Carnegie Mellon University

What would a science of software engineering look like?

Jim Herbsleb

Science of Software Engineering

- Does SE research have impact?
- Science creates impact?
- What sort of science do we need?
- How to move forward?

Does SE Research Have Impact?

No One Seems Confident . . .

- Lee Osterweil, et al, impact project (2008)
  - Bottom line: There is considerable, demonstrable impact in a number of areas, often takes many years, and seems to arise from continued interaction, not tech transfer
- Bertrand Meyer (2010):
  - "many of the advances in software engineering have come out of non-university sources . . . Academic research has had its part, honorable but limited."

Osterweil, L., Ghezzi, C., Kramer, J., & Wolf, A. (2008). Determining the Impact of Software Engineering Research on Practice. Computer, 3(41), 39-49.
Lo, D., Nagappan, N., & Zimmermann, T. (2015). *How Practitioners Perceive the Relevance of Software Engineering Research*. Paper presented at the Symposium on the Foundations of Software Engineering, pp. 415-425.
Briand, L. (2012). Embracing the Engineering Side of Software Engineering. IEEE Software, 4(29), 96.
Meyer, https://bertrandmeyer.com/2010/04/

No One Seems Confident . . .

- Lo, Nagappan, and Zimmerman (2015):
  - "We believe that embedding practitioner feedback into conferences . . . can provide great value to the software engineering community."
- Lionel Briand (2012):
  - SE should be in engineering, not computer science; hard to establish tight collaborations with industry;
  - "Software engineering isn't a branch of computer science; it's an engineering discipline relying in part on computer science, in the same way that mechanical engineering relies on physics."

Osterweil, L., Ghezzi, C., Kramer, J., & Wolf, A. (2008). Determining the Impact of Software Engineering Research on Practice. Computer, 3(41), 39-49.
Lo, D., Nagappan, N., & Zimmermann, T. (2015). *How Practitioners Perceive the Relevance of Software Engineering Research*. Paper presented at the Symposium on the Foundations of Software Engineering, pp. 415-425.
Briand, L. (2012). Embracing the Engineering Side of Software Engineering. IEEE Software, 4(29), 96.
Meyer, https://bertrandmeyer.com/2010/04/

Science Creates Impact?

Jim ←

Likes to mix things up, put them on alcohol flame
See if they catch fire or (YES!) explode
Knows nothing, cares nothing about chemistry

LDA
SVD
SVM
Deep Learning
Etc.

There's not much chemistry going on here!

This may be very useful. This is not science.

