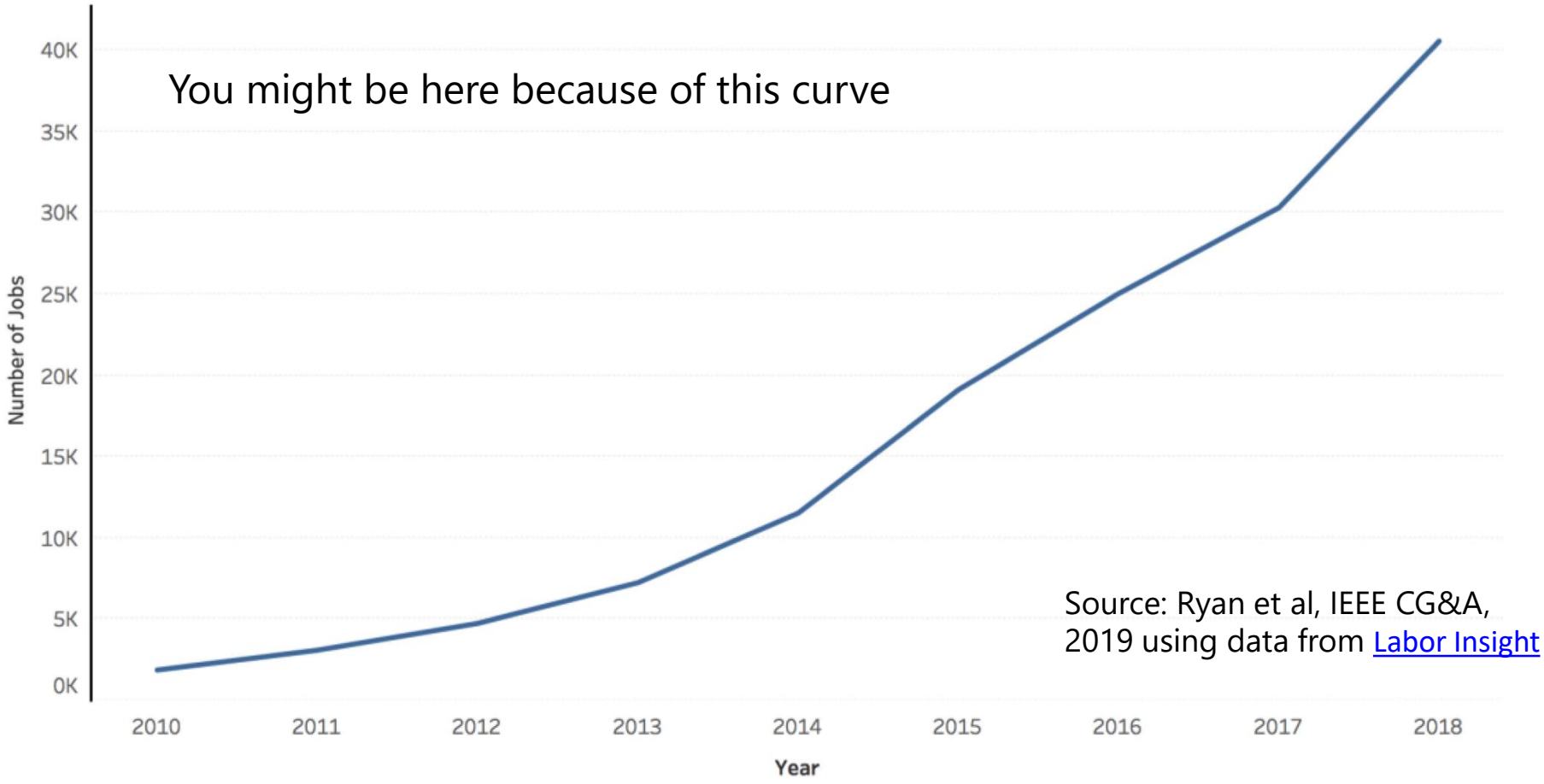
CSE/ISE 332
Introduction to Visualization

Introduction

Klaus Mueller

Computer Science Department
Stony Brook University

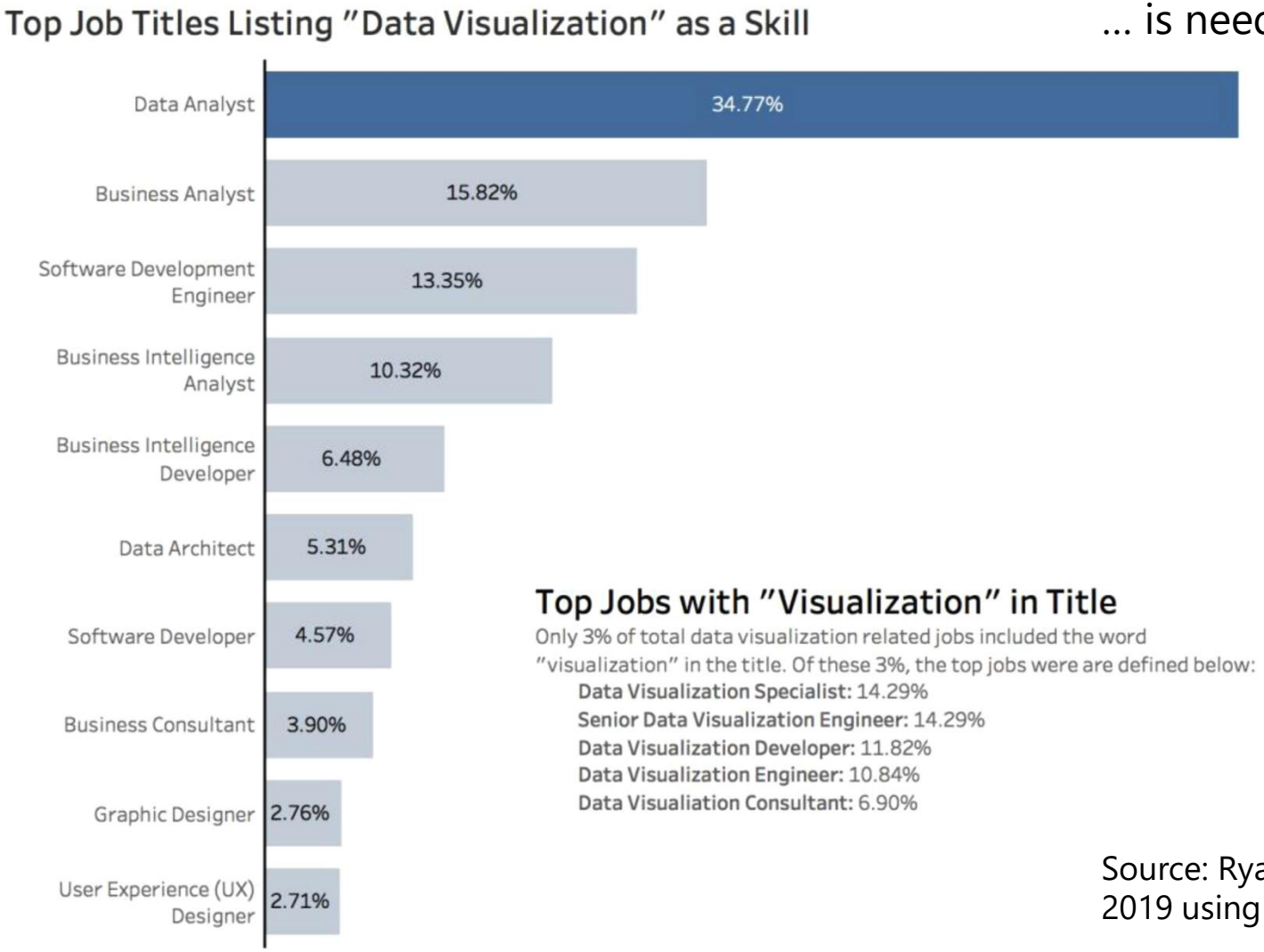
Why Are You Here?



The growth of jobs mentioning “data visualization” as a skill from 2010 through 2017 has steadily increased from only 1,888 jobs in 2010 to 30,327 jobs in 2017 (16×growth)

"Visualization" Skill…

… is needed everywhere



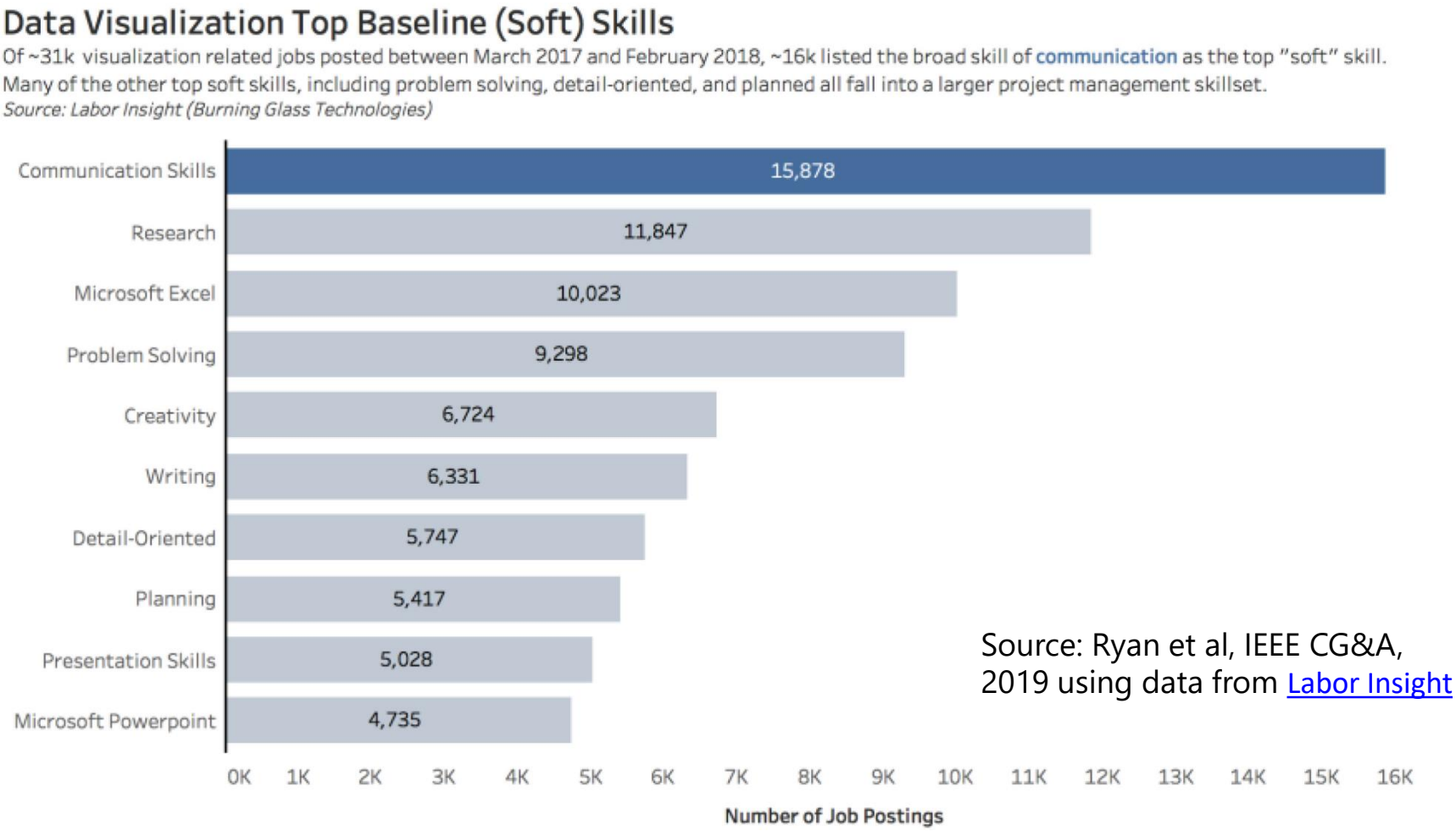
Top Jobs with "Visualization" in Title

Only 3% of total data visualization related jobs included the word "visualization" in the title. Of these 3%, the top jobs were are defined below:

- **Data Visualization Specialist:** 14.29%
- **Senior Data Visualization Engineer:** 14.29%
- **Data Visualization Developer:** 11.82%
- **Data Visualization Engineer:** 10.84%
- **Data Visualiation Consultant:** 6.90%

Source: Ryan et al, IEEE CG&A, 2019 using data from Labor Insight

What Other Skills?



Source: Ryan et al, IEEE CG&A, 2019 using data from Labor Insight

Baseline, or “soft” skills listed for these 30k “Data Visualization” jobs.

Skills, Reading Between The Lines

Communication, when mentioned in conjunction with data visualization really means:

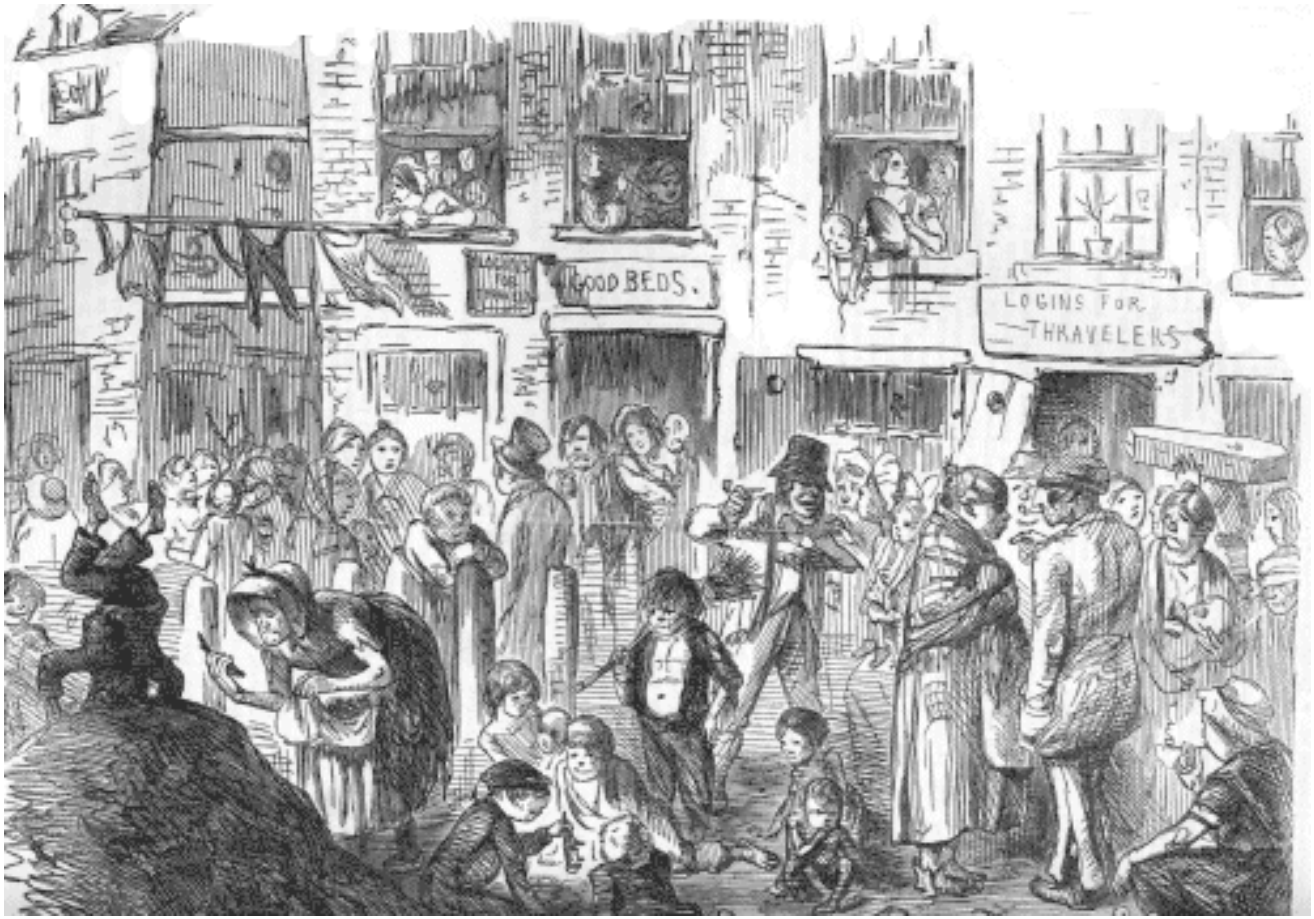
- communication of information derived from data
- *visual* story telling with data
- half of the data analytics projects fail due to poor communication*

*L. Kart, N. Neudecker, F. Buytendijk, Gartner Report GG0255160, 2013

Visualization is not new

Rich History

Let’s go back some 170 years to 1854, London, England



Newsflash, 1854

The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which is taking place in Broad Street, Golden Square, and adjoining streets.

Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there are upwards of five hundred fatal attacks of cholera in ten days.

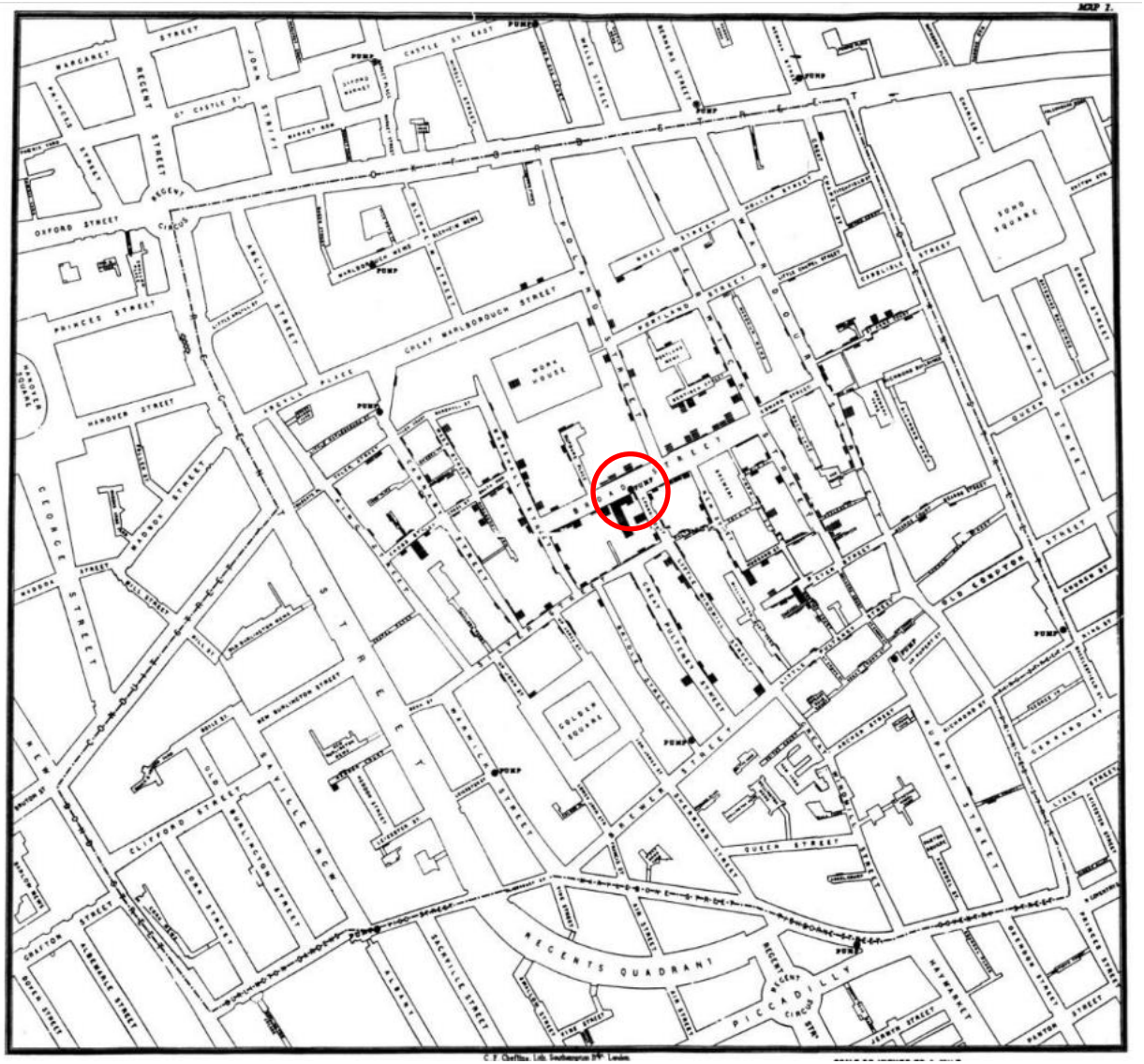
The mortality in this limited area probably equals any that was ever caused in this country, even by the plague; and it is much more sudden, as the greater number of cases terminated in a few hours.

What Can We Do?

What Is The Cause?

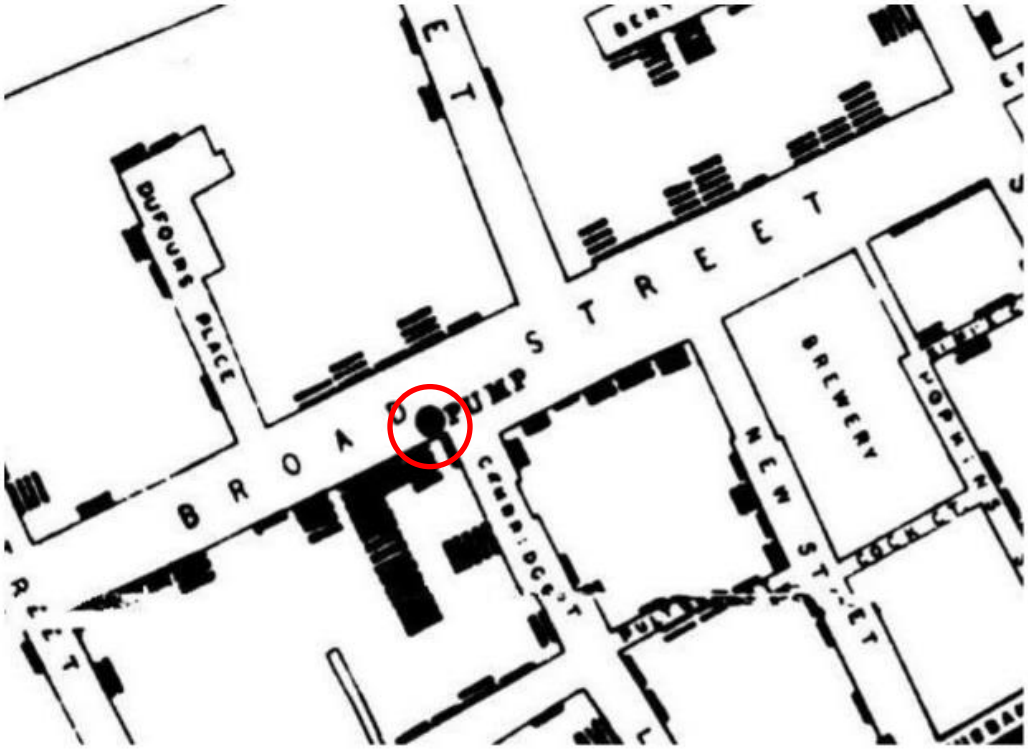
How Can We Eliminate It?

Time For "Imagination"



John Snow

Time For "Imagination"



John Snow

Proved The Hypothesis

Hypothesis: cholera spreads through water

- and not via some other fantastic causes
- one said it rose out of the burying grounds of plague victims from two centuries earlier
- the bacteria was discovered later, in 1886

A real-life experiment (often the case with observational data)

- established the mode of cholera transmission
- and consequently the method of prevention: keep drinking water, food, and hands clear of infected sewage

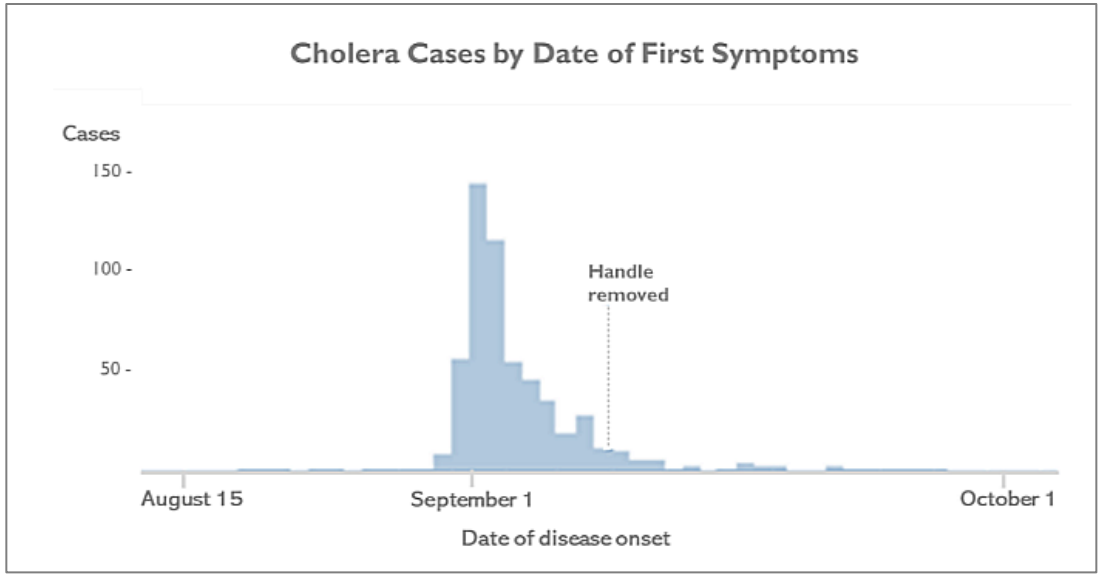
Visualization provided

- inspiration
- convincing arguments to justify actions (removing the pump handle)
- led to Dr. John Snow's historic immortality
- a bar near the old Broad Street pump bears his name (safe drinking)

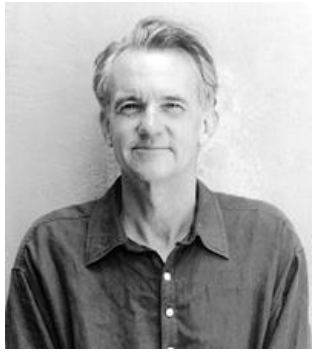
Graphed Over Time

Turns out that the handle was removed at the end of the outbreak

- graphing deaths over time revealed this
- also done by Dr. Snow but far less publicized
- but likely prevented a new outbreak



Later: Better Visual Design



Edward Tufte redrew the map many years later:

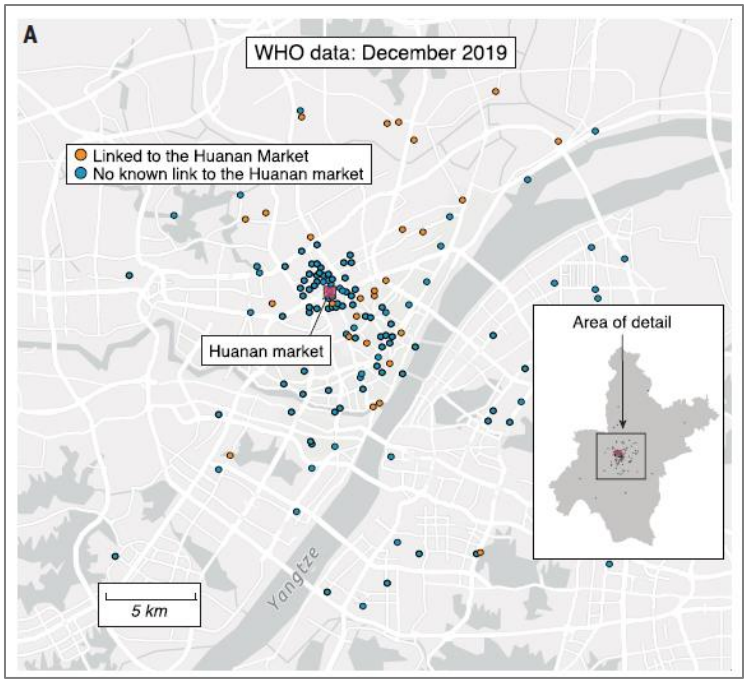
- only kept the most critical street and building details
- switched out Dr. Snow's dashes for dots
- focused the visual emphasis on Cholera victims and well locations, and not the features of the ground
- better data-to-ink ratio

John Snow, 165 years later....

Modern Version of Spatial Outbreak Analysis

1854: Where did the cholera outbreak originate?

2019: Where did the COVID-19 outbreak originate?



