Data Science for Actuaries (ACT6100)

Arthur Charpentier

Intro # 3 (Perspective Historique)

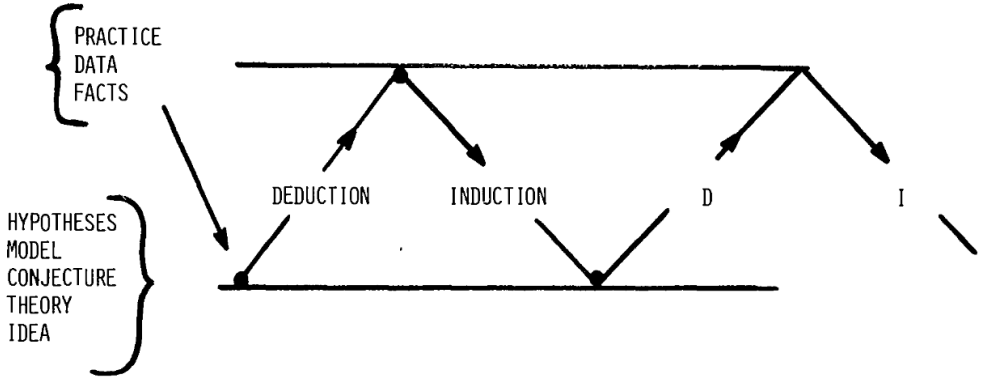
automne 2020

https://github.com/freakonometrics/ACT6100/

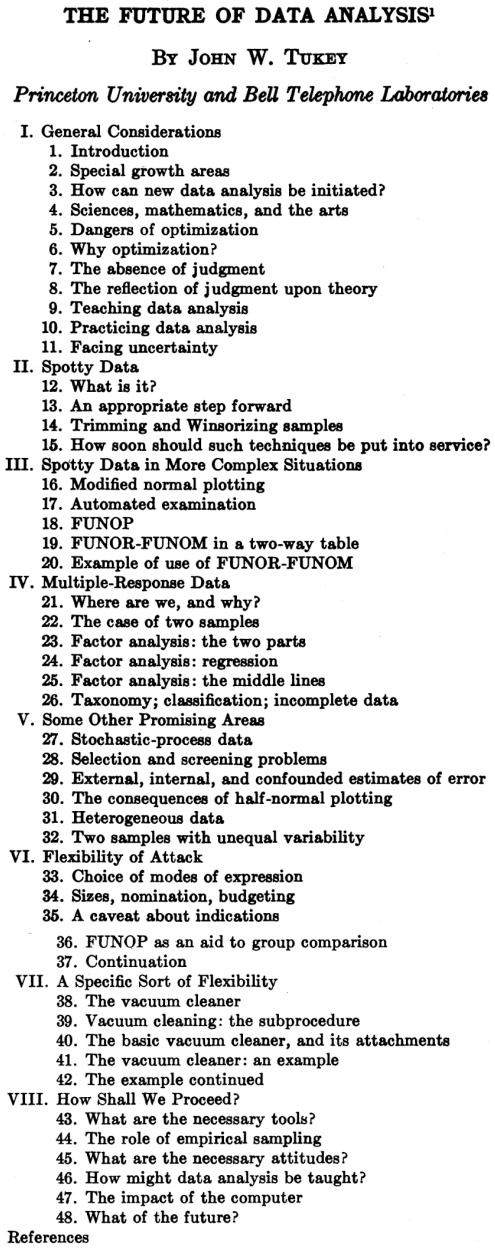
Data Science

"*The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning... Data-driven predictions can succeed–and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves*" Nate Silver (2010)