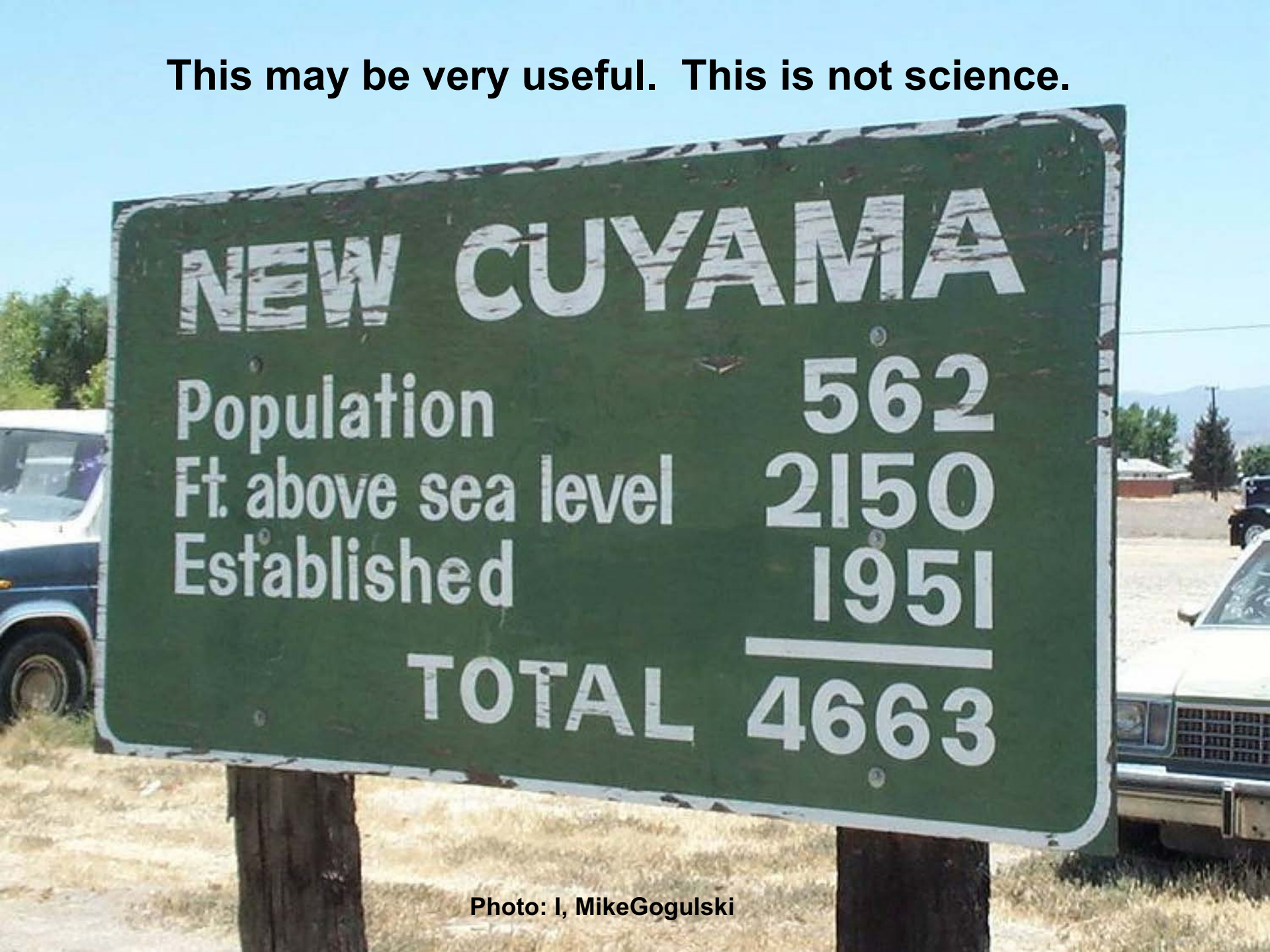


Photo: I, MikeGogulski

Predictive Analytics:
To Bleed or not to Bleed . . .

- Bleeding common medical practice
- Late 18th century
- Francois Joseph Victor Broussais
  - Promoted bleeding of “affected organ”
- Pierre-Charles-Alexandre Louis
  - Actual data collection about outcomes
  - Bleeding is not such a great idea
- The first clinical trial?

Prediction is not Good Enough

- Joseph Lister – outcomes of antiseptic surgery in Edinburgh
  - Mortality rates decreased from 45.7% to 15%
  - Technique based on Louis Pasteur's "germ theory"
- Clinical trial is important, is not enough!
  - Science to understand disease processes
- SAYS NOTHING ABOUT DEVELOPING NEW TREATMENTS!
- Left with trial-and-error

Analgesics . . .

- Tea from willow barks works!
  - A few digestive side effects ☹
- Oak bark doesn't work at all
- Hemlock bark
  - Oops, let's not try that again . . .

Science May Not Have Immediate Application

- Must be freed from demand for immediate applicability
- Suppose medical research demanded that each paper advance practice?
  - Medical research would never have had much impact
  - No germ theory, no understanding of physiological systems, etc.
- Time horizon of years, decades, more
- Gradually build deep, reliable understanding

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).

[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).

[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

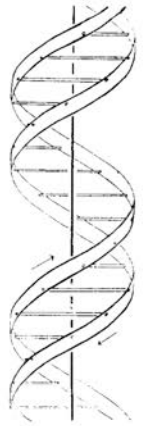
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

The demand for immediate relevance rather than overall contribution . . . a hypothetical rejection letter:

Drs. Watson and Crick:

I regret to inform you that we are unable to accept your paper.

I personally find it very interesting that the DNA molecule has the shape of a double helix held together by paired bases. But the reviewers felt that you have not demonstrated any practical application for this discovery, so it was decided that the contribution was insufficient.

isr institute for SOFTWARE RESEARCH

Science is about Theory

- What are the entities?
- What are the relationships?
- How do these entities and relationships explain the observed phenomena?

