Real-time data

. . . real-world applications

. . . almost no equations

Stijn Bierman, Jose Gonzalez-Martinez, Wayne Jones, Rakesh Paleja,
Tim Park, David Randell, Emma Ross, Mingqi Wu, Philip Jonathan

Acknowledgement

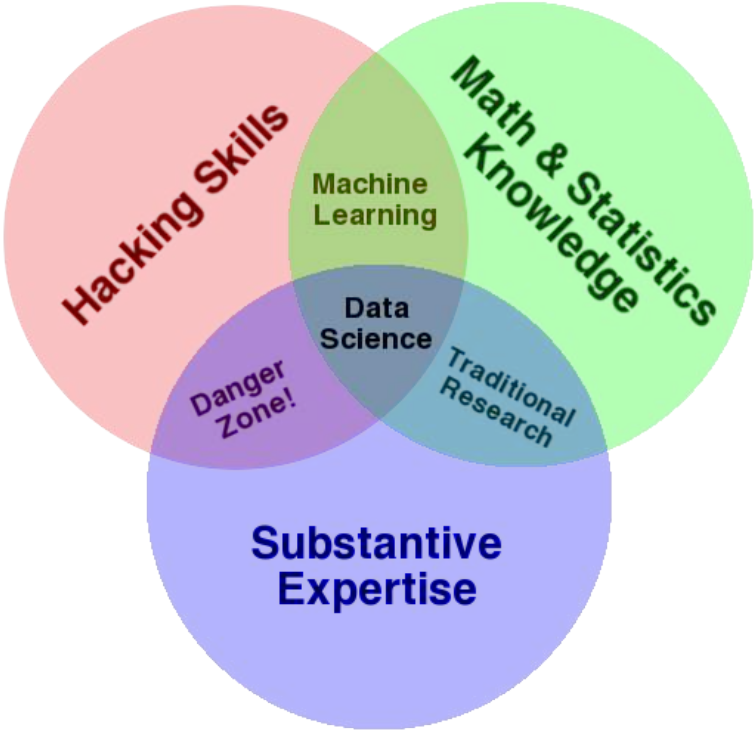
- Shell Statistics and Data Science
- Shell colleagues and clients
- Lancaster
- Delft, Durham, Glasgow, Imperial, UCL

Overview

- context
- applications
  - acoustic sensing
  - malware beaconing
  - seismic hazard monitoring
  - airborne gaseous monitoring
  - wind power forecasting
  - ...
- opportunities

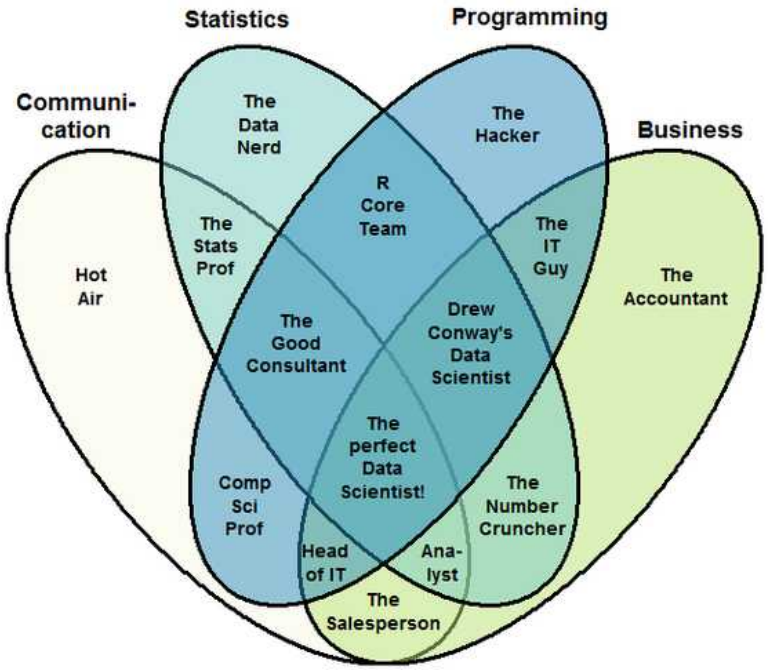
Context

the not-so-lonely statistician . . . what's changed?



- **digitalisation**
- data science, "predictive analytics"
- data: big, streamed, unstructured, connected

