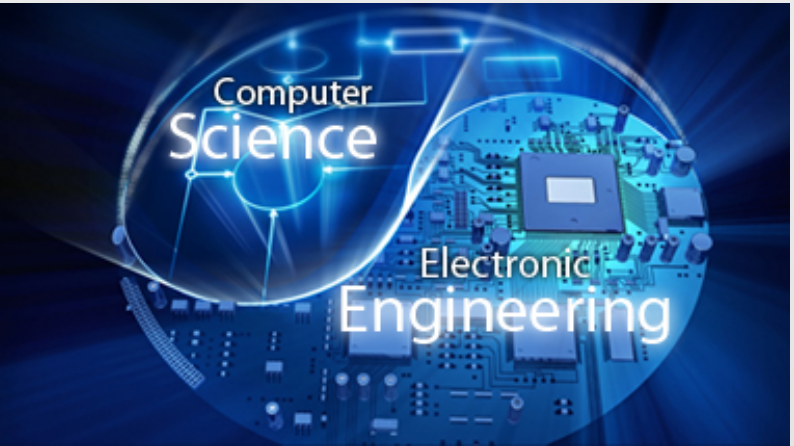
Decision Making from Data: Causes and Uncertainty

William Marsh, william@eecs.qmul.ac.uk

Risk Assessment and Decision Analysis Research Group



Queen Mary
University of London

Acknowledgements

- RSSB
  - George Bearfield, Anna Holloway
  - http://www.rssb.co.uk
- Risk Assessment group at QMUL
  - Professors Norman Fenton, Martin Neil
  - http://www.dcs.qmul.ac.uk/research/radar/

Aims

- Potential uses of Bayesian networks for *decision making from data*
- … application to analysis of incidents
- Convince you of the importance of causal modelling for *decision making from data*

- Get feedback on potential

Outline

- Introduction
- Bayesian networks and causal model
- A case study: railway safety incidents
- Wider applications
- Conclusions

Data

- What data do you have?

About us | Policy, guidance and research | Consultations | Press office | FAQs | Ministers

DfT home > Policy, guidance and research > Transport statistics > Statistics (data, tables and publications)

Policy, guidance and research

- Aviation
- Crime and public transport
- Economics and appraisal
- Transport evaluation
- Freight
- Railways
- Regional and local transport
- Transport Resilience
- Road safety
- Roads and vehicles
- Science and research
- Shipping and ports
- Social inclusion
- Sustainable travel
- Transport security
- **Transport statistics**
  - Recent publications
  - **Statistics (data, tables and**

Accidents, casualties and safety

Road Accident Statistics play a leading part in the Government's Road Safety Strategy.

In addition to the STATS19 information on road accidents, other data sources directly related to road safety have been used to compile the statistics on this web page. These include death registrations and coroners' reports as well as traffic and vehicle registration data plus Home Office data on motor vehicle offences.

Email roadacc.stats@dft.gov.uk for more information about the surveys and their findings.

Reported road casualties in Great Britain: main results

The latest bulletin containing statistics on personal injury accidents on public roads (including footways).

Published: 26 June 2003 **Last update:** 30 June 2011

Road Casualties Online

Road Casualties Online is a web based data analysis tool which provides detailed statistics about the circumstances of reported personal injury road accidents in Great Britain, including the types of vehicles involved and the consequent casualties.

Reported Road Casualties in Great Britain: Quarterly Provisional Estimates Q3 2010

Latest quarterly provisional estimates of personal injury road accidents and their casualties.

Published: 05 August 2010 **Last update:** 03 February 2011

Data

NHS

The Information Centre for health and social care

Search

Edit search options | Browse by subject | Help

- Publications calendar
- Audits and performance
- Health and lifestyles
- Hospital care
  - Cancer
  - Coronary heart disease
  - Hospital activity (Hospital Episode Statistics - HES)
  - Maternity
  - Outpatients
  - Accident and Emergency Hospital Episode Statistics (HES)
  - Patient Reported Outcome Measures (PROMS)
  - Critical care
  - Summary Hospital - level Mortality Indicator (SHMI)

Home | Statistics & data collections | Hospital care

Hospital activity (Hospital Episode Statistics - HES)

Have you ever wondered...

- How many people are treated for alcohol-related conditions?
- How many people are admitted to hospital after a dog bite?
- How many people have their tonsils removed each year?
- What the average waiting time is for hip replacements?

Hospital Episode Statistics (HES) data has been used by the NHS, government, BBC, newspapers and many other organisations and individuals to answer questions on these topics and more.

What is HES?

HES is a data warehouse that contains information about hospital admissions and outpatient attendances in England. The data in HES comes from the Secondary Uses Service (SUS), which collects data that's passed between healthcare providers and commissioners.

Hospital Episodes Statistics (HES) Training

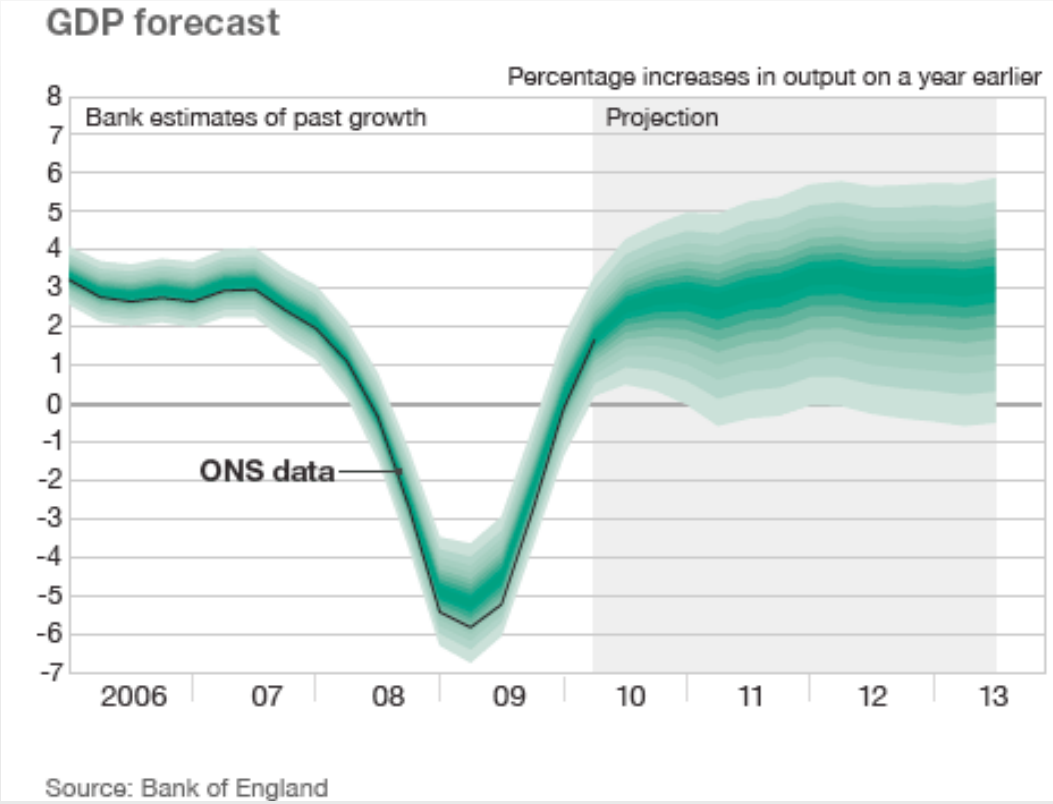
We offer a 2 day training course to NHS staff and other bodies within the NHS family which allows you to access the record level HES database.

This training focuses on how to use and analyse HES data within Business Objects, which is the software we use to access the database.

The training is intended for Analysts working within the NHS and other bodies within the NHS family who require access to HES. Delegates are expected to have some experience working in an analytical environment, using