Hannay, J. E., Sjoberg, D. I., & Dyba, T. (2007). A systematic review of theory use in software engineering experiments. *IEEE Transactions on Software Engineering, 33*(2), 87-107.

Stol, K.-J., & Fitzgerald, B. (2015). Theory-oriented software engineering. *Science of computer programming, 101*, 79-98.

What sort of science?

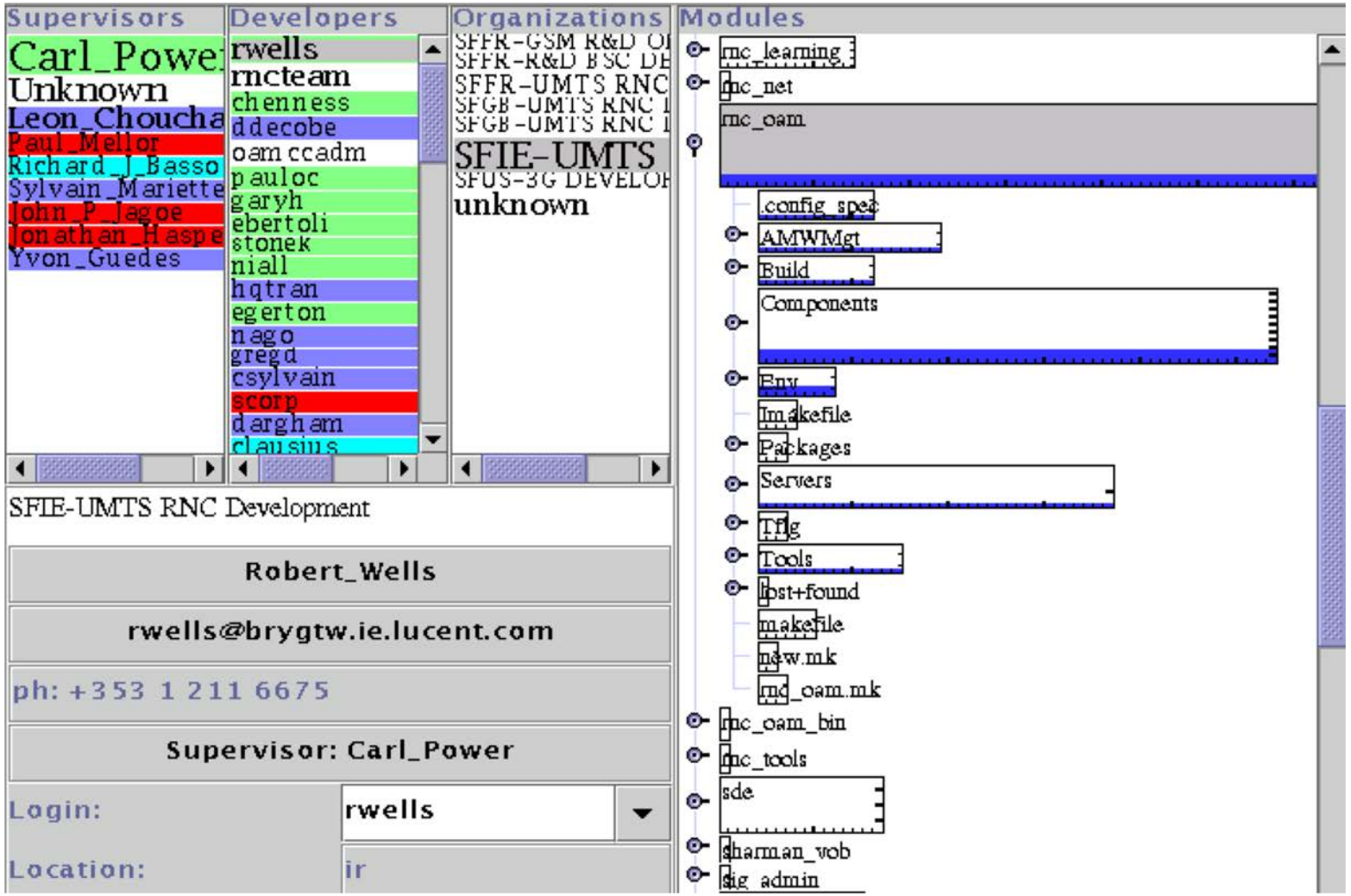
What Science Do We Need?

