

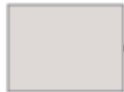


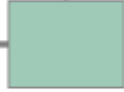





Data Quality for Index Models and Migration to Quantitative Models

Ernest Lever

- > Thursday, June 15, 2017
- > API, 1800 West Loop South, Suite 475, Houston, Texas

Agenda and Presentation Flow

Data Quality for Index Models and Migration to Quantitative Models

	Introduction	The disconnect between past performance and future results and how that limits semi-quantitative models	The correct application of distributions in models and how to utilize the distributions to model risk	How to identify data quality and how to incorporate quality information into a probabilistic risk model	How to utilize subject matter expert opinion and historic data to develop probabilities and distributions in a Probabilistic model	How to identify and optimize risk reduction activities using the probabilistic model
Complexity						
Predicting The Future						
Probability Distributions: Predicting Next Outcome						
Incorporating Data Quality metrics						
Optimization						
Subject Matter Expertise						

Complexity

Definitions: Complexity

- > The number of words we need to describe a situation
- > Complexity is closely related to information content
 - Need a lot of words to describe everything in this picture
 - We need to describe each item as they are different from the others
 - Descriptions are at a small scale



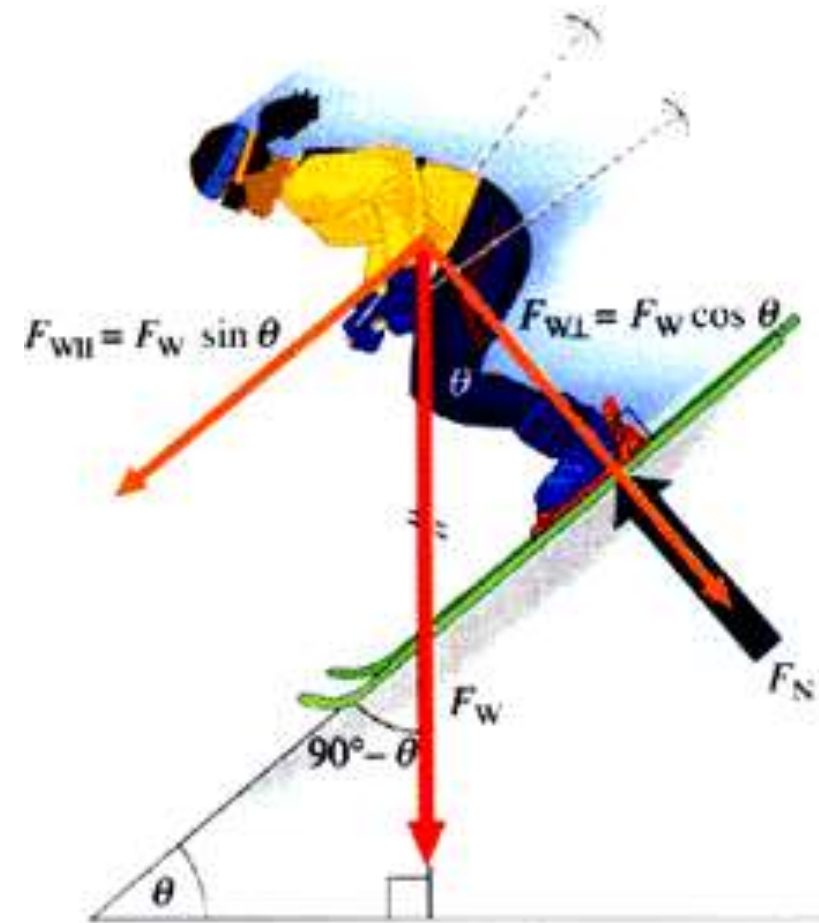
Definitions: Complexity – Less Complex

- > Need a fewer words to describe everything in this picture
- > There is more commonality
 - There are common patterns at large scale
 - Two colors
 - Two teams
 - Focus

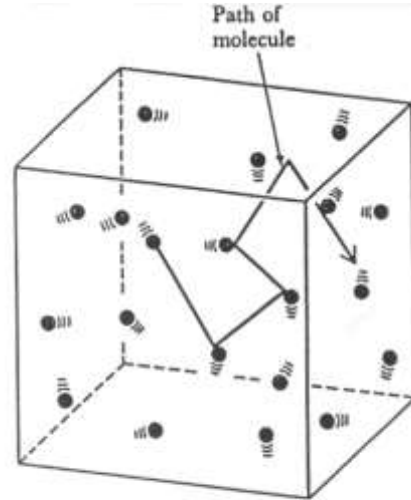


Definitions: Complexity – Simple

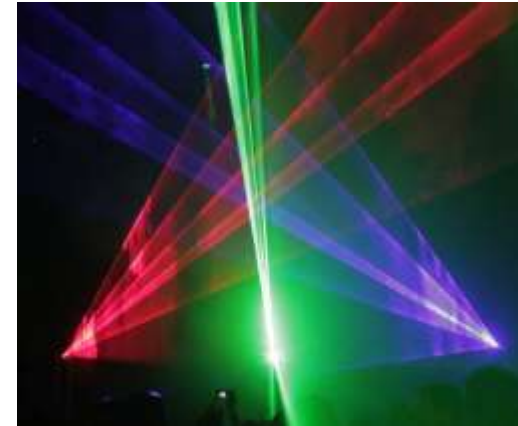
- Need even fewer words to describe this picture
- Ignore the internal details of the person
- Ignore the organization of the snow molecules
- Only need the mass of the skier, the angle of the slope and the amount of friction



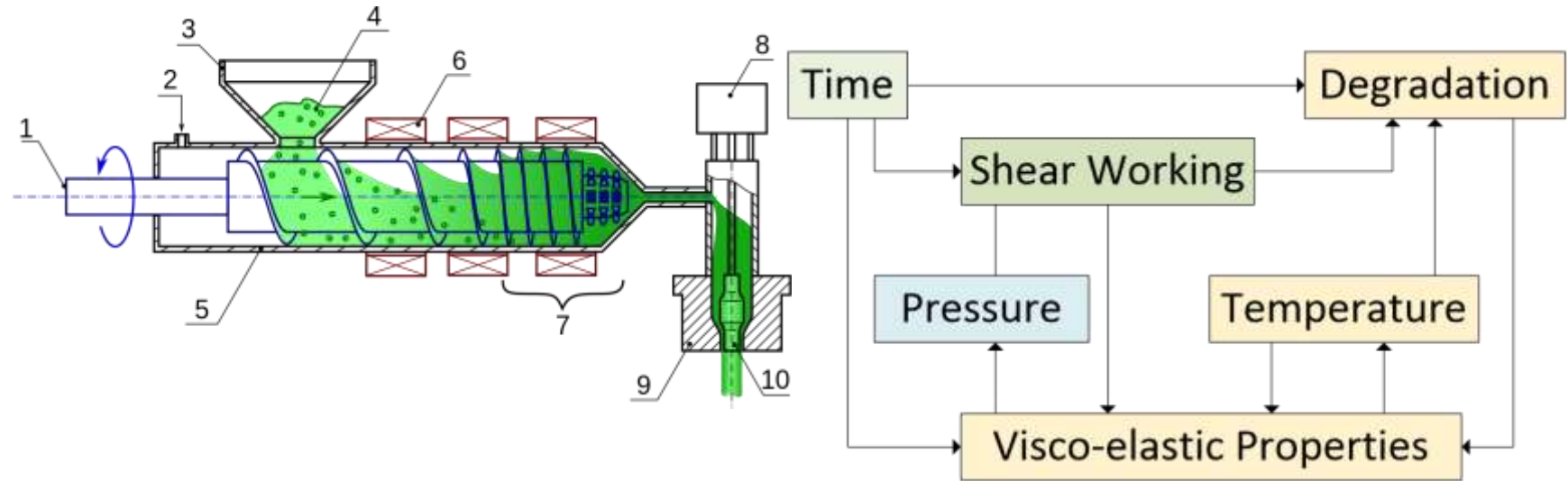
Definitions: Complexity vs. Scale



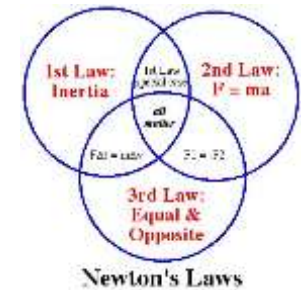
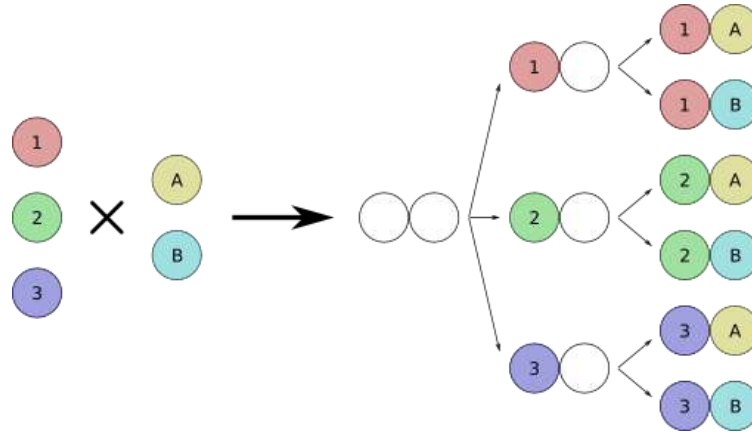
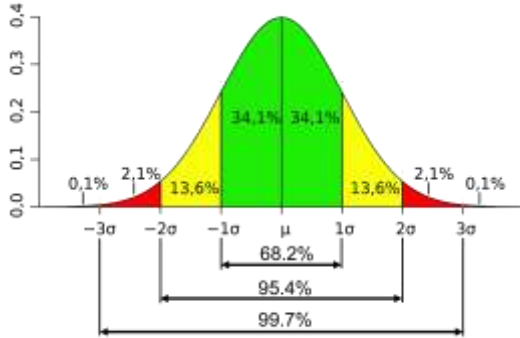
A gas may be pictured as a collection of widely spaced molecules in continuous, chaotic motion.



Highly Engineered Systems – Very Complex



Statistics → Probability → Physics



Maxwell's Equations (Differential form)	Maxwell's Equations (Integral form)
$\nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0}$	$\oint \vec{E} \cdot d\vec{s} = \frac{Q_{enc}}{\epsilon_0}$
$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$	$\oint \vec{E} \cdot d\vec{s} = -\frac{\partial}{\partial t} \oint \vec{B} \cdot d\vec{s}$
$\nabla \cdot \vec{B} = 0$	$\oint \vec{B} \cdot d\vec{s} = 0$
$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}$	$\oint \vec{B} \cdot d\vec{s} = \mu_0 \left[\oint \vec{J} \cdot d\vec{s} + \epsilon_0 \frac{\partial}{\partial t} \oint \vec{E} \cdot d\vec{s} \right]$

Predicting the Future

Why is The Past a Poor Predictor of the Future?

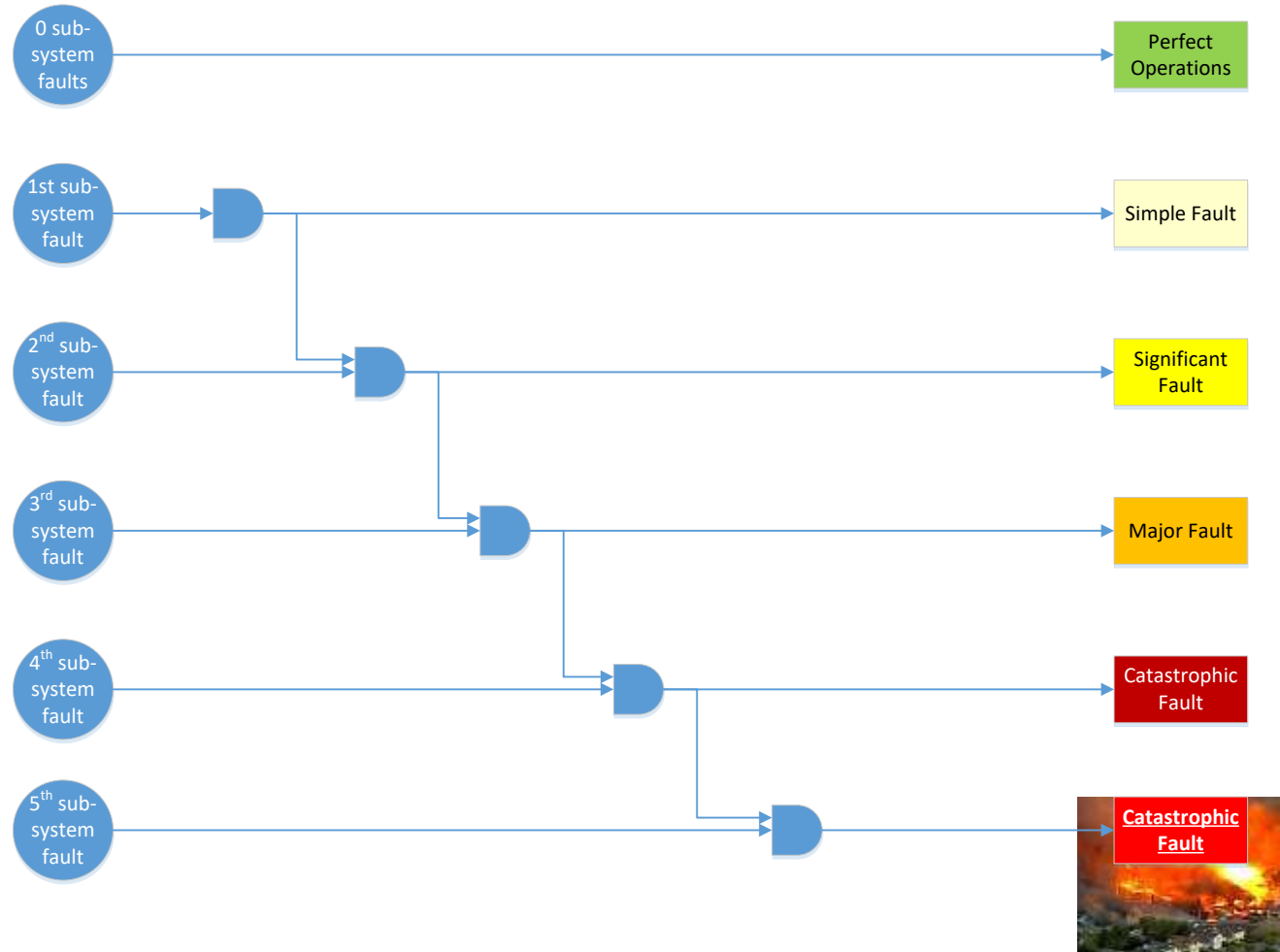
- > There are at least two reasons why the past is a poor predictor of the future:
 1. Different system states due to interactions take a long time to manifest themselves
 2. Independence of stochastic events with known probability p
 - a) The current event is not influenced by the past – the process has no memory
 - b) In reality we infer unknown p from evaluation of the past events
 - i. This is an uncertain process that we will address in upcoming slides
 - ii. We may succumb to faulty logic in thinking that the past events actually do give us the probability of occurrence p

System of Systems View

- > In systems of systems there are many possible outcome states (COMPLEXITY)
- > We need to enumerate all of these states
- > We need to count how many times each state occurs (PROBABILITY)
- > It takes many interactions over a long timeframe to see all states
- > A purely empirical approach based on reaction to what has already happened will be blind to the majority of outcome states
- > If we use a purely empirical approach we will be surprised by new lower probability outcome states
- > We will not understand how they came to be

