JACOBS UNIVERSITY

Statistical Graphics in Data Science

Adalbert F.X. Wilhelm

SDSS 2018,
Honoring Dr. Edward J Wegman
May 18th, 2018
Reston, VA

Quote of the day

"Current facilities for computing, display, and real time interaction have developed substantially beyond our understanding of how to use them effectively in data analysis.“

Tukey and Wilks, 1966

Outline

1. Data Science
2. Statistical Graphics
3. Analysis pipeline
4. Visual representatives
5. Challenges
6. Conclusion

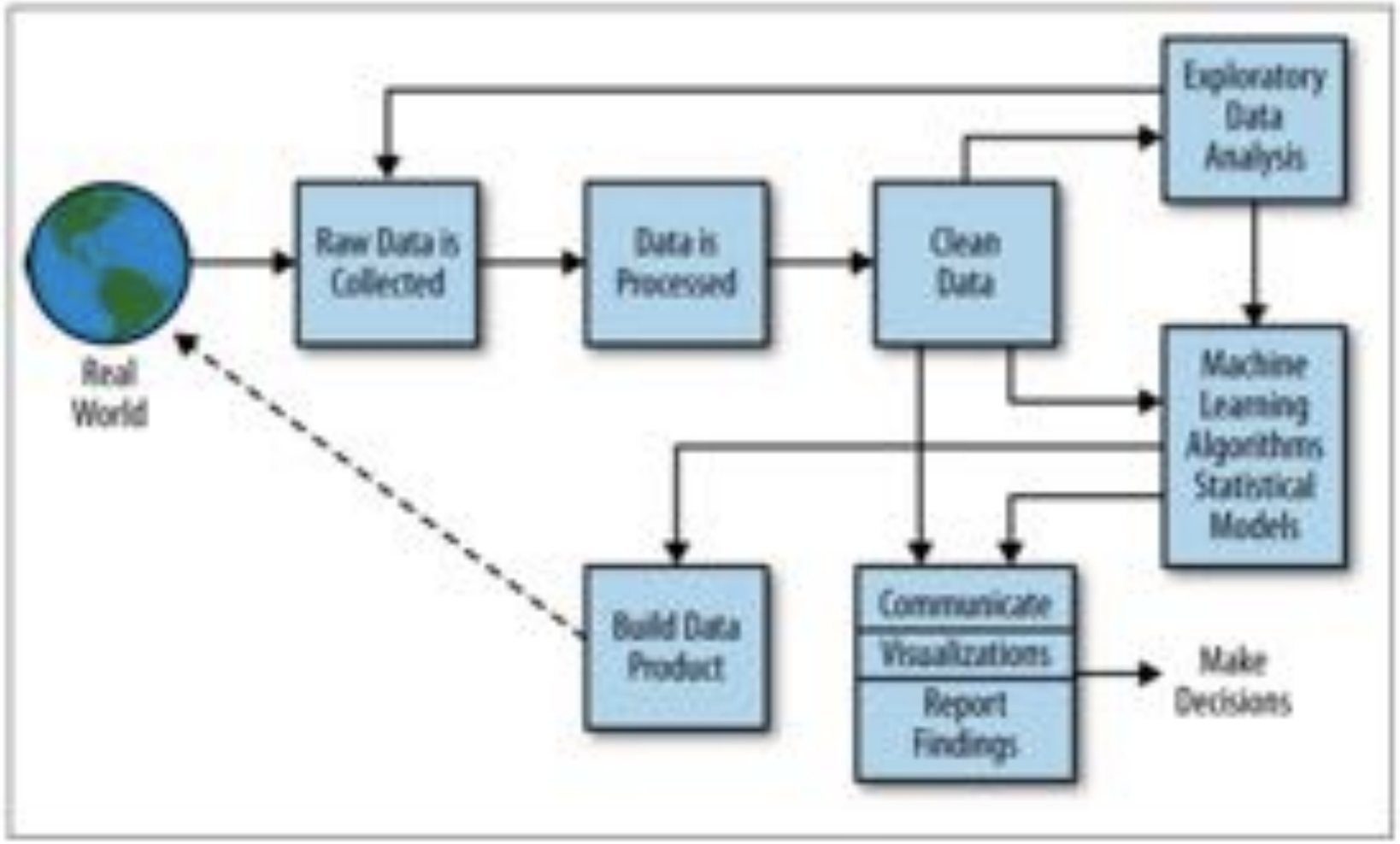
The Data Era: Data = The New Oil

- Production processes create a large amount of data from sensors, logistics, business operations and more
- Rise of cost-effective data collection gives established industries a new boost
- Producing value from data is a challenge and an opportunity
- The promise of data as the new oil is realized when we can tap into its value in a meaningful, cross-functional way to enhance decision-making which provides the competitive advantage
- Competitive advantage: lower costs, higher quality

JACOBS
UNIVERSITY

What do we do with all this data?

What is Data Science?



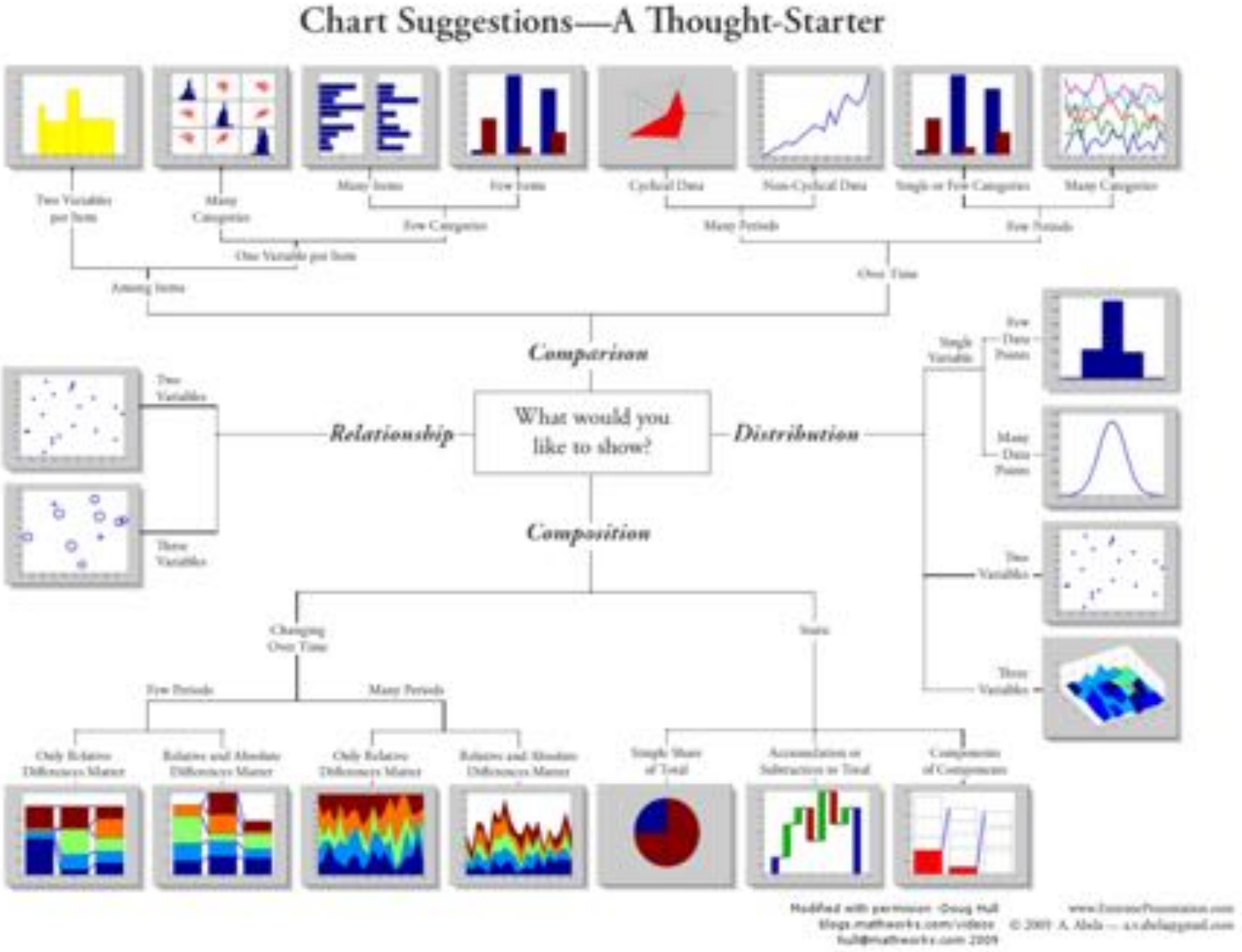
© Cathy O'Neill & Rachel Shutt: Doing Data Science

What is Data Science?

