A Proposed Method for Scaling of ~~Identification False Match Rates~~ False Positive Identification Rates (FPIR) Using Extreme Value Theory

Michael Schuckers
St. Lawrence University
schuckers@stlawu.edu

International Biometric Performance Conference
NIST
March 8, 2012

Scaling of Performance

- Known problem
- Identification FMR (FPIR)
- How to predict performance from DB of 10^3x10^3 to DB of 10^7x10^7
- How does threshold for FPIR=0.01 change?

Previous Bioauthentication Work

- Large-Scale identification System Design

Hervé]arosz,]ean-Christophe Fondeur and Xavier Dupré

Empirical (Regression line)

Ident as N Verifications

Extreme Value Theory (G. Pareto Dist)

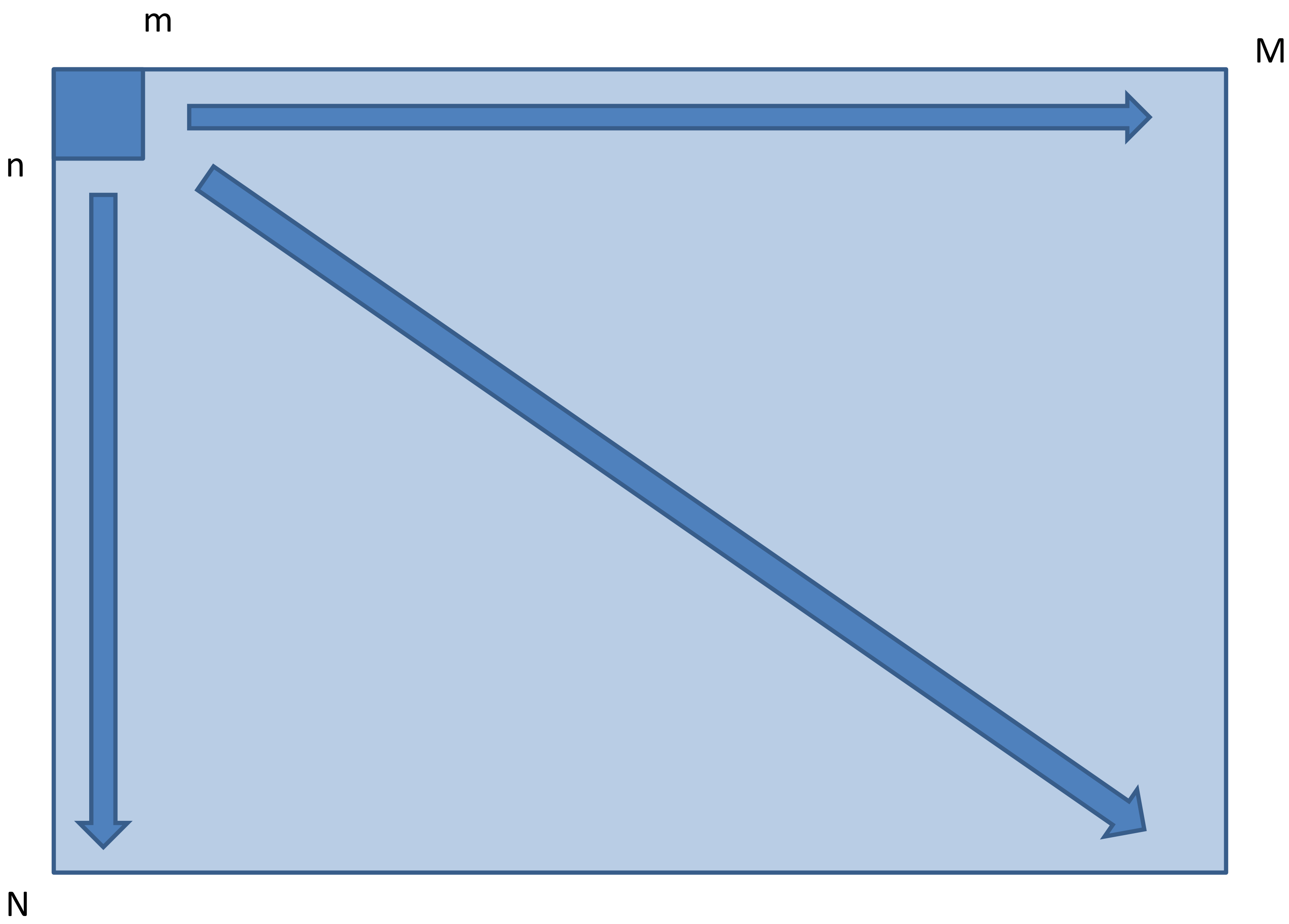
Modeling Distance

Notation

- Y_{ij} be match score from a comparison of sample from individual i compared to individual j.

- $X_{i:m} = \max_j \{Y_{ij}\}$, $1 \leq i \leq n$

- Goal is prediction of distribution of $X_{i:M}$


m
M
n
N

Extreme Value Theory

- If there exists sequences of constants $\{a_m>0\}$ and $\{b_m\}$ such that

