Is Data Science a Thing?

Raf Alvarado
Associate Professor (AGF)
UVA School of Data Science
Inaugural SDS Symposium
29 September 2022

TL;DR

Yes, data science **is a thing**

It was invented in the **1960s**

But has been **misrecognized** from the beginning

This is because the **work** of data science lies in

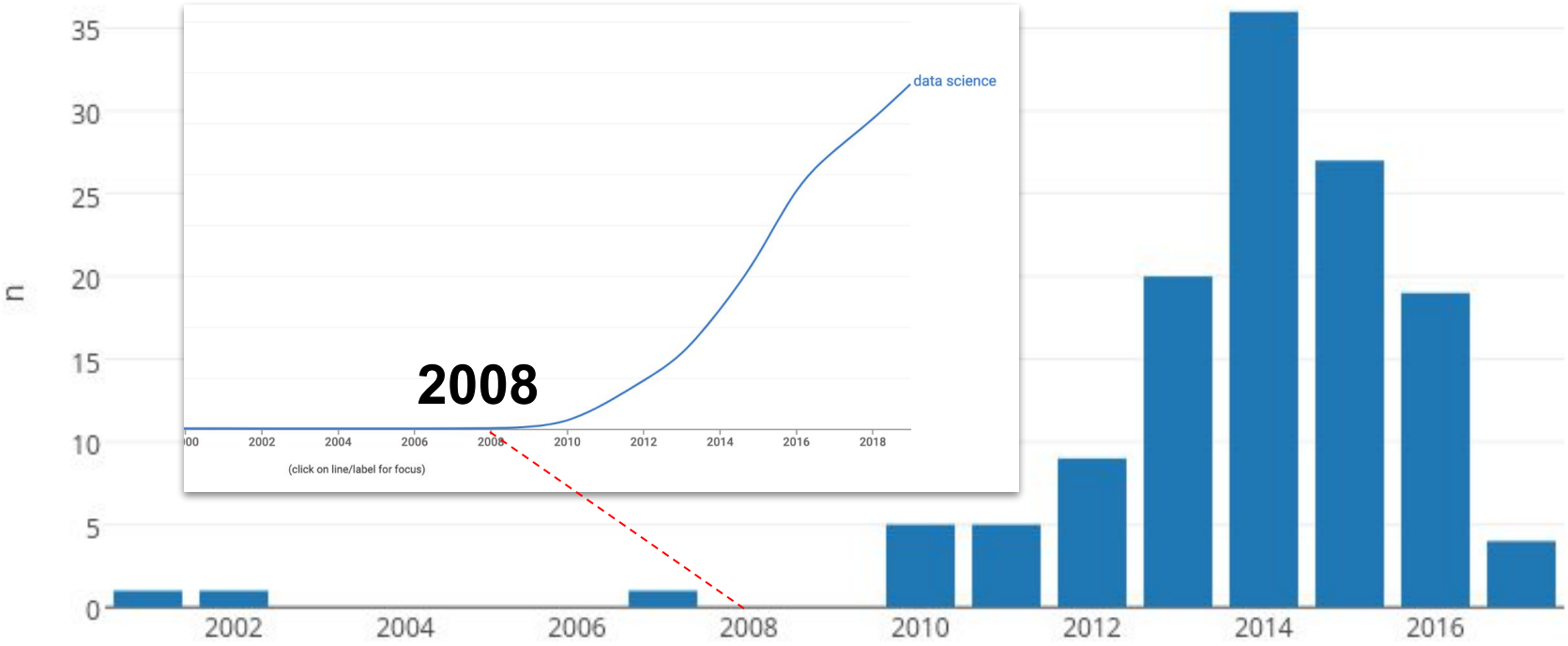
the **messy** and **impure** space

between the **production** of data

and the **communication** of results

Motivation

The **growth** of DS jobs and academic programs since around 2010 has been **massive**



Yet **consensus** on the definition of DS remains **low . . .**

Competing Definitions

– just statistics (cynical)

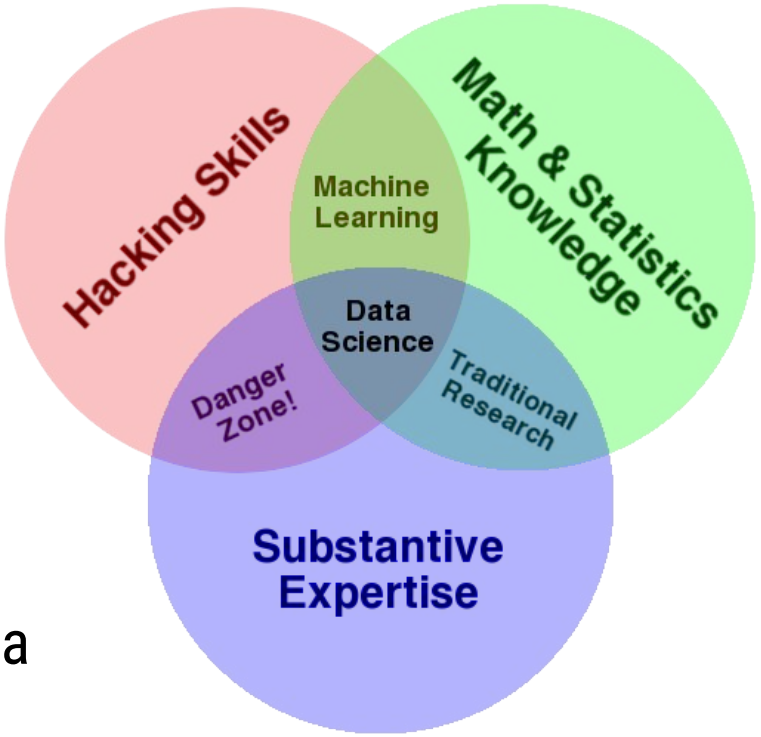
– "fourth paradigm" of science (millenarian)

– machine learning + big data

– the science of managing and processing data

– the art of turning data into actions via data products

– CS + Stats + Domain Knowledge



Conway 2010

Often "defined" as a **laundry list** of useful skills

There are not necessarily contradictory

but they are in **tension**

Leads to **turf battles** in academia

and **confusion** in labor markets

Inhibits growth as a field

Erodes respect for the field

The Disconnect

A particular tension has existed with **statistics . . .**

There has been a recognized **disconnect** in the field since 2012

Noted by **three ASA presidents** after DS trends

Addressed by Donoho in **50 Years of Data Science** (2015) →

The statistics profession is caught at a confusing moment: the activities that preoccupied it over centuries are now in the limelight, but those activities are claimed to be bright shiny new, and carried out by (although not actually invented by) upstarts and strangers. Various professional statistics organizations are reacting:

- *Aren't* **we** *Data Science?*
  Column of ASA President Marie Davidian in AmStat News, July 2013[7]
- *A grand debate: is data science just a "rebranding" of statistics?*
  Martin Goodson, co-organizer of the Royal Statistical Society meeting May 19, 2015, on the relation of statistics and data science, in internet postings promoting that event.
- *Let* **us** *own Data Science.*
  IMS Presidential address of Bin Yu, reprinted in IMS bulletin October 2014[8]

The Rub

Searching the web for more information about the emerging term "data science," we encounter the following definitions from the Data Science Association's "Professional Code of Conduct"[6]

> "Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

To a statistician, this sounds an awful lot like what applied statisticians do: use methodology to make inferences from data.

meaning = inference?

Two countries divided by a common language?

misrecognition

/ˌmisˌrekəgˈniSH(ə)n/

noun

the action of mistaking the identity of a person or thing.
"the real problem is cultural misrecognition"

The misrecognition of data science
by the field of statistics (and the academy)
is **not a new story**

It is related to the difficulty in perceiving
a kind of labor (knowledge work)

for both **social** and **epistemic** reasons

I recall being a proud young academic about 1970; I had just received a research grant to build and study a scientific database, and I had joined CODATA. I was looking forward to the future in this new exciting discipline when the head of my department, an internationally known professor, advised me that data was "a low level activity" not suitable for an academic. I recall my dismay. What can we do to ensure that this does not happen again and that data science is universally recognized as a worthwhile academic activity? Incidentally, I did not take that advice, or I would not be writing this essay, but moved into computer science. I will use my experience to draw comparisons between the problems computer science had to become academically recognized and those faced by data science.

Smith, F.J., **2006**. "Data science as an Academic Discipline." *Data Science Journal*, 5, pp.163–164. DOI: http://doi.org/10.2481/dsj.5.163

(Smith was co-editor of the *Data Science Journal*)

DATA SCIENTISTS

The interests of data scientists—the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection—lie in having their creativity and intellectual contributions fully recognized. In pursuing these interests, they have the responsibility to:

- conduct creative inquiry and analysis;
- enhance through consultation, collaboration, and coordination the ability of others to conduct research and education using digital data collections;
- be at the forefront in developing innovative concepts in database technology and information sciences, *including methods for data visualization and information discovery*, and applying these in the fields of science and education relevant to the collection;
- implement best practices and technology;
- serve as a mentor to beginning or transitioning investigators, students and others interested in pursuing data science; and design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public.

Simberloff, Daniel, B. C. Barish, K. K. Droegemeier, D. Etter, N. Fedoroff, K. Ford, L. Lanzerotti, A. Leshner, J. Lubchenco, and M. Rossmann. **2005**. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." National Science Foundation.

Thesis

Data science was **invented in the 1960s**

> The **name** and the **practice**
>
> It has had a **continuous** and **consistent** usage to the present day

This usage has been **motivated** by a persistent **situation**

> A technoscientific assemblage of **sensors, networks,** and **responders** mediated by **computational machinery** and **mathematical modeling**

This situation is characterized by **data impedance**

> The constant **disproportion** between the volume of instrument-generated **data** and the capacity of computational **machinery** to process it

Data Science designates a form of **expertise** that developed in this context

Timeline

1950s	Role of **data processing scientist** emerges in context of instrument-based science, real-time command-and-control, **data reduction**	Cold War; **SAGE** air defense system; **data deluge** coined; **NASA** formed
1960s	AFCRL forms **Data Sciences Lab** (63); "data science(s)", "data scientist" used in many military, government, and industrial contexts, e.g. **Mohawk** Data Science, the **VA**, **WSML**, etc.	**Tukey** writes on data analysis (63); **CODATA** founded (66); **Mansfield** Amendment passed (69)
1970s	**DSL** disbanded; **Peter Naur** suggests that CS be called data science given importance of data representation; replaces his term **datalogy;** Emanuel **Parzen** uses DS in stats essay (77)	**Crawford** writes on "data and things in the world"; **SQL** and **SGML** developed; interest in infology, information scientists define data, etc.
1980s	Data scientist normalized as a **job description** in **scientific research teams**, continues to be used by corporations, the military, and government agencies in US & UK	Commercial **databases** (Oracle)**, data mining** and **KDD** emerge, **DM** changes connotation; rise of computational statistics; PC revolution; Gibbs Sampling
1990s	Japan's **Hayashi** and **Ohsumi** propose **statistics** be called data science; Hayashi uses, defines DS; int'l conference on DS held; **Kettenring**, **Wu**, **Cleveland** propose same in US. No lasting effect.	Crawford describes "greater statistics" as learning from data; triumph of **ML** (SVMs); **Bayesianism** rises, neural networks revived; **CRISP-DM**; **R** and **Python**; Stats wants to rebrands as DS
2000s	**CODATA** founds **Data Science Journal** **NSF** and **JISC** issue reports on data science	**Leo Breiman**'s "Two Cultures" (2001) published (prophet in wilderness)

2001	**Cleveland**'s Action Plan; **Ohsumi**'s "From Data Analysis to Data Science"	**Google invents the data mining corporation, Breiman**'s Two Cultures, Laney 2001 defines **3D Data**
2002	**CODATA** Data Science Journal	
2005	Simberloff et al. “Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century” (**NSF**)	O'Reilly 2005 coins **Web 2.0;** rise of **blogosphere** and massive user participation data; linked to **data extraction** by O'Reilly
2006	Smith 2006 "Data Science as an Academic Discipline" (**CODATA**)	Palmer 2006 "**Data is the new oil!**" — "Data is just like crude. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals"
2008	Swan, et al. "Skills, Role and Career Structure of Data Scientists and Curators: Assessment of Current Practice and Future Needs" (**JISC**)	**Google** turns 10 ...

