Protecting patient privacy while releasing medical data for research

G. Xiang, J. O'Rawe, V. Kreinovich, J. Hajagos, S. Ferson

6th International Workshop on Reliable Engineering Computing, Chicago, Illinois, 27 May 2014

Rich data sets

- A lot of medical data is being collected
  - NIH, CDC, FDA, OSHA, EPA
  - Insurance companies
  - Hospitals, universities, clinics
  - Local health departments
  - Prison systems, cruise lines
- With big data comes big responsibility, Peter

Medical research

- These data sets would be extremely useful if they could be released to medical researchers
- But they can't be released
  - Health Information Privacy Protection Act
  - Americans with Disabilities Act
- Legal liability inhibits data custodians
  - Even if they could release, they often won't

Protecting privacy in data

- All we need to do is *anonymize* the data
- Just remove identifying details…
  - Name, social security number, address, GPS, etc.
- Unfortunately, this isn't nearly sufficient
- Just as Peter Parker discovers, doing the right thing is harder than one would think

Re-identification

- If you give me just
  - gender
  - zip code and
  - date of birth

  I can identify 63% of all Americans (Golle 2006)

- These 3 details often appear in open health archives, and are widely distributed

Many tools

- There are many sophisticated methods available for re-identification attacks
- There are many people willing to try
  - Hackers, extortionists
  - Drug companies, AARP
- Big health data is a big target

Statistical disclosure control

- Cell suppression
- Data perturbation
- Data swapping
- Replace or augment with synthetic data

Fabricates data

Alters patterns in data

Statistical disclosure control

- Variously effective at protecting privacy
- Substantial cost to utility of the data
- Information truthfulness not well preserved
- Alters statistical patterns in the data
  - Destroying some patterns
  - Amplifying others
- Produces unreliable data sets

Generalization-based anonymization

- Remove explicit identifiers
- Remaining attributes are used to re-identify
- Modify the data set by generalizing values so re-identification gets harder
  - Numbers replaced by intervals that include them
  - Categories widened to broader categories

16 ➔ [10,20]

high school student ➔ student

Privacy protection schemes

- *k*-anonymity (Sweeney 2002)
  - Every record in released data is indistinguishable from at least $k-1$ others with respect to every set of quasi-identifier attributes
  - The set of indistinguishable records is called an equivalence class
- Conceptually simple
- Many efficient algorithms

Limitation of *k*-anonymity

- Sensitive information can be revealed due to a lack of diversity in the equivalence class
  - E.g., if everyone in an equivalence class has the same disease, then sensitive information (disease status of anyone in the class) will be disclosed

Enhancements to *k*-anonymity

- *l*-diversity
  - At least *l* values for the sensitive attribute in the equivalence class
  - More protection ➔ bigger equivalence classes
- Various other enhancements suggested
  - *t*-closeness
  - (k, e)-anonymity
  - *m*-confidentiality
  - (α, k)-anonymity
  - (c, k)-safety

Many ways to skin a cat

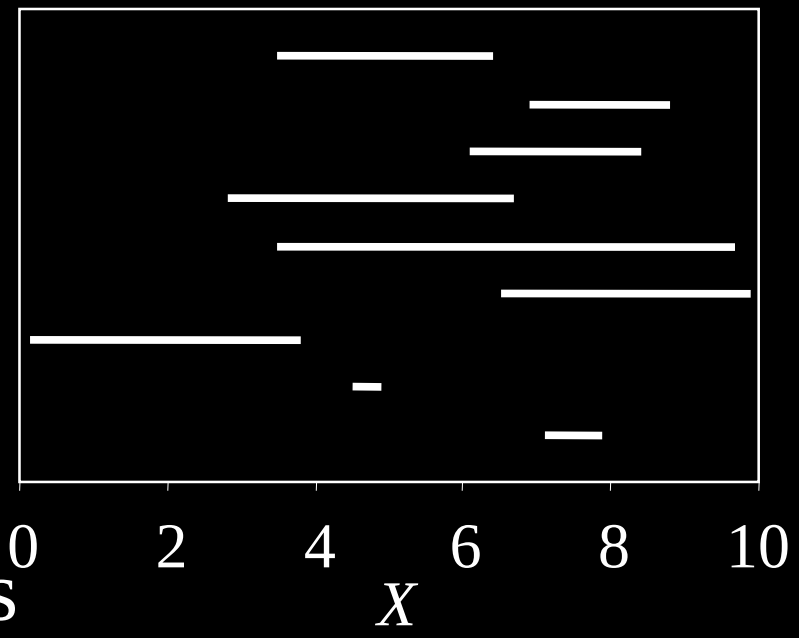
- Many ways to effectively anonymize
- Different ways yield different data utilities
- Pick the way that has the best data utility

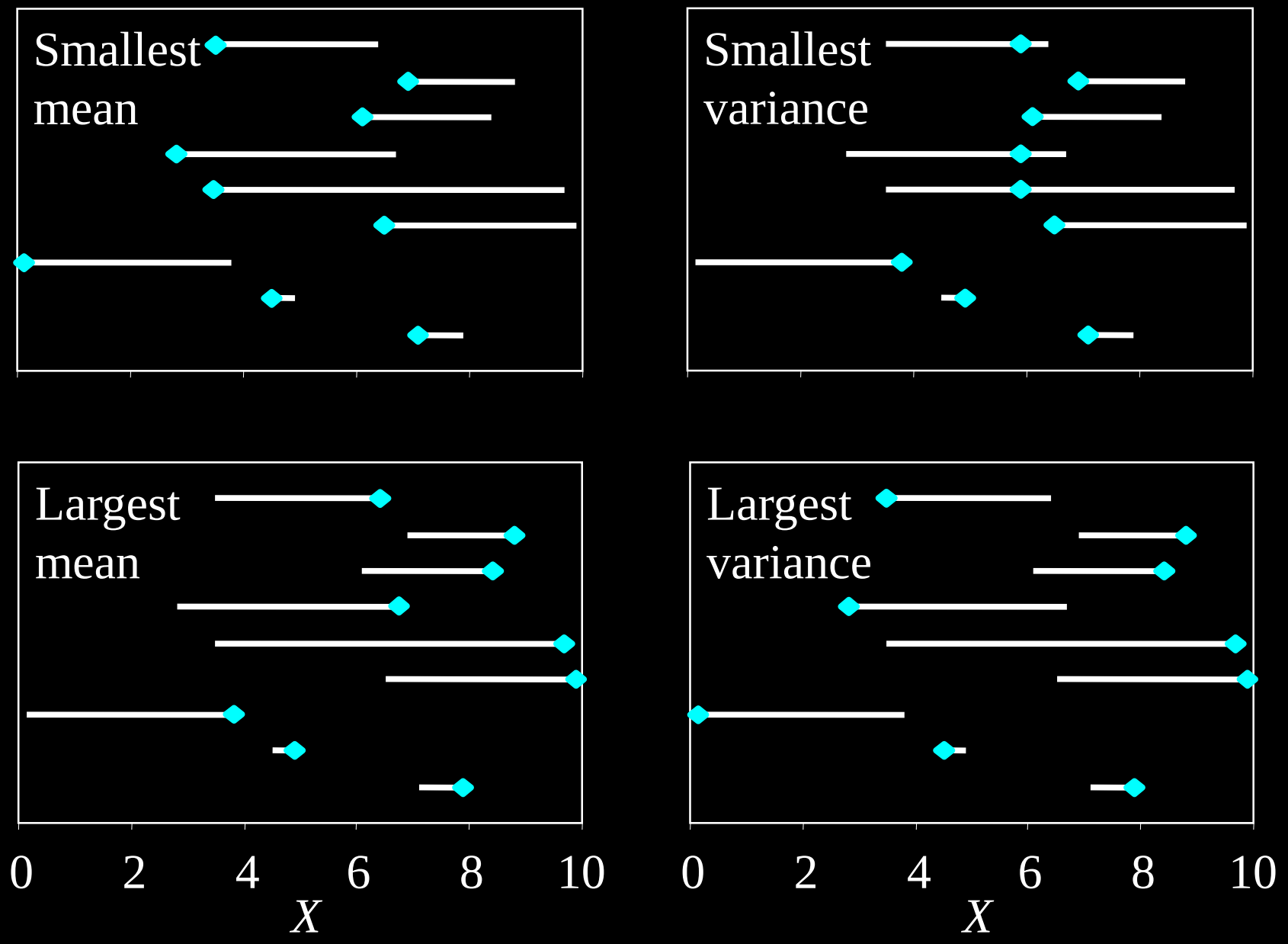
…so how do we characterize data utility?

Interval statistics gives us utility

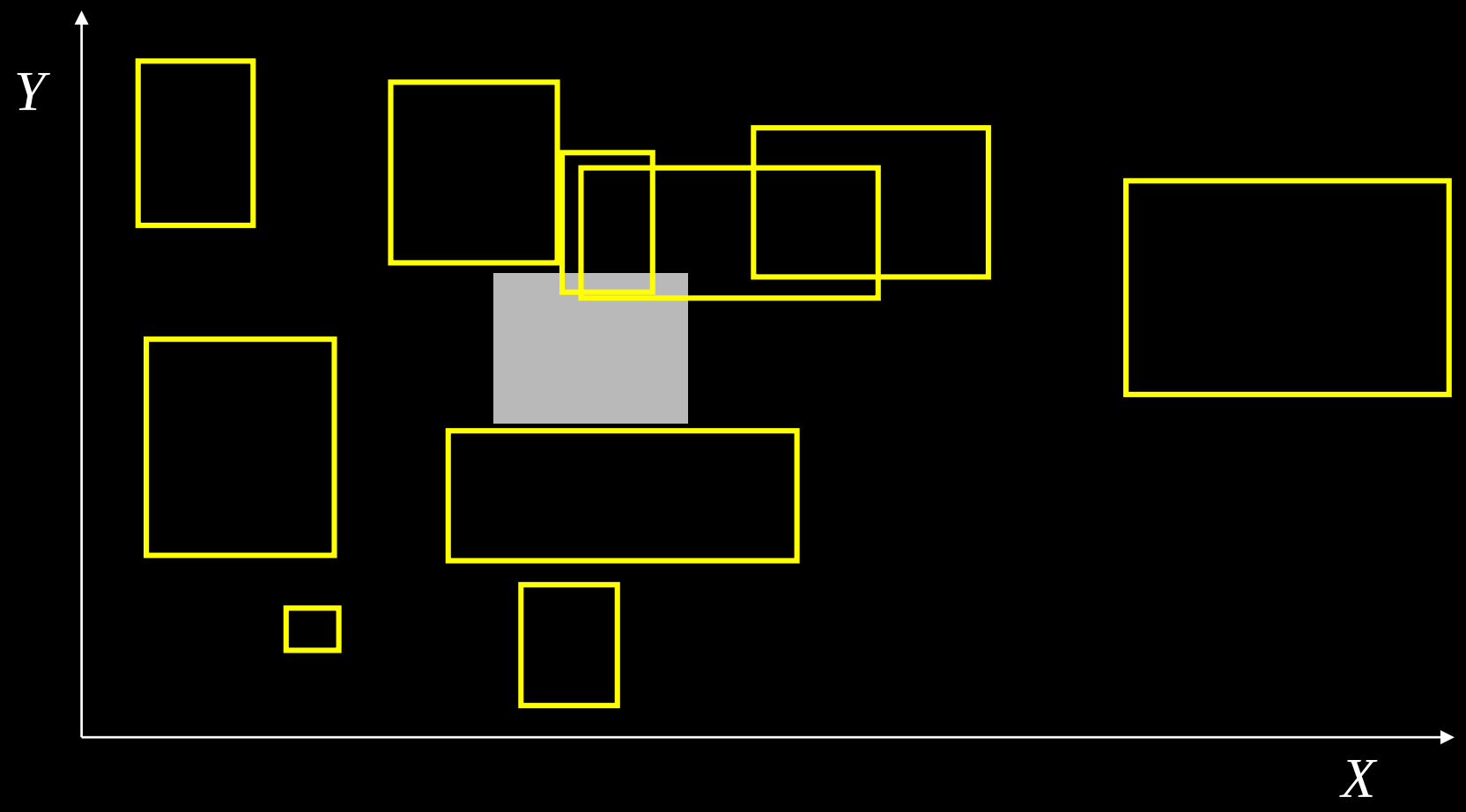
- Averages
- Dispersion
- Skewness
- Confidence limits
- Correlation
- Regression coefficients
- Distribution
- etc.

