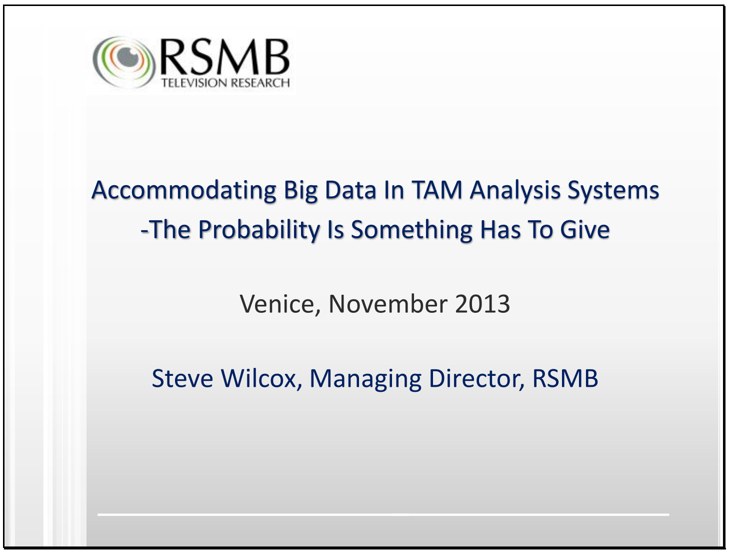
Slide 1



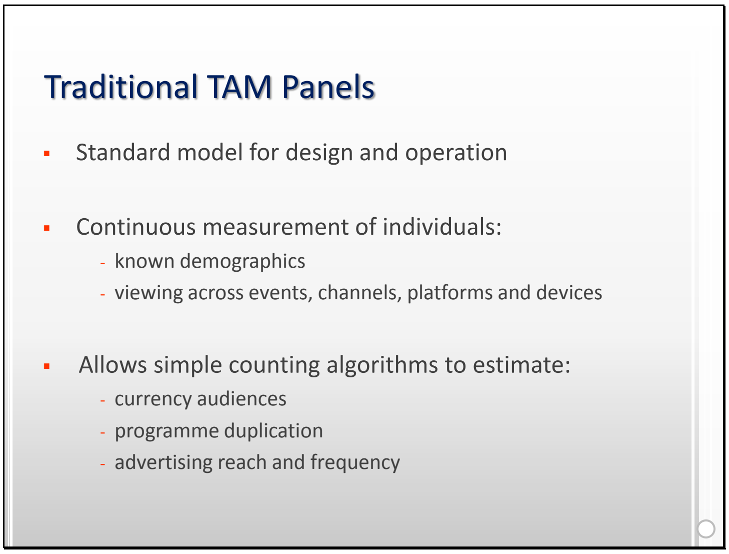
RSMB
TELEVISION RESEARCH

Accommodating Big Data In TAM Analysis Systems
-The Probability Is Something Has To Give

Venice, November 2013

Steve Wilcox, Managing Director, RSMB

Slide 2



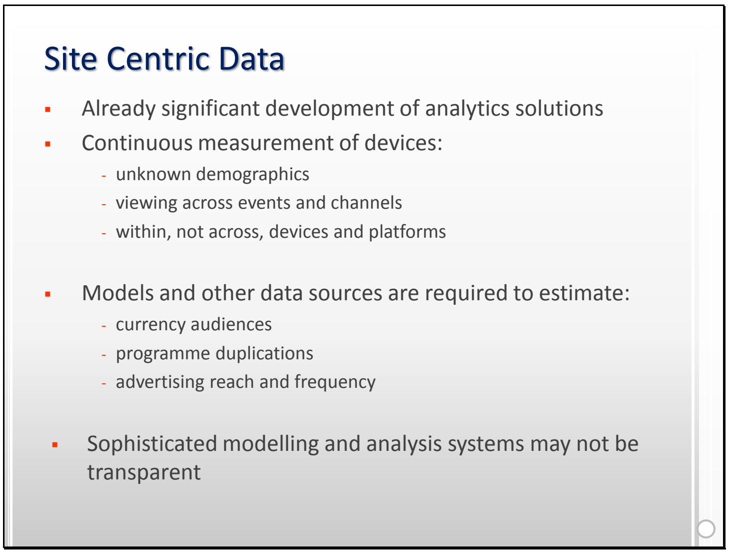
Traditional TAM Panels

- Standard model for design and operation
- Continuous measurement of individuals:
  - known demographics
  - viewing across events, channels, platforms and devices
- Allows simple counting algorithms to estimate:
  - currency audiences
  - programme duplication
  - advertising reach and frequency

Traditional TAM panels have been around for a long time and in most markets there is a fairly standard model for their design and operation.

