Thomas J. Watson Research Center

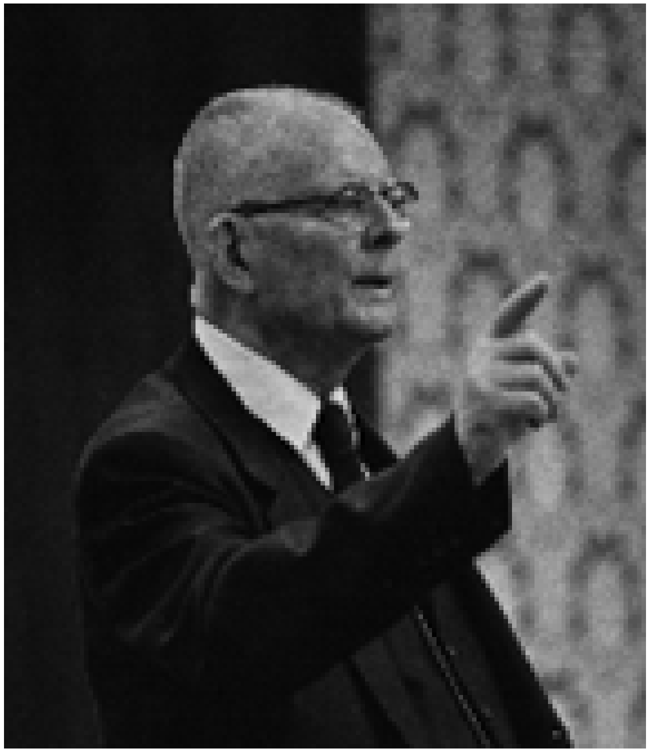
Fear, Uncertainty and Statistics

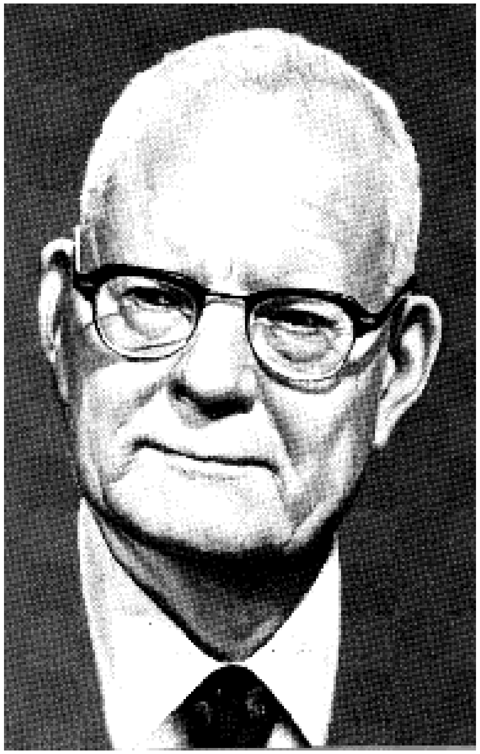
Chandu Visweswariah

Acknowledgements

- The extended statistical timing, statistical optimization, timing support and timing methodology teams at IBM Yorktown, Fishkill, Burlington, Poughkeepsie, Toronto and Waltham

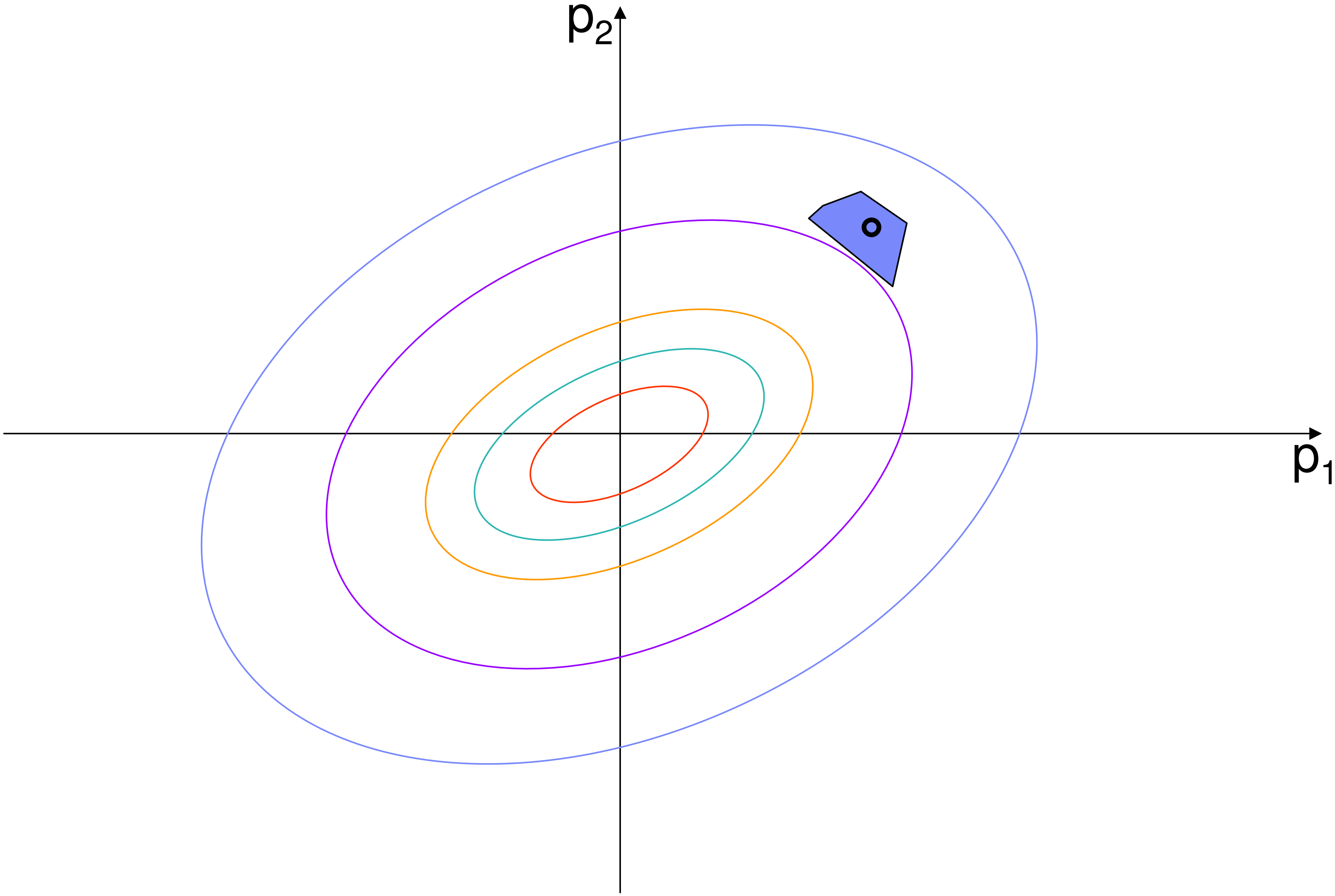
W. Edwards Deming (1900 —1993)