2008	**Jeff Hammerbacher** begins hiring "data scientists" at FB; writes essay ""Information Platforms and the Rise of the Data Scientist" explaining decision and connection to BI	**Google** turns 10 **Cheryl Sandberg** brings Google's ML/DM business model to FB ***Nature*** and ***WIRED*** celebrate Google's 10th Anniversary; view Google as paradigm for science "petabyte age"
2009	Hammerbacher's essay published **Nathan Yau** corrects Varian's usage of statistician to data scientist in "**Rise of the Data Scientist**"; **blogosphere** agrees, metabolizes Varian's interview	**Hal Varian** says "the sexy job in the next ten years will be statisticians" in McKinsey interview (published 1/2009) Google's Halevy, **Norvig**, Pereira publish "The Unreasonable Effectiveness of Data"
2010	**Conway** blogs famed Venn diagram of data science; **Mason and Wiggins** blog **OSEMI** model of DS	
2011	**O'Reilly Radar** begins posting explainers on DS; **rants** and **queries** appear in blogosphere	Russell's **Mining the Social Web**
2012	***HBR*** publishes "Data Scientist: Sexist Job of the 21st Century"	

1960s

Emergence of data science
as an official category of work

The **first usages** of the phrase "data science" and "data science" occur in the early 1960s

Data Science Corporation 1962
Mohawk Data Science 1964

Data Sciences Lab (DSL),
Air Force Cambridge Research Laboratories
(**AFCRL**) 1963

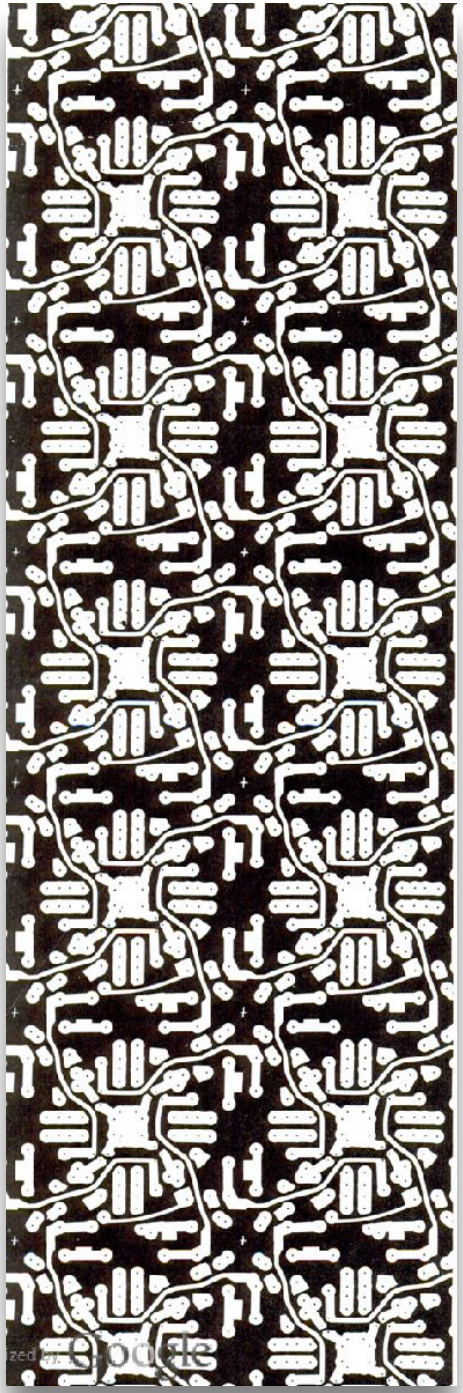
The last is particularly interesting

The Data Sciences Lab (DSL)

One of several labs within the AFCRL

The **meaning** of "data sciences" in this context may be inferred from **reports** about the lab's mission

Since the late 1950's, AFCRL has been a major supporter of research that has led to the present state of the art in the design and fabrication of integrated circuits. Shown here is a segment of an array of individual eight-neighbor elements.



These descriptions come from AFCRL Research Reports in the 1960s

VIII **Data Sciences Laboratory** **187**

Organization and Language of Computers . . . Processing of Audio-Visual Information . . . Processing of Stochastic Information . . . Artificial Intelligence

II **Data Sciences Laboratory** **13**

Recognition Processes . . . Communications . . . Man-Machine Interaction . . . Logic Networks and Circuits

XI **Data Sciences Laboratory** **318**

Computer Languages and Programming . . . Cognitive Processes . . . Speech and Data Transmission . . . Implementation

The DSL's scope form its first report (1963; excerpts)

Modern data processing and computing machinery, together improved communications, with has made it possible to ask for, collect, process and use **astronomical amounts of detailed data**.

A large number of military systems ... deal in **highly perishable information**. **Few existing computers are capable** of handling this information in "**real-time**"

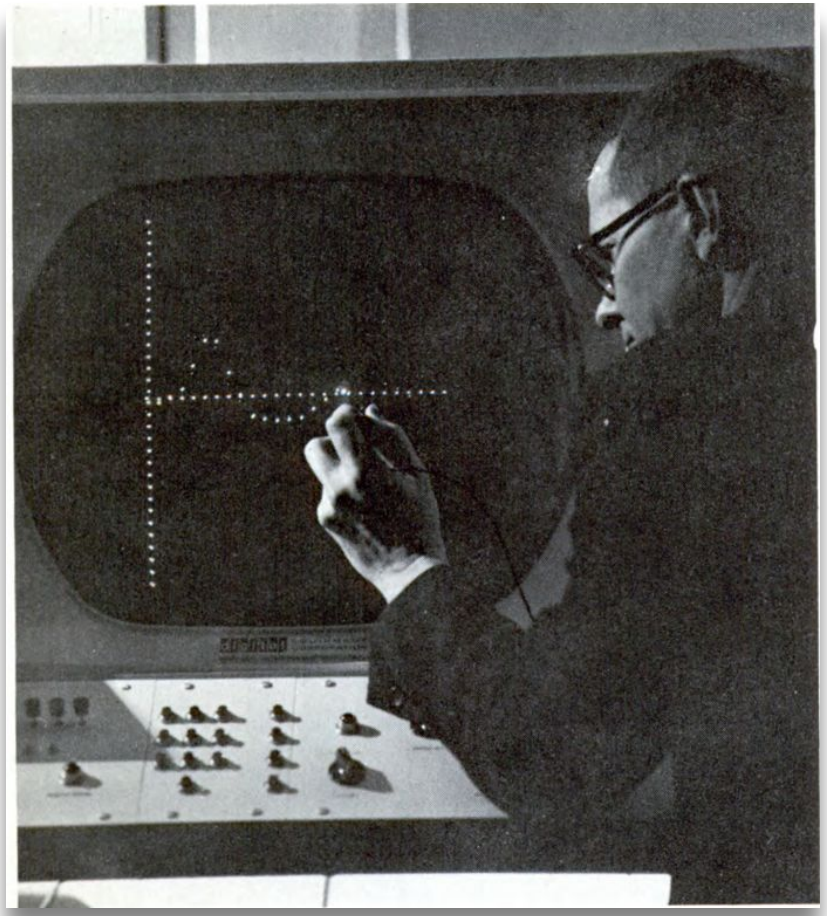
... there is impatience with the **limitations of existing machines**.

... increased speed will not overcome **fundamental shortcomings of existing computers**.

An increasing amount of **data processing research** is aimed at the creation of machines or machine programs that incorporate features of deductive and inductive **reasoning**, **learning**, **adaptation**, **hypothesis formation** and **recognition**.

Artificial intelligence is of utmost importance in **decision situations** where not all possible future events can be foreseen.

Dynamic data processing is concerned with the detection and analysis of recurring patterns and with correlating different sets of patterns on a real-time basis. Parts of AFCRL's DX-1 experimental processor are shown in the two photographs.



For example, the DSL built computational systems that **converted raw signal data into visualizations** to drive decision-making

It is no accident that the **process is gendered** . . .

It is shown that every net has associated with it a system of homogeneous polynomials which, when iterated, completely describe the probabilistic behavior of the net, and several fundamental theorems have been proved which give conditions under which a net may achieve arbitrarily high reliability. Certain special cases of these systems have been shown to be of importance in the study of the genetic characteristics of mating populations as well as in the study of neural combinatorial nets.

ORGANIZATION AND LANGUAGE OF COMPUTERS

Almost all operational computers have preserved strong characteristics of their early ancestors of the late 1940's. They are still programmed numerical calculators, only more elaborate and more difficult to use. Since many of the modern data processing tasks are nonnumerical, radical changes in the organization and the ways of communicating with computers have to be evolved.

These are two examples of the **kind of work** conducted by the DSL

The "net' (left) is a **network of computers** connected to solve a general problem

Focus on "**radical changes in the ways of communicating with computers**" due to "**modern data processing tasks**" (right)

Elements of a Definition

Most of the elements **currently considered central** to DS are here:

— a concern for processing what is later called "**big data**," clearly defined in terms of volume, velocity, and variety (and volatility)

— an embrace of **unstructured** data, including language

— a focus on **artificial intelligence** as an essential approach to extract value from data

— the goal of converting raw data into **visualizations**

The lab produced significant research

on **pattern recognition** and **classification**, **machine learning**, **neural networks**, and **spoken language processing**

Extent of Data Science Units

Other Data Science Organizations (not exhaustive)

Veterans Administration Data Science Division (1965)

NASA Space Data Science Center (1965)

Dynelectron Data Sciences Division (1967)

S Sterling Co Data Sciences Division (1967)

Data Science Ventures Inc (1968)

USAF Data Sciences Division (1968)

White Sands Missile Range Data Sciences Division (1974)

Technology Service Corp Data Sciences Division (1975) → Breiman!

USAF School of AF Medicine Data Sciences Division (1979)

Glaxo Medical Data Sciences Division (2004)

Transnoma Data Sciences Division (2005)

US Veterans Admin

Average employment

1965, 10; 1966, 10.

The Data Science Division—

Conducts basic and applied research and development in advanced data-management sciences and technology.

Develops models of VA programs and operations to provide management the capability of simulating the effect of proposed courses of administrative action and to determine quickly and accurately the effect of proposed legislation on veterans and their beneficiaries.

Provides technical support, guidance, and training in the use of mathematical, statistical, and data-transmission techniques in the field of data management.

Maintains liaison with agencies and activities in the professional, scientific, and technical fields related to data-management research.

	1965	1966
Patient care	36	40
Manpower Administration	20	5
Loan guarantee	19	17
Logistics	16	33
Facility planning and construction	7	8
Automated Reference Library	1	6
Financial benefits	8	25
Management information	13	20
Plant and facility operation and maintenance	0	2
Beneficiary identification and record locator	0	6
Insurance subsystem	13	1
Advanced planning	5	5
Data sciences (mathematical modeling, simulation, etc.)	10	10
Miscellaneous (new legislation and special requests)	23	10
Supervision, project control, and clerical support	28	29
Total man-years	199	217

Dynalectron Corp

AEROSPACE OPERATIONS DIVISION • DATA SCIENCES DIVISION

Services or products?

Field teams? Dynalectron started the whole idea of field teams back in 1951—the idea of sending especially-tailored work forces to do a job wherever aircraft or missile may be. More than 1,500 men now work in Dynalectron field teams all around the free world on aircraft maintenance, overhaul, retrofit, crash- and battle-damage repair, flight tests, systems performance analysis for military requirements.

