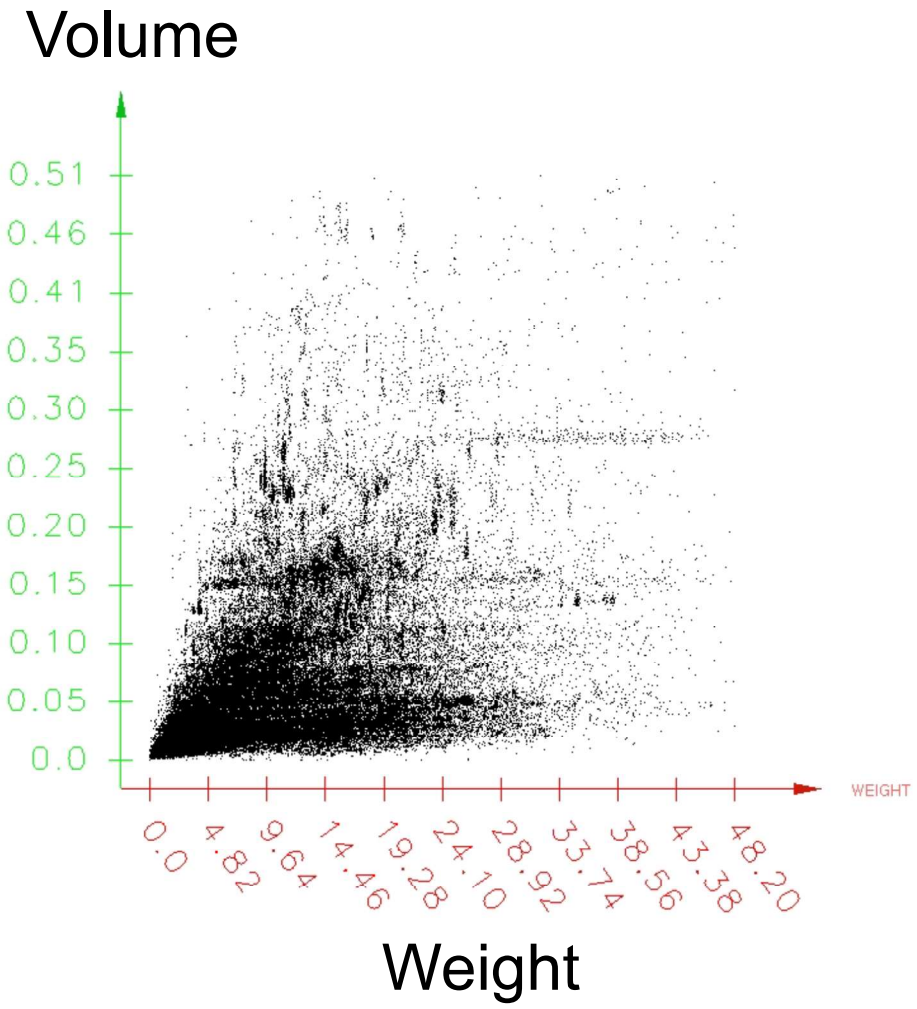
Improving 2D scatter plots effectiveness through sampling, displacement, and user perception

Enrico Bertini, Giuseppe Santucci
Dipartimento di Informatica e Sistemistica
University of Rome ”La Sapienza”

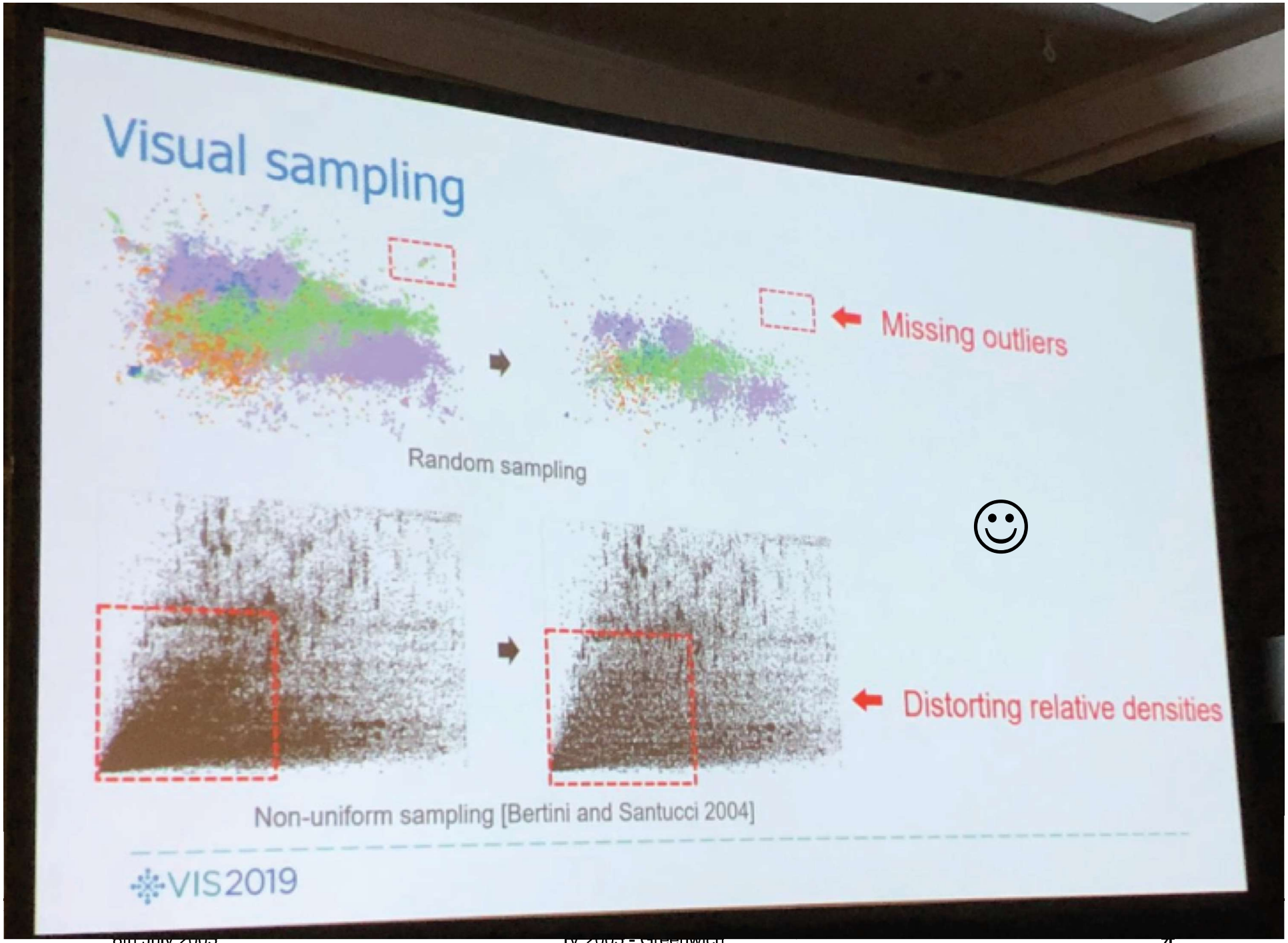
The clutter problem

- Because of **display limitations**, the **number of elements**, and **data distribution**, 2D scatter plot are affected by clutter:
  - displayed elements overlap

e.g., plotting mail parcels



- 160,000 items
- Overlapping pixels
- **Density differences** are lost, one of the most important visual features in a scatter plot, the building block for detecting:
  - clusters
  - trends
  - ...


