

**Software engineering:** the application of a systematic disciplined quantifiable approach to the development, operation and maintenance of software; that is, the application of engineering to software.

**That is: INPUT** is people, equipment, tools, methods, schedule, requirements.

**OUTPUT:** product and documentation

**PROCESS:** Manage risk to be successful.

**But we don't really understand risk:** "possibility of loss or injury." So there has to be a possibility of An event occurring and there has to be a loss if the event occurs.

**Most managers don't understand risk or are not willing to defend their decision if a risk analysis determines a project will fail. E.g., in an engineering discipline, an engineer may not undertake a project if professional judgment says it will fail. Software managers rarely make this decision.**

**Some risk analysis is subjective, so there may not be a "right" answer. But there should at least be an indication of what the risk is.**

# **EXPERIMENTAL VALIDATION IN SOFTWARE ENGINEERING**

- **Lots of technology development**
- **Rapid change today within our technological society**
- **But software failures are all too common**
- **Why such failures?**

**Often there is a lack of validation before using a new technology**

- **Anecdotal evidence that we don't validate our claims**
- **Study by Tichy (1995) that 50% of software engineering papers do not have validation;**
- **Only 15% in other scientific fields**

**We need measurements (can't have a software engineering course without this comment):**

**"I often say that when you can measure what you are speaking about, and express it in numbers, you can know something about it. But when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind." – Lord Kelvin**

## **But we also need relevant measurements:**

**“The government is very keen on amassing statistics – they collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn pleases.” – British economist Josiah Stamp, 1929**

## **Purpose of measurement?**

**“All we can ask of a theory is to predict the results of events that can be measured. This sounds like an obvious point, but forgetting it leads to the so-called paradoxes that popular writers of our culture are fond of exploiting.” – Leon Lederman, Nobel Laureate physicist**

## **What is science?**

**“Learn from science that you must doubt the experts. ... Science is the belief in the ignorance of experts.” – Richard Feynman, Nobel Laureate physicist**

## **But in Computer Science:**

- **Our theories are our tools and techniques**
- **All too often, we don't appreciate the “science” in our title**
- **Validation, experimentation, and measurement seem to be lacking**

## **Experimental Models for Software Research**

- **Recognition that we need to understand how to experiment in software engineering**
- **Problems:**
  - ◆ **Models mostly taken from social science domain**
  - ◆ **View experimentation as the replication of a hypothesis under varying controlled conditions**
- **Can we take larger view of experimentation that applies in the software domain?**

# What are Experiments?

## Different models:

- **Replicated experiments**
  - ◆ **Chemistry – Rows of test tubes**
  - ◆ **Psychology – Rows of freshmen students working on a task**
  
- **Observations of what happens**
  - ◆ **Medicine – Clinical trials**
  - ◆ **Astronomy – Observe events if and when they occur**
  
- **Data Mining of completed activities**
  - ◆ **Archaeology – Dig up the past**
  - ◆ **Forensic investigations – recreate what happened**

**How do these relate to Software?**

**What data does each method generate?**

# Basic Data Collection Models

## Impact on the process being studied:

- **Active methods** – An effect on the process being studied
- **Passive methods** – No effect on process being studied

## Classes of methods:

- **Controlled method** – Multiple instances of an observation in order to provide for statistical validity of the results. (Usually an active method.)
- **Observational method** – Collect relevant data as it develops. In general, there is relatively little control over the development process. (Weakly active, although may be passive.)
- **Historical method** – Collect data from completed projects. (Passive methods.)

These three basic methods have been classified into 12 data collection models.

(We will also consider one theoretical validation method, yielding 13 validation methods)

# Controlled Methods

**Replicated – Several projects are observed as they develop (e.g., in industry) in order to determine the effects of the independent variable. Due to the high costs of such experiments, they are extremely rare.**

**Synthetic environments -- These represent replicated experiments in an artificial setting, e.g., often in a university.**

**Dynamic analysis – The project is replicated using real project data.**

**Simulation – The project is replicated using artificial project data.**

**The first 2 of these generally apply to process experiments while the last two generally apply to product experiments.**

# **Observational Methods**

**Project monitoring – Collect data on a project with no preconceived notion of what is to be studied.**

**Case study – Data collected as a project develops by individuals who are part of the development group. (Often used in SEL.)**

**Field Study – An outside group collects data on a development. (A weaker form of case study.)**

## **Historical Methods:**

**Literature search – Review previously published papers in order to arrive at a conclusion. (e.g., Meta-analysis --- combining results from separate related studies)**

**Legacy data – Data from a completed project is studied in order to determine results.**

**Lessons-learned data – Interviews with project personnel and a study of project documentation from a completed project can be used to determine qualitative results. (A weak form of legacy data.)**