Facility operations? Dynalectron also provides complete management organizations to government and to industry, for operation of aircraft maintenance facilities and missile launch complexes, including technical and administrative services, maintenance, and other logistics management.

Engineering and technical support? Specialized engineering and technical personnel from Dynalectron are regularly dispatched on special assignments anywhere in the free world for government and industry.

Missile range operation? Dynalectron provides support of the nation's missile and space efforts at White Sands, Air Force Missile Development Center, Air Force Aeromedical Research Laboratory, Utah Missile Launch Complexes and Army Electronic Proving Ground. Dynalectron has been active in range operation activities since 1949, providing engineering, radar, telemetry, optical, communications, timing, aeromedical and photographic technical and support services.

Dynalectron has them

Electronic warfare? Products and systems developed, engineered, and manufactured by Dynalectron are in use in the F-111, B-52, B-57, B-66, and other aircraft now in service, as well as in weapon systems such as Tartar and Terrier missiles.

And beyond? Dynalectron's five divisions and two subsidiaries are also active around the free world in petroleum, petrochemical, and steel process industries; cryogenics; and the international marketing of aircraft and aviation equipment.

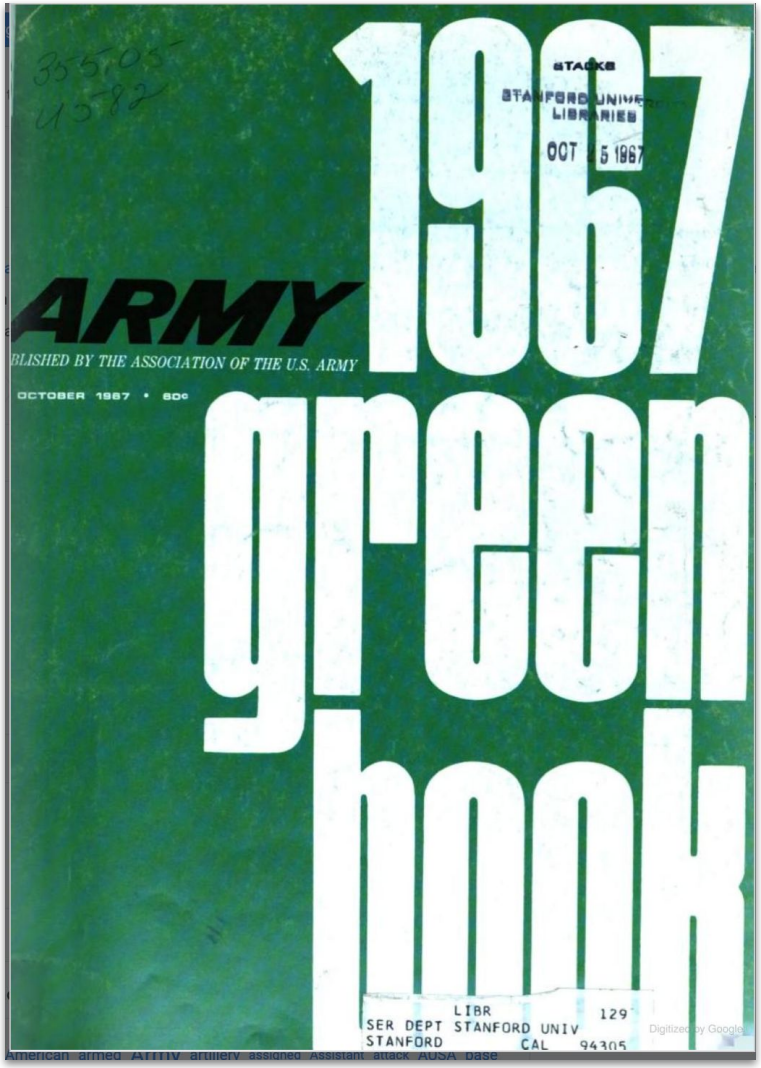
Ask us what we can do for you.

Engineers, technicians, and mechanics who can meet Dynalectron's high standards of excellence are invited to participate in these challenging assignments. Resumes to J. G. Pierce. An equal opportunity employer.

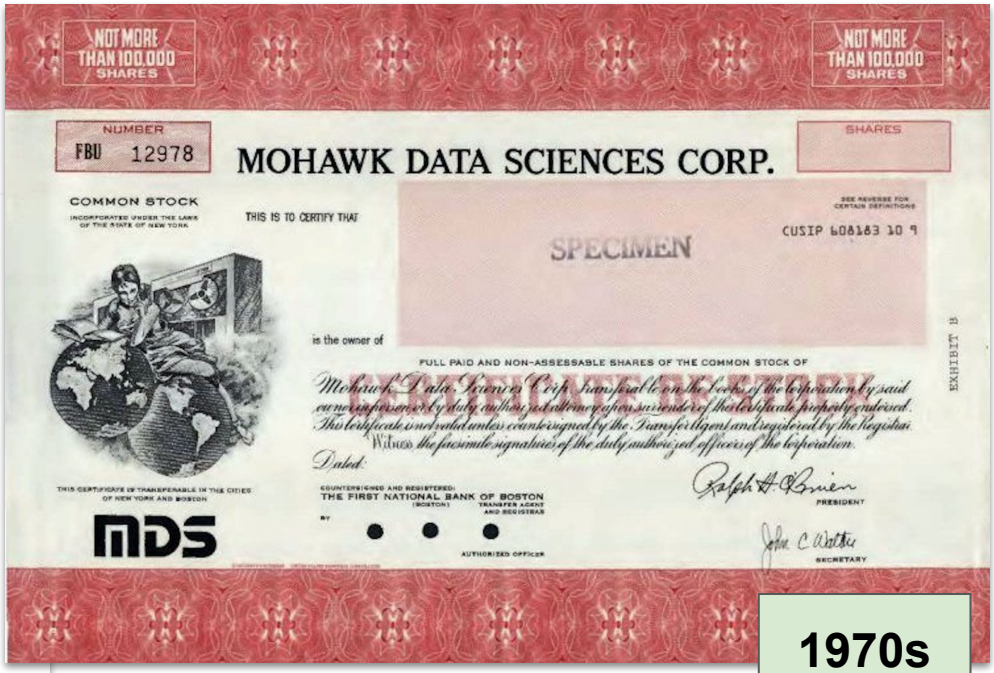
DYNALECTRON CORPORATION

2233 Wisconsin Avenue, N.W.
Washington, D.C. 20007

AEROSPACE OPERATIONS DIVISION • DATA SCIENCES DIVISION • INSTRUMENT & ELECTRONICS DIVISION • LAND-AIR DIVISION • PACIFIC DIVISION • PARADYN DIVISION • HYDROCARBON RESEARCH, INC. • AIR CARRIER SERVICE CORPORATION



Mohawk Data Sciences



1970s

Mohawk Data Sciences Elects a New Director

Herbert Roth Jr.

The Mohawk Data Sciences Corporation, a producer of electronic data input devices, announced yesterday the election of Herbert Roth Jr. as a director and chairman of the executive committee. Mr. Roth is president of the Anelex Corporation, a producer of printers for computer systems, which merged recently with Mohawk Data.

The New York Times
Published: November 2, 1967

1960s

However, he's still suffering badly in Mohawk Data Sciences, with a loss of over $5 million. Edelman had been on Mohawk's board for a couple of months, but he later withdrew. With a continuing 8 percent interest, though, he obviously has a powerful voice in the company's affairs.

The Bottom Line/Dan Dorfman

RAIDERS ON THE PROWL

20 NEW YORK/MARCH 31, 1986

1980s

Founded former **Univac** employees

First product was a **Key-to-Tape Data Entry device** that did away with Keypunch devices

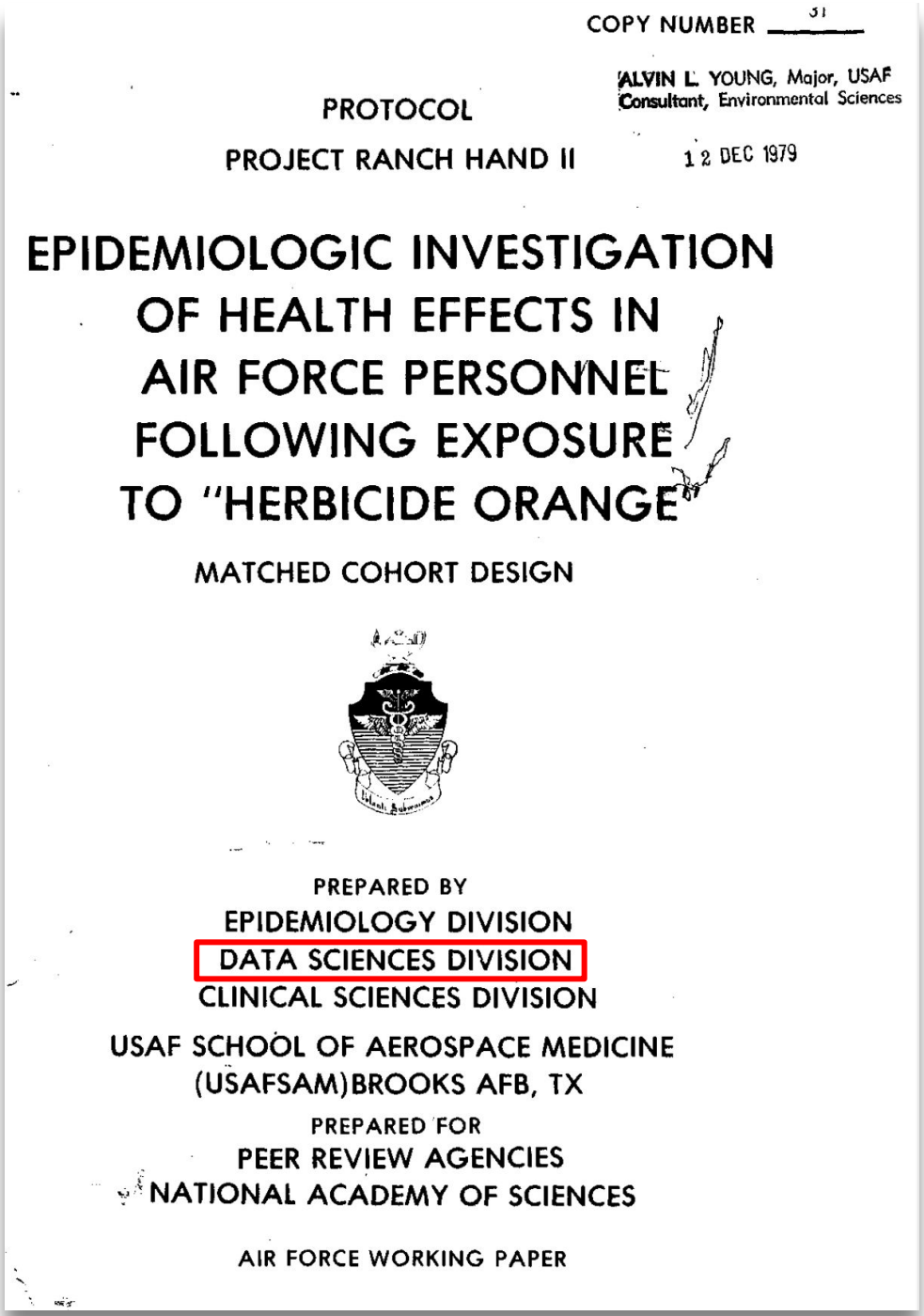
General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models

L. BREIMAN and W. S. MEISEL*

* L. Breiman is lecturer, Department of Mathematics, UCLA, and consultant, and W.S. Meisel is manager, both at Data Sciences Division, Technology Service Corporation, Santa Monica, CA. 90403. Research was sponsored by the Air Force Office of Scientific Research/AFSC, U.S. Air Force, under Contract No. F44620-71-C-0093. The authors wish to thank Mike Teener, who programmed and ran the simulation examples, and the referees for their reviewing work, which resulted in significant improvements.

