LARGE DATA IN VISUALIZATION

WHAT IS “LARGE” ?

DATABASES - PETABYTES	10^{12}
INTERNET - GIGABYTES	10^{9}
VISUALIZATION - ???	$10^{?}$

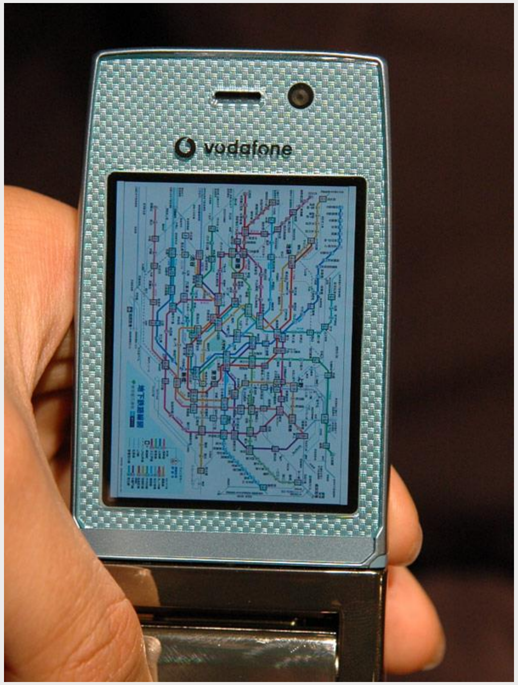
LIMITS OF VISUALIZATION:

Computer display
Human visual system

LIMITS OF COMPUTER DISPLAY

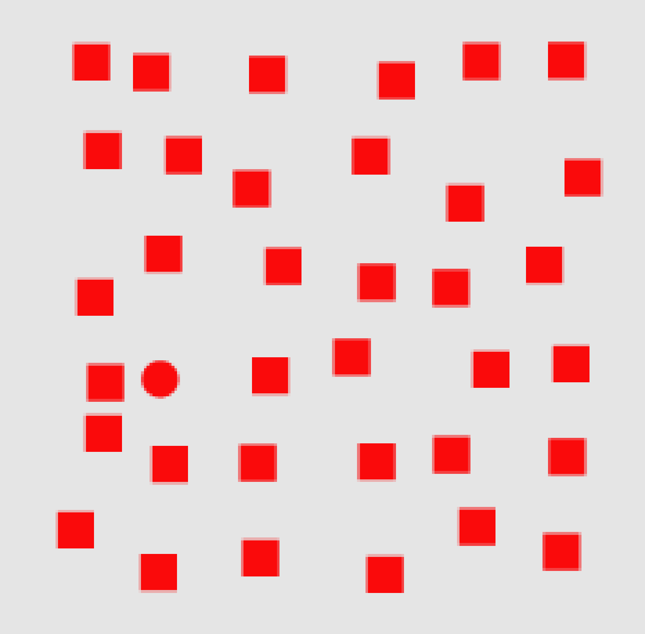
RESOLUTION

From cell phone   to Powerwall



LIMITS OF HUMAN VISUAL SYSTEM

NARROW ANGLE OF SHARP VISION (FOVEA)
SACCADIC MOVEMENTS (BROWSING)
SHORT-TERM VISUAL BUFFER
LIMITED ICONIC MEMORY



8568972698468976268976435892265986
0246299687402655762798678904567923
9083457980279075904709827908579084
98709856749068975786259845690243790472190790709811450
85689726984689762689764458922659865986554897689269898

WHAT IS “LARGE” IN VISUALIZATION?

DEPENDING ON THE TYPE OF VISUALIZATION, EACH ITEM TAKES UP SOME PIXELS:

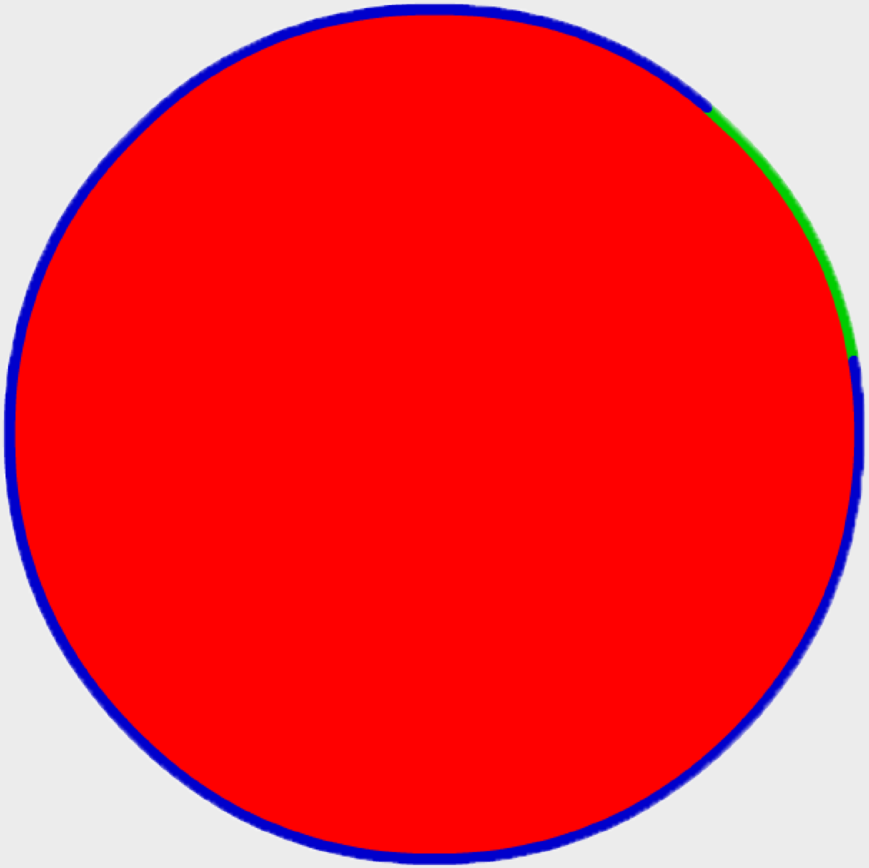
Scatterplots:	1 item ~ 1 pixel
Iconic displays:	1 item ~ 100 pixels
Parallel coordinates:	1 item ~ 1000 pixels

ALREADY HUNDREDS OF RECORDS CAN FILL THE DISPLAY

REAL WORLD DATA
~ OFTEN UP TO 10^6 - 10^7 ITEMS

PROBLEMS WITH LARGE DATA

MOTIVATION



IMDB database – 250k movies, 900k actors

ATTRIBUTES OF VISUALIZATION

STANDARD ATTRIBUTES OF ALGORITHMS:

Computational complexity
Memory demands

VISUALIZATION = ADDITIONAL ATTRIBUTES

Clarity
Truthfulness
Interactiveness

LARGE DATA VISUALIZATION

ORIGINS OF LARGE DATA

Simulation, measurement, survey, activity log…

- VISA = 6000 transaction / second

BENEFITS:

Precise description

DRAWBACKS:

Unintelligible visualization

LARGE DATA VISUALIZATION II

PROBLEMS WITH LARGE DATA VISUALIZATION:

Occlusion

} Overplotting

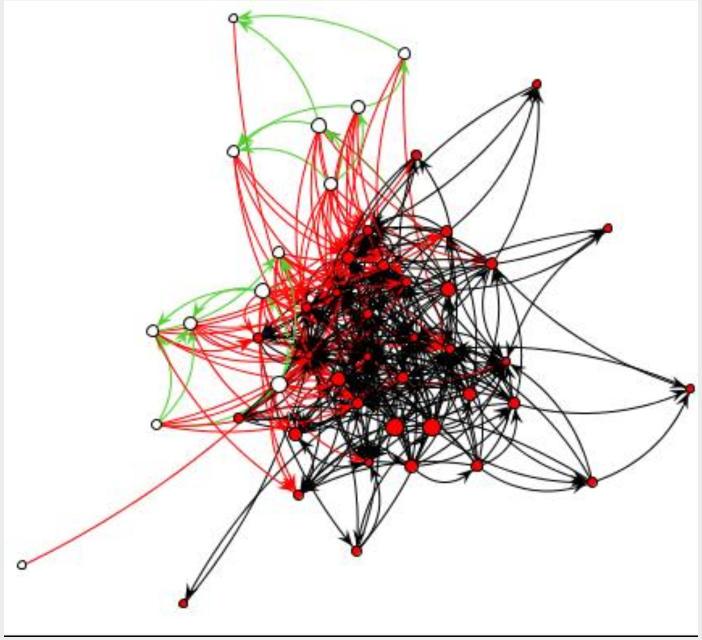
Aggregation

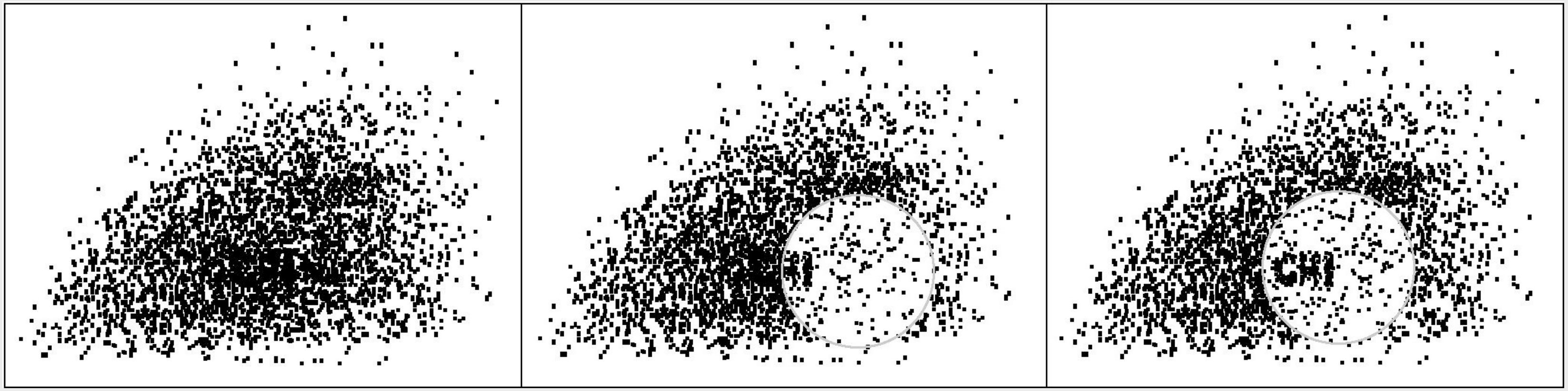
Slow response

OVERPLOTTING

OCCLUSION
- ”IS IT THERE?”

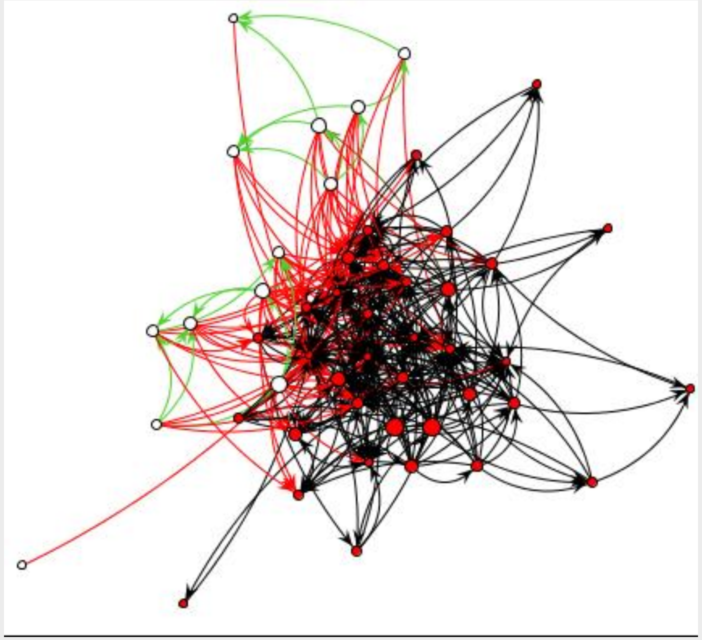
AGGREGATION
- ”HOW MANY ARE THERE?”