$$P\left(\frac{X_{i:m} - b_m}{a_m} \leq z\right) \to G(z)$$

As m $\to\infty$ then G belongs to one of three families of distributions:

Gumbel, Fréchet, Weibull.

Limiting Distributions

- Gumbel $G(z) = \exp(-\exp(-(x-b)/a))$

- Fréchet $G(z) = \exp(-((x-b)/a)^{-\alpha})$

  if z>b, G(z) = 0 o/w

- Weibull $G(z) = \exp(-(-((x-b)/a)^{\alpha}))$

  if z<b, G(z) = 1 o/w

Comment

- Stuart Coles:

“ The remarkable feature of this result is that the three types of extreme value distributions are the only possible limits for the distribution of [maximums] regardless of the distribution F for the population.”

Combining Limiting Distributions

Generalized Extreme Value (GEV) Distribution

Cdf given by

$$G(z) = \exp\{-[1+\xi((z-\mu)/\sigma)]^{-1/\xi}]\}$$

$$\text{if } 1+\xi\,(z-\mu)/\sigma>0$$

Return Level (Target)

Return Level z_p is the value that will be exceeded with probability p. i.e. $P(X_{1:M}>z_p)=p$

$$z_p=G(1-p)$$

$$z_p = \begin{cases} \mu-\sigma/\xi(1-(-\ln(1-p))^{-\xi}) & \text{if } \xi \neq 0 \\ \mu-\sigma(1-(-\ln(1-p))) & \text{if } \xi = 0 \end{cases}$$

