problem you’re familiar with
  - A problem you’re unfamiliar with

Purpose

- General description of the field
- Demonstrate challenges and techniques for:
  - A problem you’re familiar with
  - A problem you’re unfamiliar with

Some Tasks:

- Given a set of documents:
  - Find the subset that interests you
  - Categorize the set by composition
  - Compare the ‘utility’ of some subset
  - Quickly understand a single document

Purpose

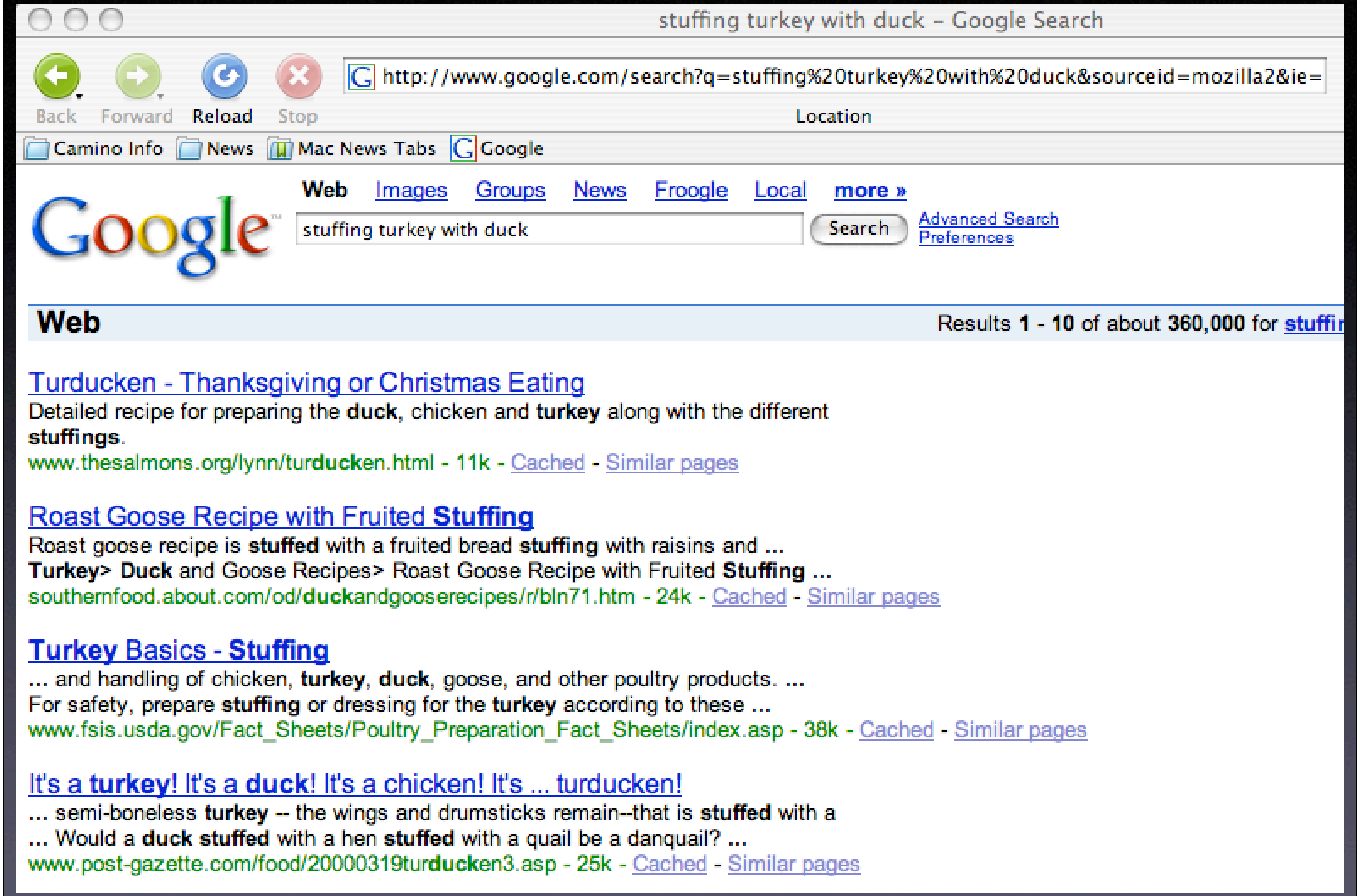
- General description of the field
- Demonstrate challenges and techniques for:
  - A problem you’re familiar with
  - A problem you’re unfamiliar with

Purpose

- General description of the field
- Demonstrate challenges and techniques for:
  - A problem you're familiar with
  - A problem you're unfamiliar with

Scenario

- It’s almost Christmas
  - We want some turkey
  - We want some duck
  - Maybe a duck... inside a turkey?


stuffing turkey with duck – Google Search
Back
Forward
Reload
Stop
http://www.google.com/search?q=stuffing%20turkey%20with%20duck&sourceid=mozilla2&ie=
Location
Camino Info
News
Mac News Tabs
Google
Web
Images
Groups
News
Froogle
Local
more »
stuffing turkey with duck
Search
Advanced Search
Preferences

Web Results **1** - **10** of about **360,000** for **stuffi**

Turducken - Thanksgiving or Christmas Eating
Detailed recipe for preparing the **duck**, chicken and **turkey** along with the different **stuffings**.
www.thesalmons.org/lynn/tur**duck**en.html - 11k - Cached - Similar pages

Roast Goose Recipe with Fruited **Stuffing**
Roast goose recipe is **stuffed** with a fruited bread **stuffing** with raisins and **...**
Turkey> **Duck** and Goose Recipes> Roast Goose Recipe with Fruited **Stuffing** **...**
southernfood.about.com/od/**duck**andgooserecipes/r/bln71.htm - 24k - Cached - Similar pages

Turkey Basics - **Stuffing**
... and handling of chicken, **turkey**, **duck**, goose, and other poultry products. **...**
For safety, prepare **stuffing** or dressing for the **turkey** according to these **...**
www.fsis.usda.gov/Fact_Sheets/Poultry_Preparation_Fact_Sheets/index.asp - 38k - Cached - Similar pages

