

Slides at <a href="http://www.lancs.ac.uk/~jonathan">www.lancs.ac.uk/~jonathan</a>

**Philip Jonathan** Statistics and Data Science

Copyright of Shell

Statistics and Data Science

Emerging Applications in Industry

November 2017 1 / 26

## 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

- $\blacksquare$   $\approx$  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?

STREET, M. BOLL	ALL DESCRIPTION OF	THE STREET			MINU CA		T B TURN		UNIT ON THE OWNER	CONTRACTOR OF
COLUMN AND	ACCOUNT OF TAXABLE PARTY.	C RELICE & C		The Local Division in which the	CANE DOWN	A REAL PROPERTY.	CONTRACTOR OF	A CONTRACTOR OF	CARL ST. DO. LT.	Contrast of
			1000	The Party of		A CONTRACTOR OF				
	TAXABLE CONTRACTOR									
	Statistics in the second									
						State Set		Con Stars	AN INCOME	
	A PROPERTY OF TAXABLE	18232								
		2489 B			Charles and star					
States and	and the lot of	address of	April 1 and	all and the second	a second	and a start of	1.00	about the second		ALC: N
Contraction of the local division of the loc	States of the states of the	COLUMN D	No. of Concession, Name	SUCCESSION OF THE OWNER	Bernard Street, Street		Courses on the	NAME OF BRIDE	CONTRACTOR OF STREET,	THE OWNER ADDRESS
<b>DECORDOR</b>	E.S. Constant	10000	San Colora	201040	2020 A 18	S. O. L. R. L.	0000000		100010000	10.000
in landide a	ALC: NOT THE OWNER	A designation of	V STATISTICS	· ALLOWING	The Paralities of	The second second second	privates parted	the sectors	and the Real Property lies	a hand
and the second	Southern Street Street	1000	Could be the second	Salara and	mminn	432.00 2 2		Trating	of the surgery of the local division of the	1000
INTERNAL COLUMN	COMPANY OF STREET,	COLUMN 1	A STATISTICS	Contraction of the	IN LOCAL DR.	ENGLACIONAL	10 00 00 00 m	ALCOHOLD FUEL	Transferration and	ATT DESCRIPTION OF
THE OTHER DESIGN	DISCHARGE THE	omer h	INCOMENT	TO DO DO	COLUMN 1	THE R. LEWIS CO., LANSING MICH.	CORDON.	NUMBER OF STREET, STRE	00000	A COLUMN
				X MARINE	A PROPERTY AND	Sector Labor	C BARCOLLA		A MARCHINE OF	ALC: NO

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

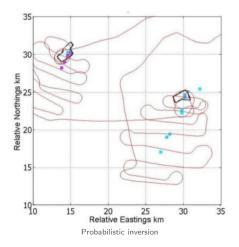
### More accessible data ... "digitalisation"

- $n_{2017} >> n_{1080}, p_{2017} >> p_{1080}$
- 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
- Altervx, Apache Spark, SQL, NoSQL.

Context: What's changing for us?



### 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**

Emerging Applications in Industry

Statistics and Data Science

### Context: More, faster, better measurement

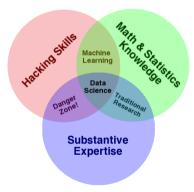


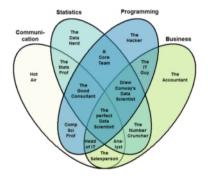
An ocean drifter, credit diydrones.com Diameter  $\approx$  20cm, 1000s in the ocean

- 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?

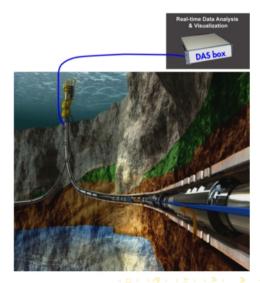


- "Crude" data from any available source ingested into an "unstructured data store"
- "Unstructured" data "refined" and extracted to a structured data store, the "data mart"
- Inference using data mart and "analytics"

#### Copyright of Shell

Statistics and Data Science

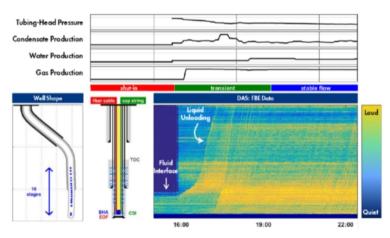
- Fibre optic cable; Pulsed infra-red light from DAS box
- Acoustic noise causes optical properties of cable to change and reflect light
- Reflected light detected at DAS box;
- Inferred flow rates, instabilities, composition;
- Continuous 10kHz data over network; FFT to estimate f(z, t);
- Simple stats,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