Visual sampling
Missing outliers
Random sampling
Distorting relative densities
Non-uniform sampling [Bertini and Santucci 2004]
VIS2019

Most common techniques to deal with clutter

- Interactive commands (up to the user)
  - zoom, pan, etc.
- Mapping collisions to color (or to other visual attributes)
- Clustering (or other aggregation techniques)
- Pixel displacement
- Data sampling (automatic, non uniform)

Our approach

1. Provide a formal model to characterize and forecast clutter
2. Formalize density
3. Alter data density to preserve density differences
4. Perceptual studies to refine our approach

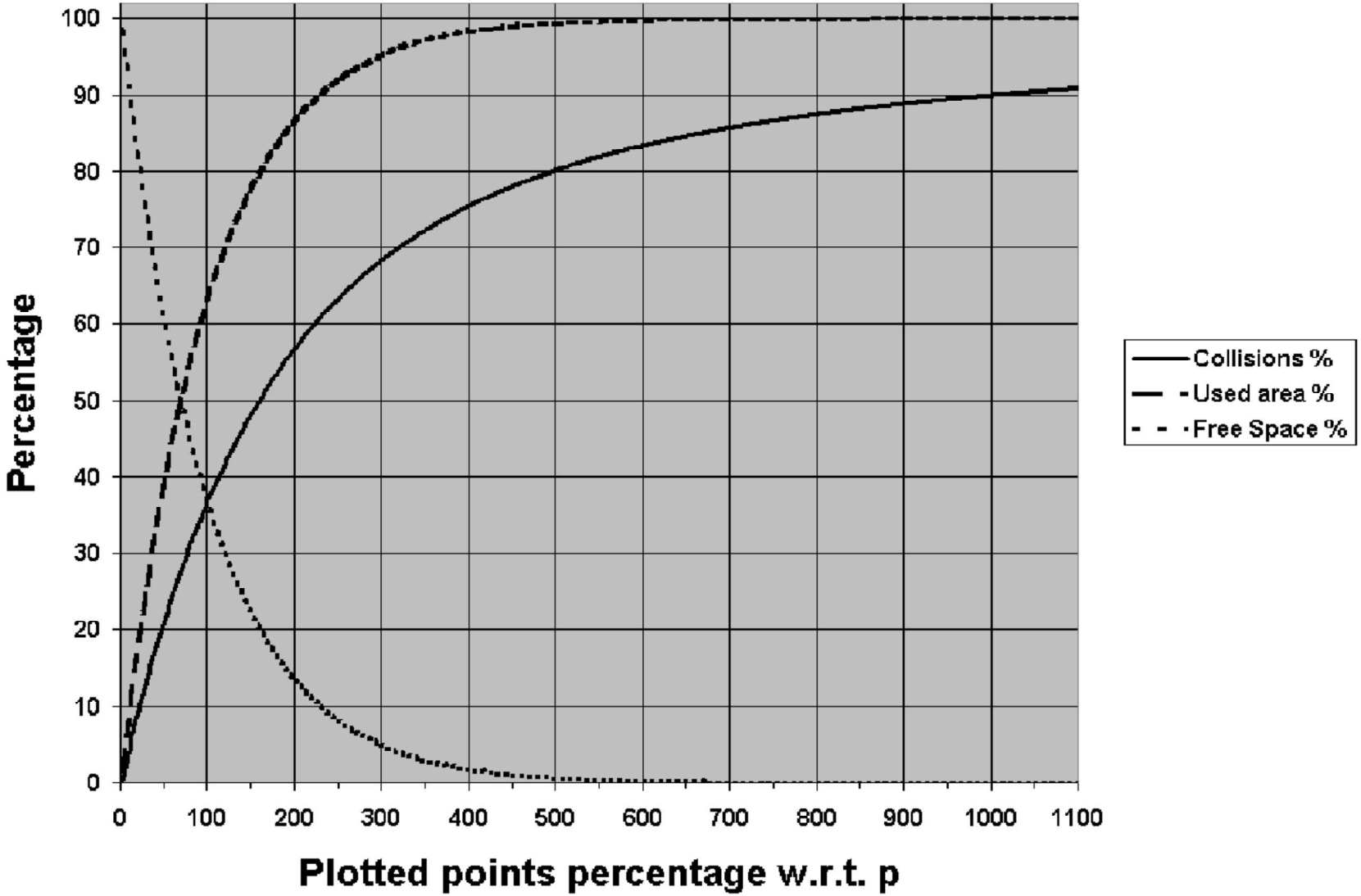
Our approach

1. Provide a formal model to **characterize** and **forecast** clutter
2. Formalize density
3. Alter data density to preserve density differences
4. Perceptual studies to refine our approach

1-Main assumptions

- 1 data element = 1 pixel
- a collision happens when two elements fall into the same pixel
  - rounding issues
  - same values
- we split the plane in little squares spanning **p** pixels that we call **sample areas**
  - typical dimension: 8x8 pixels, **p**=64

1-Overplotting behavior (forecasting clutter)



Our approach

1. Provide a formal model to characterize and forecast clutter
2. Formalize density
3. Alter data density to preserve density differences
4. Perceptual studies to refine our approach