It's a **turkey**! It's a **duck**! It's a chicken! It's ... turducken!
... semi-boneless **turkey** -- the wings and drumsticks remain--that is **stuffed** with a
... Would a **duck stuffed** with a hen **stuffed** with a quail be a danquail? **...**
www.post-gazette.com/food/20000319tur**duck**en3.asp - 25k - Cached - Similar pages

Observation:

- It’s hard to tell which documents:
  - Are talking about turkey and duck
  - Are talking about a turducken
  - Contain recipes or histories

The Big Problem:

- Have to scan each document to see if it’s appropriate
- Ranking procedure is *opaque*

Hearst’s Idea

- When searching fulltext, user should know:
  - Relative lengths of documents
  - Frequency of query terms
  - *Distribution of terms* in the document

Proposed Solution

- CHI ‘96 video

Method

- Split document into chunks (topical, paragraph)
- Each chunk represented by a TextTile
  - Darkness = frequency

Encodings

- Relative length of document = bar length
- Term frequency = TextTile darkness
- Term distribution = TextTile position

Strengths

- Simple and intuitive to use
- Based on accepted vis principles
- Can be integrated with ranking
- Documents can be ‘pre-chunked’

Weaknesses

- Long documents cause problems
- No discussion of how to chunk document
- No user study
- TextTiles give no indication of passage length
- Colour?

Applications

- Was incorporated in Berkeley’s e-library
    - Clunky applet solution
    - Addressed length problem

Applications

- Good candidate to supplement web search
- Would be interesting to see reactions

Before it can be used:

- Need to deal with length problem
- Decide how to position in search interface

Summary

- TileBars good summary of document structure and term frequency distribution
- Supplements ranked search
- Lightweight solution to a common problem

Purpose

- General description of the field
- Demonstrate challenges and techniques for:
  - A problem you're familiar with
  - A problem you're unfamiliar with

The Problem

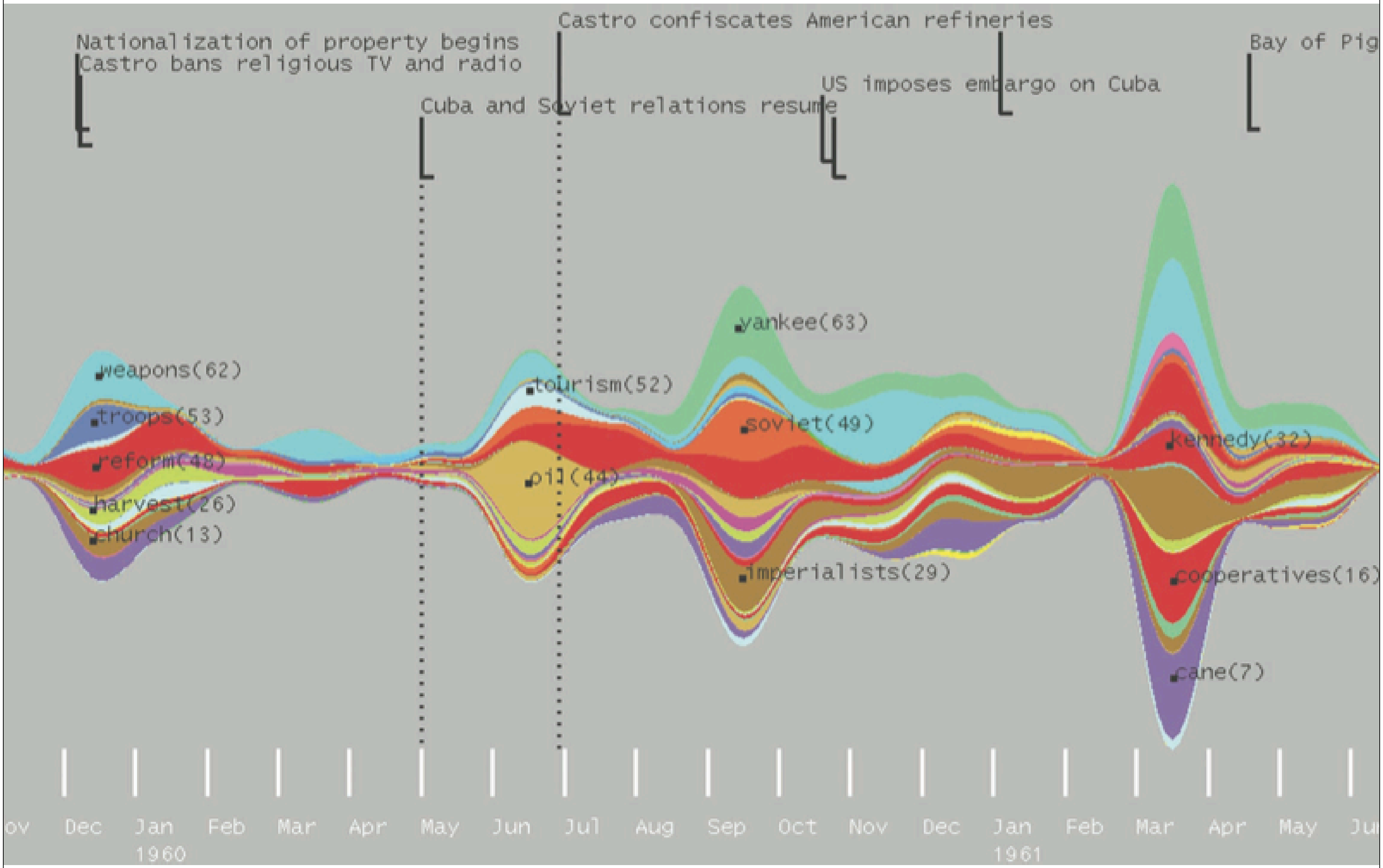
- Understanding how the thematic composition of document collection *changes over time*

Observation

- People tend to abstract time as *relative motion*
- Consider pithy sayings:
  - “Don’t let your life pass you by...”
  - “Your time will come...”

