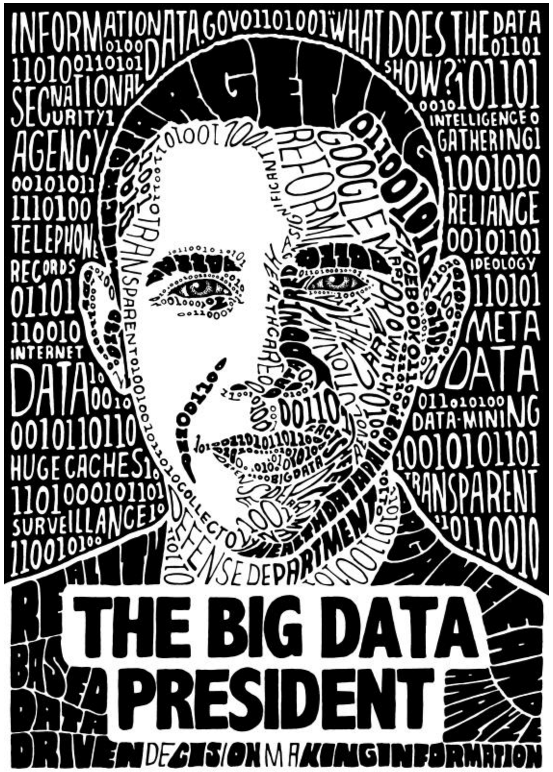
ESRA Norge
Norsk forening for risiko-
og pålitelighetsanalyse

Discovery and prediction of change

Arnoldo Frigessi
frigessi@medisin.uio.no

www.biginsight.no



Office of Science and Technology Policy
Executive Office of the President

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

A WORLD THAT COUNTS

MOBILISING THE DATA REVOLUTION FOR SUSTAINABLE DEVELOPMENT

DATA REVOLUTION GROUP

The UN Secretary General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

Big Data, Big Impact: New Possibilities for International Development

A flood of data is created every day. (…)

(We) are beginning to realise the potential for channelling these torrents of data into actionable information that can be used to identify needs, provide services, predict and prevent crises.

Navigating the next industrial revolution

Revolution	Year	Information
1	1784	Steam, water, mechanical production equipment
2	1870	Division of labour, electricity, mass production
3	1969	Electronics, IT, automated production
4	?	Cyber-physical systems

Die Bundesregierung

DIE NEUE HIGHTECH STRATEGIE
Innovationen für Deutschland

Industrie 4.0

Die Wirtschaft steht an der Schwelle zur vierten industriellen Revolution. Durch das Internet getrieben, wachsen reale und virtuelle Welt immer weiter zu einem Internet der Dinge zusammen. Die Kennzeichen der künftigen Form der Industrieproduktion sind die starke Individualisierung der Produkte unter den Bedingungen einer hoch flexibilisierten (Großserien-)Produktion, die weitgehende Integration von Kundinnen und Kunden sowie Geschäftspartnerinnen und -partnern in Geschäfts- und Wertschöpfungsprozesse und die Verkopplung von Produktion und hochwertigen Dienstleistungen, die in sogenannten hybriden Produkten mündet. Die deutsche Industrie hat jetzt die Chance, die vierte industrielle Revolution aktiv mitzugestalten. Mit dem Zukunftsprojekt Industrie 4.0 wollen wir diesen Prozess unterstützen.

Digital is the main reason just over half of the companies on the Fortune 500 have disappeared since the year 2000

Pierre Nanterme
CEO of Accenture

Statistics plays a central role in the analysis of big data.

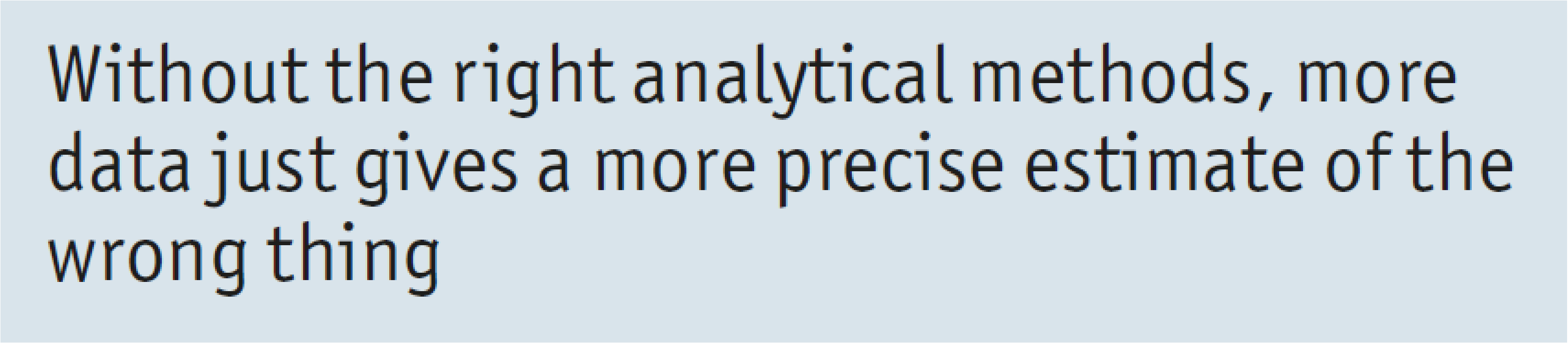
> “That the dry world of statistics is becoming a battleground of ideas and commercial interests, affecting the future of people around the world, may shock some.”

The Economist

Technology Quarterly

March 7th 2015

Often, it is not enough to crunch data!



> Without the right analytical methods, more data just gives a more precise estimate of the wrong thing

The Economist

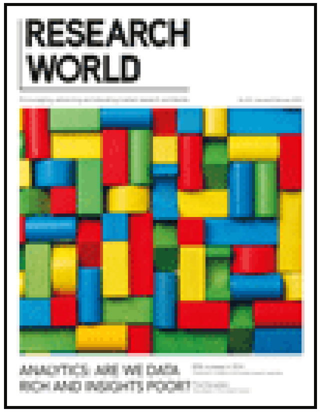
Technology Quarterly

March 7th 2015

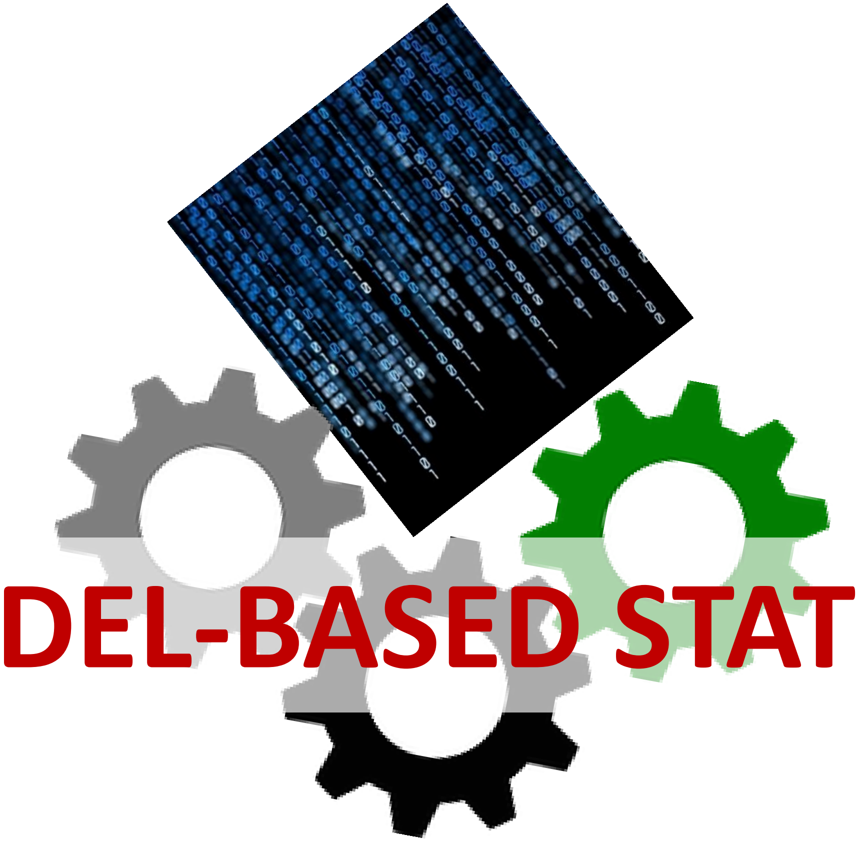
COVER STORY

JO BOWMAN

Analytics: are we data rich and insights poor?



RESEARCH WORLD | January/February 2015



MODEL-BASED STATISTICS

Model-based methods exploit knowledge and structure in the new data,
To understand, discover, predict, control, quantifying uncertainty.

TRACKING UNEMPLOYMENT USING MOBILE PHONE DATA

Toole, J. L., Lin, Y. R., Muehlegger, E., Shoag, D., González, M. C., & Lazer, D. *Journal of The Royal Society Interface, 2015*

- Real time estimate of changes in unemployment, at arbitrarily fine spatial scale, using existing mobile phone data.

- Ahead traditional indicators in European countries

Data - mobile phone calls:

- caller -> receiver
- location
- time

Training: Case of a large factory closing down

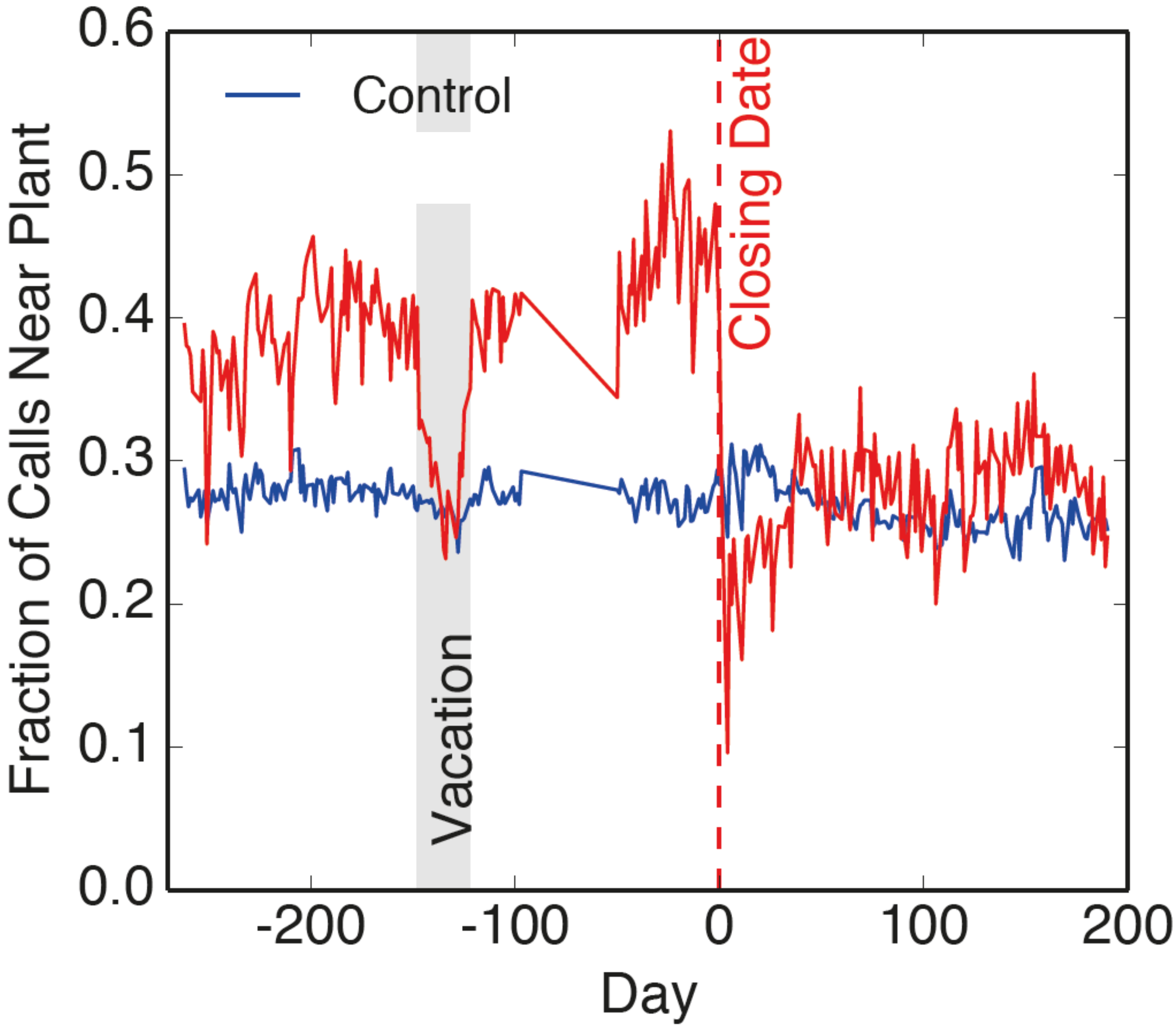
- Compare individual signal before vs. after closure
- Find special features of the signal when jobs are lost

Calibrating: A region with official unemployment estimates

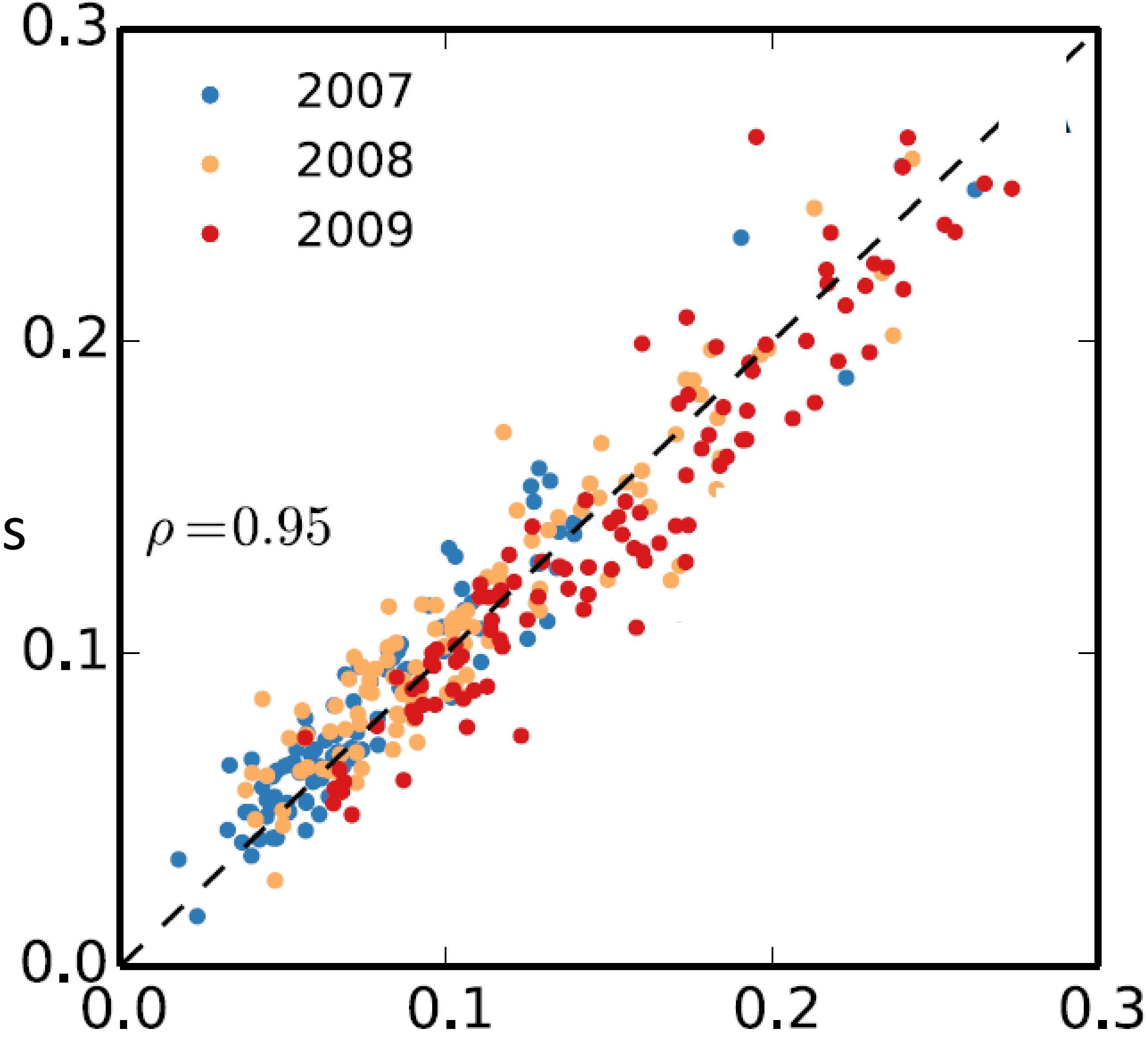
- Match "lost-job" mobile phone signal to unemployment rates