© Journal of the American Statistical Association
June 1976, Volume 71, Number 354
Applications Section

USAF School of Aerospace Medicine (1979)



COPY NUMBER 31

ALVIN L. YOUNG, Major, USAF
Consultant, Environmental Sciences

PROTOCOL

PROJECT RANCH HAND II

12 DEC 1979

EPIDEMIOLOGIC INVESTIGATION OF HEALTH EFFECTS IN AIR FORCE PERSONNEL FOLLOWING EXPOSURE TO "HERBICIDE ORANGE"

MATCHED COHORT DESIGN

PREPARED BY

EPIDEMIOLOGY DIVISION

DATA SCIENCES DIVISION

CLINICAL SCIENCES DIVISION

USAF SCHOOL OF AEROSPACE MEDICINE
(USAFSAM) BROOKS AFB, TX

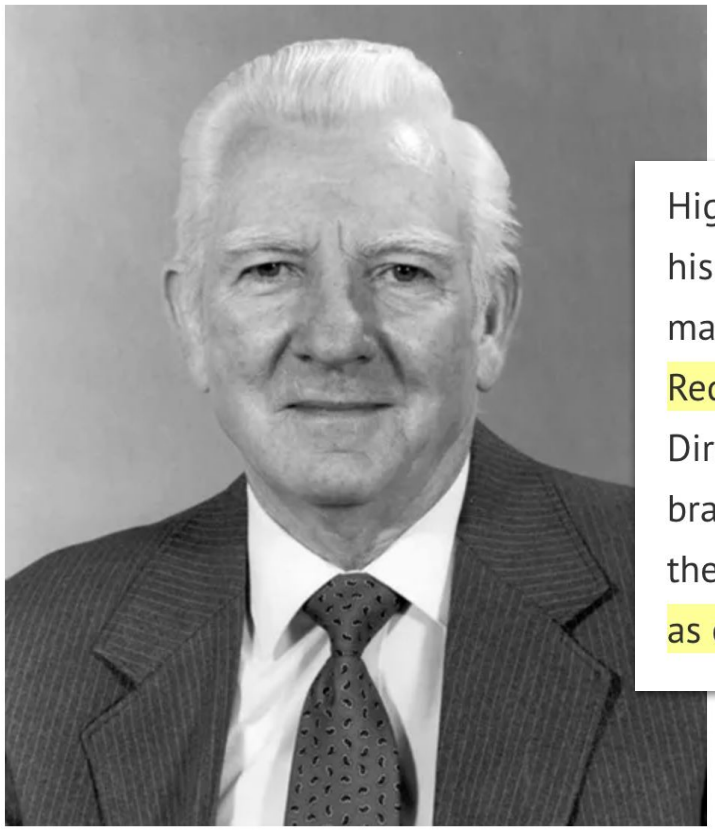
PREPARED FOR

PEER REVIEW AGENCIES
NATIONAL ACADEMY OF SCIENCES

AIR FORCE WORKING PAPER

White Sands Missile Range (1974)

Patrick Higgins
Chief of Data Sciences
Served 1950 – 1981
Inducted 1988



WHITE SANDS MISSILE RANGE MUSEUM
Atomic. Missile. Space.
Birthplace of American Ages.

Higgins came to White Sands in 1950 as a physical science aide. As his career progressed, he worked as a mathematician, supervisory mathematician and supervisory physical scientist in the Data Reduction Division of what is now the National Range Operations Directorate. In the late 1960s he served as chief of the support branch of the Analysis and Computation Division and later as chief of the Operations branch of the same division. He assumed the duties as chief of the Data Sciences Division in 1974.

White Sands Missile Range

The Data Sciences Division had responsibility for both real-time and post-test data acquisition and processing at White Sands. The real-time responsibility included the critical data (radar, telemetry, optics, etc.) acquisition, development of real-time algorithms, data processing, and display support for missile flight safety officers and project engineers - a Range Control Center operation.

Burkett, Ron. **2003**. "Burkett Announces His Retirement as Director of Museum." *A Newsletter for the White Sands Missile Range Historical Foundation*, 2003.

Context

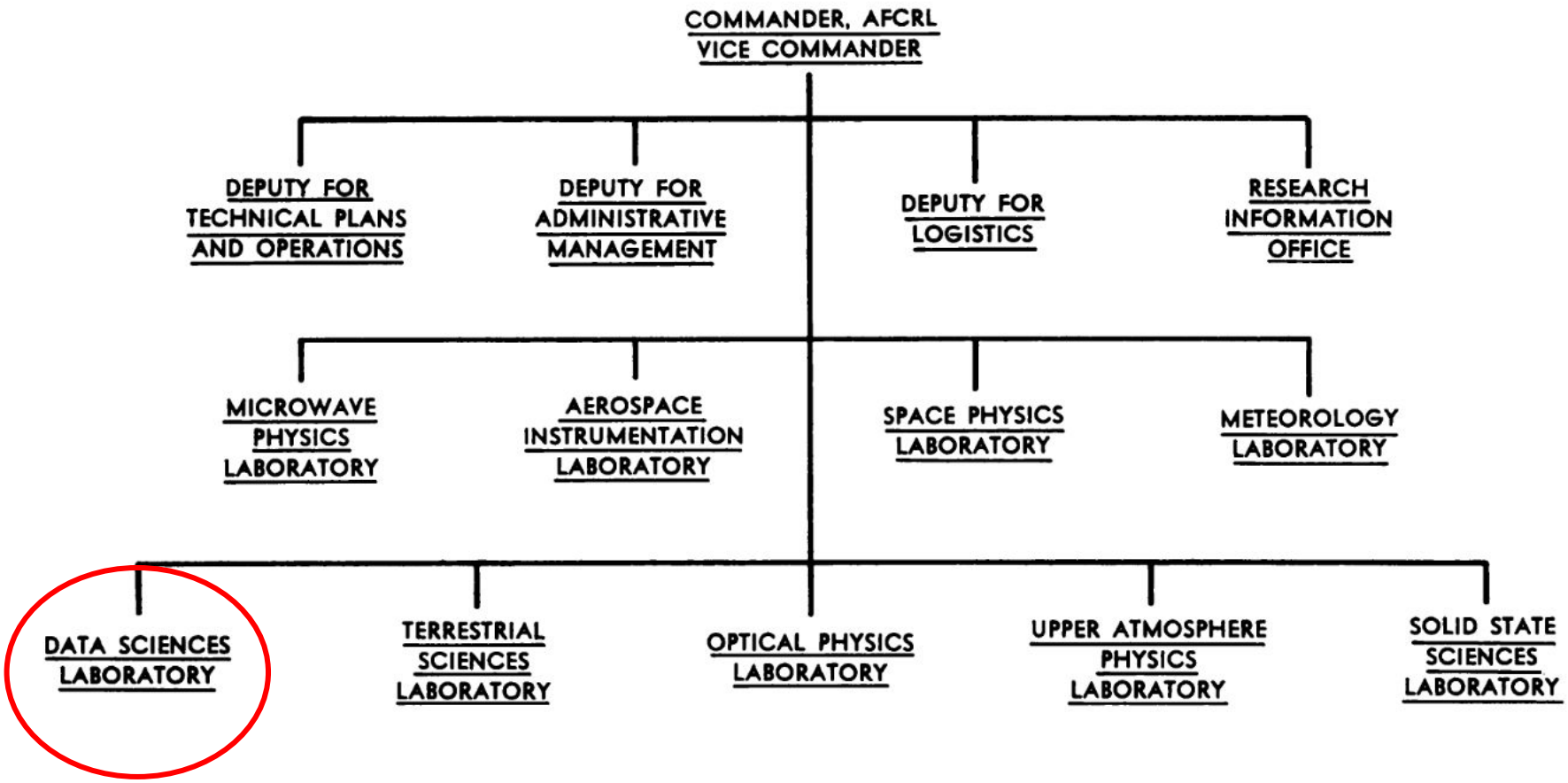
Air Force Cambridge Research Lab (AFCRL)

Established in **1945** as the **Cambridge Field Station**

> Created to hold onto the Harvard, MIT, and BU scientists and engineers who performed significant research on **radar** and **electronics** in **WWII**

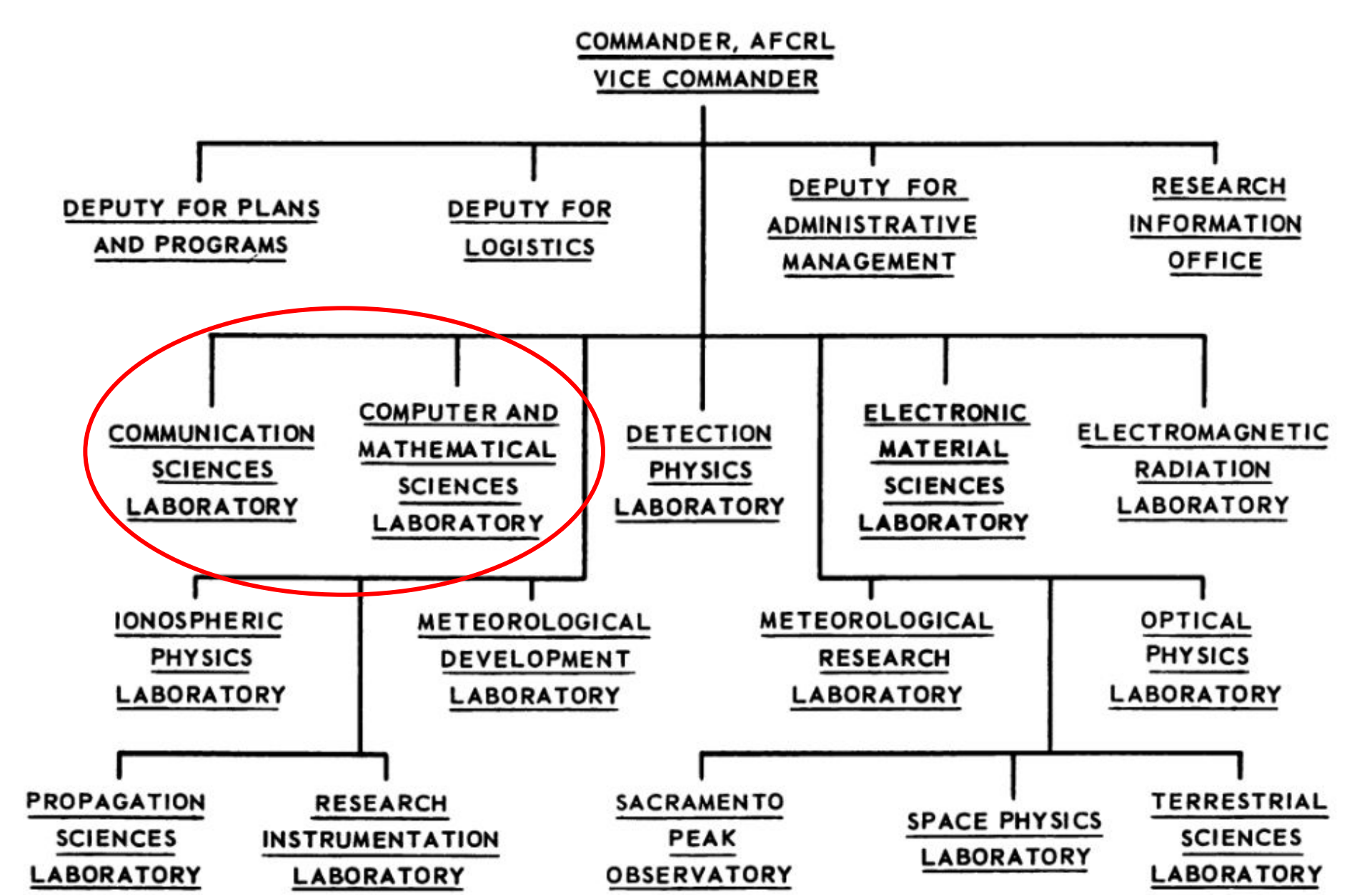
During the **1950s**, the lab focused on **Project Lincoln**

> PL led to the creation of the Semi-Automatic Ground Environment (**SAGE**), a **real-time command-and-control** system developed to counter to perceived threat of an airborne nuclear attack by the Soviet Union
>
> At the heart of the system was **a network of large computers** that **coordinated the data retrieved from radar** sites over phone lines and processed them to produce **a single unified image**—literally displayed on a monitor—of the airspace over a wide area



Org chart for the AFCRL for **1963**

Note the theme large-scale **non-point systems** — space, atmosphere, weather, earth . . .