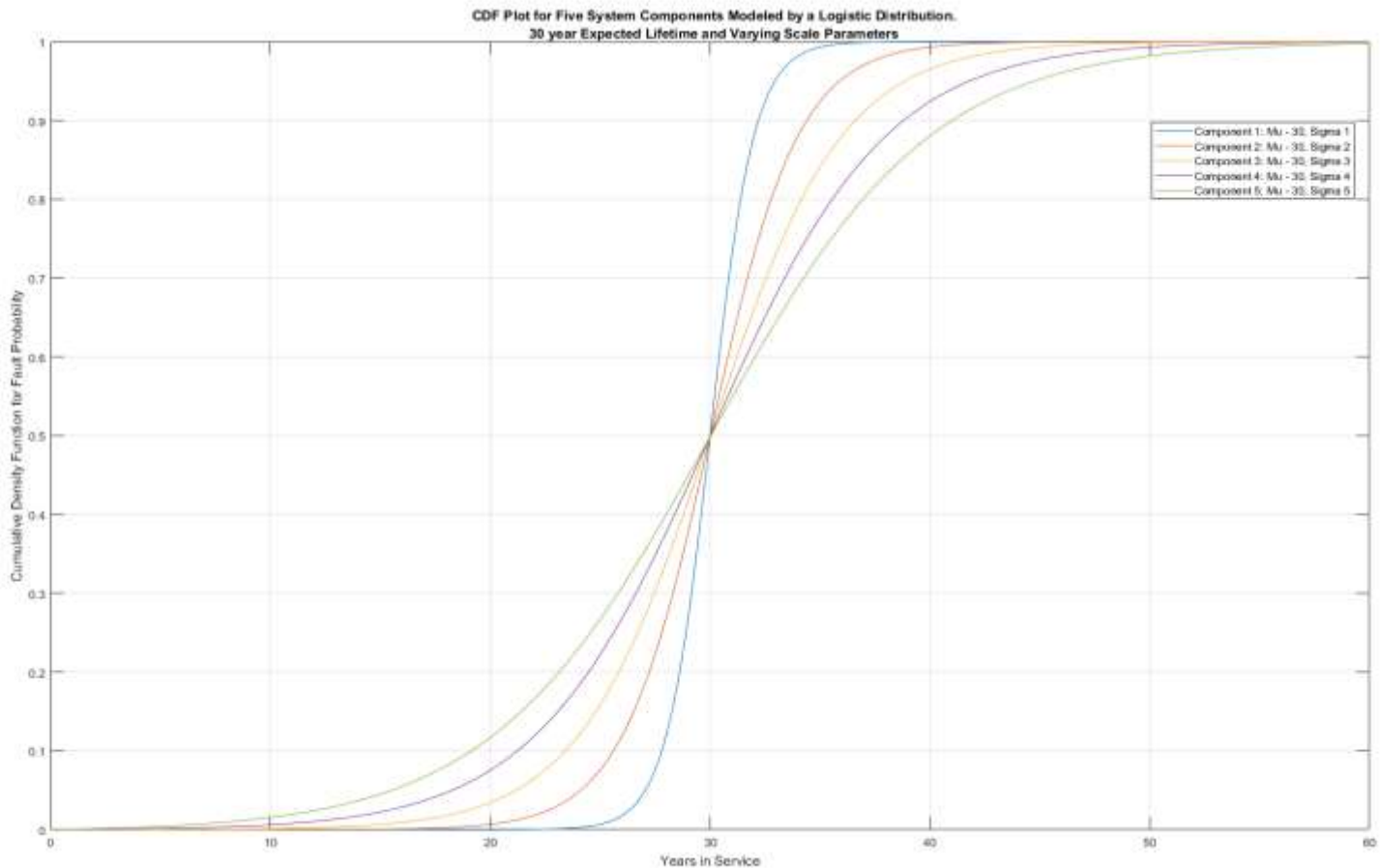
Progression of System Fault Severity Due to Sub-System Interactions

- > Working assumption that more simultaneous sub-system faults that interact leads to greater consequence
- > Backed up by NTSB and other investigations of industrial disasters



Spatio-Temporal Simulation of Interactions

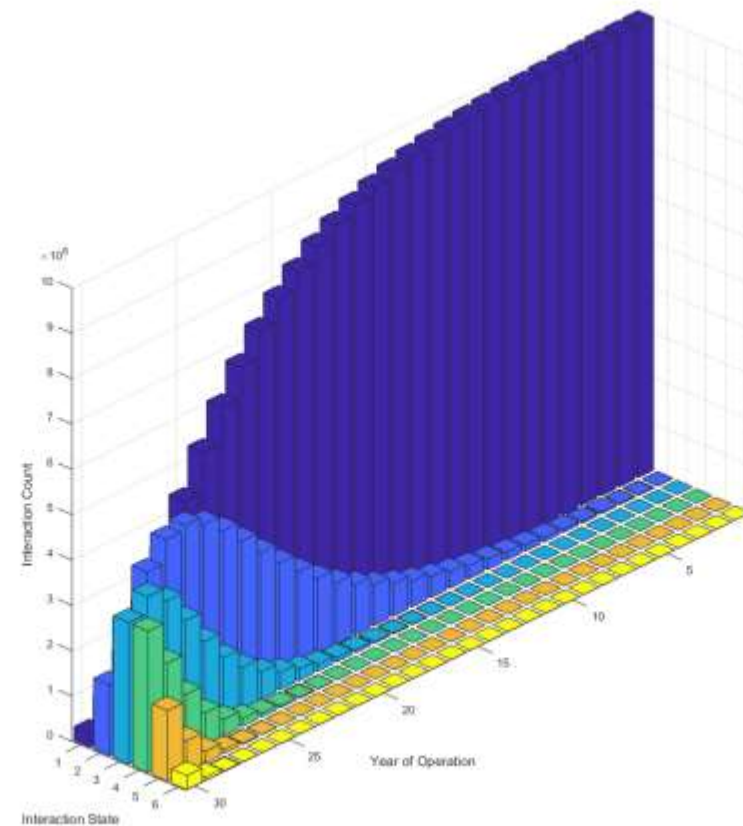
- Simulate a system with five sub-systems
- Each has a 30 year expected lifetime (50% likelihood of system anomaly)
- Each has a different rate of reaching 50% likelihood of system anomaly in 30 years
- Use logistic distributions to model each sub-system



Cumulative Interactions Over 30 Years

- 10,000 spatial locations where five sub-systems can interact
- 1,000 simulations of possible combinations of the sub-system states at each spatial location
- 10,000,000 total interactions evaluated
- Count the number occurrences of 0, 1, 2, 3, 4 or 5 faults at a location

Interaction States:
1 - No faults
2 - One fault
3 - Two interacting faults
4 - Three interacting faults
5 - Four interacting faults
6 - Five interacting faults

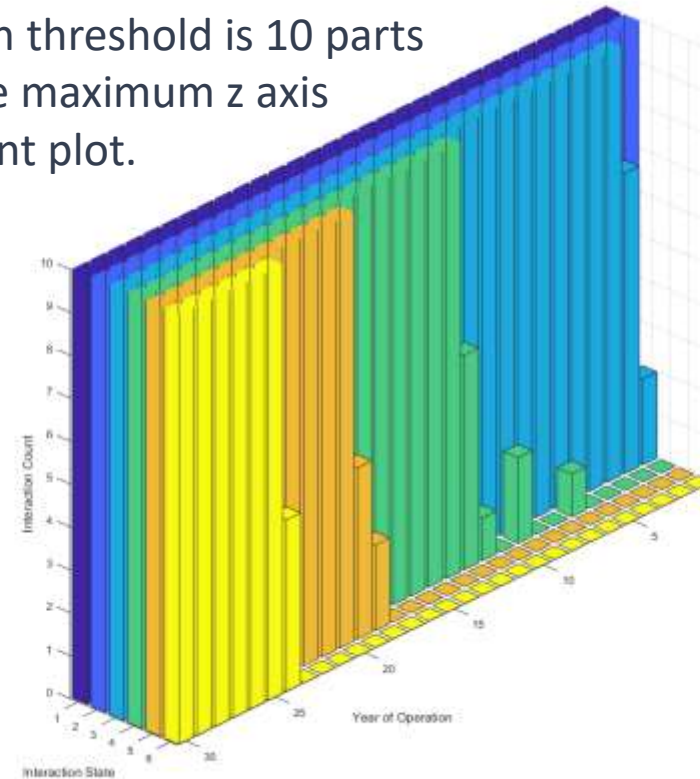


One in a Million Threshold for Cumulative Interactions

- 1 system fault always exceeds 1×10^{-6} likelihood of occurrence
- 2 interacting faults – in year 3
- 3 interacting faults – in year 13
- 4 interacting faults – in year 19
- 5 interacting faults – in year 25

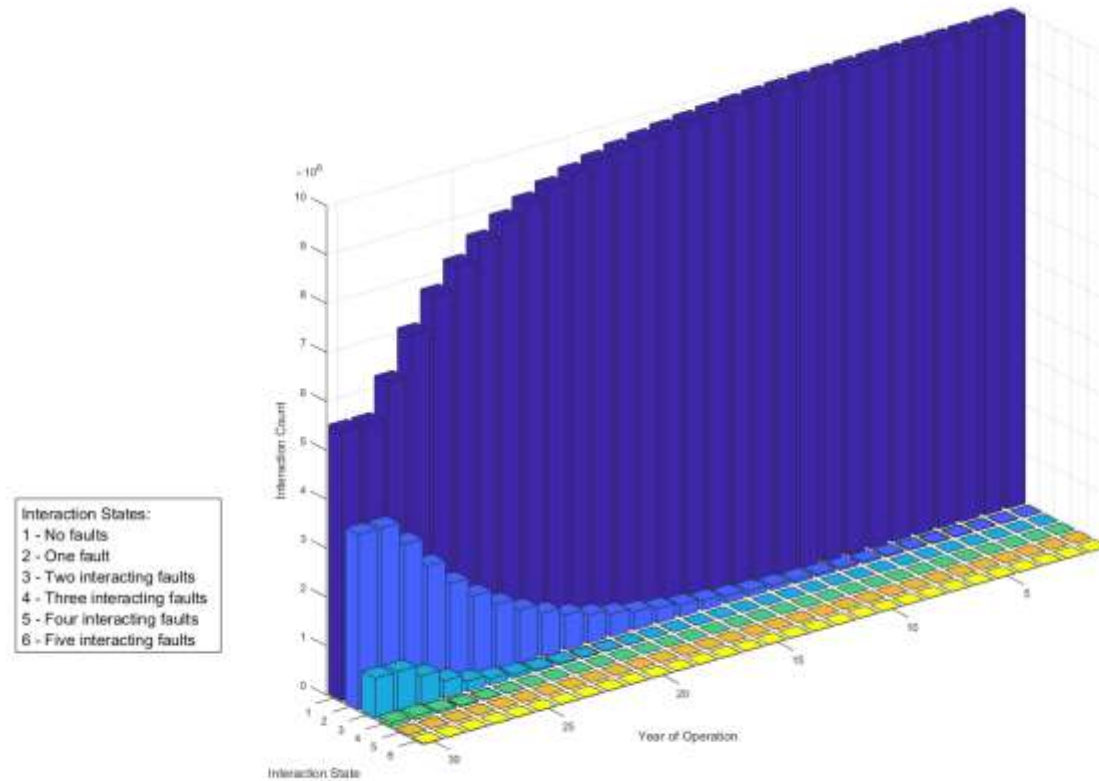
10,000,000 interactions were evaluated.
The one in a million threshold is 10 parts in 10,000,000 – the maximum z axis bound in the current plot.

Interaction States:
1 - No faults
2 - One fault
3 - Two interacting faults
4 - Three interacting faults
5 - Four interacting faults
6 - Five interacting faults



Interactions in Each One Year Timeframe

- > Count the number of occurrences of 0, 1, 2, 3, 4 or 5 faults at a location in each one year timeframe

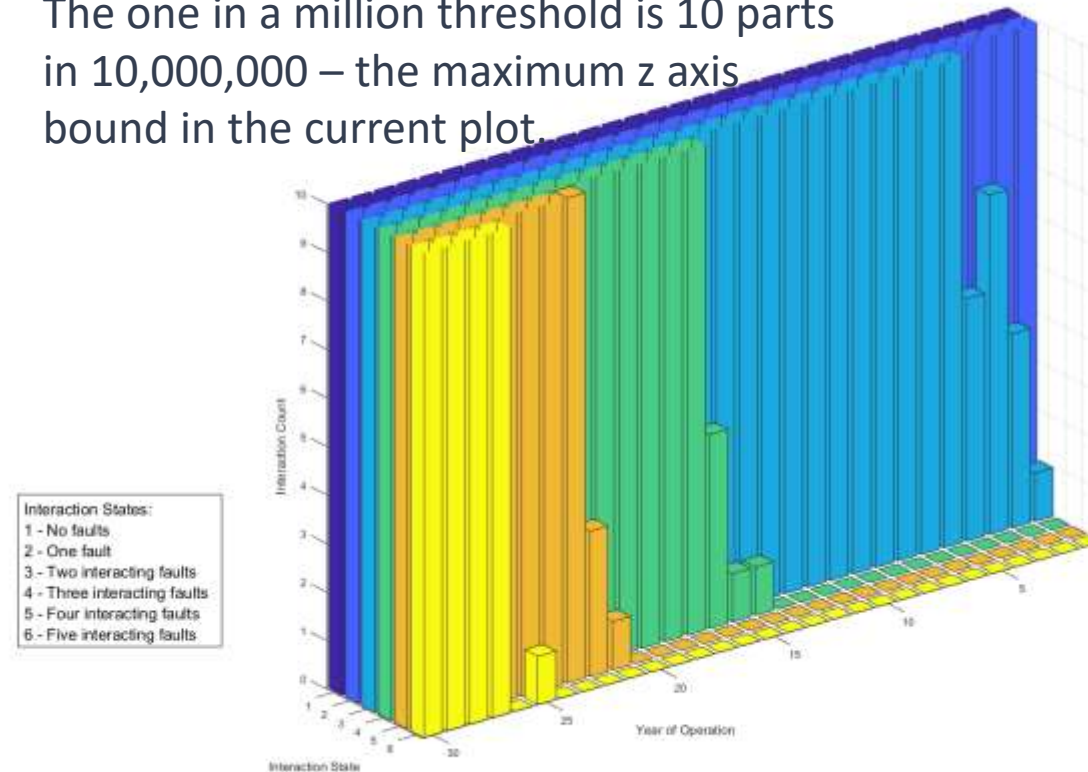


One in a Million Threshold for Interactions in a Specific Year

- 1 system fault always exceeds 1×10^{-6} likelihood of occurrence
- 2 interacting faults – in year 5
- 3 interacting faults – in year 17
- 4 interacting faults – in year 23
- 5 interacting faults – in year 27

10,000,000 interactions were evaluated.

The one in a million threshold is 10 parts in 10,000,000 – the maximum z axis bound in the current plot.



Interactions in Individual Years									Cumulative Interactions								
Year	Interactions: 0	Interactions: 1	Interactions: 2	Interactions: 3	Interactions: 4	Interactions: 5	Sum	Row Distribution	Year	Interactions: 0	Interactions: 1	Interactions: 2	Interactions: 3	Interactions: 4	Interactions: 5	Sum	Row Distribution
1	99.9116%	0.0884%	0.0000%	0.0000%	0.0000%	0.0000%	100%	█	1	99.9290%	0.0710%	0.0000%	0.0000%	0.0000%	0.0000%	100%	█
2	99.8878%	0.1122%	0.0000%	0.0000%	0.0000%	0.0000%	100%	█	2	99.8377%	0.1622%	0.0001%	0.0000%	0.0000%	0.0000%	100%	█
3	99.8627%	0.1372%	0.0001%	0.0000%	0.0000%	0.0000%	100%	█	3	99.7286%	0.2712%	0.0002%	0.0000%	0.0000%	0.0000%	100%	█
4	99.8323%	0.1677%	0.0001%	0.0000%	0.0000%	0.0000%	100%	█	4	99.5924%	0.4072%	0.0003%	0.0000%	0.0000%	0.0000%	100%	█
5	99.7908%	0.2091%	0.0001%	0.0000%	0.0000%	0.0000%	100%	█	5	99.4249%	0.5745%	0.0006%	0.0000%	0.0000%	0.0000%	100%	█
6	99.7352%	0.2647%	0.0002%	0.0000%	0.0000%	0.0000%	100%	█	6	99.2087%	0.7900%	0.0012%	0.0000%	0.0000%	0.0000%	100%	█
7	99.6725%	0.3274%	0.0002%	0.0000%	0.0000%	0.0000%	100%	█	7	98.9447%	1.0531%	0.0022%	0.0000%	0.0000%	0.0000%	100%	█
8	99.5908%	0.4088%	0.0005%	0.0000%	0.0000%	0.0000%	100%	█	8	98.6218%	1.3738%	0.0044%	0.0000%	0.0000%	0.0000%	100%	█
9	99.4910%	0.5084%	0.0007%	0.0000%	0.0000%	0.0000%	100%	█	9	98.2216%	1.7714%	0.0070%	0.0000%	0.0000%	0.0000%	100%	█
10	99.3711%	0.6281%	0.0008%	0.0000%	0.0000%	0.0000%	100%	█	10	97.7132%	2.2746%	0.0122%	0.0000%	0.0000%	0.0000%	100%	█
11	99.2142%	0.7842%	0.0016%	0.0000%	0.0000%	0.0000%	100%	█	11	97.0918%	2.8873%	0.0209%	0.0000%	0.0000%	0.0000%	100%	█
12	99.0143%	0.9833%	0.0024%	0.0000%	0.0000%	0.0000%	100%	█	12	96.3107%	3.6538%	0.0355%	0.0001%	0.0000%	0.0000%	100%	█
13	98.7788%	1.2170%	0.0043%	0.0000%	0.0000%	0.0000%	100%	█	13	95.3453%	4.5980%	0.0566%	0.0001%	0.0000%	0.0000%	100%	█
14	98.4681%	1.5246%	0.0073%	0.0000%	0.0000%	0.0000%	100%	█	14	94.1601%	5.7510%	0.0886%	0.0004%	0.0000%	0.0000%	100%	█
15	98.0881%	1.9005%	0.0114%	0.0000%	0.0000%	0.0000%	100%	█	15	92.6698%	7.1842%	0.1451%	0.0009%	0.0000%	0.0000%	100%	█
16	97.6044%	2.3771%	0.0185%	0.0000%	0.0000%	0.0000%	100%	█	16	90.8416%	8.9159%	0.2408%	0.0018%	0.0000%	0.0000%	100%	█
17	97.0058%	2.9655%	0.0287%	0.0001%	0.0000%	0.0000%	100%	█	17	88.5611%	11.0462%	0.3887%	0.0041%	0.0000%	0.0000%	100%	█
18	96.2597%	3.6929%	0.0471%	0.0003%	0.0000%	0.0000%	100%	█	18	85.8232%	13.5455%	0.6229%	0.0084%	0.0000%	0.0000%	100%	█
19	95.3271%	4.5979%	0.0745%	0.0005%	0.0000%	0.0000%	100%	█	19	82.4149%	16.5727%	0.9941%	0.0183%	0.0000%	0.0000%	100%	█
20	94.1586%	5.7194%	0.1208%	0.0012%	0.0000%	0.0000%	100%	█	20	78.2841%	20.1037%	1.5725%	0.0397%	0.0001%	0.0000%	100%	█
21	92.7001%	7.1021%	0.1957%	0.0021%	0.0000%	0.0000%	100%	█	21	73.3514%	24.0689%	2.4918%	0.0872%	0.0007%	0.0000%	100%	█
22	90.8786%	8.8082%	0.3089%	0.0043%	0.0000%	0.0000%	100%	█	22	67.4605%	28.4699%	3.8764%	0.1909%	0.0023%	0.0000%	100%	█
23	88.6488%	10.8461%	0.4952%	0.0098%	0.0001%	0.0000%	100%	█	23	60.5933%	33.0227%	5.9733%	0.4030%	0.0078%	0.0000%	100%	█
24	85.8199%	13.3694%	0.7894%	0.0211%	0.0002%	0.0000%	100%	█	24	52.7594%	37.3374%	9.0233%	0.8544%	0.0254%	0.0000%	100%	█
25	82.1861%	16.4833%	1.2823%	0.0475%	0.0008%	0.0000%	100%	█	25	44.0107%	40.8315%	13.2922%	1.7827%	0.0823%	0.0005%	100%	█
26	77.4109%	20.3721%	2.1082%	0.1061%	0.0026%	0.0000%	100%	█	26	34.6258%	42.6248%	18.8931%	3.5880%	0.2643%	0.0042%	100%	█
27	70.8316%	25.3304%	3.5790%	0.2497%	0.0091%	0.0001%	100%	█	27	25.0283%	41.5021%	25.5620%	7.0650%	0.8160%	0.0267%	100%	█
28	62.4215%	31.1295%	5.8958%	0.5304%	0.0225%	0.0002%	100%	█	28	15.8983%	36.3955%	31.8854%	13.1573%	2.4975%	0.1661%	100%	█
29	55.1829%	35.7709%	8.1601%	0.8448%	0.0404%	0.0009%	100%	█	29	8.2257%	26.9949%	34.9249%	22.1517%	6.8731%	0.8298%	100%	█
30	55.1381%	35.7929%	8.1784%	0.8507%	0.0393%	0.0007%	100%	█	30	3.1357%	15.6546%	31.2686%	31.2221%	15.6024%	3.1165%	100%	█
Column Distribution									Column Distribution								