	Goals and Means	Data selection	Data Analysis	Business decision
Purpose	Decide on objectives, identify business levers, define key measures	Provide appropriate data for the given objective	Using and adapting best suitable analysis tools	Improving business processes. Generate value
Central questions	What do we need to improve? What can we change? How can we measure change?	Which data is already available? How to merge and connect data? Can we collect additional valuable data?	How to clean data? Which modeling technique to use? Can we assess reliability of our resulting predictions? How robust are our results?	How to implement results? Can we automatise this analysis? Which changes are most relevant?
Potentials/ Benchmarks	Include information of context experts. Give data science team clear objectives!	Create a data inventory. Increase data awareness among employees to increase data quality.	What are state-of-the-art algorithms? Cross-check with current expertise within company	Enhanced understanding of processes. Improved data culture.

JACOBS UNIVERSITY

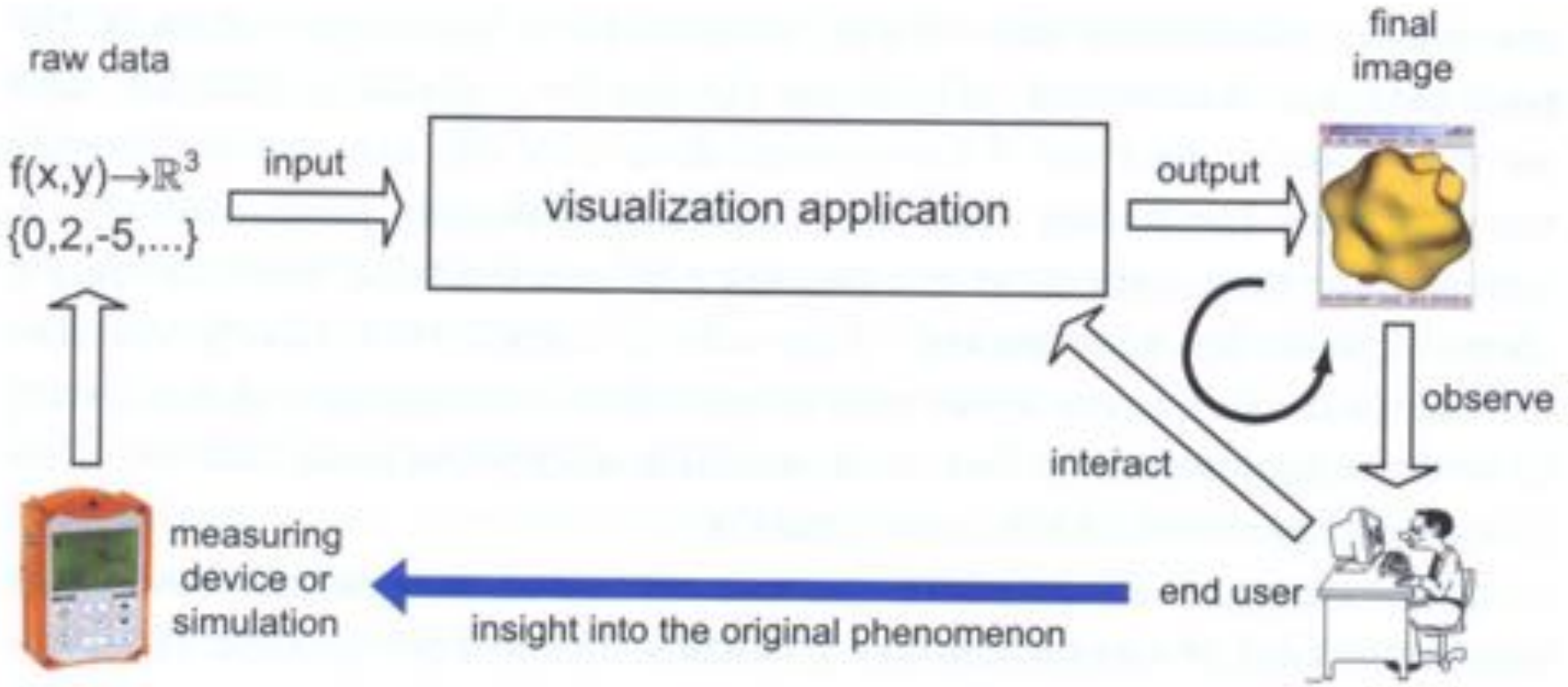
What are statistical graphics?



https://apandre.wordpress.com/dataviews/choiceofchart/

JACOBS UNIVERSITY

The visualization process



Graphics system

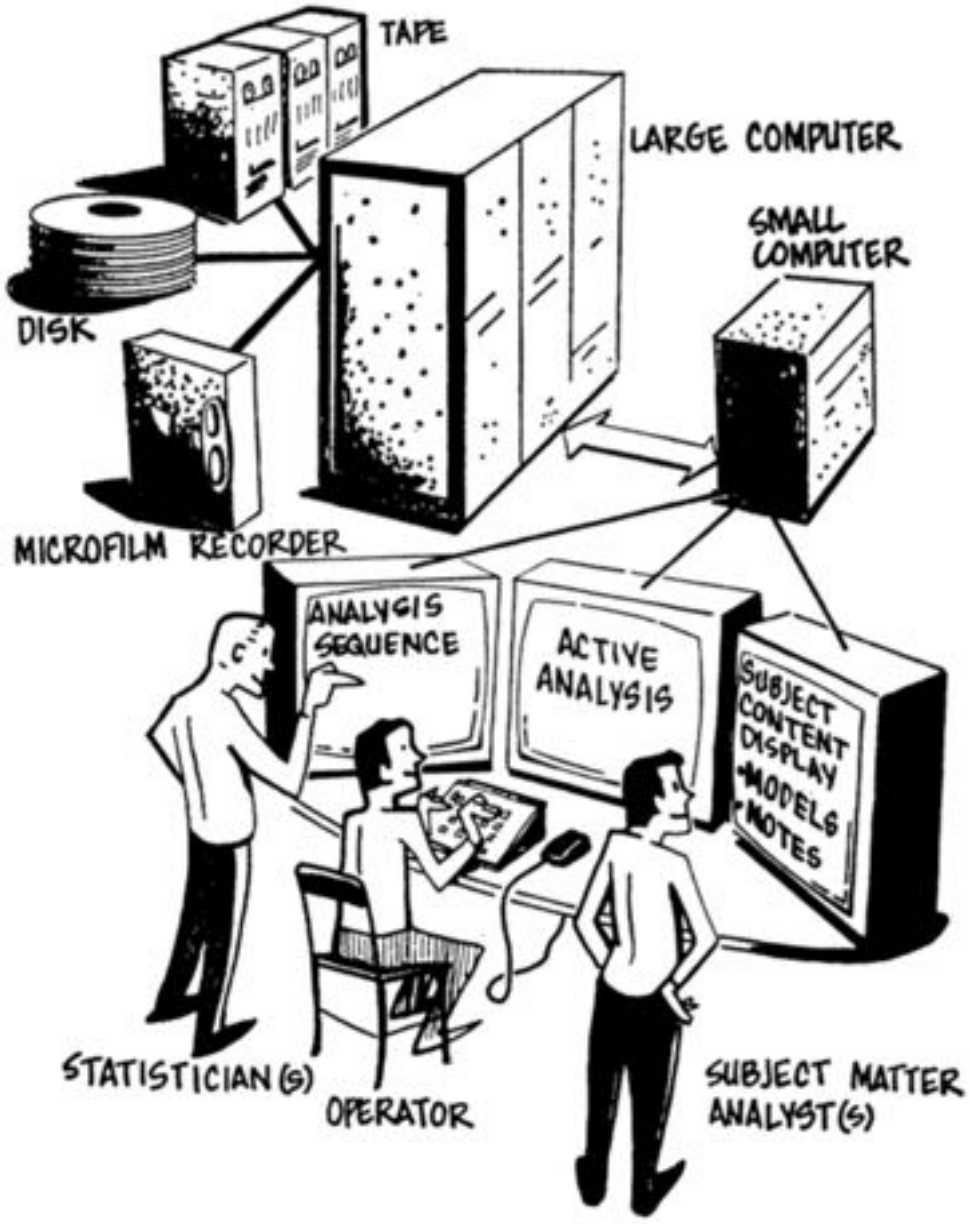


FIGURE 10
Future Interactive Graphic Data Analysis Facility.