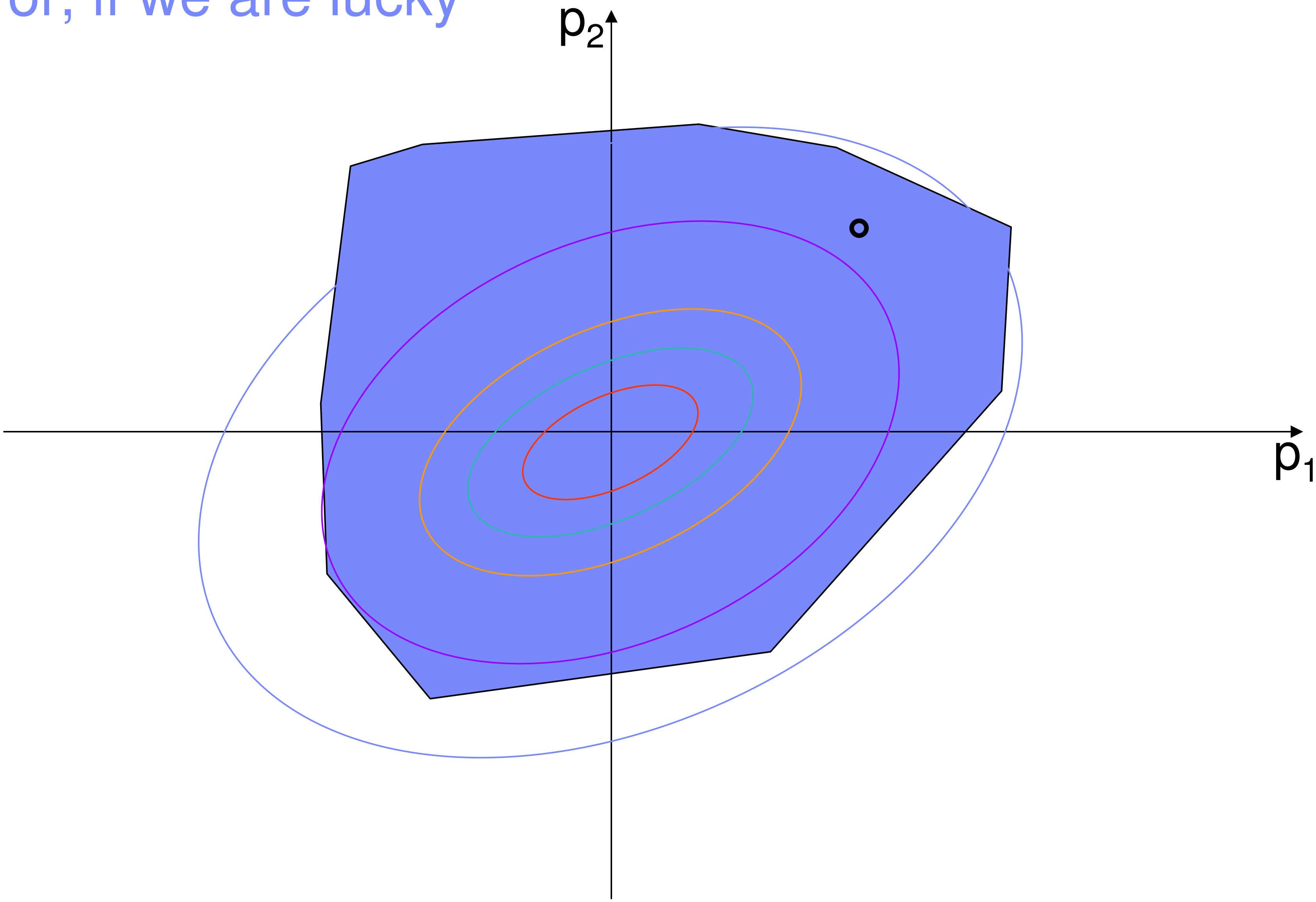


“Statistical theory has changed practice in almost every [discipline].”

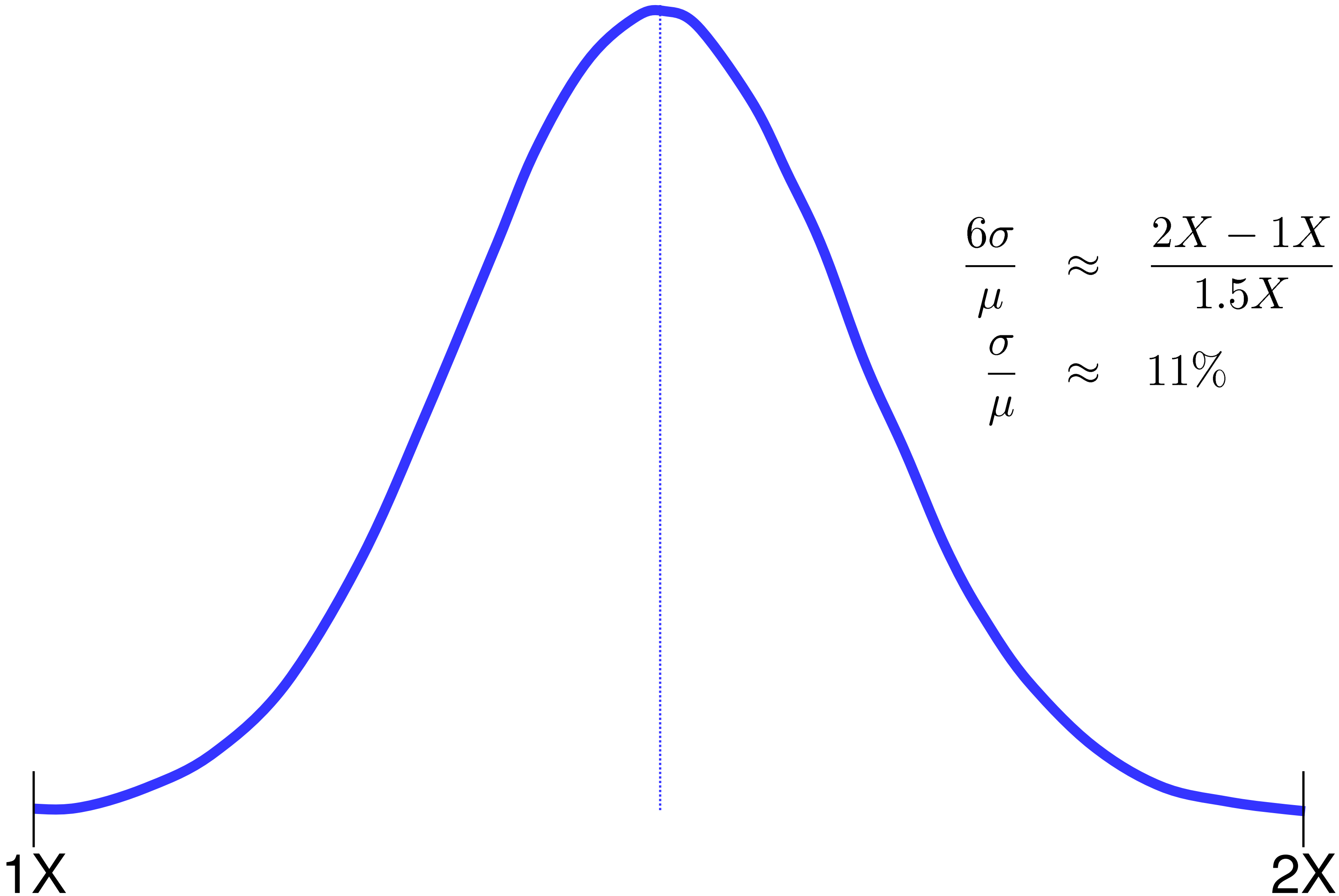
What’s wrong with what we do today?