Idea

- Use visual metaphor of a river

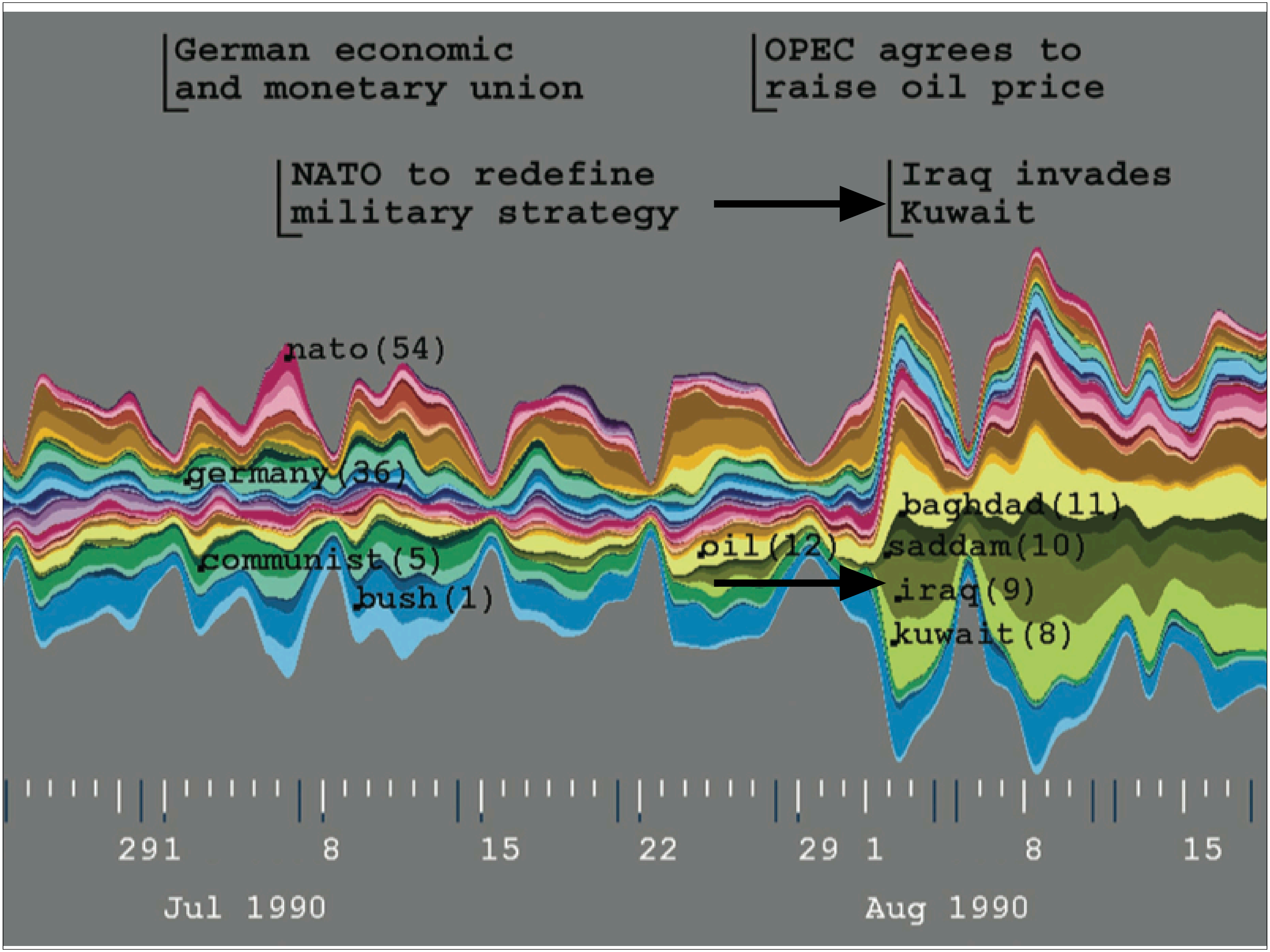

Castro confiscates American refineries
Nationalization of property begins
Castro bans religious TV and radio
Bay of Pig
US imposes embargo on Cuba
Cuba and Soviet relations resume
yankee(63)
weapons(62)
tourism(52)
troops(53)
soviet(49)
kennedy(32)
reform(48)
oil(44)
harvest(26)
church(13)
imperialists(29)
cooperatives(16)
cane(7)
ov
Dec
Jan
1960
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
1961
Feb
Mar
Apr
May
Ju

Encodings

- Time = horizontal axis
- Thematic strength = height of stream
- Salient events plotted with time scale