- Many fields of engineering
  - Need a science to describe, explain, and predict the properties of materials and compositions
- In software engineering
  - What does our science need to do?
  - Our materials are abstractions: programs, patterns, etc.
  - Describe, explain, and predict behavior of artifacts
    - Computer science
  - Describe, explain, and predict behavior of people creating artifacts
    - Human Science of Software Engineering

If Only We Had Known . . .

- Problem: people finding the right experts at a remote site
- Solution: Expertise Browser

Expertise Browser



Mockus, A., & Herbsleb, J.D. (2002). Expertise Browser: A quantitative approach to identifying expertise. In Proceedings of *International Conference on Software Engineering*, Orlando, FL, May 19-25, pp. 503-512.

What Didn't We Know?

- Transactive Memory Systems
- Theory from Organizational Behavior

Transactive Memory Systems (TMS)

- Group level phenomenon
- Arises naturally
- Specialization + index
  - People take responsibility for group knowledge and memory in some area
  - Everyone shares an index of “who knows what”
  - Origins in people watching each other work
- Very powerful impacts on how well groups function

TMS: Benefits and Conditions

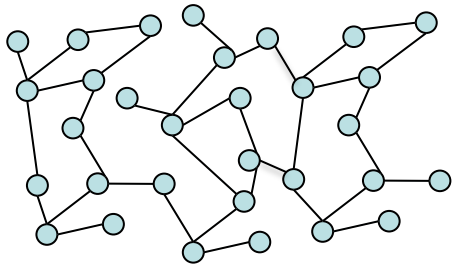
- Specialization gives better performance
- Better coordination, agree on responsibilities
- Facilitates adaptation to new situations or tasks
- Facilitates creativity
- Develops under right conditions
  - Observe each other working
  - Communication

Argote, L. and Ren, Y. Transactive memory systems: A microfoundation of dynamic capabilities. *Journal of Management Studies*, 49, 8 (2012), 1375-1382.

If We Had Known?

- Rather than support isolated search for one individual on one occasion
- Build a system that would effectively provide TMS for the whole organization
- What would we call it?
  - Maybe . . . GitHub?
  - Activity traces, profiles, consistent across repositories

Socio-Technical Coordination



Decisions and Constraints

Technical coordination is a Constraint satisfaction problem (CSP) over decisions

Decisions distributed over people (DCSP)

Social algorithm to solve DCSP

Herbsleb, J.D., & Mockus, A. (2003). Formulation and preliminary test of an empirical theory of coordination in software engineering. In Proceedings, *ACM SIGSOFT Symposium on the Foundations of Software Engineerin*g, Helsinki, Finland, September 1-5, pp. 112-121

Herbsleb, J.D., Mockus, A., Roberts, J.A. (2006). Collaboration in Software Engineering Projects: A Theory of Coordination. *International Conference on Information Systems*, Milwaukee, WI.

Distributed Constraint Satisfaction

- Decisions are represented as *n* variables $x_1, x_2, \ldots, x_n$
- Values from finite, discrete domains $D_1, D_2, \ldots, D_n$.
- A set of constraints that operate over the variables serve to limit possible values that can be assigned to other variables.
- Formally, constraints $p_k(x_{k1}, x_{k2}, \ldots, x_{kn})$ can be represented as predicates defined on the Cartesian product $D_{k1} \times D_{k2} \times \ldots \times D_{kj}$.
- *Distributed* constraint satisfaction problem, two relations
- Each variable x_j belongs to one agent *i*, represented as the relation *belongs*(x_j, i).
- Agents only know about a subset of the constraints:
- *known*(P_l, k), meaning agent *k* knows about constraint P_l.

