NSF Big Data Publications

Story
Slides

NITRD
NITRD and the NCO
Member Agencies
NITRD PCAs
NITRD SSGs
FY 2012 Budget Estimates
NITRD Strategic 2012 Plan
The Future is Bright

Accelerating Innovation in Big Data: From Data to Knowledge to Action

The Promise of Big Data
Era of Data and Information
Era of “Big Data” in Healthcare
Conceptualizing Big Data for Science
From Hypothesis-driven to Data-driven Discovery
Smart Sensing, Reasoning and Decision
What is possible?

What's Different? Why Now?
Why Now?
Three Shifts in the Way We Analyze Information

Federal Big Data R&D Initiative
Federal Big Data R&D Initiative
NSF Framework for Investments
NSF 14-543 Due June 9, 2014
Building a Big Data R&D Pipeline
Complex Policy Setting
Public Access: Memorandum
Public Access: Implementation Plans
Data to Knowledge to Action
Materials Genome Initiative
Cognitive Science and Neuroscience

Some Success Stories...
Enabling the Patient, Curing Diseases, and Saving Lives
Data at the Forefront of Diagnosis
Transformative Implications for Commerce
UPS Uses Data Analytics to Deliver Faster and More Sustainability
Helping Small Businesses
Big Data for a Sustainable Future
Bringing NASA Data Down to Earth
Accelerating the Pace of Discovery in Science and Engineering
Extracting Knowledge from Data
Social Media and Big Data

Barriers, Challenges and Opportunities

Barriers, Challenges...

Big Data and Security
Evolution of Cyber Threats

Big Data and Privacy

Predictive Analytics
The Power and Dangers of Predictive Analytics
Dangers: Predictions Based on Correlations to Make Causal Decisions

Education and Workforce Development

NSF Research Traineeship (NST)
Imagine a Day...

Credits
Spotfire Dashboard
Slides
Slide 1 Baseball Salaries for 2013
Slide 2 2013 MLB Standard Pitching
Slide 3 2013 Pitcher Statistics
Slide 4 Data Ecosystem
Research Notes

STM Innovations Seminar U.S. 2014

Program
Results for #stminnovations2014

Professional Baseball Pitchers’ Performance and its Effect on Salary

Abstract
1 Introduction

2 Methods
   2.1 Sampling
   2.2 Software Used

3 Results
   Figure 1: Model candidates and their BIC Values
   Table 1: Predictor summary statistics
   Table 2: Coefficients for the regression of the model given in (4)
   Figure 2: Plot of the regression
   Figure 3: Residual analysis for (4)
   Table 3: Coefficients for the regression of the model given in (4) with minimum wage players excluded
   Figure 4: Plot of the regression with minimum wage players excluded
   Figure 5: Residual analysis after removing minimum wage players

4 Conclusion

References
   [1]
   [2]
   [3]
   [4]
   [5]
   [6]
   [7]
   [8]

5 Code

Story

NSF Big Data Publications

Dr. George O. Strawn, NITRD Co-chair and NCO Director, and Dr. Farnam Jahanian, Assistant Director, National Science Foundation, have the NITRD Big Data Senior Steering Group and the Federal Big Data R&D Initiative in common. Their recent presentations at the 2014 Ontology Summit are shown below.

Dr. Jahanian said: "Implementation plans for public access (to scientific research data) could vary by discipline, and new business models for universities, libraries, publishers, and scholarly and professional societies could emerge."

At the summit, they asked me about digital objects and meetups, respectively. First, what I told Dr. Jahanian about our meetups:
Our next Meetups are May 6th and 20th, 6:30-9 p.m. at Excelerate Solutions, 8405 Greensboro Dr., Suite 930, McLean, VA 22102.

Brief History: I founded this meetup because of my experience working with George Strawn (and Susi Iacona and Wendy Wigen) on a presentation to their Big Data Senior Steering Group and my work with Congressional staff on the Data Act.

I worked for many years while a government employee on Federal CIO Council activities under George’s direction and most recently on what he thought was “the killer semantic web application for the government” called Semantic Medline. We now have Semantic Medline running on the new YarcData Graph Appliance and it was the demonstration for our kickoff Meetup in January.

George and NIH (Phil Bourne and Peter Lyster) hosted the March Meetup at NSF on Joint NSF-NIH Biomedical Big Data Research: Euretos BRAIN with BarendMons.

Our Meetup mission statement is:

- Federal: Supports the Federal Big Data Initiative, but not endorsed by the Federal Government or its Agencies;
- Big Data: Supports the Federal Digital Government Strategy which is "treating all content as data", so big data = all your content;
- Working Group: Data Science Teams composed of Federal Government and Non-Federal Government experts producing big data products; and
- Meetup: The world's largest network of local groups to revitalize local community and help people around the world self-organize like MOOCs (Massive Open On-line Classes) being considered by the White House.

Our framework for working with CODATA is:

- Organize a Community of Data Scientists and Related Fields to focus on treating all of your content as "Big Data"
  - Example: Federal Big Data Working Group Meetups
- Follow the Cross Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000) consisting of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment
  - Example: Semantic Community Data Science Knowledge Base and Big Data Science for CODATA.
- Mine prominent scientific journals for data policy, data bases, and data results that can be reused.
  - Example: CODATA Data Science Journal and International Journal of Digital Earth
- Provide data stories and presentation materials for public education and conferences
  - Example: CODATA International Workshop on Big Data for International Scientific Programmes, June 8-9, in Beijing

Second, what I told Dr. Strawn about digital objects:

Re your statement Existing identity management systems (eg, VIVO), each of which uses its own version of unique identifiers as part of the process by which identity is established, may be accommodated within the Digital Object Architecture (DOA).

