



# Efficient Implementation of Large-Scale Multi-Structural Databases

**Ravi Kumar**

**Yahoo! Research**

**[ravikumar@yahoo-inc.com](mailto:ravikumar@yahoo-inc.com)**



## Joint work with

---

- **Ronald Fagin, IBM Almaden**
- **Phokion Kolaitis, IBM Almaden**
- **Jasmine Novak, Yahoo! Research**
- **D. Sivakumar, Google**
- **Andrew Tomkins, Yahoo! Research**

**Work done at IBM Almaden Research Center**



# Outline of the talk

---

- **Multi-structural databases (MSDB)**
- **Queries**
- **A new algorithm**
- **Conclusions**



## A motivating example

---

Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

There seem to be a lot of documents talking about politics – do they come predominantly from any particular geography and/or time?

Ans: Eg, documents about politics are much more likely to come from Europe, published in 2005, or from Korea, published in June of 2004



## A motivating example

---

Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

**Among documents that mention finance what are the topics that are strongly correlated with a particular geography?**

**Ans: Eg, globalization is strongly correlated with California, India, and China; and currency is strongly correlated with Europe**



## A motivating example

---

Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

**What are the three combinations of geography and media type that have grown most significantly over the last year?**

**Ans: Eg, Japanese press releases, Asian newspapers, and Iranian blogs**



## Some observations

---

### Japanese press releases, Asian newspapers, and Iranian blogs

- **Consist of combinations of geography and media type**
  - **Express “clusters” in easy-to-understand terms**
- **Select different levels of the geographic hierarchy in the same result set**
  - **May select one or more appropriate granularities for the answer**
- **Japan is part of Asia, but fortunately newspapers and press releases are disjoint**
  - **Will not return overlapping regions of the multi-dimensional space**



A framework ...

