So much data…

…who needs probability?

Have we been here before?

BFF5

May 7, 2018

Glenn Shafer

RANDOM HOUSE UNABRIDGED DICTIONARY
1987

Statistics, *n*. The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.

WIKTIONARY
2018

Statistics, *n*. A mathematical science concerned with data collection, presentation, analysis, and interpretation.

Data Science, *n*. An interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data.

Random House / David Moore

Statistics, *n*. The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.

Who invented this definition of statistics as a field?

1. Science
2. Data of any kind on any topic
3. Uses probability theory

Statistik was invented in Germany

- First used in Latin.

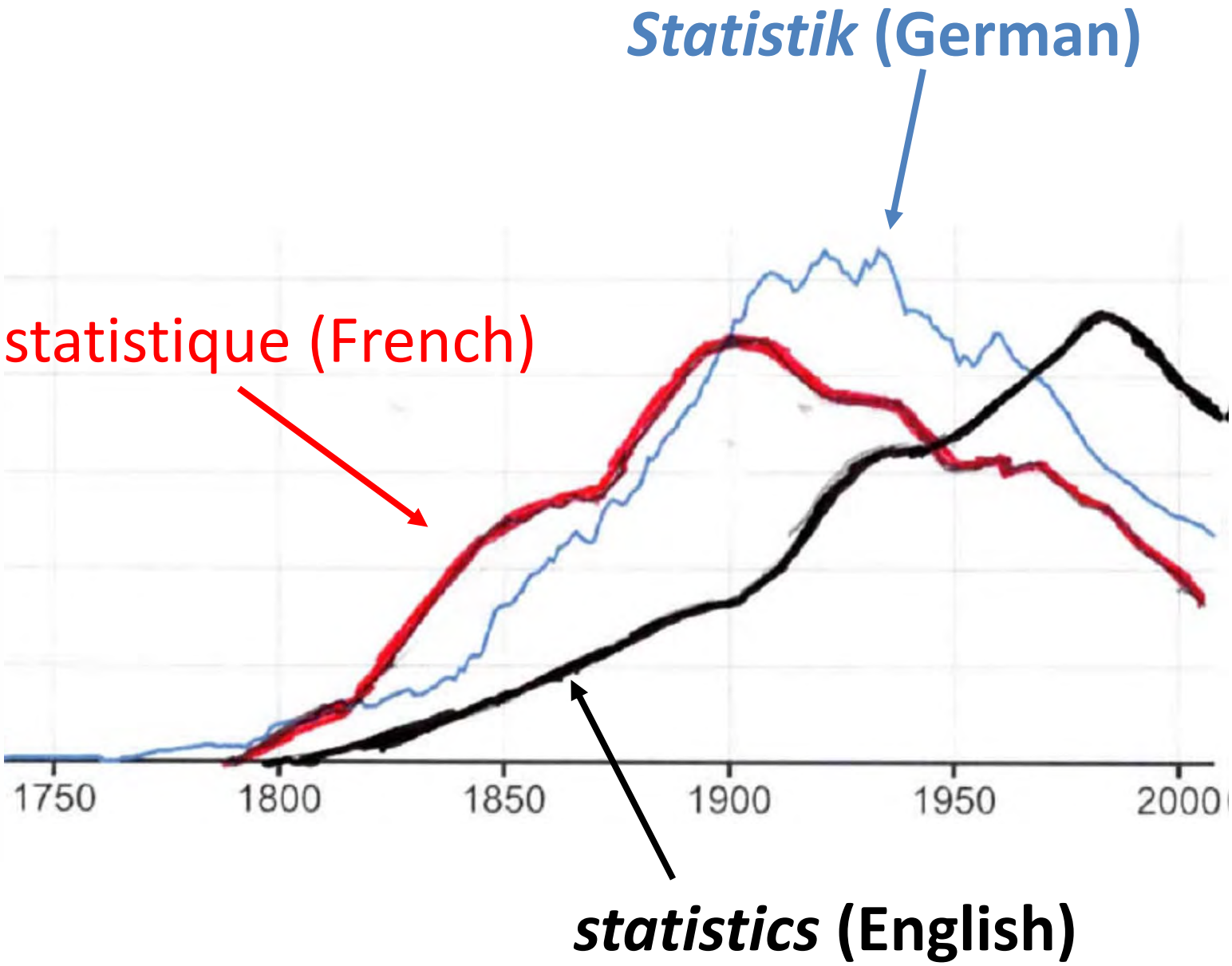
Statistik popularized by Göttingen professor Gottfried Achenwall in 1749.

