Emerging Applications in Industry

Slides at *www.lancs.ac.uk/∼jonathan*

Philip Jonathan
Statistics and Data Science

Acknowledgement

- Shell Statistics and Data Science
- Shell colleagues and clients
- Lancaster
- Delft, Durham, Glasgow, Imperial, UCL

Overview

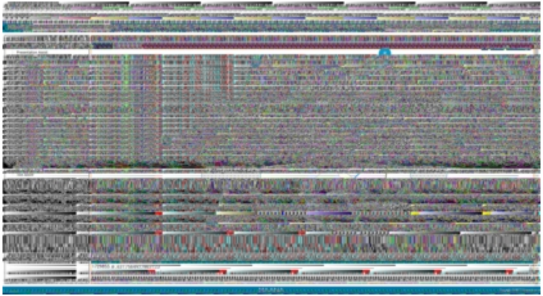
- Context
- Measurement; Connectivity and streaming; Data science
- Modelling the physical environment
- Opportunities

Context: Shell's Statistics and Data Science group

- $\approx$ 20 statisticians, modellers, chemometricians and data scientists
- Based in the Netherlands, UK and USA
- World-wide client base within Shell
- Upstream: Seismic hazards; Acoustic and remote sensing; Extreme value analysis
- Downstream: Manufacturing support; New chemicals, fuels, lubricants; Retail; Inventory management
- Corporate: Economic modelling; Safety; HR; Wind power
- Training: Introductory; Design of experiments; Visualisation; Machine learning
- Academic: R&D; Maintaining expertise base; Recruitment

- In a rapidly growing "analytics" community

Context: What's changing for us?



MAANA "Knowledge Graph" interface, credit maana.io, "turns human expertise and data into digital knowledge for employees to make better and faster decisions"

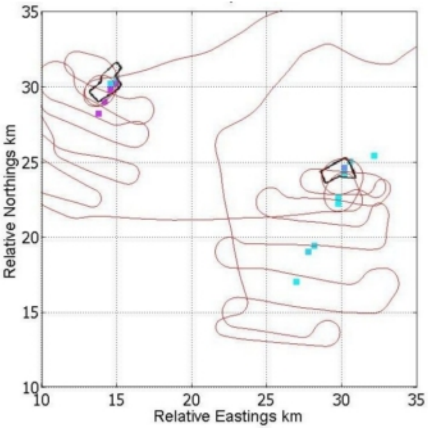
More accessible data ... "digitalisation"

- $n_{2017} >> n_{1980}$, $p_{2017} >> p_{1980}$
- Streaming
- Connected data sources
- Text, images, sound, speech

Better computing and storage

- Parallelism: multiple cores, cheap memory; Cloud
- Freeware: R, PYTHON, C, JAVA taking over from SAS, MATLAB
- Graphical interfaces e.g. SHINY R
- Alteryx, Apache Spark, SQL, NoSQL, ...

Context: What's changing for us?



Probabilistic inversion

More Bayesian

- Awareness, acceptance, interpretation
- Approach of choice in many applications
- Compromise between best of frequentist and Bayesian perspectives
- Uncertainty quantification (emulation)
- Decision theory, hierarchical models, dynamic linear models
- "Approximate" Bayesian methods

Client expectations

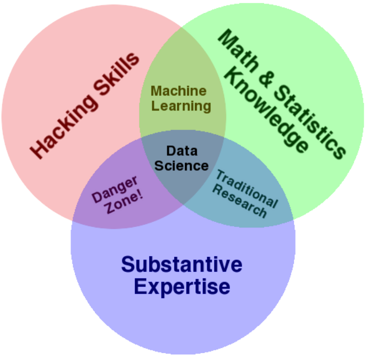
Context: More, faster, better measurement

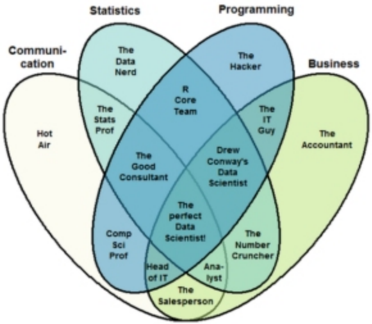
An ocean drifter, credit diydrones.com
Diameter $\approx$ 20cm, 1000s in the ocean