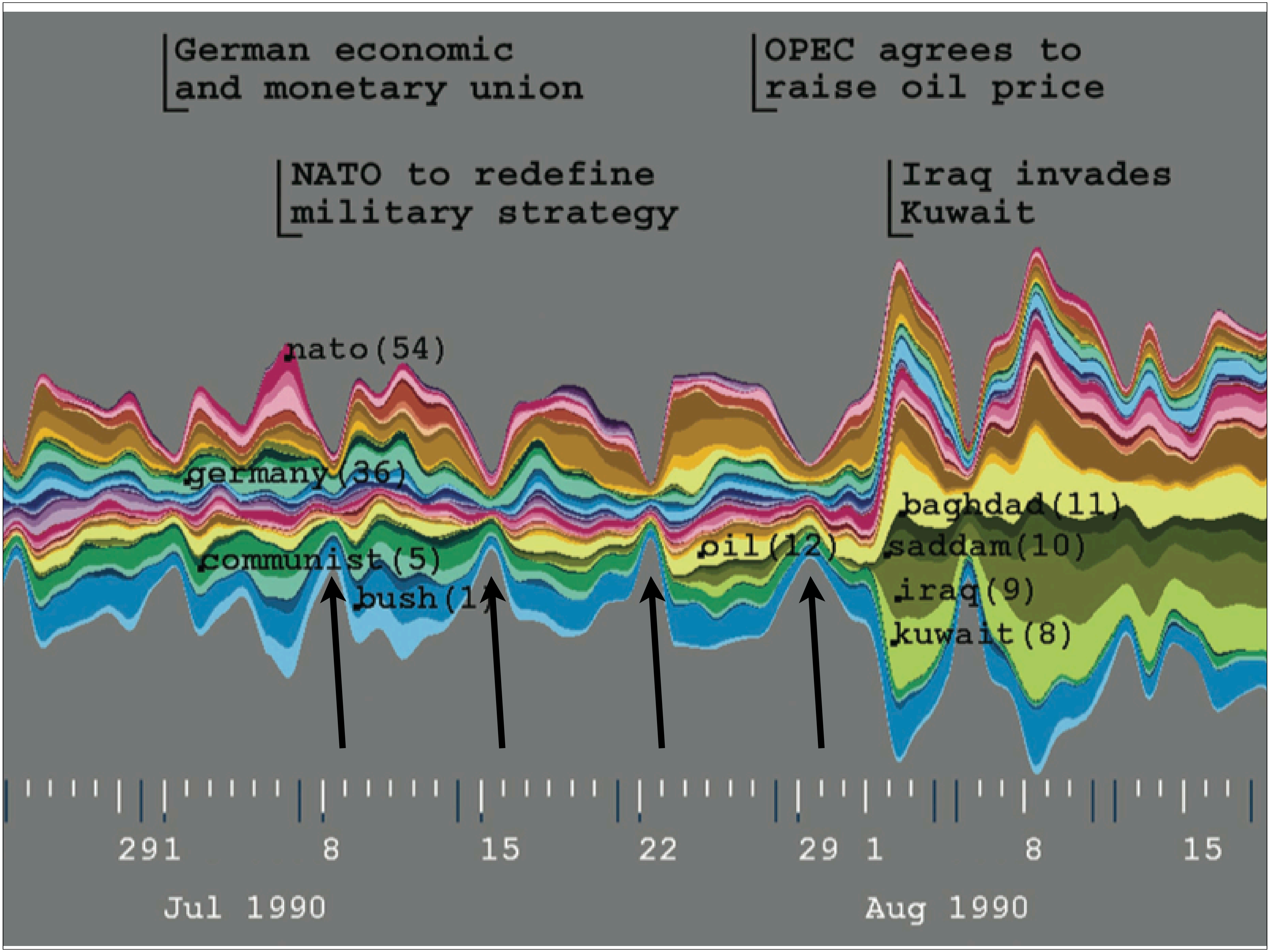
Example

- We have 100,000 Associated Press articles from 1990
- How did media react to Iraqi invasion of Kuwait?


German economic
and monetary union
OPEC agrees to
raise oil price
NATO to redefine
military strategy
Iraq invades
Kuwait
nato(54)
germany(36)
communist(5)
bush(1)
oil(12)
baghdad(11)
saddam(10)
iraq(9)
kuwait(8)
29
1
8
15
22
29
1
8
15
Jul 1990
Aug 1990

Sample analysis

- What are those periodic ‘chokepoints’?


German economic and monetary union
OPEC agrees to raise oil price
NATO to redefine military strategy
Iraq invades Kuwait
nato(54)
germany(36)
communist(5)
bush(1)
oil(12)
baghdad(11)
saddam(10)
iraq(9)
kuwait(8)
29
1
8
15
22
29
1
8
15
Jul 1990
Aug 1990

Sample analysis

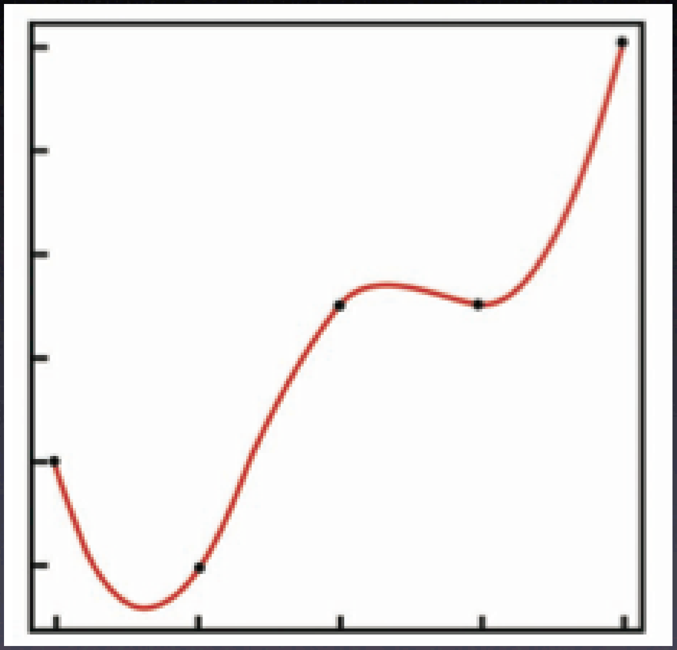
- On Sundays:
  - Is overall volume reduced?
  - Or are our themes less popular?

Plotting streams

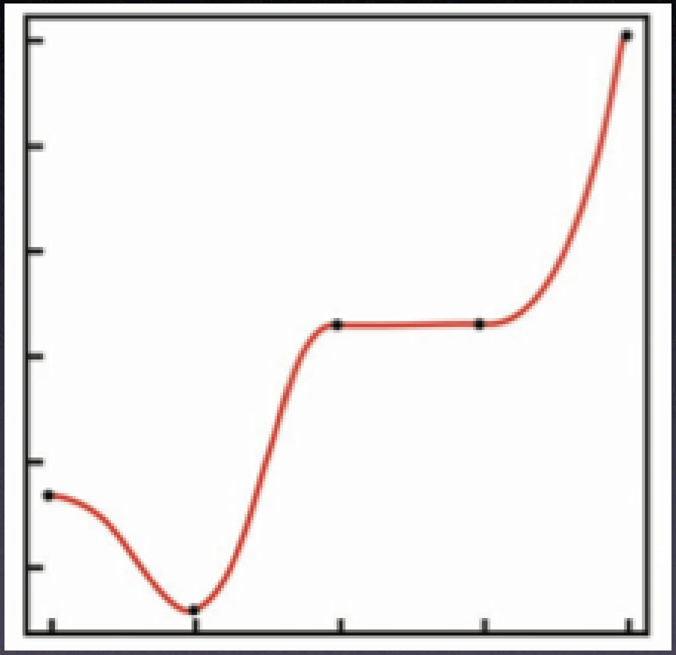
- For a theme *T*:
  - Articles of *T* are sorted into bins by time
  - For each bin, plot the *number* of articles

Plotting streams

- To keep curve smooth:
  - Interpolate between points
- At t' between bin t and $t+1$:
  - $y(t) <= y(t') <= y(t+1)$
  - Or vice-versa if decreasing



