Advanced Data Management (CSCI 640/490)

Spatial Data

Dr. David Koop





Data Exploration Through Visualization







Data Exploration Through Visualization









Why do we visualize data?

Total Bandwidth

(millions of bits per second)

















						IV		
X	У	X	У	Х	У	X	У	
10.0	8.04	10.0	9.14	10.0	7.46	8.0		
8.0	6.95	8.0	8.14	8.0	6.77	8.0		
13.0	7.58	13.0	8.74	13.0	12.74	8.0		
9.0	8.81	9.0	8.77	9.0	7.11	8.0		
11.0	8.33	11.0	9.26	11.0	7.81	8.0		
14.0	9.96	14.0	8.10	14.0	8.84	8.0		
6.0	7.24	6.0	6.13	6.0	6.08	8.0		
4.0	4.26	4.0	3.10	4.0	5.39	19.0	1	
12.0	10.84	12.0	9.13	12.0	8.15	8.0		
7.0	4.82	7.0	7.26	7.0	6.42	8.0		
5.0	5.68	5.0	4.74	5.0	5.73	8.0		

D. Koop, CSCI 680/490, Spring 2022









						IV		
X	У	Х	У	Х	У	X	У	
10.0	8.04	10.0	9.14	10.0	7.46	8.0		
8.0	6.95	8.0	8.14	8.0	6.77	8.0		
13.0	7.58	13.0	8.74	13.0	12.74	8.0		
9.0	8.81	9.0	8.77	9.0	7.11	8.0		
11.0	8.33	11.0	9.26	11.0	7.81	8.0		
14.0	9.96	14.0	8.10	14.0	8.84	8.0		
6.0	7.24	6.0	6.13	6.0	6.08	8.0		
4.0	4.26	4.0	3.10	4.0	5.39	19.0	1	
12.0	10.84	12.0	9.13	12.0	8.15	8.0		
7.0	4.82	7.0	7.26	7.0	6.42	8.0		
5.0	5.68	5.0	4.74	5.0	5.73	8.0		

D. Koop, CSCI 680/490, Spring 2022



Mean of x	9
Variance of x	11
Mean of y	7.50
Variance of y	4.122
Correlation	0.816



















Mean of x	9
Variance of x	11
Mean of y	7.50
Variance of y	4.122
Correlation	0.816











Visual Pop-out











Supporting Scalable Visualization

- Two Problems:
 - Lots of data, how to display (encode) it
 - User interaction is key to gaining insight, requires low latency
- Addressing big data:
 - Encoding should focus on available resolution, not size of data
 - Approaches:
 - Sampling
 - Modeling
 - Binning
 - Bin \rightarrow Aggregate (\rightarrow Smooth) \rightarrow Plot









Effects of Latency

- **Higher latency** leads to...
 - Reduced user activity and data set coverage
 - Significantly fewer brushing actions
 - Less observation, generalization & hypothesis
- Interaction effect: Exposure to delay reduces subsequent performance in low-latency interface.
- Different interactions exhibit varied sensitivity to latency. Brushing is highly sensitive!













Interactive Scalability Solutions

- Use Database Technology: databases built for scalability
- Client-side Indexing (Data Cubes): take advantage of data structure
- Prefetching: load data before requests based on predictions
- Approximation: show estimates early but with error information











Sampling vs. Aggregation





Full 5-D Data Cube



D. Koop, CSCI 680/490, Spring 2022

Northern Illinois University

Break into Tiles



D. Koop, CSCI 680/490, Spring 2022





Data Cube Decomposition



D. Koop, CSCI 680/490, Spring 2022

Northern Illinois University



Aggregation of Tiles for Results



D. Koop, CSCI 680/490, Spring 2022

NIU



Prefetching (ForeCache)

- Predict which tiles a user will need next and prefetch those
 - Use common patterns (zoom, pan)
 - Use regions of interest (ROIs)