Probability Distributions: Predicting the Next Outcome

How to Develop Distributions for Risk Models

- > We need to identify possible system states
- > Each time we inspect the system it is either in a particular state or not
- > We count each outcome separately –is in state k , is not in state k
- > This is a Bernoulli process

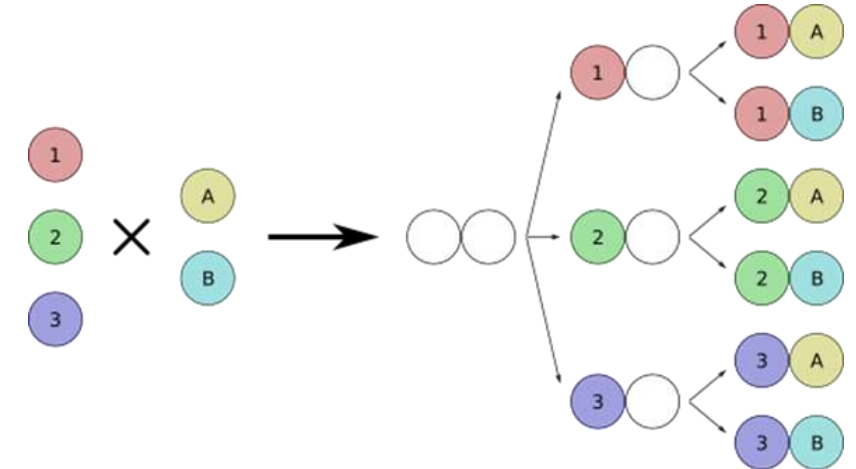
A **Bernoulli process** is a finite or infinite sequence of independent random variables X_1, X_2, X_3, \dots , such that

- For each i , the value of X_i is either 0 or 1;
- For all values of i , the probability that $X_i = 1$ is the same number p .

In other words, a Bernoulli process is a sequence of independent identically distributed Bernoulli trials.

Independence of the trials implies that the process is memoryless. **Given that the probability p is known, past outcomes provide no information about future outcomes.** (If p is unknown, however, the past informs about the future indirectly, through inferences about p .)

https://en.wikipedia.org/wiki/Bernoulli_process



$$\begin{matrix} b(p) & & b(p) & & b(p) & & b(p) & & b(p) & & b(p) \\ \bullet & + & \circ & + & \circ & + & \bullet & + & \bullet & + & \bullet \\ 1 & & 0 & & 0 & & 1 & & 1 & & 1 \end{matrix} = \begin{matrix} B(6,p) \\ \bullet \bullet \bullet \bullet \bullet \bullet \\ 4 \end{matrix}$$

$$f(k; p) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

The Past Informs About the Future Indirectly, Through Inferences About p

- > Counting the number of times a system is in state k gives an estimate of the probability p of being in state k
- > How certain are we about this inference?
- > Bayesian Inference using the Beta distribution quantifies the uncertainty in our inference

In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behavior of percentages and proportions.

https://en.wikipedia.org/wiki/Beta_distribution

$$\begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \circ \\ 0 \end{array} + \begin{array}{c} b(p) \\ \circ \\ 0 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} = \begin{array}{c} B(6,p) \\ \bullet \bullet \bullet \bullet \\ 4 \end{array}$$

$$P \approx 4/6 = 2/3. \quad (1-p) \approx 1/3$$

Beta(4,2) and Beta(5,3)

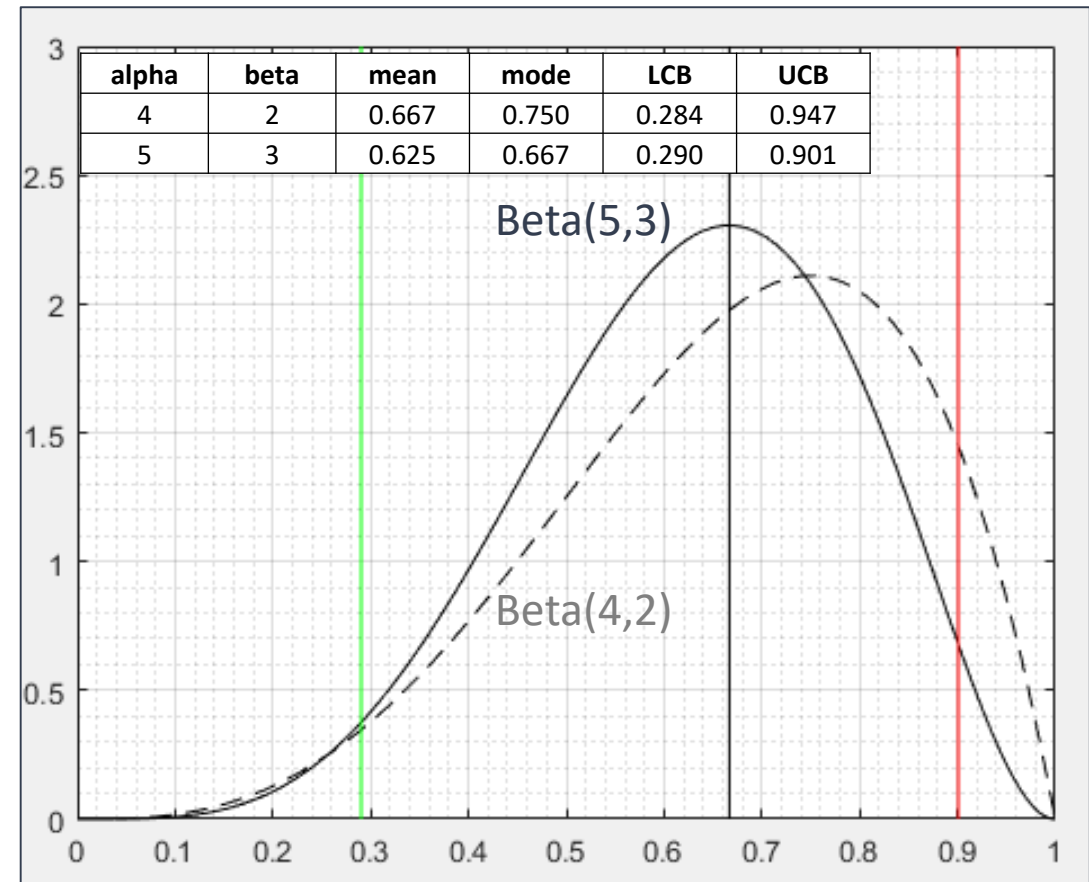
> The Beta distribution has two parameters:

- Beta(alpha, beta)
- Alpha – count of successes
- Beta – count of non-successes

$$\text{mean} = \frac{\alpha}{\alpha + \beta} \quad \text{mode} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

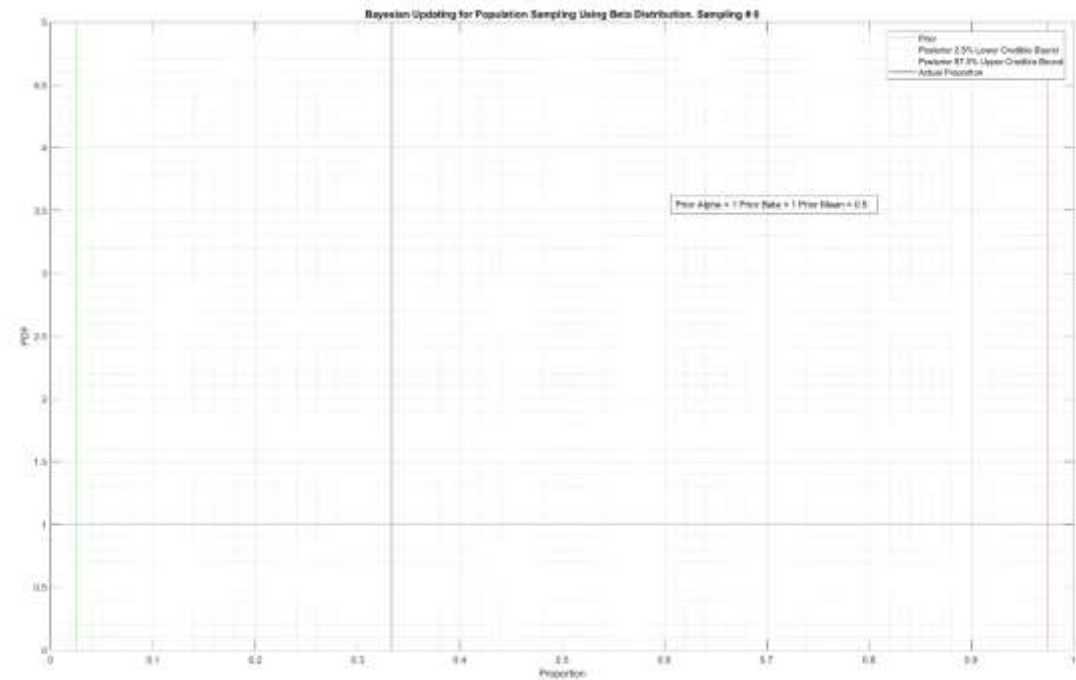
$$\begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \circ \\ 0 \end{array} + \begin{array}{c} b(p) \\ \circ \\ 0 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} + \begin{array}{c} b(p) \\ \bullet \\ 1 \end{array} = \begin{array}{c} B(6,p) \\ \bullet \bullet \bullet \bullet \\ 4 \end{array}$$

$$P \cong 4/6 = 2/3. \quad (1-p) \cong 1/3$$



How to Develop Beta Distributions from Data

1. Start by defining our initial belief about the probability of success or failure
 - a) In the absence of solid knowledge use an ignorant prior: $\text{Beta}(1,1)$
 - b) An ignorant prior defines the state where we have a 50% probability of success or failure, but any proportion of success or failure is equally likely
 - c) In other words we do not have a clue



Bayesian Updating for the Beta Distribution is Simple and Painless

2. Setup a simple spreadsheet
3. Enter an ignorant prior in the first cells
4. Use the Excel function `BETA.INV(D$1,B4,C4)` to calculate the credible bounds
5. Calculate the mean directly from the data

Credible Bounds		95%	0.025	0.975		
Data Point	Alpha	Beta	LCB	Mean	UCB	TRUE / FALSE
0	1	1	0.025	0.500	0.975	N/A
1						
2						

Credible Bounds		0.95	=(1-C1)/2	=1-D1		
Data Point	Alpha	Beta	LCB	Mean	UCB	TRUE / FALSE
0	1	1	$\text{=BETA.INV(D$1,B4,C4)}$	$\text{=IF(G4<>CHAR(32),B4/(B4+C4),CHAR(32))}$	$\text{=BETA.INV(E$1,B4,C4)}$	N/A
1	$\text{=IF(G5<>CHAR(32),B4+G5,CHAR(32))}$	$\text{=IF(G5<>CHAR(32),C4+(1-G5),CHAR(32))}$	$\text{=IF(G5<>CHAR(32),BETA.INV(D$1,B5,C5),CHAR(32))}$	$\text{=IF(G5<>CHAR(32),B5/(B5+C5),CHAR(32))}$	$\text{=IF(G5<>CHAR(32),BETA.INV(E$1,B5,C5),CHAR(32))}$	=CHAR(32)
2	$\text{=IF(G6<>CHAR(32),B5+G6,CHAR(32))}$	$\text{=IF(G6<>CHAR(32),C5+(1-G6),CHAR(32))}$	$\text{=IF(G6<>CHAR(32),BETA.INV(D$1,B6,C6),CHAR(32))}$	$\text{=IF(G6<>CHAR(32),B6/(B6+C6),CHAR(32))}$	$\text{=IF(G6<>CHAR(32),BETA.INV(E$1,B6,C6),CHAR(32))}$	=CHAR(32)

Bayesian Updating for the Beta Distribution is Simple and Painless

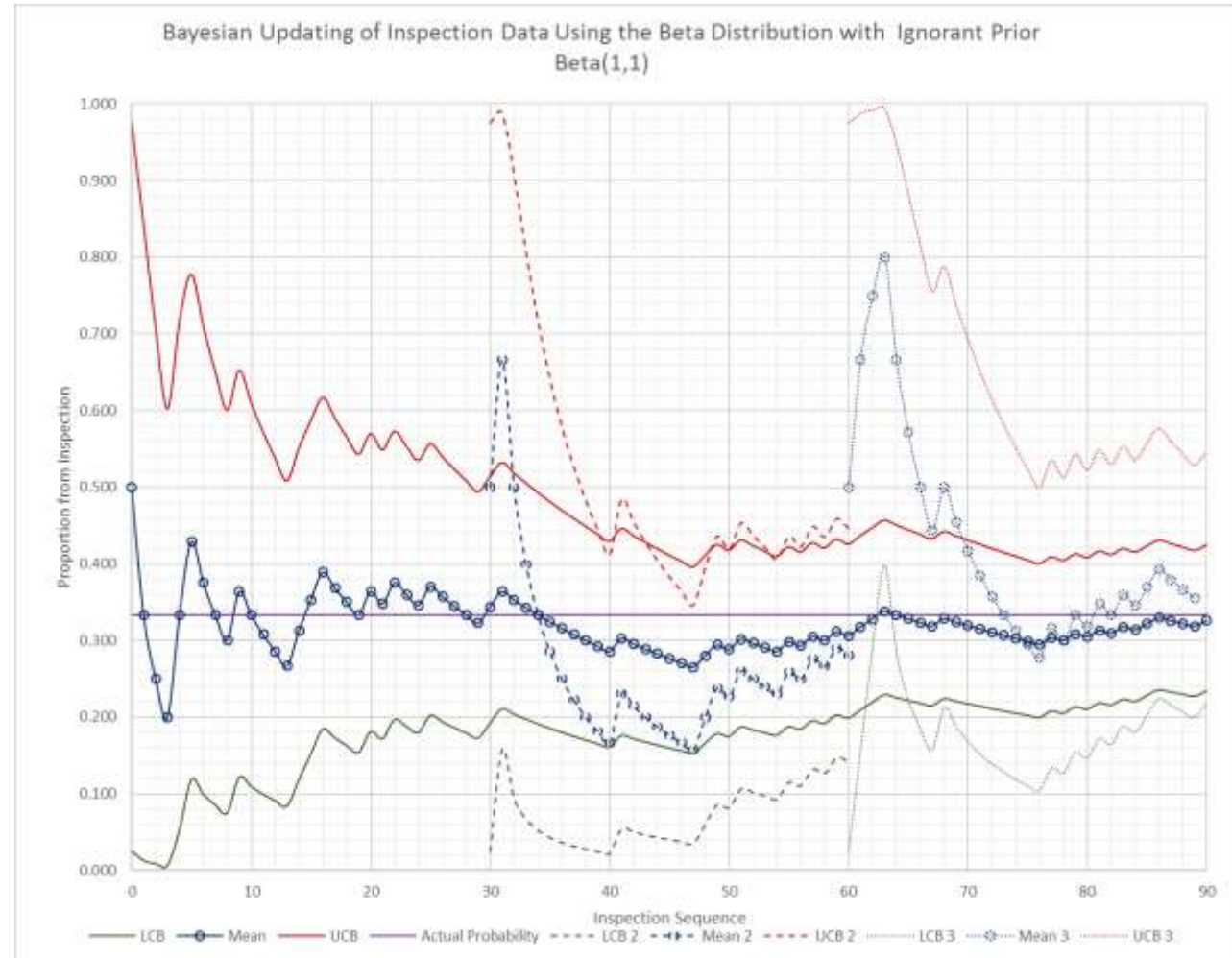
- 6. Increment the Alpha column by 1 for a successful outcome
- 7. Increment the Beta column by one for an unsuccessful outcome

