

Report on NISS Affiliates Planning Meeting: March 2, 2001

Alan F. Karr
March 8, 2001

1 Summary

The annual Planning Meeting of the NISS Affiliates was held on March 2, 2001, at NISS headquarters in Research Triangle Park, NC.

Present were David Banks (Bureau of Transportation Statistics), James Berger (Duke University), Kevin Coakley (National Institute of Standards and Technology), Sid Dalal (Telcordia Technologies), William Dunsmuir (University of Minnesota), Stephen Eick (Visual Insights), John Eltinge (Bureau of Labor Statistics), Thomas Ferryman (Pacific Northwest National Laboratory), Nancy Flournoy (American University), Alan Karr (NISS), Lisa LaVange (Quintiles Transnational), Russell Lenth (University of Iowa), Joseph McCloskey (National Security Agency), Andrew Nobel (University of North Carolina at Chapel Hill), Sastry Pantula (North Carolina State University), Rick Picard (Los Alamos National Laboratory), Jeffrey Robinson (General Motors), Robert Rodriguez (SAS Institute), Avi Singh (Research Triangle Institute), Keith Soper (Merck & Company), Jackson Stenner (MetaMetrics), Charles Taylor (Procter & Gamble), Tommy Wright (Census Bureau), Colin Wu (Johns Hopkins University) and Stanley Young (GlaxoSmithKline).

2 Program Review

Karr reported on past activities, current status and plans for the affiliates program,¹ emphasizing:

- The mission and vision (high value to both affiliates and NISS) for the program;
- Management of the program: a director is being recruited, an Affiliates Advisory Council is in place, and an Affiliates Committee of the NISS Board of Trustees is being formed;
- Communications initiatives: monthly updates, an annual report² and the NISS newsletter;³
- Planning meetings in 2000 (March 3 at NISS and August 12 at the JSM);
- Workshops to data: on bioinformatics (July 13, 2000, at NISS and February 11-12, 2001, at Amgen) and data quality (November 30 – December 1, 2000, at Telcordia);

¹The report is available at www.niss.org/affiliates/planning200103/afkpresentation.pdf.

²Available at www.niss.org/affiliates/annualreports/affiliates-2000.pdf.

³Available at www.niss.org/newsletter.html.

- Affiliate–stimulated research on computer model evaluation (funded by GM), drug discovery (funded by GlaxoSmithKline) and Web data (funded by Visual Insights), as well as the NSF–funded digital government project, in which BLS, Census and NCES are partners;
- Planned (or in the process of being planned) workshops on network data (March 9–10, 2001, at NISS), data confidentiality, bioinformatics, Web data and large data sets;
- The NISS Affiliates Internship Program (NAIP), planned to begin full operation in 2002;
- The proposed Statistical and Applied Mathematical Sciences Institute (SAMSI), and its relationship to the NISS affiliates;
- Challenges facing NISS and the affiliates program.

3 Four Minute Madness (FMM)

3.1 Summary

Every attendee took (approximately) four minutes to describe concerns, needs and capabilities of his or her organization that are relevant to the affiliates program. Corporate and government affiliates concentrated on personnel needs (see §3.2 below) and on statistical problems they face, while university affiliates emphasized the interests and experience of their faculties.

3.2 Human Resource Needs and Opportunities

Implicit and often explicit in FMM presentations were a number of human resources concerns that present new opportunities for the program to serve affiliates:

Recruiting: Several corporate and (especially) government affiliates stressed their need for new employees. NISS can help in this process by means of the NAIP (§2), which can create contacts between students and potential employers, as well as by simply creating greater awareness among university affiliates of employment opportunities at corporate and government affiliates. Other potential activities include NISS’ assembling resume books and a workshop or other event (possibly at the JSM) to bring students and employers into contact with each other.

Dissertation Topics: Flournoy stated a need for dissertation topics for graduate students in statistics at American University. The affiliates program could assist such processes on a larger scale, by channeling problems from corporate and government to university affiliates.

3.3 Other Items

For completeness and visibility to affiliates not at the meeting, other items raised in FMM but not necessarily captured in the topics selected for breakout discussions included: aviation safety (Ferryman), small area estimation (Eltinge), common scales for language measurement (Stenner), software quality (Dalal), Internet search engines (Dalal), informational retrieval (McCloskey), computer models (Picard, Robinson), industrial process data (Rodriguez), metrology (Coakley), net-

work data (Banks, Dalal) and computer security (Banks, McCloskey). LaVange raised the question of whether NISS could function as a (trusted third party) sharing mechanism for proprietary data.