D. Koop, CSCI 680/490, Spring 2022

next and prefetch those









Latency Differences in Tasks











Task-Prioritized Prefetching



D. Koop, CSCI 680/490, Spring 2022



NIU

Northern Illinois University

Reading Quiz





<u>Assignment 5</u>

- Divvy Bikes Data
- Spatial, Graph, and Temporal Data Processing





Data Cubes

J. Han, M. Kamber, and J. Pei





Data Cube: A Lattice of Cuboids



D. Koop, CSCI 680/490, Spring 2022

0-D (*apex*) cuboid

4-D (base) cuboid







Cube Operations

- Roll-up: aggregate up the given hierarchy
- Drill-down: refine down the given hierarchy
- Roll-up and drill-down are "inverses"









Spatial Data Exploration Motivation

L. Battle





Nanocubes for Real-Time Exploration of Spatiotemporal Datasets

L. Lins, J. T. Klosowski, and C. Scheidegger





Goal: Interactive Exploration of Data Cubes



Linked view of tweets in San Diego, US











Move to Another Location



Linked view of tweets in Ensenada, Mexico









iPhone vs. Android Map







Zoom into Chicago













SuperBowl in Indianapolis













New Year's Eve in Manhattan











Aggregations on Spatiotemporal Data

- Spatial: e.g. counting events in a spatial region (world or San Fran.)
- Temporal: e.g. queries at multiple scares (hour, day, week, month)
- Seek to address Visual Information Seeking Mantra:
- Overview first, zoom and filter, details-on-demand
- Multidimensional:
 - Latitude, Longitude, Time + more







	ult. If we	auow en	tries in the re	coxan	••			U	
	d repres	sent this a	gereation as		Country	y Device	Language	Count	
				2 WOU	All	All	All	5	
Dala Uul					All	Android	All	2	
	theureratik	SH above	eould be to se	electaaut 5 follo	All	iPhone	All	3	
	se using c		fl ^a this case, f	ive wat	All	All	eu	4	
	tds in ave	allowina	tiontainthere	perial	All	All	ru	1	
			BUIGHEORAS		All	iPhone	ru	1	
	CT CESCIT		ations for a	Ve wol	All	Android	en	2	
	Gountry	Devigeor	- Language on	Gaunty	All	iPhone	en	2	
	All_1	All	a lice All for the state						
				BULES a					
	XONIDINEL S			Praticipa					operation where
									Il un on Device
	Country	Device	Language	Count					un by's on (1)
	All	Android	en ettrib					on	$ap \ by \ box{on.} (1)$
	All	iPhone	en attrib	2^{2}		all possible s	subsets of		e. Note that the
	All	iPhone	group by	y on Ly		Device, Lan	guage}		As the results
			of group	by's,e	Hater	ation is "r"	vuii uv uv	VII UD IV	lations, we can
		ETHERE 10		F268 P16	statuch	the pations	As we wi	ll descri	be nanocubes is
				reatedat	dettio	the restore	and query	cubes o	f roll ups.
			Langhage			adivalont to			
SVILIWE PETER TO TECOTO	S III and att	OTHEREDUCE	TLATAULTON A	LOC NAL	SAUD	SCHOMPACT	SPATIO	UNS. TEMPOI	RAL ROLL-UP
tes. stributes and hall		PHOLAMON	Tels easy top	THE SE	deelte		subsets or		
les verivediftomabe	asestation		ist Riattribut	estanda	naggi	Device, Lan	guage}		
H, its fun totat the fork	avanteditte		matentaty	latesati	esentities	computer a	are necessa	rily bou	nded by display
thes with the solution		on an as	EKE S (SUKE) HABIG		velocal do	lithe to be a	ble to quic	kly colle	ect subspaces of
D. Koop, CSCI 680/490	J, Spring 202	Entreomh	storatilite datas	et that	Wallan	end up in th	e same pix	cel on th	e screen. How-



