Where Are We Going? My Reflections on Beckman and the Past Five Years

Key Message

- **It was the best of times, it was the worst of times**
- **We have unprecedented opportunities**
  - data, data everywhere; everyone needs data management
  - Beckman report made this point very clearly
- **But we also face serious crises**
  - many things in our community are broken; we don't know how to fix
  - Beckman report briefly discussed this, but offered no solution
- **Both got exacerbated in the past five years**
- **If don't take actions soon, we risk becoming irrelevant**
  - will just be one among many communities doing data
  - not even the biggest nor the most influential one
  - there are cautionary tales: communities that are becoming irrelevant

Research

- **Correctly identified Big Data as a big theme**
  - now morphed into data science, which poses big challenges
- **Missed the AI/ML trend**
- **Promoted five directions**
  - scalable data infrastructure
  - diversity in data management
  - end-to-end processing and understanding of data
  - cloud services
  - roles of humans in the data life cycle
- **Made good progress, also branched out**
  - e.g., into AI/ML

Research

- **Beckman predicted the rise of a data-driven world**
- **Correctly observed that this gives us unprecedented opportunities**
  - extremely exciting time for database research
  - golden opportunity for us to play a central role in this emerging world
  - an abundance of research opportunities
- **All of these have been true, but there are deep concerns that we have failed to exploit this wealth of opportunities**
  - while other communities have moved far more decisively

Research

- **We often select problems by how cleanly and quickly we can solve them, not by how important they are**
- **Or we chase buzzword problems**
  - our field often feels very reactive, we seem to have no vision

- **Too many incremental (yet complex) solutions**
- **Same solution templates blindly applied over and over**
  - e.g., “declarative specification / optimization / execution” everywhere
- **Hard to publish innovative solutions**
  - because reviewers are conservative, want perfect papers

- **Too much “la la land” research, far divorced from reality**
  - Things seem to have gotten much worse since Beckman

Reality = Customers with Data Problems

- **There are far more of them now**
  - enterprises, governments, non-profits, citizen scientists
  - domain sciences have become especially hungry customers
- **They need**
  - tools and systems that can be applied to their data problems
  - solution ideas (that they can quickly implement)
  - consulting advice, warm bodies
- **We don't supply much of these**
  - conf reviews seem to suggest our goal is to write perfect papers
- **Other communities have moved into the void**
- **We think they do data management poorly**
  - but with a few exceptions, we just criticize, don't do much

Reality = Customers with Data Problems

- **As a result, customers have gone elsewhere for help**
  - our conferences have some whales, but not dolphins, sardines, etc.
- **My conversation with a Fortune-500 chief data scientist**
  - you guys used to be good, now your conferences feel sleepy
  - I go to KDD, NIPS, ICML, PyData, AnacondaCon, etc. instead
- **My work with 7 domain science teams at UW-Madison**
  - they used tools from PyData, R, and others, barely used ours
- **We can't claim any population as "our customer"**
- **This is worrisome**
  - we need customers to keep us honest, fund our work, train our students, and increase our impacts
  - we seem to get into a negative reinforcement cycle
  - this was barely discussed at Beckman, a big issue now

System/Tool Building

- **Critical for many reasons**
  - our field is empirical by nature
  - needs system/tool building to evaluate research, keep it relevant, help customers, train our students, get funding, etc.
- **Discussed at Beckman, agreed to do more, but no solution**
- **Except a few successes, nothing much has been done**
  - few long-term system building projects (5 years or more)
  - few well-known systems/tools that customers can use
  - little if any work on how to build data systems, share experience
  - no incentives for system work
  - much harder to get papers published
  - system building knowledge is not passed to our junior people
- **Meanwhile other communities are moving into the void**