Predict: Current (and near future) unemployment

Training

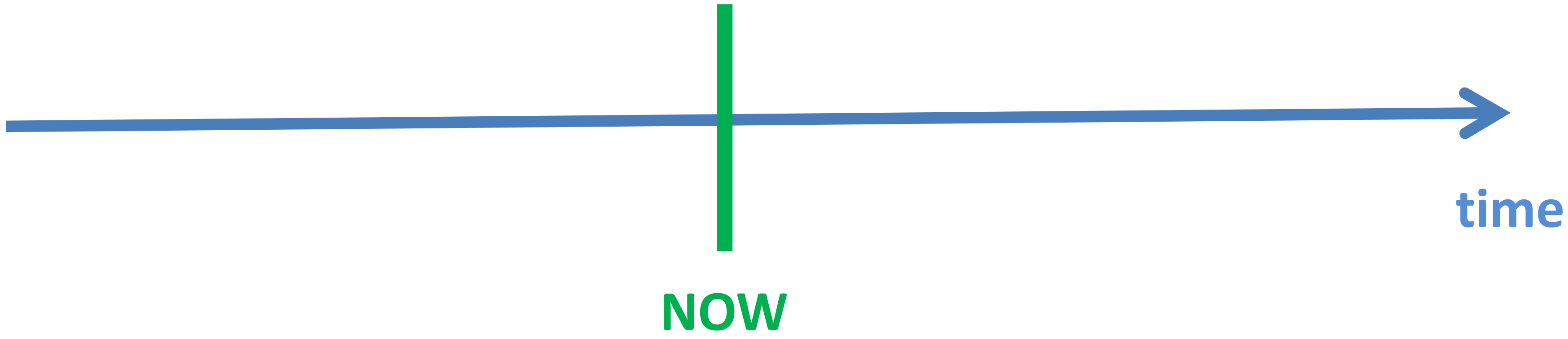

0.6
0.5
0.4
0.3
0.2
0.1
0.0
Fraction of Calls Near Plant
Control
Vacation
Closing Date
-200
-100
0
100
200
Day

Prediction


0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
● 2007
● 2008
● 2009
$\rho = 0.95$

Based on mobile phones
Official rates

Hindcast **Nowcast** **Forecast**



future

a statistical calculation determining probable present conditions

past

Oslo University Hospital

UiO : University of Oslo

UNIVERSITETET I BERGEN

CANCER Registry of Norway
INSTITUTE OF POPULATION-BASED CANCER RESEARCH

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

BigInsight

DNV·GL

ABB

nav

Gjensidige

telenor

Skatteetaten

DNB

HYDRO

Oslo University Hospital

UiO : University of Oslo

UNIVERSITETET I BERGEN

CANCER Registry of Norway
INSTITUTE OF POPULATION-BASED CANCER RESEARCH

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Personalised solutions

Forecasting the transient

DNV·GL

ABB

Gjensidige

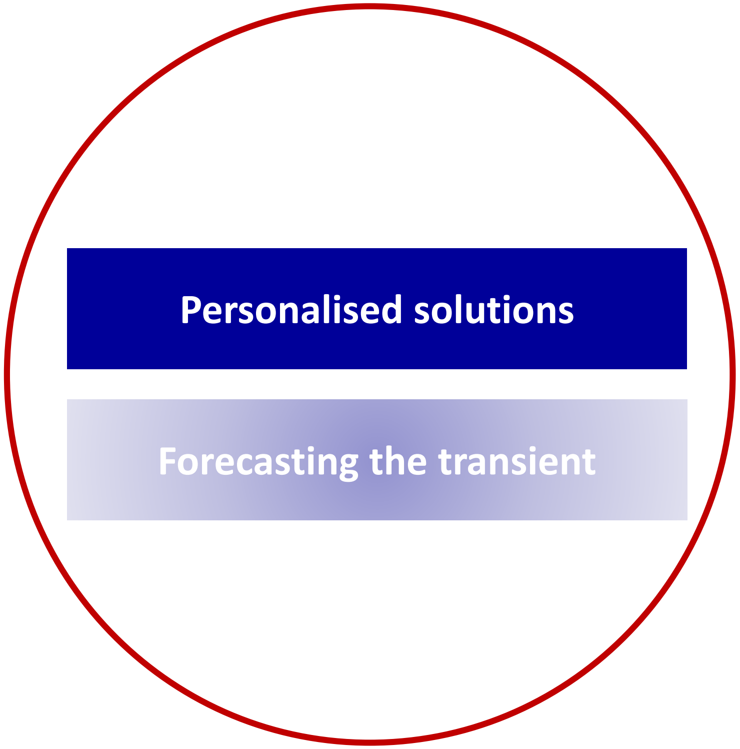
nav

telenor

Skatteetaten

DNB

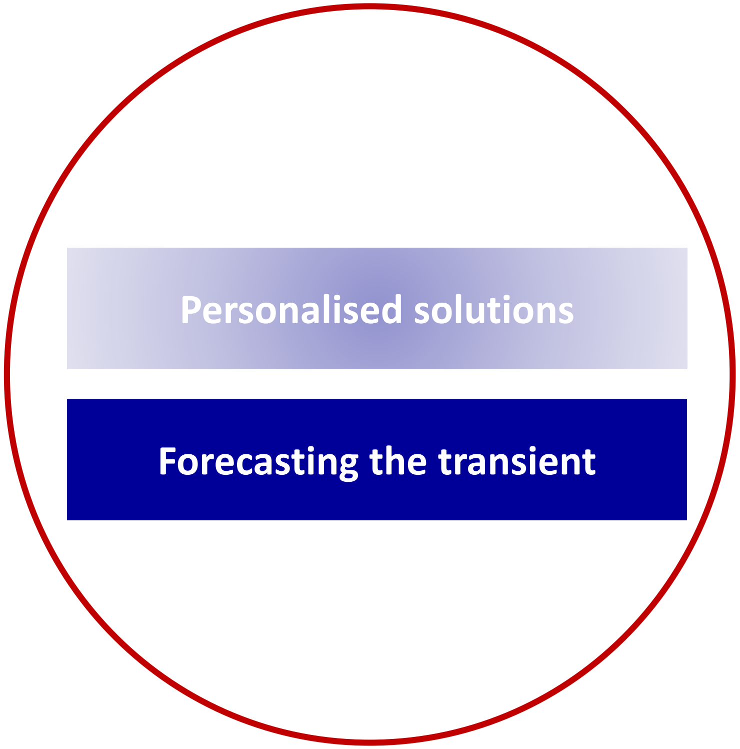
HYDRO



Personalised solutions

Forecasting the transient

- personalised marketing,
- personalised products,
- personalised prices,
- personalised risk assessments,
- personalised fraud assessment,
- personalised screening,
- personalised therapy,
- personalised patient safety,
- individualised maintenance schemes,
- individualised communication



Personalised solutions

Forecasting the transient

High frequency data allow to measure processes in time while they are not in a stable situation, not in equilibrium.

- **Predict** the dynamics, the next events.
- Optimal **intervention** while real time monitoring.
- **Causa**l understanding of the factors which affect the process.

- Move away from operations based on average and typical behaviour towards individualised actions.

- Intervene in real time and *while it happens,* to improve performance and gain control.

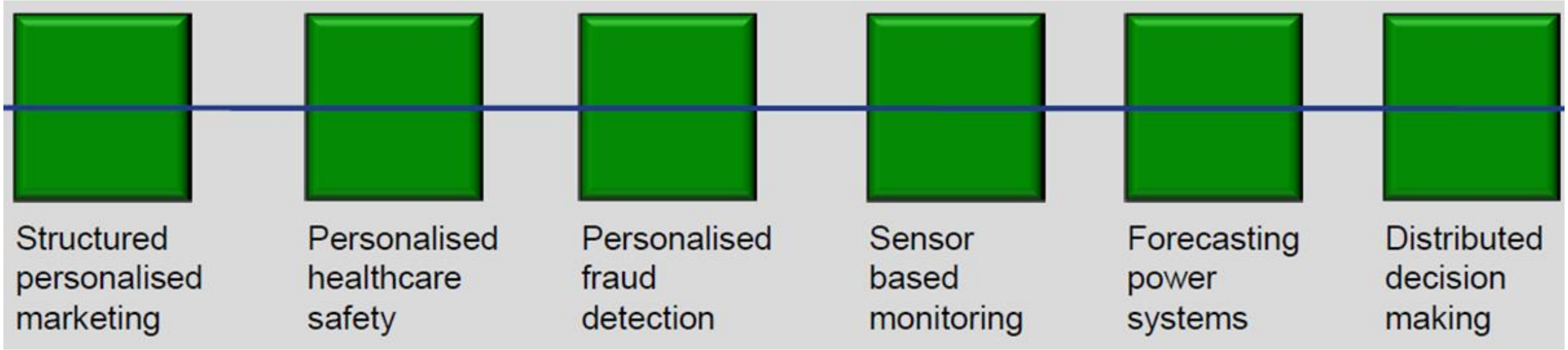
- data must be **integrated** with soft and hard substantive industrial knowledge;

- industrial decisions and risk assessment require a precise quantification of the **uncertainty** inherent in predictions and segmentations;

- interventions require an understanding of the **causal** mechanisms behind behaviours.

→ **MODEL BASED ANALYTICS**

6 Innovation Objectives



Structured personalised marketing

Personalised healthcare safety

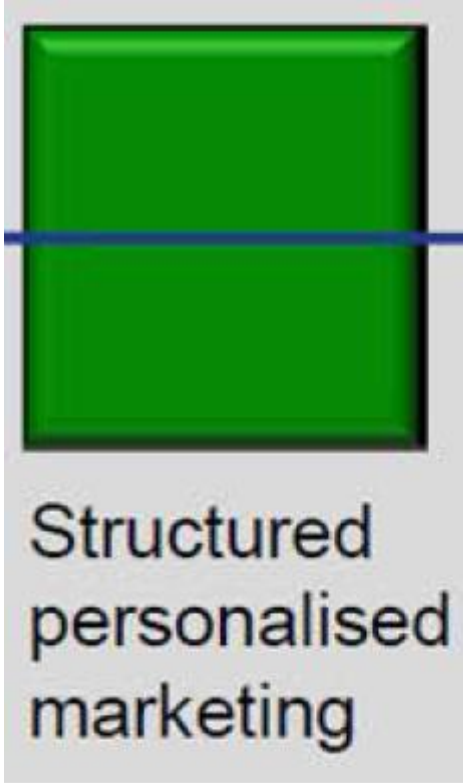
Personalised fraud detection

Sensor based monitoring

Forecasting power systems

Distributed decision making

BigInsight


Structured personalised marketing

1. Personalised communication

- Multiple communication channels with customers / users
- To and from
- Determine for each customer, the best channel

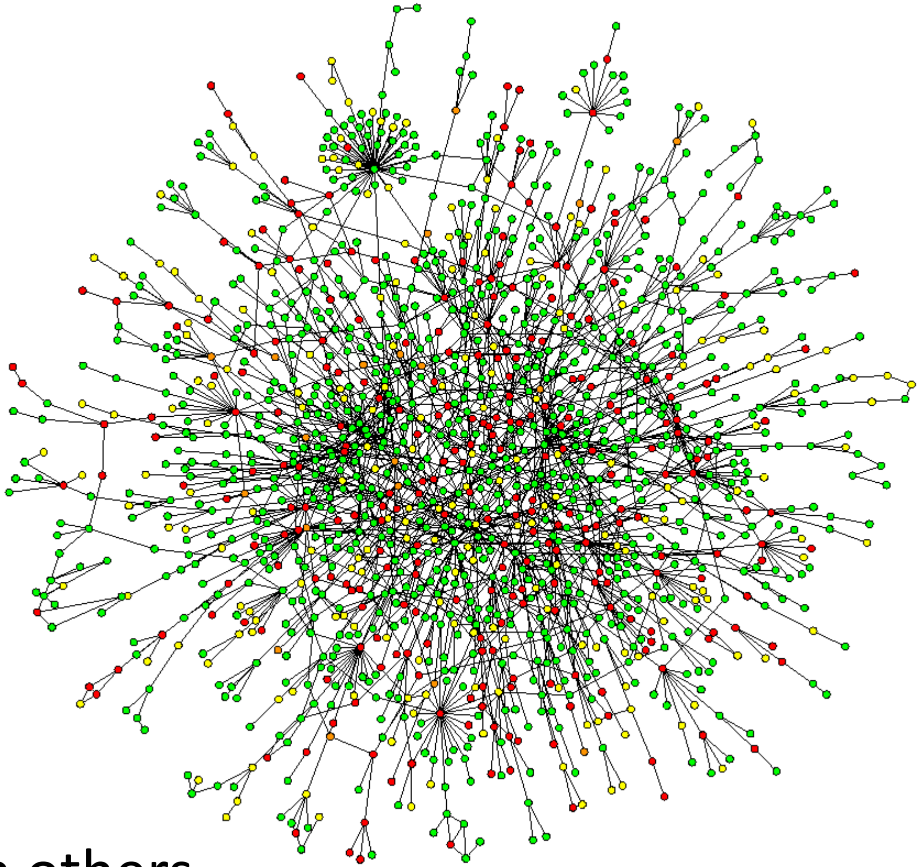
- Challenges:
  - heterogeneous data, from demography to browsing history, from text to phone calls
  - high dimensional (factor selection)
  - counterfactual

2. Personalised marketing

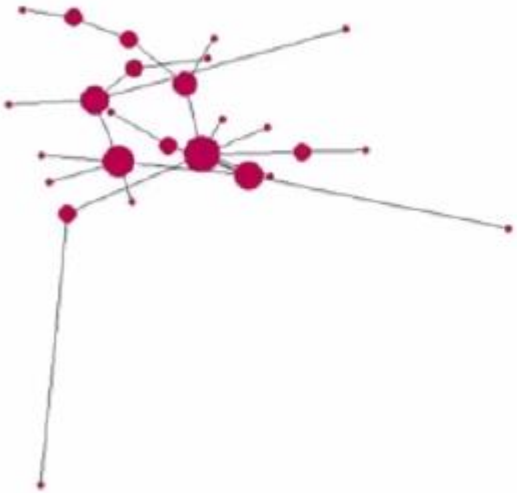
- Favour specific customer behaviour (buy, loan, no churn, …)
- Interventions: recommend, change price/product, new product …
- Model customer behaviour,
- Exploit known/latent relations between customers (network)
- Simulation: play interventions, observe behaviour *in silico*

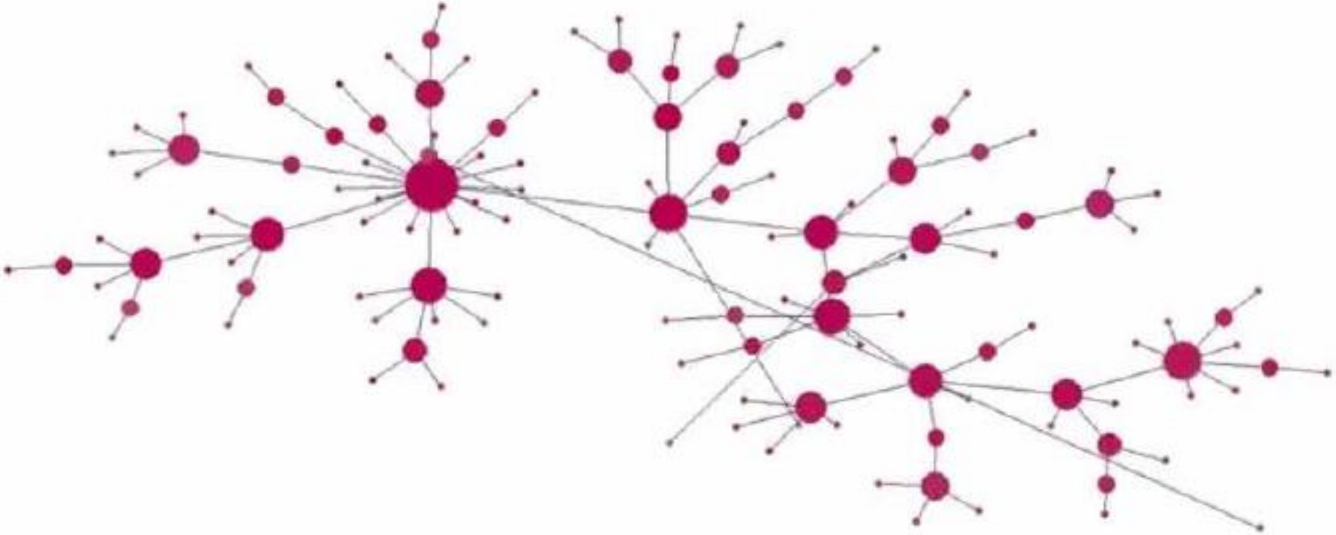
- Challenges:
  - heterogeneous data, incl. ratings and network topology
  - high dimensional (factor selection)
  - counterfactual
  - real time
  - prediction, as early as possible

