Ten tips for applying artificial intelligence in data science

Chen-Hsiang Yeang
Institute of Statistical Science
Academia Sinica
Taipei, Taiwan

Outline

- **Lecture.**
- Assignments and discussion.
- Presentation.

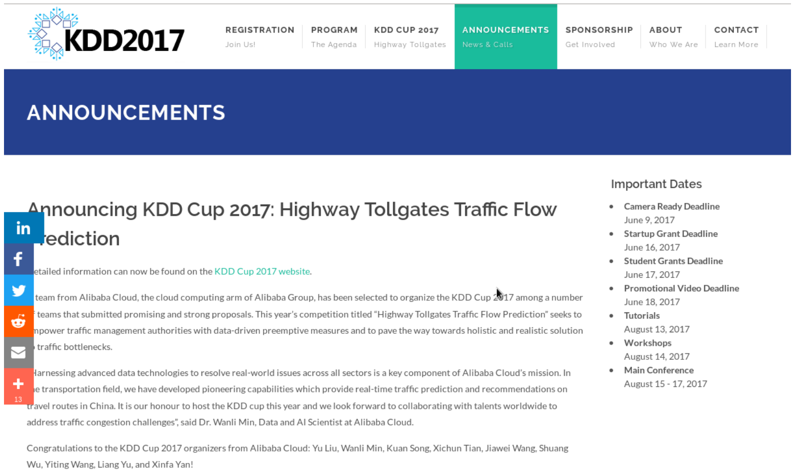
Hypes will go away, but something will stay

- Dotcom bubbles → Internet infrastructure
- Web 2.0 → gig economy
- Smart phones → pervasive connection
- AI and data science → ?

To leave something tenacious, I found some tips might be useful.

Tip 1: Thou shall know the game you are playing

Game 1: Task-oriented competitions



ANNOUNCEMENTS

Announcing KDD Cup 2017: Highway Tollgates Traffic Flow rediction

etailed information can now be found on the KDD Cup 2017 website.

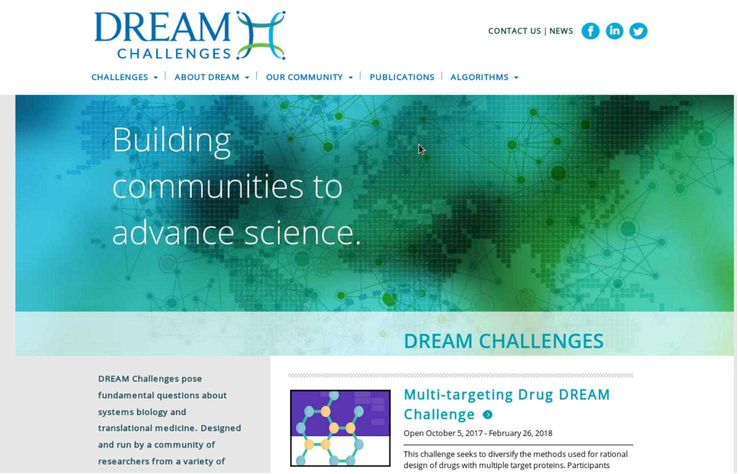
team from Alibaba Cloud, the cloud computing arm of Alibaba Group, has been selected to organize the KDD Cup 2017 among a number f teams that submitted promising and strong proposals. This year's competition titled "Highway Tollgates Traffic Flow Prediction" seeks to mpower traffic management authorities with data-driven preemptive measures and to pave the way towards holistic and realistic solution traffic bottlenecks.

Harnessing advanced data technologies to resolve real-world issues across all sectors is a key component of Alibaba Cloud's mission. In e transportation field, we have developed pioneering capabilities which provide real-time traffic prediction and recommendations on ravel routes in China. It is our honour to host the KDD cup this year and we look forward to collaborating with talents worldwide to address traffic congestion challenges", said Dr. Wanli Min, Data and AI Scientist at Alibaba Cloud.

Congratulations to the KDD Cup 2017 organizers from Alibaba Cloud: Yu Liu, Wanli Min, Kuan Song, Xichun Tian, Jiawei Wang, Shuang Wu, Yiting Wang, Liang Yu, and Xinfa Yan!

Important Dates

- Camera Ready Deadline
  June 9, 2017
- Startup Grant Deadline
  June 16, 2017
- Student Grants Deadline
  June 17, 2017
- Promotional Video Deadline
  June 18, 2017
- Tutorials
  August 13, 2017
- Workshops
  August 14, 2017
- Main Conference
  August 15 - 17, 2017



Building communities to advance science.

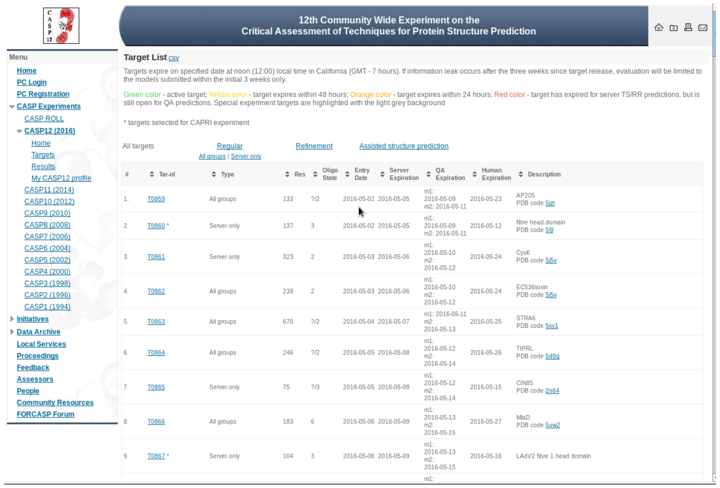
DREAM CHALLENGES

DREAM Challenges pose fundamental questions about systems biology and translational medicine. Designed and run by a community of researchers from a variety of

Multi-targeting Drug DREAM Challenge

Open October 5, 2017 - February 26, 2018

This challenge seeks to diversify the methods used for rational design of drugs with multiple target proteins. Participants



12th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

Target List csv

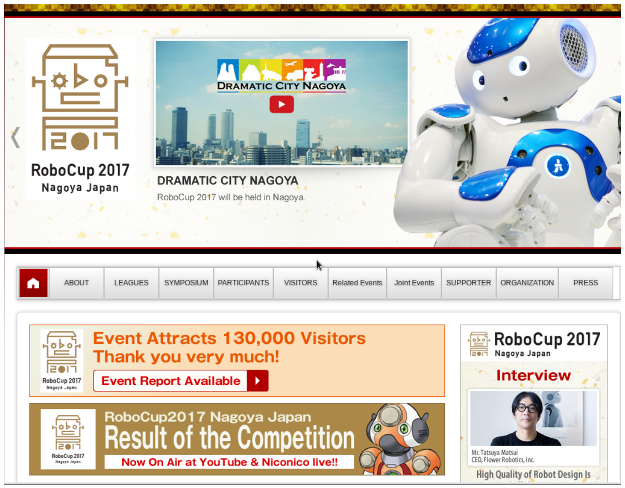
Targets expire on specified date at noon (12:00) local time in California (GMT - 7 hours). If information leak occurs after the three weeks since target release, evaluation will be limited to the models submitted within the initial 3 weeks only.

Green color - active target; Yellow color - target expires within 48 hours; Orange color - target expires within 24 hours; Red color - target has expired for server TS/RR predictions, but is still open for QA predictions. Special experiment targets are highlighted with the light grey background

* targets selected for CAPRI experiment

All targets   Regular   Refinement   Assisted structure prediction
All groups | Server only

#	Tar-id	Type	Res	Oligo State	Entry Date	Server Expiration	QA Expiration	Human Expiration	Description
1.	T0859	All groups	133	?/2	2016-05-02	2016-05-05	m1: 2016-05-09 m2: 2016-05-11	2016-05-23	AP205 PDB code 5jzr
2.	T0860 *	Server only	137	3	2016-05-02	2016-05-05	m1: 2016-05-09 m2: 2016-05-11	2016-05-12	fibre head domain PDB code 5fjl
3.	T0861	Server only	323	2	2016-05-03	2016-05-06	m1: 2016-05-10 m2: 2016-05-12	2016-05-24	CysK PDB code 5j5v
4.	T0862	All groups	239	2	2016-05-03	2016-05-06	m1: 2016-05-10 m2: 2016-05-12	2016-05-24	EC536toxin PDB code 5j5v
5.	T0863	All groups	670	?/2	2016-05-04	2016-05-07	m1: 2016-05-11 m2: 2016-05-13	2016-05-25	STRA6 PDB code 5sy1
6.	T0864	All groups	246	?/2	2016-05-05	2016-05-08	m1: 2016-05-12 m2: 2016-05-14	2016-05-26	TIPRL PDB code 5d9q
7.	T0865	Server only	75	?/3	2016-05-05	2016-05-08	m1: 2016-05-12 m2: 2016-05-14	2016-05-15	CIN85 PDB code 2n64
8.	T0866	All groups	183	6	2016-05-06	2016-05-09	m1: 2016-05-13 m2: 2016-05-15	2016-05-27	MlaD PDB code 5uw2
9.	T0867 *	Server only	104	3	2016-05-06	2016-05-09	m1: 2016-05-13 m2: 2016-05-15	2016-05-16	LAdV2 fibre 1 head domain
							m1:		



DRAMATIC CITY NAGOYA

RoboCup 2017 will be held in Nagoya.

Event Attracts 130,000 Visitors
Thank you very much!

RoboCup2017 Nagoya Japan
Result of the Competition
Now On Air at YouTube & Niconico live!!

Mr. Tatsuya Matsui
CEO, Flower Robotics, Inc.

