Big data in cancer research: dangers and opportunities

ACC Coolen
King's College London

- Introduction
- Big data
- Overfitting
- AI and machine learning
- The future

MRC Medical Research Council
CANCER RESEARCH UK | CITY OF LONDON CENTRE
UCL • King's • Barts • Crick

Introduction
- Data analysis in cancer research
- Complexities of modern cancer data

Big data
- What do we mean?

Overfitting
- Phenomenology
- Strategies to deal with overfitting

AI and machine learning

The future

Introduction
Data analysis in cancer research
Complexities of modern cancer data

Big data
What do we mean?

Overfitting
Phenomenology
Strategies to deal with overfitting

AI and machine learning

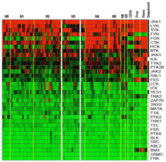
The future

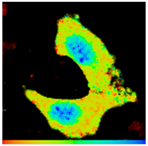
Data analysis in modern cancer research

predict clinical outcomes ...
(OS/PFS, treatment response, side effects)

... from observed patient data
(genome, blood, environment, images)

acid test: predict outcomes for *unseen* data





new problems

- complexity of patterns
- diversity of covariates
- curse of dimensionality

new ambitions

- personalised cancer medicine
  – use *all* information available
  – hence *multivariate* models

Introduction
- Data analysis in cancer research
- Complexities of modern cancer data

Big data
- What do we mean?

Overfitting
- Phenomenology
- Strategies to deal with overfitting

AI and machine learning

The future

Merging data sets

response rates for treatments A and B (Simpson's paradox)

confounding factors

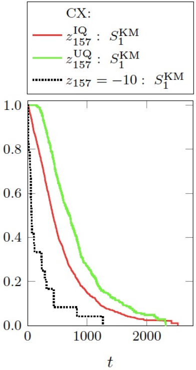
	response to A	response to B
centre 1	40/100 (40%)	150/500 (30%)
centre 2	36/200 (18%)	12/80 (15%)
combined	76/300 (25%)	162/580 (28%)

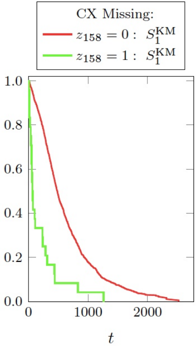
Missing covariate values

red herrings
or white sharks?

sophisticated imputation
not enough:

guard against
informative missingness





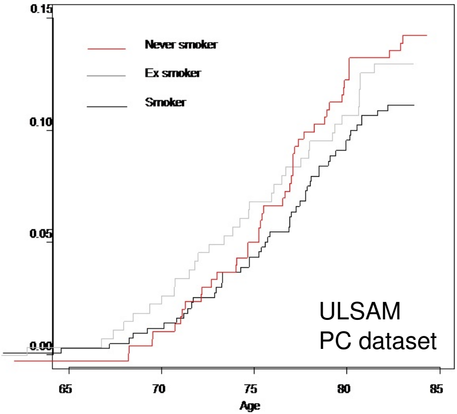
Disease interactions

If we assume censoring risks
<u>uncorrelated</u> with primary risk:

informative censoring
can give nonsensical results ...

– harmful drugs look beneficial
– beneficial drugs look harmful
– false protectivity of covariates

would we have spotted this
if the covariate represented
expression of a specific gene?



Latent heterogeneity

latent: not visible in covariates

say two covariates,
hazard ratios HR_1 and HR_2

consequences:

- proportional hazards X
- *interpreting* time dependencies X
  even if associations time-*indep*:
  cohort level values time-*dep*
- *interpreting* survival curves X
  (Kaplan-Meier, Cox, Fine+Gray, ...)

HR_1 HR_2

n=1000

$\mathrm{HR}_1\uparrow$ $\mathrm{HR}_2\uparrow$ *650*	$\mathrm{HR}_1\downarrow$ $\mathrm{HR}_2\uparrow$ *100*
$\mathrm{HR}_1\uparrow$ $\mathrm{HR}_2\downarrow$ *200*	$\mathrm{HR}_1\downarrow$ $\mathrm{HR}_2\downarrow$ *50*

Interventions

Suppose: gene X ok: low risk $\rightarrow$ few cancers
gene X mutated: high risk $\rightarrow$ more cancers

clear link, easily detected ✓

- but we usually don't observe untreated patients ...

  once we know about gene X:

  gene X ok: low risk $\rightarrow$ few cancers
  gene X mutated: high risk $\rightarrow$ treatment $\rightarrow$ few cancers

  link no longer visible ...

  targeted treatment <u>undermines</u> patterns

Introduction
- Data analysis in cancer research
- Complexities of modern cancer data

Big data
- What do we mean?