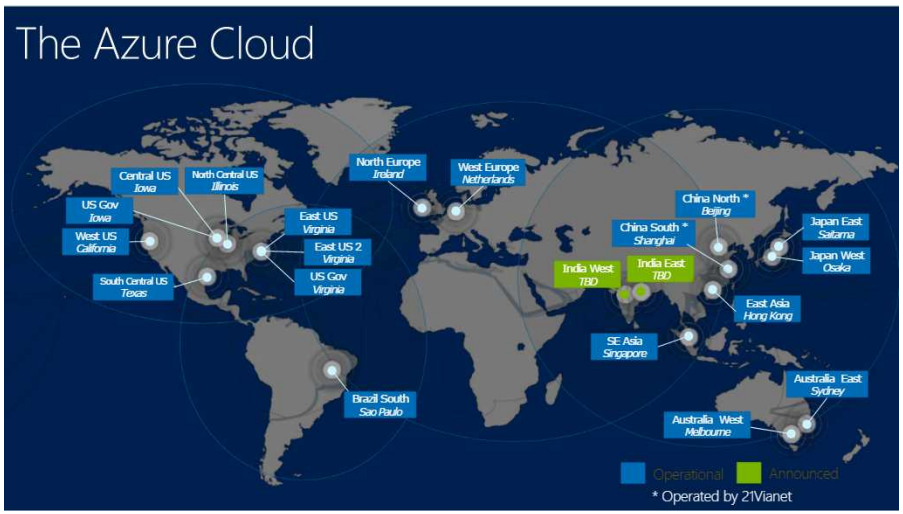
credit Drew Conway and Yanir Seroussi



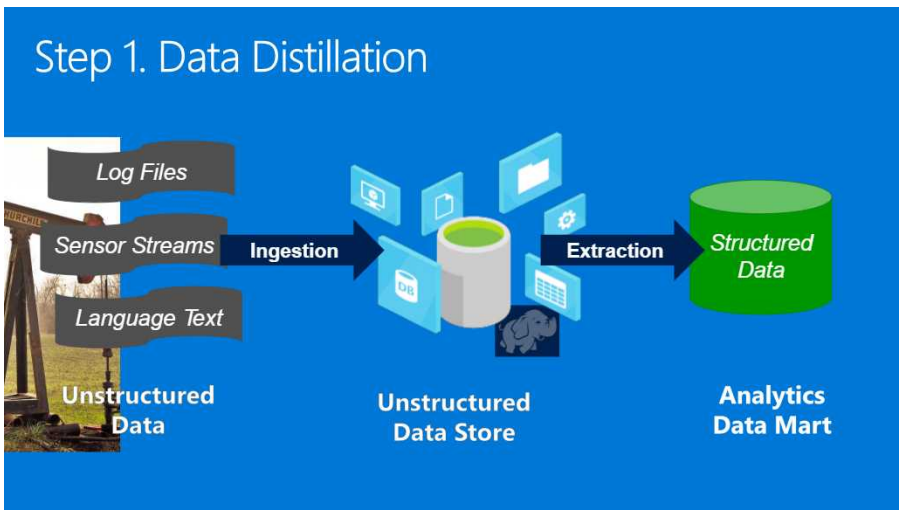
- statistical expertise
- problem domain knowledge
- scientific programming expertise
- communication and consultancy skills
- computer science, IT and "hacking nouse"

Connected

everyone and everything digitally inter-connected; everything is feasible source data for statistical inference
. . . whether we like it or not



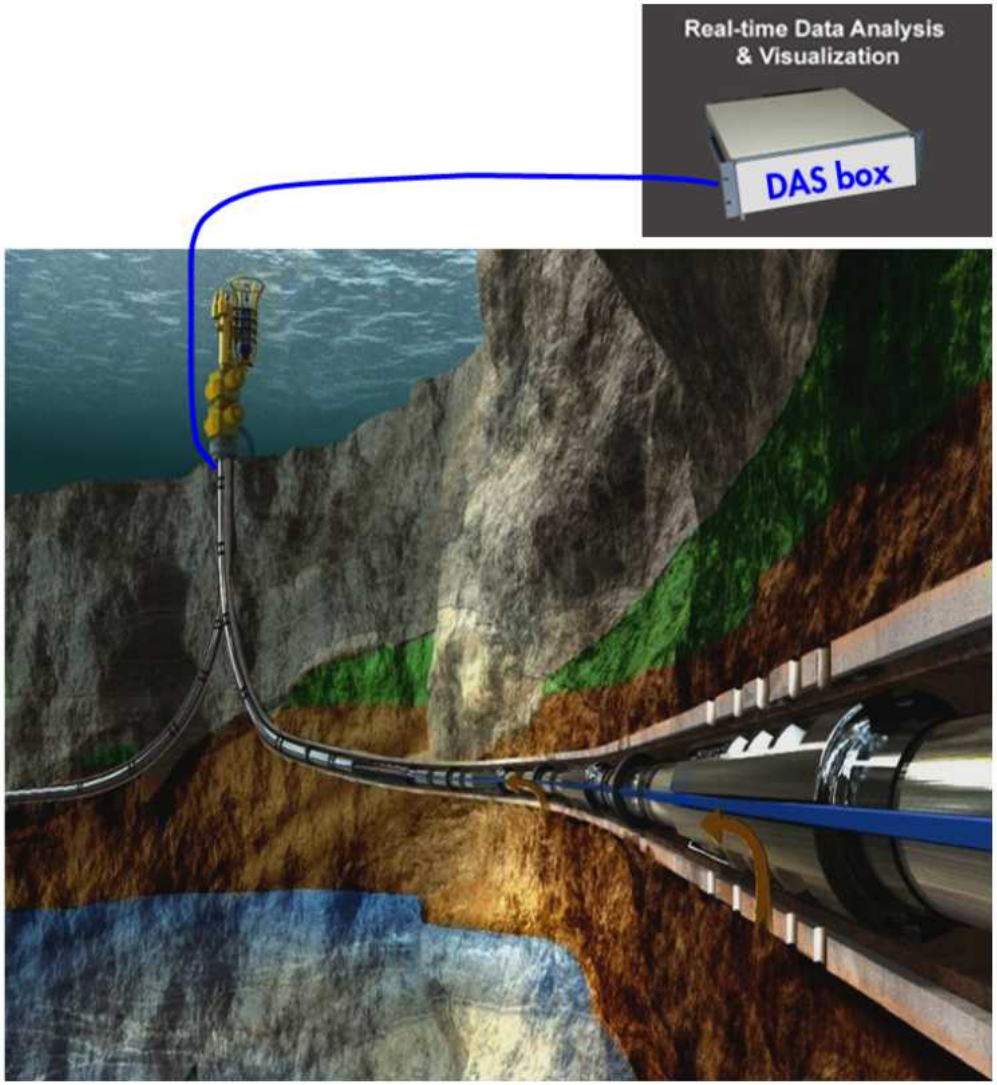
- global computing resources
- credit Microsoft for slides



- "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"
- millions of transactions per second