I wrote a recent story on VIVO's Evolution from Semantic Web Application to DuraSpace and then submitted an Abstract to the upcoming VIVO 2014 Conference about Duraspace doing essentially the same thing as Semantic Community about trying to make all of the VIVO and IVMOOC “research objects” well-defined digital objects.
In addition, I prepared NIH Data Publication 1 for my upcoming discussion with Dr. Phil Bourne and I am thinking about doing NSF Data Publication 1 using your and Dr. Jahanian’s slides from the 2014 Ontology Summit as the basis. END OF DR. STRAWN COMMUNICATION

Now I have already done a number of pilot data publications with NITRD, NSF, CODATA, EarthCube, and VIVO content, and could mine for some more content. However the recent STM Annual Event on Bright Research, Smart Articles and the new Author Ego-System gave me a new perspective. STM is the International Association of Scientific, Technical & Medical Publishers: The Voice of Academic and Professional Publishing. STM is at the leading edge of the latest technology trends within publishing. This annual US-event brings together the industry’s most established thinkers and bright up-and-coming future stars to gives attendees an insight into the hottest innovations and vital technological trends and developments which will define STM publishing for years to come.

The Plenary Session on The Smart Article: The role of Publishers in Semantic Annotation and Knowledge Representation and Searching for Data and Finding New Science description really interested me:

Increasingly the research article becomes computable, adding research data, algorithms and smart searching. How intelligent will the article become; Can it find you so you no longer need to search for it? Can it test assertions? Generate new hypotheses? Can articles generate new articles without human interference? Will human analysis be eliminated and, if so, up to what point….where are the new opportunities for publishers. Come and listen to two experts in data mining and actionable articles, both well known from FORCE11.

I mined the Tweets and selected the following:

- Common thread running through talks: metadata is extremely valuable, more valuable than the data/article/research itself
- L Hunter: "With enough data you don’t need semantic search. You can just use statistics."
- Publishers could (and should?) deliver knowledge representation thru broad knowledge, multiple information sources, etc
- L Hunter: Knowledge Representation (publishers) look at Alzforum collaborative knowledge sharing
- STM Tech trend 2: return of the author, Scholarly Author Ego System
- Kevin Boyak notes that books are cited more than journal articles
- Kevin Boyack of SciTech shares data that shows books are 2 to 4x more cited than journal articles in sciences
- Content vs context in data analysis
- Kevin Boyack up now on the mapping of science and the analytics of publishing
- "Looking for Data: Finding New Science": http://t.co/eok3ma37vO
- Tech trend 3: new players changing the game. see http://ow.ly/3jPdvY
- Tech trend 2: the return to the author
- Tech trend 1: the machine is the new reader. Highlights from the Future Lab team
- A baseball metrics talk to open. With perfect timing, the latest submission to the @writelatex gallery is an article on baseball: https://www.writelatex.com/articles/...ect-on-salary/
- How publishers can be worth the money? Also, joined a thought provoking “Future Lab Committee” this morning

This made me recall that I had sent knowledge representations on Graph Databases, Semantic Search, Data Science for VIVO, etc. to Dr. Strawn, which he really liked, and realized this was a way in which senior science managers and I...
liked information distilled and presented. In fact I have a Data Science Knowledge Base that contains a number of Data Science Books, especially for teaching Data Science. So I need to do more. The state-of-the-art wiki (MindTouch) I use supports book and publishing where I can export one or more pages to PDF and MySQL formats.

The STM Annual Event Opening Keynotes: Analytics and Metrics, used one of the innovators, writeLaTeX, as an example of advanced authoring: Professional Baseball Pitchers' Performance and its Effect on Salary. I have decided to use this an example of the advanced scientific data analysis and publishing tools I use, namely MindTouch, Excel, and Spotfire. ADD LINK AND SCREEN CAPTURES BELOW This example supports STEM education at the high school and college level. The data came from [2], [3], [4], [5], and [6], but the Excel was not provided, so I created a spreadsheet. In addition, one cannot interact with the data in their publication and their publication is not reusable, like mine is here.

Stephen Wolfram's recent Blog: Something Very Big Is Coming: Our Most Important Technology Project Yet said:


  There’ll be the Wolfram Data Science Platform, that allows one to connect to all sorts of data sources, then use the kind of automation seen in Wolfram|Alpha Pro, then pick out and modify Wolfram Language programs to do data science—and then use CDF to set up reports to generate automatically, on a schedule, through an API, or whatever.

  There’ll be the Wolfram Publishing Platform that lets you create documents, then insert interactive elements using the Wolfram Language and its free-form linguistics—and then deploy the documents, on the web using technologies like CloudCDF, that instantly support interactivity in any web browser, or on mobile using the Wolfram Cloud App.

  We’ve also been building the Wolfram Course Authoring Platform, that does major automation of the process of going from a script to all the elements of an online course—then lets one deploy the course in the cloud, so that students can have immediate access to a Wolfram Language sandbox, to be able to explore the material in the course, do exercises, and so on.

I am anxious to see what comes out, but glad to know I have a way to do advanced big data publications for NSF and other agencies and organizations now!

MORE TO FOLLOW
NITRD Networking and ITRD IT R&D
CIC computing, info and comm
HPCC and communication
HPC high-performance computing

George O. Strawn
NITRD co-chair and NCO director

NITRD and the NCO

NITRD and the NCO

- NITRD: an interagency program to enhance coordination and collaboration of the IT R&D that is performed and supported by Federal agencies
- NCO: National Coordination Office--provides support for the NITRD Program, reports to OSTP, and interfaces for NITRD with OMB, GAO, Congress, etc.