2-Abstract and real space

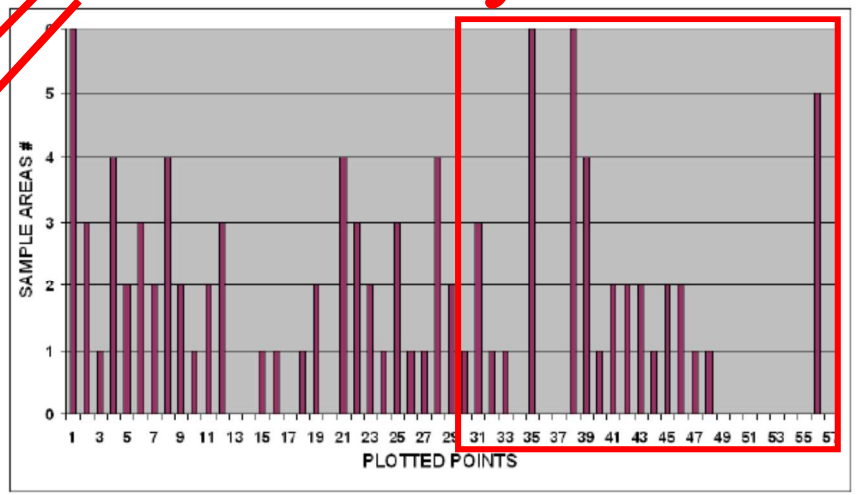
- we misure density in a continuous **abstract space** (sample areas measured in inches)
  - **Data Density**: number of data points plotted within a sample area
- we misure density in a discrete **real space** (sample areas measured in pixels)
  - **Represented Density**: number of active pixels within a sample area (values range from **0** to **p**)
- because of collisions we have that in a sample area: Represented Density<=Data Density

2- Example

Lost density differences!

Abstract space , 100 sample areas

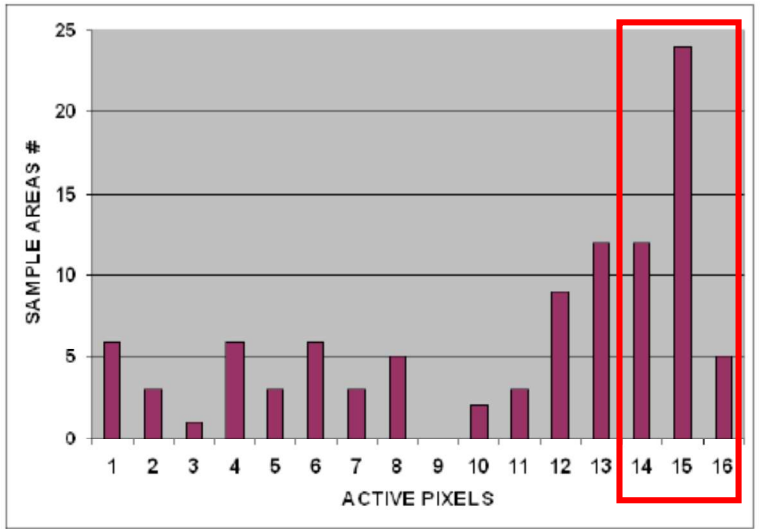
46	1	11	25	1	31	19	21	18	1
4	3	12	38	45	56	26	35	12	38
38	38	35	56	28	8	21	42	22	7
35	24	4	1	30	9	41	31	21	23
41	47	45	6	35	22	7	56	38	28
8	4	39	56	28	27	42	22	43	29
1	35	21	39	44	35	1	6	5	25
32	2	31	29	8	28	33	39	5	40
43	4	12	16	23	2	9	48	39	8
11	25	56	46	15	38	2	15	6	19



Data Density (DD) distribution

Real space , 100 4x4 sample areas

15	1	8	13	1	14	11	12	11	1
4	3	8	15	15	16	13	14	8	15
15	15	14	16	13	6	12	15	12	6
14	13	4	1	14	7	15	14	12	12
15	15	15	5	14	12	6	16	15	13
6	4	15	16	13	13	15	12	15	13
1	14	12	15	15	14	1	5	4	13
14	2	14	13	6	13	14	15	4	15
15	4	8	10	12	2	7	15	15	6
8	13	16	15	7	15	2	10	5	11



DD values from 30-56 mapped on RD values from14-16

Represented Density (RD) distribution

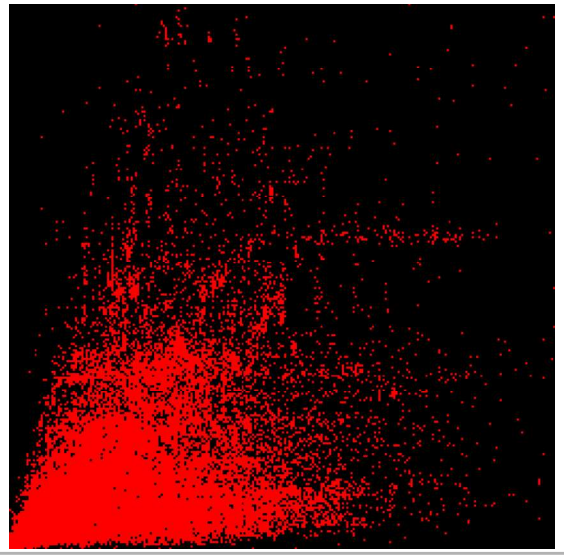
Our approach

1. Provide a formal model to characterize and forecast clutter
2. Formalize density
3. **Alter** data density to **preserve** density differences (This is a presentation issue)
4. Perceptual studies to refine our approach

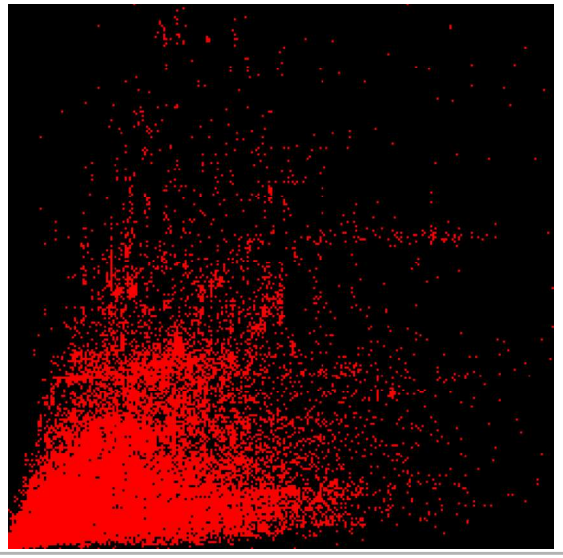
3-Sampling

- Sampling is effective but…
  - If the sampling is **too weak** high density areas can not be disambiguated
  - If the sampling is **too strong** low density areas disappear
  - in some cases we need to **increase** the represented density