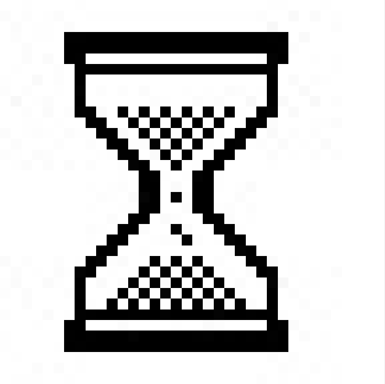
INTERACTION

SELECTING MESSY DATA IS ARDUOUS

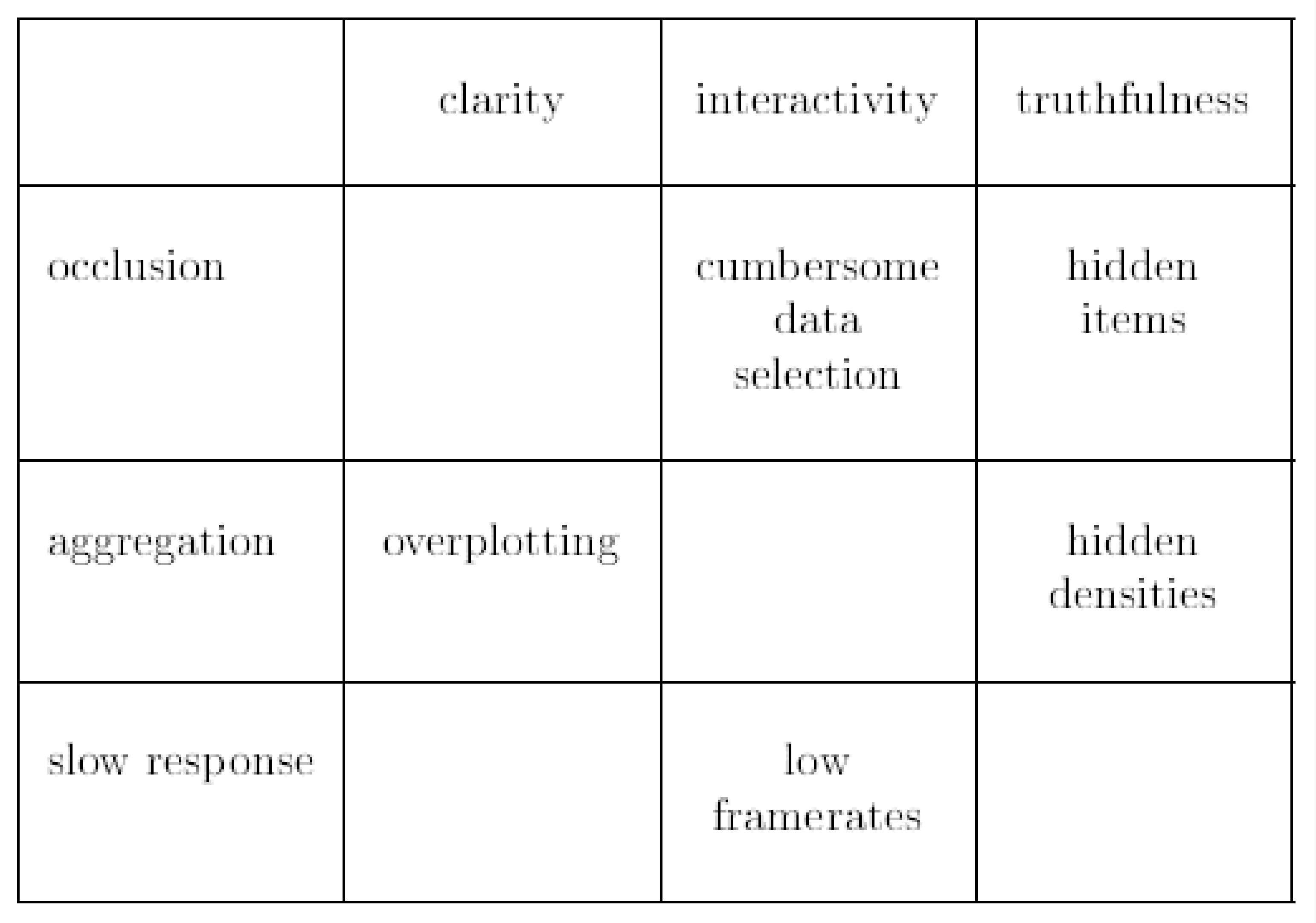


LONG RESPONSE TIME

Remember: direct manipulation and interactive display should react within a fraction of a second



EFFECT ON VIS. ATTRIBUTES



	clarity	interactivity	truthfulness
occlusion		cumbersome data selection	hidden items
aggregation	overplotting		hidden densities
slow response		low framerates	

SOLUTIONS

REDUCE DATA

INCREASE DISPLAY CAPACITY

METAMORPHOSES

SOLUTION:
DATA REDUCTION

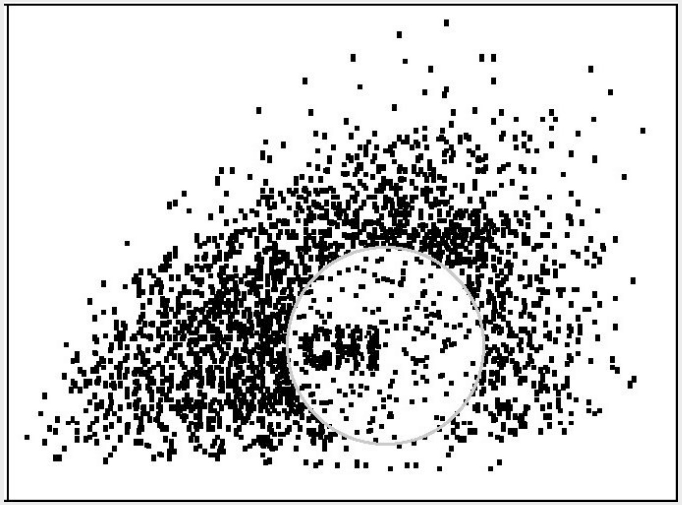
SOLUTIONS IN DATA SPACE

DATA SUBSETTING

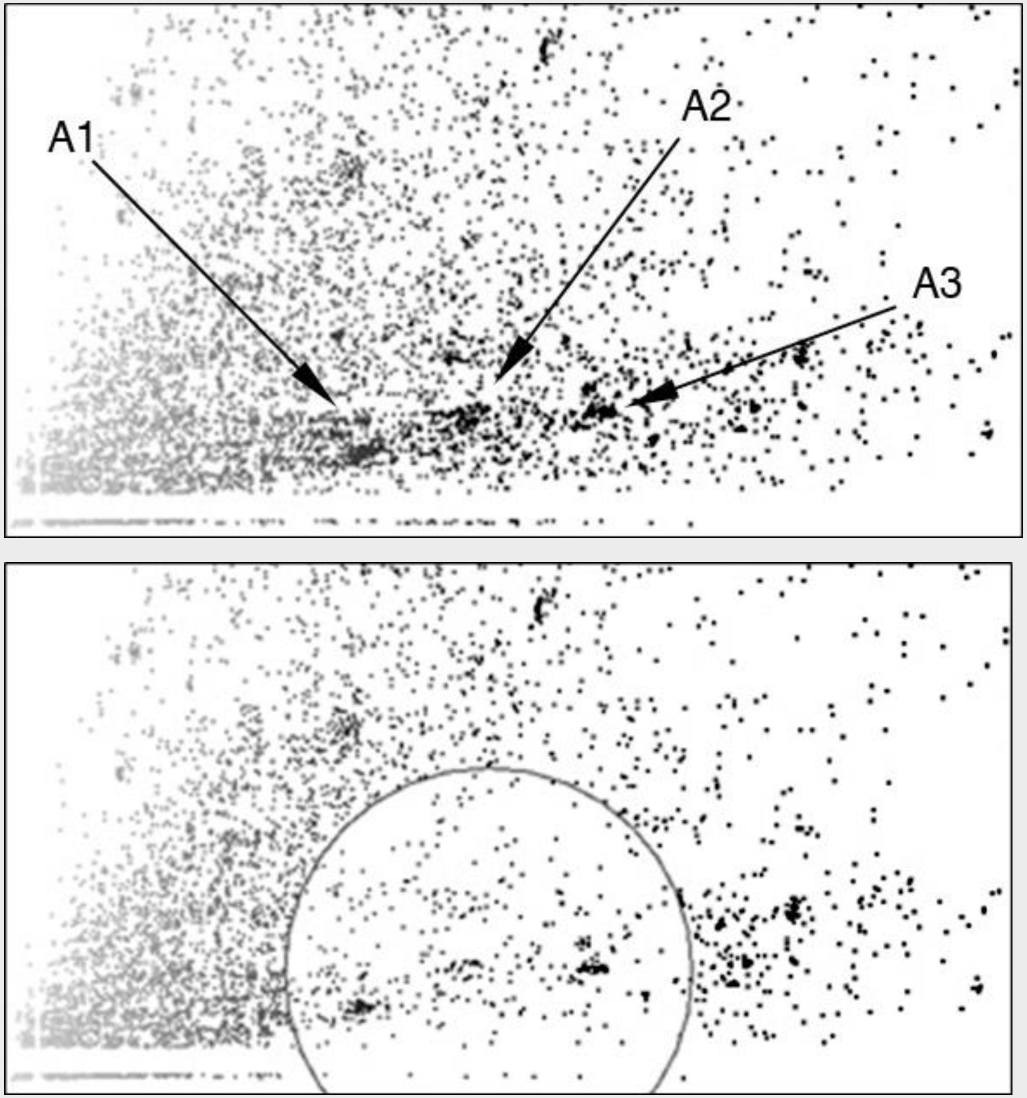
Some experts claim that only first 10.000 records are relevant, rest can be interpolated if necessary

DATA SUBSAMPLING

Subsetting ☹
Systematic sampling ☹
Random sampling ☺



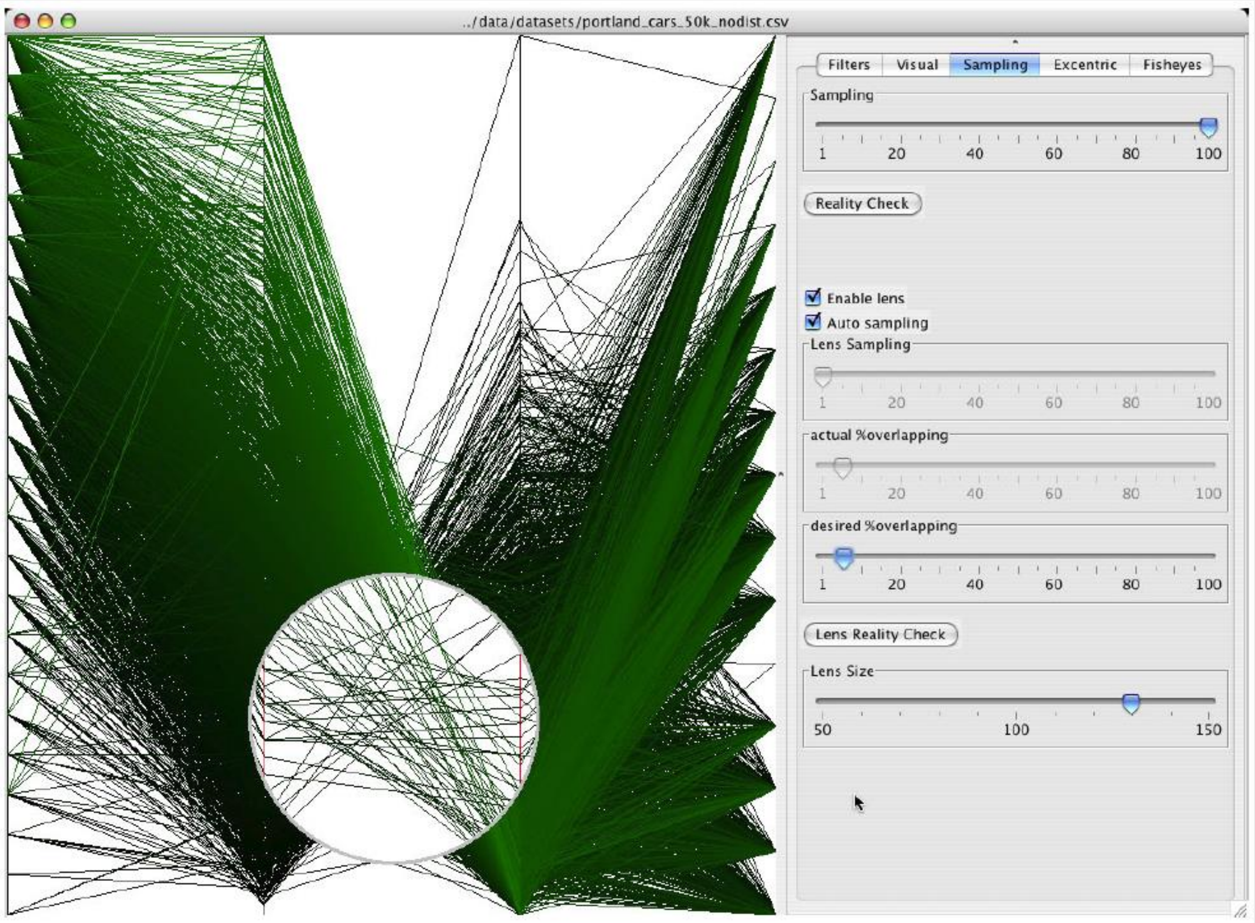
EXAMPLE OF SAMPLING



Sampling Lens
[Ellis, Bertini, Dix]

