Design and Performance of a Scalable, Parallel Statistics Toolkit

Philippe Pébay, Janine Bennett,
David Thompson and Ajith Mascarenhas

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Context

- Statistical engines are part of VTK, which is used in many analytics and visualization applications.
- Freely available under BSD license:

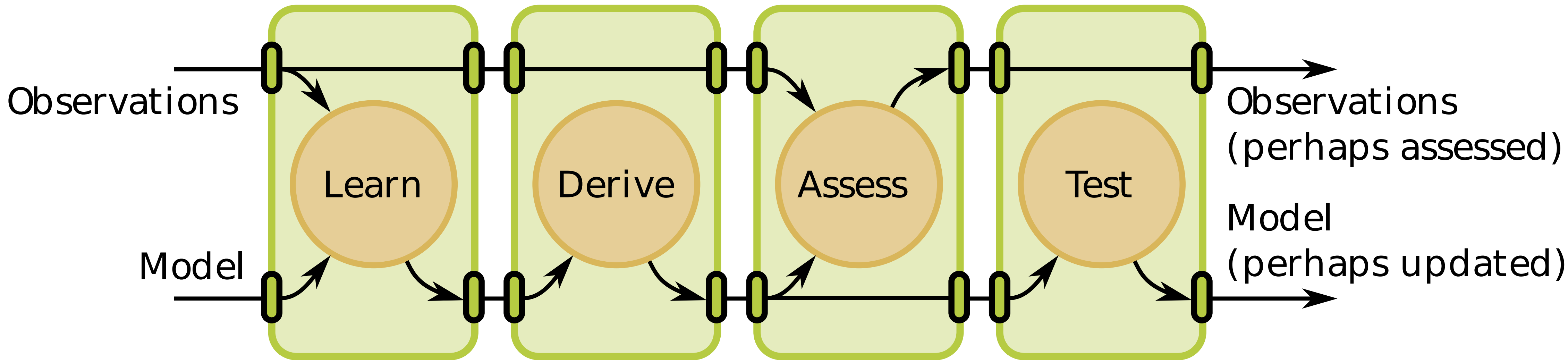
  http://www.vtk.org/

Outline

- Design model for scalable data analysis
- Available parallel workflows
- Example of a moment-based workflow
- Moment vs. quanta-based statistics
- Parallel example
- Class hierarchy
- Scaling study
- Observations and future work

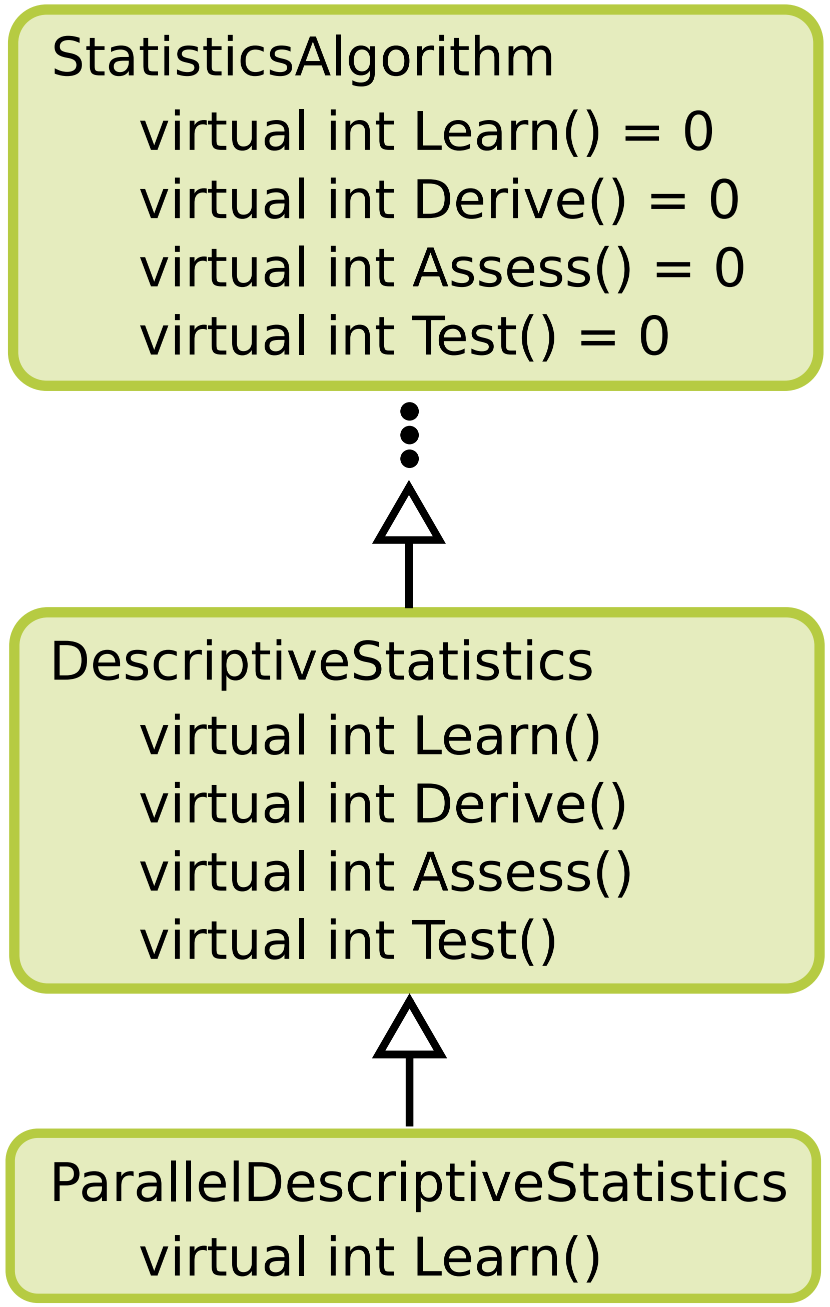
Design Decisions

(i) Mimic predominant types of data analysis workflow
(ii) Conducive to scalable parallel implementation



- Learn a minimal model,
- Derive quantities of interest,
- Assess observations with the model, and
- Test the validity of the model hypothesis.

Implementation



Statistics algorithms all conform to a simple class hierarchy:

- Abstract base class with virtual modeling operations
- Concrete serial implementation
- Derived parallel implementation

Available Workflows

- Descriptive statistics
- Order statistics
- Linear and multi-linear correlation
- Contingency statistics
- Principal Component Analysis (PCA)
- *k*-means clustering

Moment-based Analysis

- Descriptive statistics
  - **Learn**: minimum, maximum, mean, centered M2, M3, M4 aggregates
  - **Derive**: variance, standard deviation, skewness, kurtosis
  - **Assess**: mark each datum with relative deviation (1-D Mahalanobis distance)
  - **Test**: Jarque-Bera statistic and χ^2 test of goodness of fit

Quanta-based Analysis

- Order statistics
  - **Learn**: histogram
  - **Derive**: arbitrary quantiles
  - **Assess**: mark each datum with quantile index
  - **Test**: Kolmogorov-Smirnov test

Moments vs. Quanta

- Only Learn requires communication
- Moment-based analysis: few quantities to be communicated and updated
  - *Model:* AllGather with local update
  - *Caveat:* numerical stability
- Quanta-based analysis: variable number of quantities to be communicated and updated depending on data
  - *Model:* FullReduce + broadcast
  - *Caveat:* scalability