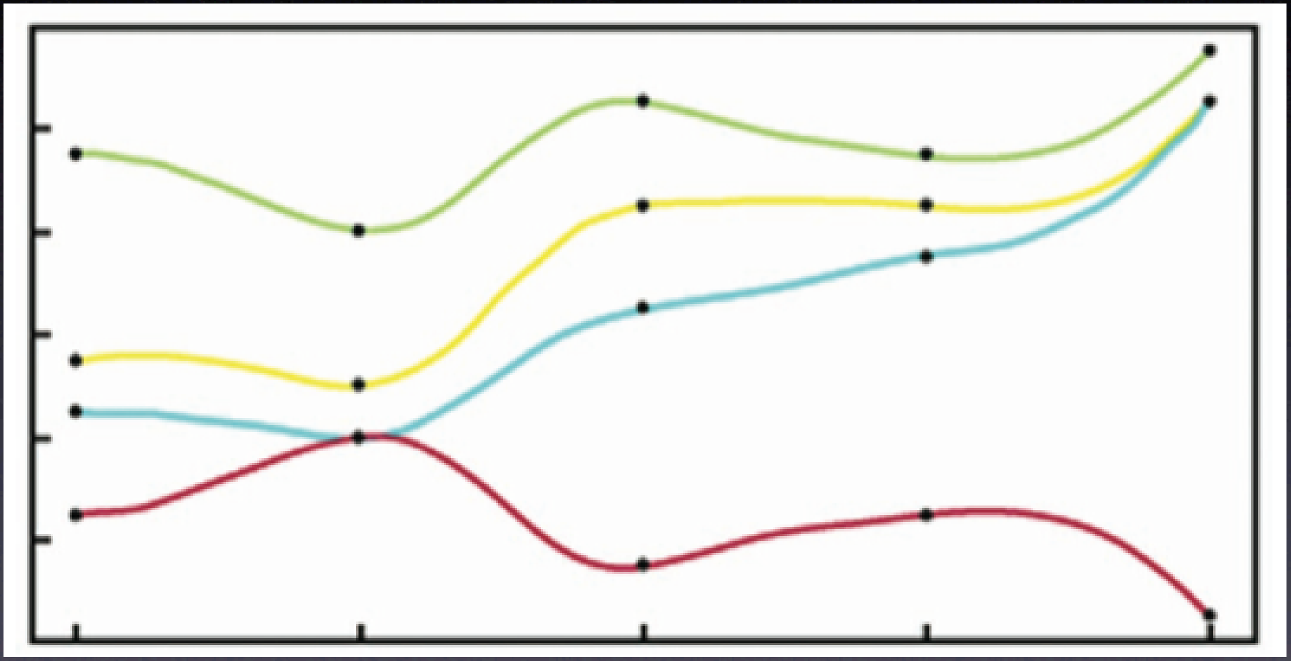
Bad!



Good!

Plotting streams

- Streams are then ‘stacked’ to create river

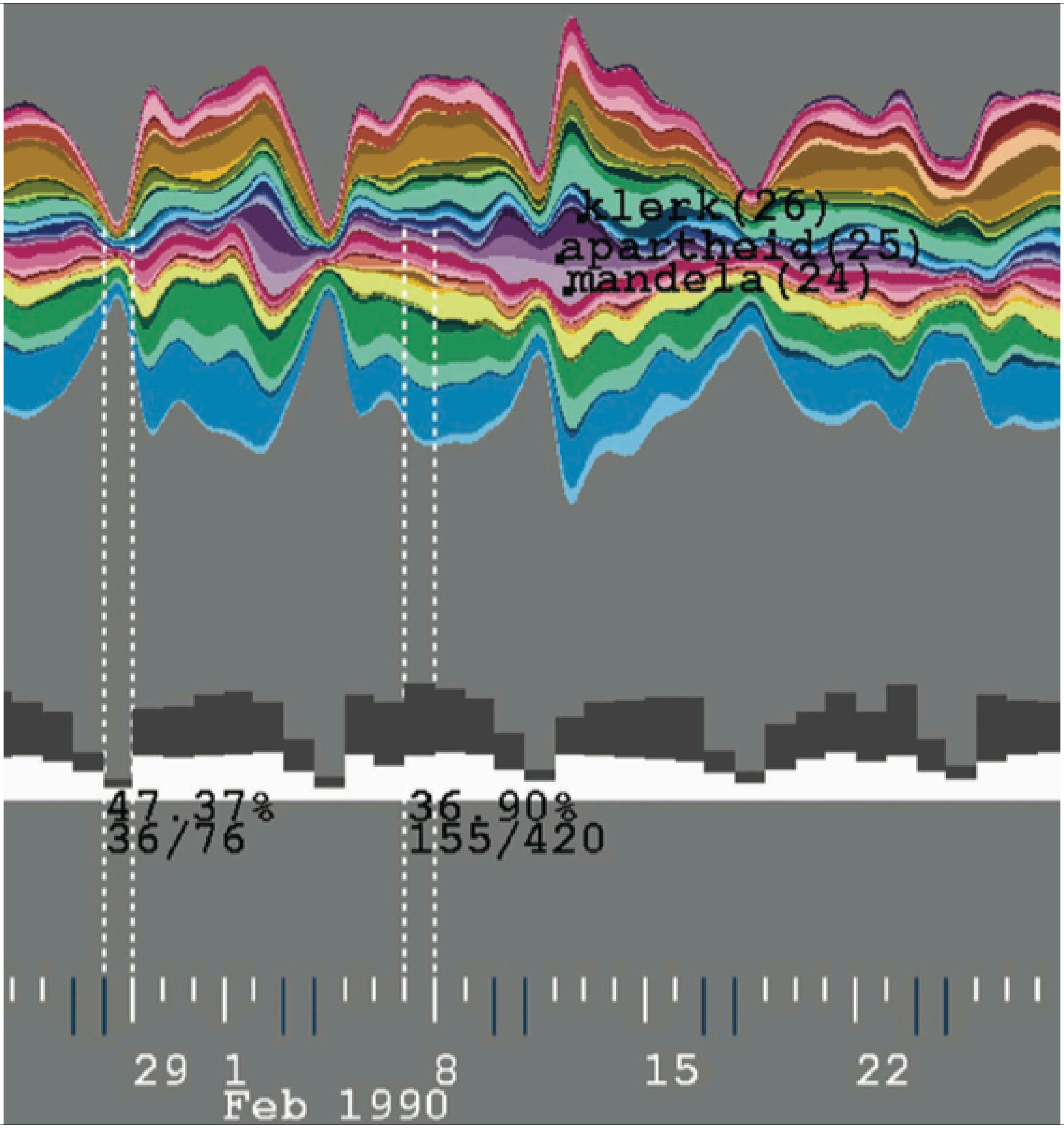


So what about Sundays?

