

# Metrics and Measurement

17-313: Foundations of Software Engineering

<https://cmu-313.github.io>

Michael Hilton and Josh Sunshine

Spring 2026

# Administrivia

- Project 1(b) is due Friday, January 23, 2026 @ 11:59PM
- If you haven't: **PLEASE FILL OUT TEAMWORK SURVEY!**
- Get started early, ask for help, and check the **#technical-support** channel; chances are your questions have been asked by others!

# Smoking Section

- Last full row



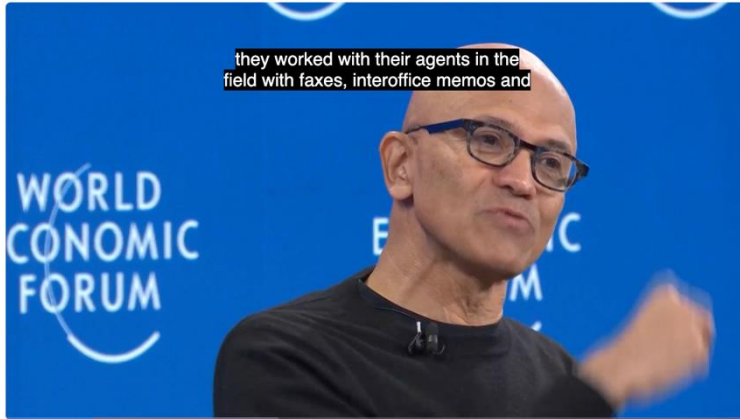
# Learning Goals

- Explain the importance of measurement and metrics in Software Engineering
- Provide examples of metrics for software qualities and process
- Apply goal-based frameworks for decision making using metrics
- Identify the limitations and dangers of decisions and incentives based on measurements

# From Today's news:

BENZINGA

**Microsoft CEO Satya Nadella Warns AI Could Lose 'Social Permission' If It Burns Energy Without Real-World Benefits In Health And Education**



- “We will quickly lose even the social permission to actually take something like energy, which is a scarce resource, and use it to generate these tokens,” he said. “If these tokens are not improving health outcomes, education outcomes, public sector efficiency, private sector competitiveness across all sectors, small and large.”

# Measurement in everyday life

- Economics
  - price, inflation rate, stock price, volume
- Medicine
  - heart rate, blood pressure, body temperature, ECG
- Engineering
  - force, torque, heat transfer coefficient, thermal efficiency
- Natural sciences
  - AQI, carbon footprint, soil pH



NS New Scientist

## Ants use pedometers to find home

An experiment that involves attaching stilts to ants' legs reveals that the insects somehow keep a record of how many steps they take.

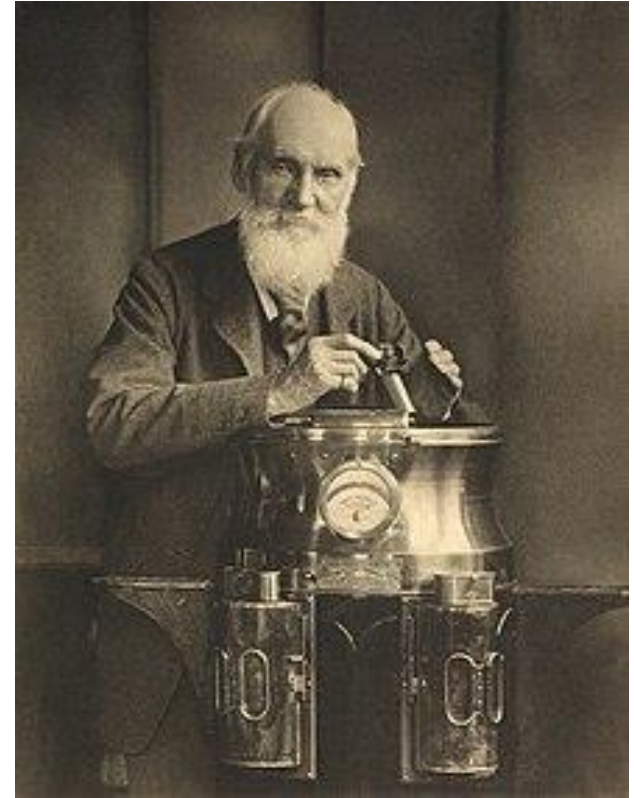
Jun 29, 2006



***“To measure is to know; if  
you can not measure it, you  
can not improve it”***

William Thomson, Lord Kelvin

$$K = \left( \frac{5}{9} (F - 32) \right) + 273.15$$



# Software Engineering

Principles and practices (technical and non-technical) for confidently building **high-quality software**