Special thanks to
Wayne Oldford

The Data Team

Data Scientist

Main responsibilities:

- Data munging
- Modeling
- Machine Learning
- Reporting and Presenting

Data Science Manager

Main responsibilities:

- Group morale and support
- Business development
- Research & Development

Data Analyst

Main responsibilities:

- Acquiring data
- Developing and implementing data analysis
- Interpreting data
- Analysing results

Data Engineer

Main responsibilities:

- Data ingesting
- Data architecture
- Data formatting

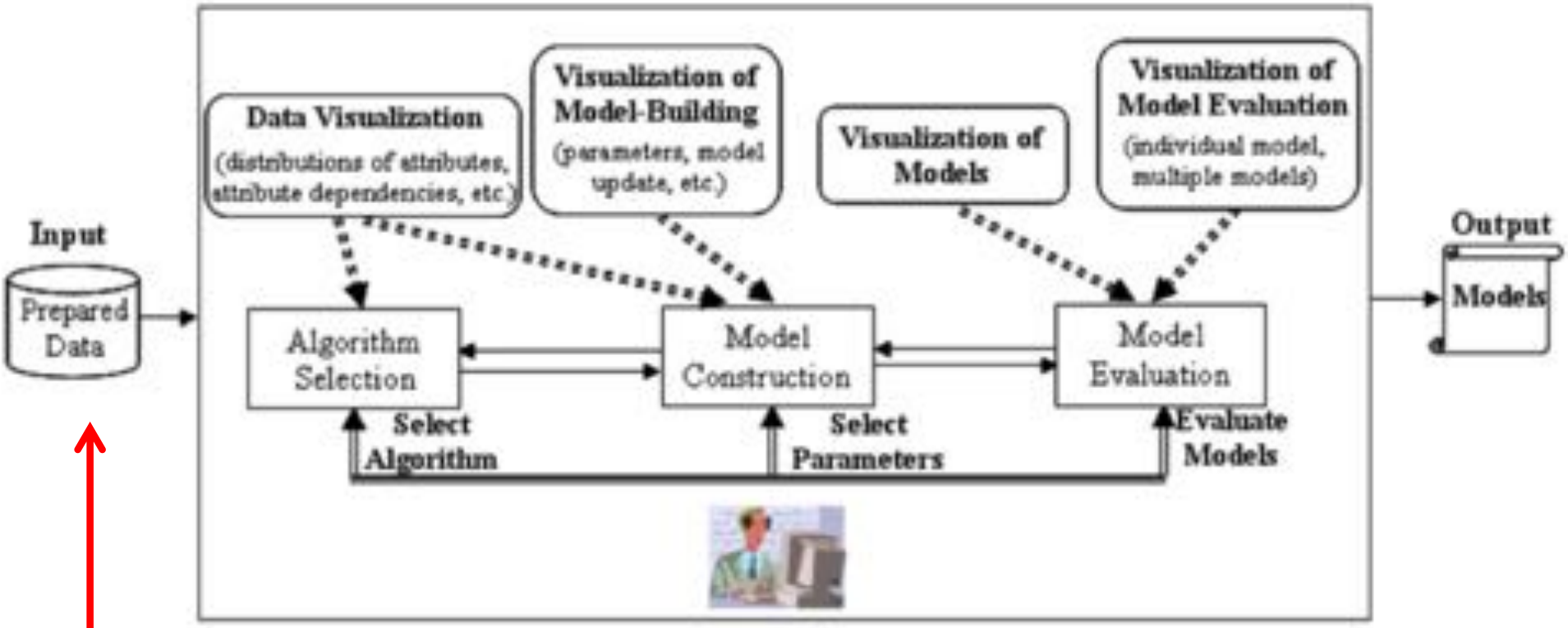
Data Visualizer

Main responsibilities:

- Dashboard creation
- Story telling / effective communication
- Programmatic visualisation

JACOBS UNIVERSITY

Conceptual model of visualisation support (Liu & Salvendy, 2007)



Graphics for data preparation

Data exploration

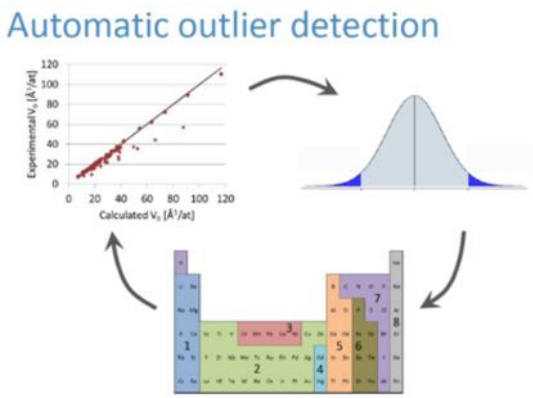
Aspects

bias → noise → abnormality → meaningful → relevant → uptodate

Data cleaning

Plausibility checks

Outlier detection



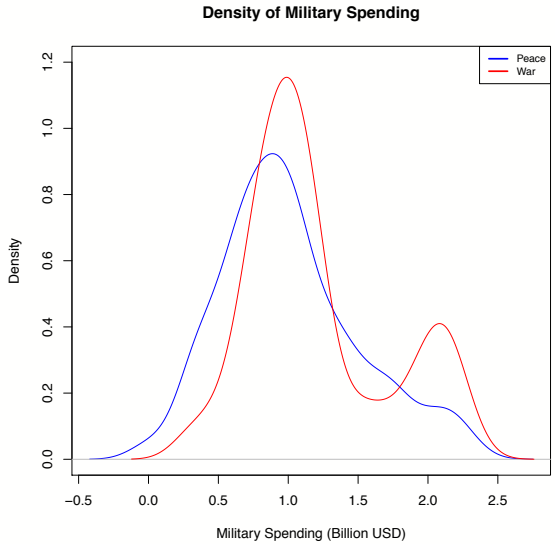
Data understandability

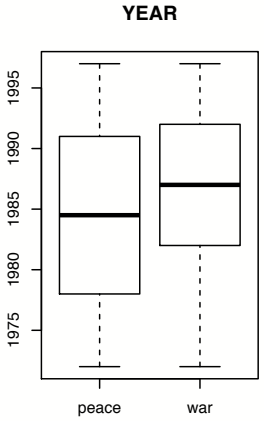
Junk in – Junk out

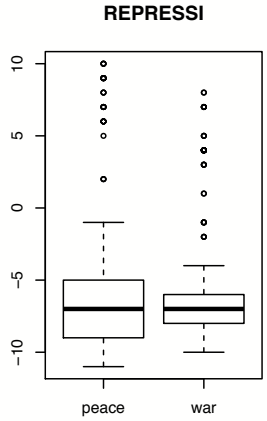
Efficient time synchronization required

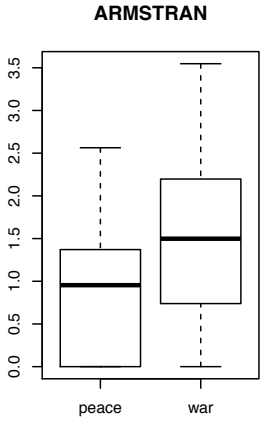
JACOBS UNIVERSITY

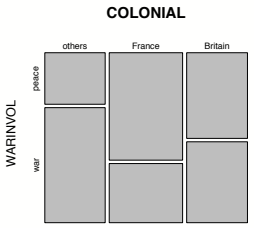
Exploratory Visualisation Prior to Modelling