---

## Multi-Structural Databases (MSDB)



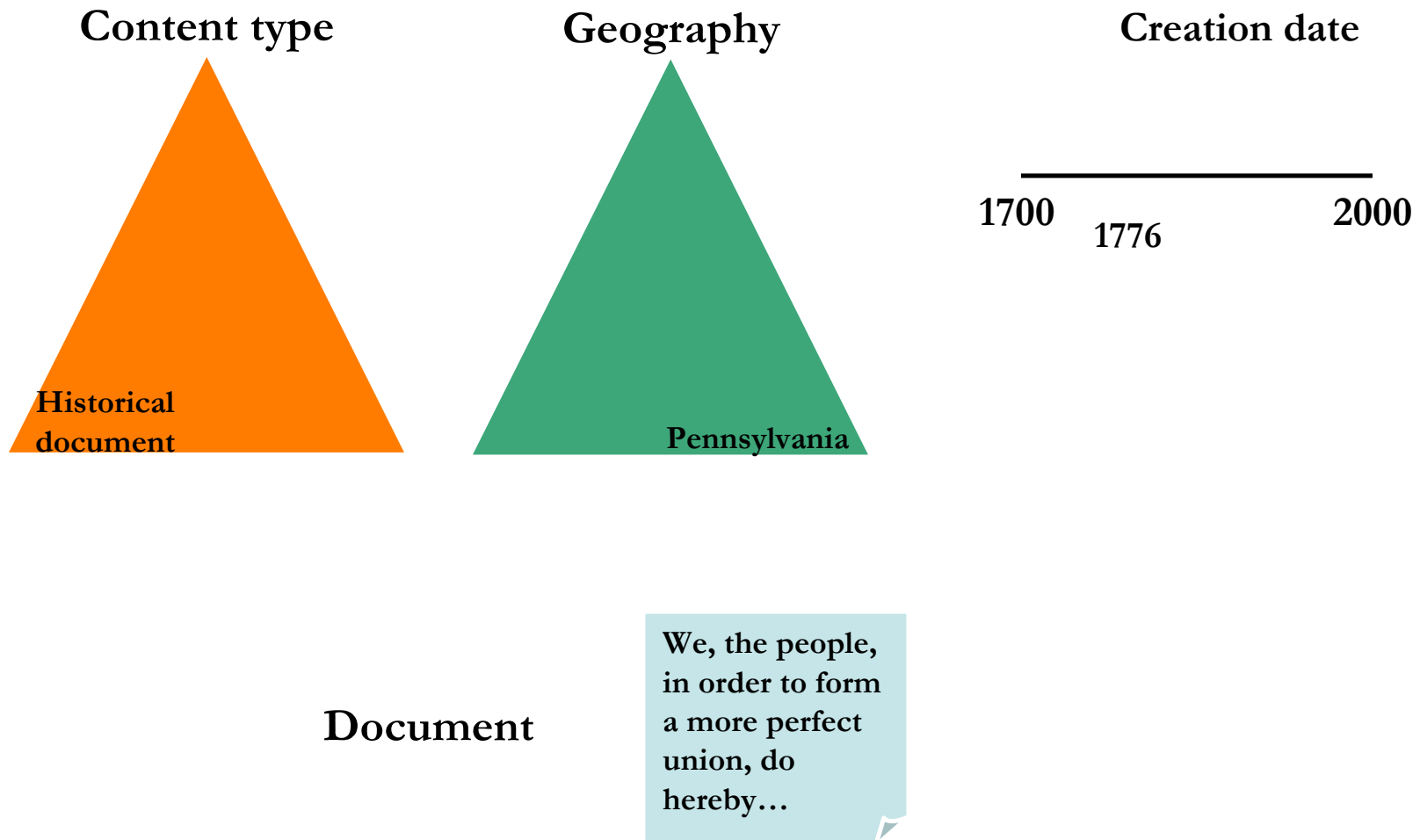
# Baltimore vs Trondheim

---

- **PODS 2005: “Multi-Structural Databases”**
  - Described basic framework
  - Gave three example query types
  - Small study
- **This talk: “Efficient Implementation of Large-Scale Multi-Structural Databases”**
  - Broader family of queries based on the framework
  - New algorithms for certain cases
  - Larger study: real-time queries on 4B web pages



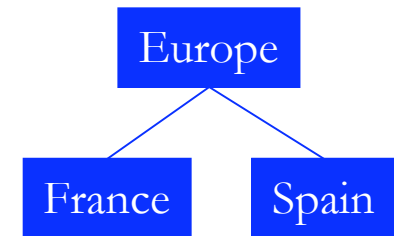
# MSDB model via an example





# Dimensions

- **Hierarchical dimensions**
  - Eg, a tree of geographic locations, topics
  - Restrictions are nodes of the tree  
Eg, restrict to documents from Europe
- **Numerical dimensions**
  - Eg, timestamp, price, temperature
  - Restrictions are intervals (ranges)  
Eg, documents from 1720 to 1860
- Most generally, **lattice dimensions**
  - Dimension is a bounded lattice
  - Restrictions are lattice elements



We consider only hierarchical and numerical dimensions



# An example MSDB

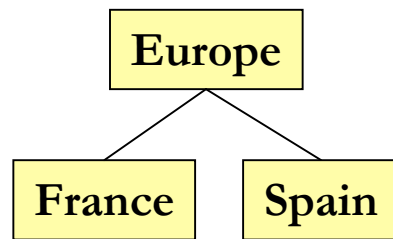
---

<b>Doc ID</b>	<b>Name</b>	<b>Content type</b>	<b>Geography</b>	<b>Topic</b>	<b>Date</b>
1	Declaration of Independence	Historical document	Pennsylvania	Politics	1776
2	The Dilbert Principle	Book	US	Humor	1995
3	Yahoo! announces earnings	News article	Washington, DC	Financial	2004