- In all European languages by 1800.
- Only data about the condition of the state.
- No probability theory.

Frequency of *Statistik* in German books, per Google's Ngram Viewer

Peaked in early 20th century.

About 1 in every 40,000 words.


Statistik (German)
statistique (French)
1750
1800
1850
1900
1950
2000
statistics (English)

Theodore Porter, 1986
The Rise of Statistical Thinking, 1820-1900

Ian Hacking, 1990
The Taming of Chance

The rise of statistics in the 1800s resulted from a flood of data in demography, medicine, meteorology, manufacturing, economics, etc.

- Probability not the dominant influence.
- André-Michel Guerry and Auguste Comte sneered at probability.

In 1800, *statistics* and *probability* were very separate.

- *Statistics* = the science of the condition of the state.
- *Probability* = the reason we have to believe.

The separation endured through the 1800s even as the scope of statistics broadened.

- No *statistique* in Laplace (1812) or Poisson (1837).
- No *statistics* in Thomas Galloway's *Treatise on Probability* (1839).
- *Statistics* appears just two times in De Morgan's essay on probabilities and insurance (1838).

As Steve Stigler explained in his 1986 book, *The History of Statistics: The Measurement of Uncertainty before 1900*:

> …Statistics, as we now understand the term, has come to be recognized as a separate field only in the twentieth century.

But who planted the seeds?

What 19th century mathematician invented our modern idea of statistics?

Who invented our identity as statisticians?

1) The science dealing with collection, classification, analysis, and interpretation

2) of data on any topic,

3) using probability to impose order.

Who first gave *statistics* this three-fold meaning?

Antoine Augustin Cournot
1801-1877

1828: In *Le Lycée*

1834: Appendix to his French translation of John Herschel's *Treatise on Astronomy*

French Revolution and aftermath

1770-1815

Enlightenment and Revolution.

Probability blossoms.

Condorcet, Laplace:

Probability = reason we have to believe.

Births, deaths, marriages recorded by mayors.

1815

Napoleon defeated.

Bourbon kings restored.

Property restored to aristocrats and church.

Mathematics and probability suspect.

But administrative state continues to expand.

Jean Baptiste Joseph Fourier
1768-1830

Revolutionary.

Administrator under Napoleon.

Lost pension when Napoleon fell.

Royalist friend gave him job with the Paris statistical bureau.

Introduced probability into statistics.

Antoine Augustin Cournot
1801-1877

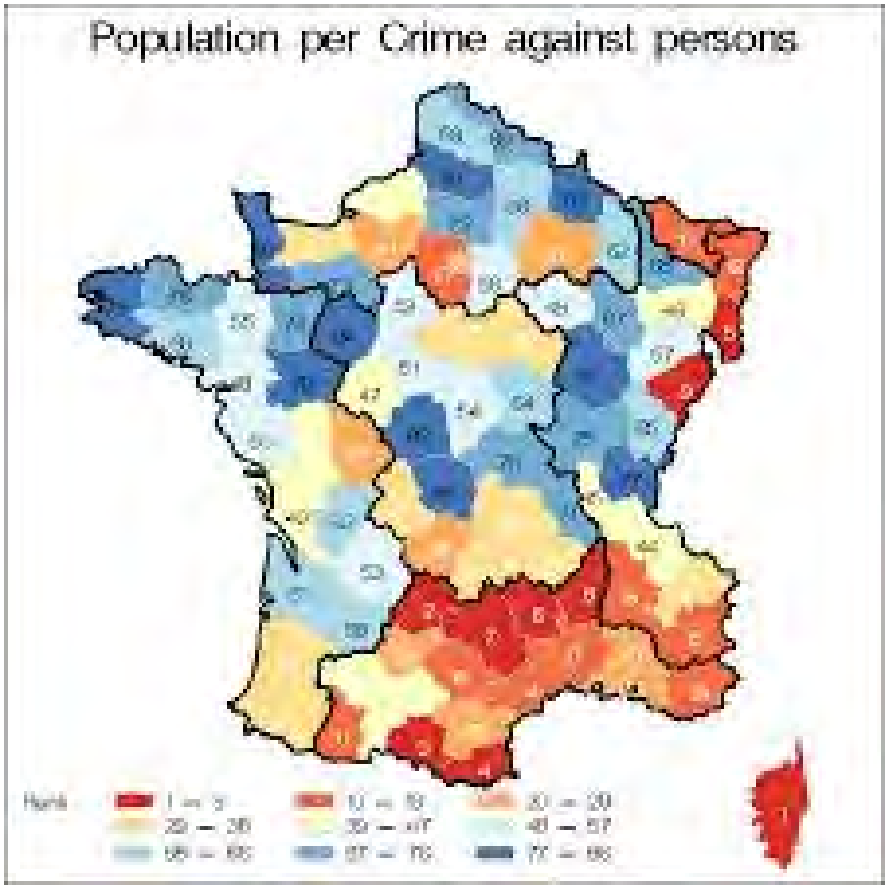
Roman Catholic, not too pious.