Parallel computation

Process 0	Process 1	Process 2	
a,c,d,c, d,a,e	c,c,b,e,c,c,e, c,b,e,c,e,c	b,b,d,b,a, b,b,a,b,a	Raw observations
			Count observations
a,c,d,e 2,2,2,1	b,c,e 2,7,4	a,b,d 3,6,1	Process-local contingency table

```
    a,c,d,e
∪   b,c,e
∪   a,b,d
-----------
    a,b,c,d,e
```

Union all keys

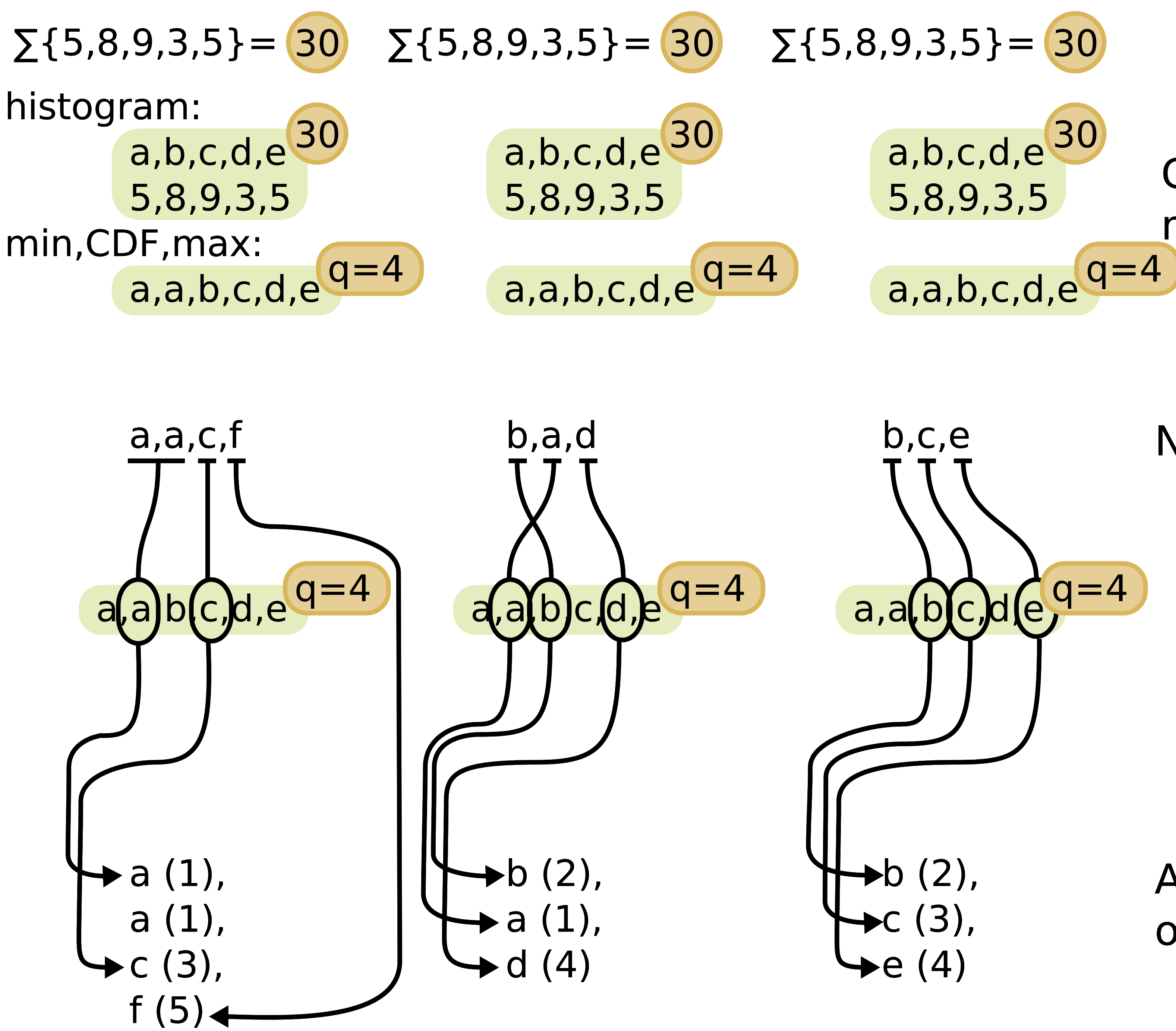
```
    a,b,c,d,e
    2,  ,2,2,1
+    ,2,7,  ,4
+   3,6,  ,1
-----------
    5,8,9,3,5
```

Sum values for each key

Process 0	Process 1	Process 2	
a,b,c,d,e 5,8,9,3,5			Global ordering (histogram)
	a,b,c,d,e 5,8,9,3,5	a,b,c,d,e 5,8,9,3,5	

Learn

Parallel computation



Derive

Count total observations and compute CDF.

Global histogram, min, max, and CDF.