High Quality of Robot Design Is

Game 1: Task-oriented competitions

Successful teams often

- Combine multiple techniques (learning methods).
- Craft the methods toward the well-defined tasks.
- Know the rival teams well.

Game 2: Method invention and analysis

delineating the absolute indigeneity of amino acids in fossils. As AMS techniques are refined to handle smaller samples, it may also become possible to date individual amino acid enantiomers by the ^{14}C method. If one enantiomer is entirely derived from the other by racemization during diagenesis, the individual D- and L-enantiomers for a given amino acid should have identical ^{14}C ages.

Older, more poorly preserved fossils may not always prove amenable to the determination of amino acid indigeneity by the stable isotope method, as the prospects for complete replacement of indigenous amino acids with non-indigenous amino acids increases with time. As non-indigenous amino acids undergo racemization, the enantiomers may have identical isotopic compositions and still not be related to the original organisms. Such a circumstance may, however, become easier to recognize as more information becomes available concerning the distribution and stable isotopic composition of the amino acid constituents of modern representatives of fossil organisms. Also, AMS dates on individual amino acid enantiomers may, in some cases, help to clarify indigeneity problems, in particular when stratigraphic controls can be used to estimate a general age range for the fossil in question.

Finally, the development of techniques for determining the stable isotopic composition of amino acid enantiomers may enable us to establish whether non-racemic amino acids in some carbonaceous meteorites[27] are indigenous, or result in part from terrestrial contamination.

M.H.E. thanks the NSF, Division of Earth Sciences (grant EAR-8352055) and the following contributors to his Presidentia Young Investigator Award for partial support of this research: Arco, Exxon, Phillips Petroleum, Texaco Inc., The Upjohn Co. We also acknowledge the donors of the Petroleum Research Fund, administered by the American Chemical Society (grant 16144-AC2 to M.H.E., grant 14805-AC2 to S.A.M.) for support. S.A.M. acknowledges NSERC (grant A2644) for partial support.

Received 19 May; accepted 15 July 1986.

1. Bada, J. L. & Protsch, R. *Proc. natn. Acad. Sci. U.S.A.* **70**, 1331–1334 (1973).
2. Bada, J. L., Schroeder, R. A. & Carter, G. F. *Science* **184**, 791–793 (1974).
3. Boulton, G. S. *et al. Nature* **298**, 437–441 (1982).
4. Wehmiller, J. F. in *Quaternary Dating Methods* (ed. Mahaney, W. C.) 171–193 (Elsevier, Amsterdam, 1984).
5. Engel, M. H., Zumberge, J. E. & Nagy, B. *Analyt. Biochem.* **82**, 415–422 (1977).
6. Bada, J. L. *A. Rev. Earth planet. Sci.* **13**, 241–268 (1985).
7. Chisholm, B. S., Nelson, D. E. & Schwarcz, H. P. *Science* **216**, 1131–1132 (1982).
8. Ambrose, S. H. & DeNiro, M. J. *Nature* **319**, 321–324 (1986).
9. Macko, S. A., Estep, M. L. F., Hare, P. E. & Hoering, T. C. *Yb. Carnegie Instn Wash.* **82**, 404–410 (1983).
10. Hare, P. E. & Estep, M. L. F. *Yb. Carnegie Instn Wash.* **82**, 410–414 (1983).
11. Engel, M. H. & Hare, P. E. in *Chemistry and Biochemistry of the Amino Acids* (ed. Barrett, G. C.) 462–479 (Chapman and Hall, London, 1985).
12. Johnstone, R. A. W. & Rose, M. E. in *Chemistry and Biochemistry of the Amino Acids* (ed. Barrett, G. C.) 480–524 (Chapman and Hall, London, 1985).
13. Weinstein, S., Engel, M. H. & Hare, P. E. in *Practical Protein Chemistry—A Handbook* (ed. Darbre, A.) 337–344 (Wiley, New York, 1986).
14. Bada, J. L., Gillespie, R., Gowlett, J. A. J. & Hedges, R. E. M. *Nature* **312**, 442–444 (1984).
15. Mitterer, R. M. & Kriausakul, N. *Org. Geochem.* **7**, 91–98 (1984).
16. Williams, K. M. & Smith, G. G. *Origins Life* **8**, 91–144 (1977).
17. Engel, M. H. & Hare, P. E. *Yb. Carnegie Instn Wash.* **81**, 425–430 (1982).
18. Hare, P. E. *Yb. Carnegie Instn Wash.* **73**, 576–581 (1974).
19. Pillinger, C. T. *Nature* **296**, 802 (1982).
20. Neuberger, A. *Adv. Protein Chem.* **4**, 298–383 (1948).
21. Engel, M. H. & Macko, S. A. *Analyt. Chem.* **56**, 2598–2600 (1984).
22. Dungworth, G. *Chem. Geol.* **17**, 135–153 (1976).
23. Weinstein, S., Engel, M. H. & Hare, P. E. *Analyt. Biochem.* **121**, 370–377 (1982).
24. Macko, S. A., Lee, W. Y. & Parker, P. L. *J. exp. mar. Biol. Ecol.* **63**, 145–149 (1982).
25. Macko, S. A., Estep, M. L. F. & Hoering, T. C. *Yb. Carnegie Instn Wash.* **81**, 413–417 (1982).
26. Vallentyne, J. R. *Geochim. cosmochim. Acta* **28**, 157–188 (1964).
27. Engel, M. H. & Nagy, B. *Nature* **296**, 837–840 (1982).

Learning representations by back-propagating errors

David E. Rumelhart[*], Geoffrey E. Hinton[†] & Ronald J. Williams[*]

[*] Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
[†] Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure[1].

There have been many attempts to design self-organizing neural networks. The aim is to find a powerful synaptic modification rule that will allow an arbitrarily connected neural network to develop an internal structure that is appropriate for a particular task domain. The task is specified by giving the desired state vector of the output units for each state vector of the input units. If the input units are directly connected to the output units it is relatively easy to find learning rules that iteratively adjust the relative strengths of the connections so as to progressively reduce the difference between the actual and desired output vectors[2]. Learning becomes more interesting but more difficult when we introduce hidden units whose actual or desired states are not specified by the task. (In perceptrons, there are 'feature analysers' between the input and output that are not true hidden units because their input connections are fixed by hand, so their states are completely determined by the input vector: they do not learn representations.) The learning procedure must decide under what circumstances the hidden units should be active in order to help achieve the desired input–output behaviour. This amounts to deciding what these units should represent. We demonstrate that a general purpose and relatively simple procedure is powerful enough to construct appropriate internal representations.

The simplest form of the learning procedure is for layered networks which have a layer of input units at the bottom; any number of intermediate layers; and a layer of output units at the top. Connections within a layer or from higher to lower layers are forbidden, but connections can skip intermediate layers. An input vector is presented to the network by setting the states of the input units. Then the states of the units in each layer are determined by applying equations (1) and (2) to the connections coming from lower layers. All units within a layer have their states set in parallel, but different layers have their states set sequentially, starting at the bottom and working upwards until the states of the output units are determined.

The total input, x_j, to unit j is a linear function of the outputs, y_i, of the units that are connected to j and of the weights, w_{ji}, on these connections

$$x_j = \sum_i y_i w_{ji} \qquad (1)$$

Units can be given biases by introducing an extra input to each unit which always has a value of 1. The weight on this extra input is called the bias and is equivalent to a threshold of the opposite sign. It can be treated just like the other weights.

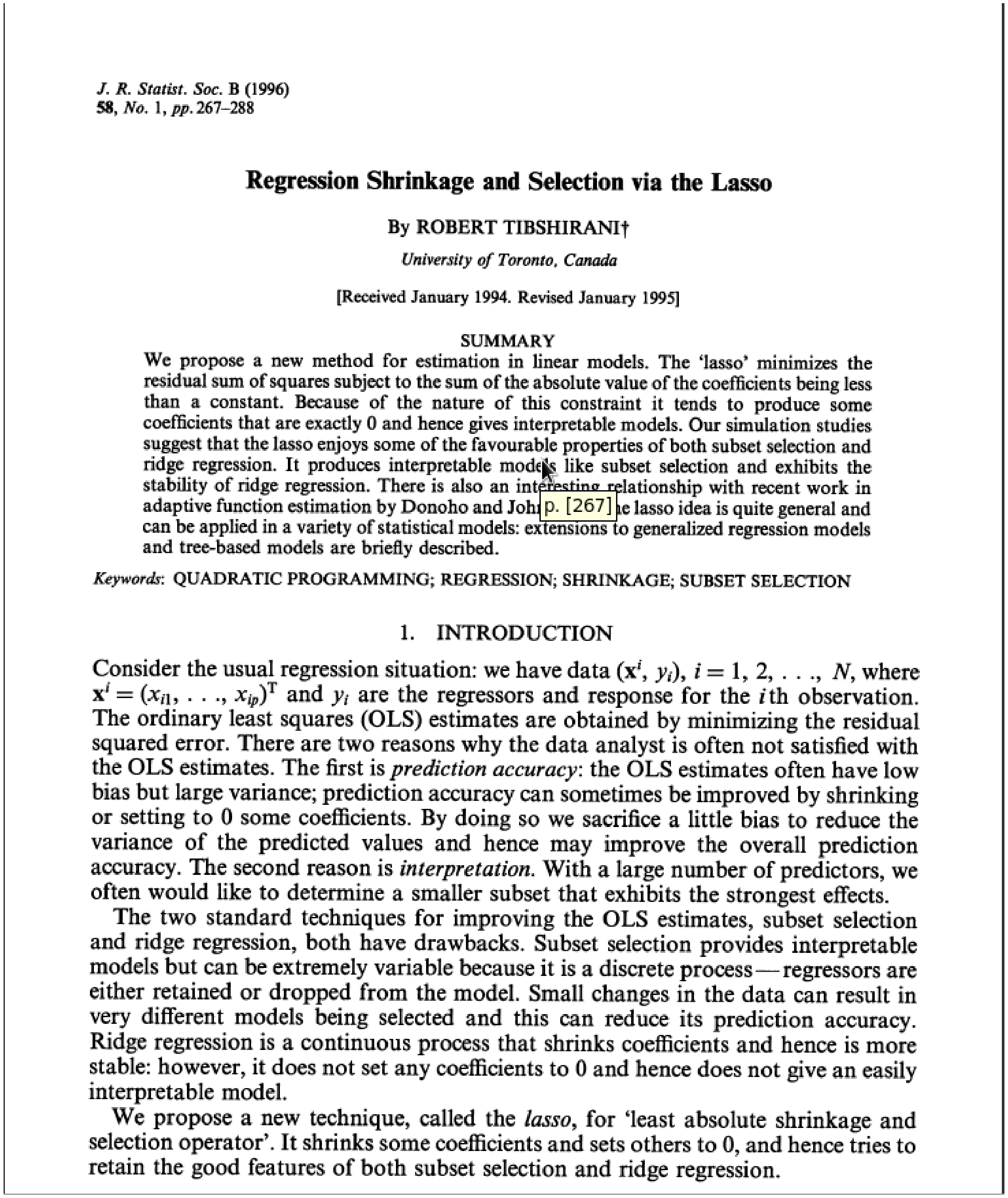
A unit has a real-valued output, y_j, which is a non-linear function of its total input