Overfitting
- Phenomenology
- Strategies to deal with overfitting

AI and machine learning

The future

Introduction
- Data analysis in cancer research
- Complexities of modern cancer data

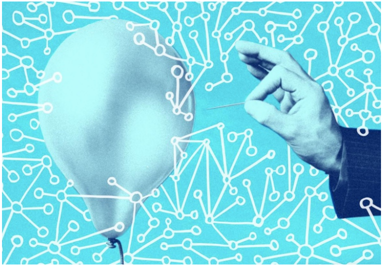
Big data
- What do we mean?

Overfitting
- Phenomenology
- Strategies to deal with overfitting

AI and machine learning

The future

What is big data?

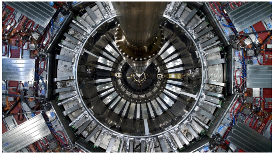


the hypnotizing power
of clever slogans ...

‘modernization’, ‘take back control’,
‘deep learning’, ‘big data’ ...

‘big data’ are themselves not new ...

just new in medicine ...

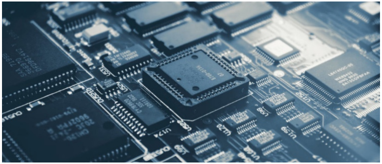


Types of 'big data'

- Very many samples, relatively few variables per sample

  problems mostly of a *practical* nature

  (solved by larger disks, faster computers, parallelization of existing algorithms)



- Very many variables per sample, relatively few samples

  problems of a *conceptual* nature

  – lack of intuition
  – lack of appropriate methods

  genomic data, images, ...

here conventional multi-variate methods
break down due to overfitting

Introduction
Data analysis in cancer research
Complexities of modern cancer data

Big data
What do we mean?

Overfitting
Phenomenology
Strategies to deal with overfitting

AI and machine learning

The future

Introduction
Data analysis in cancer research
Complexities of modern cancer data

Big data
What do we mean?

Overfitting
Phenomenology
Strategies to deal with overfitting

AI and machine learning

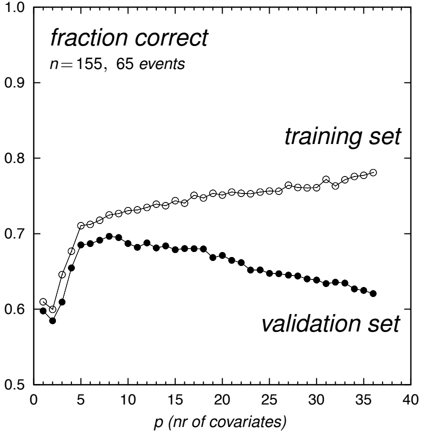
The future

Phenomenology of overfitting

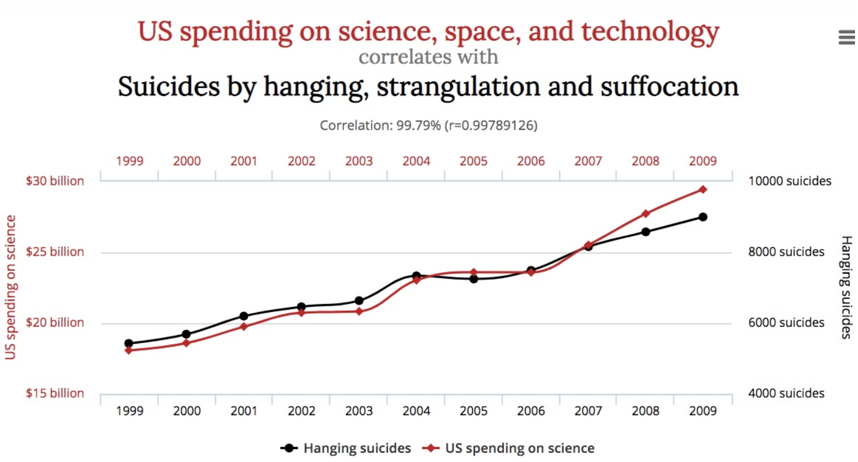
deteriorating outcome
prediction performance
on unseen data ...