Defended mathematics.

Purged probability of dangerous elements (especially errors of Bayes, Condorcet, and Laplace).

Broadened concept of statistics.

André-Michel Guerry
1802-1866



Essai sur la Statistique Morale de la France, 1833

Emphatically rejected relevance of probability to statistics.

Antoine-Augustin Cournot
1801-1877

Exposition de la théorie des chances et des probabilités, 1843

Statistics is usually taken to mean (as the etymology indicates), the collection of facts arising from the clustering of people in civil societies. But for us the word will take on a more extended meaning.

By statistics, we mean the science that collects and systematizes numerous facts of every kind, so as to obtain numerical ratios that are reasonably independent of random anomalies and indicate the existence of regular causes whose influence is combined with that of random causes.

Antoine-Augustin Cournot
1801-1877

Cournot's contributions included…

- critique and rejection of Bayes' rule,
- first discussion of multiple testing,
- clear explanation of why rates (such as rates of crime) could be even more stable than the Bernoulli model suggests.
- Cournot's principle: The only way to apply probability theory is to predict an event by giving it a high probability.

In 1835, the role in probability in medicine was debated on the floor of the Academy of Sciences.

In 1840, Jules Gavarret published *Principes généraux de statistique médicale*.

Claude-Louis Navier
1785-1836

Defended probability in medicine.

By the 20th century, ***Statistik, statistique, statistics*** meant data on any topic:

- medical statistics
- social statistics
- astronomical statistics
- even statistical mechanics
  (where data is not collected!)

But role of probability was still contested.

George Udny Yule
1871-1951

Assistant to Karl Pearson

Hero of Stigler's *History of Statistics*

By **statistics** we mean quantitative data affected to a marked extent by a multiplicity of causes.

By **statistical methods** we mean methods especially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

By **theory of statistics** we mean the exposition of statistical methods.

The Germans invented two new terms for (statistics + probability).

- *Mathematische Statistik* (*mathematical statistics*)
  Theodor Wittstein 1867

- *Stochastik* (*stochastics*)
  Laudislaus von Bortkiewicz 1917

Origin of "mathematical statistics"

Theodor Wittstein (1816-1894)

Hanover

Mathematische Statistik und deren Anwendung auf National-Oekonomie und Versicherungs-Wissenschaft (1867)

Gustav Zeuner (1828-1907)

In 1869, in exile in Zürich:
Abhandlung aus der Mathematischen Statistik.

The study of data oriented to probability theory and thus to the "the law of large numbers" may be called stochastics.

Stochastics is not probability theory by itself but rather probability theory in application, be it to empirical data in general or to empirical data of a certain kind.

Laudislaus von Bortkiewicz
1868-1931
Polish by birth,
Russian by education,
German by choice.

Few university statistics departments in German universities.

Stochastik is taught in secondary schools.