$$y_j = \frac{1}{1+e^{-x_j}} \qquad (2)$$

[†] To whom correspondence should be addressed.

Cited by 13558.

Game 2: Method invention and analysis



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Joh e lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

1. INTRODUCTION

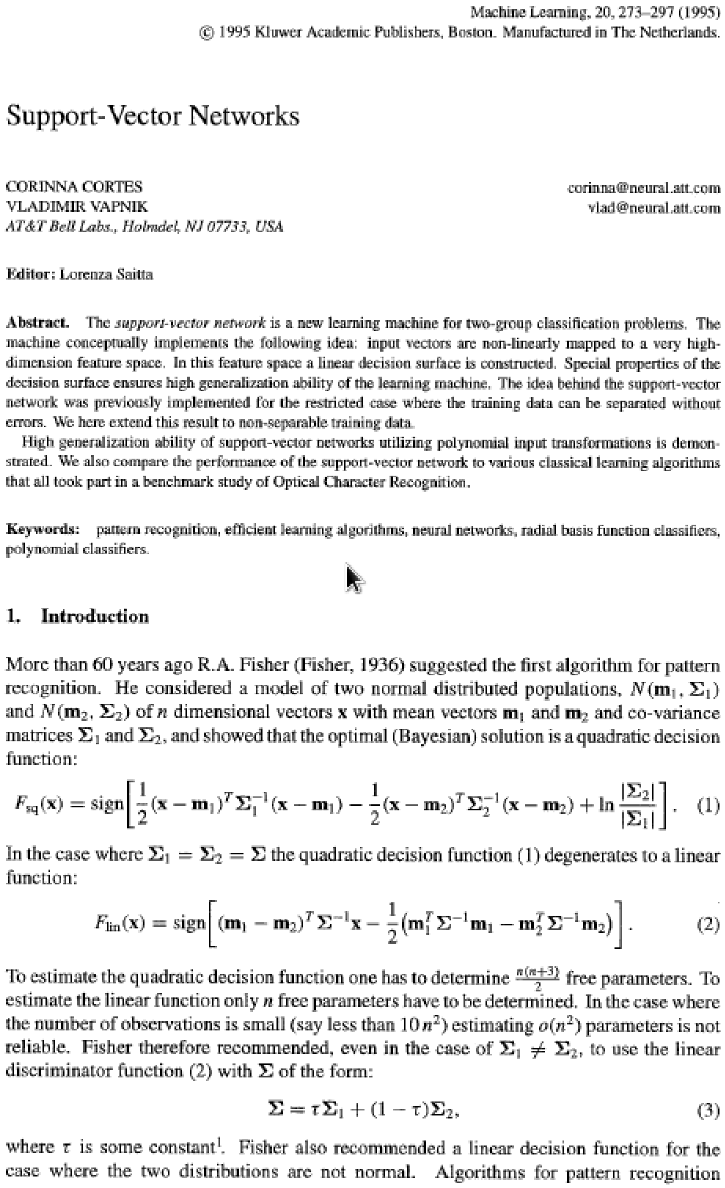
Consider the usual regression situation: we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and y_i are the regressors and response for the ith observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is *prediction accuracy*: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

The two standard techniques for improving the OLS estimates, subset selection and ridge regression, both have drawbacks. Subset selection provides interpretable models but can be extremely variable because it is a discrete process—regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. Ridge regression is a continuous process that shrinks coefficients and hence is more stable: however, it does not set any coefficients to 0 and hence does not give an easily interpretable model.

We propose a new technique, called the *lasso*, for 'least absolute shrinkage and selection operator'. It shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression.

Cited by 22275.

Game 2: Method invention and analysis



Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

Editor: Lorenza Saitta

Abstract. The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data.

High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

Keywords: pattern recognition, efficient learning algorithms, neural networks, radial basis function classifiers, polynomial classifiers.

1. Introduction

More than 60 years ago R.A. Fisher (Fisher, 1936) suggested the first algorithm for pattern recognition. He considered a model of two normal distributed populations, $N(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $N(\mathbf{m}_2, \mathbf{\Sigma}_2)$ of n dimensional vectors $\mathbf{x}$ with mean vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ and co-variance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, and showed that the optimal (Bayesian) solution is a quadratic decision function:

$$F_{\mathrm{sq}}(\mathbf{x}) = \mathrm{sign}\left[\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1) - \frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mathbf{m}_2) + \ln\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}\right]. \quad (1)$$

In the case where $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ the quadratic decision function (1) degenerates to a linear function:

$$F_{\mathrm{lin}}(\mathbf{x}) = \mathrm{sign}\left[(\mathbf{m}_1-\mathbf{m}_2)^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\mathbf{m}_1^T\mathbf{\Sigma}^{-1}\mathbf{m}_1 - \mathbf{m}_2^T\mathbf{\Sigma}^{-1}\mathbf{m}_2\right)\right]. \quad (2)$$

To estimate the quadratic decision function one has to determine $\frac{n(n+3)}{2}$ free parameters. To estimate the linear function only n free parameters have to be determined. In the case where the number of observations is small (say less than $10\,n^2$) estimating $o(n^2)$ parameters is not reliable. Fisher therefore recommended, even in the case of $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, to use the linear discriminator function (2) with $\mathbf{\Sigma}$ of the form:

$$\mathbf{\Sigma} = \tau\mathbf{\Sigma}_1 + (1-\tau)\mathbf{\Sigma}_2, \quad (3)$$

where τ is some constant[1]. Fisher also recommended a linear decision function for the case where the two distributions are not normal. Algorithms for pattern recognition

Cited by 27860.

Game 2: Method invention and analysis

Successful researchers often

- Solve key problems important in the areas.
- Substantially outperform existing methods.
- Undergo rigorous theoretical analysis.

Game 3: Prototype system building

Game 3: Prototype system building

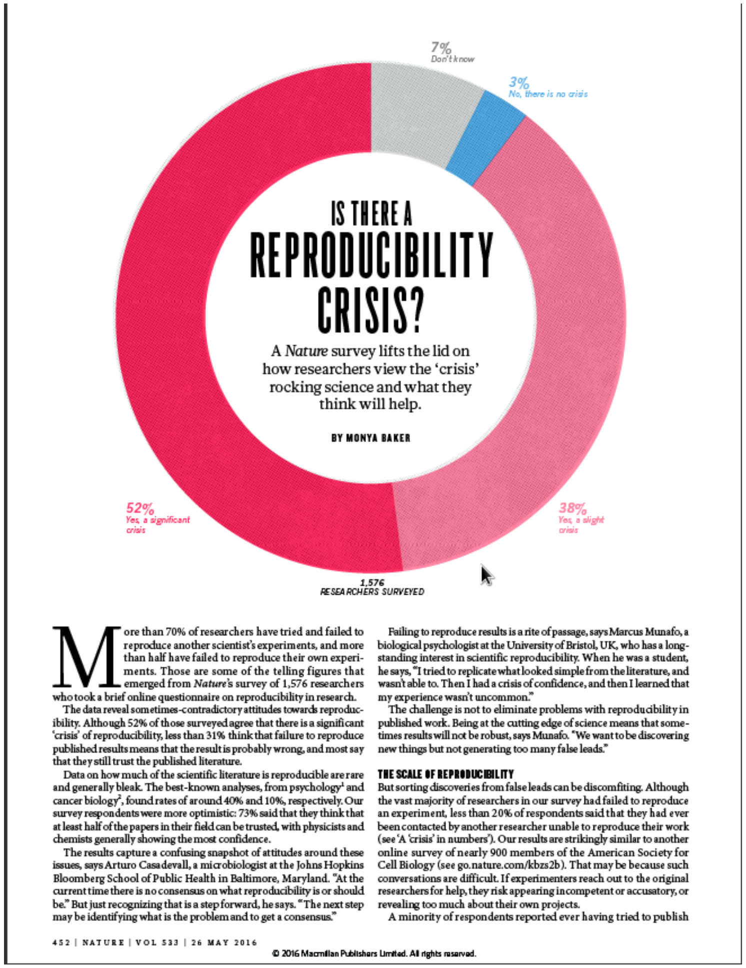
Successful teams often

- Invest considerable efforts for a long time without obvious returns.
- Consist of multi-disciplinary members.
- Take many possible factors into account.