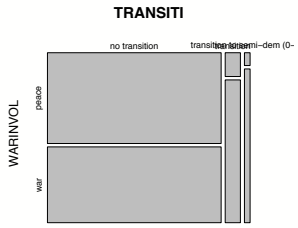


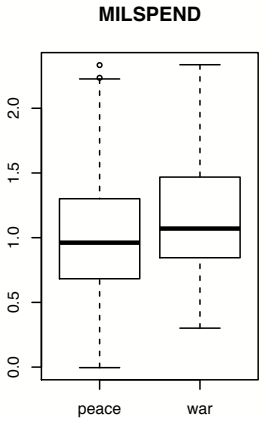


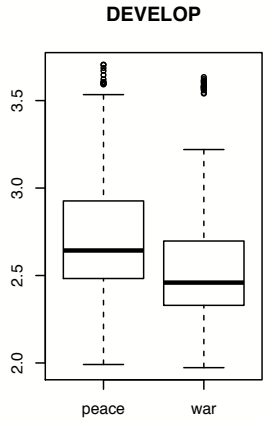


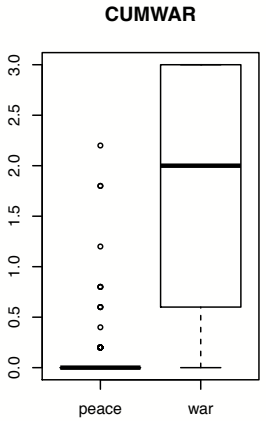


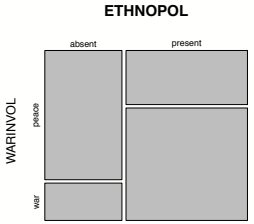


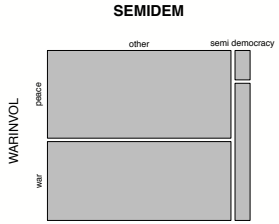






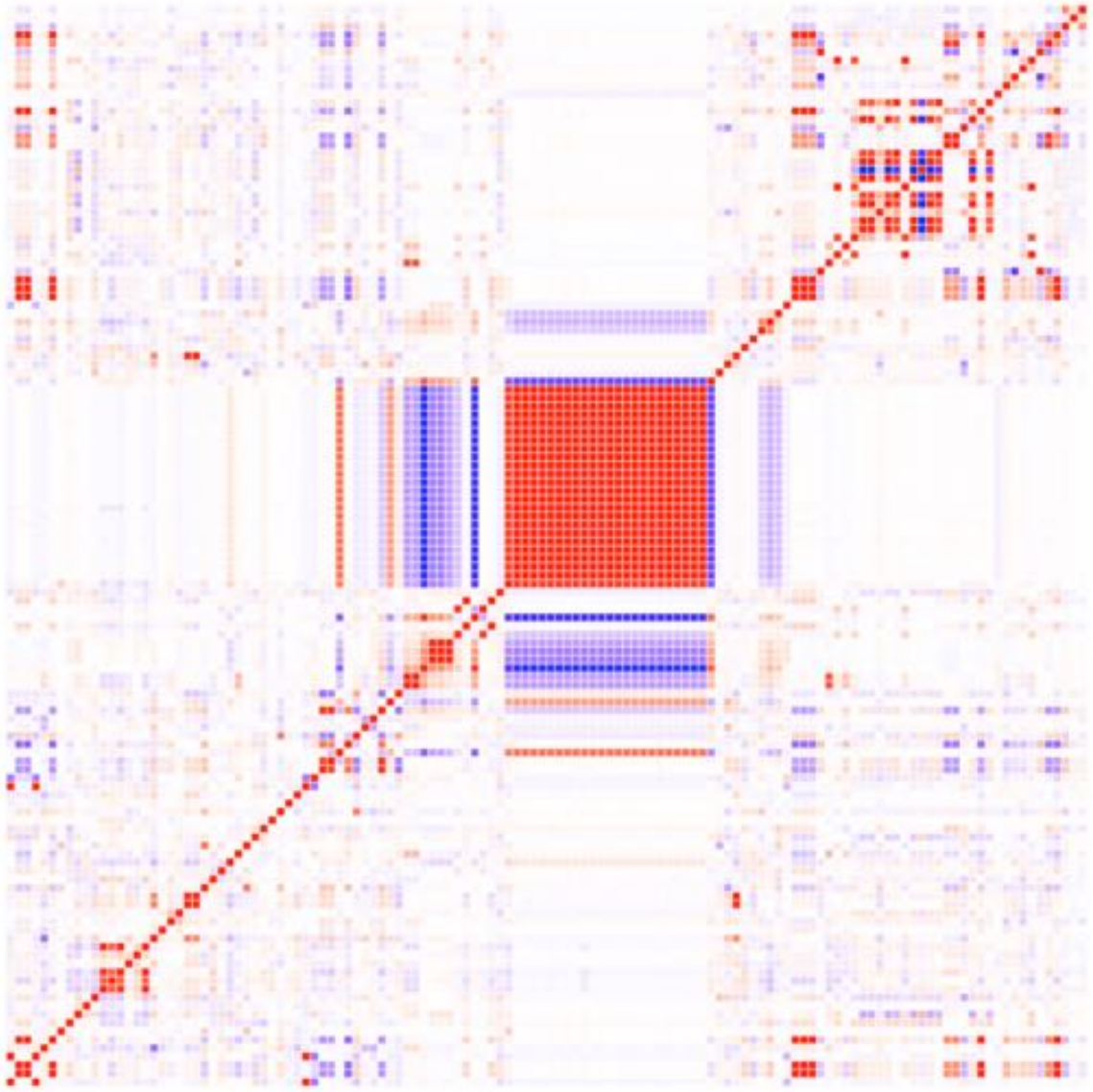






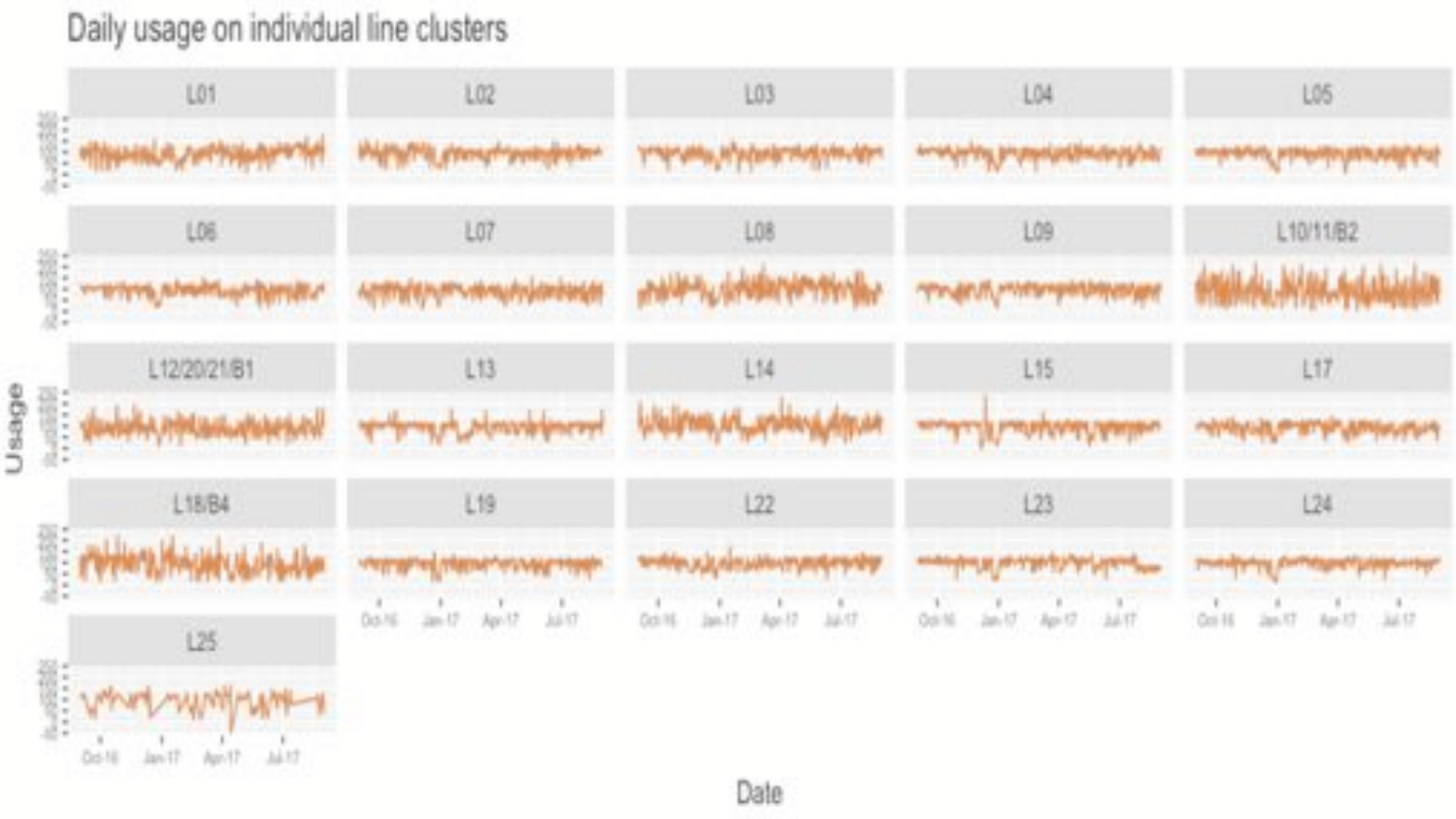
JACOBS UNIVERSITY

Exploratory Visualisation Prior to Modelling



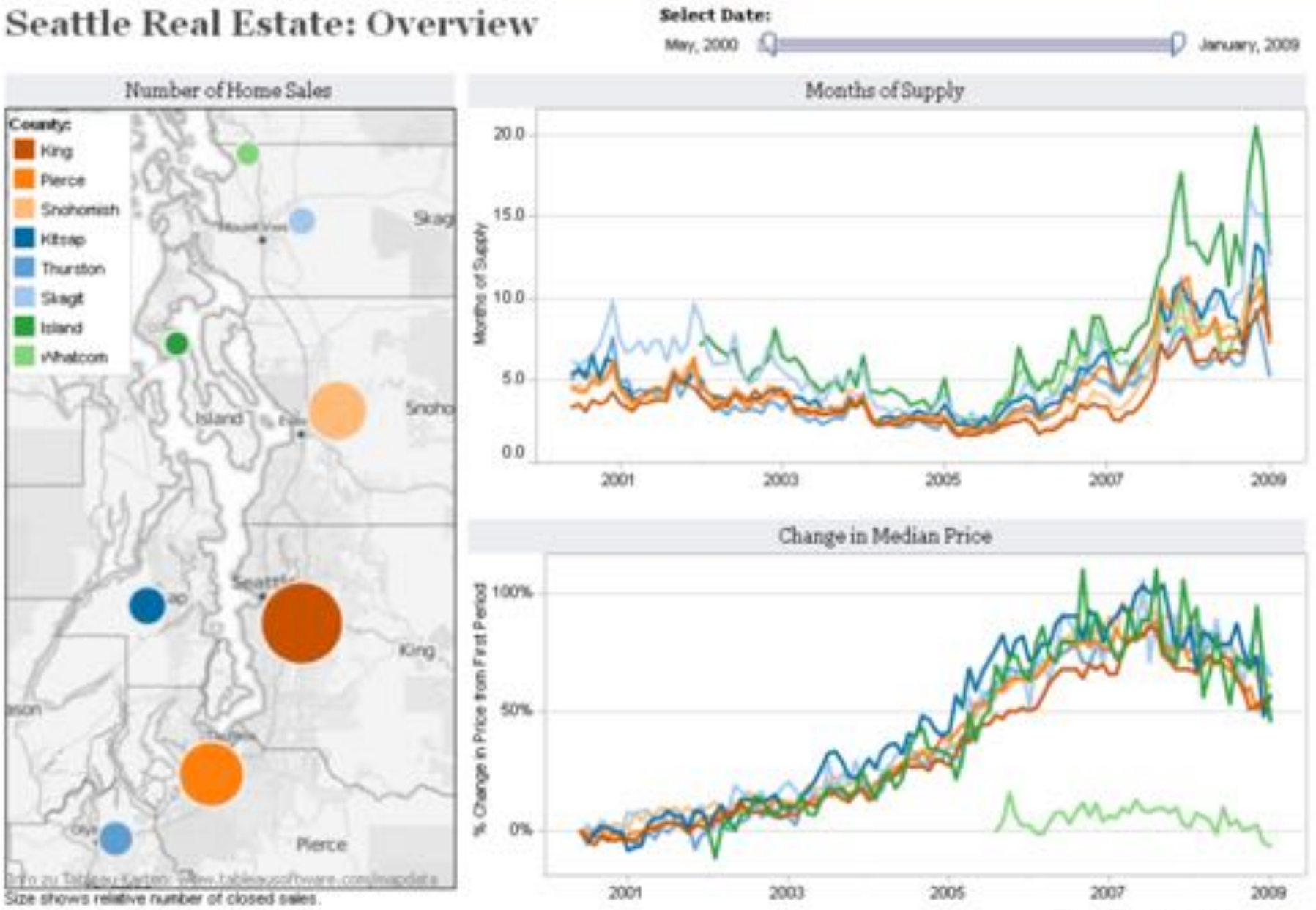