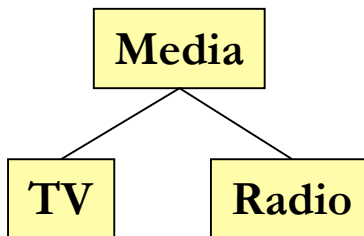
# Combining dimensions

## Geography

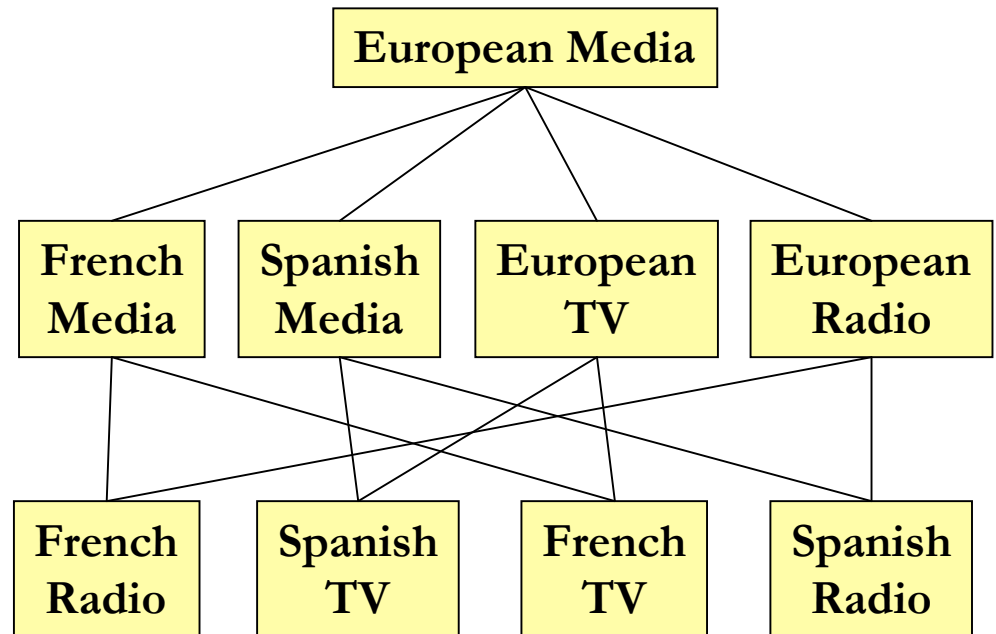


**X**

## Content type



**=**

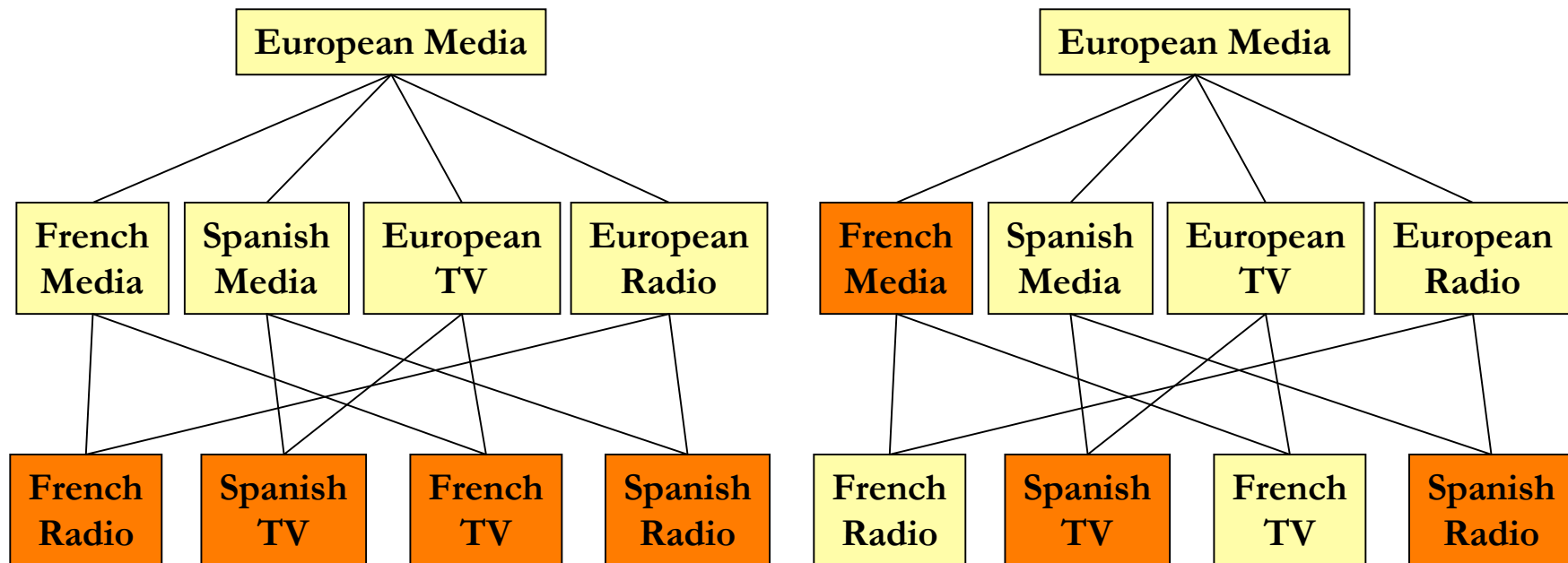


**Multi-Dimensions:** Easy to name and communicate elements and is based on user's view of the world  
Eg, ({Time, Topic}): 2003 Politics, 1990s Music