Tip 2: Thou shall not feed garbage to your powerful black box

How reliable are the data you feed into the black box?

Reproducibility crisis



IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

The data reveal sometimes-contradictory attitudes towards reproducibility. Although 52% of those surveyed agree that there is a significant 'crisis' of reproducibility, less than 31% think that failure to reproduce published results means that the result is probably wrong, and most say that they still trust the published literature.

Data on how much of the scientific literature is reproducible are rare and generally bleak. The best-known analyses, from psychology[1] and cancer biology[2], found rates of around 40% and 10%, respectively. Our survey respondents were more optimistic: 73% said that they think that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence.

The results capture a confusing snapshot of attitudes around these issues, says Arturo Casadevall, a microbiologist at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "At the current time there is no consensus on what reproducibility is or should be." But just recognizing that is a step forward, he says. "The next step may be identifying what is the problem and to get a consensus."

Failing to reproduce results is a rite of passage, says Marcus Munafo, a biological psychologist at the University of Bristol, UK, who has a long-standing interest in scientific reproducibility. When he was a student, he says, "I tried to replicate what looked simple from the literature, and wasn't able to. Then I had a crisis of confidence, and then I learned that my experience wasn't uncommon."

The challenge is not to eliminate problems with reproducibility in published work. Being at the cutting edge of science means that sometimes results will not be robust, says Munafo. "We want to be discovering new things but not generating too many false leads."

THE SCALE OF REPRODUCIBILITY

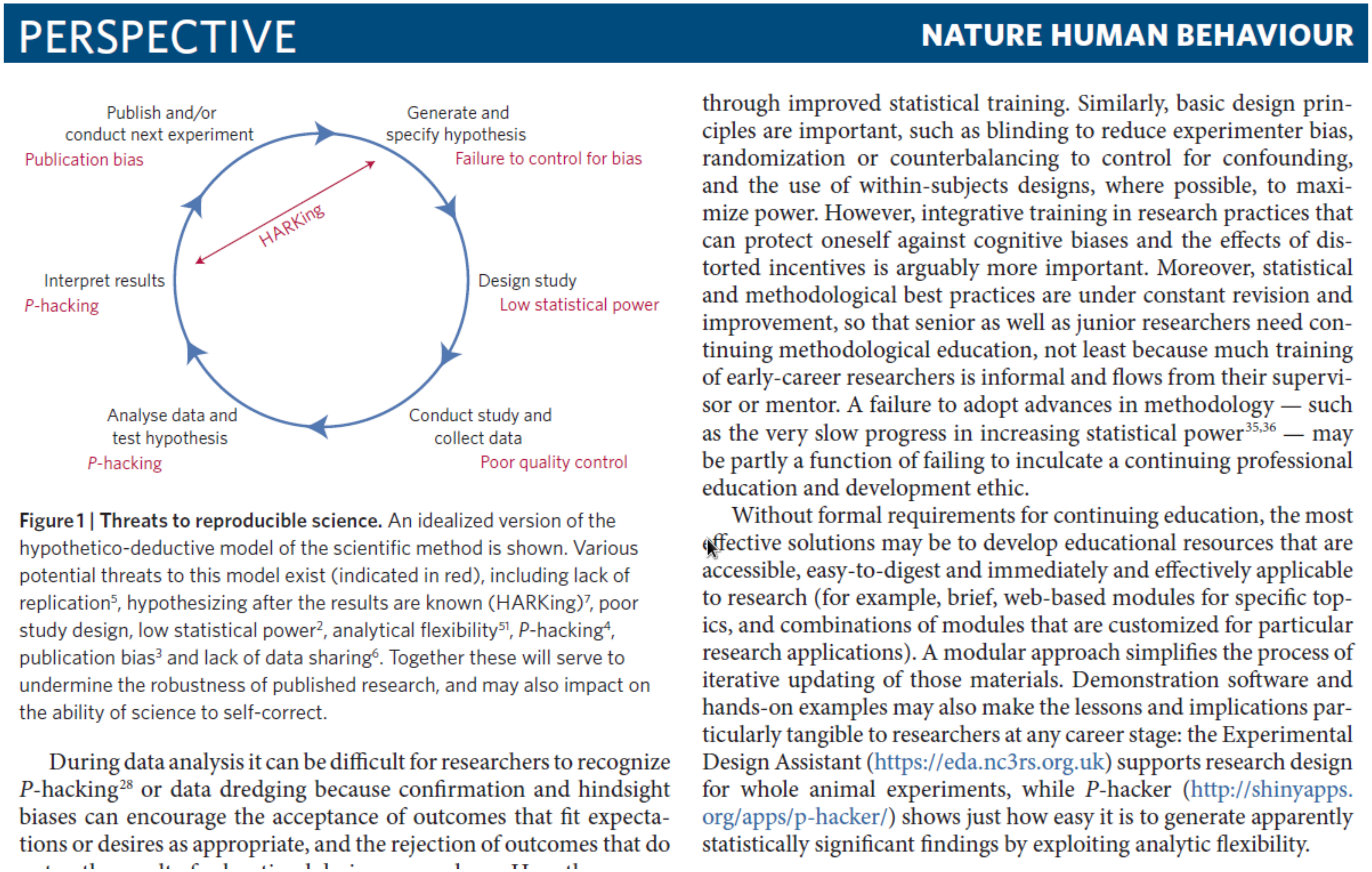
But sorting discoveries from false leads can be discomfiting. Although the vast majority of researchers in our survey had failed to reproduce an experiment, less than 20% of respondents said that they had ever been contacted by another researcher unable to reproduce their work (see 'A 'crisis' in numbers'). Our results are strikingly similar to another online survey of nearly 900 members of the American Society for Cell Biology (see go.nature.com/kbzs2b). That may be because such conversations are difficult. If experimenters reach out to the original researchers for help, they risk appearing incompetent or accusatory, or revealing too much about their own projects.

A minority of respondents reported ever having tried to publish

Nature 2016.

How reliable are the data you feed into the black box?

A manifesto for reproducible science



PERSPECTIVE **NATURE HUMAN BEHAVIOUR**

Publish and/or conduct next experiment
Publication bias
Generate and specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and test hypothesis
P-hacking
Conduct study and collect data
Poor quality control

Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

During data analysis it can be difficult for researchers to recognize *P*-hacking[28] or data dredging because confirmation and hindsight biases can encourage the acceptance of outcomes that fit expectations or desires as appropriate, and the rejection of outcomes that do

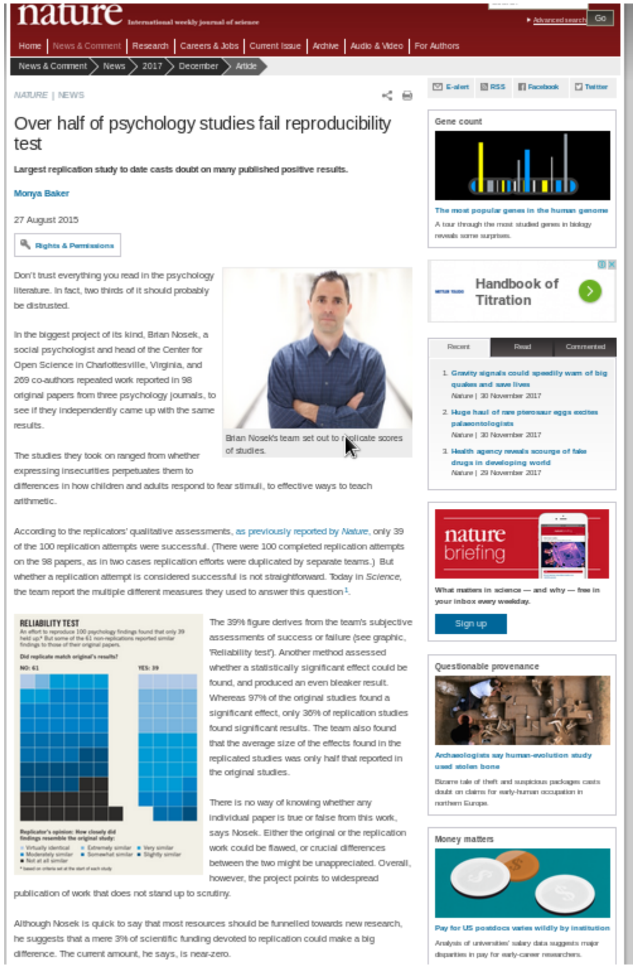
through improved statistical training. Similarly, basic design principles are important, such as blinding to reduce experimenter bias, randomization or counterbalancing to control for confounding, and the use of within-subjects designs, where possible, to maximize power. However, integrative training in research practices that can protect oneself against cognitive biases and the effects of distorted incentives is arguably more important. Moreover, statistical and methodological best practices are under constant revision and improvement, so that senior as well as junior researchers need continuing methodological education, not least because much training of early-career researchers is informal and flows from their supervisor or mentor. A failure to adopt advances in methodology — such as the very slow progress in increasing statistical power[35,36] — may be partly a function of failing to inculcate a continuing professional education and development ethic.

Without formal requirements for continuing education, the most effective solutions may be to develop educational resources that are accessible, easy-to-digest and immediately and effectively applicable to research (for example, brief, web-based modules for specific topics, and combinations of modules that are customized for particular research applications). A modular approach simplifies the process of iterative updating of those materials. Demonstration software and hands-on examples may also make the lessons and implications particularly tangible to researchers at any career stage: the Experimental Design Assistant (https://eda.nc3rs.org.uk) supports research design for whole animal experiments, while *P*-hacker (http://shinyapps.org/apps/p-hacker/) shows just how easy it is to generate apparently statistically significant findings by exploiting analytic flexibility.