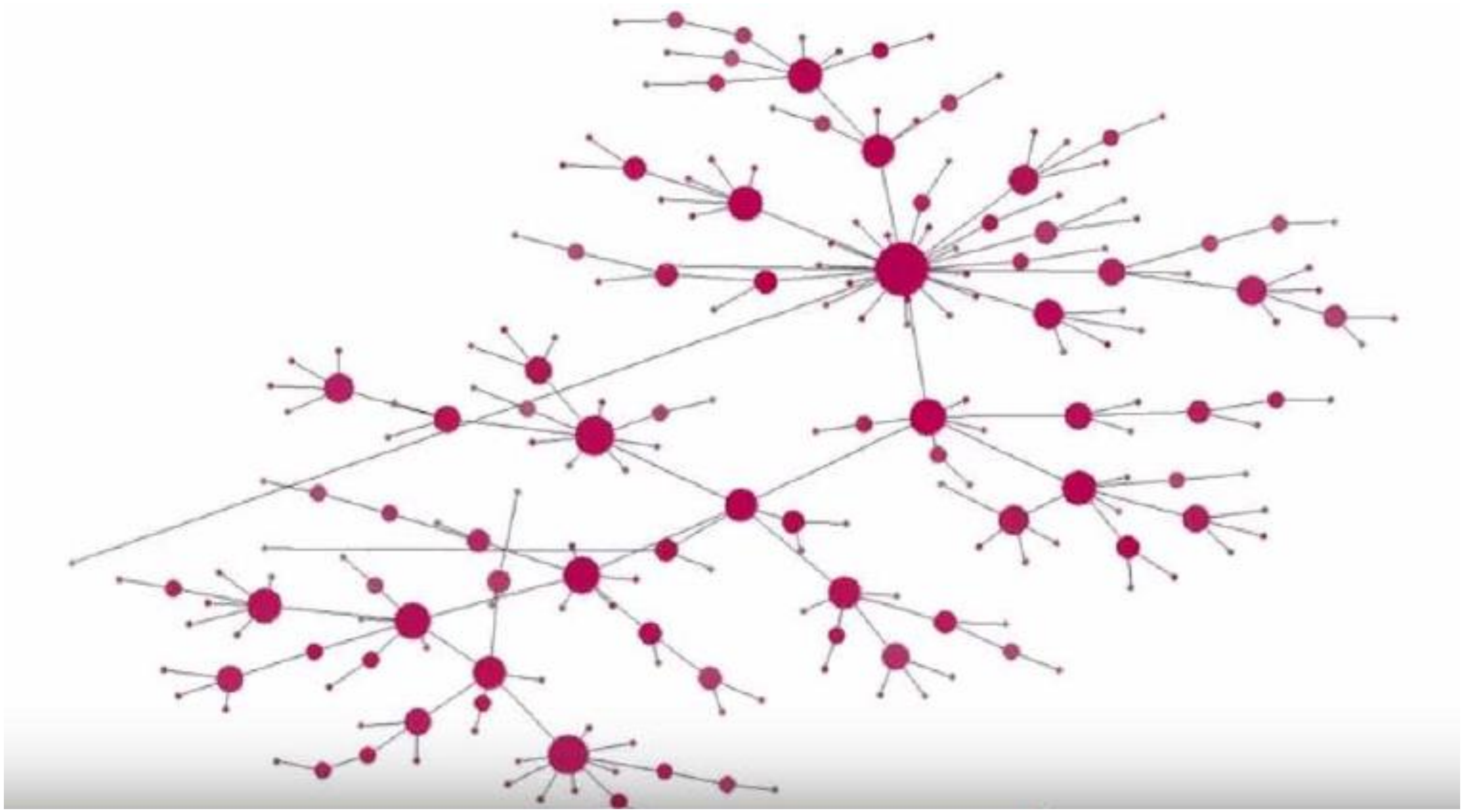
vipps



- Service to send money to each others
- app
- Nodes = people
- Edges = transaction
- New people; new transactions







GROWTH is change

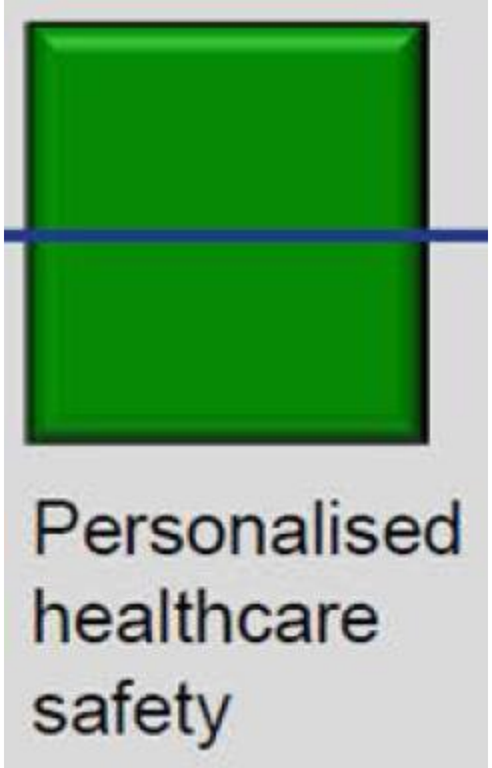
GOAL:

How does the network grows?

  - Understand dynamics
  - Predict growth
  - Help marketing = interventions

- Some information on people
- Transactions

- Non-identifiable data


Personalised healthcare safety

1. Personalized cancer statistics

- Cancer registries publish survival statistics by gender, stage, cancer site
- New clinical registries include treatment and later events.
- Can be linked to other registries on comorbidity, income, education.
- Produce survival predictions using all such individualized information

- Challenges:
  - heterogeneous data
  - high dimensional (factor selection)
  - counterfactual
  - prediction, with uncertainty

Cases and data from Kreftregisteret

2. Personalized cancer treatments

- Cancer is a collection of different diseases, that call for different treatment.
- Genomic profile determines which treatment works.
- There are extremely many treatment combinations and plans.
- Produce a computer model of cancer, adaptable to each genome.
- Simulate all treatments and determine the optimal for each patient

- Challenges:
  - Approximating complexity
  - high dimensional (factor selection)
  - Fast computation
  - prediction, with uncertainty

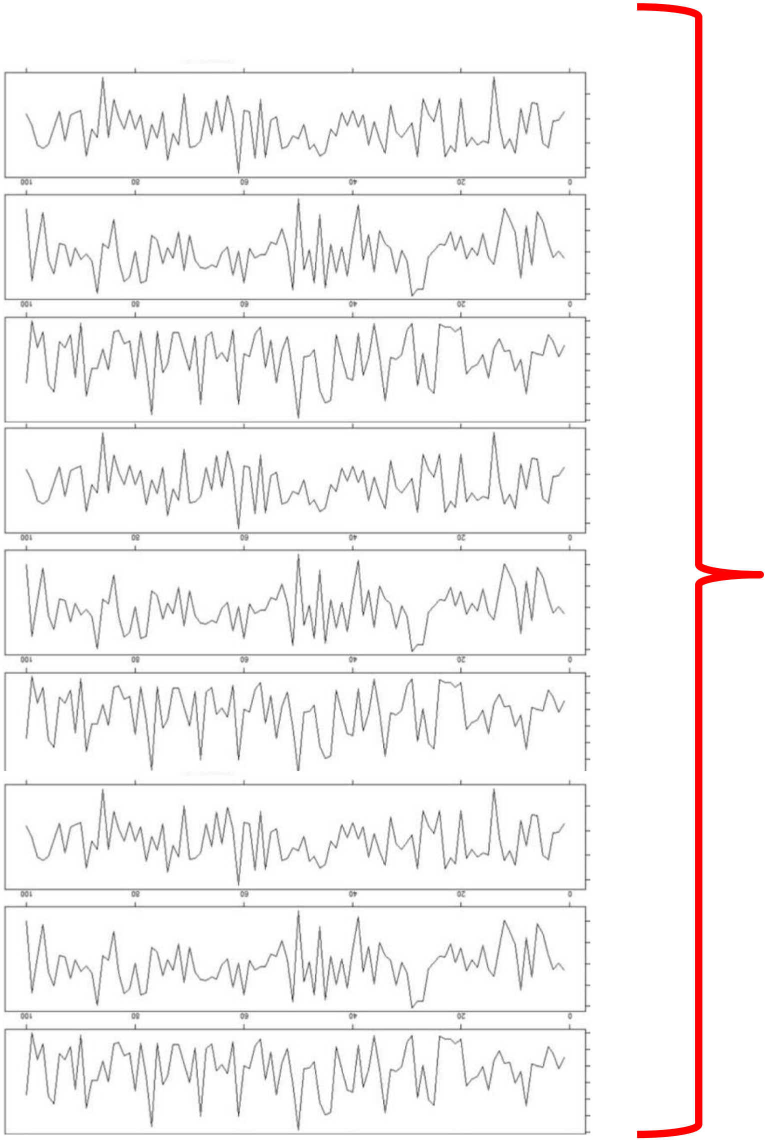
Cases and data from OUS

3. Healthcare safety management

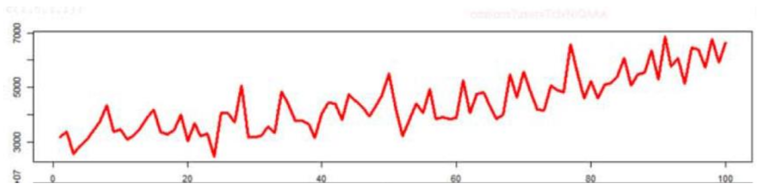
- Electronic Health Records
- Predict harms to patients (or efficiency of hospitals) from EHR, hospital data and many other variables
- Understand causes and thus allow prevention or mitigation

- Challenges:
  - Approximating complexity
  - high dimensional (factor selection)
  - Real time computation
  - prediction, with uncertainty

Cases and data from DNV-GL and OUS



harm



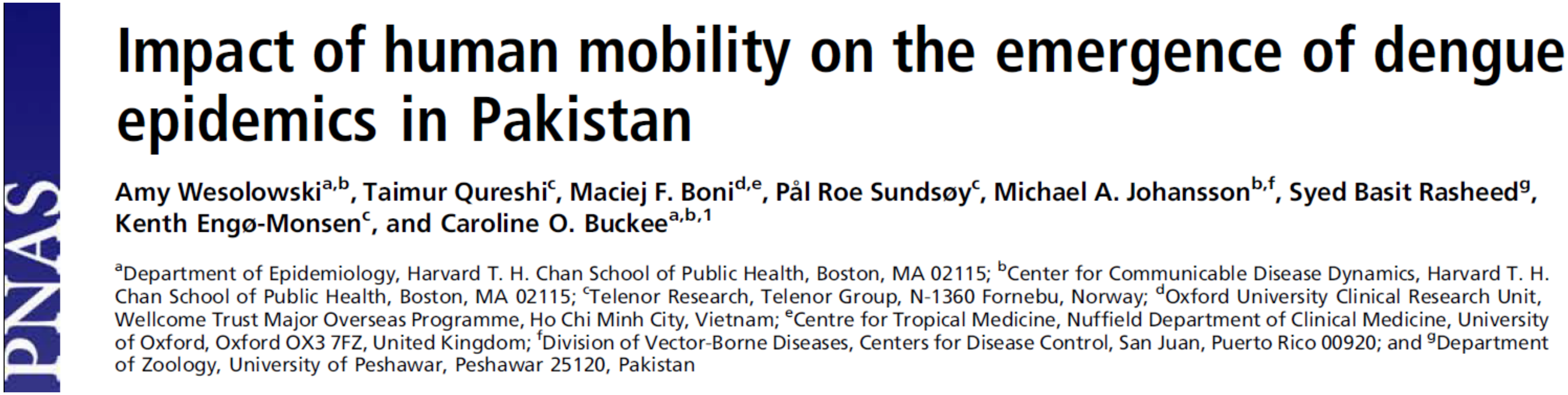
- detection
- prediction

4. Telecom data for epidemics control

- Infectious diseases spread by social mixing and mobility
- Mobile phone geo-temporal information allows to observe movements of people and contacts.
- Model epidemic spread based on estimated mobility
- Optimise vaccination strategies

- Challenges:
  - approximating complexity
  - high dimensional data
  - real time computation
  - uncertainty

Cases and data from Telenor and NIPH



Impact of human mobility on the emergence of dengue epidemics in Pakistan

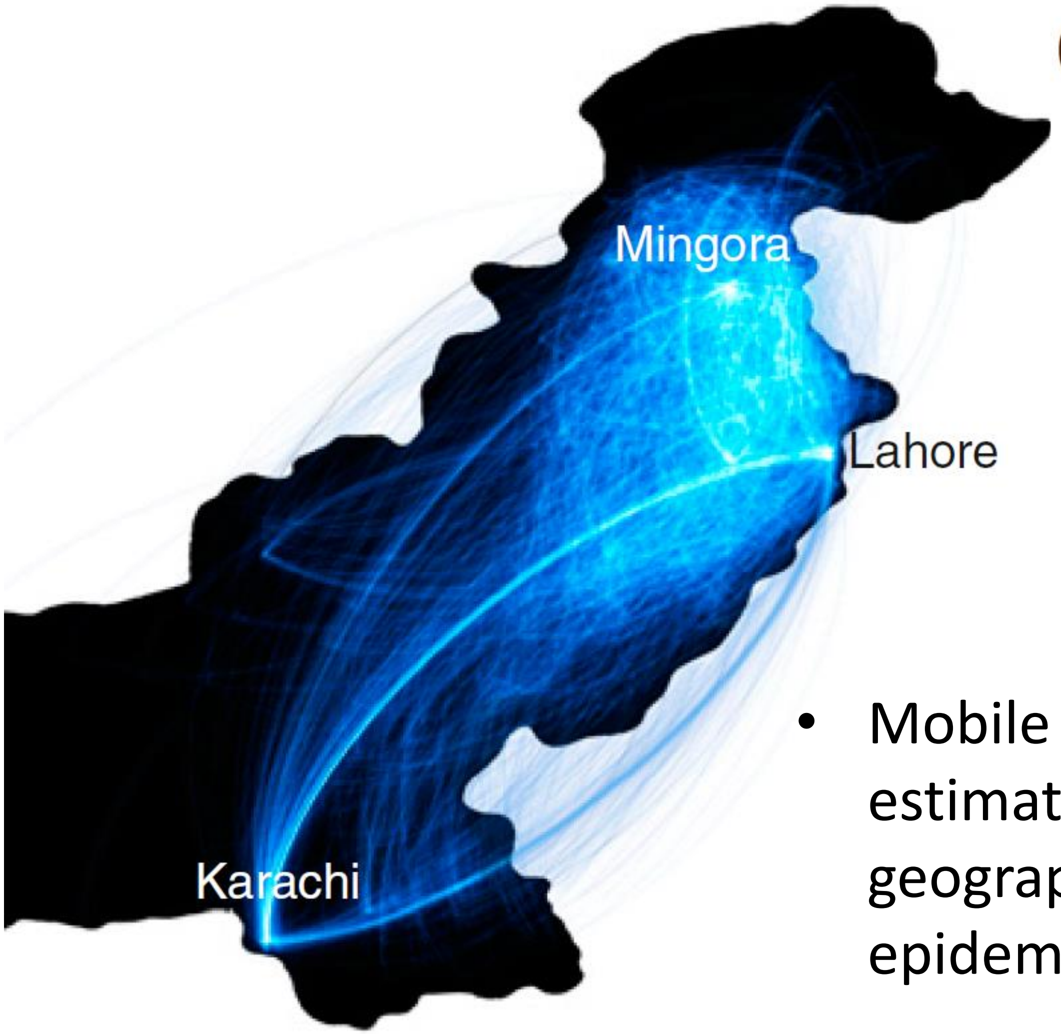
Amy Wesolowski[a,b], Taimur Qureshi[c], Maciej F. Boni[d,e], Pål Roe Sundsøy[c], Michael A. Johansson[b,f], Syed Basit Rasheed[g], Kenth Engø-Monsen[c], and Caroline O. Buckee[a,b,1]

[a]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [b]Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [c]Telenor Research, Telenor Group, N-1360 Fornebu, Norway; [d]Oxford University Clinical Research Unit, Wellcome Trust Major Overseas Programme, Ho Chi Minh City, Vietnam; [e]Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 7FZ, United Kingdom; [f]Division of Vector-Borne Diseases, Centers for Disease Control, San Juan, Puerto Rico 00920; and [g]Department of Zoology, University of Peshawar, Peshawar 25120, Pakistan

- dengue data from a large outbreak in Pakistan in 2013

GOAL:

Epidemiological model of dengue virus transmission based on climate and mobility data from ~40 million mobile phone subscribers.