## Pairwise-Disjoint Collections (PDCs)

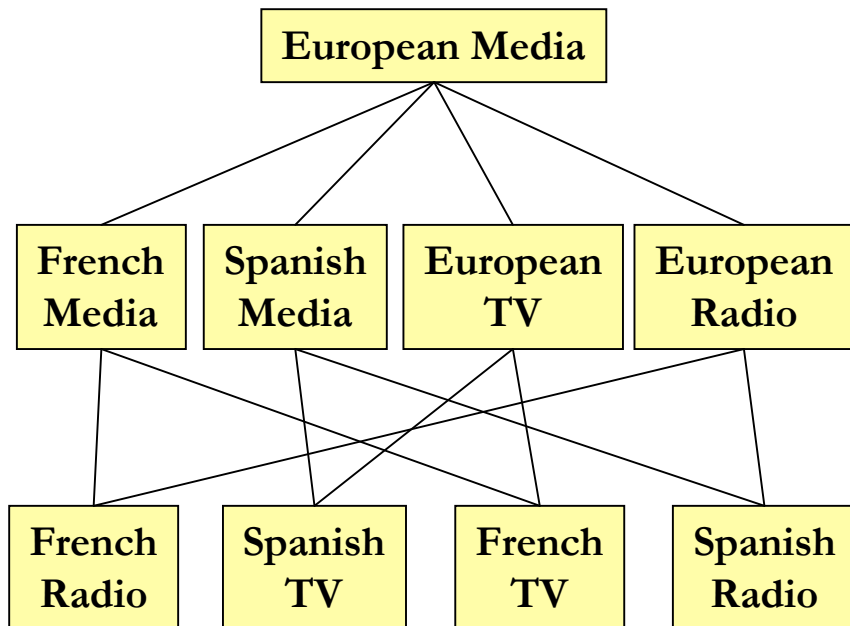
- Elements of a PDC represent **disjoint** parts of the concept space
- No **document overlap** unless documents appear at multiple locations