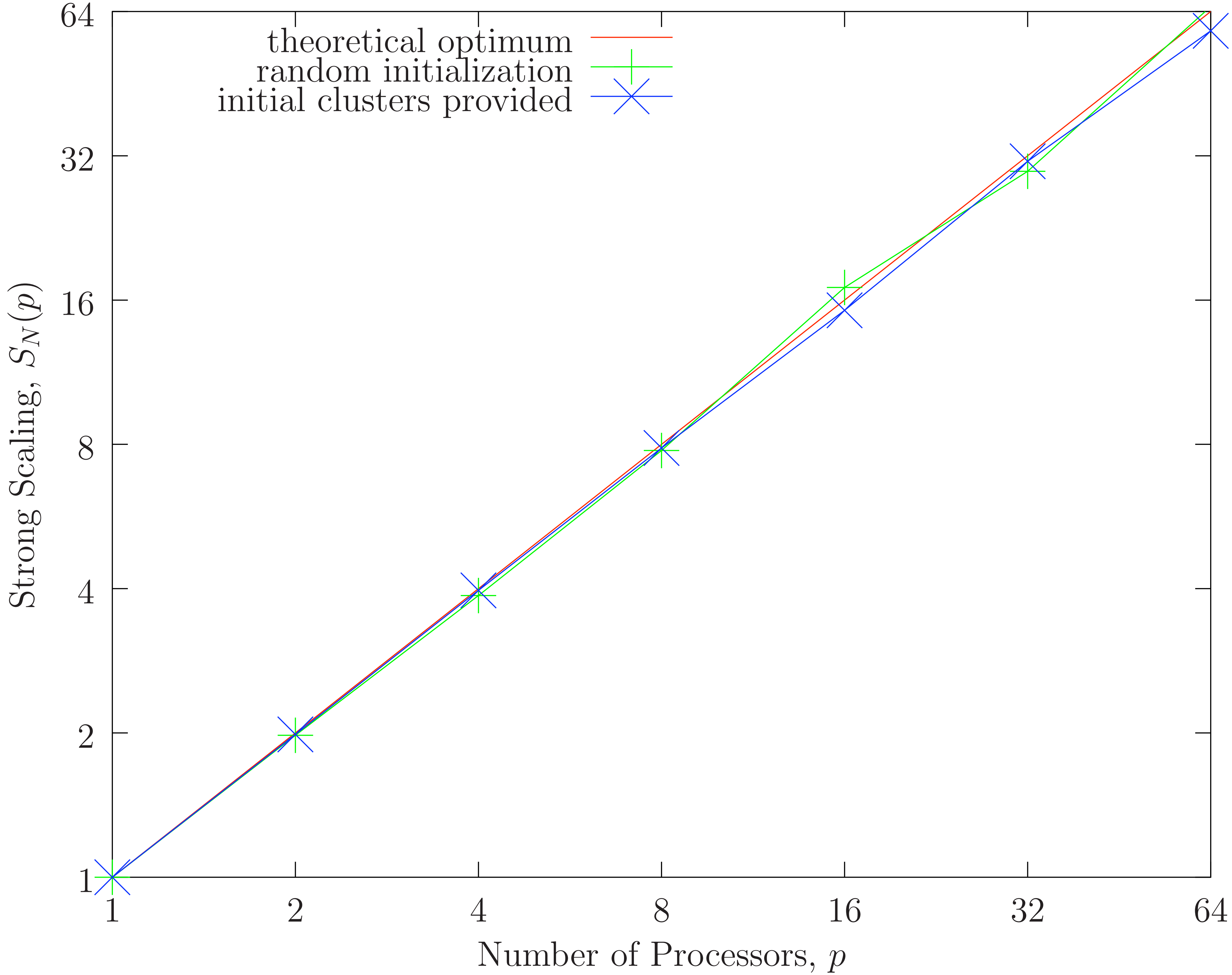
Assess

New raw observations

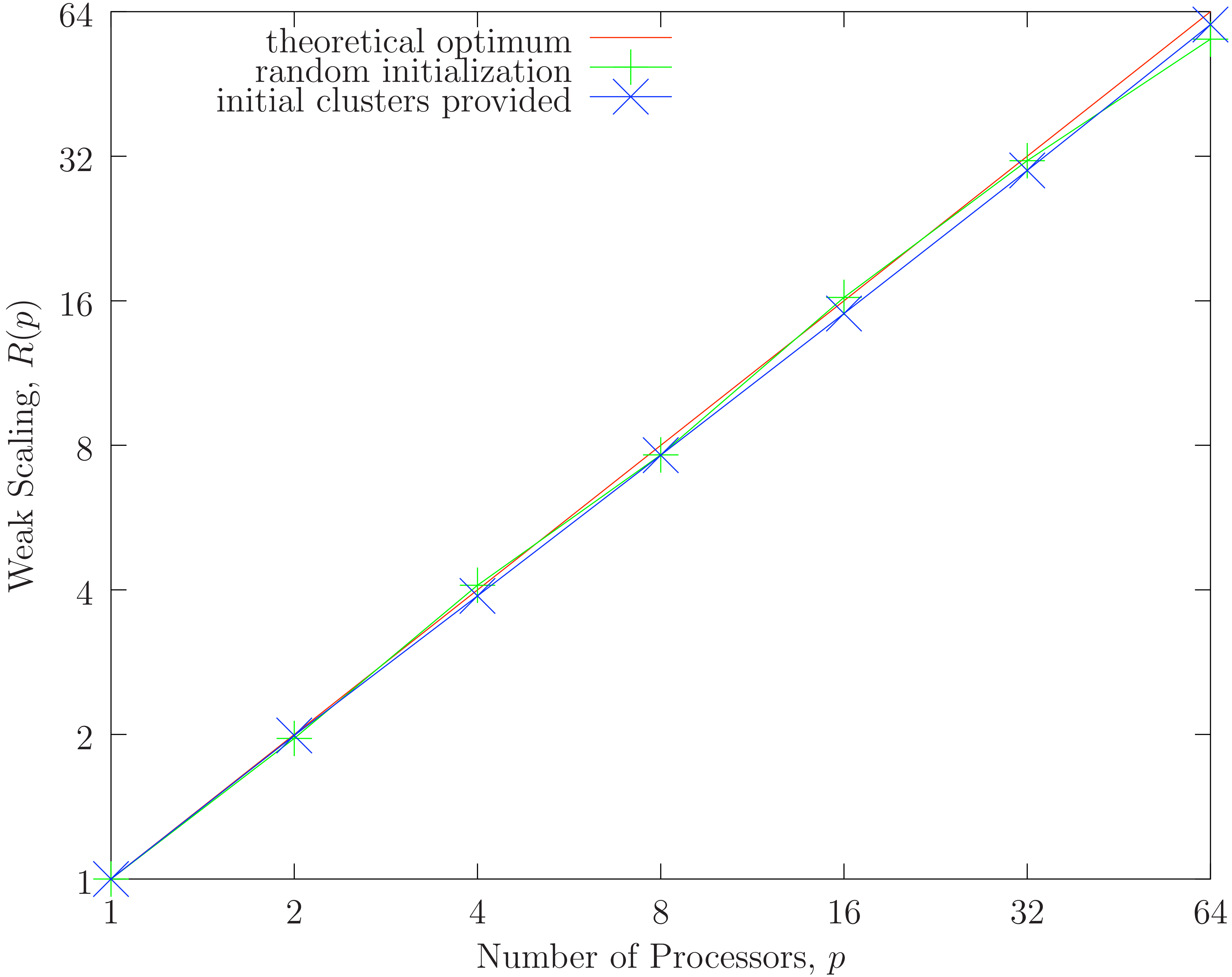
Quantile in which each observation lies (or 0 or q+1 if outside domain of CDF)