JACOBS UNIVERSITY

Exploratory Visualisation Prior to Modelling



JACOBS UNIVERSITY

Exploratory Visualisation: Linked views

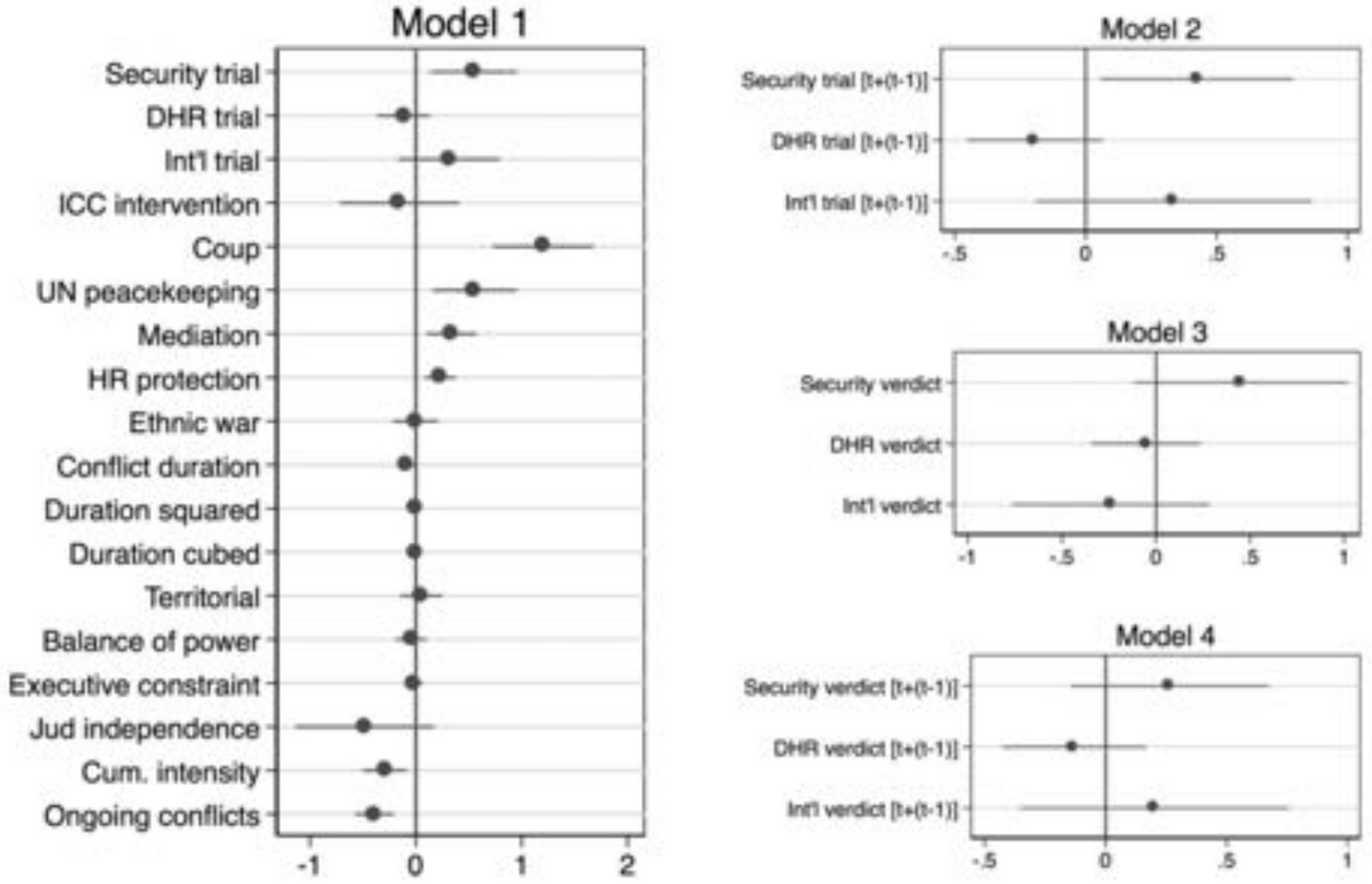


Visualisation for modeling support

- Depends on purpose
  - Prediction
  - Explanation
  - Pattern recognition

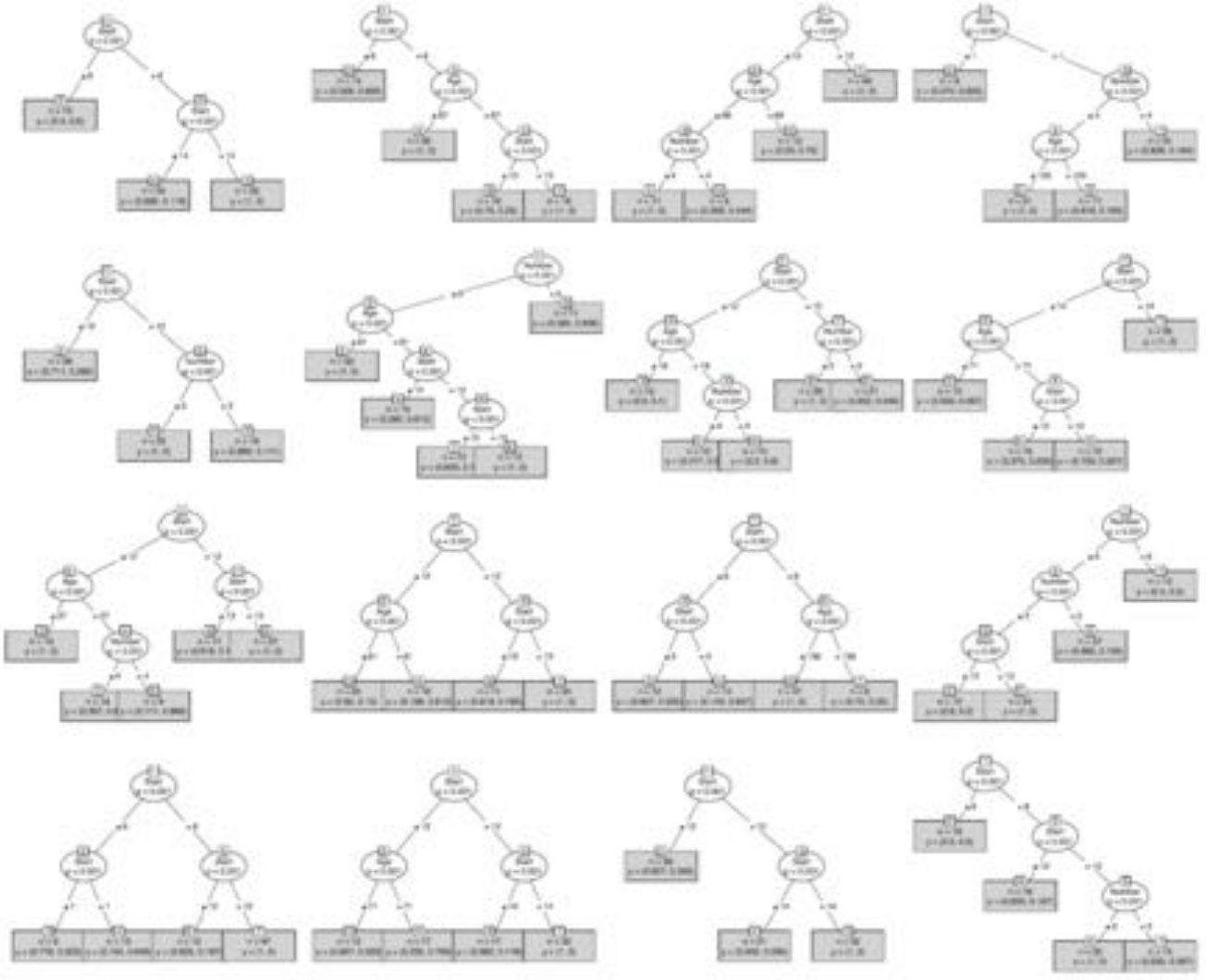
- Depends on data type and structure
- Depends on modeling approach
- Depends on audience and their standards
- Depends on software ecosystem