The panel provides a continuous measurement of households and individuals. There are two key attributes: We know the demographics of the people and we have a single source measurement of viewing across events, channels, platforms and devices. Then, in theory at least, we have simple counting algorithms to estimate currency audiences, programme duplication and advertising reach and frequency.

Slide 3



Site Centric Data

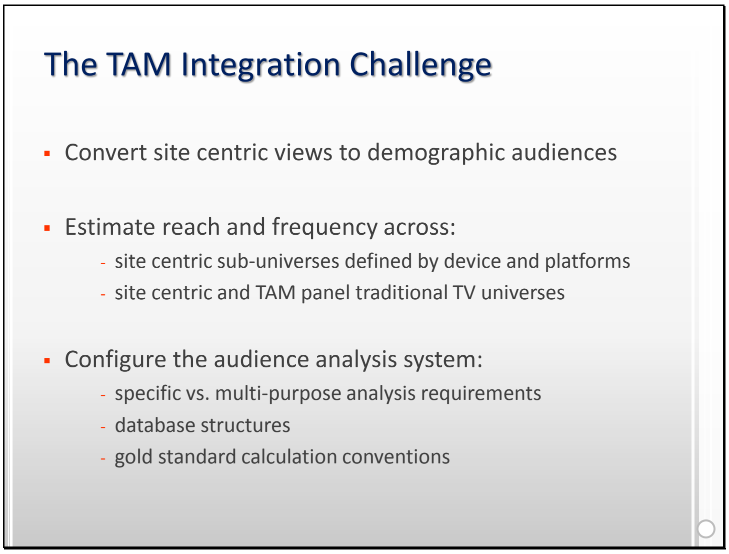
- Already significant development of analytics solutions
- Continuous measurement of devices:
  - unknown demographics
  - viewing across events and channels
  - within, not across, devices and platforms
- Models and other data sources are required to estimate:
  - currency audiences
  - programme duplications
  - advertising reach and frequency
- Sophisticated modelling and analysis systems may not be transparent

So what about big data? First I want to acknowledge that broadcasters have made significant developments in analytics solutions designed to maximize insight from their site centric data. But we all know its important to recognize what these data are.

They are based upon continuous measurement of devices, not people, so we don't know the demographics of the audiences. They do measure viewing across events and channels, however this measurement is only single source within devices and platforms – they don't allow you to track people across different site centric databases. Traditional TV viewing is outside of the loop.

As we've heard, models and other data sources are required to provide context in terms of currency audiences, programme duplications and advertising reach and frequency. The associated modeling and analysis solutions are necessarily complex. This could lead to a lack of transparency and I think this is contradictory to the objectives of an industry measurement system.

Slide 4



The TAM Integration Challenge

- Convert site centric views to demographic audiences
- Estimate reach and frequency across:
  - site centric sub-universes defined by device and platforms
  - site centric and TAM panel traditional TV universes
- Configure the audience analysis system:
  - specific vs. multi-purpose analysis requirements
  - database structures
  - gold standard calculation conventions