EXAMPLE OF SAMPLING II

SAMPLING LENS, AUTOMATED CLUTTER REDUCTION [DIX & ELLIS,06]

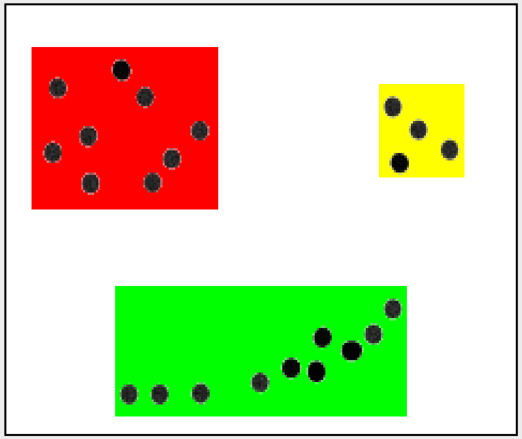


../data/datasets/portland_cars_50k_nodist.csv

☑ Enable lens

☑ Auto sampling

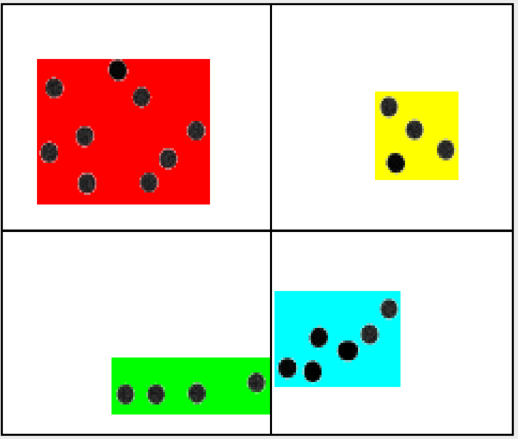
DATA AGGREGATION



CLUSTERING

Items of similar properties are grouped together

Clusters replace original records



BINNING

n-D intervals

DATA IS REDUCED

MUCH INFORMATION IS PRESERVED

HIGH N => CURSE OF DIMENSIONALITY

X^N BINS

BINNING

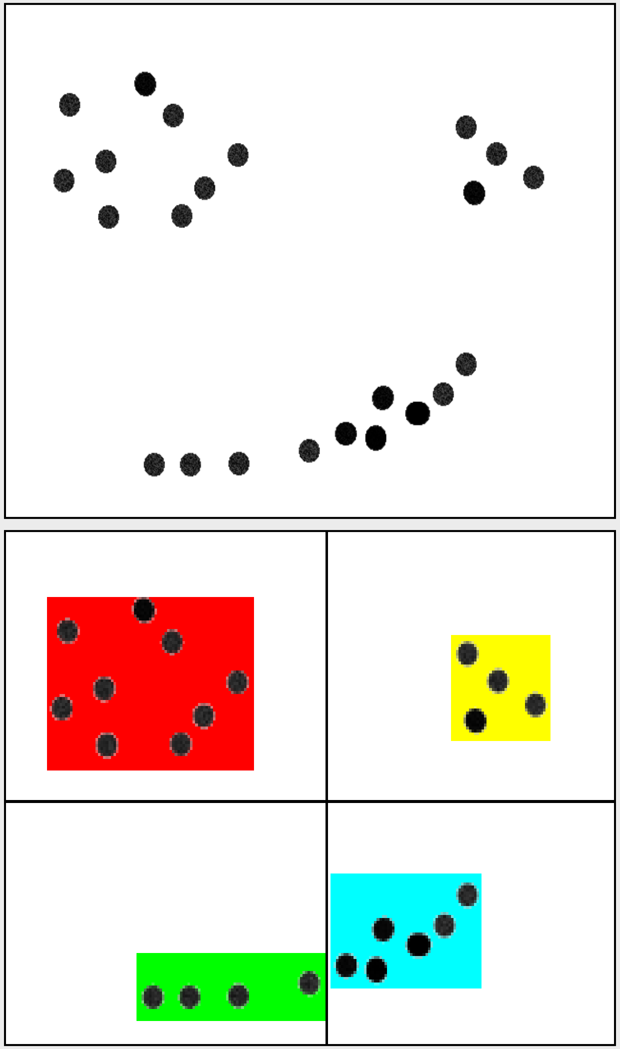
DATA RECORDS ARE REPLACED BY *N*- DIMENSIONAL BOXES (BINS)

RECURSIVE SUBDIVISION

TOP-DOWN APPROACH

Computationally cheap ☺

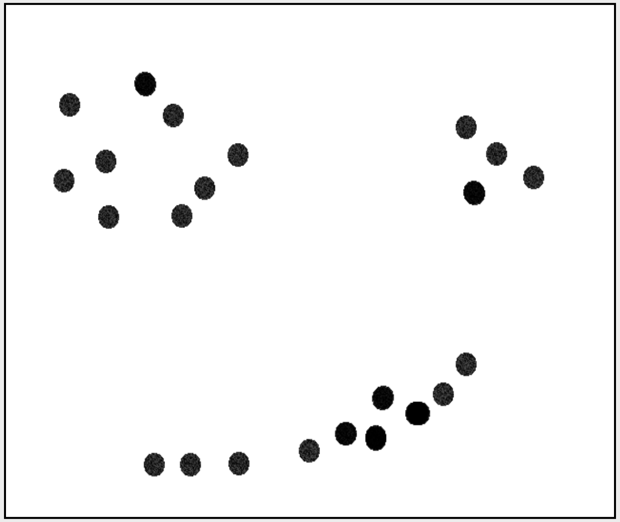
Ignores data ☹

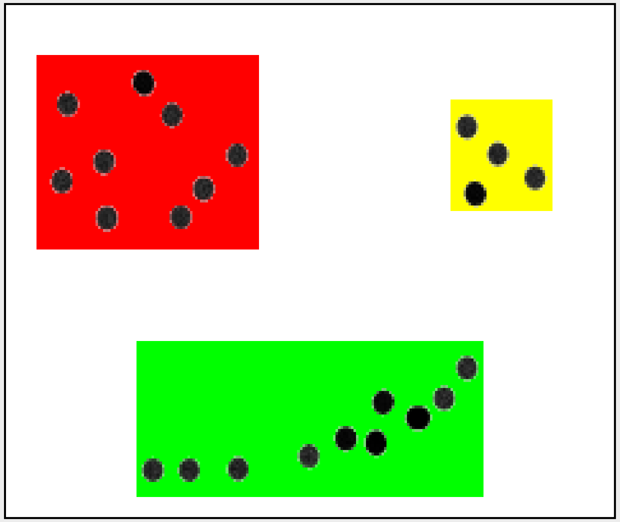


CLUSTERING

DATA RECORDS ARE REPLACED BY *N* - DIMENSIONAL CLUSTERS

i.e. cluster centroid, min/max values population, density etc….