Intervals

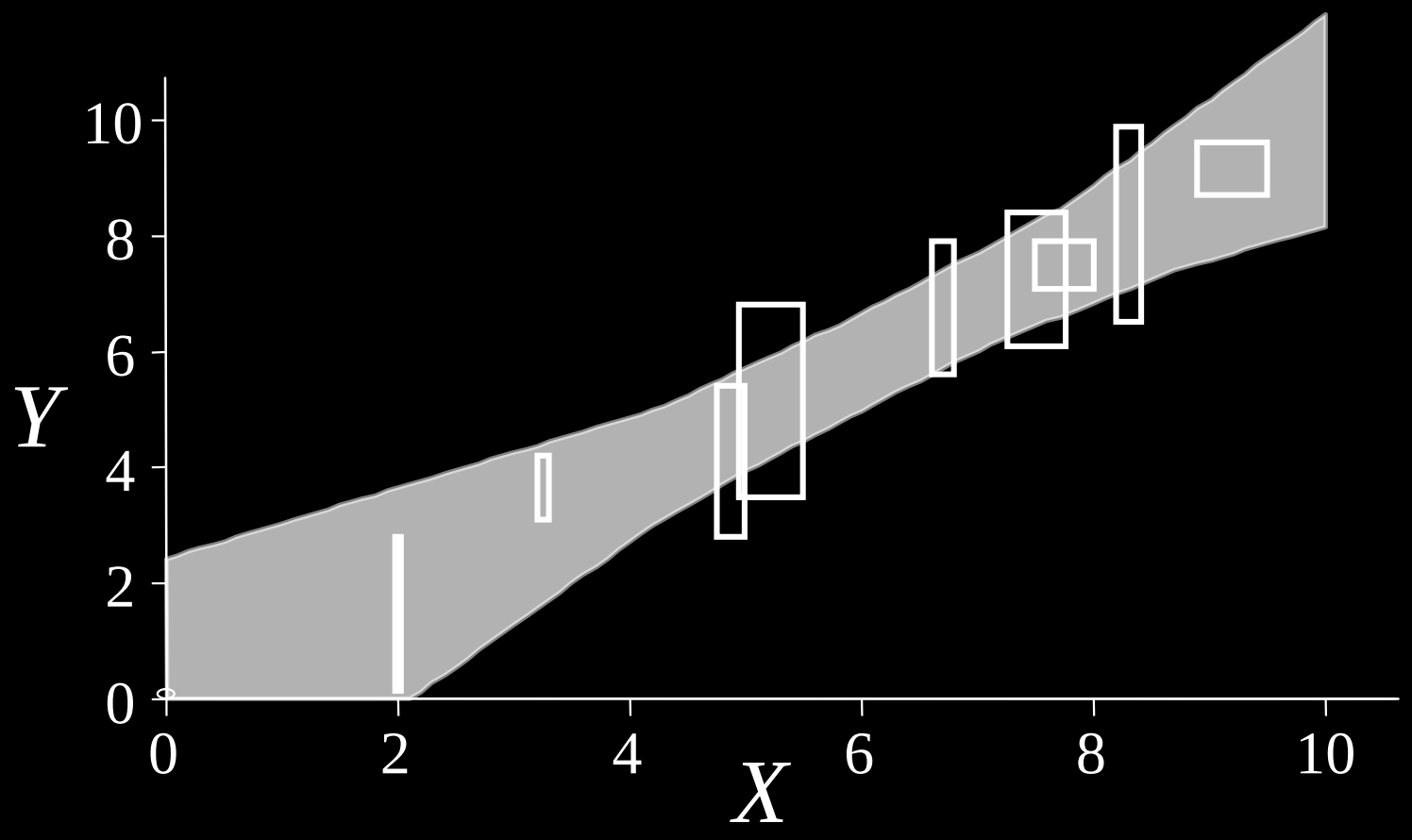



Smallest mean
Smallest variance
Largest mean
Largest variance
0
2
4
6
8
10
X
0
2
4
6
8
10
X

Bivariate data blur to boxes



Envelope of linear regressions



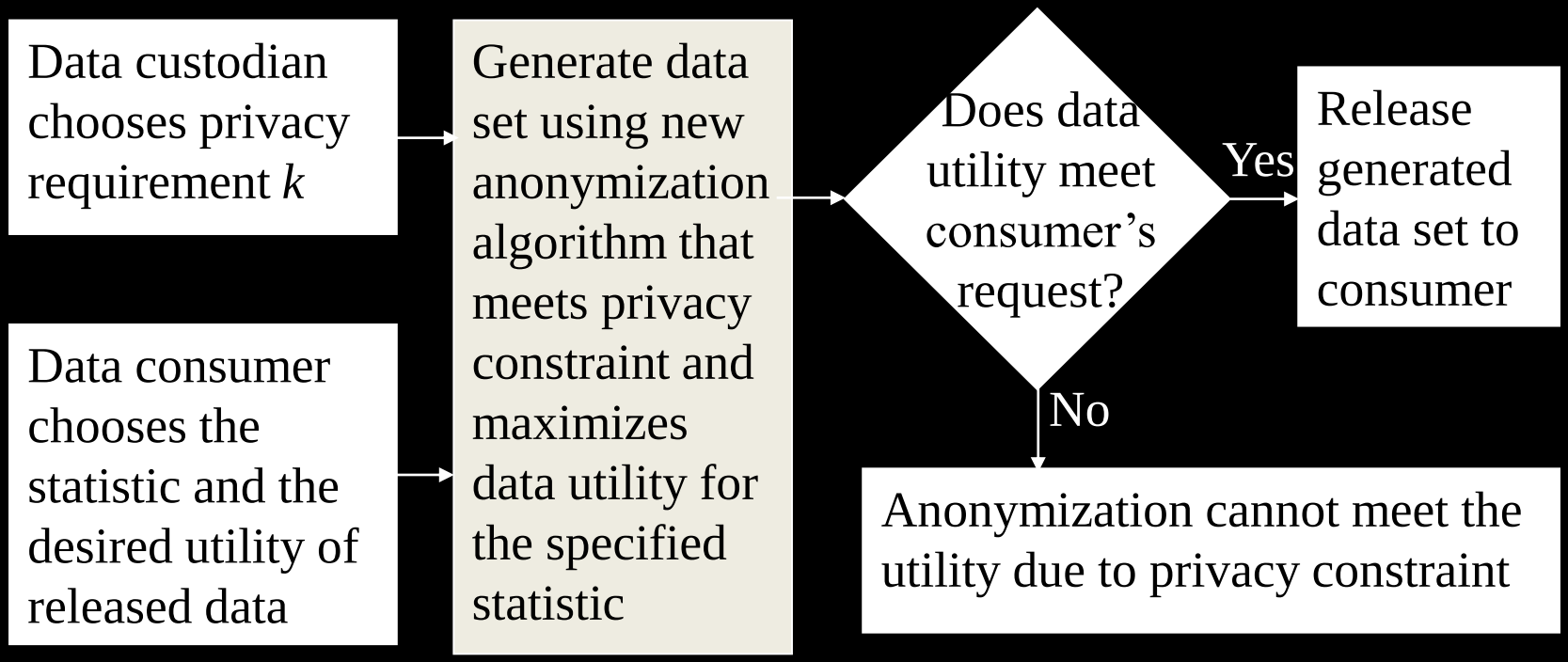
We're developing software

- Anonymization algorithms observing privacy-constraint with improved control on data utility
- Breadth of output statistics measures utility
  - Mean, variance, covariance, correlation, etc.
- Algorithms theoretically proven to generate datasets meeting both the privacy requirement (*k*-anonymity) and specified utility (Xiang et al. 2013)

Two kinds of users

- Data custodian
  - Keeper of data, charged with protecting privacy
  - Needs to protect the privacy of people in the data
- Data consumer
  - Researcher wanting to access anonymized data
  - Needs to do statistics on interval data released by the data custodian