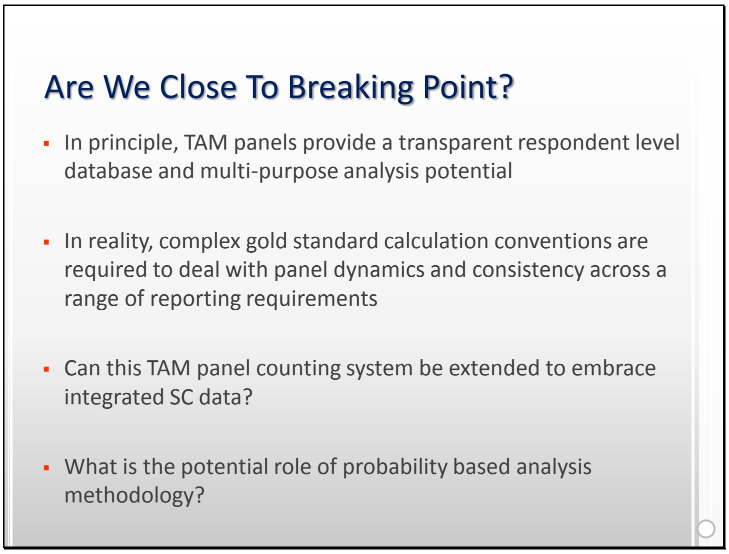
So this is the integration challenge.

First we have to convert site centric views to demographic audiences. In simple terms this could mean multiplying site centric views by TAM panel viewers per view factors.

The bigger challenge is to estimate reach and frequency across all the sub-universes measured by the different site centric databases and also across traditional TV.

And amidst all this theory, we must not forget that at some point we have to configure the audience analysis systems. Pity the poor IT man who has to deal with a typical statistician's specification, completely divorced from reality. We need to consider if we can live with analysis specific solutions, which are often easier to achieve, or if we need the transparent, multi-purpose solutions that we have at present. The all important product is the output database structures and the associated gold standard calculation conventions.

Slide 5



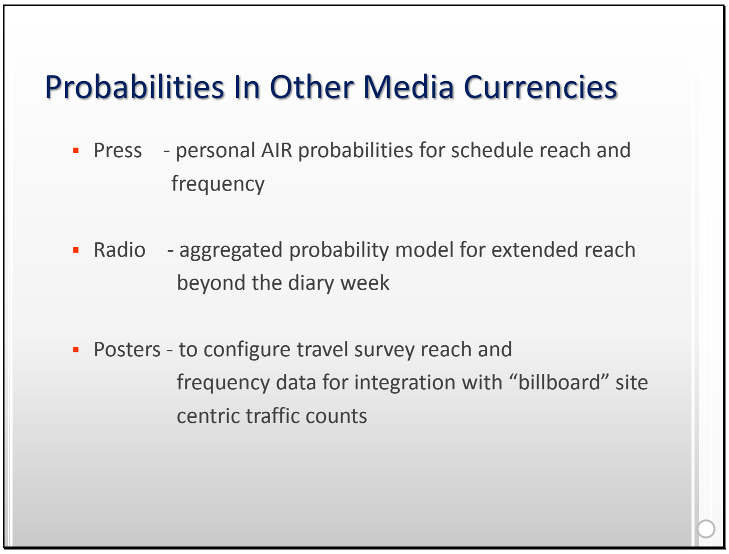
Are We Close To Breaking Point?

- In principle, TAM panels provide a transparent respondent level database and multi-purpose analysis potential
- In reality, complex gold standard calculation conventions are required to deal with panel dynamics and consistency across a range of reporting requirements
- Can this TAM panel counting system be extended to embrace integrated SC data?
- What is the potential role of probability based analysis methodology?

So I have to ask the question, are we close to breaking point? In principle, TAM panels do provide a transparent respondent level database and multi-purpose analysis potential. In reality we have complex set of calculation conventions which are required to deal with panel dynamics, like turnover, and to provide consistency across a range of reporting requirements.

Can we extend the TAM counting system to embrace integrated site centric data? If not, maybe we need to make more use of probabilities and probability modeling in the audience analysis systems.

Slide 6



Probabilities In Other Media Currencies

- Press - personal AIR probabilities for schedule reach and frequency
- Radio - aggregated probability model for extended reach beyond the diary week
- Posters - to configure travel survey reach and frequency data for integration with “billboard” site centric traffic counts