BOTTOM-TOP APPROACH

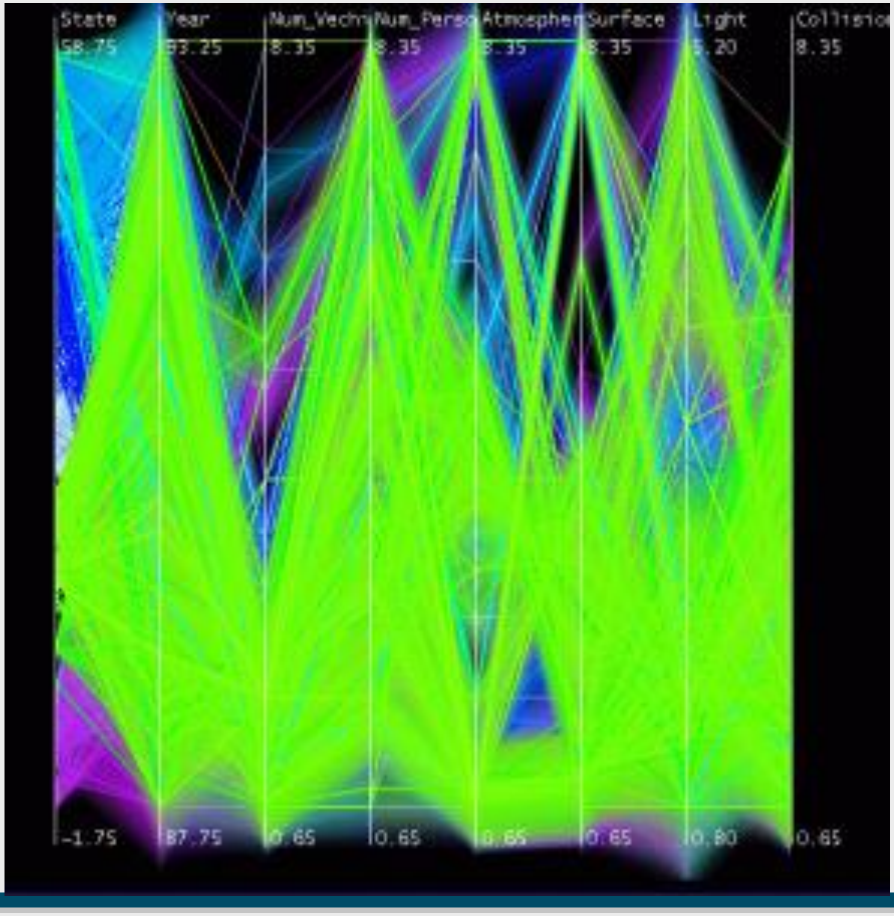
Computationally expensive ☹

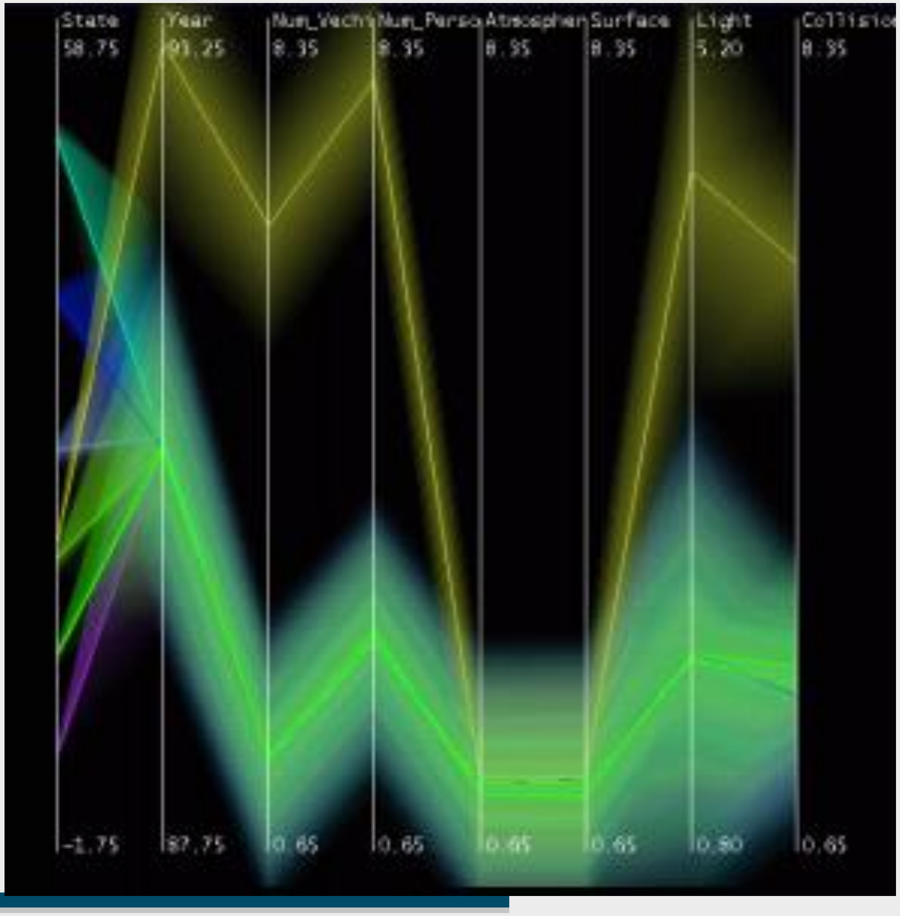
Data-aware ☺

e.g. k-means algorithm

DATA-ORIENTED EXAMPLES

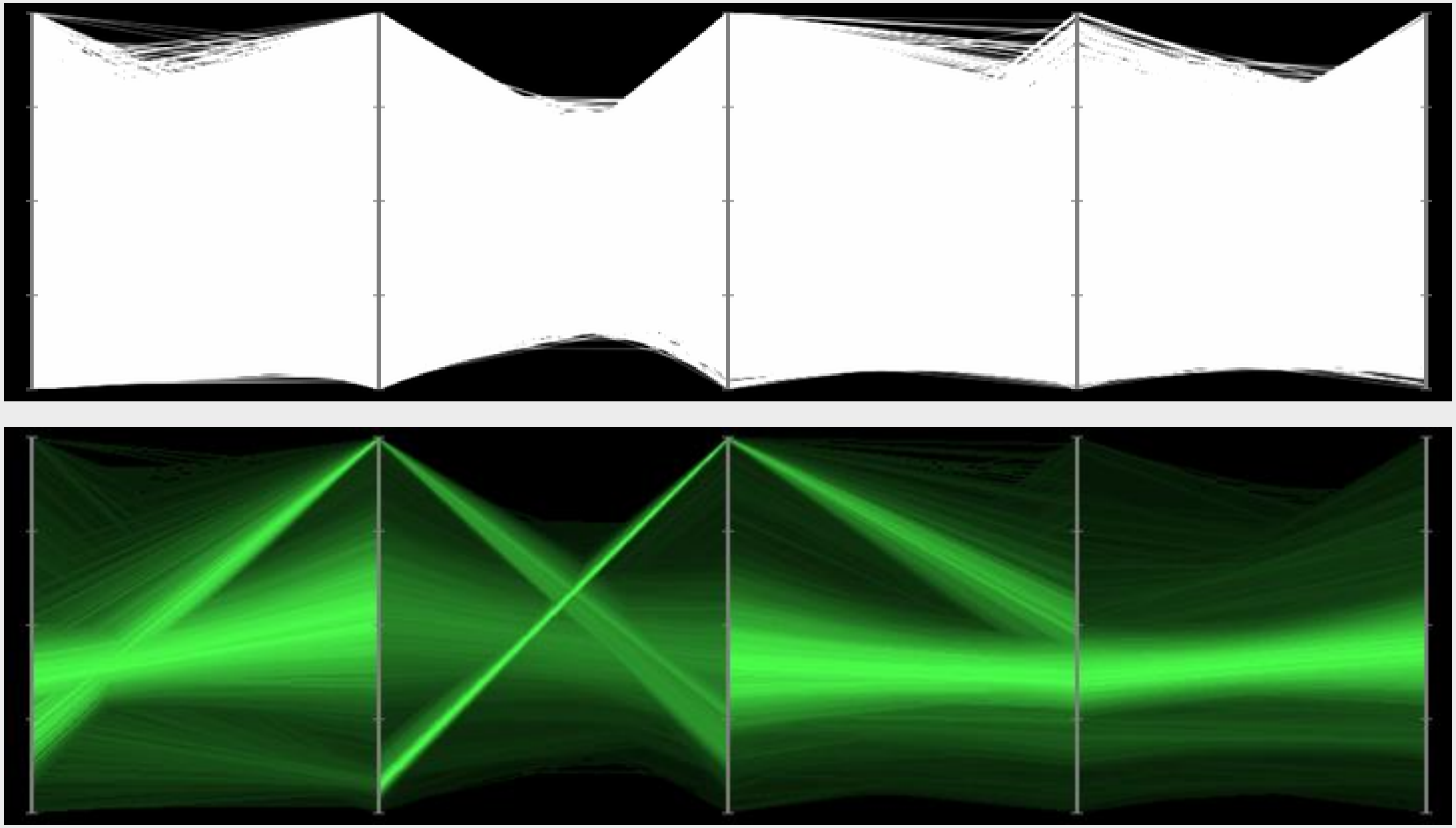
CLUSTERING IN PARALLEL COORDINATES [FUA, WARD, RUNDENSTEINER]



