150
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
THE 1890 CENSUS

Herman Hollerith (1860-1929)
EM 1879, PhD 1890

Hollerith's work on the census was a great example of seeing how a process as complex as the US Census could be abstracted, rationalized, and improved through a technological innovation. A century later Columbia Engineers are continuing to advance this tradition as technological breakthroughs make more and more of our world data-driven.

As one example, in recent work with Columbia University Medical School virologists we considered the problem of identifying the host organism of origin in the presence of a pandemic virus. While this is not the problem for the current Ebola epidemic in many episodes of viral outbreak the source of the virus, like the local pig or bird population, is unknown. Hollerith's work on the census was a great example of seeing how a process as complex as the US Census could be abstracted, rationalized, and improved through a technological innovation. A century later Columbia Engineers are continuing to advance this tradition as technological breakthroughs make more and more of our world data-driven. As one example, in recent work with Columbia University Medical School virologists we considered the problem of identifying the host organism of origin in the presence of a pandemic virus. While this is not the problem for the current Ebola epidemic in many episodes of viral outbreak the source of the virus, like the local pig or bird population, is unknown.

150
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
THE 1890 CENSUS

Herman Hollerith (1860-1929)
EM 1879, PhD 1890

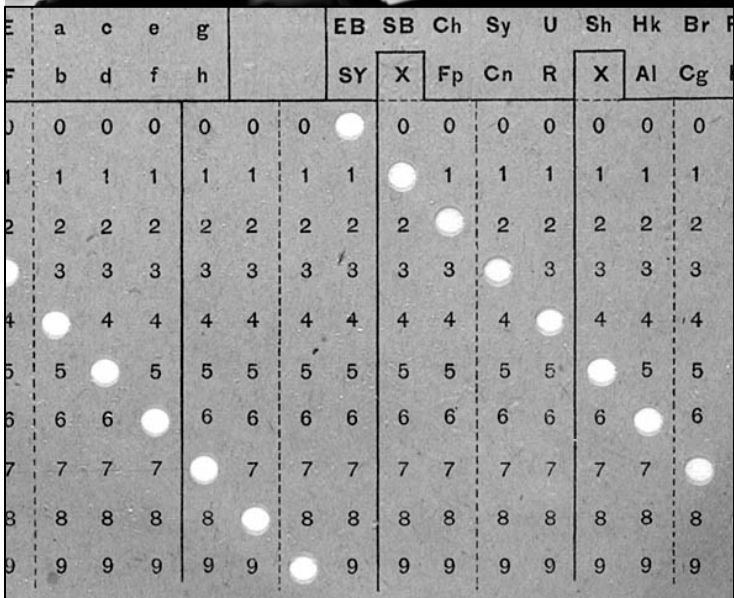
Invents tabulating machine that dramatically reduces time to process 1890 census data (from years to just a few months)

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
THE 1890 CENSUS

Herman Hollerith (1860-1929)
EM 1879, PhD 1890

First use of punch cards and electromagnetic counters



DATA SCIENCE
THE 1890 CENSUS

Herman Hollerith (1860-1929)
EM 1879, PhD 1890

Forms start-up called Tabulating Machine Company that grows into International Business Machines (IBM)

DATA SCIENCE
THE 1890 CENSUS

Our vision was to take the available training data, meaning viral genomes which had evolved in known hosts, and, instead of taking the usual route of trying to infer the entire hidden phylogentic tree of ancestor viral genomes, simply to try to predict which organism a target genome originated from.

DATA SCIENCE
THE 1890 CENSUS

Our approach was a machine learning approach called boosted decision trees, in which we learn a combinatorial predictive algorithm assembled from individually interpretable features corresponding to the presence of genomic or amino acid sequence elements.

The impact of this work is to suggest to the clinicians which parts of the genome or resulting proteins might be mechanistically important to the functioning of the virus, as well as to provide an engineering tool in the form of an algorithm that can be deployed on novel target genomes in the case of a novel outbreak.

