

Challenges in Business Analytics

An industrial research perspective

30 October 2008

Eleni Pratsini

IBM Zurich Research Laboratory

Alexis Tsoukiàs

LAMSADE-CNRS, Université Paris Dauphine

Fred Roberts

DIMACS, Rutgers University

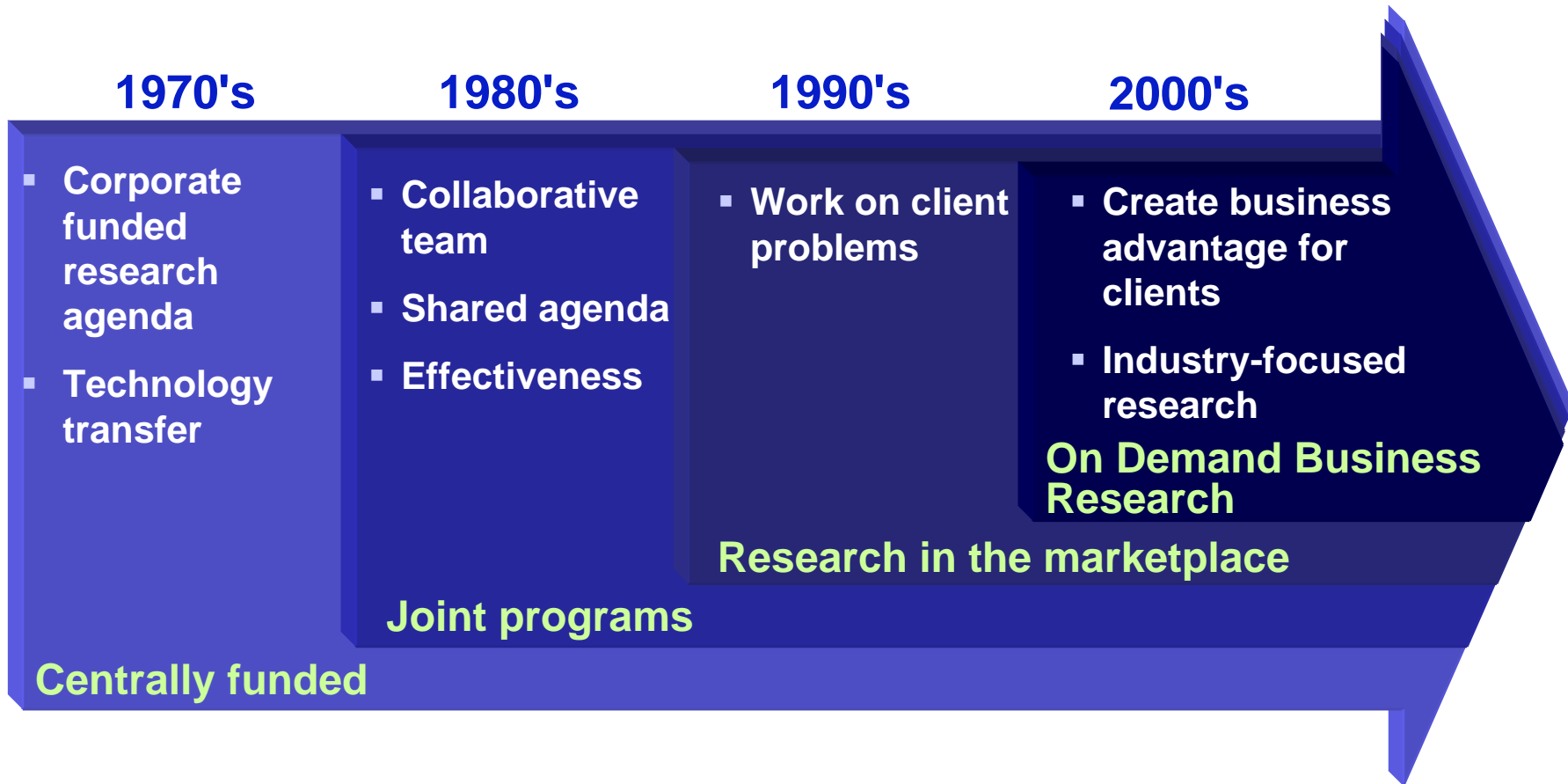
60 Years of IBM Research

The Sun Never Sets at IBM Research



Evolution of Role

Work on client-specific technology, business problems



Overview

- **Business Environment** → need for analytics in a world of increasing complexity

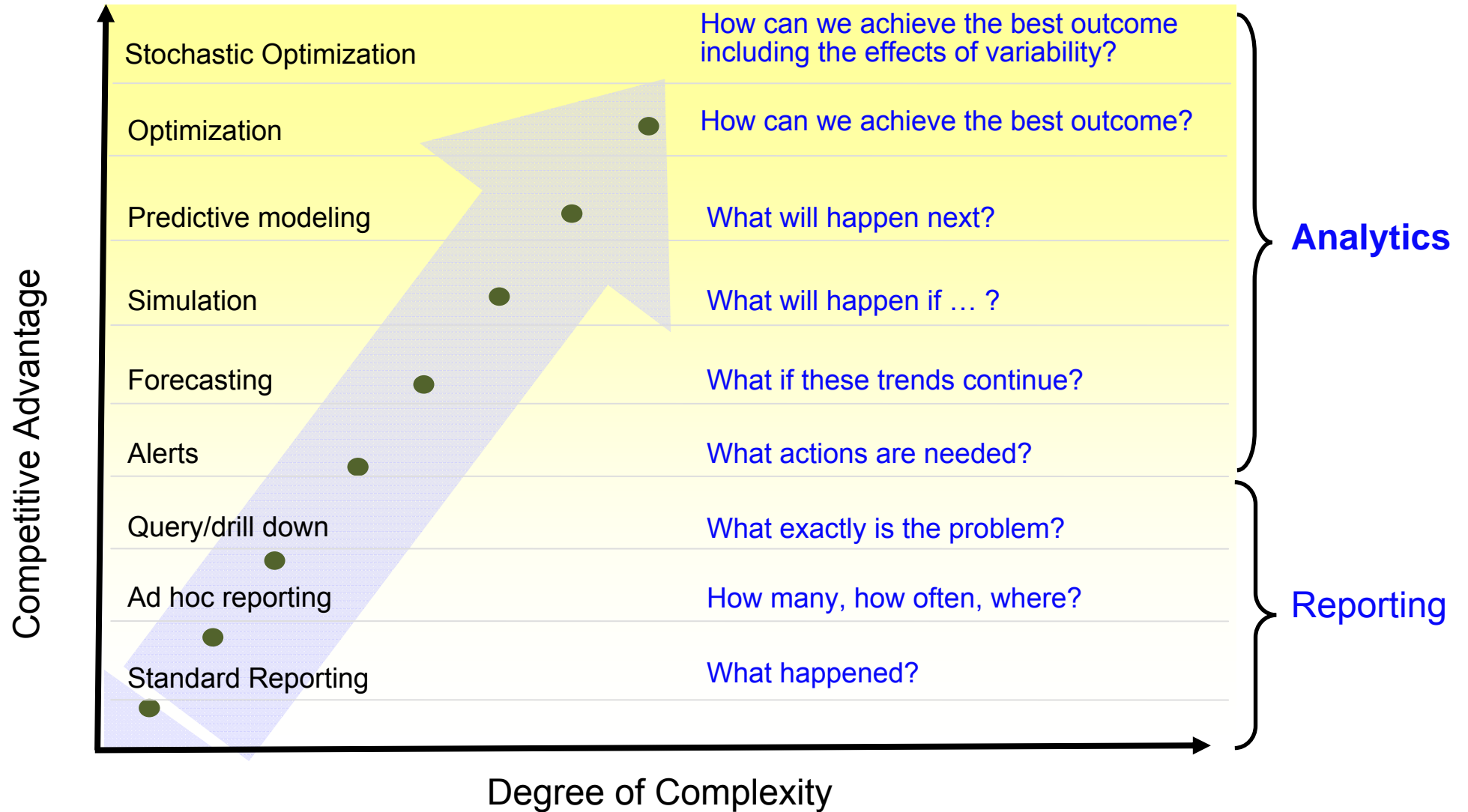
- **Technical Challenges:** *Example from the Pharma Industry*
 - Solution architecture
 - Data availability – a pleasure and a plague
 - quality, uncertainty, missing
 - Prescriptive models depend on predictive models
 - Optimization under uncertainty

- **Business Challenges**
 - Client external projects
 - Client internal projects
 - Academic projects