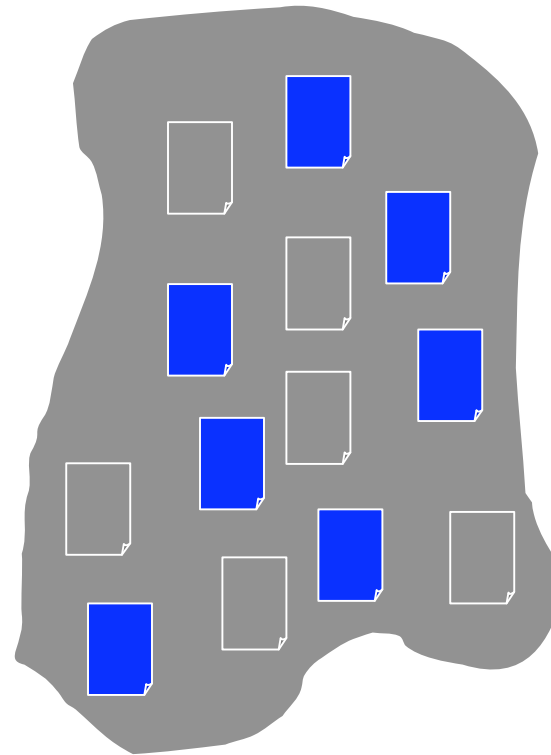


# Inputs to an MSDB query

Multi-Dimension



Document subset



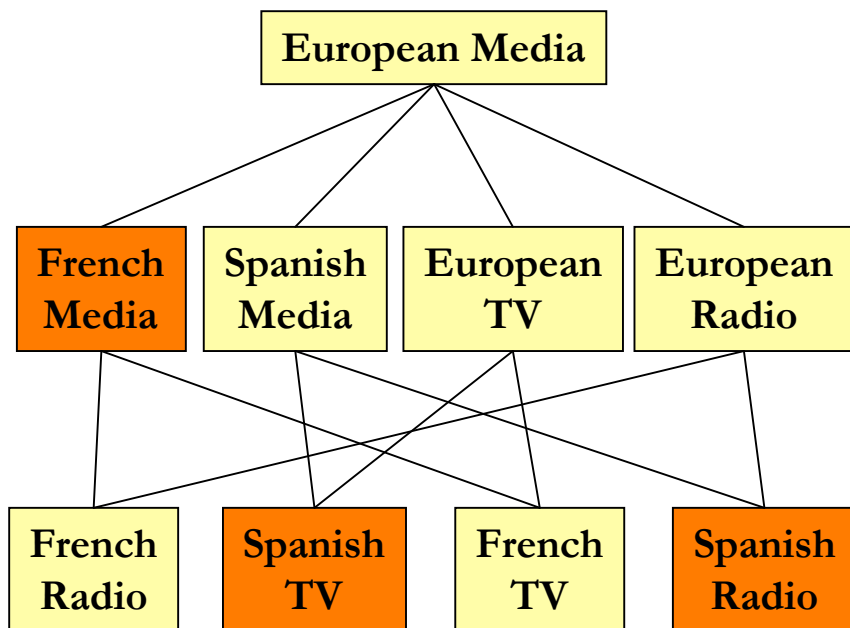
PDC size

3

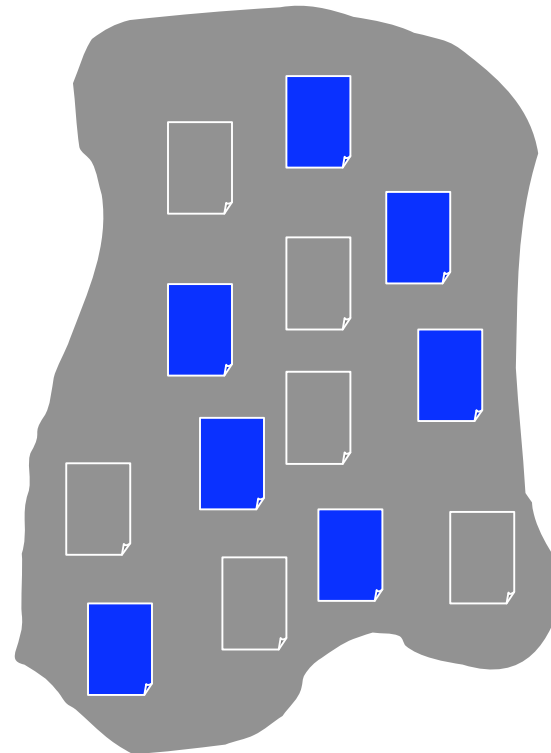


# Outputs of an MSDB query

Multi-Dimension



Document subset



PDC size

3

Output is PDC of size at most 3 that is maximal under some measure

Particular query is determined by the measure

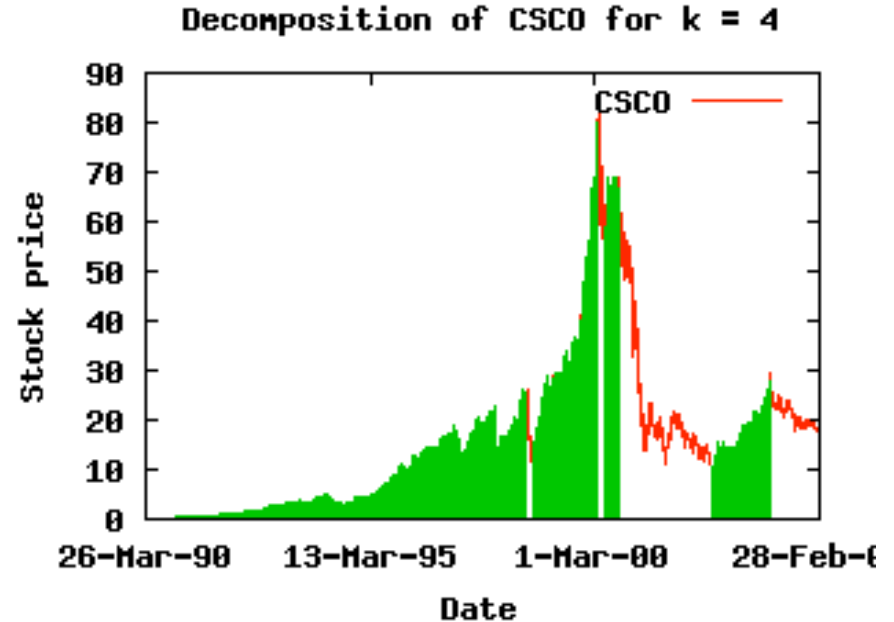


# What is the measure/query?

Queries correspond to solving an optimization problem

**Growth:** Break  $X$ ' into  $k$  pieces of maximum growth

Eg, In stock price time-series, show 4 intervals of maximum growth





## $(f, \circ)$ query type

---

Given  $D' \subseteq D$ ,  $X' \subseteq X$ , parameter  $k$ , find PDC

$\{ L_1, \dots, L_k \}$  such that

$$f(X', L_1) \circ \dots \circ f(X', L_k)$$

is maximized

Eg,  $f(X', L) = \#\{ x \in X' \mid x \text{ belongs to } L \}$ ;  $\circ = +$

**Sum-additive query type**

$$A \cap B = \emptyset \Rightarrow f(A \cup B, L) = f(A, L) + f(B, L)$$

$$\circ = +$$



# Growth in (f, o) language

---

- Find 4 regions of most rapid growth
- Numerical dimension
- A candidate f

$$g(t) = \#docs@t / \#docs@(t-1)$$

$$f([a, b]) = \sum_{t \in [a, b]} \log (g(t))$$

$$o = +$$

- Sum-additive type



# More query types

---

- **Divide**: Break  $X'$  into pieces that partition the space and have roughly equal cardinality
- **Differentiate**: Find parts of multi-dimension that occur more often in a foreground set of documents than a background set
- **Discover**: Find parts of  $X'$  defined in terms of multi-dimension that are cohesive with respect to another set of “measurement” dimensions
- **Recency**: Find regions of  $X'$  that have shown significant recent growth
- **Value**: Find parts of  $X'$  that maximize a “value” function



# Algorithms for sum-additive queries

---

## Single hierarchical dimension

- $n$  = number of nodes in the tree
- $O(nk^2)$  algorithm

## Single numerical dimension

- $n$  = number of time units
- $O(n^2 k)$  algorithm

Dynamic programming!



# Single numerical dimension

---

Given a plot of a stock price, and constrained to hold the stock during only  $k$  distinct regions, find the regions which maximize your total profit