Decision Making from Data

- What has happened?
  - Observe patterns in the data



- What should we do?
  - Estimate effect of change

Causal Modelling with Bayesian Networks

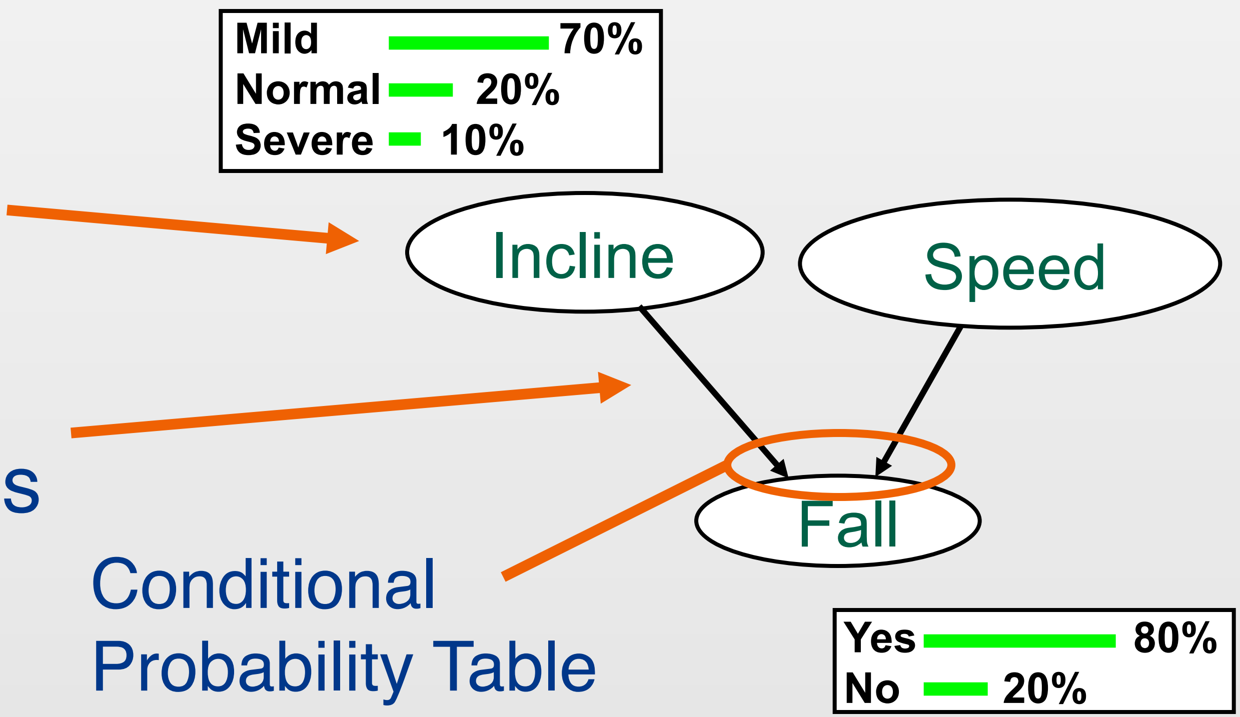
- What’s a BN
- Why Causal Models

Bayesian Networks

$$P(A \mid B).P(B) = P(B \mid A).P(A)$$

Bayes' Theorem

- Uncertain variables
- Probabilistic dependencies

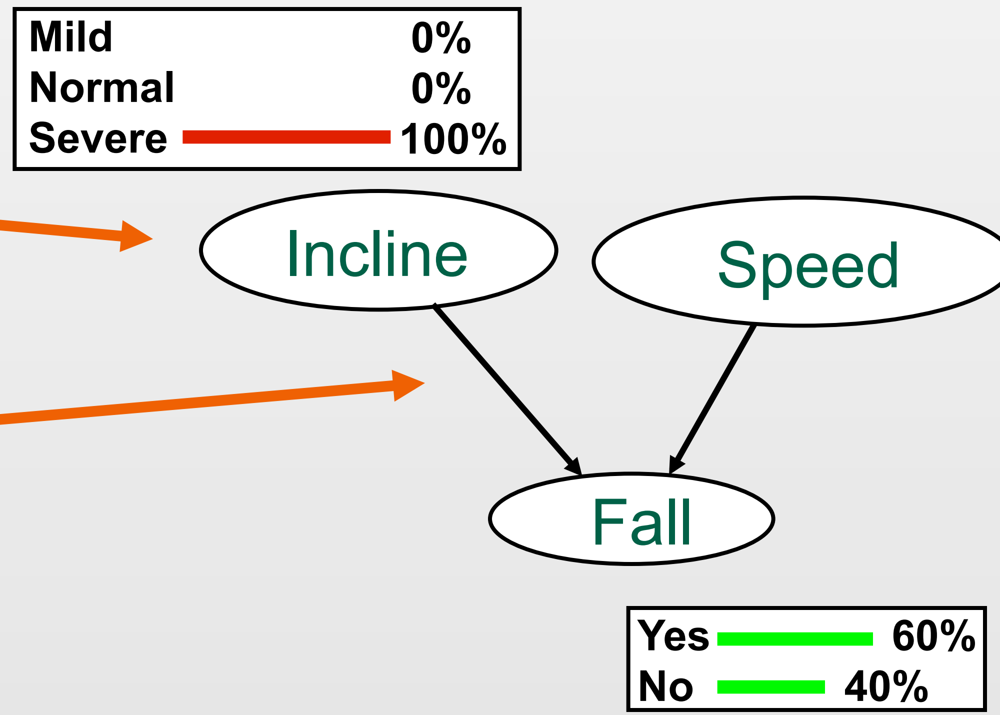


Bayesian Networks

$$P(A \mid B).P(B) = P(B \mid A).P(A)$$

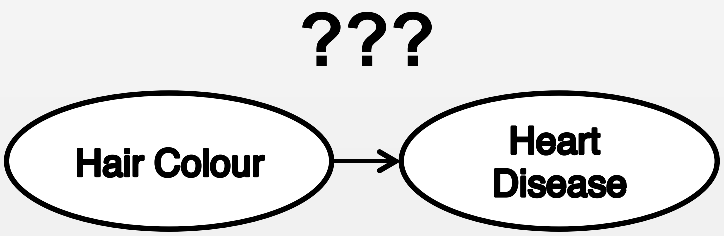
Bayes' Theorem

- Uncertain variables
- Probabilistic dependencies
- Efficient inference algorithms

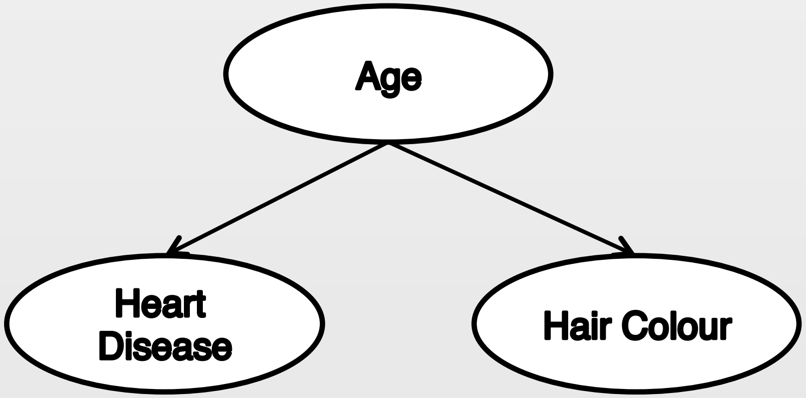


Association, Causality & Interventions

- Need for causal relations
  - Cause → Effect



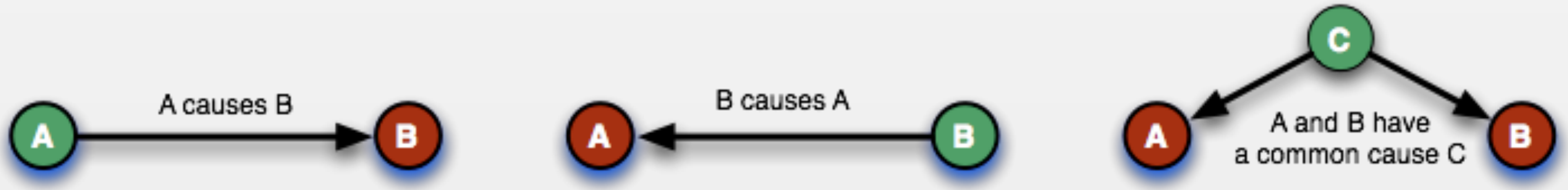
- Association vs. Causation
  - Grey hair predicts heart disease
  - Colouring hair to reduce risk?



- Identifying causes
  - Experiment (e.g. medical trials)
  - Domain Knowledge + Observational Data

Causality from Data

- In general, hard to distinguish causal relations from data



- Our approach
  - Causal relationships from knowledge
- Example ‘systems engineering’ causal models
  - Fault trees
  - Simulations

Why Does Causality Matter?

- Change cause … change consequences
- What a cause is!

Causal claim

Step right up! It's the miracle cure we've all been waiting for.

It can reduce your risk of major illnesses, such as heart disease, stroke, diabetes and cancer by up to 50% and lower your risk of early death by up to 30%.

It's free, easy to take, has an immediate effect and you don't need a GP to get some. Its name? Exercise.

Case Study: Railway Incidents

- Background and aims
- BN model and data analysis
- Uses of the model
- Further work

Safety Management Information System (SMIS)

- SMIS – database of safety related events that
  - UK rail network
  - Use is mandatory on Network Rail managed infrastructure
- Purpose
  - Analysing risk
  - Predicting trends

> *"key to successful management, planning and decision making within the industry"*

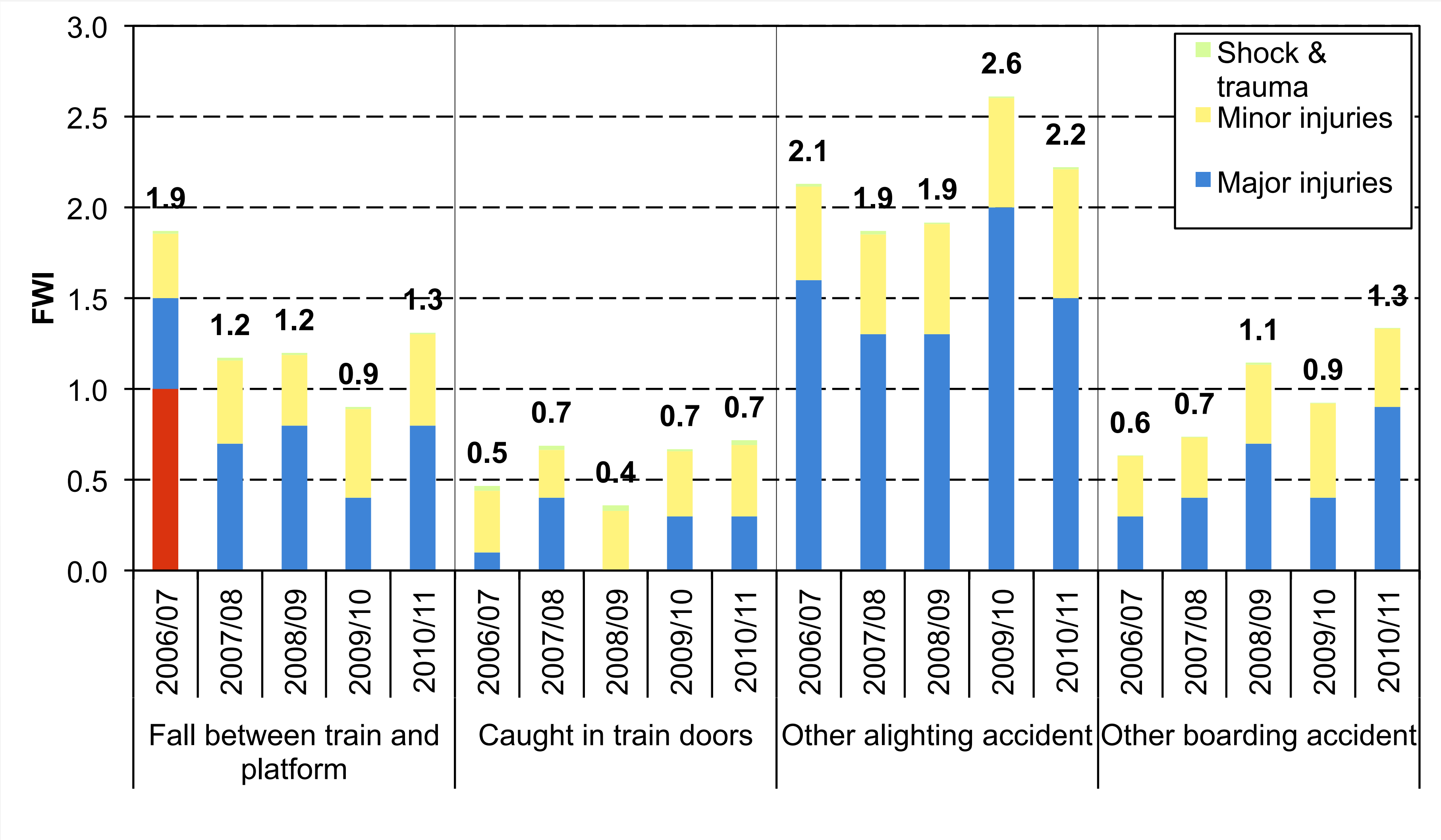
- Development began in 1997
- Over 1.5 million events have been recorded