… or, if we are lucky



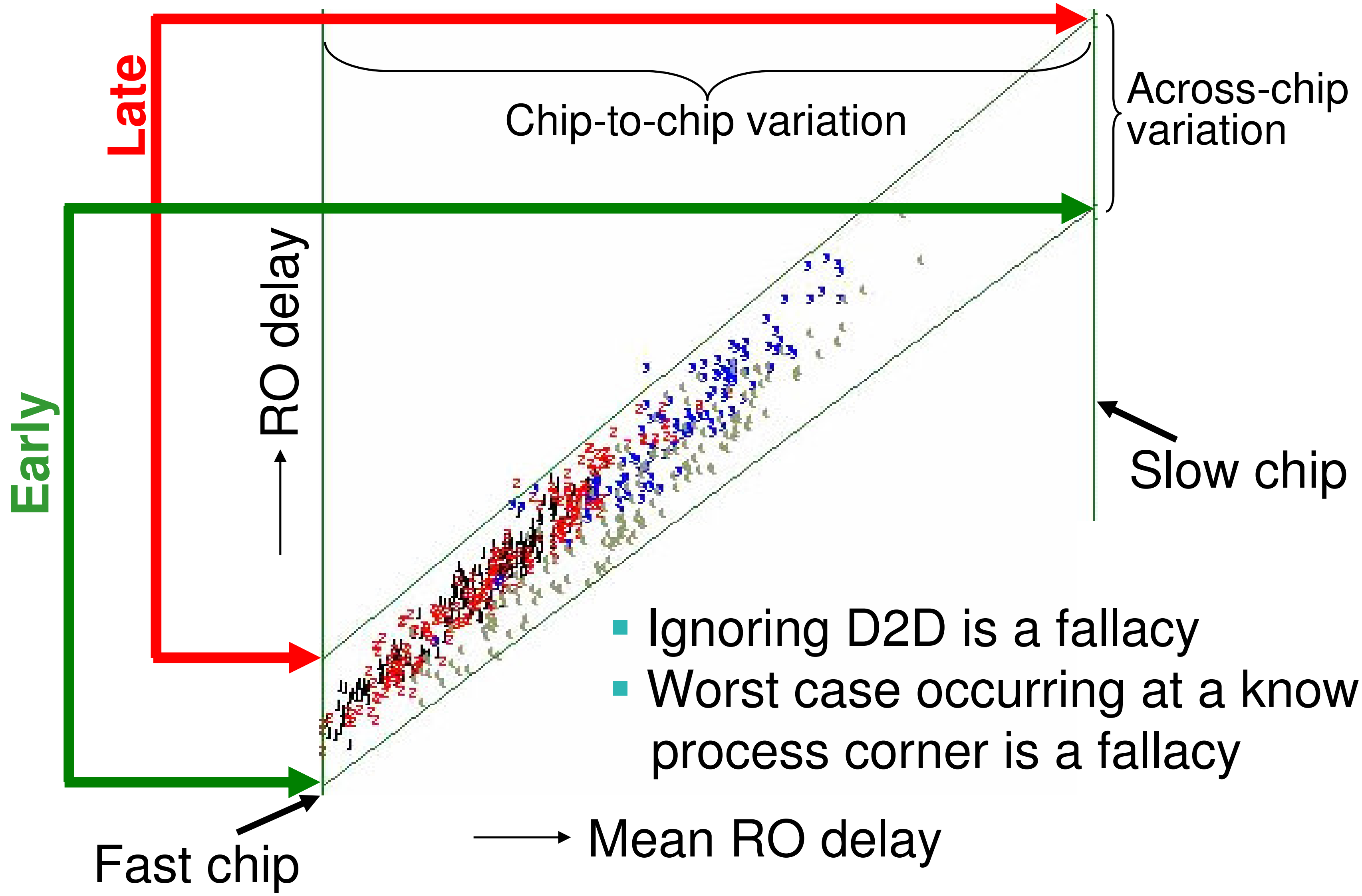
Hardware performance is a distribution



$$\frac{6\sigma}{\mu} \approx \frac{2X - 1X}{1.5X}$$

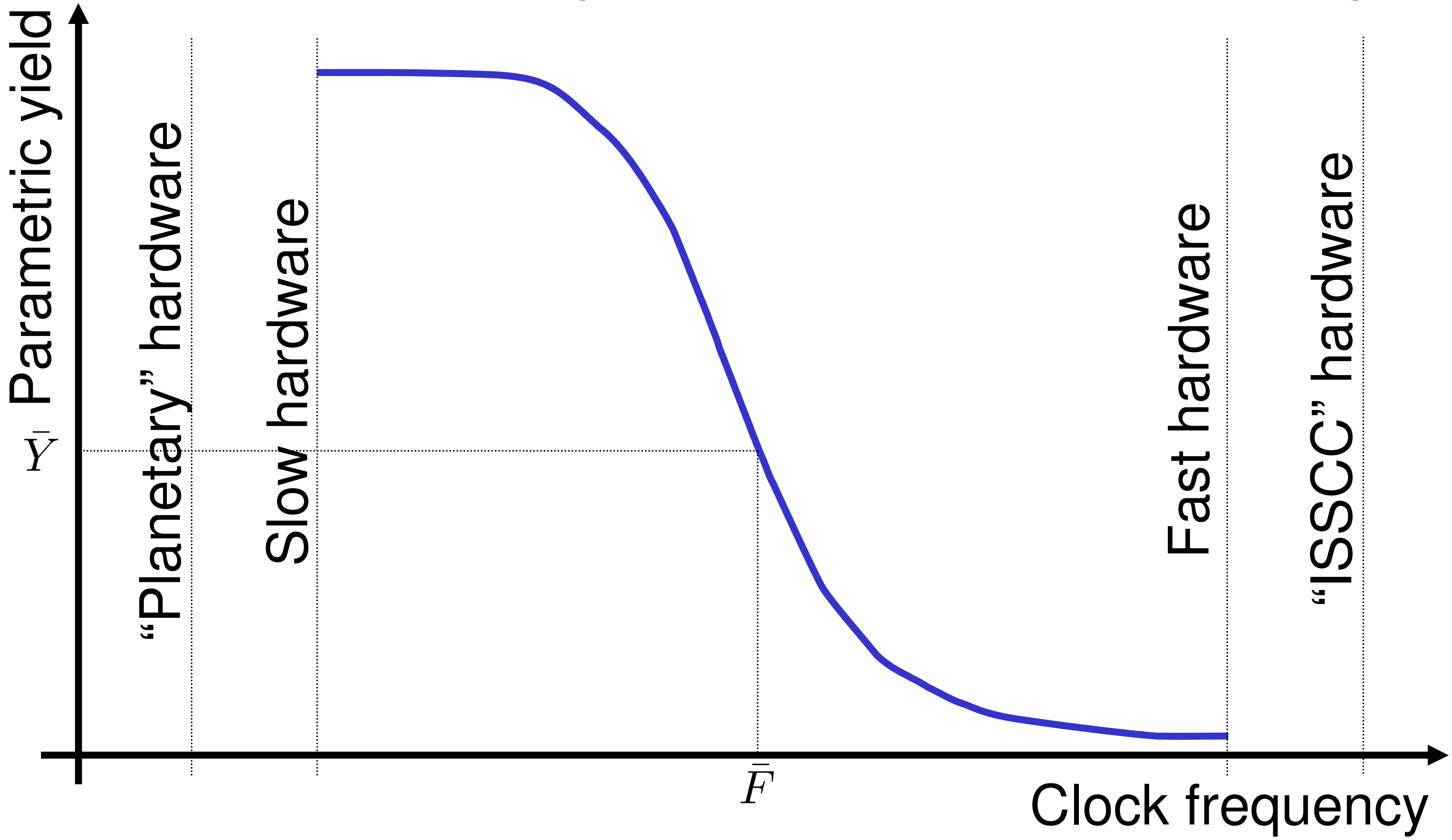
$$\frac{\sigma}{\mu} \approx 11\%$$

Die-to-die vs. within-die variability



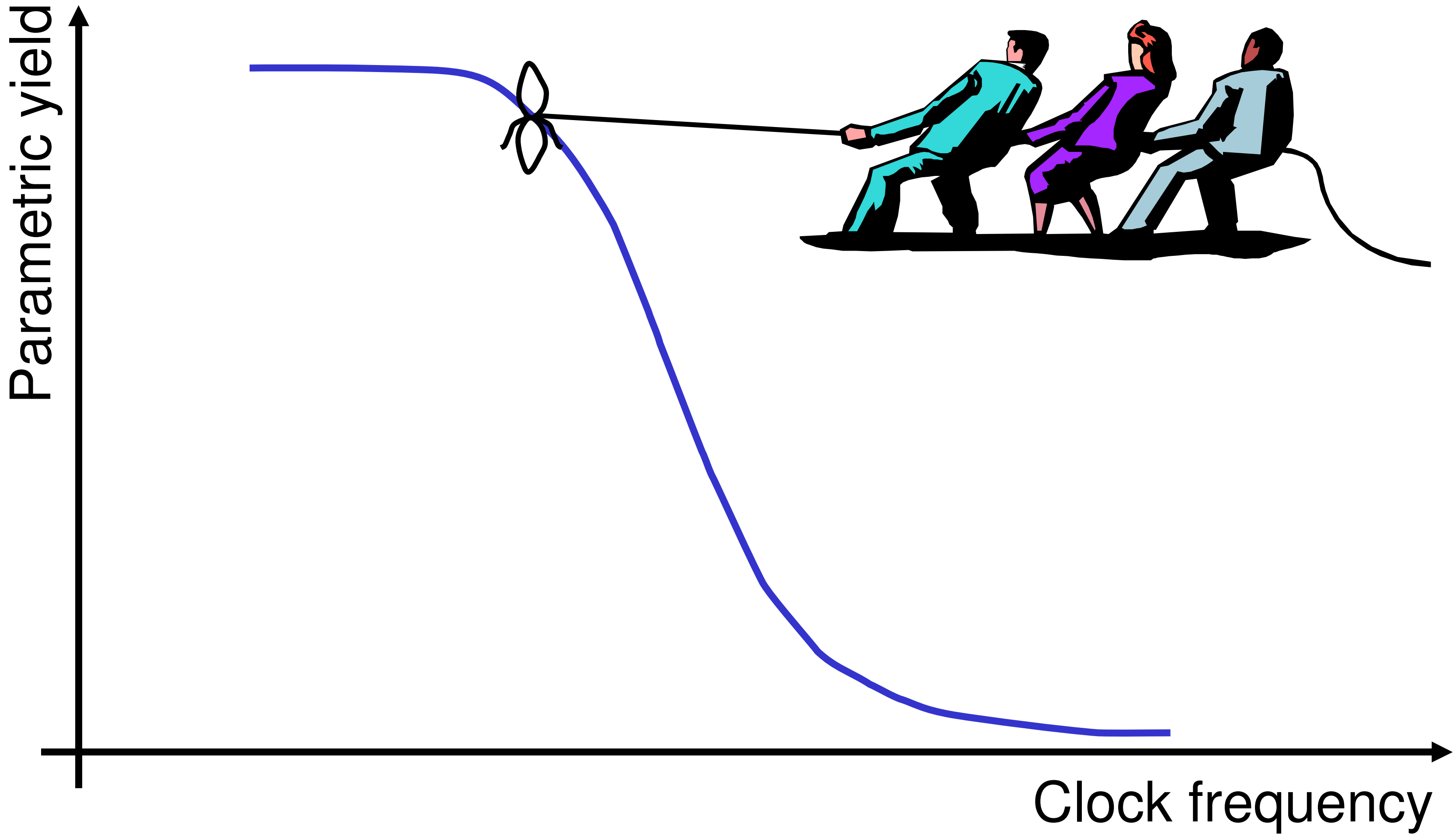
- Ignoring D2D is a fallacy
- Worst case occurring at a known process corner is a fallacy

More understandable representation of hardware spread

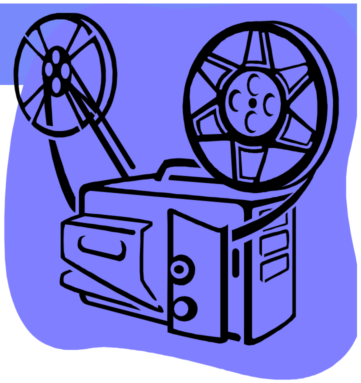


$$\bar{F} = \text{Frequency at a given yield}$$
$$\bar{Y} = \text{Yield at a given frequency}$$

The lasso method: a scalar at any cost!