Assessed observations

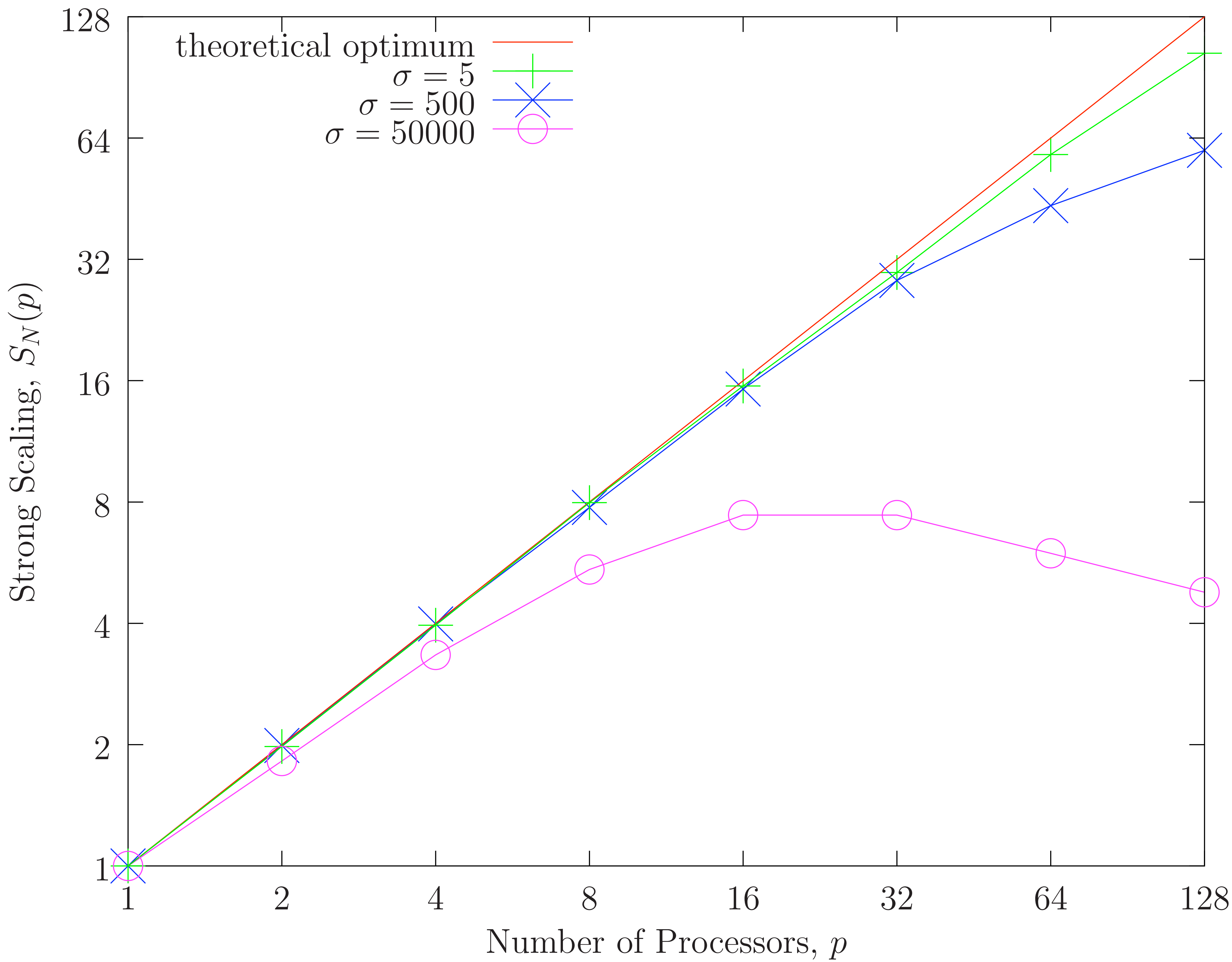
Strong Scaling: *k*-means



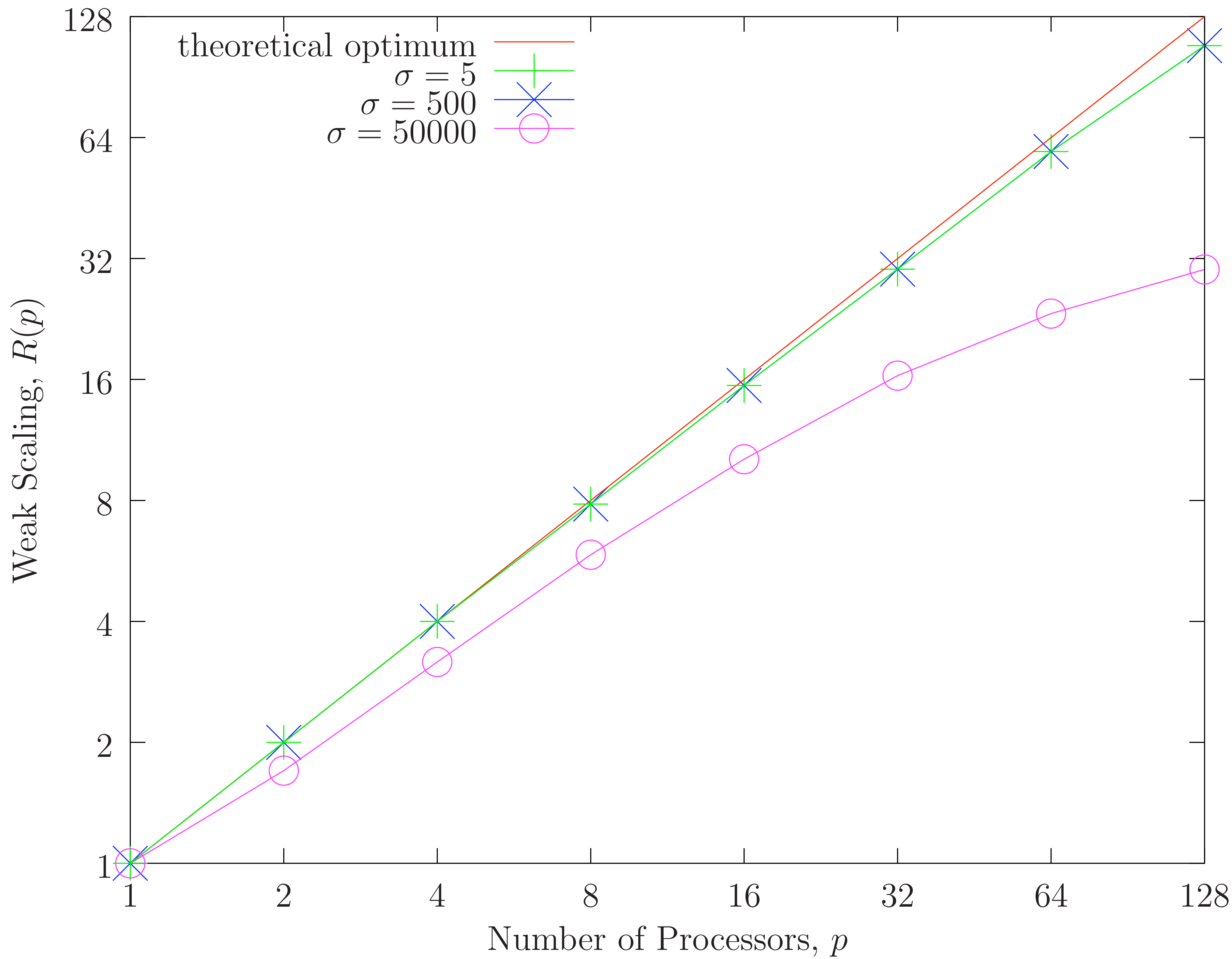
Weak Scaling: *k*-means



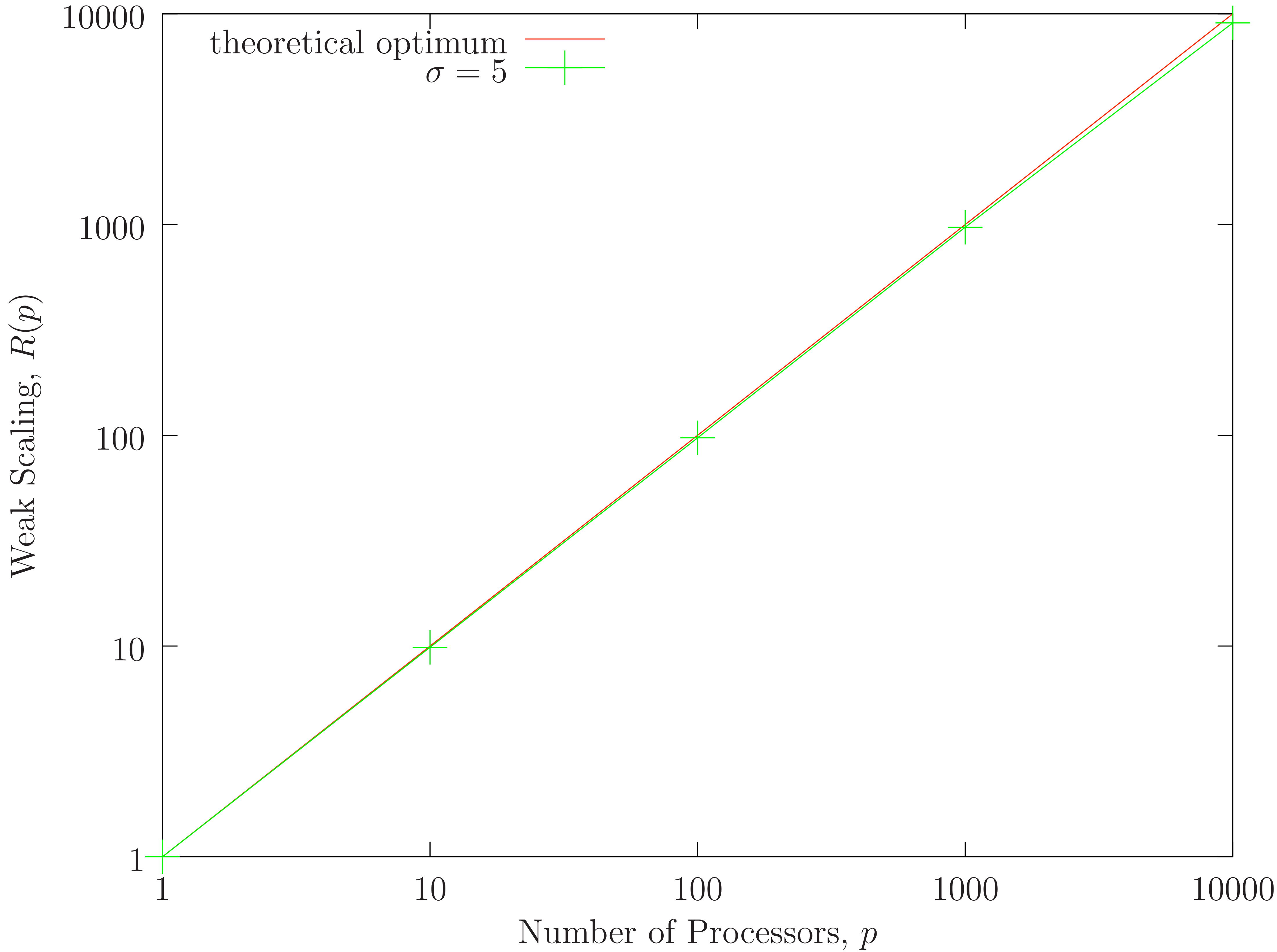
Strong Scaling: order



Weak Scaling: order



Weak Scaling: order



Efficiency

- Efficiency of quantizing techniques is

$$\mathcal{E}_{\text{global}} = 1 - \frac{N_q}{N}$$