multivariate
Cox regression:

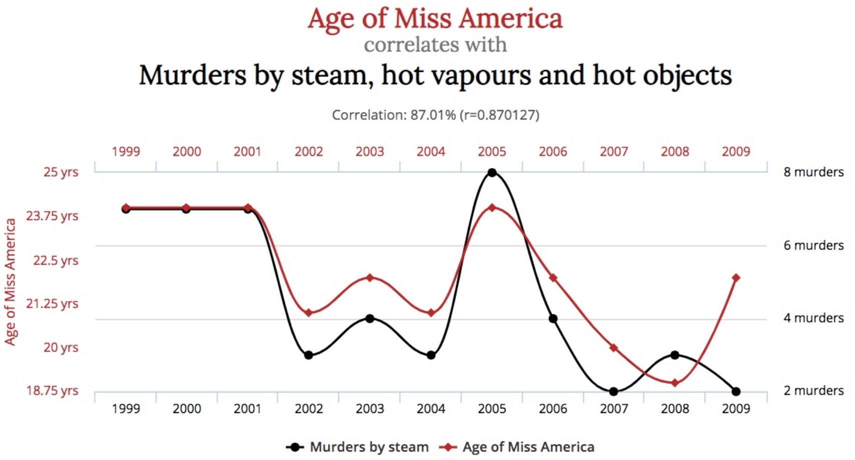
predict whether event before
or after a cutoff time point



so what exactly is going wrong?



(www.tylervigen.com)



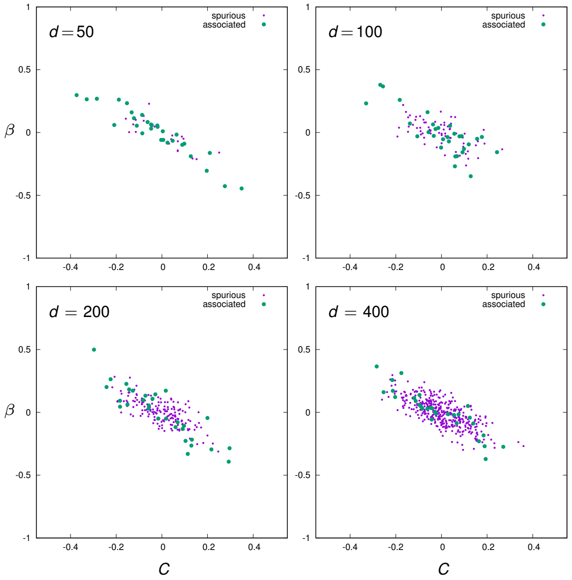
(www.tylervigen.com)

false positive associations ...

$n = 100$

d covariates:

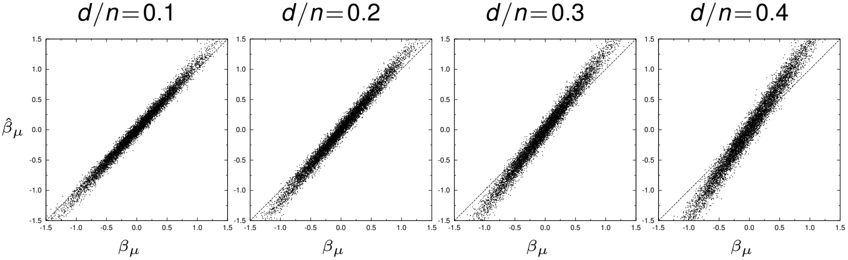
30 true associations
$d-30$ spurious ones



C: correlation between covariate and event time,

β: univariate Cox parameter

bias in inferred association parameters ...



β_μ: true associations
$\hat{\beta}_\mu$: multivariate regression

synthetic survival data, n=400
figures independent of base hazard rate ...

Introduction
- Data analysis in cancer research
- Complexities of modern cancer data

Big data
- What do we mean?

Overfitting
- Phenomenology
- Strategies to deal with overfitting

AI and machine learning

The future

Strategies to deal with overfitting

in multivariate regression

- *'Back off'*

  find 'safe' ratio covariates/samples,
  construct risk 'signatures' or 'scores'

- *Eliminate redundant information*

  improve covariates/samples ratio via
  intelligent dimension reduction

- *'Integrate out' overfitting effects*

  fully Bayesian analysis of parameter uncertainty,
  while keeping computations feasible

- *Model overfitting effects*

  Overfitting correction theory for multivariate regression,
  based on theoretical physics techniques

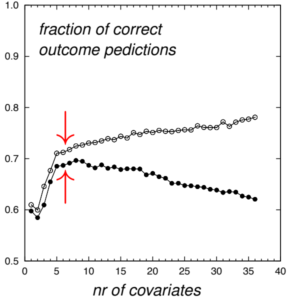
Know when to 'back off'

iterative pipelines and optimised risk scores:
devil is very much in the detail ...