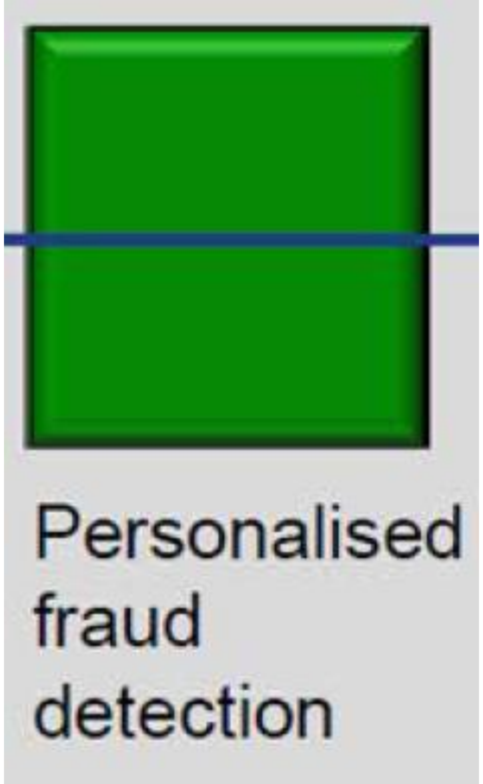


- Mobile phone-based mobility estimates predict the geographic spread and timing of epidemics

- Fine-scale dynamic risk maps for epidemic preparedness.

What data?

- Mobile phone subscribers location (nearest phone tower).
- Daily locations and movements were aggregated to measure travel between 356 small areas



Personalised fraud detection

1. Ensemble methods for personalised fraud detection

- Develop new approach to fraud/money laundering detection
- Depends on many covariates, and their interactions
- Combining results from many methods, to exploit each strength.

- Challenges:
  - Merging of different predictions
  - high dimensional data
  - few known cases
  - efficient computation
  - uncertainty

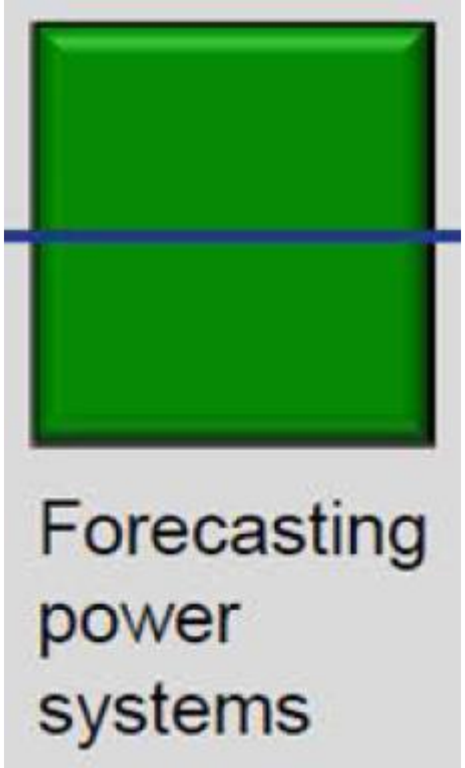
Case and data from Skatteetaten, DNB, Gjensidige, NAV.

2. Network analysis for personalised fraud detection

- Fraud is viral, spreading directly or indirectly from one fraudster to others.
- Network relations between persons, businesses and groups thereof.
- Understand how these networks evolve over time
- Exploit for better fraud forecasts
- Allow other preventive interventions

- Challenges:
  - approximating complexity
  - high dimensional data
  - Multiple relations
  - uncertainty

Case and data from Skatteetaten, DNB, Gjensidige, NAV.

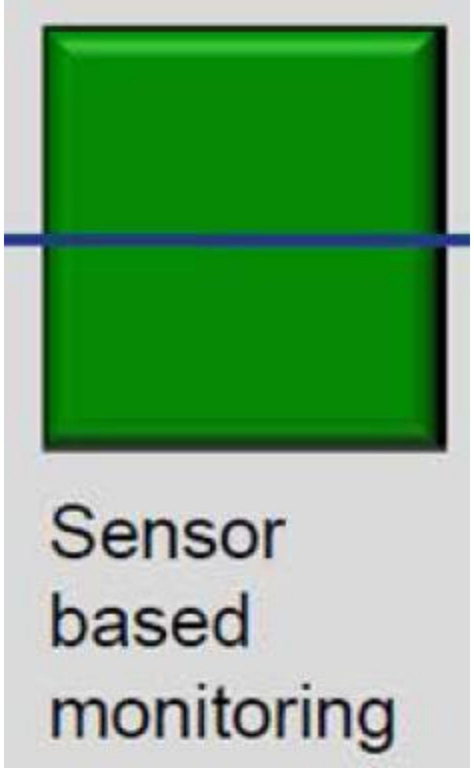

Forecasting power systems

2. Optimal power match for smart grid

- The smart grid combines the optimization of use and production of electricity with forecasted prices.
- In planning each device’s demand for electricity, the future state of the system must be predicted.
- Stochastic dynamic optimisation

- Challenges:
  - Complex system
  - high dimensional data
  - real time computation
  - uncertainty

Case and data from DNV-GL.



Sensor
based
monitoring

1. Condition monitoring: Fault and anomaly detection and prediction

- Sensors on ships control operations
- Detect as fast as possible faults and anomalies
- Predict faults as early as possible
- Filter away effects of weather, sea conditions … on the sensors
- Exploit knowledge about machines and sensor networks
- Quantification of uncertainty to control false warnings

- Challenges:
  - approximating complexity
  - high dimensional data
  - real time computation
  - uncertainty

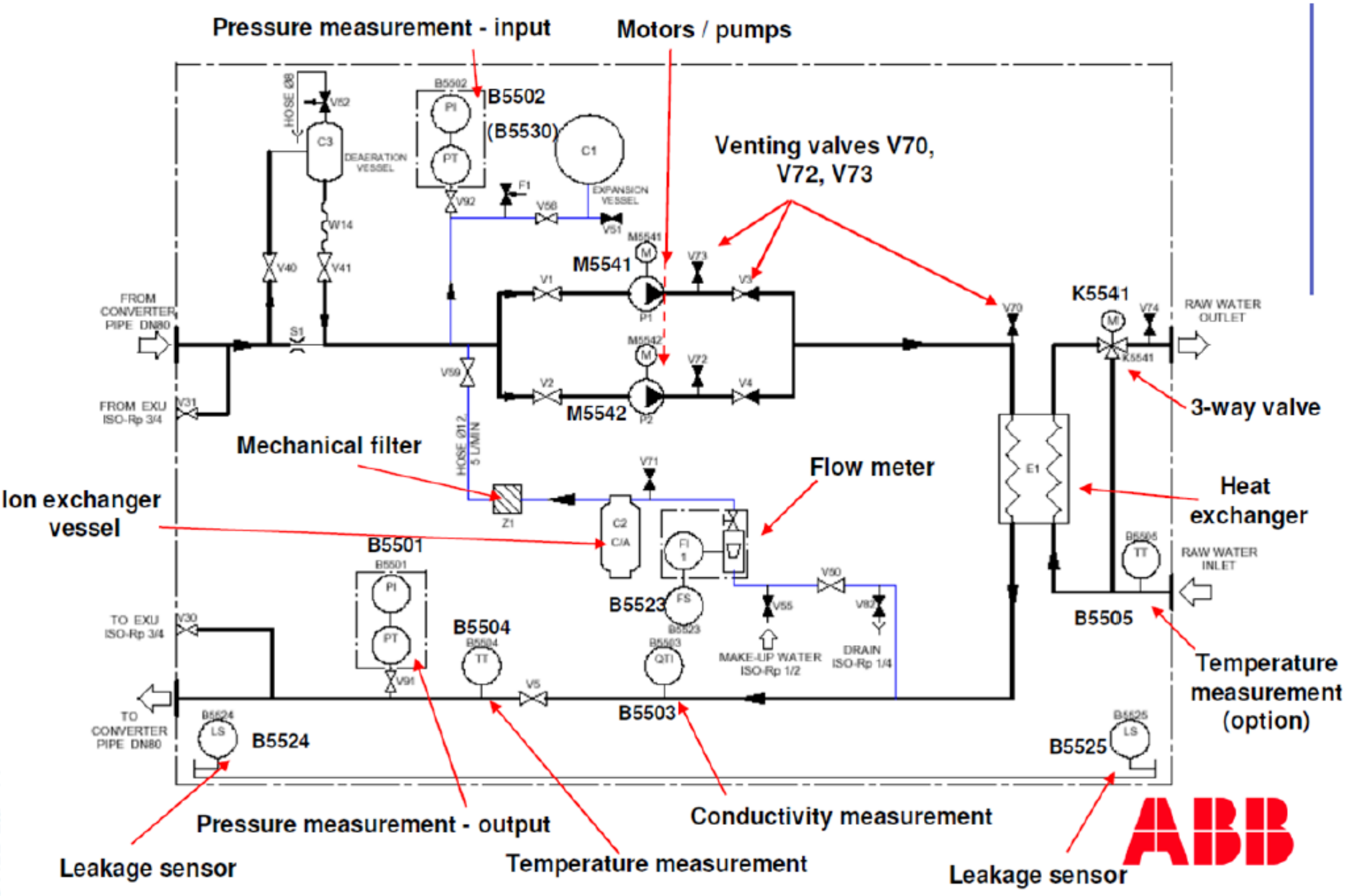
Case and data from ABB and DNV-GL

2. Performance monitoring and optimisation

- Develop a multisensor multiscale approach to describe and optimise performance of the ship operation under given weather and ocean conditions and other operation constrains.

- Challenges:
  - approximating complexity
  - high dimensional data
  - real time computation
  - uncertainty

Case and data from ABB and DNV-GL



Pressure measurement - input

Motors / pumps

B5502

(B5530)

Venting valves V70, V72, V73

M5541

K5541

M5542

3-way valve

Mechanical filter

Flow meter

Ion exchanger vessel

Heat exchanger

B5501

B5523

B5504

B5505

Temperature measurement (option)

B5503

B5524

B5525

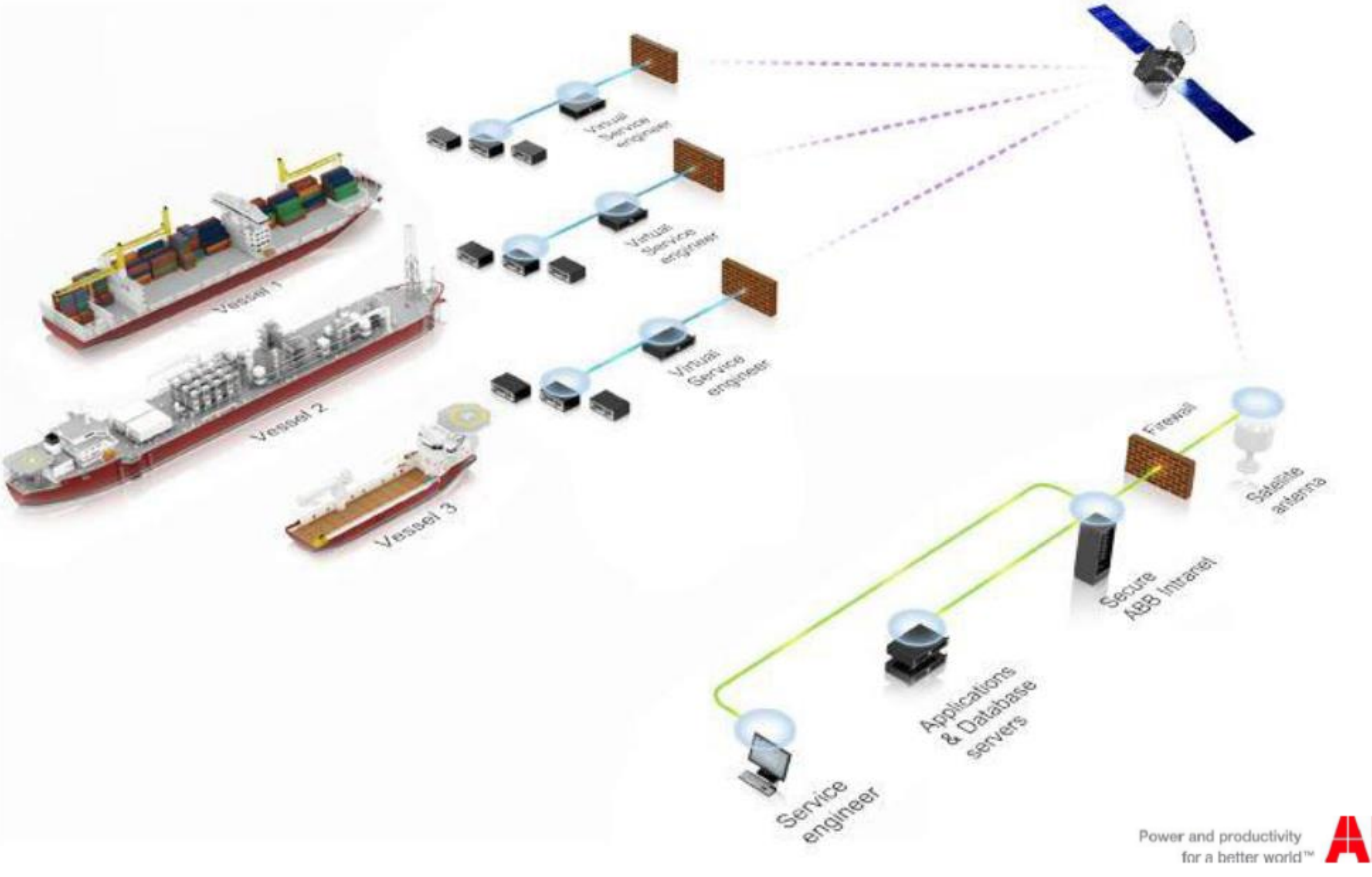
Conductivity measurement

Pressure measurement - output

Leakage sensor

Temperature measurement

Leakage sensor



Power and productivity
for a better world™

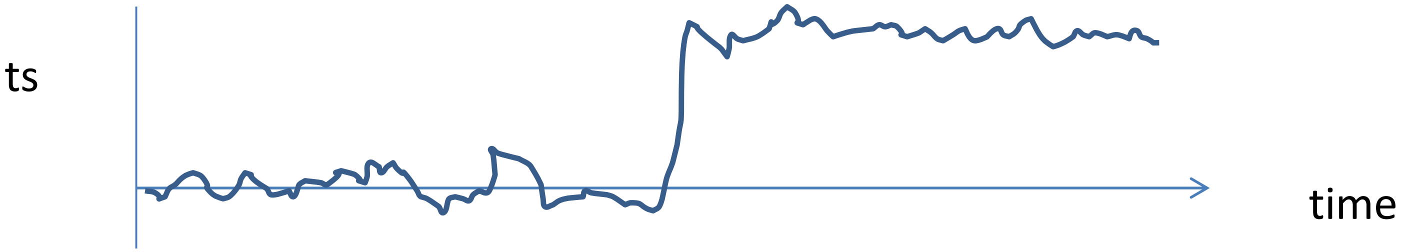
■ Surprise and changepoint prediction allows control.

- Measures of surprise quantify the level of incompatibility of data with a given model, without any reference to alternatives.

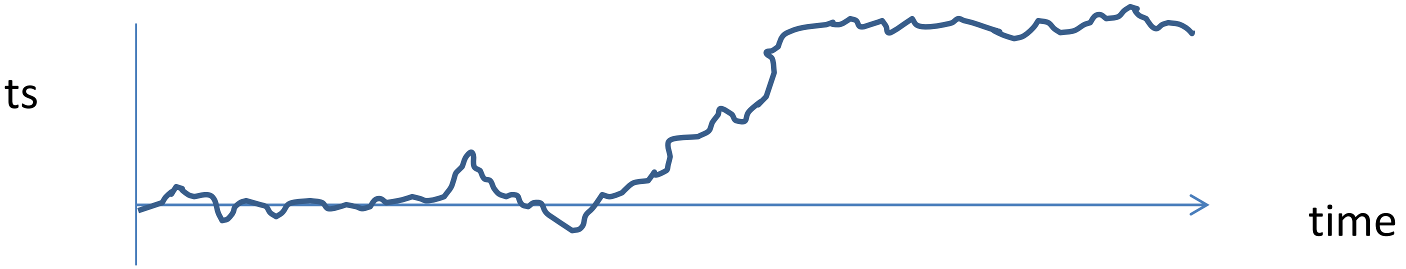
- Surprise plays an important role in dynamic situations, where the reference is the past trajectory.

