# Data and Associations

**Jim Warren**
Professor of Health Informatics

# Outline

- Big Data
- Bayes' Theorem and associations
- Looking at associations
- Looking at data over time

# 'Big Data'

- Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**

# Some domains

- Some domains swimming in Big Data
  - Astronomy
    - SKA will generate a few Exabytes per day and 300-1500 Petabytes of data per year to be stored
  - Weather and climate modelling
  - Biomedicine
    - Genomics, proteomics, metabolomics (-omics)
  - Healthcare delivery
  - Retail and marketing
  - Finance and economic modelling

# Bayes Theorem

- Associations affect our expectations
- This can be quantified with conditional probability
  - Consider the probability, P, of a diagnosis, Dx, being valid, given a patient exhibiting a symptom, Sy:
    - P(Dx|Sy)= [P(Sy|Dx) × P(Dx)] / P(Sy)
    - *Posterior* probability can be quite different than the *a priori* P(Dx)

**Bayes' Theorem**

  - So we might have P(flu)=0.05, P(fever)=0.04
    - With P(fever given flu)=0.5,
      P(flu given fever) = [(0.5)(0.05)]/(0.04) = 62.5%

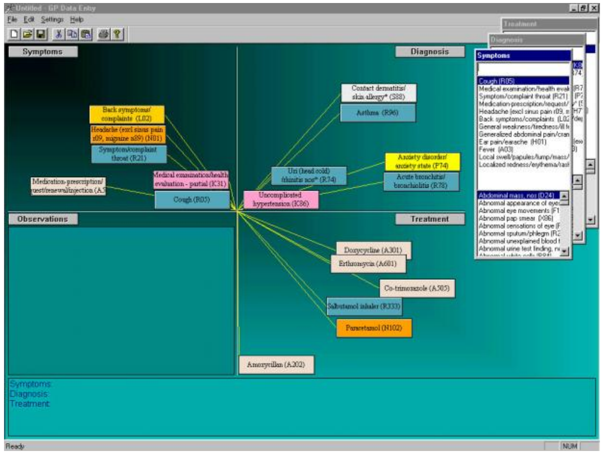
# Using conditional probability

- Conditional probability is very context dependent
  - Won't be the same in Poland as South Africa, or in winter as summer
- Can learn from data the number to apply Bayes Theorem
  - Count number of flu cases and number of patients with fever symptoms
    - Divide by total for P(Dx) and P(Sy), *aka* 'prevalence' of each
  - Count number of cases with flu *and* fever
    - Divide by number of cases with flu to get P(Sy | Dx)
- But your estimation is only as good as your data
  - Did fever always get recorded? Was every flu recorded *and* correctly diagnosed?
  - And you have to assume the new context is similar to the one where you 'learned' (estimated) the parameters

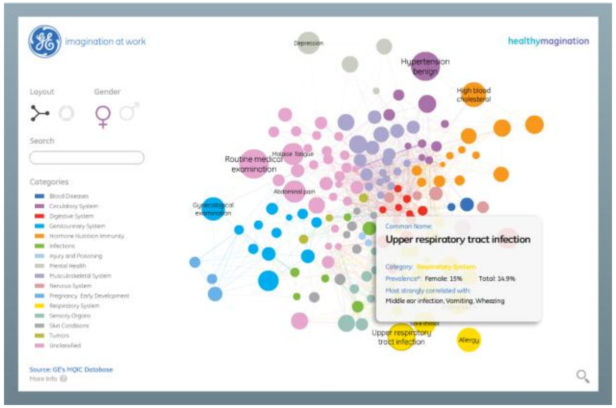
# Probability in user interaction

- Can use *a priori* prevalence and posterior probability as basis for layout decisions
  - E.g. intelligent split menu: offer most likely item selections at top
  - MS Word does a heuristic split menu with a few common and/or recently used fonts at top
  - Can estimate contextually-likely actions for right-click options, or to offer help topics
- I developed *Mediface* a few years ago
  - Used General Practice electronic medical records to estimate prevalence and conditional probabilities on diagnoses, symptoms and treatments

# *Mediface*



## GE / MIT unlocking big data

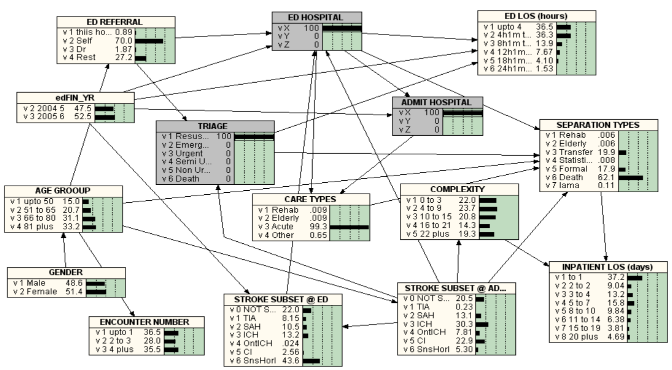



- http://www.gereports.com/the-magic-of-big-data-ge-mit-unveil-new-way-of-visualizing-disease/
- http://visualization.geblogs.com/visualization/network/