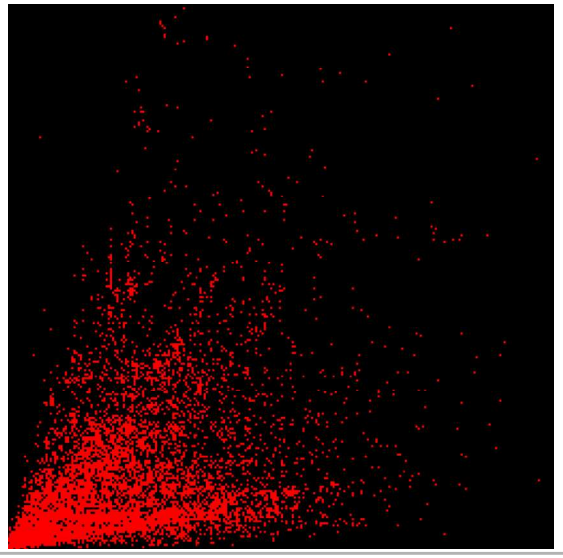
No sampling



Too weak (80%)



Too strong (20%)



3-Non-uniform sampling & displacement

- We treat in a different manner **each sample area** trying to preserve the difference in densities existing in the abstract space:
  - we **decrease** the represented density through **sampling**
  - we **increase** the represented density through **displacement**

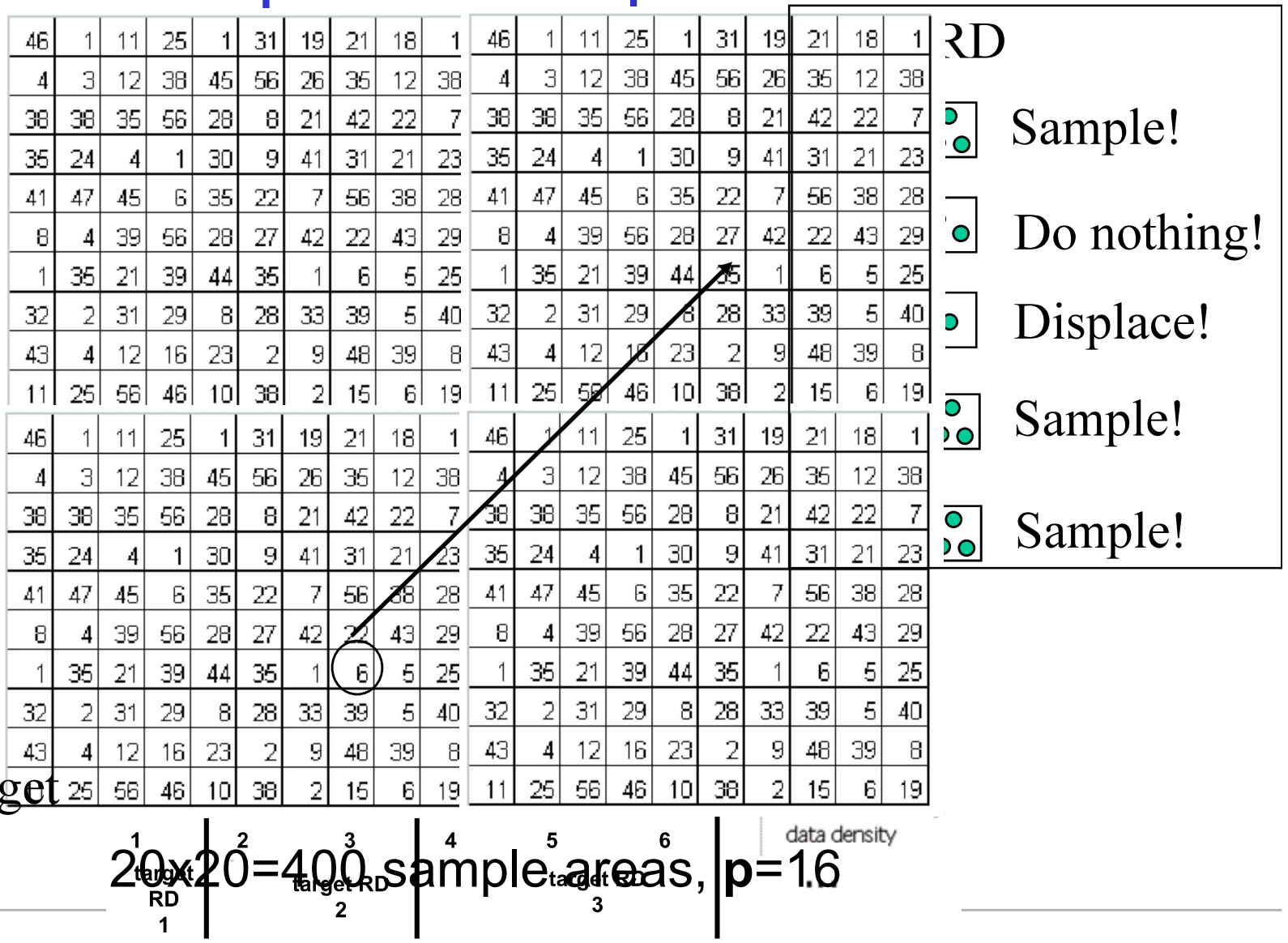
3-The algorithm

1. Split the data density distribution in **p** intervals $i_1,\ldots,i_p$ in a way that each interval contains the **same number** of sample areas (**p** is the number of different represented densities 4x4->16)
2. Sample or displace the data in the interval i_k as much as to achieve a represented density equal to k

3 - Example: 800x800 pixels split in 4x4 pixels sample areas

we split the density distribution 16 intervals each one containing 400/16=25 sample area

sample or displace to reach the target RD

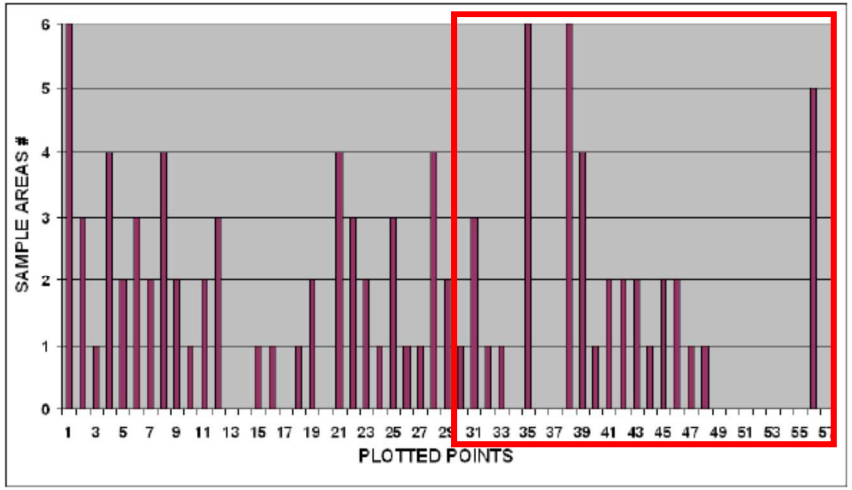


20x20=400 sample areas, p=1.6

3-Before applying the algorithm

Abstract space , 100 sample areas

46	1	11	25	1	31	19	21	18	1
4	3	12	38	45	56	26	35	12	38
38	38	35	56	28	8	21	42	22	7
35	24	4	1	30	9	41	31	21	23
41	47	45	6	35	22	7	56	38	28
8	4	39	56	28	27	42	22	43	29
1	35	21	39	44	35	1	6	5	25
32	2	31	29	8	28	33	39	5	40
43	4	12	16	23	2	9	48	39	8
11	25	56	46	10	38	2	15	6	19