$N_q \longleftarrow$ #(quanta)

$N \longleftarrow$ #(observations)

Engine	order statistics			contingency statistics		
σ	5	500	50000	5	50	200
$\mathcal{E}_g$	0.9999	0.9994	0.9619	0.9977	0.9866	0.8732

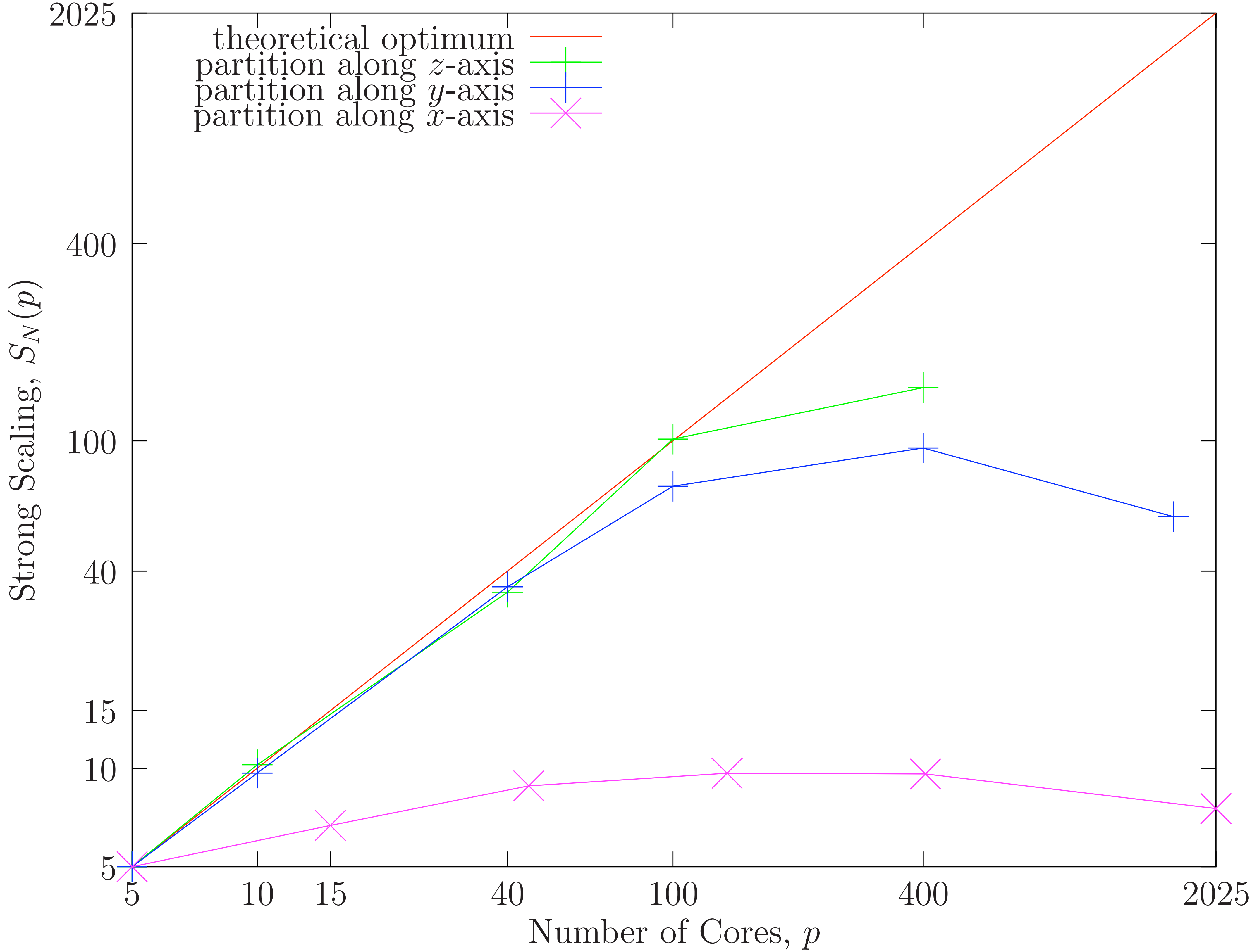
- Can be estimated by per-process efficiency

$$\mathcal{E}_{\text{global}} \approx \min_{p \in N_p} \left(\mathcal{E}_p\right)$$

Questions?