Canonical parameterized slack → single value?



Projection methods

$$\text{Canonical parameterized slack} = a_0 + \sum_{i=1}^{n} a_i \Delta X_i + a_R \Delta X_R$$

$$k \text{ sigma value} = a_0 - k \left(\sum_{i=1}^{n} a_i^2 + a_R^2 \right)^{1/2}$$

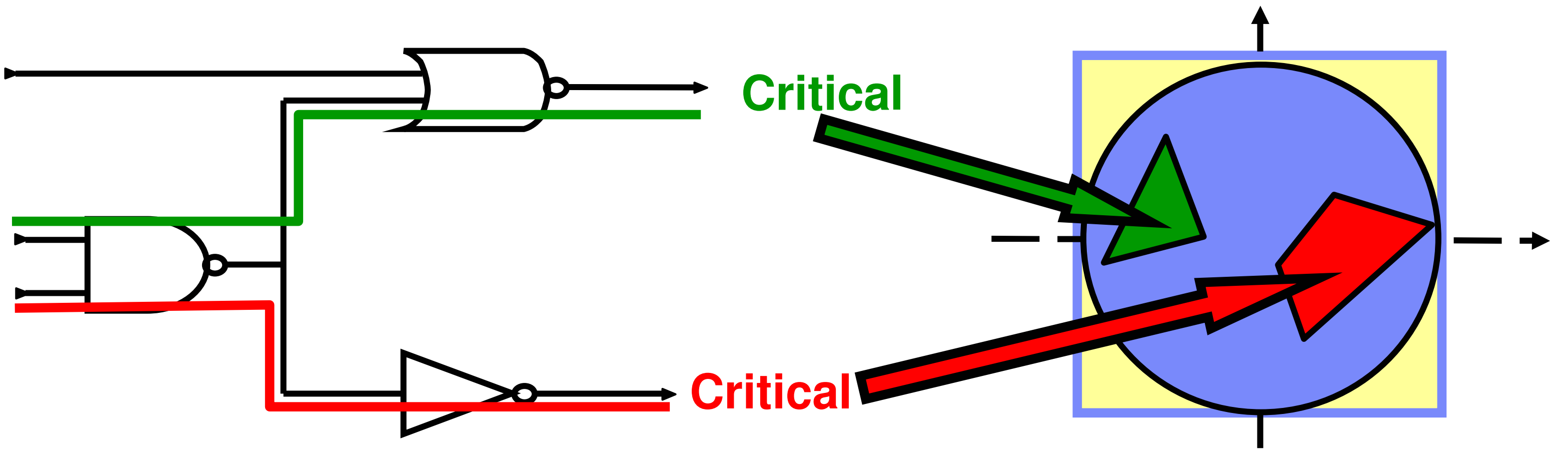
$$\text{Worst–case projection} = a_0 - k \left(\sum_{i=1}^{n} |a_i| + |a_R| \right)$$

$$\text{Per–variable worst–case projection} = a_0 - \sum_{i=1}^{n} k_i |a_i| - k_R |a_R|$$

$$\text{Per–corner worst–case projection} = a_0 + \sum_{i=1}^{n} \min\left(k_i^1 a_i, k_i^2 a_i\right) - k_R |a_R|$$

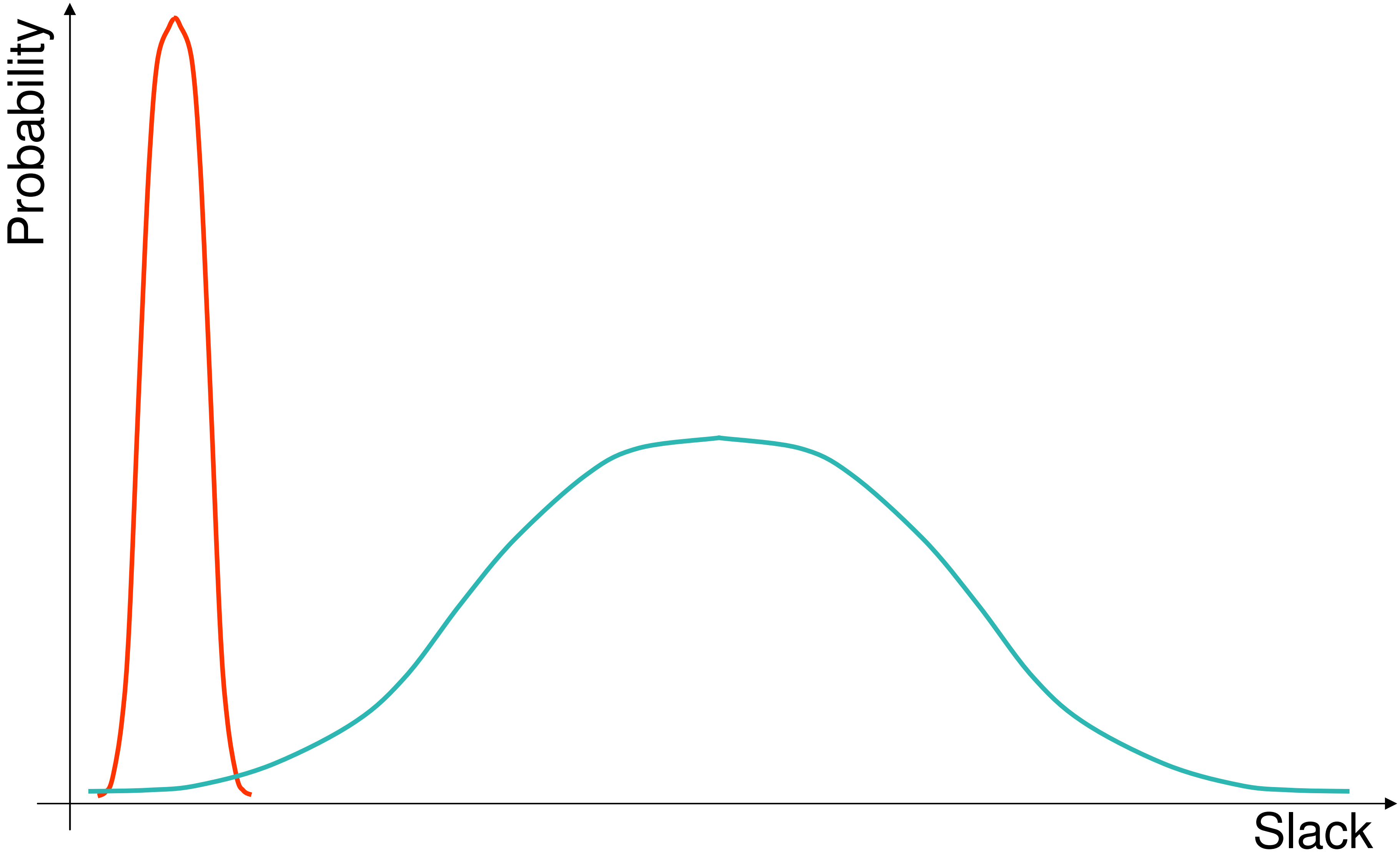
$$\text{Hybrid projection} = a_0 - k_1 \sum_{i=1}^{m} |a_i| - k_2 \sum_{i=m+1}^{n} \left(a_i^2 + a_R^2\right)^{1/2}$$