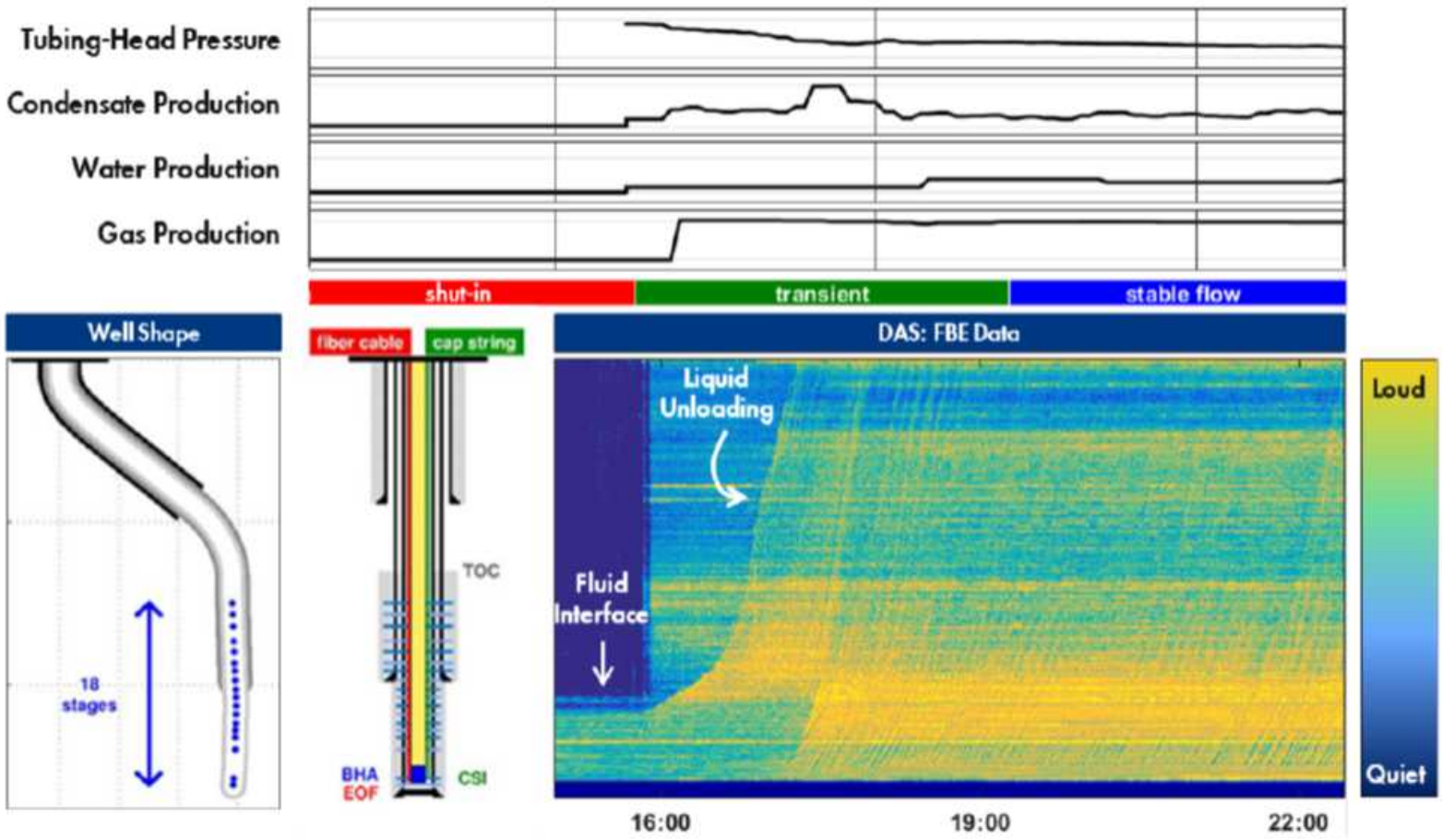
Distributed acoustic sensing (DAS)

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; infer flow rates, instabilities, composition; **continuous 10Hz data over network**; preprocessed data via FFT to $f(z, t)$; **simple stats, automated large scale**



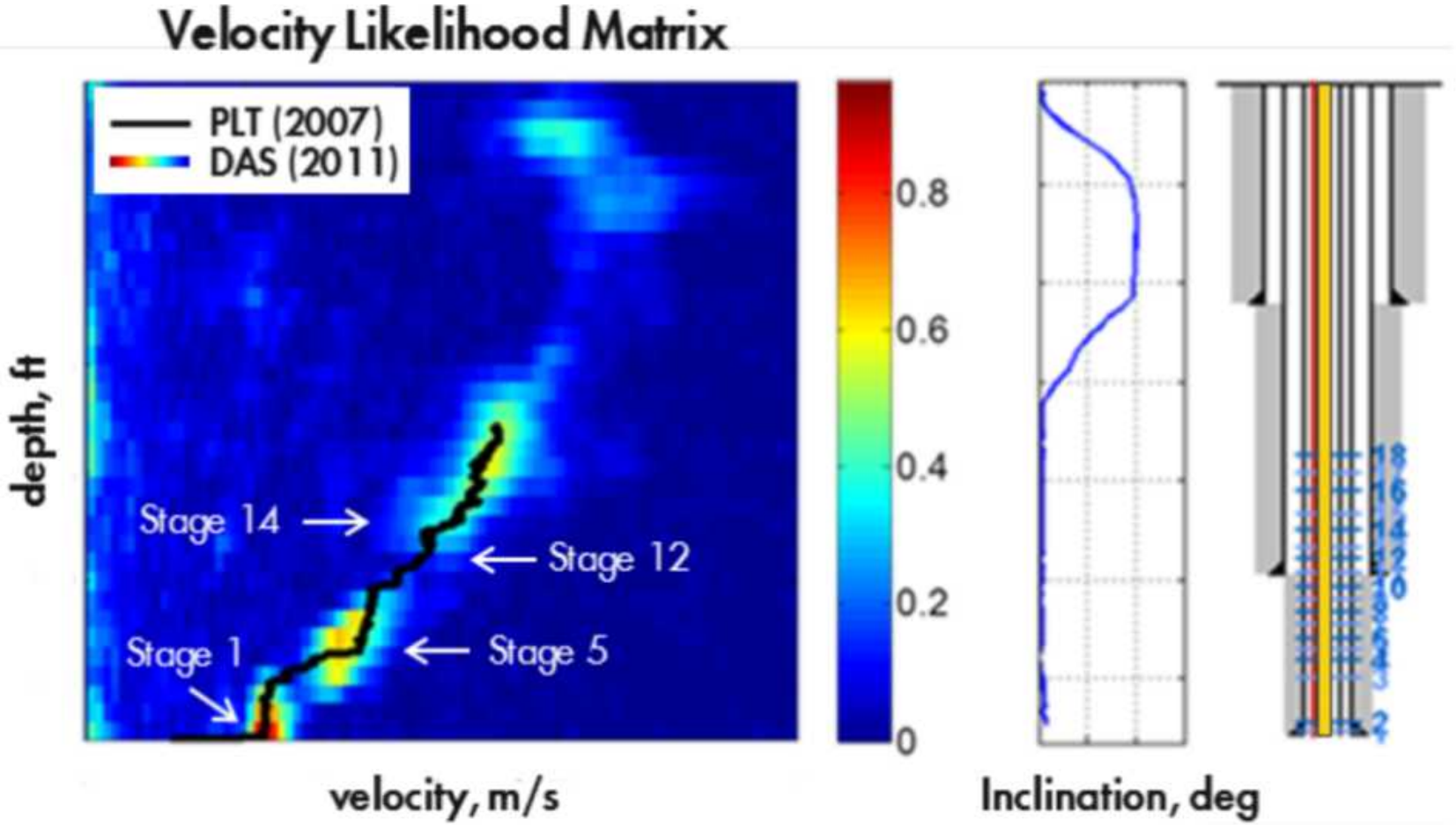
Distributed acoustic sensing (DAS)