Questions?

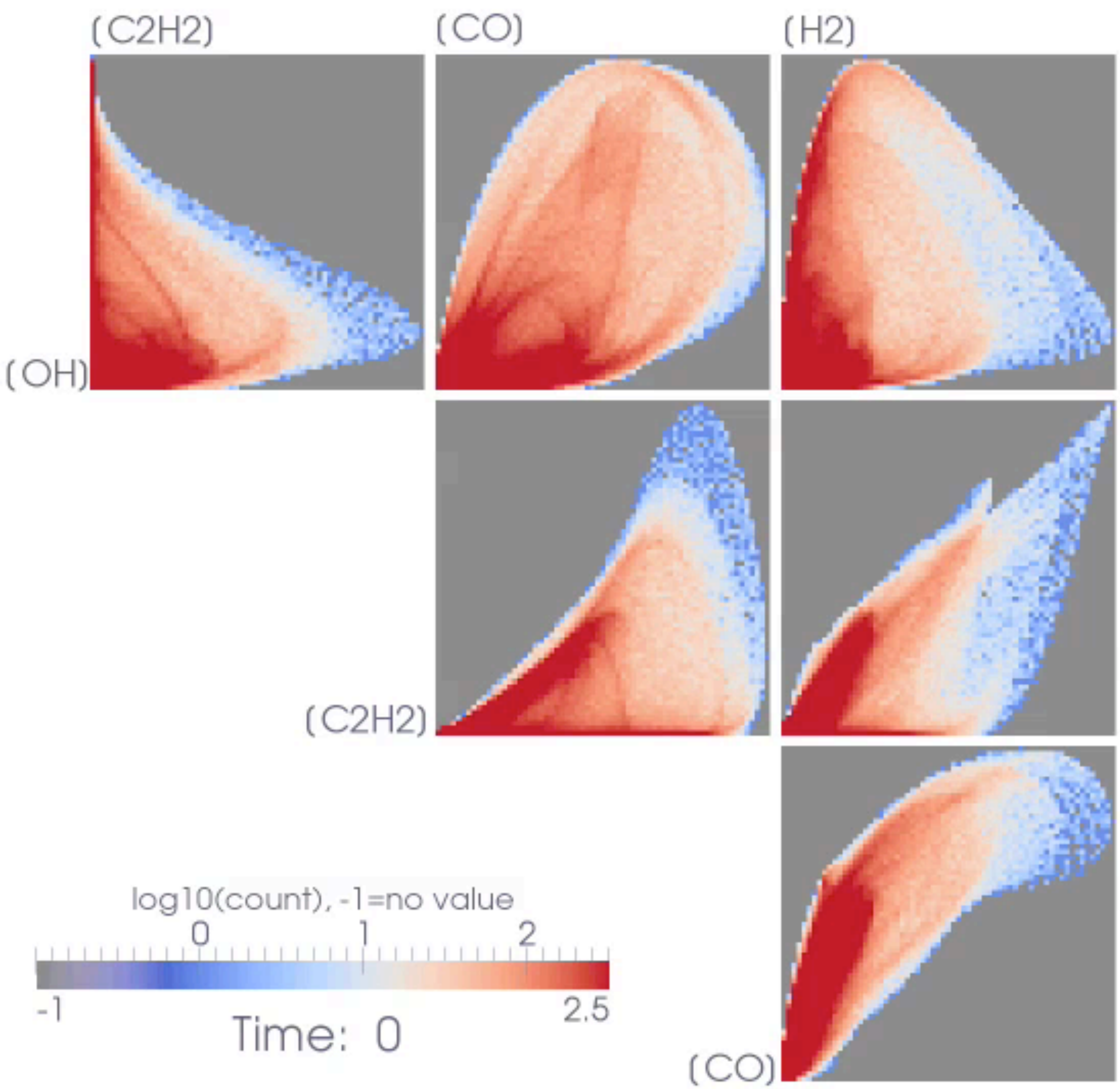
Weak scaling... with I/O



An application

- Combustion simulation
  - 21 chemical species
  - 2025x1600x400 grid; 15000 processes
  - Moment-preserving time steps
- Contingency statistics
  - 4 species in one reaction of interest
  - Concentrations normalized and bucketed
  - Want to test whether process is Poisson