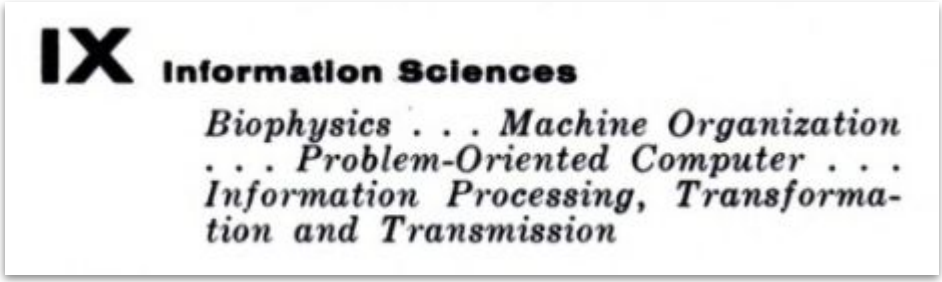


AFCRL org chart for **1962**

AFCRL and DSL

DSL was formed from the **Computer and Mathematical Sciences** Laboratory and the **Communications Sciences** Laboratory of the Electronics Research Directorate

These labs were essential to the construction of computational machinery at the heart of the **SAGE** project . . .



IX **Information Sciences**

Biophysics . . . Machine Organization . . . Problem-Oriented Computer . . . Information Processing, Transformation and Transmission

Electronics Research Directorate

The Electronics Research Directorate evolved from the Cambridge Field Station, established in 1945, which was staffed largely by scientists who had engaged in electronics research during World War II at the MIT Radiation Laboratory and at Harvard's Radio Research Laboratory. Many of the large command and control systems that are now an important part of the national defense program had their inception in projects conceived and carried out by this group of scientists in the late 1940's and early 50's. Research projects of the present Electronics Research Directorate are conducted and monitored by one of the seven following laboratories.

SAGE

SAGE (Semi-Automatic Ground Environment) was a networked system of computers, radars, and other elements

- Designed to detect Soviet bombers carrying nuclear weapons into North America in the 1950s
- Decommissioned equipment used in *Dr Strangelove*
- **Real-time command-and-control system**

Notable for many reasons

- More expensive than the **Manhattan** project
- Drove development of **numerous computer and network technologies**
- Birthed **systems engineering**
- Became a model for **business processes**

By almost any measure—scale, expense, technical complexity, or influence on future developments—the single most important computer project of the postwar decade was MIT's Whirlwind and its offspring, the SAGE computerized air defense system.

SAGE was the first large-scale, computerized command, control, and communications system. Although it was obsolete before it was completed, it unleashed a cascading wave of command-control projects from the late 1950s onwards, tied largely to nuclear early warning systems. These systems eventually formed the core of a worldwide satellite, sensor, and communications web that would allow global oversight and instantaneous military response. Enframing the globe, this web formed the technological infrastructure of closed-world politics.

Edwards, 1996, *The Closed World*, p. 75

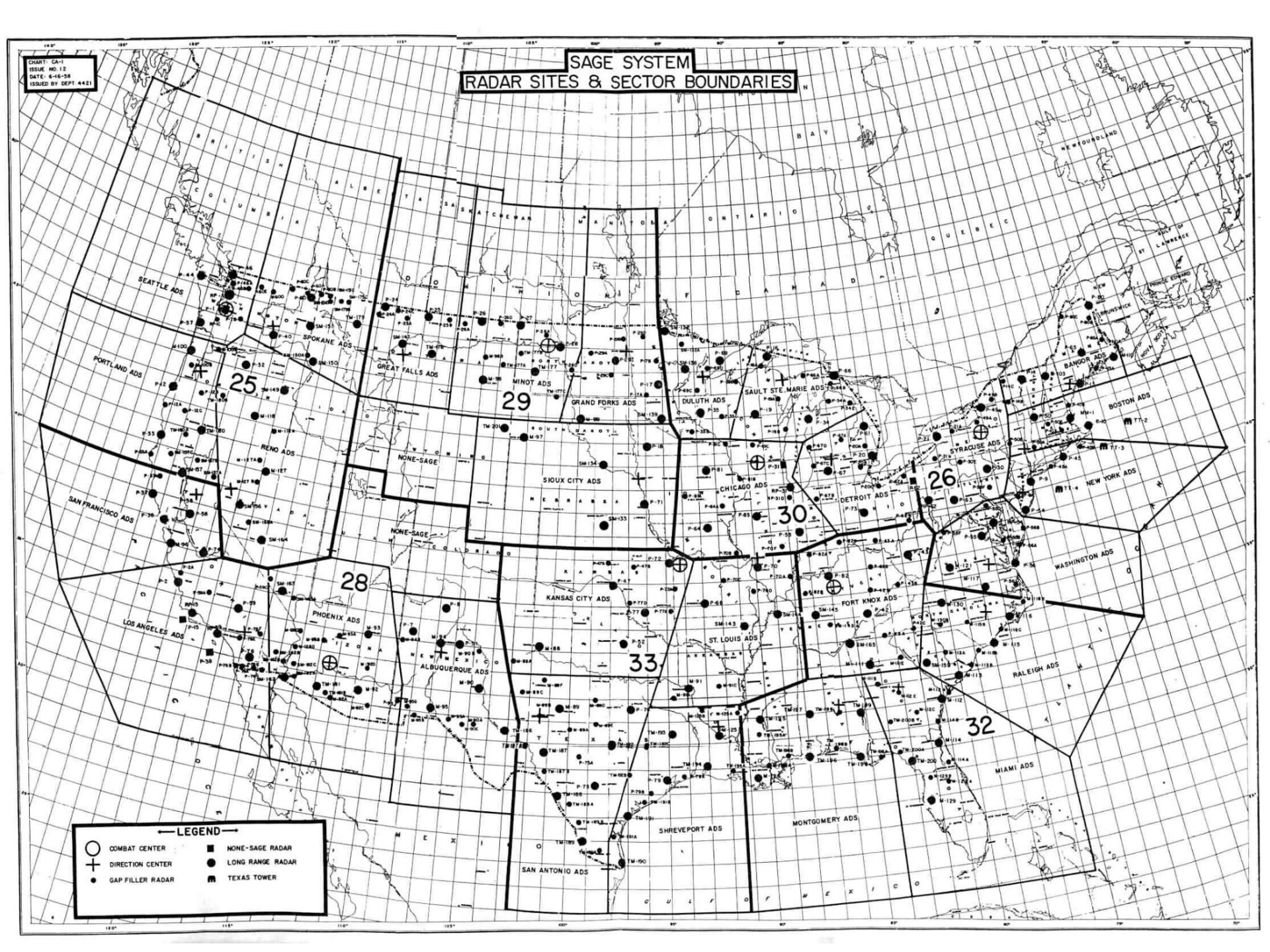


CHART: CA-1
ISSUE NO. 12
DATE- 6-16-58
ISSUED BY DEPT 4421
SAGE SYSTEM
RADAR SITES & SECTOR BOUNDARIES
SEATTLE ADS
SPOKANE ADS
PORTLAND ADS
GREAT FALLS ADS
MINOT ADS
GRAND FORKS ADS
DULUTH ADS
SAULT STE. MARIE ADS
BANGOR ADS
BOSTON ADS
NEW YORK ADS
SYRACUSE ADS
RENO ADS
SIOUX CITY ADS
CHICAGO ADS
DETROIT ADS
SAN FRANCISCO ADS
NONE-SAGE
KANSAS CITY ADS
FORT KNOX ADS
WASHINGTON ADS
LOS ANGELES ADS
PHOENIX ADS
ST. LOUIS ADS
ALBUQUERQUE ADS
RALEIGH ADS
MIAMI ADS
SHREVEPORT ADS
MONTGOMERY ADS
SAN ANTONIO ADS
25
29
30
26
28
33
32
LEGEND
COMBAT CENTER
DIRECTION CENTER
GAP FILLER RADAR
NONE-SAGE RADAR
LONG RANGE RADAR
TEXAS TOWER