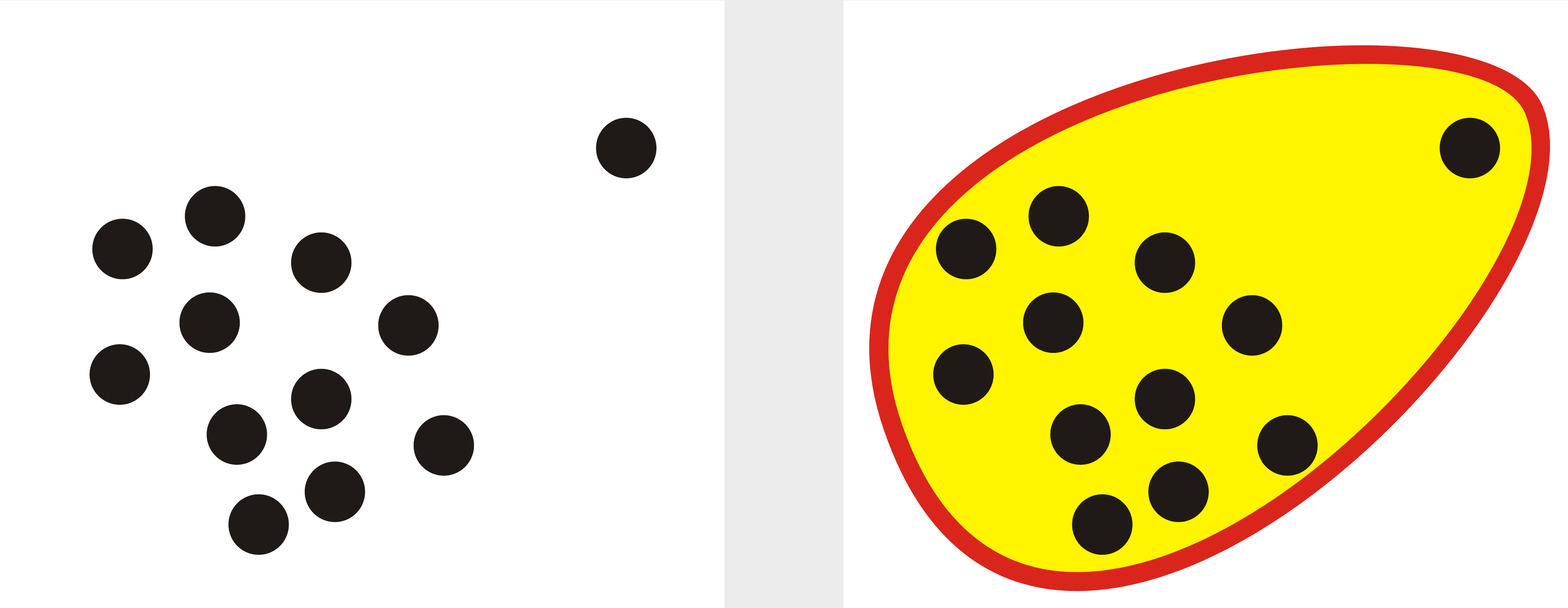


DATA-ORIENTED EXAMPLES

BINNING



OUTLIERS IN DATA AGGREGATION



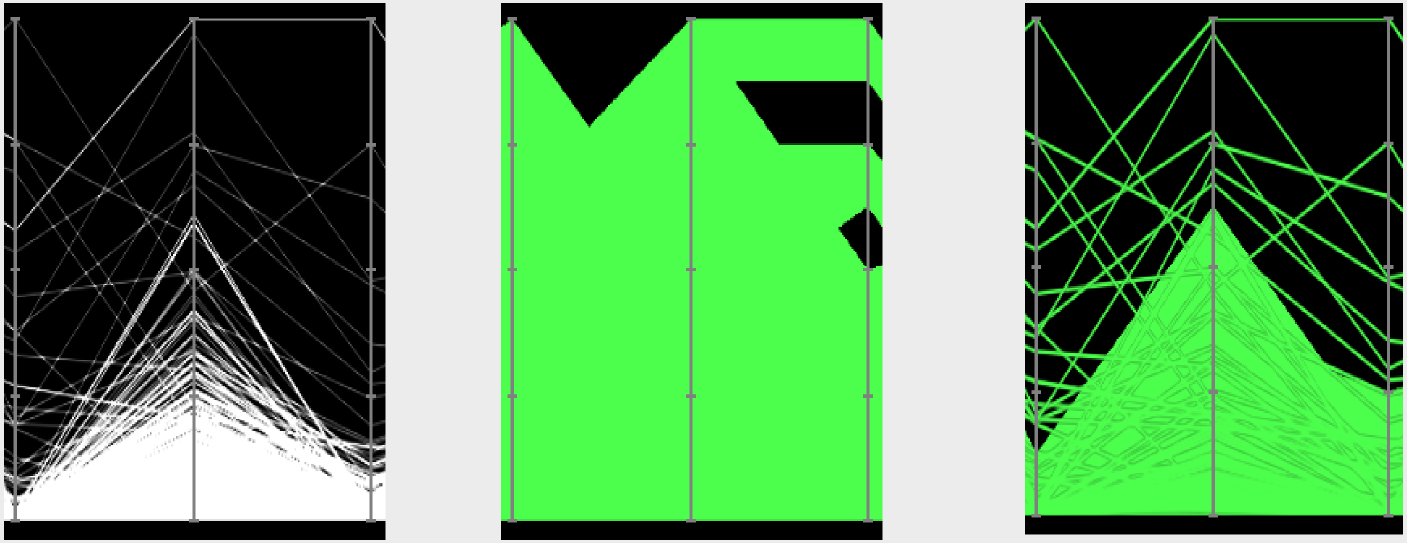
OUTLIERS IN DATA AGGREGATION

AVOID LOSING THEM IN VISUALIZATION

e.g. due to transparency or abstraction

IMPROVE DATA ABSTRACTION OR F+C

e.g. remove outliers from clustering



SOLUTION: INCREASE DISPLAY CAPACITY

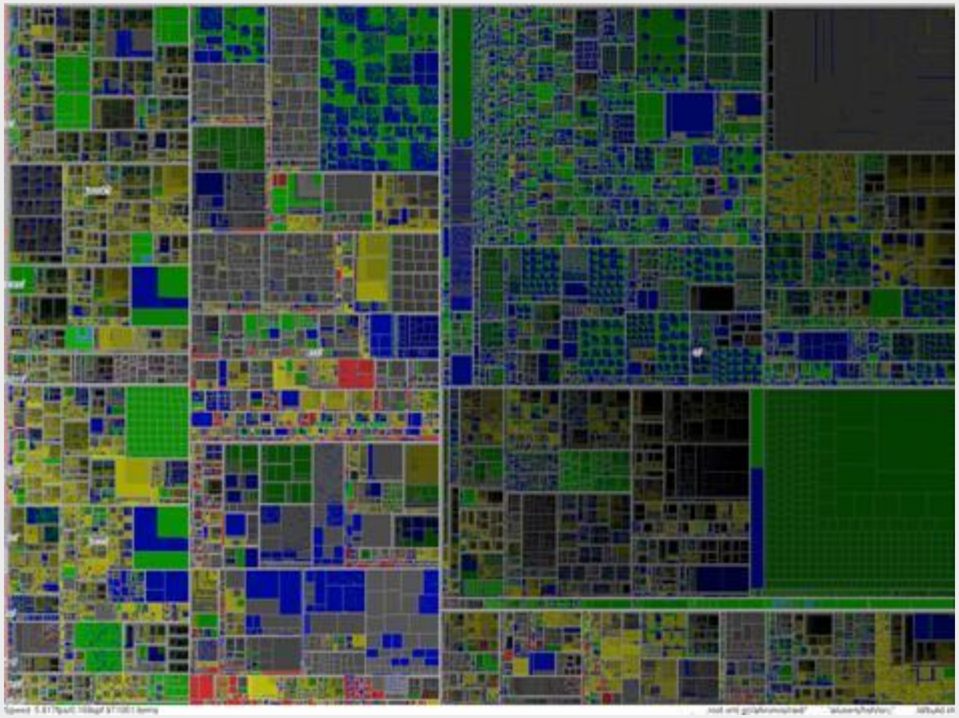
TECHNOLOGY ENHANCEMENTS

LARGE DISPLAYS

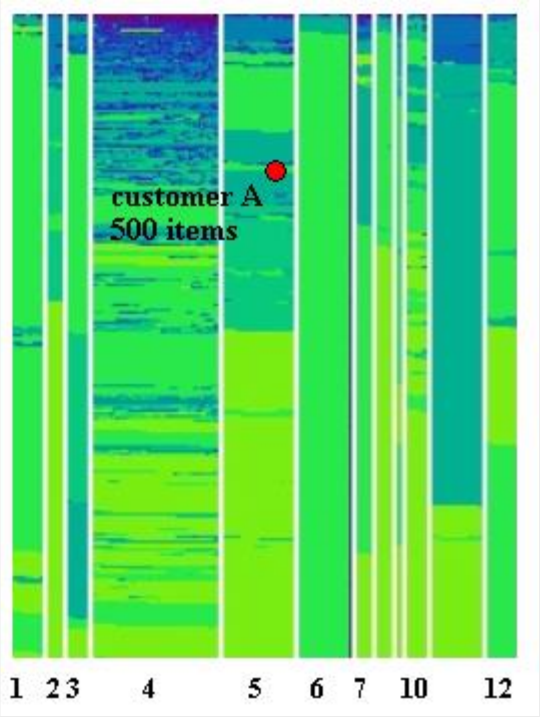
VISUALIZATION ENHANCEMENTS

SMALL VISUAL ELEMENTS

Pixel-based techniques



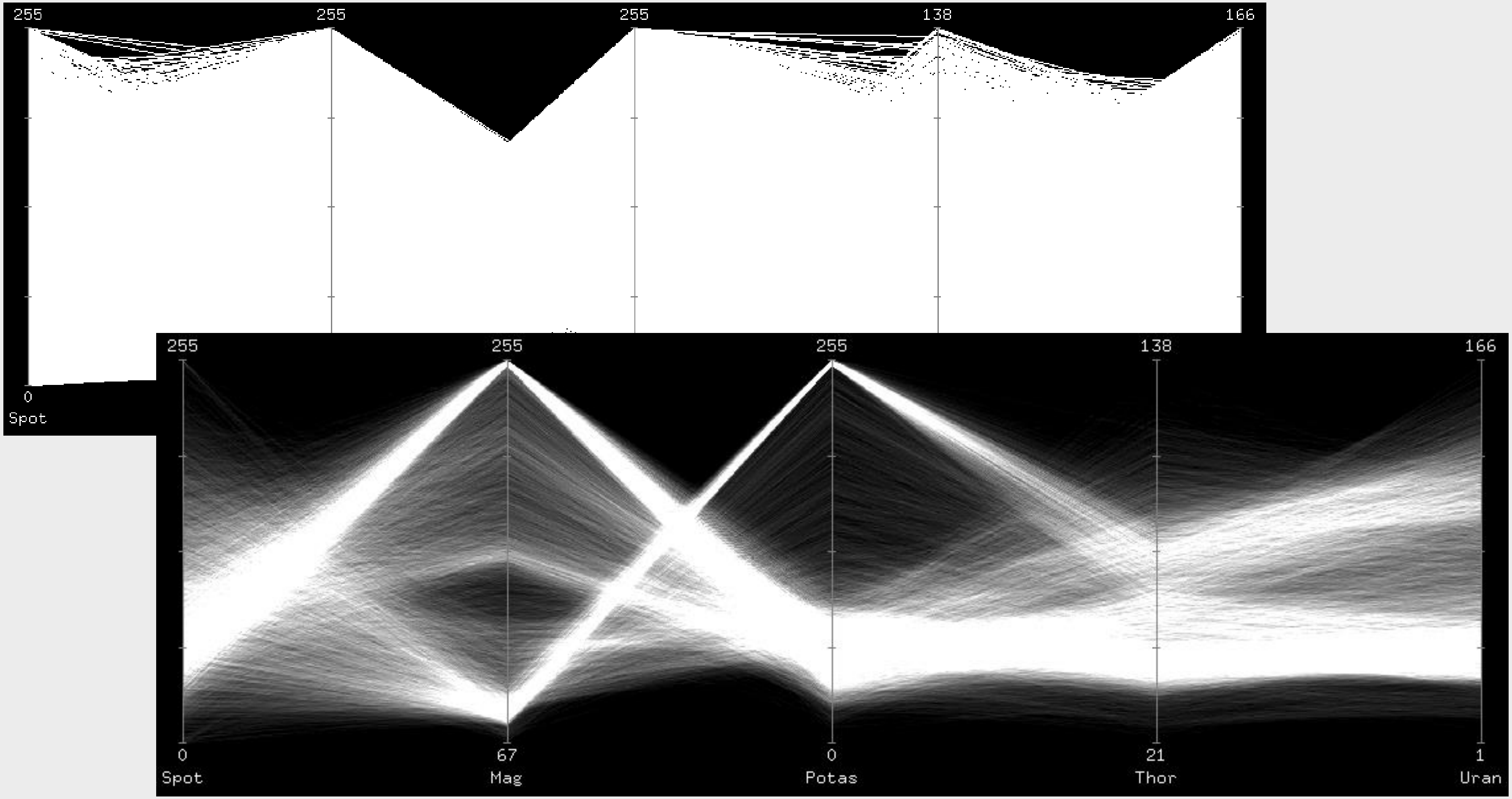
[Fekete & Plaisant, 02]



[Keim et al., 01]

VISUALIZATION ENHANCEMENT

TRANSPARENCY



SOLUTIONS IN SCREEN SPACE

RESOURCES TO MANIPULATE:

COLOR

MAPPING FUNCTIONS
(geometry, alpha...)