Member Agencies

Member Agencies

- DoC
  - NOAA
  - NIST
- DoD
  - OSD
  - HPCMOD
  - DARPA
  - Service Res Orgs
- DoE
- - SCI
- - NNSA
- - OE
- - DHS
- EPA
- HHS
- - AHRQ
- - NIH
- - ONC
- NARA
- NASA
- NRO
- NSA
- NSF

http://semanticommunity.info/Data_Science/NSF_Big_Data_Publications
Updated: Wed, 23 Sep 2015 02:57:44 GMT
Powered by mindtouch
NITRD PCAs

NITRD PCAs
(program component areas)

• Cyber Security and Information Assurance
• High-End Computing (R&D and I&A)
• High Confidence Software and Systems
• Human Computer Interaction and Info Mgmt
• Large Scale Networking
• Social, Economic, and Workforce Implications
• Software Design and Productivity

NITRD SSGs

NITRD SSGs (senior steering groups)

• Big Data -> HCl&IM
• CPS -> HCSS
• Cybersecurity -> CSIA
• Health IT R&D ->
• Wireless Spectrum R&D -> LSN
FY 2012 Budget Estimates

<table>
<thead>
<tr>
<th>Agency</th>
<th>HECia</th>
<th>IECrd</th>
<th>CSIA</th>
<th>HClm</th>
<th>LSN</th>
<th>HCSS</th>
<th>SDP</th>
<th>SEW</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF</td>
<td>250</td>
<td>103</td>
<td>98</td>
<td>292</td>
<td>122</td>
<td>85</td>
<td>78</td>
<td>110</td>
<td>1,138</td>
</tr>
<tr>
<td>DoD</td>
<td>211</td>
<td>49</td>
<td>145</td>
<td>111</td>
<td>112</td>
<td>36</td>
<td>30</td>
<td>694</td>
<td></td>
</tr>
<tr>
<td>NIH</td>
<td>222</td>
<td>18</td>
<td>215</td>
<td>12</td>
<td>10</td>
<td>54</td>
<td>22</td>
<td>553</td>
<td></td>
</tr>
<tr>
<td>DOE</td>
<td>317</td>
<td>92</td>
<td>34</td>
<td>74</td>
<td>4</td>
<td>16</td>
<td>6</td>
<td>543</td>
<td></td>
</tr>
<tr>
<td>DARPA</td>
<td>75</td>
<td>223</td>
<td>138</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>NIST</td>
<td>14</td>
<td>5</td>
<td>47</td>
<td>15</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>NASA</td>
<td>61</td>
<td>14</td>
<td>1</td>
<td>18</td>
<td>9</td>
<td></td>
<td></td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>DHS</td>
<td>43</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>AHRQ</td>
<td>25</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>NOAA</td>
<td>19</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>DOE</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>NARA</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,107</td>
<td>347</td>
<td>590</td>
<td>814</td>
<td>385</td>
<td>158</td>
<td>196</td>
<td>143</td>
<td>3,739</td>
</tr>
</tbody>
</table>

NITRD Strategic 2012 Plan

- “We Compute”
- Trust and Confidence
- Cyber-ready Society
The Future is Bright

- NIT R&D
  - IST R&D?
  - Computing R&D?
- Informatics R&D?
- Whatever we call it, science and society depend on it more every year!
- Enhanced NIT R&D is critical
  - to fulfilling the promise of NIT for society
  - to enhancing American competitiveness
  - to furthering the missions of NITRD Agencies

Accelerating Innovation in Big Data: From Data to Knowledge to Action

The Promise of Big Data
Era of Data and Information

Era of "Big Data" in Healthcare

- Large volumes of data currently collected
  - EHRs and PHRs
  - Multi-scale and multi-source
  - During hospitalizations
  - For safety and diagnosis
  - On an out-patient basis
  - Typically event monitors
  - Via ubiquitous mobile sensors
  - Behavior, physiology, environment
  - As part of clinical studies
  - To evaluate safety and efficacy
  - From growing body of scientific knowledge in biomedical research literature

- Gigabytes/patient/day
- High sampling rates
- Multiple signals
- Accumulating data is getting easier, but using data is hard
Conceptualizing Big Data for Science

- Science gathers data at an ever-increasing rate across all scales and complexities of natural phenomena
- Sloan Digital Sky Survey in 2000, collected more data in its 1st few weeks than had been amassed in the entire history of astronomy
  - Within a decade, over 140 terabytes of information collected
- The Large Synoptic Survey Telescope due in Chile in 2016 will amass double that quantity of data every week

From Hypothesis-driven to Data-driven Discovery

- The Economist: The data deluge and how to handle it. A 24-page special report (Feb 25, 2010).

Data are motivating a profound transformation in the culture and conduct of scientific research.
Smart Sensing, Reasoning and Decision

What is possible?

From data to knowledge to discovery by
- Enabling extraction of knowledge from very large, heterogeneous data sets
- Providing novel approaches to driving discovery and decision-making
- Yielding increasingly more accurate predictions, potentially saving lives
- Providing deeper understanding of causal relationships based on advanced data analysis

What’s Different? Why Now?
Why Now?

Why Now?
Confluence of Social, Technical and Policy Interests

- Decades of advances in technology
- Plunging costs of computation, communication and storage
- Consumer technologies, broadband connections and social media
- Data is no longer regarded as static:
  - now a raw material or a corporate asset, used to create economic value, and foundation for new business models
- Platform economy and innovation ecosystem
- Increasing transparency of democratic governance (e.g., gov)
  - Public access to high value datasets (data.gov)
- Democratization of data and tools

Three Shifts in the Way We Analyze Information

Three Shifts in the Way We Analyze Information

- **More**: ability to collect, manage and analyze far more data rather than be artificially limited by sampling

- **More**: loss of accuracy and exactness at micro-level, but gain insight at the macro-level

- **Good enough**: discover patterns and correlations rather than causality

Federal Big Data R&D Initiative

Federal Big Data R&D Initiative
Federal Big Data R&D Launch
March 29, 2012

- Led by cross-agency “Big Data” Senior Steering Group – chartered in spring 2011 by the White House OSTP:
  - Co-chaired by NSF and NIH
  - Significant research-community input
  - Members from 15 agencies
  - Charged with developing a framework and a plan

- Major Announcements: NSF, NIH, USGS, DOD, DARPA, DOE

- Cornerstone Announcement: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BD Data) Initiative:
  - 18 NSF Directors and 8 NIH Institutes
  - Research Themes: Collectors, Storage, and Management, Data Analytics, Research in Data Sharing and Collaboration

NSF Framework for Investments

**NSF Framework for Investments**

- **Foundational Research**: to develop new techniques and technologies to derive knowledge from data
- **New Cyberinfrastructure**: to manage, curate, and serve data to research communities
- **New Approaches for Education and Workforce Development**: New types of interdisciplinary Collaborations, grand challenges, and competitions
- **Policy**:
Critical Techniques and Technologies for Advancing Big Data Science and Engineering (NSF 14-543)

- Two categories for submission
  - Foundational: Encourages fundamental, novel techniques, theories, methodologies and technologies of broad applicability
  - Innovative Applications: Encourages novel techniques, theories, methodologies, and technologies of interest to at least one specific application
- Due Date: June 9, 2014
- Size: up to $500K per year for up to 4 years

Building a Big Data R&D Pipeline

Building a Big Data R&D Pipeline

- Foundational Research
- Domain Specific Application Research
- Cyberinfrastructure Pilots
Complex Policy Setting

- Researchers want data.
- Public policy requires access to data.
- Public policy also requires protection of privacy, intellectual property, and sensitive information.
- Business model challenges for publishers and societies.
- White House Memo on Feb. 22 directs United States federal agencies to develop a plan to support “increased public access” of results from federally funded research.