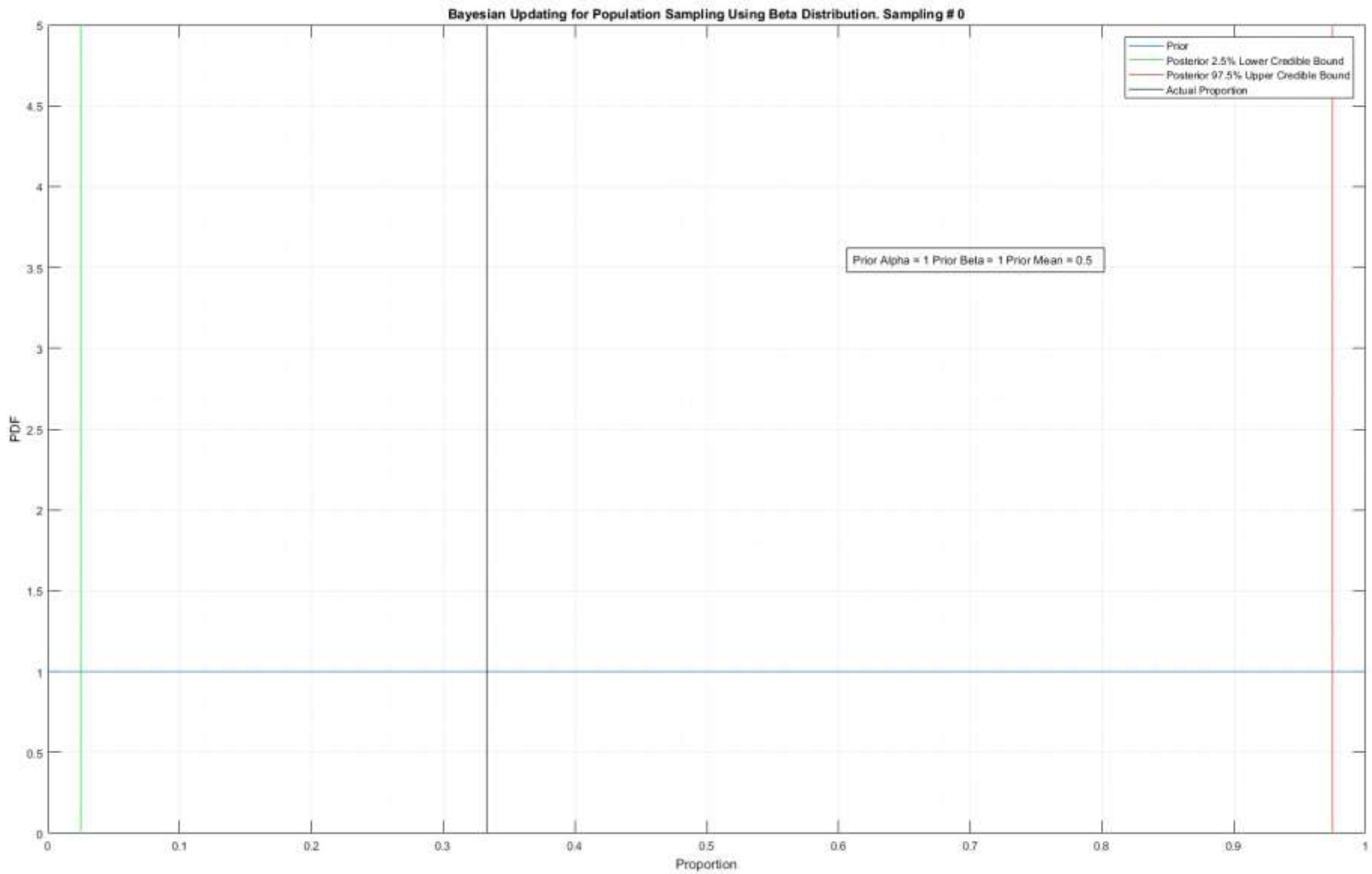
Credible Bounds		95%	0.025	0.975		
Data Point	Alpha	Beta	LCB	Mean	UCB	TRUE / FALSE
0	1	1	0.025	0.500	0.975	N/A
1	2	1	0.158	0.667	0.987	TRUE
2	2	2	0.094	0.500	0.906	FALSE

Credible Bounds		0.95	=(1-C1)/2	=1-D1		
Data Point	Alpha	Beta	LCB	Mean	UCB	TRUE / FALSE
0	1	1	=BETA.INV(D\$1,B4,C4)	=IF(G4<>CHAR(32),B4/(B4+C4),CHAR(32))	=BETA.INV(E\$1,B4,C4)	N/A
1	=IF(G5<>CHAR(32),B4+G5,CHAR(32))	=IF(G5<>CHAR(32),C4+(1-G5),CHAR(32))	=IF(G5<>CHAR(32),BETA.INV(D\$1,B5,C5),CHAR(32))	=IF(G5<>CHAR(32),B5/(B5+C5),CHAR(32))	=IF(G5<>CHAR(32),BETA.INV(E\$1,B5,C5),CHAR(32))	TRUE
2	=IF(G6<>CHAR(32),B5+G6,CHAR(32))	=IF(G6<>CHAR(32),C5+(1-G6),CHAR(32))	=IF(G6<>CHAR(32),BETA.INV(D\$1,B6,C6),CHAR(32))	=IF(G6<>CHAR(32),B6/(B6+C6),CHAR(32))	=IF(G6<>CHAR(32),BETA.INV(E\$1,B6,C6),CHAR(32))	FALSE

Bayesian Updating for an inspection Process

Credible Bounds		95%	0.025	0.975	Inspection Proportion			0.333
Data Point	Alpha	Beta	LCB	Mean	UCB	TRUE / FALSE	Inspections	Actual Probability
0	1	1	0.025	0.500	0.975	N/A	N/A	0.333
1	1	2	0.013	0.333	0.842	FALSE	0	0.333
2	1	3	0.008	0.250	0.708	FALSE	0	0.333
3	1	4	0.006	0.200	0.602	FALSE	0	0.333
4	2	4	0.053	0.333	0.716	TRUE	1	0.333
5	3	4	0.118	0.429	0.777	TRUE	1	0.333
6	3	5	0.099	0.375	0.710	FALSE	0	0.333
7	3	6	0.085	0.333	0.651	FALSE	0	0.333
8	3	7	0.075	0.300	0.600	FALSE	0	0.333
9	4	7	0.122	0.364	0.652	TRUE	1	0.333
10	4	8	0.109	0.333	0.610	FALSE	0	0.333
11	4	9	0.099	0.308	0.572	FALSE	0	0.333
12	4	10	0.091	0.286	0.538	FALSE	0	0.333
13	4	11	0.084	0.267	0.508	FALSE	0	0.333
14	5	11	0.118	0.313	0.551	TRUE	1	0.333
15	6	11	0.152	0.353	0.587	TRUE	1	0.333
16	7	11	0.184	0.389	0.617	TRUE	1	0.333
17	7	12	0.173	0.368	0.590	FALSE	0	0.333
18	7	13	0.163	0.350	0.566	FALSE	0	0.333
19	7	14	0.154	0.333	0.543	FALSE	0	0.333
20	8	14	0.181	0.364	0.570	TRUE	1	0.333
21	8	15	0.172	0.348	0.549	FALSE	0	0.333
22	9	15	0.197	0.375	0.573	TRUE	1	0.333
23	9	16	0.188	0.360	0.553	FALSE	0	0.333
24	9	17	0.180	0.346	0.535	FALSE	0	0.333
25	10	17	0.202	0.370	0.557	TRUE	1	0.333
26	10	18	0.194	0.357	0.540	FALSE	0	0.333
27	10	19	0.186	0.345	0.524	FALSE	0	0.333
28	10	20	0.179	0.333	0.508	FALSE	0	0.333
29	10	21	0.173	0.323	0.494	FALSE	0	0.333
30	11	21	0.192	0.344	0.514	TRUE	1	0.333





Example of Beta Distributions and Bayesian Updating Used to Evaluate Risk Under Uncertainty

- > 22 out of 365 fusions failed in service
- > Random sampling plan developed to assess quality of non-failed fusions
- > Hypergeometric distribution used to predict findings based on prior data – equivalent to an exact Fisher test for significance

Size	Quantity on PDF 1/12/2017	Proportion	Samples	Prior Probability of Failure	Prior Probability of Finding Zero Failures in Sampled Saddles	Prior Probability of Finding One Failure in Sampled Saddles	Prior Probability of Finding Two Failures in Sampled Saddles	Prior Probability of Finding Three Failures in Sampled Saddles	Sum of Prior Probabilities of Detection
16	67	18.36%	5	1.41%	81.56%	17.34%	1.08%	0.02%	100%
14	26	7.12%	2	0.00%	100.00%	0.00%	0.00%	0.00%	100%
12	28	7.67%	2	37.93%	64.40%	31.33%	4.12%	0.14%	100%
10	46	12.60%	3	13.33%	33.50%	50.25%	16.26%	0.00%	100%
8	58	15.89%	4	0.00%	100.00%	0.00%	0.00%	0.00%	100%
6	140	38.36%	9	2.17%	92.96%	7.04%	0.00%	0.00%	100%
Sum	365	100.00%	25						

Size	Count	Field Failures	Field Failure Percent	Prior Beta Distribution alpha	Prior Beta Distribution beta	Prior Probability of failure	Sampling Plan Failures	Sampling Plan Non-Failures	Posterior Beta Distribution alpha	Posterior Beta Distribution beta	Posterior Probability of Failure
16	71	1	1.41%	0.028	1.972	1.41%	5	0	5.028	1.972	71.83%
14	22	0	0.00%	1.000	1.000	50.00%	2	0	3.000	1.000	75.00%
12	29	11	37.93%	0.759	1.241	37.93%	2	0	2.759	1.241	68.97%
10	45	6	13.33%	0.267	1.733	13.33%	3	0	3.267	1.733	65.33%
8	60	0	0.00%	1.000	1.000	50.00%	4	0	5.000	1.000	83.33%
6	138	3	2.17%	0.043	1.957	2.17%	9	0	9.043	1.957	82.21%
Total	365	21	5.75%	0.115	1.885	5.75%	25	0	25.115	1.885	93.02%

Credible Bounds for Sub-Populations and Entire Population

- > Credible bounds for each sub-population heavily impacted by number of samples
- > All results are statistically significant at the 0.05 level per the hypergeometric prior probabilities
- > Can combine sub-populations as the installations were performed by a single contractor, using a single process in the same time frame.

Size	Posterior Beta Distribution alpha	Posterior Beta Distribution beta	Posterior Probability of Failure	2.5% Lower Credible Bound	97.5% Upper Credible Bound
16	5.028	1.972	71.83%	37.1%	95.8%
14	3.000	1.000	75.00%	30.2%	99.2%
12	2.759	1.241	68.97%	24.0%	98.0%
10	3.267	1.733	65.33%	24.7%	95.5%
8	5.000	1.000	83.33%	48.7%	99.5%
6	9.043	1.957	82.21%	56.7%	97.6%
Entire Population (Not additive with results above)	25.115	1.885	93.02%	81.4%	99.2%

Incorporating Data Quality Metrics

Simple Scoring of Data Quality Attributes

- > At the March meeting of the Risk Model Work Group we discussed methods for assigning data quality scores
- > We can subjectively choose a line of demarcation for the Weights Score :
 - ≥ 12.5 Good
 - < 12.5 Bad

Pedigree Level	Score
A	15
B or Default	10
C	5
D	3
No Info on Data Field	1

Integrity Agreement	Score			
	Authenticity	Compliance	Transparency	Reliability
Yes	15	15	15	15
Partial	10	10	10	10
Default Value	5	5	5	5
No	3	3	3	3
No Info on Data Field	1	1	1	1

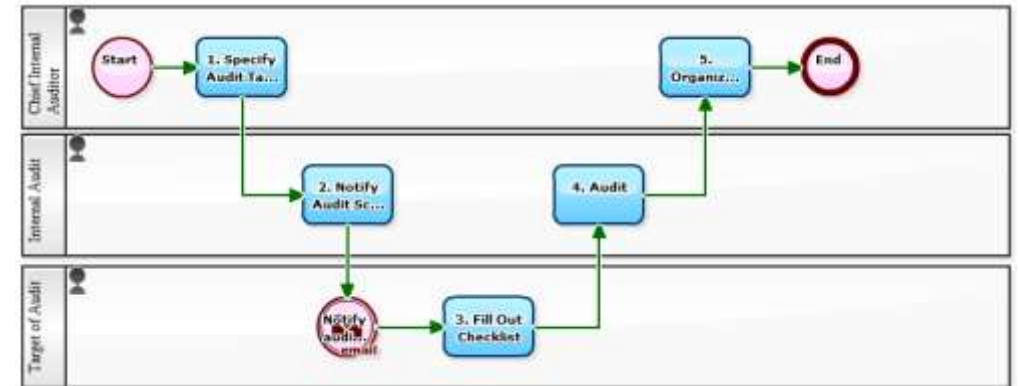
Component		Score / Level	Roll-Ups
Integrity	Authenticity	10	12.5
	Compliance	15	
	Transparency	15	
	Reliability	10	
Pedigree		10	10.0
Weighted Score		75% Pedigree + 25% Integrity	10.6

How to Implement the Data Quality Score

- > For an poorly understood system the data quality score is arrived at through an audit process



- > How do we extrapolate the audit results to the system as a whole?