Worobey et al. "The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic," Science, August 2022

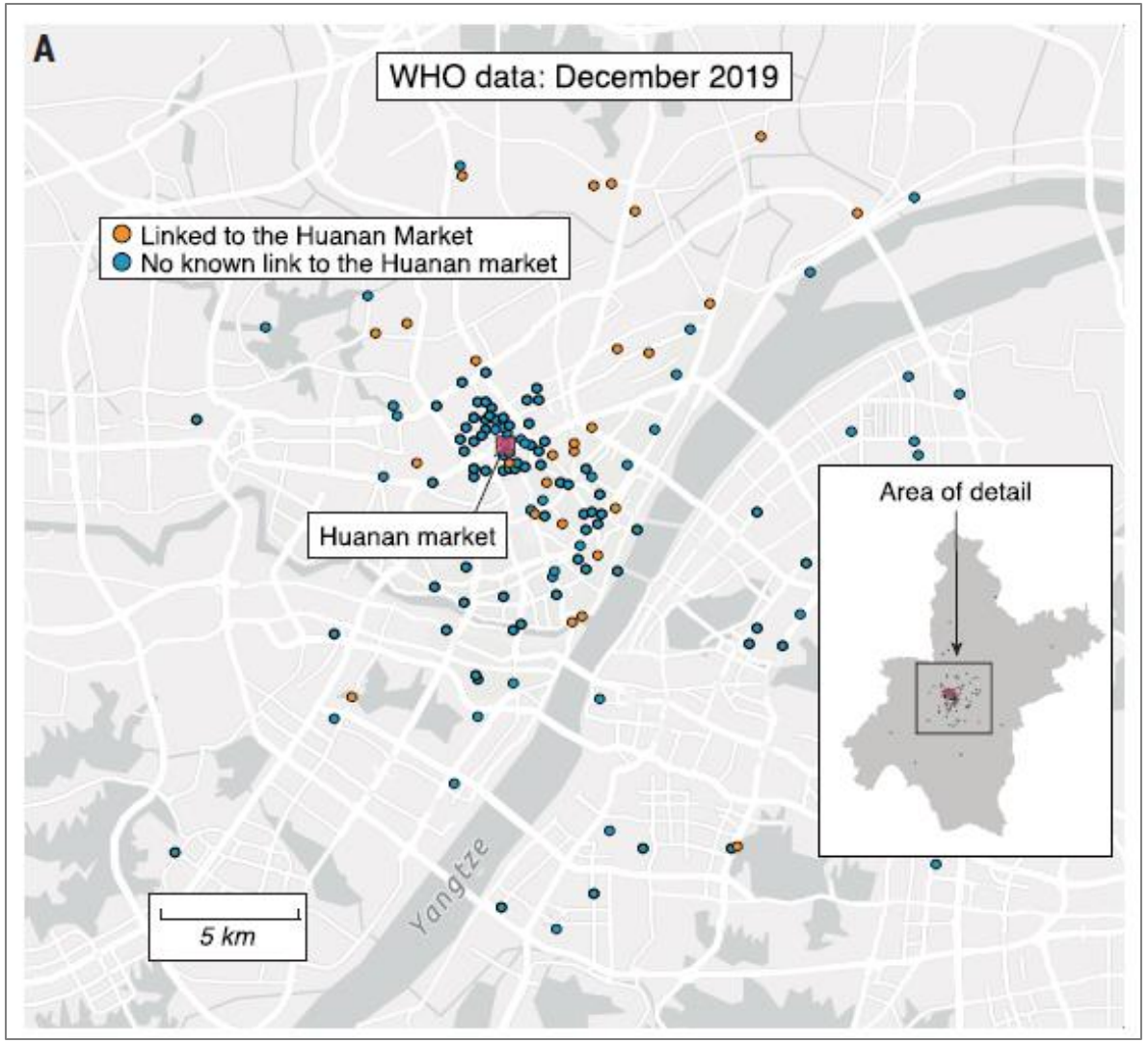
The First Map

155 onset cases:

- Dots: residences of December 2019 cases
- Orange: known market link (35)
- Blue: no known market link (120)
- Red square: Huanan market

What do you conclude?

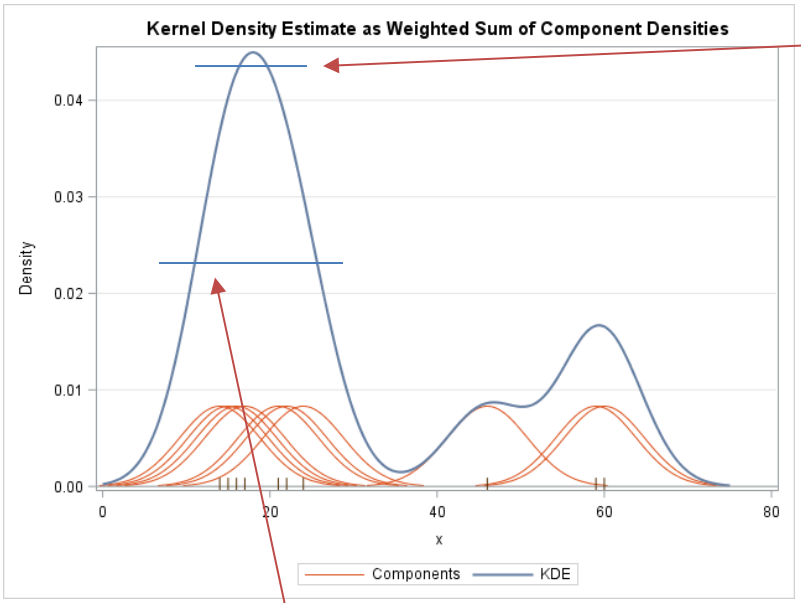
But… is seeing a cluster enough to conclude that the market was the epicenter?



From Points to Density

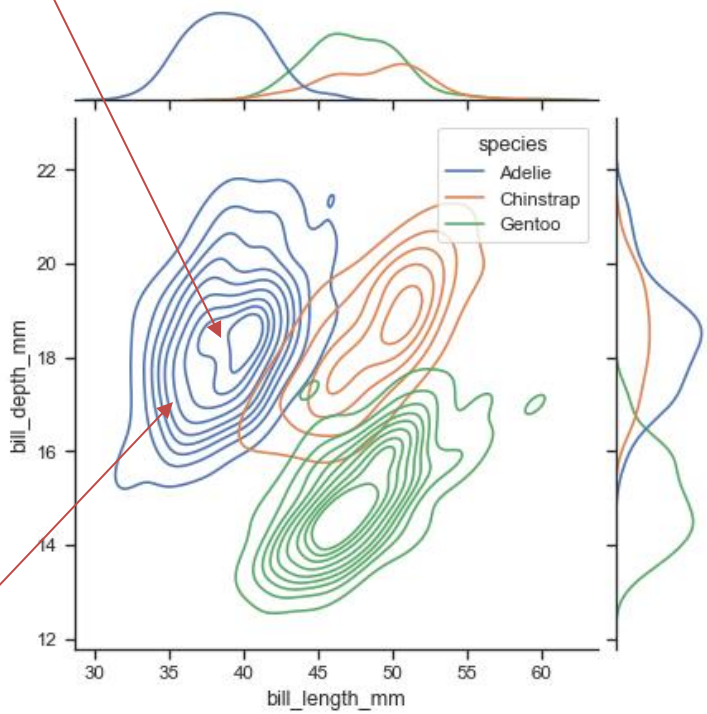
Need to use **Kernel Density Estimation (KDE)**

- Turns discrete case locations → estimated spatial density

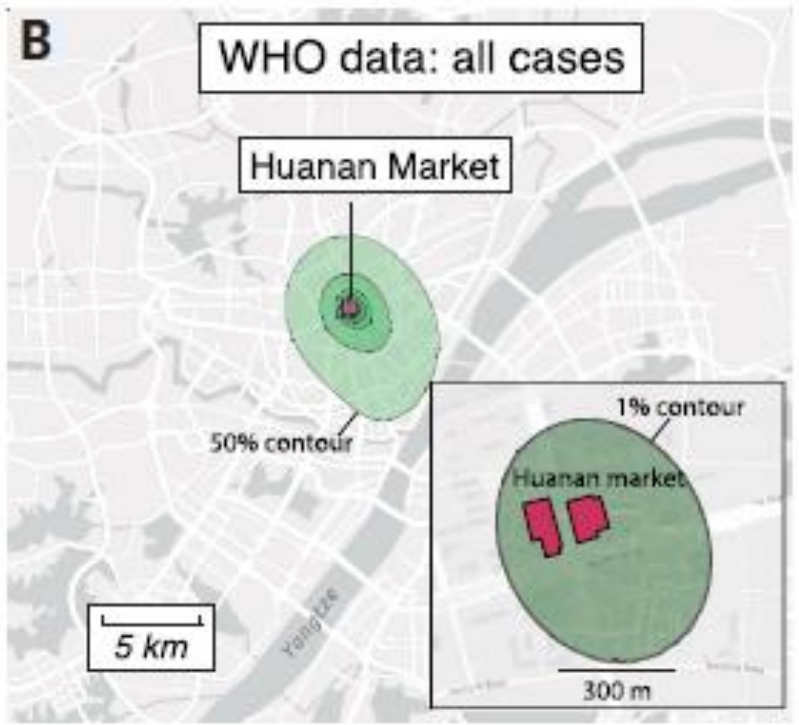


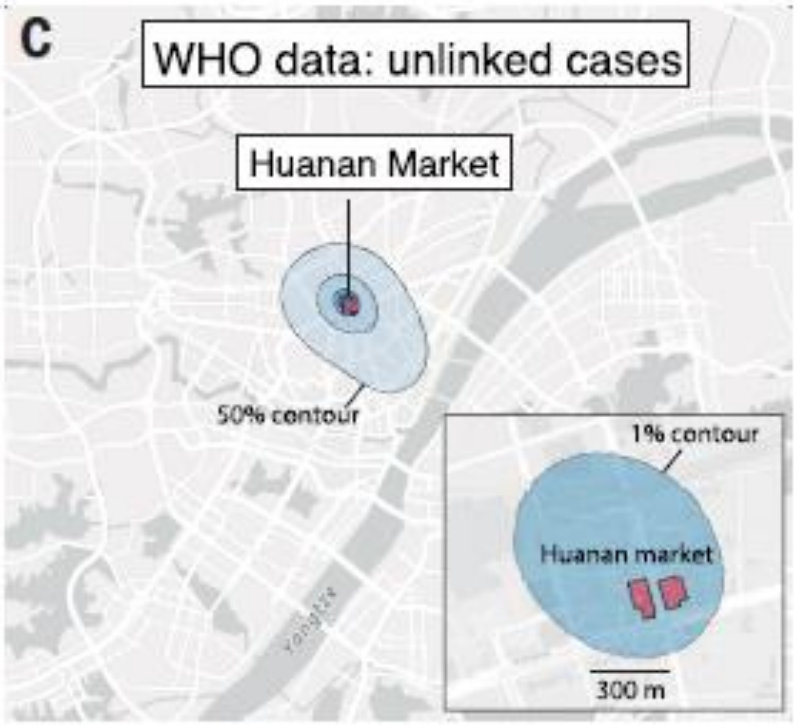
1% contour: the very peak of the density surface

50% contour: contains the densest region accounting for 50% of the estimated case-location probability



Applied to the 155 Onset Cases





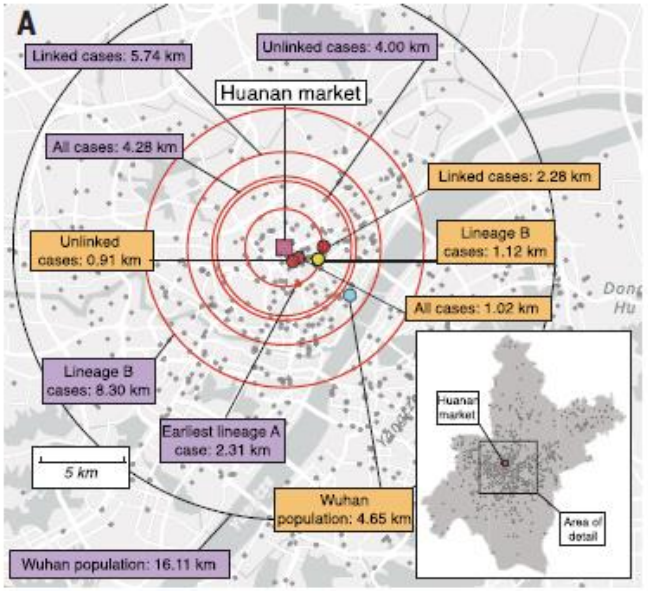
C is consistent with the outbreak radiating outward from the Huanan market into the surrounding community: even the 120 cases with no known market connection remain strongly centered on the market

Suggests the spread mechanism: person lives 500 m away → encounters an infected person locally → infected → no obvious connection to Huanan

Skeptic's Argument: "But Wuhan isn't Uniformly Populated"

Introducing the **null model**:

- Observed pattern vs. where would cases occur if they simply followed Wuhan's population distribution?



Median distance to Huanan

- "Are the cases unusually NEAR the market?"
- Wuhan population null: **16.11 km**
- all December cases: **4.28 km**
- market-linked: **5.74 km**
- unlinked: **4.00 km**
- All 3 case groups were significantly closer than expected

Mean distance of case center points to Huanan

- Asks: "are the cases unusually CENTERED ON the market?"
- Similar conclusions than for the median distance

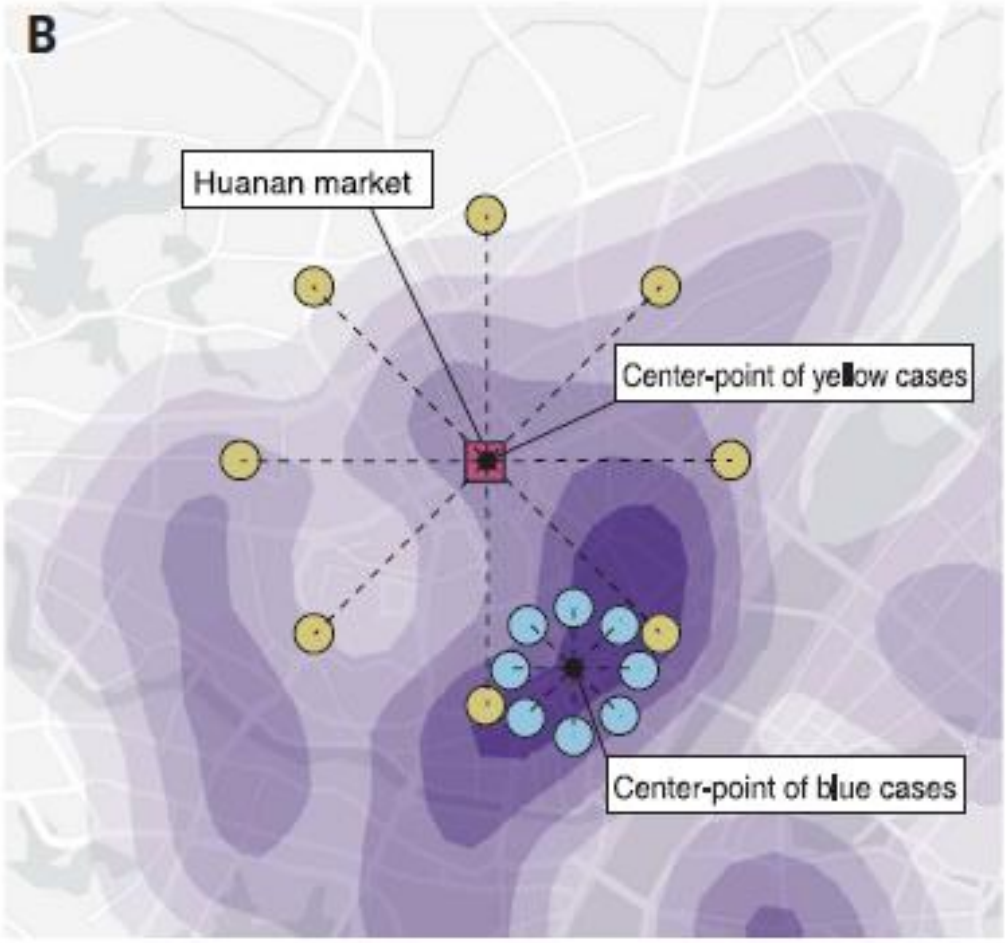
"Near" is Not the Same as "Centered"

Demonstration case:

Yellow cases are both near the market **and** centered on the market

Blue cases are near the market but not centered on it

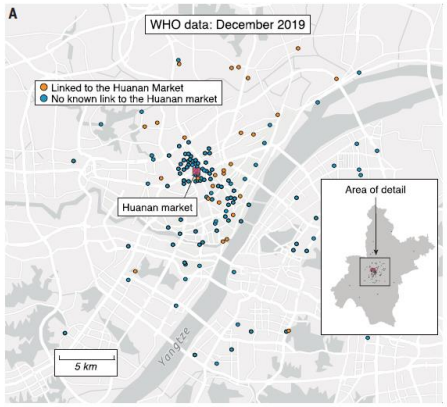
- If it could be shown it would provide much weaker evidence that Huanan itself was the epicenter.
- Perhaps the actual source was that neighborhood, a hospital, another market, etc.
- But it is not the case given the data

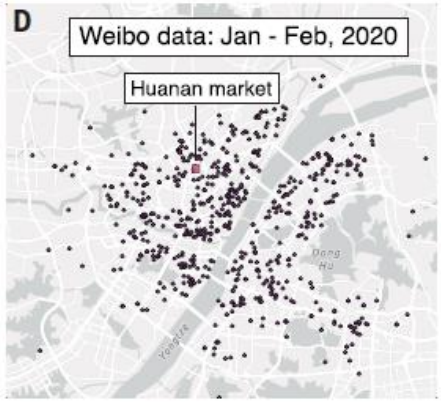


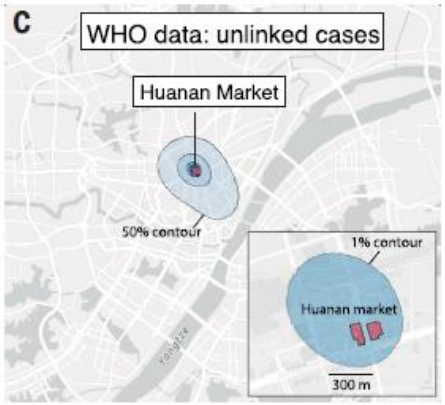
A Few Months After...