Public Access: Memorandum

- White House Memo on Feb. 22 directs United States federal agencies to develop a plan to support “increased public access” of results from federally funded research.

- Peer-reviewed publications should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the output and accountability of the federal research investment.

- Digitally formatted scientific data resulting from unclassified research should be stored and publicly accessible to search, retrieve, and analyze.
Public Access: Implementation Plans

- White House Memo on Feb. 22 directs United States federal agencies to develop plans to:

  Implementation plans for public access could vary by discipline, and new business models for universities, libraries, publishers, and scholarly and professional societies could emerge.

- Digitally formatted scientific data resulting from unclassified research "should be stored and publicly accessible to search, retrieve, and analyze."

Data to Knowledge to Action

Data to Knowledge to Action:
White House event encouraging public-private partnerships across the country
November 12, 2013
Materials Genome Initiative

Enabling the Patient, Curing Diseases, and Saving Lives

Cognitive Science and Neuroscience

Some Success Stories...

http://semanticommunity.info/Data_Science/NSF_Big_Data_Publications
Updated: Wed, 23 Sep 2015 02:57:44 GMT
Powered by mindtouch
Data at the Forefront of Diagnosis

Transformative Implications for Commerce

UPS Uses Data Analytics to Deliver Faster and More Sustainability
Helping Small Businesses

- The NYC Mayor’s Office of Data Analytics (MODA) is working across city government to use city data to improve daily operations, help to prepare for and respond to disasters, and support economic growth.
- Now working with NYC’s New Business Acceleration Team (NBAT) to help new restaurants get through red tape and open their doors to customers.
- Using data on construction permits (Department of Buildings), restaurant inspections (Department of Health and Mental Hygiene), and NBAT counseling notes to see how free NBAT services can reduce a new business’ time to open.

Big Data for a Sustainable Future

Bringing NASA Data Down to Earth

- Amazon Web Services (AWS) and NASA are providing a significant amount of NASA’s Earth science data and models to the public.
- Giving everyone access to data and analytic techniques previously only available to NASA researchers.
- Enables calculation of the next National Climate Assessment on the AWS cloud.
- Hosting NASA NEX data in the cloud also enables crowd-sourced citizen science applications like those found on Zooniverse (zooniverse.com).

Accelerating the Pace of Discovery in Science and Engineering
Extracting Knowledge from Data

Social Media and Big Data

Barriers, Challenges and Opportunities
Barriers, Challenges...

- Technological Solutionism
- Big data security
- Big data and privacy
- Dangers of predictive analytics
- Education and Workforce Development

Big Data and Security

Evolution of Cyber Threats

Evolution of Cyber Threats

Future security challenges will follow technology and Internet adoption patterns

- Proliferation of mobile devices and wireless networks exposes new vulnerabilities.
- Social media platforms open new avenues for hackers.
- Protecting cloud infrastructure has become key to long-term adoption.
- Increasingly cyber-enabled systems expand the scope of attacks to physical infrastructure — manufacturing, energy production, healthcare and transportation.

Big Data and Privacy
Big Data and Privacy

- Informed consent, opting out, and anonymization are not as effective to ensure privacy with big data:
- How can companies provide notice for a purpose that has yet to exist?
- How can one give informed consent if new secondary uses of information haven’t even been imagined?
- What if anonymized data can be mined with other (public) datasets to identify private information?

Predictive Analytics

The Power and Dangers of Predictive Analytics

- Data mining and pattern recognition to make predictions:
  - Flu outbreaks throughout the world
  - Predicating election outcomes with high accuracy
  - Suggestions on customer preferences for books
  - Predicting customer behavior and managing inventory
  - Detecting insurance fraud
- Who would object to Preventing unhealthy or dangerous behavior?
- Who would object to improving quality of customer experience?
- Who would object to improving business efficiency?
- Who would object to efficient use of resources in crime prevention?
Dangers: Predictions Based on Correlations to Make Causal Decisions

- Correlation does not imply causation
- “Fooled by randomness” as suggested by Nassim Taleb
- Crossing the line from improvement to profiling
- Crossing the line from prevention to penalizing
- Idea of the “presumption of innocence” is foundational to our legal system

Education and Workforce Development

NSF Research Traineeship (NST)

NSF Research Traineeship (NRT)

Preparing professionals in emerging STEM fields vital to the nation

Priority research theme: Data-enabled science and engineering

- **Purpose**: create and promote new, innovative, effective, and scalable models for STEM graduate student training and prepare scientists and engineers of the future, particularly in emerging STEM fields vital to the nation.
- **Anticipated award amount**: up to $3M over 5 yrs.

NSF-wide Initiative
Imagine a Day...

- By integrating biomedical, clinical, and scientific data, we can predict the onset of diseases and identify unwanted drug interactions.
- By coupling roadway sensors, traffic cameras, and individuals’ GPS devices, we can reduce traffic congestion and generate significant savings in time and fuel costs.
- By accurately predicting natural disasters such as hurricanes and tornadoes, we can employ life-saving and preventative measures that mitigate their potential impact.
- By integrating emerging technologies, such as MOOCs and inverted classrooms, with knowledge from research about how people learn, we can transform formal and informal education.
- By correlating disparate data streams through text mining, image analysis, and face recognition, we can enhance public safety and security.