- We don’t know how many articles were published *in total* on a given day
- So we can’t tell if
  - our themes were unpopular, or
  - Sundays have reduced distribution

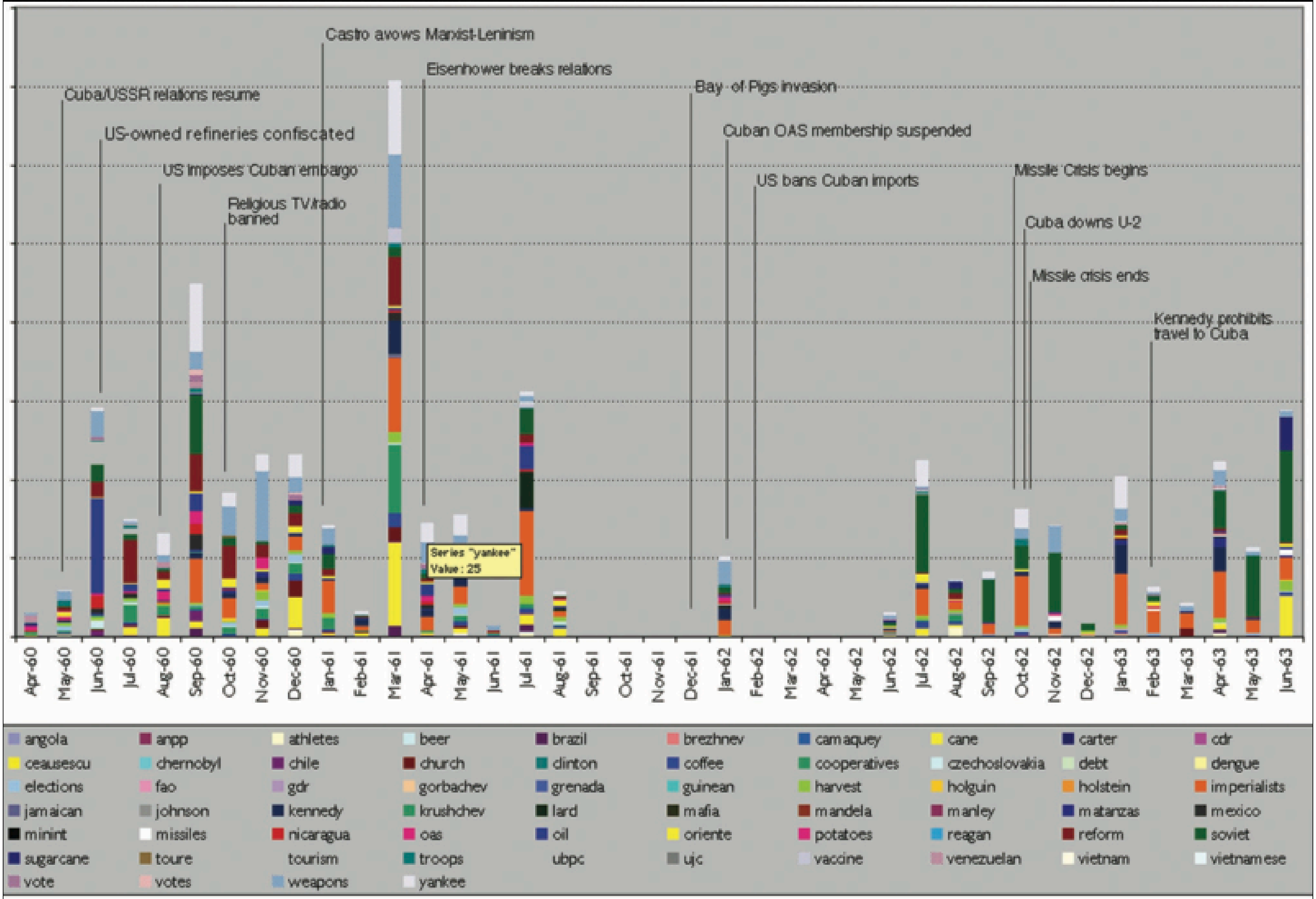
Need more information

- Forced to plot histograms


klerk(26)
apartheid(25)
mandela(24)
47.37%
36/76
36.90%
155/420
29
1
8
15
22
Feb 1990

Other problems?

- User study done to identify other problems
  - Used collection of Castro’s speeches
  - Compared to a plain histogram (!)
  - Two users: No mention of competency


Castro avows Marxist-Leninism
Eisenhower breaks relations
Bay of Pigs invasion
Cuba/USSR relations resume
US-owned refineries confiscated
Cuban OAS membership suspended
US imposes Cuban embargo
US bans Cuban imports
Missile Crisis begins
Religious TV/radio banned
Cuba downs U-2
Missile crisis ends
Kennedy prohibits travel to Cuba
Series "yankee"
Value : 25
Apr-60
May-60
Jun-60
Jul-60
Aug-60
Sep-60
Oct-60
Nov-60
Dec-60
Jan-61
Feb-61
Mar-61
Apr-61
May-61
Jun-61
Jul-61
Aug-61
Sep-61
Oct-61
Nov-61
Dec-61
Jan-62
Feb-62
Mar-62
Apr-62
May-62
Jun-62
Jul-62
Aug-62
Sep-62
Oct-62
Nov-62
Dec-62
Jan-63
Feb-63
Mar-63
Apr-63
May-63
Jun-63
angola
anpp
athletes
beer
brazil
brezhnev
camaquey
cane
carter
cdr
ceausescu
chernobyl
chile
church
clinton
coffee
cooperatives
czechoslovakia
debt
dengue
elections
fao
gdr
gorbachev
grenada
guinean
harvest
holguin
holstein
imperialists
jamaican
johnson
kennedy
krushchev
lard
mafia
mandela
manley
matanzas
mexico
minint
missiles
nicaragua
oas
oil
oriente
potatoes
reagan
reform
soviet
sugarcane
toure
tourism
troops
ubpc
ujc
vaccine
venezuelan
vietnam
vietnamese
vote
votes
weapons
yankee