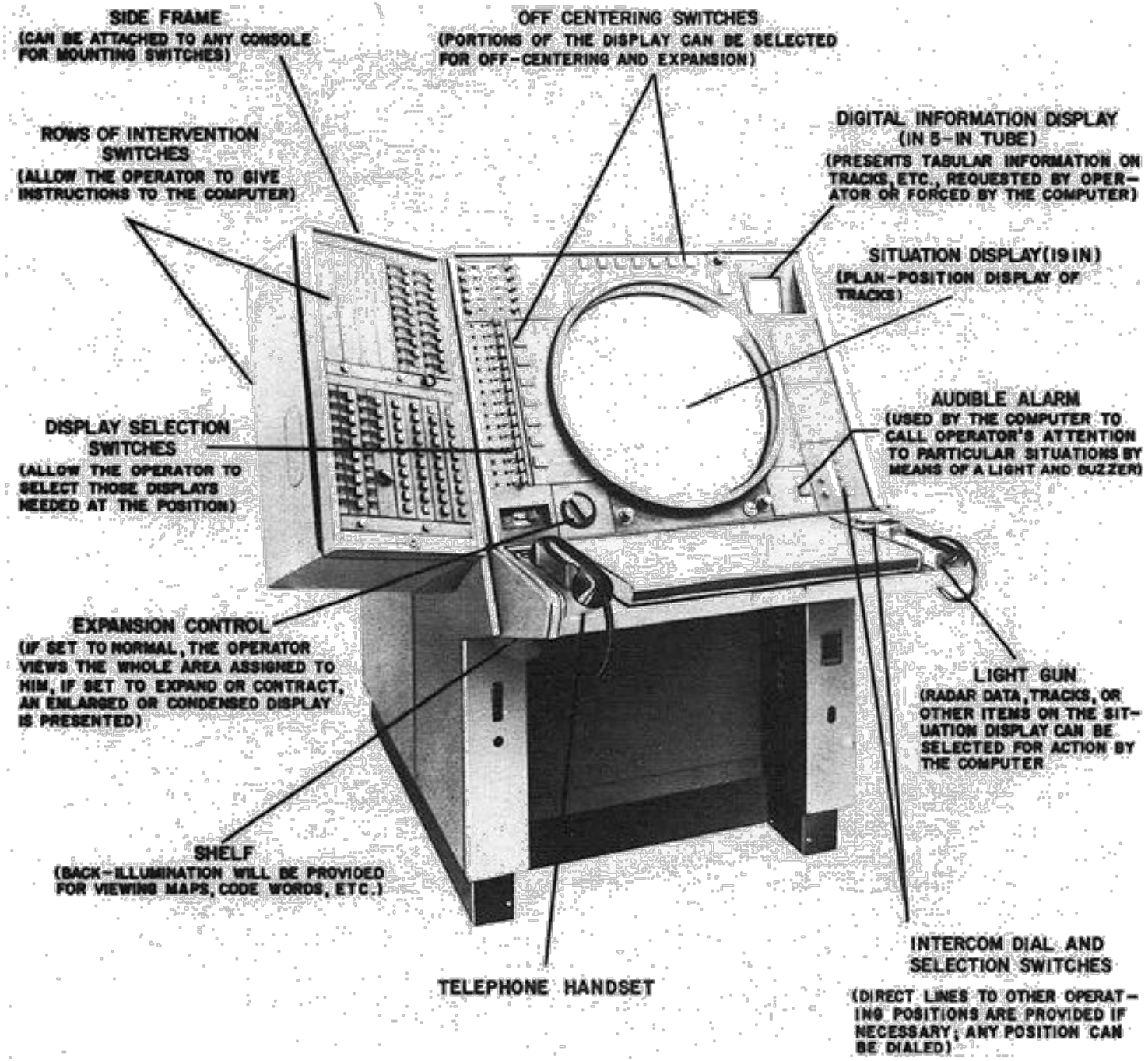
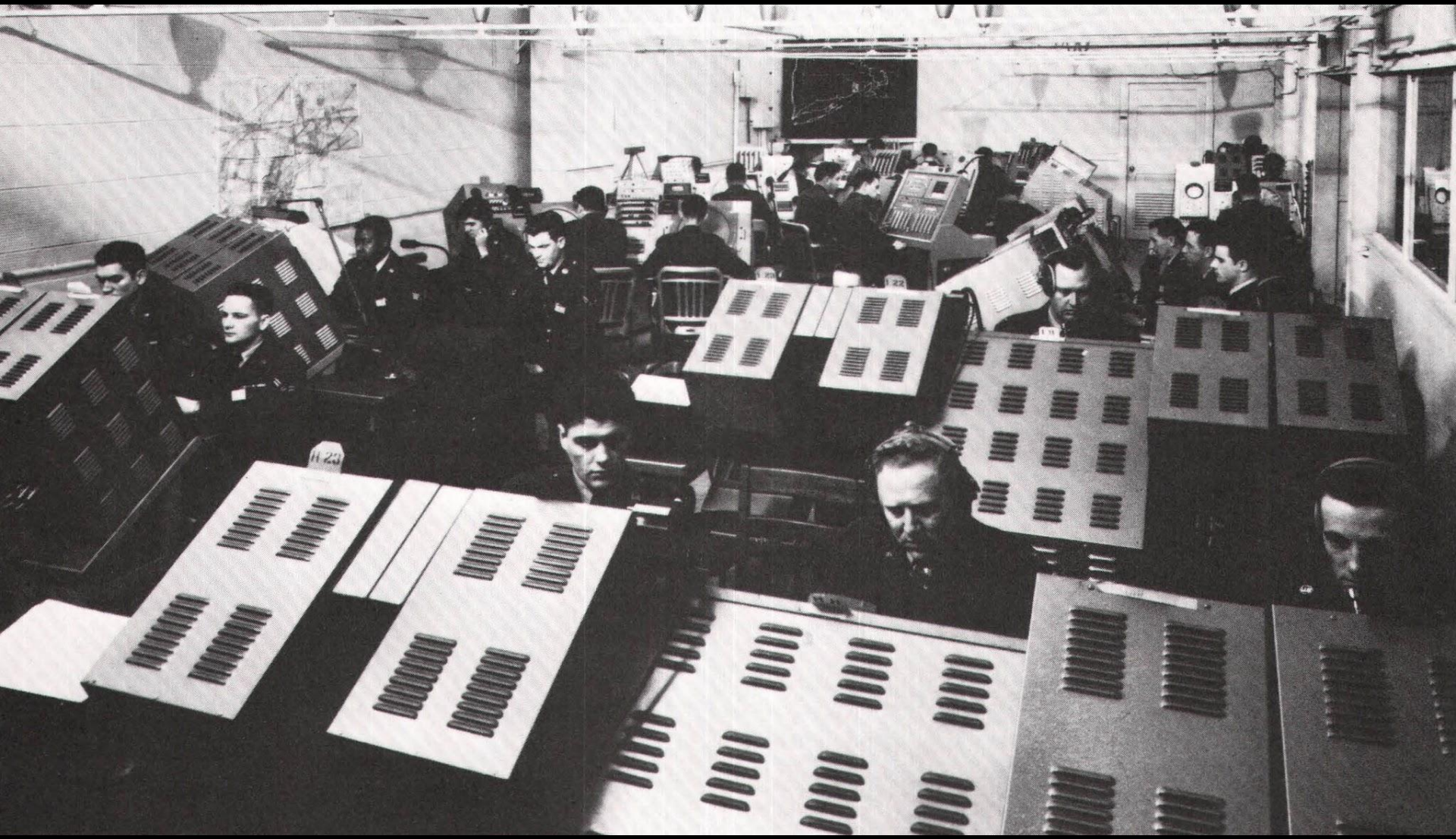
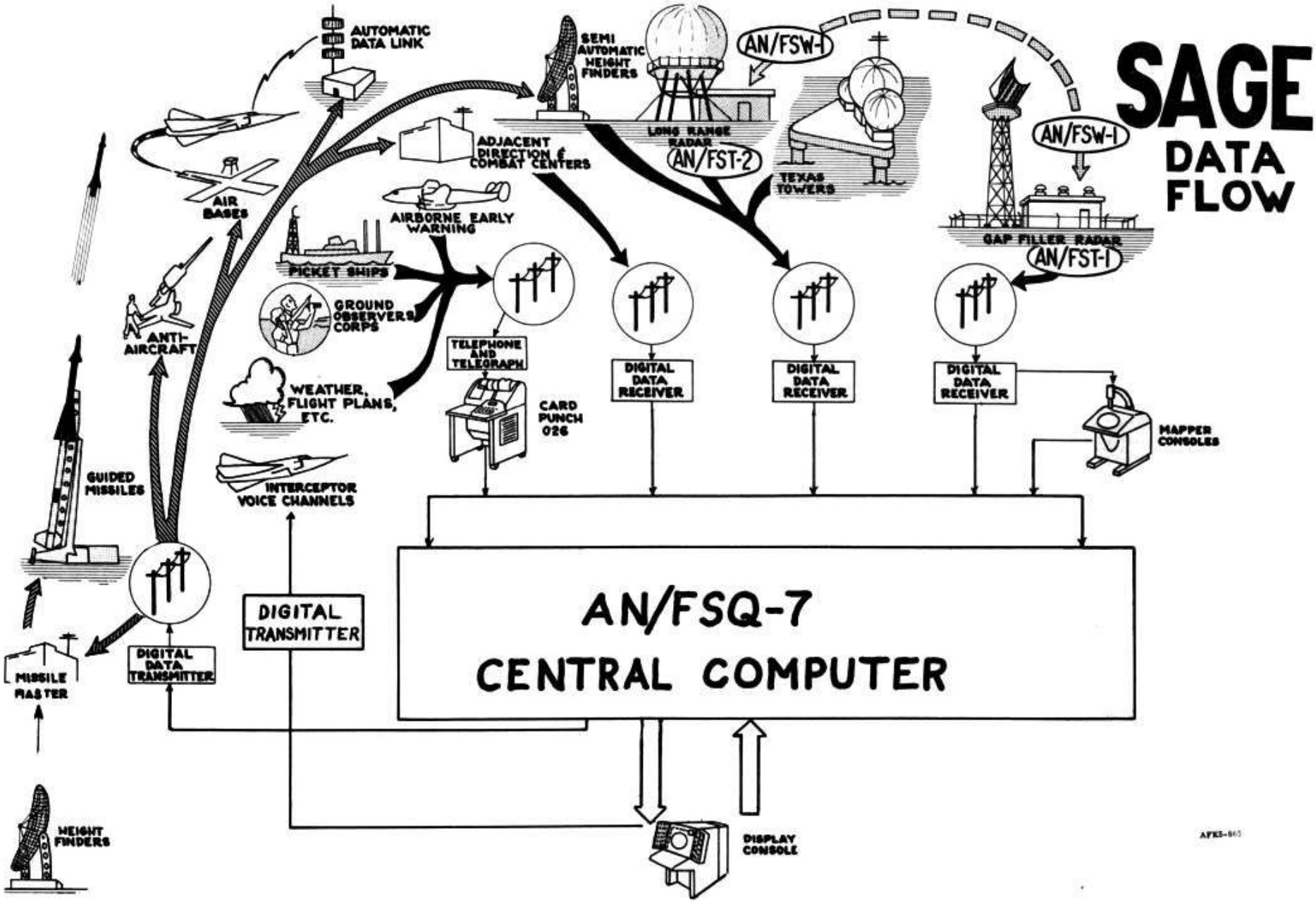


Figure 3–7. Facilities at a Typical Situation Display Console




SAGE
DATA
FLOW
AUTOMATIC
DATA LINK
SEMI
AUTOMATIC
HEIGHT
FINDERS
AN/FSW-1
LONG RANGE
RADAR
AN/FST-2
TEXAS
TOWERS
AN/FSW-1
GAP FILLER RADAR
AN/FST-1
ADJACENT
DIRECTION &
COMBAT CENTERS
AIR
BASES
AIRBORNE EARLY
WARNING
PICKET SHIPS
GROUND
OBSERVERS
CORPS
ANTI-
AIRCRAFT
WEATHER,
FLIGHT PLANS,
ETC.
TELEPHONE
AND
TELEGRAPH
CARD
PUNCH
026
DIGITAL
DATA
RECEIVER
DIGITAL
DATA
RECEIVER
DIGITAL
DATA
RECEIVER
MAPPER
CONSOLES
GUIDED
MISSILES
INTERCEPTOR
VOICE CHANNELS
AN/FSQ-7
CENTRAL COMPUTER
DIGITAL
TRANSMITTER
MISSILE
MASTER
DIGITAL
DATA
TRANSMITTER
HEIGHT
FINDERS
DISPLAY
CONSOLE
AFKS-865

General Topology of SAGE

A Real-Time Command-and-Control System

S1

S2

S3

S4

S5

SN

Aggregate and reduce

Analyze and Model

Report and Influence

R1

R2

R3

RN

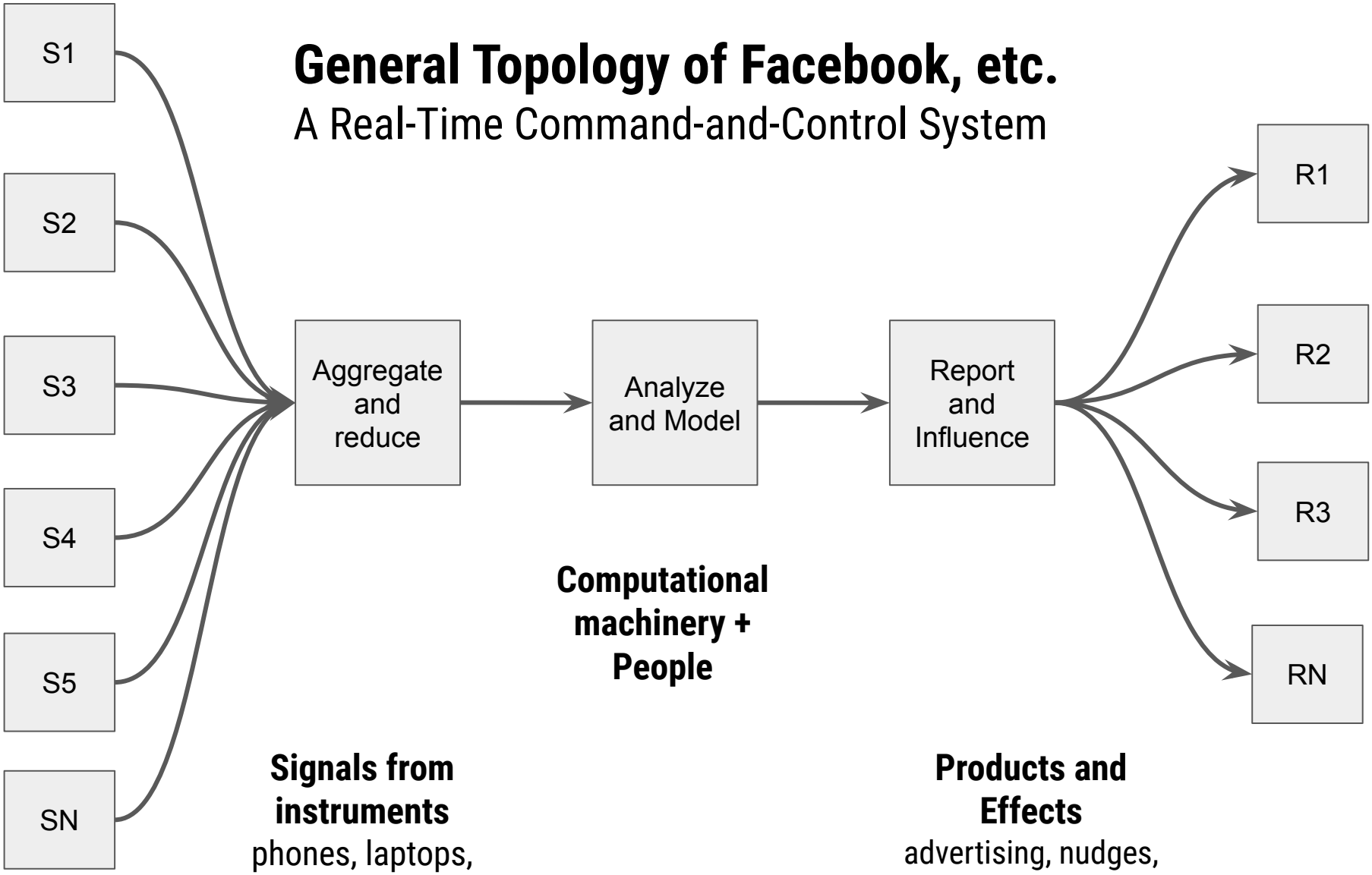
Computational machinery + People

Signals from instruments
radar, telemetry, audio, etc.