Education

- **Agreed at Beckman to modernize DB teaching**
  - to catch up with the many changes of DB technology
  - no consensus on how, little progress
- **At grad level, we have problems deciding what to teach**
  - e.g., teaching data cleaning/integration/wrangling is tough
- **Beckman observed an opportunity to influence data science curricula**
  - this has become even more obvious and acute now
  - many universities are designing DS courses, DS ugrad/grad degrees
  - we do have a golden opportunity to influence, but no game plan
    - RDBMSs, Big Data systems, ML, DI, data/system tools?
- **Meanwhile other folks are doing education/training in DS**

Conferences

- **The past ten years give rise to the impression that**
  - our community's goal is to produce perfect papers
  - so we will make authors revise & run all imaginable experiments
- **Results**
  - 3-5 papers accepted outright, rest goes into revision
  - "let's flood the system", revisions are sometimes 10 pages long
  - it takes way too much work now to get a paper in
    (probably twice the amount of work required 10 years ago)
  - and way too long
- **Collectively, as a community**
  we are wasting a colossal amount of time
  - not to talk about re-formatting the rejected paper for a new conf
  - this time can be more productively used for many other purposes

Conferences

- **This drives away our senior people**
  - "too much work, I can better use my limited time for something else"
- **This drives away our junior people**
  - they need to make tenure, so look elsewhere to publish
  - and there are many other data-centric conferences now
  - run the risk that they will permanently abandon our community
- **Other harms if our people can't publish**
  - they can't get promotion (such as to full professors)
  - their influence and ours will be diminished at universities
  - they can't get good students, nor good grants
- **So we are trying to keep our conferences elite, and at the same time harm our junior people in many ways**

And Our Junior Folks Are Going Elsewhere

- **One paper was rejected from SIGMOD**
  - because paper has novel ideas but is not polished
  - same paper accepted at WWW with 2 full accepts / 2 weak accepts
- **Another paper was rejected from VLDB**
  - because proposed system can only handle English text
  - "It is the first time I encounter that argument. Apparently, having a novel solution for English is no longer enough to get a publication in VLDB. I never met that argument in our community or the NLP community. I decided together with xxx to skip SIGMOD and to send it to WWW in Nov."
- **Another paper on named entity extraction**
  - "This paper is an example where I gave up on submitting to VLDB or SIMOD. I sent it directly to EMNLP and got it in the first attempt."
  - "These examples show 1) the rigidity of VLDB/SIGMOD communities and 2) that people may consider skipping them all together, at least me and xxx did it twice."

All of These Make for Sparse Conferences

"Very sad. Half of the attendees have already left on the 3rd day of the conference"

Relationship with Other Communities

- **This is a very tricky time**
- **There are many more data-centric communities**
- **All clamoring for a piece of action**
  - have data needs: e.g., domain sciences
  - or want to help: e.g., data tool building communities
  - or smell money: e.g., domain sciences
- **Example: political fights to control the DS agenda**
- **We do not appear to have a game plan**
  - are we now just ONE among many communities working with data?
  - or are we the dominating one, in what sense?
  - what is our game plan for data science, especially at universities?
  - what is our angle? how can we ensure a seat at the table?
- **These other communities have been quite aggressive**

OK. But …

- Haven't we heard all of these before?
- Sure. They seem more serious now.
- But if we can't do much, then what is the point of wasting our time on them?
- Well. Inaction has a serious price.
- There are cautionary tales, suggesting that it is urgent to act now.

Cautionary Tale: The Statistics Community

- **The goal of their field is great**
  - collect, manage, process, analyze, visualize, and interpret data
- **But for 50 years they focused mostly on narrow topics**
  - develop and reason with mathematical models of data
- **While their elders repeatedly exhorted them to do more**
- **Tukey: "The Future of Data Analysis" (1962)**
  - urged them to reduce focus on statistical theory and engage with the entire data analysis process
  - "We need to face up to more realistic problems"
- **Chambers: "Greater or Lesser Statistics: A Choice for Future Research" (1993)**
  - same recommendation
  - "If statisticians remain aloof, others will act. Statistics will lose."