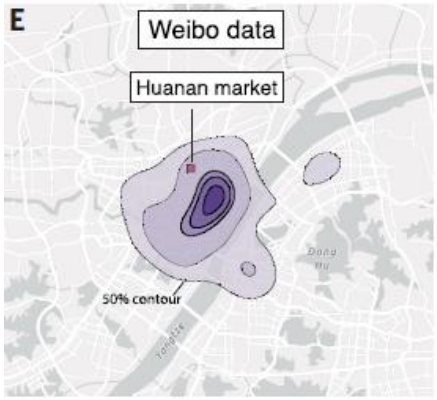
The pattern disappears as the epidemic spreads

- First localized around Huanan
- Later: distributed across Wuhan









December 2019

January-February 2020

Probing Question

Q: If Huanan were merely one arbitrary superspreading location, what spatial pattern might we expect?

A: If Huanan were merely a superspreading location rather than the epidemic's epicenter, the spatial pattern could look quite different—especially for the unlinked cases.

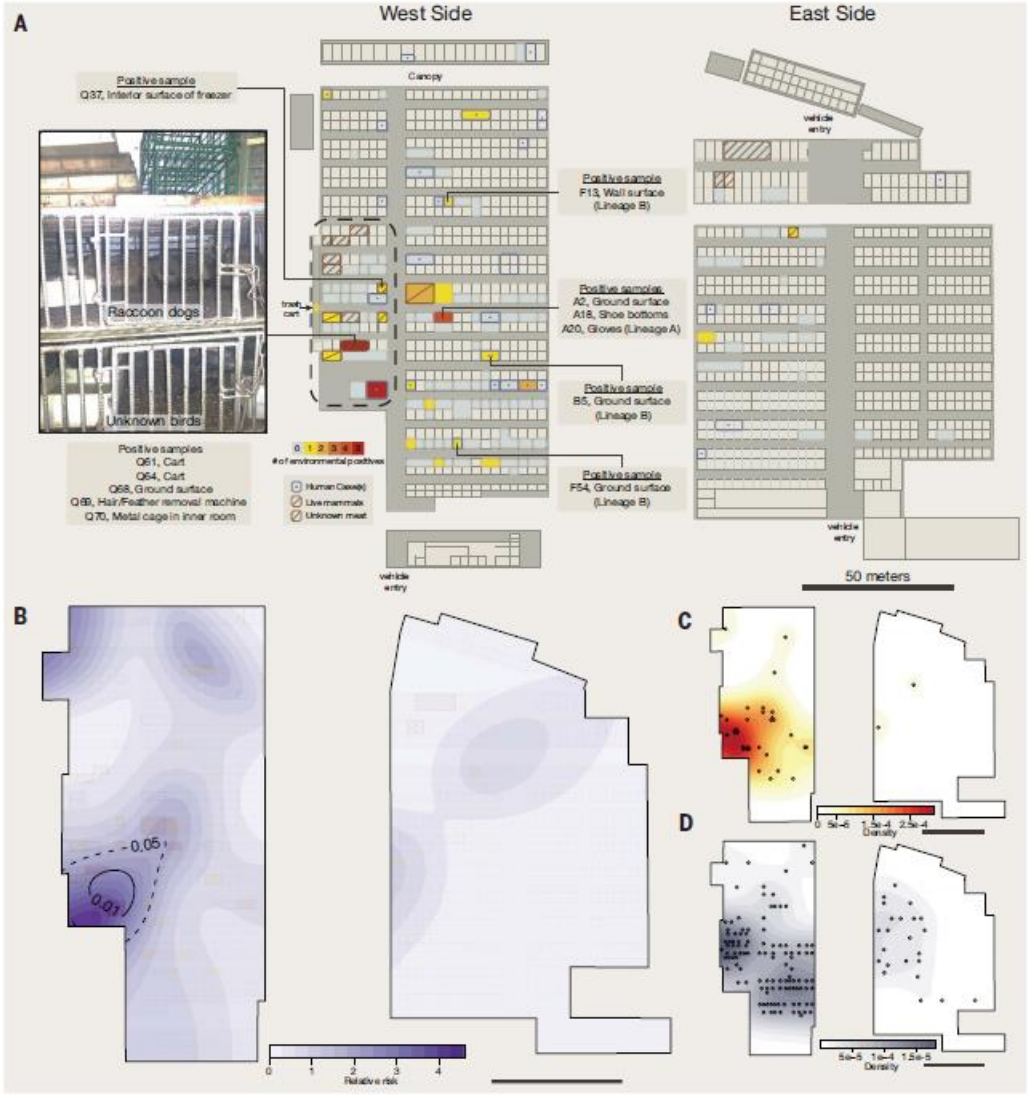
- **Huanan-linked cases** → strong association with the market, because workers/customers were infected there.
- But **unlinked cases** → no particular reason to be geographically centered on Huanan. Their residences might instead reflect wherever the virus was already circulating across Wuhan. That is why unlinked analysis is so important.

Zoom in: From the City to the Market

Now the visual evidence layers are:

- live mammal vendors
- human cases
- environmental SARS-CoV-2 samples
- locations of specific positive samples
- spatial relative-risk analysis

Spatial visualization can integrate heterogeneous evidence



Wuhan Analysis Conclusions

Visualization supports:

- Early cases clustered around Huanan.
- Unlinked cases also centered around Huanan.
- The pattern differed from population density.
- Environmental positives clustered near live-mammal vendors.
- Multiple independent spatial datasets point toward the same location.

Visualization does NOT establish:

- which animal was infected (no direct evidence exists for this)
- precisely when transmission occurred
- the upstream source of the virus
- the complete causal chain

Value of the spatial analysis

- The map did not prove the cause. It made a hypothesis visible—and testable.
- The study also shows how far spatial epidemiological visualization has evolved since Snow while retaining essentially the same intellectual idea

Can We Anticipate COVID Infection Risk?

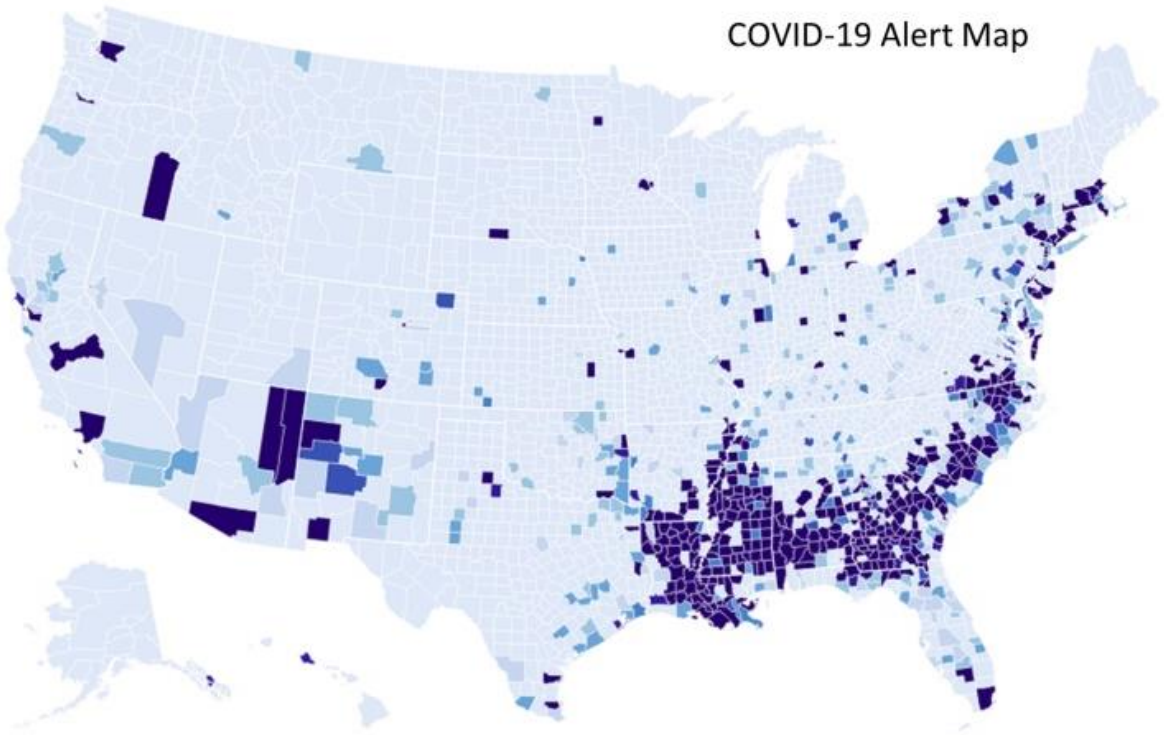
COVID-19 Risk Map

Color mapping:

- the number of times a U.S. county is part of a "high risk" set
- the higher level of risk a county has for high COVID-19 death rates, the darker the color

Only counties with at least 1 death on May 10, 2020, are shown

This map was created on May 10, 2020



The map tells us WHERE infection risk is high

- But it does not tell us WHY

D. Coelho, N. Gupta, E. Papenhausen, K. Mueller, "Patterns of Social Vulnerability – An Interactive Dashboard to Explore Risks to Public Health on the US County Level," *Workshop on Visual Analytics in Healthcare (at AMIA). 2022*

WHY IS COVID RISK HIGH IN CERTAIN PLACES AND NOT IN OTHERS?

This vulnerability is defined by **socioeconomic risk factors,** for example:

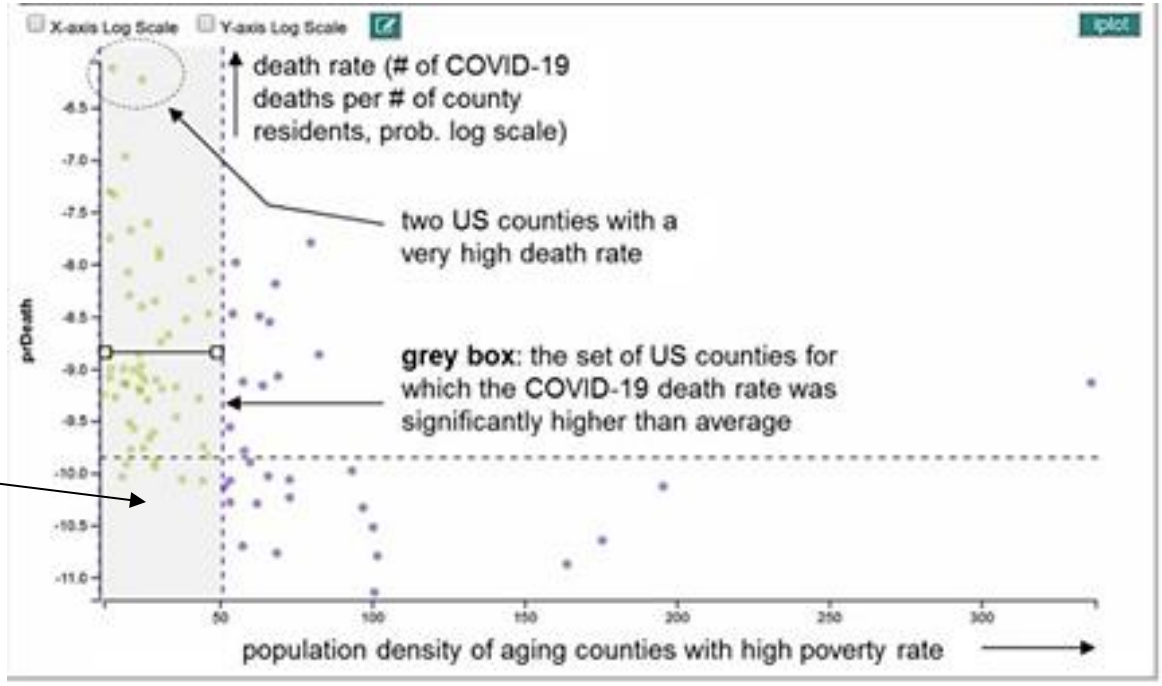
- high poverty + crowded housing + limited healthcare access → high risk
- older population + chronic disease + limited healthcare access → high risk

Identify these risk patterns via **pattern mining** over US counties (3,300)

A COVID risk pattern is a region in feature space where the average of COVID mortality is > than the global average