Products and Effects
aircraft, anti-aircraft, missiles

General Topology of Facebook, etc.

A Real-Time Command-and-Control System



What kind of knowledge?

So, **what kind of knowledge did the data science represent**?

> Not **computer science**, since it was focused on **particular kinds of data**
>
> Not **statistics**, since it involves **active construction of machinery**

One cluse is the expression "**data processing research**" from the first report

> An increasing amount of **data processing research** is aimed at the creation of machines or machine programs that incorporate features of deductive and inductive **reasoning**, **learning**, **adaptation**, **hypothesis formation** and **recognition**.
>
> This looks more like **machine learning**

The expression echoes an earlier one: "**data processing scientist" . . .**

1950s

Data processing scientists
and the work of **data reduction**

Data Processing Scientist • Expert familiar with current developments in data processing. Systems planning and practical computer experience desirable. Knowledge of present techniques in computer usage and available peripheral equipment necessary. Graduate degree in electrical engineering, physics, or applied mathematics.

Journal of the Society for Industrial and Applied Mathematics, Vol. 11, No. 2 (Jun., **1963**)

Data processing scientists were in demand in the 1950s and 60s

They "**devised techniques**" to apply computers to research problems

They were usually scientists

Data Processing Scientist

M.S. or equivalent in chemistry, or biochemistry, with extensive experience in information handling, to devise technique for application of electronic computer to research and development activities, Philadelphia location. Liberal benefit program. Send complete résumé. Box 254, SCIENCE.

Science, N.S., Vol. 126, No. 3270 (Aug. 30, **1957**), pp. 417-422

Rise of the Data Processing Scientist

The phrase emerges in the **1950s**

> Associated with the work of **data reduction** and data processing in the context of **scientific work**

Data reduction associated with the need manage the "**data deluge**" of **instrument-generated data**

> INSTRUMENTS: Radar, missile telemetry, satellites, wind tunnels, particle accelerators, etc.
>
> REDUCTION: Converting analog signal data into digital and to analytical form, i.e. the production of *representational data*
>
> "Data deluge" and "information explosion" gain currency to describe this situation

This **assemblage of instruments** are associated with the Cold War project of nuclear defense and large-scale systems of **real-time command-and-control**

General **Doolittle**, head of NACA, explains role of **data-processing scientist** to Congress in 1958:

The data processing function is much more complex than the mere production-line job of translating raw data into usable form. Each new research project must be reviewed to determine how the data will be obtained, what type and volume of calculations are required, and what modifications must be made to the recording instruments and data-processing apparatus to meet the requirements. It may even be necessary for the data-processing scientist to design and construct new equipment for a new type of problem. Some projects cannot be undertaken until the specific means of obtaining and handling the data have been worked out. In some research areas, on-line service to a data processing center saves considerable time by allowing the project engineer to obtain a spot check on the computed results while the facility is in operation. This permits him to make an immediate change in the test conditions to obtain the results that he wants.

Appropriations, United States Congress House. 1958. *Second Supplemental Appropriation Bill: 1958, Hearings ... 85th Congress, 2d Session*, p. 147.

The role was fundamental to the proposed **data reduction** center. This was not a data-entry and reporting job . . .

The role figures prominently in the **budget** for staff

The staff for operation of the data reduction center will comprise 104 personnel, as follows:

Data processing systems scientists	26
Machine programing mathematicians	51
Tabulating equipment operators	3
Computing equipment operators	14
Card punch operators	3
Secretary	1
Electronic instrument mechanic	7
Maintenance mechanic	1
Janitor	1

These numbers form a power distribution (Zipf's Law); the role has a rank of 2.

1	51
2	**26**
3	14
4	7
5	3
6	1

Christine Darden

Mary Jackson

Katherine Johnson

Dorothy Vaughan

Mathematician **Katherine Johnson** performed data reduction at **Langley**, the same place that **Doolittle** received funding to build a data reduction center

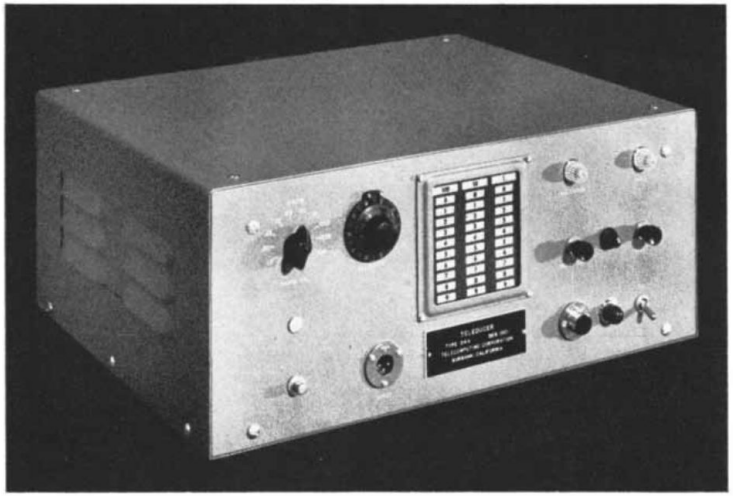
She learned that the NACA Langley Aeronautical Laboratory was hiring a group of African-American mathematicians with teaching experience to perform **mathematical calculations that transformed raw data that had been obtained using instrumentation into final engineering parameters.** She began her career at Langley in the segregated West Computing section in the summer of 1953 under the supervision of fellow West Virginian **Dorothy Vaughan. The pool of women mathematicians performing data reduction calculations were known as "computers."** Just two weeks into Katherine's tenure in the office, Dorothy Vaughan assigned her to a project in the Maneuver Loads Branch of the Flight Research Division, where her position soon became permanent. She spent the next four years analyzing data from flight tests, and worked on the investigation of a plane crash caused by an encounter with wake turbulence. She was assertive, asking to be included in editorial meetings (where no women had gone before). It was during this time that her husband James died of cancer in December 1956.

https://www.nasa.gov/feature/katherine-g-johnson

Again Telecomputing announces a new advance in automatic data reduction...

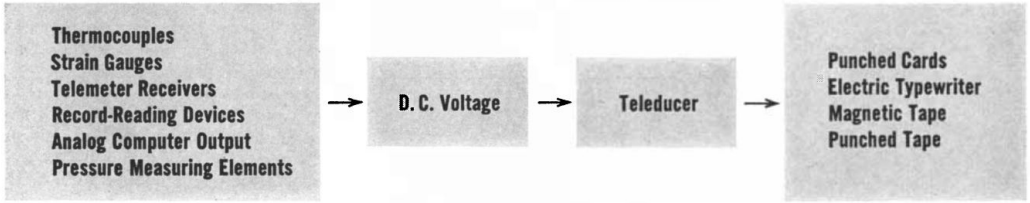
THE TELEDUCER

AUTOMATICALLY CONVERTS ANALOG VOLTAGE INTO DECIMAL DIGITS



It offers these important advantages:

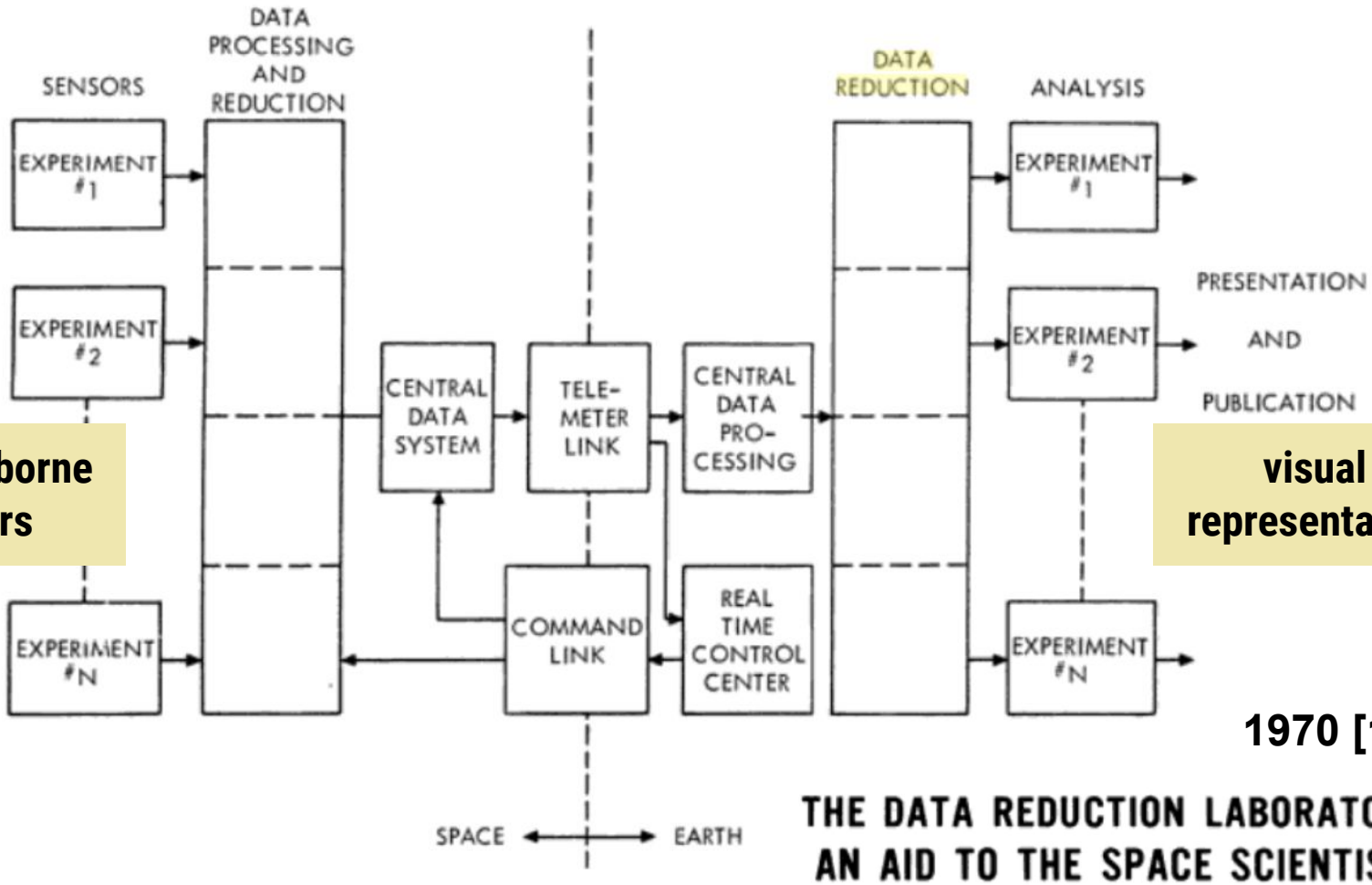
- Does not hunt or oscillate.
- Reads low voltage without D. C. amplification.
- Digitizes higher voltages by means of attenuators.
- Provides 0.1% accuracy (1,000 counts full scale).
- Requires only 0.8 second or less for balancing.
- Uses a simple bridge-balancing circuit.
- Relays digital output to punched cards, an electric typewriter, magnetic tape or punched tape.
- Provides for minimum full-scale input of 20 millivolts (20 microvolts per count).