Nature Human Behaviour 2017.

How reliable are the data you feed into the black box?

Reproducibility in psychology research


nature
International weekly journal of science
Advanced search
Go
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors
News & Comment › News › 2017 › December › Article
E-alert
RSS
Facebook
Twitter
RELIABILITY TEST
An effort to reproduce 100 psychology findings found that only 39 held up.* But some of the 61 non-replications reported similar findings to those of their original papers.
Did replicate match original's results?
NO: 61
YES: 39
Replicator's opinion: How closely did findings resemble the original study:
Virtually identical
Extremely similar
Very similar
Moderately similar
Somewhat similar
Slightly similar
Not at all similar
* based on criteria set at the start of each study
nature briefing

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Rights & Permissions

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

Brian Nosek's team set out to replicate scores of studies.

The studies they took on ranged from whether expressing insecurities perpetuates them to differences in how children and adults respond to fear stimuli, to effective ways to teach arithmetic.

According to the replicators' qualitative assessments, as previously reported by *Nature*, only 39 of the 100 replication attempts were successful. (There were 100 completed replication attempts on the 98 papers, as in two cases replication efforts were duplicated by separate teams.) But whether a replication attempt is considered successful is not straightforward. Today in *Science*, the team report the multiple different measures they used to answer this question[1].

The 39% figure derives from the team's subjective assessments of success or failure (see graphic, 'Reliability test'). Another method assessed whether a statistically significant effect could be found, and produced an even bleaker result. Whereas 97% of the original studies found a significant effect, only 36% of replication studies found significant results. The team also found that the average size of the effects found in the replicated studies was only half that reported in the original studies.

There is no way of knowing whether any individual paper is true or false from this work, says Nosek. Either the original or the replication work could be flawed, or crucial differences between the two might be unappreciated. Overall, however, the project points to widespread publication of work that does not stand up to scrutiny.

Although Nosek is quick to say that most resources should be funnelled towards new research, he suggests that a mere 3% of scientific funding devoted to replication could make a big difference. The current amount, he says, is near-zero.

Gene count

The most popular genes in the human genome

A tour through the most studied genes in biology reveals some surprises.

Recent | Read | Commented

1. Gravity signals could speedily warn of big quakes and save lives
Nature | 30 November 2017
2. Huge haul of rare pterosaur eggs excites palaeontologists
Nature | 30 November 2017
3. Health agency reveals scourge of fake drugs in developing world
Nature | 29 November 2017

What matters in science — and why — free in your inbox every weekday.

Sign up

Questionable provenance

Archaeologists say human-evolution study used stolen bone

Bizarre tale of theft and suspicious packages casts doubt on claims for early-human occupation in northern Europe.

Money matters

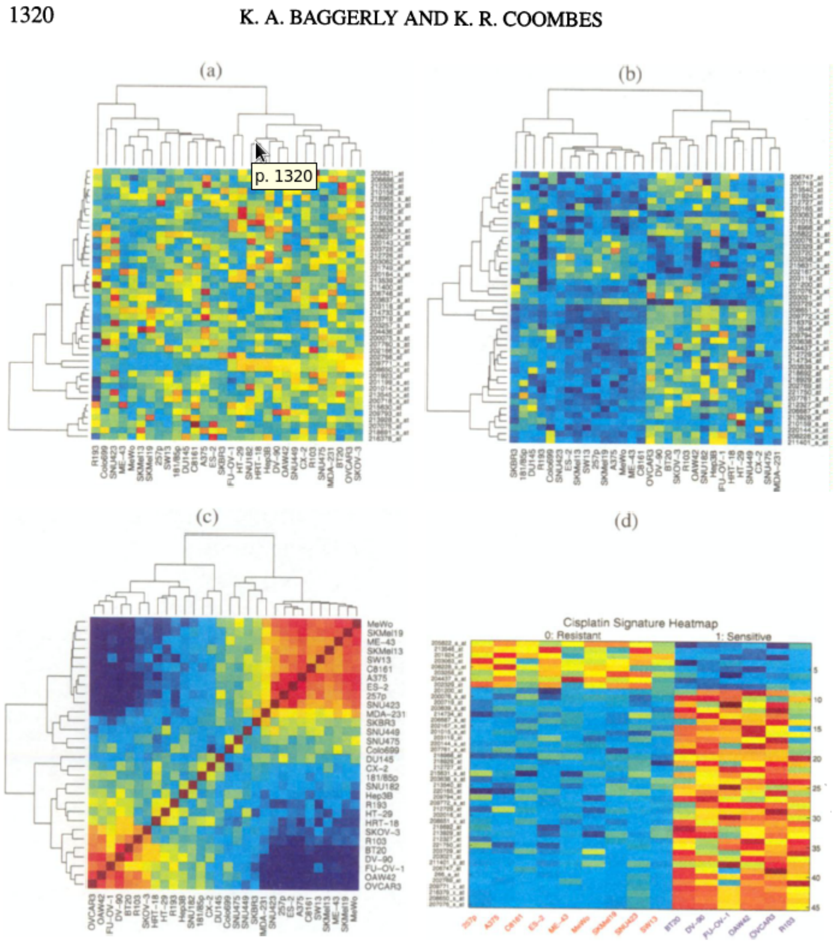
Pay for US postdocs varies wildly by institution

Analysis of universities' salary data suggests major disparities in pay for early-career researchers.

Nature News 2015.

How reliable are the data you feed into the black box?

Forensic bioinformatics



1320 K. A. BAGGERLY AND K. R. COOMBES

Baggerly and Coombes, The Annals of Applied Statistics, 2009.

How reliable are the data you feed into the black box?

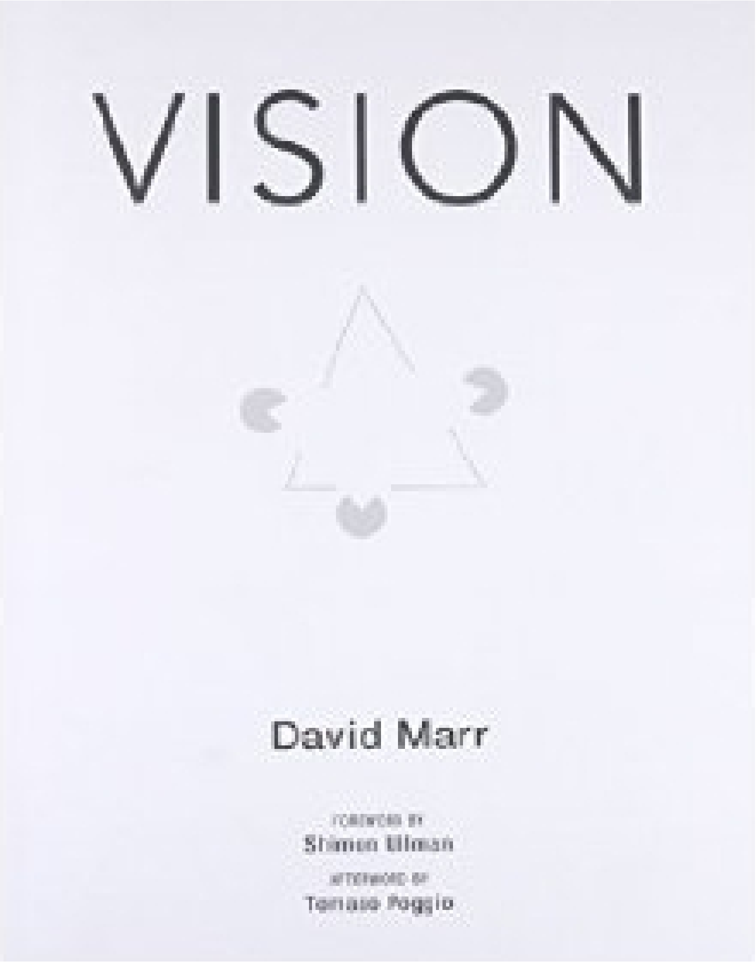
My own experience: About 1/5 of the acquaintance network data are mis-annotated.

How shall we handle the error-prone data?

- Always do quality check.
- Always do sanity check.
- Draw your conclusion from multiple studies.
- Validate your findings on external datasets.

Tip 3: Thou shall understand the level of the problem

David Marr's three levels



VISION

David Marr

FOREWORD BY
Shimon Ullman

AFTERWORD BY
Tomaso Poggio

Vision, by David Marr, W.H. Freeman and Company, New York, 1982.

David Marr’s three levels

An information processing problem can be decomposed into problems at three levels:

1. Computational.

2. Representational and algorithmic.

3. Implementational.

David Marr’s three levels

Example 1: A cash register.

1. Computational: Arithmetic theory.

2. Representational and algorithmic: Decimal system, binary system, additions and subtractions you learned from the grade school, etc.

3. Implementational: Pencil and paper (brain), abacus, digital computer, etc.

David Marr's three levels

Example 2: Edge detection in vision.

1. Computational: Detecting edges in an image.
2. Representational and algorithmic: Laplacian operator.
3. Implementational: A neural network.

Understand the level of the problem you are facing

- Which level is the major bottleneck to the progress?
- Which level is tractable or intractable with the current technology?
- Which level will you put the effort on?