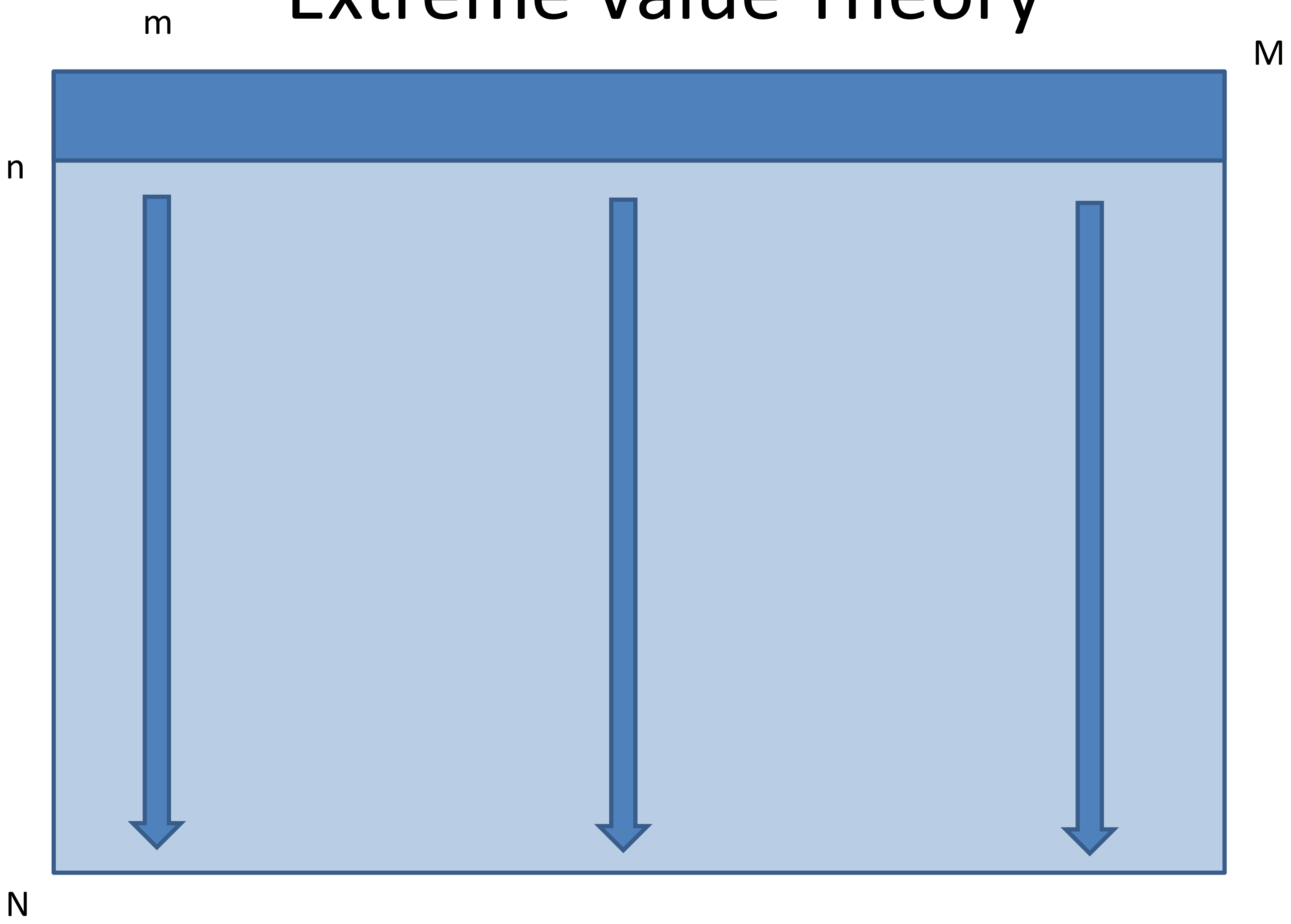
Goals

Take $X_{1:m}$, ..., $X_{n:m}$

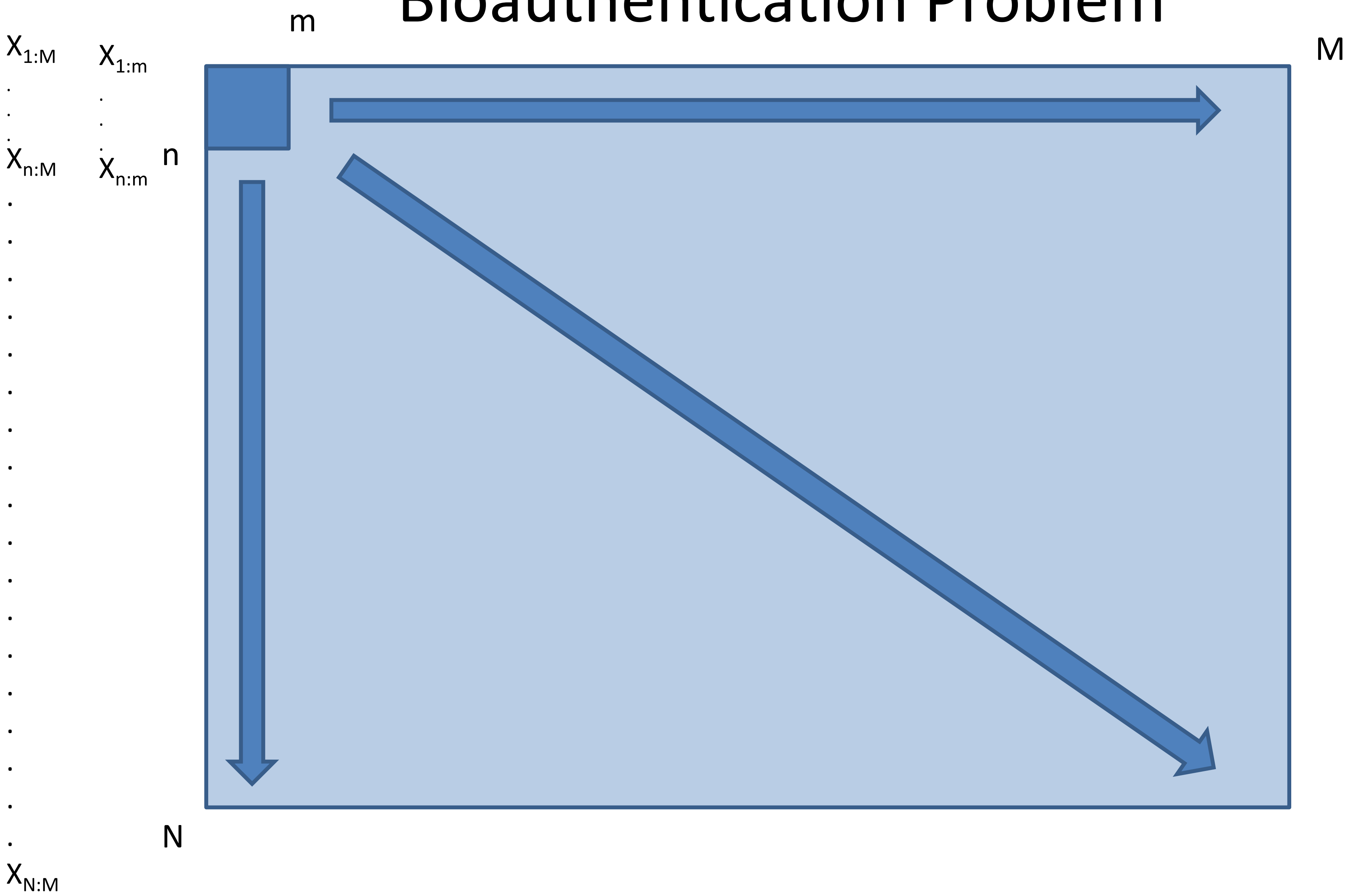
Estimate distribution of $X_{i:M}$'s and find z_p from this.

Note: $z_{0.01}$ is value of scores that gives iFMR of 0.01.