- early pipelines used covariate-outcome correlations, reproducibility poor ...

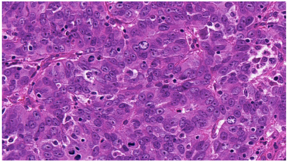
  (e.g. MammaPrint BC gene signature, 70 genes, FDA approved in 2007)

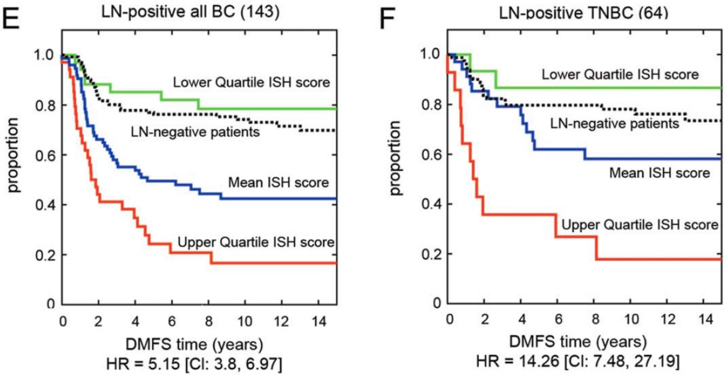
- modern pipelines
  - multivariate regression
  - MAP inference with adaptive Bayesian prior
  - deal with informative missingness
  - probabilistic predictions
  - iterative covariate removal, information-theoretic criterion
  - detection of overfitting transition
  - many randomisations per iteration
  - identification of optimal covariate set
  - . . . . . .



multivariate ‘immune-stroma-histological risk score’ (ISH)

(prevent unnecessary chemotherapies for LN-positive BC patients)





(Grigoriadis et al, 2018)

Eliminate redundant information

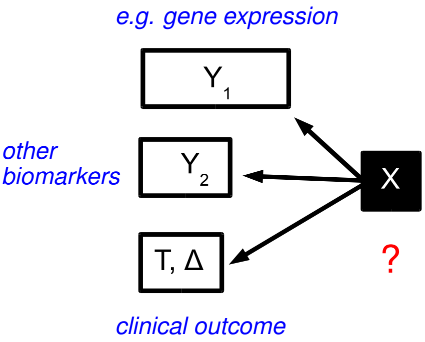
Bayesian latent variable methods

Assume:

(a) data Y_k are *high-dim windows* on q-dim latent variables X

(b) X actually drives outcome

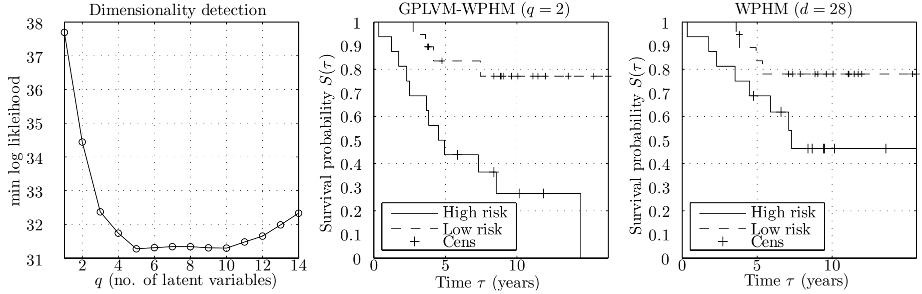
(c) dimension of X less than dimension of Y_k



- nonlinear stochastic relations $Y_k = f_k(X) + \textit{noise}$
- dimension detection: optimal q?
- find most probable latent variables X
- use X to predict clinical outcome

Application to METABRIC
BC gene signature data

data Y: scores of 28 gene signatures
outcome: overall survival time



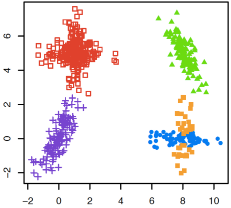
extract dimension of X from training set ($n = 74$)

predict risk from X ($q = 2$) in validation set ($n = 74$)

predict risk from Y in validation set ($n = 74$)