- Change point prediction, rather than locating changepoints after occurrence.

CHANGEpoint --- Abrupt discontinuity of a feature of a time series



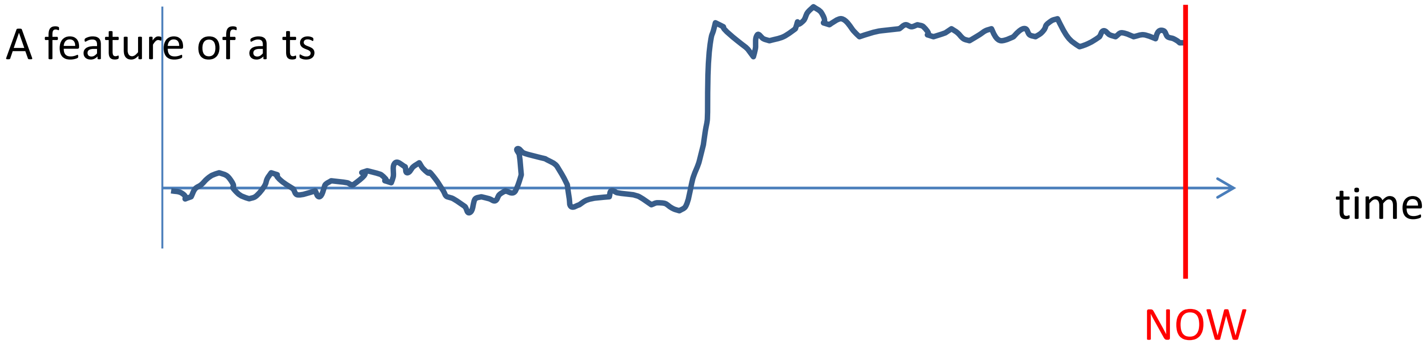
CHANGEregion --- slow change in some feature, from a status to an other one



What is changing?

- Feature of ts: mean, st dev, spectrum, first derivative, second derivative, etc etc (VERY many options!)

CHANGEpoint **DETECTION**



Looking back to the past, find as fast as possible:

- If there has been a CP
- Where it is?
- What is changing?
- How big is the change?

with uncertainty

Univariate: one ts, one feature at the time

Multivariate: many ts's or many features of one ts, as a time

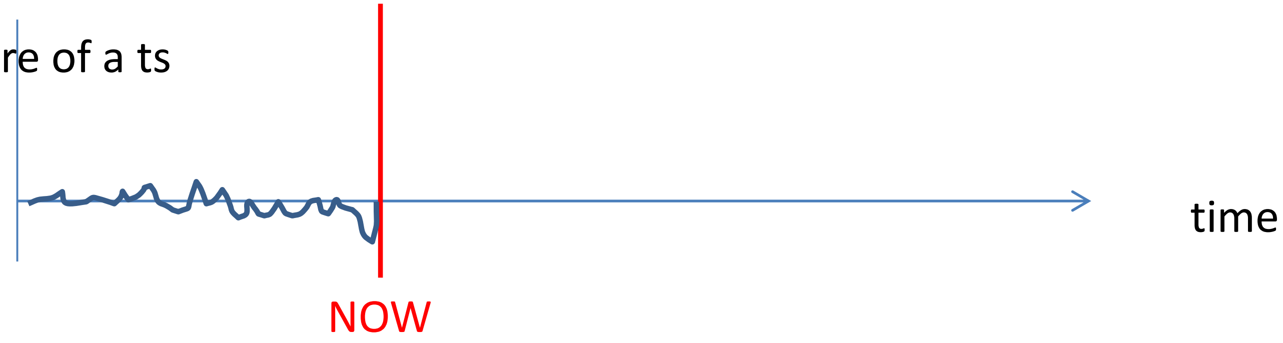
Dependence between time series and CPs

Coherence of CPs

[There are methods, but much remains to be done]

CHANGEpoint **PREDICTION**

A feature of a ts



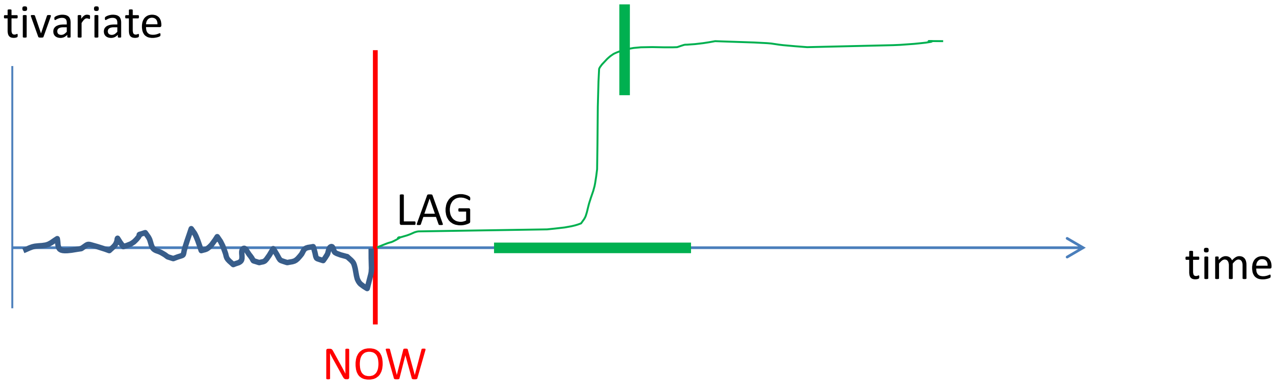
Looking **forward**, find as early as possible:

- If there will be a CP
- Where it will be?
- What will change changing?
- How big will the change be?

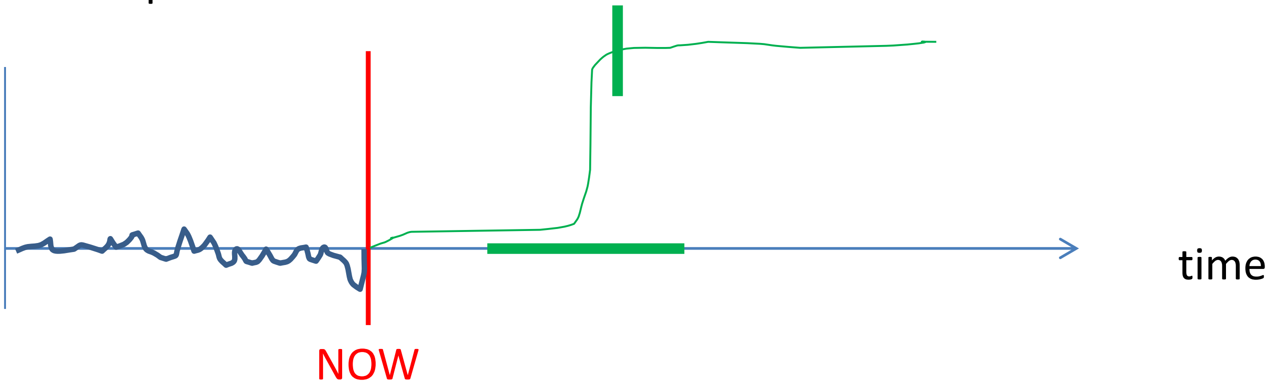
with uncertainty

Univariate

Multivariate



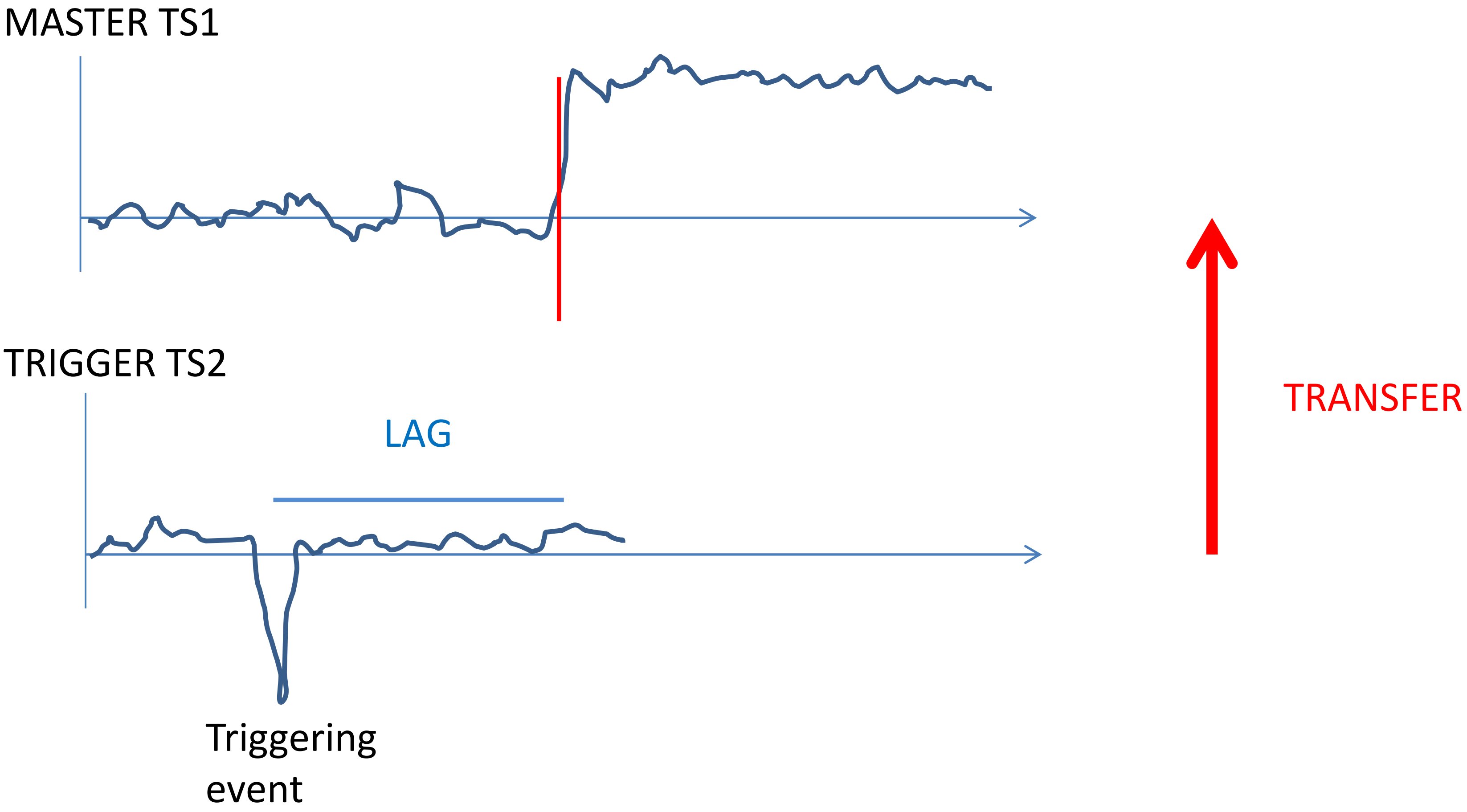
CHANGEpoint PREDICTION



To predict we must understand/learn the mechanism that generates the CPs

- Experiments in the lab
- Statistics
  - A series with a lot of CPs, to learn the process under which the occur (we do not need to know why the happen, but can predict them)
  - One ts causes a CP in another ts. TRIGGERING