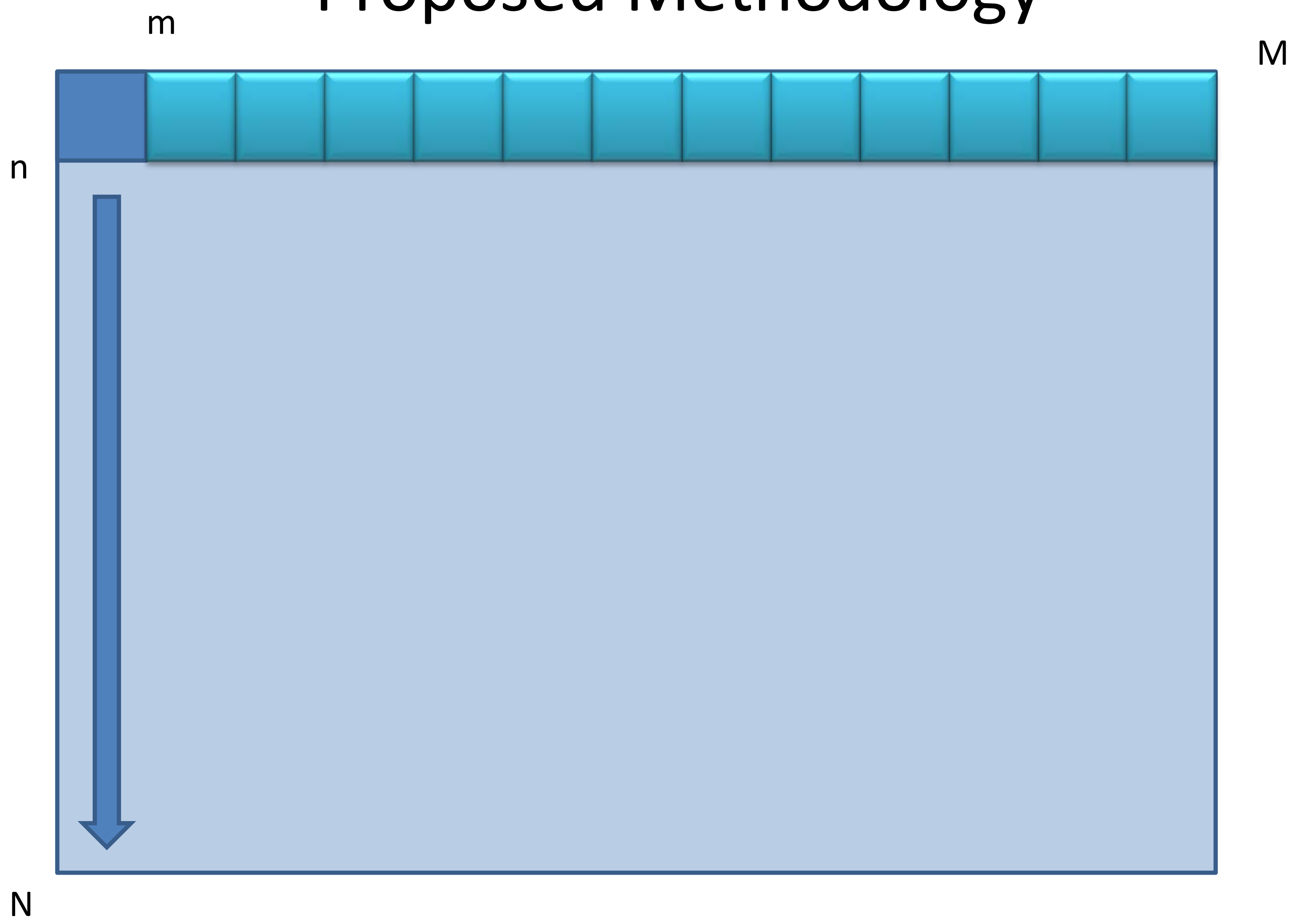
Extreme Value Theory



Bioauthentication Problem



Proposed Methodology



Proposed Methodology

1. Fit a GEV to the n observed $X_{i:m}$'s to get MLE's of μ, σ, ξ.

