Topic Models for Understanding History

David M. Blei
Columbia University

joint work with Allison Chaney, Hanna Wallach, and Matt Connelly
▶ ORGANIZE
▶ VISUALIZE
▶ SUMMARIZE
▶ SEARCH
▶ PREDICT
▶ UNDERSTAND
TOPIC MODELING

1. **Discover** the thematic structure
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, ...
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Game</td>
<td>Life</td>
<td>Film</td>
<td>Book</td>
<td>Wine</td>
</tr>
<tr>
<td>Season</td>
<td>Know</td>
<td>Movie</td>
<td>Life</td>
<td>Street</td>
</tr>
<tr>
<td>Team</td>
<td>School</td>
<td>Show</td>
<td>Books</td>
<td>Hotel</td>
</tr>
<tr>
<td>Coach</td>
<td>Street</td>
<td>Life</td>
<td>Novel</td>
<td>House</td>
</tr>
<tr>
<td>Play</td>
<td>Man</td>
<td>Television</td>
<td>Story</td>
<td>Room</td>
</tr>
<tr>
<td>Points</td>
<td>Family</td>
<td>Films</td>
<td>Man</td>
<td>Night</td>
</tr>
<tr>
<td>Games</td>
<td>Says</td>
<td>Director</td>
<td>House</td>
<td>Place</td>
</tr>
<tr>
<td>Giants</td>
<td>House</td>
<td>War</td>
<td>Author</td>
<td>Restaurant</td>
</tr>
<tr>
<td>Second</td>
<td>Children</td>
<td>Story</td>
<td>House</td>
<td>Park</td>
</tr>
<tr>
<td>Players</td>
<td>Night</td>
<td>Says</td>
<td>War</td>
<td>Garden</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bush</td>
<td>Building</td>
<td>Won</td>
<td>Yankees</td>
<td>Government</td>
</tr>
<tr>
<td>Campaign</td>
<td>Street</td>
<td>Team</td>
<td>Game</td>
<td>War</td>
</tr>
<tr>
<td>Clinton</td>
<td>Square</td>
<td>Second</td>
<td>Mets</td>
<td>Military</td>
</tr>
<tr>
<td>Republican</td>
<td>Housing</td>
<td>Race</td>
<td>Season</td>
<td>Officials</td>
</tr>
<tr>
<td>House</td>
<td>House</td>
<td>Round</td>
<td>Run</td>
<td>Iraq</td>
</tr>
<tr>
<td>Party</td>
<td>Buildings</td>
<td>Cup</td>
<td>League</td>
<td>Forces</td>
</tr>
<tr>
<td>Democratic</td>
<td>Development</td>
<td>Open</td>
<td>Baseball</td>
<td>Iraqi</td>
</tr>
<tr>
<td>Political</td>
<td>Space</td>
<td>Game</td>
<td>Team</td>
<td>Army</td>
</tr>
<tr>
<td>Democrats</td>
<td>Percent</td>
<td>Play</td>
<td>Games</td>
<td>Troops</td>
</tr>
<tr>
<td>Senator</td>
<td>Real</td>
<td>Win</td>
<td>Hit</td>
<td>Soldiers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>Stock</td>
<td>Church</td>
<td>Art</td>
<td>Police</td>
</tr>
<tr>
<td>School</td>
<td>Percent</td>
<td>War</td>
<td>Museum</td>
<td>Yesterday</td>
</tr>
<tr>
<td>Women</td>
<td>Companies</td>
<td>Women</td>
<td>Show</td>
<td>Man</td>
</tr>
<tr>
<td>Family</td>
<td>Fund</td>
<td>Life</td>
<td>Gallery</td>
<td>Officer</td>
</tr>
<tr>
<td>Parents</td>
<td>Market</td>
<td>Black</td>
<td>Works</td>
<td>Officers</td>
</tr>
<tr>
<td>Child</td>
<td>Bank</td>
<td>Political</td>
<td>Artists</td>
<td>Case</td>
</tr>
<tr>
<td>Life</td>
<td>Investors</td>
<td>Catholic</td>
<td>Street</td>
<td>Found</td>
</tr>
<tr>
<td>Says</td>
<td>Funds</td>
<td>Government</td>
<td>Artist</td>
<td>Charged</td>
</tr>
<tr>
<td>Help</td>
<td>Financial</td>
<td>Jewish</td>
<td>Paintings</td>
<td>Street</td>
</tr>
<tr>
<td>Mother</td>
<td>Business</td>
<td>Pope</td>
<td>Exhibition</td>
<td>Shot</td>
</tr>
</tbody>
</table>
Figure B.4: In a traditional ideal point model, lawmakers' ideal points are static. In the issue-adjusted ideal point model, lawmakers' ideal points change when they vote on certain issues, such as taxation (top panel) and health (bottom panel). A line segment connects select lawmakers' ideal points (top row of each panel) to their issue-adjusted ideal points (bottom row of each panel). Unlabeled lawmakers are illustrated by the remaining, faint line segments. We have colored Democrats blue and Republicans red.
Historians want to identify important events from primary sources.