Copyright of Shell

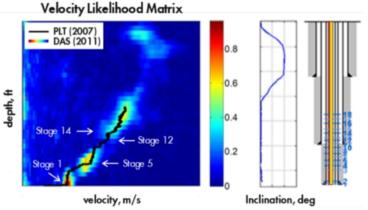
Statistics and Data Science

- 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"

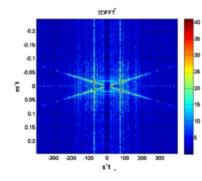


#### Copyright of Shell

Empirical modelling of "slugs"; Regression



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

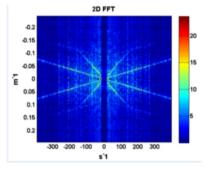


• 2DFFT: 
$$F(\omega, k) = \sum_t \sum_z f(t, z) \exp[-2\pi i(\omega t - kz)]$$

- spectrum:  $S(\omega, k) = |F(\omega, k)|^2$
- **phase speed:**  $\omega/k$
- Radon transform

Copyright of Shell

Statistics and Data Science

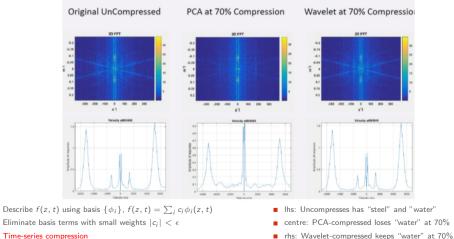


- Ihs: Sound transmitted through steel only  $5500 m s^{-1}$
- rhs: Sound transmitted through water also  $1600 m s^{-1}$
- Non-dispersive regime:  $\omega$  varies linearly with k

Emerging Applications in Industry

November 2017 13 / 26

Compression; Principal components versus wavelet



Time-series compression 

Copyright of Shell

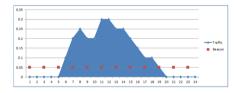
Statistics and Data Science

Emerging Applications in Industry

November 2017 14 / 26

### Malware beaconing

Computer infected with malware; Malware seeks instructions from command server on internet; Spot beacon  $\Rightarrow$  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	16061		1	TDIFF	BytesReceived	BytesSevt
	small.largerhope.org	2014-07-01	02:16:19	2 5465		82
144,199,166,69	small, largerhope, org	2014-07-01	02:14:49	10 5465	283	
144,199,166.69	small, largerhope, org	2014-07-01	02:29:16	8 5405	816	
	small.largerhope.org		02:29:17	1 5465	28	176
	small.largerhope.org			34 secs		
144.199.166.69	small, largerhope, org	2014-07-01	02:82:21	10 secs		
144.199.166.69	small.largerhope.org	2014-07-01	02:82:58	12 \$865		
	small.largerhope.org			12 secs	28.28	
	small.largerhope.org		02:37:09	<b>JL secs</b>		
144.199.166.69	small.largerhope.org	2014-07-01	02:37:42	11 secs		
	small.largerhope.org			12 secs	298	
	small.largerhope.org			13 secs	298	
	small.largerhope.org		02:42:30	31 secs		
	small.largerhope.org			31 secs		
	small.largerhope.org			31 secs		
	small.largerhope.org		02:47:15	32 secs		
	small.largerhope.org			32 secs		
	small.largerhope.org			31 secs		
	small.largerhope.org		05:14:20	29 secs		
	small.largerhope.org			2 secs		82
	small.largerhope.org			2 14C5		176
	small.largerhope.org			29 secs		23 82
	small.largerhope.org			2 14C5		
144.199.166.69	small.largerhope.org	2014-07-01	07:02:28	30 secs		
144.199.166.69	small.largerhope.org	2014-07-01	07:07:31	42 secs		

#### 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

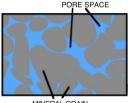
- Ihs: Simplest beacon is regular "background" pulse
- Ihs: Need to detect pulse within "normal" traffic
- rhs: Beacon with 30 second pulse in infected system

#### Copyright of Shell

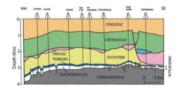
Statistics and Data Science

### 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



MINERAL GRAIN





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

- Ihs: Pore pressure drop causes compaction
- centre: Compaction causes faults to "slip"
- rhs: Surface fault in sandstone rock