Credits

Copyrighted material used under Fair Use. If you are the copyright holder and believe your material has been used unfairly, or if you have any suggestions, feedback, or support, please contact: ciseitsupport@nsf.gov.

Except where otherwise indicated, permission is granted to copy, distribute, and/or modify all images in this document under the terms of the GNU Free Documentation license, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation license” at http://commons.wikimedia.org/wiki/Co...tation_License.

The inclusion of a logo does not express or imply the endorsement by NSF of the entities' products, services, or enterprises.

Spotfire Dashboard

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Slides
STM Innovations Seminar U.S. 2014


International Association of Scientific, Technical & Medical Publishers  
The Voice of Academic and Professional Publishing

Bright Research, Smart Articles and the new Author Ego-System  
Annual US event

STM is at the leading edge of the latest technology trends within publishing. This annual US-event brings together the industry's most established thinkers and bright up-and-coming future stars to gives attendees an insight into the hottest innovations and vital technological trends and developments which will define STM publishing for years to come.

Don’t miss the Flash Session with a line-up of crazy-5-minute talks that provide the newest of what publishers are preparing to launch.

Program

13.15  Welcome and Introduction by [Gerry Grenier](http://www.ieee.org), IEEE, chair of the STM Future Lab

Opening Keynotes; Analytics and Metrics

**Moderator:** [Dave Martinsen](http://www.acs.org), ACS

Science in Sports,
Prof David W. Smith, Professor of Biological Sciences and baseball researcher and historian, University of Delaware

Mapping of Science and the Analytics of Publishing,

Kevin Boyack, President SciTech Strategies, Inc.

If we say ‘data’ we also automatically say ‘metrics and analytics’. New analytics tools have surged innovative new applications using the increasing abundance of data. In sports the use of new metrics and analytics creates a whole new area of science, as shown by baseball historian and researcher professor David W. Smith (who is also professor of Biological Sciences). In the area of STM publishing, analysis of publication data helps visualize the mapping of science. Pioneer Kevin Boyack of Scitech Strategies shows several real life examples.

14:30 Panel: STM Tech Trends 2014

Moderator Christopher Kenneally (CCC)

The new technology trends for 2014 and further, as they will impact STM publishers, explained by a panel of 4 from STM’s Future Lab Committee. With Andrea Powell (CABI), Howard Ratner (Chorus), Heather Ruland Staines (SIPX), Dave Martinsen (ACS),

15:10 Break

15:30 Plenary: The Smart Article

Moderator: Eefke Smit, STM

The role of Publishers in Semantic Annotation and Knowledge Representation

Professor Larry Hunter, University of Colorado, Denver, Computational BioScience Program

Searching for Data and Finding New Science

Anita de Waard, Elsevier, Vice-President Research Data Collaborations

Increasingly the research article becomes computable, adding research data, algorithms and smart searching. How intelligent will the article become; Can it find you so you no longer need to search for it? Can it test assertions? Generate new hypotheses? Can articles generate new articles without human interference? Will human analysis be eliminated and, if so, up to what point….where are the new opportunities for publishers. Come and listen to two experts in data mining and actionable articles, both well known from FORCE11.
16:15  **Panel: The new Author Ego-System**

**Moderator:** Ann Michael, DeltaThink

Within the STM scholarly ecosystem, publishers must consider and address an evolving set of author expectations. This panel-session will explore four types of innovative author-focused services: Collaborative Authoring Tools, Readership Maximization, Alternative Metrics, and Author Payment Systems. Panelists will briefly describe the underlying author need these services are meant to address before engaging in a lively and participatory discussion with each other and the audience. With Mike Buschman (Plum Analytics), Sarah Tegen (ACS), Jake Kelleher (CCC), Melinda Kenneway (GrowKudos.com).

17:00  **Flash Session**

**Moderator:** Terry Hulbert, consultant.

Six ultra short talks on new start-ups, new launches, new services that help improve scholarly communication, with Colwiz, Newstex, Digimarc, WriteLatex, Algosharing, CHORUS:

Stuart Cooper, Colwiz

Ed Colleran, Newstex

Burt Slavin, Digimarc

John Hammersley, WriteLatex

Simon Adar, AlgoSharing

Howard Ratner, CHORUS

17:30  **Welcome Reception** for STM Spring Conference & Innovations Seminar attendees

---

Results for **#stminnovations2014**

**Source:** https://twitter.com/search?f=realtime&s2014&src=typd
Thanks to everyone who made #stminnovations2014 and #stmspring2014 a success.

2.

Howard Ratner @hratner  Apr 30

Griffiths: Funding-Intellectual property ownership-Access-Validation-Provenance-Compliance, notification, registration #stminnovations2014

3.

John Hammersley @DrHammersley  Apr 29

#stminnovations2014 With perfect timing, the latest submission to the @writelatex gallery is an article on baseball! https://www.writelatex.com/read/xndqqqynynm#.U2BvIqI6V6A.twitter …

4.

John Hammersley @DrHammersley  Apr 29
What a great reaction to my @writelatex flash talk earlier :) Looking forward to more STM events! http://www.stm-assoc.org/events/stm-innovations-seminar-u-s-2014/... #stminnovations2014

Jeff Lewandowski @JeffLewandowski Apr 29

Common thread running through #stminnovations2014 talks: metadata is extremely valuable, more valuable than the data/article/research itself

Alex Humphreys @abhumphreys Apr 29

L Hunter: "With enough data you don't need semantic search. You can just use statistics." Interested to hear more. #stminnovations2014
Jayne Marks @jaynedmarks  Apr 29

Publishers could (and should?) deliver knowledge representation thru broad knowledge, multiple information sources, etc #stminnovations2014

Expand

- Reply
- Retweet
- Favorite
- More

8.

Adrian @AdrianStanley13  Apr 29

L Hunter: Knowledge Representation (publishers) look at Alzforum collaborative knowledge sharing
http://www.alzforum.org/about-us/mission ... #stminnovations2014

Expand

- Reply
- Retweet
- Favorite
- More

9.