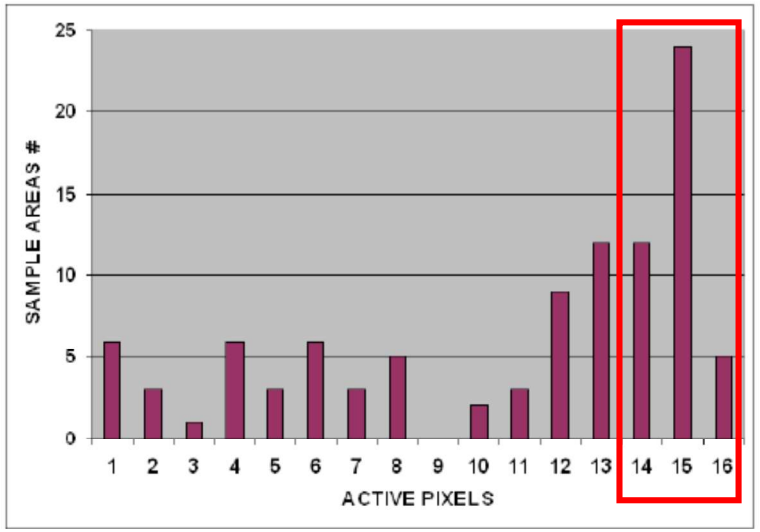


Data Density (DD) distribution

DD values from 30-56 mapped on RD values from14-16

Real space , 100 4x4 sample areas

15	1	8	13	1	14	11	12	11	1
4	3	8	15	15	16	13	14	8	15
15	15	14	16	13	6	12	15	12	6
14	13	4	1	14	7	15	14	12	12
15	15	15	5	14	12	6	16	15	13
6	4	15	16	13	13	15	12	15	13
1	14	12	15	15	14	1	5	4	13
14	2	14	13	6	13	14	15	4	15
15	4	8	10	12	2	7	15	15	6
8	13	16	15	7	15	2	10	5	11

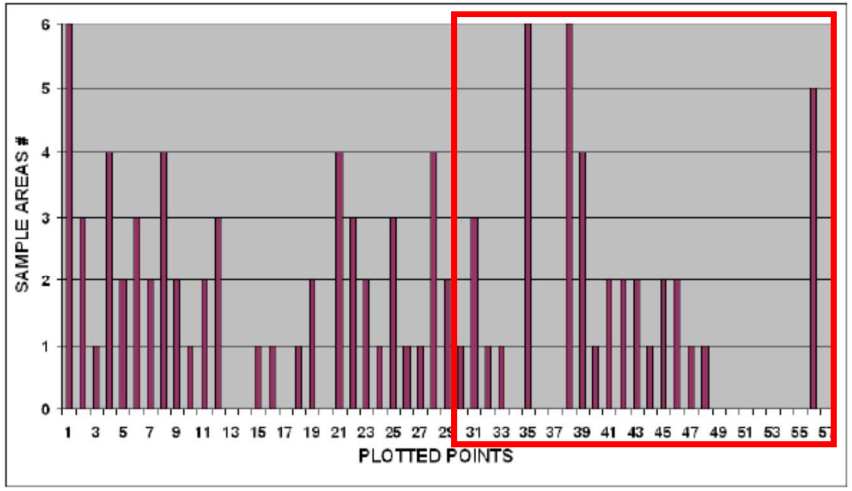


Represented Density (RD) distribution

3-After applying the algorithm

Abstract space , 100 sample areas

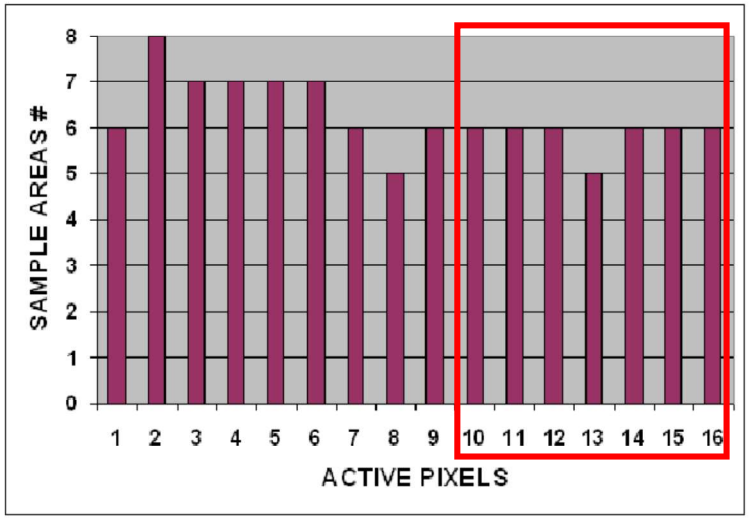
46	1	11	25	1	31	19	21	18	1
4	3	12	38	45	56	26	35	12	38
38	38	35	56	28	8	21	42	22	7
35	24	4	1	30	9	41	31	21	23
41	47	45	6	35	22	7	56	38	28
8	4	39	56	28	27	42	22	43	29
1	35	21	39	44	35	1	6	5	25
32	2	31	29	8	28	33	39	5	40
43	4	12	16	23	2	9	48	39	8
11	25	56	46	10	38	2	15	6	19



Data Density (DD) distribution

Real space , 100 4x4 sample areas

15	1	5	8	1	10	6	6	6	1
2	2	5	12	15	16	8	11	5	12
12	12	11	16	9	4	6	14	7	3
11	7	2	1	10	4	14	10	6	7
14	15	15	3	11	7	3	16	12	9
4	2	13	16	9	8	14	7	14	9
1	11	6	13	15	11	1	3	3	8
10	2	10	9	4	9	10	13	3	13
14	2	5	5	7	2	4	16	13	4
5	8	16	15	4	12	2	5	3	6



