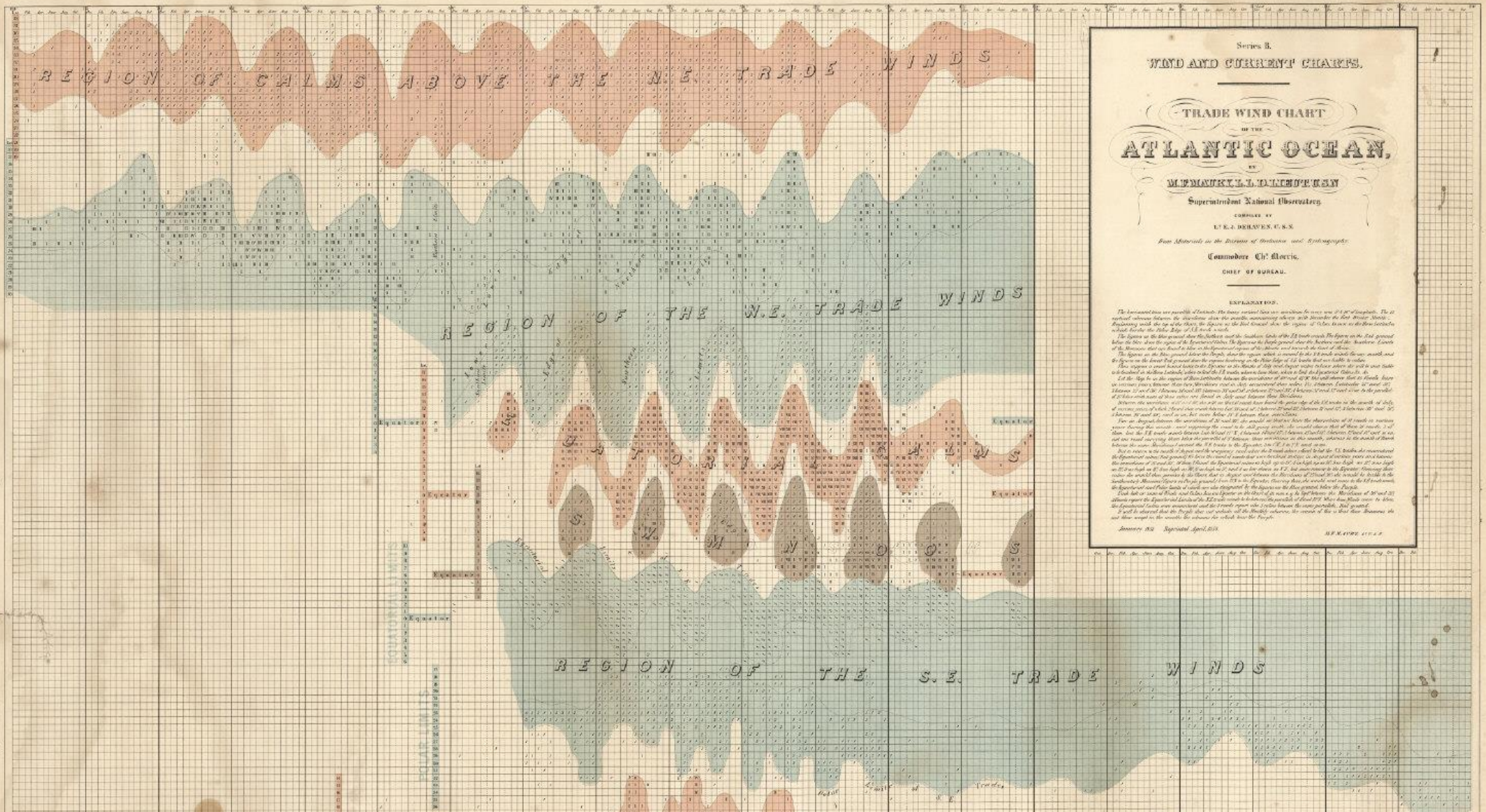
Big Data

ID 413: Information Graphics and Data Visualization

Spring 2016

Venkatesh Rajamanickam (@venkatrajam)
venkatra@iitb.ac.in
http://info-design-lab.github.io/ID413-DataViz/


REGION OF CALMS ABOVE THE N.E. TRADE WINDS
REGION OF THE N.E. TRADE WINDS
EQUATORIAL CALMS
S.W. MONSOONS
REGION OF THE S.E. TRADE WINDS
REGION OF CALMS BELOW THE S.E. TRADE WINDS
Equator
Equator
Equator
Equator
Equator
Equator

Series B.