Visualizing model results: model coefficients & CI



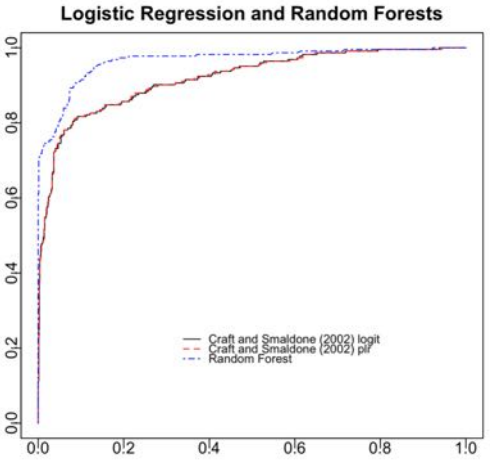
Let‘s practice what we preach, Gelman et al. (2002)

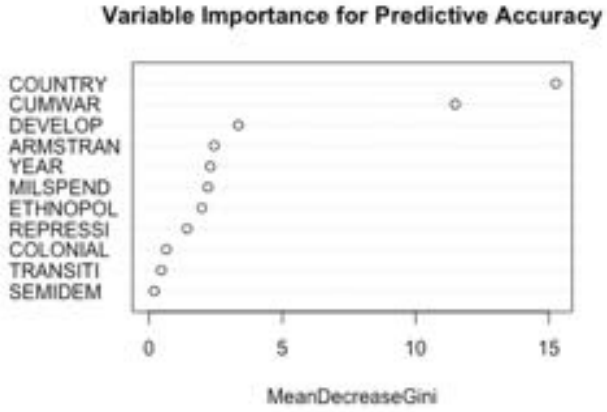
Visualizing model results: e.g trees, random forests

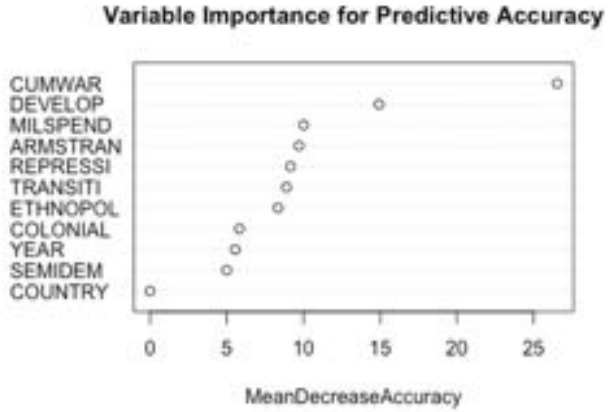


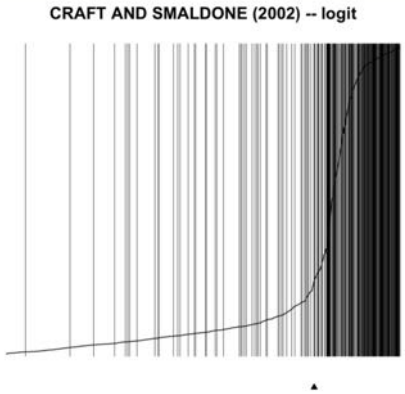
JACOBS UNIVERSITY

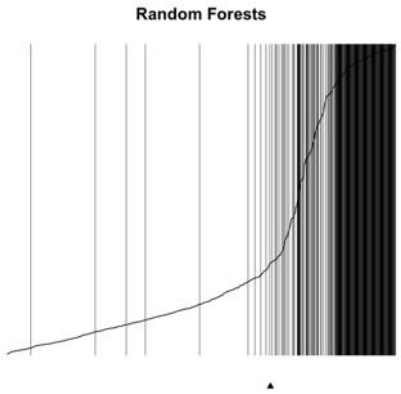
Visualizing model results: ROC, Importance plots, separation plots