Data custodian: anonymization



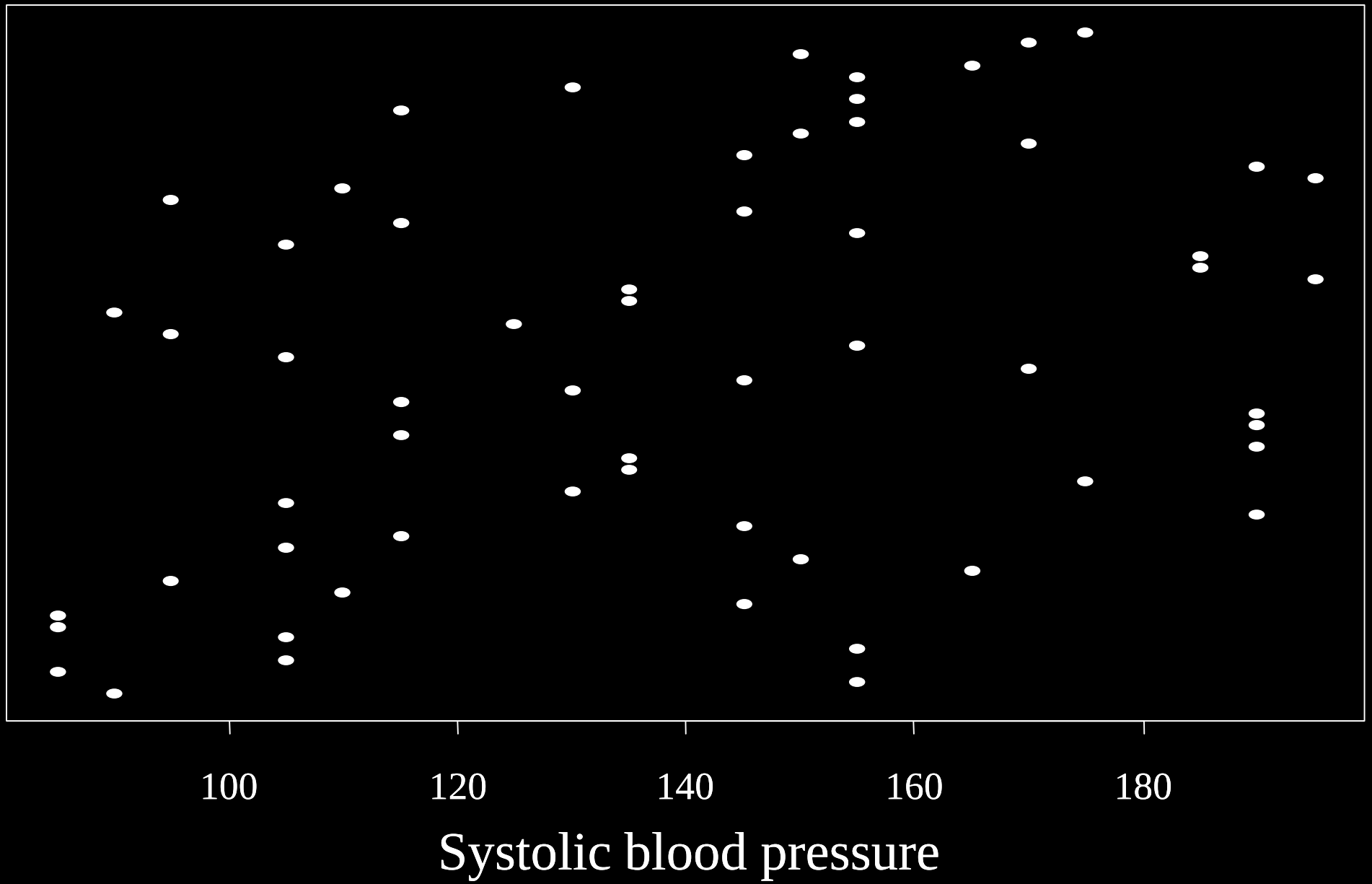
Data consumer: interval statistics

- Most basic descriptive statistics
  - Median; arithmetic, geometric, harmonic means;
  - Confidence intervals on the mean
  - Variance, standard deviation, cv, interquartile range
  - Skewness, quantiles, exceedance risks
  - Distribution function, confidence bands on cdf
- Consumers who get anonymized data sets can extract the statistical information still present

Intervals or p-boxes

Identifier	Sex	Treatmen	Diabetes	Smoker	Age years	Total Chol	HDL-Chole	Systolic Bl
1397	M	N	Y	Y	36	159	60	90
1419	M	Y	N	Y	36	141	63	155
1446	M	N	Y	N	34	141	77	85
1449	M	Y	Y	Y	34	141	83	105
1470	M	N	Y	N	34	159	80	155
1500	M	Y	N	N	36	159	103	105
1538	M	Y	Y	N	34	189	40	85
1542	M	N	Y	N	36	189	43	85
1563	M	Y	N	Y	36	180	43	145
1597	M	N	Y	Y	35	189	60	110
1640	M	N	N	N	34	180	80	95
1672	M	N	N	N	34	189	80	165
1705	M	Y	Y	Y	34	171	100	150
1738	M	Y	Y	N	35	191	43	105
1742	M	N	Y	N	34	209	37	115
1757	M	N	Y	Y	34	209	40	145
1774	M	N	Y	N	36	200	37	190
1785	M	Y	Y	Y	36	191	63	105
1798	M	N	Y	N	34	209	60	130
1815	M	N	N	Y	35	209	57	175

Data



Data Anonymization Administration Tool

File

Source Data File: Browse

Select Columns To Be Released

Requested Statistics To Be Computed:

Mean

Columns For Computing Statistics:

- [ ] Identifier
- [ ] Sex
- [ ] Treatment of High Blood Pressure
- [ ] Diabetes
- [ ] Smoker
- [ ] Age years
- [ ] Total Cholesterol
- [ ] HDL-Cholesterol
- [x] Systolic Blood Pressure

Columns Used As Category In Computing Statistics:

- [ ] Identifier
- [ ] Sex
- [ ] Treatment of High Blood Pressure
- [ ] Diabetes
- [ ] Smoker
- [ ] Age years
- [ ] Total Cholesterol
- [ ] HDL-Cholesterol
- [ ] Systolic Blood Pressure

Select Privacy Protection Parameters

k-Anonymity:

k = 3

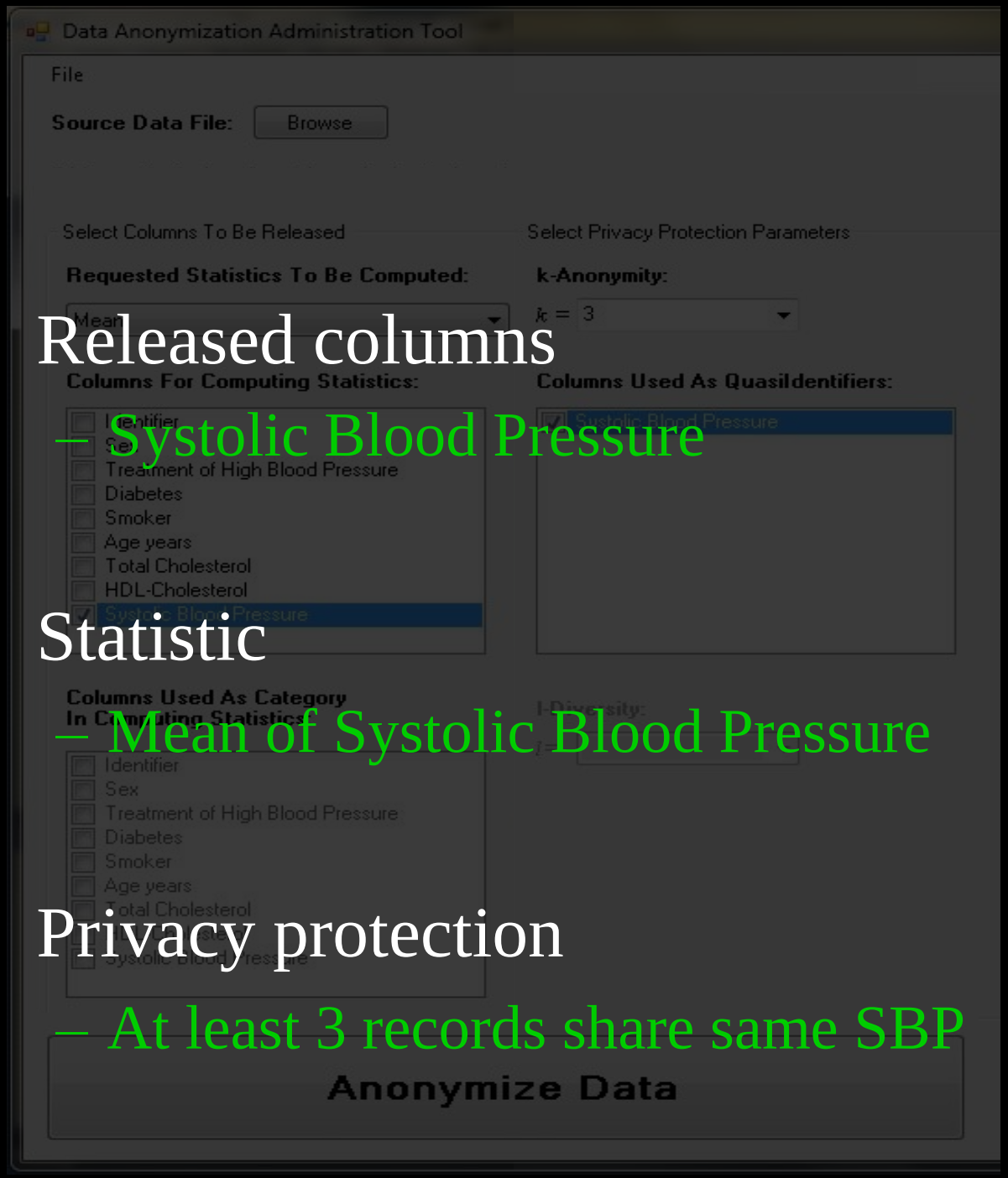
Columns Used As Quasiidentifiers:

- [x] Systolic Blood Pressure

l-Diversity:

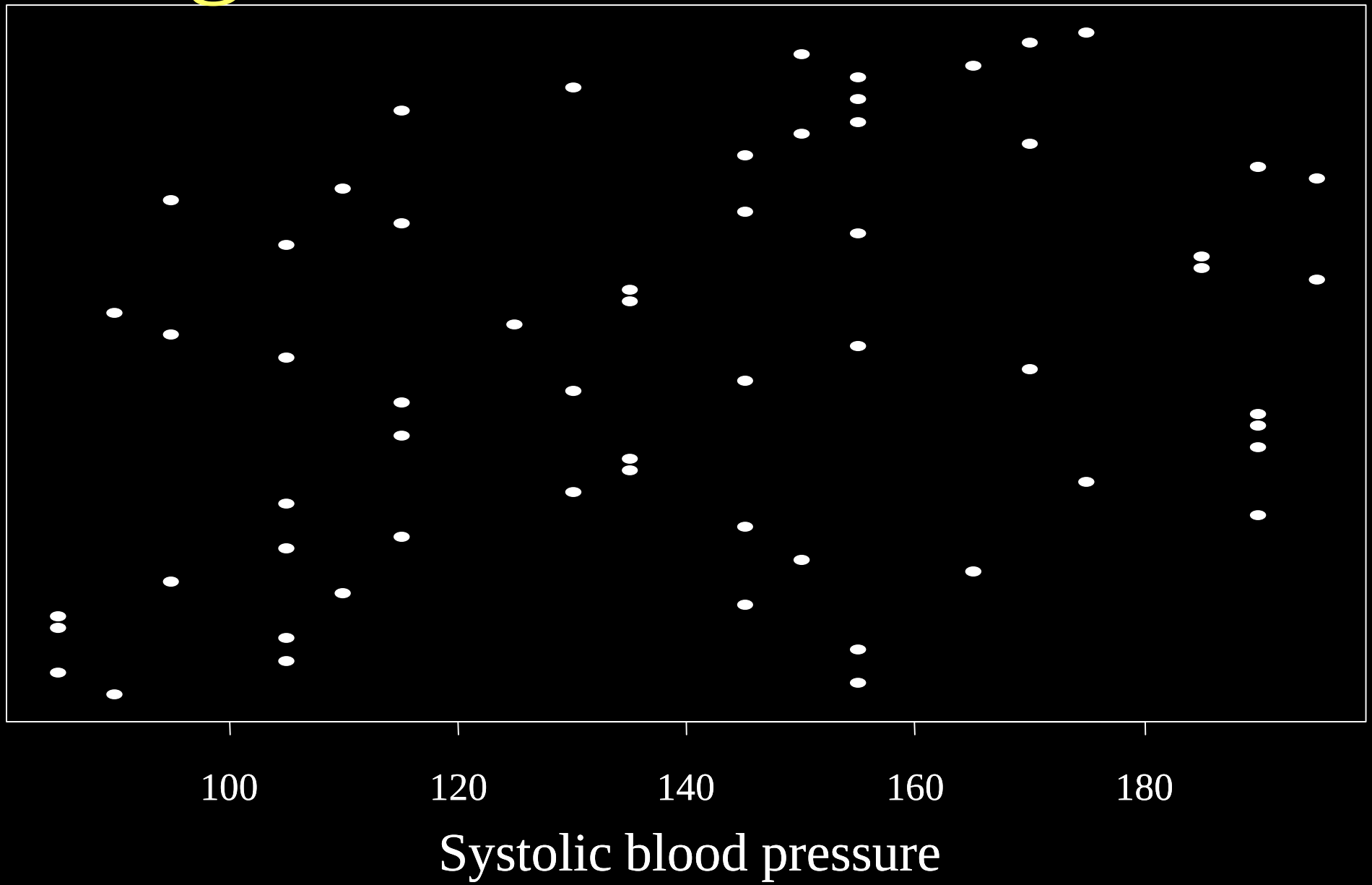
l =

Anonymize Data

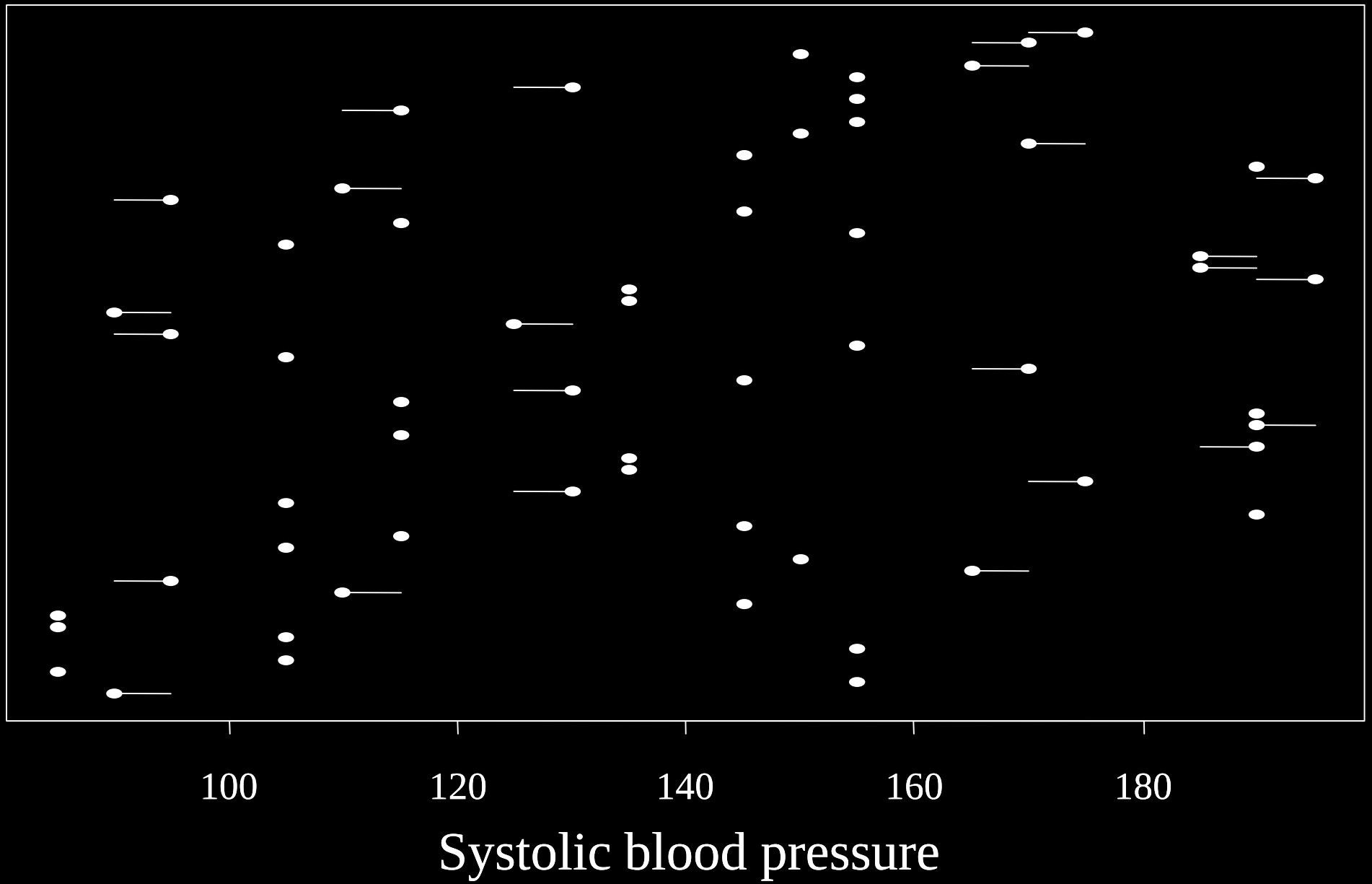


- Released columns
  - Systolic Blood Pressure
- Statistic
  - Mean of Systolic Blood Pressure
- Privacy protection
  - At least 3 records share same SBP