Warren Persons's 1923 ASA presidential address:

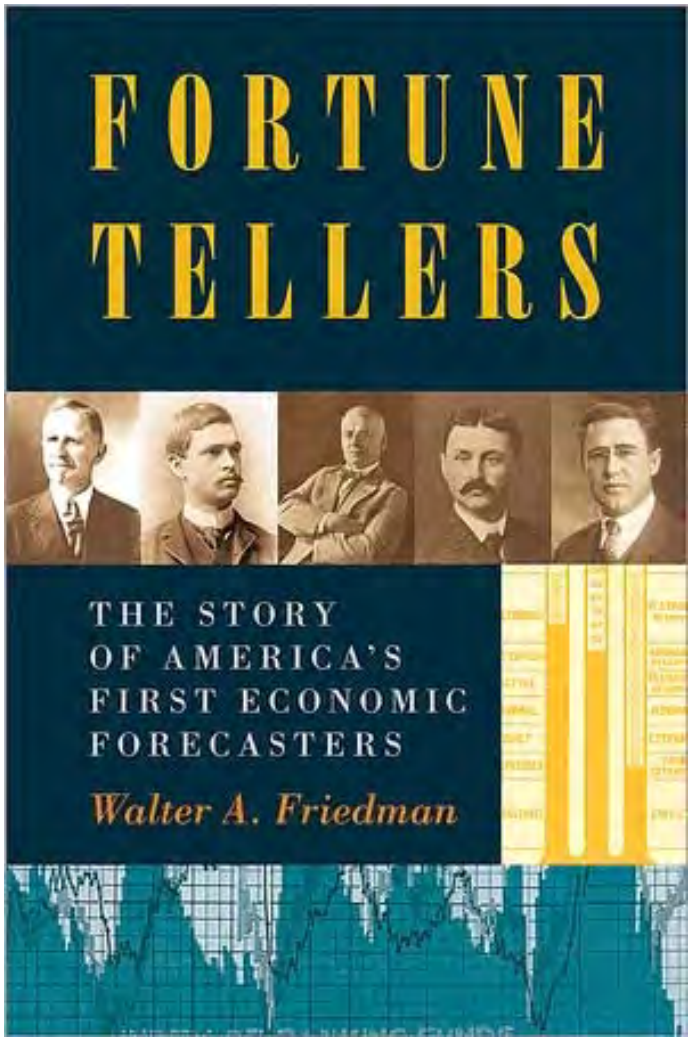
- The view that the mathematical theory of probability provides a method of statistical induction … is wholly untenable.
- The probabilities of the economic statistician are not the numerical probabilities which arise from the application of the theorems of Bernoulli and Bayes; they are, rather, non-numerical statements of the conclusions of inductive arguments.

Warren Persons
1878-1937

Why are Bernoulli and Bayes inapplicable?

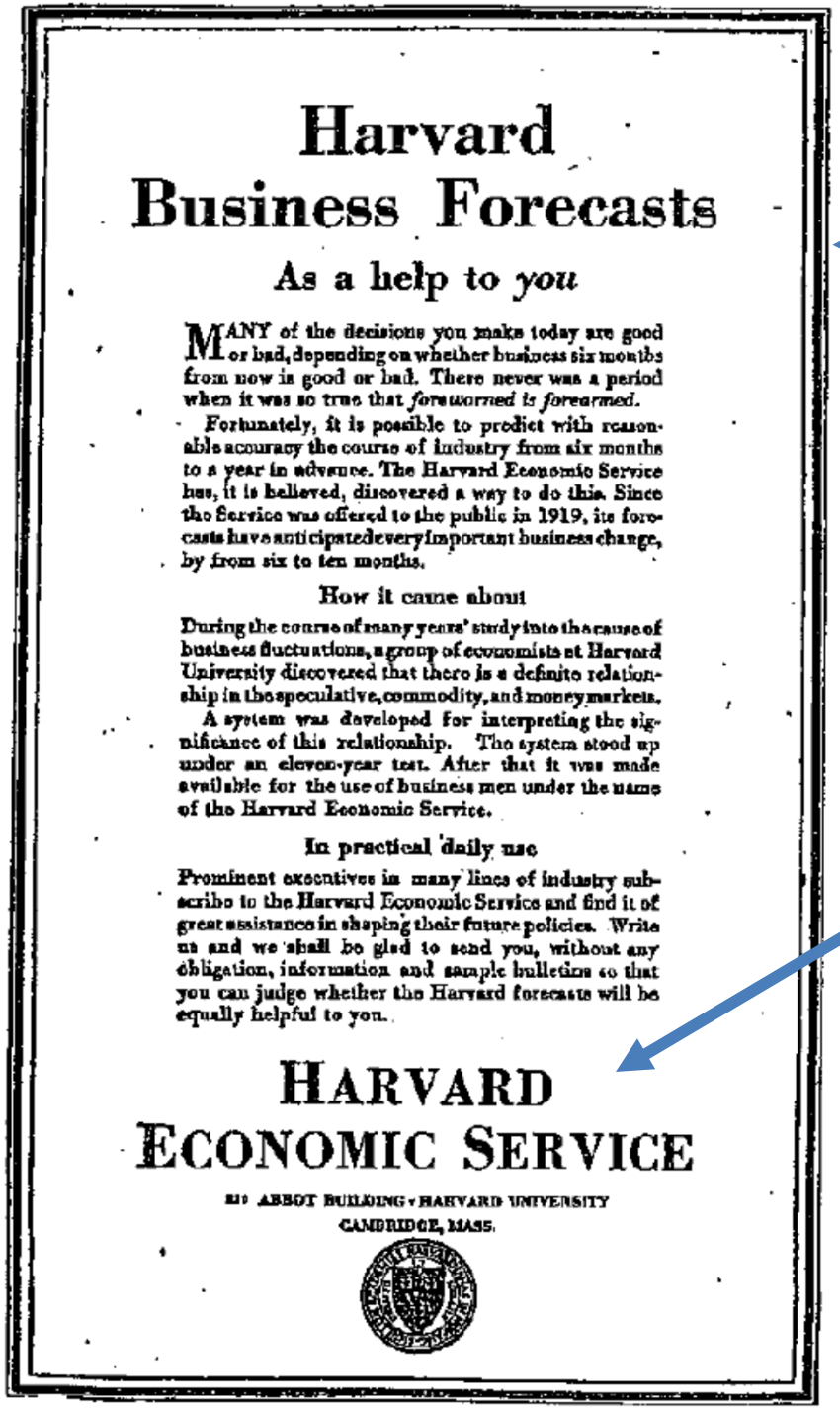
Persons's explanation:

Economic observations not iid!



Princeton University Press
2013

Friedman lectures at the Harvard Business School.



Harvard Business Forecasts

As a help to *you*

MANY of the decisions you make today are good or bad, depending on whether business six months from now is good or bad. There never was a period when it was so true that *forewarned is forearmed.*

Fortunately, it is possible to predict with reasonable accuracy the course of industry from six months to a year in advance. The Harvard Economic Service has, it is believed, discovered a way to do this. Since the service was offered to the public in 1919, its forecasts have anticipated every important business change, by from six to ten months.

How it came about

During the course of many years' study into the cause of business fluctuations, a group of economists at Harvard University discovered that there is a definite relationship in the speculative, commodity, and money markets.

A system was developed for interpreting the significance of this relationship. The system stood up under an eleven-year test. After that it was made available for the use of business men under the name of the Harvard Economic Service.

In practical daily use

Prominent executives in many lines of industry subscribe to the Harvard Economic Service and find it of great assistance in shaping their future policies. Write us and we shall be glad to send you, without any obligation, information and sample bulletins so that you can judge whether the Harvard forecasts will be equally helpful to you.

HARVARD ECONOMIC SERVICE

310 ABBOT BUILDING • HARVARD UNIVERSITY
CAMBRIDGE, MASS.

Advertisement, *New York Times*, October 1923.

Profit-making arm of the Economics Department at Harvard in 1920s.

Headed by Charles Bullock and Warren Persons.