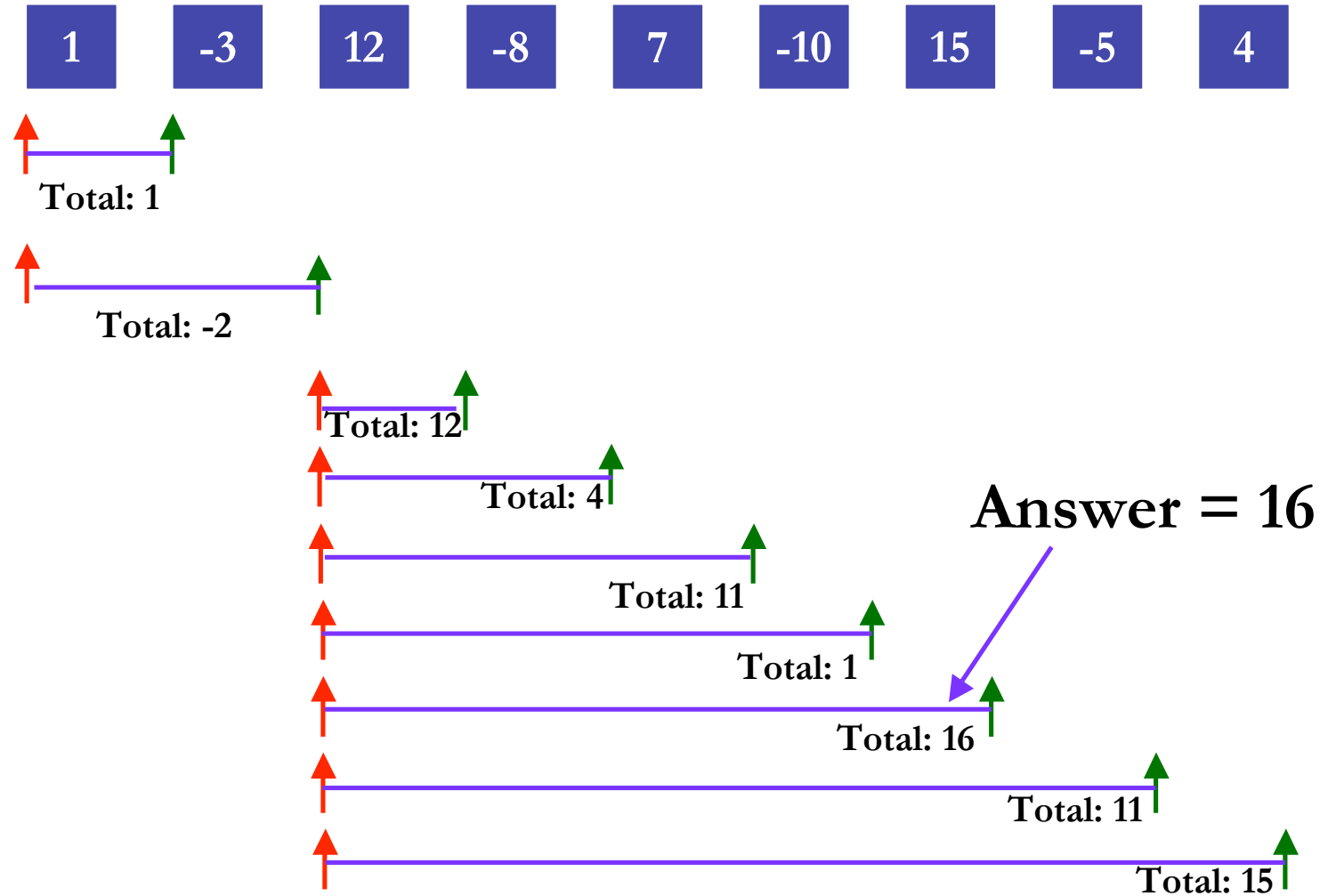
A simple abstraction: Given

Maximal subinterval problem: Find  $k$  sub-intervals of this sequence whose sum is maximal





# Folklore algorithm for $k=1$





## An $O(n^2k)$ algorithm

---

$P([1, i], k)$  = Best solution of  $x_1, \dots, x_i$  with  $k$  intervals

$P([j, i], 1)$  = Best solution of  $x_j \dots x_i$  with 1 interval

Solvable in  $O(|i-j|)$  time

$$P([1, i], k) = \max\{j < i\} P([1, j-1], k-1) + P([j, i], 1)$$

Prohibitive even when  $n$  is only large (eg, #days)

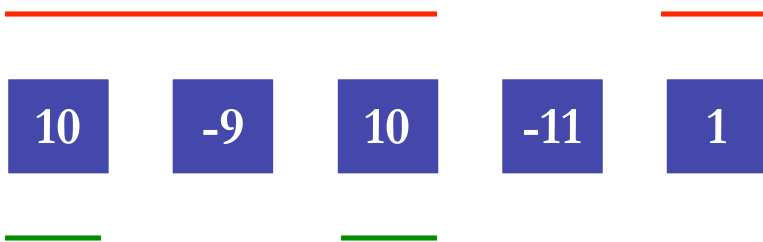
Can we do better?



# A bad idea

---

- For  $k = 2$ 
  - Pick the best interval ( $k = 1$ )
  - Remove the interval
  - Pick the best interval in the rest ( $k = 1$ )
- Doesn't work!