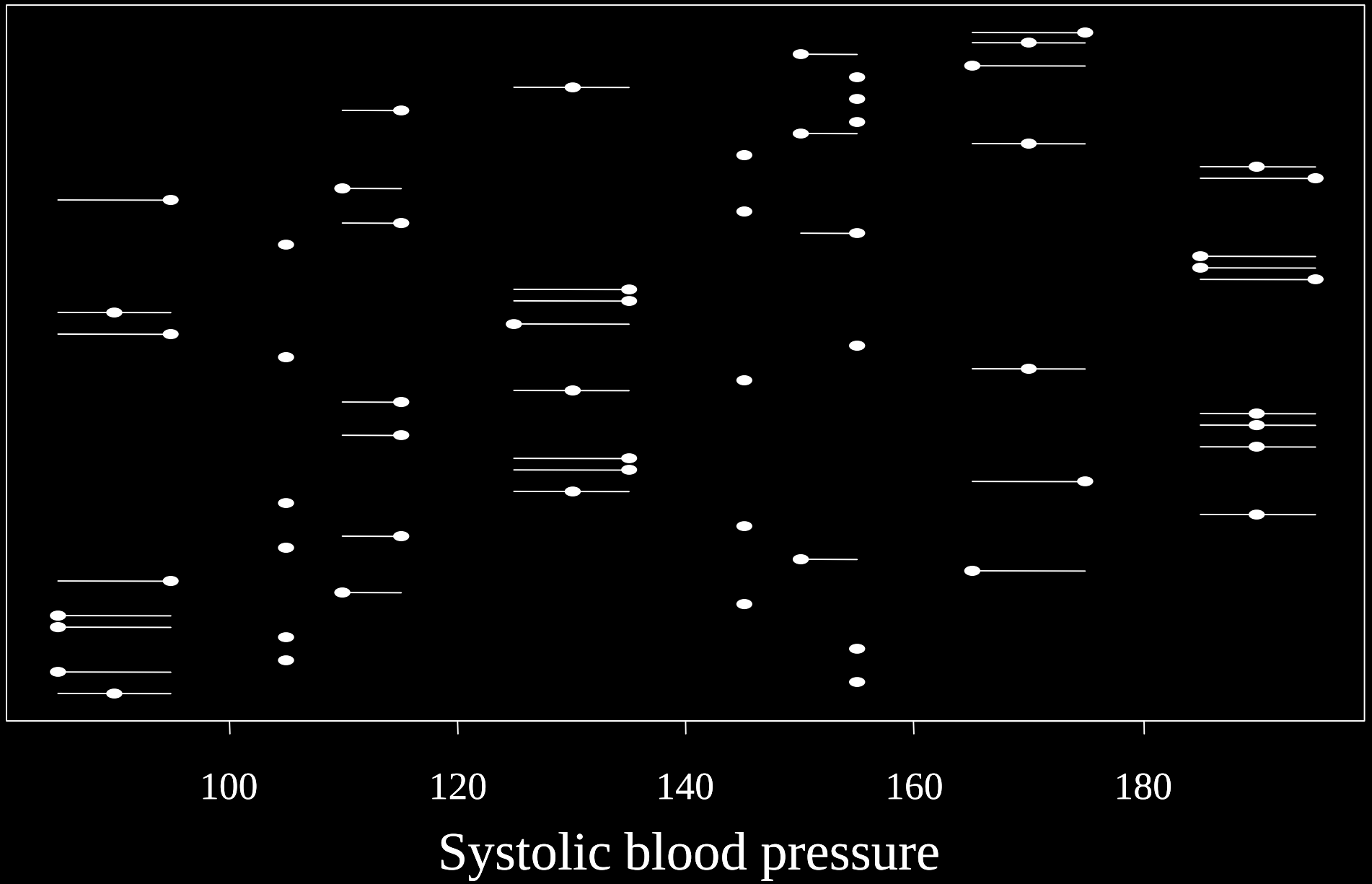
Original data



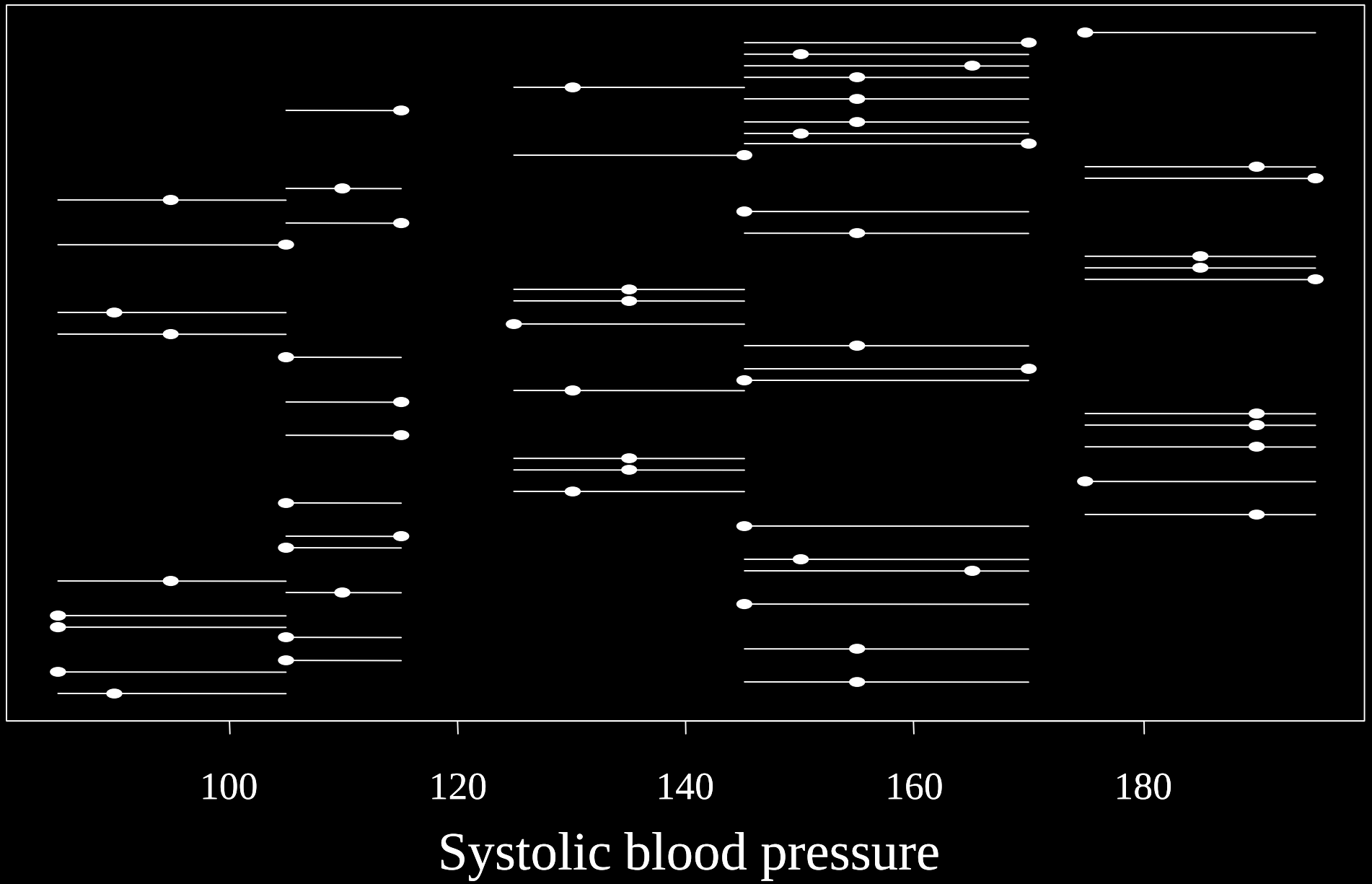
$k = 3$

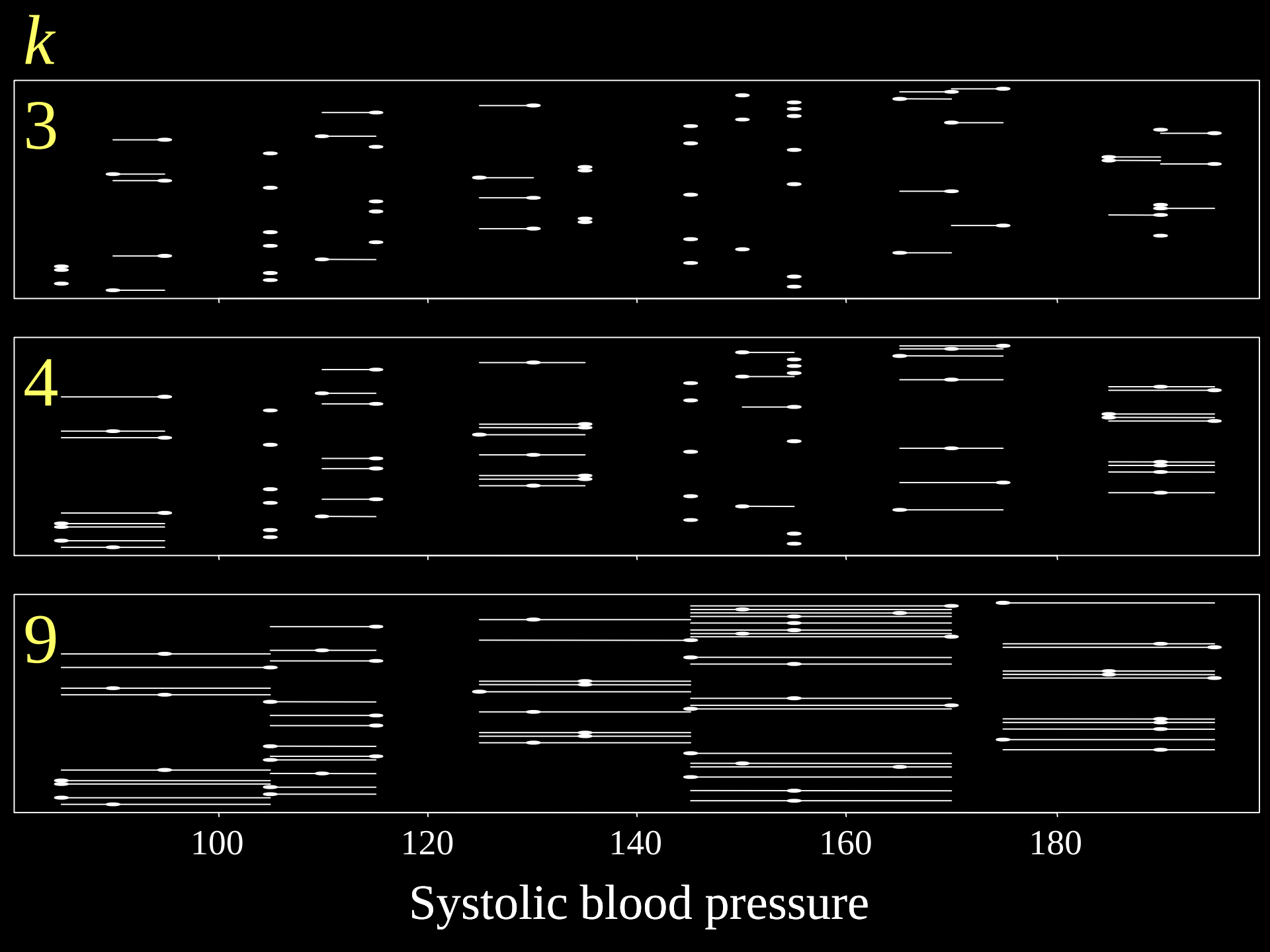


$k = 4$

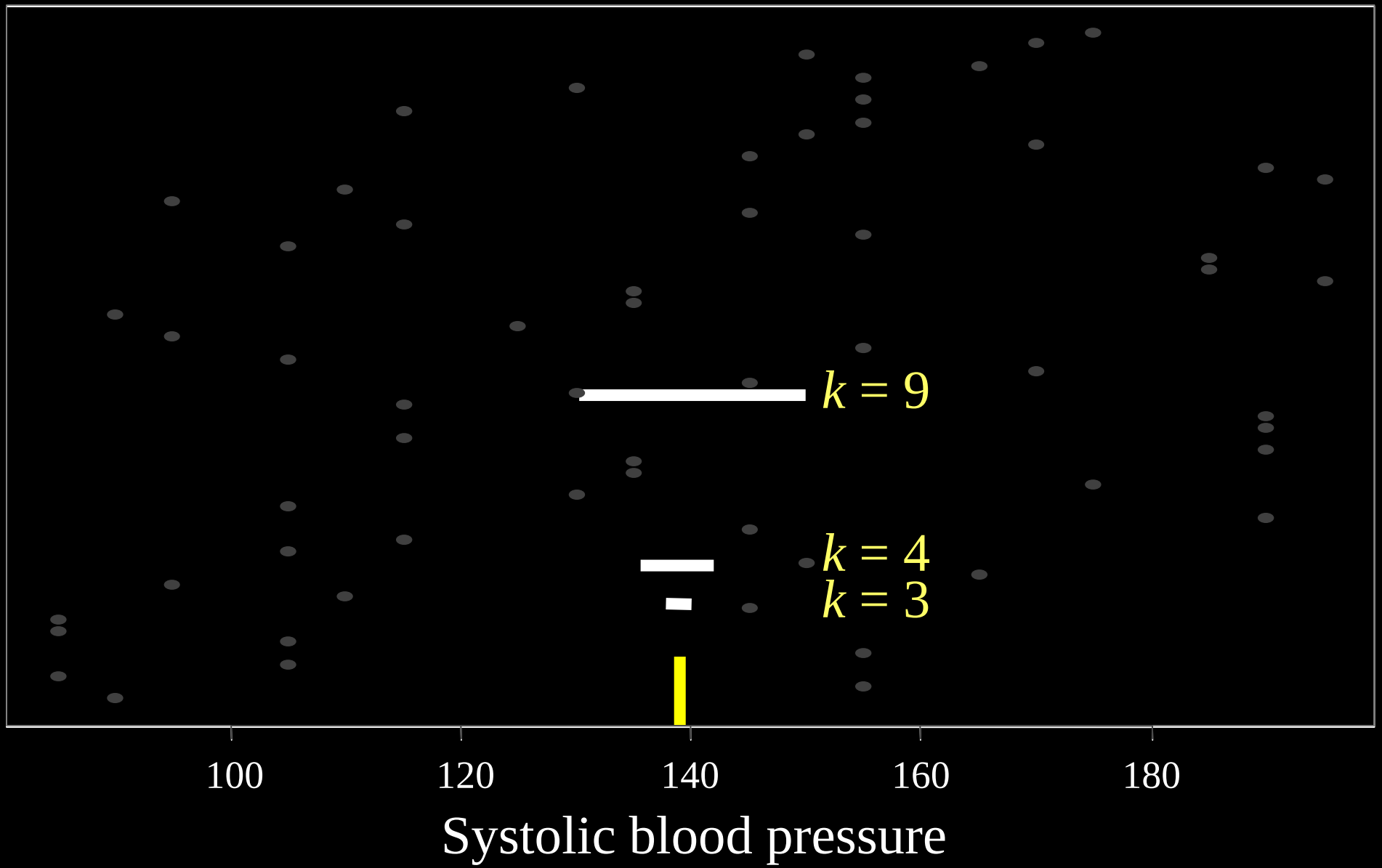


$k = 9$

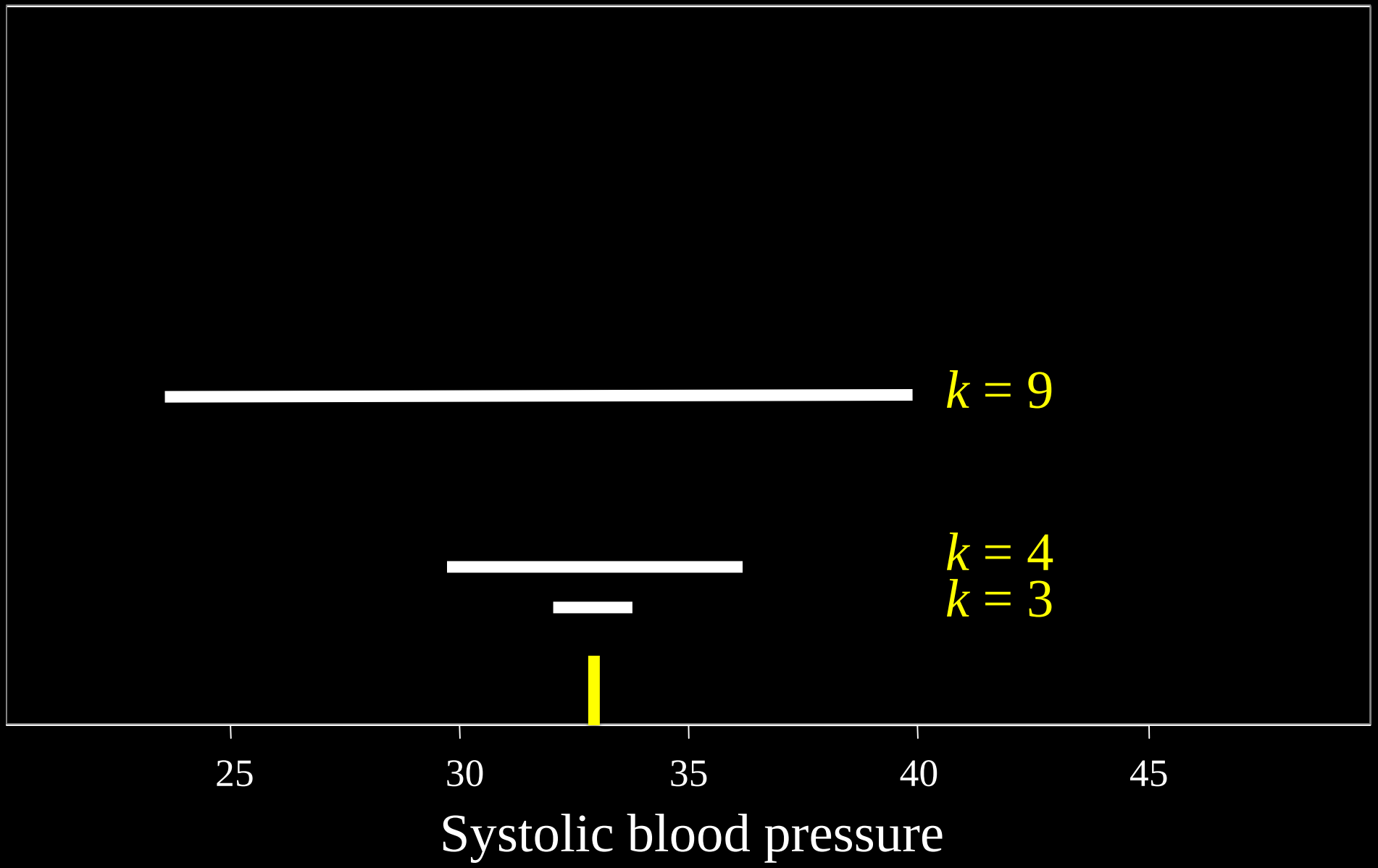



k
3
4
9
100
120
140
160
180
Systolic blood pressure