Boarding and Alighting from Trains

- Accidents to passengers getting on and off trains

Boarding and Alighting

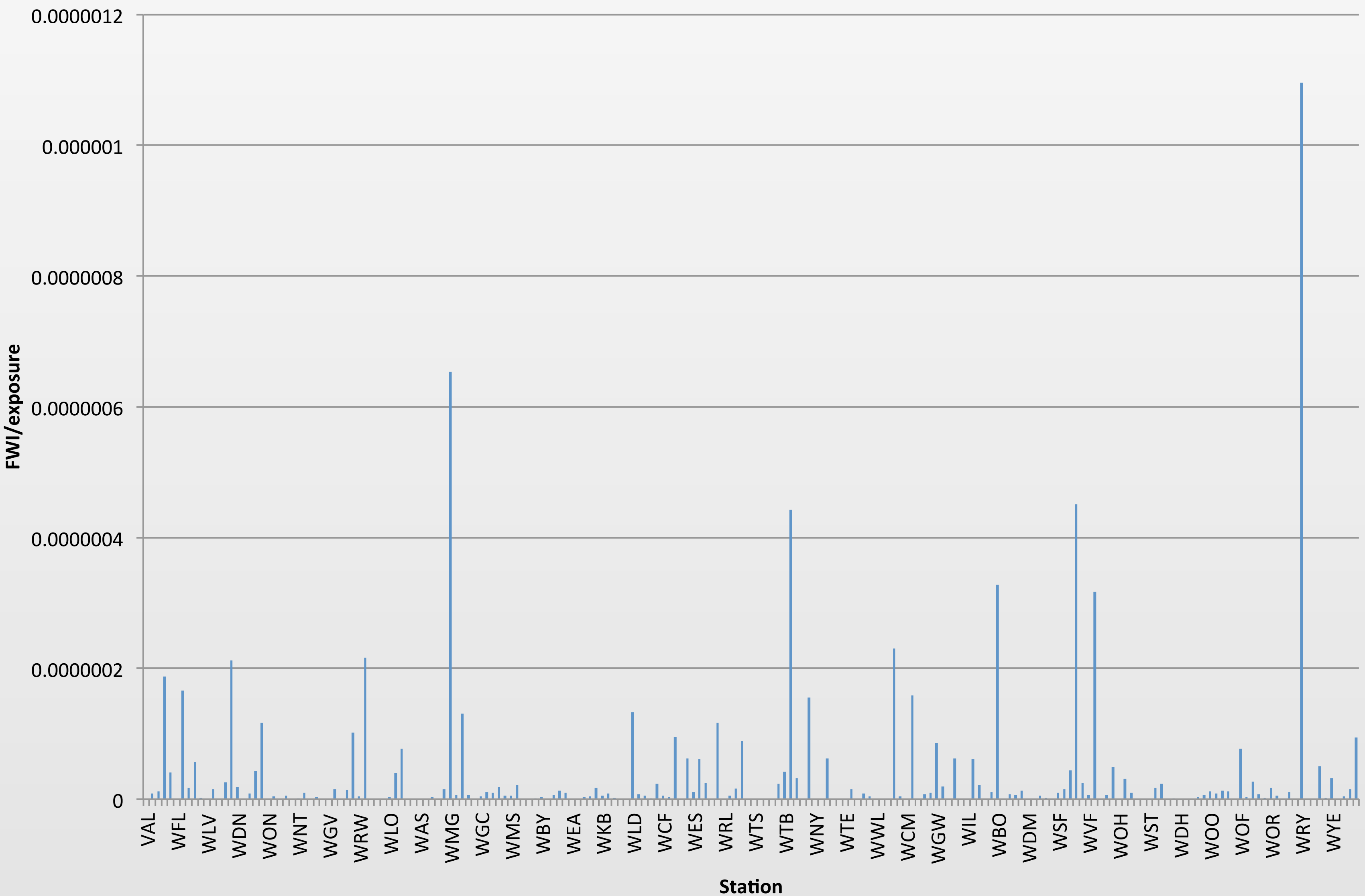
- From 2011 Annual Safety Performance Report



Problem To Solve

- Categorisation of data
  - Network average risk figures
- Risk Management is local
  - E.g. at stations or platform
  - Local estimates of the risk are needed
- Few safety incidents at most locations
- How do we use the data to estimate local risk?
  - Current data + assumptions
  - More data in future

Observed Normalised FWI



Modelling Aims

- National average and local risk estimates
  - Train operating company
  - Region
  - Station
- Understand the risk contribution of causes
- Estimate the change in risk associated with changes to operations, assets
  - Improvements
  - Acceptable savings

Case Study: Railway Incidents

- Background and aims
- **BN model** and data analysis
- Uses of the model
- Further work

Modelling Concept

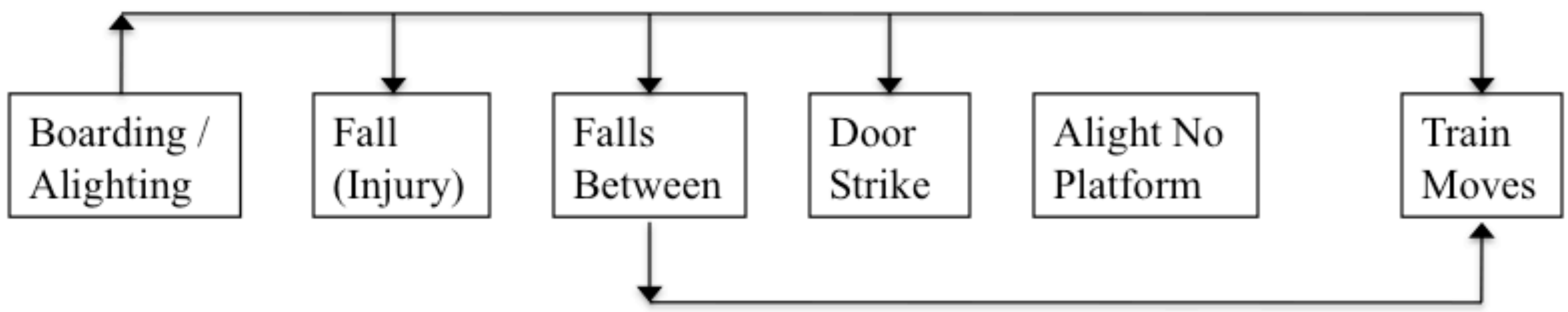
- Incident data
  - Categorize events
  - Presence of causes in events (e.g. ice, crowding)