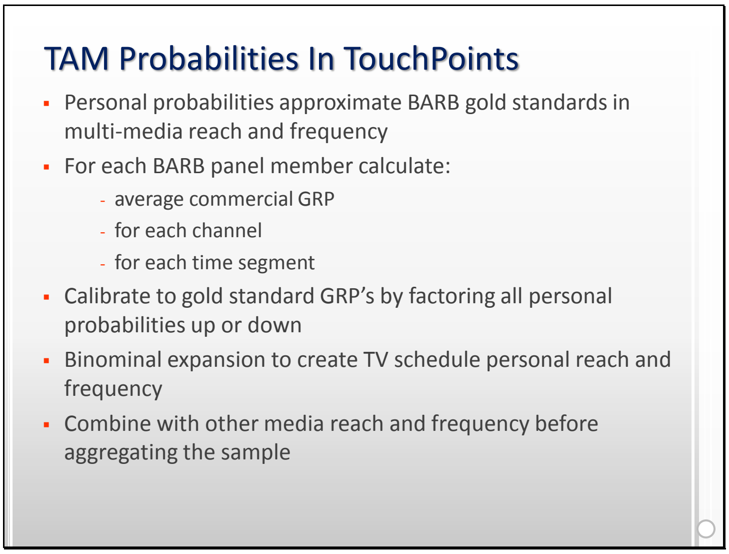
I feel it’s important to remind people that probability models are fundamental to other media currencies and already exist in TAM analysis methodology. Along the way I hope you will see some analogies with the TAM/big data situation.

Most readership systems convert claimed recency and frequency to personal probabilities for schedule reach and frequency.

Radio use formula based probability models for extended reach beyond the diary week.

Poster measurements use probabilities to configure travel survey reach and frequency for integration with site centric traffic counts. This is probably the closest analogy to our big data projects.

Slide 7



TAM Probabilities In TouchPoints

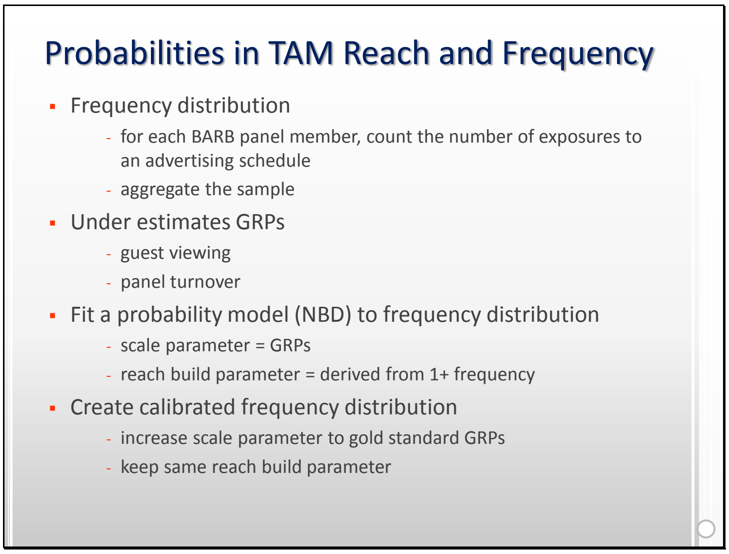
- Personal probabilities approximate BARB gold standards in multi-media reach and frequency
- For each BARB panel member calculate:
  - average commercial GRP
  - for each channel
  - for each time segment
- Calibrate to gold standard GRP's by factoring all personal probabilities up or down
- Binominal expansion to create TV schedule personal reach and frequency
- Combine with other media reach and frequency before aggregating the sample

Personal probabilities provide a common denominator for multi-media reach and frequency in TouchPoints. It is impossible to replicate the BARB gold standards in the content of multi-media.

For each BARB panel member, we calculate the average, personal, GRP for each TV channel and time segment. These are calibrated to BARB gold standard audiences by factoring all personal probabilities up or down. This factoring might be an appropriate tool for calibrating TAM panel to site centric data.

Then a binominal expansion is applied to create a TV schedule reach and frequency analysis for each person. This is then combined with the equivalent analysis for all other media in the fused TouchPoints database.

Slide 8



Probabilities in TAM Reach and Frequency

- Frequency distribution
  - for each BARB panel member, count the number of exposures to an advertising schedule
  - aggregate the sample
- Under estimates GRPs
  - guest viewing
  - panel turnover
- Fit a probability model (NBD) to frequency distribution
  - scale parameter = GRPs
  - reach build parameter = derived from 1+ frequency
- Create calibrated frequency distribution
  - increase scale parameter to gold standard GRPs
  - keep same reach build parameter

Most TAM systems already have a probability model in the reach and frequency methodology.