Two Sub-Systems: 5,000 Data Aggregations Each

- > Sub-system 1:
- > Use the following stationary distribution :

- 90 Audits

- > 60 Bad

- > 30 Good

- > Sub-system 2:

- 6 Audits

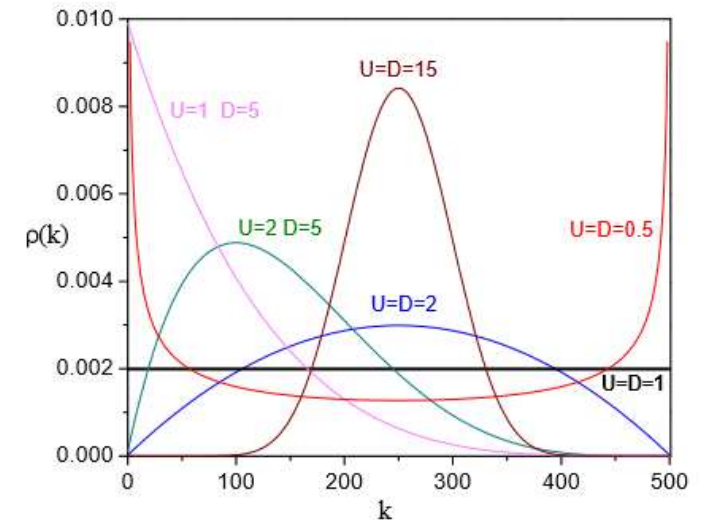
- > 2 Bad

- > 4 Good

- > This distribution estimates the number of system components k out of a total of N that are in a good state given D negative (Down) and U positive (Up) influences (Bad and Good audit results respectively)

- > The distribution behaves very similarly to the Beta distribution

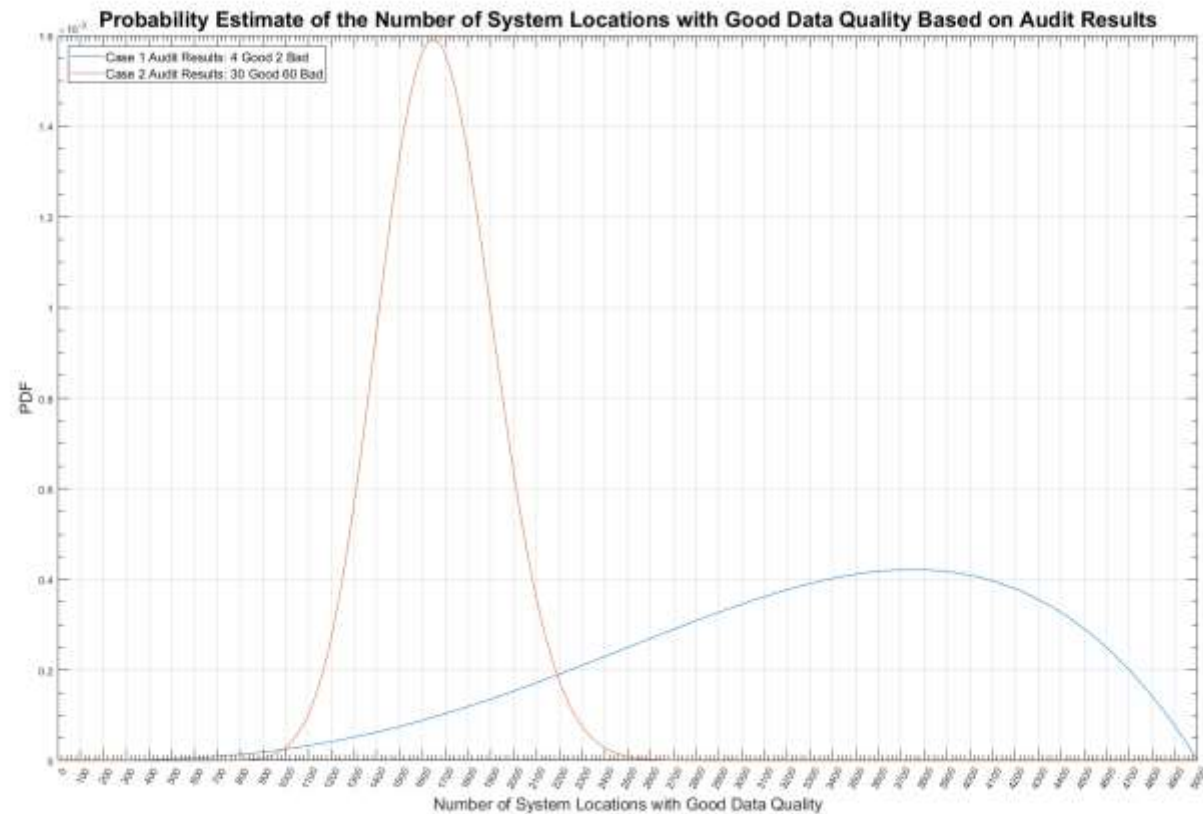
$$\rho(k) = \frac{\binom{U+k-1}{k} \binom{N+D-k-1}{N-k}}{\binom{N+D+U-1}{N}}.$$



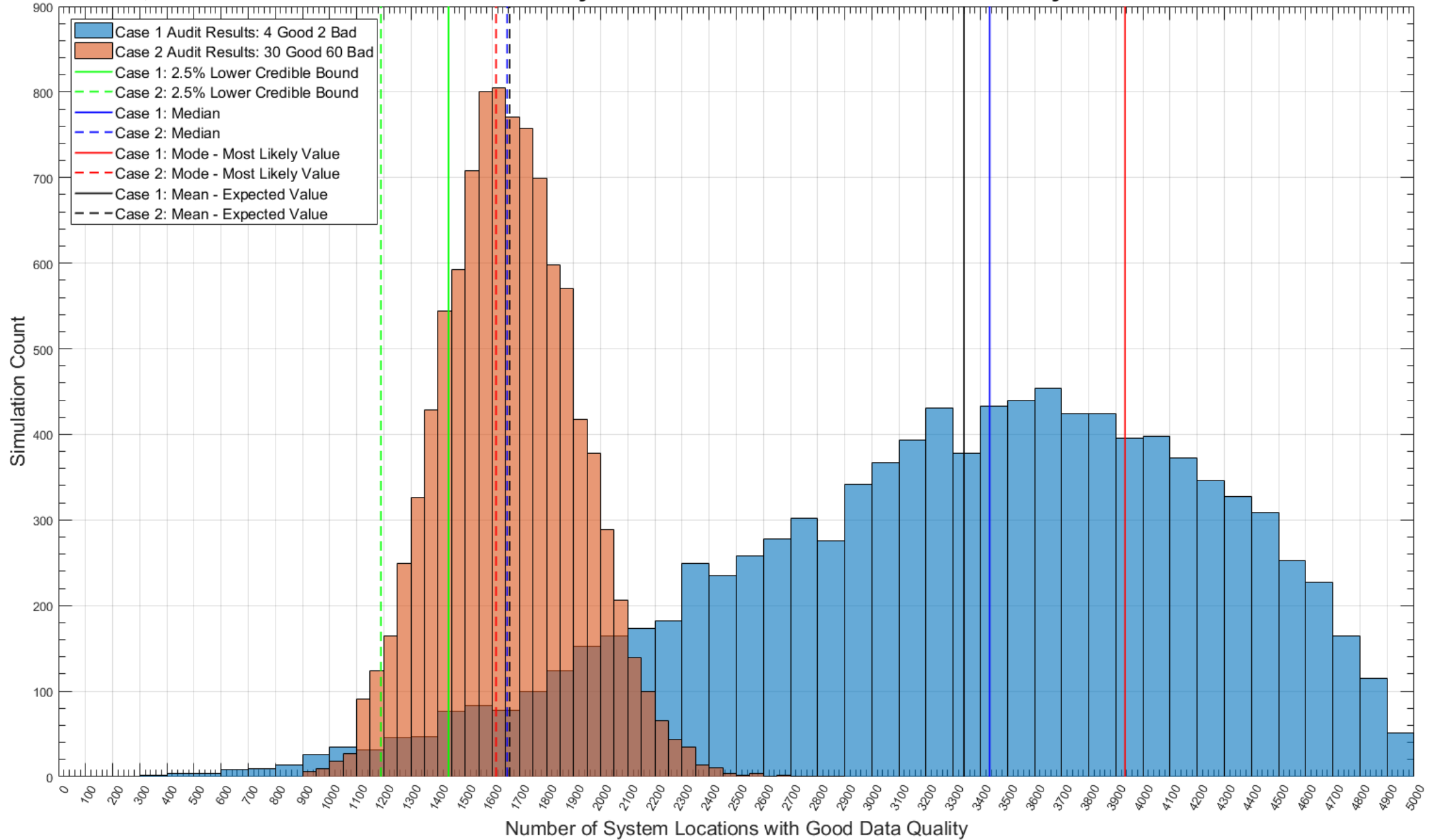
Harmon, D., et al., Predicting economic market crises using measures of collective panic. 2011.

Probability Distribution for The Audit Results Example

- > The two probability density functions obtained for the two audit cases presented can be used to simulate locations with good data quality based on the audit results

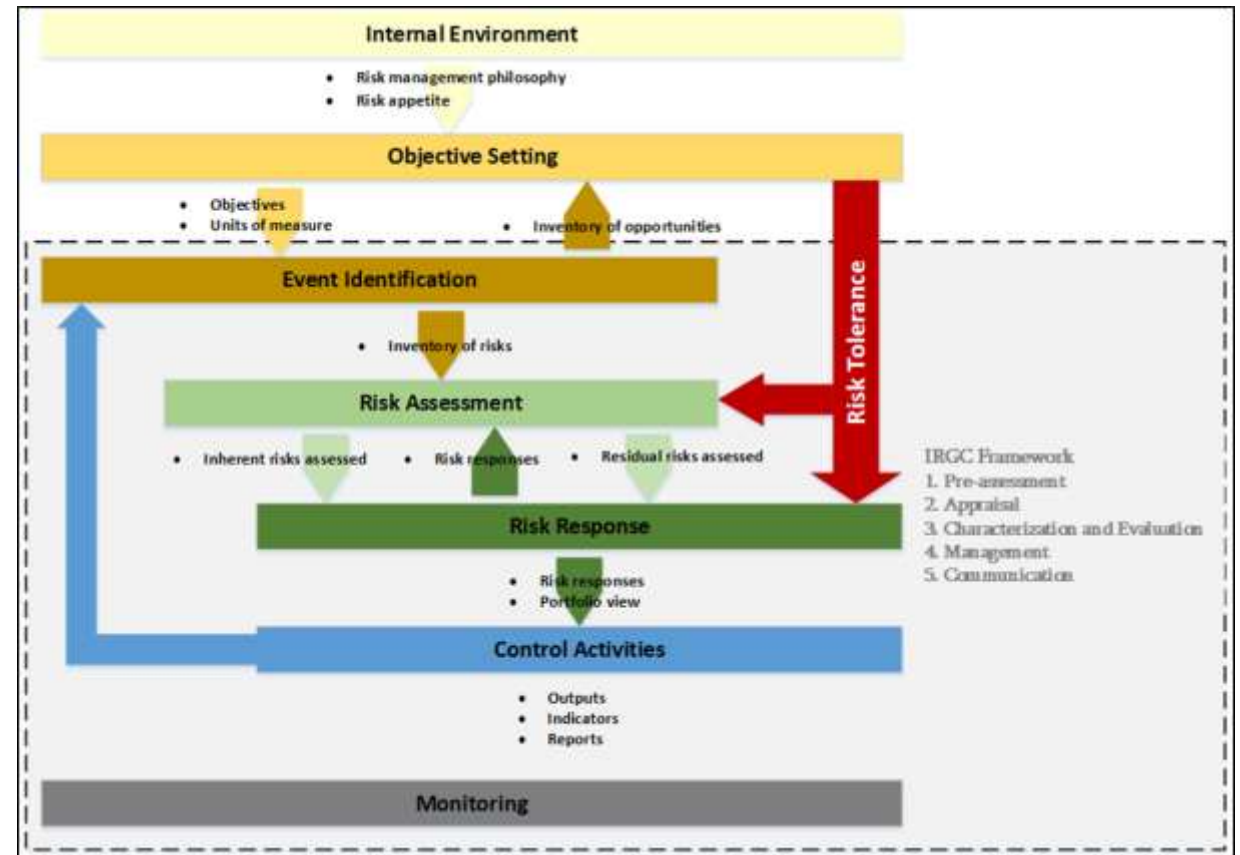
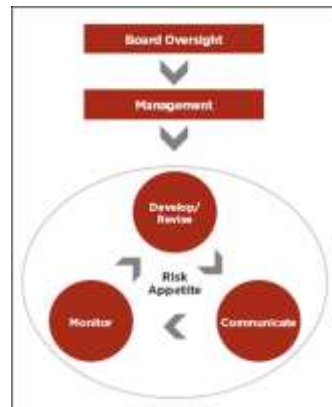


10,000 Simulations of the Number of System Locations with Good Data Quality Based on Audit Results



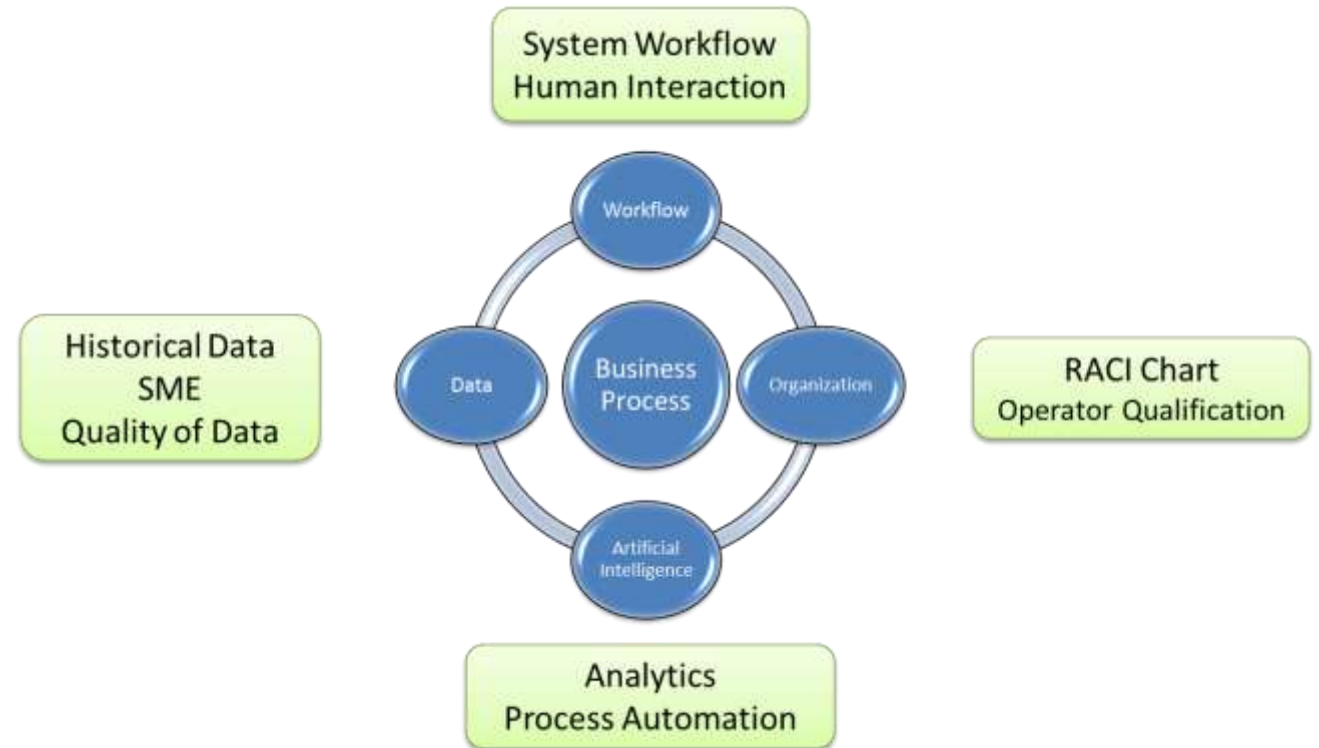
What to do with the Data Quality Distribution Results

- This is a risk tolerance discussion that has to tie back in to the **Risk Governance/ Risk Management/ Risk Assessment** framework of the organization.