See Tukey (1962)'s vision on data analysis



Box (1976)'s inductive-deductive reasoning cycle



THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

I. General Considerations
1. Introduction
2. Special growth areas
3. How can new data analysis be initiated?
4. Sciences, mathematics, and the arts
5. Dangers of optimization
6. Why optimization?
7. The absence of judgment
8. The reflection of judgment upon theory
9. Teaching data analysis
10. Practicing data analysis
11. Facing uncertainty
II. Spotty Data
12. What is it?
13. An appropriate step forward
14. Trimming and Winsorizing samples
15. How soon should such techniques be put into service?
III. Spotty Data in More Complex Situations
16. Modified normal plotting
17. Automated examination
18. FUNOP
19. FUNOR-FUNOM in a two-way table
20. Example of use of FUNOR-FUNOM
IV. Multiple-Response Data
21. Where are we, and why?
22. The case of two samples
23. Factor analysis: the two parts
24. Factor analysis: regression
25. Factor analysis: the middle lines
26. Taxonomy; classification; incomplete data
V. Some Other Promising Areas
27. Stochastic-process data
28. Selection and screening problems
29. External, internal, and confounded estimates of error
30. The consequences of half-normal plotting
31. Heterogeneous data
32. Two samples with unequal variability
VI. Flexibility of Attack
33. Choice of modes of expression
34. Sizes, nomination, budgeting
35. A caveat about indications
36. FUNOP as an aid to group comparison
37. Continuation
VII. A Specific Sort of Flexibility
38. The vacuum cleaner
39. Vacuum cleaning: the subprocedure
40. The basic vacuum cleaner, and its attachments
41. The vacuum cleaner: an example
42. The example continued
VIII. How Shall We Proceed?
43. What are the necessary tools?
44. The role of empirical sampling
45. What are the necessary attitudes?
46. How might data analysis be taught?
47. The impact of the computer
48. What of the future?
References

Data vs. Algorithmic Modeling

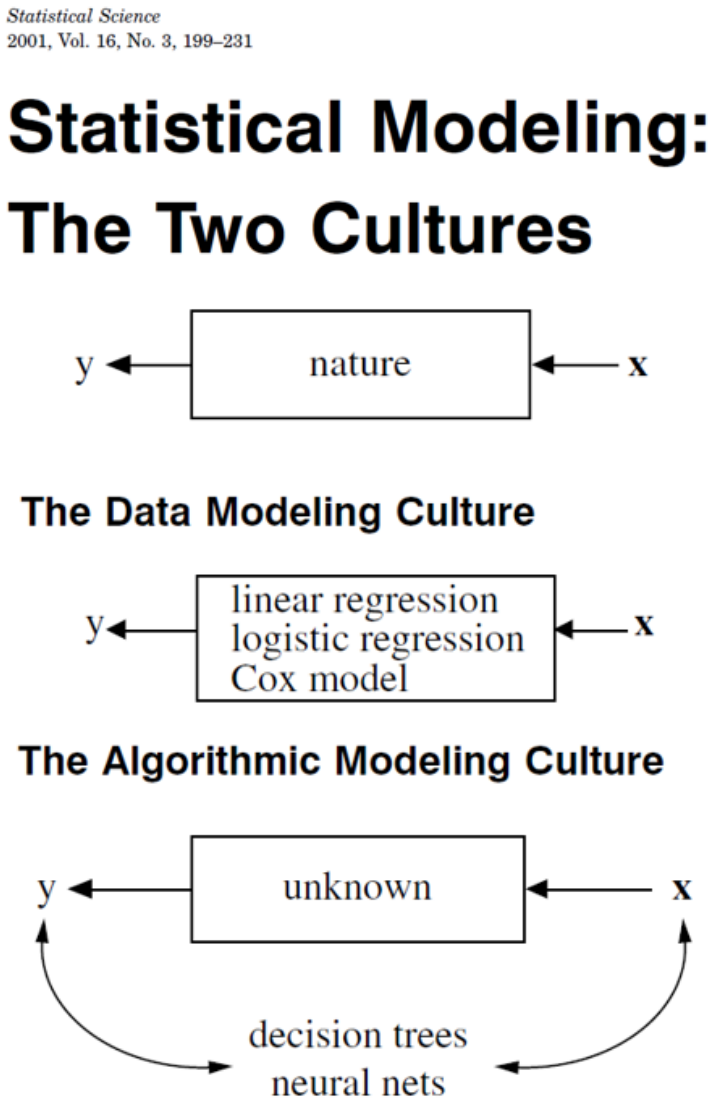
Breiman (2001)'s two cultures

- Data Modeling

choose a simple (linear) model based on intuition about the data-generating mechanism. Emphasis is on model interpretability and validation, if done at all, is done through goodness-of-fit.