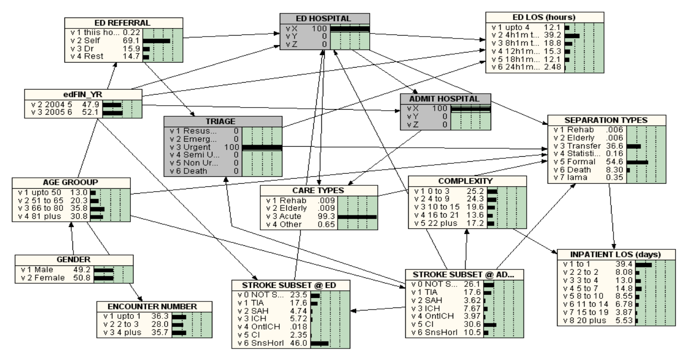
## Working with Bayesian Networks

- You can visualise a series of Bayes Theorem based associations
  - Tools like Nettica will learn these from data and give you a GUI to explore the data
    - You can provide some initial network structure (hypothesized associations) or let it guess (but it might get causality the wrong way around)
- E.g. we looked at Victorian (i.e. Melbourne area) hospital discharges for patients admitted to emergency departments (ED) with stroke
  - [next 2 slides]: note comparison of 'death' discharge/separation outcome for cases with priority of 'resus' (needing to be resuscitated) versus merely 'semi-urgent' at hospital 'X'
    - 62.1% versus 8.3% death rather than other separation code
    - Also note different input distribution of stroke type – about 4 times as many Intracerebral hemorrhage (ICH) in the Resus cases; and very different ED LOS (length of stay) distribution

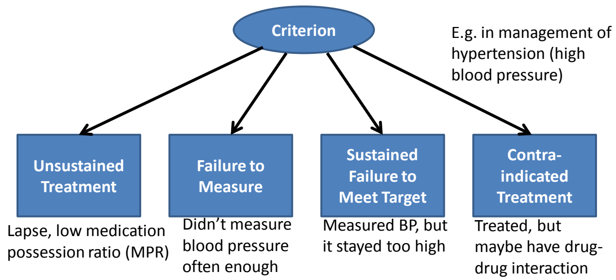
## Triaged as 'Resuscitation required'




## Triaged as 'Semi-urgent'

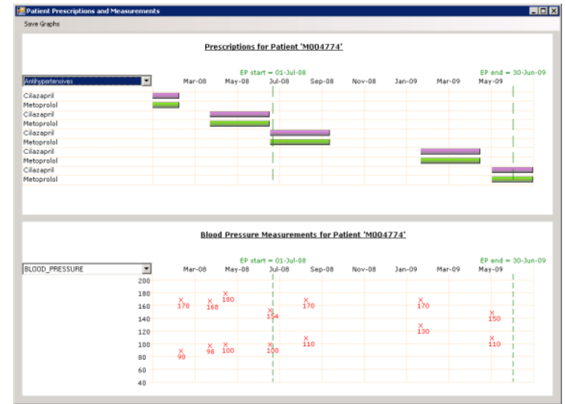


## ChronoMedIt: Assessing suboptimal long-term condition management

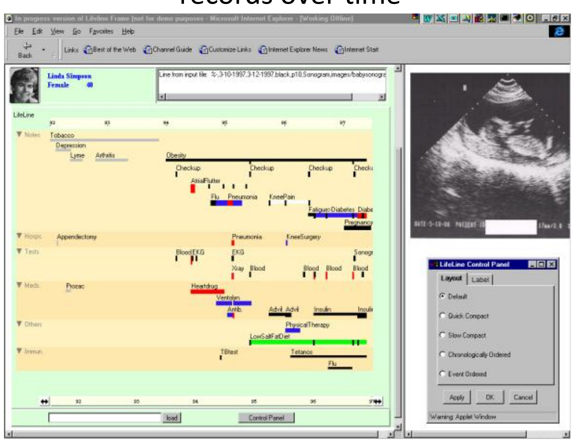



- Model of criteria for long-term treatment
  - Use an ontology (in Protégé/OWL) to hold parameters of treatments, problems and measurements