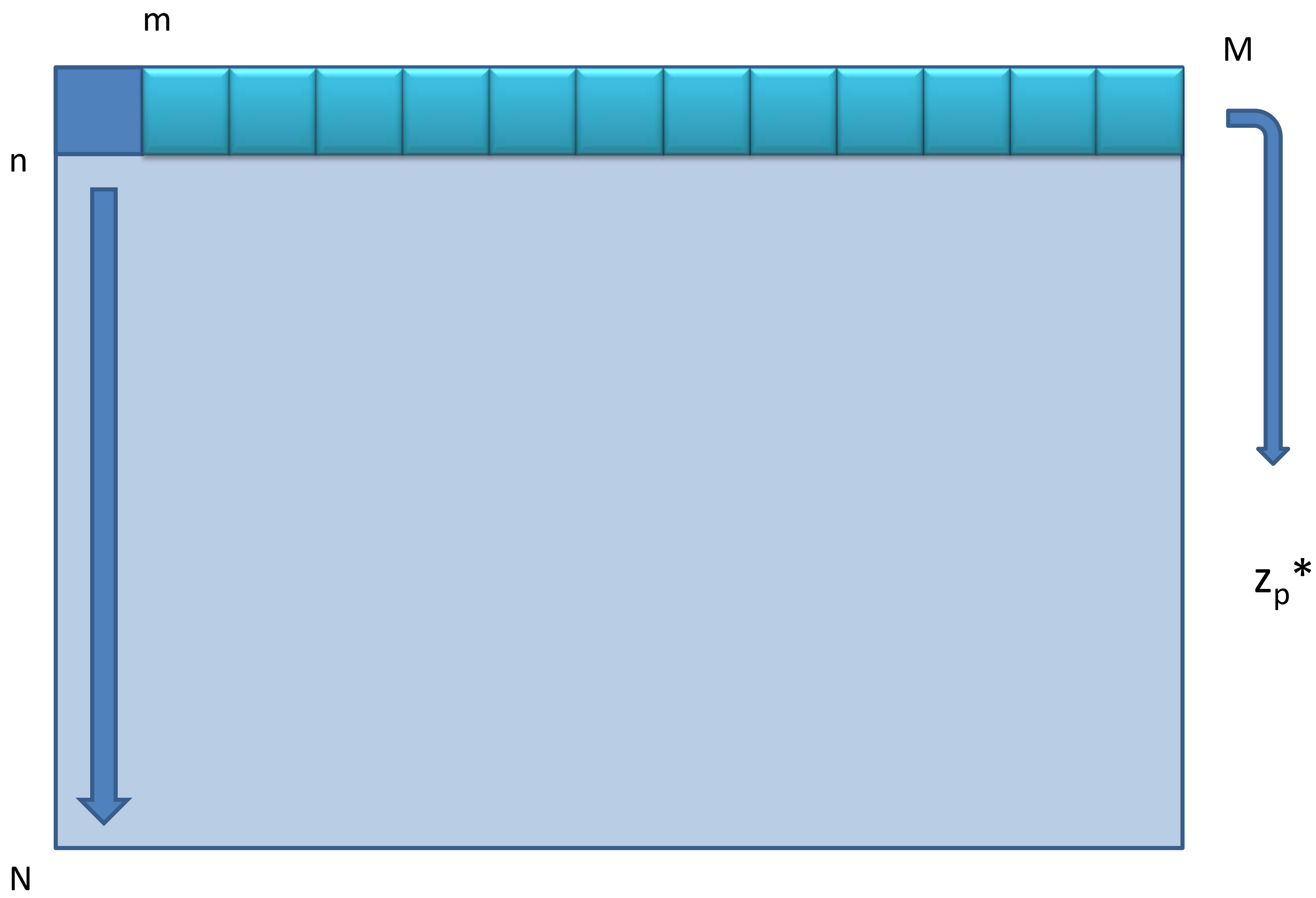
2. For each i, $1 \le i \le n$, generate k = ceil(M/m)-1 values from a GEV with parameters $\hat{\mu} + e_1 s_{\hat{\mu}}, \hat{\sigma} + e_2 s_{\hat{\sigma}}, \hat{\xi} + e_3 s_{\hat{\xi}}$.