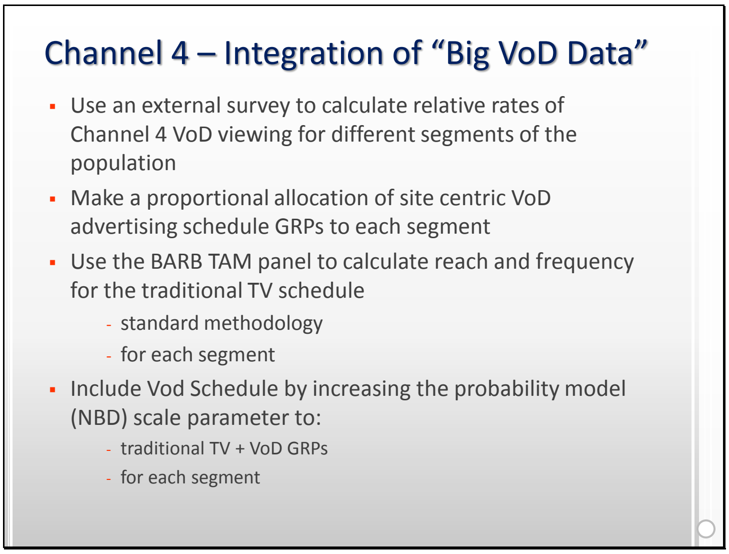
For each person we count the number of exposures to an advertising schedule and then add up the sample to create the frequency distribution. This is the number of people who viewed the schedule once, twice, three times and so-on.

This will under estimate gold standard GRPs because it doesn’t take account of guest viewing and panel turnover.

So we fit a probability model which describes the frequency distribution. This has a scale parameter which is the schedule total GRPs and a reach build parameter derived from the 1+ cover.

By increasing the scale parameter to the gold standard GRPs, but keeping the reach build parameter constant, the model produces a calibrated reach and frequency analysis.
Maybe you can see the potential for this to provide a different kind of calibration from TAM panel analysis to a set of site centric derived GRPs.

Slide 9



Channel 4 – Integration of “Big VoD Data”

- Use an external survey to calculate relative rates of Channel 4 VoD viewing for different segments of the population
- Make a proportional allocation of site centric VoD advertising schedule GRPs to each segment
- Use the BARB TAM panel to calculate reach and frequency for the traditional TV schedule
  - standard methodology
  - for each segment
- Include Vod Schedule by increasing the probability model (NBD) scale parameter to:
  - traditional TV + VoD GRPs
  - for each segment

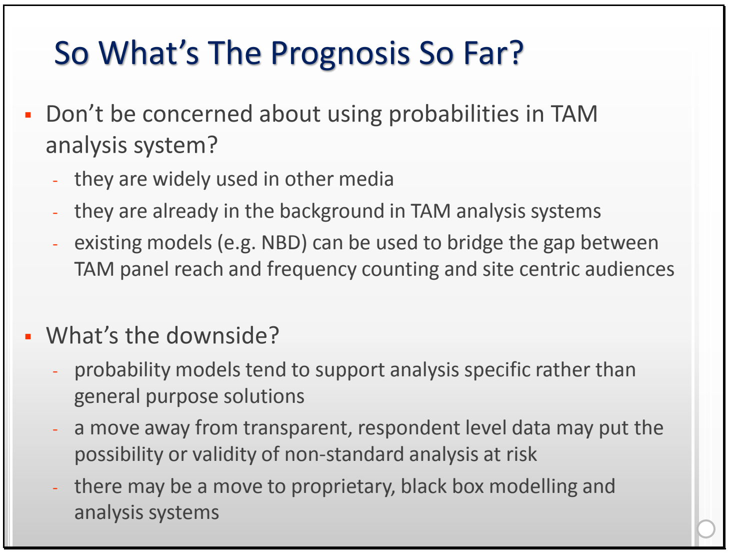
This example is taken from the paper which David Harrison and I presented last year. The requirement is to include Channel 4’s site centric video on demand data
Basically we used Kantar’s landscape survey to calculate relative rates of Channel 4, video on demand viewing for different segments of the population. For a particular advertising schedule, these relative rates are used to make a proportional allocation of site centric video on demand GRPs to each segment.