1



MASTER TS1

TRIGGER TS2

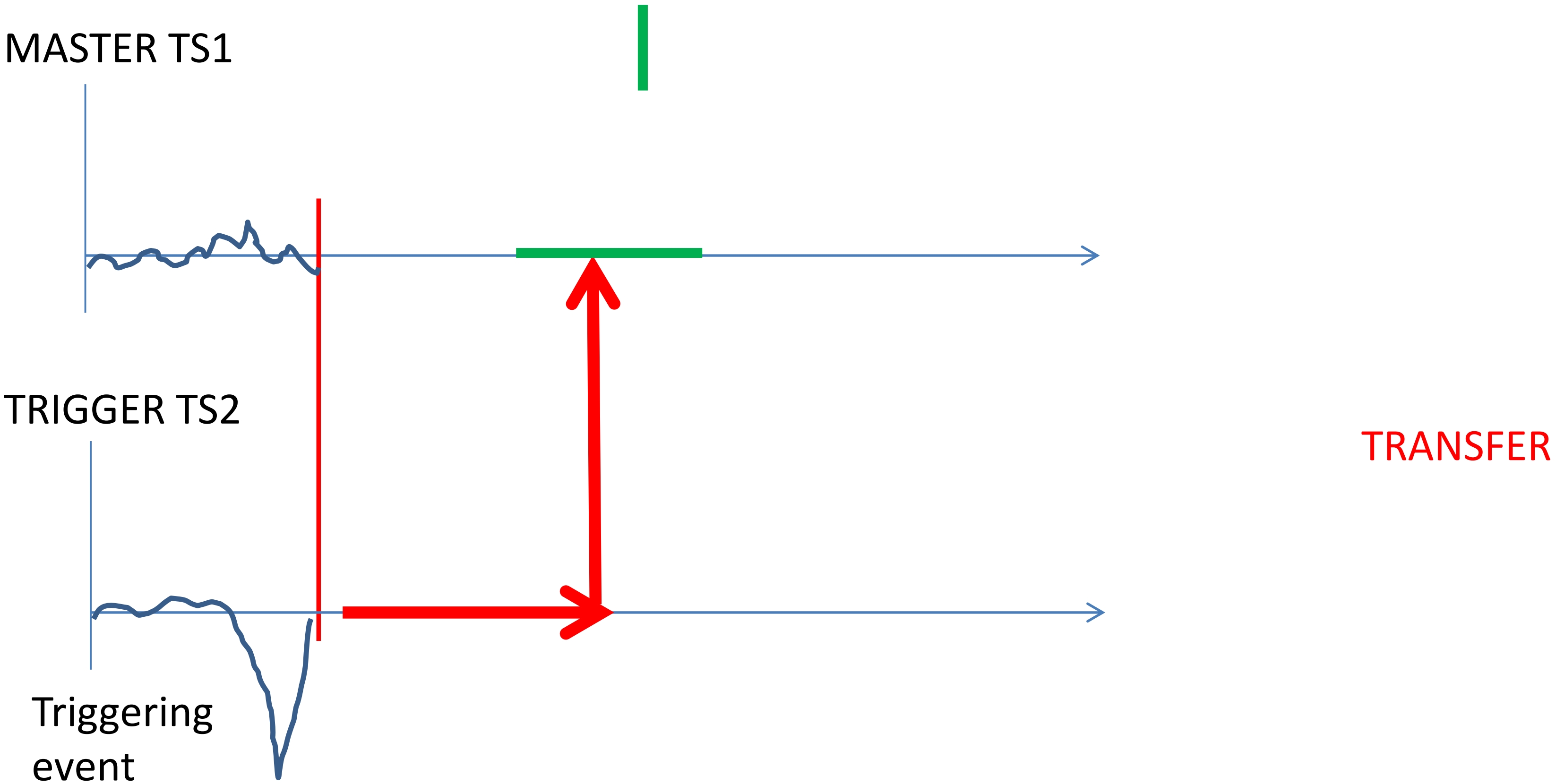
LAG

Triggering event

TRANSFER

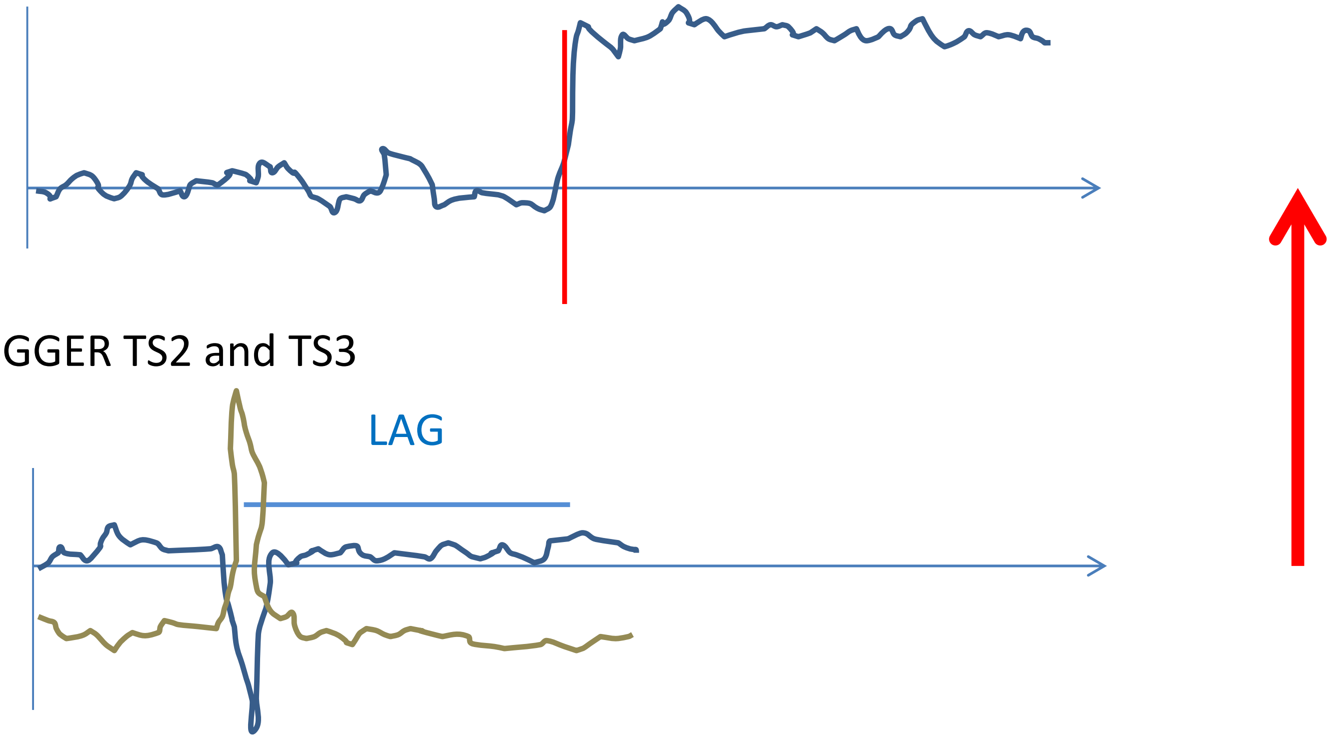
- Assume: We know which ts is master and which ts is trigger
- Maybe we know which two *features* are involved.
- Maybe not: find which feature of TS2 triggers TS1
- Estimate the lag and transfer function, with uncertainty

1



- At the time point when the triggering event in TS2 is detected, predict the CP in TS1.

2



MASTER TS1

TRIGGER TS2 and TS3

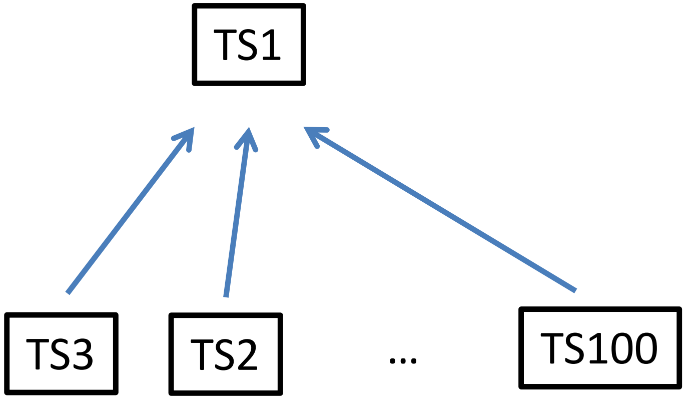
LAG

TRANSFER

Triggering events

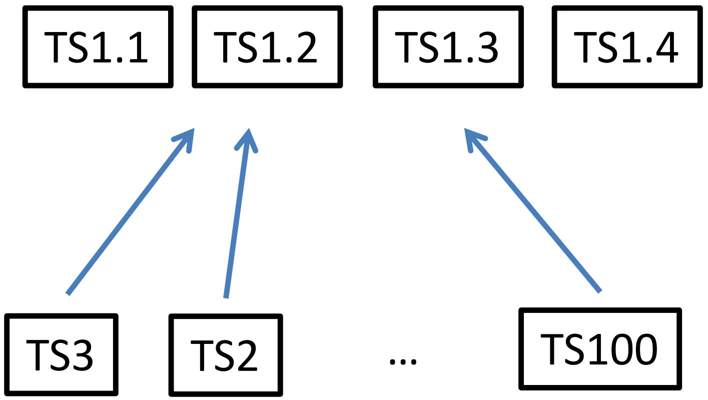
- Assume: We know which ts is master and which TS2 and TS3 are triggers
- The triggering events might be simultaneous or not, but co-ordinated in some way
- The triggering events do not need to be CPs, but just events which co-occur rarely.
- Estimate the lag and transfer function, with uncertainty

3



Selection of triggers

- Many possible triggers: which ones are really triggering?
- Few – sparsity
- Can be many features, combinatorics.



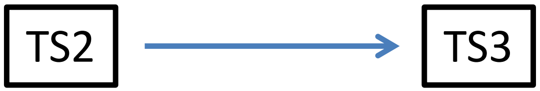
TS1.1 TS1.2 TS1.3 TS1.4

TS3 TS2 … TS100

More than one master

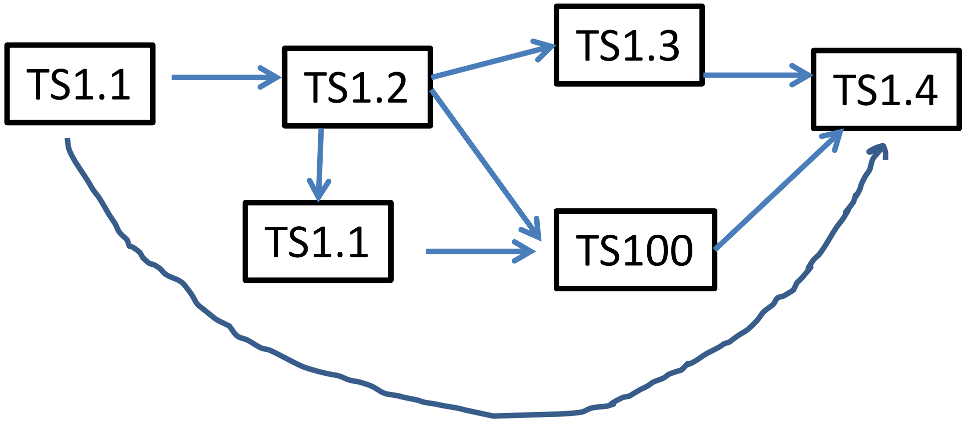
- We know the masters (or features of the same master ts)
- Co-occurrence or coherence of the CP in the masters
- Also: a master can be a trigger of another master.

5


TS2
TS3

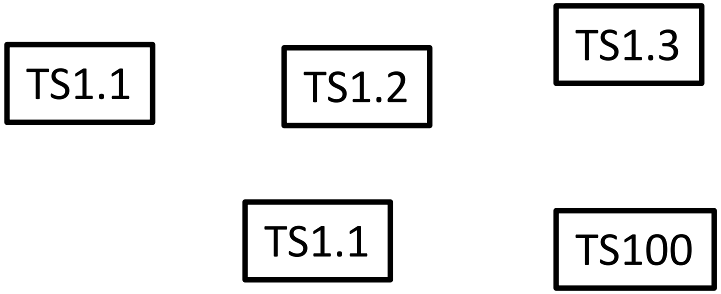
means triggering, or “effect”, with some delay

A known network of ts


TS1.1
TS1.2
TS1.3
TS1.4
TS1.1
TS100

- no master
- multiple triggering paths

Network not known exactly



TS1.1

TS1.2

TS1.3

TS1.4

TS1.1

TS100

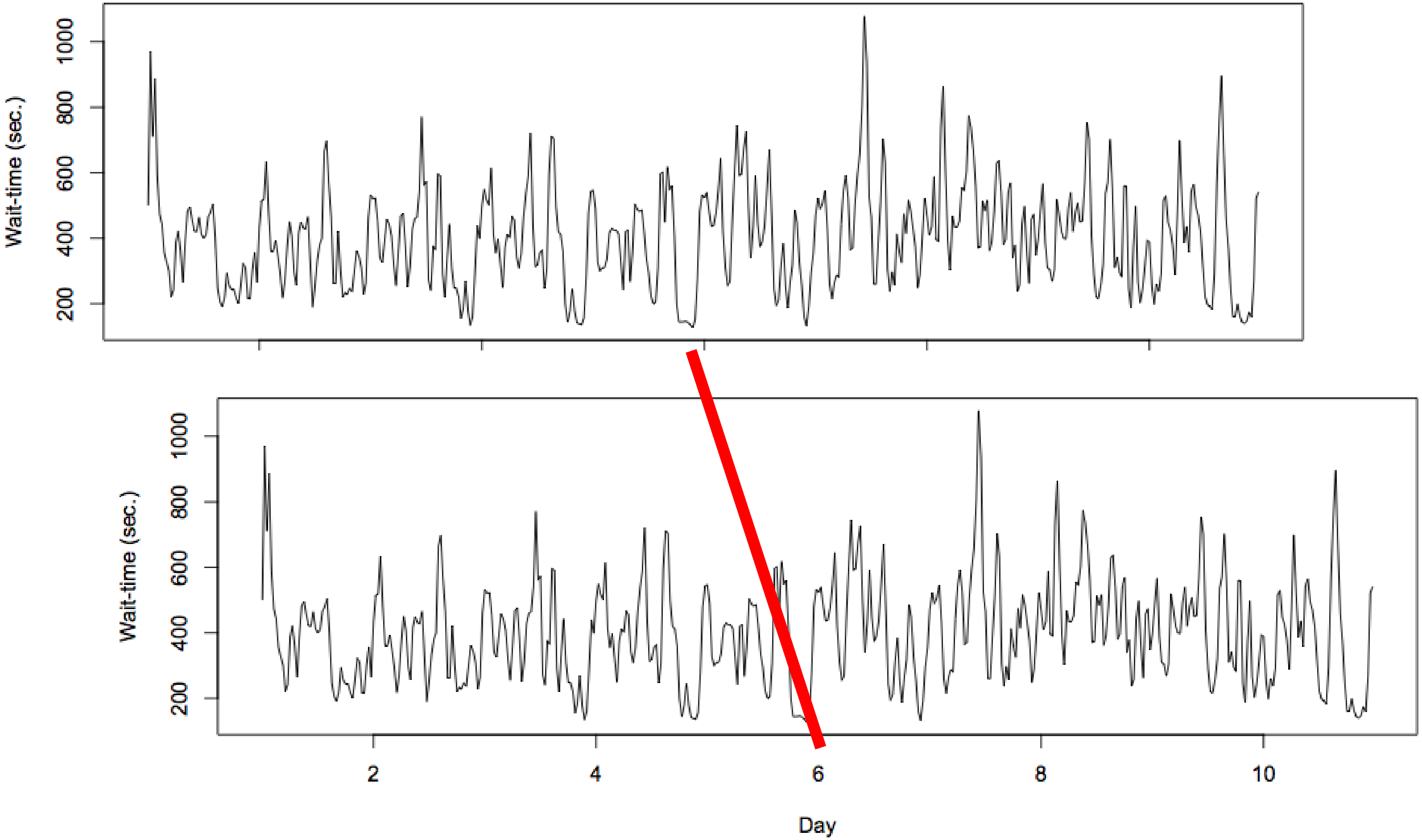
- estimate the network
- for the purpose of CP triggering, via transfer functions (not correlation of the signal)’
- Learn the way CPs are produced in the network, in which order
- After estimation of network, predict future CPs in some ts, given observed triggering events.
- More uncertain

Detecting Changes in Covariance

Jamie-Leigh Chapman[1], Idris Eckley[1], Arnoldo Frigessi[2], Rebecca Killick[1]

X(t)

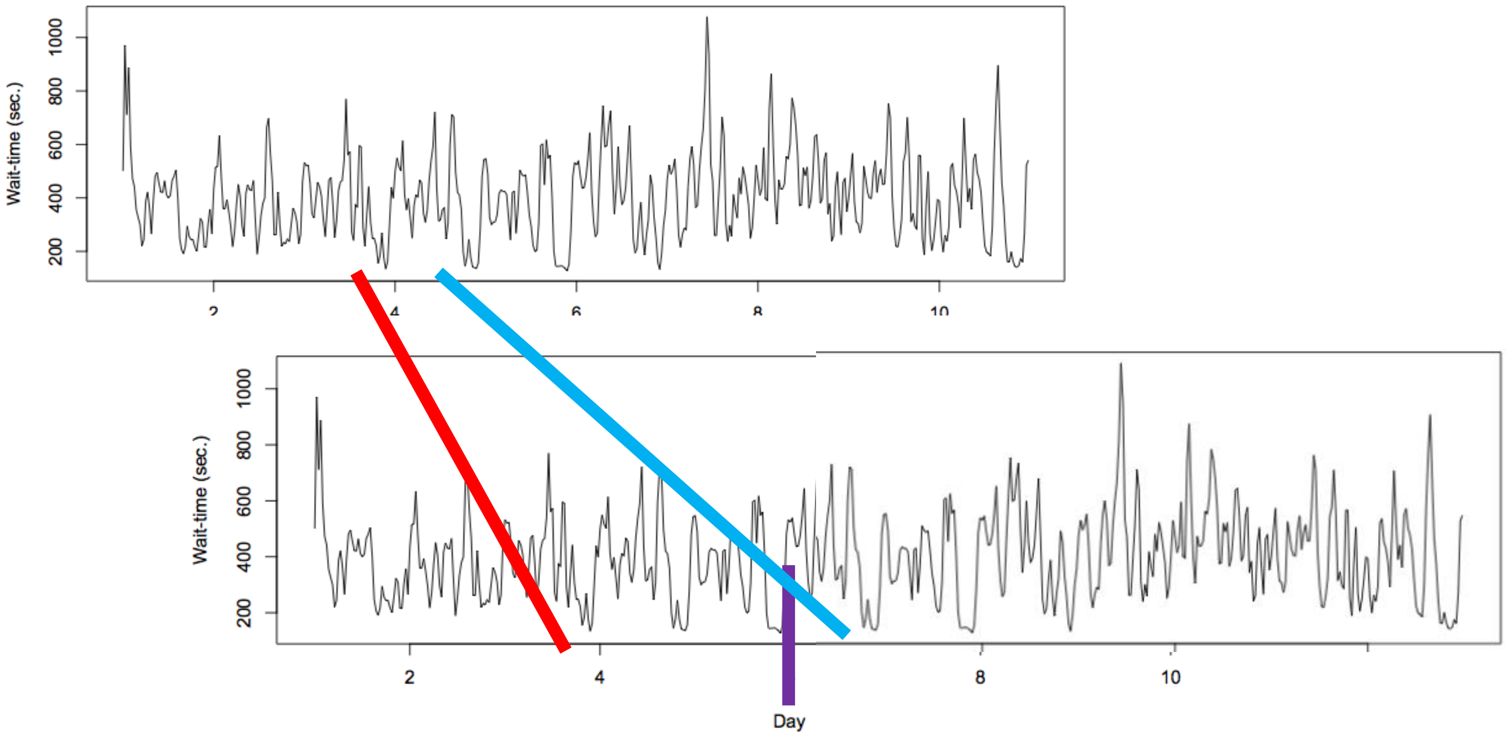
Y(t) = X(t-D)



cov (X(t), Y(t)) is small, while cov (X(t-D), Y(t)) =1

we can look to all cov (X(t-lag), Y(t))

X(t)