Herbsleb, J.D., & Mockus, A. (2003). Formulation and preliminary test of an empirical theory of coordination in software engineering. In Proceedings, *ACM SIGSOFT Symposium on the Foundations of Software Engineering*, Helsinki, Finland, September 1-5, pp. 112-121

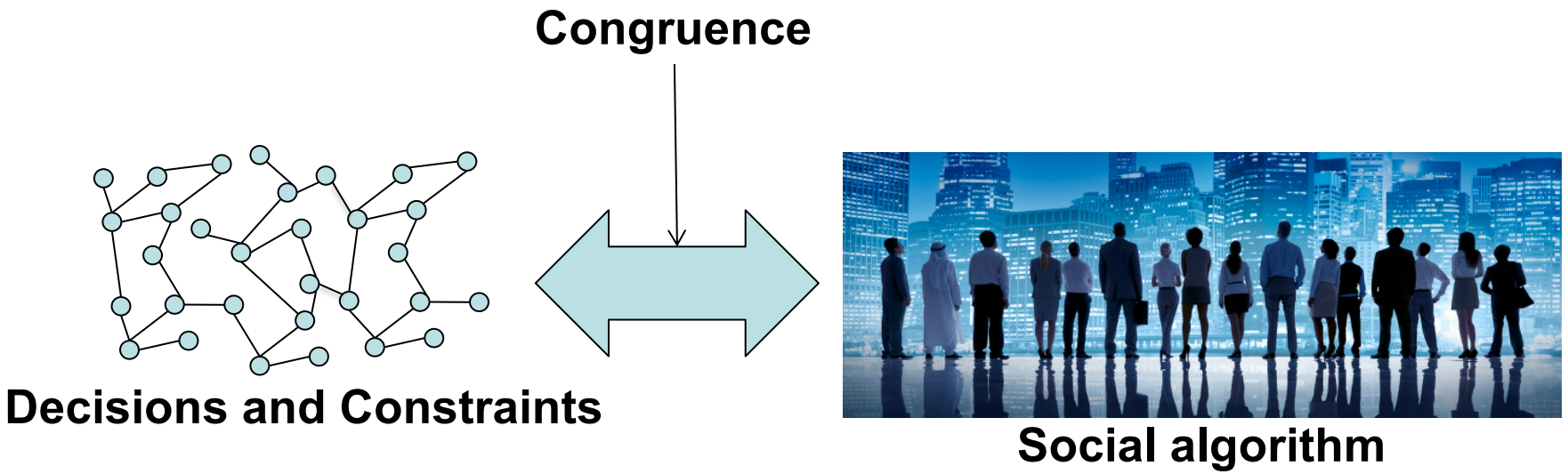
Herbsleb, J.D., Mockus, A., Roberts, J.A. (2006). Collaboration in Software Engineering Projects: A Theory of Coordination. *International Conference on Information Systems*, Milwaukee, WI.

Yokoo, M. Distributed Constraint Satisfaction: *Foundations of Cooperation in Multi-agent Systems*. Springer, New York, 2001.

Solving a DCSP

- Computational agents' actions
  - Make decisions, backtrack
  - Send message (decision, constraint)
  - Create link (change network topology)
  - Edit a shared object
  - Predict other agents' behavior
- When agents are human
  - Execute a social algorithm