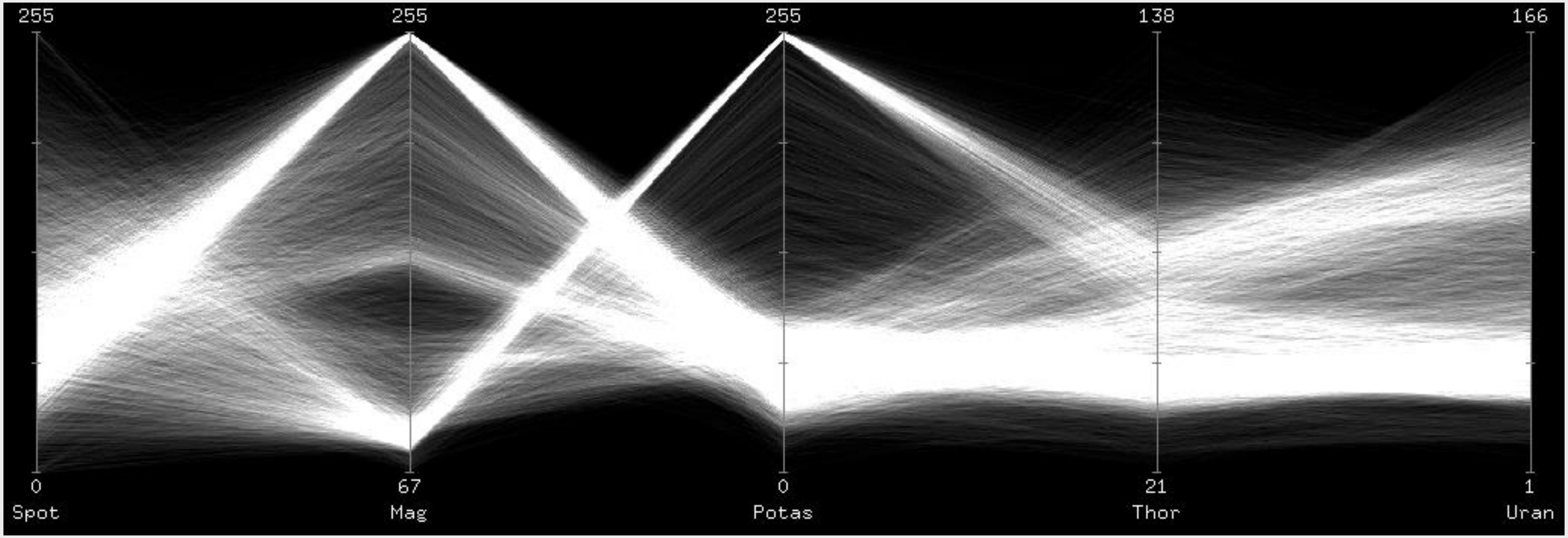
LAYOUT OF ITEMS

PROJECTION

GENERAL CONFIGURATION OF THE VIEW

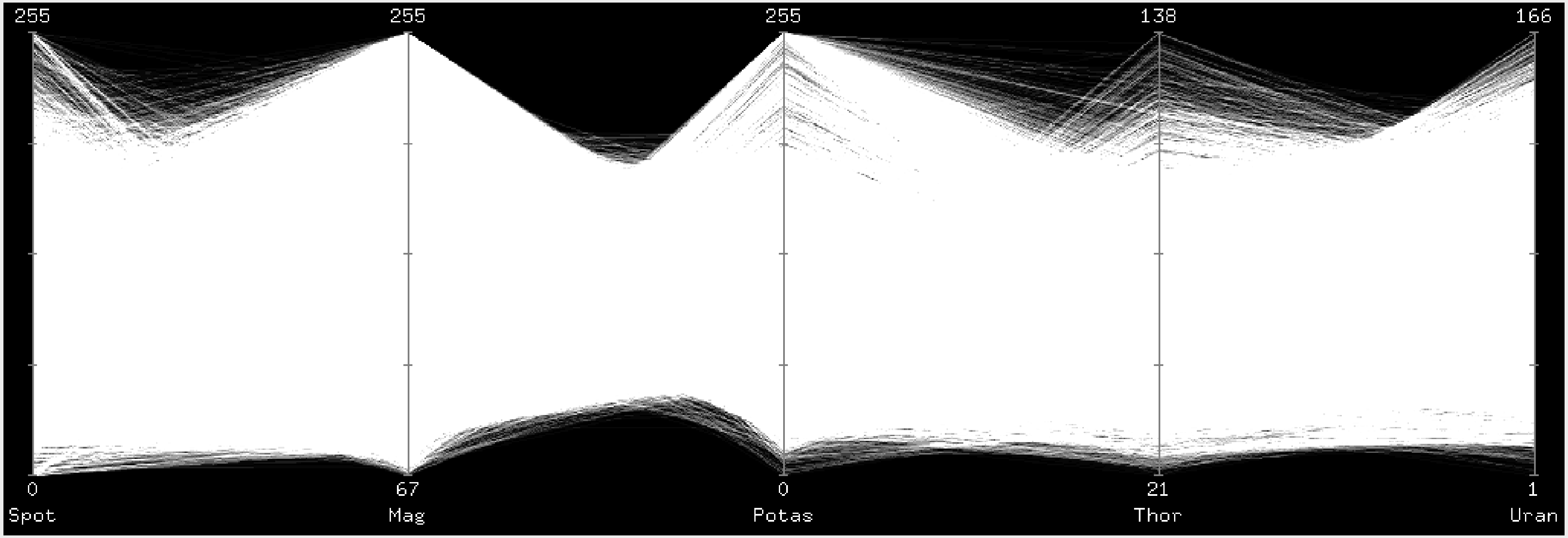
USING TRANSPARENCY

16.000 RECORDS IN PARALLEL COORDINATES



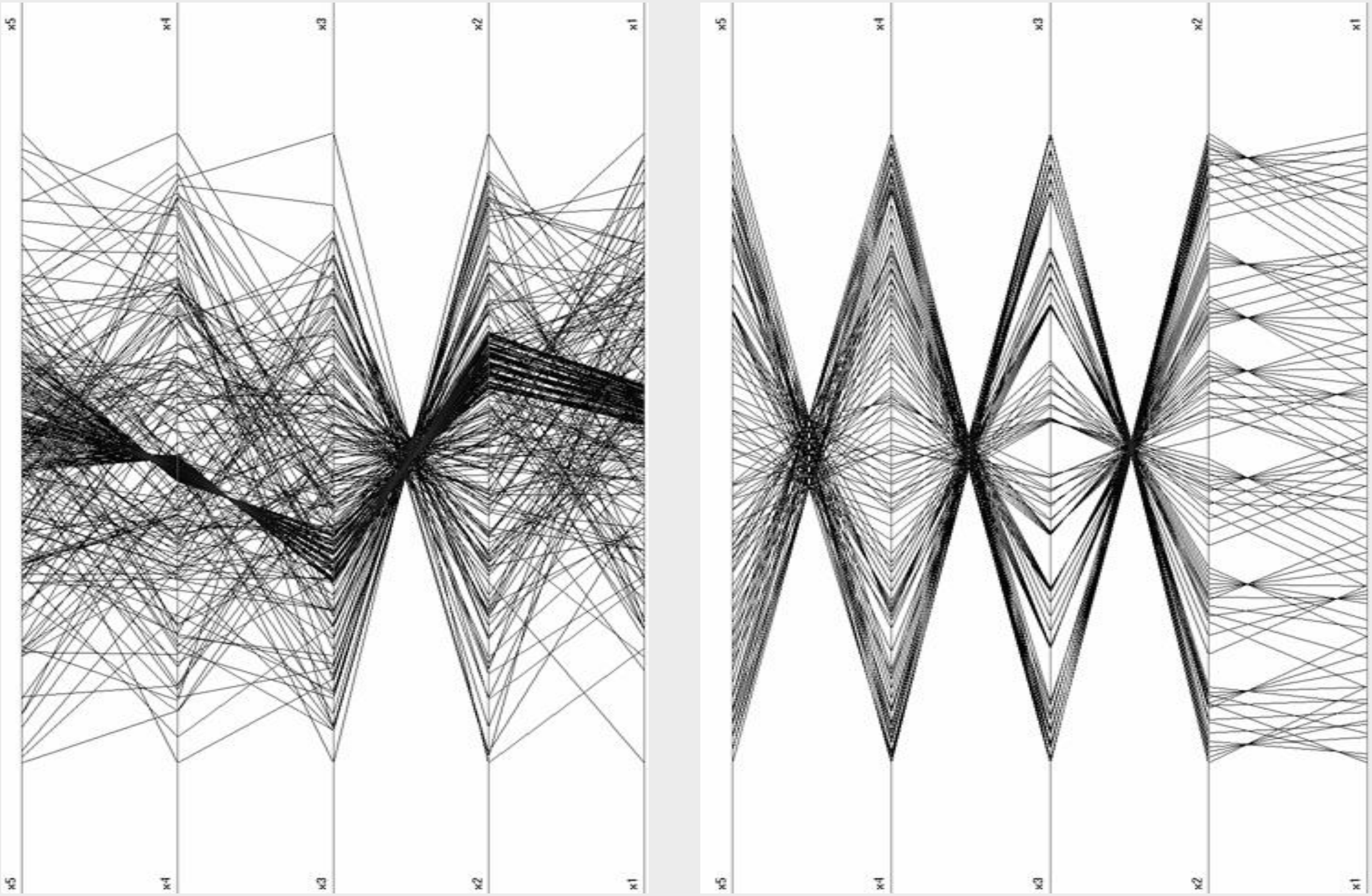
LIMITATIONS OF TRANSPARENCY

64.000 RECORDS IN PARALLEL COORDINATES



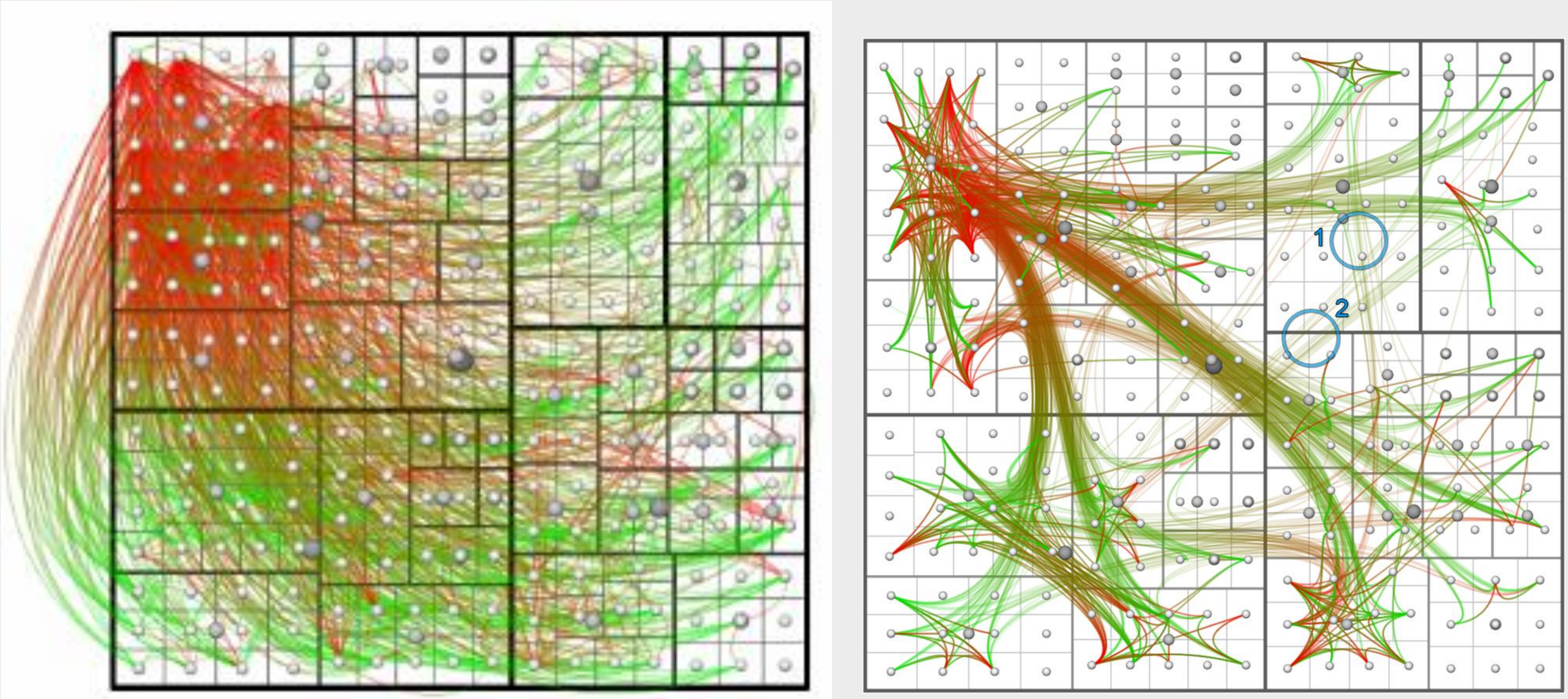
TRANSPARENCY ADDS 1-2 ORDERS OF MAGNITUDE TO CAPACITY

CHANGING THE PROJECTION



Parallel coordinates and the grand tour [Wegman]

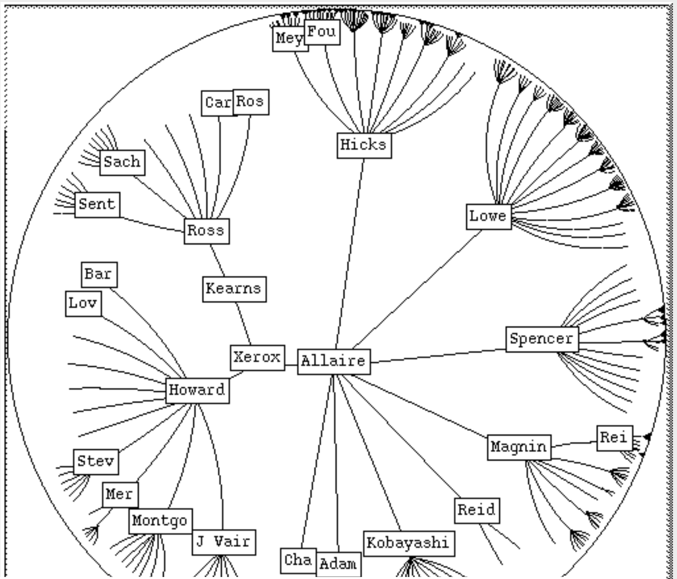
CHANGES IN GEOMETRY

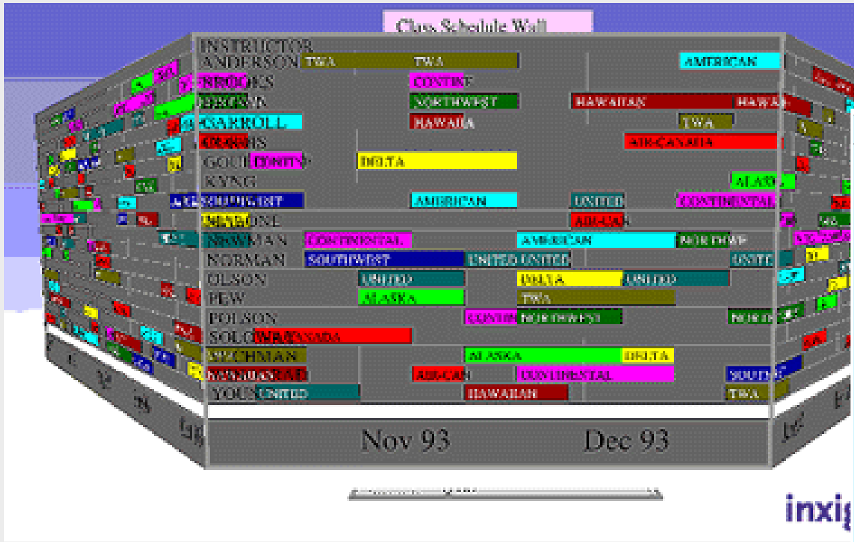


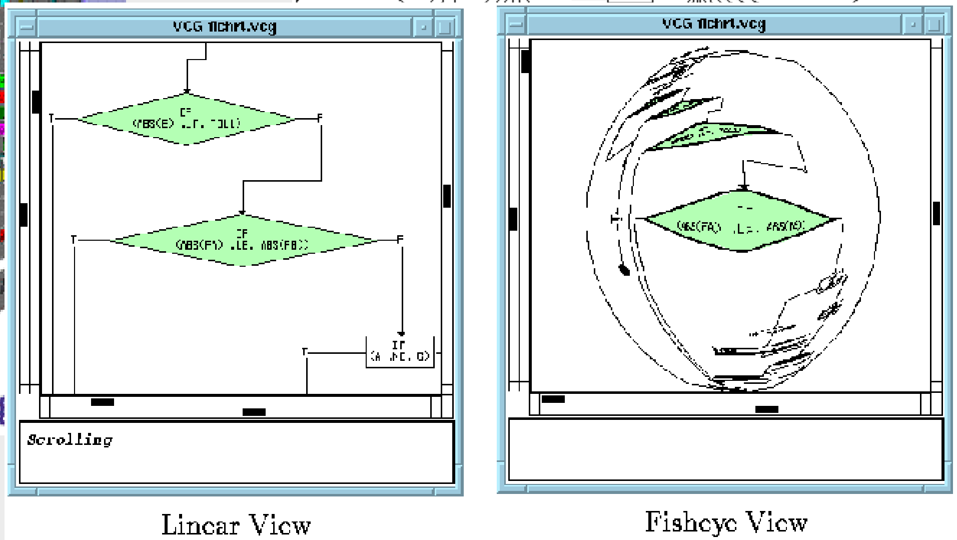
HIERARCHICAL EDGE BUNDLES [HOLTEN, 06]

DISTORTION OF THE SCREEN SPACE

FISHEYE VIEWS
HYPERBOLIC PROJECTION
PERSPECTIVE







Scrolling

Linear View

Fisheye View

FOCUS+CONTEXT PRINCIPLE

CONTEXT = ALL DATA, FOCUS = INTEREST

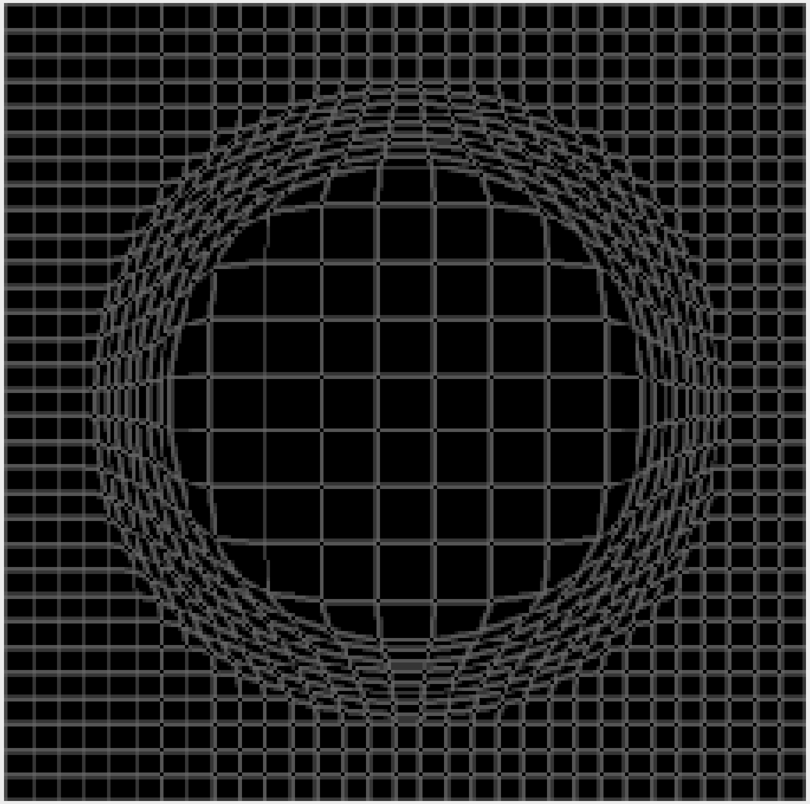
Goal:
DISPLAY FOCUS IN DETAIL WHILE STILL SHOWING THE CONTEXT