- Good, cheap, widget sensors
  - Environmental; Inspection and maintenance; Drifters
- Satellites
  - Ocean; Seismic; Greenhouse gases; Economic; Telemetry
- Drones, autonomous vehicles
- Sophisticated sensing
  - Spectroscopy; Optics
- Processes heavily monitored, data recorded
  - Manufacturing; Retail; Financial; Economic; Internet of things

Context: Emergence of data science

credit Drew Conway and Yanir Seroussi





Context: Dramatically improved connectivity

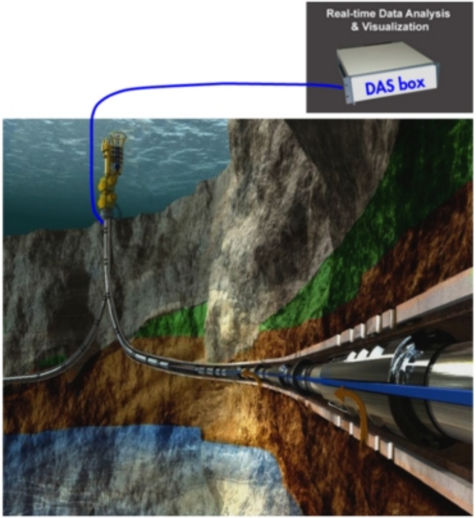
Everyone and everything digitally inter-connected; Everything is feasible source data for empirical inference . . . whether we like it or not
Credit Microsoft for images

- Global computing resources
- Millions of transactions per second
- "New state" for humanity?

- "Crude" data from any available source ingested into an "unstructured data store"
- "Unstructured" data "refined" and extracted to a structured data store, the "data mart"
- Inference using data mart and "analytics"

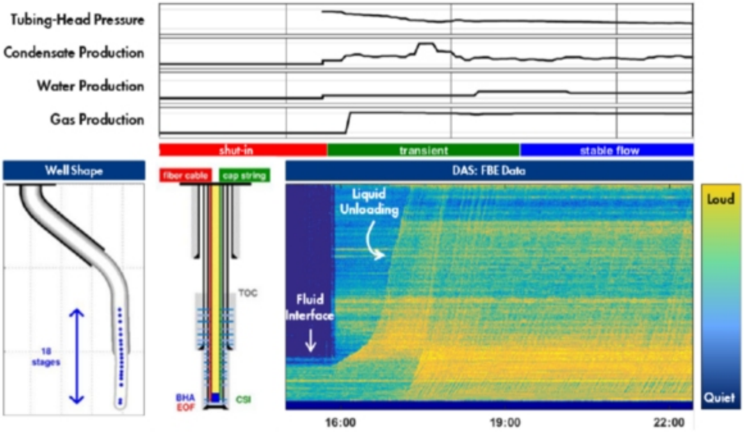
Distributed acoustic sensing (DAS)

- Fibre optic cable; Pulsed infra-red light from DAS box
- Acoustic noise causes optical properties of cable to change and reflect light
- Reflected light detected at DAS box;
- Inferred flow rates, instabilities, composition;
- **Continuous 10kHz data over network**; FFT to estimate $f(z, t)$;
- **Simple stats, automated large scale**



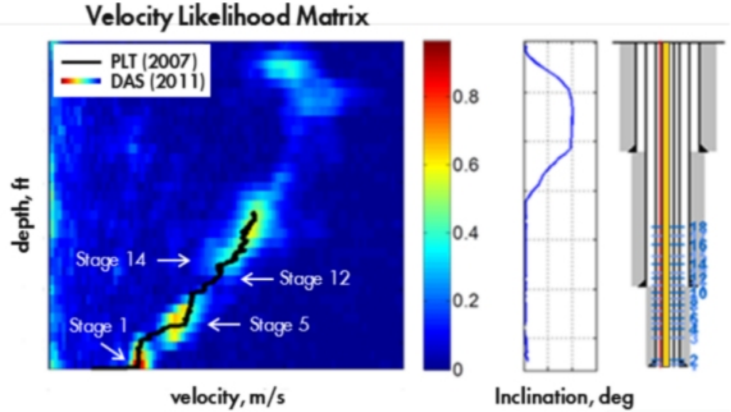
Distributed acoustic sensing

- **Up-front processing to z, t space**
- Well operation: In- and out-flows of oil, water, gas; Some flow control; Signal drops with distance;
- **"Velocity tracking" of multi-phase and inhomogeneous flow "slugs"**