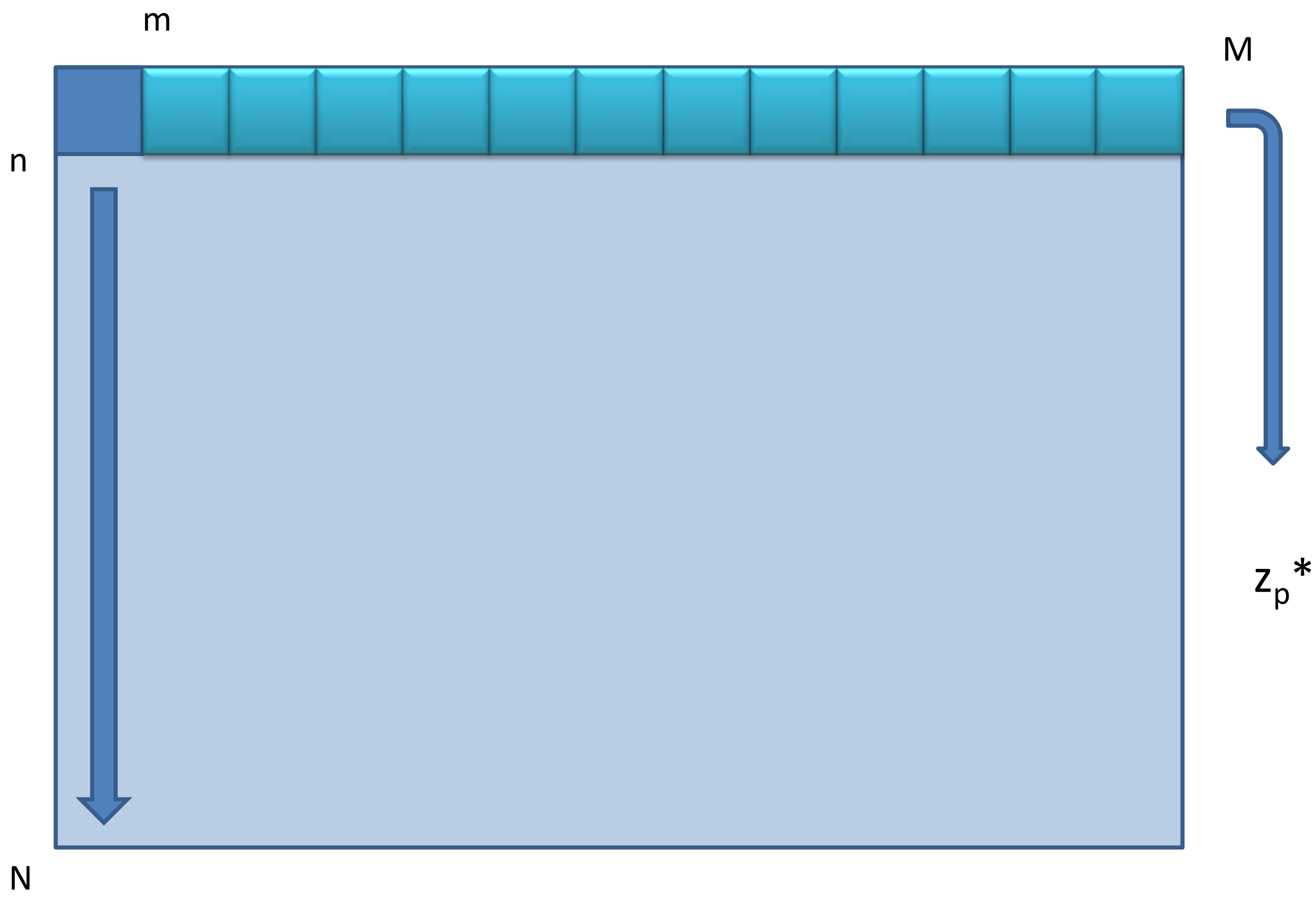
3. Find $X^*_i = \max\{X_{i:m}, X^*_{i1}, X^*_{i2}, \ldots, X^*_{ik}\}$