**Static analysis – Artifacts of a completed project are processed to determine characteristics.**

# But List of Methods is Incomplete

- **Assertions: What do software engineers often do?**
  - ◆ **For a new technology validation often consists of:  
“I tried it and I like it”**
  - ◆ **Validation often consists of a few trivial examples of using the technology to show that it works.**
  - ◆ **Added this validation as a weak form of case study under the “Observational Method:”**
  - ◆ **Assertion – A simple form of case study that does not meet rigorous scientific standards of experimentation.**
- **Theoretical validation – A form of validation based upon mathematical proof.**
- **Summary: 13 methods**
  - ◆ **11 experimental methods**
  - ◆ **assertion (weak experimental validation)**
  - ◆ **theoretical validation**

# Evaluation of this classification

## Review of 1995 Tichy study:

- Reviewed 403 papers
- Sources: ACM journals and conferences, IEEE TSE
- Classification of papers

Formal theory	Proofs
Design and modeling	Designs which are not formal
Empirical study	Evaluation of existing technology
Hypothesis testing	Experiments to test a hypothesis
Other	Anything else, e.g., surveys

## Conclusions from Tichy study

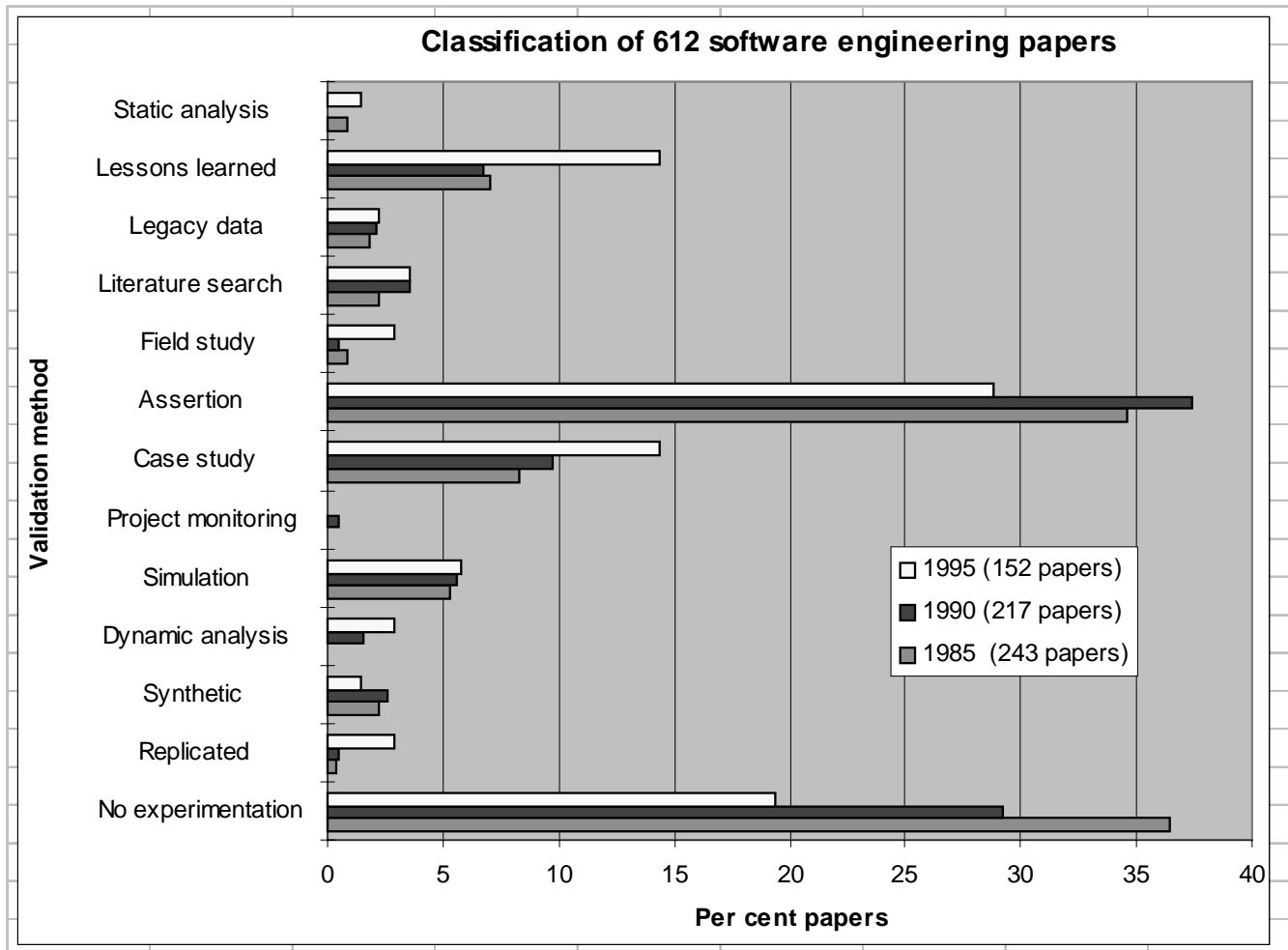
### Those relevant to current study:

- 40% of computer science papers without validation
- 50% of software engineering papers without validation
- Comparable numbers are neuroscience (12%) and optical engineering (15%)
- But only considered design and modeling papers. Perhaps too narrow a view

# **NIST Evaluation –**

- **Performed by Zelkowitz and Dolores Wallace**
  
- **New literature search:**
  - ◆ **Papers published in 1985, 1990, 1995**
  - ◆ **Sources:**
    - **IEEE Software --- a technical magazine**
    - **Transactions on Software Engineering --- an archival research journal**
    - **ICSE proceedings --- a conference**
  - ◆ **612 papers reviewed**
  - ◆ **Can we detect changing trends over 10 years?**
  
- **Added 2 more classifications to above 13:**
  - ◆ **Not applicable --- The paper does not discuss a new technology, e.g., a survey paper.**
  
  - ◆ **No experimentation --- The paper presents a new technology, but makes no claims as to experimental validity. These are the papers that SHOULD have validation of some form.**

SUMMARY TOTALS Method	85			90			95			Ttl
	ICSE	Soft	TSE	ICSE	Soft	TSE	ICSE	Soft	TSE	
Not applicable	6	6	3	4	16	2	5	7	1	50
Theoretical	3	1	18	1	0	19	3	0	7	52
No experimentation	13	10	38	7	8	22	7	3	7	115
Replicated	1	0	0	0	0	1	1	0	3	6
Synthetic	3	1	1	0	1	4	0	0	2	12
Dynamic analysis	0	0	0	0	0	3	0	0	4	7
Simulation	2	0	10	0	0	11	1	1	6	31
Project monitoring	0	0	0	0	1	0	0	0	0	1
Case study	5	2	12	7	6	6	4	6	10	58
Assertion	12	13	54	12	19	42	4	14	22	192
Field study	1	0	1	0	0	1	1	1	2	7
Literature search	1	1	3	1	5	1	0	3	2	17
Legacy data	1	1	2	2	0	2	1	1	1	11
Lessons learned	7	5	4	1	4	8	5	7	8	49
Static analysis	1	0	1	0	0	0	0	0	2	4
<b>Yearly totals</b>	<b>56</b>	<b>40</b>	<b>147</b>	<b>35</b>	<b>60</b>	<b>122</b>	<b>32</b>	<b>43</b>	<b>77</b>	<b>612</b>



# Quantitative Observations

- **Most prevalent validation mechanisms were lessons learned and case studies, each about 10%**
- **Simulation was used in about 5% of the papers, while the remaining techniques were each used in under 3% of the papers**
- **About one-fifth of the papers had no experimental validation**
- **Assertions (a weak form of validation) were about one-third of the papers**
- **But percentages of no experimentation dropped from 26.8% in 1985 to 19.0% in 1990 to only 12.2% in 1995. (Perhaps a favorable trend?)**

# Qualitative Observations

- **We were able to classify every paper according to our 13 categories, although somewhat subjective (e.g., assertion versus case study).**
- **Some papers can apply to 2 categories. We chose what we believed to be the major evaluation category.**
- **Authors often fail to clearly state what their paper is about. Its hard to classify the validation if one doesn't know what is being validated.**
- **Authors fail to state how they propose to validate their hypotheses.**
- **Terms (e.g., experiment, case study, controlled experiment, lessons learned) are used very informally.**
- **MAJOR CAVEAT: The papers that appear in a publication are influenced by the editor of that publication or program committee. The editors and program committees from 1985, 1990, and 1995 were all different. This then imposes a confounding factor in our analysis process that may have affected our outcome.**

# Overall Observations

- **Many papers have no experimental validation at all (about one-fifth), but fortunately, this number seems to be dropping.**
- **BUT too many papers use an informal (assertion) form of validation. Better experimental design needs to be developed and used.**
- **Lessons learned and case studies each are used about 10% of the time, the other techniques are used only a few percent at most.**
- **Terminology of how one experiments is sloppy. We hope a classification model, such as ours, can help to encourage more precision in the describing of empirical research.**