Then the BARB panel is used to calculate reach and frequency for the traditional TV schedule, using standard methodology and including the probability model.

Video on demand is included in the reach and frequency by increasing the probability model scale parameter to the combined traditional TV plus video on demand GRPs.

Of course this model will be enhanced when the BARB measurement is extended to all video on demand.

Slide 10



So What's The Prognosis So Far?

- Don't be concerned about using probabilities in TAM analysis system?
  - they are widely used in other media
  - they are already in the background in TAM analysis systems
  - existing models (e.g. NBD) can be used to bridge the gap between TAM panel reach and frequency counting and site centric audiences
- What's the downside?
  - probability models tend to support analysis specific rather than general purpose solutions
  - a move away from transparent, respondent level data may put the possibility or validity of non-standard analysis at risk
  - there may be a move to proprietary, black box modelling and analysis systems

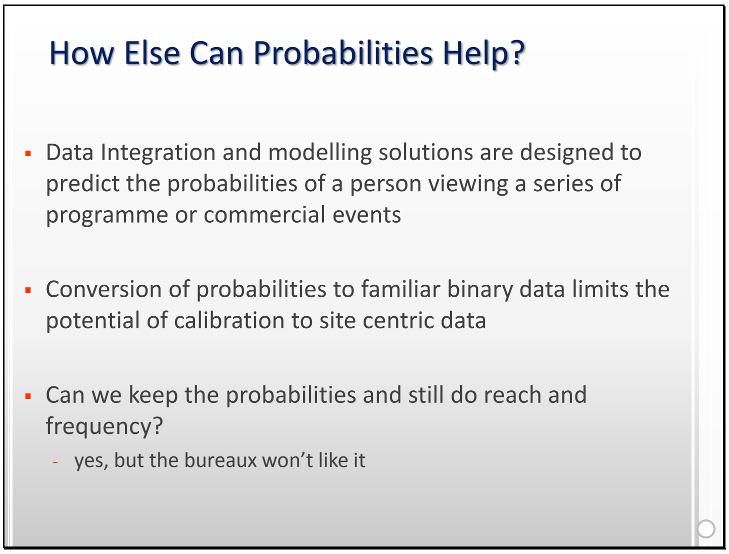
So what does all that tell us?

Well, we shouldn't be concerned about the principle of using probabilities. They are widely used in other media and in fact they are already in the background in TAM analysis systems. And I hope the examples suggest ways in which the existing models – and I'm really talking about the negative binominal or a personal Poisson process – can be used to bridge the gap between TAM panel reach and frequency counting and site centric audiences.

Is there a downside? Apart from cynicism.

Probability models tend to support analysis specific rather than general purpose solutions – the example is schedule reach and frequency. I'm concerned that a move away from transparent, respondent level data may put the possibility or validity of non-standard analysis at risk. There is a danger that if you don't consider the whole requirement, bits of methodology get bolted on one by one and eventually you realise you've gone in the wrong direction.
And I'm concerned that sophistication may drive a move to proprietary, black box modeling and analysis systems – not the BARB way at least.

Slide 11



How Else Can Probabilities Help?

- Data Integration and modelling solutions are designed to predict the probabilities of a person viewing a series of programme or commercial events
- Conversion of probabilities to familiar binary data limits the potential of calibration to site centric data
- Can we keep the probabilities and still do reach and frequency?
  - yes, but the bureaux won’t like it