(Barrett & Coolen, 2015)

‘Integrate out’ overfitting effects



data: $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$

$\boldsymbol{x}_i$: covariates
y_i: outcome class labels

$$ML: \quad p(y|\boldsymbol{x}, D) \approx p(y|\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{ML}}), \qquad \boldsymbol{\theta}_{\mathrm{ML}}: \quad \textit{maximize} \;\; p(D|\boldsymbol{\theta})$$

$$MAP: \quad p(y|\boldsymbol{x}, D) \approx p(y|\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{MAP}}), \qquad \boldsymbol{\theta}_{\mathrm{MAP}}: \quad \textit{maximize} \;\; p(\boldsymbol{\theta}|D)$$

$$Bayes: \quad p(y|\boldsymbol{x}, D) = \int \mathrm{d}\boldsymbol{\theta} \; p(y|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D)$$

$$p(\boldsymbol{\theta}|D) = \frac{p(\boldsymbol{\theta}) p(D|\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}' \; p(\boldsymbol{\theta}') p(D|\boldsymbol{\theta}')}$$

keep track fully & precisely of parameter uncertainty

large d:

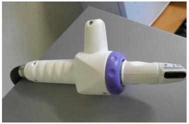
- in view of overfitting:
  full Bayesian parameter estimation
- computational feasibility:
  evaluate d-dimensional integrals *analytically*

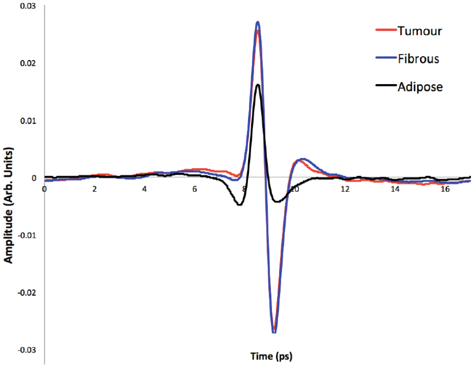
(Shalabi et al, 2016, Sheikh & Coolen, 2019)

Tissue classification during BC surgery

using handheld Terahertz device

$n = 257, \;\; d = 301$

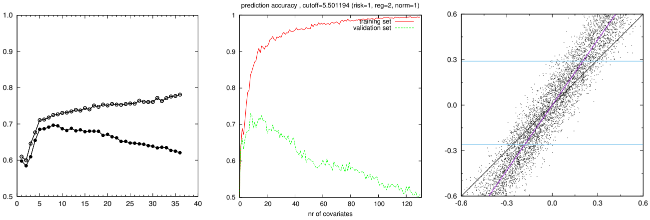




	Number of patients/samples	Accuracy	Sensitivity	Specificity
Frozen section analysis (FSA)	46-1327	84-98 %	78-91%	92-98%
Specimen radiography	12-119	33-84%	45-61%	77-89%
Intraoperative ultrasound	81-225	62-80%	36-79%	66-91%
Touch imprint cytology	27-510	78-99%	71-97%	90-98%
Optical spectroscopy	20-179	75-94%	74-91%	65-96%
Support Vector Machine	257	75%	86%	66%
Naive Bayesian method	257	69%	89%	53%
New Bayesian method	257	95%	96%	95%

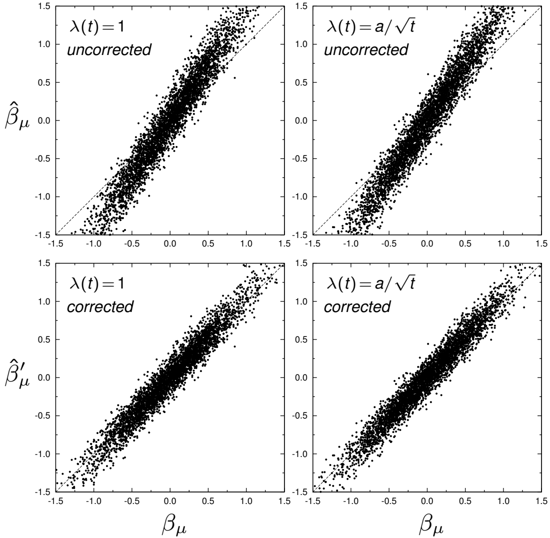
Overfitting correction
in multivariate survival analysis

can we model what happens in the overfitting regime?