up-front processing to z, t space (Jean-Philippe's talk yesterday); 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"**



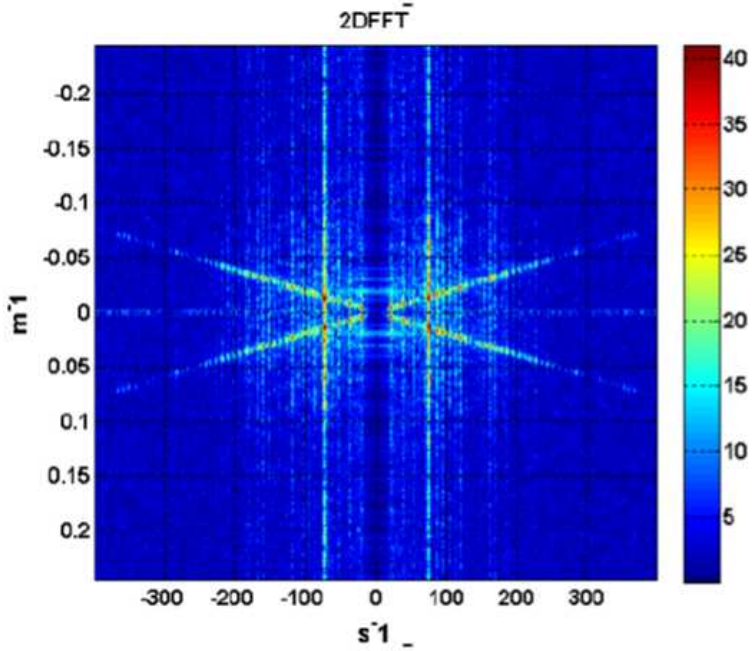
Distributed acoustic sensing (DAS)

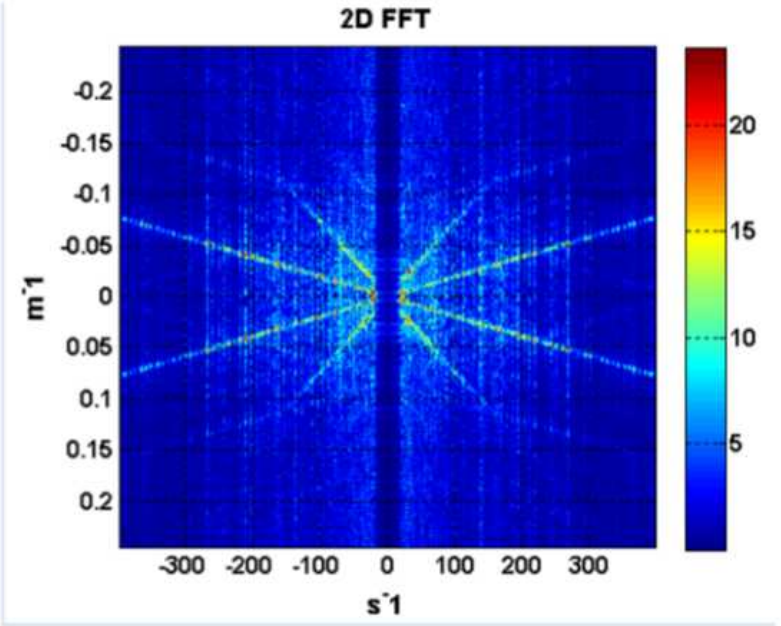
empirical modelling of "slugs"



Distributed acoustic sensing (DAS)