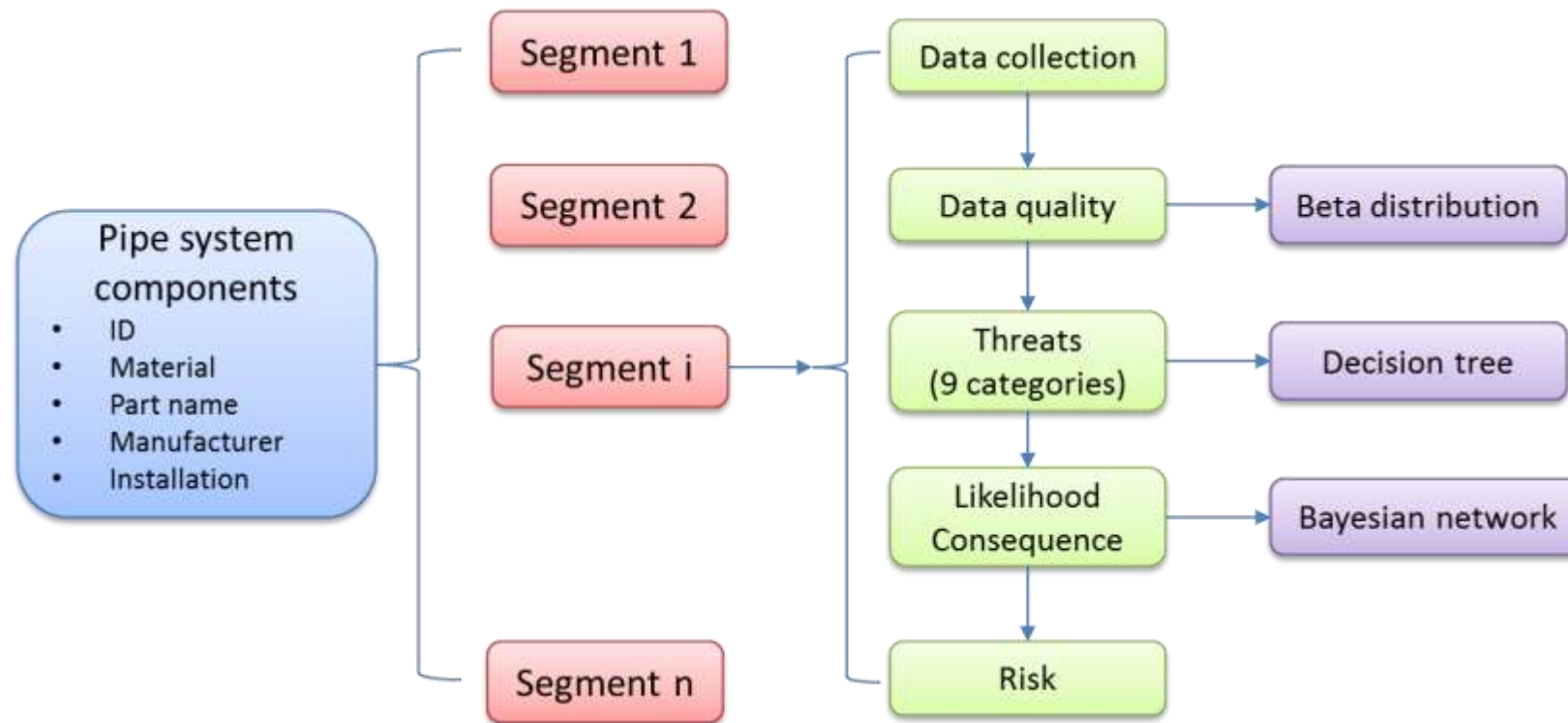


How to Address the Risk Tolerance Problem

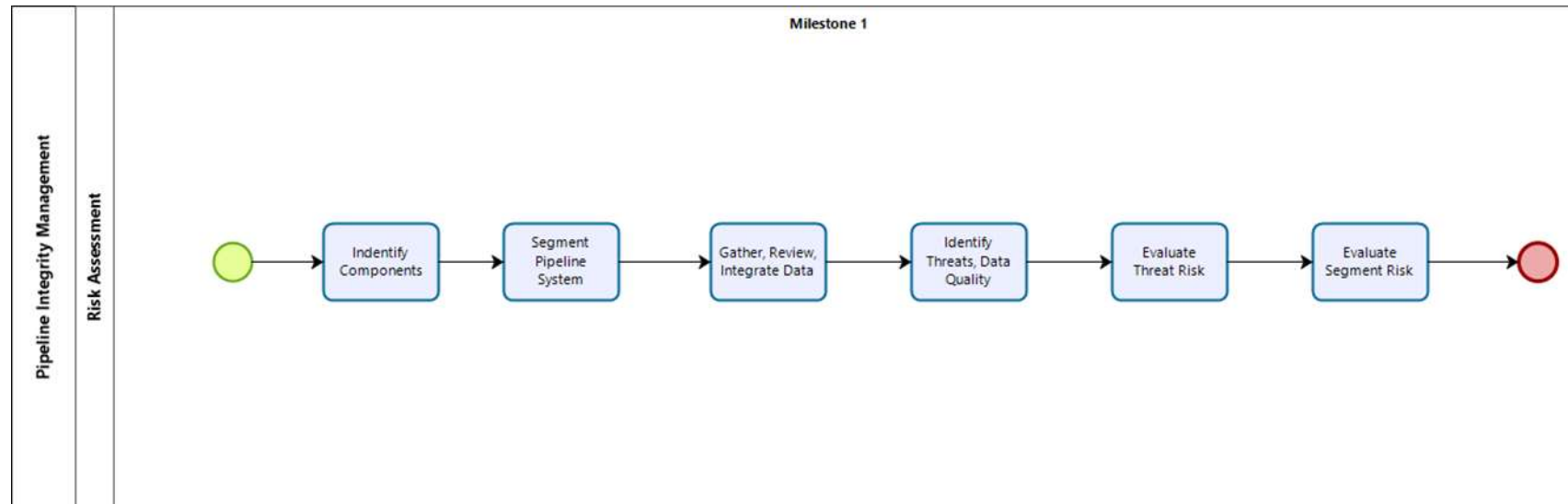
- > Address the organization, processes, and the physical system
- > Select program/data quality measures and periodically evaluate them
- > Quantify the causal relationship of root causes of each threat
- > Incorporate the historical data, subject matter expert opinion, and belief about collected data
- > Consider the interactive nature of threats



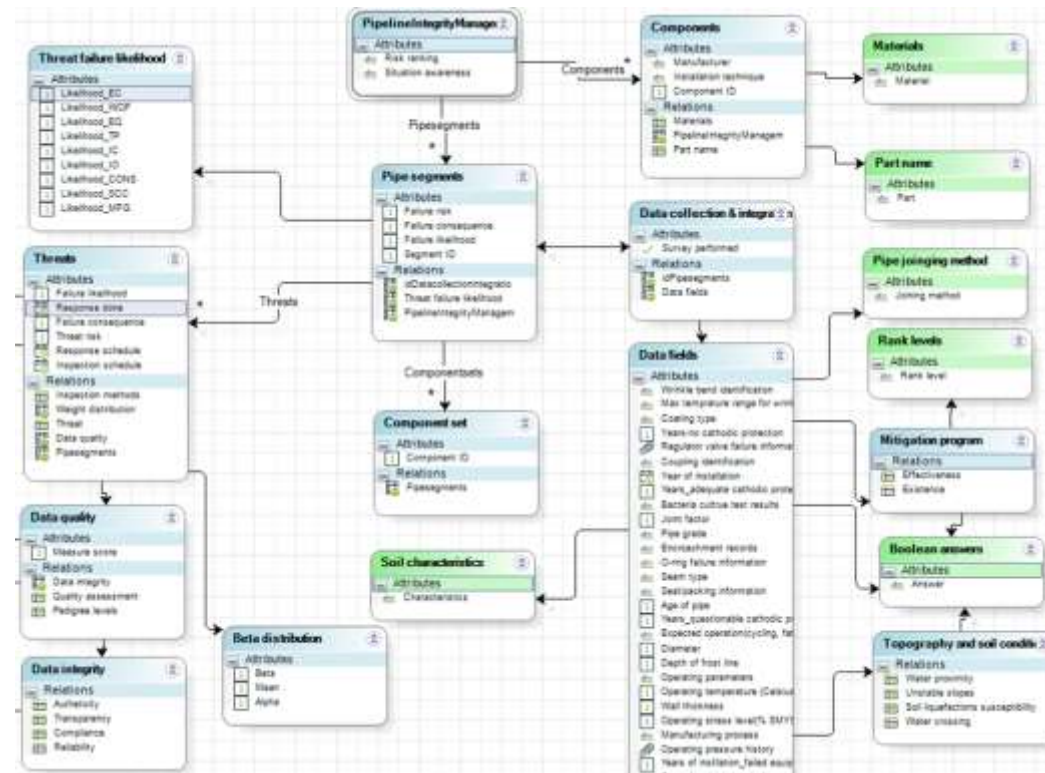
Risk Assessment of a Pipeline System



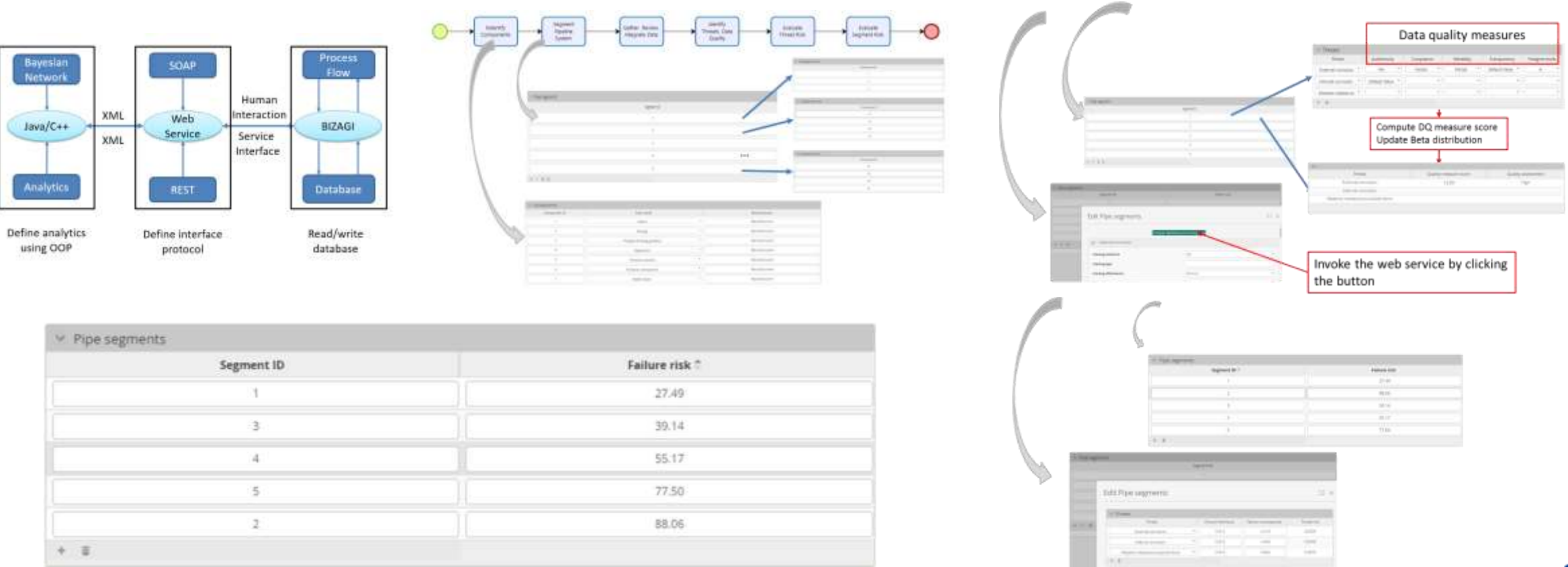
Simplified Process Workflow



Data Entity Relationships



Process is Readily Automated



Node by Node Data Quality Consideration

- > Conceptually, greater data quality measure value means higher confidence about the data, which is equivalent to smaller σ value in likelihood function
- > Assume the data quality of each threat is determined by a Beta distribution $Beta(\alpha_i, \beta_i)$

$$DQ_i = mean = \frac{\alpha_i}{\alpha_i + \beta_i}$$

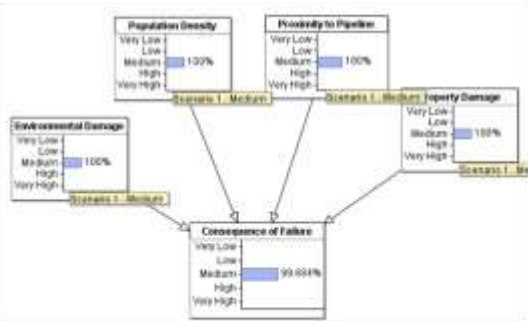
- > α_i and β_i will be updated based on the value of data quality measures such as authenticity, compliance, reliability, transparency, and pedigree

	Authenticity		Compliance		Reliability		Transparency		Pedigree	
Measure Value	High		Low		High		Low		Low	
Prior	$\alpha=1$	$\beta=1$	2	1	2	2	3	2	3	3
Updated	2	1	2	2	3	2	3	3	3	4
Final	Beta($\alpha=3, \beta=4$)—DQ=mean=3/7									

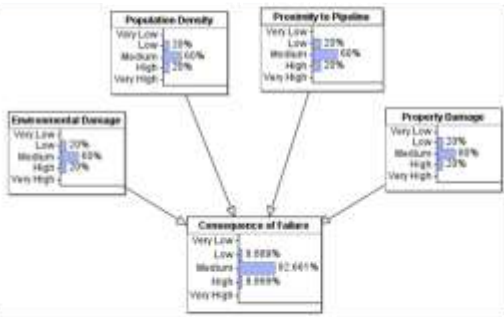
- > For continuous nodes, introduce a measurement error $\sigma(1-\text{"DQ"})$
- > For ranked nodes, adjust probability histogram to increase/decrease uncertainty

Data Quality Effect – Can be Extended to Consequence Measures for Benefit/Cost Analysis of Improved Data Quality

Data Quality	Environment Damage			Population Density			Proximity to Pipeline			Property Damage		
	Medium			Medium			Medium			Medium		
High DQ Score	Medium (100%)			Medium (100%)			Medium (100%)			Medium (100%)		
Low DQ Score	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
	20%	60%	20%	20%	60%	20%	20%	60%	20%	20%	60%	20%



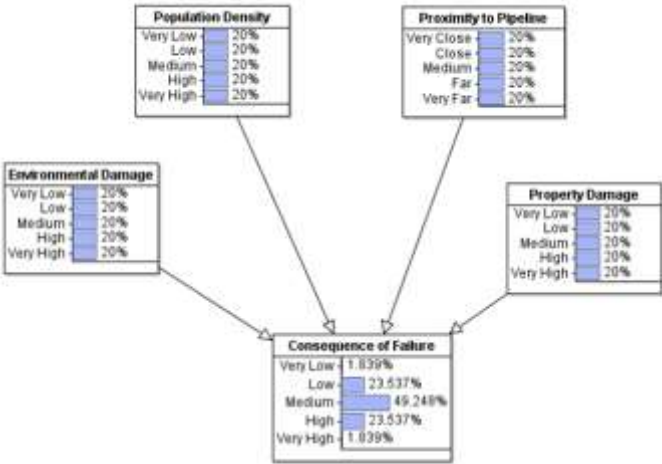
Lower Uncertainty



Higher Uncertainty



Likelihood of Failure



Consequence of Failure

Direct Tie in to Six Sigma Approaches

- > To investigate the effect of likelihood function, it is assumed that the pressure of a pipeline system follow a normal distribution $\theta \sim N(\mu_0, \sigma_0^2)$
- > Assume the value from the measurement system x'
- > $x' = \theta + \varepsilon$
- > where the measurement error term $\varepsilon \sim N(0, \sigma^2)$
- > Therefore, the posterior distribution of the pressure given the measurement x' is given as

$$p(\theta|x) \propto \exp\left(-\frac{(\theta - \mu_0)^2}{2 \times \sigma_0^2}\right) \exp\left(-\frac{(x' - \theta)^2}{2 \times \sigma^2}\right) \propto \exp\left(-\frac{(\theta - \frac{\sigma^2 \mu_0 + \sigma_0^2 x'}{\sigma_0^2 + \sigma^2})^2}{2 \times \frac{\sigma^2 \sigma_0^2}{\sigma_0^2 + \sigma^2}}\right)$$

- > The uncertainty of the posterior distribution depends on

$$\frac{\sigma^2 \sigma_0^2}{\sigma_0^2 + \sigma^2}$$

- > It is trivial to show that values of σ produce higher uncertainty in the posterior distribution θ

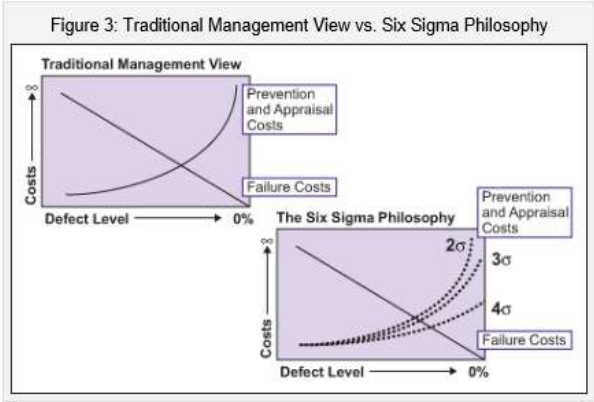


Table 1: Sigma Level and the Cost of Quality		
Sigma Level	DPMO	Cost of Quality as Percentage of Sales
2	298,000	More than 40%
3	67,000	25-40%
4	6,000	15-25%
5	233	5-15%
6	3.4	Less than 1%

Subject Matter Expertise

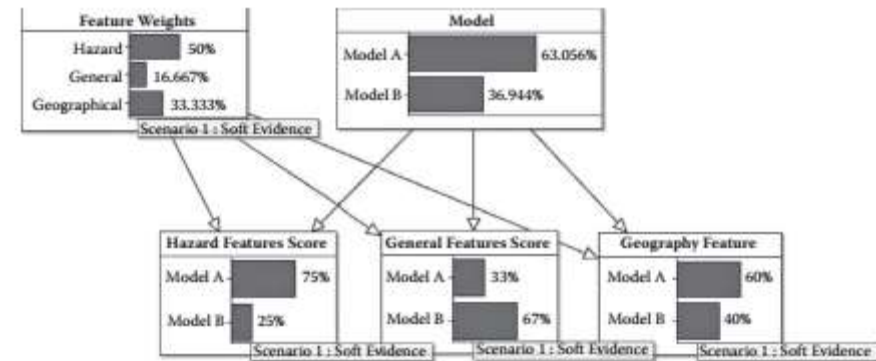
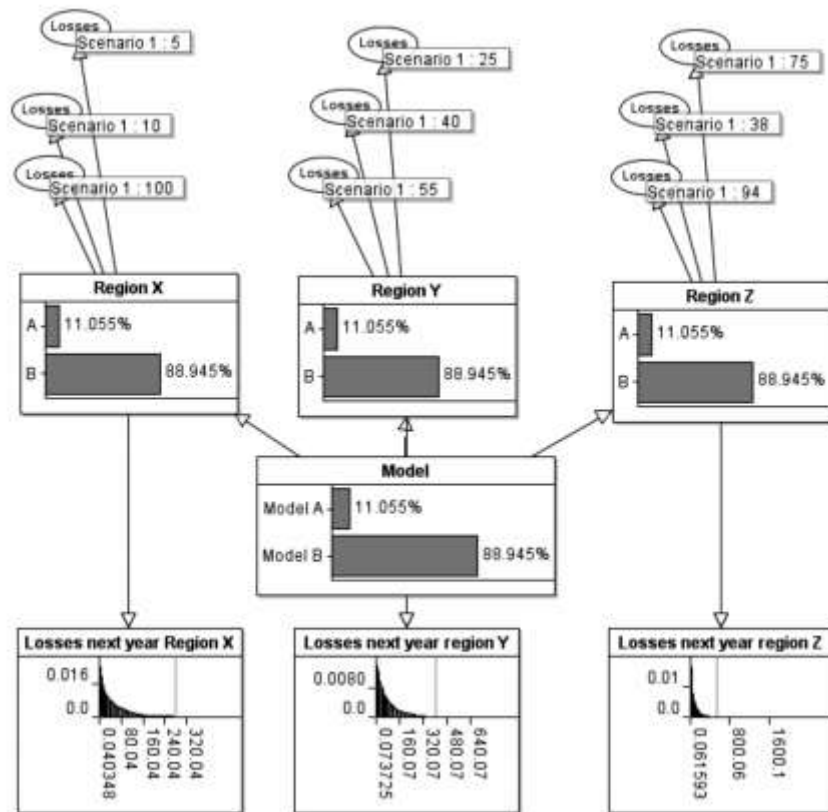
Biases and Heuristics (Kahneman and Tversky 1982)

■ **Ambiguity effect**—The avoidance of options for which missing information makes the probability seem unknown. ■ **Attentional bias**—Neglect of relevant data when making judgments of a correlation or association. ■ **Availability heuristic**—Estimating what is more likely by what is more available in memory, which is biased toward vivid, unusual, or emotionally charged examples. ■ **Base rate neglect**—Failing to take account of the prior probability. This was at the heart of the common fallacious reasoning in the Harvard medical study described in Chapter 1. It is the most common reason for people to feel that the results of Bayesian inference are nonintuitive. ■ **Bandwagon effect**—Believing things because many other people do (or believe) the same. Related to groupthink and herd behavior. ■ **Confirmation bias**—Searching for or interpreting information in a way that confirms one's preconceptions. ■ **Déformation professionnelle**—Ignoring any broader point of view and seeing the situation through the lens of one's own professional norms.