Business Environment

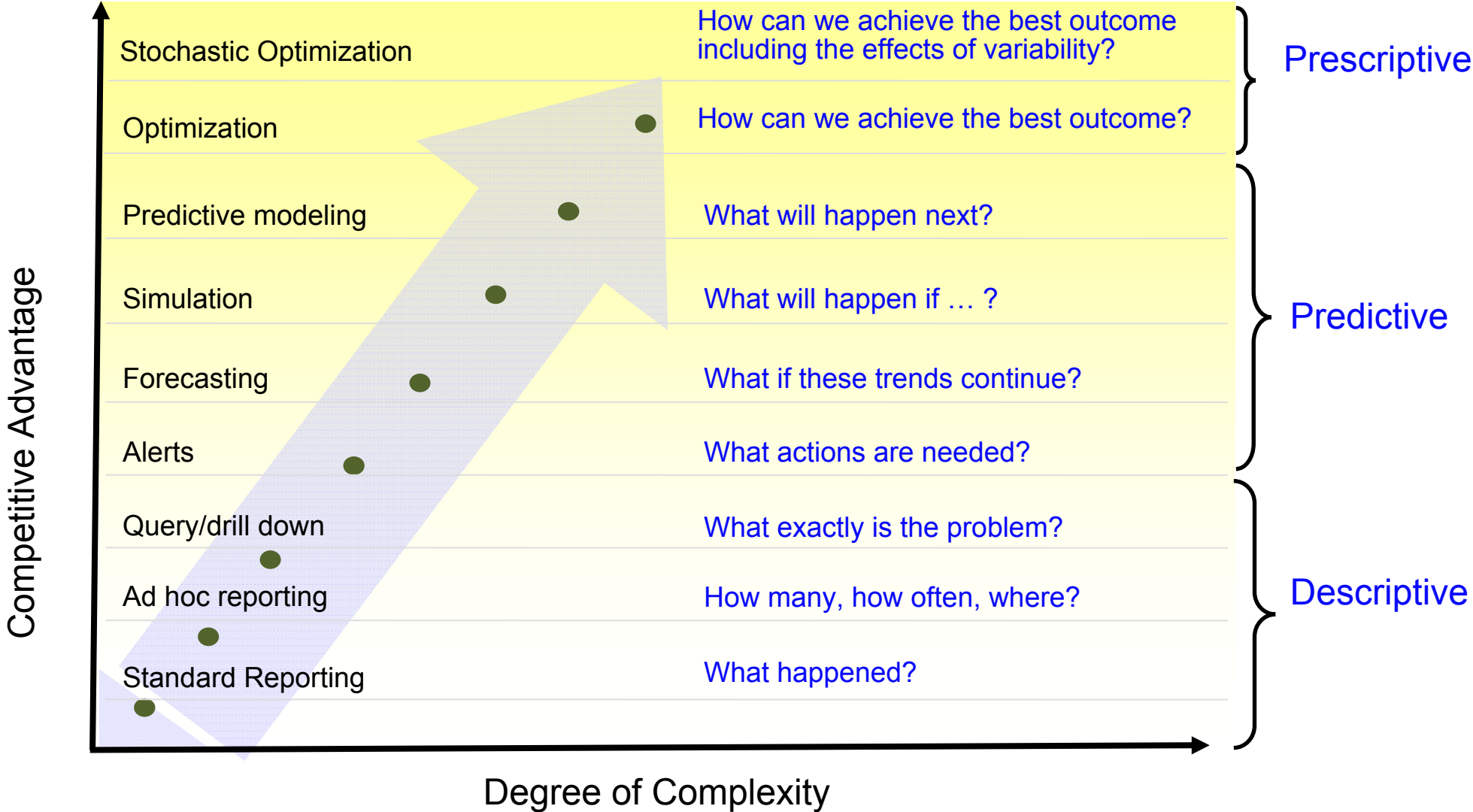
- We live in a highly constrained resource world
- Rapidly reaching capacity for many resources
 - natural, infrastructure, human
- Study: What do you see as the primary business challenges currently affecting your organization?
 - Improving operational effectiveness
- Analytics are essential to address the constrained resources challenge. The historical barriers to the use of analytics are rapidly disappearing, enabling us to tackle complex high-value problems.

Analytics Landscape



Based on: Competing on Analytics, Davenport and Harris, 2007

Analytics Landscape



Based on: Competing on Analytics, Davenport and Harris, 2007

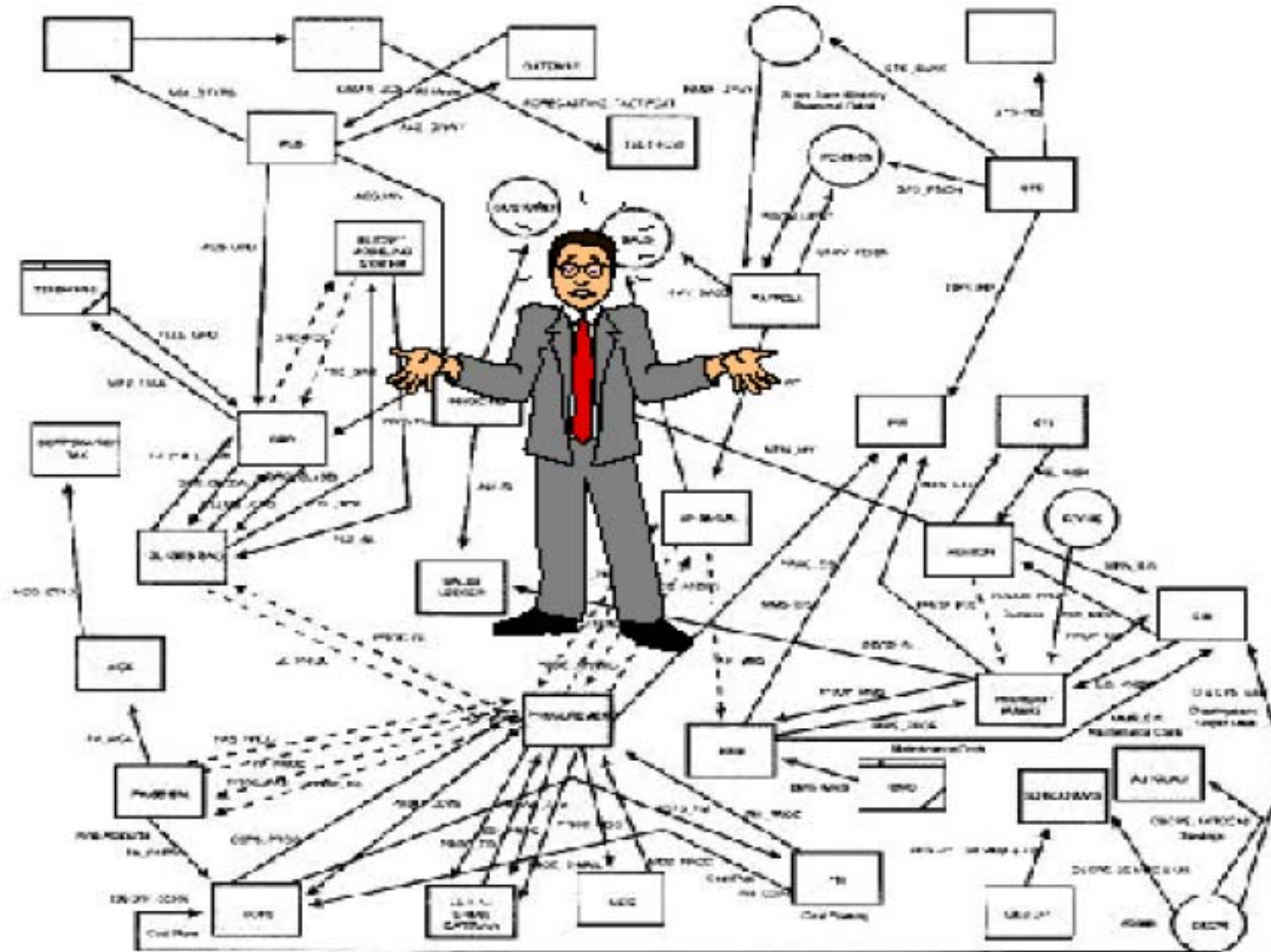
Challenges in a world of increasing complexity

- Advances in computer power bring capabilities (and expectations) for solving larger problem sizes
 - Problem decomposition not always obvious
 - Sequential optimization
 - Exact vs approximate solution or mixed
 - Hybrid methods
- Information Explosion
 - There is a lot of data (too much!) but how about data quality?
 - Not enough of the “right” data
 - Missing values
 - Outdated information
 - Data in multiple formats
- Can our optimization models handle the information?



“Looks like you’ve got all the data – what’s the holdup?”

Where do we start?



Overview

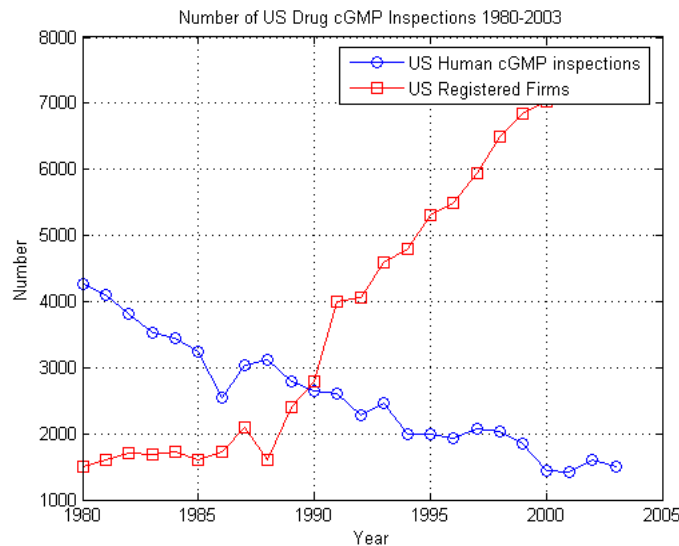
- **Business Environment** → need for analytics in a world of increasing complexity

- **Technical Challenges:** *Example from the Pharma Industry*
 - Solution architecture
 - Data availability – a pleasure and a plague
 - quality, uncertainty, missing
 - Prescriptive models depend on predictive models
 - Optimization under uncertainty

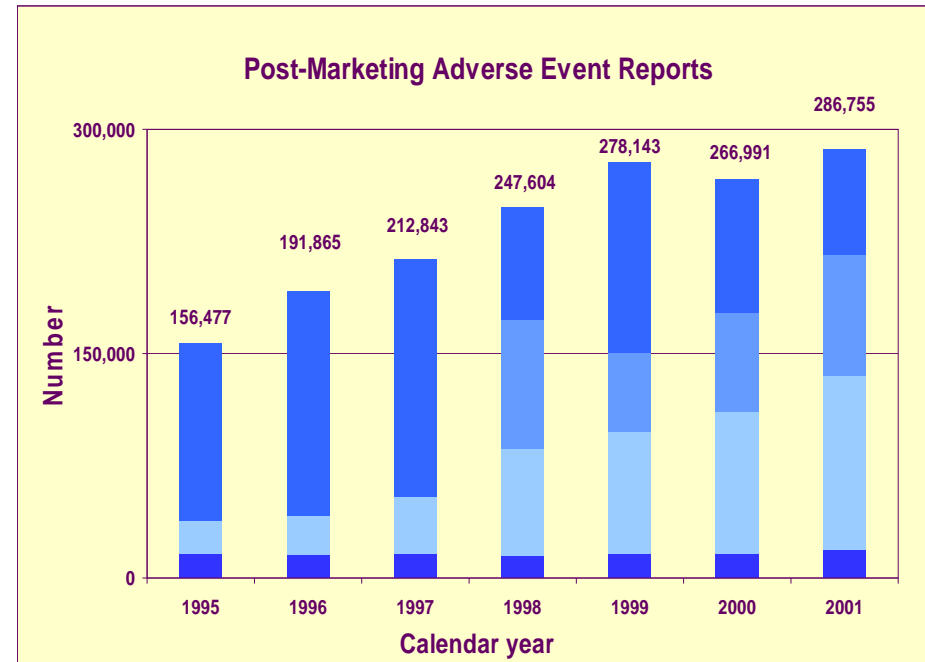
- **Business Challenges**
 - Client external projects
 - Client internal projects
 - Academic projects