**What does this mean?  
How do we know?**

**We need metrics  
and measurement!**





## Executing our safety case

A safety case is a logical argument, supported by evidence, intended to justify that a system is acceptably safe for a specific application in a specific operating environment. A structured argument includes a specific claim—such as that our self-driving vehicles are acceptably safe to operate on public roads—that is broken down into sub-claims, which are then ultimately supported by evidence. We believe a safety case is the most effective and efficient path to safe driverless operations, and is imperative for any company looking to safely deliver commercial-ready self-driving vehicles at scale.

The diagram illustrates the safety analysis process for an infusion pump, structured as a Goal Structing Notation (GSN) diagram. It begins with the identification of a hazard (J0) and its elimination (S0). This leads to a goal (G1) to describe the hazard in terms of air in the line. The process then moves to identifying hazard controls (J1) and analyzing them (A1). If the controls are not effective, corrective controls are identified (G2) and analyzed (A2). The process concludes with the completion of the safety analysis (A3). The diagram includes various artifacts such as hazard logs, system description discs, medical procedures, and training manuals, which are used and produced at different stages of the analysis.

```

graph TD
    J0([J0 One hazard identified in this simplified argument]) --> S0[/S0 Safety demonstrated by mitigating infusion hazards/]
    S0 --> G1[/G1 Air in line hazard description/]
    G1 --> J1([J1 Hazard controls have been identified during adequate hazard analysis])
    J1 --> A1[/A1 Safety analysis of the hazard is complete/]
    A1 --> G2[/G2 Corrective controls are effective/]
    G2 --> A2[/A2 Corrective controls have been identified during adequate hazard analysis/]
    A2 --> G3[/G3 Corrective controls are effective/]
    G3 --> A3[/A3 Safety analysis of the hazard is complete/]
    A3 --> J2([J2 Defect controls decomposition])
    J2 --> S1[/S1 Defect controls decomposition/]
    S1 --> G4[/G4 Defect controls decomposition/]
    G4 --> A4[/A4 Defect controls decomposition/]
    A4 --> J3([J3 Defect controls decomposition])
    J3 --> S2[/S2 Defect controls decomposition/]
    S2 --> G5[/G5 Defect controls decomposition/]
    G5 --> A5[/A5 Defect controls decomposition/]
    A5 --> J4([J4 Defect controls decomposition])
    J4 --> S3[/S3 Defect controls decomposition/]
    S3 --> G6[/G6 Defect controls decomposition/]
    G6 --> A6[/A6 Defect controls decomposition/]
    A6 --> J5([J5 Defect controls decomposition])
    J5 --> S4[/S4 Defect controls decomposition/]
    S4 --> G7[/G7 Defect controls decomposition/]
    G7 --> A7[/A7 Defect controls decomposition/]
    A7 --> J6([J6 Defect controls decomposition])
    J6 --> S5[/S5 Defect controls decomposition/]
    S5 --> G8[/G8 Defect controls decomposition/]
    G8 --> A8[/A8 Defect controls decomposition/]
    A8 --> J7([J7 Defect controls decomposition])
    J7 --> S6[/S6 Defect controls decomposition/]
    S6 --> G9[/G9 Defect controls decomposition/]
    G9 --> A9[/A9 Defect controls decomposition/]
    A9 --> J8([J8 Defect controls decomposition])
    J8 --> S7[/S7 Defect controls decomposition/]
    S7 --> G10[/G10 Defect controls decomposition/]
    G10 --> A10[/A10 Defect controls decomposition/]
    A10 --> J9([J9 Defect controls decomposition])
    J9 --> S8[/S8 Defect controls decomposition/]