Nanocube Queries

ine Submission ID: 276

• Representing natural language queries as data cube queries

$\triangleright n > 0$								
ty podenissio	National dange (anotyo, $d, S, \ell_{\text{time}}, updat$	ed_snodes)	c		t		URL	
	Count of all Delta flightsN(S, d)		R	{ Delta }	R	U	/where/carrier=Delta	
~ 1	Eount Statk Del Raffie BRISHER RATME COVES the	(OR,, Midwest	R	{ Delta }	R	U	/region/Midwest/where/carri	er=Delta
aen Submissio	modint-offalt flights-in 2010	R U	D		R	2010	/field/carrier/when/2010	
y node	15 prosections of a protect of the sine 200 dat	ed R odes) U	R	{ United }	D	2009	/tseries/when/2009/where/ca	rrier=United
$\sim n \sim 0 n \sim 276$	² chieatmap & Berta Hights 12910	D tile0	R	{ Delta }	R	2010	/tile/tile0/when/2010/where/	carrier=Delta
v node	37: stackupdate AILEROPERPATH(root, [l] 1: procedure ADD(root. o. d. S. luima updat	$\{o\}, \dots, \ell_k(o)\}$ ed_nodes	•					
• S	48: Child Chode has a single child then 25 Store Charles a single child then	$t = time_{D}$						
1. procedu	3. A DESTACK TARAL PROPER PATH (root, 1	$(O), \dots, (k_k(O))$						
ty Stack 200								
ission ID: 270		I						
4: <i>chil</i>	A null nodes BASSACK ONTENT node	CONTENT (Mild)						
v Stack	CUBLES STUDY PLANE OUTER SULLY	Fen OF OIPPIE	SU15	SION S OC		ain, i	J = universe	
??!. proced	$\mathcal{F}_{k} \leftarrow CHARMOGGEHBS a SHIELE CHHELLINENInter DUITE CHELLINE STORE CHELLINGE$) and more (1.1.1.1)	•		c			
$ssion_1 + 2$		Ed Sales then	in a	an arrav	OŤ	CUM	iulative counts	
sional D: 200	SEAST PROPERTY AND A CONTENT OF	SERIES(): NODE	())					
ssion5ID: 246	RAUPROPRE ASHI AGE, ON CO.P.Y. (CONTEN	Г(<i>node</i>)))						
24: proceou g.6: [ℓ₁.if	PARTE PORTES ID CLEAN HERE SHARRED (NODE	and SERIES(1): NODE(
()) 5:- White W	STACK IS NOT CROPY TO A PERIOD TO THE PROPERTY OF THE PROPERTY) then						
242 protect	The the second she have been the sheep of the second	andie						
	A CALLER AND A C	ated_nodes then						.
on-H2: 27.0	TOSERSELARD DER MARKEN CHILD	then						[Lins et. al,
	CONTENT	T(node)))						

D. KOOP, CONTRACTOR (Match) $(\ell_{ime}(o))$









Building a Nanocube









Summed-area Table

• Every node in the previous figure stores an array of timestamped co this:



bin: 1	bin: 3	bin: 4	l b
accum: 2	accum: 3	accum: 4	aco

A Summed Table Sparse Representation for Counts

- query/tseries/1/3/4
- start at bin 1, use buckets of 3 bins each, and collect 4 of these buckets















NIU



D. Koop, CSCI 680/490, Spring 2022



 o_1



Northern Illinois University





D. Koop, CSCI 680/490, Spring 2022

 O_5









Nanocubes Discussion

- sparseness
- Limited to one spatial dimension, one temporal dimension Precompute once, then exploration has low latency

D. Koop, CSCI 680/490, Spring 2022

Save space by organizing the data in a manner that takes advantage of data







Example: American vs. Delta



D. Koop, CSCI 680/490, Spring 2022





[Lins et. al, 2013]



Example: Cell Data Records



D. Koop, CSCI 680/490, Spring 2022





Example: Cell Data Records



D. Koop, CSCI 680/490, Spring 2022

Total count: 1,490,178 of 1,043,884,027

Duration over_60_min 30_to_59_min 10_to_29_min 5_to_9_min 1_to_4_min under_1_min



[Lins et. al, 2013]





Big Spatial Data Management

A. Eldawy