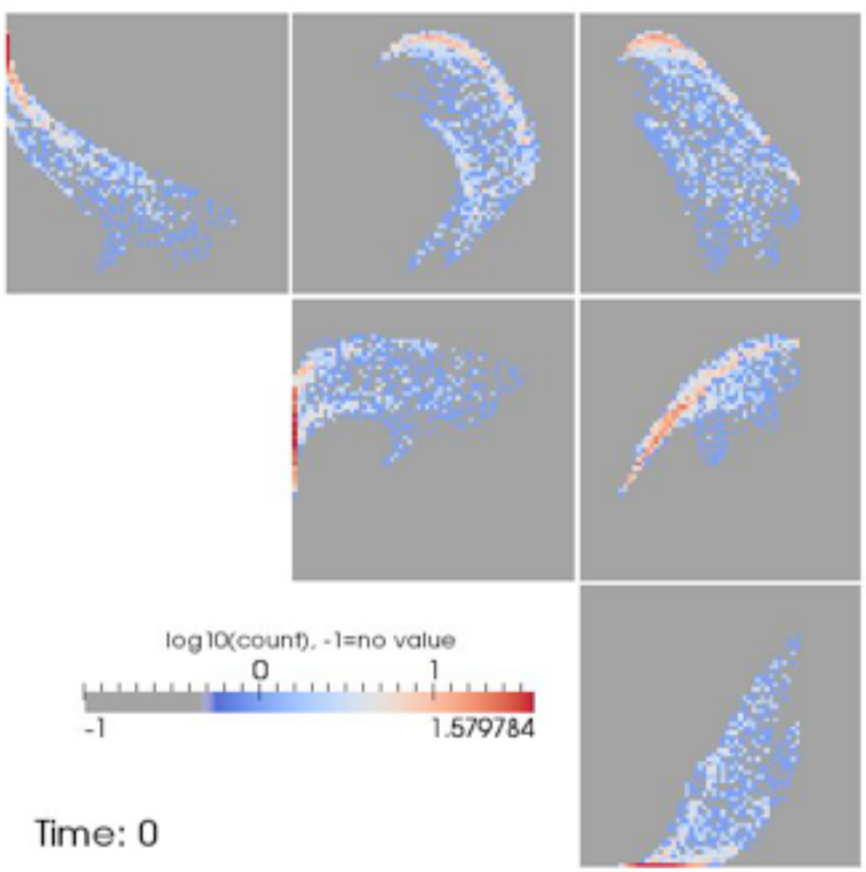
An application

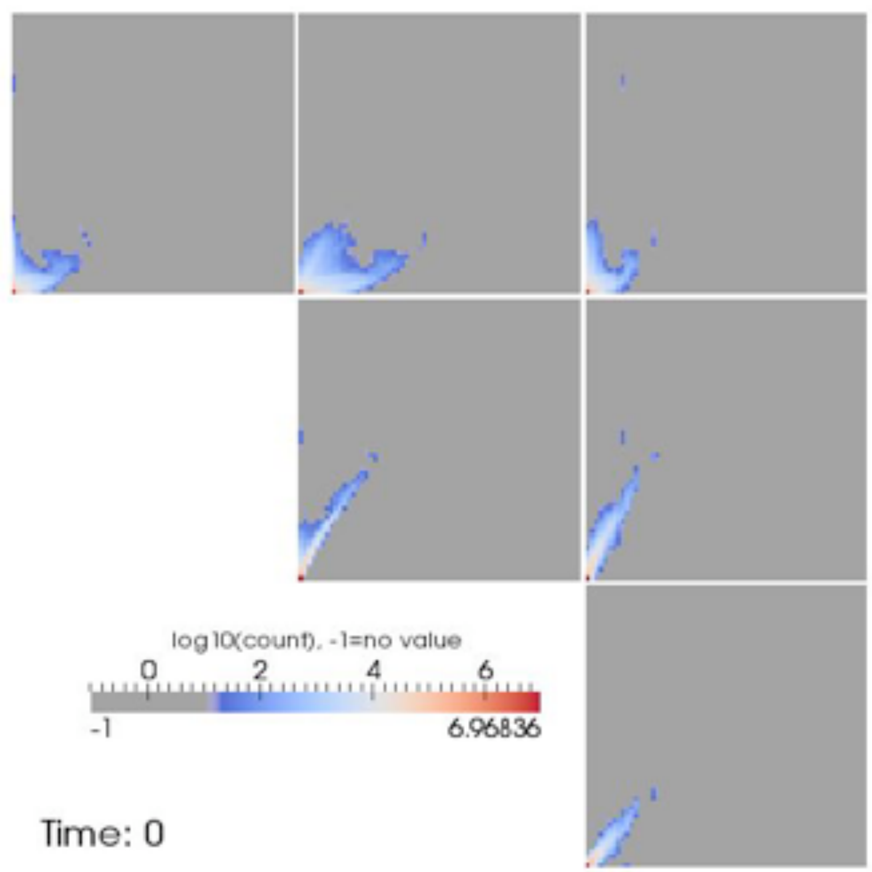


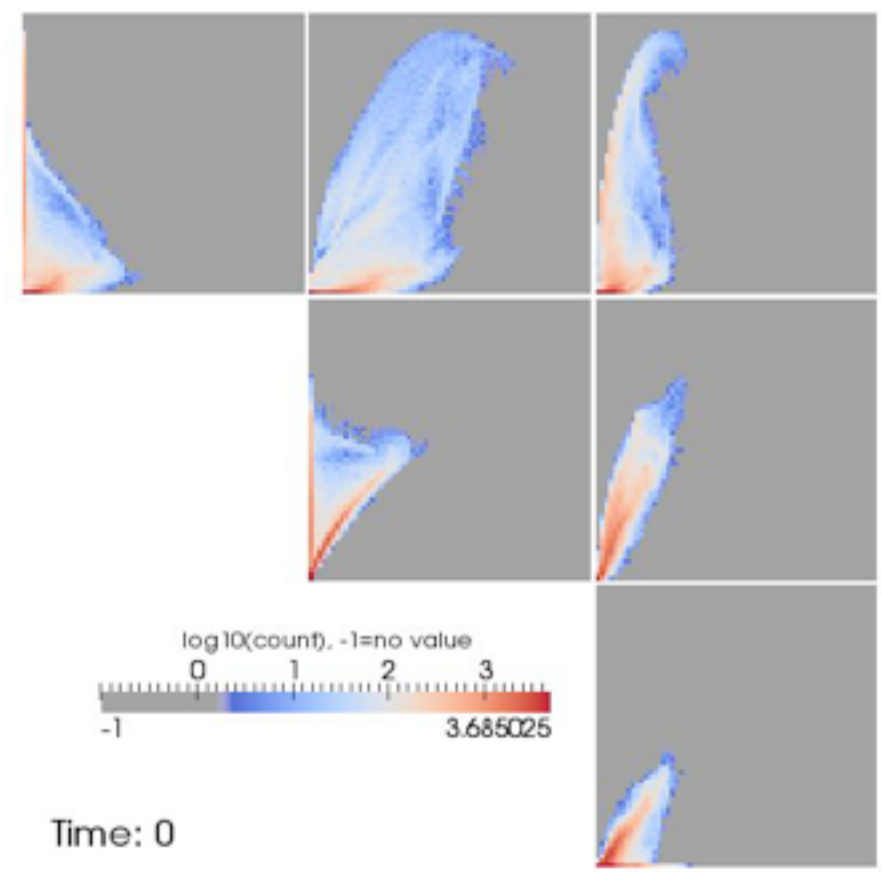
Data courtesty of Drs. Jacqueline Chen, Chun Sang Yoo, and Ray Grout.

An application

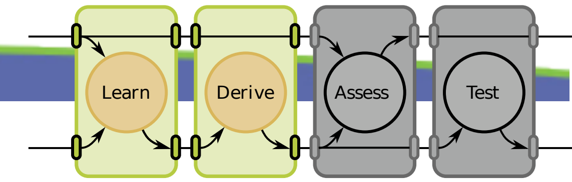
Are distributions over features meaningful?







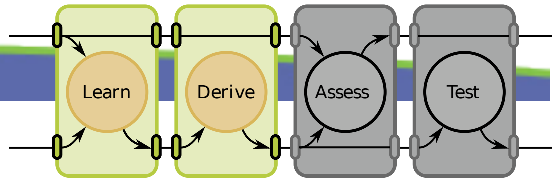
Learn and Derive



- Joint probability: $p_{X,Y}(x,y)$
- Marginal probabilities: $p_X(x),\ p_Y(y)$
  - Eliminates variables by summing.
  - For n variables, 2^n-1 marginal distributions.

port	protocol	port	protocol	port	protocol
80	HTTP	1122	HTTP	80	SMTP
80	HTTP	80	HTTP	20	FTP
80	HTTP	25	SMTP	20	FTP
80	HTTP	25	SMTP	20	FTP
80	HTTP	25	SMTP	122	FTP
80	HTTP	25	SMTP	20	FTP
8080	HTTP	25	SMTP	20	FTP

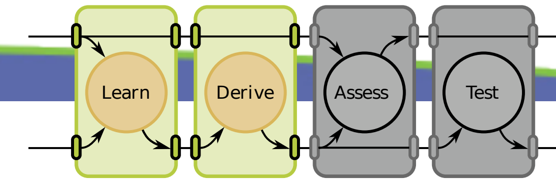
Learn and Derive



- Joint probability: $p_{X,Y}(x,y)$
- Marginal probabilities: $p_X(x),\ p_Y(y)$
  - Eliminates variables by summing.
  - For n variables, 2^n-1 marginal distributions.

count	HTTP	FTP	SMTP
20	0	5	0
25	0	0	5
80	7	0	1
122	0	1	0
1122	1	0	0
8080	1	0	0

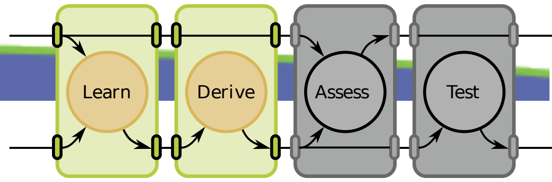
Learn and Derive



- Joint probability: $p_{X,Y}(x,y)$
- Marginal probabilities: $p_X(x),\ p_Y(y)$
  - Eliminates variables by summing.
  - For n variables, 2^n-1 marginal distributions.