Non-uniqueness of critical path

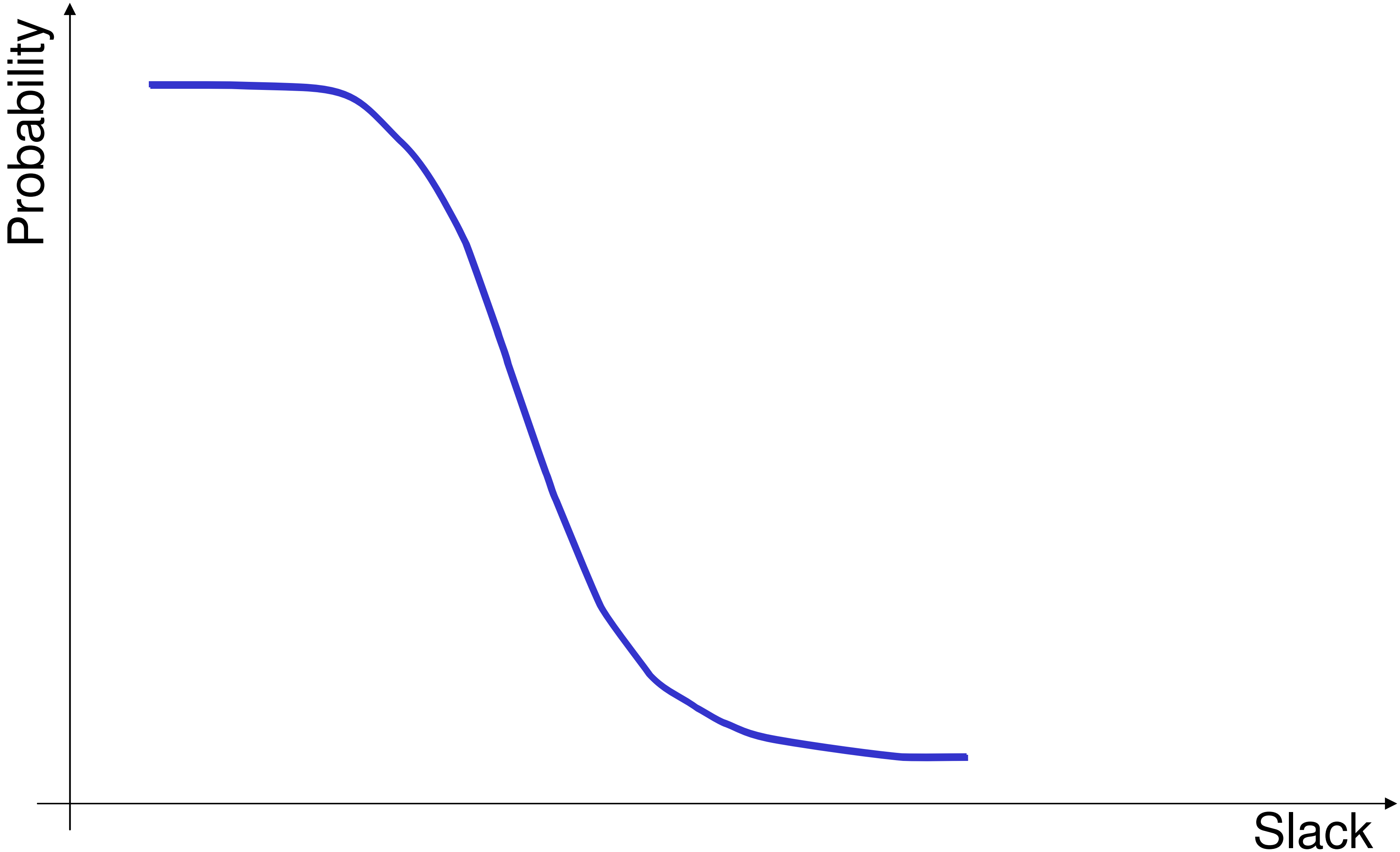


- **Which path to attack?**
- **Slack continuity on the critical path?**
- **How to judge improvement?**