User comments

+ Users comfortable with river metaphor
+ Smooth curves aid navigation
+ Easily identified macro trends
+ Became inquisitive about theme changes

User comments

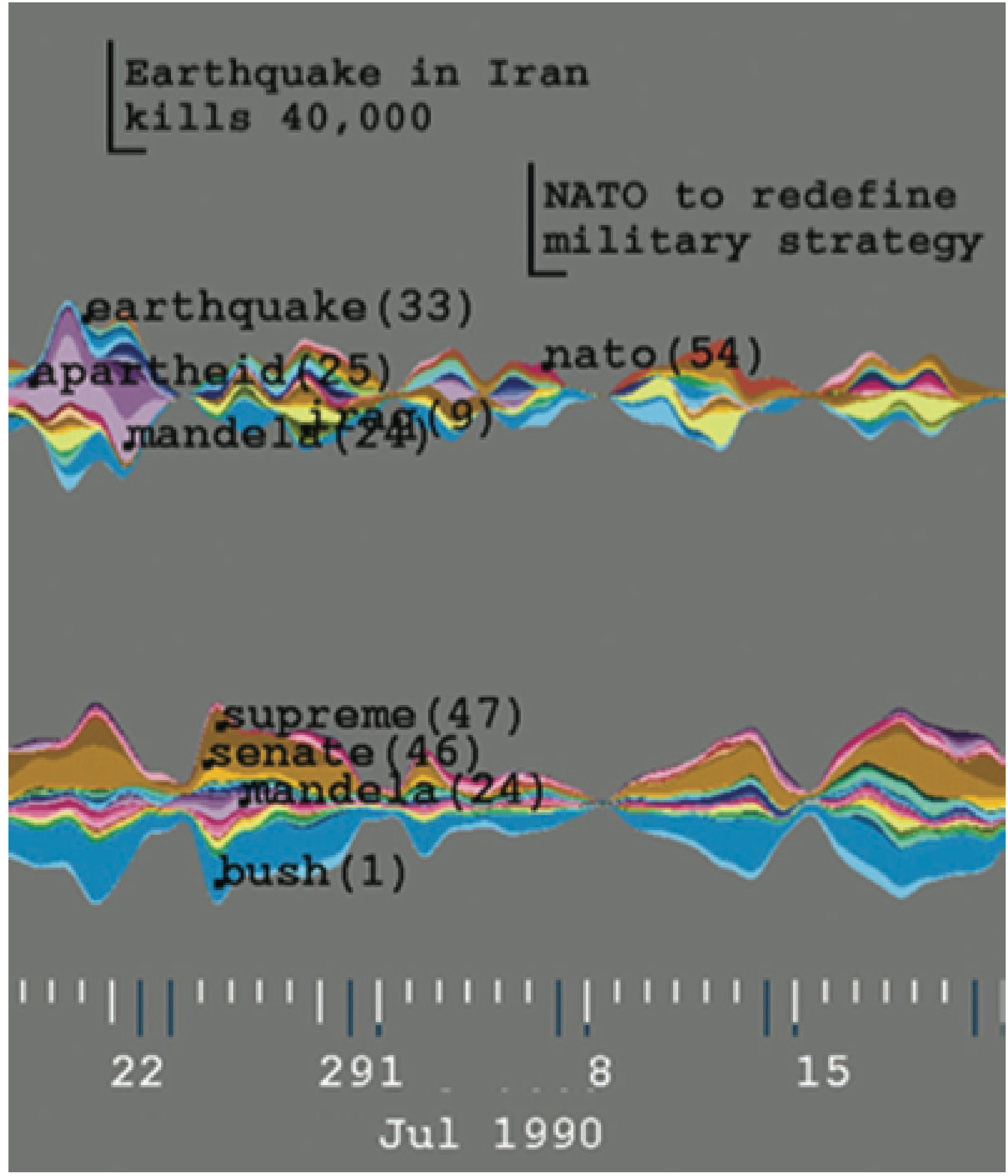
- Hard to compare micro changes
- Interpolation breeds mistrust
- Theme ordering
- Want access to more detail

User study

- User study highlighted some key strengths and weaknesses of approach

Strengths

- Comparing between small set of collections is fairly easy (without histograms)

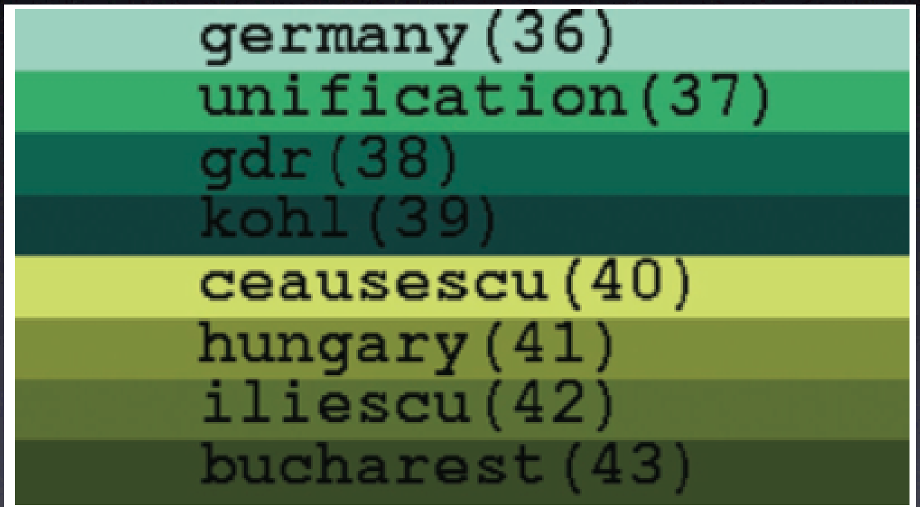

Earthquake in Iran
kills 40,000
NATO to redefine
military strategy
earthquake(33)
apartheid(25)
iraq(9)
mandela(24)
nato(54)
supreme(47)
senate(46)
mandela(24)
bush(1)
22
291
8
15
Jul 1990