Cautionary Tale: The Statistics Community

- **Breiman: Statistical Modeling: The Two Cultures (2001)**
  - recommended to focus less on theory and more on data
  - if nothing changes, three major opportunity costs:
    - led to irrelevant theory and questionable scientific conclusions
    - kept statisticians from using more suitable algorithmic models
    - prevented statisticians from working on exciting new problems
  - "Don't go into statistics. Academic statistics may have lost its way."
  - If you take statistics, "[…] remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems."
- **Cleveland: Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics (2001)**
  - recommended pretty much the same thing
- **All of these had relatively little effect**

Cautionary Tale: The Statistics Community

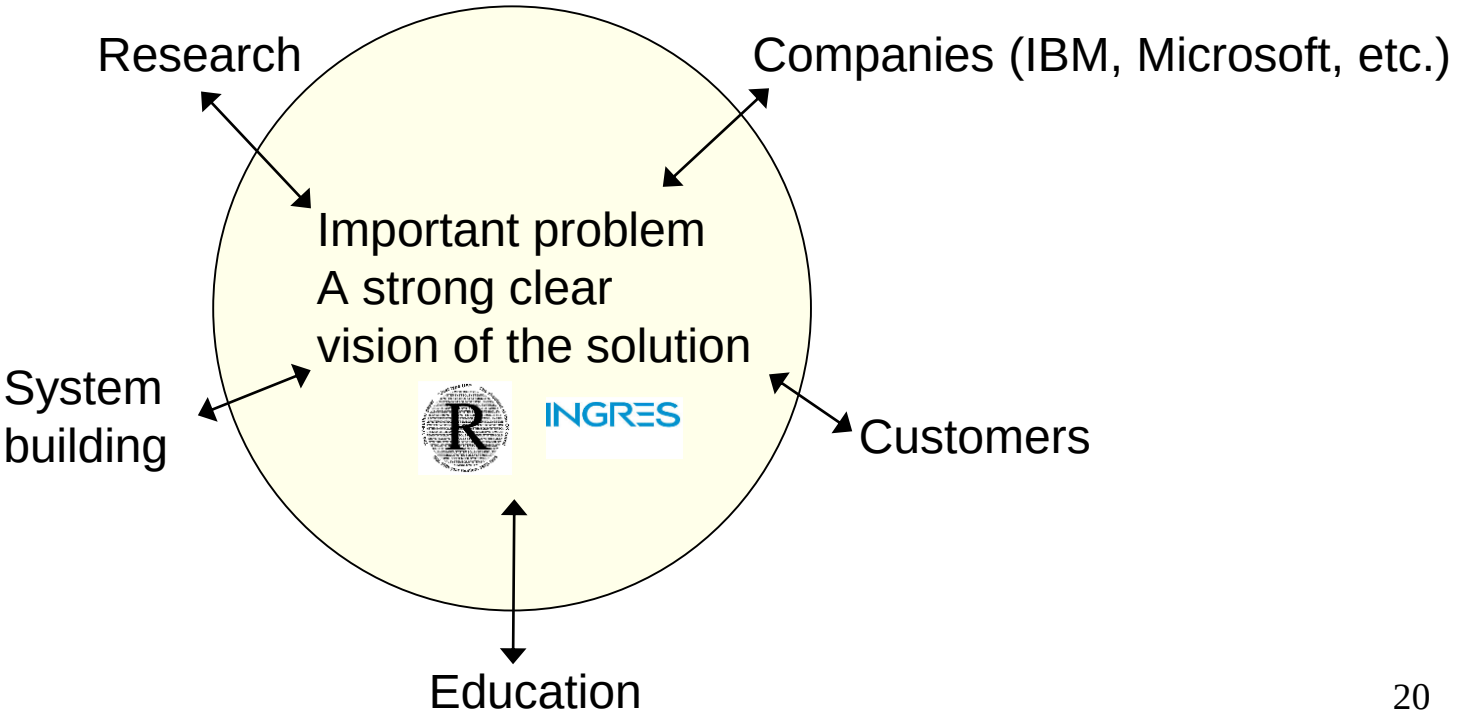
- **Today statistics is having a major identity crisis**
- **Initial "confusion" article headlines …**
  - Aren't we Data Science?
  - A grand debate: is data science just a "rebranding" of statistics?
  - Let us own Data Science
  - Why do we need data science when we've had statistics for centuries?
  - Data science is statistics
- **… Followed by "gloom and doom" headlines**
  - statistics is the least important part of data science
  - data science without statistics is possible, even desirable
  - Data Science: the evolution or the extinction of statistics?
  - The identity of statistics in Data Science
    - "This conversation about data science betrays an anxiety about our (statisticians') identity."
- **Most common solutions: hire more CS people**