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



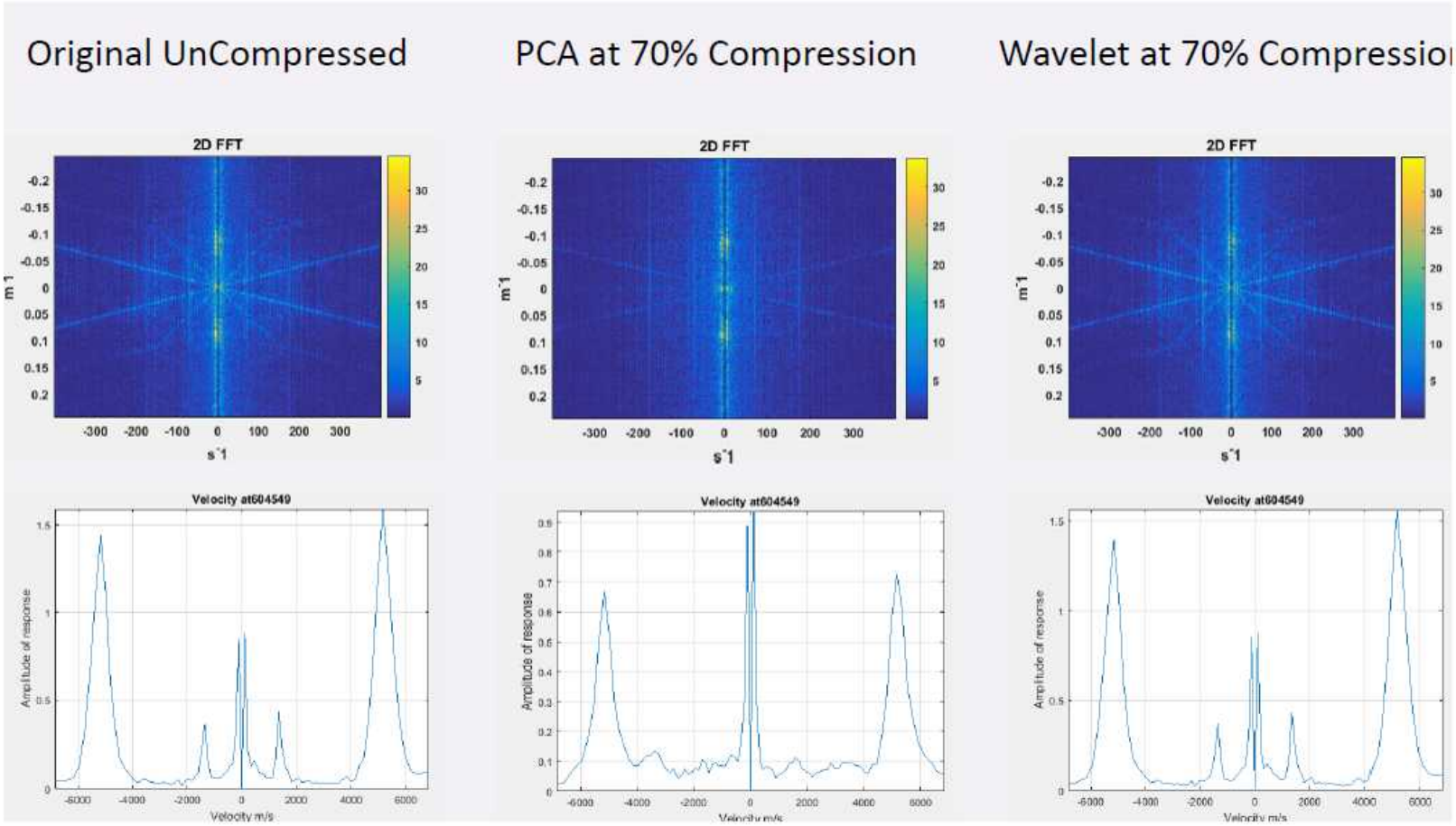


- 2DFFT: $F(\omega, k) = \sum_t \sum_z f(t, z) \exp[-2\pi i(\omega t - kz)]$
- spectrum: $S(\omega, k) = |F(\omega, k)|^2$
- phase speed: ω/k
- Radon transform

- 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
- Adam's talk yesterday.

Distributed acoustic sensing (DAS)

compression; principal components versus wavelet



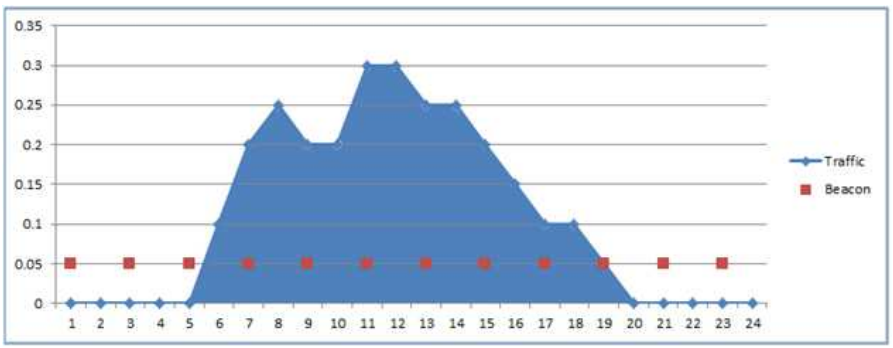
- describe $f(z,t)$ using basis $\{\phi_i\}$, $f(z,t) = \sum_i c_i \phi_i(z,t)$
- eliminate basis terms with small weights $|c_i| < \epsilon$

- 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 $\Rightarrow$ 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	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

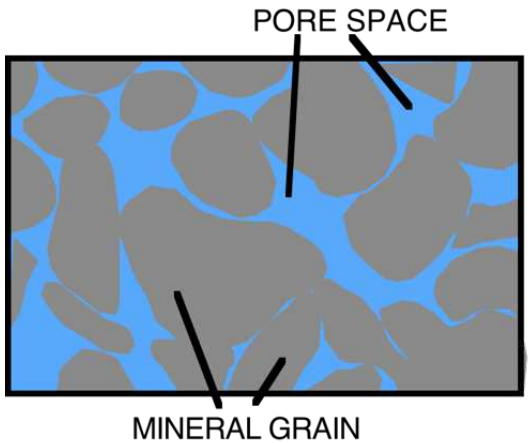
- **pattern recognition** in time-series, change-point partitioning
- web sessions with "unusual" mix of web-browsing metadata
- beaconing can be minor component of traffic
- Nick's talk yesterday

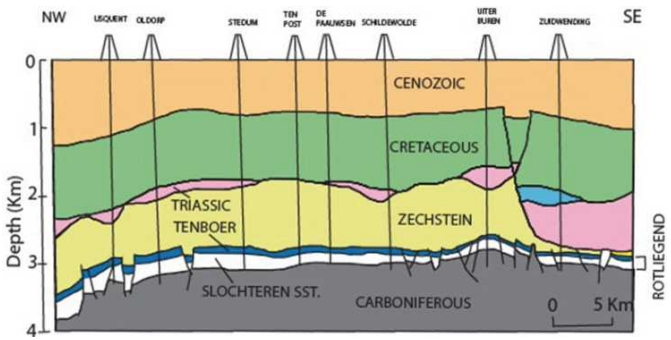
- lhs: simplest beacon is regular "background" pulse
- lhs: need to detect pulse within "normal" traffic
- rhs: beacon with 30 second pulse in infected system