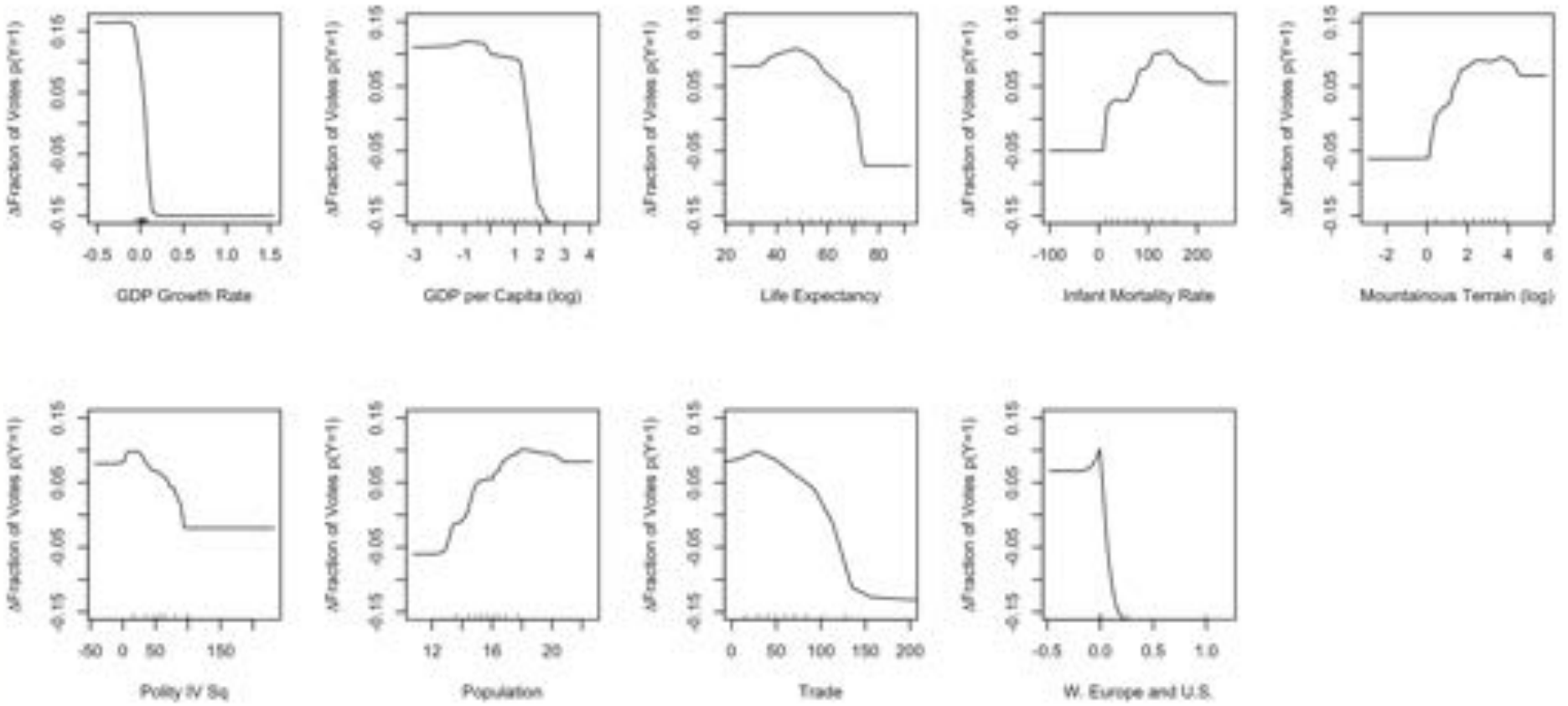






JACOBS UNIVERSITY

Visualizing model results: partial dependency plots

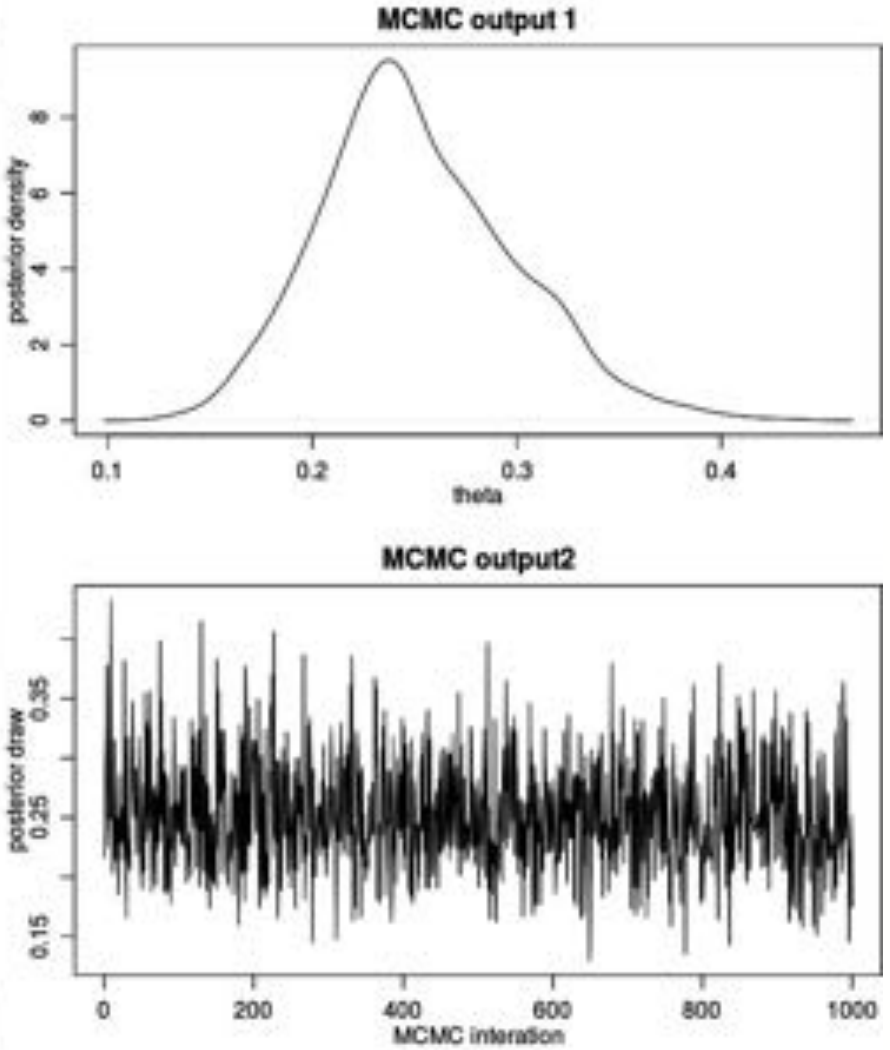


JACOBS UNIVERSITY

Visualizing model results: MCMC and Bayesian models

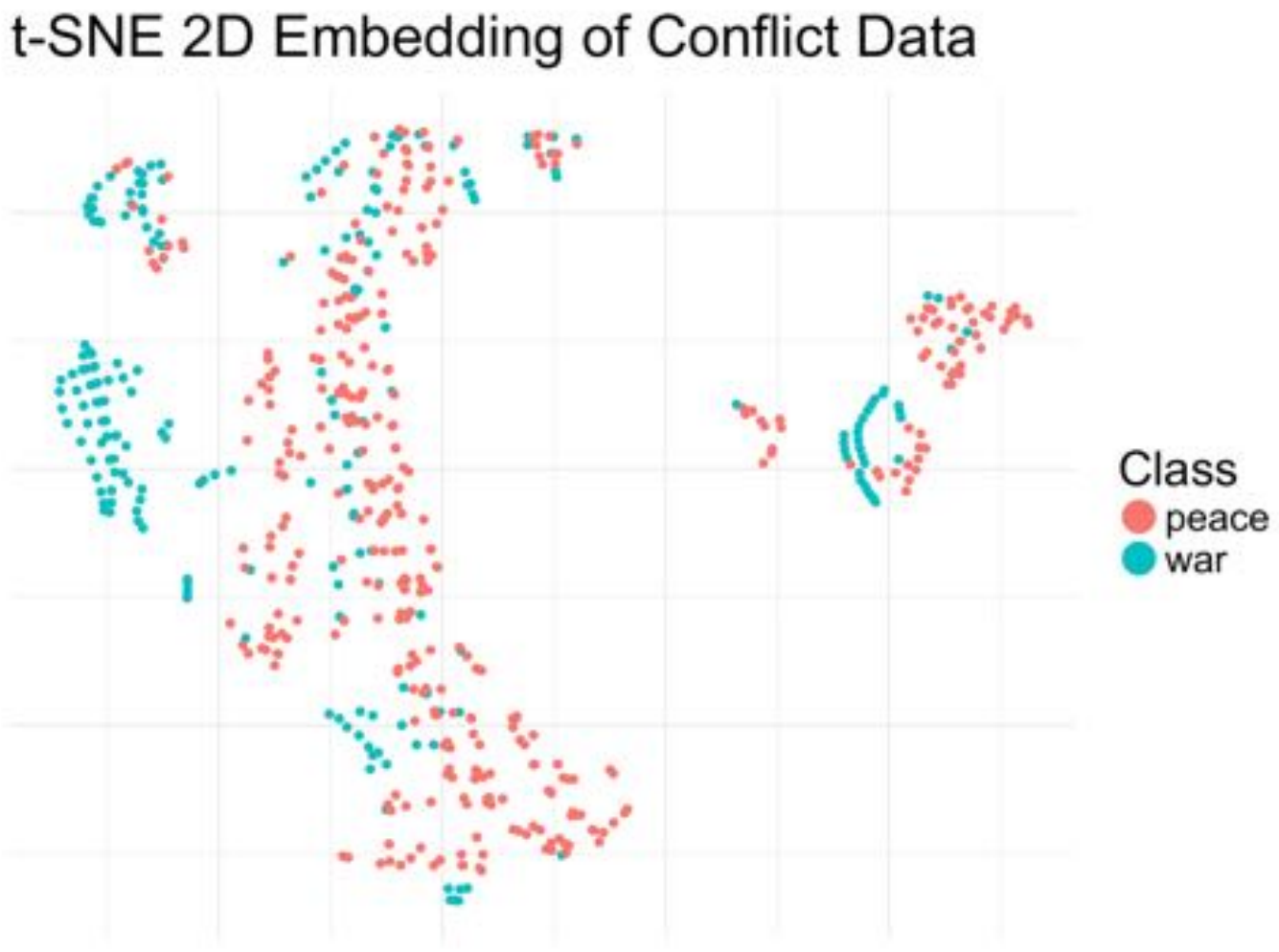
```
# RUN JAGS MODEL AND SAMPLE FROM POSTERIOR
  m3 <- jags.model(file=modname3,data=jags.data,ini
# burn-in phase
  update(m3,nburn) # discard nburn iterations (ensu
# which variables to summarize
  variable.names <- c("sigma.phi","sigma.g1","sigma
# sample from posterior
  samp <- coda.samples(m3,variable.names=variable.n

  summary(samp) # summarize output
  plot(samp, ask =TRUE) # inspect chains for adequa
  gelman.diag(samp) # inspect chains for convergenc
# DONE PART 1
```

```
> plot(density(sigma2),main="MCMC output 1",xlab="sigma",ylab="posteri
or density")
> plot(sigma2,type="l",main="MCMC output2",ylab="posterior draw",xlab=
"MCMC interation")
> par(mfrow=c(2,1),mar=c(3.5,3.5,2.0),mgp=c(1.7,0.8,0))
> plot(density(sigma2),main="MCMC output 1",xlab="theta",ylab="posteri
or density")
> plot(sigma2,type="l",main="MCMC output2",ylab="posterior draw",xlab=
"MCMC interation")
>
```



JACOBS UNIVERSITY

Exploratory Visualisation after modeling

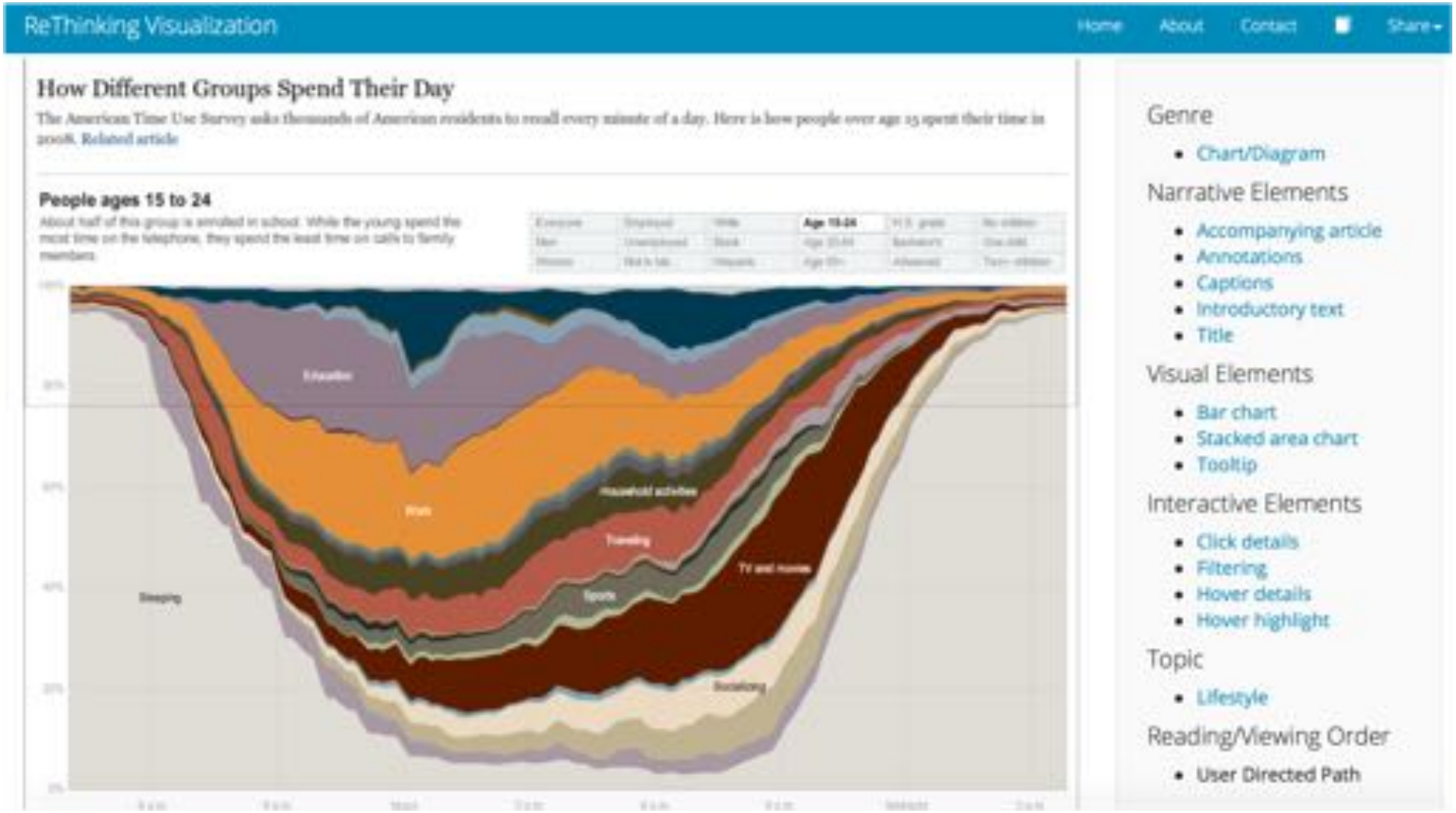


JACOBS UNIVERSITY

Continuing challenges:

- Overcoming 2D/3D – limitation
  - Dynamic plots
  - Conditioning (C.Hurley: condvis)