Example: Risk Based Approach to Manufacturing

'FDA's GMPs for the 21st Century – A Risk-Based Approach'

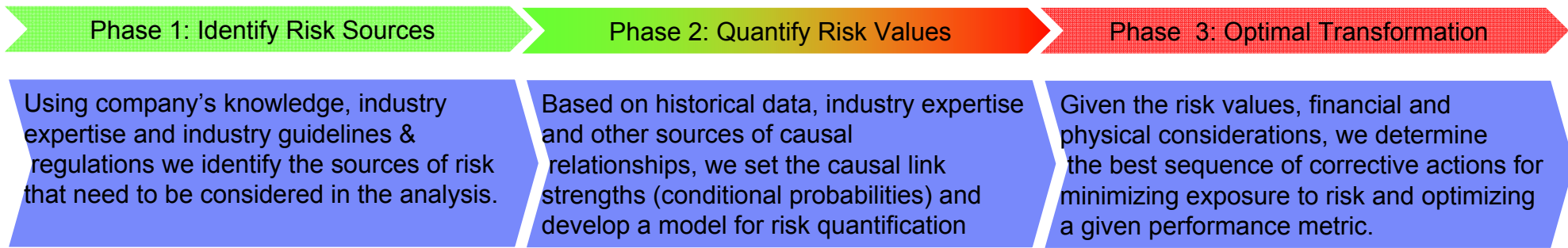
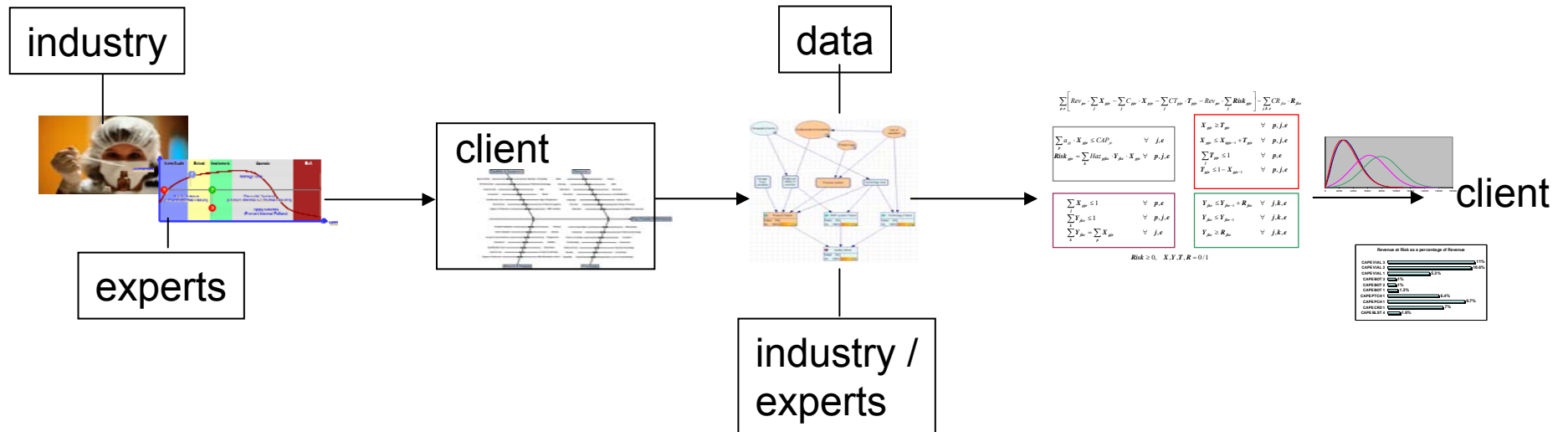


*Horowitz D.J. *Risk-based approaches for GMP programs*, PQRI Risk management Workshop, February 2005



- **Manage risk to patient safety – “Critical to Quality”**
- **Apply science in development and manufacturing**
- **Adopt systems thinking – build quality in**

Risk Management and Optimization



Definition of “Risk”

■ Fitness for Use

- **Quality**: product conforms to specifications
- **Efficacy**: ensures product has a positive effect (drugs treats a patient)
- **Safety**: end product is not harmful (patient does not suffer from fatal side effects)

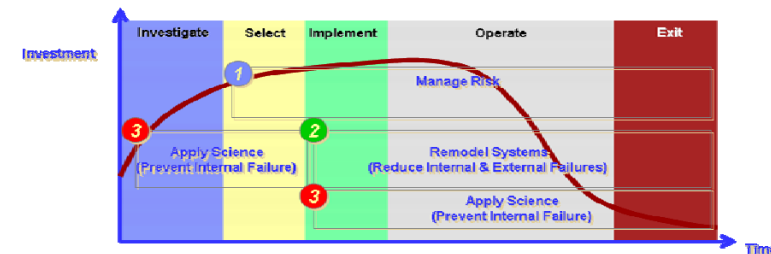
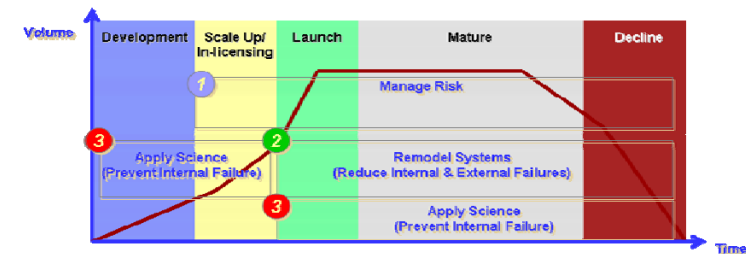
- **Compliance**: failure to comply with regulations (quality, efficacy and safety not fully under control), e.g. lack of documentation

Risk: Three sources of hazards

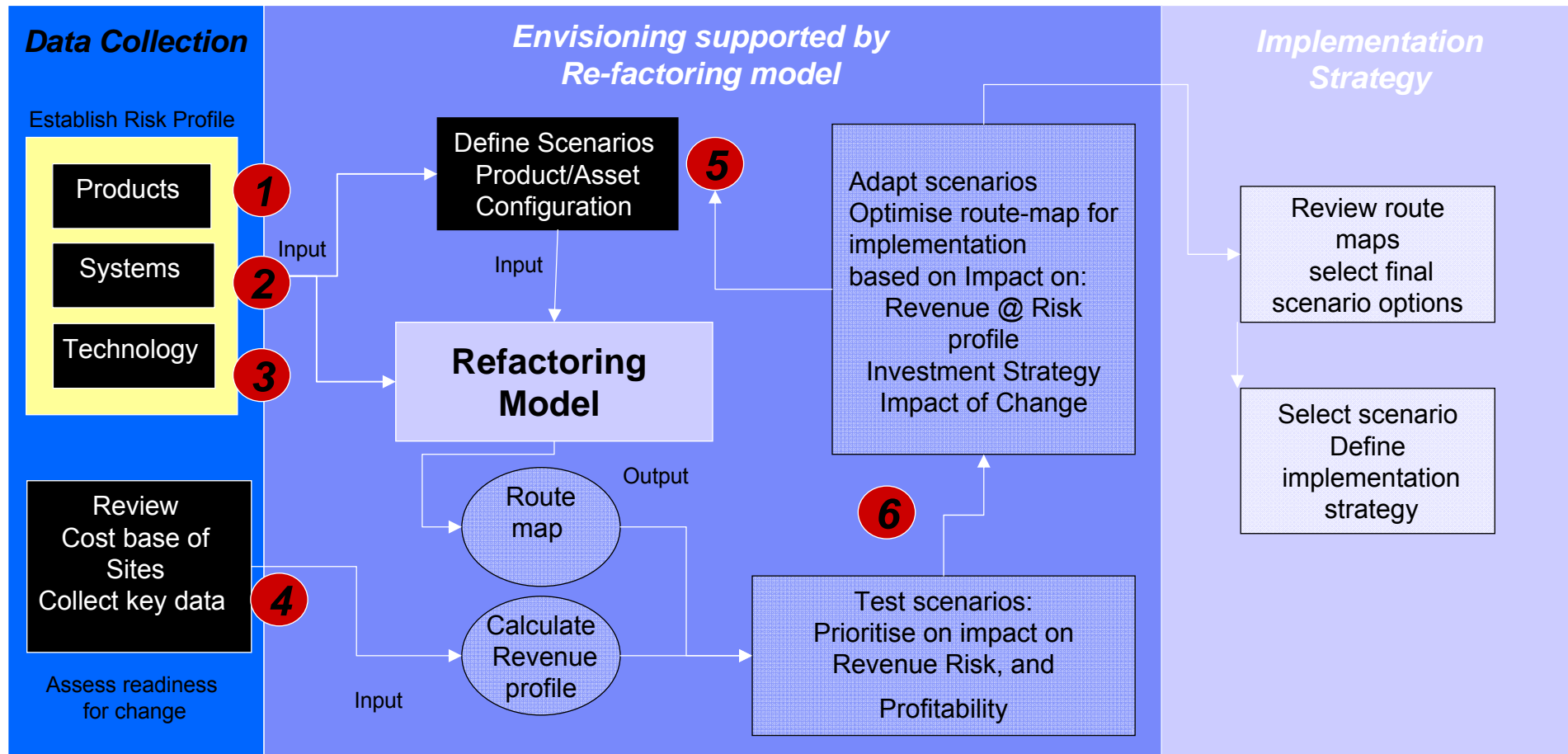
- Products: Traditional focus; Based on portfolio analysis
 - product is unstable or contaminated

- Technologies: some technologies may not be appropriate given risk they create.
 - a new blender does not work well and the mixing is not done properly