- yes, using techniques from many-particle physics (the replica method)
- leads to correction formulae for overfitting bias in association parameters and base hazard rates
- can be rolled out to arbitrary generalized linear models (logistic regression, frailty models, latent class models, ...)

(Coolen et al, 2017, Sheikh & Coolen, 2019)



overfitting correction of association pars,
slope predicted by variational replica theory

$n = 200$, $p = 80$

Introduction
Data analysis in cancer research
Complexities of modern cancer data

Big data
What do we mean?

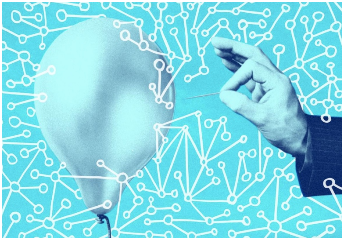
Overfitting
Phenomenology
Strategies to deal with overfitting

AI and machine learning

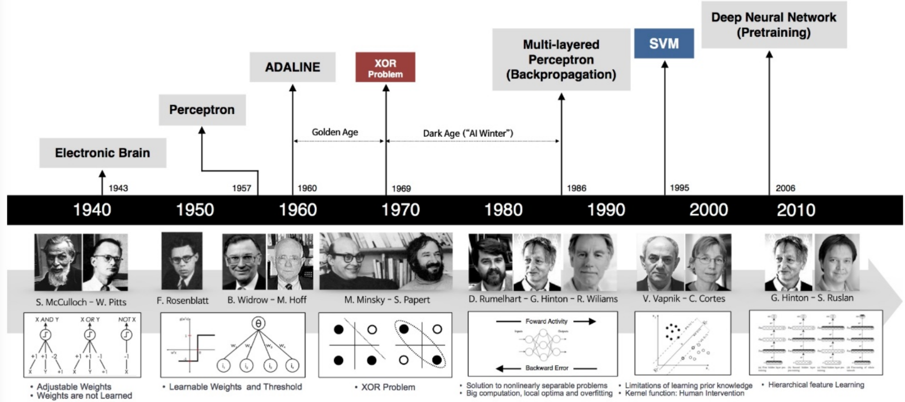
The future

What is new?

- faster and bigger computers
- more data
- intense marketing
- inflation of terminology:
  data + computers = AI

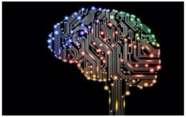


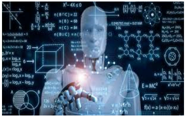
history of machine learning:

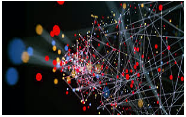


AI and Deep Learning

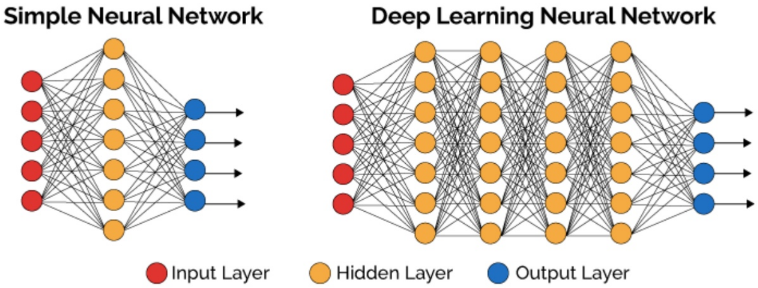
fancy names,
fancy pictures ...



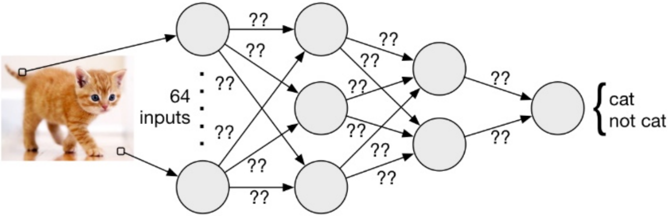




let's open the box:
1980s architectures, 1980s learning rules ...



- suitable problems for standard ML approaches



– many data of the type (question,answer)
– we can trust the answers
– we're not interested in knowing the underlying rules

e.g. speech recognition, detection of anomalies in images

- limitations of standard ML approaches

– 'black box' decision making
– often no reliable error bars
– cannot handle complexities of cancer data, such as confounders, informative missingness, disease interactions, ...