- Context: how railway is used
  - Presence of causal factors

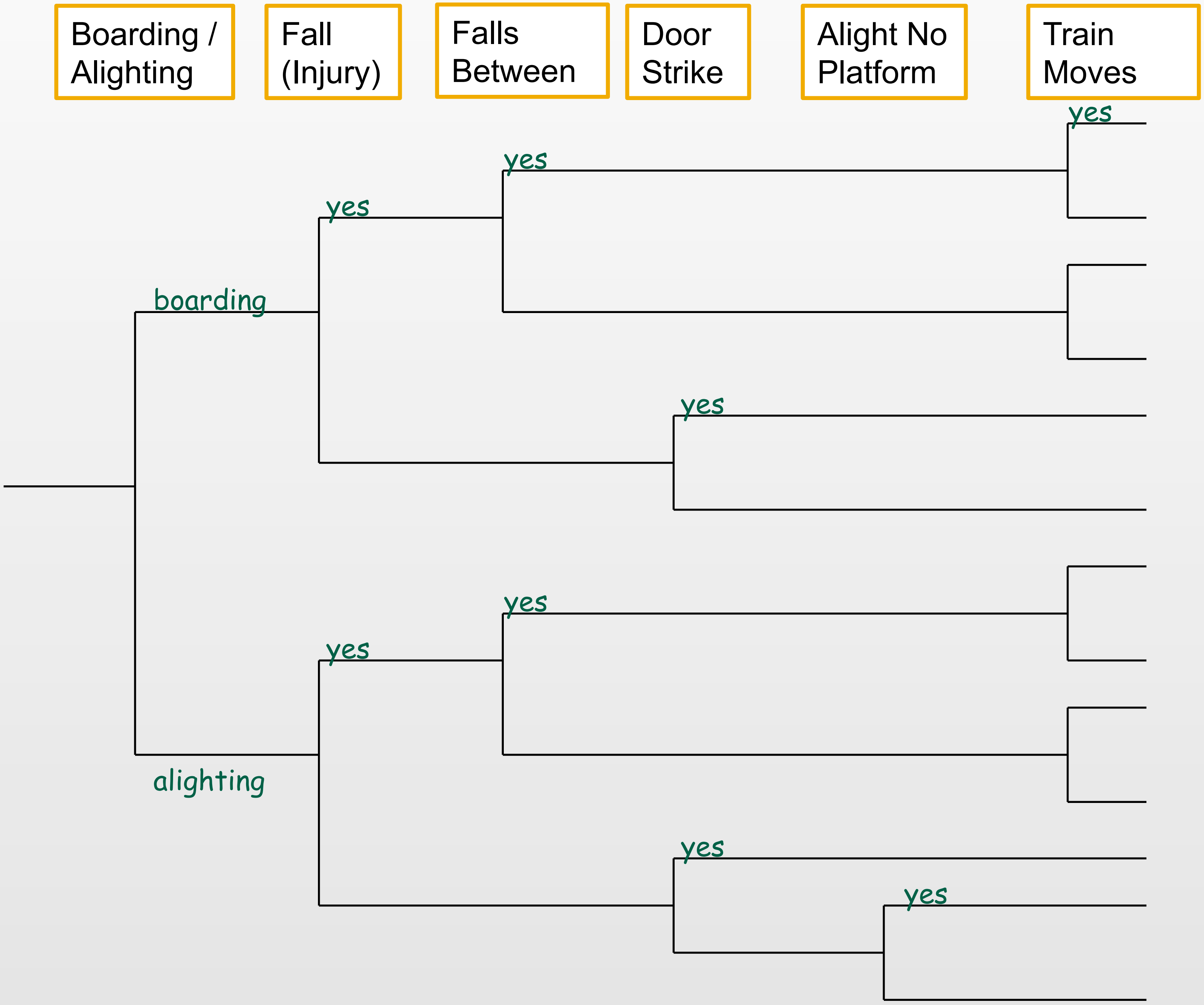
- Estimate effect of causes on the probability of incidents

Events Sequence

- Model the event sequence
  - Align to existing categories

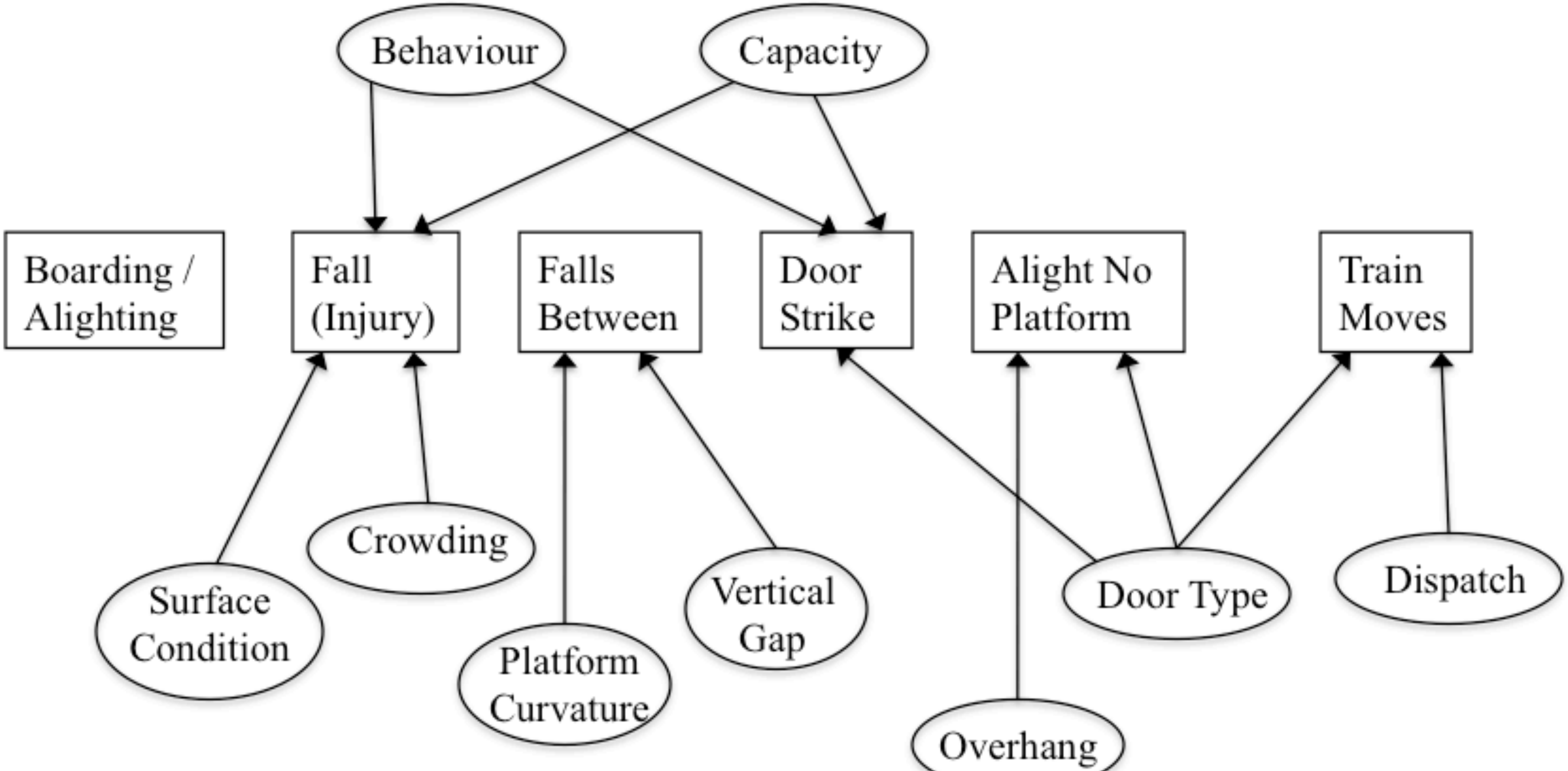


- Model direct causes of each event


Boarding / Alighting
Fall (Injury)
Falls Between
Door Strike
Alight No Platform
Train Moves
yes
yes
yes
boarding
yes
yes
yes
alighting
yes
yes