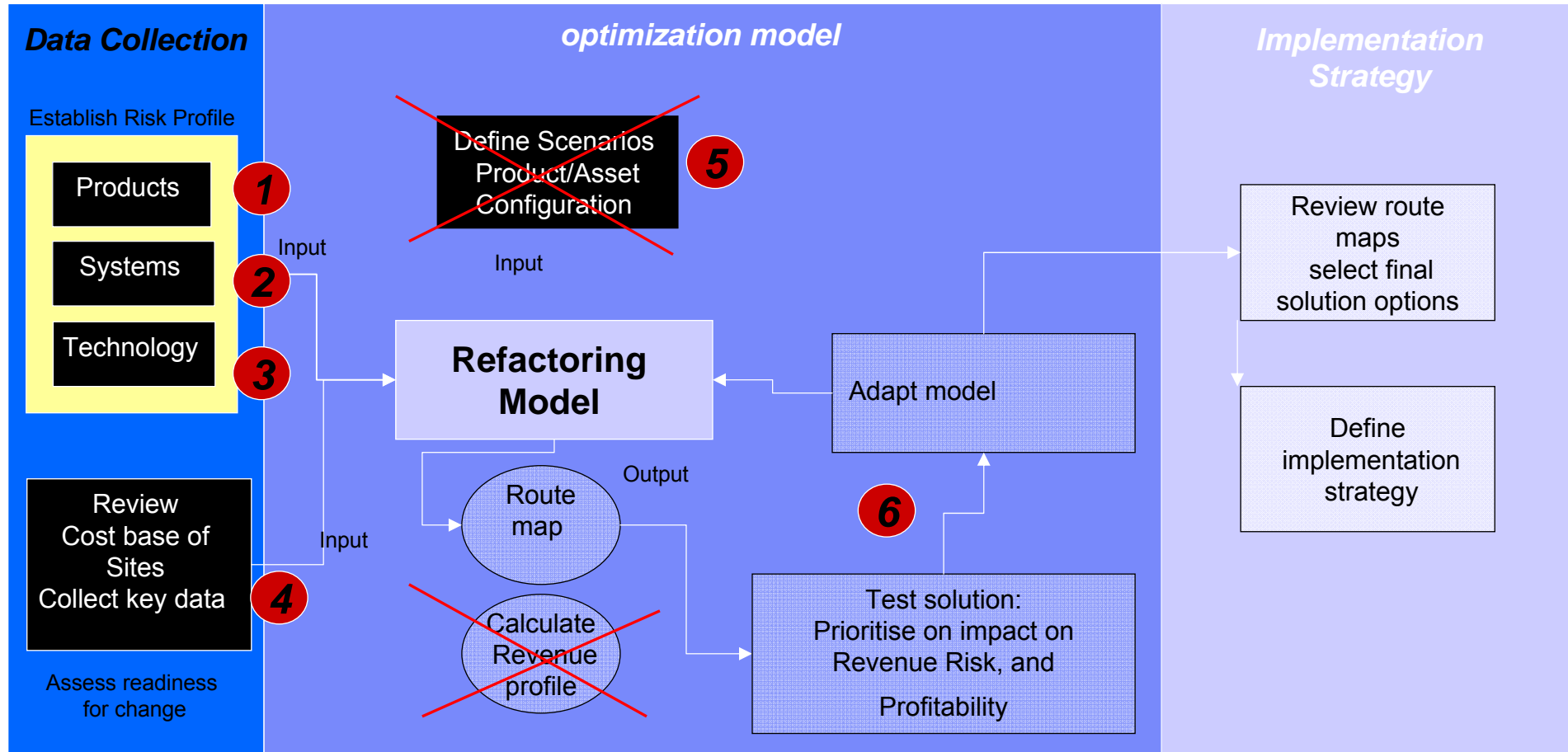
- Systems: basis for regulatory inspection. Create transfer risks between product profile classes. Systems capabilities & technologies.
 - a biotech product is developed and there is no control of the specification of its components



Overview of Approach



Overview of Approach



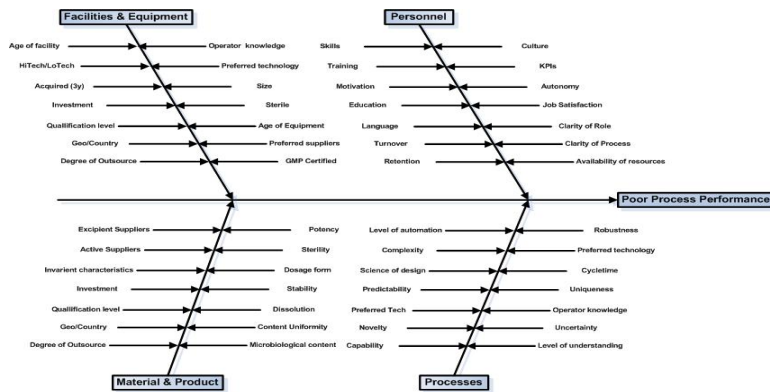
Effect of uncontrolled data collection

- A pharmaceutical firm is assessing the risk of two of their sites: Country A and Country X. Engineers at both sites are asked to collect non-compliant and defective batches in the last two years. Analysis of the data indicates that site A has 50% more defective / non-compliant batches than site X. They deduce that they should move their manufacturing processes to X or drastically change their operations at A.
- Just before launching an audit of their site A, they receive a letter from the FDA mentioning many recalls of batches all coming from site X.
- Further investigation indicates that site X does not report the many defects they notice every week and they are indeed more risky.

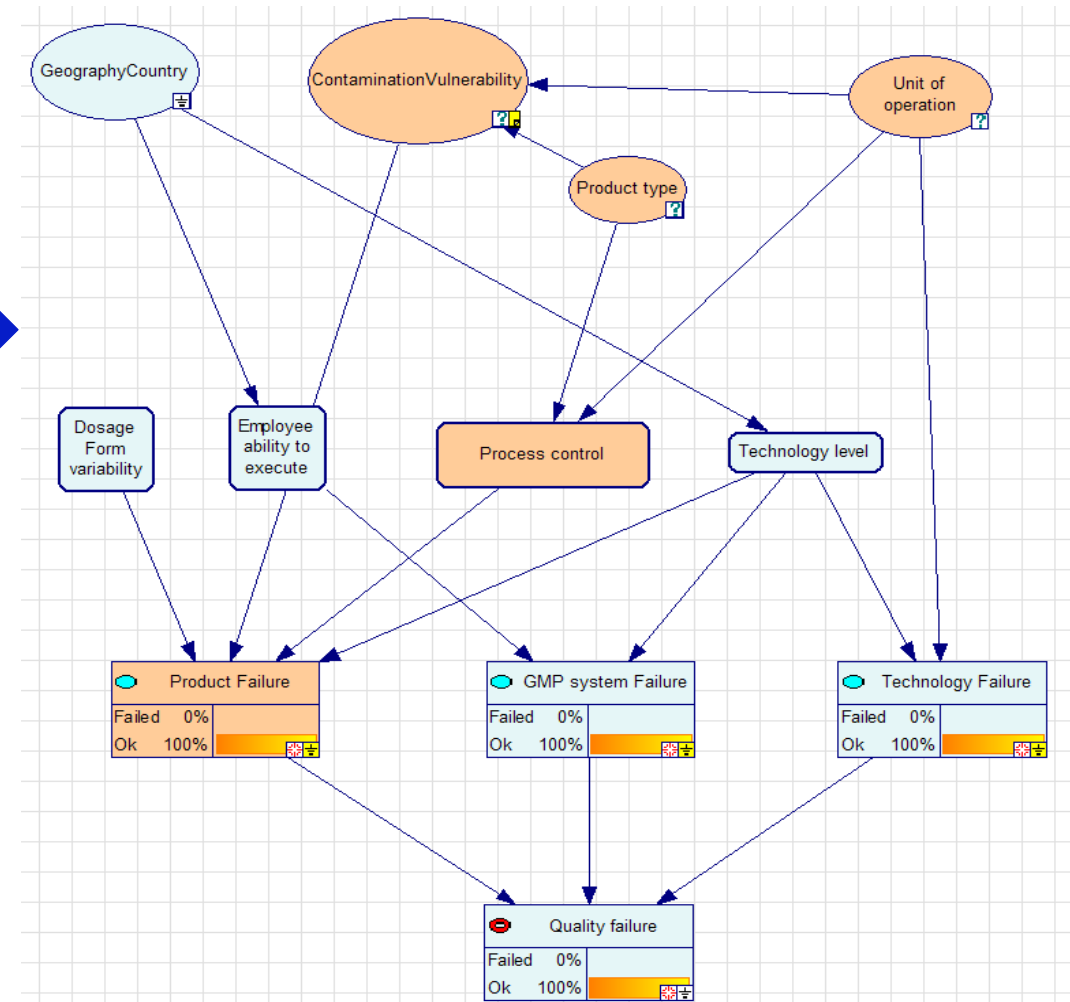
Lack of observations does not mean no events; it means no reporting of events
Extract knowledge from non statistical sources, e.g. experts

Technology, product and system failure are all linked via common root causes. The risk model starts from building a dependency graph of all these root causes.