- Algorithmic Modeling

choose the model with the highest predictive validation accuracy with no consideration for model explainability.



Machine Learning or Data Mining ?

Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305
jhf@stat.stanford.edu

Abstract

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets. In spite of (or perhaps because of) the somewhat exaggerated hype, this field is having a major impact in business, industry, and science. It also affords enormous research opportu-

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. - Zekulin.

Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. - Ferruzza.

Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. - Parsaye

Data Mining is ...

- Decision Trees
- Neural Networks
- Rule Induction
- Nearest Neighbors
- Genetic Algorithms

  - Mehta

Friedman (1997)

Regression

Legendre (1805) introduced the "*méthode des moindres carrés* " and Gauss (1821)

> Pour cet effet, la méthode qui me paroît la plus simple et la plus générale, consiste à rendre *minimum* la somme des quarrés des erreurs. On obtient ainsi autant d'équations qu'il y a de coëfficiens inconnus ; ce qui achève de déterminer tous les élémens de l'orbite.
>
> Comme la méthode dont je viens de parler, et que j'appelle *Méthode des moindres quarrés*, peut être d'une grande utilité

Galton (1870, 1886) and Pearson & Lee (1896, 1903) studied genetic transmission of characterisitcs, e.g. the heigth.

On average the child of tall parents is taller than other children, but less than his parents.

"*I have called this peculiarity by the name of regression* ", Galton (1886).

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	5	5	10	4	9	2	2	45	11	69.9
70.5	1	—	1	—	1	1	5	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	158	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1888a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties; he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

Soit p la population : représentons par dp l'accroissement infiniment petit qu'elle reçoit pendant un temps infiniment court dt. Si la population croissait en progression géométrique, nous aurions l'équation $\frac{dp}{dt} = mp$. Mais comme la vitesse d'accroissement de la population est retardée par l'augmentation même du nombre des habitans, nous devrons retrancher de mp une fonction inconnue de p; de manière que la formule à intégrer deviendra

$$\frac{dp}{dt} = mp - \varphi(p).$$

L'hypothèse la plus simple que l'on puisse faire sur la forme de la fonction φ, est de supposer $\varphi(p) = np^2$. On trouve alors pour intégrale de l'équation ci-dessus

$$t = \frac{1}{m}\left[\log. p - \log.(m - np)\right] + \text{constante},$$

et il suffira de trois observations pour déterminer les deux coefficiens constans m et n et la constante arbitraire.

En résolvant la dernière équation par rapport à p, il vient

$$p = \frac{mp'\, e^{mt}}{np'\, e^{mt} + m - np'} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

en désignant par p' la population qui répond à $t = o$, et par e la base des logarithmes népériens. Si l'on fait $t = \infty$, on voit que la valeur de p correspondante est $P = \frac{m}{n}$. Telle est donc *la limite supérieure de la population.*

Logistic Regression

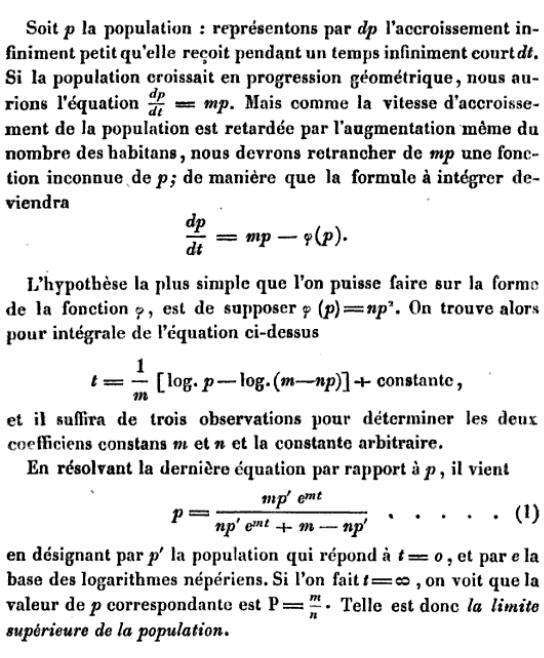
Historically, the idea was to model the *odds ratio*,

$$\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y\neq 1|\boldsymbol{X}=\boldsymbol{x})}=\exp[\boldsymbol{x}^{\mathsf{T}}\beta]$$