Can include counties with lower mortality

- these counties may not have been 'visited' by the pandemic yet

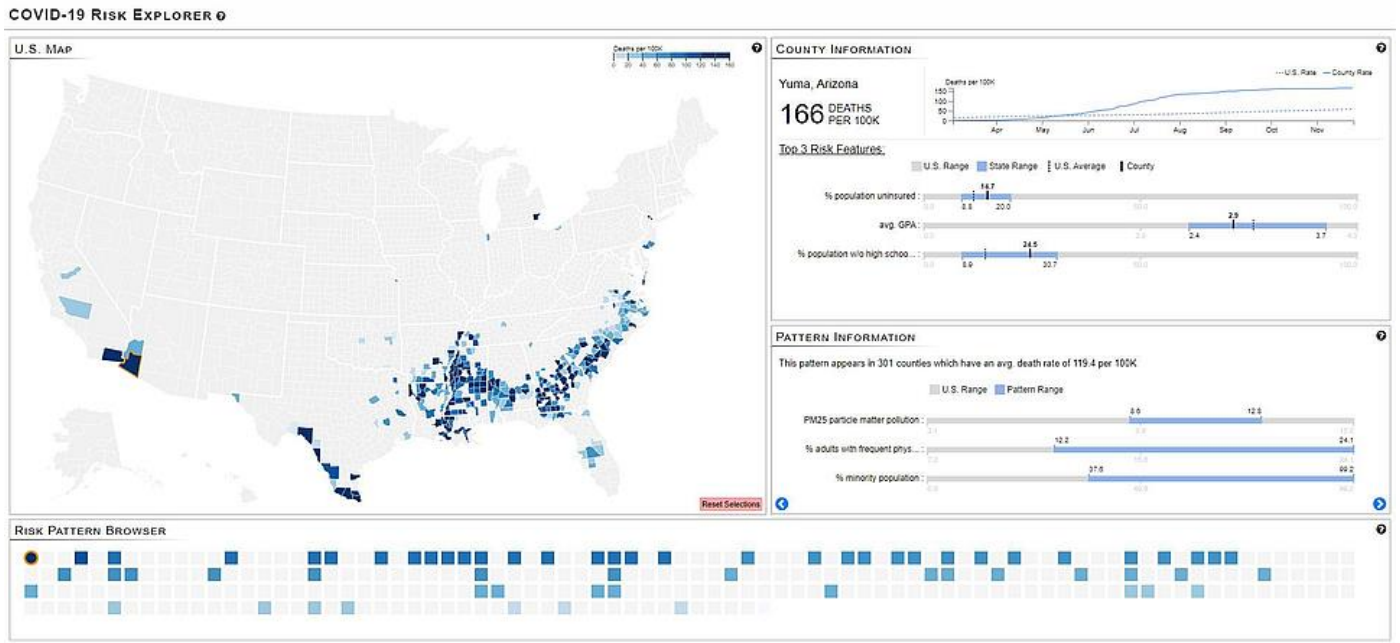


Interactive COVID Risk Explorer

Visualization — Snow: *See* patterns in data

Analytics — Wuhan: *Test* whether those patterns hold up through filtering, modeling, comparison, and uncertainty analysis

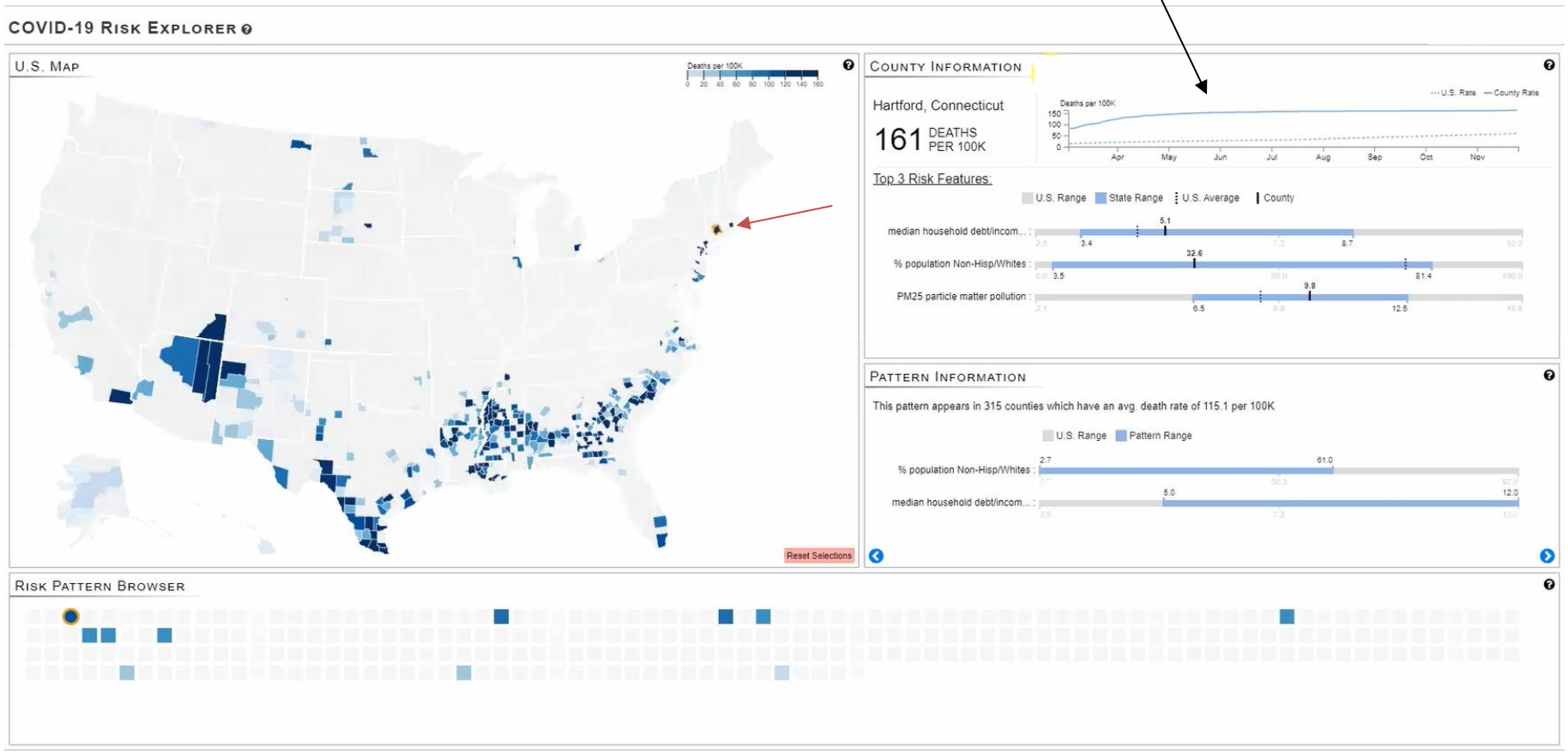
Exploration — COVID Risk Explorer: *Interactively ask new questions* that weren't necessarily specified in advance → **Visual Analytics**



COVID Risk Explorer: Let's See Some Examples

Hartford County, CT

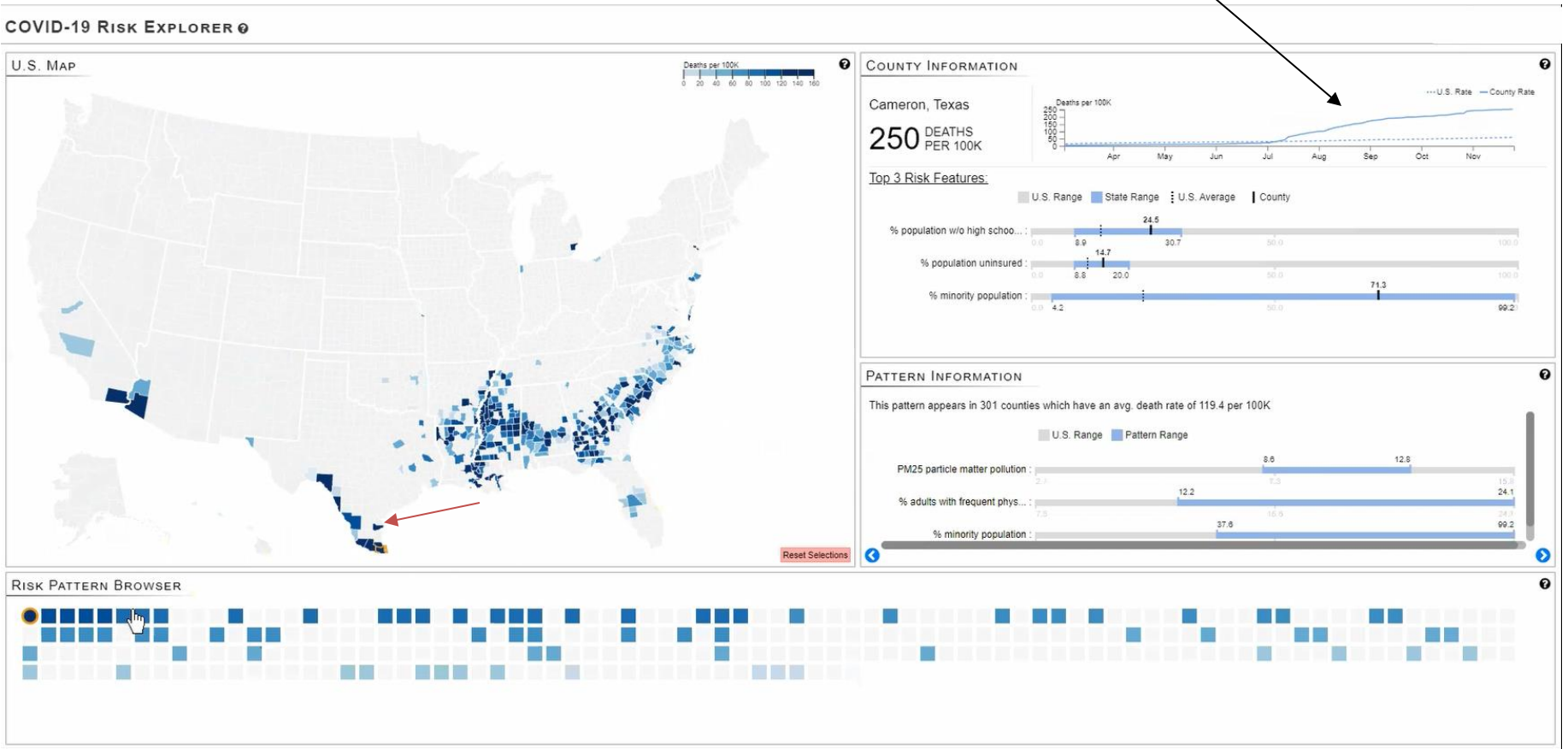
Was hit by COVID early on but stabilized quickly



Camron County, TX

Has many more patterns → more risk

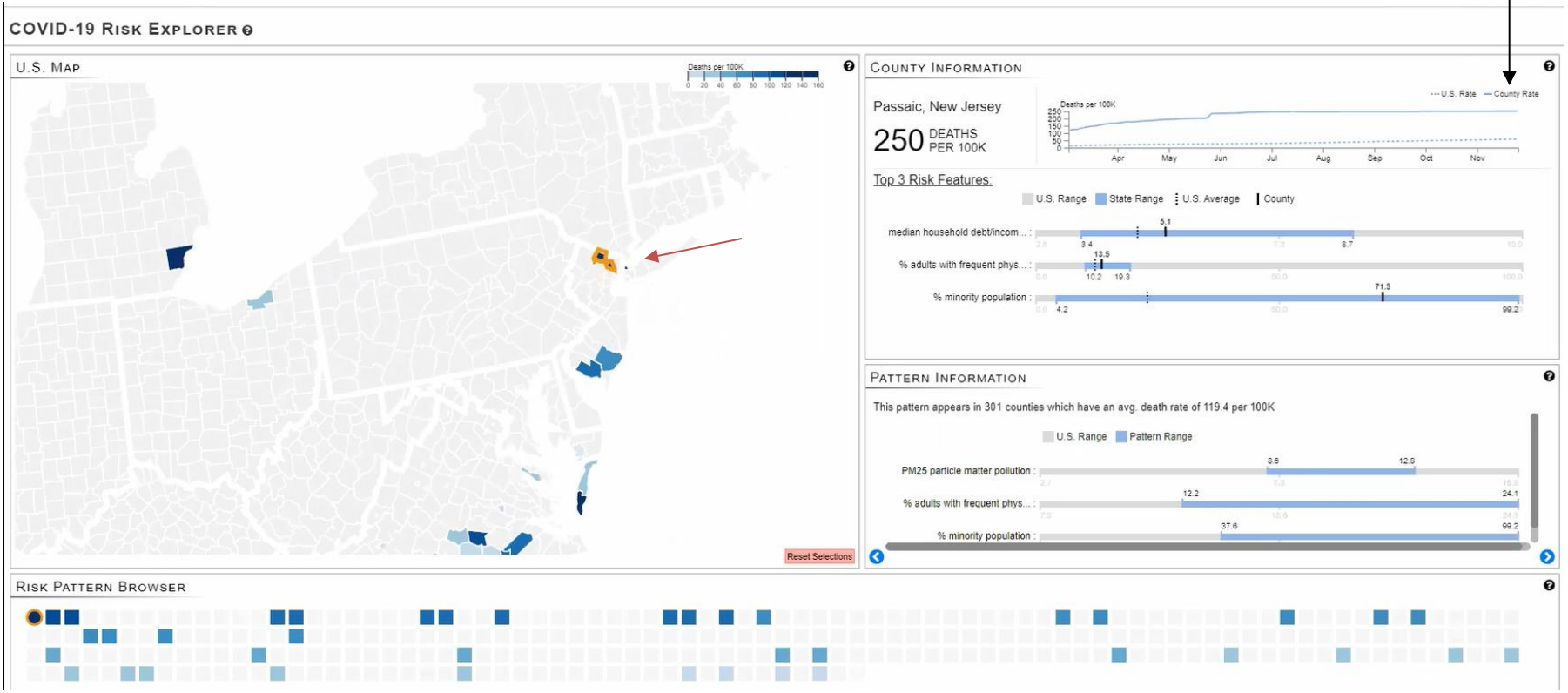
Was reached by pandemic much later & death toll still rising



Passaic County, NJ

Shares most dominant risk factor pattern with Cameron, TX

- but was reached by pandemic much earlier
- mitigated well → death curve flattened quickly
- could have provided (1) early warning and (2) mitigation advice to Cameron, TX

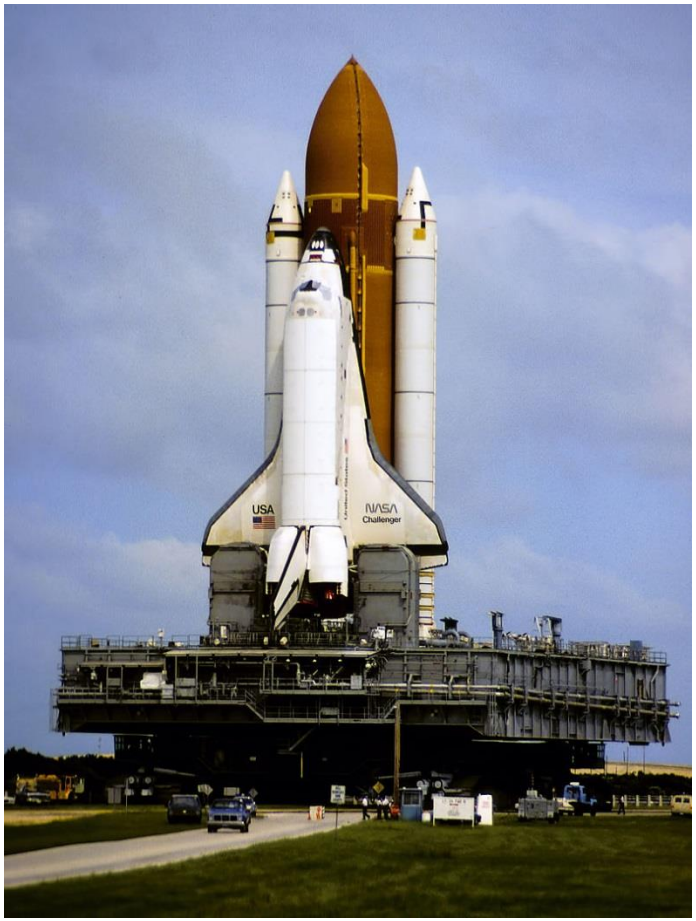


One More Example

Let's go Back Some 40 Years, 1986

Location: JFK Space Center, FL

The crew of Space Shuttle mission STS-51-L 11/15/85. Back row, left to right: Ellison S. Onizuka, Sharon Christa McAuliffe, Greg Jarvis, Judy Resnik. Front row, left to right: Michael J. Smith, Dick Scobee, Ron McNair.