Socio-Technical Coordination

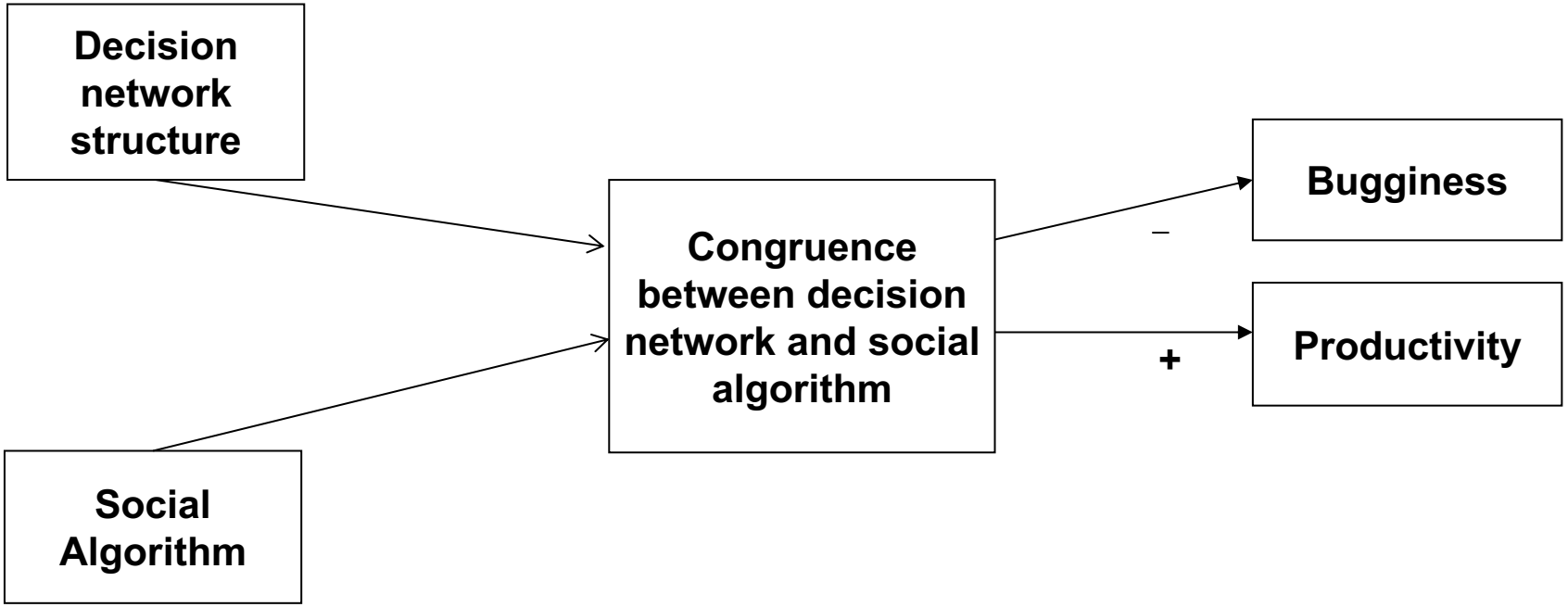


Cataldo, M., Wagstrom, P. A., Herbsleb, J. D. and Carley, K. M. (2006). Identification of coordination requirements: implications for the Design of collaboration and awareness tools. In Proceedings, Computer supported cooperative work, Banff, Alberta, Canada, pp. 353-362.

Cataldo, M., Herbsleb, J. D. and Carley, K. M. (2008). Socio-Technical Congruence: A Framework for Assessing the Impact of Technical and Work Dependencies on Software Development Productivity. In Proceedings, International Symposium on Empirical Software Engineering and Measurement, Kaiserslautern, Germany, pp. 2-11.

Cataldo, M. and Herbsleb, J. D. Coordination Breakdowns and Their Impact on Development Productivity and Software Failures. IEEE Transactions on Software Engineering 39, 3 (2013), 343-360.