(the odds ratio is a positive number). Hence,

$$\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})=H(\boldsymbol{x}^{\mathsf{T}}\beta)$$

where $H(\cdot)=\dfrac{\exp[\cdot]}{1+\exp[\cdot]}$ is the c.d.f. of the logistic variable, popular in demography, see Verhulst (1845), and then Pearl & Reed (1920) or Yule (1925)



Soit p la population : représentons par dp l'accroissement infiniment petit qu'elle reçoit pendant un temps infiniment court dt. Si la population croissait en progression géométrique, nous aurions l'équation $\frac{dp}{dt}=mp$. Mais comme la vitesse d'accroissement de la population est retardée par l'augmentation même du nombre des habitans, nous devrons retrancher de mp une fonction inconnue de p; de manière que la formule à intégrer deviendra

$$\frac{dp}{dt}=mp-\varphi(p).$$

L'hypothèse la plus simple que l'on puisse faire sur la forme de la fonction φ, est de supposer $\varphi(p)=np^2$. On trouve alors pour intégrale de l'équation ci-dessus

$$t=\frac{1}{m}\left[\log.p-\log.(m-np)\right]+\text{constante},$$

et il suffira de trois observations pour déterminer les deux coefficiens constans m et n et la constante arbitraire.

En résolvant la dernière équation par rapport à p, il vient

$$p=\frac{mp'e^{mt}}{np'e^{mt}+m-np'}\quad\ldots\ldots\quad(1)$$

en désignant par p' la population qui répond à $t=o$, et par e la base des logarithmes népériens. Si l'on fait $t=\infty$, on voit que la valeur de p correspondante est $P=\frac{m}{n}$. Telle est donc *la limite supérieure de la population*.

Probit Regression

Bliss (1934) suggested a model such that

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = H(\boldsymbol{x}^{\mathsf{T}}\beta)$$

where $H(\cdot) = \Phi(\cdot)$ is the c.d.f. of the $\mathcal{N}(0, 1)$ distribution.

This is the probit model.
This yields a latent model,

$$y_i = \mathbf{1}(y_i^\star > 0)$$

where $y_i^\star = \boldsymbol{x}_i^{\mathsf{T}}\beta + \varepsilon_i$.

Table 3.2 Transformation of percentages to probits

%	0	1	2	3	4	5	6	7	8	9
0	—	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
—	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

Mathematical Statistics

The philosophy of the classical parametric paradigm is based on the following three beliefs:

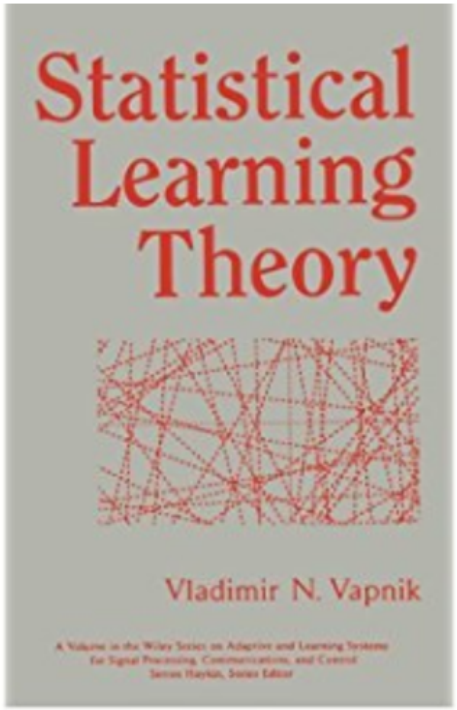
1. *To find a functional dependency from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation to the desired function. The number of free parameters describing this set is small.*

2. *The statistical law underlying the stochastic component of most real-life problems is the normal law.*

3. *The induction engine in this paradigm—the maximum likelihood method —is a good tool for estimating parameters.*

Note that these three beliefs were also supported by the philosophy:

If there exists a mathematical proof that some method provides an asymptotically optimal solution, then in real life this method will provide a reasonable solution for a small number of data samples.



in Vapnik (1998)

