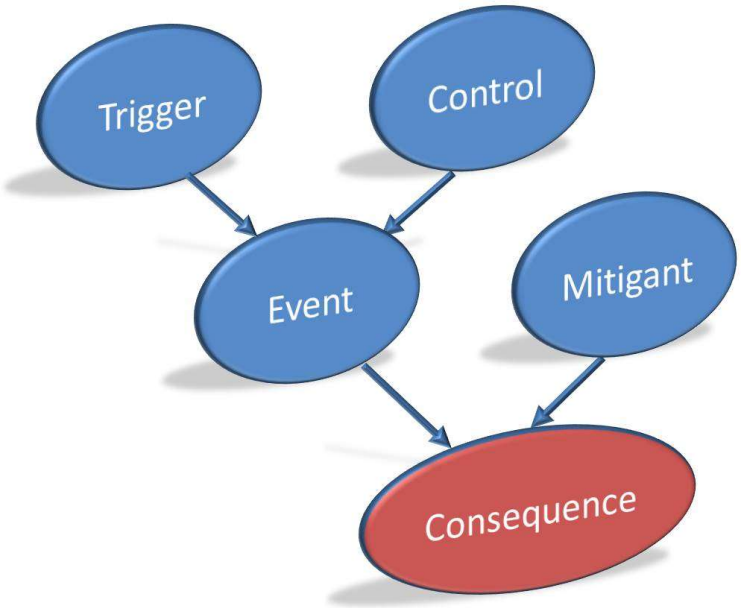


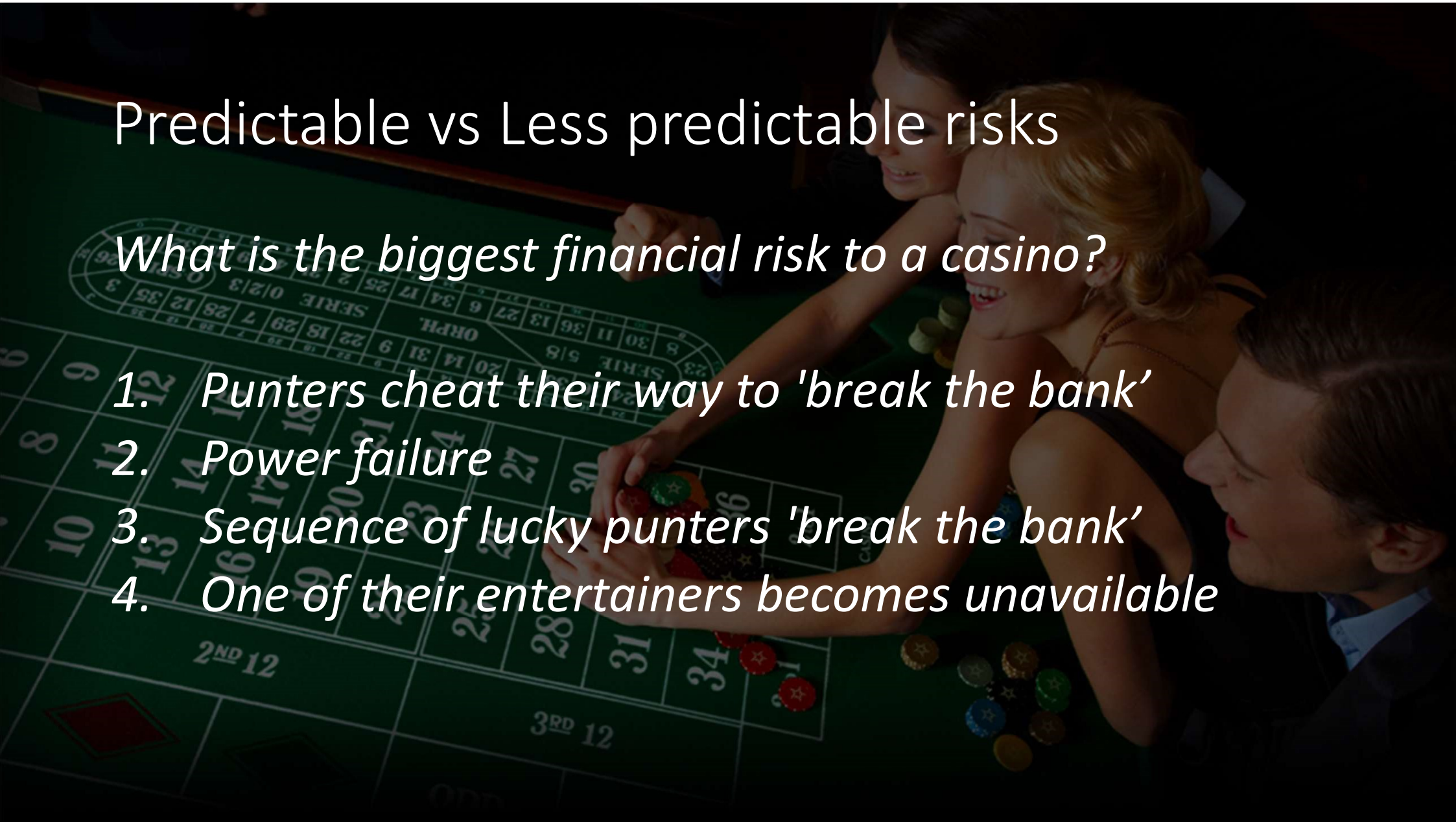
Risk and Decision Making for Data Science and AI