Example: Embassies send cables to each other during the 1970s

Goal: Use topic models to discover events in this data set
This talk

1. Introduction to topic modeling
2. Topic models for understanding history
3. The bigger picture: Using probability models to solve problems with data
Introduction to Topic Modeling
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Documents exhibit multiple topics.
 Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Anderson, a professor at the University in St. Paul, Minn. Arrived at the 800 number, but coming up with a consensus answer may be more than just a matter of numbers, he notes, particularly as more and more genomes are completed, mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12

SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation
Latent Dirichlet Allocation
LDA as a graphical model

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.
**LDA as a graphical model**

- Encodes independence assumptions about the variables
- Defines a factorization of the joint probability distribution
- Connects to algorithms for computing with data
The joint defines a posterior, $p(\theta, z, \beta | w)$.

From a collection of documents, infer

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions $\theta_d$
- Per-corpus topic distributions $\beta_k$

Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.
Mean field variational methods (Blei et al., 2001, 2003)
Expectation propagation (Minka and Lafferty, 2002)
Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
Collapsed variational inference (Teh et al., 2006)
Stochastic inference (Hoffman et al., 2010, 2013; Mimno et al., 2012)
Factorization inference (Arora et al., 2012; Anandkumar et al., 2012)
- LDA in R  [https://cran.r-project.org/web/packages/lda/]
- GenSim   [https://radimrehurek.com/gensim]
- Mallet    [http://mallet.cs.umass.edu]
- Vowpal Wabbit    [http://hunch.net/~vw/]
- Apache Spark   [http://spark.apache.org/]
- SciKit Learn   [http://scikit-learn.org/]
Data: The OCR’ed collection of *Science* from 1990–2000
- 17K documents
- 11M words
- 20K unique terms (stop words and rare words removed)

Model: 100-topic LDA model using variational inference.
**Seeking Life’s Bare (Genetic) Necessities**

**COLD SPRING HARBOR, NEW YORK—** How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Sir Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcade Mushgian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

---


**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.
<table>
<thead>
<tr>
<th>human</th>
<th>evolution</th>
<th>disease</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>information</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>data</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>computers</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>system</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>network</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>systems</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>model</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infectious</td>
<td>parallel</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>methods</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>networks</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>software</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td>new</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>tuberculosis</td>
<td>simulations</td>
</tr>
</tbody>
</table>
Game
Season
Team
Coach
Play
Points
Games
Giants
Second
Players

Life
Know
School
Street
Man
Family
Says
House
Children
Night

Film
Movie
Show
Life
Television
Films
Director
Man
Story
Says

Book
Life
Books
Novel
Story
Man
Author
House
War
Children

Bush
Campaign
Clinton
Republican
House
Party
Democratic
Political
Democrats
Senator

Building
Street
Square
Housing
House
Buildings
Development
Space
Percent
Real

Won
Team
Second
Race
Round
Cup
Open
Game
Play
Win

Yankees
Game
Mets
Season
Run
League
Baseball
Team
Games
Hit

Government
War
Military
Officials
Iraq
Forces
Iraqi
Army
Troops
Soldiers

Children
School
Women
Family
Parents
Child
Life
Says
Help
Mother

Stock
Percent
Companies
Fund
Market
Bank
Investors
Funds
Financial
Business

Church
War
Women
Life
Black
Political
Catholic
Government
Jewish
Pope

Art
Museum
Show
Gallery
Works
Artists
Street
Artist
Paintings
Exhibition

Police
Yesterday
Man
Officer
Officers
Case
Found
Charged
Street
Shot
How does LDA “work”?

- LDA trades off two goals.
  1. In each document, allocate its words to few topics.
  2. In each topic, assign high probability to few terms.

- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document’s words, it must assign many topics to it.

- Trading off these goals finds groups of tightly co-occurring words.
Organizing and finding patterns in text is important in the sciences, humanities, industry, and culture.

LDA is a simple building block that enables many applications. Topic modeling is an active field of research.

Algorithmic improvements let us fit models to massive data.
Case study in text analysis with probability models

Topic modeling research
- develops new models.
- develops new inference algorithms.
- develops new applications, visualizations, tools.
Topic Models for Understanding History
Historians want to identify important events from primary sources.