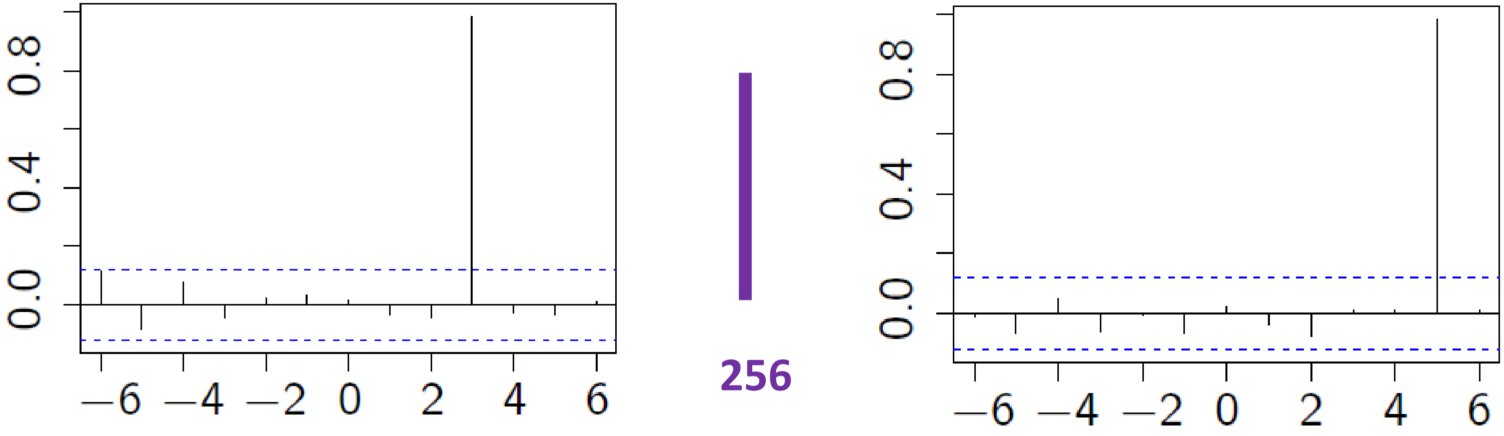
$Y(t) = X(t-3)$ for $t<256$

$\quad = X(t-5)$ for $t>256$

$X(t)$

$Y(t) = X(t-3)$ for $t<256$
$\quad = X(t-5)$ for $t>256$

Plot of cov ($X(t\text{-lag}), Y(t)$) as a function of lag changes in $t=256$



Change in cross covariance

Rolling window estimate of the cross correlation function at lag

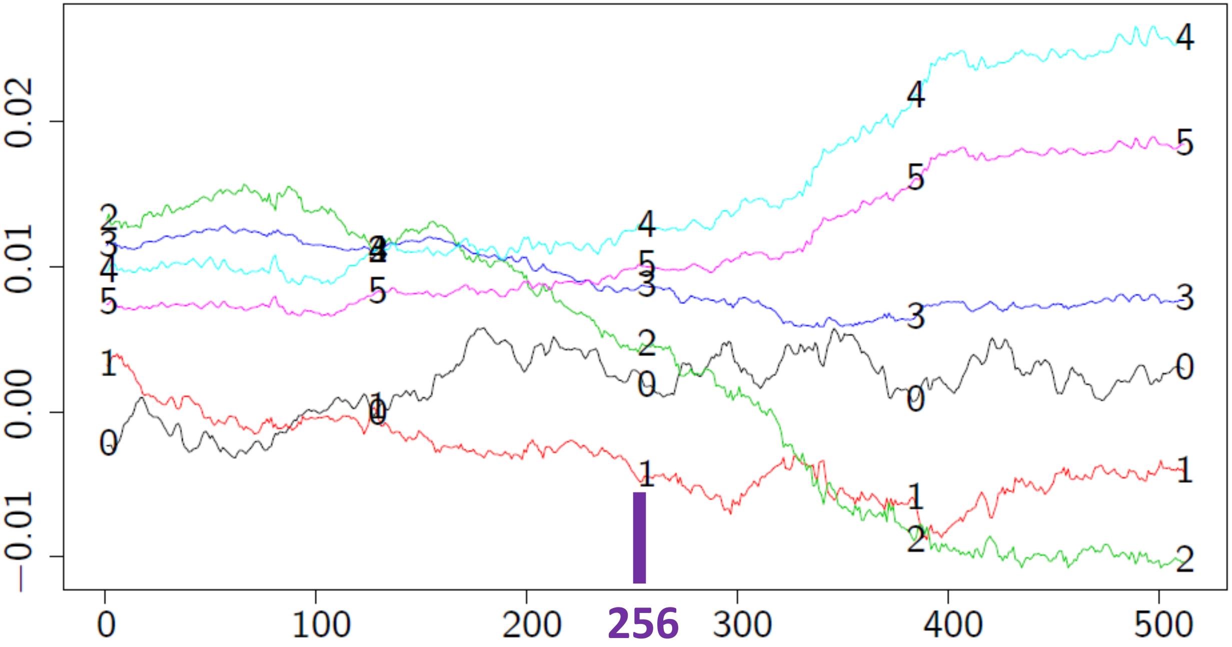
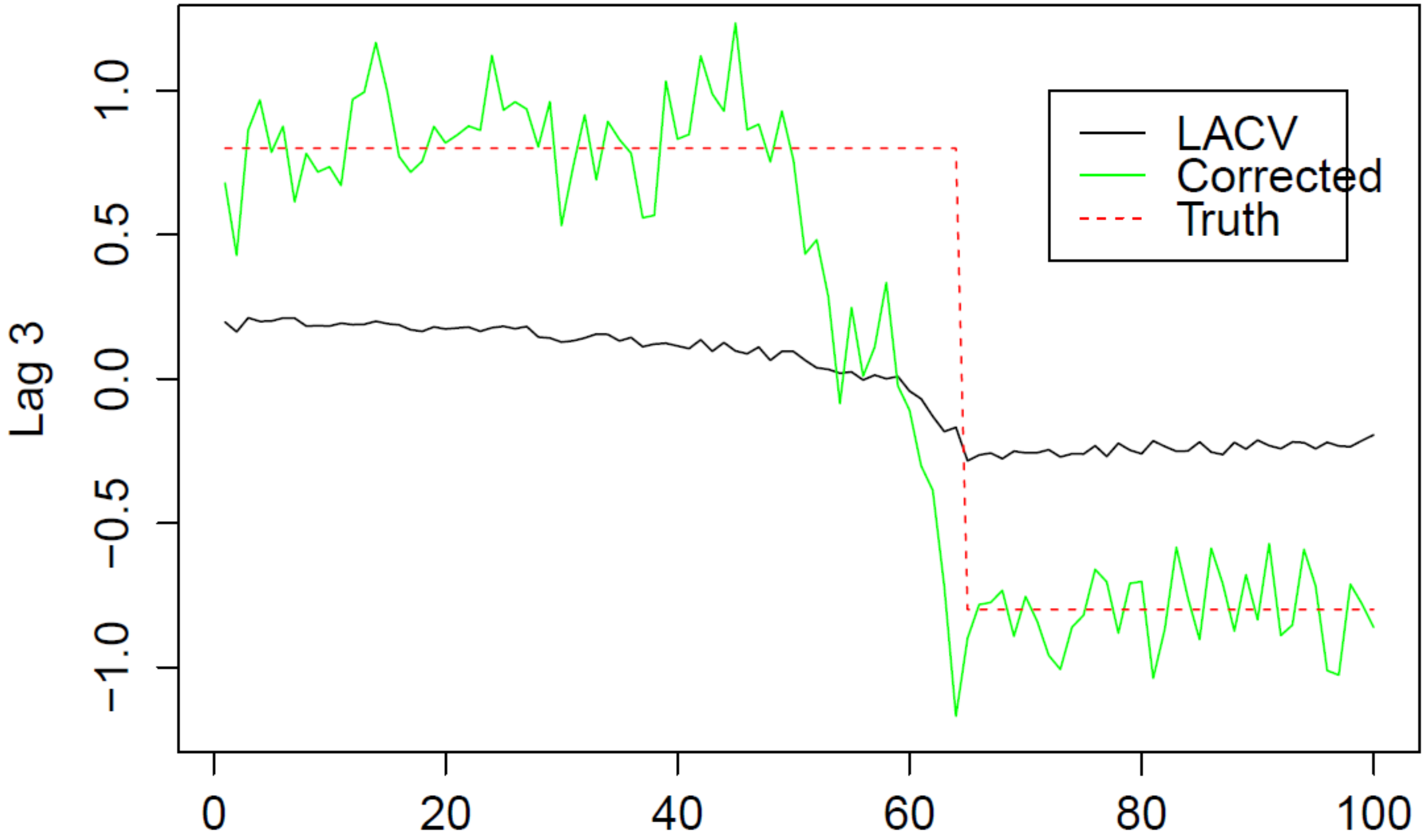


Figure: Local cross covariance function between X and Y for lags zero to five.

- Change in 256 for lag 3 and lag 5
- Also for other lags we see a change!
- Eckley has a method to correct for this bias, and catch just the right lags


LACV
Corrected
Truth
Lag 3
1.0
0.5
0.0
−0.5
−1.0
0
20
40
60
80
100

Some observations from recent literature:

- Multi sensor, multi stream, high dimensional time series, …
- Both in statistics journals and IEEE journals
- Either classical time series approaches, or taking time into consideration in other ways, like sequential analysis, or ignoring it!
- Motivated by signal processing problems, speech recognition, nuclear power plants ++

- Mainly (multiple) change point detection <<as soon as possible>> after it has happened, Expected detection delay (EDD) and Average run length (ARL), CUSUM

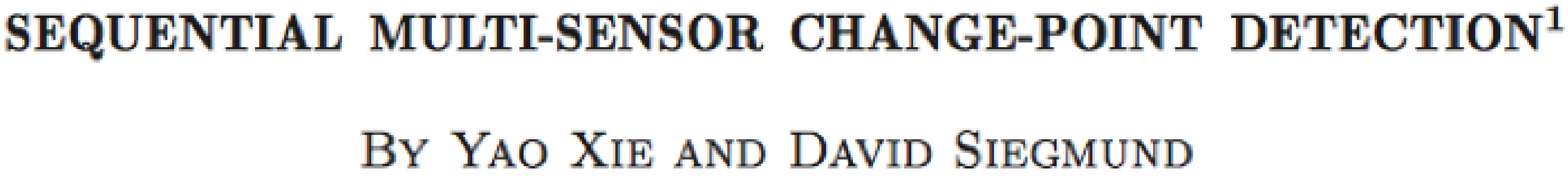
The Annals of Statistics
2013, Vol. 41, No. 2, 670–692
DOI: 10.1214/13-AOS1094

SEQUENTIAL MULTI-SENSOR CHANGE-POINT DETECTION[1]

BY YAO XIE AND DAVID SIEGMUND

Duke University and Stanford University

We develop a mixture procedure to monitor parallel streams of data for a change-point that affects only a subset of them, without assuming a spatial structure relating the data streams to one another. Observations are assumed initially to be independent standard normal random variables. After a change-point the observations in a subset of the streams of data have nonzero mean values. The subset and the post-change means are unknown. The procedure we study uses stream specific generalized likelihood ratio statistics, which are combined to form an overall detection statistic in a mixture model that hypothesizes an assumed fraction p_0 of affected data streams. An analytic expression is obtained for the average run length (ARL) when there is no change and is shown by simulations to be very accurate. Similarly, an approximation for the expected detection delay (EDD) after a change-point is also obtained. Numerical examples are given to compare the suggested procedure to other procedures for unstructured problems and in one case where the problem is assumed to have a well-defined geometric structure. Finally we discuss sensitivity of the procedure to the assumed value of p_0 and suggest a generalization.



SEQUENTIAL MULTI-SENSOR CHANGE-POINT DETECTION[1]

BY YAO XIE AND DAVID SIEGMUND

N sensors, each giving observations

$y_{n,t}$ n=1,2,…, N
t=1,2,…

At certain time K, there are changes in the distributions of observations from a subset M of the sensors. This changetime K, the subset M and its size #M are unknown.

Goal: to detect K as soon as possible after it occurs (minimizing EDD) while keeping the frequency of false alarms as low as possible (maximizing ARL) .

N is large, #M is relatively small

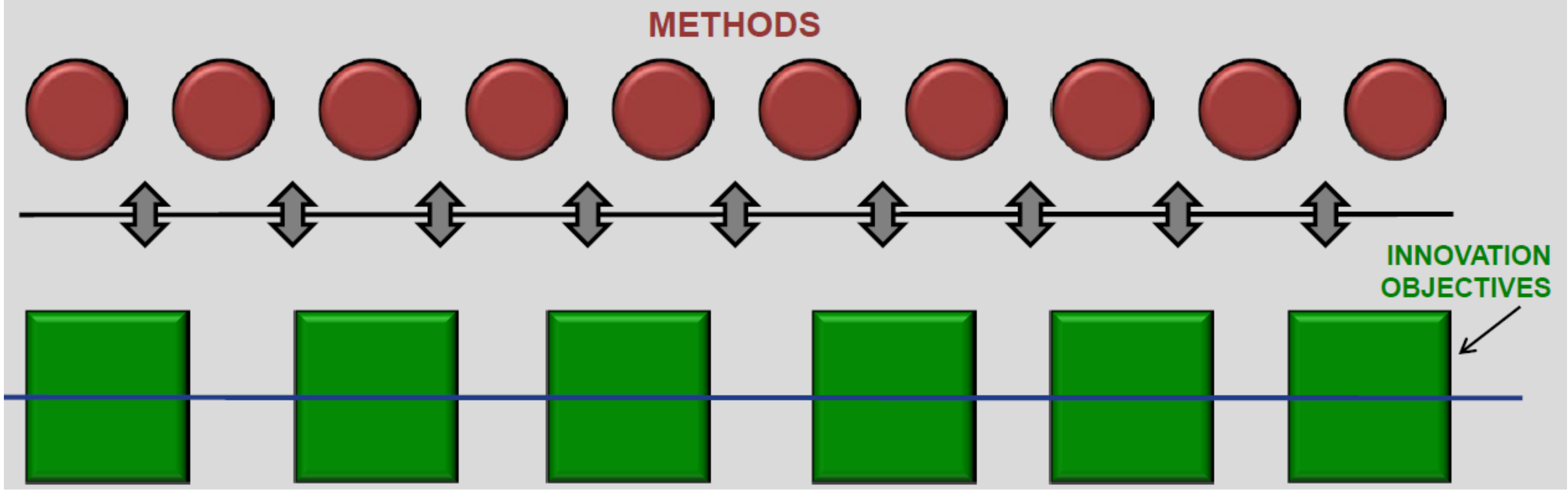
Submitted to Ann Oper Res manuscript No.
(will be inserted by the editor)

Multi-Sensor Slope Change Detection

Yang Cao · Yao Xie · Nagi Gebraeel

Received: date / Accepted: date

Abstract We develop a mixture procedure for multi-sensor systems to monitor parallel streams of data for a change-point that causes a gradual degradation to a subset of data streams. Observations are assumed initially to be normal random variables with known constant means and variances. After a change-point the observations in a subset of the streams of data have increasing or decreasing mean values. The subset and the slope changes are unknown. Our procedure uses a mixture statistics which assumes that each sensor is affected with probability p_0. Analytic expressions are obtained for the average run length (ARL) and the expected detection delay (EDD) of the mixture procedure, which are demonstrated to be quite accurate numerically. We establish asymptotic optimality of the mixture procedure. Numerical examples


METHODS
INNOVATION OBJECTIVES

- Sampling bias and missing values take new dimensions.
- There are new possibilities in using all data.
- High frequency time series data allow intervention in real time.
- Accounting for uncertainty of estimates is fundamental in decision making.
- Causal effects enable effective actions.
- Real time computations mean model approximation.
- Surprise and changepoint prediction allows control.
- Sparsity and knowledge integration allow dimension reduction and sharper inference.
- Network based decision theory.
- Large Scale Optimisation.