Lesson 4a Modelling rare events - and confidence intervals

Norman Fenton

@ProfNFenton

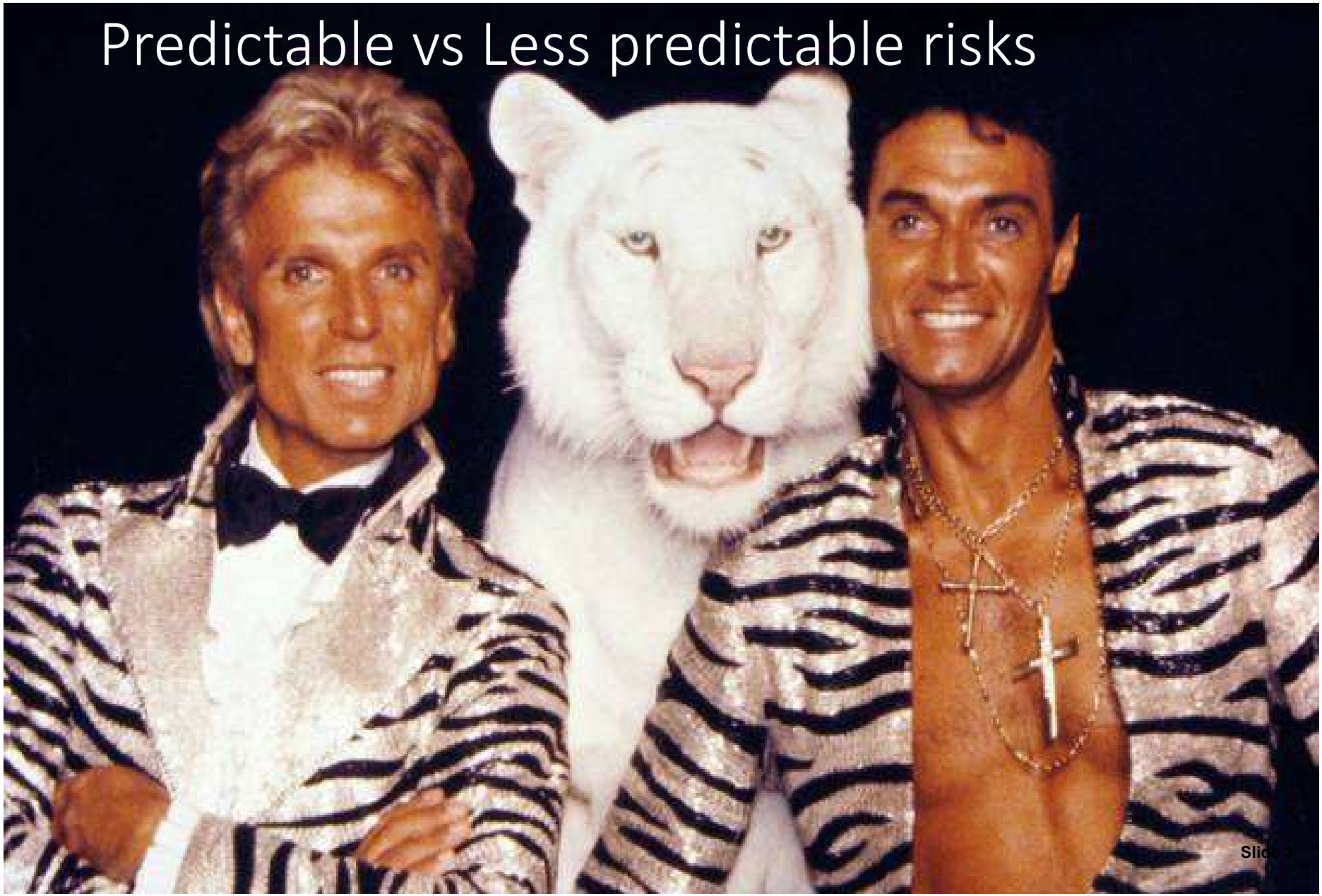
Predictable vs Less predictable risks



What is the biggest financial risk to a casino?

1. *Punters cheat their way to 'break the bank'*
2. *Power failure*
3. *Sequence of lucky punters 'break the bank'*
4. *One of their entertainers becomes unavailable*