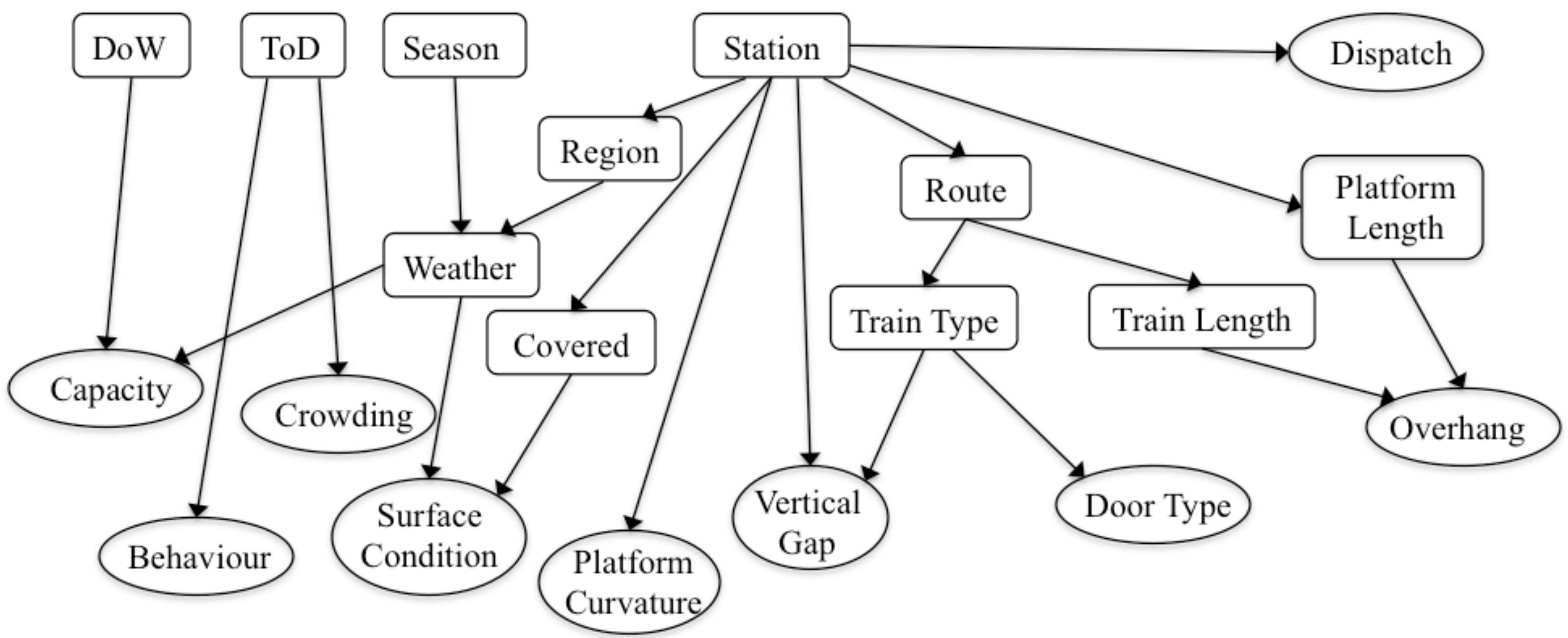
Direct Causal Factors

- Elicit possible causes for each event
  - Assumes knowledge



Top-Level Factors

- Determine the occurrence of the causal factors

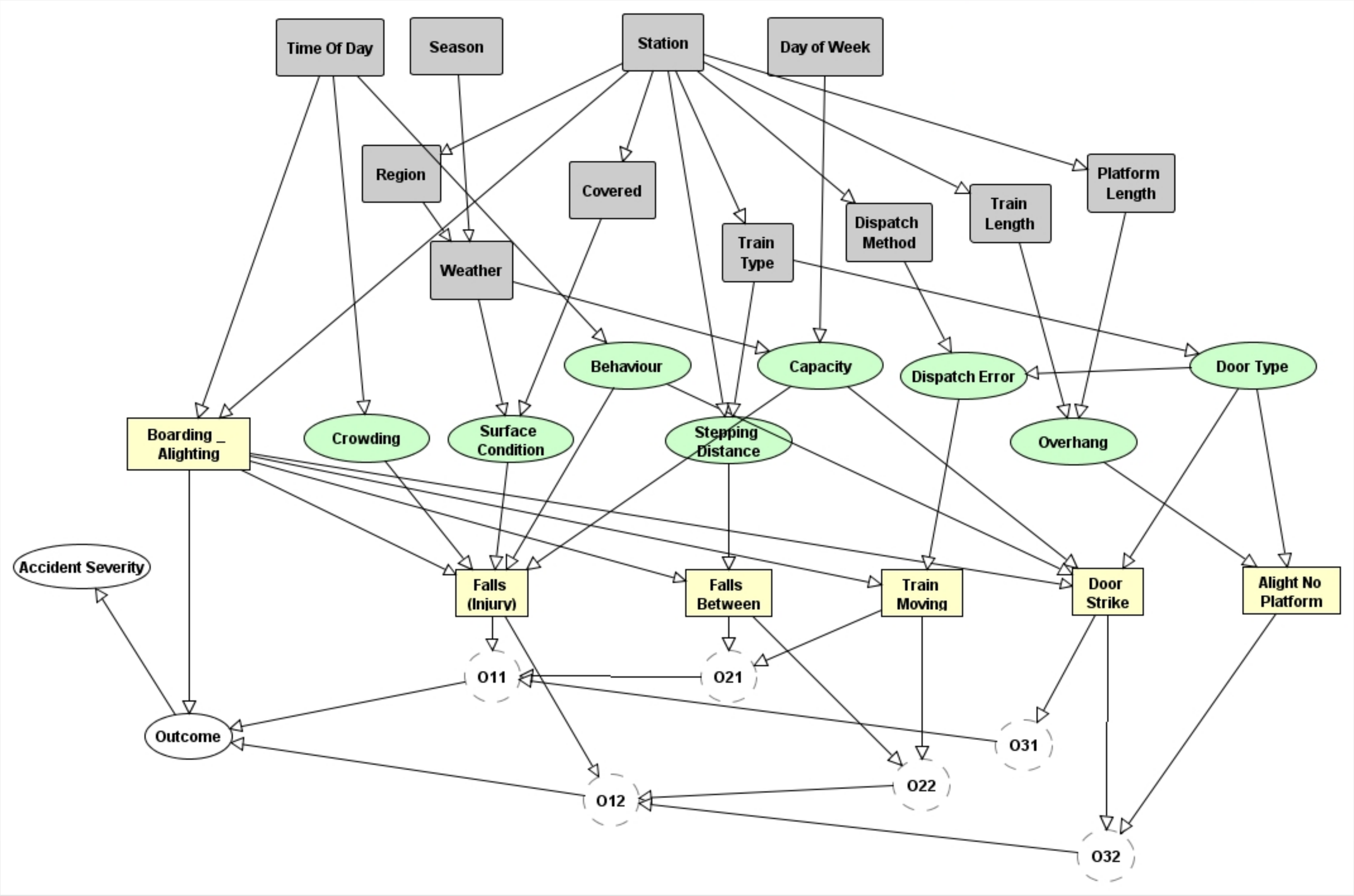


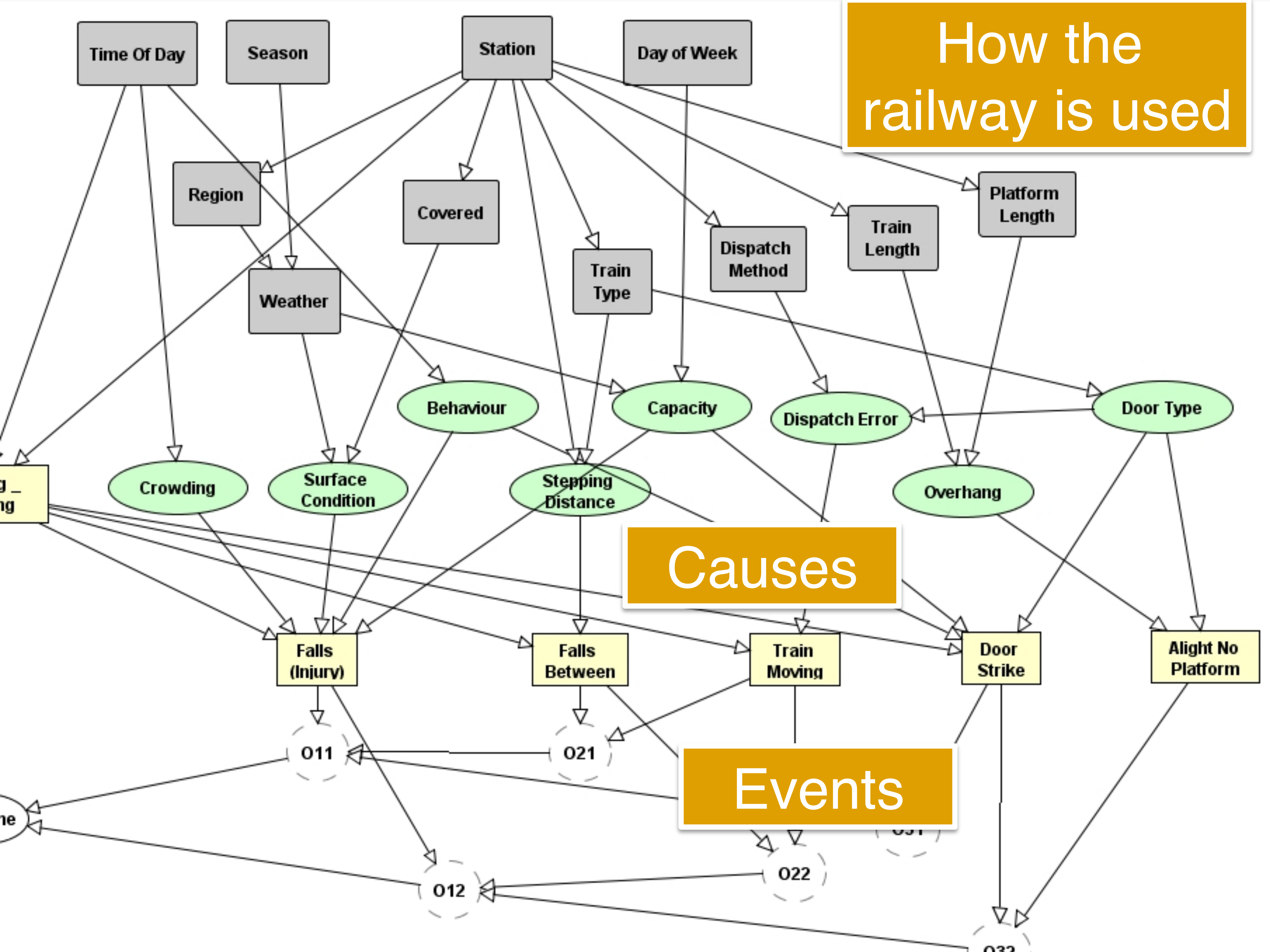
Summary of Model Structure

- Overall problem
  - Model probability of outcomes at each station
- Three levels
  - Level 1: the sequence of events
  - Level 2: immediate causes
  - Top-level: usage, i.e. exposure to risk
- Example of reasoning

X% of boarding and alighting events are made on curved platforms but a greater proportion of of incidents of falling between platform and train occur on curved platforms, so curvature increases the probability of these events

Final Structure





How the railway is used
Causes
Events
Time Of Day
Season
Station
Day of Week
Region
Covered
Train Type
Dispatch Method
Train Length
Platform Length
Weather
Behaviour
Capacity
Dispatch Error
Door Type
Crowding
Surface Condition
Stepping Distance
Overhang
Falls (Injury)
Falls Between
Train Moving
Door Strike
Alight No Platform
O11
O21
O12
O22