  - Linked views (W. Oldford: loon, H. Hofmann: cranvas, altair)
- Resolution (e.g. imbalanced data)
- Reproduction (J. Harner) and Automation (P. Hall: H2O, L. Wilkinson: next session)
- Transporting to future computational ecosystems
- Interpretation
- Storytelling
- Data Literacy and visual literacy

Storytelling



JACOBS UNIVERSITY

The Data Team

Data Scientist

Main responsibilities:

- Data munging
- Modeling
- Machine Learning
- Reporting and Presenting

Data Science Manager

Main responsibilities:

- Group morale and support
- Business development
- Research & Development

Data Analyst

Main responsibilities:

- Acquiring data
- Developing and implementing data analysis
- Interpreting data
- Analysing results

Data Engineer

Main responsibilities:

- Data ingesting
- Data architecture
- Data formatting

Data Visualizer

Main responsibilities:

- Dashboard creation
- Story telling / effective communication
- Programmatic visualisation

JACOBS UNIVERSITY

Conclusion:

- Many ingredients have been around since long
- Power of statistical graphics is widely acknowledged
- Visualisation throughout the data analysis process
- Interactive graphics in R!
- Visual communication and storytelling
- Liars know how to figure
- Visual literacy
- Interpretation
- Increasing number of specialised graphics
- Reproducibility and Automation

Thank you very much for your attention!

Questions?

Comments?