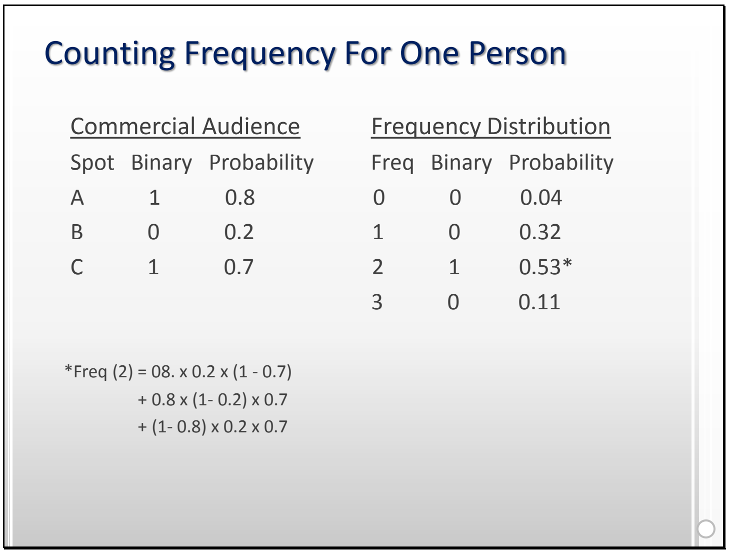
But let’s plough on and consider how else probabilities can help. This example is analogous to the TouchPoints model I showed a few charts ago.

First we need to recognise that data integration solutions are designed to predict the probabilities of a person viewing a series of programme or commercial events.

There is a danger that if we try to convert these probabilities to familiar binary data formats, we will limit the potential for calibration to site centric data.

So the question is, can we keep the probabilities and still do reach and frequency? The answer is probably yes, but I don’t think the TAM analysis bureaux will like it!

Slide 12



Counting Frequency For One Person

Commercial Audience			Frequency Distribution		
Spot	Binary	Probability	Freq	Binary	Probability
A	1	0.8	0	0	0.04
B	0	0.2	1	0	0.32
C	1	0.7	2	1	0.53*
			3	0	0.11

*Freq (2) = 08. x 0.2 x (1 - 0.7)
+ 0.8 x (1- 0.2) x 0.7
+ (1- 0.8) x 0.2 x 0.7

This is the sort of thing we might be asking them to do.

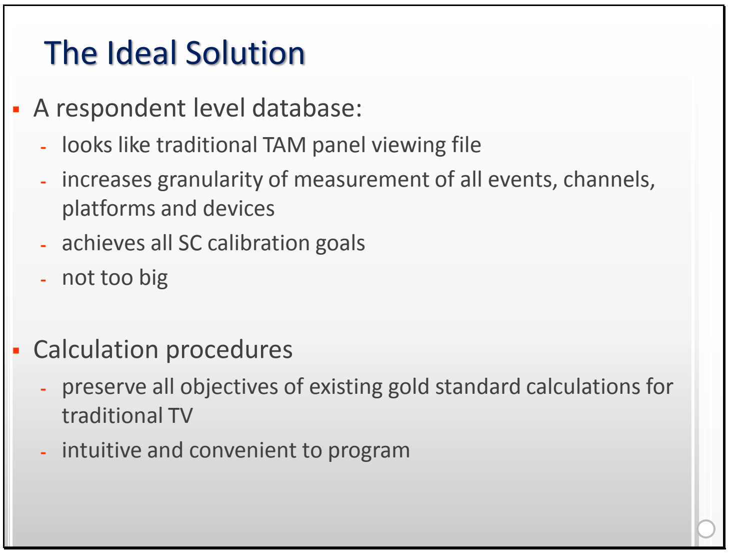
The table on the left shows the two ways that the data would be held for one person. We've got three advertising spots A, B and C. In the binary world our person either sees the spots or not, zero or one. In the probability world, our person has an 80% chance of viewing spot A.

The table on the right shows the contribution to the frequency distribution for our person. In the binary world we just count the ones; our person has a frequency of two.

In the probability world, we have to multiply probabilities together and our person had a probability for every frequency of two.

In the probability world, we have to multiply probabilities together and our person has a probability for every frequency. I've asterisked one number. Our person has a probability of 0.53 (53 %) - of seeing two of the three spots. There are six multiplications and two additions to get this number. Imagine if you had a thousand spots!