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



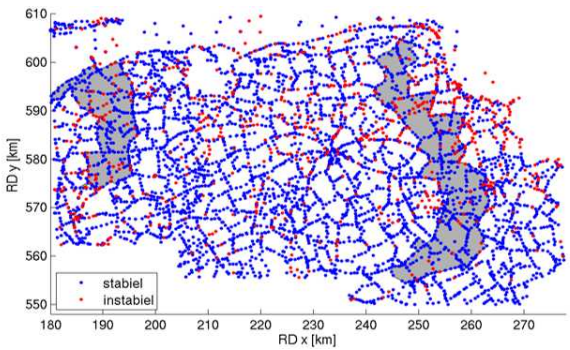


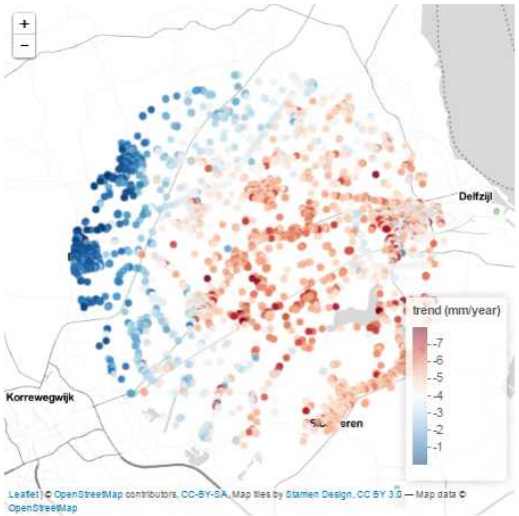
- $Pr(E) = f(C; \Theta)$
- E earthquake, C compaction, Θ reservoir parameters
- $S = S(C; \Theta)$, S subsidence
- $C = C(P; \Theta)$, S subsidence, P pore pressure

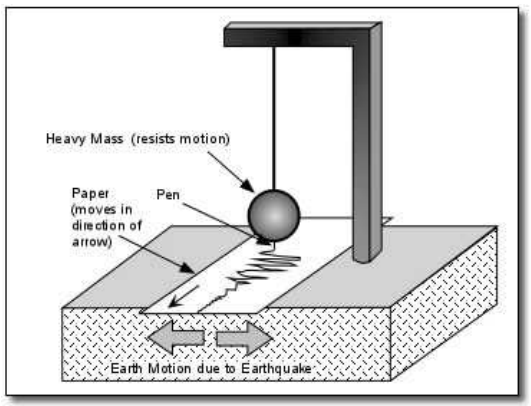
- lhs: pore pressure drop causes compaction
- centre: compaction causes faults to "slip"
- rhs: surface fault in sandstone rock

Seismic hazard monitoring

multiple data sources, spatio-temporal hierarchical models, real-time monitoring







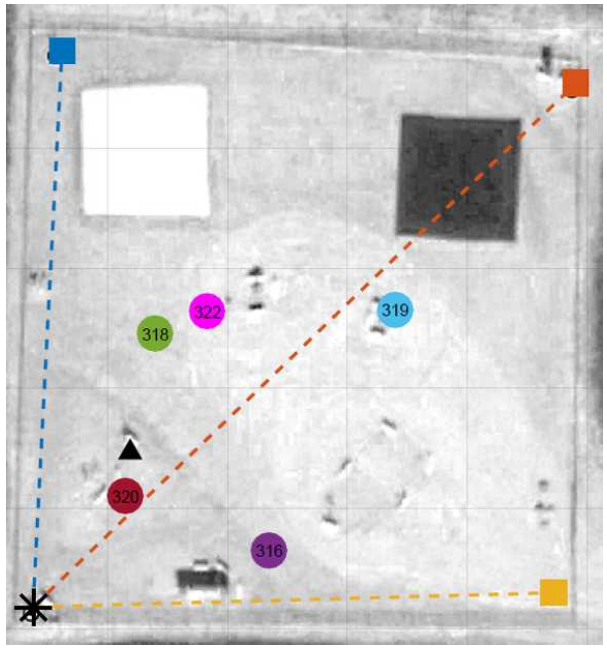
- $Pr(E) = f(C; \Theta)$
- E earthquake, C compaction, Θ reservoir parameters
- $S = S(C; \Theta)$, S subsidence
- $C = C(P; \Theta)$, S subsidence, P pore pressure
- real-time monitoring
- random fields, non-stationary extremes

- lhs: optical leveling network measurements
- centre: interferometric synthetic aperture radar (InSAR) measurements
- rhs: seismograph
- also, more recently: GPS