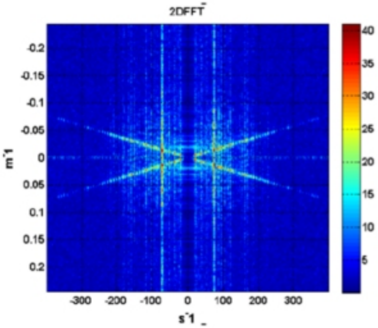
Distributed acoustic sensing

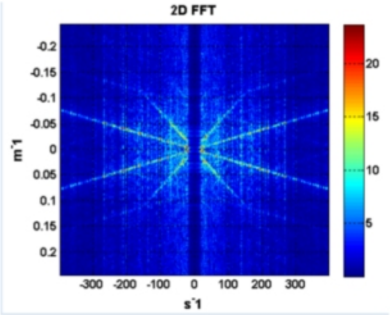
Empirical modelling of "slugs"; Regression



Distributed acoustic sensing

2D-FFT; Rays indicate sounds travelling at different speeds (ie phases) $\Rightarrow$ flow composition



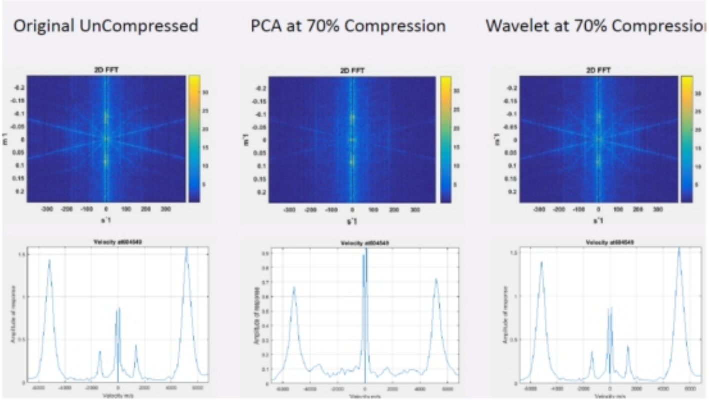


- 2DFFT: $F(\omega, k) = \sum_t \sum_z f(t, z) \exp[-2\pi i(\omega t - kz)]$
- spectrum: $S(\omega, k) = |F(\omega, k)|^2$
- phase speed: ω / k
- Radon transform

- lhs: Sound transmitted through steel only $5500 ms^{-1}$
- rhs: Sound transmitted through water also $1600 ms^{-1}$
- Non-dispersive regime: ω varies linearly with k

Distributed acoustic sensing

Compression; Principal components versus wavelet

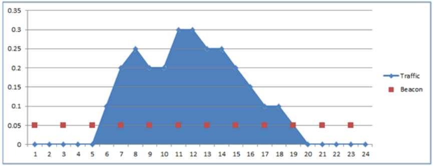


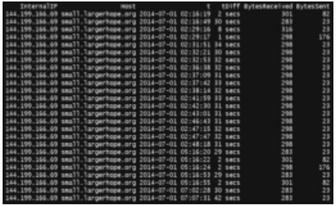
- Describe $f(z, t)$ using basis $\{\phi_i\}$, $f(z, t) = \sum_i c_i \phi_i(z, t)$
- Eliminate basis terms with small weights $|c_i| < \epsilon$
- Time-series compression

- lhs: Uncompresses has "steel" and "water"
- centre: PCA-compressed loses "water" at 70%
- rhs: Wavelet-compressed keeps "water" at 70%