Example: Embassies send cables to each other during the 1970s

Goal: Use topic models to find events in this data set
Topic models (by themselves) are a start. But they don’t identify events.
- Embassies typically discuss their *usual business*.
- When a cable is about an **event**:
  - It diverges from the usual business of the sender
  - Multiple embassies discuss it
- *Usual business* is framed in terms of topics; **events** are framed in terms of words.
Hidden variables

- Topics
- Event description (per week)
- Topic description (per entity)
- Topic strength, event strength (per cable)

Observed variables

- Cables (per-week, per-entity)
To find events:

- Calculate the posterior of the hidden variables given the observed variables
- Examine cables where the event strength is high

2. THE FRENCH PASSED A SECOND MESSAGE, THIS ONE FROM AMIN: AMIN TOLD THE FRENCH AMBASSADOR THAT HE EXPECTS ALL FOUR COUNTRIES TO COMMUNICATE TO HIM THE FLIGHT NUMBERS AND ETA OF ALL AIRCRAFT BRINGING PRISONERS TO UGANDA BEFORE THE END OF THE ULTIMATUM. (HE DID NOT SPECIFY AN HOUR.)

3. THE FRG CRISIS CENTER TOLD US THE FRG IS CONSULTING WITH THE OTHER GOVERNMENTS INVOLVED AT THE HIGHEST LEVEL. IT HAS NOT RPT NOT REACHED A DECISION ON RELEASE OF PRISONERS. HILLENBRAND
1. ACCORDING TO AS YET PROVISIONAL INFORMATION OBTAINED FROM IDF SPOKESMAN AND PASSENGERS, THERE WERE FOUR AMERICAN CITIZENS AMONG PASSENGERS FROM HIJACKED AIR FRANCE FLIGHT BROUGHT TO ISRAEL ON JULY 4. THEY ARE GEORGE AND RENE GARFUNKEL OF NEW YORK CITY; MRS. JANETTE ALMOG (HUSBAND ESRA ALMOG IS ISRAELI CITIZEN) OF MADISON WISCONSIN; AND MOSHE PERES OF NEW HAVEN CONNECTICUT. ALL RETURNED PASSENGERS ARE REPORTED WELL IN TEL AVIV. THOSE IN TRANSIT ARE BEING HOUSED AT PLAZA HOTEL IN TEL AVIV.

2. PASSENGERS KILLED DURING LIBERATION WERE JEAN-JACQUES MIMOUNI REPORTEDLY OF FRENCH NATIONALITY AND MRS IDA BOROCHOWITZ, AN ISRAELI OF RUSSIAN ORIGIN. ONE ISRAELI SOLDIER DIES IN FIGHT. NINE PERSONS REQUIRING MEDICAL CARE WERE LEFT IN NAIROBI DURING BRIEF STOP-OVER.

3. ACCORDING TO THE GARFUNKELS, FRENCH PILOT SAID THAT HIJACKERS ENTERED EMBARKATION AREA WITH SIX PACKAGES CLAIMING THAT THEY CONTAINED CANDY. AT TIME OF THEIR ENTRY ELECTRICITY ALLEGEDLY WENT OUT AND STOPPED SCREENING DEVICES FROM WORKING. RATHER THAN DELAY THE PLANE FOR EXAMINATION OF PACKAGES, THE SIX WERE HURRIED ON BOARD.

4. GARFUNKELS ARE LEAVING JULY 5 ON EL AL...
The way it really happened - The tense, action-packed story of the raid that startled the world.

"OPERATION THUNDERBOLT"

An INTER-OCEAN film "OPERATION THUNDERBOLT" A Golan-Globus Film of a G.S. Films Production
 starring YEHORAM GAON, ASSAF DAYAN, KLAUS KINSKY
 SYDIL DANNING, ORI LEVY, ARIK LAVI and MARK HEATH as Idi Amin
 Produced by MENAHEM GOLAN and YORAM GLOBUS
 Music by DOV SELTZER  Directed by MENAHEM GOLAN
 Eastmancolour  Distributed by EMI Films Limited

FROM THURS. OCT. 20th  ABC1 Shaftesbury Ave Licensed Bar
 AND AT SELECTED ABC AND OTHER LEADING CINEMAS FROM OCT. 27th.
 SEE LOCAL PRESS FOR DETAILS
1. THE GENERAL ACCOUNTING OFFICE IS ACTING UPON A REQUEST, BY THE HOUSE GOVERNMENT OPERATIONS COMMITTEE, TO GATHER INFORMATION ON THE OPERATION OF GAMING DEVICES IN ALL EMBASSY AND CONSULATE FACILITIES. GAMING DEVICES INCLUDE SLOT MACHINES, WHEELS OF CHANCE, DICE GAMES, ETC. BUT SPECIFICALLY EXCLUDE BINGO.