Cause Effect Analysis



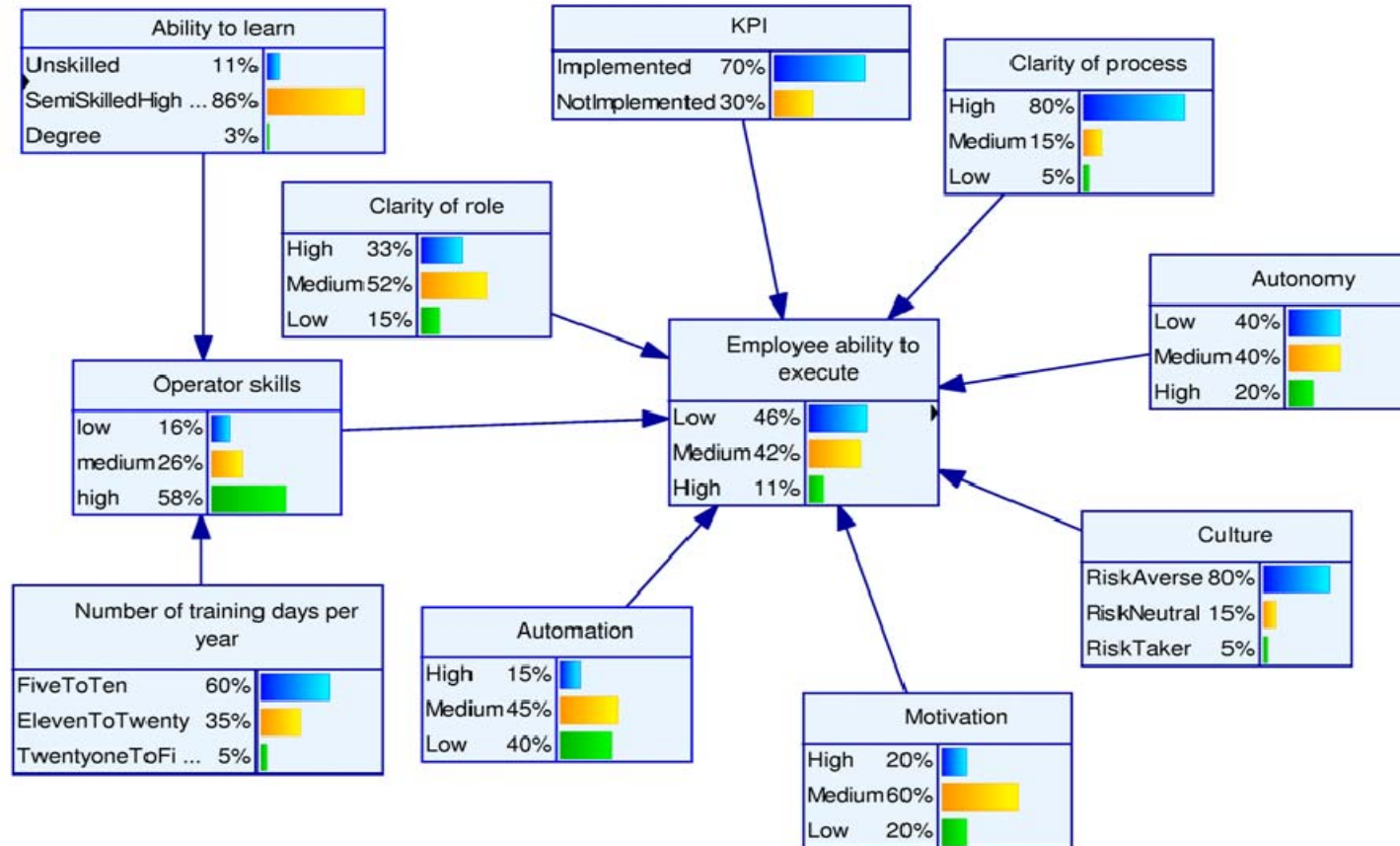
- industry Expertise
 - Client Constraints
- ➔
- Past Experiences



Different sub-models created:

- Dosage form variability
- Employee execution
- Process control
- Technology level

Employee Ability to Execute Sub-model



Directly linked to GMPs. Example events: reporting not done properly, quality control protocol not implemented, etc.

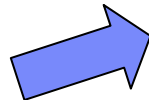
Elicitation of Expert Opinion - Challenges

- Network parameters can be estimated from data and manual entry of probabilities.
- Using expert knowledge in Bayesian networks can be an intractable task due to the large number of probabilities that need to be elicited

We need

- an *elicitation method* that is fast and gives not only a probability but also a confidence in the probability
- an *aggregation method* for combining the elicited probabilities taking into account the experts' confidence in their assessments
- A *probability estimation* method to reduce the number of probabilities to elicit in a noisy MAX node

Elicitation / Aggregation / Estimation



$$\alpha = \frac{h+l}{h-l}, \quad N = \frac{100}{h-l} \quad \text{and} \quad \beta = N - \alpha.$$



Calibration

Please note that the following calibration will only affect the answer given on this page.

1 2 3 4 5
0% 50% 100%

Summary

1: rare	0% - 10%
2: occasionally	10% - 30%
3: common	30% - 70%
4: typical	70% - 90%
5: always	90% - 100%

Answer

rare
 occasionally
 common
 typical
 always

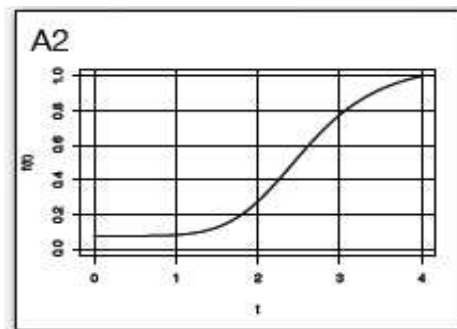
(you may adjust your answer below to reflect your confidence)

Adjustment

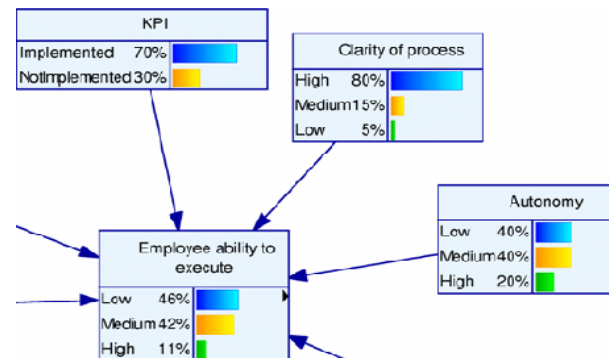
Please adjust your initial answer here to reflect your confidence. You may either narrow the range if you are more confident, or widen the range if you are less confident on your answer.

0% 100%

Range: 70% - 90%

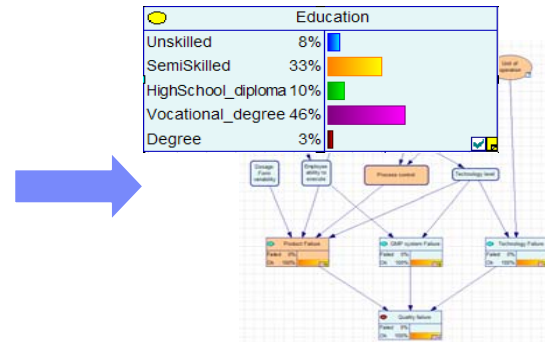


	A1		A2	
B	1	0	1	0
4	1.0	0.0	1.0	0.0
3		0.0		0.0
2	0.850	0.0	0.275	0.0
1		0.0		0.0
0	0.275	1.0	0.075	1.0



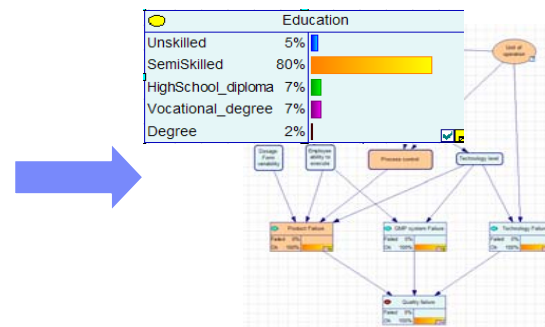
Based on location, technologies or any previously identified components involved in the drug production process, the network model computes an estimate of the quality risk. Changing the way a drug is produced will (a priori) induce a change in the quality risk estimate.

Site A



Quality failure: 2.4%

Site B



Quality failure: 1.9%

Optimization model

Risk Exposure

- Performance measure for exposure to risk:

$$\mathbf{Revenue@risk}_{pe} = \text{Revenue}_{pe} \cdot \sum_j \mathbf{Risk}_{pje}$$

(Insurance premium)

- Risk value:

$$\mathbf{Risk}_{pje} = \sum_k \mathbf{Haz}_{pjke} \cdot \mathbf{Y}_{jke} \cdot \mathbf{X}_{pje}$$

Formulation

$$\text{MAX} \sum_{p,e} \left[\text{Rev}_{pe} \cdot \sum_j X_{pje} - \sum_j C_{pje} \cdot X_{pje} - \sum_j CT_{pje} \cdot T_{pje} - \text{Rev}_{pe} \cdot \sum_j \text{Risk}_{pje} \right] - \sum_{j,k,e} CR_{jke} \cdot R_{jke}$$

ST

$$\sum_p a_{pj} \cdot X_{pje} \leq CAP_{je} \quad \forall j, e$$

$$\text{Risk}_{pje} = \sum_k \text{Haz}_{pjke} \cdot Y_{jke} \cdot X_{pje} \quad \forall p, j, e$$

Capacity, Risk

$$X_{pje} \geq T_{pje} \quad \forall p, j, e$$

$$X_{pje} \leq X_{ipje-1} + T_{pje} \quad \forall p, j, e$$

$$\sum_j T_{pje} \leq 1 \quad \forall p, e$$

$$T_{pje} \leq 1 - X_{pje-1} \quad \forall p, j, e$$

Transfer

$$\sum_j X_{pje} \leq 1 \quad \forall p, e$$

$$\sum_k Y_{jke} \leq 1 \quad \forall p, j, e$$

$$\sum_k Y_{jke} = \sum_p X_{pje} \quad \forall j, e$$

State, Level

$$Y_{jke} \leq Y_{jke-1} + R_{jke} \quad \forall j, k, e$$

$$Y_{jke} \leq Y_{jke-1} \quad \forall j, k, e$$

$$Y_{jke} \geq R_{jke} \quad \forall j, k, e$$

Remediation

$$\text{Risk} \geq 0, \quad X, Y, T, R = 0/1$$

Problem characteristics

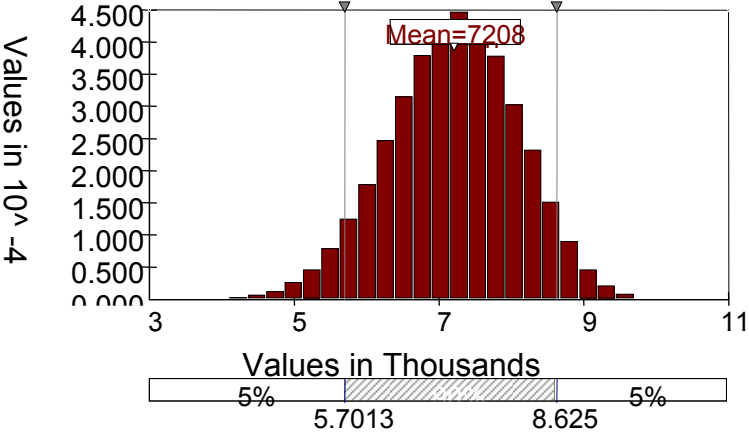
- 17 product families
- 3 systems plus one for subcontracting
- 14 technology platforms – most available in all systems
- Planning horizon: 5 years split into 10 periods of 6 months
- Restriction on rate of change
- Statistical analysis gave risk values based on historical data
- Revenue projections available
- Aggregated costs per scenario – later disaggregated
- Simulation of given scenarios (ending result); optimization