Jayne Marks @jaynedmarks  Apr 29

Larry Hunter on “How publishers can be worth the money” #stminnovations2014

Expand

- Reply
- Retweet
- Favorite
- More

10.
Adrian @AdrianStanley13 Apr 29

STM Tech trend 2: return of the author, Scholarly Author Ego System #stminnovations2014

Expand

- Reply
- Retweet
- Favorite
- More

11.

Ann Michael @annmichael Apr 29

Kevin Boyak notes that books are cited more than journal articles - would love to hear more about that
#stminnovations2014

Expand

- Reply
- Retweet
- Favorite
- More

12.

Anita de Waard @anitawaard Apr 29

Relevant to #stminnovations2014 https://docs.google.com/document/d/1uC5wmbcL90qyrkJ27fWAvDcDNrO9o_APkicwRkOKc/edit ...

Expand

- Reply
- Retweet
- Favorite
- More

13.

Alex Humphreys @abhumphreys Apr 29

Kevin Boyack of SciTech shares data that shows books are 2 to 4x more cited than journal articles in sciences.
#stminnovations2014
14. John Hammersley @DrHammersley Apr 29

Content vs context in data analysis at #stminnovations2014 pic.twitter.com/6hVlDhX5QC

View photo

15. Jayne Marks @jaynedmarks Apr 29

Kevin Boyack up now on the mapping of science and the analytics of publishing. #stminnovations2014

16. Anita de Waard @anitawaard Apr 29

Just posted my slides for #STMinnovations2014 "Looking for Data: Finding New Science" on @slideshare
http://www.slideshare.net/anitawaard/looking-for-data-finding-new-science-34091601 ...

View media
17. Melinda @MelindaKenneway  Apr 29

Looking forward to my panel session - representing Kudos - on the author egosystem at #stminnovations2014

18. Jayne Marks @jaynedmarks  Apr 29

Tech trend 3: new players changing the game. see http://ow.ly/3jPdvY for more info. #stminnovations2014

19. Jayne Marks @jaynedmarks  Apr 29

Tech trend 2: the return to the author. #stminnovations2014
20.

Jayne Marks @jaynedmarks Apr 29

Tech trend 1: the machine is the new reader. Highlights from the Future Lab team at the STM Assoc #stminnovations2014 Expand

21.

Jeff Lewandowski @JeffLewandowski Apr 29

A baseball metrics talk to open #stminnovations2014 ... quoting #TheodosiusDobzhansky on slide 3... Yeah, this suits me well! #allears

22.

Larry Hunter @ProfLHunter Apr 29

My talk at #stminnovations2014: How publishers can be worth the money? Also, joined a thought provoking “Future Lab Committee” this morning

http://semanticommunity.info/Data_Science/NSF_Big_Data_Publications
Updated: Wed, 23 Sep 2015 02:57:44 GMT
Powered by mindtouch
23. 

Anita de Waard @anitawaard Apr 29

#stminnovations2014 needs to get a better #tag!

Expand

24. 

John Hammersley @DrHammersley Apr 29

The shiny new Overleaf banner gets its first outing at #stminnovations2014 :) http://writelatex.com/overleaf
pic.twitter.com/9lE5WKnx3

View photo

25. 

writeLaTeX @writelatex Apr 29

Our @drhammersley is presenting in the final session at STM Innovations today - updates via #stminnovations2014 http://www.stm-assoc.org/events/stm-innovations-seminar-us-2014/#.U1-H8fGIQ04.twitter ...

Expand
John Hammersley @DrHammersley  Apr 27


writeLaTeX @writelatex  Mar 19

Latest news! Our @DrHammersley is presenting Overleaf at #stminnovations2014 in Washington in April - full line up: http://www.stm-assoc.org/events/stm-innovations-seminar-u-s-2014/ …

STM Association @STMAssoc  Feb 26


---

Professional Baseball Pitchers’ Performance and its Effect on Salary

Source: https://www.writelatex.com/articles/…dsgqnnrnym.pdf (PDF)
Abstract

In this study we identify factors that affect a Major League Baseball (MLB) pitcher’s salary. We are interested in knowing whether ability is a good indicator of compensation. To test this we created a model to predict the salaries of pitchers in the MLB.

1 Introduction

Money is a major driving factor in professional baseball and a major consideration for team managers looking to make changes to their rosters. Baseball is not a fair game: in most professional sports, teams are limited to a salary cap (e.g., the NFL has a salary cap of $133 million per team [1]). In baseball, however, there are no such limitations; team payrolls are limited only by their owners’ willingness to pay.

These payrolls may be determined by the amount of money generated by ticket sales or by the sale of team paraphernalia and royalties. There is no set amount required for ticket sales by the MLB, therefore each team can choose to charge as much or as little as they want for tickets. Popular teams with large fanbases are generally able to charge more for tickets or sell tickets in greater volume than less popular teams. Additionally, team payroll may be correlated by the market size of their home city [?].

This leads to major discrepancies in the amount teams are able to pay their players and the caliber of players they are able to recruit. In 2013, the Houston Astros had the lowest payroll in baseball at $26.1 million. The New York Yankees, the highest paying team in the league, paid out a staggering $228.1 million - over eight times as much. Alex Rodriguez, the highest paid player on the Yankees and in the league, earned $28 million in 2013: more than every player on the Astros team combined!

With this in mind, it is clear that low-budget teams (often called “small market teams”) should be seeking out players who will play for a lower salary but still perform. Contrastingly, large market teams should only accept the best players, and will lure them in with exorbitant salaries. We evaluated the salaries of 345 pitchers in 2013 to see if this is true.

2 Methods

2.1 Sampling

We used a sample of 345 pitchers for this study. Pitchers were chosen as they play a crucial role on the team and are relatively easy to compare to one another. We only considered pitchers who have been playing for three consecutive years (i.e. 2011, 2012, and 2013), as the salaries of rookie pitchers cannot be predicted without statistics from prior years. We also removed from consideration pitchers who make less than $700,000; these players’ salaries are dictated by the MLB price floor, not their (not-so-great) performances.
2.2 Software Used

All statistics and plots were done in R. Charts and formatting were done in LaTeX. Data was gathered and arranged in Microsoft Excel.