BigInsight

Oslo University Hospital

UiO : University of Oslo

UNIVERSITETET I BERGEN

CANCER Registry of Norway
INSTITUTE OF POPULATION-BASED CANCER RESEARCH

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

DNV·GL

ABB

Gjensidige

nav

telenor

DNB

HYDRO

Skatteetaten

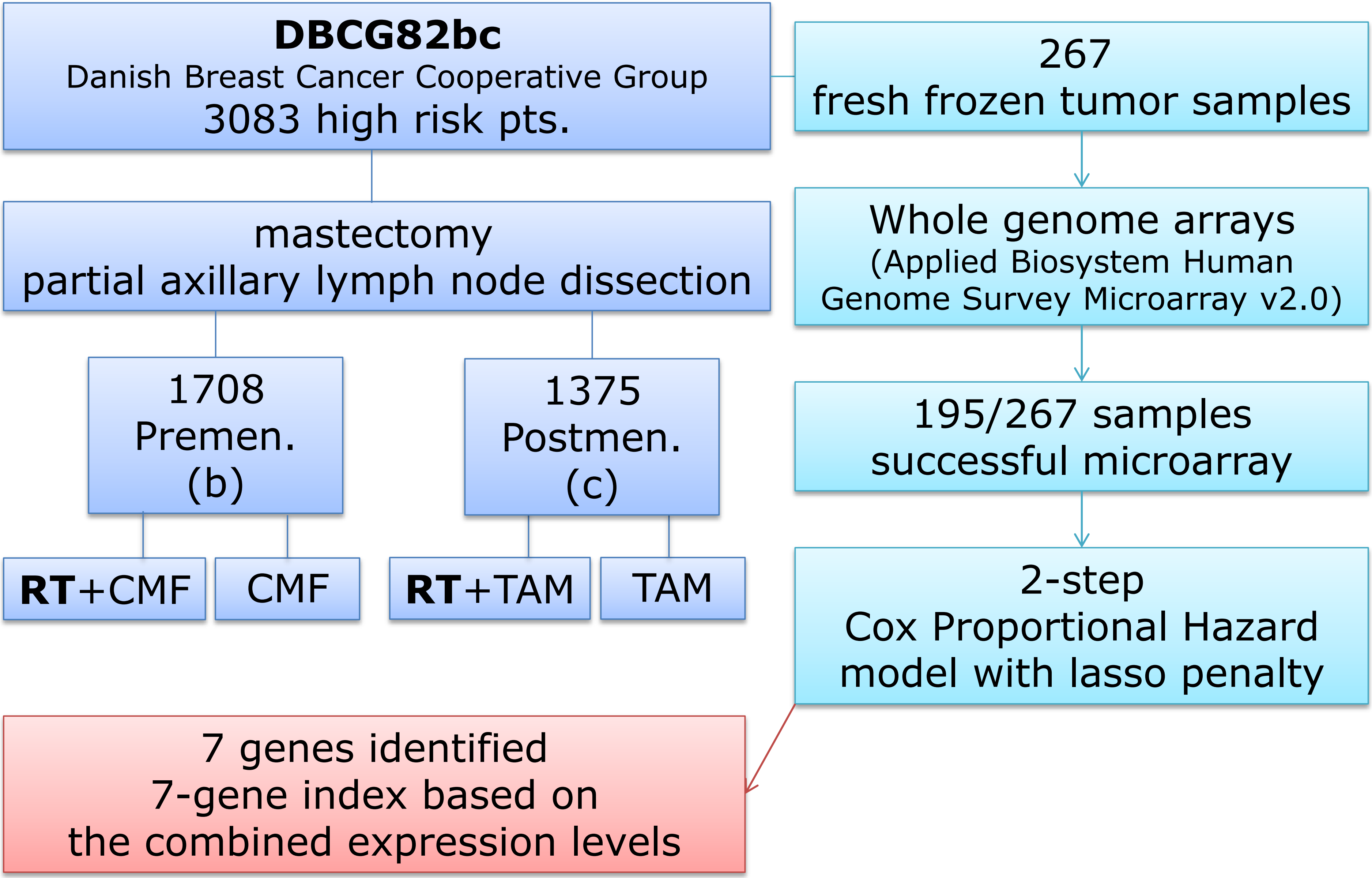
A seven-gene signature predicting benefit of postmastectomy radiotherapy in high risk breast cancer

- Postmastectomy radiotherapy (PMRT) is recommended to breast cancer patients estimated to have high risk of loco-regional recurrence (LR)
- Lancet 2005: substantial survival benefit after PMRT also in patients with low risk of LR

Aim

- To identify genes, whose transcription interacts with PMRT to modify the hazard of LR

Material and Methods



DBCG82bc
Danish Breast Cancer Cooperative Group
3083 high risk pts.

mastectomy
partial axillary lymph node dissection

1708
Premen.
(b)

1375
Postmen.
(c)

RT+CMF

CMF

RT+TAM

TAM

267
fresh frozen tumor samples

Whole genome arrays
(Applied Biosystem Human
Genome Survey Microarray v2.0)

195/267 samples
successful microarray

2-step
Cox Proportional Hazard
model with lasso penalty

7 genes identified
7-gene index based on
the combined expression levels

- **Cox proportional hazard**

T_i = 1 if patient did receive RT
= 0 if patient did NOT receive RT

Hazard of LR for person i at time t =

$$h_o(t) \cdot \exp(\, \theta Z_i + \alpha \mathrm{T}_i + \sum_{g=1}^{17910} \beta_g X_{g,i} + \sum_{g=1}^{17910} \gamma_g \mathrm{T}_i X_{g,i} \,)$$

baseline hazard

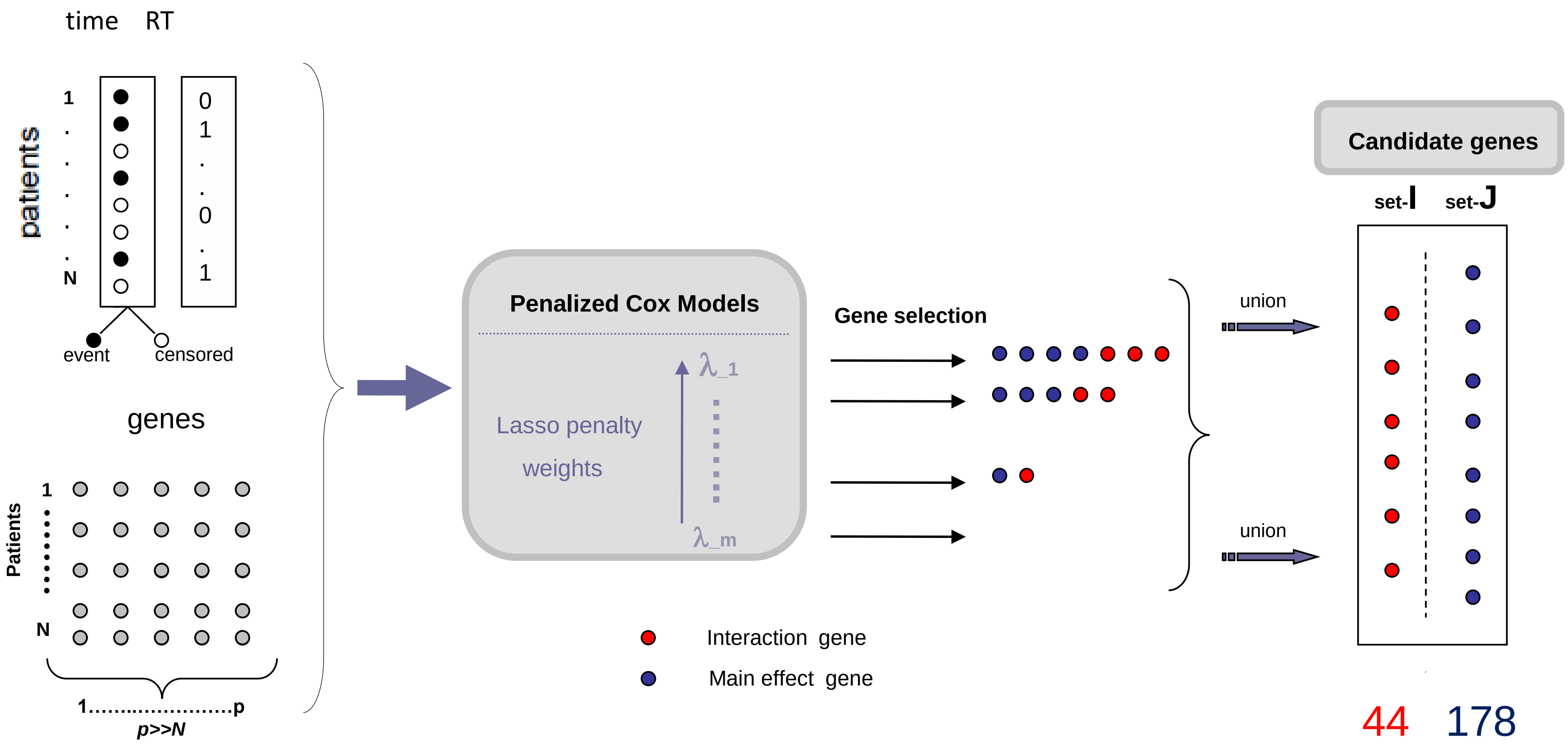
effect of other clinical factors (ER, menop.status, tumor size, posNode)

overall benefit of RT

effect of the expression of the genes

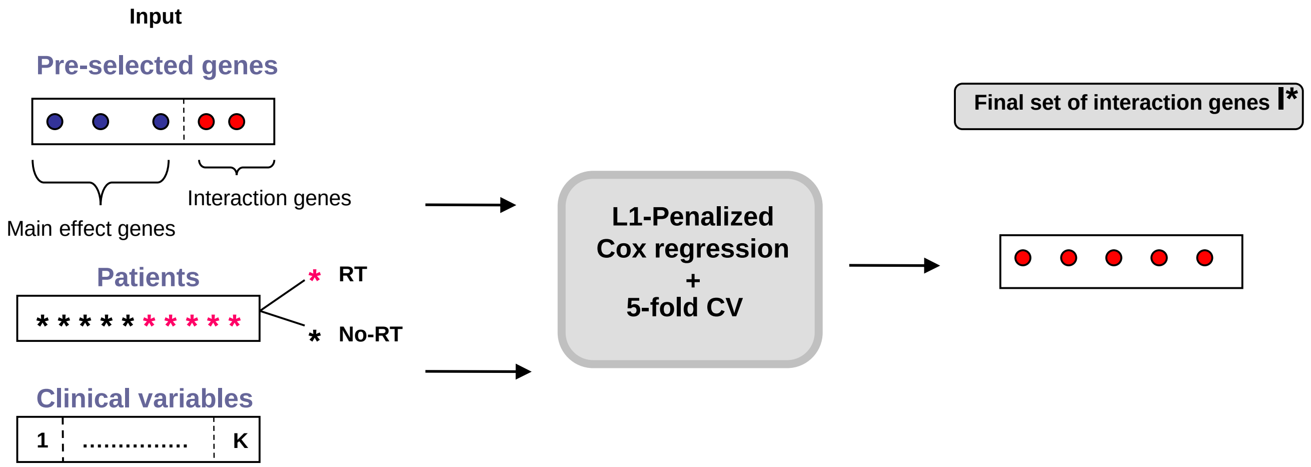
joint effect of the expression of the genes and the RT: **RT/gene interaction**

Pre-selection step



- 44+178=206, as 16 genes were in both lists

Final selection



interaction genes: 7
main effect genes: 46
(3 genes in common)

AB-ID	Gene Symbol	$(\hat{\gamma})$	RR=exp$(\hat{\gamma})$	Description
hCG2042724	HLA-DQA1	0.0699 (0.0440)	1.0724 (0.0412)	major histocompatibility complex, class II, DQ alpha 1 chr. 6p21.3
hCG1980528.1	IGKC	-0.0646 (0.0426)	0.9375 (0.0455)	immunoglobulin kappa constant chr. arm 2p12
hCG39901.3	RGS1	0.2810 (0.1323)	1.3244 (0.1115)	regulator of G-protein signalling 1 chr. 1q31
hCG41484.2	ADH1B	-0.0314 (0.0305)	0.9691 (0.0325)	alcohol dehydrogenase IB (class I), beta polypeptide chr. 4q21-q23
hCG25678.3	DNAL1	0.3763 (0.1429)	1.4568 (0.1130)	dynein, axonemal, light intermediate polypeptide 1 chr.arm 1p35.1
hCG2032658	OR8G2	-0.1266 (0.0636)	0.8811 (0.0699)	olfactory receptor, family 8, subfamily G member 2 chr. 11q24
hCG2023290		0.0452 (0.0186)	1.0462 (0.0179)	Unknown, chr 7,

- **Is RT worth doing? Yes, all the time, because of large main effect of RT.**
- **But the relative advantage depends on our 7 genes.**

SCORE of LR for person i at time t =

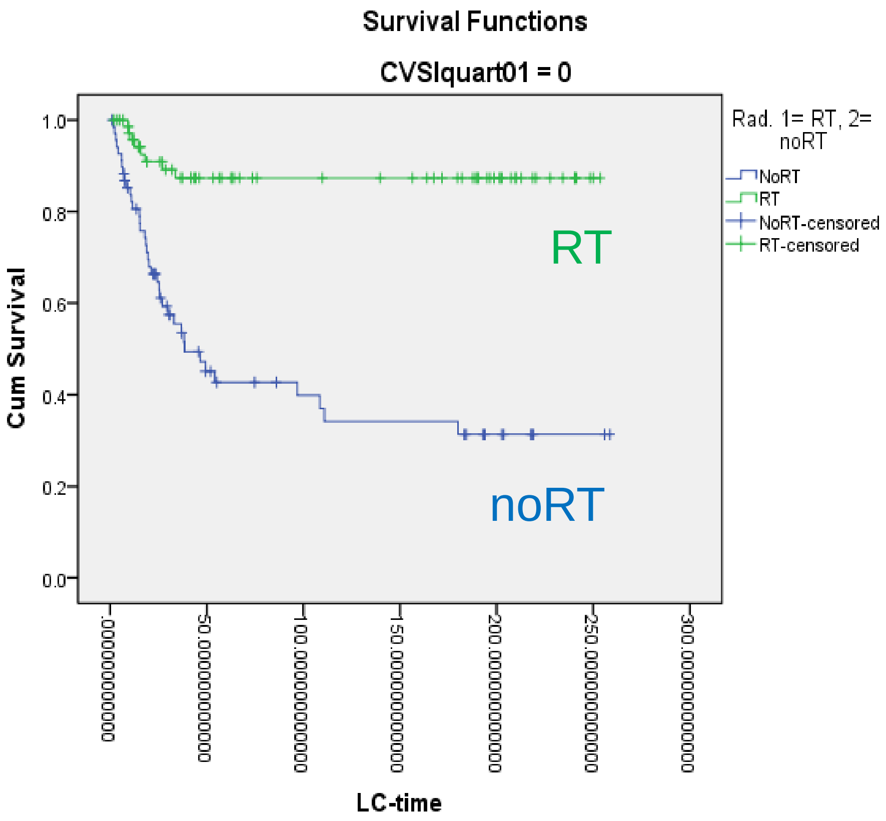
$$\exp\left(\alpha \mathrm{T}_i + \qquad + \sum_{\text{7 genes}} \gamma_g \mathrm{T}_i X_{g,i} \right)$$

overall benefit of RT

RT/gene interaction

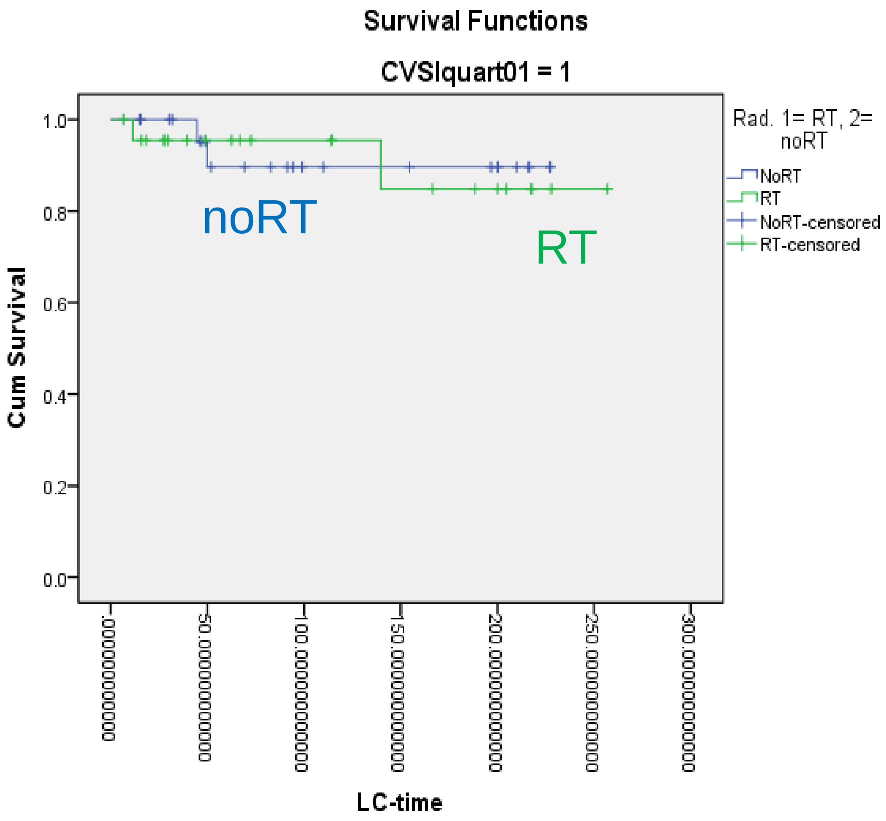
Low CVSI score (lowest 3/4)

N = 144



High CVSI score (highest 1/4)

N = 48


Survival Functions
CVSIquart01 = 1
Cum Survival
LC-time
Rad. 1= RT, 2= noRT
NoRT
RT
NoRT-censored
RT-censored
noRT
RT