From Harvard advertisement

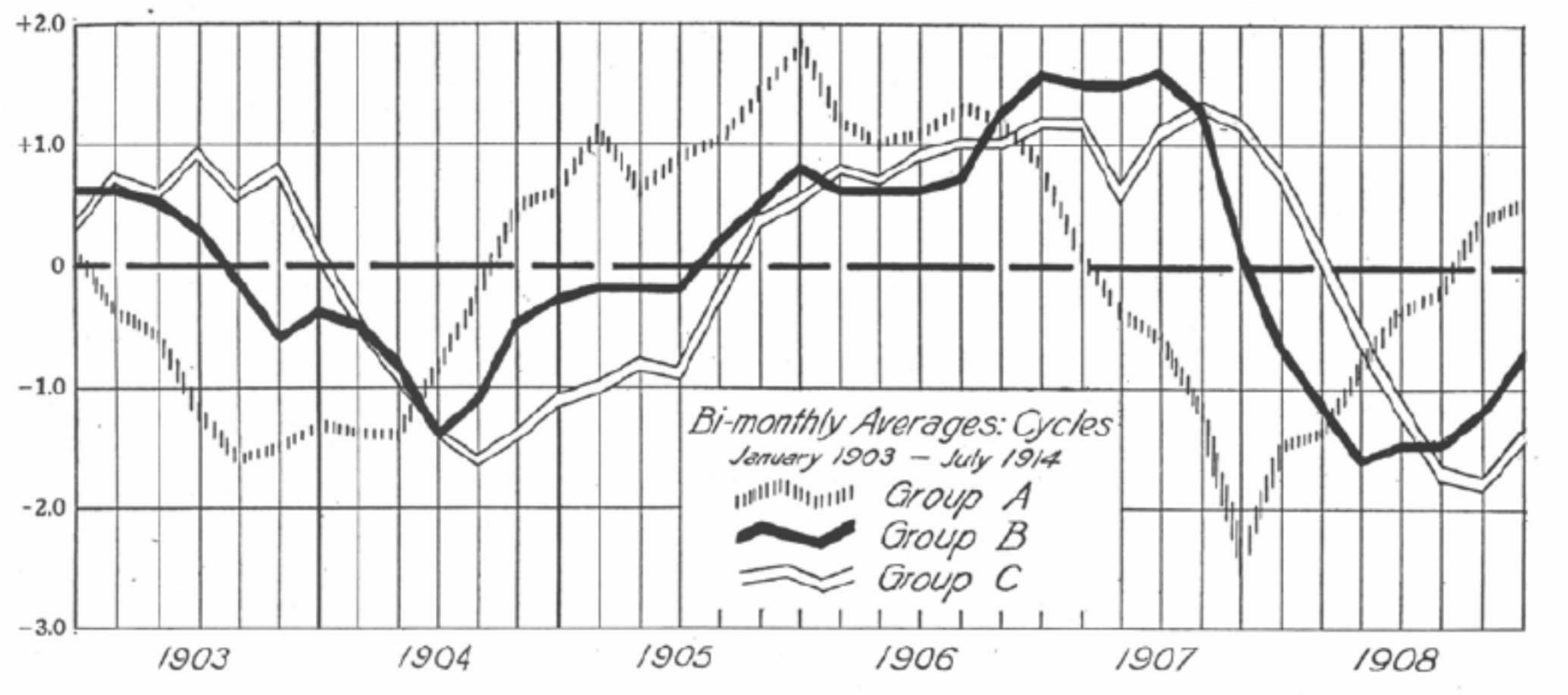
Many of the decisions you make today are good or bad, depending on whether business six months from now is good or bad. There never was a period when it so true that forewarned is forearmed.

Fortunately, it is possible to predict the course of industry from six months to a year in advance. The Harvard Economic Service, it is believed, discovered a way to do this. Since the Service was offered to the public in 1919, its forecasts have anticipated every important business change, by from six to ten months.

How it came about

During the course of many years' study into the causes of business fluctuations, a group of economists at Harvard university discovered that there is a definite relationship in the speculative, commodity and money markets.

A system was developed for interpreting the significance of this relationship. The system stood up under an eleven-year test. After that it was made available for the use of business men under the name of the Harvard Economic Service.



A = stocks

B = business activity

C = banking

Stock market predicts the business cycle!!

"The Probability Approach in Econometrics" **1944:**

The reluctance among economists to accept probability models as a basis for economic research has, it seems, been founded upon a very narrow concept of probability and random variables.

Probability schemes, it is held, apply only to … those series of observations where each observation may be considered as an independent drawing from one and the same 'population'…

Trygve Haavelmo
(1911-1999)

Nobel Prize 1989

Haavelmo 1944:

It is sufficient to assume that the *whole set* of, say *n*, observations may be considered as *one* observation of *n* variables…

The class of scientific statements that can be expressed in probability terms … contains all the `laws' that have, so far, been formulated. For such `laws' say no more and no less than this:

The probability is almost 1 that a certain event will occur.

How can statistics conquer data science?

Yes, compete in computation.

But emphasize assessment of uncertainty.

Embrace probability going beyond Fisher.

- stochastic processes
- support vector machines
- probabilistic machine learning
- evidence theory
- probability forecasting with physical models
- game-theoretic probability