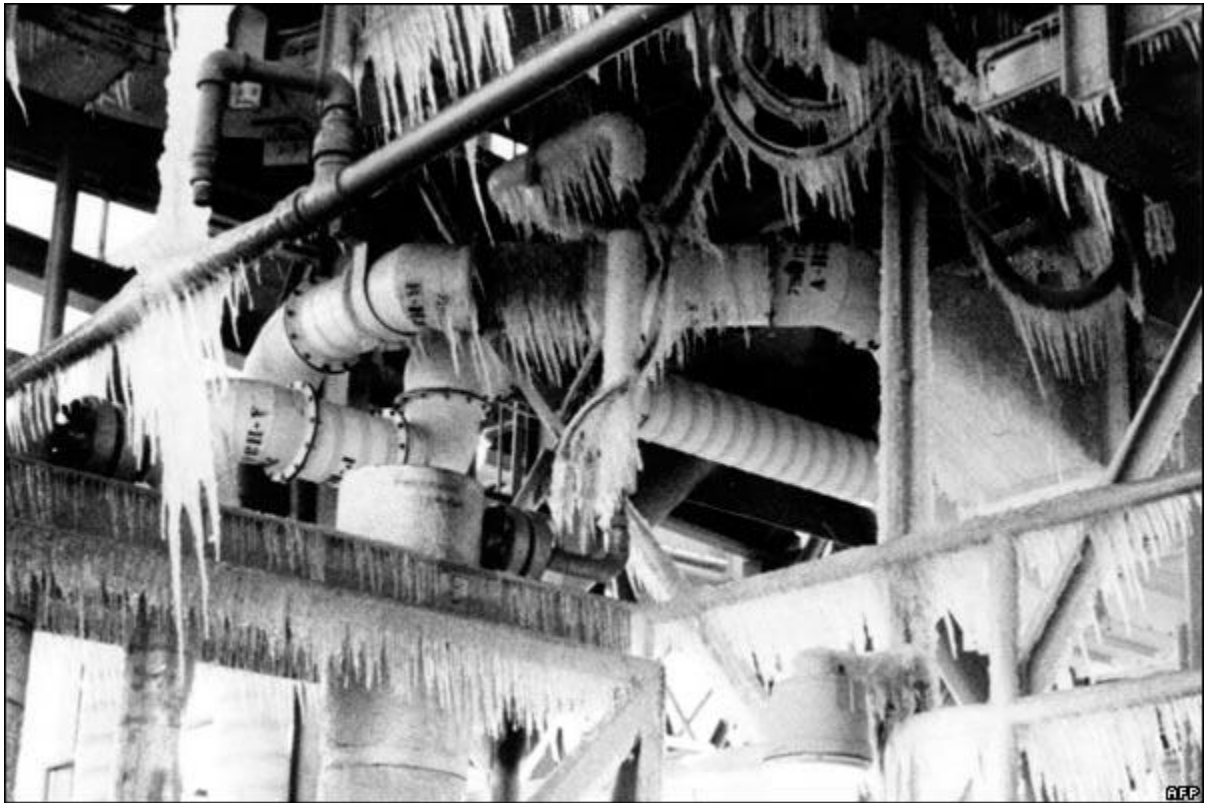
73 Seconds After Lift-off

What Happened?

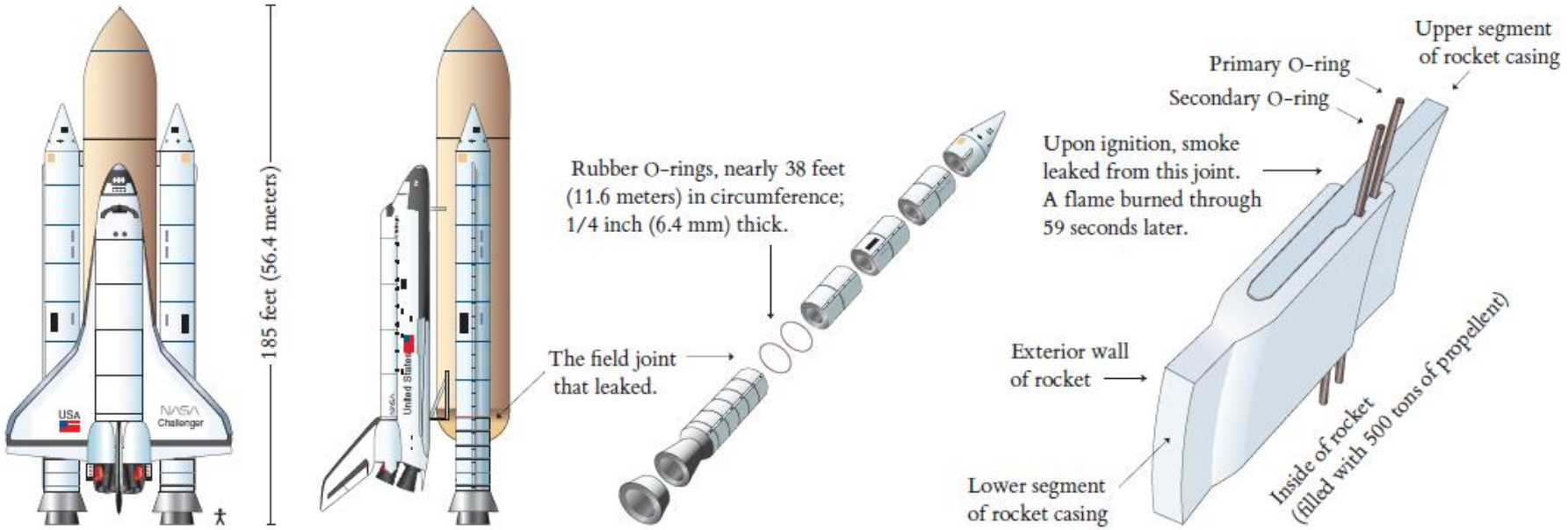
What Was The Cause?

The Day of The Launch

36 degrees F on Launch Pad 39



Space Shuttle 101



Fast Forward
58 Seconds After Ignition

What Happened?

What Was The Cause?

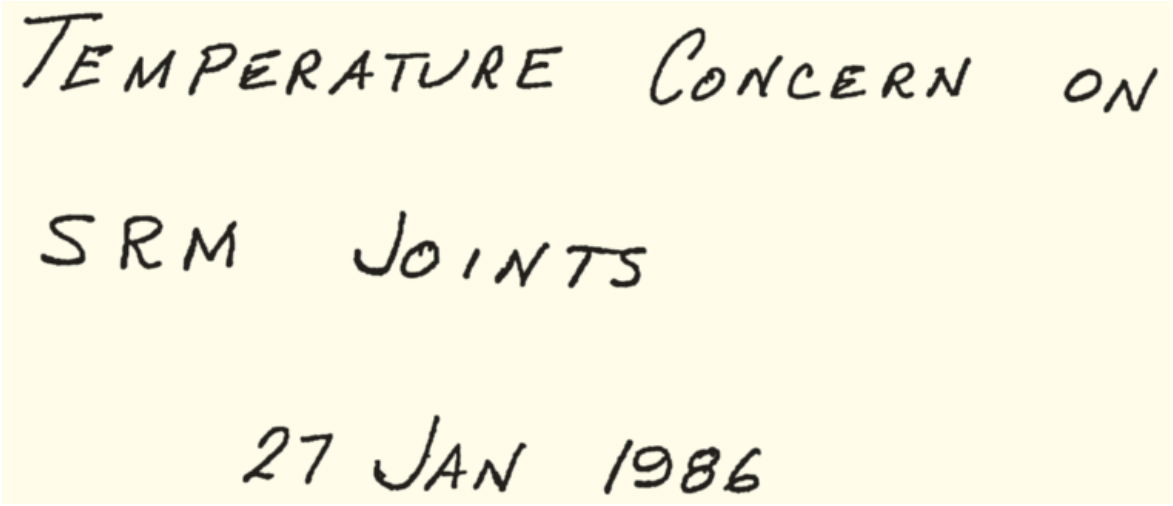
Could It Have Been Prevented?

Engineers At Thiokol Had A Hunch

Two days before launch they presented their concerns

- created 13 charts to make their case

Slide #1:



- SRM – Solid Rocket Motor

SLIDE #2

Teaches about past damages to O-ring

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36° --66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

SLIDES #2 AND 3

Teaches about O-ring damage mechanics and erosion

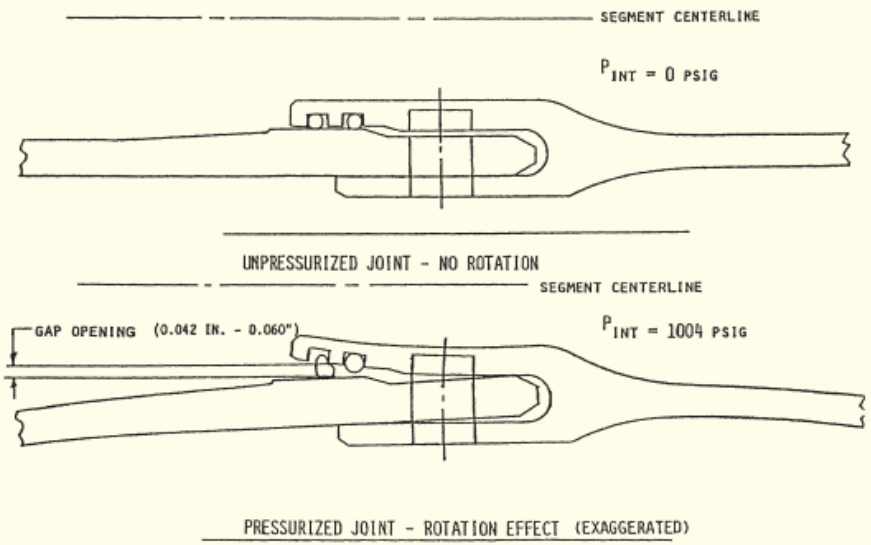
PRIMARY CONCERNS -

FIELD JOINT - HIGHEST CONCERN

- EROSION PENETRATION OF PRIMARY SEAL REQUIRES RELIABLE SECONDARY SEAL FOR PRESSURE INTEGRITY
  - IGNITION TRANSIENT - (0-600 MS)
    - (0-170 MS)HIGH PROBABILITY OF RELIABLE SECONDARY SEAL
    - (170-330 MS) REDUCED PROBABILITY OF RELIABLE SECONDARY SEAL
    - (330-600 MS) HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY

- STEADY STATE - (600 MS - 2 MINUTES)
  - IF EROSION PENETRATES PRIMARY O-RING SEAL - HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY
    - BENCH TESTING SHOWED O-RING NOT CAPABLE OF MAINTAINING CONTACT WITH METAL PARTS GAP OPENING RATE TO MEOP
    - BENCH TESTING SHOWED CAPABILITY TO MAINTAIN O-RING CONTACT DURING INITIAL PHASE (0-170 MS) OF TRANSIENT

PRIMARY CONCERNS - CONT



SLIDES #4 AND 5

Lists temperature and blow-by history for two SRMs

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

Assume You're a NASA Manager

Given the information provided in the company slides

- would you vote for a launch?
- ignore you know about the consequences

Be keenly aware of the immense PR pressures

- President Reagan's upcoming State of the Union speech
- the first civilian in space
- NASA's funding problems

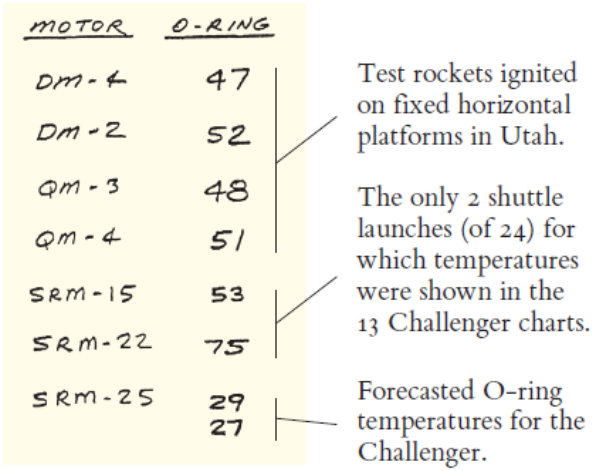
Launch:

- No: OK with a PR disaster & possible budget cuts down the road
- Yes: the rocket company is too cautious & concerns are unproven

Why The Recommendation Failed

Presentation only has exactly two shuttle flights

- one with two blow-by's and high temperature
- one with two blow-by's and low temperature
- ignores all other 22 shuttle flights (SRM)



Statistically weak

Recommendation

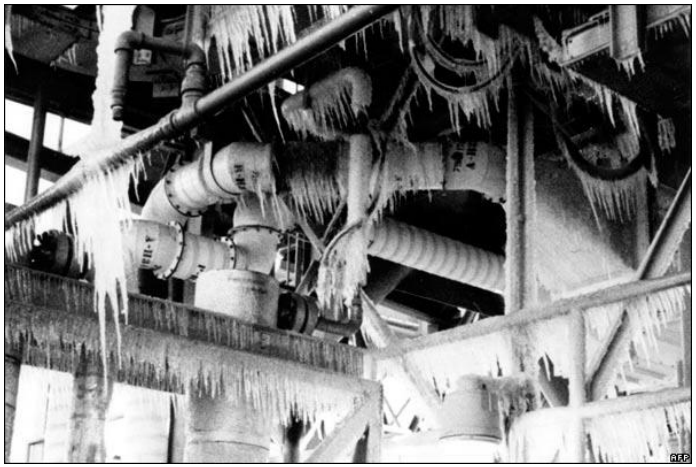
- "O-ring temp must be >53°F at launch"

- is only based on a sample size of 1
- context of other flights is missing
- no statistical leverage