WIND AND CURRENT CHARTS.

TRADE WIND CHART

OF THE

ATLANTIC OCEAN,

BY

M.F. MAURY, LL.D. LIEUT. U.S.N.

Superintendent National Observatory.

COMPILED BY

LT E.J. DEHAVEN, U.S.N.

From Materials in the Bureau of Ordnance and Hydrography

Commodore Chs. Morris.

CHIEF OF BUREAU.

EXPLANATION.

The horizontal lines are parallels of Latitude. The heavy vertical lines are meridians for every 5° [illegible] of Longitude. The 12 vertical columns between the meridians show the months, commencing always with December [illegible]. Beginning with the top of the Chart, the figures on the Red Ground show the region of Calms [illegible] which border the Polar Edge of N.E. trade winds.

The figures on the blue ground show the Northern and the Southern limits of the N.E. trade winds. [illegible] below the blue show the region of the Equatorial Calms. The figures on the Purple ground show the Northern [illegible] of the Monsoons that are found to blow in the Equatorial regions of the Atlantic and towards the Coast [illegible]

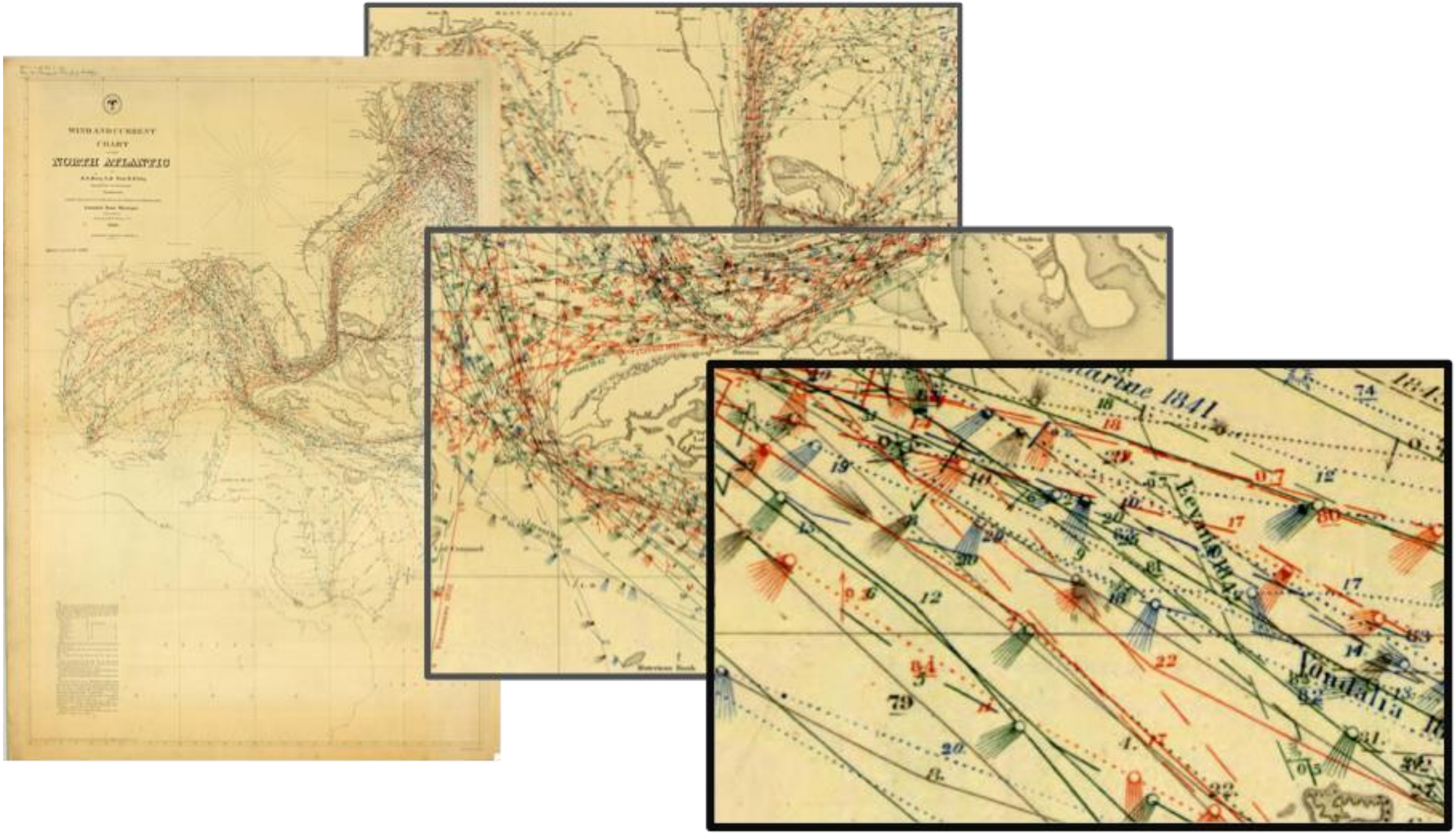
The figures on the Blue ground below the Purple, show the region which is covered by the S.E. trade [illegible] the figures on the lowest Red ground show the regions bordering on the Polar Edge of S.E. trades that [illegible]