DD values from 30-56 mapped on RD values from 10-16

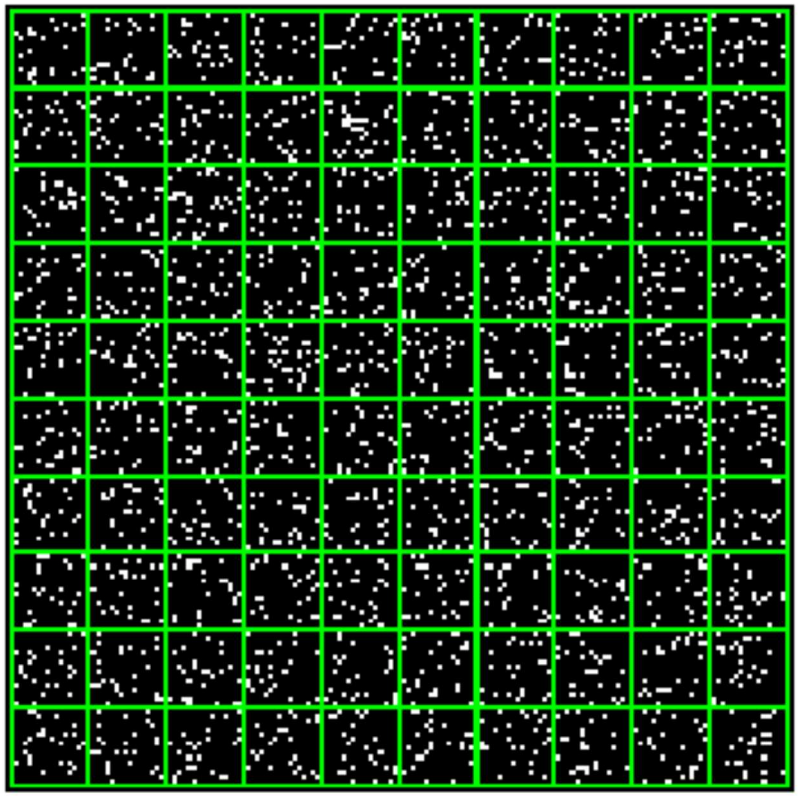
NEW

Represented Density (RD) distribution

Our approach

1. Provide a formal model to characterize and forecast clutter
2. Formalize density
3. Alter data density to preserve density differences
4. Perceptual studies to refine our approach

4 - Are numerical differences adequate?



100 sample areas

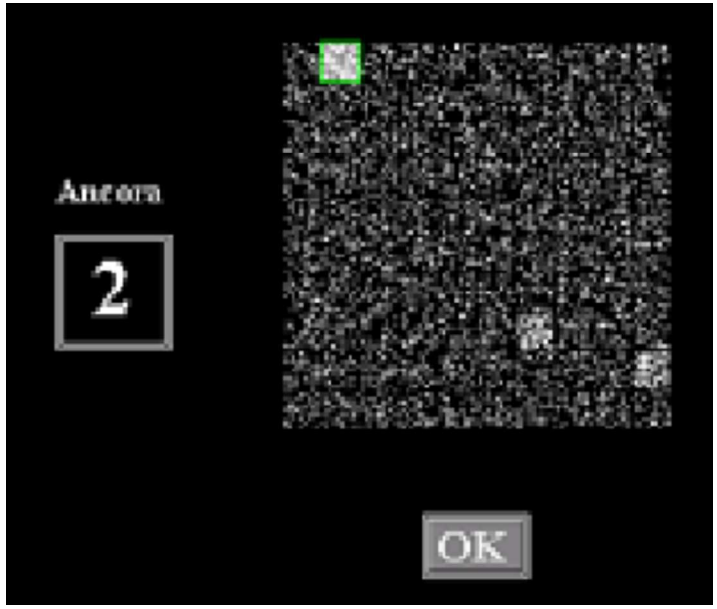
97 contain 25 pixels

3 contains 38 pixels

Which ones ?

4 - Perceptual studies: the user test

- What is the smallest difference in pixels between two sample areas that produce the perception of density difference? JND ?
- Generic step:
  - images with 100 sample areas
  - 97 with the same number of pixels (basis)
  - 3 filled with extra (delta) pixels
  - the user has to recognize the three more dense areas



- repeat for different basis and deltas

4 - Perceptual studies: results

Basis\Delta	D1	RP1	D2	RP2	D3	RP3	D4	RP4	D5	RP5	DMIN
5	70	53	90	73	110	89	130	93	150	98	**87**
8	60	42	80	69	100	84	120	93	140	95.5	**81**
10	55	62	65	77	75	82	85	92	95	97	**60**
20	35	41	40	64	45	70	50	77	55	87	**45**
30	30	62	35	56	40	74	45	95	40	97	**39**
40	26	67	30	77	34	85	38	90	42	100	**27**
50	20	59	22.5	79	25	77	27.5	92	30	95	**21**
60	22	72	25	92	28	100	31	97	34	100	**22**
70	12	64	13.5	67	15	73	16.5	77	18	90	**14**
80	10	70	11.5	87	13	95	14.5	97	16	97	**10**
90	6	77	7	92	8	100	9	97	10	100	**6**