4 Breakout Discussions

The purpose of these discussions was to identify paths (workshops, tutorials, research proposals, . . .) to approach topics arising in FMM that span multiple affiliate interests (as well as support those of NISS).

4.1 Topics and Groups

Building on FMM, four topic areas were identified:

Bioinformatics and Genomics, a set of issues raised by Soper (proteomics), Taylor and Young, and of interest to all universities represented. *Group Members*: Dunsmuir, Flournoy, LaVange, McCloskey, Nobel, Soper, Young, Wu.

Confidentiality, a continuing concern of Federal agencies (Eltinge, Wright) and an emerging one of many corporations, in such contexts as medical records (LaVange) and post–marketing surveillance (Soper). *Group Members*: Eltinge, Picard, Singh.

Customer Behavior, encompassing problems mentioned by Dalal (Internet quality of service), Eick (Web site analysis, e.g., promotional effectiveness), Robinson (forecasts of demand for specific features, in part from “clickstream” data), Rodriguez (CRM — customer relationship management) and Taylor (assessment of consumer–product relationships). *Group Members*: Dalal, Eick, Robinson, Rodriguez, Wright.

Data Quality, raised by Banks (US DOT legacy databases), Dalal, Eltinge (the relative roles of science and technology), LaVange (medical transaction data), McCloskey (hardware issues), Rodriguez (corporate data warehouses) and Soper (robotic assays). *Group Members*: Banks, Coakley, Ferryman, Lenth.

Although not the subject of a breakout discussion, the topic of **large/massive data sets** also received repeated mention, including by: LaVange (who described organizations “drowning in data”), McCloskey (stressing heterogeneity and scale), Soper (“large d , small n ” problems) and Young.

4.2 Reports

Each breakout discussion group reported results, accompanied by general discussion.

Bioinformatics and Genomics. Young summarized four aspects of the group discussion.

Problems: Microarray analysis and drug design/development.

Steps: Posted data sets, data–centric workshops, visiting experts (at NISS), paired subject matter and statistical sciences experts (potentially, through NISS), and involvement of computer science.

Strategy: Pursuit of multiple grants, including those from organizations that support career development.

Confidentiality. Reflecting human resource concerns raised in FMM (§3.2), Eltinge outlined a possible NISS training component.

Personnel aspects: Postdoctoral or dissertation fellowships that accommodate significant learning curves; adequate mentorship, through NISS, agencies or contractors.

Funding mechanisms: NSF grants; add-ons to agency contracts; private sector (medical and pharmaceutical).

Possible topics concerning disclosure limitation: Public use microdata (from either government or private sources); nominally aggregated analyses.

Cautions included the heterogeneity of technical goals and of organizations' interests, as well as intellectual property considerations.

Customer Behavior. Eick focused on technical issues, including:

- Noisiness of Web data;
- Integration of Web data with other (e.g., demographic) data;
- How to understand and influence customer behavior (and, implicitly, whether E-commerce really is different from other forms of commerce and whether the Web is different from other means of gathering information).

Several of these, especially the first and second, have strong data quality components.

Data Quality. Coakley presented a number of specifically statistical issues associated with data quality:⁴

- Multivariate missing data;
- Statistical planning for data quality;
- Real-time robust methods (for massive data sets) to measure and improve data quality;
- Benchmark data sets;
- Cost aspects of data collection and decisions based on data (What is the impact of data quality?);
- Quality of scientific data;
- Consistency studies (Do multiple approaches yield compatible inferences?); and
- Graphical methods.

Banks volunteered BTS efforts as one component of a case study-oriented data quality workshop.

⁴These echo issues identified in report from the November – December, 2000, workshop on data quality, available at www.niss.org/dqworkshop.html.

5 Wrap-Up

Karr outlined follow-up to the meeting by NISS:

- This report;
- Formation of Affiliate Working Groups to assist and advise NISS in pursuit of the topic areas from §4;
- Implementation of activities such as workshops (both educational and those directed at preparation of proposals), research projects and the NAIP.

Karr closed the meeting by thanking attendees for their participation and support.