Tip 4: Thou shall not mis-recognize random patterns as true signals

Einstein's picture is found in random images

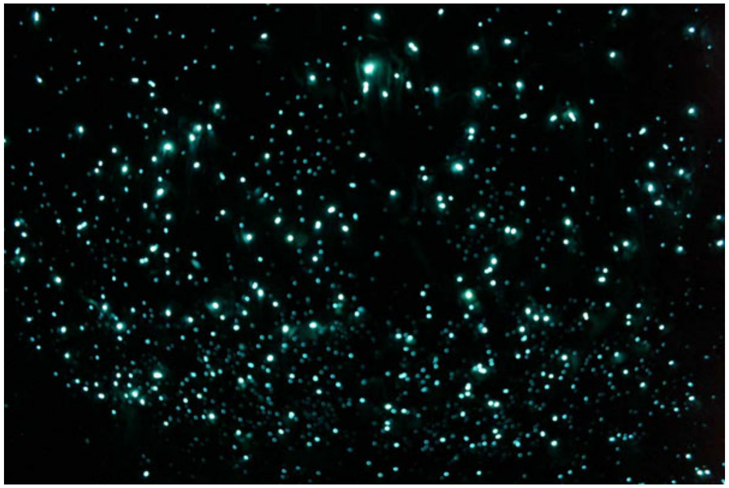
An Einstein's photo can be found from random images.

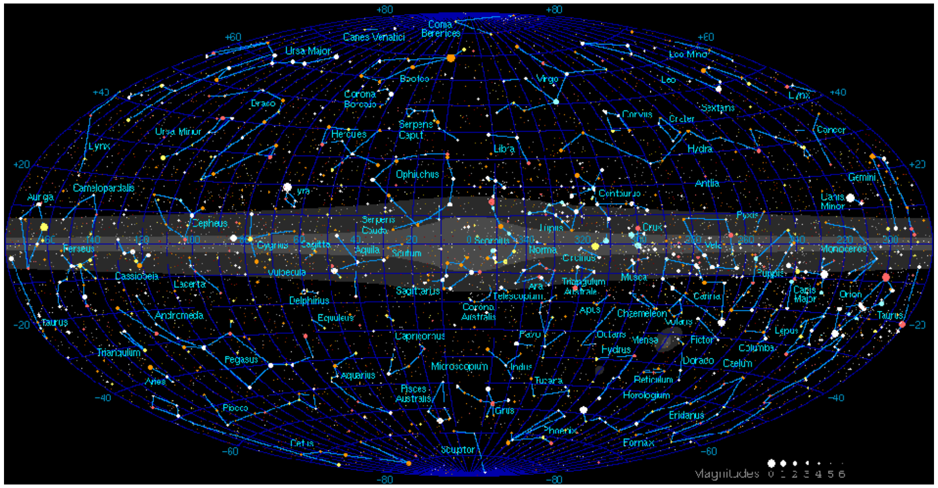
Henderson, PNAS, 2013.

cf Prof. I-Ping Tu's talk.

Glowworms and constellations

Why do the glowworm light spots from Waitomo, New Zealand not form constellations?





Glow, Big Glowworm, by Stephen Jay Gould, Natural History, 1986.

Ramsey's theorem

(Two-colored case) Any graph of $R(k, l)$ nodes contains at least k mutually adjacent nodes and l mutually disconnected nodes.

In a large enough party, you will sure to find k persons who know each other and l persons who don't know each other.

How to avoid recognizing random patterns?

That's the job of statisticians. There are many variations of this task, such as

- Hypothesis testing.
- Model selection.
- Generalization error.
- Non-parametric tests.
- Cross validation.

Tip 5: Thou shall select the right tools for the problem

A big pool of machine learning toolkits

- Multilayer neural networks.
- Support vector machines.
- Probabilistic graphical models.
- Dimensional reduction and decomposition.
- Manifold learning.

Which tool to pick?

Different tools have different properties

- Multilayer neural networks – large sample size.
  - Convolutional neural networks – image data containing similar components.
  - Recurrent neural networks – sequential data such as speech and texts.
  - Generative adversarial networks – unsupervised learning such as density estimation.
- Support vector machine – classification with high dimensional data.
- Probabilistic graphical models – capturing dependency of many random variables.
- Dimensional reduction and decomposition – many redundant or mutually dependent random variables with relatively small sample size.
  - PCA, SVD, NNF – linear decomposition of numerical data.
  - Topical models – decomposition of categorical data (such as texts).
  - ICA – nonlinear decomposition of numerical data.
- Manifold learning – nonlinear approximation of high dimensional data.

Tip 6: Thou shall not confuse correlation with causation

Polio and ice cream

- Polio is one of the most devastating diseases of children before mid 20th century.
- Epidemiologists found polio infection surged during the summer.
- Ice cream sales also peaked in the summer.
- Some drew the conclusion that eating ice cream could cause polio.
- There was a public campaign against eating ice cream to prevent polio infection.
- Source: Freakonomics, https://www.youtube.com/watch?v=lbODqslc4Tg.

Big data, correlation and causation

- The fallacy of confusing correlation with causation is more liable to happen in the era of big data.

- Many confounding variables probed, unstructured data, less or no control over collected sample.

- Need to be much more cautious about causality inference from big data.

Probing correlation from big data

- Most machine learning (and statistical) methods pertain to discover correlation.
- Regression.
- Classification.
- Clustering.
- Density estimation.
- Dependency tests.

Inferring causation from big data

- The best (and probably the only affirmative) way to verify causality is through explicit intervention vs control on randomized populations.
- Randomized trials are often expensive (e.g., biology, medicine), intractable (e.g., economics, politics), or impossible (e.g., astronomy, evolution).
- Various statistical techniques have been developed to infer causality from passively collected data.
- Instrumental variables.
- D-separation.
- Structural equation models.
- Mediation analysis.
- Co-integration.
- Nonlinear functional forms.
- Have to be aware of their underlying assumptions.

Tip 7: Thou shall exploit the power of automation

Example 1: Robot scientist

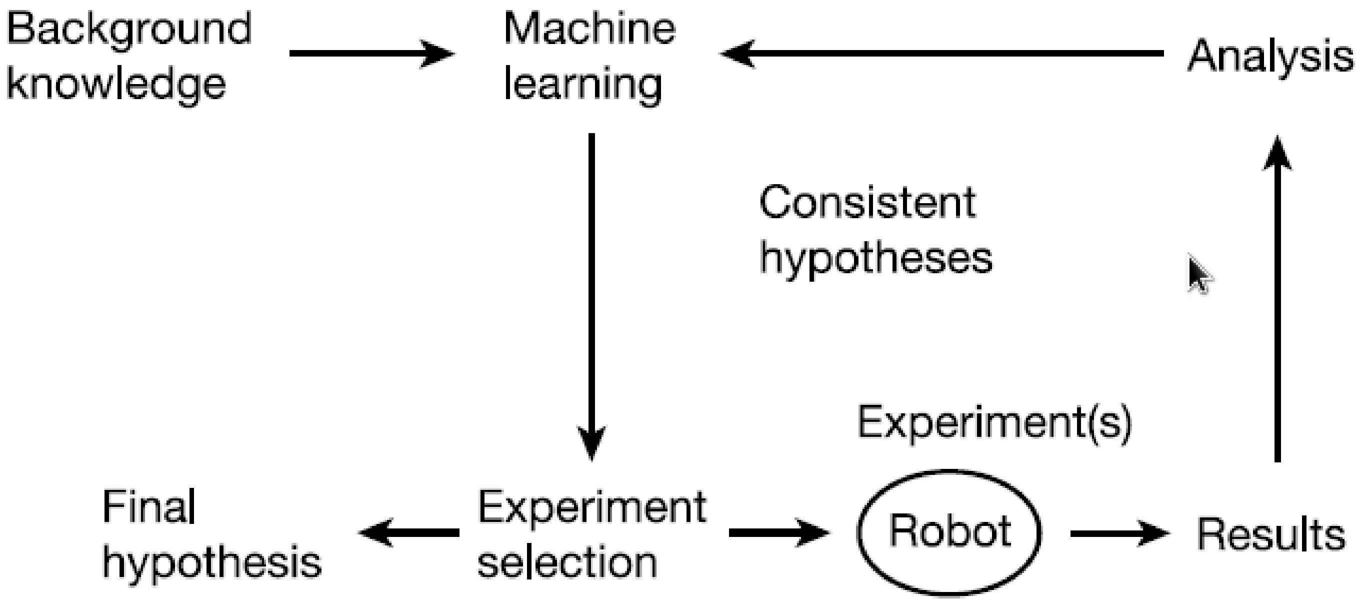


Figure 1 The Robot Scientist hypothesis-generation and experimentation loop.

King et al., Nature, 2004.

Example 2: Composing Bach-style music

https://www.youtube.com/watch?v=PczDLl92vlc

Experiments in musical intelligence, by David Cope.