Data reduction was also a process of converting **signals** into discrete **numbers**

It's how **instruments** produce **data** as used by a **computer**

Generalized space information system



satellite-borne sensors

visual representation

1970 [1969]

THE DATA REDUCTION LABORATORY: AN AID TO THE SPACE SCIENTIST*

Figure 1—Generalized space information system (Reference 2).

The Data Reduction Laboratory provides a means for rapid presentation of processed data to experimenters and other users, either in real time or from stored data. The laboratory uses a computer with a large storage capacity and associated display and output devices (Figure 2). The

Data Reduction

Referred to a **variety of methods** to make data manageable and intelligible

Discretization — converting continuous signals into discrete numbers
Parameterization — replace data points with formulae (regression, etc.)
Filtering — eliminating data or simplifying formulae
Compression — application of information theory
Visualization — reduced representations, 2D projections on ND data

Also, later (and today) includes:

Vectorization, database normalization, PCA, tSNE, etc. etc

The **data processing scientist** had to get **the computer** to do these things

Under conditions of "**data deluge**"

The work of data reduction was prompted by the **rise of scientific instruments** in the post-war period

These instruments produced a **surplus of data**

Here the author predicts "**great advances** in the techniques of data reduction" (1951)

Stalking the Guided Missile

New Instruments Track and Report the Performance of Long-Range Rockets

Dr. Dirk Reuyl and L. G. de Bey
Ballistic Research Laboratories, Aberdeen Proving Ground, Md.

In conclusion, it may be well to call attention to the field of data reduction which until now has lagged considerably behind the development of field instrumentation. It would seem safe to predict great advances in the techniques of data reduction including such improvements as refined tracking controls and film-measuring devices.

Ordnance, Vol. 36, No. 188 (September–October, **1951**), pp. 237-241.

Data Reduction and Data Impedance

The work of data reduction emerges is a solution to **a problem that emerges** in this situation

> The ever-present disproportion between the **surplus data** produced by **instruments** and the need to **model** that data to guide decision-making
>
> Surplus data is signified by the expression "**data deluge**" (which later becomes "**big data**")

I call this the problem of **data impedance**

> **Data impedance** motivates the development of machinery, methods, and personnel . . .

The Paradigm

Data science was invented to manage and leverage **data impedance**

Data scientists applied the emerging field of **artificial intelligence** (AI) to this problem

> The applied and pushed the development of **pattern recognition**, **classification**, **machine learning**, etc.

But, to apply AI computational machinery had to "**understand**" data

> In a sense, **data had to be invented**

The **data deluge**, information flood, or whatever you choose to call it, is hard to measure in common terms. An Observatory-class satellite may spew out more than 10^{11} data words during its lifetime, the equivalent of several hundred thousand books. **Data-rate projections, summed for all scientific satellites, prophesy hundreds of millions of words per day** descending on Earth-based data processing centers. **These data must be translated to a common language**, or at least a language **widely understood by computers** (viz, PCM), **then edited, cataloged, indexed, archived, and made available to the scientific community upon demand.** Obviously, the vaunted information explosion is not only confined to technical reports alone, but also to the data from which they are written. In fact, the quantity of raw data generally exceeds the length of the resulting paper by many orders of magnitude (Corliss 1967: 157).

Corliss, William R. **1967**. *Scientific Satellites*. Scientific and Technical Information Division, National Aeronautics and Space Administration.

Why AI?

Another division of the Laboratory's effort is that of using existing computers for more sophisticated tasks. One such task is the automatic identification, recognition, and classification of sensor inputs of all kinds. The inputs may be of great variety—photographs, human speech, radar, or infrared signals. Techniques for the real-time extraction of meaningful data—signatures otherwise buried in the flood of data from sensors—are of fundamental Air Force importance.

In 1980 contingency planning in a crisis can be performed in near-real time and alternatives quickly analyzed to assist the decision-makers. Question: Does this guarantee better decisions?

Continuous surveillance of global air and ocean traffic by satellites will be possible and even large scale movement patterns on land may be detectable.

The critical uncertainty lies in the magnitude of advances in pattern recognition. We want this for near-automatic analysis of the data so that the output gives us only anomalies. Otherwise we will be faced with a data deluge. Another major problem is the presentation of the information to the decisionmakers.

With more of our "nervous system" in space, its defense may become a valid mission in the 1980's. Passive defense by redundancy, increased power levels, and more distant orbits would be the choice over active defense in view of both technical problems and the 1966 space treaty.

Data deluge related to AI . . . AI viewed as solution to problem of DD.

"Strategy and Science: Toward a National Security Policy for the 1970's." March 11-12, 1969. In *Hearings of United States Congress, House Committee on Foreign Affairs*, 1969.

As the infrastructure of impedance became
generalized
(via the Internet)

and **evolved**
(evolution and growth of databases),

so too did the practice of **data science** . . .

Role of Data Scientist

Data Scientist

Marine Information and Advisory Service

Birkenhead, Merseyside **Up to £10 500**

The Institute of Oceanographic Sciences Marine Information and Advisory Service is the UK's national oceanographic data centre. It is responsible for providing an up-to-date archive of high-quality data for the use of industry research workers and government departments, and has extensive involvement in international oceanographic data exchange.

There is currently an opportunity at Bidston Observatory in Birkenhead to join a small team developing and operating the databank on a Honeywell level 66 computer. The work is varied and the successful candidate will be responsible for preparing and screening physical oceanographic data prior to banking, liaising with scientists collecting data at sea, and assisting in the servicing of customer requests.

Candidates should ideally have a degree in Mathematics, Physics or Earth Sciences, an interest in environmental sciences and postgraduate experience in the computer processing of environmental data, including Fortran programming.

Starting salary if you have two or more years' postgraduate experience will be in the range £7788–£10 541. Otherwise it will be in the range £6190–£8561, depending on age, qualifications and experience.

For further details and an application form contact: Mr T. E. Dugdale, Institute of Oceanographic Sciences, Bidston Observatory, Birkenhead, Wirral, Merseyside L43 7RA. Tel 051-653 8633.

Closing date for completed application forms: 20 February, 1986.

Natural Environment Research Council

New Scientist 6 February 1986

SEVERN-TRENT WATER AUTHORITY
TAME DIVISION

Senior Scientific Assistant

(Data Processing)

£7722-£8553 pa

Based in Central Birmingham, the successful applicant will work in the data information section of the divisional scientist's department and duties will include:

I Operating and maintaining programs and suites of programs currently in use.

II Producing and supporting new programs and packages as required by the data scientist.

III Providing general mathematical and statistical support to the data section and the scientists department generally.

The post involves the preparation of data and production of reports using clear and unambiguous presentations, and it is necessary to meet the standards required both within the section and the data processing department in general as regards documentation.

Experience of programming in basic and/or Fortran is essential, preferably through the use of mini computers, although experience of M.A.C. mainframe systems would be advantageous.

The successful applicant must be able to communicate clearly and work closely with other members of staff, and should possess a good degree in a mathematics or scientific discipline and have relevant experiecne in an appropriate field.

Application forms quoting reference I.A.850 are obtainable from the Personnel Office, Severn-Trent Water Authority, Tame Division, Tame House, 156/170 Newhall Street, Birmingham B3 1SE. Telephone 021 233 1616 ext 2190.

Closing date for the return of completed application forms: 8 November, 1982.

This post is open to both men and women.

New Scientist 28 October 1982

From a 2008 piece in the London *Times* that quoted "Nathan Cunningham, 36, data scientist, **British Antarctic Survey**":

When I am on the ship I am part of a team of scientists collecting data about everything from the biomass in the ocean to the weather patterns. … Our monitoring equipment is always on and sends us 180 pieces of information every second. *My role is to make sure that each person can find the exact data that they want among all this, so I write programs to help them to do this*. Another one of my field responsibilities is getting the information that we collect back to Cambridge via satellite link so that other researchers can use the data (Chynoweth 2008; emphasis added).

Hammeracher at Facebook in 2008

When Facebook opened registration to all users, the user population grew at disproportionately rapid rates in some countries. At the time, however, we were not able to perform granular analyses of clickstream data broken out by country. Once our Hadoop cluster was up, we were able to reconstruct how Facebook had grown rapidly in places such as Canada and Norway by loading all of our historical access logs into Hadoop and writing a few simple MapReduce jobs.

At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, we created the role of "Data Scientist."

Hammerbacher, Jeff. 2009. "Information Platforms and the Rise of the Data Scientist." In *Beautiful Data: The Stories Behind Elegant Data Solutions*, 73–84. O'Reilly Media Sebastopol, CA.

Outside of industry, I've found that grad students in many scientific domains are playing the role of the Data Scientist. One of our hires for the Facebook Data team came from a bioinformatics lab where he was building data pipelines and performing offline data analysis of a similar kind. The well-known Large Hadron Collider at CERN generates reams of data that are collected and pored over by graduate students looking for breakthroughs (Hammerbacher 2009: 84).

What is Data Science, then?

It was born from the specific need to get computers to transform instrument data into actionable representations

It contributed significantly to the development of programming languages, databases, and machine learning

- Databases both in form and content

When the field was inflected in 2008, it retained this fundamental role, but in the context of business

- A delayed effect of the commercialization of the Internet

What is Data Science, then?

Most important, data science has been **misrecognized** throughout its career for its **liminal status**

> Because it is **both and neither** computing and statistics it gets misclassified as one of the other
>
> Or in terms of a **simple relationship** — data science as support for computational efficiencies in statistics (Donoho, etc.)

It is much more than that

> At its origin, both historically and structurally, data science represents **a way of thinking with technology where the primary language of that thought is data**

Rise of the Data Scientist

FLOWINGDATA

June 4, 2009

Topic

Design, Statistics

As we've all read by now, Google's chief economist Hal Varian commented in January that the next sexy job in the next 10 years would be statisticians. Obviously, I whole-heartedly agree. Heck, I'd go a step further and say they're sexy now – mentally *and* physically.

Photo by majamarko

However, if you went on to read the rest of Varian's interview, you'd know that by *statisticians*, he actually meant it as a general title for someone who is able to extract information from large datasets and then present something of use to non-data experts.

Sexy Skills of Data Geeks

As a follow up to Varian's now-popular quote among data fans, Michael Driscoll of Dataspora, discusses the three sexy skills of data geeks. I won't rehash the post, but here are the three skills that Michael highlights:

1. Statistics – traditional analysis you're used to thinking about
2. Data Munging – parsing, scraping, and formatting data
3. Visualization – graphs, tools, etc.

When **Nathan Yau** effectively launched the term data science in 2009, he did so by emphasizing its connection to **design** . . .

These skills actually fit tightly with Ben Fry's dissertation on Computational Information Design (2004). However, Fry takes it a step further and argues for an entirely new field that combines the skills and talents from often disjoint areas of expertise:

1. **Computer Science** – acquire and parse data
2. **Mathematics, Statistics, & Data Mining** – filter and mine
3. **Graphic Design** – represent and refine
4. **Infovis and Human-Computer Interaction (HCI)** – interaction

Design is the thing.