# Comparison to Other Fields

- **We decided to look at several other disciplines for comparison, An informal study. No attempt at choosing the “best” journal in each field.**
  
- **Journals:**
  - ◆ **J – Measurement Science and Technology, (Devices to perform measurements)**
  - ◆ **J 2 – American Journal of Physics, (Theory and application of new physical theories)**
  - ◆ **J 3 – Journal of Research of NIST, (Research on measurement and standardization issues)**
  - ◆ **J 4 – Management Science, (Queueing theory and scheduling problems)**
  - ◆ **J 5 – Behavior Therapy, (Clinical therapies)**
  - ◆ **J 6 – Journal of Anthropological Research, (Study of human cultures)**

<b>Method</b>	<b>J1</b>	<b>J2</b>	<b>J3</b>	<b>J4</b>	<b>J5</b>	<b>J6</b>	<b>TTL</b>	<b>%</b>
	<b>%</b>	<b>%</b>	<b>%</b>	<b>%</b>	<b>%</b>	<b>%</b>		
<b>NA</b>		<b>2</b>	<b>5</b>			<b>1</b>	<b>8</b>	<b>---</b>
<b>None</b>	<b>16</b>	<b>58</b>	<b>7</b>	<b>21</b>	<b>6</b>	<b>31</b>	<b>26</b>	<b>20</b>
<b>Replicated</b>		<b>5</b>	<b>4</b>	<b>4</b>	<b>12</b>		<b>5</b>	
<b>Synthetic</b>			<b>4</b>	<b>11</b>	<b>29</b>		<b>9</b>	
<b>Dynamic anal.</b>	<b>32</b>	<b>5</b>	<b>19</b>	<b>11</b>			<b>17</b>	
<b>Simulation</b>			<b>15</b>	<b>32</b>			<b>13</b>	
<b>Proj. Mon.</b>								
<b>Case study</b>	<b>40</b>	<b>16</b>	<b>41</b>		<b>6</b>	<b>8</b>	<b>26</b>	
<b>Assertion</b>	<b>8</b>	<b>4</b>	<b>11</b>			<b>8</b>	<b>7</b>	<b>5</b>
<b>Field study</b>				<b>4</b>	<b>18</b>		<b>4</b>	
<b>Liter. Search</b>	<b>4</b>	<b>11</b>	<b>7</b>	<b>7</b>	<b>24</b>	<b>23</b>	<b>14</b>	
<b>Legacy data</b>					<b>6</b>	<b>23</b>	<b>4</b>	
<b>Lessons learn</b>		<b>5</b>				<b>8</b>	<b>2</b>	
<b>Static anal.</b>								
<b>Paper count(#)</b>	<b>25</b>	<b>21</b>	<b>32</b>	<b>28</b>	<b>17</b>	<b>14</b>	<b>137</b>	

**Note clustering of techniques across journals**

**No attempt to summarize across fields, except for experimentation and assertions**

# Results from Other Fields

- **No experimentation plus assertion data much lower than in software engineering (25% versus 55%)**
- **Each field has a characteristic data collection model:**
  - ◆ **Physics --- dynamic analysis and simulation (repeated experiments)**
  - ◆ **Psychology --- replicated and synthetic (repeated trials of individuals)**
  - ◆ **Anthropology --- legacy data (historical data)**
- **Literature search more accepted model for publication. (Does this refer to publication of similar studies that are frowned upon in computer science?)**

## **In conclusion ...**

- **We have proposed a 13-way approach toward developing a quantitative model of software experimentation. It seems applicable to the software engineering literature.**
- **In a 1992 report from the National Research Council the Panel on Statistical Issues and Opportunities for Research in the Combination of Information recommended:**

**“The panel urges that authors and journal editors attempt to raise the level of quantitative explicitness in the reporting of research findings, by publishing summaries of appropriate quantitative measures on which the research conclusions are based ...”**

- **In general, software engineering experimental validation is probably not as bad as folklore says, but could stand to do a better job.**

## **WHY DOESN'T INDUSTRY "BUY" THIS?**

### **Industry:**

- **Ignores results from archival journals**
- **Believes in unsubstantiated rumors**

### **Research community:**

- **Doesn't require validation**
- **Doesn't perform validations as thorough as necessary**

**There is a "disconnect" between these 2 cultures**

### **Homework assignment:**

- **How hard is it to do these evaluations?**
- **How effective are they in answering these necessary questions?**

**For each method, write down the difficulty in using that method to answer the proposed question.**

**Do not discuss this with other class members. There is no "right" answer since this is an evaluation of your subjective opinions on the effectiveness of each technique.**

**Class results will be summarized and discussed in a few weeks.**