Deficiencies

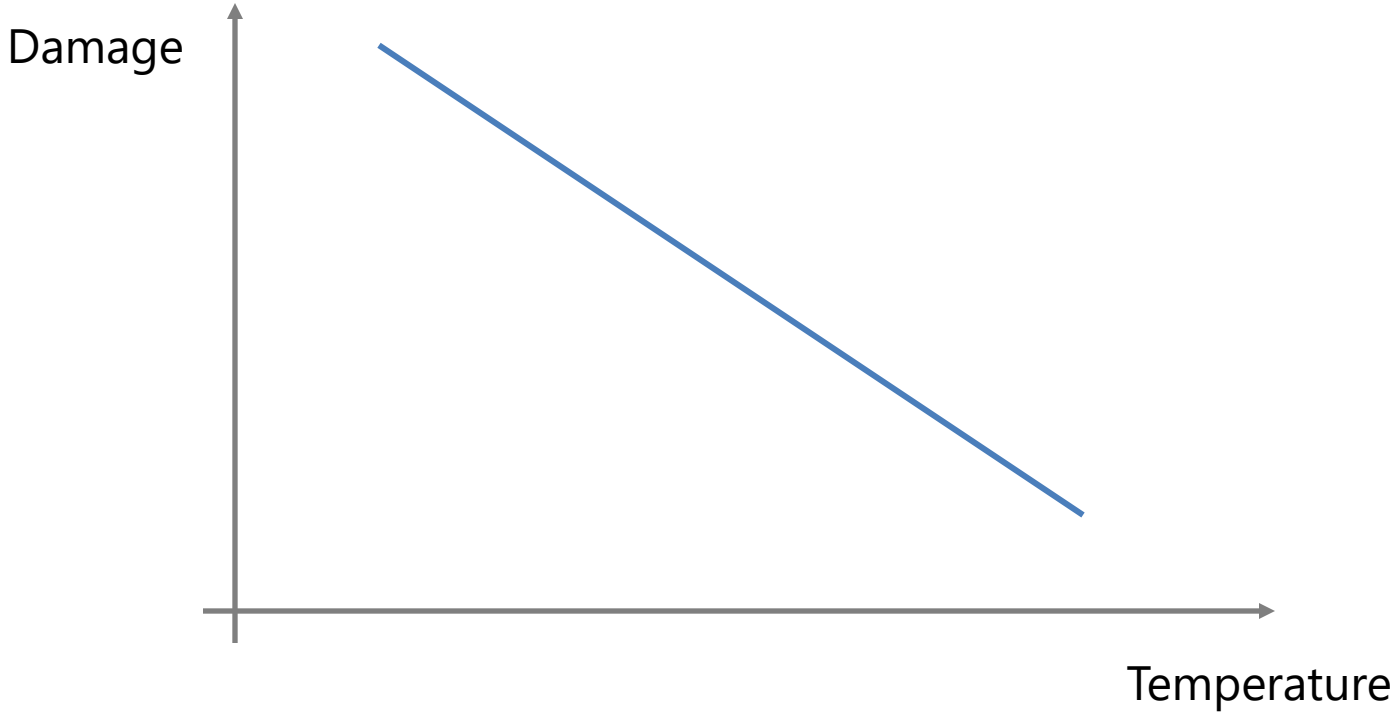
Lots of numbers and facts

But no causal evidence that could predict



What is needed?

What Is Needed?

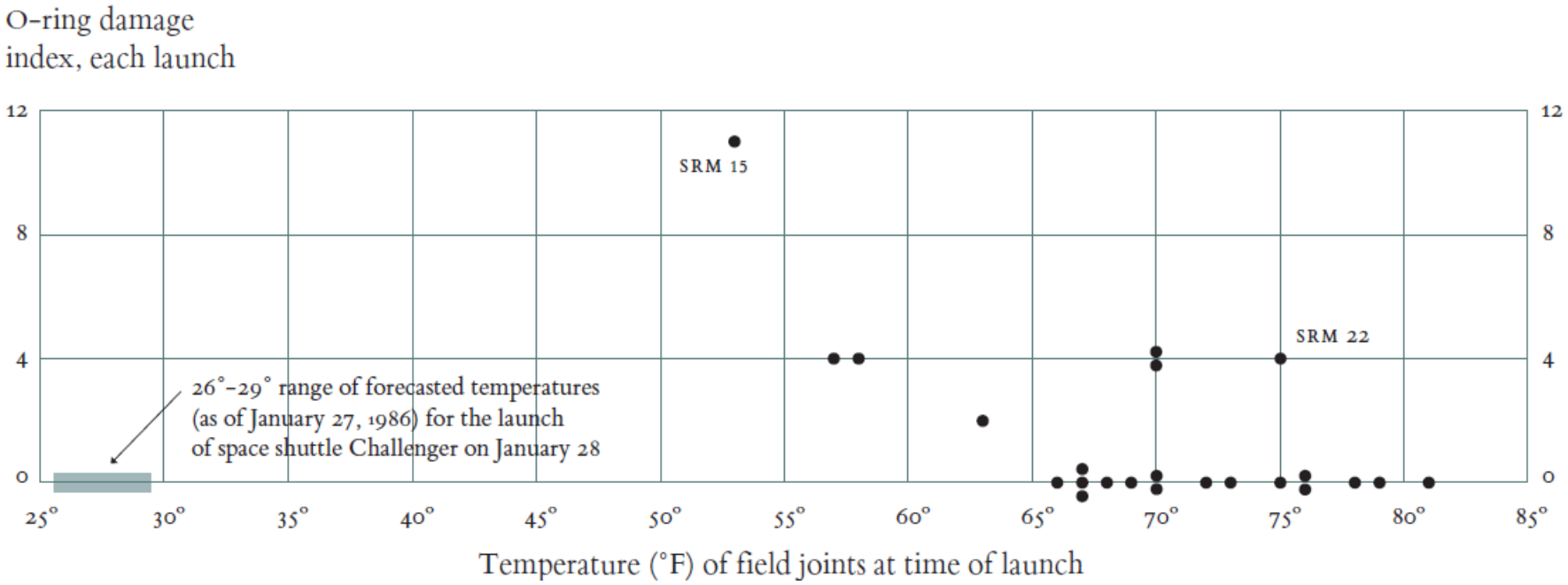


Need a measure for damage

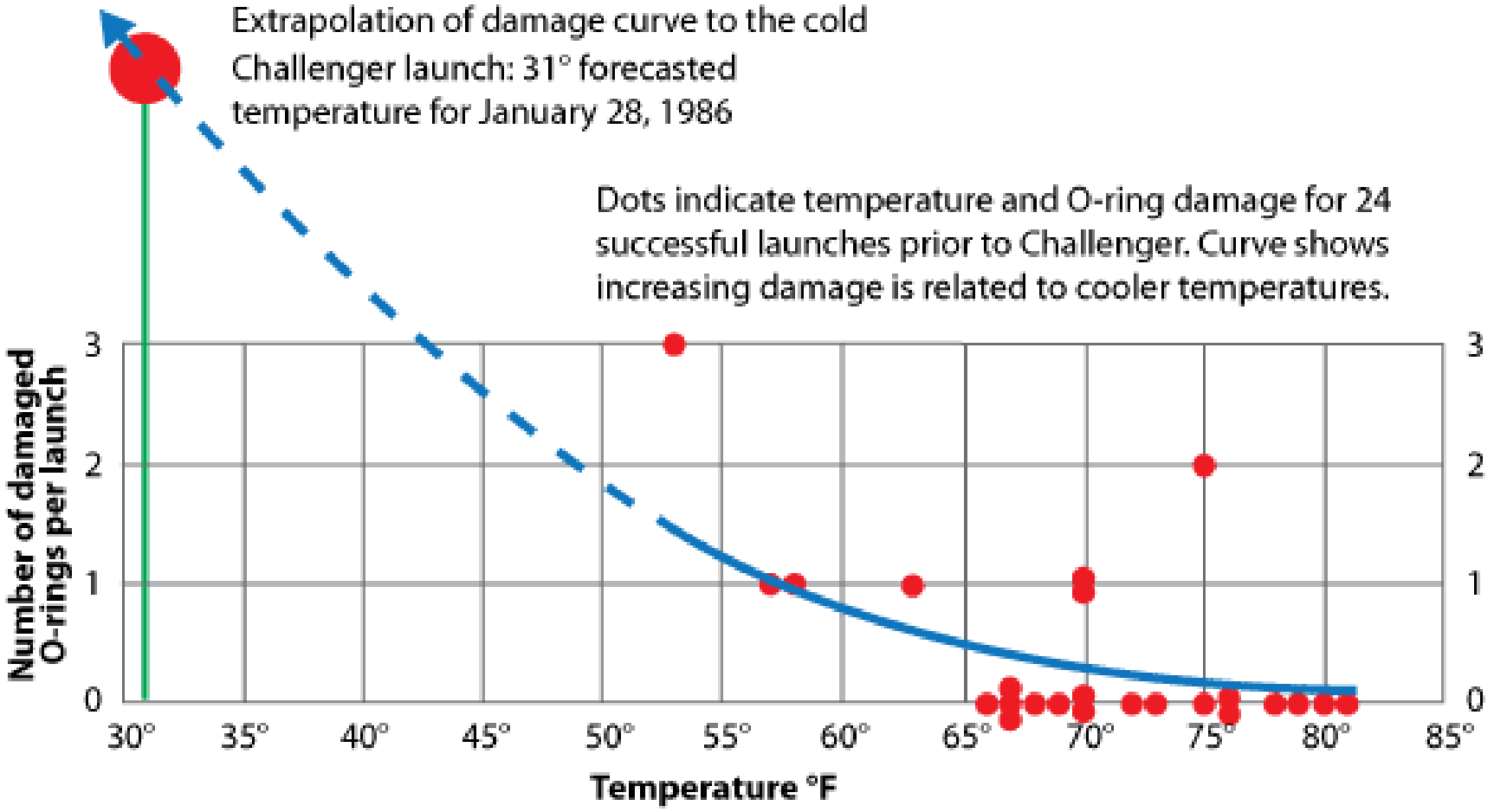
Damage Index

Flight	Date	Temperature °F	Erosion incidents	Blow-by incidents	Damage index	Comments
51-C	01.24.85	53°	3	2	11	Most erosion any flight; blow-by; back-up rings heated.
41-B	02.03.84	57°	1		4	Deep, extensive erosion.
61-C	01.12.86	58°	1		4	O-ring erosion on launch two weeks before Challenger.
41-C	04.06.84	63°	1		2	O-rings showed signs of heating, but no damage.
1	04.12.81	66°			0	Coolest (66°) launch without O-ring problems.
6	04.04.83	67°			0	
51-A	11.08.84	67°			0	
51-D	04.12.85	67°			0	
5	11.11.82	68°			0	
3	03.22.82	69°			0	
2	11.12.81	70°	1		4	Extent of erosion not fully known.
9	11.28.83	70°			0	
41-D	08.30.84	70°	1		4	
51-G	06.17.85	70°			0	
7	06.18.83	72°			0	
8	08.30.83	73°			0	
51-B	04.29.85	75°			0	
61-A	10.30.85	75°		2	4	No erosion. Soot found behind two primary O-rings.
51-I	08.27.85	76°			0	
61-B	11.26.85	76°			0	
41-G	10.05.84	78°			0	
51-J	10.03.85	79°			0	
	06.27.82	80°			?	O-ring condition unknown; rocket casing lost at sea.
51-F	07.29.85	81°			0	

Visualize It – Just The Facts

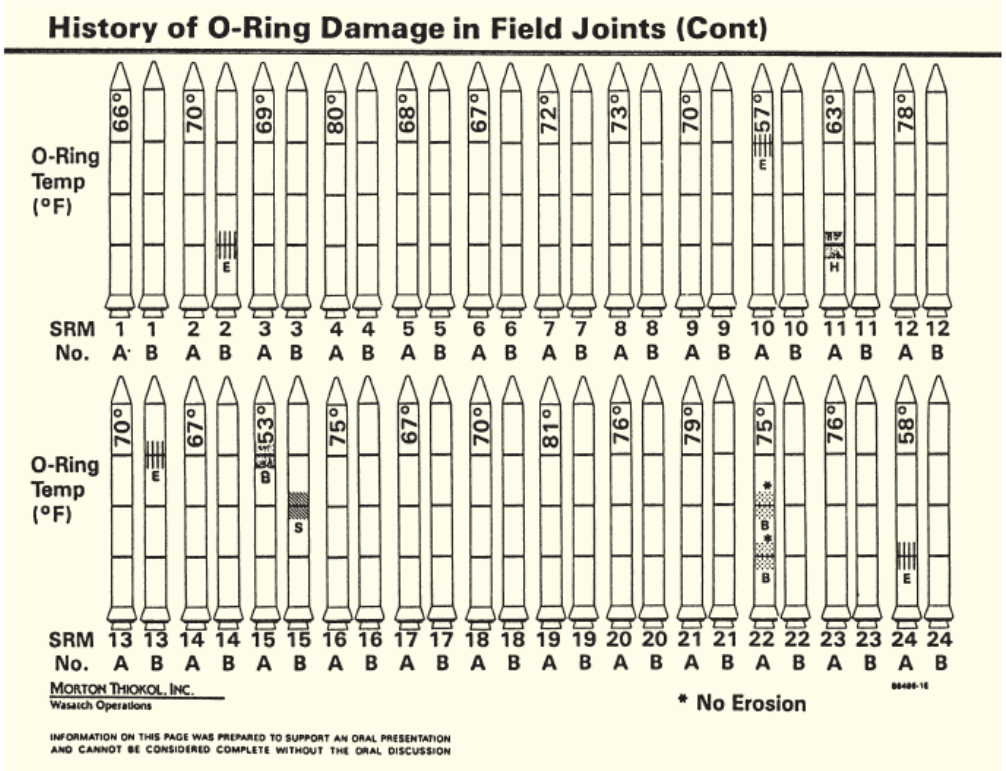


Visualize It – Tell The Story



Shown at Congressional Hearings

Used these charts



All information is there

- but very hard to identify and assimilate
- why?

Quiz Time

What Is Needed for Visualization?

Data (wide variety)

Algorithms

- data mining
- data analytics

Computer

- run those algorithms
- data storage

Humans

- with a purpose/need to understand their data
- endowed with cognitive faculties, creative thought, intuition
- domain expertise

Understanding of humans

- perception, cognition, HCI issues
- we can gain it through experimentation with humans

What Is Needed for Visualization?