Malware beaconing

Computer infected with malware; Malware seeks instructions from command server on internet; Spot beacon ⇒ spot infection; **Beaconing signal can be very sophisticated** bypassing best anti-virus defences; Beacons use any protocol, HTTPS increasingly used
Simple stats, automated large scale





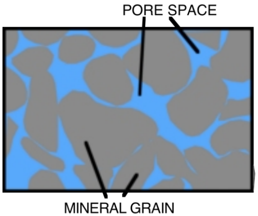
InternalIP	Host	t		tDiff		BytesReceived	BytesSent
144.199.166.69	small.largerhope.org	2014-07-01	02:16:19	2	secs	301	82
144.199.166.69	small.largerhope.org	2014-07-01	02:16:49	30	secs	283	23
144.199.166.69	small.largerhope.org	2014-07-01	02:29:16	8	secs	316	23
144.199.166.69	small.largerhope.org	2014-07-01	02:29:17	1	secs	298	176
144.199.166.69	small.largerhope.org	2014-07-01	02:31:51	34	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:32:21	30	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:32:53	32	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:36:38	32	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:37:09	31	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:37:42	33	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:38:14	32	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:41:59	33	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:42:30	31	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:43:01	31	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:46:43	31	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:47:15	32	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:47:47	32	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	02:48:18	31	secs	298	23
144.199.166.69	small.largerhope.org	2014-07-01	05:16:20	29	secs	283	23
144.199.166.69	small.largerhope.org	2014-07-01	05:16:22	2	secs	301	82
144.199.166.69	small.largerhope.org	2014-07-01	05:16:24	2	secs	298	176
144.199.166.69	small.largerhope.org	2014-07-01	05:16:53	29	secs	283	23
144.199.166.69	small.largerhope.org	2014-07-01	05:16:55	2	secs	301	82
144.199.166.69	small.largerhope.org	2014-07-01	07:02:28	30	secs	283	23
144.199.166.69	small.largerhope.org	2014-07-01	07:07:31	42	secs	283	23

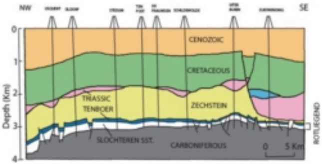
- Pattern recognition in time-series, change-point partitioning, anomaly detection
- Web sessions with "unusual" mix of web-browsing metadata
- Beaconing can be minor component of traffic

- lhs: Simplest beacon is regular "background" pulse
- lhs: Need to detect pulse within "normal" traffic
- rhs: Beacon with 30 second pulse in infected system

Seismic hazard monitoring

Gas extraction $\Rightarrow$ reduced pore pressure $\Rightarrow$ “compaction” $\Rightarrow$ subsidence and seismic activity
Multiple data sources; Spatio-temporal hierarchical models; Real-time monitoring

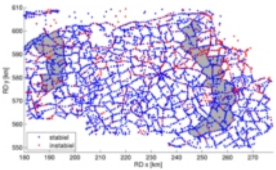


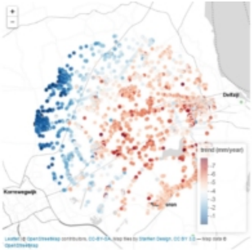


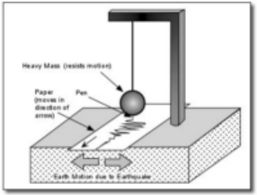
- $Pr(E) = f(C; \Theta)$
- E earthquake, C compaction, Θ reservoir parameters
- $S = S(C; \Theta)$, S subsidence
- $C = C(P; \Theta)$, P pore pressure

- lhs: Pore pressure drop causes compaction
- centre: Compaction causes faults to “slip”
- rhs: Surface fault in sandstone rock