Watson Oncology,
the dangers of hyping ...

FEBRUARY 23, 2017

MD Anderson Cancer Center's IBM Watson project fails, and so did the journalism related to it

From Hero to Has-Been in Just 4 Years

If you're at all interested in technology and healthcare, by now you've probably heard about IBM Watson, the artificial intelligence technology that went from winning on Jeopardy in 2(healthcare organizations for a variety of p

EDITOR'S PICK | 214,282 views | Feb 19, 2017, 03:48pm

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine

In total, the project cost MD Anderson more than $62.1 million.

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

IBM are now turning towards Bayesian statistical modelling ...

Introduction
Data analysis in cancer research
Complexities of modern cancer data

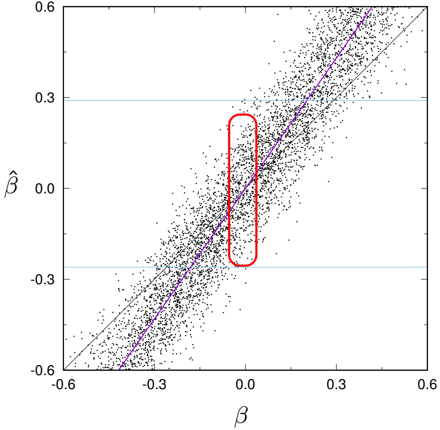
Big data
What do we mean?

Overfitting
Phenomenology
Strategies to deal with overfitting

AI and machine learning

The future

back to the false positive associations ...



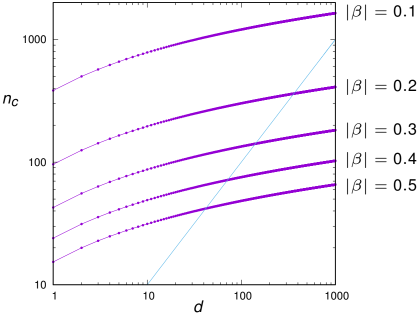
multivariate regression with more covariates than samples in principle possible

if n large enough

- $n \uparrow$: prob of false positive associations $\downarrow$
- $d \uparrow$: prob of false positive associations $\uparrow$

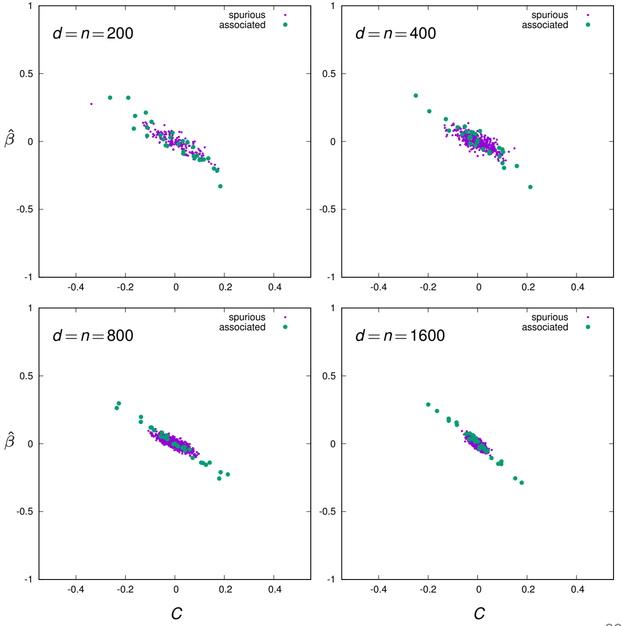
uncorrelated covariates:

> *prob of finding one or more spurious univariate associations of strength* $\geq |\beta|$ *is less than* 5% *if* $n > n_c$



synthetic survival data, with $d = n$

(30 true associations)



Future of big data analytics in cancer research

short term, 1-5 years

- refinement of ML methods for anomaly detection in images and natural language processing (patient records)
- outcome prediction: 'purge' of black-box ML approaches, leaving algorithms with transparent statistical interpretations
- increased parallelization of algorithms, to run on dedicated hardware
- reliable statistical regression in overfitting regime

longer term, 5-10 yrs

- longitudinal survival analysis:

  rigorous methods/standards for handling time-dependent covariates and observation-triggered clinical interventions
- from association to causality:

  further development of general theory of causal inference, and application in (cancer) medicine