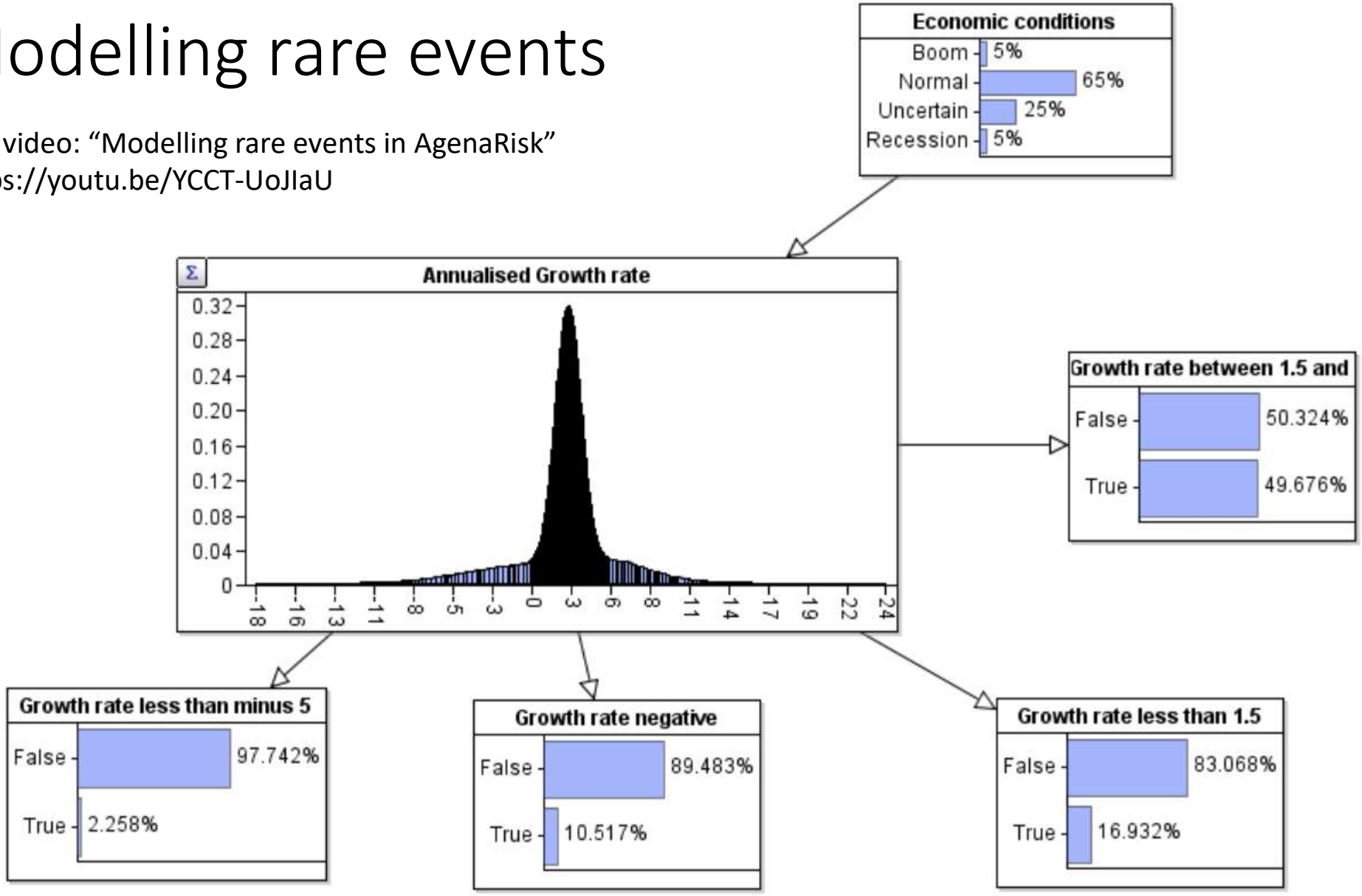
Predictable vs Less predictable risks



Predictable vs Less predictable risks

Modelling rare events

See video: “Modelling rare events in AgenaRisk”
https://youtu.be/YCCT-UoJIaU



Adding confidence intervals to risk predictions

"Support for candidate Joe Bloggs now stands at 43%. The margin of error is plus or minus three percent."

Which of the following do you think ***most closely*** represents your understanding of the above statement:

A. It is certain that 43% support Bloggs
B. There is a high probability that 43% support Bloggs
C. It is certain support for Bloggs is between 40-46% with the most likely being 43%
D. It is certain that support for Bloggs is between 40-46%
E. There is a high probability that support for Bloggs is between 40-46% with the most likely being 43%
F. There is a high probability that support for Bloggs is between 40-46%

Most people assume the 'correct' answer is F.

Indeed a statement like

"There is a 95% probability that support for Bloggs is between 40-46%"

is a meaningful statement expressing our uncertainty about the unknown proportion of voters who support Bloggs

But NONE of the above are even close to the 'true meaning' because such statements are invariably based on 'classical' frequentist statistical analysis

“Support for candidate Joe Bloggs now stands at 43%.
The margin of error is plus or minus three percent,
with 95% confidence”

Surely this means

“There is a 95% probability that support for Bloggs is between 40-46%”

… no, again, it does not

Adding confidence intervals to risk predictions

Whereas a statement about the probability of an unknown value is natural for Bayesians, it is simply not allowed (because it has no meaning) in the classical (frequentist) approach – which has to assume that data comes from 'repeated experiments'.

Hence the exact frequentist meaning of the statement
"Support for candidate Joe Bloggs now stands at 43%. The margin of error is plus or minus three percent."
is
If we could repeat the sampling many times and each time record the exact proportion of votes for Bloggs in the sample then every time the proportion would be in the range [40-46]
If an additional 'confidence value' was added to the statement, such as
"Support for candidate Joe Bloggs now stands at 43%. The margin of error is plus or minus three percent with 95% confidence"
Then the exact meaning is
If we could repeat the sampling many times and each time record the exact proportion of votes for Bloggs in the sample then for 95% of the times, the proportion would be in the range [40-46]
These statements do not capture any intuitive notion of uncertainty that lay people can understand.