Interval mean



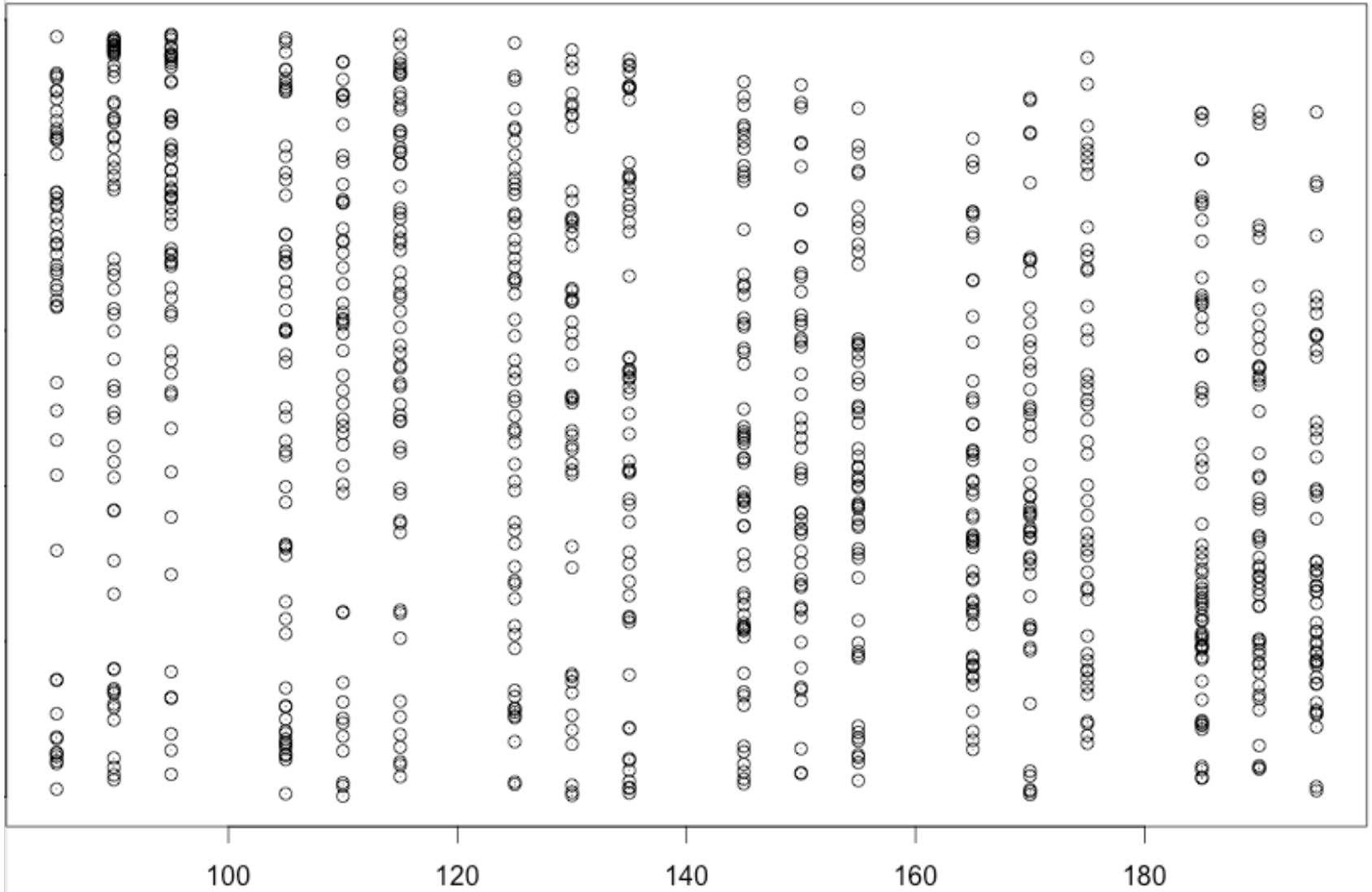
Interval standard deviation



Sometimes a little is enough

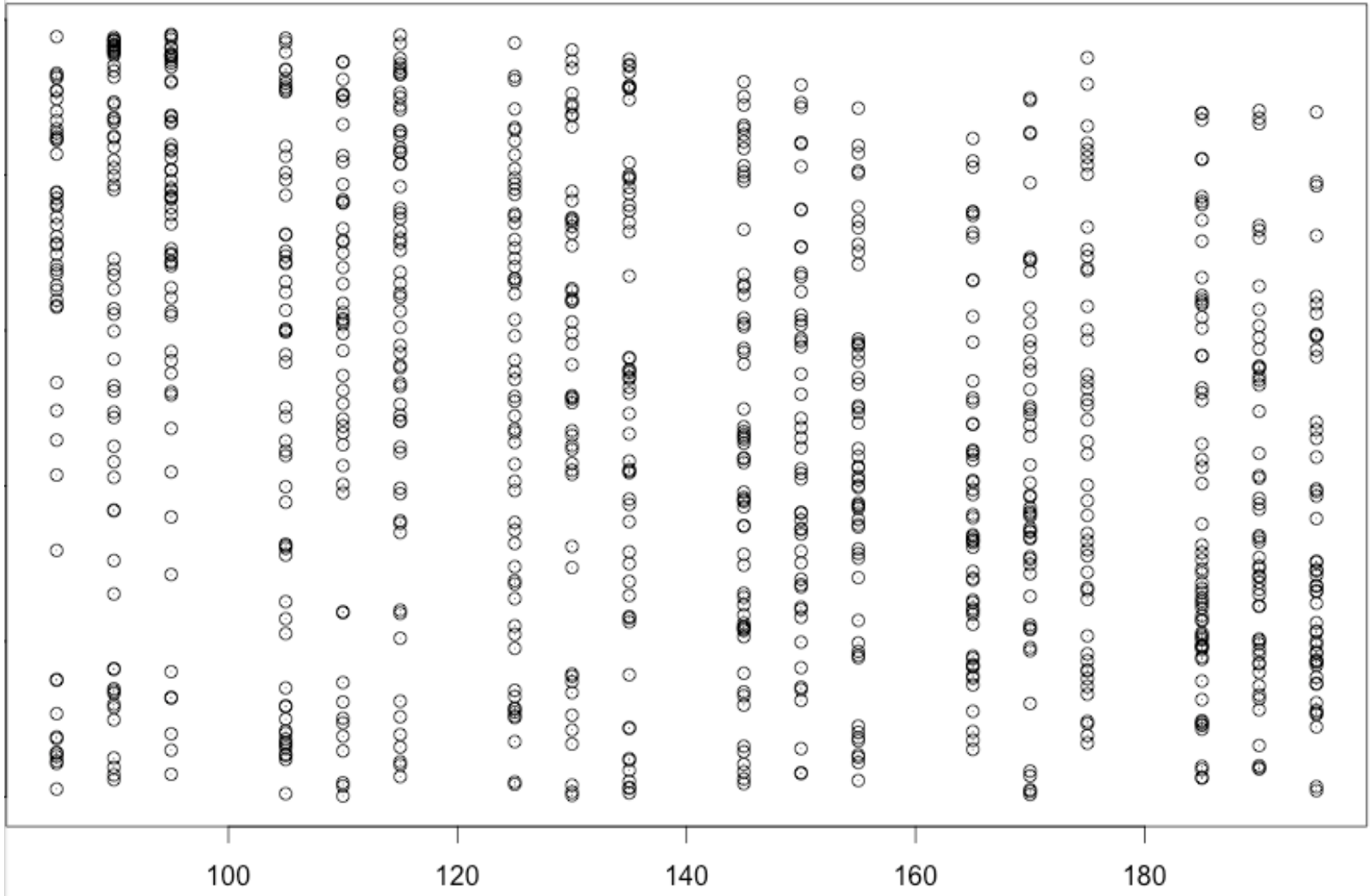
- The bigger the data set, the less blurring is needed to achieve a given *k*-diversity
- Similar economies for *l*-diversity, etc.