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**

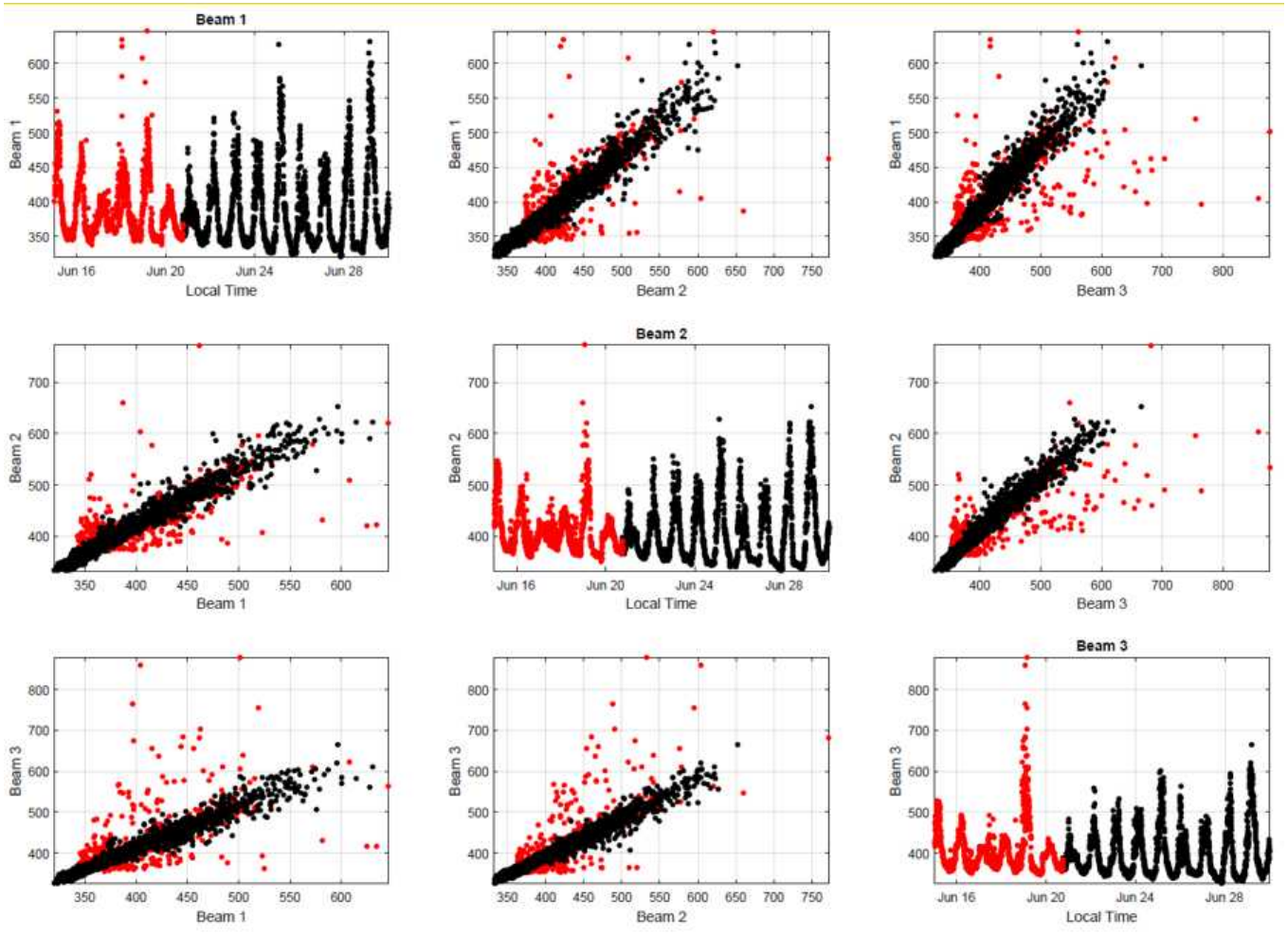


- path-integrated $C(t, P_i) = \int_{P_i} c(\boldsymbol{r}(p), t)dp$ for paths $\{P_i\}$
- $c(\boldsymbol{r}, t) = A(\{S_j\}, W(\boldsymbol{R}, t)) + B(\boldsymbol{r}, t) + \epsilon(t)$
- smooth B, "rougher" A, $B \gg A$

- lhs: laser source
- centre: retro-reflector
- rhs: layout of sensors (source and 3 "retros")

Airborne gas monitoring

sample; **web-based implementation**

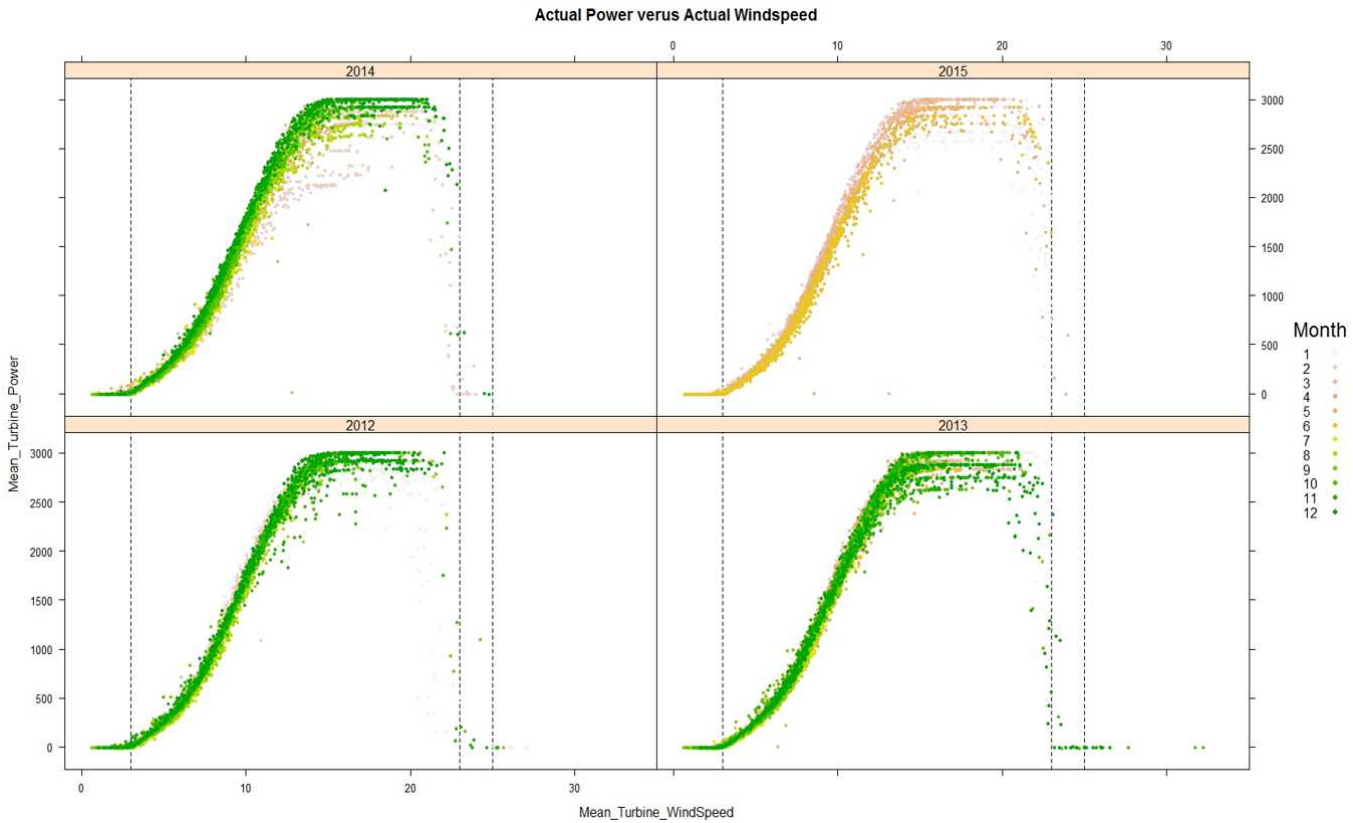


- path-integrated concentrations for 3 paths
- red: controlled release of gas
- black: natural variability
- cross-correlation important
- strong diurnal effect (mean, variance), sensor **anomalies**
- multivariate dynamic linear modelling