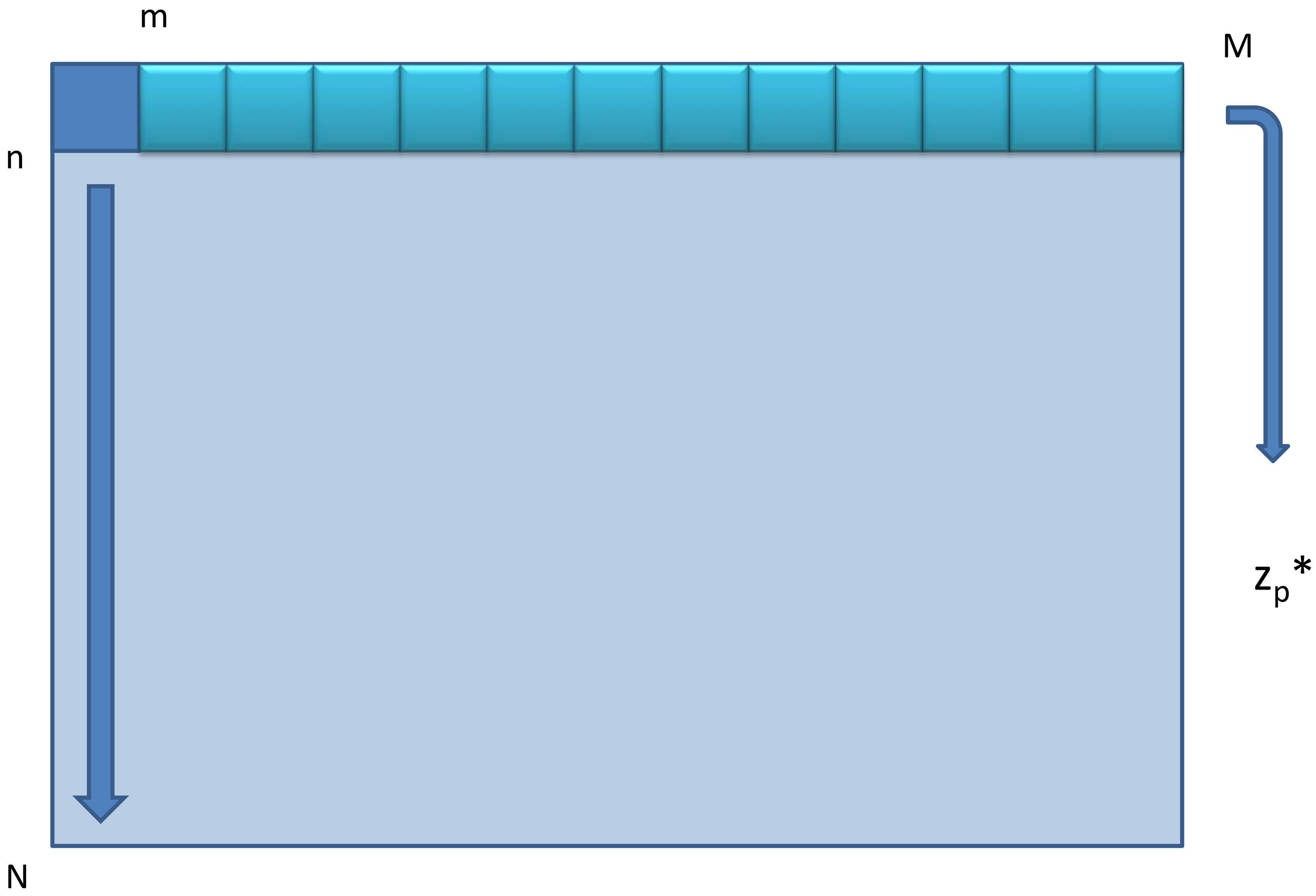
4. Fit a GEV distribution to the *n* X^*_i's and get MLE's. Call these estimates $\hat{\mu}^*, \hat{\sigma}^*, \hat{\xi}^*$. Use these estimates to get z_p^*

5. Repeat steps 2 to 4, saving z_p^* each time.

6. Use distribution of z_p^* to make inference for FPIR=p.


m
M
n
N
z_p*


m
M
n
N
z_p*


m
M
n
N
z_p*

Testing via Simulation/Synthetic

1. Create N x N database of scores
   – Calculate known z_p
2. Randomly sample n from {1, …., N}
3. Run model on selected n x n database
4. Create 95% CI for z_p
5. Repeat Steps 2) to 4) 100 times
6. Determine % of times z_p inside CI's.

Gaussian (mean =35)

Database Size	Sample Size	Stand. Deviation	P (FPIR)	Coverage
5000	500	10	0.01	0.95
5000	500	10	0.005	0.94
5000	500	5	0.01	0.92
5000	100	10	0.01	0.94
5000	100	5	0.01	0.93
5000	100	10	0.005	0.92*
10000	100	10	0.005	0.90

Gamma(mean=500,stddev=225)

Database Size	Sample Size	p(FPIR)	Coverage
5000	500	0.01	0.96
5000	500	0.001	0.94
5000	100	0.01	0.96
5000	100	0.005	0.96

Summary

New method, theoretically grounded

Up to 2(?) orders of magnitude, good performance

Need to test on ‘real’ data

Need to look at more orders of magnitude.

Limitations

- Issues with correlation (?)
- Needs more testing
  - Synthetic: different distributions
  - ‘Real’ data
- More orders of magnitude to test
- Changing data collection process
- Address binning, etc.

Thank You!

schuckers@stlawu.edu