Case Study: Railway Incidents

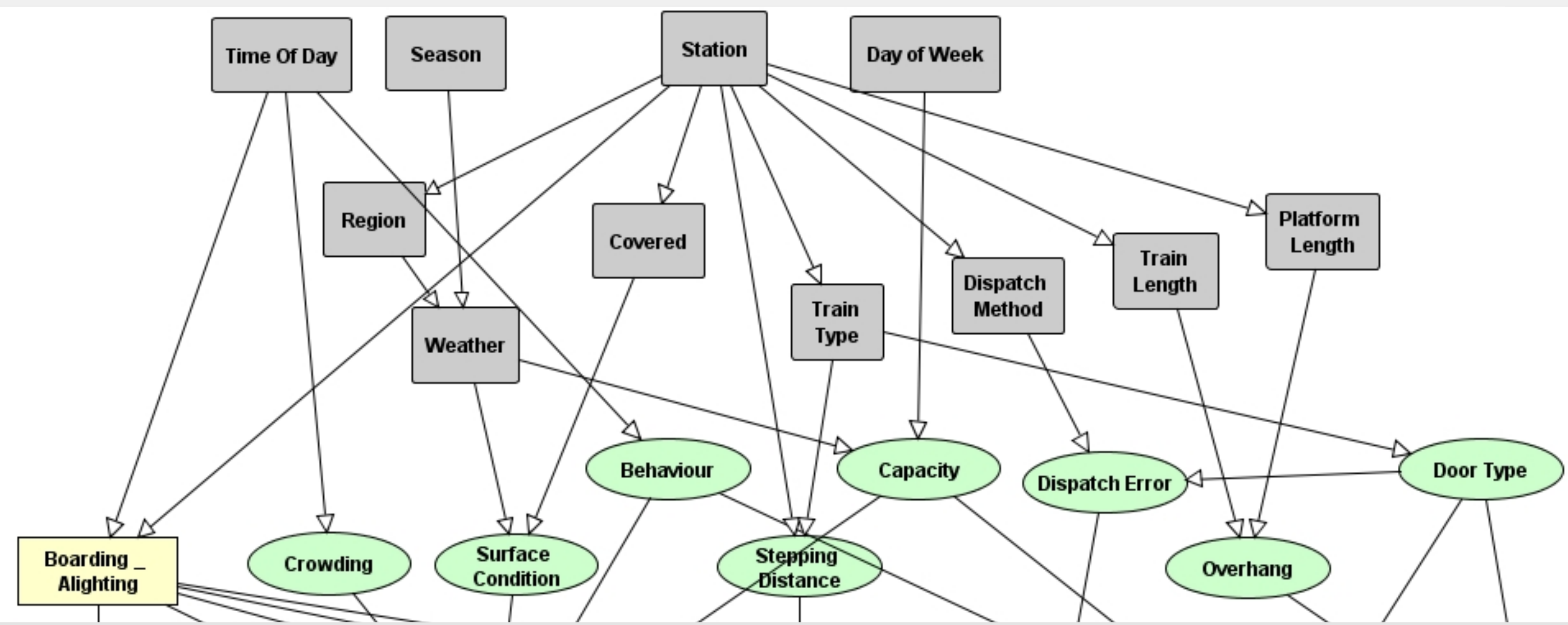
- Background and aims
- BN model and data analysis
- Uses of the model
- Further work

Priors versus Causes Seen

- Example: crowding
  - (Prior) probability of boarding/alighting when crowded?
  - How many incidents occur when crowded?
- If crowding a cause then
  - Expect more crowding in incidents than in normal use
  - Step 1: incidents while crowded
  - Step 2: how much crowding
- When / where crowded?
  - Time of day → crowded (Step 2)
  - Step 3: proportion of boarding / alighting by time of day

Usage Model

- How many correlations?
  - Time of Day, Station assumed independent
  - Time of day → Boarding / Alighting



Data on Usage

- Multiple sources
- Probabilistic approximations

ORR Station Usage
Train Service Database (TSDB)
Locomotives and Coaching Stock 2007
T866 Platform Investigation to Support Research into the Reduction in Passenger Stepping Distance
DfT – Significant Steps Research
DFT National Travel Survey
SRM Normalisers
MET Office
Assisted Passenger Request System (APRS)
T763 dispatch data

Example: Train Length

- Data available: deterministic

Location Name	TLC	Platform	BRAND_NAME	TC1	TrainLength Cars	Number of stops per week	Length of train (m)
Abbey Wood	ABW		Southeastern	376	5	129	100
Abbey Wood	ABW		Southeastern	376	10	105	200
Abbey Wood	ABW		Southeastern	465	4	174	80
Abbey Wood	ABW		Southeastern	465	6	135	120
Abbey Wood	ABW		Southeastern	465	8	495	160
Abbey Wood	ABW		Southeastern	465	10	20	200
Aber	ABE		Arriva Trains Wales	142	2	20	30
Aber	ABE		Arriva Trains Wales	142	4	40	60
Aber	ABE		Arriva Trains Wales	143	2	5	30
Aber	ABE		Arriva Trains Wales	143	4	90	60
Aber	ABE		Arriva Trains Wales	150	2	105	40
Aber	ABE		Arriva Trains Wales	150	4	10	80

Example: Train Length

- Model of proportion of train stops with a given carriage length
  - Probability weights by usage

		Train Length											
Location Name	**TLC**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
Abbey Wood	ABW	0.00	0.00	0.00	0.16	0.12	0.13	0.00	0.47	0.00	0.12	0.00	0
Aber	ABE	0.05	0.44	0.48	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
Abercynon South	ACY	0.09	0.66	0.23	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
Aberdare	ABA	0.09	0.73	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
Aberdeen	ABD	0.00	0.18	0.36	0.26	0.05	0.06	0.01	0.00	0.00	0.04	0.05	0

Example: Passenger Capacity

- Based on many factors:
  - Alcohol → Incident data
  - Age → NTS data
  - Luggage /large objects → assumptions
  - Illness → assumptions
  - Disability → ATOC data

Case Study: Railway Incidents

- Background and aims
- BN model and data analysis
- Uses of the model
- Further work

Types of queries and results

- Profile
  - Risk per exposure event
  - Aggregate
- Change of risk
  - Lengthening trains
  - Station staffing
  - Curvature
- *Explanation of incident*

Profile: Region

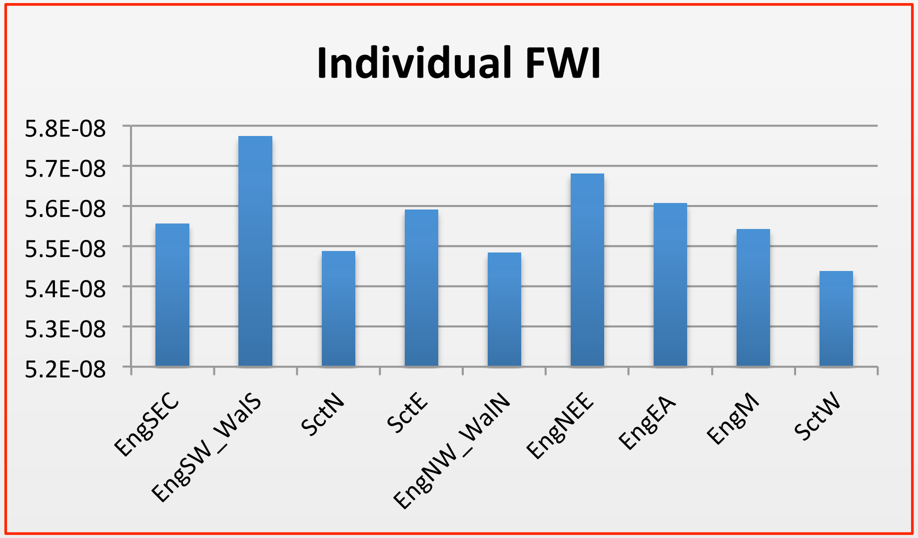
- Query

Profile	Region
Marginal	Severity

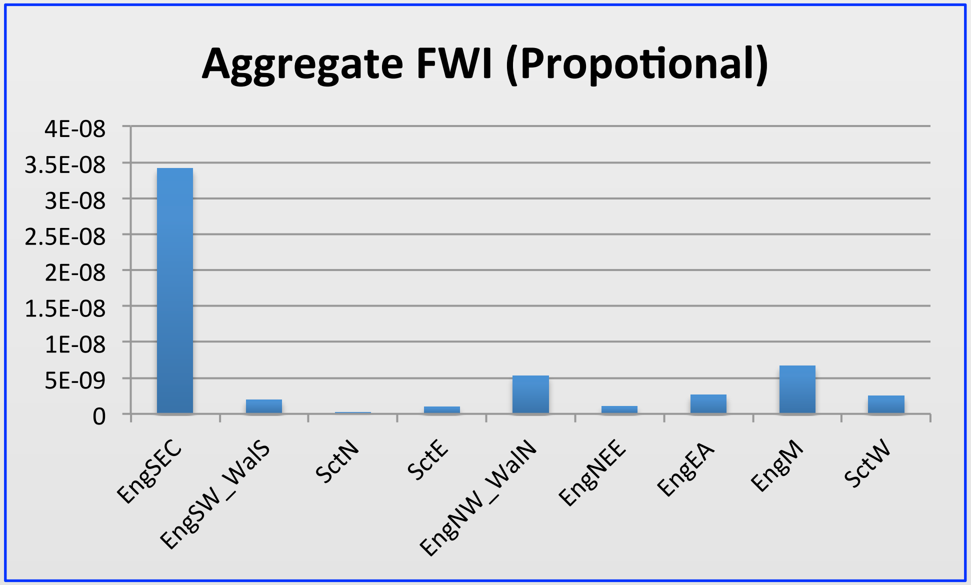
- Result

Region	Probability	FT	MA	MR	MN	ST
EngSEC	0.614634996	1.24E-09	1.10E-07	7.94E-07	3.45E-06	2.00E-07
EngSW_WalS	0.03403162	1.21E-09	1.13E-07	8.20E-07	4.00E-06	2.39E-07
SctN	0.003524171	1.21E-09	1.08E-07	7.84E-07	3.48E-06	2.02E-07
SctE	0.018044351	1.30E-09	1.10E-07	7.98E-07	3.52E-06	2.07E-07
EngNW_WalN	0.096698011	1.23E-09	1.08E-07	7.84E-07	3.42E-06	1.99E-07
EngNEE	0.018532217	1.27E-09	1.11E-07	8.08E-07	3.78E-06	2.25E-07
EngEA	0.047264548	1.25E-09	1.11E-07	8.00E-07	3.57E-06	2.09E-07
EngM	0.120774015	1.27E-09	1.09E-07	7.92E-07	3.47E-06	2.02E-07
SctW	0.046496071	1.24E-09	1.07E-07	7.78E-07	3.36E-06	1.94E-07