Example 3: Brain-machine interface

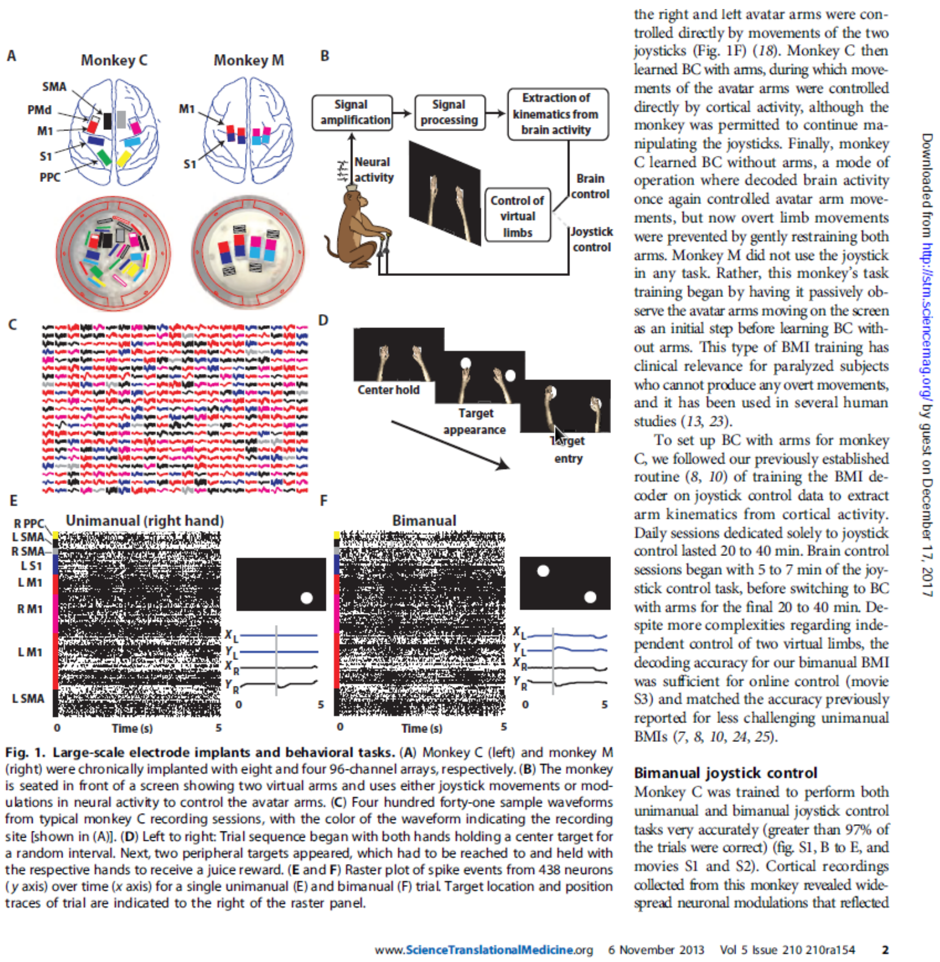


Fig. 1. Large-scale electrode implants and behavioral tasks. (**A**) Monkey C (left) and monkey M (right) were chronically implanted with eight and four 96-channel arrays, respectively. (**B**) The monkey is seated in front of a screen showing two virtual arms and uses either joystick movements or modulations in neural activity to control the avatar arms. (**C**) Four hundred forty-one sample waveforms from typical monkey C recording sessions, with the color of the waveform indicating the recording site [shown in (A)]. (**D**) Left to right: Trial sequence began with both hands holding a center target for a random interval. Next, two peripheral targets appeared, which had to be reached to and held with the respective hands to receive a juice reward. (**E** and **F**) Raster plot of spike events from 438 neurons (*y* axis) over time (*x* axis) for a single unimanual (E) and bimanual (F) trial. Target location and position traces of trial are indicated to the right of the raster panel.

the right and left avatar arms were controlled directly by movements of the two joysticks (Fig. 1F) (*18*). Monkey C then learned BC with arms, during which movements of the avatar arms were controlled directly by cortical activity, although the monkey was permitted to continue manipulating the joysticks. Finally, monkey C learned BC without arms, a mode of operation where decoded brain activity once again controlled avatar arm movements, but now overt limb movements were prevented by gently restraining both arms. Monkey M did not use the joystick in any task. Rather, this monkey's task training began by having it passively observe the avatar arms moving on the screen as an initial step before learning BC without arms. This type of BMI training has clinical relevance for paralyzed subjects who cannot produce any overt movements, and it has been used in several human studies (*13*, *23*).

To set up BC with arms for monkey C, we followed our previously established routine (*8*, *10*) of training the BMI decoder on joystick control data to extract arm kinematics from cortical activity. Daily sessions dedicated solely to joystick control lasted 20 to 40 min. Brain control sessions began with 5 to 7 min of the joystick control task, before switching to BC with arms for the final 20 to 40 min. Despite more complexities regarding independent control of two virtual limbs, the decoding accuracy for our bimanual BMI was sufficient for online control (movie S3) and matched the accuracy previously reported for less challenging unimanual BMIs (*7*, *8*, *10*, *24*, *25*).

Bimanual joystick control

Monkey C was trained to perform both unimanual and bimanual joystick control tasks very accurately (greater than 97% of the trials were correct) (fig. S1, B to E, and movies S1 and S2). Cortical recordings collected from this monkey revealed widespread neuronal modulations that reflected

Ifft et al., Science Translational Medicine, 2013.

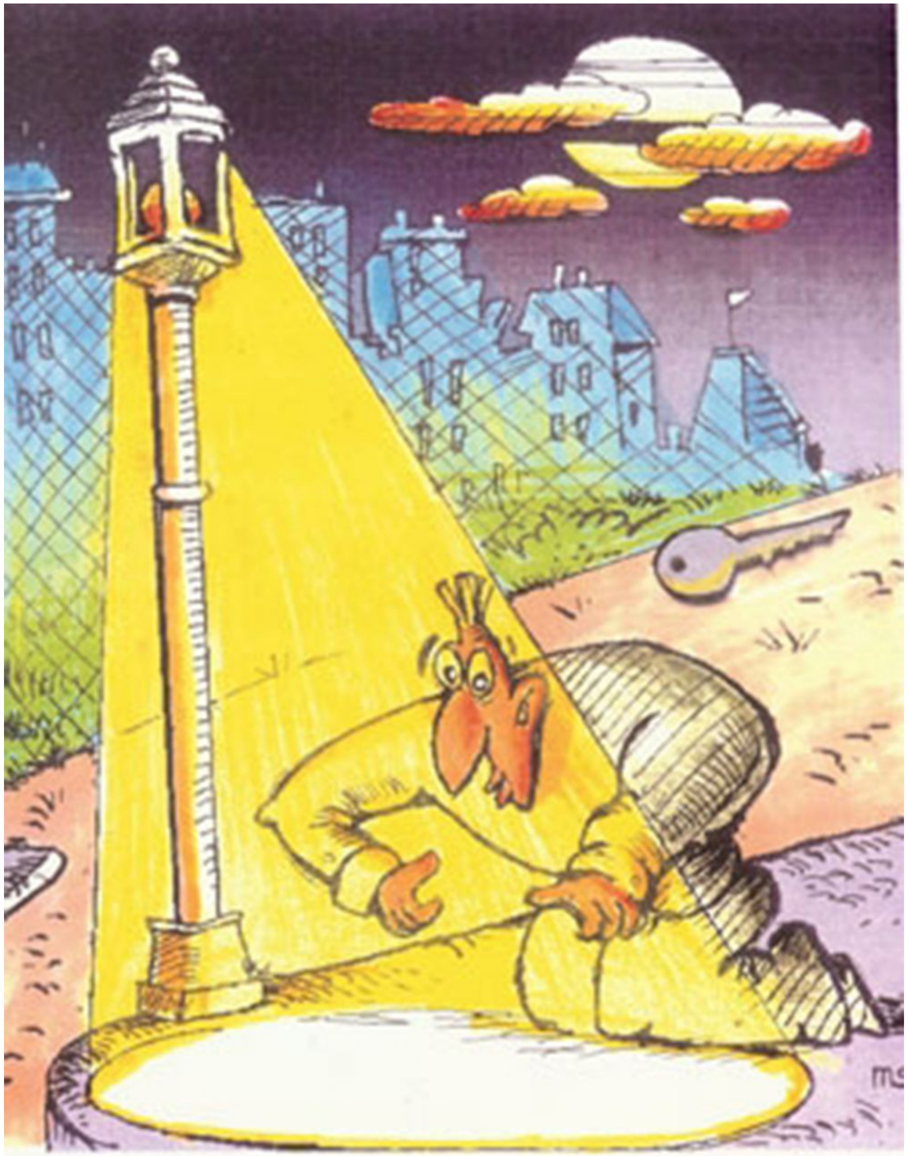
What aspects of the job(s) can be automated?

- Data processing.
- Inference.
- Pattern finding and generation.
- Movements.

What aspects cannot be automated?