My Note: The data came from [2], [3], [4], [5], and [6], but the Excel was not provided, so I created a spreadsheet

3 Results

Since we are want to know if ability is the driving factor behind a pitcher’s salary, we chose earned run average (ERA) to quantify this. ERA is defined as

\[ \text{ERA} = 9 \times \frac{\text{Earned runs allowed}}{\text{Innings pitched}} \]  

where “earned runs” are runs not scored on a fielding error [7]. A lower score signifies a pitcher who allows less runs, so a lower ERA is better. Since runs are all that matter at the end of a game, this is a good indicator of a pitcher’s performance. It is also worth noting that this is not a count statistic, since it is an average per nine innings. The only major downfall here is ERA does not take quality of opponents into account, but since every pitcher are pitching to hundreds of opponents across different teams, this isn’t significant.

We defined our testing hypotheses using ERA:

\[ H_0 : \beta_1 = 0 \]
\[ H_1 : \beta_1 \neq 0 \]  

where \( \beta_1 \) is the coefficient for ERA predicting salary.

In order to test these hypotheses, we created by choosing from 36 predictors, including ERA. We choose from all possible models with 5 or fewer predictors based on their Bayesian information criterion (BIC). This selection method helped us deal with multicolinearity and the computational time required to deal with a large number of potential models. BIC is defined as

\[ \text{BIC} = -2 \ln(\hat{L}) + k \cdot (\ln(n) + \ln(2\pi)) \]  

where \( n \) is the number of data points (in our case 345), \( k \) is the number of regressors (this penalizes models using many regressors), and \( \hat{L} \) is the maximum value of the likelihood function for the model. Minimizing BIC, we found several good candidates:

**Figure 1: Model candidates and their BIC Values**

ERA was not chosen in any of the models. Note that models were generated using an exhaustive algorithm (i.e. during each step all models were considered).
Based on this criterion, we chose the predictor variables age (AGE), age squared \((AGE^2)\), games starting (GS), innings pitched as a starter (IP.S), and saves (SV). In order to test our hypothesis, include ERA.

Table 1: Predictor summary statistics

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>Std. Dev.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERA</td>
<td>1.44</td>
<td>3.75</td>
<td>8.11</td>
<td>1.04</td>
<td>345</td>
</tr>
<tr>
<td>AGE</td>
<td>21</td>
<td>29</td>
<td>41</td>
<td>3.71</td>
<td>345</td>
</tr>
<tr>
<td>(AGE^2)</td>
<td>441</td>
<td>854.9</td>
<td>961</td>
<td>224.50</td>
<td>345</td>
</tr>
<tr>
<td>GS</td>
<td>0</td>
<td>12.4</td>
<td>34</td>
<td>7.856</td>
<td>345</td>
</tr>
<tr>
<td>IP.S</td>
<td>0</td>
<td>75.83</td>
<td>236</td>
<td>80.29</td>
<td>345</td>
</tr>
<tr>
<td>SV</td>
<td>0</td>
<td>3.518</td>
<td>46</td>
<td>7.86</td>
<td>345</td>
</tr>
</tbody>
</table>

Using these variables we predict salary using a regression in the form \(\ln(SALARY) = Beta_0 + Beta_2AGE + Beta_3AGE^2 + Beta_4GS + Beta_5IP.S + Beta_6SV\) (4)

We use \(\ln(SALARY)\) to deal with heteroscedasticity.
There are no negative salaries, so we expect $\beta_0$ to be positive.

As previously discussed, a low ERA indicates a more skilled pitcher, so we expect $\beta_1$ to be negative.

Since professional athletes tend to get better after their rookie year up until a "peak", and then decline with age, we expect the age predictors will create a concave-down parabola peaking somewhere in the mid to late twenties. Therefore, we expect $\beta_2$ to be positive and $\beta_3$ to be negative.

Valuable pitchers will start more games, so we expect $\beta_4$ to be positive. This will also increase the value of IP$_S$ (along with the stamina required to pitch more innings per game), so we predict $\beta_5$ will be positive.

A pitcher who finishes a game records a record a save if at least one of three conditions are satisfied: 
• his team is ahead by less than four runs when he enters the game and he pitches for an entire inning
• he enters the game when the enemy team has the potential to tie the game with the next at-bat
• he pitches for at least three innings

A pitcher cannot record a win and a save in the same game [7]. Since a good relief pitcher will rack up more saves and have more opportunities to do so, $\beta_6$ is positive.

Running the regression, we found the following coefficient values:

**Table 2: Coefficients for the regression of the model given in (4)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>27.254145</td>
<td>1.933914</td>
<td>14.093</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>ERA</td>
<td>0.042576</td>
<td>0.028177</td>
<td>1.551</td>
<td>0.13172</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.736468</td>
<td>0.131323</td>
<td>-5.608</td>
<td>4.26e-08</td>
</tr>
<tr>
<td>( \text{AGE}^2 )</td>
<td>0.010664</td>
<td>0.002214</td>
<td>4.817</td>
<td>2.21e-06</td>
</tr>
<tr>
<td>GS</td>
<td>0.042898</td>
<td>0.038828</td>
<td>1.105</td>
<td>0.270020</td>
</tr>
<tr>
<td>IP$_S$</td>
<td>-0.017646</td>
<td>0.006203</td>
<td>-2.845</td>
<td>0.004715</td>
</tr>
<tr>
<td>SV</td>
<td>-0.068462</td>
<td>0.006258</td>
<td>-10.940</td>
<td>&lt; 2e-16</td>
</tr>
</tbody>
</table>

Here, $\beta_3$ went against our intuition. Stranger yet, $\beta_3$ and $\beta_4$ have opposite signs, although IP$_S$ is directly dependent on GS. This combination of indicators provides an interesting metric that values pitchers who pitch many innings with few starts - these are pitchers either have the endurance to pitch deeper into the game or are not pulled early from the game as often.
We also predicted $\beta_6$ incorrectly. This is likely negative because savers earn less on average than starting pitchers.