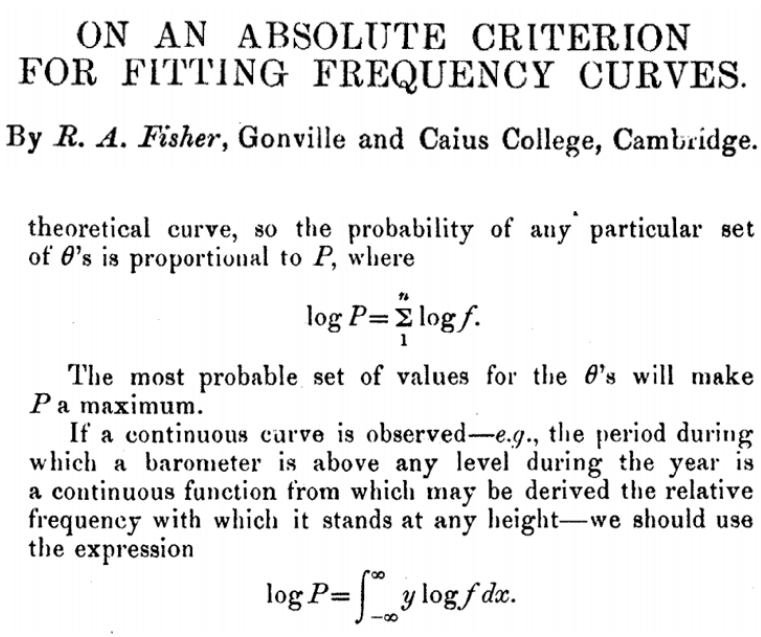
Mathematical Statistics

Consider observations $\{y_1, \cdots, y_n\}$ from iid random variables $Y_i \sim F_\theta$ (with "density" f_θ).
Likelihood is - Fisher (1912)

$$\mathcal{L}(\boldsymbol{\theta}\,;\boldsymbol{y}) \mapsto \prod_{i=1}^{n} f_\theta(y_i)$$

Maximum likelihood estimate is

$$\widehat{\boldsymbol{\theta}}^{\text{mle}} \in \arg\max_{\boldsymbol{\theta}\in\Theta}\{\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})\}$$



ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

theoretical curve, so the probability of any particular set of θ's is proportional to P, where

$$\log P = \sum_1^n \log f.$$

The most probable set of values for the θ's will make P a maximum.

If a continuous curve is observed—*e.g.*, the period during which a barometer is above any level during the year is a continuous function from which may be derived the relative frequency with which it stands at any height—we should use the expression

$$\log P = \int_{-\infty}^{\infty} y \log f \, dx.$$

Computational Issues

In multiple regression, it is shown that least square parameter estimates can be unsatisfactory if the prediction vectors are not orthogonal. Proposed is a procedure based on adding small positive quantities to the diagonals of the normal equations to obtain estimates with smaller mean square error. [The *Science Citation Index*® (*SCI*®) and the *Social Sciences Citation Index*® (*SSCI*®) indicate that this paper has been cited in over 310 publications since 1970.]

It would be great to report that we had profound discussions on the foundations of statistics, but such was not the case. Much of the time was spent trying to find ways to do regression computations economically and to come up with solutions that made engineering sense. We were charging $90/day for our time, but had to charge $450/hour for computer time on a Univac I that had 1,000 words of memory. With this machine, it took 75 processing minutes to invert a 40 × 40 matrix through a 4 × 4 partition of 10 × 10 submatrices, using magnetic tapes for temporary storage.

This Week's Citation Classic CC/NUMBER 35 AUGUST 30, 1982

Hoerl A E & Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**:55-67, 1970.
[University of Delaware, Newark, and E.I. du Pont de Nemours & Co., Wilmington, DE]

"In these discussions, we found that we had both encountered the same phenomenon, one that had caused some embarrassment with clients. We found that multiple linear regression coefficients computed using least squares didn't always make sense when put into the context of the process generating the data. The coefficients tended to be too large in absolute value, some would even have the wrong sign, and they could be unstable with very small changes in the data.