■ **Expectation bias**—The tendency for people to believe, certify, and publish data that agrees with their expectations. This is subtly different to confirmation bias, because it affects the way people behave before they conduct a study. ■ **Framing**—By using a too narrow approach or description of the situation or issue. Also framing effect, which is drawing different conclusions based on how data are presented. ■ **Need for closure**—The need to reach a verdict in important matters; to have an answer and to escape the feeling of doubt and uncertainty. The personal context (time or social pressure) might increase this bias. ■ **Outcome bias**—Judging a decision by its eventual outcome instead of based on the quality of the decision at the time it was made. ■ **Overconfidence effect**—Excessive confidence in one's own answers to questions. For example, for certain types of question, answers that people rate as 99% certain turn out to be wrong 40% of the time. ■ **Status quo bias**—The tendency for people to like things to stay relatively the same.

Fenton, Norman. Risk Assessment and Decision Analysis with Bayesian Networks (Page 262). Taylor and Francis CRC ebook account. Kindle Edition

Comparing Expert Models



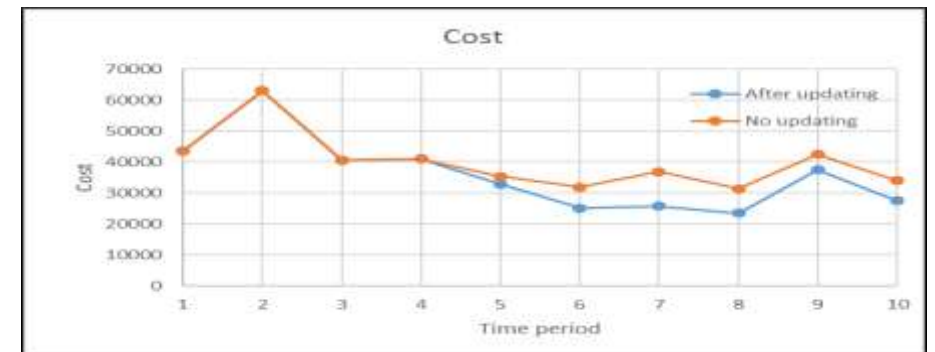
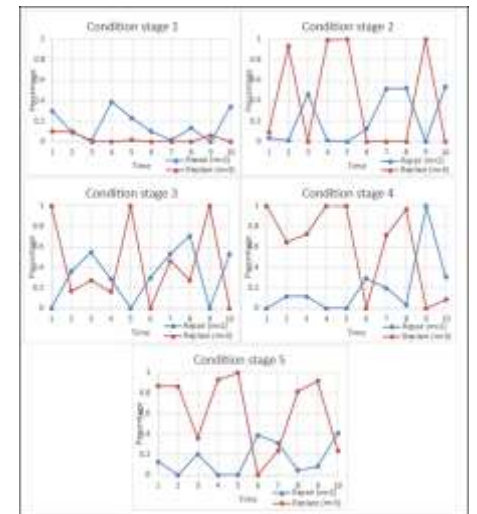
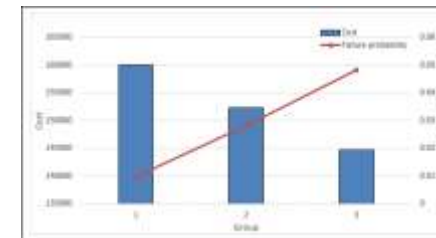
- > It is possible to construct hybrid Bayesian networks that compare the performance of competing expert models given different data

Fenton, Norman. Risk Assessment and Decision Analysis with Bayesian Networks (Page 320 - 324). Taylor and Francis CRC ebook account. Kindle Edition

Optimization

How to Optimize Risk Reduction Activities

- > This is a vast topic with many well defined solution approaches that we cannot even begin to address adequately in a single session
- > Monte Carlo simulation can be used to good effect in evaluating a large number of alternative approaches
- > It is important to evaluate many diverse approaches and identify which approach is best in particular situations (SITUATIONAL AWARENESS)
- > Constant Bayesian updating of the underlying probability distributions and re-running simulations is helpful
- > Causal networks will help define the optimization problem better
- > Approaches that reduce system variance will have great effect
- > Prompt mitigation and repair of identified system anomalies will help reduce the likelihood of higher order system interactions that can be problematic



Applying the Causal Framework to Armageddon

Fenton, Norman. Risk Assessment and Decision Analysis with Bayesian Networks (Page 46). Taylor and Francis CRC ebook account. Kindle Edition.

■ Risk measurement is more meaningful in the context; the BN tells a story that makes sense. This is in stark contrast with the simple “risk equals probability times impact” approach where not one of the concepts has a clear unambiguous interpretation. ■ Uncertainty is quantified and at any stage we can simply read off the current probability values associated with any event. ■ It provides a visual and formal mechanism for recording and testing subjective probabilities. This is especially important for a risky event that you do not have much or any relevant data about (in the Armageddon example this was, after all, mankind’s first mission to land on a meteorite).

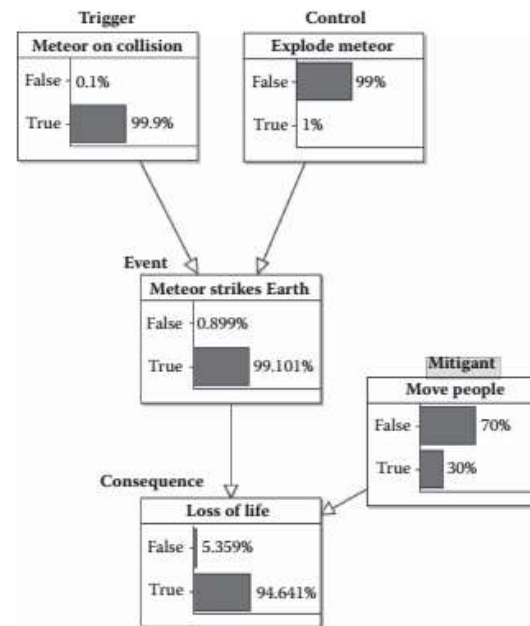


Figure 2.18 Initial risk of meteor strike.

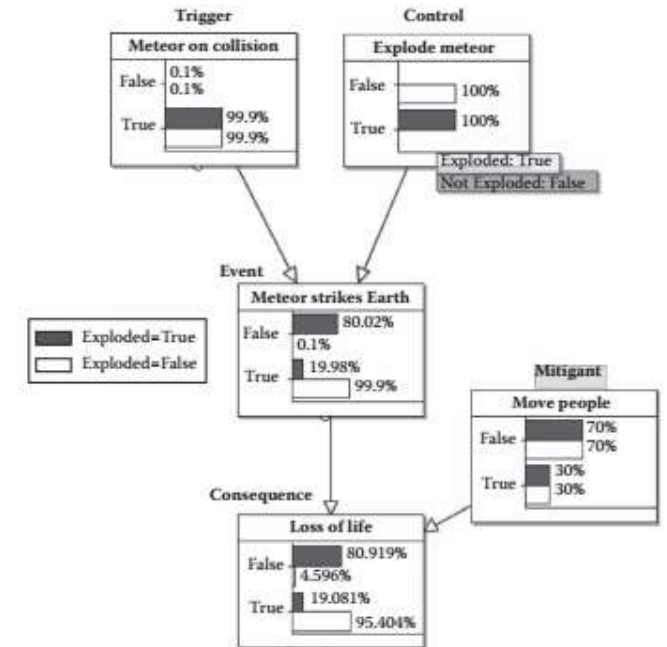


Figure 2.19 The potential difference made by Bruce Willis and crew.

KUUUB Factors and Other operational Risks

Fenton, Norman. Risk Assessment and Decision Analysis with Bayesian Networks (Chapter 11). Taylor and Francis CRC ebook account. Kindle Edition.

- > Known Unknowns
- > Unknown Unknowns
- > Bias

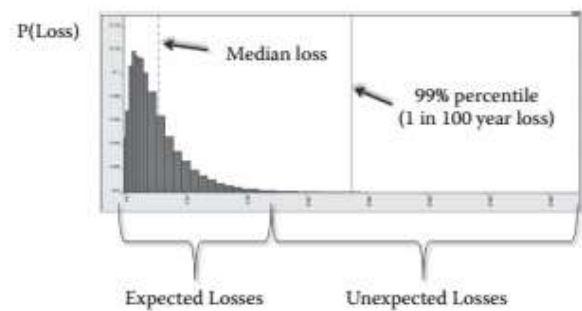


Figure 11.19 Loss distribution with 99% VaR and median losses.

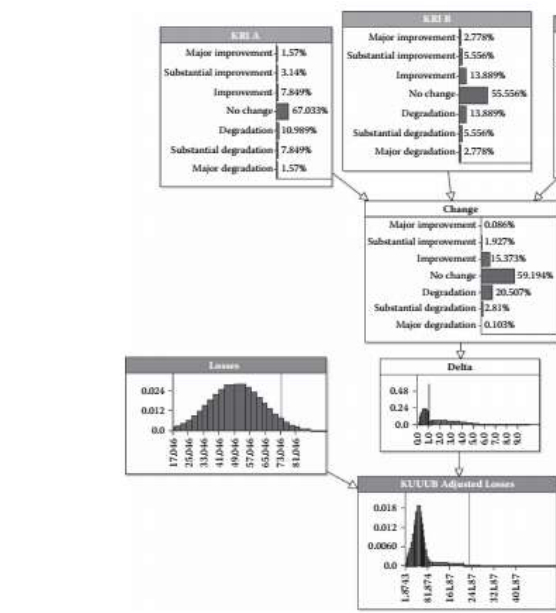


Figure 11.15 KUUUB model example, adjusting a financial loss estimate.

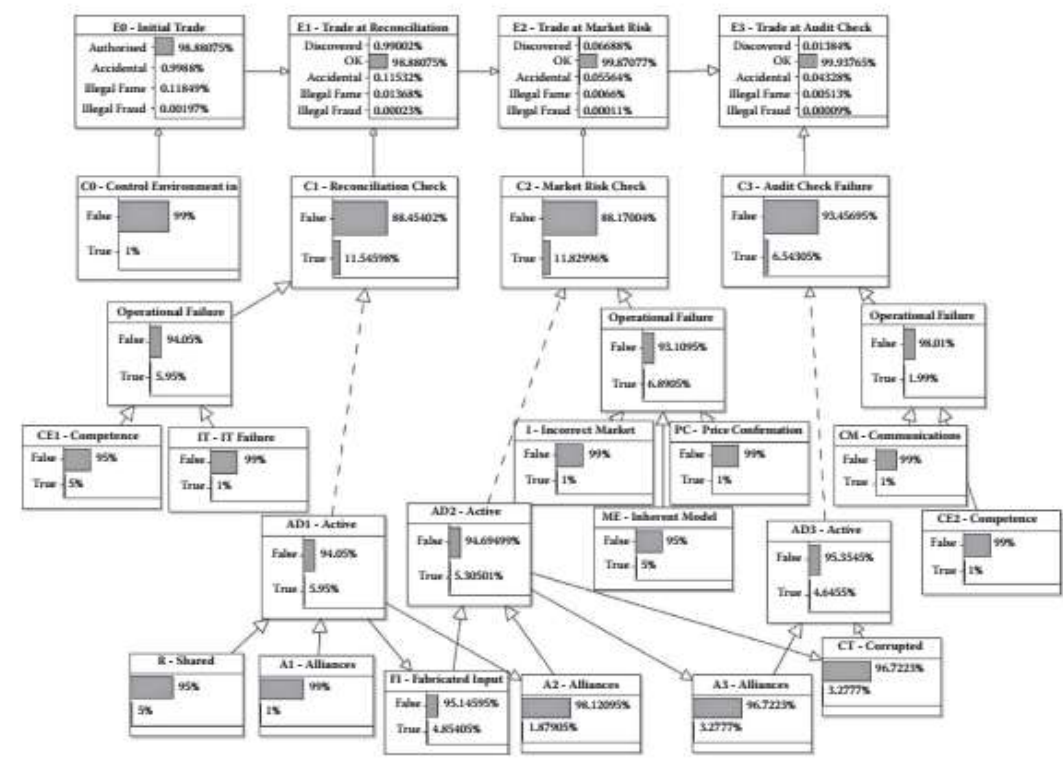


Figure 11.18 BN loss event model for rogue trading process with superimposed marginal probability distributions.

Discussion and Questions

Additional Slides On Integrating Data Quality into Risk Assessment

Stochastic Risk Analysis of Gas Pipeline with Bayesian Statistics

Deterministic Risk Analysis

- > Most common quantitative risk assessment method
- > Estimates single-point value for discrete scenarios such as worst case, best case, and most likely outcomes
- > Considers only a few outcomes, ignoring all other possibilities
- > Disregards interdependence between inputs
- > Ignores uncertainty in input variables

Stochastic Risk Analysis with Bayesian Statistics

- > Advanced quantitative risk assessment method
- > Estimates probability distribution of the risk
- > By using probability distributions, variables can have different probabilities of different outcomes
- > Uncertainties in variables are described by probability distributions
- > Probabilistic results show not only what could happen, but how likely each outcome is
- > Allows scenario analysis and sensitivity analysis
- > Possible to model interdependent relationships between input variables

Bayesian Statistics

- > Quantitative tool to rationally update subjective prior beliefs in light of new evidence.