$p_{\text{prt,prtcl}}$	HTTP	FTP	SMTP	p_{prt}
20	0	0.238	0	0.238
25	0	0	0.238	0.238
80	0.333	0	0.0476	0.381
122	0	0.0476	0	0.0476
1122	0.0476	0	0	0.0476
8080	0.0476	0	0	0.0476
p_{prtcl}	0.429	0.286	0.286	

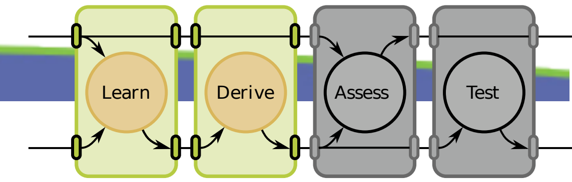
Learn and Derive



- Conditional probability: $p_{Y|X}(y|x)$
  - Normalizes joint probability by marginal.
  - Undefined when marginal probability is 0.

$p_{\mathrm{prt,prtcl}}$	HTTP	FTP	SMTP	p_{prt}
20	0	0.238	0	0.238
25	0	0	0.238	0.238
80	0.333	0	0.0476	0.381
122	0	0.0476	0	0.0476
1122	0.0476	0	0	0.0476
8080	0.0476	0	0	0.0476
p_{prtcl}	0.429	0.286	0.286	

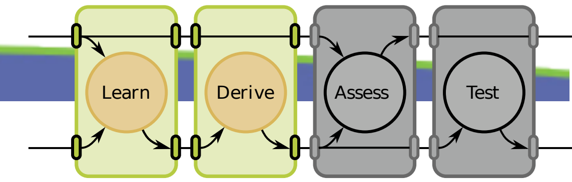
Learn and Derive



- Conditional probability: $p_{Y|X}(y|x)$
  - Normalizes joint probability by marginal.
  - Undefined when marginal probability is 0.

$p_{\text{prtcl}\vert\text{prt}}$	HTTP	FTP	SMTP
20	0	1	0
25	0	0	1
80	0.875	0	0.125
122	0	1	0
1122	1	0	0
8080	1	0	0

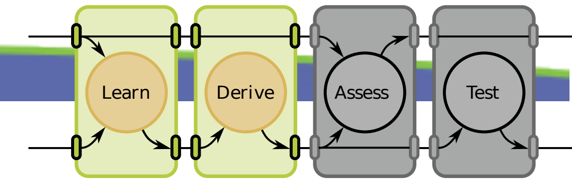
Learn and Derive



- Conditional probability: $p_{Y|X}(y|x)$
  - Normalizes joint probability by marginal.
  - Undefined when marginal probability is 0.

$p_{\mathrm{prt,prtcl}}$	HTTP	FTP	SMTP	p_{prt}
20	0	0.238	0	0.238
25	0	0	0.238	0.238
80	0.333	0	0.0476	0.381
122	0	0.0476	0	0.0476
1122	0.0476	0	0	0.0476
8080	0.0476	0	0	0.0476
p_{prtcl}	0.429	0.286	0.286	

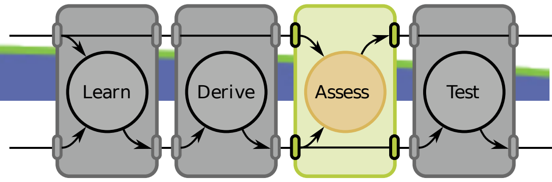
Learn and Derive



- Conditional probability: $p_{Y|X}(y|x)$
    - Normalizes joint probability by marginal.
    - Undefined when marginal probability is 0.

$p_{\mathrm{prt}\|\mathrm{prtcl}}$	HTTP	FTP	SMTP
20	0	0.833	0
25	0	0	0.833
80	0.778	0	0.167
122	0	0.167	0
1122	0.111	0	0
8080	0.111	0	0

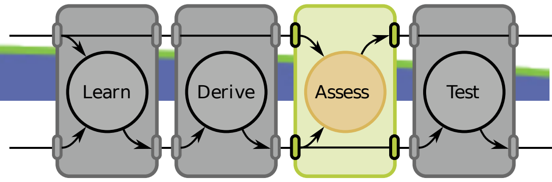
Assess



- Pointwise mutual information:
  - $\mathrm{PMI}(x, y) = \log \dfrac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}$
  - Association of x and y relative to model

$p_{\mathrm{prt,prtcl}}$	HTTP	FTP	SMTP	p_{prt}
20	0	0.238	0	0.238
25	0	0	0.238	0.238
80	0.333	0	0.0476	0.381
122	0	0.0476	0	0.0476
1122	0.0476	0	0	0.0476
8080	0.0476	0	0	0.0476
p_{prtcl}	0.429	0.286	0.286	

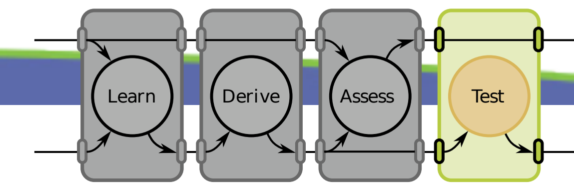
Assess



- Pointwise mutual information:
  - $$\mathrm{PMI}(x, y) = \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)}$$
  - Association of x and y relative to model

PMI(prt, prtcl)	HTTP	FTP	SMTP
20	$-\infty$	1.26	$-\infty$
25	$-\infty$	$-\infty$	1.26
80	0.714	$-\infty$	-0.827
122	$-\infty$	1.26	$-\infty$
1122	0.847	$-\infty$	$-\infty$
8080	0.847	$-\infty$	$-\infty$

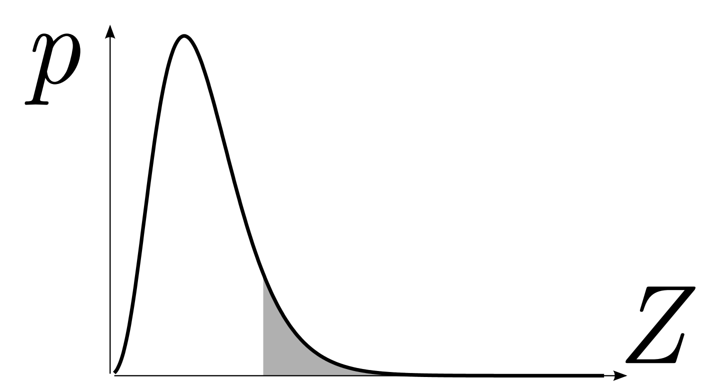
Test



- Null hypothesis: variables are independent
- Compare difference between joint and product of marginal frequencies to test independence

$$Z = N \sum_{x,y} \frac{(N N_{x,y} - N_x N_y)^2}{N_x N_y}$$

$$Z \sim \chi^2_{(m-1)(n-1)}$$



- Integrate tail to get probability of observed Z statistic of at least this size.
- Small probability indicates independence is unlikely.