Example Analysis

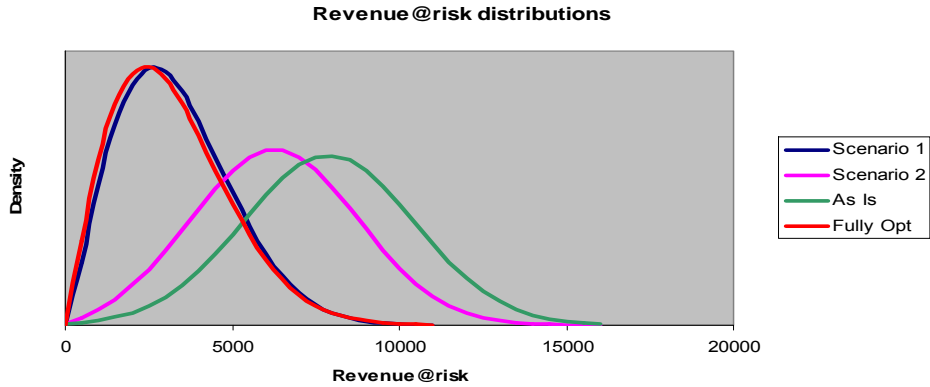
Optimal sequence of actions for minimizing risk exposure

Work center	Current	2005	2006	2007	2008	2009	Scenario X
Tech 1	Site A						Site B
Tech 2							
Tech 3	Site B						
Tech 4							
Tech 5	Site C						
Tech 6							
Tech 7	Site A / Site C						
Tech 8							
Tech 9	Site A						
Tech 10							
Tech 11	Site C						
Tech 12							
Tech 13	Site C						
Tech 14							

Calculate Business Exposure




Scenario Comparison



$$E(exposure) = \sum rev_i \cdot p_i$$

$$Var(exposure) = \sum_i rev_i^2 \cdot var(p_i) = \sum_i rev_i^2 \cdot p_i(1 - p_i)$$

IBM Pharma Refactoring Control Center



User name:

Password:

Please enter userid and password

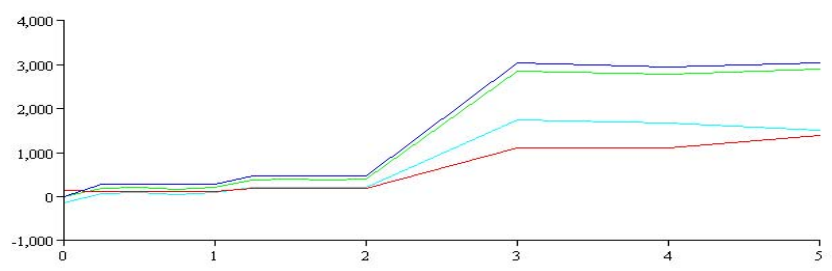
IBM Pharma Refactoring Control Center

User: test
Data used: test-21-l

Data Entry Scope Selection Analysis Selection Scenarios Selection **Scenarios Comparison** Exit

Financials Sites Products Technologies

Financials

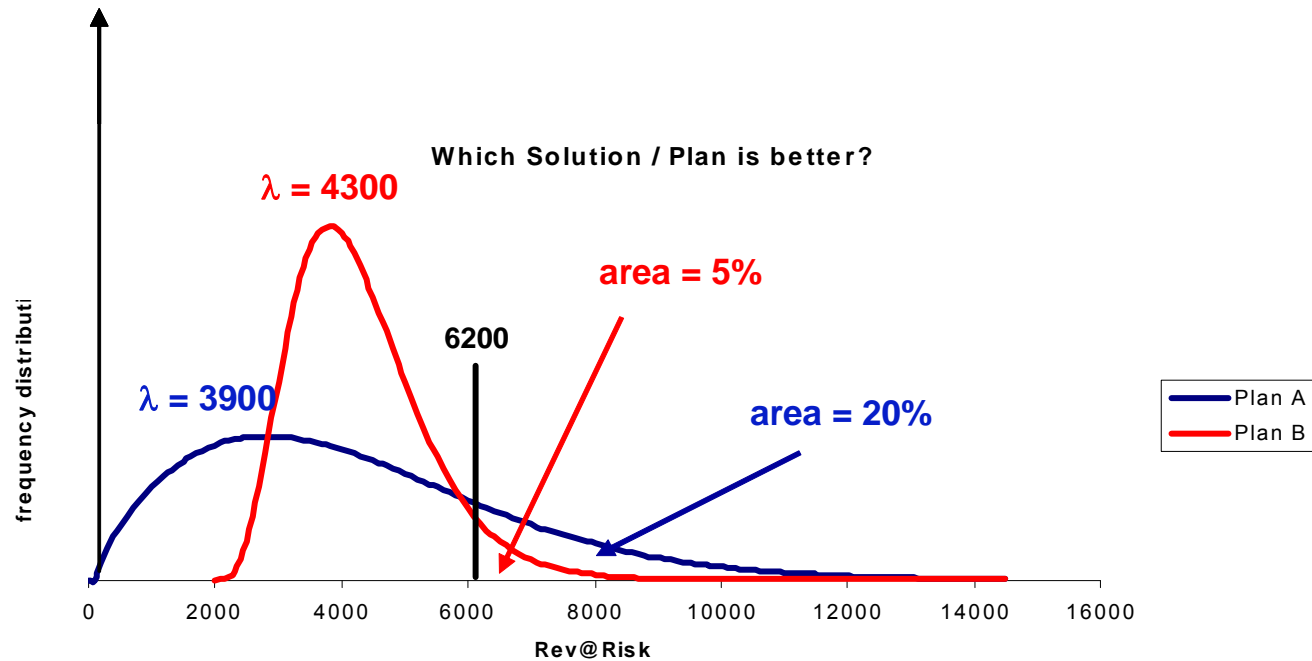


	Scenario 0	Scenario 1	Scenario 2	Scenario 3
aa_bb	aa_bb	No selection	No selection	No selection
Revenue	<input checked="" type="checkbox"/>	12030.0		
Operating Cost	<input type="checkbox"/>	1050.0		
Decision Cost	<input type="checkbox"/>	1153.0		
Revenue@Risk	<input checked="" type="checkbox"/>	4769.87		
Profit	<input checked="" type="checkbox"/>	10877.0		
Secured Profit	<input checked="" type="checkbox"/>	6107.13		

Scenarios Comparison

Data Variability

- The consideration of the frequency distribution of each solution is essential when there is uncertainty in the input parameters



- Robust Optimization techniques to determine the solution that is optimal under **most** realizations of the uncertain parameters

Possible Approaches

- Stochastic Programming
 - Average Value certainty equivalent
 - Chance constraints (known CDF – continuous)
- Robust Optimization
 - Soyster (1973)
 - Bertsimas & Sim (2004)
- CVaR
 - Rockafeller & Uryasev (2000)

Modeling with the approach of Bertsimas and Sim

- Formulation adjusted to incorporate a constraint that could be violated with a certain probability:

$$\begin{aligned} \max_{x \in X} & \quad E(\text{SecuredPro fit}(x)) \\ \text{s.t.} & \quad \text{Rev@Risk}(x) \leq \text{critical value} \\ & \quad \sum_{t,s} x_{p,t,s,e} \leq 1 \quad \forall p, e \\ & \quad x_{p,t,s,e} \in \{0,1\} \quad \forall p, t, s, e \end{aligned}$$

- Uncertain parameters have unknown but symmetric distributions
- Uncertain parameters take values in bounded intervals:

$$\text{Haz}_{t,s,e} \in \left[\hat{H}_{t,s,e} - \bar{H}_{t,s,e} ; \hat{H}_{t,s,e} + \bar{H}_{t,s,e} \right]$$

- Uncertain parameters are independent

Bertsimas & Sim cont'd

- Provided guarantee:

$$Pr(\text{Rev@Risk}(\mathbf{x}) > \text{critical value}) \leq \gamma$$

$$\gamma = \frac{1}{2^n} \left\{ (1 - \mu) \binom{n}{\lfloor \nu \rfloor} + \sum_{l=\lfloor \nu \rfloor+1}^n \binom{n}{l} \right\}$$

$$n = \text{total number of uncertain parameters}, \quad \nu = \left(\frac{\Gamma + n}{2} \right), \quad \mu = \nu - \lfloor \nu \rfloor$$

- Example Values of Γ :

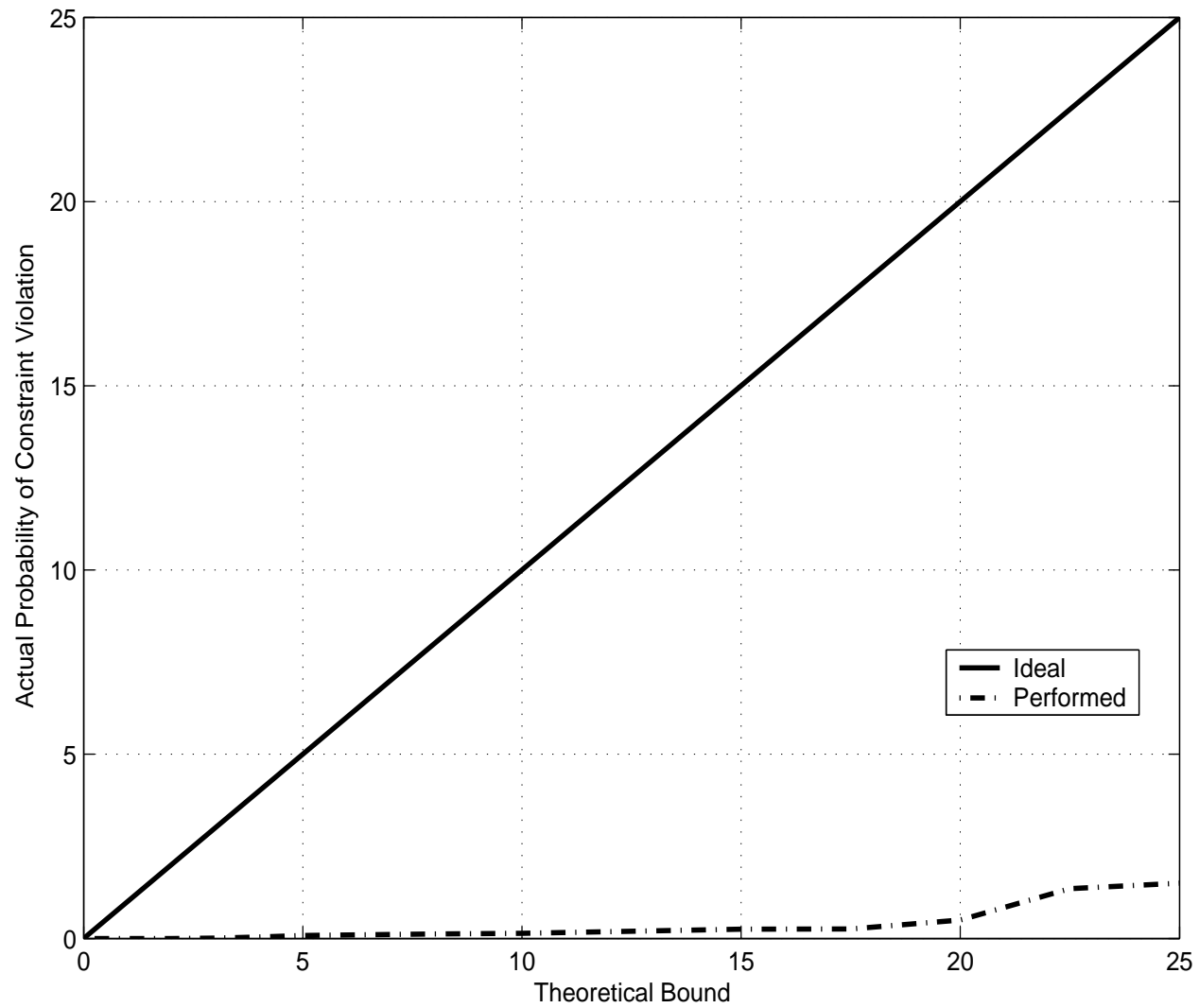
$n \setminus \gamma$	5%	1%	0.1%
100	17.502	24.232	31.764
1000	53.2	74.6	98.8
5000	118	166	220
10000	166	234	311