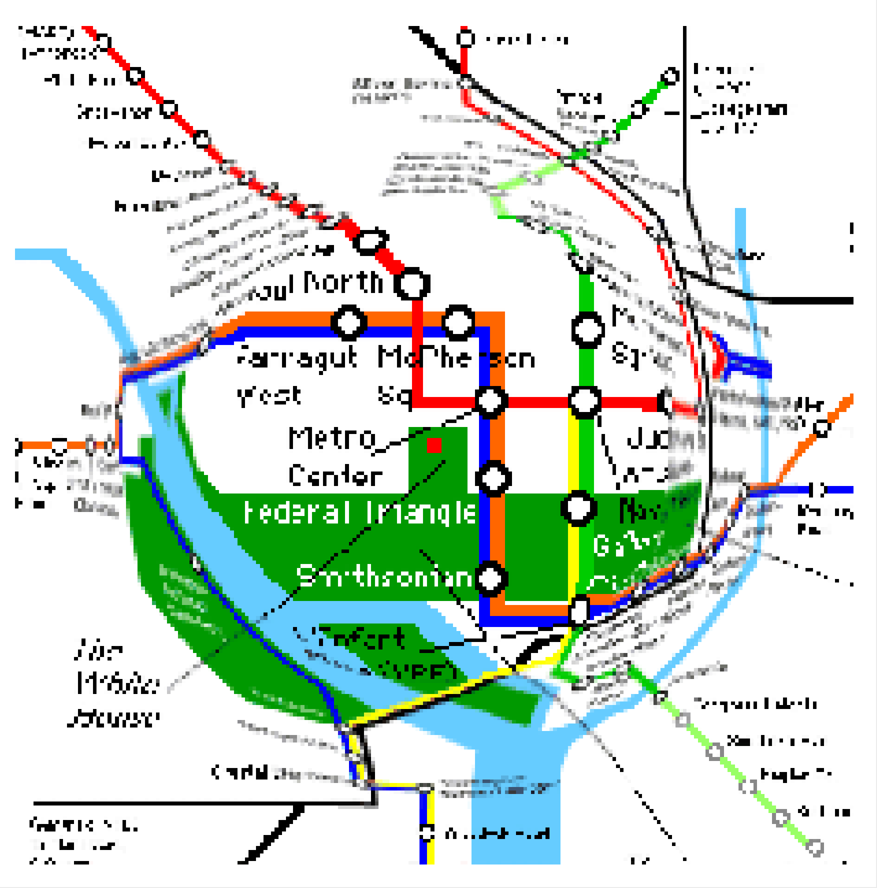
CONTEXT
displayed in lower details
taking up less screen space and less attention

FOCUS
high details, more screen space, more attention

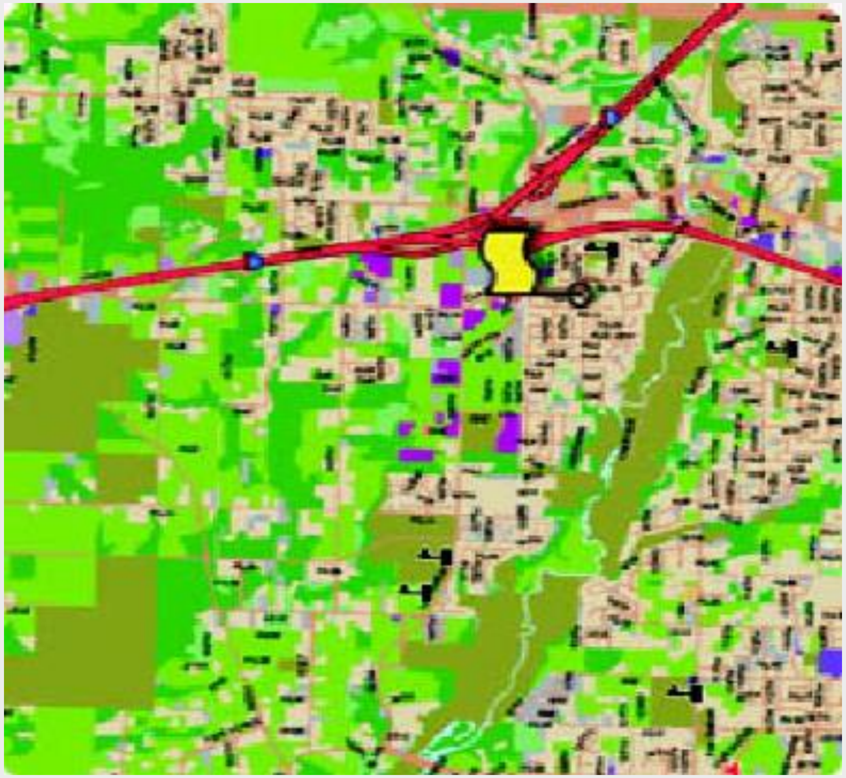
FOCUS+CONTEXT

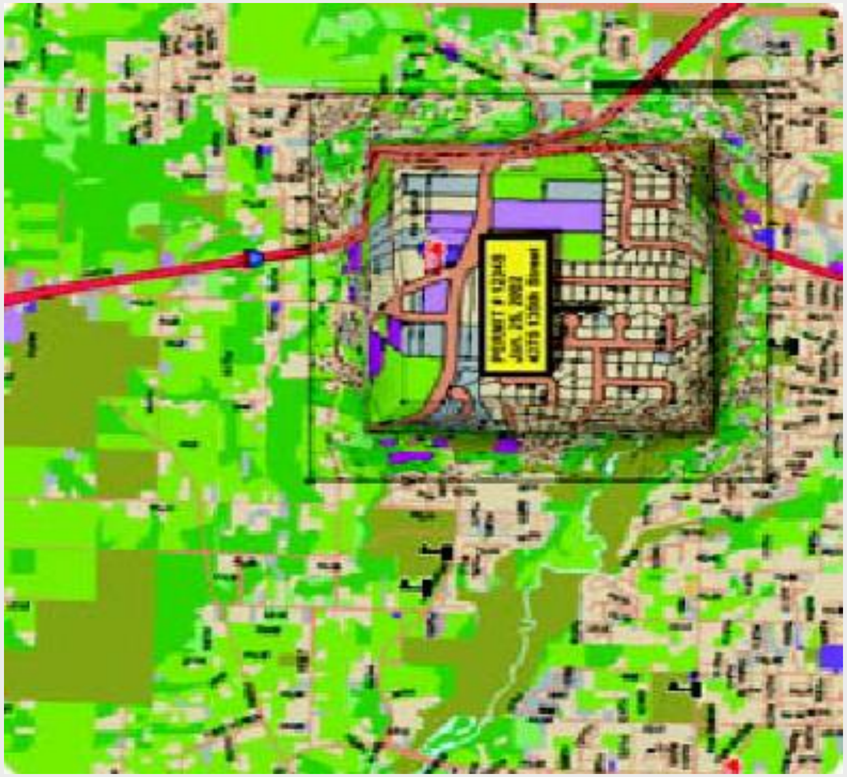
THE CONTEXT IS SUPPRESSED BY EITHER LESS SPACE OR LESS DETAILS





FOCUS+CONTEXT EXAMPLES

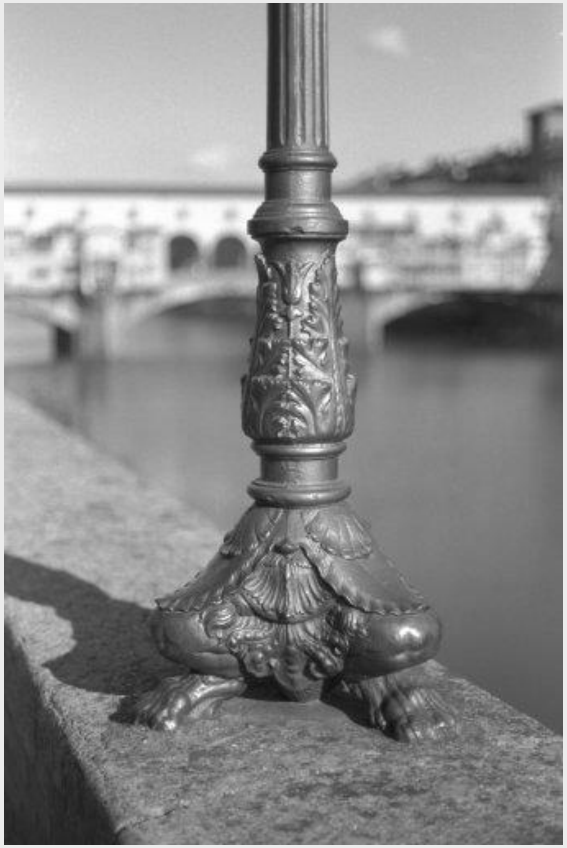


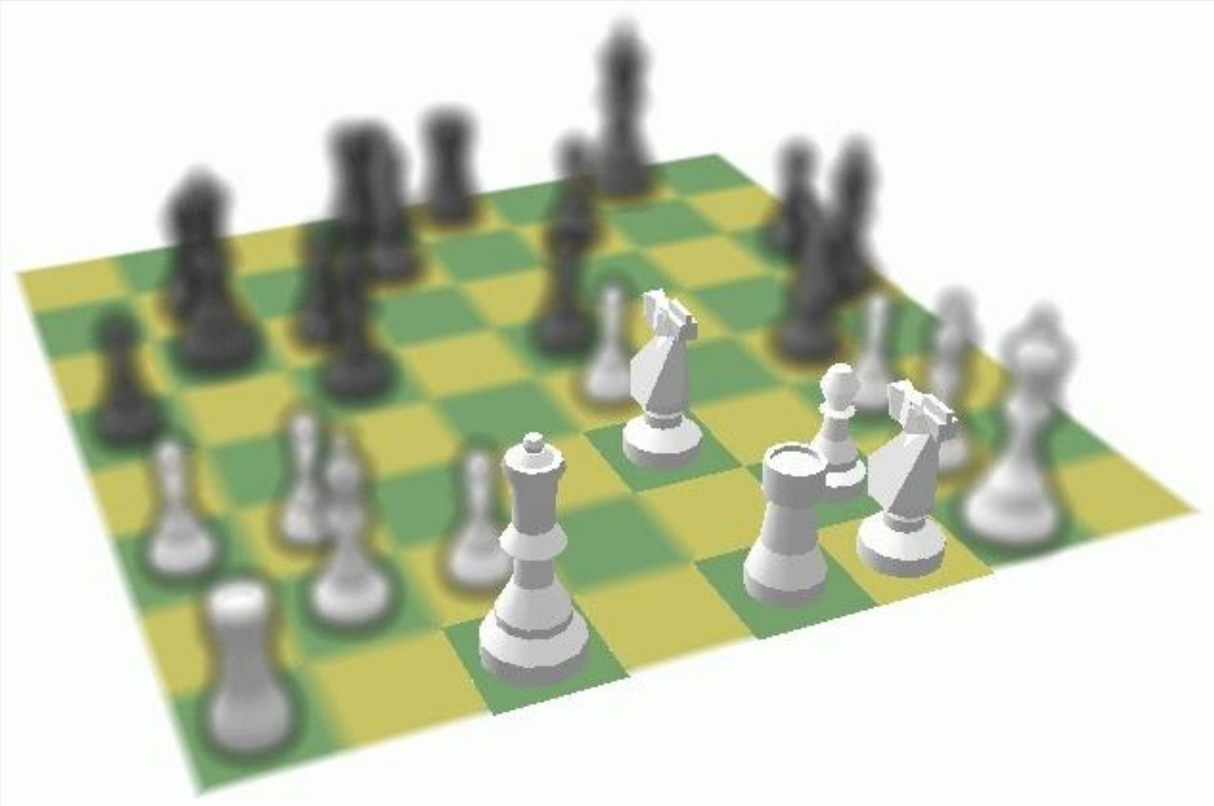


[WWW.IDELIX.COM]