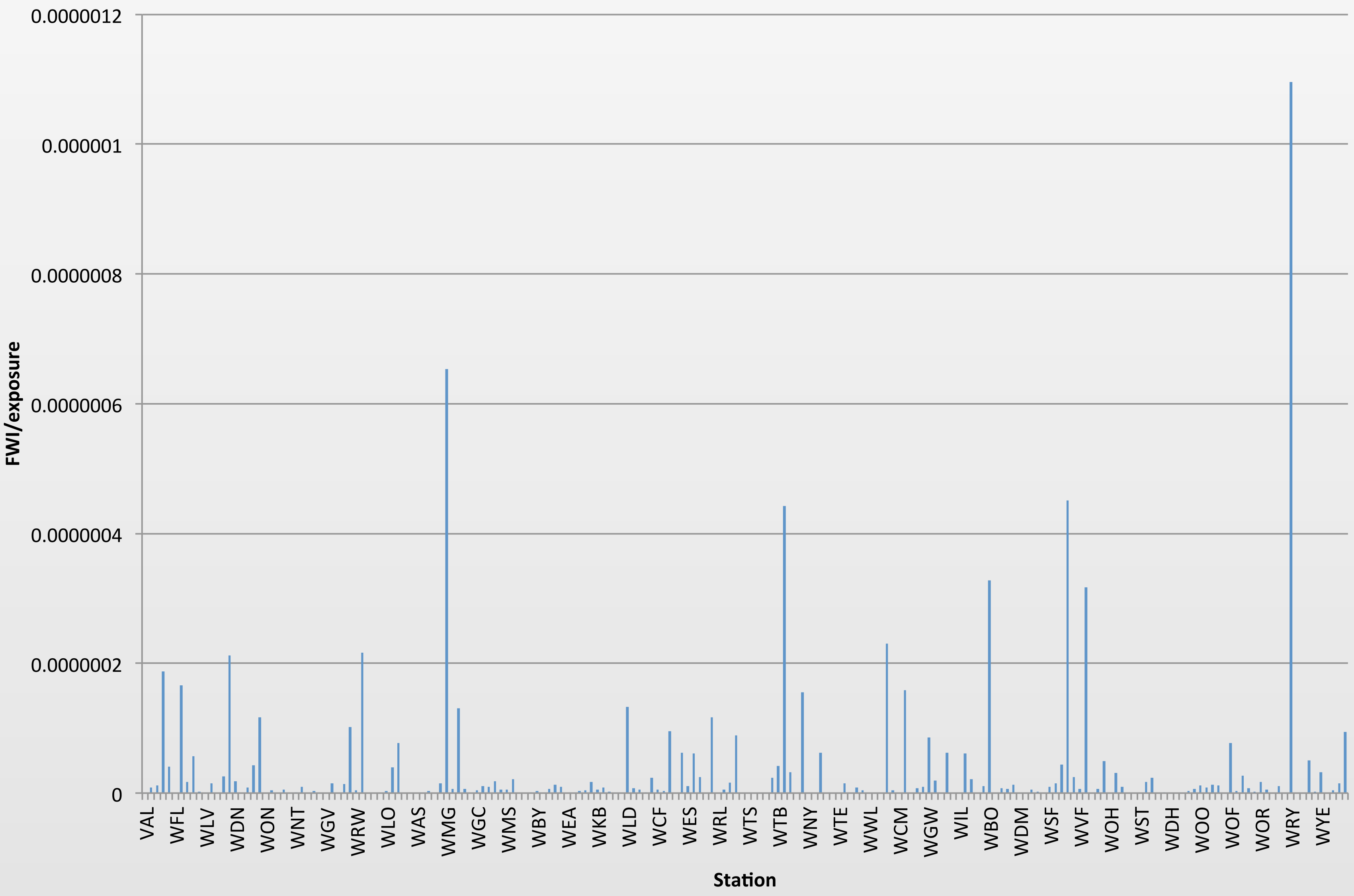
Profile: Region



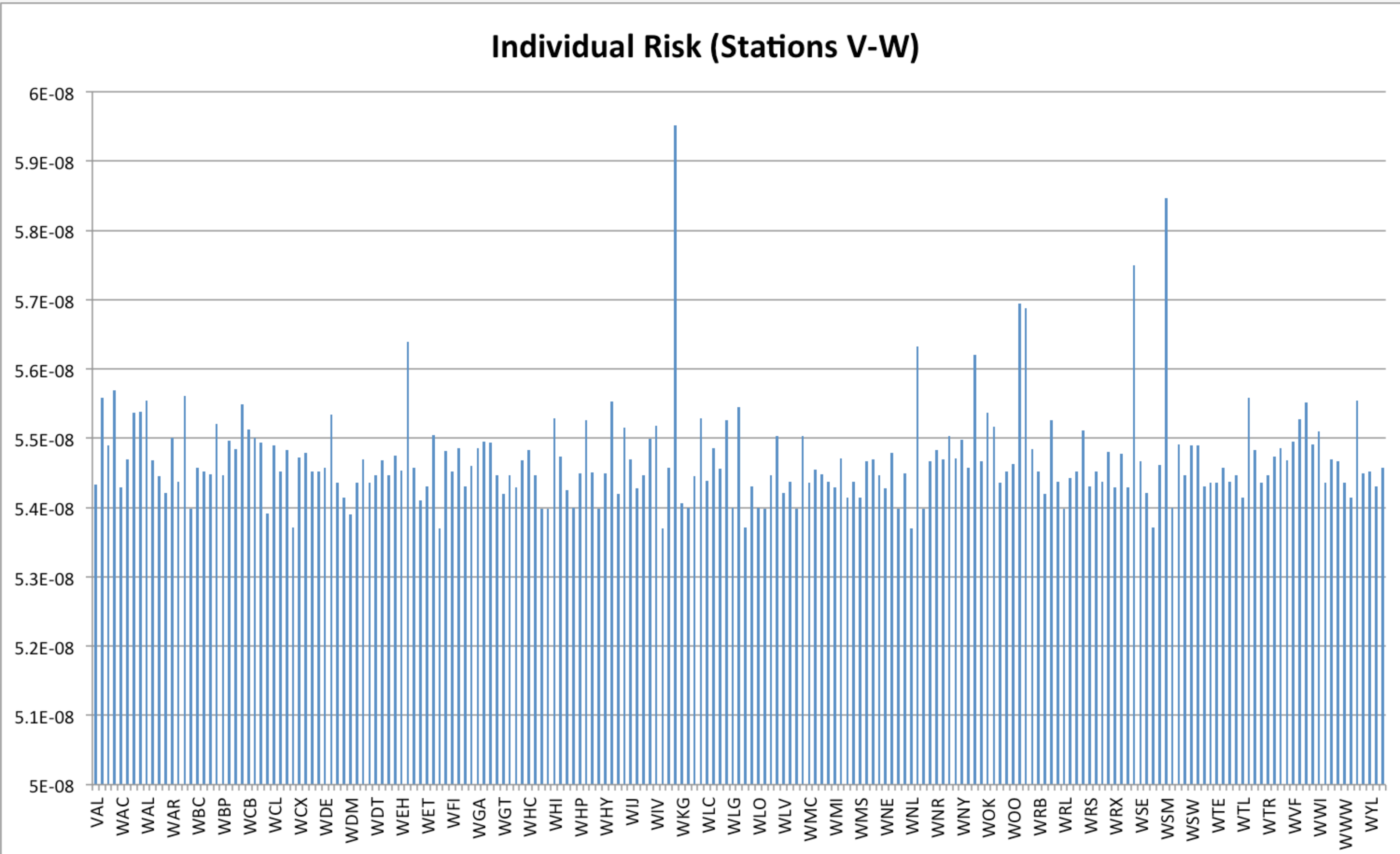
- Profile of several variable possible
- Calculates probability of scenario