Application to pharma example

- Robust Formulation:

$$\begin{aligned}
 & \max_{\mathbf{x} \in X} E(\text{SecuredProfit}(\mathbf{x})) \\
 & s.t. \quad \sum_{p,t,s,e} \hat{H}_{t,s,e} \cdot \text{Rev}_{p,e} \cdot X_{p,t,s,e} + \mathbf{Z} \cdot \Gamma + \sum_{t,s,e} \mathbf{V}_{t,s,e} \leq \text{critical value} \\
 & \quad \mathbf{Z} + \mathbf{V}_{t,s,e} \geq \mathbf{Y}_{t,s,e} \quad \forall t,s,e \\
 & \quad -\mathbf{Y}_{t,s,e} \leq \left(\sum_p X_{p,t,s,e} \cdot \text{Rev}_{p,e} \right) \cdot \bar{H}_{t,s,e} \leq \mathbf{Y}_{t,s,e} \quad \forall t,s,e \\
 & \quad \sum_{t,s} X_{p,t,s,e} \leq 1 \quad \forall p,e \\
 & \quad X_{p,t,s,e} \in \{0,1\} \quad \forall p,t,s,e \\
 & \quad \mathbf{V}_{t,s,e} \geq 0 \quad \forall t,s,e \\
 & \quad \mathbf{Y}_{t,s,e} \geq 0 \quad \forall t,s,e \\
 & \quad \mathbf{Z} \geq 0
 \end{aligned}$$

Results!



Observations

- Provides solutions that are much more conservative than expected $Pr(Rev@Risk(x) > critical\ value) \ll \gamma$

- Plans satisfying risk requirement are discarded by the optimization - suboptimal solutions

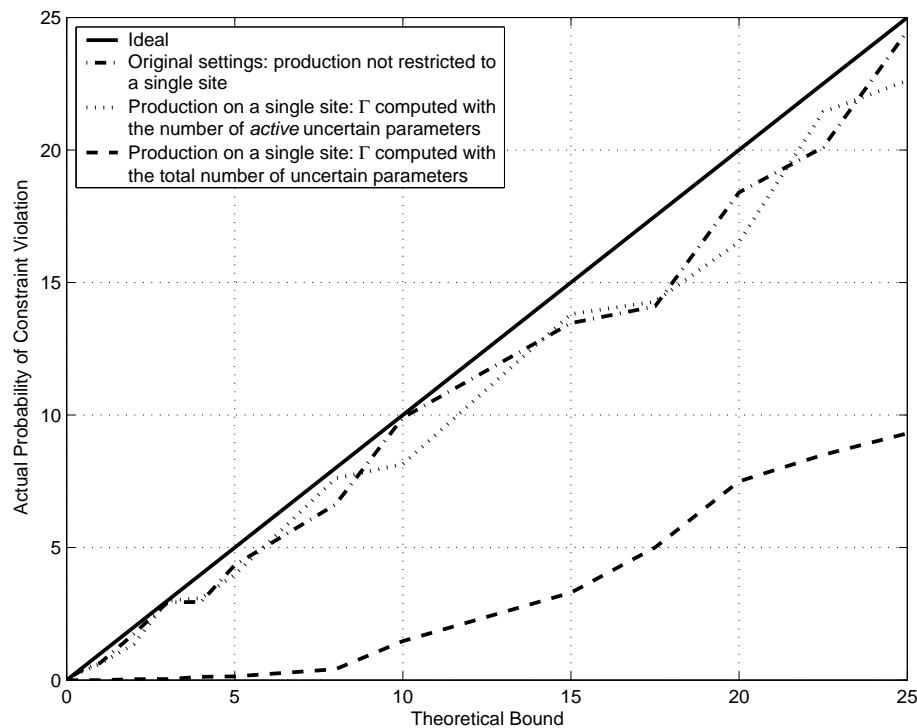
- Adjustments to calculation of parameters to improve results

- Extreme distribution
- Consider the number of *expected basic* variables and not the total number of variables affected by uncertain coefficients
 - e.g. production of 100 products, 3 possible sites, uncertainty in capacity absorption coefficient:
 - 300 possible variables (value of n)
 - each product can only be produced on one site i.e. only 100 variables > 0 (use this value for the calculation of Γ)

Bertsimas & Sim approach

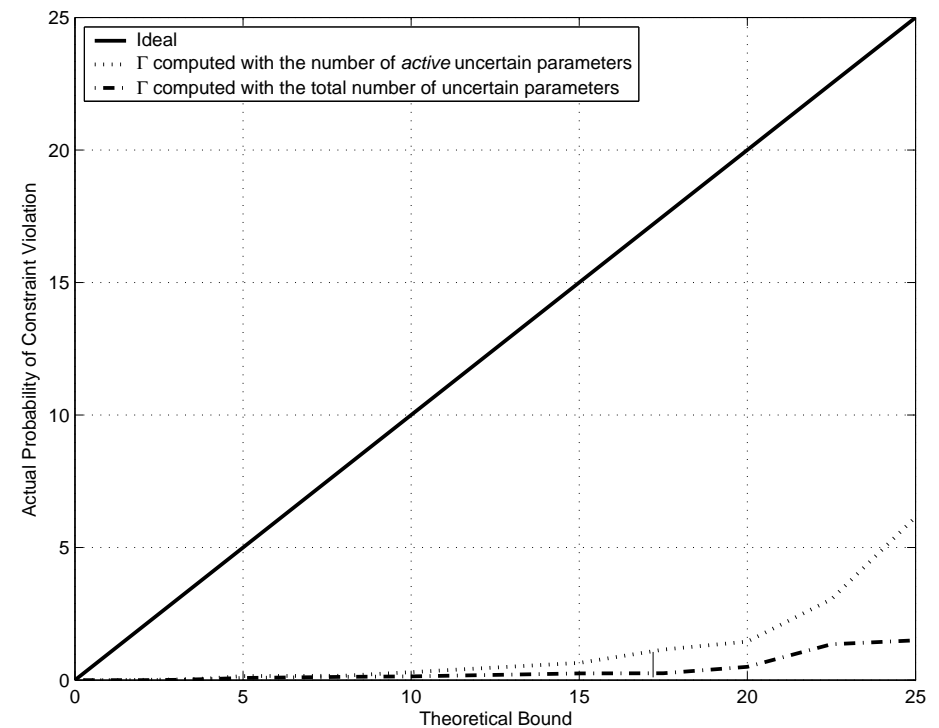
Production – Distribution

continuous variables



Pharma Regulatory Risk

binary variables



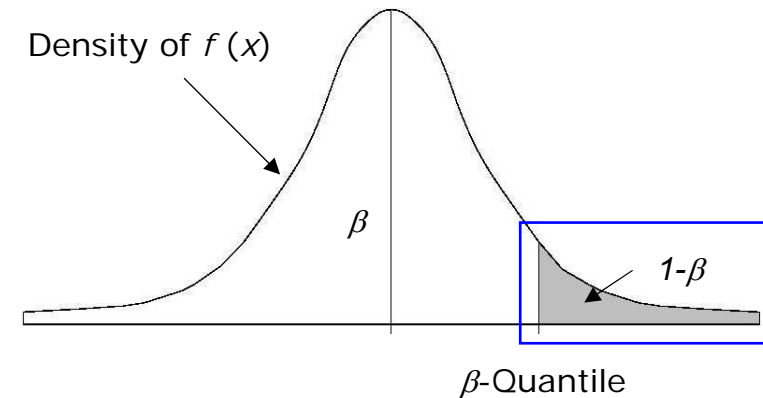
Related work

- Robust optimization Models for Network-based Resource Allocation Problems, Prof. Cynthia Barnhart, MIT
 - Robust Aircraft Maintenance Routing (RAMR) minimizing expected propagated delay to get schedule back on track (taken from Ian, Clarke and Barnhart, 2005)
 - Among multiple optimal solutions, no incentive to pick the solution with “highest” slack
 - Relationship of Γ (per constraint) to overall solution robustness?
 - Comparison with Chance constrained Programming
 - Explicitly differentiates solutions with different levels of slack
 - *Extended* Chance Constrained Programming
 - Introduce a “budget” constraint setting an upper bound on acceptable delay
 - Improved results