See Section 10.5 of: Fenton, N.E. and M. Neil, Risk Assessment and Decision Analysis with Bayesian Networks. 2018,

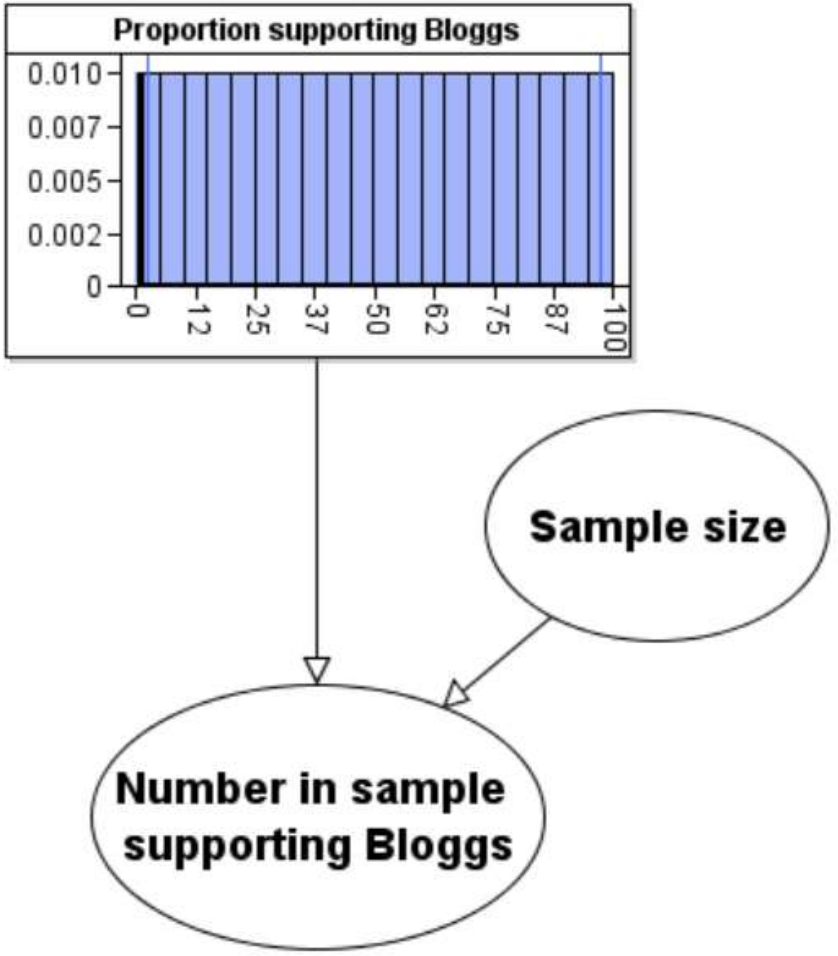
Adding confidence intervals to risk predictions