[illegible]

January 1852. Reprinted April 1853.

M.F. MAURY LT U.S.N.

Matthew Fontaine Maury (1806 –1873) *Winds and Currents Chart*



What is Big Data?

“The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value.... Big data refers to things one can do at a large scale that cannot be done at a smaller one”

Viktor Mayer-Schönberger and Kenneth Cukier
Big Data, A Revolution That Will Transform How We Live, Work and Think, 2013

What is Big Data?

“The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value.... **Big data refers to things one can do at a large scale that cannot be done at a smaller one**”

Viktor Mayer-Schönberger and Kenneth Cukier
Big Data, A Revolution That Will Transform How We Live, Work and Think, 2013

Characteristics of Big Data?

In the book, Mayer-Schönberger and Cukier outline 3 perspectives (or “shifts”) which are inherent in big data:

1. The obsolescence of sampling
2. The acceptance of increased measurement error in return for more data
3. A “move away from the age-old search for causality”

Big Data stories

1. Building Inspections, NYC
2. Target, Netflix, Amazon
3. Google Flu Trends
4. Nathan Eagle’s work in Kenya & Rwanda
5. LinkedIn college rankings

What would a school data look like?

What would a school data look like?

- which questions generate most thought
- at exactly what point in a lesson learners switch off
- who will be bullied and when
- which concepts are actually most difficult
- the link between diet and progress
- how friendship networks impact on exam performance
- the emotional state of a group as they approach an exam
- which pupils thrive on what sorts of feedback
- the amount of time actually spent on homework
- the ideal design of a learning task
- which book a child should read next
- how room temperature affects learning
- who runs most in the playground

Big Data: Three challenges

Volume
– the size of the data

Velocity
– the latency of data processing relative to the growing demand for interactivity

Variety
– the diversity of sources, formats, quality, structures

Big Data Visualization

http://www.nytimes.com/interactive/science/space/keplers-tally-of-planets.html

http://demographics.coopercenter.org/DotMap/index.html

Where does big data come from?

- “data exhaust” from customers
- new and pervasive sensors
- the ability to “keep everything”

Building Inspections, NYC

For example, members of his group have created a visualization for use by oncologists that connects gene sequence and activation data with cancer types and stages, treatments, and clinical outcomes. That allows the data to be viewed in a way that shows which particular gene expression pattern is associated with high mortality regardless of cancer type, for example, giving an important, actionable insight for how to devise new treatments.

Building Inspections, NYC

New York City gets roughly 25,000 illegal-conversion complaints a year, but it has only 200 inspectors to handle them. There seemed to be no good way to distinguish cases that were simply nuisances from ones that were poised to burst into flames.

They started with a list of every property lot in the city—all 900,000 of them. Next they poured in datasets from 19 different agencies indicating, for example, if the building owner was delinquent in paying property taxes, if there had been foreclosure proceedings, and if anomalies in utilities usage or missed payments had led to any service cuts.

They also fed in information about the type of building and when it was built, plus ambulance visits, crime rates, rodent complaints, and more. Then they compared all this information against five years of fire data ranked by severity and looked for correlations in order to generate a system that could predict which complaints should be investigated most urgently.

Google Flu Trends

Five years ago, a team of researchers from Google announced a remarkable achievement in one of the world’s top scientific journals, Nature. Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What’s more, they could do it more quickly than the Centers for Disease Control and Prevention (CDC). Google’s tracking had only a day’s delay, compared with the week or more it took for the CDC to assemble a picture based on reports from doctors’ surgeries. Google was faster because it was tracking the outbreak by finding a correlation between what people searched for online and whether they had flu symptoms.

Not only was “Google Flu Trends” quick, accurate and cheap, it was theory-free. Google’s engineers didn’t bother to develop a hypothesis about what search terms – “flu symptoms” or “pharmacies near me” – might be correlated with the spread of the disease itself. The Google team just took their top 50 million search terms and let the algorithms do the work.