Value at Risk & Conditional Value at Risk

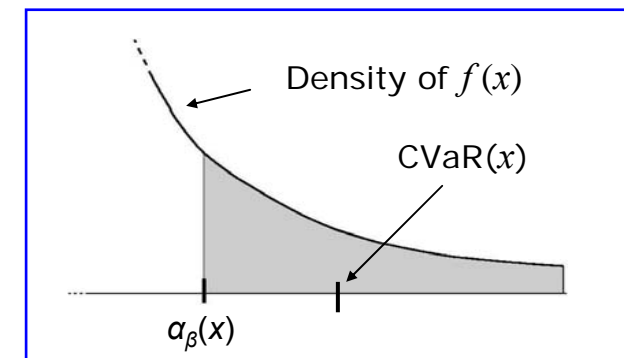
VaR

- Maximum loss with a specified confidence level
 - Non sub-additive
 - Non-convex



CVaR

- β -Quantile $\alpha_\beta(x)$ ($\beta=95\%$)
- $E [f(x) | f(x) > \alpha_\beta(x)]$
- Advantages
 - Sub-additive
 - Convex



CVaR in Math Programming: Linear Program with Random Coefficients

Consider a LP where, coefficient vector, ξ is random:

$$\text{Min } \mathbf{c}^T \mathbf{x}$$

$$\xi^T \mathbf{x} \leq b$$

$$D\mathbf{x} = \mathbf{d}, \mathbf{x} \geq \mathbf{0}$$

CVaR can be used to ensure 'robustness' in the stochastic constraint:

$$\text{Min } \mathbf{c}^T \mathbf{x}$$

$$\text{CVaR}_\alpha(\xi^T \mathbf{x}) \leq b$$

$$D\mathbf{x} = \mathbf{d}, \mathbf{x} \geq \mathbf{0}$$

Reformulation of Rockafellar & Uryasev (2000)

Reformulation

Each CVaR expression, say a constraint:

$$\text{CVaR}_\alpha(\xi^T \mathbf{x}) \leq b$$

is reformulated as:

$$t + \frac{1}{N(1-\alpha)} \sum_{i=1}^N (\xi_i^T \mathbf{x} - t)^+ \leq b$$

ξ_i : i^{th} sample of random vector ξ
(out of N total samples)

Linear Reformulation

$$t + \frac{1}{N(1-\alpha)} \sum_{i=1}^N z_i \leq b$$
$$z_i \geq \xi_i^T \mathbf{x} - t$$
$$z_i \geq 0$$

- **Good news: Fully linear reformulation, so presentable to any LP solver**
- **Bad news: New variables, and New constraints = O (Number of Samples)**
 - Reformulated LP has grown in both dimensions!
 - Large number of CVaR constraints & large number of samples => Very Large LP

Observations

- Number of scenarios necessary for a good estimation of $CVaR(\alpha, x)$ and thus meaningful results?

Number of Scenarios	Minimum Value	Maximum Value	Variation Around the Mean	Standard Deviation
300	4621.47	4987.77	4.89%	45.06
600	4653.56	4896.76	2.99%	31.54
1000	4663.60	4863.39	2.29%	24.48
2000	4693.98	4835.27	1.69%	17.28
5000	4710.09	4800.66	0.95%	11.01

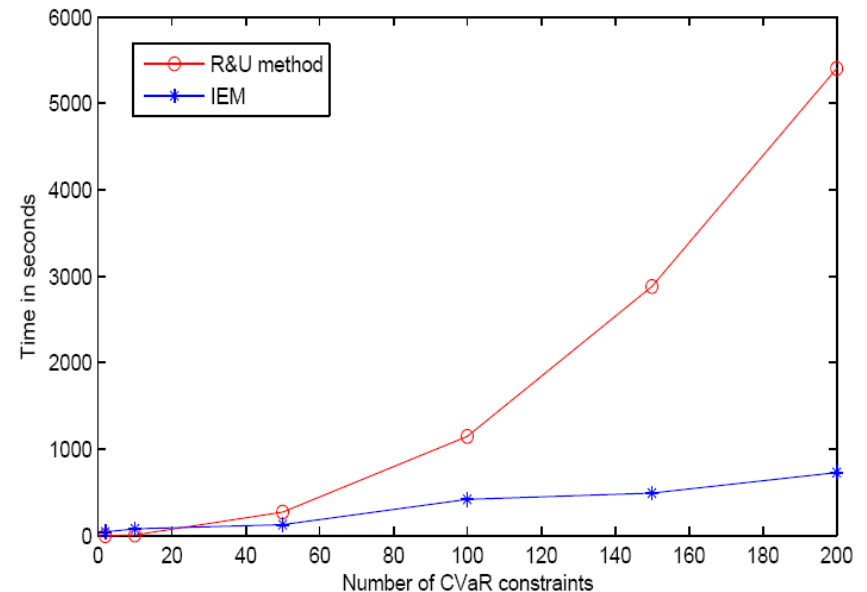
- >20,000 scenarios necessary for stable results (Kuenzi–Bay A. and J. Mayer. "Computational aspects of minimizing conditional value–at–risk",NCCR FINRISK Working Paper No. 211, 2005).
- Can handle up to 300 scenarios – computational complexity

Recent advances in CVaR

→Kunzi-Bay & Mayer (2006) presented a 2-stage interpretation if CVaR appears only in the Objective Function & their algorithm CVaRMin led to an order of magnitude speed-up

→P. Huang and D. Subramanian. "Iterative Estimation Maximization for Stochastic Programs with Conditional-Value-at-Risk Constraints". IBM Technical Report, RC24535, 2008

- They exploit a different linear reformulation, motivated by Order Statistics, and this leads to a new and efficient algorithm.
- The resulting algorithm addresses CVaR in both the Objective Function, as well as Constraints



→Interesting work on Robust Optimization techniques & CVaR

- David B. Brown, Duke University:
 - Risk and Robust Optimization (presentation)
 - Constructing uncertainty sets for robust linear optimization (with Bertsimas)
 - Theory and Applications of Robust Optimization (with Bertsimas and Caramanis)

Overview

- **Business Environment** → need for analytics in a world of increasing complexity

- **Technical Challenges:** *Example from the Pharma Industry*
 - Solution architecture
 - Data availability – a pleasure and a plague
 - quality, uncertainty, missing
 - Prescriptive models depend on predictive models
 - Optimization under uncertainty

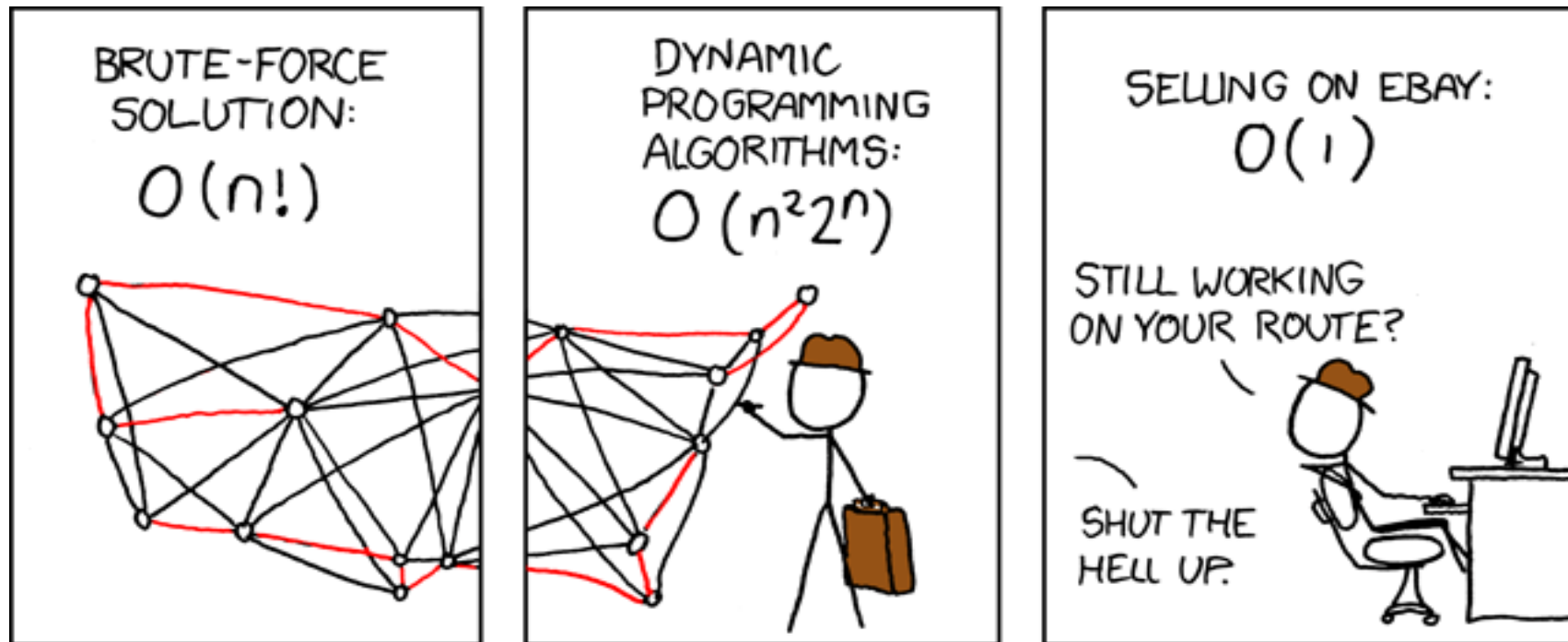
- **Business Challenges**
 - Client external projects
 - Client internal projects
 - Academic projects

Client external projects

- Very interesting and challenging problems
- We might speak English / German / French / ... but we do not speak their language!
 - *“Have you used simulation to prove your LP gives the optimal solution?”*
- Important to have industry knowledge
- Often rely on consultants to “sell” our work
 - Oversell / undersell
- Time horizon
 - They want the “optimal” solution now
- Prefer matured technology
 - They do not want to be the first to try a new technique

All they need is a simple solution ...

The fastest traveling salesman solution



**<http://xkcd.com/>*

They don't want to collect the data ...

You get what you pay for ...

*the amount and type of information used in the analysis affects the output**



**<http://www.mneylon.com/blog/2008/04/>*

Client internal projects

- Interesting real life application and client eager to use advanced analytics, often deal with very technical business partners so we understand each other
- Limited in scope
 - often consider a small aspect of the whole problem
- Medium time horizon
 - Usually 1 year but often continues for 2 or more years
- Internal politics!

Academic projects

- Life is good!!
- Academic collaborations:
 - Often “killed” by the lawyers because of IP (dis)agreement
- It does not pay the bills!!



Thank You!