Validated Congruence Model



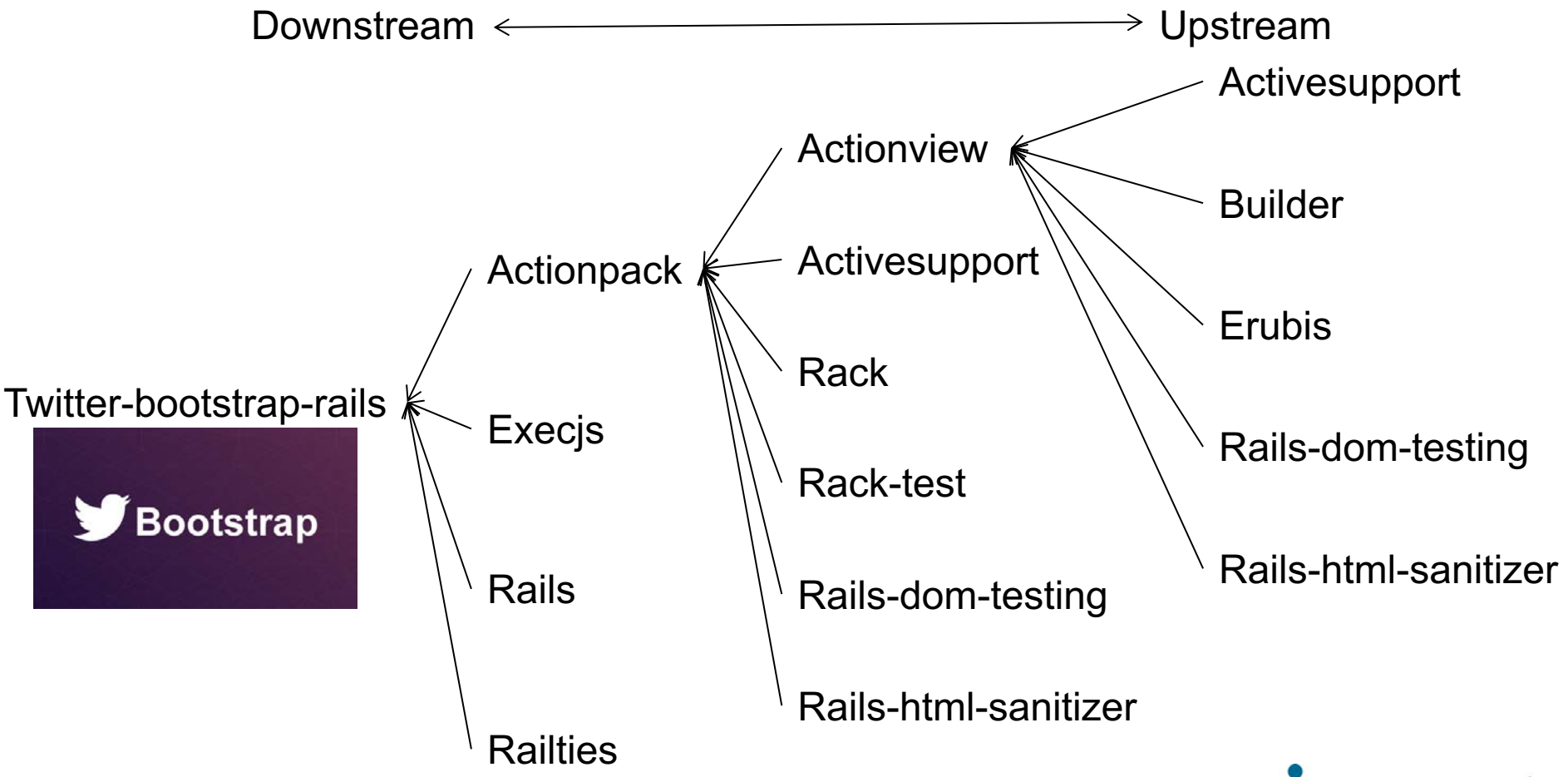
Many Questions Remain . . .

- We only showed that for a few types of social algorithm, it works when the right people use it
- What about match of mechanisms to dependency types?
- What about match of mechanisms to decision pace?

Scale Up . . .

- Looked at coordination in relatively small tasks (a few people, 1-2 weeks)
- How about coordination across an ecosystem?

Dependency Graph



Socio-Technical Ecosystems

- Constraints: changes that break code
- Study showed several different social algorithms
  - Snapshot consistency (R/CRAN)
  - Rigid backward compatibility (Eclipse)
  - Semantic versioning (node.js/npm)

Bogart, C., Kästner, C., Herbsleb, J., & Thung, F. (2016). *How to Break an API: Cost Negotiation and Community Values in Three Software Ecosystems*. Paper presented at the Foundations of Software Engineering, Seattle, WA.

The Science We Need

- Software engineering is in need of a science beyond computer science
- I nominate “human science of software engineering” to fill the role
- We are moving in this direction anyway, let’s acknowledge it and speed it up!

How to move forward?

Barriers to Human Science

- The universal principle of interdisciplinary contempt
- DPHB* principle: everything I don't understand is simple
- Intellectual worth is evaluated on a single dimension
  - From math to BS
- Not all statistical models are just about prediction
  - Theory seen as mere decoration and distraction on top of statistical model
  - Statistics used to test relations between theoretical constructs
  - Not just associations among variables
- Border defense, antibodies
  - Is that really computer science?
- Necessity to argue for practical application of each result

*Dilbert's pointy-haired boss

What Next?

- I'm "preaching to the choir" in this room
- The kinds of things we are all doing are the future of the field
- Remember, science is for the longer term, years, decades, generations
- Push back on demand for immediate impact!
- Make theory central!
- Push for funding a portfolio of research

Q&A