Slide 13



The Ideal Solution

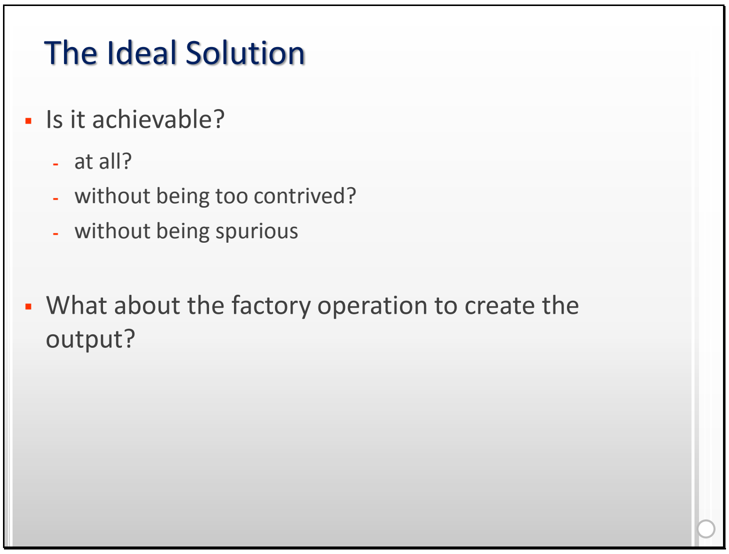
- A respondent level database:
  - looks like traditional TAM panel viewing file
  - increases granularity of measurement of all events, channels, platforms and devices
  - achieves all SC calibration goals
  - not too big
- Calculation procedures
  - preserve all objectives of existing gold standard calculations for traditional TV
  - intuitive and convenient to program

So let’s try and pull this together.

We know that the ideal solution is a respondent level database which looks like a traditional TAM panel viewing file. We need increased granularity of measurement of all events, channels, platform and devices, for both audiences and reach and frequency. And at the same time as achieving all our site centric calibration goals, the database mustn't be too big.

Moreover, the calculation procedures must preserve all the objectives of the existing gold standard for traditional TV and they must be intuitive and convenient to programme.

Slide 14



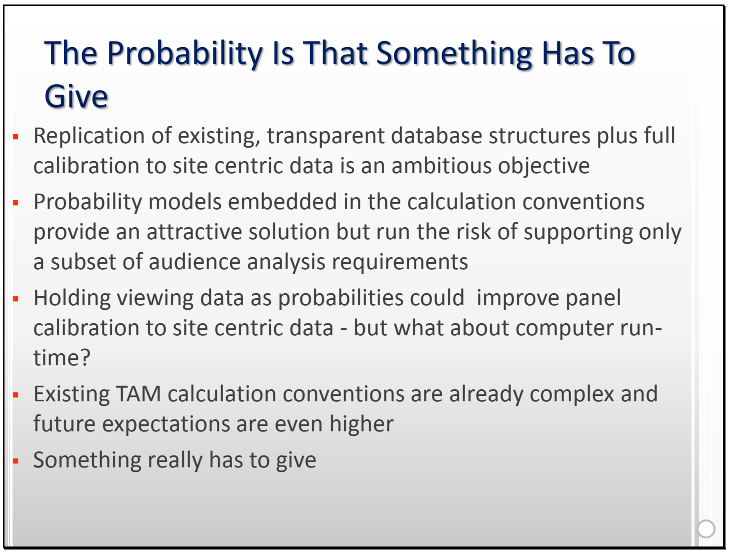
The Ideal Solution

- Is it achievable?
  - at all?
  - without being too contrived?
  - without being spurious
- What about the factory operation to create the output?

We have to ask if this is achievable at all, and if so, will it be too contrived or even spurious?

And I haven’t even been talking about the factory operation which is necessary to create the ideal output.

Slide 15



The Probability Is That Something Has To Give

- Replication of existing, transparent database structures plus full calibration to site centric data is an ambitious objective
- Probability models embedded in the calculation conventions provide an attractive solution but run the risk of supporting only a subset of audience analysis requirements
- Holding viewing data as probabilities could improve panel calibration to site centric data - but what about computer run-time?
- Existing TAM calculation conventions are already complex and future expectations are even higher
- Something really has to give

So to conclude.

Replication of existing, transparent database structures plus full calibration to site centric data is an ambitious objective, in both statistical and IT terms.

Probability models embedded in the calculation conventions provide an attractive solution, but run the risk of being too analysis specific.

Holding viewing data as probabilities could improve calibration of panel data to site centric data, but we have to be sure that we are not creating an IT monster.

Against the background that calculation conventions are already complex and that future expectations are even higher, the probability really is that something has to give.

Slide 16

RSMB