Seismic hazard monitoring





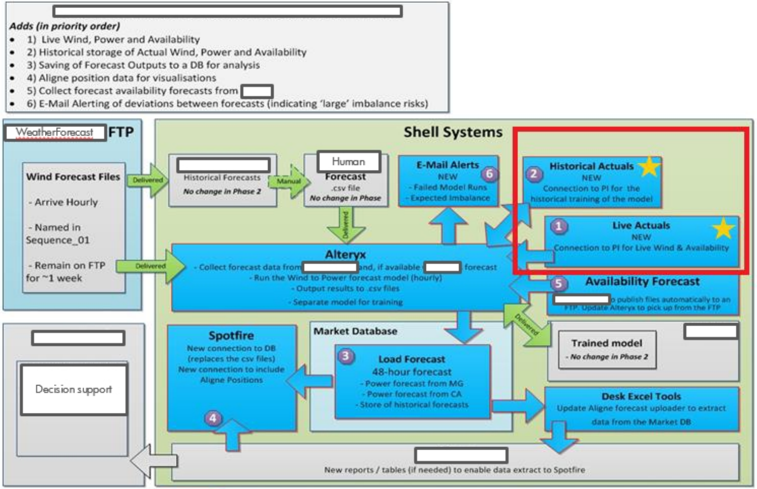


- $Pr(E) = f(C;\Theta)$
- E earthquake, C compaction, Θ reservoir parameters
- $S = S(C;\Theta)$, S subsidence
- $C = C(P;\Theta)$, P pore pressure
- Real-time monitoring
- Spatio-temporal modelling; Non-stationary extremes

- lhs: Optical leveling network measurements
- centre: Interferometric synthetic aperture radar (InSAR) measurements
- rhs: Seismograph
- More recently: GPS

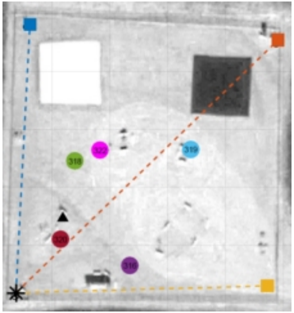
Wind power forecasting

- Wind + turbine $\Rightarrow$ electrical energy $\Rightarrow$ $$
- Plan energy production; Forecast wind field in time; Forecast production in time
- Extreme gusts $\Rightarrow$ turbine damage $\Rightarrow$ shutdown forecast
- Regression; Dynamic linear modelling
- Integrated model



Airborne gas monitoring

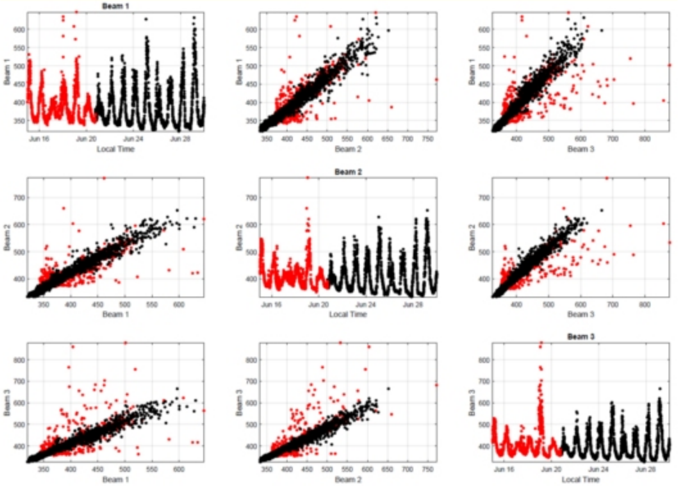
Carbon sequestration; Pump CO_2 underground; Need to ensure nothing escapes; On-line laser monitoring. Detection of unusual characteristics of multivariate time-series; **Web-based on-line implementation**



- Path-integrated $C(t, P_i) = \int_{P_i} c(\boldsymbol{r}(p), t)dp$ for paths $\{P_i\}$
- $c(\boldsymbol{r}, t) = A(\{S_j\}, W(\boldsymbol{R}, t)) + B(\boldsymbol{r}, t) + \epsilon(t)$
- Smooth B, "rougher" A, $B \gg A$

- lhs: Laser source
- centre: Retro-reflector
- rhs: Layout of sensors (source and 3 "retros")