Remarkably Similar Complaints

- **From our elders …**
  - "We have become a community that looks for problems with a clean theoretical foundation that beget mathematical solutions, not one that tries to solve important real world problems."
  - "Obviously, this attitude will drive us toward long-term irrelevance."
- **… And from our junior people**
  - "As someone put it to me very recently, VLDB/SIGMOD communities seem to repeat the mistake of the Stats community years back: elitism."
  - "Nowadays, Stats are left behind and DM and ML communities move on with solving big data problems while Stats struggle to make themselves relevant again."
  - "It was a time when papers that were accepted in NIPS could not make it into Stats main conferences. Nowadays, people do not even submit there, but Stats community submit to NIPS lately."

So What Should We Do?

- **What did we do right in our RDBMS days?**
  - worked on the most important data management problem of the time
  - a strong clear vision of what the end-to-end solution should look like
  - with broad community buy-in
- **This created a virtuous cycle, everything "clicked"**



What is the Root Cause of Today Problems?

- **It is not**
  - too many papers: if they all push the field forward, then no problem
  - incremental papers: steady incremental advances are the norm for problem of engineering nature
  - chasing buzzwords: most researchers will chase buzzwords
  - theoretical papers: again, if they push the field forward, then …
  - too little system building: most DB groups do not have resource nor expertise to build systems (but they can extend existing ones)
- **The root cause is likely because**
  - we no longer agree on what important problems to work on
  - **we don't have a strong clear vision with community buy-in**
  - **put differently, the field is having an identity crisis**
  - **in the absence of a vision to follow, people do random stuff**
  - there is no virtuous cycle, things feel "off"

How to Develop a Vision for Our Field?

- **This is surprisingly difficult**
- **Beckman**
  - we are "the community that has traditionally dealt with all things related to data"
- **SIGMOD homepage**
  - "concerned with the principles, techniques and applications of database management systems and data management technology."
- **There are many other data-centric communities now**
  - many seem to have similar goals
    - "gathering and using data to solve interesting and important real world problem" (Breiman, statistics)
  - are we just one of these now? what sets us apart?
- **We need a broad "big tent" vision**
- **But field is too diverse now, for a single vision to work**

Proposal: Empower Sub-Communities

- **Empower folks so that they can build vibrant communities**
  - core DB technology, data integration/wrangling, HILDA/visualization, mobile/spatial/temporal data, AI/ML, etc.
- **Hold them accountable but allow them to experiment on how best to grow their communities**
- **Help them to**
  - focus on important real world problems
  - develop strong clear visions
  - generate community-wide buy-in
  - stay close to customers, police their own conf reviewing process
- **Each of these communities is cohesive enough that we may be able to do the same thing as in RDBMS days**
  - agreement on what to work on, strong clear vision for solutions
  - re-establish the virtuous cycle, obtain community-wide buy-in

Proposal: Empower Sub-Communities

- **Add multiple tracks to SIGMOD**
  - one for each community, with independent PCs
- **SIGMOD chairs help coordinate and empower the PCs**
  - big help, as running a stand-alone conference is a lot of logistic work
- **Tracks operate as "mini-conferences"**
  - free to invite speakers, create panels, solicit industrial papers, and more
- **Allow tracks to**
  - transfer memory between conferences, develop their own visions, enforce their own cultures
- **This can instantly make our conferences more vibrant, reviewing process less random**
- **We don't need to do this in "one shot"**
  - can just experiment with a single new track