all figures are percentages

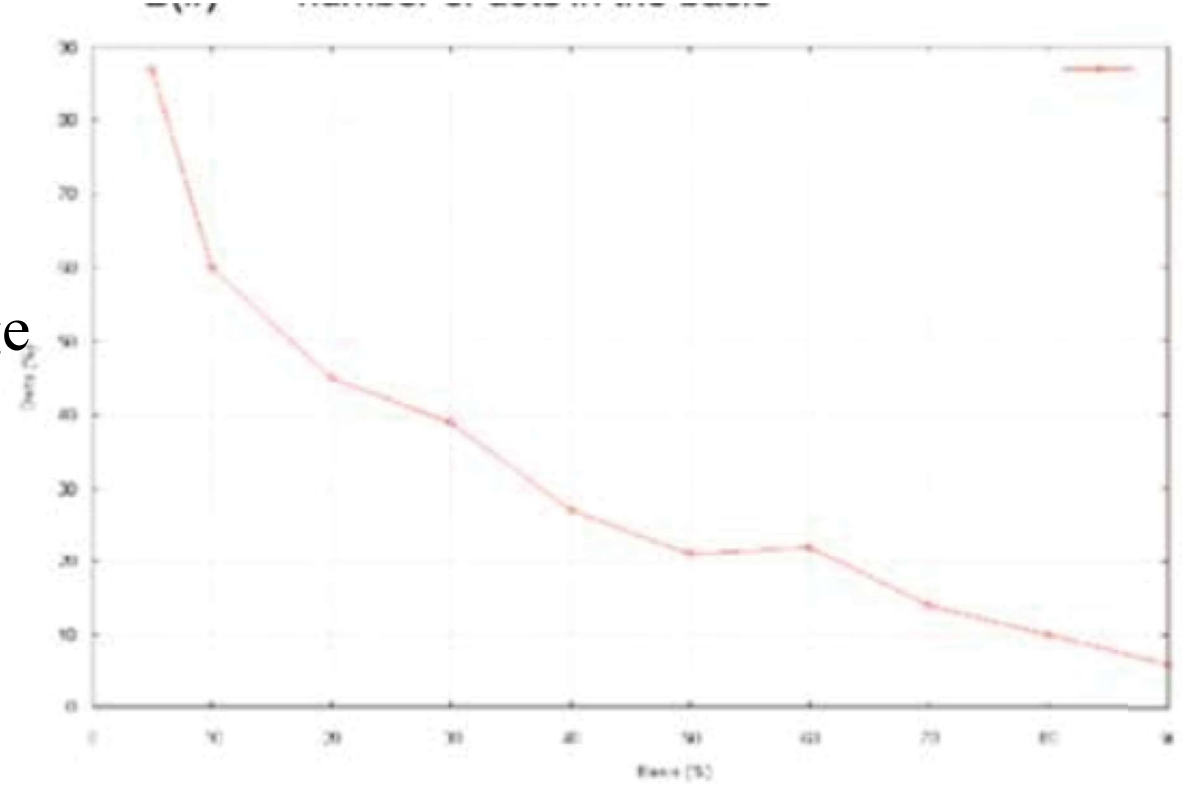
D1..D5 delta increments applied during the test

RP1..RP5 corresponding percentage of people recognizing correctly the three denser sample areas

Dmin minimum delta needed to guarantee a perception of density difference (70%)

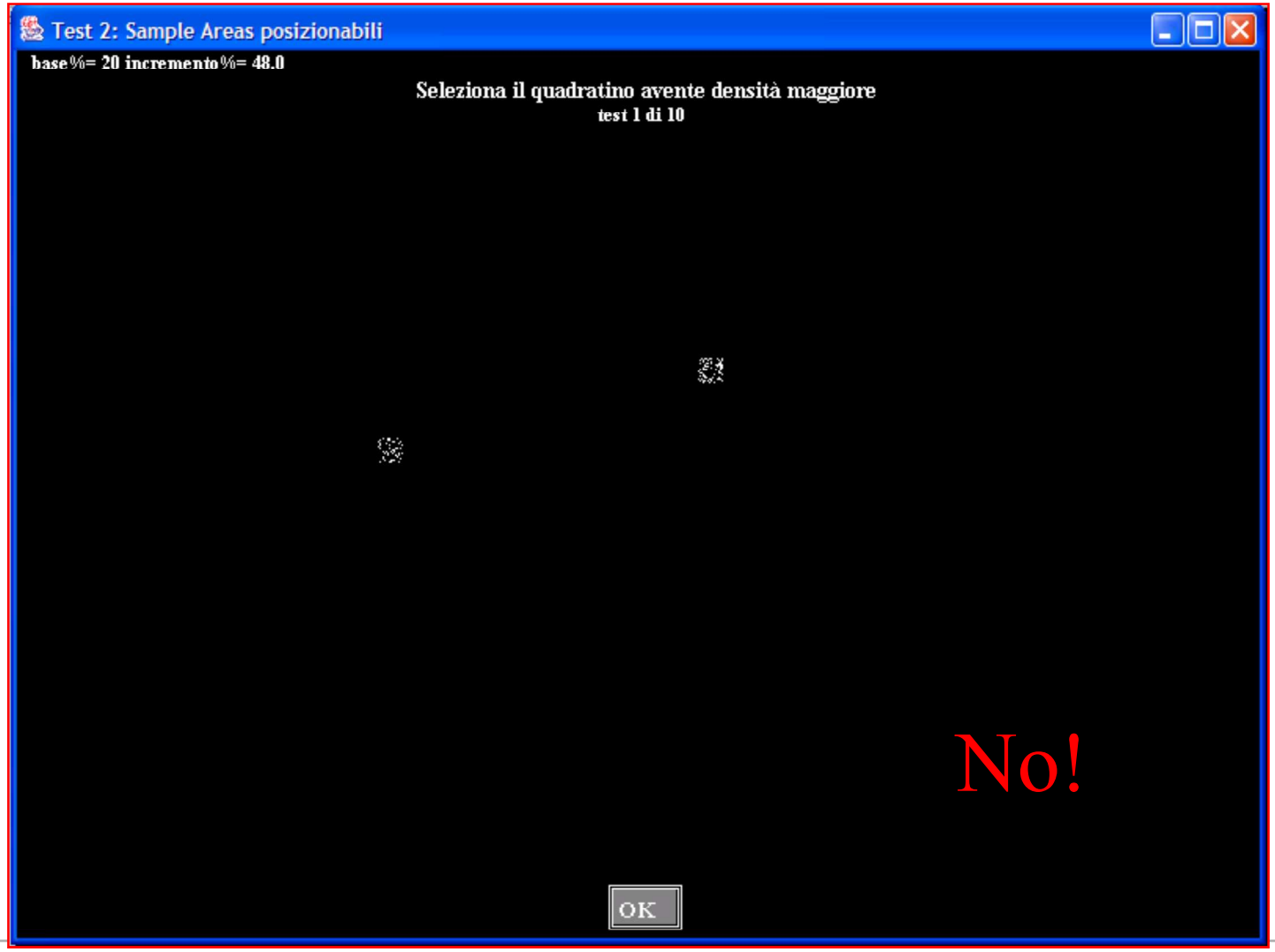
Perceptual perception of JND in densities

percentage



base

4 - Perceptual studies: does the distance matters?