$$\textit{Posterior} \propto \textit{Prior} \times \textit{Likelihood}$$

$$P(\theta|D) = P(D|\theta) P(\theta)/P(D)$$

θ is the parameter

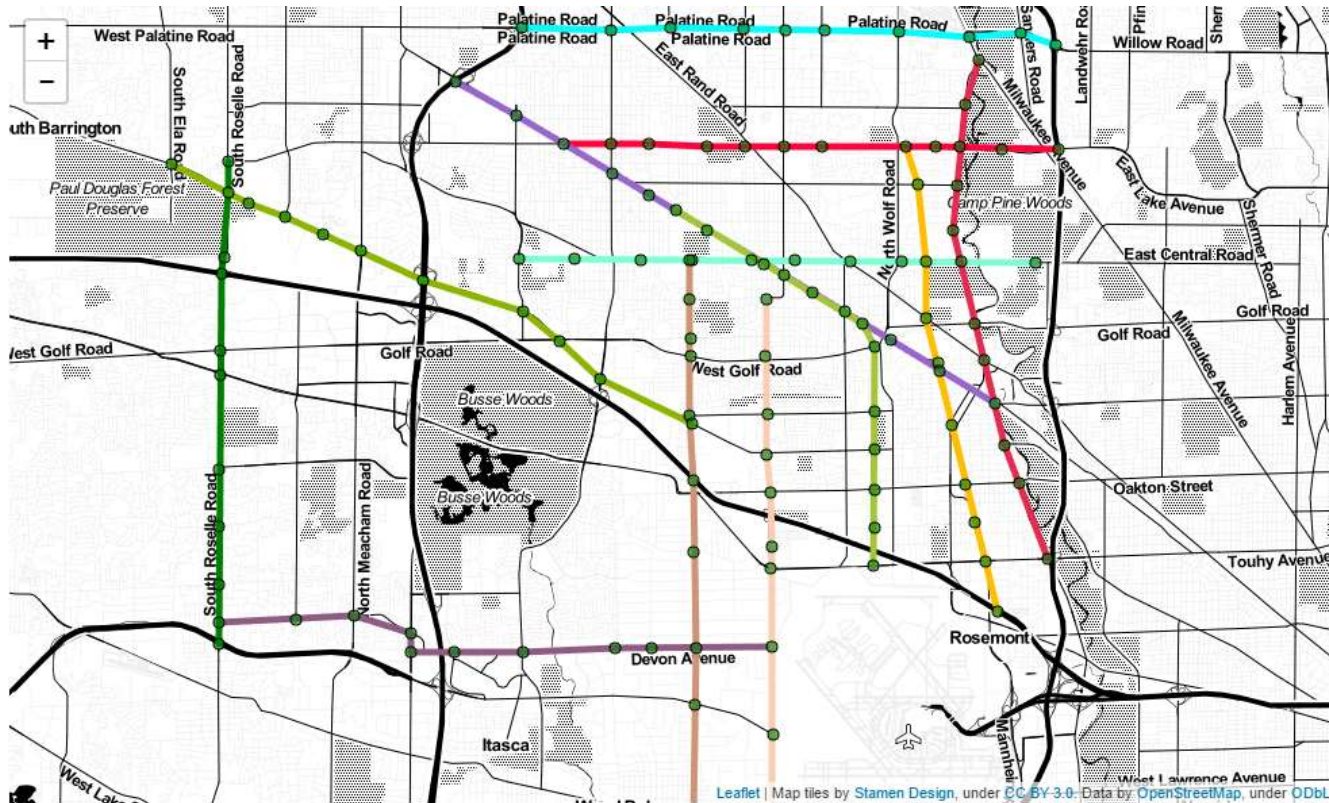
$P(\theta)$ is the prior

$P(\theta|D)$ is the posterior

$P(D|\theta)$ is the likelihood

$P(D)$ is the evidence

Illustrative Pipeline Network for Example in Following Slides



- ☐ Synthetic gas pipeline network of 120 components (e.g., joints, piping, etc) divided into:
 - 3 regions
 - 4 segments per region
 - 10 components per segment

- ☐ Risk analysis can be performed at component level, segment level, or region level

Component Failure Threat Types from ASME B31.8S Standard

> Time-Dependent Threats

- Internal Corrosion
- External Corrosion
- Stress Corrosion Cracking

> Resident Threats

- Manufacturing Defects
- Construction or Fabrication Defects
- Equipment Failure

• Time-Independent Threats

- Incorrect Operations Procedure
- Weather and Outside Force
- Third Party Damage

Probability of Component Failure

> **Probability of Failure over time** = $\int_{t=t_1}^{t_p} \int_{j=1}^n \int_{k=1}^m w_{j,k} f(t_i, x_j, y_k) dy dx dt$

- x , y , and t represents threat type, input variable, and time instance respectively.
- $[t_1, t_p]$ is the time interval for which likelihood of failure is calculated
- n is the total number of threat type (=9 for gas pipeline per AMSE B31.8S)
- m is the total number of input variables responsible for a given threat type
- where $w_{j,k}$ is the weight applied for each input variable from threat model

> $f(t, x, y)$ is calculated from the Beta distribution of component failure attribute as shown in the next few slides.

Probability of Component Failure

1. **Determine prior belief** $P(\theta)$ on parameters $\theta \in [0,1]$ where θ is always either success (1) or failure (0)

- **Beta distribution** quantifies the prior beliefs for binomial outcome.
- The probability density function of the beta distribution is

$$P(\theta|\alpha, \beta) = \theta^{\alpha-1}(1 - \theta)^{\beta-1}/B(\alpha, \beta)$$

where $B(\alpha, \beta)$ acts a normalizing constant so that the area under PDF sums to one.

- Initially, ignorant prior of $B(\alpha = 1, \beta = 1)$ is used in the very first run.

Probability of Component Failure

2. Determine likelihood function.

- **Bernoulli distribution** is well-suited for the Boolean-valued outcome usually labelled as ‘success’ (1) and ‘failure’ (0), in which it takes the value 1 with probability **p** and the value 0 with probability **1-p**.
- The probability mass function of the distribution is

$$f(k, p) = p^k (1 - p)^{1-k}$$

where $k \in \{1, 0\}$.

Probability of Component Failure

3. Determine posterior probability function of $\theta \in [0,1]$.

> $Posterior \propto Prior \times Likelihood$

> $P(\theta|D) = P(D|\theta) \frac{P(\theta)}{P(D)}$ where, $P(D) = \int_{\theta} P(D, \theta) d\theta$

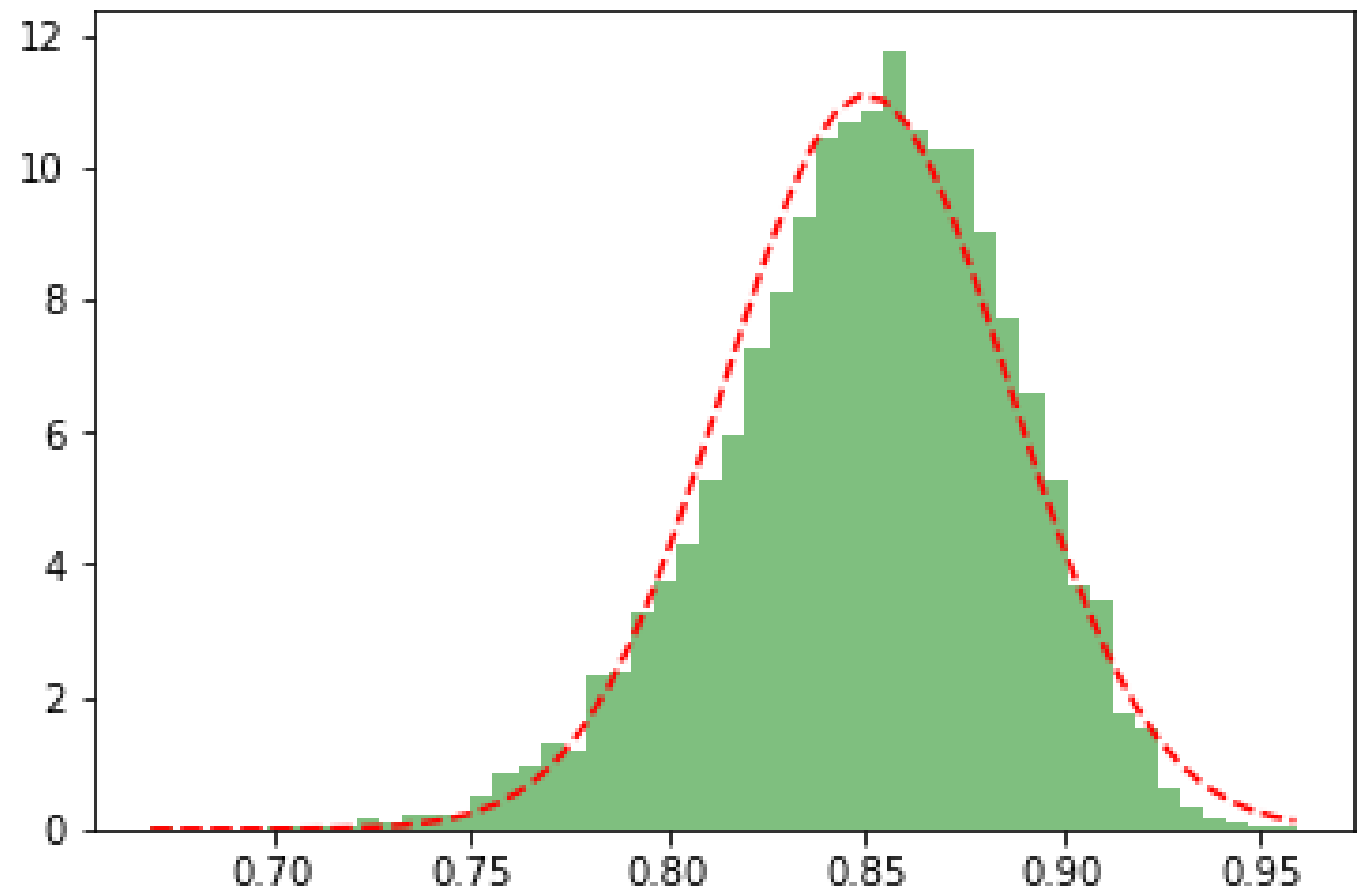
- Bernoulli likelihood and Beta prior are conjugate pairs – as a result, the **posterior** is a **Beta distribution**. The conjugate priors simplifies the calculation of posterior distribution.
- The computation of posterior distribution is a complex process (example, integration for $P(D)$) and therefore a numerical approximation method instead such as **Markov Chain Monte Carlo** (MCMC) is needed.
- $Posterior\ Beta\ distribution = MCMC (Beta\ prior, Bernoulli\ likelihood)$

Probability of Component Failure using Markov Chain Monte Carlo

Metropolis-Hastings algorithm

1. Begin the algorithm at the *current* position in parameter space (θ_{current})
2. Propose a "jump" to a new position in parameter space (θ_{new})
3. Accept or reject the jump probabilistically using the prior information and available data
4. If the jump is accepted, move to the new position and return to step 1
5. If the jump is rejected, stay at current position and return to step 1
6. After a set number of jumps have occurred, return all of the *accepted* positions

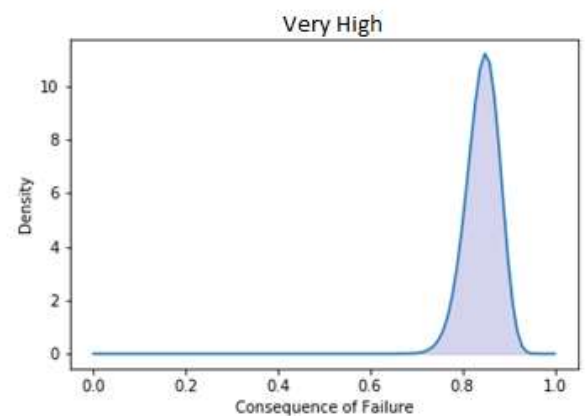
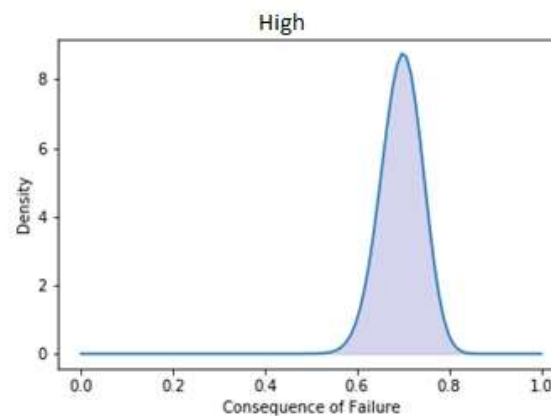
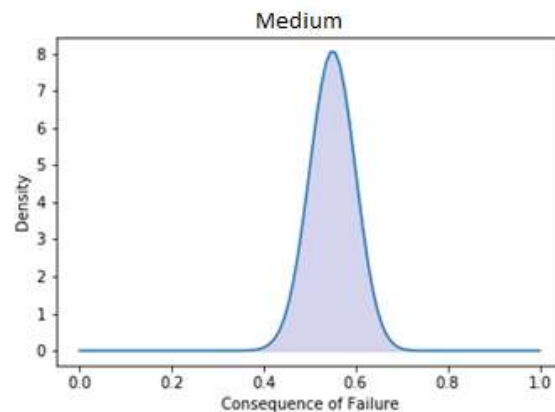
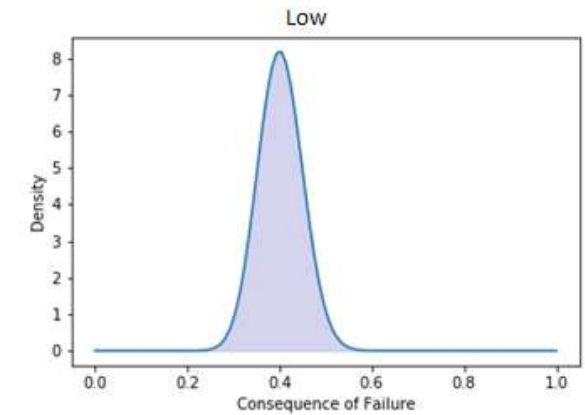
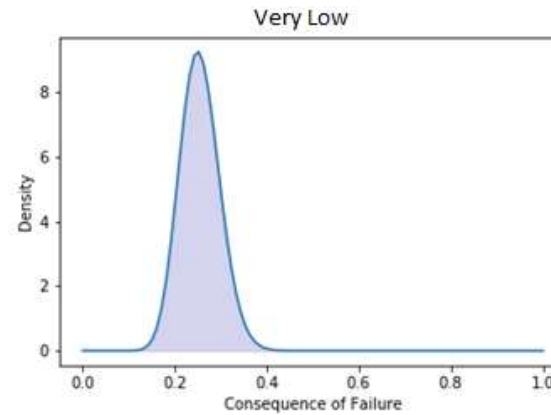
Probability Distribution of Component Failure



Consequence of Failures

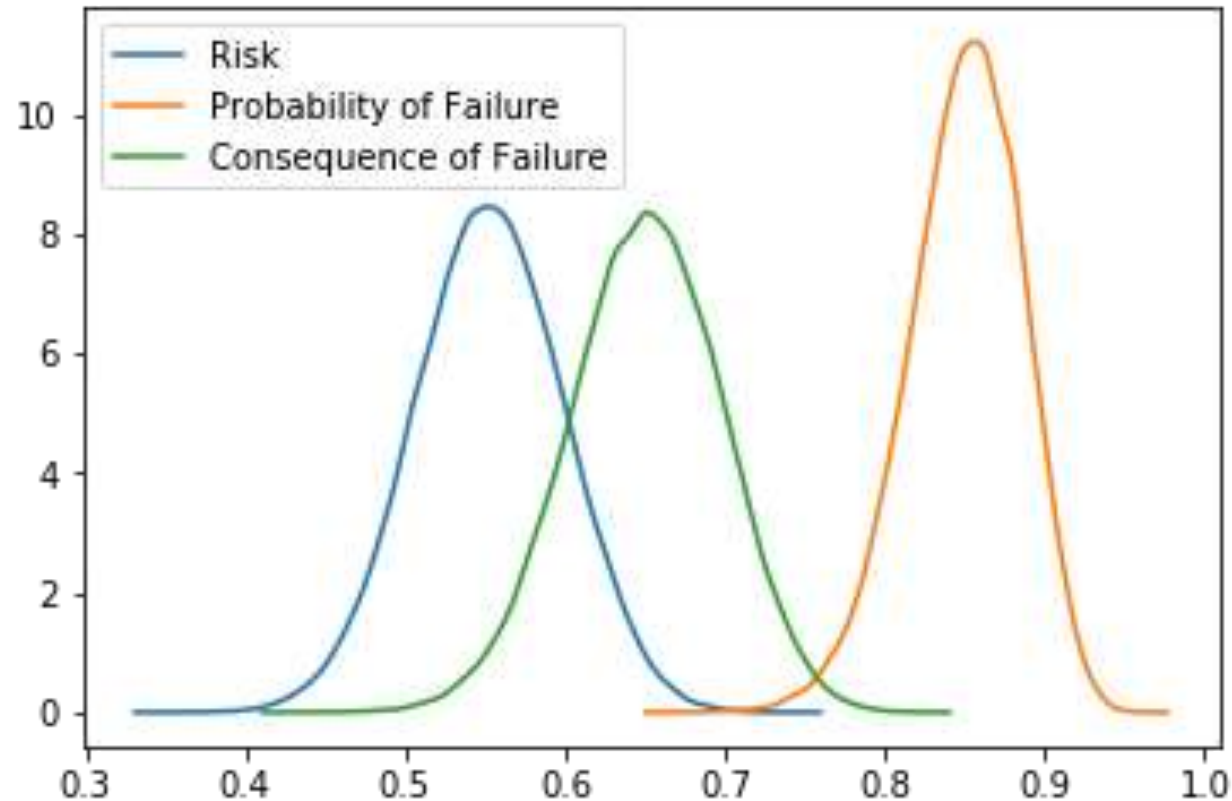
Five Categories of Consequences

- Very Low
- Low
- Medium
- High
- Very High



Stochastic Risk Analysis

$$\text{Risk} = \text{Probability of Failure} \times \text{Consequence of Failure}$$



Implication of Data Quality in Risk Analysis

> Integrity

- Authenticity
- Compliance
- Transparency
- Reliability

> Pedigree

- **Beta Distribution** $B(\alpha, \beta)$

- Start with $\alpha = 1, \beta = 1$
- $\alpha = \alpha + 1$ (if Authenticity = *TRUE*) +
1 (if Compliance = *TRUE*) +
1 (if Transparency = *TRUE*) +
1 (if Reliability = *TRUE*) + 1 (if Pedigree = *TRUE*)
- $\beta = \beta + 1$ (if Authenticity = *FALSE*) +
1 (if Compliance = *FALSE*) +
1 (if Transparency = *FALSE*) +
1 (if Reliability = *FALSE*) +
1 (if Pedigree = *FALSE*)