$k = 10$



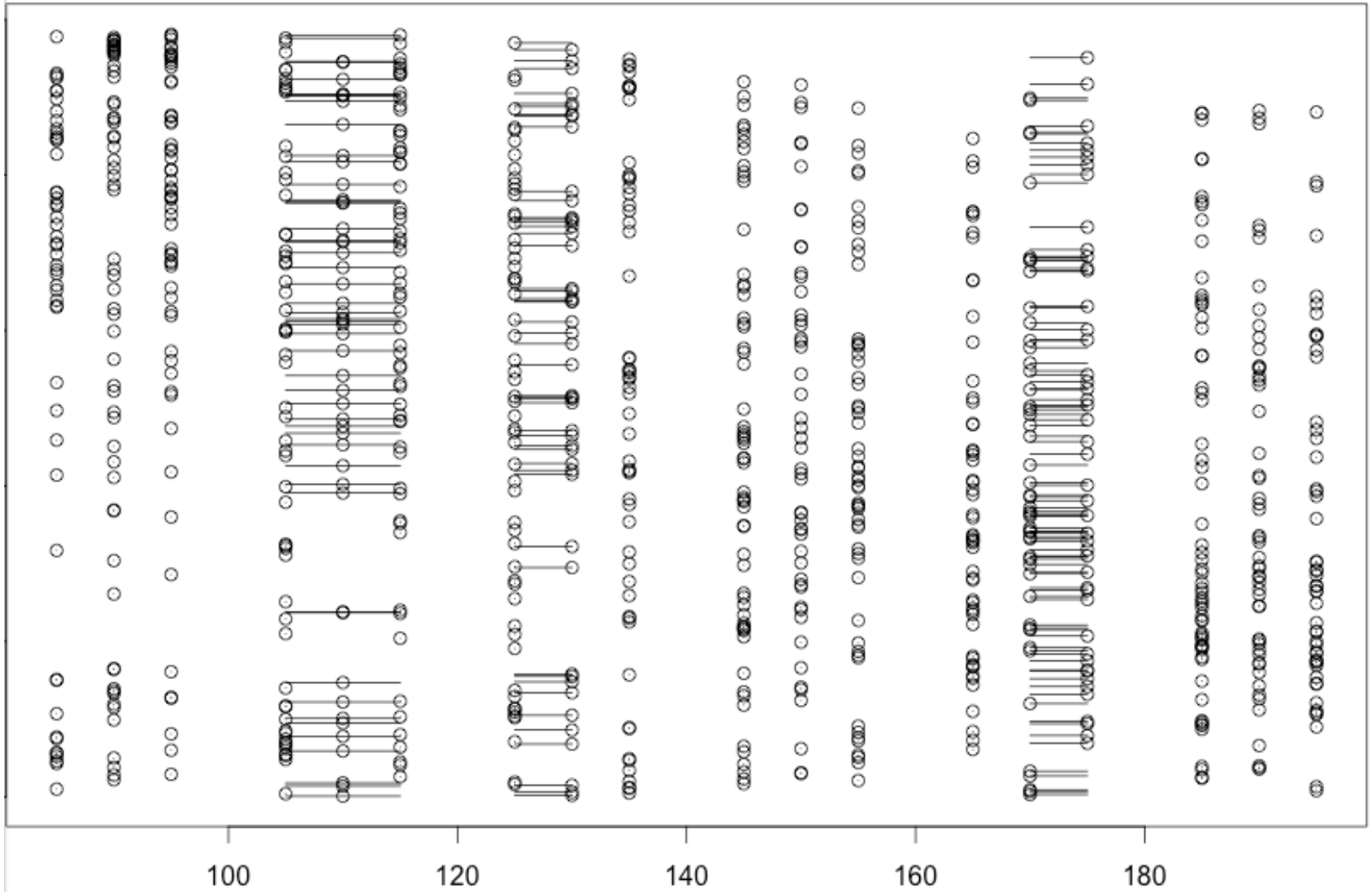
Systolic blood pressure (n=981)

$k = 20$



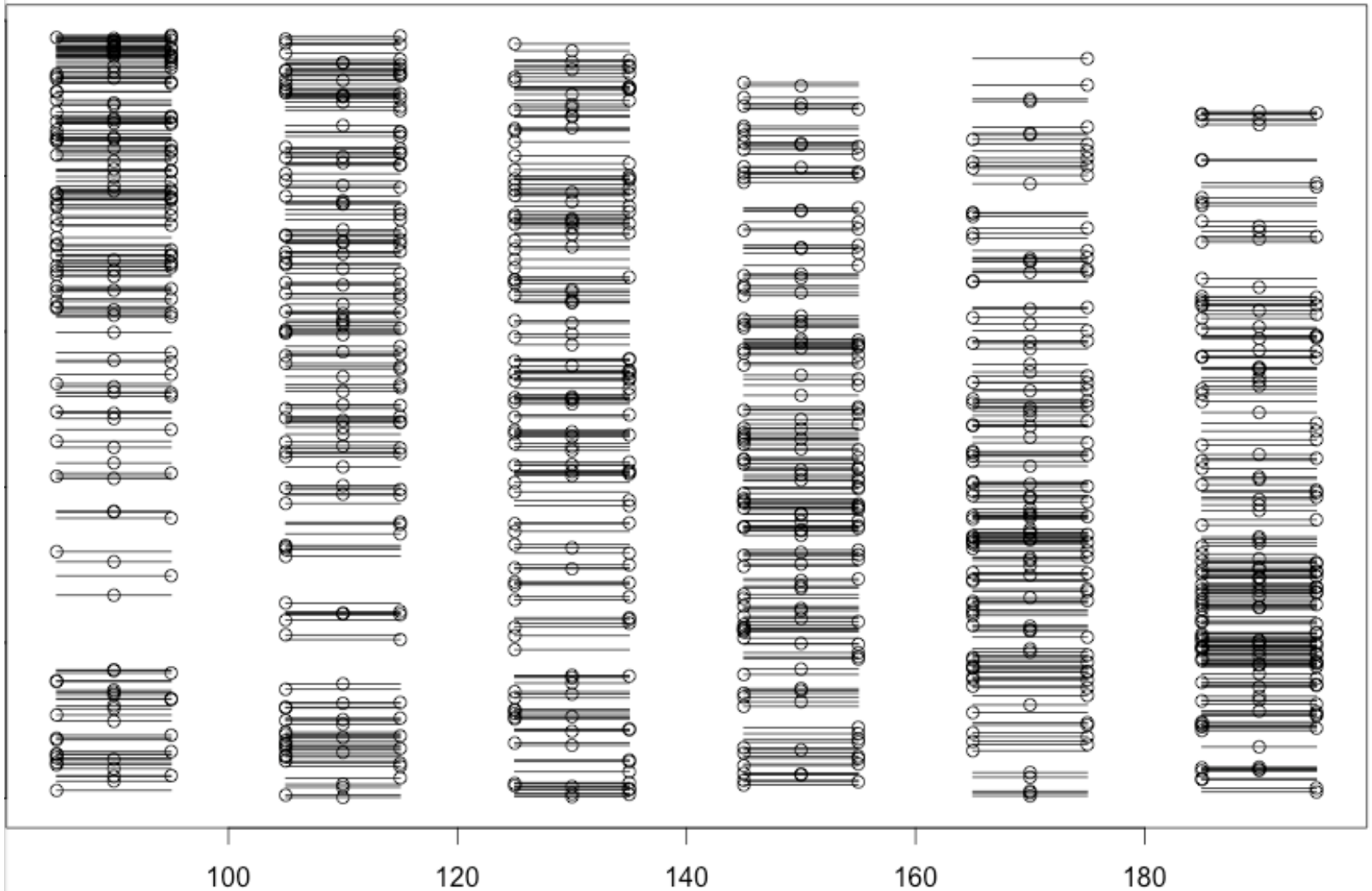
Systolic blood pressure (n=981)

$k = 50$



Systolic blood pressure (n=981)

$k = 100$



Systolic blood pressure (n=981)

RAMAS Interval Statistics: Basic statistics for interval and censored data

File Edit Page Help

Possible range 0 infinity

	Left	Right
1	0.16	34.60
2	1.17	12.09
3	4.53	17.28
4	5.34	20.23
5	7.47	10.75
6	9.84	12.66
7	14.04	16.43
8	15.51	52.20
9	17.35	48.34
10	18.99	46.90
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		

Univariate statistics | Two-sample comparisons | Bivariate statistics | Spatial statistics | Options

Central tendancy statistics

[9.440, 27.148]	Arithmetic mean
[5.468, 22.844]	Geometric mean
[1.251, 19.495]	Harmonic mean
[7.470, 17.280]	Median
[13.157, 36.292]	95 % UCL on mean
[4.772, 18.004]	95 % LCL on mean

Dispersion statistics

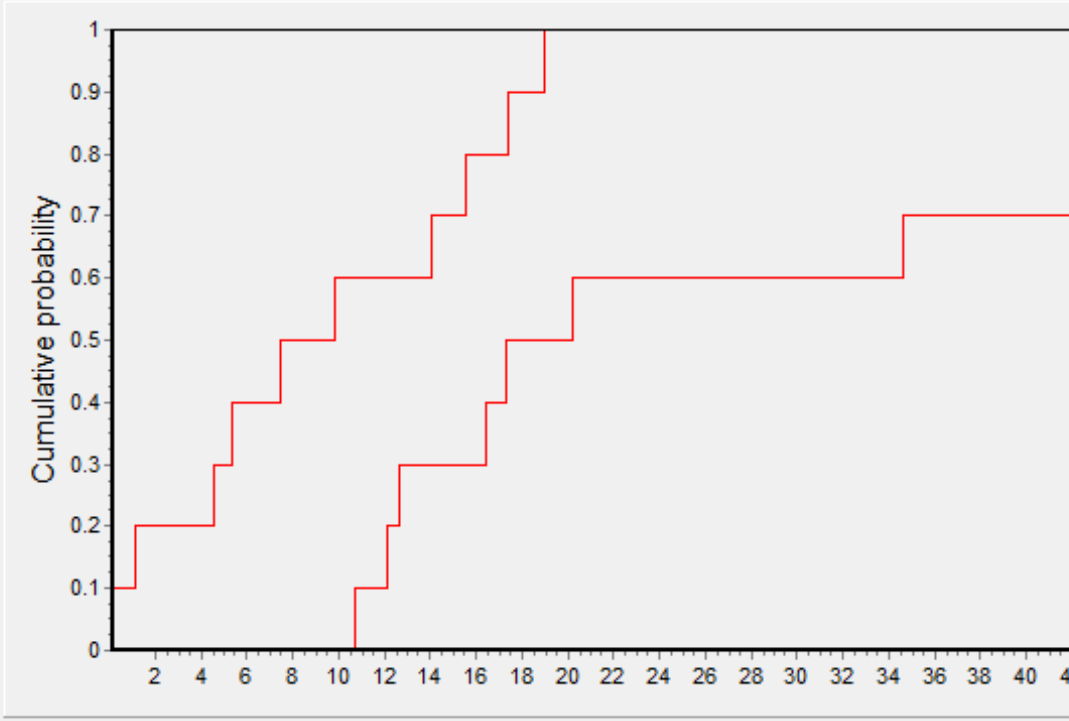
[1.719, 6.365]	Standard error
[5.435, 20.129]	Standard deviation
[29.544, 405.176]	Variance
[0.362, 1.107]	Coefficient of variation
[4.530, 46.900]	Interquartile range
[0.160, 52.200]	Range

Shape statistics

[-1.433, 1.896]	
[0.160, 10.750]	
[0.160, 10.750]	
[17.350, 48.340]	
[18.990, 52.200]	
Confidence for KS li	

Counts

10	Total values
0	Point values
10	Interval values



Compute

Edit graph | Print graph

Help | Exceedance

Extremal configurations

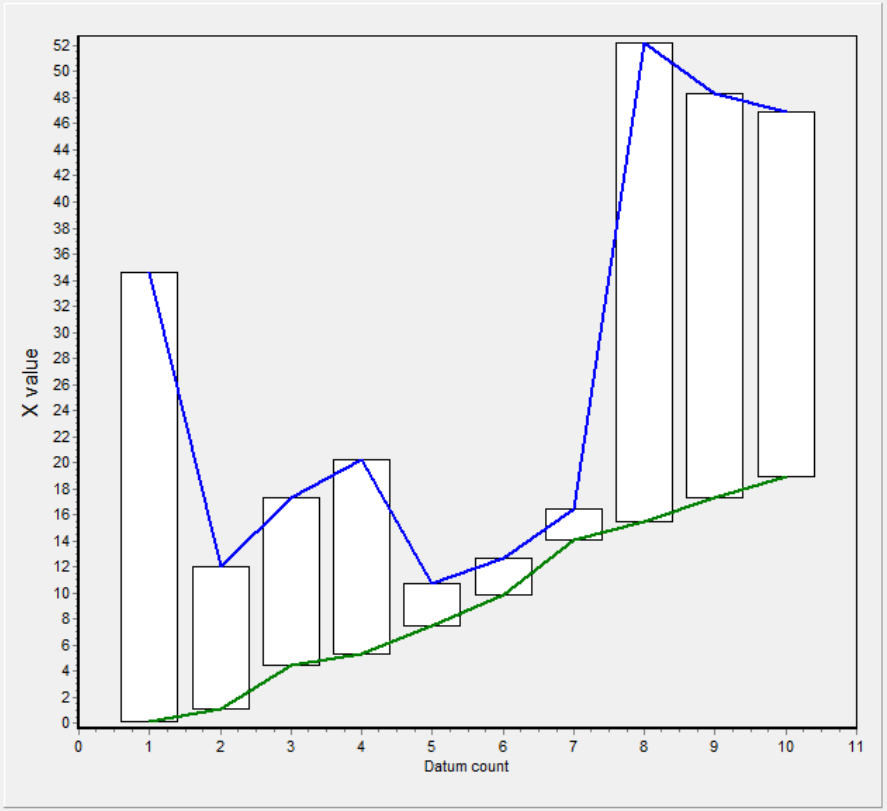
Arithmetic mean [9.440, 27.148]