Airborne gas monitoring



- Path-integrated concentrations for 3 paths
- Red: controlled release of gas; Black: natural variability
- Strong diurnal effect (mean, variance), sensor **anomalies**
- Dynamic linear modelling
- Inversion

Probabilistic inversion

Model

$$y = As + b + \epsilon$$

- y: Measured concentrations
- A: Assumed known from plume model
- s: Sources to be estimated
- b: Background to be estimated
- ϵ: Measurement error (assumed Gaussian), variance to be estimated

Inference

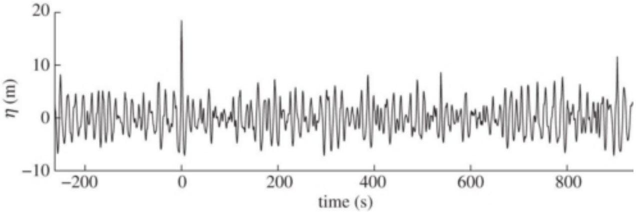
- Infer sources, background, measurement error, wind–field parameters
- Sources: Spiky; Gaussian mixture model
- Background: Smooth; Gaussian Markov random field, wind covariate
- Reversible jump MCMC inference over number of sources

Extreme environments

- Marine structures: Reliability, safety
- Extreme storms: Wave, wind and current fields
- Loading on structure, wave in deck
- Extreme value analysis: Non-stationary; Multivariate

Typical northern North Sea event

- Draupner New Year wave
  - 01.01.1995
- Different physics?
  - Higher-order effects
  - Directional spreading



Opportunities: Application areas

Physical sciences

- Spatio-temporal
- Inversion
- Multivariate time-series
- "Odd likelihoods"
  - Extreme value analysis
- Statistics and physical sciences

Data science (all with $n, p >> 1$)

- Text analytics
- Speech analytics
- Computer vision
- Statistics and "automatic control"
- Huge data
- Real-time analysis

Opportunities: Data science

What clients want?

- "Simple"
- Off the shelf: "self service analytics"
- Automatic: Effectively no human intervention; Stable algorithms
- Globally-connected
- Large scale: Huge numbers of concurrent analyses
- Real-time

Impact on the statistician

- Modelling with different data types: Numeric, text, image, language
- IT know-how: Databases, software, cluster, cloud, "hacking nouse"
- End-to-end involvement in projects: All "traditional" statistical skills needed
- Data "quality control", data cleaning
- Linear model, experimental design
  - $var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- Model assessment
- Custodian of responsible practice

Thank-you!

Spatial extremes

- Locations $\{s_k\}_{k=1}^p$, maxima $\{X_k\}$, covariates $\{\mathcal{C}_k\}$, density $\dot{f}$, cdf $\dot{F}$
- $\dot{f}(x_1, x_2, ..., x_p) = \left[\dot{f}(x_1)\dot{f}(x_2)...\dot{f}(x_p)\right] f(x_1, x_2, ..., x_p)$
- $X_k \sim \text{GEV}(\xi_k, \beta_k, \mu_k)$, so $\dot{f}, \dot{F}$ known
- GEV parameters ξ_k, β_k, μ_k vary smoothly between locations, and with $\mathcal{C}_k$
- Frechet scale: $x \to z$; $\dot{f}, \dot{F} \to f, F$
- $F(z_1, z_2, ..., z_p) = \exp\{-V(z_1, z_2, ..., z_p)\}$
- $V_{kl}(z_k, z_l; h(\Sigma)) = \frac{1}{z_k}\Phi(\frac{m(h)}{2} + \frac{\log(z_l/z_k)}{m(h)}) + \frac{1}{z_l}\Phi(\frac{m(h)}{2} + \frac{\log(z_k/z_l)}{m(h)})$
- $h = s_l - s_k$, $m(h) = (h'\Sigma^{-1}h)^{1/2}$, Φ is Gaussian
- Covariate effects $\mathcal{C}$ in Σ
- Joint Bayesian inference for $\{\xi_k(\mathcal{C}), \sigma_k(\mathcal{C}), \mu_k(\mathcal{C})\}$ and $\Sigma(\mathcal{C})$