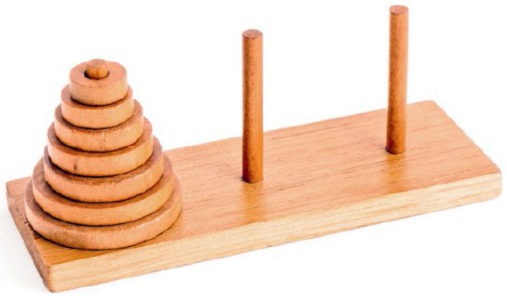
"Since the method proposed attacked one of the sacred cows of linear regression—least squares—there was considerable resistance. However, the solid theoretical basis and the practical usefulness of the method gradually overcame most objections.

Arthur Hoerl in 1982, back on Hoerl & Kennard (1970)

Algorithms

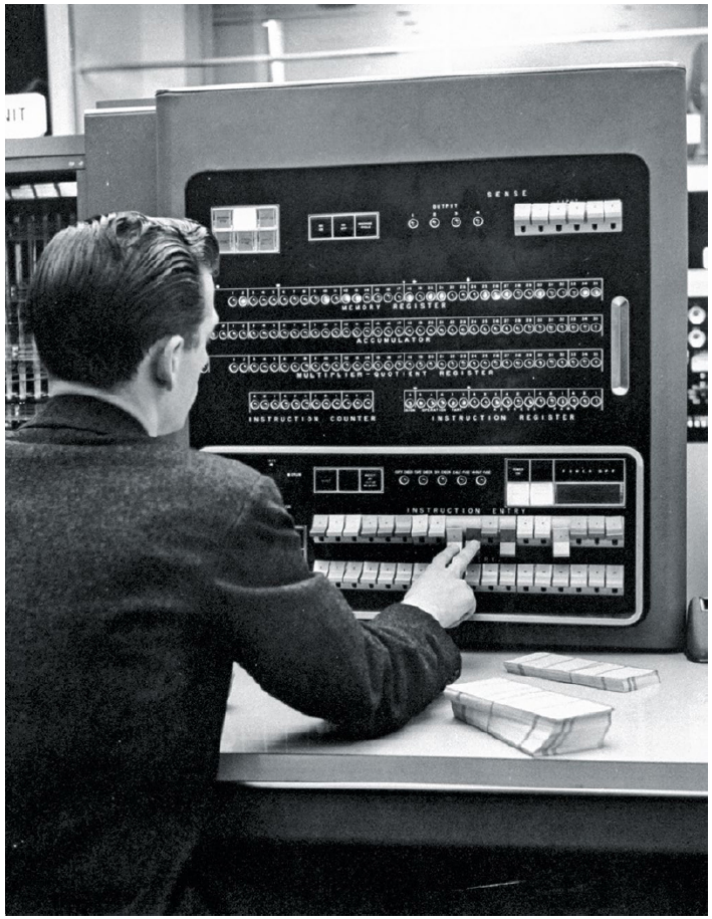
McCulloch & Pitts (1943) : neural network
Rosenblatt (1957) : perceptron
Fukushima (1980) : neocognitron
& convolutional neural network
Hopfield (1982) : recurrent neural network
LeCun (1998) : gradient-based learning

Michie (1963) tic-tac-toe (RL)
1997: Deep Blue (IBM) Chess
2011: Watson (IBM) Jeopardy
2013: Atari Games (RL)
2016: Google's AlphaGo (Deepmind)



Computers

the biggest difference between AI then and now is that the necessary computational capacity, raw volumes of data, and processing speed are readily available so the technology can really shine, Hammond (2015)



Machine Learning ?

RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

The main contribution of this paper is that it shows that it is possible to design *learning machines* that have all three of the following properties:

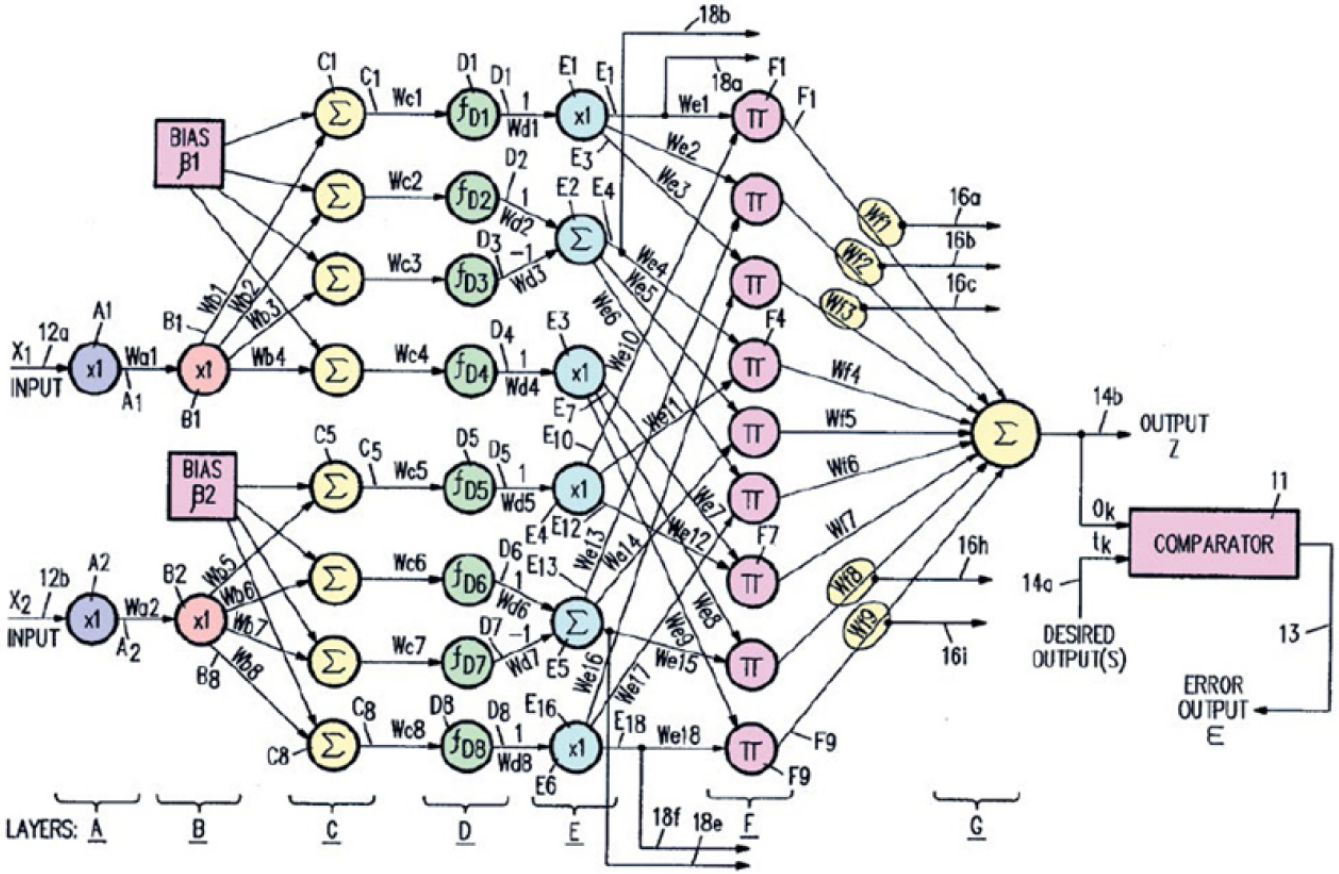
1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized.
2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge.
3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

More precisely we say that a class X of programs is *learnable* with respect to a given learning protocol if and only if there exists an algorithm A (the deduction procedure) invoking the protocol with the following properties:

1. The algorithm runs in time polynomial in an adjustable parameter h, in the various parameters that quantify the size of the program to be learned, and in the number of variables t.
2. For all programs $f \in X$ and all distributions D over vectors v on which f outputs 1, the algorithm will deduce with probability at least $(1 - h^{-1})$ a program $g \in X$ that never outputs one when it should not but outputs one almost always when it should. In particular, (i) for all vectors v, $g(v) = 1$ implies $f(v) = 1$, and (ii) the sum of $D(v)$ over all v such that $f(v) = 1$, but $g(v) \neq 1$ is at most h^{-1}.

via Vailiant (1984)

Neural Nets & Deep Learning



via Khan (1993), patent # US5579439A