# A better idea



1. Solve  $k=1$



2. Invert subinterval

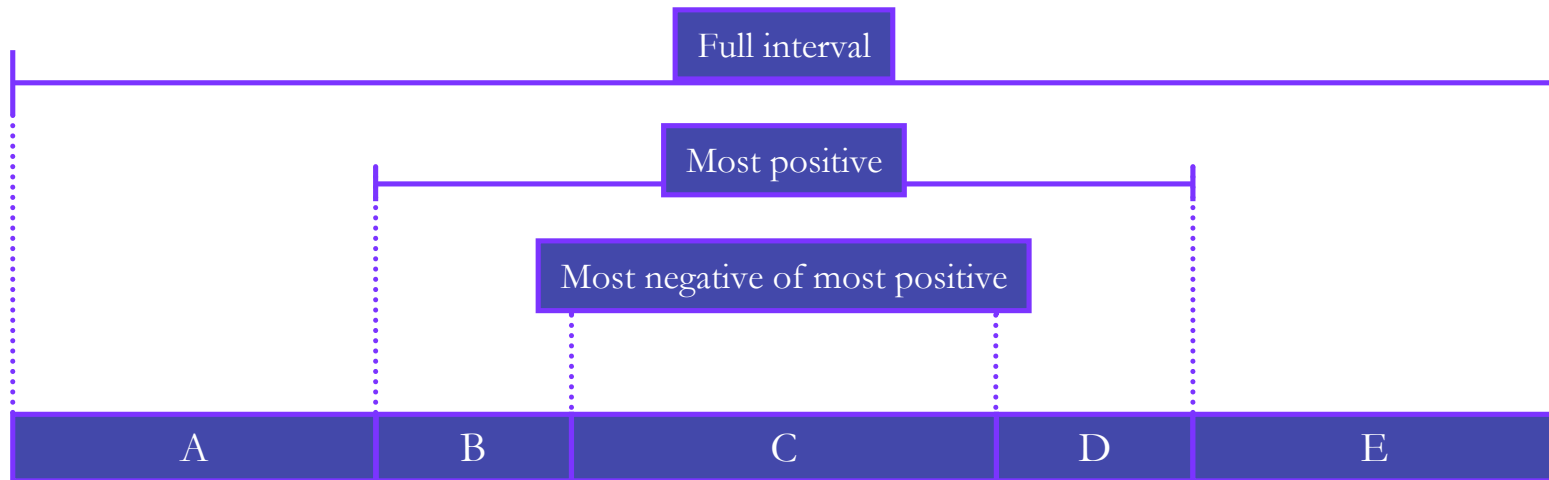


3. Solve for  $k=1$



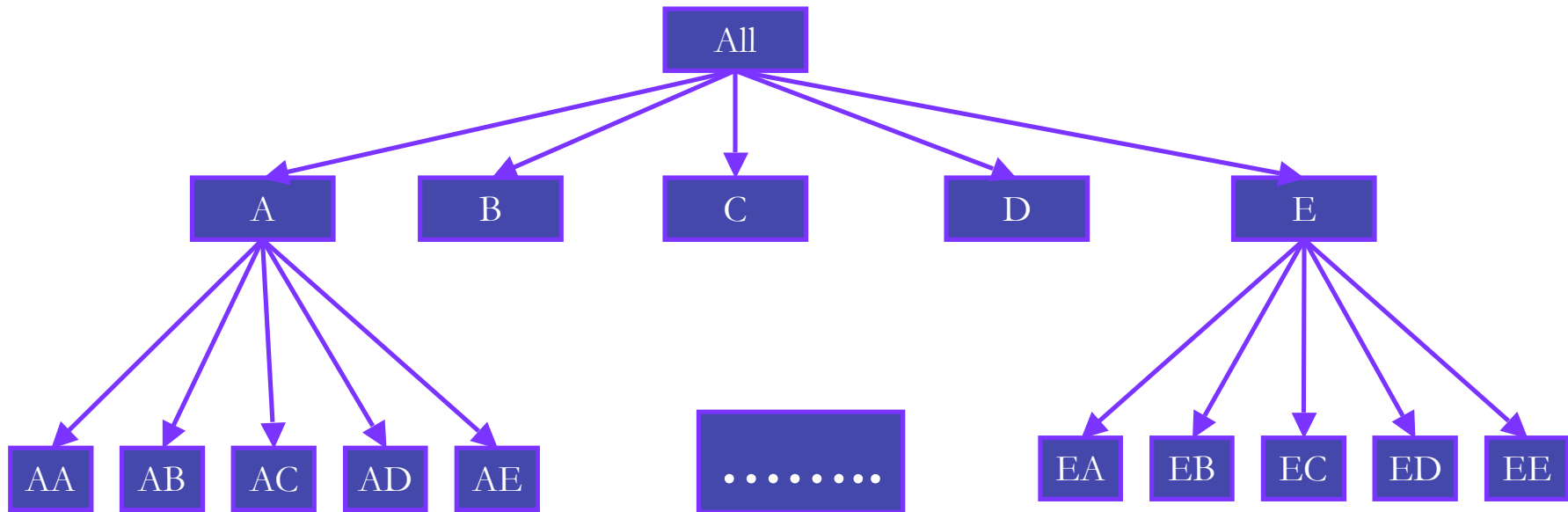


# Five-way decomposition





# Five-way decomposition tree



Continue for k levels



# Theorems

---

**Theorem:** There is an optimal  $k$ -element solution such that every interval in the solution is a node of the tree

**Proof idea:** Structural analysis of the optimal solution

**Theorem:** There is an algorithm on the tree to find the best set of  $k$  nodes with running time

$$\min (nk^2, \max (nk, k5^k))$$

**Proof idea:** Dynamic programming on the five-way decomposition tree



# Experiments and graphs

---

**See the paper!**



## Summary

---

- An abstraction of query types to capture many interesting queries for MSDB
- An almost optimal algorithm for the sum-additive query type for numerical dimension
- Large-scale experiments



# Open problems

---

- **New types of queries**
- **Broadening of the PDC concept**
- **Many interesting algorithmic open problems**
  - Variants of many of these problems have been studied before



*Thank you!*

**[ravikumar@yahoo-inc.com](mailto:ravikumar@yahoo-inc.com)**