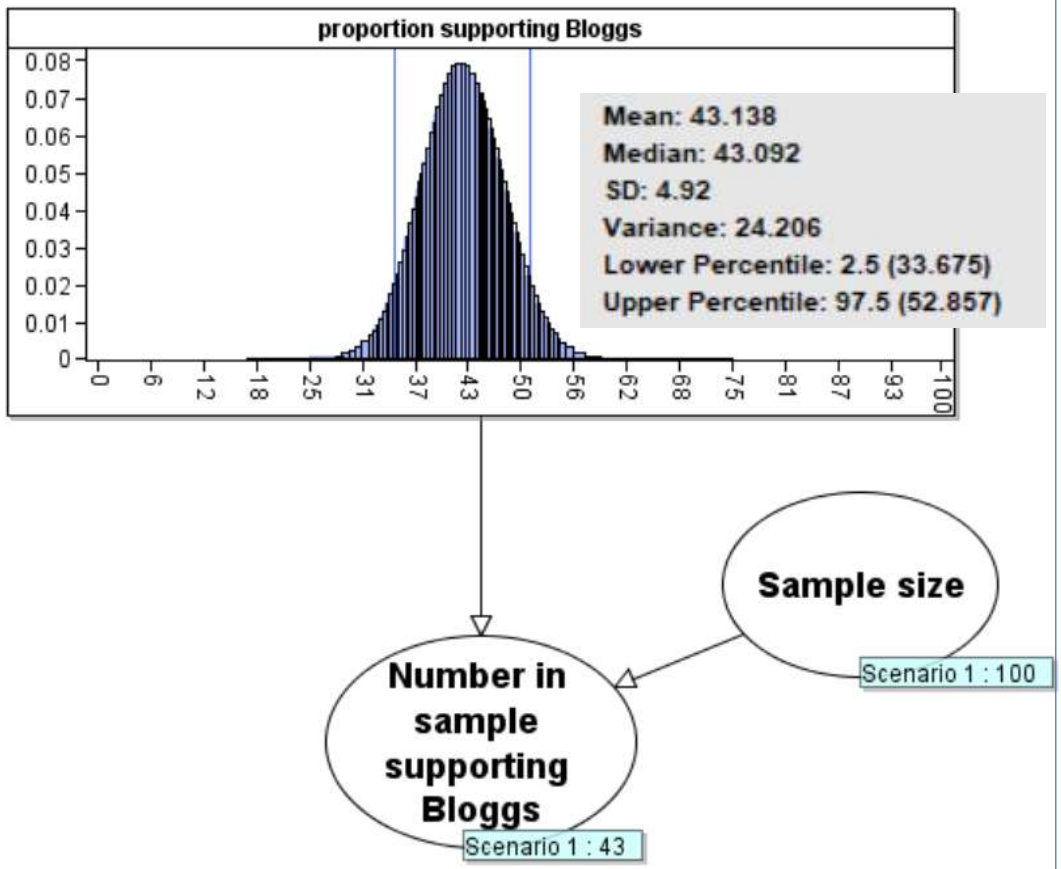
With Bayes we start with an assumption about the 'true' population.
For example, an 'ignorant' prior assumption like "Support for Bloggs could just as likely be anywhere between 0 and 100% (this is a so-called Uniform[0, 100] distribution prior)
We then use the observed data – such as 43 people in a sample of 100 saying they support Bloggs - to revise our belief about the 'true' population
Then we can make meaningful probability statements about the learnt population
Classical statisticians (frequentists) do not like the fact that Bayes must assume a prior probability for an unknown. Yet it turns out the classical analysis makes all kind of implicit prior assumptions.

See Section 10.5 of Fenton, N.E. and M. Neil, Risk Assessment and Decision Analysis with Bayesian Networks, Second Edition. 2018,