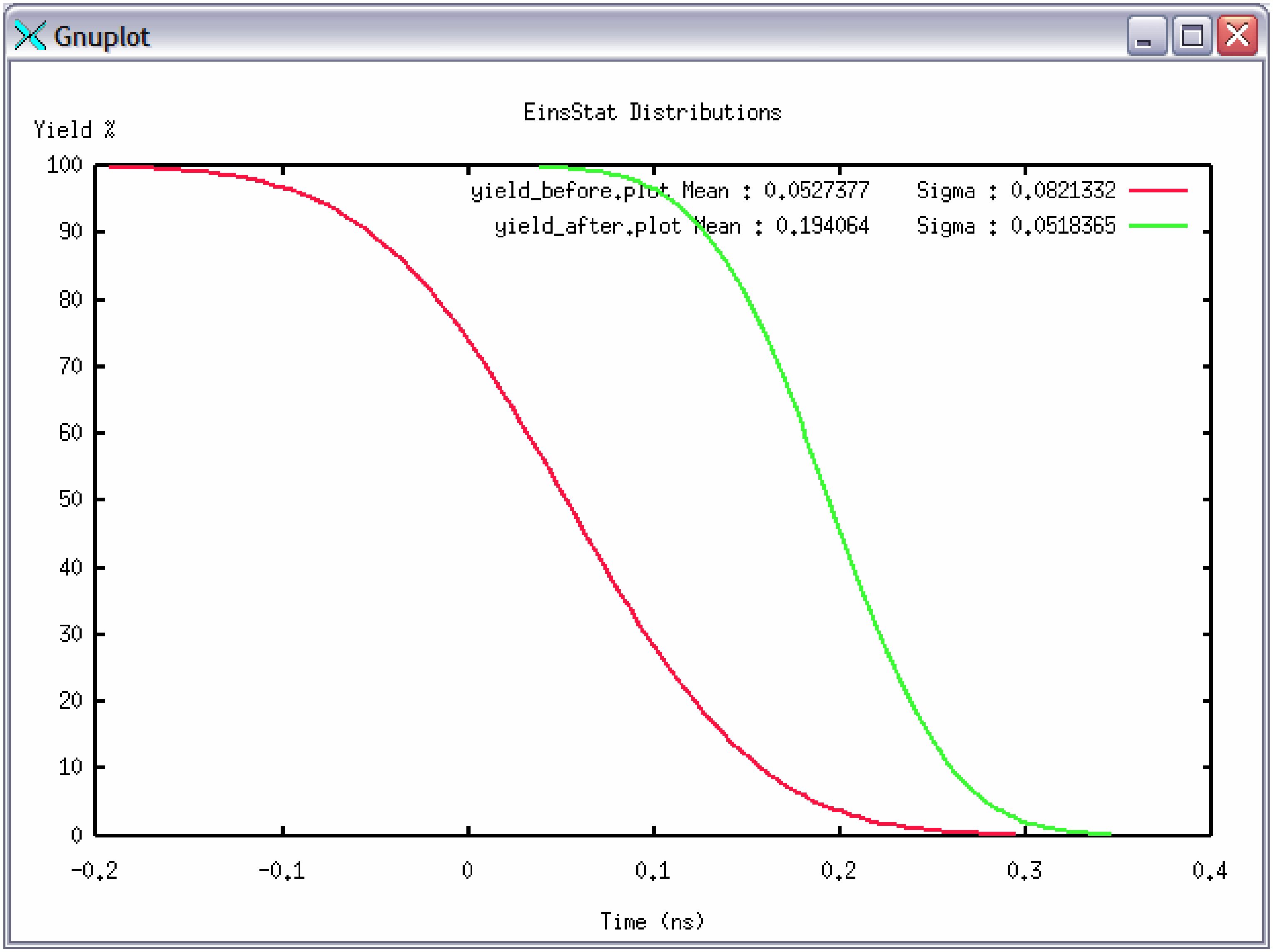
Shift mean? Reduce sensitivities?



Traditional delay optimization



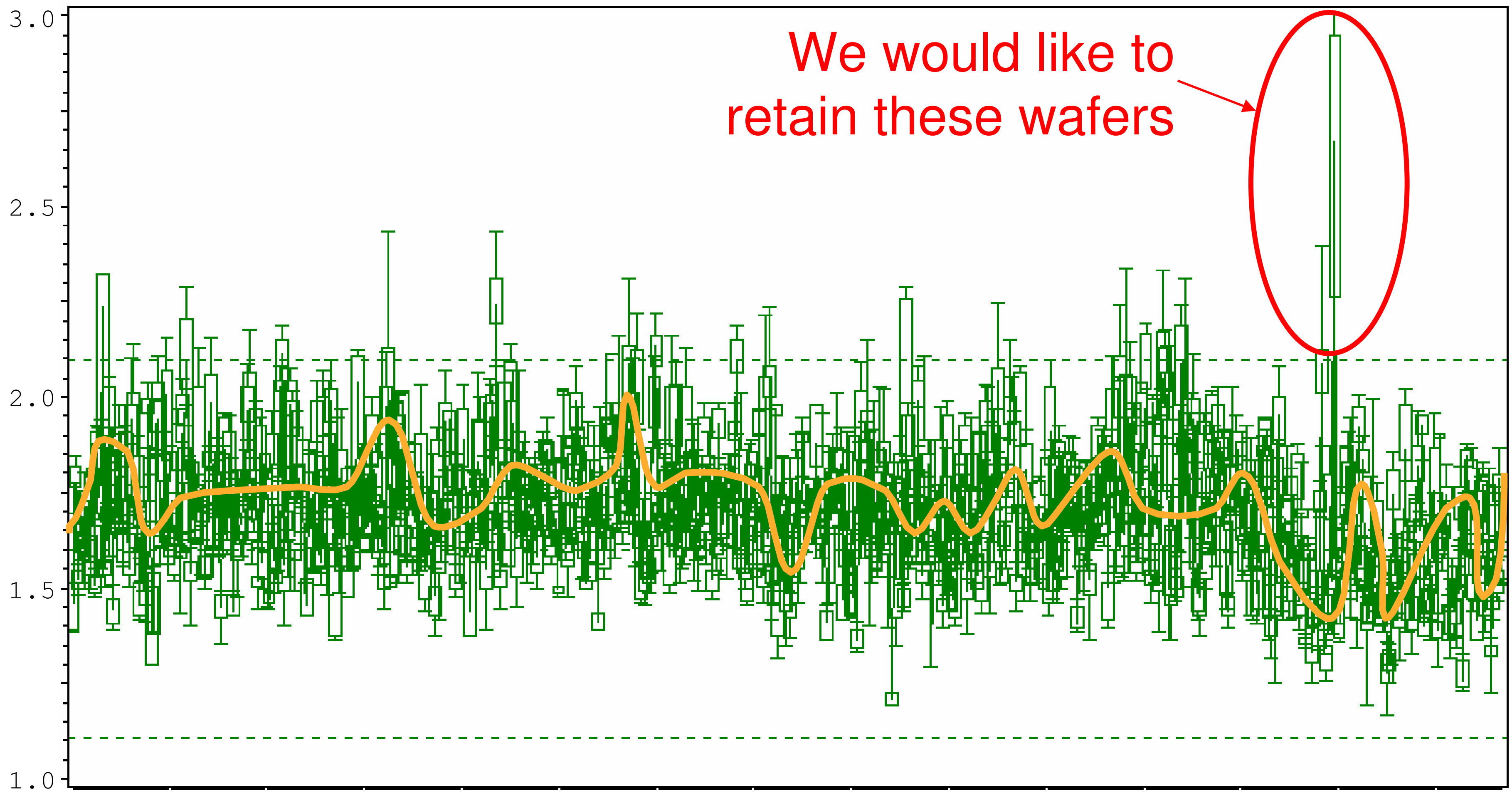
What is needed



Relative ordering of paths

- **Path A slack = -60 ps, path B slack = -50 ps**
  - Correlated case
  - Uncorrelated case
- **Other problems**
  - One end-point at a time: no notion of yield!