FOCUS+CONTEXT EXAMPLES

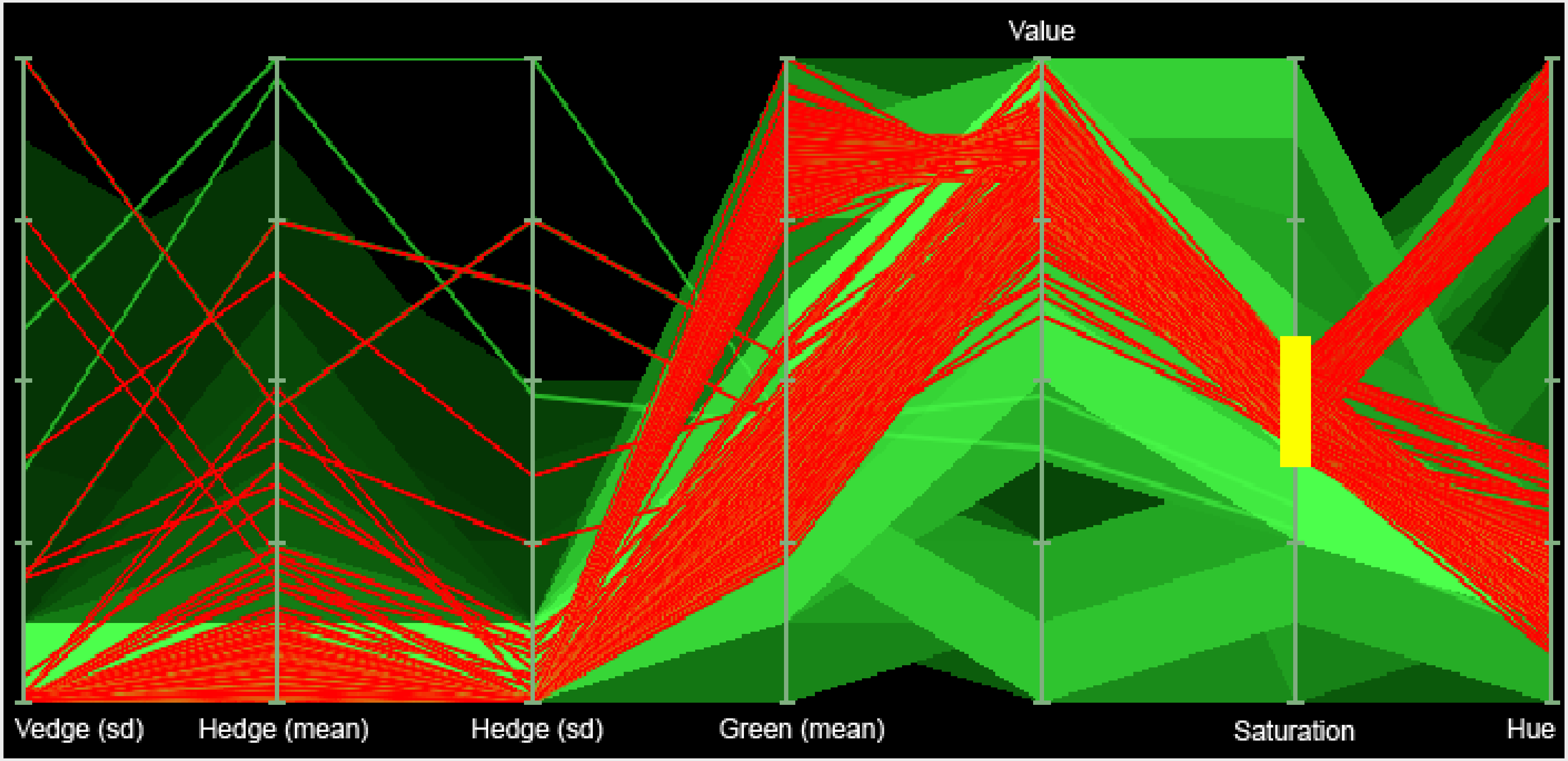
SEMANTIC DEPTH OF FIELD
[KOSARA, MIKSCH, HAUSER]





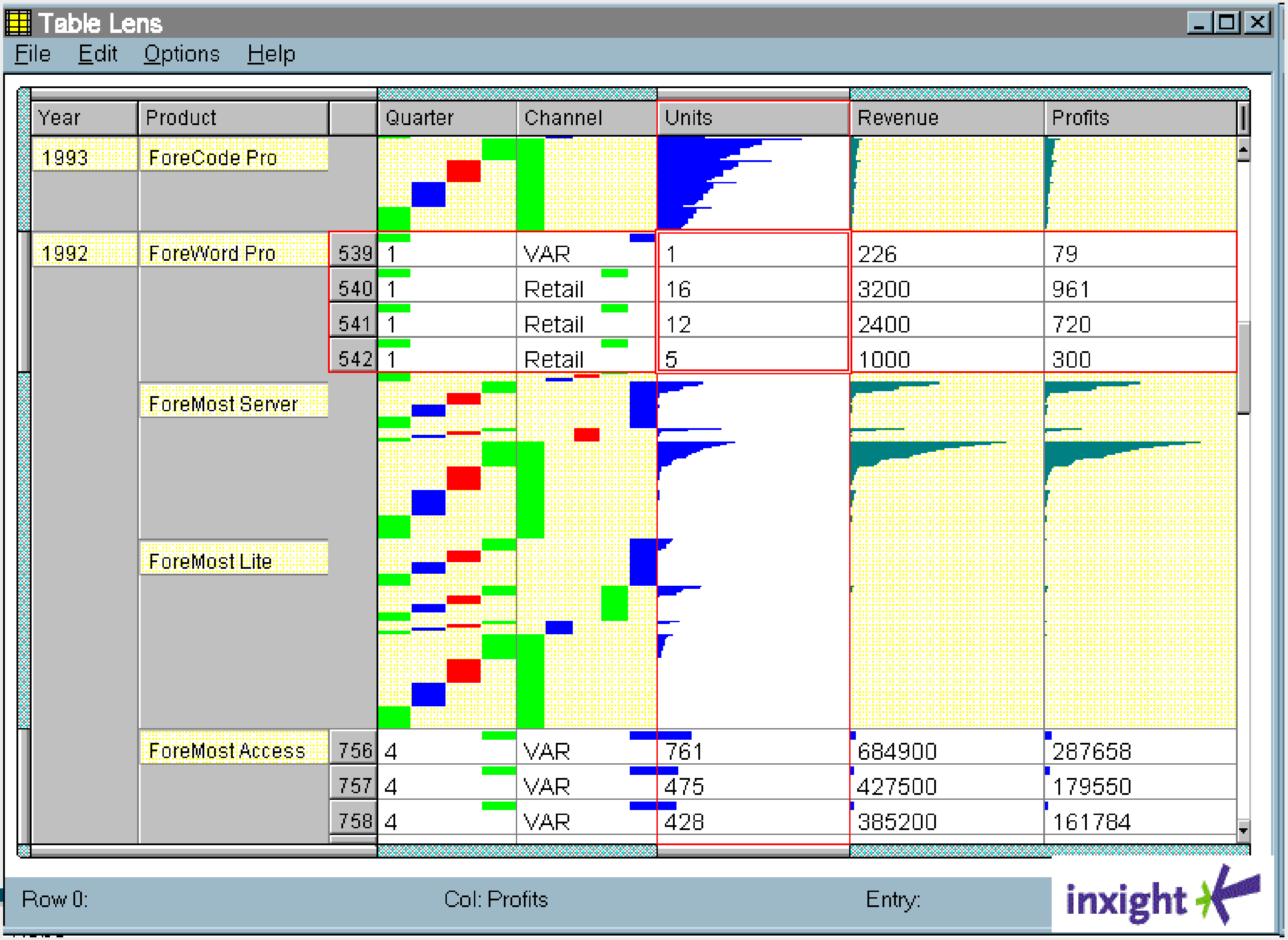
FOCUS+CONTEXT EXAMPLES

OUTLIER-PRESERVING F+C IN PAR.COORDS.



FOCUS+CONTEXT EXAMPLES

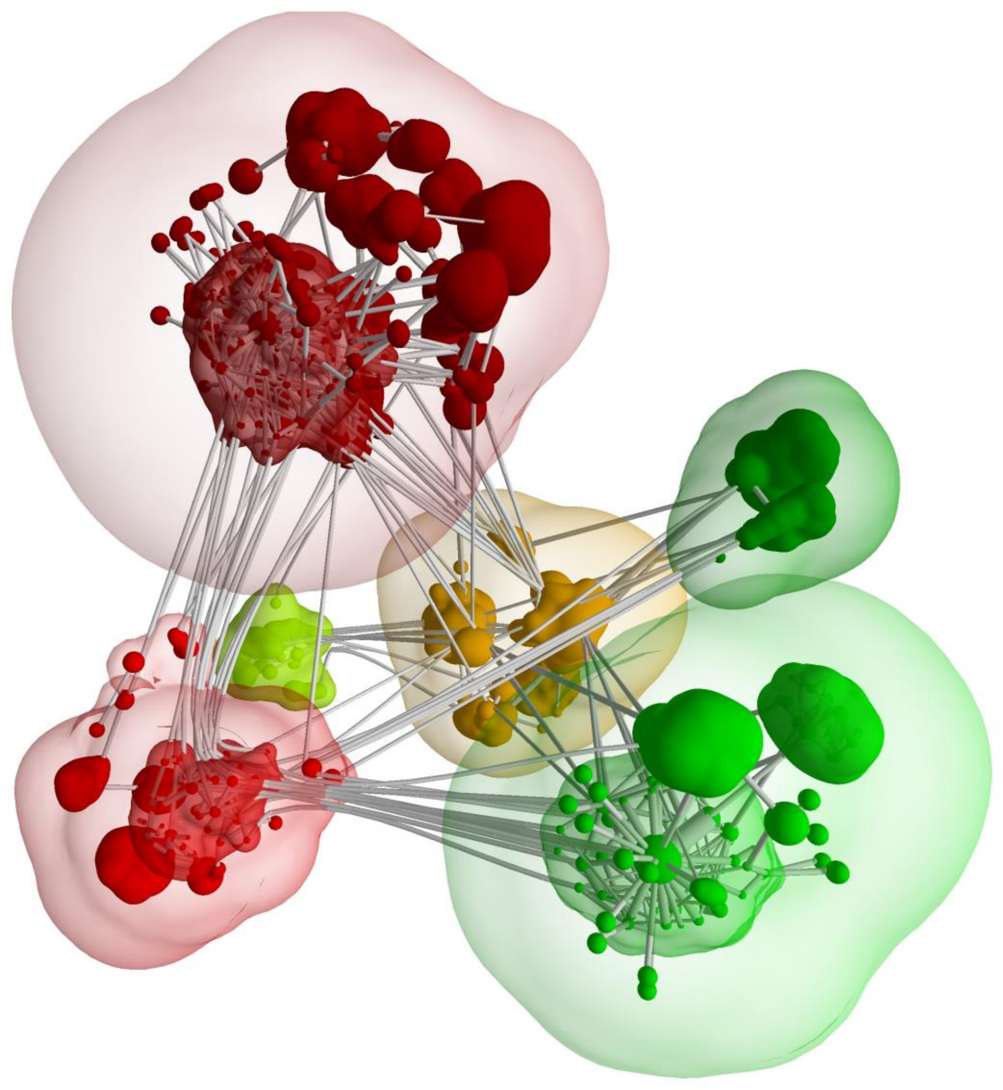
TABLE LENS

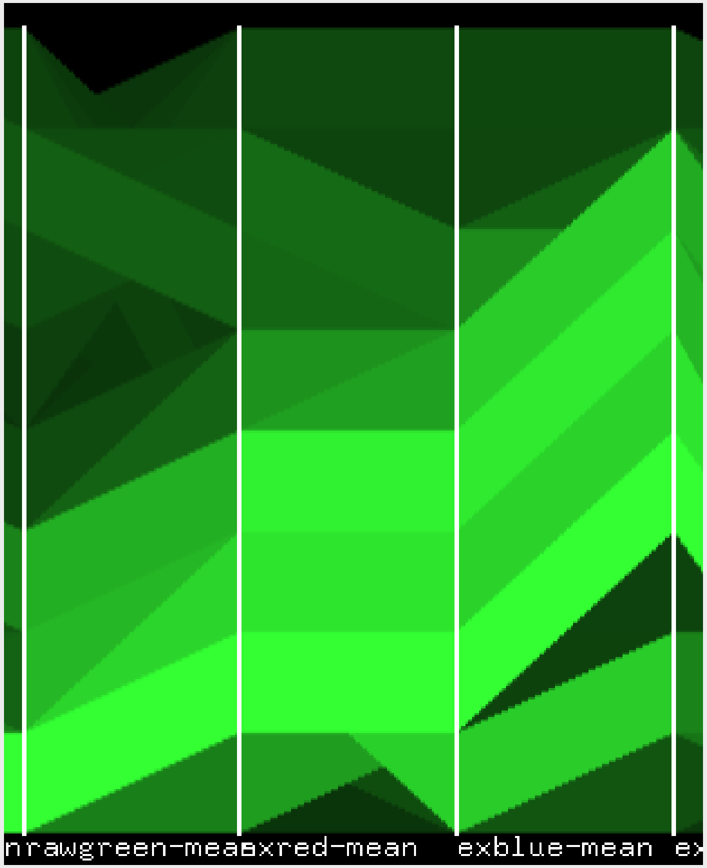


Year	Product		Quarter	Channel	Units	Revenue	Profits
1993	ForeCode Pro						
1992	ForeWord Pro	539	1	VAR	1	226	79
		540	1	Retail	16	3200	961
		541	1	Retail	12	2400	720
		542	1	Retail	5	1000	300
	ForeMost Server						
	ForeMost Lite						
	ForeMost Access	756	4	VAR	761	684900	287658
		757	4	VAR	475	427500	179550
		758	4	VAR	428	385200	161784

Row 0: Col: Profits Entry:

LEVEL OF DETAIL TECHNIQUES





SUMMARY

WHAT LARGE DATA CAUSES:

Occlusion, aggregation, bad interaction

HOW TO FIX IT:

Reduce data (sampling, clustering, aggregation)
Tweak the view (layout, mapping, shapes)
Distort the view (hyperbolic, fisheye, perspective)
Use different levels of detail (F+C, L.O.D)