Data (wide variety)

Algorithms

- data mining
- data analytics

Computer

- run those algorithms
- data storage

= Visual Analytics

Humans

- with a purpose/need to understand their data
- endowed with cognitive faculties, creative thought, intuition
- domain expertise

Understanding of humans

- perception, cognition, HCI issues
- we can gain it through experimentation with humans

Dr. John Snow: A Visual Analytics Pioneer

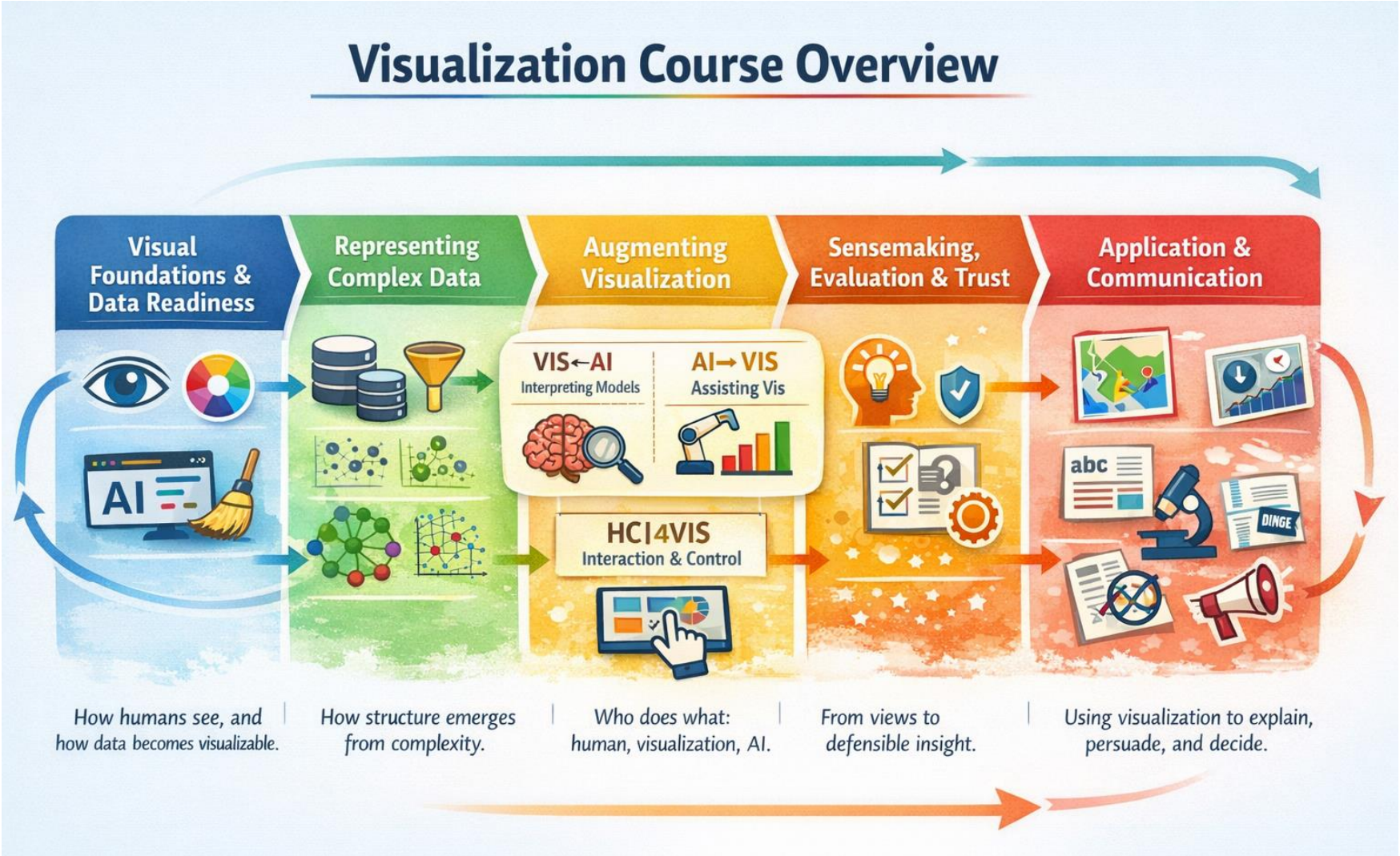
Dr. John Snow's London Cholera Map of 1854

- data collection
- data assimilation
- statistical testing
- visualization
- computational analysis (brain)
- domain knowledge

Very early example of visual analytics

This Course – Specifics

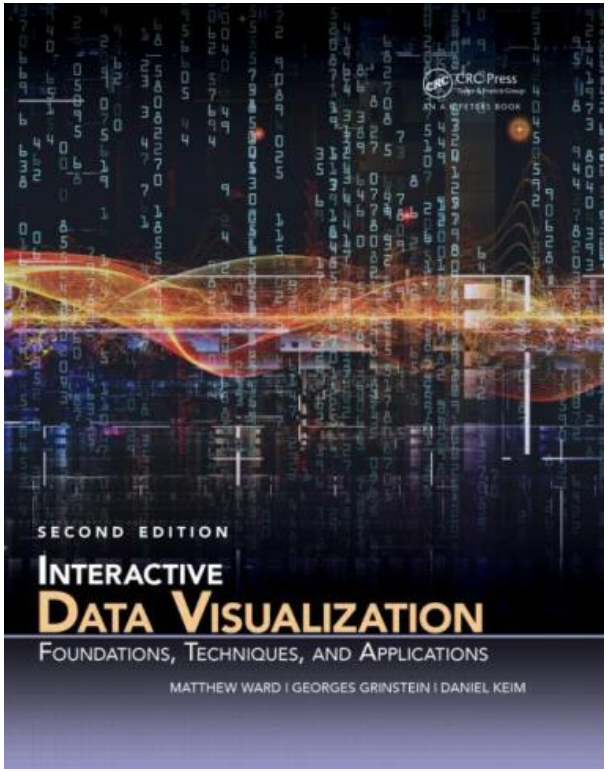
What You Will Learn



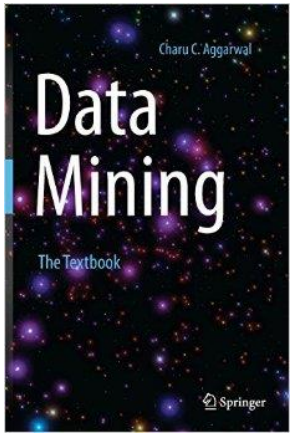
How We Will Get There

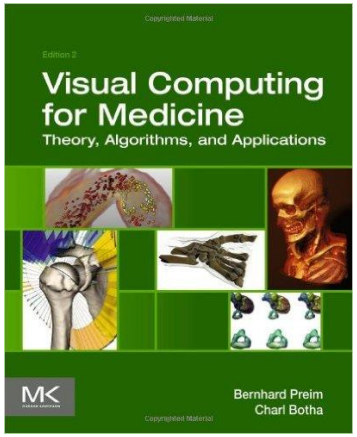
Lecture	Topic	Projects
1	Intro and logistics	
2	Basic visualizations and tasks, data types, examples, ethical considerations	
3	Data preparation (cleaning, imputation, data set integration)	Project #1 out
4	AI-assisted coding for VIS applications (design, debugging, refactoring)	
5	Big data and data reduction (distance/sim metrics, intro to clustering)	
6	High-D data: concept, subspaces, dimension reduction, PCA	
7	Cluster analysis: hierarchical, density, model, embedding, temporal	Project #2(a) out
8	Perception and cognition (human visual system, color, contrast)	
9	Visual design and aesthetics	
10	Visualization of multivariate and high-D data: linear methods, projections	
11	Vis. of multivariate and high-D data: non-linear methods, embeddings	Project #2(b) out
12	Visualization and AI: mutual support and capabilities (VIS4AI, AI4VIS)	
13	Principles of interaction: drive what is visualized, analyzed & how (HCI4VIS)	
14	Midterm recap	
15	Midterm	
16	Visual analytics, human-centered AI, mixed-initiative & collaborative VA	Capstone proposal due
17	Visualization system design & evaluation: nested model, study designs	
18	Visualization of hierarchical data	
19	Visualization of maps and data with geo-reference	
20	Vis. of time-varying, time-series, streaming data, progressive visualization	
21	Critiquing visualization: Tufte's views, integrity, uncertainty, responsibility	Capstone prelim report due
22	Design of effective infographics	
23	Foundations scientific & medical visualization, computer graphics	
24	Intro to volume rendering	
25	Scientific visualization	
26	Capstone project demos (4-5 minutes + 2 minutes Q&A)	All capstone materials due
27	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
28	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
Final	Final exam	

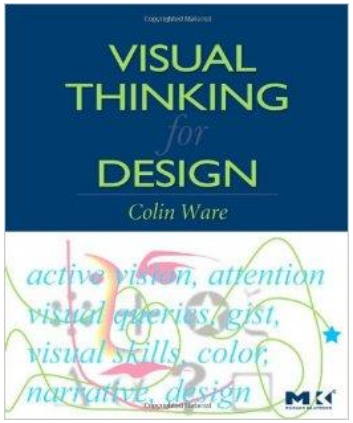
Textbooks

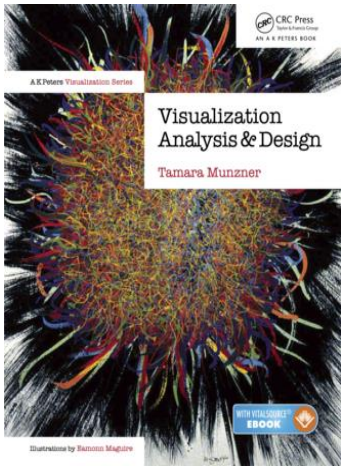


Required









Optional

Online Presence

Course website:

- https://www3.cs.stonybrook.edu/~mueller/teaching/cse332/

Everything you need is there:

- syllabus
- course notes (slides) posted shortly after the lecture
- lab assignments
- course policy
- grades (anonymized)

Additionally:

- Brightspace for lab submissions and links to echo and zoom
- Piazza for online support – please register if not already invited
- https://piazza.com/class/msno5omyrgm9o#

Grading

Labs (3): 25% (**skill** aspect, combined into letter grade ABCF)

Midterm/final: 25% each (**theory** aspect, each gets a letter grade ABCF)

Capstone Project: 25% (**synthesis + design** aspect, letter grade ABCF)

- proposal: 2.5%, prelim report: 2.5%, final demo: 20%

Extra credits

- will be given for labs but can only be applied in labs grade

Final course grade

- rewards balanced strength in all aspects (skill, theory, synthesis)

Participation

- not graded, but I hope you will attend regularly and participate
- lectures will live-stream on echo and be recorded
- will use zoom when lectures are fully online (occasionally)

For late submission policy, etc. see website

Capstone Project

A **visual analytics dashboard** that enables analytical tasks

- has synergy with one or more datasets you will identify
- a collection of data visualizations linked together
- needs to support brushing and linking and fit on the screen

Your path to this visual analytics dashboard:

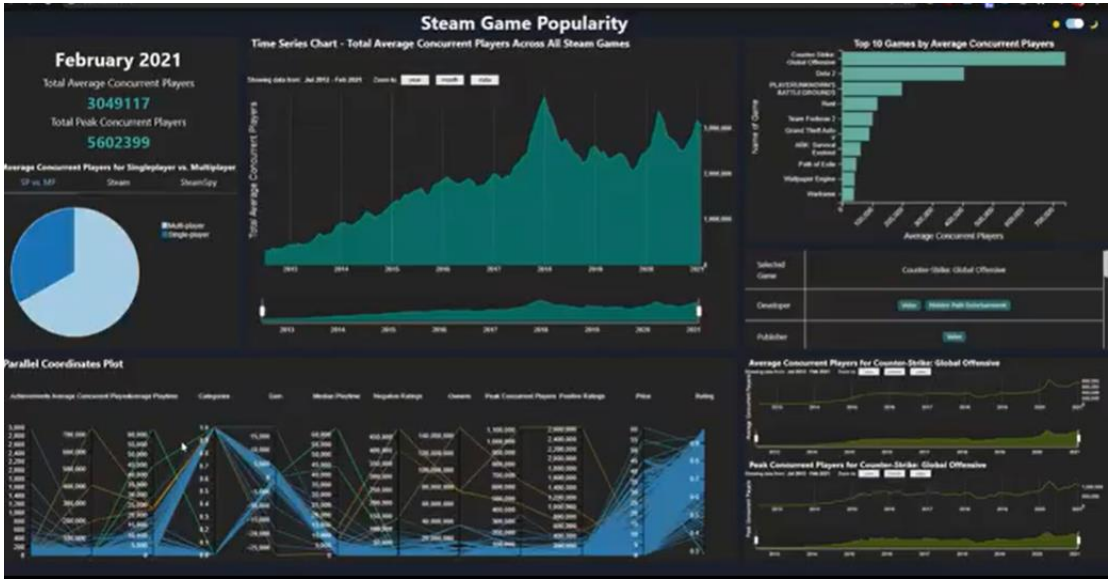
- you will program most (but not all) of the individual dashboard components in labs 1 and 2ab
- the dashboard puts them all on one webpage and connects them in a meaningful way so they together can support users in interactive data analysis explorations

What's A VA Dashboard?

Good example is here, more on the CSE 332 playlists here

Programmed with:

- python
- html
- JavaScript
- D3 API



Capstone project

- Can be done in groups of up to 2 students
- In-class demo in the final 3 lecture periods
- Best demos will be highlighted on the 'Stars of CSE 332' YT playlist

Code Editors

Several free code development environments are available

Visual Studio Code (recommended)

- specifically developed for JS / TS / HTML / CSS tooling
- built-in AI coding support: completion, chat & coding agents
- great for D3 / Vega / Observable-style workflows
- python support is strong
- AI integration is mature: GitHub Copilot + multiple AI models well integrated

Browsers to run and develop your code

- Chrome
- Firefox
- IE and Edge are not overly suitable
- Chrome and Firefox also have panels where you can see and edit your code
- comes in handy when you want to change values of variables

AI Coding Support

You are welcome to make use of AI-assisted coding

Guiding principle: AI is a tool, but not an author – that's **you**

- Use AI tools (e.g., Copilot, Cursor, Claude Code, Codex) to assist with coding
- You will do the design, interaction, and analytic decisions
- AI will give you more time to think about effective visual analytics
- It raises expectations on what you will deliver

AI assistance is encouraged for:

- Boilerplate codes, syntax, debugging, error explanations, cleanup
- Translating intent into initial code drafts
- This reflects how visualization systems are built in practice

AI may **not** replace:

- Justification of visual encodings, explanation of interaction logic
- If you cannot explain it, you did not write it

Coding Languages

Even with AI, you will need to know html and js

- better get ready now if you do not know it

Fortunately, there is a great and easy resource

- W3schools html
- W3schools JavaScript

HTML part, focus on:

- HTML Tutorial (specifically the sections *Home* to *Layout*)
- HTML Graphics
- will take you 2 days max

JavaScript part, focus on:

- JS Tutorial
- JS Objects, JS Functions, JS Async
- JS HTML DOM (Document Object Model)
- JS JSON (JavaScript Object Notation)
- will take you 2 weeks (one hour each day, ~15-20 hours total)

from here

AI Policy – Document All AI Code

In your submission, include:

- which AI tool(s) you used
- what they were used for (e.g., “initial D3 scaffold”)
- what you modified or added yourself
- short annotations are sufficient

Projects are not graded on code quality, rather they are graded on:

- visual design quality
- correctness and clarity
- interaction and analytic insight
- demonstrated understanding of visual analytics and its effective/creative use

Using AI does not lower or raise your grade by itself

AI can help you write code faster — it cannot decide what a good visualization is