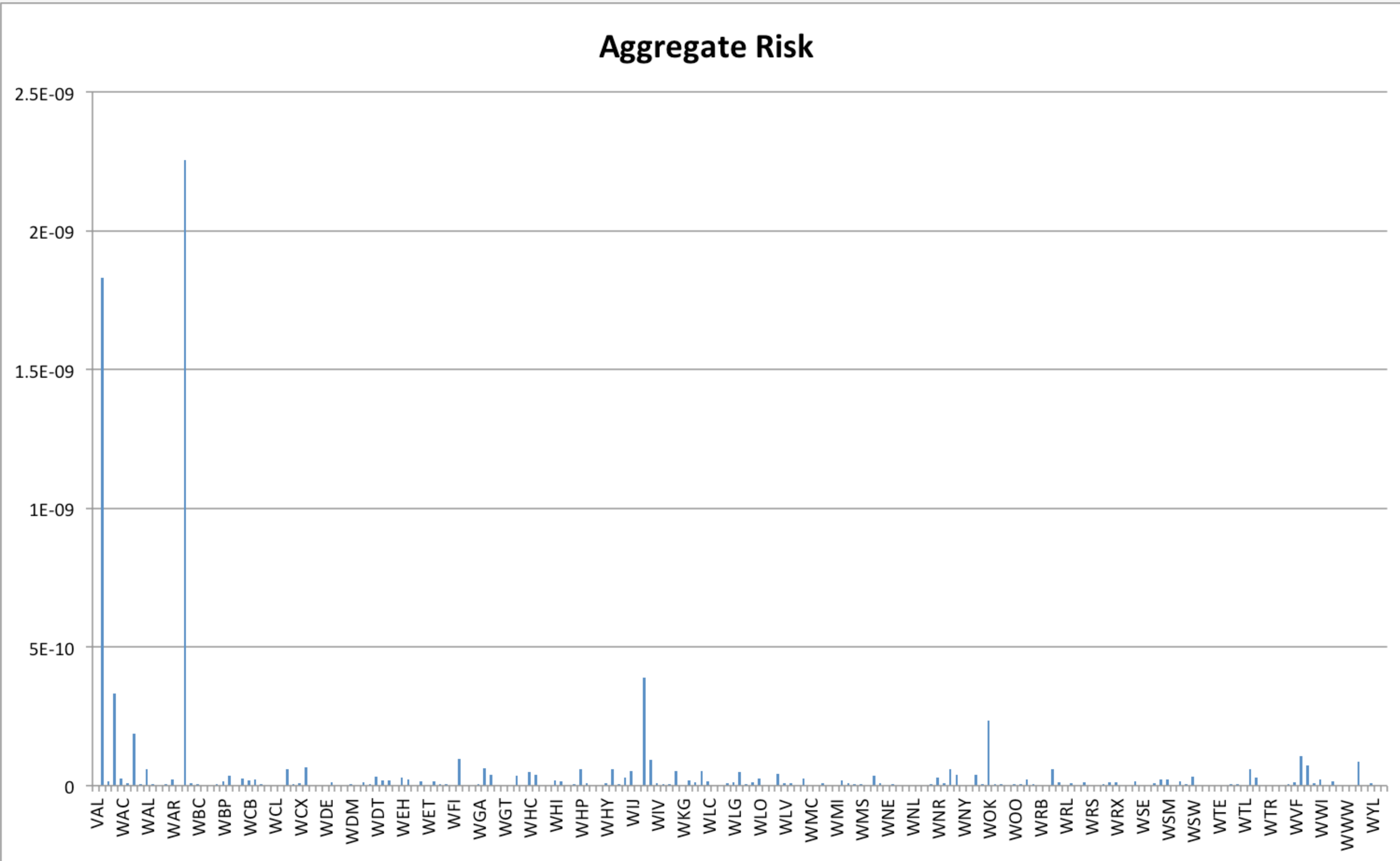
Observed Normalised Risk



Calculated Station Profile: Individual



Station Profile: Aggregate



Case Study: Railway Incidents

- Background and aims
- BN model and data analysis
- Uses of the model
- Further work

Assumptions Made: Event Probabilities

- Calculation steps
  - Priors of causes, from BN
  - Conditional probability of causes, given incident
  - Derive probability of event given causes
  - *Complex!*
- Assumptions
  - Independence assumed
  - Alternatives?
  - How to check?
- Similar assumptions elsewhere

Data Analysis Lessons Learnt

- Need to combine data sources
  - Some data sources are old/static
  - Inconsistent coding e.g. stations
- Expert judgement
  - Needed where data was unavailable e.g. passenger behaviour
- Automation
  - Spreadsheets (MS Excel)
  - Databases not very flexible

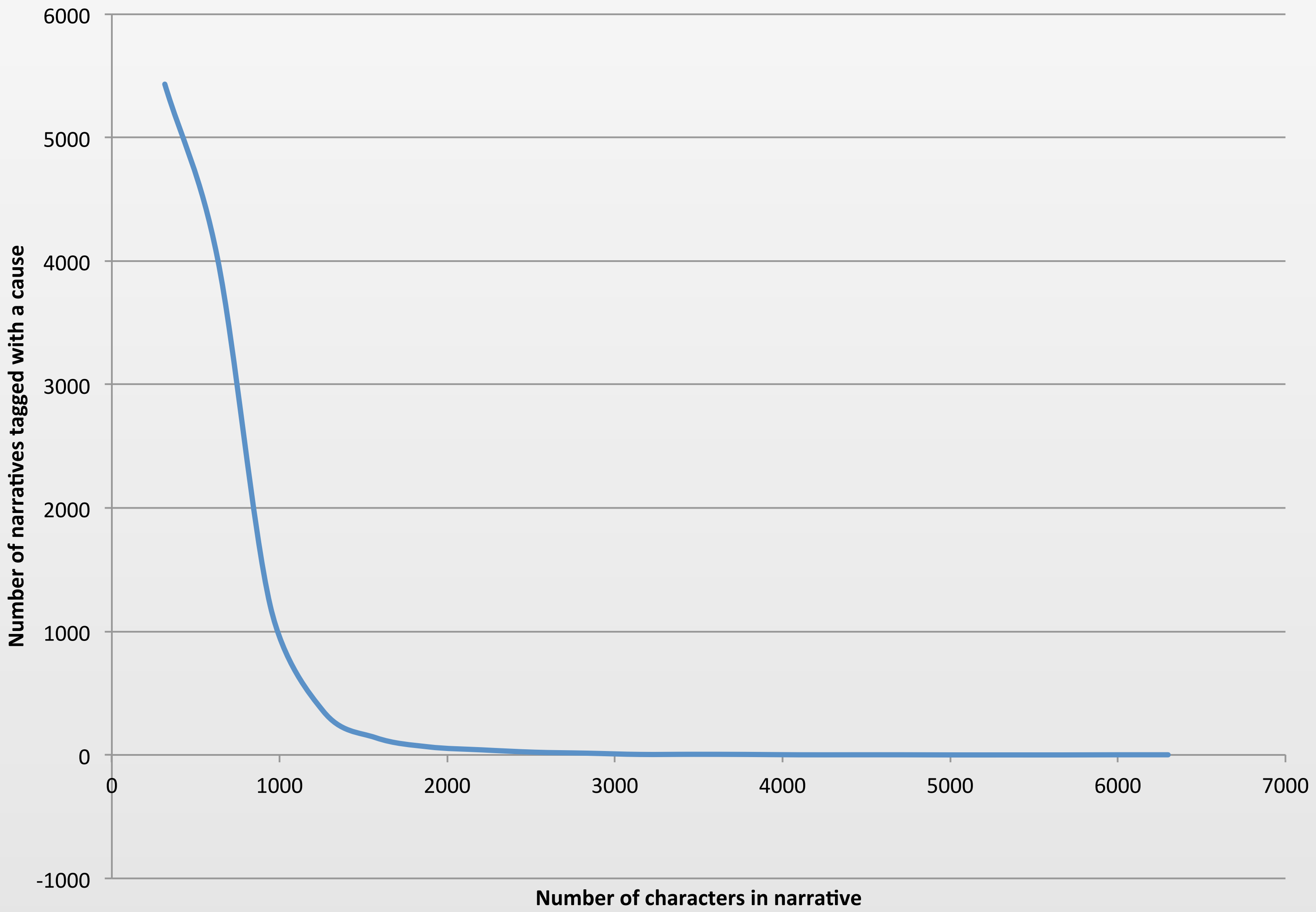
Search Narrative Text for ‘Cause’

- Search used to tag the incidents with causes

```
INJURY_ID|SRM_PRECURSOR_CODE|Adjusted_precursor|EVENT_DATE|TRAIN_CLASS|
    INTOXICATED_IND|APPARENT_AGE_DESC


isIcy  NARR_TEXT \b(snow|ice|icy|freezing|frozen|frost|snowing|slippery|slippy)\b
isNotIcy  NARR_TEXT (\Wnot|\Wno|\wn't).{1,10}\b(snow|ice|icy|freezing|frozen|frost|snowing|slippery|
    slippy)\b
isRush NARR_TEXT \b(run|running|rushing|sprinting|rushed|sprinted|hurrying|hurried|rush|sprint|tip|
    hurry|hustle|late.{1,20}(boarding|aboard|board|boarded)|(boarding|aboard|board|boarded|ran).
    {1,20}late)\b
isWet  NARR_TEXT \b(wet|water|damp|rain|raining)\b
isNotWet  NARR_TEXT (\Wnot|\Wno|\wn't).{1,10}\b(wet|water|damp|rain|raining)\b
isCrowd NARR_TEXT \b(crowd|crowds|crowding|crowded|busy|overcrowded|overcrowding)\b
isGap NARR_TEXT \s(gap|stepping\s*(distance|height)|step\s(up|down)|platform.{1,15}height|height.
    {1,15}platform)
isSlam  NARR_TEXT \b(slam)
isOverhang NARR_TEXT \b(slope|sloped|fully|ramp|stopped\s*short|short\splatform)\b
```

How Good is the Narrative?



How Good is the Detection of Causes?

- Overhang, Door type → Alight No Platform
- Prior

Overhang	Overhang	Overhang	No_overhang	No_overhang
Door type	Slam	Power	Slam	Power
	1.95%	43.34%	2.13%	52.58%

- Incident data

Overhang	Overhang	Overhang	No_overhang	No_overhang
Door type	Slam	Power	Slam	Power
100.0000%	13.78%	71.94%	2.30%	11.99%

Looks reasonable

How Good is the Detection of Causes?

- Stepping distance → Falls between
- Prior

Boarding_or_Alighting	Boarding	Boarding	Alighting	Alighting
Stepping distance	Significant	Not_Significant	Significant	Not_Significant
	33.38%	66.62%	33.38%	66.62%

- Incident data

	Boarding	Alighting
Stepping distance		
Significant	17.88%	17.59%
Not_Significant	82.12%	82.41%

Is this reasonable? Perhaps incident causes incorrect?

Potential Applications?

- Applications of Causal Modelling
- Conclusions

Potential Applications: Safety

- Major disasters preceded by minor failures
- Modelling further back in causal chain

Potential Application: Operations

- Many times of incident other than safety
- Causes of operational incidents
  - Maintenance
  - Staff
- Evaluating changes to maintenance regime

Summary

- Causal model allow effect of changes to be estimated
- Incident data can be used to estimate strength of causes
- … combined with data on the usage
- Bayesian networks flexible
  - Approximations
- Improvements to practicality

Questions?