## Example visual presentation of a case with low Medication Possession Ratio (MPR)

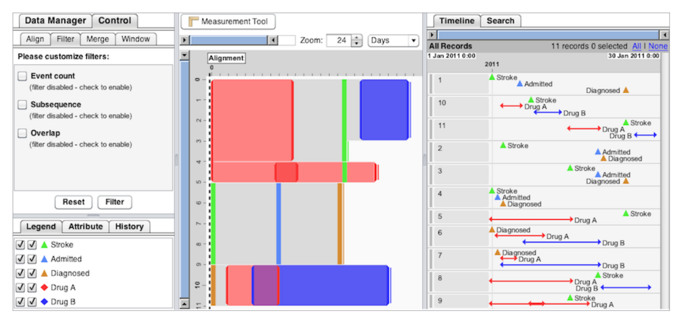



## LifeLines (2nd half of 1990's): visualising patient records over time

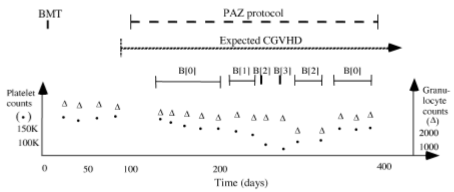


## EventFlow

- Exploring Point and Interval Event Temporal Patterns over multiple patients

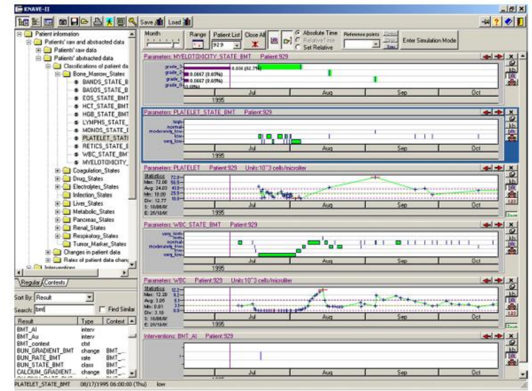


## Temporal abstraction

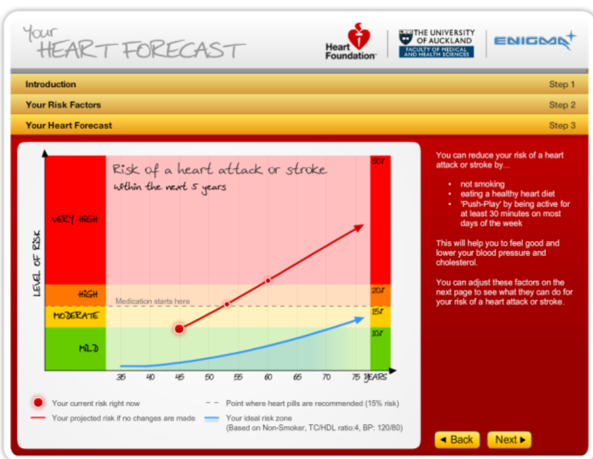



- Process individual data points to infer semantics on time intervals
  - E.g. levels of bone marrow toxicity (B(x)) following a Bone Marrow Transplant (BMT) as computed on a time series of platelet count and granulocyte count measures over the duration of a treatment protocol (PAZ) for graft rejection (chronic graft versus host disease, CGVHD)

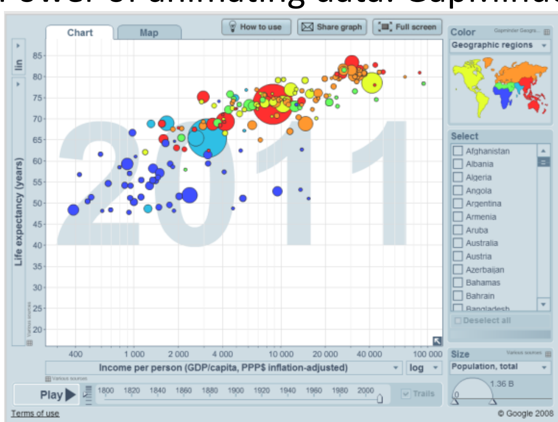
## KNAVE-II: interface to distributed knowledge-based interpretation and summarisation



## Prediction over time with option for 'what if'




## Power of animating data: GapMinder




http://www.gapminder.org/ http://www.ted.com/talks/hans_rosling_at_state.html

## 3D/VR renderings

- Visible Human project involved CT, MR and cryosection images of representative recently deceased individuals
  - Can be rendered as 3D models
  - Can be navigated for medical education as alternative (or in addition to) using real cadavers

## Conclusion

- The world is increasingly ‘drowning’ in data
  - Well, not ‘drowning’ – but at least there’s a lot of missed opportunity from data not being reviewed
- Interactive visualisation lets us filter, do ‘what-if?’ scenarios and review slices of time
  - Animations and 3D reconstructions give us dimensional (time-space) experience of data
- Statistical models can add inference to the raw data
  - Putting semantic labels on time intervals and adding predictions