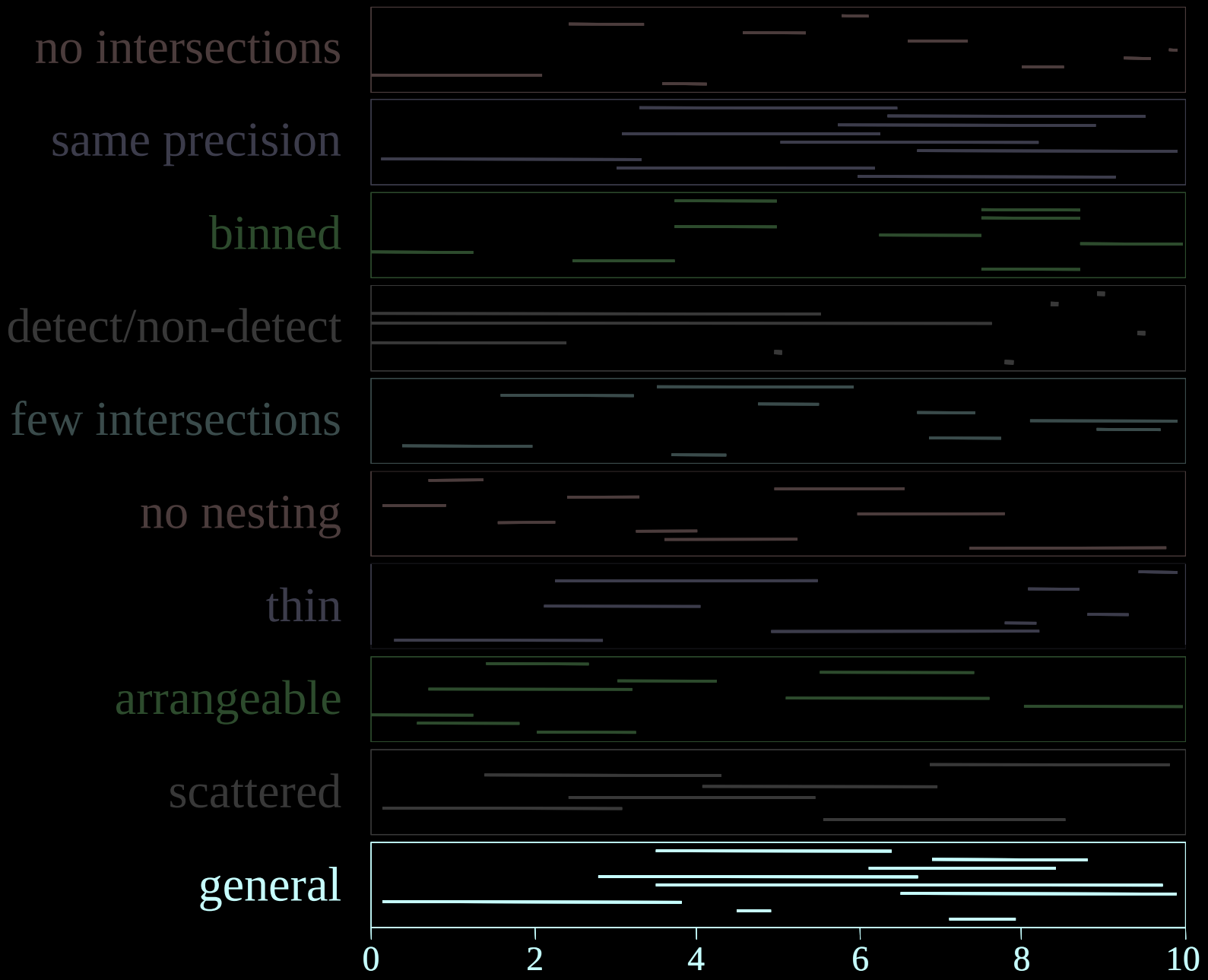
Close

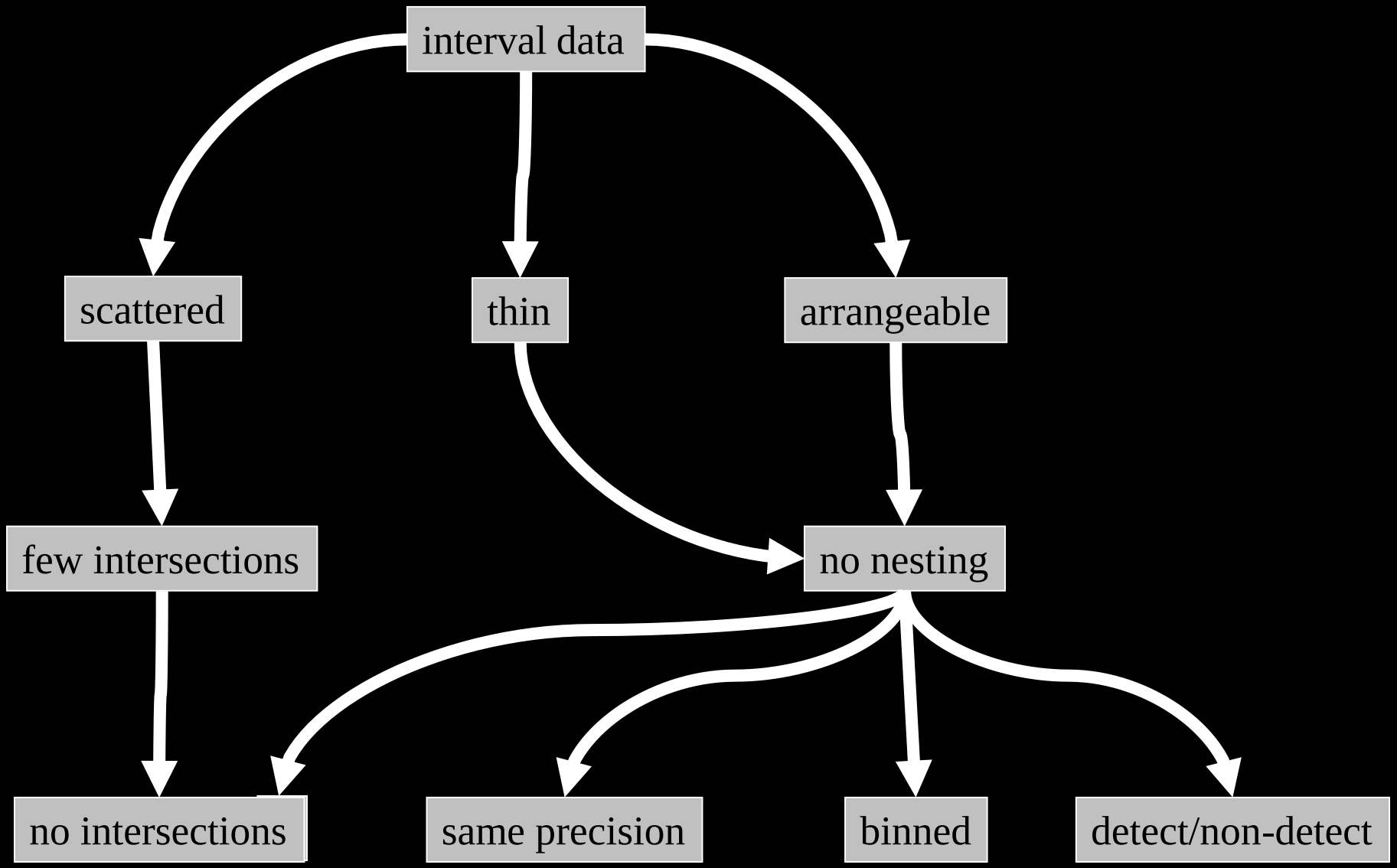
Edit graph

Print graph

	Left bound	Configuration 1	Configuration 2	Right bound
1	0.160	0.160	34.600	34.600
2	1.170	1.170	12.090	12.090
3	4.530	4.530	17.280	17.280
4	5.340	5.340	20.230	20.230
5	7.470	7.470	10.750	10.750
6	9.840	9.840	12.660	12.660
7	14.040	14.040	16.430	16.430
8	15.510	15.510	52.200	52.200
9	17.350	17.350	48.340	48.340
10	18.990	18.990	46.900	46.900




no intersections
same precision
binned
detect/non-detect
few intersections
no nesting
thin
arrangeable
scattered
general
0
2
4
6
8
10


interval data
scattered
thin
arrangeable
few intersections
no nesting
no intersections
same precision
binned
detect/non-detect

Computational complexity

Case	$[\underline{V},\overline{V}]$	$[\underline{L},\overline{L}]$ and $[\underline{U},\overline{U}]$	$[\underline{S},\overline{S}]$
No intersections	$O(n)$	$O(n \log n)$	$O(n^2)$
Same precision	$O(n)$	$O(n \log n)$	$O(n^3)$
Binned data	$O(n)$	$O(n \log n)$	$O(n^2)$
Few intersections	$O(n \log n)$	$O(n \log n)$	?
No nesting	$O(n)$	$O(n \log n)$	$O(n^2)$
Arrangeable	$O(n^m)$	$O(n^m)$	$O(n^{2m})$
General	NP-hard	NP-hard	?

Computability

- We get to decide what shape of intervals to use
- We can elect to pick shapes for which statistics are relatively easy to compute
- Important consideration in very large data sets

Next steps

- Currently exploring the scalability of the algorithms for large and very large data sets which are common in health care
- Generalizing for other privacy constraints (*l*-diversity, etc.)
- Generalizing for other statistics

Acknowledgments

- National Institutes of Health SBIR program
- Jody Sacks & Olga Brazhnick, NCATS
- Lev Ginzburg, Stony Brook University

Thanks

Abstract

Patient health records possess a great deal of information that would be useful in medical research, but access to these data is impossible or severely limited because of the private nature of most personal health records. Anonymization strategies, to be effective, must usually go much further than simply omitting explicit identifiers because even statistics computed from groups of records can often be leveraged by hackers to re-identify individuals. Methods of balancing the informativeness of data for research with the information loss required to minimize disclosure risk are needed before these private data can be widely released to researchers who can use them to improve medical knowledge and public health. We are developing an integrated software system that provides solutions for anonymizing data based on interval generalization, controlling data utility, and performing statistical analyses and making inferences using interval statistics.