**Figure 2: Plot of the regression**

![Pitcher Salaries vs Predictions](image)

The model fits the data well and has an $R^2$ value of 0.64, but there are multiple problems which are illustrated by the residual plots:

**Figure 3: Residual analysis for (4)**

Note the straight line on the residual plot and the "V shape" on the scale-location plot. Both indicate systematic residuals.
Both of these problems can be explained by the large group of players seen in Figure 2 on the bottom end of salaries. These players are earning the MLB minimum wage ($500000) [8]. Since these players’ salaries are not dictated by the salary price floor and not skill, we removed them and reran the model:

**Table 3: Coefficients for the regression of the model given in (4) with minimum wage players excluded**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.293866</td>
<td>1.784664</td>
<td>4.647</td>
<td>5.51e-06</td>
</tr>
<tr>
<td>ERA</td>
<td>-0.051700</td>
<td>0.039049</td>
<td>-1.324</td>
<td>0.186755</td>
</tr>
<tr>
<td>AGE</td>
<td>0.394701</td>
<td>0.119750</td>
<td>3.296</td>
<td>0.001126</td>
</tr>
<tr>
<td>AGE$^2$</td>
<td>-0.005752</td>
<td>0.001978</td>
<td>-2.907</td>
<td>0.003979</td>
</tr>
<tr>
<td>GS</td>
<td>-0.062597</td>
<td>0.033955</td>
<td>-1.844</td>
<td>0.066467</td>
</tr>
<tr>
<td>IP.S</td>
<td>0.019358</td>
<td>0.005356</td>
<td>3.614</td>
<td>0.000366</td>
</tr>
<tr>
<td>SV</td>
<td>0.047817</td>
<td>0.00501</td>
<td>49.537</td>
<td>&lt; 2e-16</td>
</tr>
</tbody>
</table>

Figure 3: Residual analysis for (4). Note the straight line on the residual plot and the “V shape” on the scale-location plot. Both indicate systematic residuals.
Figure 4: Plot of the regression with minimum wage players excluded

The transformation did not significantly improve our $R^2$ value, but this is a much more valid model:

Figure 5: Residual analysis after removing minimum wage players

The only potential problem is the slight movement in standardized residuals, but this is not a big issue considering the data. A transformation other than log might fix this.
We failed $H_0$ at a 0.05 significance level in this model for both data sets.

4 Conclusion

ERA was not chosen as a strong predictor in our BIC predictor selection, and it was not found to be significant in our model. Every other predictor selected is either a count statistic or age. The count statistics are related to the amount a player is chosen to play, instead of directly measuring his ability; while a better player is certainly likely to play more often, there is could an additional effect of coaches trying to "get their money's worth" out of highly-paid players.

We hypothesized that high ability should yield a high salary, but this is not strictly the case for the data we observed. One explanation for this could be contract restrictions: contracts can sometimes block players from being paid a salary deserve.

Another variable likely to be significant which we omitted is attendance per game. Since ticket sales generate revenue for teams, a more likable or exciting pitcher may be worth more than a highly skilled one. This might help to explain why age was so significant, since older players have had more time to gather a large fanbase.

We can see that the data does indicate some correlation between salary performance, but this effect is not as direct as we had expected.

References

My Note: See Spreadsheet for Data from these References
5 Code

1 # Get predictor variables and salary data:
2 X <- read .csv ( " predictors .csv ")
3 d <- read .csv ( " PitcherData . csv ")
4 SALARY <- d$ SALARY
5
6 # Choose predictor variables
7 library ( alr3 )
8 library ( leaps )
9 library ( car )
10 ss <- regsubsets ( as . matrix (X),Y, nvmax =5)
11 rs <- summary (ss)
12 subsets (ss , statistic =c(" bic "),legend =FALSE , xlim =c(1 ,6))
13 title (" BIC values ")
# Create linear model
attach(X)
lm <- lm( SALARY ~ AGE + AGE2 + GS + IP. Start + SV)
sink (" lmoutput1 . txt ")
summary(lm)

# Plot this model
par ( mar =c(4 ,4 ,2 ,2))
plot ( SALARY [ order ( fit)], ylim =c (0 ,25000000) , xlab =" Pitcher ", ylab =" Salary (Millions of Dollars )", axes =FALSE )
box ()
axis (2)
par ( new =t)
plot ( fit [ order ( fit )], ylim =c (0 ,25000000) , col =" red ", type ="l", lwd =2, axes =F,ylab ="", xlab ="")
title (" Pitcher Salaries vs Predictions ")
legend (" topleft ", legend =c(" Actual Salaries ", " Predicted Salaries "), pch =c(1 ,26) , lty =c(0 ,1) , lwd=c(0 ,2) , col =c(" black "," red"))

# plot residuals
par ( mfrow =c(2 ,2) , mar =c(4 ,4 ,2 ,2))
plot (lm)
powerTransform (lm)
bcSALARY <- 4*(Y ^0.25 -1)
bcclm <- lm( bcSALARY ~ AGE + AGE2 + GS + IP. Start + SV)
sink (" lmoutput2 . txt ")
summary ( bclm )

# Plot bc model
par ( mfrow =c(1 ,1) , mar =c(4 ,4 ,2 ,2))
plot ( bcSALARY [ order ( bcfit )], ylim =c (111 ,279) , xlab =" Pitcher ", ylab =" Salary (Transformed Dollars )", axes =F)
box ()
axis (2)
par ( new =t)
plot ( bcfit [ order ( bcfit )], ylim =c (111 ,279) , col =" red ", type ="l", lwd =2, axes =F,ylab ="", xlab ="")
title (" Transformed Pitcher Salaries vs Predictions ")
legend (" topleft ", legend =c(" Actual Salaries ", " Predicted Salaries "), pch =c(1 ,26) , lty =c(0 ,1) , lwd=c(0 ,2) , col =c(" black "," red"))
=c(" black "," red"))

52

53 # plot bc residuals
54 par ( mfrow =c(2 ,2) , mar =c(4 ,4 ,2 ,2))
55 plot ( bclm )