Tip 8: Thou shall not look for answers from the wrong data

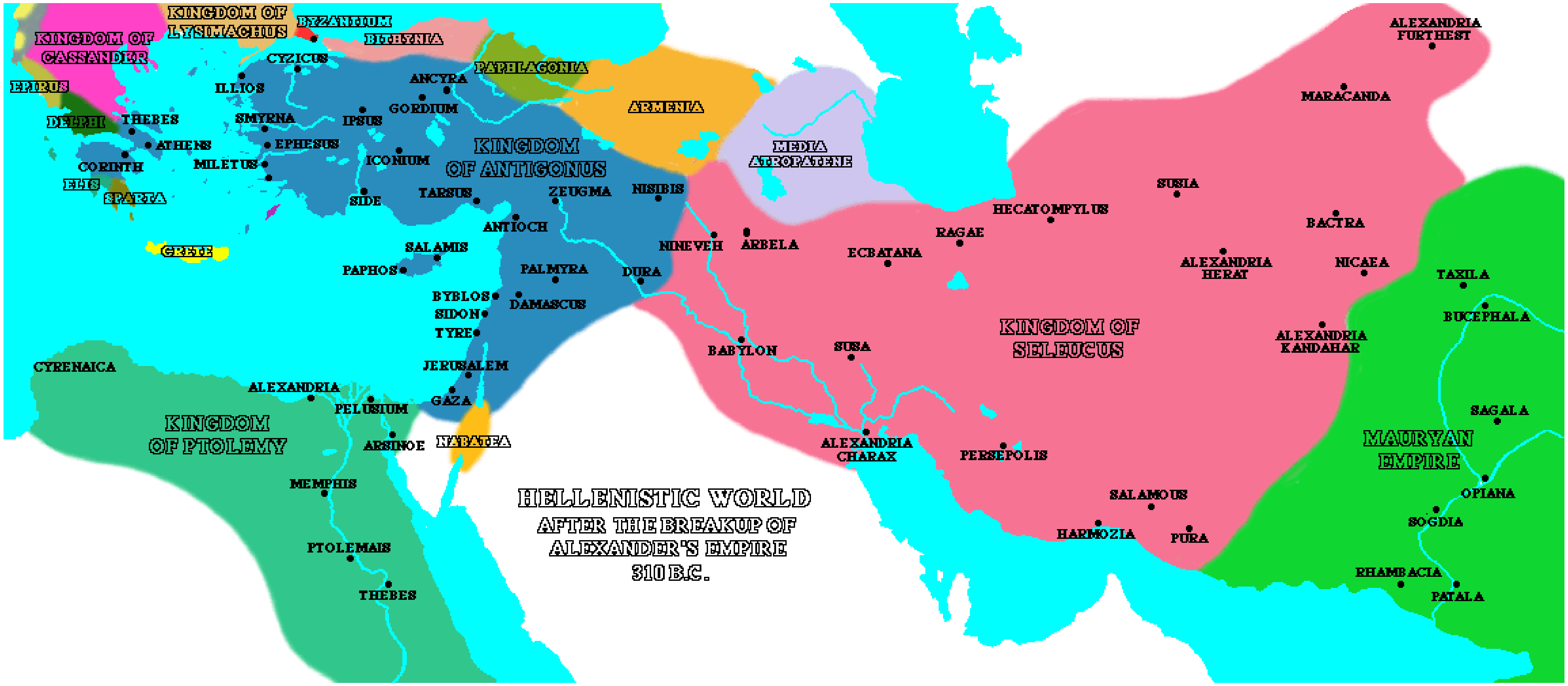
Streetlight Effect



Lernmark, Diabetes, 2015

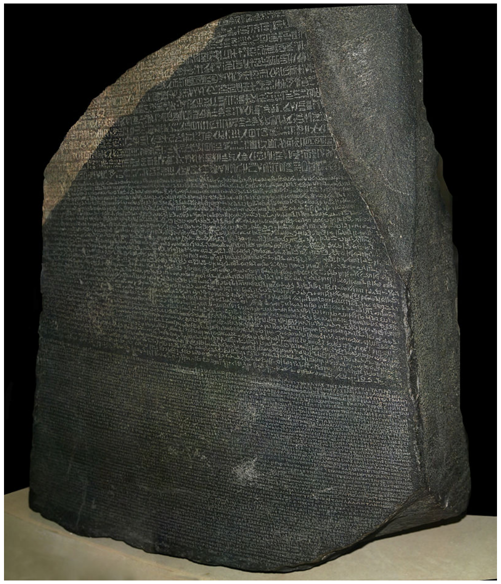
People are searching for something and look only where it is easiest.

Two heirs of the Macedonian empire



http://www.houseofptolemy.org/housemap.htm

Observational bias in studying the two kingdoms



- Seleucid Kingdom had more influence in the Hellenistic world but is less studied due to the shortage of scripts.

- Ptolemy Kingdom was less influential but is intensively studied due to the abundant papyrus.

A study of history, by Arnold Toynbee, 1934-1961.

Observational bias can be a major impediment to “big data”

- All data (including big data) are restricted to ethical, social, economic and technological conditions.
  - Web documents retrieved from keyword search are highly redundant.
  - Unlabeled data are far more abundant than labeled data.
  - Social interactions and influences are segregated and self-reinforced.
  - Longitudinal data are highly biased toward recent collections.
  - Nature (OMICs) is much more emphasized than nurture (environment).
- After correcting the observation biases big data may be no longer big.
- In science, the data driven approach is still a screening tool to identify viable hypotheses for further verification.
- Data collection through designed experiments is still the best way to test scientific hypotheses, yet is often expensive and time-consuming.
- Perhaps a compromise is an interative approach mixing data mining and hypothesis testing.

Tip 9: Thou shall take commonsense into account

Commonsense reasoning: The ultimate frontier of artificial intelligence?

- Judgements about the physical properties, purpose, intentions and behavior of people and objects.

- Possible outcomes of actions and interactions.

- Very difficult since human beings inherit and learn those knowledge constantly, and they are highly context dependent.

- Do Eliza and Siri really understand human words?

Different levels of commonsense in the models

- Domain knowledge is critical.
- Computer programs in 1950s proved all theorems in Principia Mathematica by Whitehead and Russell.
- Robot Scientist in early 2000s inferred and tested a small metabolic network.
- AlphaGO in 2017 beat human players by learning chess playing from scratch.
- More “soft” knowledge is more difficult to formalize, thus require more inputs from domain experts.

Tip 10: Thou shall not overlook various aspects of computation

A brief survey of computational literacy, part 1

How many of you have

- Implemented a sort program (instead of calling the sort function)?
- Calculated the shortest distance between every pair of nodes in a graph?
- Transposed a large (say $10^6 \times 10^6$) matrix?
- Inverted a large (say $10^6 \times 10^6$) matrix?

A brief survey of computational literacy, part 2

How many of you have programming experience in

- PCs (including Linux, Windows and Mac OS)?
- Cluster servers?
- Cloud computing?
- GPUs?
- Distributed databases?
- Mobile APPs?

Computing becomes more critical as the data size grows bigger

Statisticians tend to have profound understanding and ideas about data but less skillful in realizing them in the big data era.

- What's the space and time complexity of the problem?
- Are there polynomial (or linear, sublinear) time algorithms?
- If not, are there good approximation, randomized algorithms or heuristics?
- What programming languages and platforms to choose?
- How to translate your programs in a parallel and distributed environment?
- Are you willing to outsource these tasks?

Conclusion – Episteme and Techne

Heidegger's remarks about episteme and techne:

- Episteme, in Greek term, is knowledge about the world.
- Techne has the meaning beyond instrumentality.
- It is also a mode of revealing.
- In this regard techne is not only a mean to achieve better life, but also a process of acquiring episteme.
- Likewise poeisis (poetry) is a way of bringing forth unrevealed.
- Perhaps one should ponder about what can be revealed by the techne of AI in data science.

The question concerning technology, by Martin Heidegger, 1954.

Outline

- Lecture.
- **Assignments and discussion.**
- Presentation.

Exercise: Proposing action plans for data analysis

- You've learned a lot about data analysis. Now it's time to employ your skills in solving real problems.

- We cannot really do that since real world problems typically take months or even years to solve.

- Instead this exercise gives you a flavor about data analysis jobs in practice.

- Divide students into five teams. Each team consists of 4-6 members.

- Each team picks one designated task below.

- Team members spend one hour in discussion and come up with an "action plan" of data analysis.

- After one hour, each team gives a 12-minutes presentation about the action plan. Other students raise questions and comment on the plans.

What will be covered in an action plan?

- What's the goal(s) of the project?
- What are the required data? How will you acquire those data?
- How do you control and maintain the quality of the data?
- How will you analyze the data? Which methods will you choose?
- How will you validate and justify the findings from data analysis?
- What computing infrastructure do you need?

Task 1: Mining the Taiwan Biobank data

- You are biostatisticians/bioinformaticians jointly hired by Ministry of Health and Welfare and Academia Sinica.

- You are given a vague task of “understanding and improving the health condition of Taiwanese populations using the Taiwan Biobank data”.

- The (future) Taiwan Biobank collects the genomic, transcriptomic and epigenomic data of 100,000 Taiwanese subjects. It also collects general health information of those subjects, their geographic locations and basic socioeconomic information.

- It is possible to link the Biobank data with other datasets such as the patients records from Taiwan’s universal health plan. Yet you need another proposal for passing the Institutional Review Board (IRB) review.

Task 2: Profiting from the volumous transaction records in Alibaba

- You are in an elite data analysis team in the Alibaba Conglomerate, and have the priviledge to access all the transaction records on the Taubau Platform.

- Your boss, Jack Ma, asks the team leader what you can do with the data to benefit the corporate.

- The data consists of trillions of transactions records including the personal identifications of buyers and sellers, the items and prices of the purchases, and locations and times of those transactions.

Task 3: Making sense of the data from the Sloan Digital Sky Survey

- You are in a team composed of dedicated astronomers, data scientists, and amateur star watchers.

- The team has a firm conviction that important astronomical discoveries can be drawn from publically available data.

- The Sloan Digital Sky Survey provides detailed three-dimensional maps of the visible universe, including the images of stars, galaxies, and interstellar materials.

- You can also access other public data of a wide range of electromagnetic wave spectra.

- There are no issues about privacy or conflict of interests.

Task 4: Drafting the long-term global warming response plan for New York City

- You are hired by the New York City government to draft a global warming response plan for the Big Apple.

- NYC is deeply impacted recently by extreme meterological events (e.g., hurricane Sandy, large snow storms, deadly heat waves in summer), and the Federal government is unlikely to take actions to alleviate climate change.

- Thus, the mayor decides to act independently to “plan for the worst” in response to global warming.

- The goal of the team is to assess the impacts of climate change in all aspects of the city, and proposes proper plans to alleviate those impacts.

- The government provides full support for access of all public data (at federal, state and municipal levels) and purchases for necessary private data. You may also propose plans to collect additional data.

Task 5: Alleviating poverty in African countries

- You are in a joint team sponsored by World Bank, African Development Bank, and private foundations for a special task force of poverty alleviation in African countries.

- The top-100 billionaires in the world decide to jointly donate 10 billion USD to engage a "war on poverty" in Africa. WB and ADB decide to match up.

- The goal of the project is to drastically reduce (or eliminate) the poor population in African countries.

- 20 billion USD is a lot but not sufficient to boost everyone's income level above the poverty line.

- Your team has to inform the committee how to spend the money to optimize the outcome.

- The committee provides full support for access of all public data and purchases for necessary private data. You may also propose plans to collect additional data.

Outline

- Lecture.
- Assignments and discussion.
- **Presentation.**