Strengths

- Can easily relate events to macro trends (when close to related event!)
- Smooth curves easy to follow
- Performed (minimal) user study, and very forthcoming about weaknesses

Weaknesses

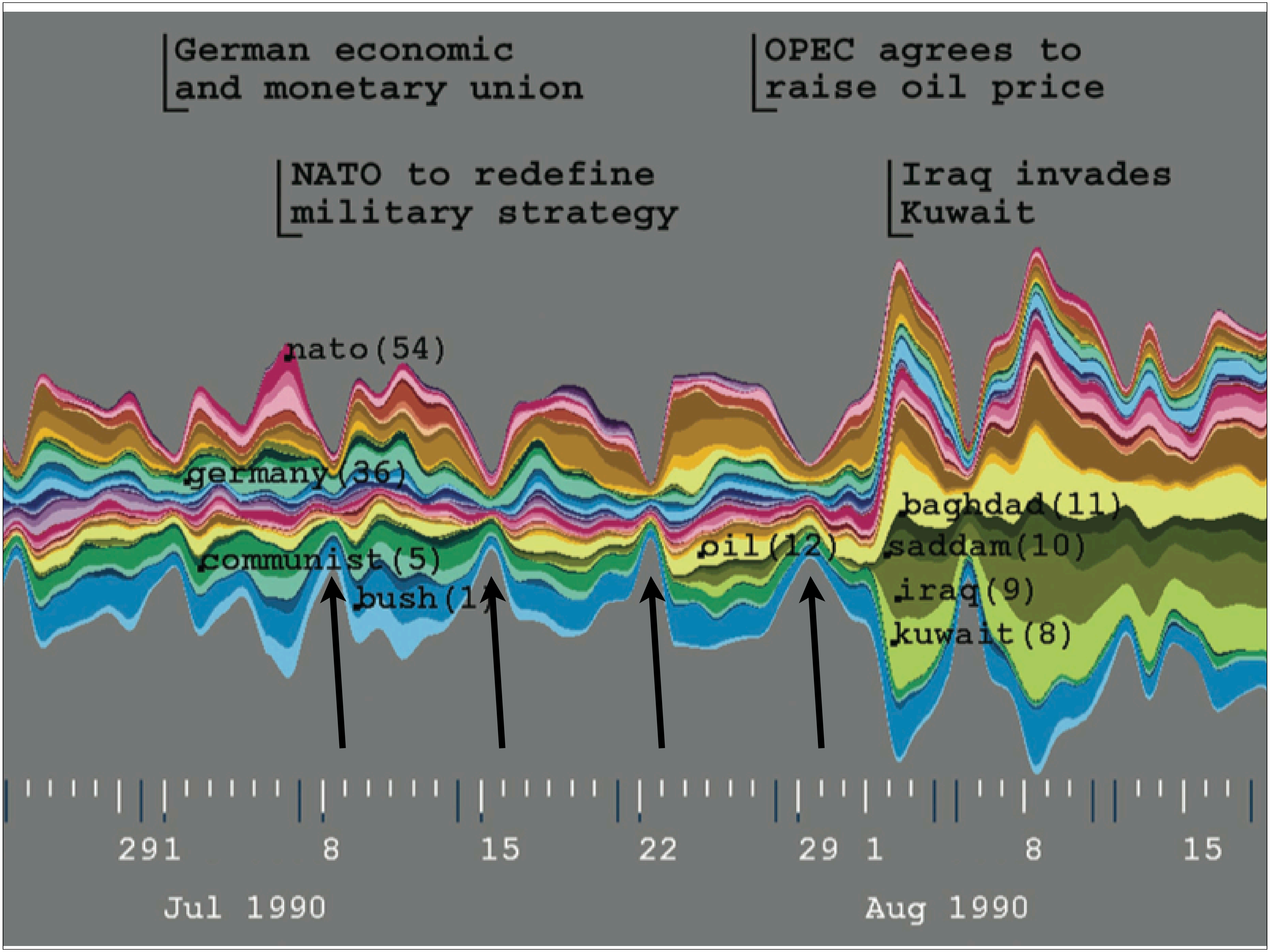
- Colour selection
  - Differences in colour groups imply an ordering



germany(36)
unification(37)
gdr(38)
kohl(39)
ceausescu(40)
hungary(41)
iliescu(42)
bucharest(43)

Weaknesses

- Theme labels are hard to resolve


German economic and monetary union
OPEC agrees to raise oil price
NATO to redefine military strategy
Iraq invades Kuwait
nato(54)
germany(36)
communist(5)
bush(1)
oil(12)
baghdad(11)
saddam(10)
iraq(9)
kuwait(8)
29
1
8
15
22
29
1
8
15
Jul 1990
Aug 1990

Weaknesses

- Theme ordering matters
  - In how river looks (stacking)
  - In ease of theme comparison
- No discussion of theme selection
  - How to select good theme set?

Weaknesses

- Micro trends difficult to perceive
- Hard to select good “thematic strength”
  - The “Sunday problem”

Weaknesses

- How to enable details-on-demand?
  - Picking specific documents
  - Seeing what types of documents compose streams

Applications

- Patterns of media coverage
- Discussion flows in meetings (minutes)
- Changes in patient records
- Juvenile criminal records
- Politics
- Assessing tech trends through patents

Summary

- Wanted to view thematic change over time
- Use river as visual metaphor
- Effective at identifying macro trends
- Need to address colour selection, ambiguities in thematic strength, enabling details on demand

In Closing

- Understanding Document Collections:
  - Search
  - Thematic composition
  - Document structure comparison
  - Document browsing

In Closing

- We saw two domain problems:
  - Understanding search results on full-length documents
  - Understanding thematic change over time

In Closing

- We saw two (attempted) solutions:
    - TileBars - succinct representation of document length and structure
    - ThemeRiver - visual metaphor to describe how themes change over time