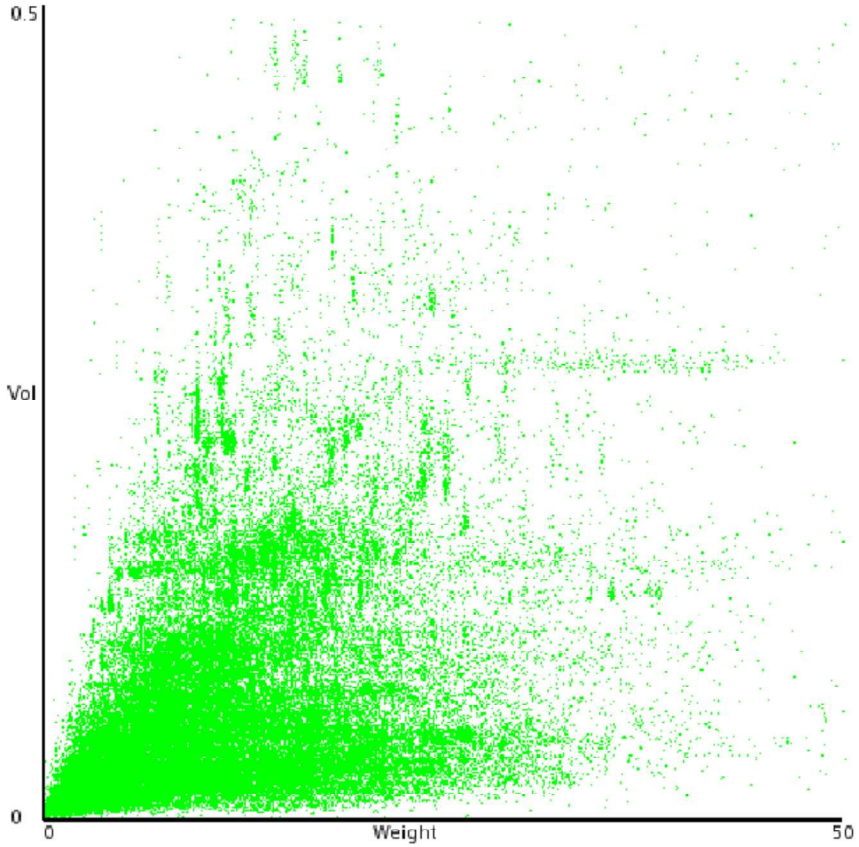


4 - Perceptual studies: tune algorithm

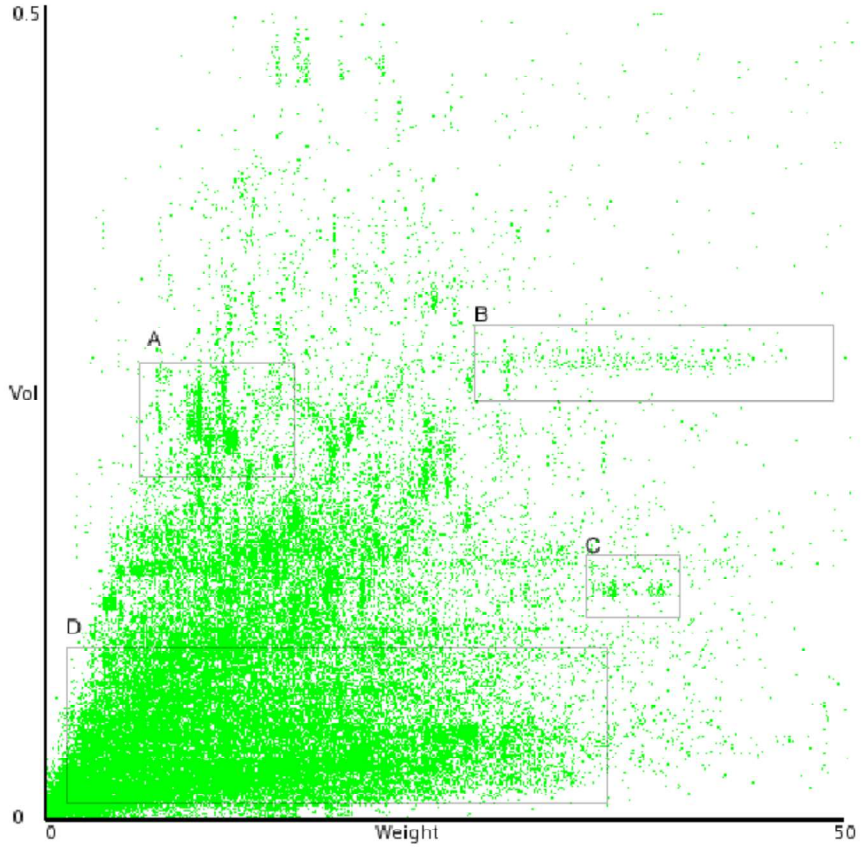
- we use a subset of the available represented density
- for 8x8 sample areas we use 14 out of 64 represented densities
- JND!!!

Represented Density	Perceptual Density
1	1
2, 3	2
4, 5, 6	4
7,8,9,10	7
11,12,13,14,15,16	11
17,18,19,20,21,22,23	17
24,25,26,27,28,29,30,31	24
32,33,34,35,36,37,38	32
39,40,41,42,43,44,45,46	39
47,48,49,50,51,52	47
53,54,55,56,57	53
58,59,60	58
61,62,63	61
64	64

Last but non least :does our technique visually works?



(a) original visualization



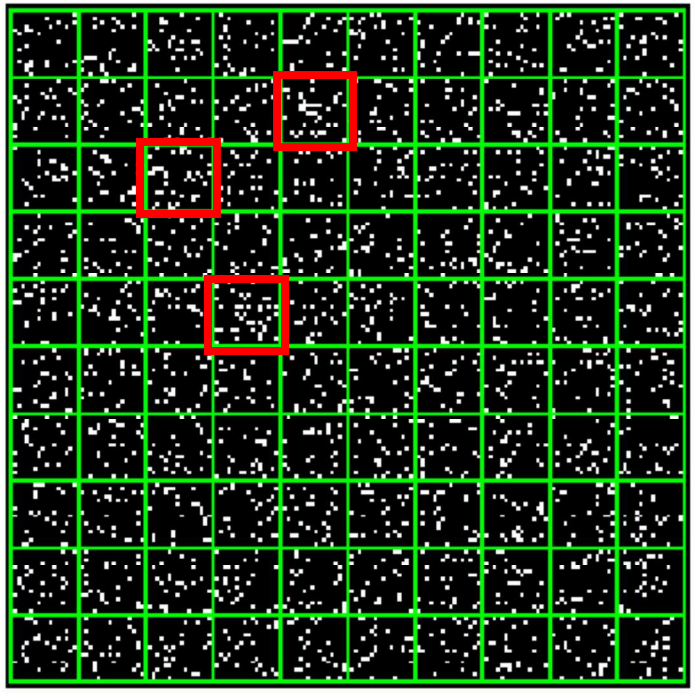
(b) decluttered visualization

Conclusions

- Clutter affects 2D scatter plots
- We developed a model to
  - measure and forecast clutter
  - characterize density
- Using this model we applied a non uniform sampling technique plus displacement preserving **perceptual** density differences

That's all (folks), thanks…

- Questions?



- For people still warring about the three denser areas....