2. WE NEED TO HAVE FOLLOWING INFORMATION BY PRIORITY CABLE NO LATER THAN DECEMBER 30. REPLIES SHOULD BE DIRECTED TO LEAMON R. HUNT, DEPUTY ASSISTANT SECRETARY FOR OPERATIONS. NEGATIVE RESPONSES ARE REQUIRED. A. NUMBER OF GAMING DEVICES, WHERE LOCATED, AND WHO MANAGES THEM.
   B. OPERATING POLICIES.
   C. WHETHER LEASING OR CONcessions ARE PERMITTED, TO INCLUDE IDENTIFYING MACHINES LEASED OR BELONGING TO CONcessionaires.
   D. NUMBER OF DEVICES PURCHASED AND PROCUREMENT POLICIES.
   E. AMOUNT OF PROFITS DERIVED.

3. WE APPRECIATE YOUR PROMPT ATTENTION TO THIS REQUEST. KISSINGER
1976-12-21 | BANJUL | STATE | GAMING DEVICES
THERE ARE NO GAMING DEVICES IN EMBASSY FACILITIES. WYGANT

1976-12-22 | BREMEN | STATE | GAMING DEVICES
NO RPT NO GAMING DEVICES OF ANY TYPE AT AMCONSUL BREMEN. LONGMYER

1976-12-21 | BANGUI | STATE | GAMING DEVICES
THERE ARE NO RPT NO GAMING DEVICES IN EMBASSY FACILITIES. QUAINTON

1976-12-22 | BELIZE | STATE | GAMING DEVICES
NEGATIVE RESPONSE. WALSH

1976-12-21 | ABIDJAN | STATE | GAMING DEVICES
FOR: LEAMON R. HUNT, DEPUTY ASSISTANT SECRETARY FOR OPERATIONS NO GAMING DEVICES EXIST AT THIS POST. STEARNS

1976-12-21 | ASMARA | STATE | GAMING DEVICES
CONSULATE GENERAL ASMARA NEITHER OPERATES OR OWNS ANY GAMING DEVICES. WAUCHOPE

1976-12-21 | AMMAN | STATE | GAMING DEVICES
EMBASSY AMMAN SUBMITS NEGATIVE REPORT ON AVAILABILITY OF GAMING DEVICES WITHIN MISSION. PICKERING
Topic models can help us find events

Extensions:

- Network characteristics
- Better characterize an event for fewer “false positives”
- Word embeddings
- Autocorrelated time series
Discussion: Modern Probabilistic Modeling
I. Assume our data come from a model with hidden patterns at work

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions —at not all that far apart,” especially in comparison to the 75,000 genes in the human genome—suggest that the still-incomplete genomes at 100 genes. But coming up with a consensus answer may be more than just a genetic numbers game, particularly—more and more genomes are being sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing such minimally gene contents should show the states of ancient life.

II. Discover those patterns from data

\[ \nu^* = \arg \max_{\nu} \mathbb{E}_q [\log p(x, z, \beta | \alpha)] + \mathbb{H}[q(z, \beta | \nu)] \]
III. Use the discovered patterns to predict about and explore the data
Our perspective:

- Customized data analysis is important to many fields.
- This pipeline separates assumptions, computation, and application.
- It facilitates solving data science problems.
Figure S2: Population structure inferred from the TGP data set using the TeraStructure algorithm at three values for the number of populations $K$. The visualization of the ✓’s in the Figure shows patterns consistent with the major geographical regions. Some of the clusters identify a specific region (e.g. red for Africa) while others represent admixture between regions (e.g. green for Europeans and Central/South Americans). The presence of clusters that are shared between different regions demonstrates the more continuous nature of the structure. The new cluster from $K=7$ to $K=8$ matches structure differentiating between American groups. For $K=9$, the new cluster is unpopulated.

What we need in probabilistic ML:

- **Flexible** and **expressive** components for building models
- **Scalable** and **generic** inference algorithms
- **Easy to use** software to stretch probabilistic modeling into new areas
Figure S2: Population structure inferred from the TGP data set using the TeraStructure algorithm at three values for the number of populations $K$. The visualization of the ✓'s in the Figure shows patterns consistent with the major geographical regions. Some of the clusters identify a specific region (e.g. red for Africa) while others represent admixture between regions (e.g. green for Europeans and Central/South Americans). The presence of clusters that are shared between different regions demonstrates the more continuous nature of the structure. The new cluster from $K=7$ to $K=8$ matches structure differentiating between American groups. For $K=9$, the new cluster is unpopulated.
We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.

(John Tukey, *The Future of Data Analysis*, 1962)