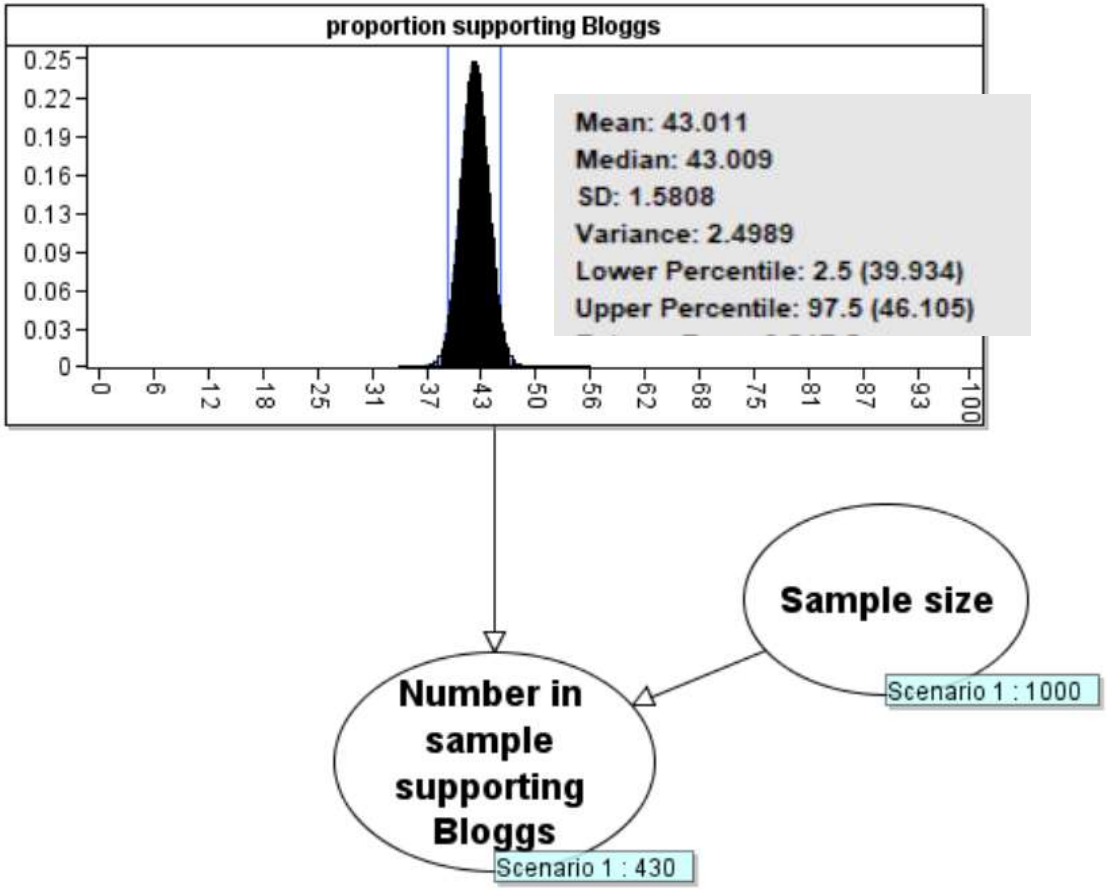
Adding confidence intervals to risk predictions

With Bayes we start with an assumption about the ‘true’ population.
For example, an ‘ignorant’ prior assumption like “Support for Bloggs could just as likely be anywhere between 0 and 100% (this is a so-called Uniform[0, 100] distribution prior).
We then use the observed data – such as 43 people in a sample of 100 saying they support Bloggs - to revise our belief about the ‘true’ population.
Then we can make meaningful probability statements about the learnt population
Classical statisticians (frequentists) do not like the fact that Bayes must assume a prior probability for an unknown. Yet it turns out the classical analysis makes all kind of implicit prior assumptions.





"There is a 95% probability that support for Bloggs is between 33.7 and 52.9%"



"There is a 95% probability that support for Bloggs is between 39.9 and 46.1%"

Adding confidence intervals to risk predictions

With Bayes we start with an assumption about the 'true' population.
For example, an 'ignorant' prior assumption like "Support for Bloggs could just as likely be anywhere between 0 and 100% (this is a so-called Uniform[0, 100] distribution prior)
We then use the observed data – such as 43 people in a sample of 100 saying they support Bloggs - to revise our belief about the 'true' population
Then we can make meaningful probability statements about the learnt population
Classical statisticians (frequentists) do not like the fact that Bayes must assume a prior probability for an unknown. Yet it turns out the classical analysis makes all kind of implicit prior assumptions.

See Section 10.5 of Fenton, N.E. and M. Neil, Risk Assessment and Decision Analysis with Bayesian Networks, Second Edition. 2018,

Key points

- Standard probability and statistics are excellent for modelling and predicting ‘predictable risks’
- But they rely on the frequentist notion of repeated observations
- These methods are hopelessly inadequate when it comes to predicting ‘rare’ risks
- Classical statistical models – using regression and short-tailed distributions cannot model or predict rare events
- Bayesian network causal models that incorporate knowledge with data can model and predict rare events
- The confidence intervals used in classical statistics do not mean what most people assume they mean. The Bayesian equivalents do.