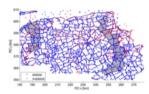
### Copyright of Shell

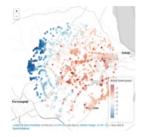
#### Statistics and Data Science

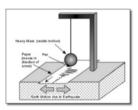
#### Emerging Applications in Industry

#### November 2017 16 / 26

### Seismic hazard monitoring







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

Copyright of Shell

Statistics and Data Science

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

Emerging Applications in Industry

November 2017 17 / 26

### Wind power forecasting

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

#### Adds (in priority order) 1) Live Wind Power and Availability 2) Historical storage of Actual Wind, Power and Availability 3) Saving of Forecast Outputs to a DB for analysis . 4) Aligne position data for visualisations 5) Collect forecast availability forecasts from 6) E-Mail Alerting of deviations between forecasts (indicating 'large' imbalance risks) Shell Systems Weatherforecast FTP Human E-Mail Alerts **Historical Actuals** Wind Forecast Files Forecast NEW cay file Connection to PI for the No change in Phase 2 No change in Phase Arrive Hourly Live Actuals Named in Sequence 01 Alterva Remain on FTP for ~1 week Availability Forecast Market Database Spotfire Trained model New connection to DR (replaces the cost files) No change in Phase 3 48-hour forecast Decision support Desk Excel Tools Update Aligne forecast uploader to extract data from the Market DB

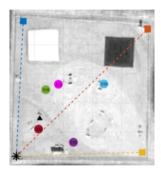
New reports / tables (if needed) to enable data extract to Spotfire

### 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







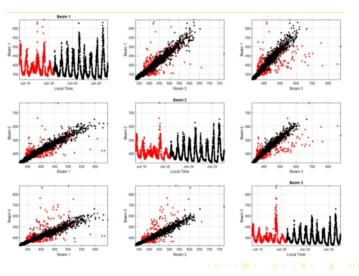
- Path-integrated  $C(t, P_i) = \int_{P_i} c(\mathbf{r}(p), t) dp$  for paths  $\{P_i\}$
- $c(r, t) = A(\{S_j\}, W(R, t)) + B(r, t) + \epsilon(t)$
- Smooth B, "rougher" A,  $B \gg A$

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

Copyright of Shell

### 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

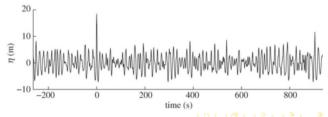
- Infer sources, background, measurement error, wind–field parameters
- Sources: Spiky; Gaussian mixture model
- Background: Smooth; Gaussian Markov random field, wind covariate
- Reversible jump MCMC inference over number of sources

### **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

Copyright of Shell

Statistics and Data Science

### **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(β̂) = σ²(X'X)<sup>-1</sup>
- Model assessment
- Custodian of responsible practice



# Thank-you!

Copyright of Shell

Statistics and Data Science

Emerging Applications in Industry

November 2017 25 / 26

### **Spatial extremes**

• Locations  $\{s_k\}_{k=1}^p$ , maxima  $\{X_k\}$ , covariates  $\{\mathcal{C}_k\}$ , density  $\dot{f}$ , cdf  $\dot{F}$ 

• 
$$\dot{f}(x_1, x_2, ..., x_p) = \left[\dot{f}(x_1)\dot{f}(x_2)...\dot{f}(x_p)\right]\dot{f}(x_1, x_2, ..., x_p)$$

- $X_k \sim \text{GEV}(\xi_k, \beta_k, \mu_k)$ , so  $\dot{f}, \dot{F}$  known
- GEV parameters  $\xi_k, \beta_k, \mu_k$  vary smoothly between locations, and with  $C_k$
- Frechet scale:  $x \rightarrow z$ ;  $\dot{f}, \dot{F} \rightarrow f, F$

• 
$$F(z_1, z_2, ..., z_p) = \exp\{-V(z_1, z_2, ..., z_p)\}$$

- $V_{kl}(z_k, z_l; h(\Sigma)) = \frac{1}{z_k} \Phi(\frac{m(h)}{2} + \frac{\log(z_l/z_k)}{m(h)}) + \frac{1}{z_l} \Phi(\frac{m(h)}{2} + \frac{\log(z_k/z_l)}{m(h)})$
- $h = s_l s_k$ ,  $m(h) = (h' \Sigma^{-1} h)^{1/2}$ ,  $\Phi$  is Gaussian
- Covariate effects C in  $\Sigma$
- Joint Bayesian inference for  $\{\xi_k(\mathcal{C}), \sigma_k(\mathcal{C}), \mu_k(\mathcal{C})\}\$  and  $\Sigma(\mathcal{C})$

Copyright of Shell