Wind power forecasting

wind energy logistics; optimal turbine location; production forecasting

simple stats, "resistant" on-line implementation

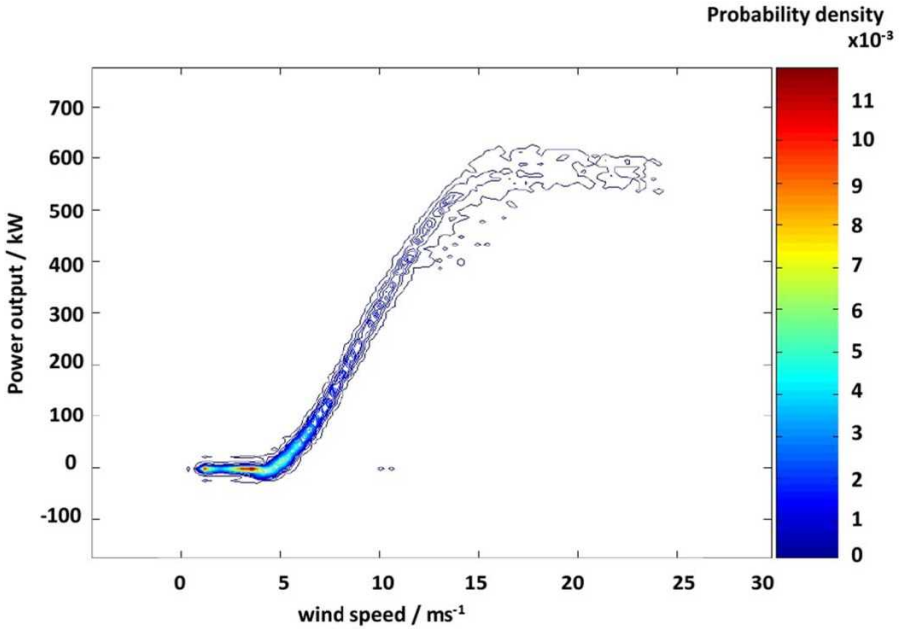


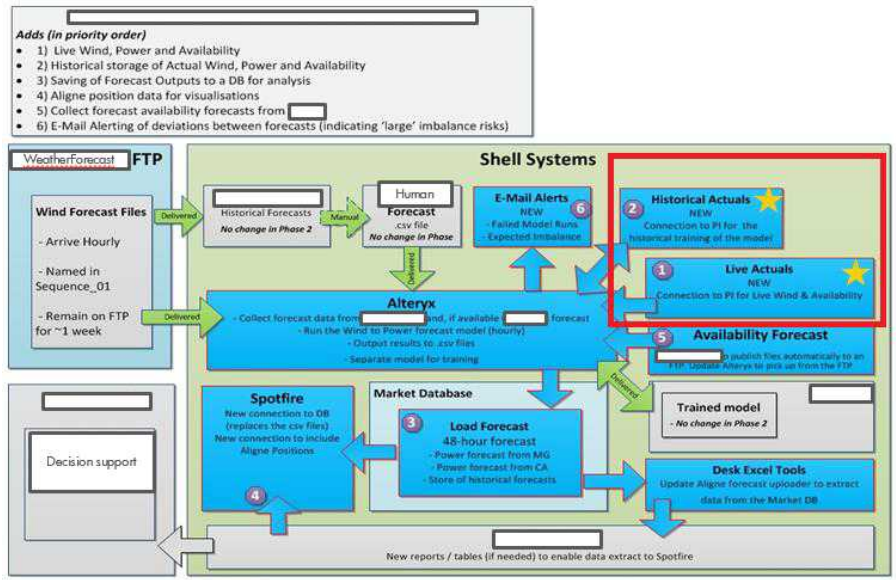
- wind + turbine $\Rightarrow$ electrical energy $\Rightarrow$ \$\$
- need to plan energy production (trading, supply assurance)
- $\Rightarrow$ need to forecast wind field $\boldsymbol{w}(\boldsymbol{r}, t)$

- wind forecasting problematic; literally **chaotic**
- extreme gusts $\Rightarrow$ turbine damage $\Rightarrow$ shutdown forecast

Wind power forecasting

simple non-parametric model; on-line implementation





- filtering of 3^{rd} party wind field predictions
- lhs: prediction of power given wind speed
- rhs: on-line implementation

Other applications

Long-standing

- process monitoring in manufacturing
  - long history ($\gg$ 40 years) of "chemometrics"
  - modelling of messy multivariate time-series
  - on-line monitoring, off-line trouble-shooting, "SPC"
- severe environments
  - spatio-temporal modelling, computationally challenging
    - Jean-Philippe's talk (c.f. comp. tricks, scale trans.)
  - "data rich": observations, forecasts, hindcasts (Adam's talk)
- cash forecasting ("spurious correlation", Chao's talk)

More recent

- vehicle telemetry
  - optimal product (fuel, lubricant) design
  - combining on-board data with other sources in road tests
- text analytics
  - "competitive intelligence", breakthrough technology
  - NLP: unstructured $\Rightarrow$ structured

Opportunities

What clients want in terms of real-time analysis

- simplicity
- automation
  - effectively no human intervention
  - "strong and stable" algorithms
- huge numbers of concurrent analyses
- "at-line" if not on-line real-time execution

Impact on the statistician

- involvement from "solution design" to "end implementation"
  - all "traditional" statistical skills still needed
- processing multiple "unstructured" (input) data types
- IT interfaces
  - databases, software, cluster/cloud ...
  - "hacking nouse"