Robust optimization: normalized 90-day M3 resistance



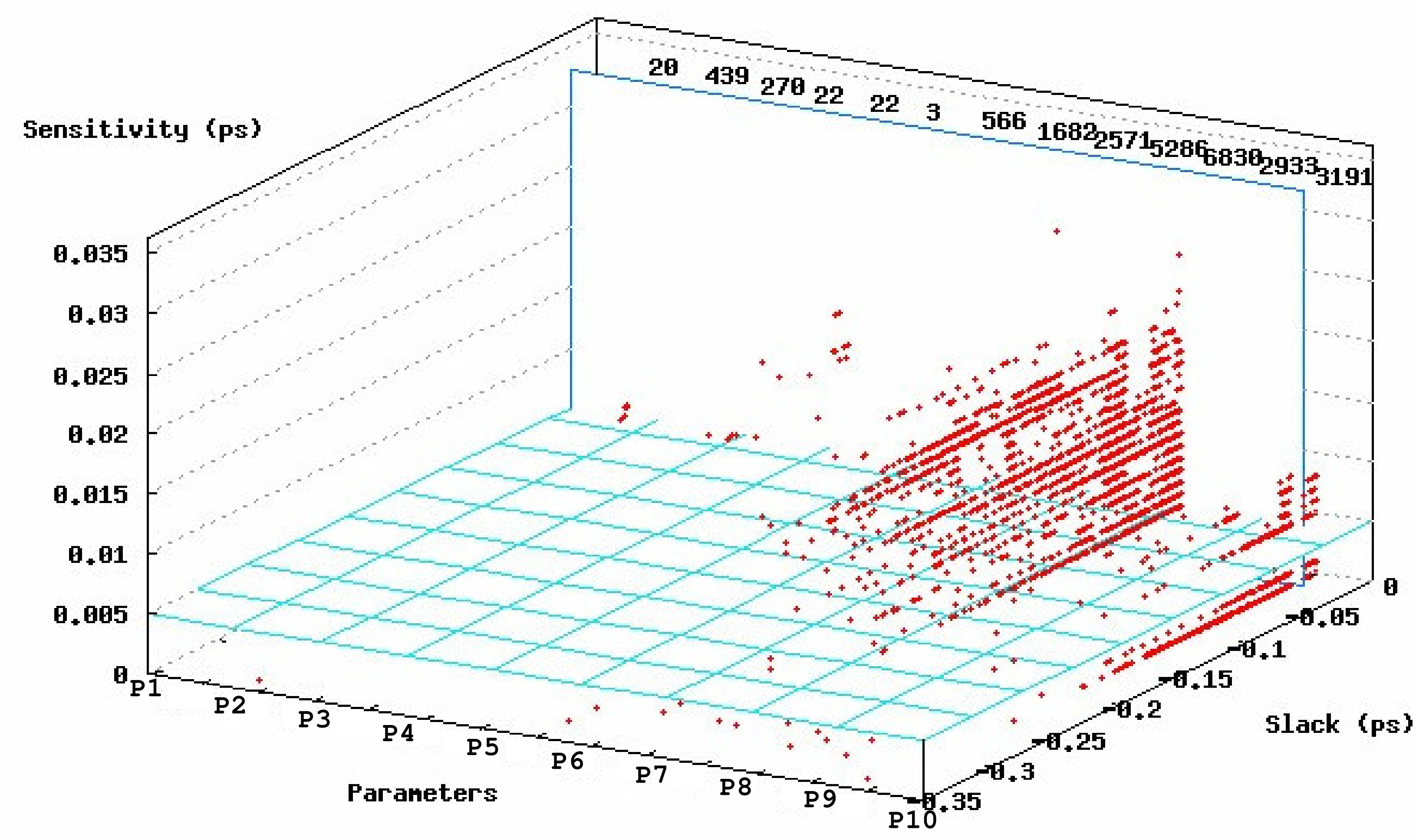
Robust circuit design

First order model

$$P(x, y) = mean + \frac{\partial p}{\partial x}\Delta x + \frac{\partial^2 p}{\partial x^2}\Delta x^2 + \cdots + \frac{\partial p}{\partial y}\Delta y + \frac{\partial^2 p}{\partial y^2}\Delta y^2 + \cdots$$

- **Its the sensitivities, stupid!**

Slack/sensitivity plots



Courtesy Xiaoyue Wang, IBM Toronto

Criticality as a diagnostic

- **Definition: probability of a {path, edge, node} being on the critical path**
- **Pros**
  - Single number reflecting entire process space
  - Captures relative effects between paths
- **Cons**
  - Dominant path masks other paths
  - Unity-sum metrics have this limitation
- **A combination of slack, sensitivity and criticality will probably be the best metric**

Continuous yield optimization

$$\max_{W_i} \quad \text{yield at required slack}$$
s.t. β ratio constraints
s.t. slew constraints
s.t. input pincap constraints
s.t. area/power constraints

$$\max_{W_i} \quad slack$$
s.t. yield requirement
s.t. β ratio constraints
s.t. slew constraints
s.t. input pincap constraints
s.t. area/power constraints

$$\min_{W_i} \quad area/power$$
s.t. yield requirement @ required slack
s.t. β ratio constraints
s.t. slew constraints
s.t. input pincap constraints

See V. Zolotov et al, "Computation of yield gradients from statistical timing analysis," TAU 2006

Sensitivity reduction examples

- **Routing: changing metal levels**
- **Metal: buffering, double wide wires**
- **N/P mistracking: resynthesis**
- **Gate replacement based on sensitivity signature**
- **Layout tricks to bring capturing/launching paths closer**
- **Vt families: balanced use**

Phased adoption of statistical optimization

- **Deterministic optimization + manual fix-up**
- **Single-valued projections for fix-up**
- **Sensitivity fix-up**
- **Truly distribution-based optimization**
- **Optimization of adaptive circuits**
  - Performance is probabilistic even in the absence of process variations
  - Go after "expected value" of performance

Conclusions

- **Single-valued optimization will run out of steam**
  – even if projected or sampled from a yield curve
- **Need robustness to be considered**
  – Special targeted changes to tame sensitivity
  – Including layout and routing of critical paths
- **Blindly improving slack costs power/area**
- **Prediction: industry will see a phased adoption of statistical optimization methods between now and the 23 nm node**

The power of statistical formulas