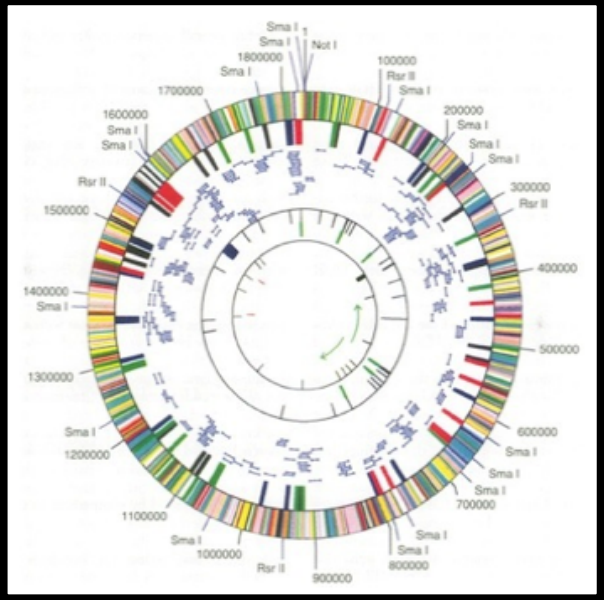
It's worth noting that, as with IBM, this approach of reframing domain questions as predictive computational tasks is spawning novel companies all over New York City as well as providing novel insights and products in established companies. Recently I've been trying to help the New York Times learn more about the genome of its readers using similar tools, where the impact is to suggest changes in their digital products and marketing decisions based on the usage patterns of its readers at web scale.

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
TRANSFORMING THE 21ST CENTURY

Vision:
Armed with abundant data about complex, real-world systems, build predictive and interpretable models

Consider making a short explicit reference to figure to the right, e.g. (example:" identifying the host organism in a pandemic virus."



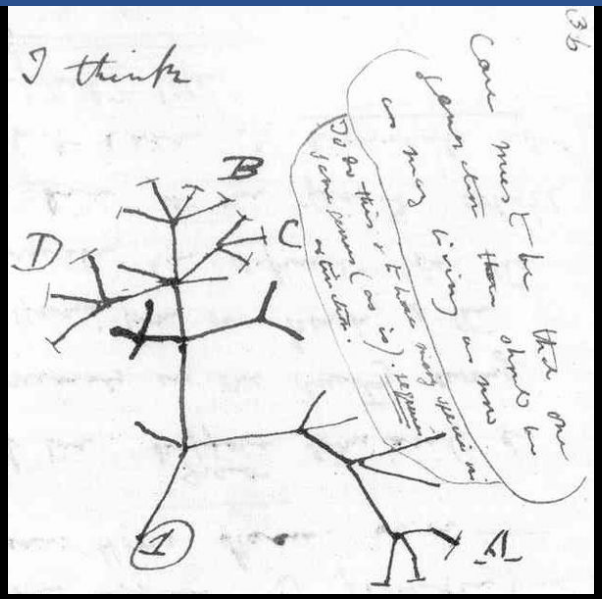
What does image show ? brief figure caption may be helpful! – Viral Genome

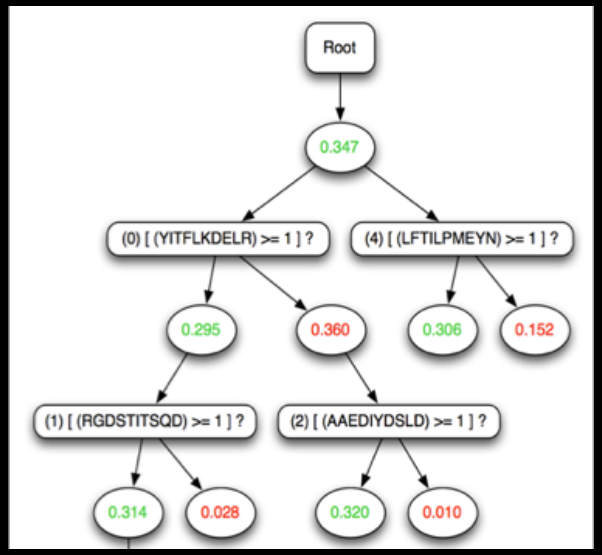
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
TRANSFORMING THE 21ST CENTURY

Vision:
Armed with abundant data about complex, real-world systems, build predictive and interpretable models

Approach:
Reframe domain questions as predictive machine learning tasks





COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

DATA SCIENCE
TRANSFORMING THE 21ST CENTURY

Vision:
Armed with abundant data about complex, real-world systems, build predictive and interpretable models

Approach:
Reframe domain questions as predictive machine learning tasks

Impact:
Learn models that both predict, but also suggest novel experiments in natural sciences and product or marketing changes when applied in technology companies and startups