    S8 --> G11[/G11 Defect controls decomposition/]
    G11 --> A11[/A11 Defect controls decomposition/]
    A11 --> J10([J10 Defect controls decomposition])
    J10 --> S9[/S9 Defect controls decomposition/]
    S9 --> G12[/G12 Defect controls decomposition/]
    G12 --> A12[/A12 Defect controls decomposition/]
    A12 --> J11([J11 Defect controls decomposition])
    J11 --> S10[/S10 Defect controls decomposition/]
    S10 --> G13[/G13 Defect controls decomposition/]
    G13 --> A13[/A13 Defect controls decomposition/]
    A13 --> J12([J12 Defect controls decomposition])
    J12 --> S11[/S11 Defect controls decomposition/]
    S11 --> G14[/G14 Defect controls decomposition/]
    G14 --> A14[/A14 Defect controls decomposition/]
    A14 --> J13([J13 Defect controls decomposition])
    J13 --> S12[/S12 Defect controls decomposition/]
    S12 --> G15[/G15 Defect controls decomposition/]
    G15 --> A15[/A15 Defect controls decomposition/]
    A15 --> J14([J14 Defect controls decomposition])
    J14 --> S13[/S13 Defect controls decomposition/]
    S13 --> G16[/G16 Defect controls decomposition/]
    G16 --> A16[/A16 Defect controls decomposition/]
    A16 --> J15([J15 Defect controls decomposition])
    J15 --> S14[/S14 Defect controls decomposition/]
    S14 --> G17[/G17 Defect controls decomposition/]
    G17 --> A17[/A17 Defect controls decomposition/]
    A17 --> J16([J16 Defect controls decomposition])
    J16 --> S15[/S15 Defect controls decomposition/]
    S15 --> G18[/G18 Defect controls decomposition/]
    G18 --> A18[/A18 Defect controls decomposition/]
    A18 --> J17([J17 Defect controls decomposition])
    J17 --> S16[/S16 Defect controls decomposition/]
    S16 --> G19[/G19 Defect controls decomposition/]
    G19 --> A19[/A19 Defect controls decomposition/]
    A19 --> J18([J18 Defect controls decomposition])
    J18 --> S17[/S17 Defect controls decomposition/]
    S17 --> G20[/G20 Defect controls decomposition/]
    G20 --> A20[/A20 Defect controls decomposition/]
    A20 --> J19([J19 Defect controls decomposition])
    J19 --> S18[/S18 Defect controls decomposition/]
    S18 --> G21[/G21 Defect controls decomposition/]
    G21 --> A21[/A21 Defect controls decomposition/]
    A21 --> J20([J20 Defect controls decomposition])
    J20 --> S19[/S19 Defect controls decomposition/]
    S19 --> G22[/G22 Defect controls decomposition/]
    G22 --> A22[/A22 Defect controls decomposition/]
    A22 --> J21([J21 Defect controls decomposition])
    J21 --> S20[/S20 Defect controls decomposition/]
    S20 --> G23[/G23 Defect controls decomposition/]
    G23 --> A23[/A23 Defect controls decomposition/]
    A23 --> J22([J22 Defect controls decomposition])
    J22 --> S21[/S21 Defect controls decomposition/]
    S21 --> G24[/G24 Defect controls decomposition/]
    G24 --> A24[/A24 Defect controls decomposition/]
    A24 --> J23([J23 Defect controls decomposition])
    J23 --> S22[/S22 Defect controls decomposition/]
    S22 --> G25[/G25 Defect controls decomposition/]
    G25 --> A25[/A25 Defect controls decomposition/]
    A25 --> J24([J24 Defect controls decomposition])
    J24 --> S23[/S23 Defect controls decomposition/]
    S23 --> G26[/G26 Defect controls decomposition/]
    G26 --> A26[/A26 Defect controls decomposition/]
    A26 --> J25([J25 Defect controls decomposition])
    J25 --> S24[/S24 Defect controls decomposition/]
    S24 --> G27[/G27 Defect controls decomposition/]
    G27 --> A27[/A27 Defect controls decomposition/]
    A27 --> J26([J26 Defect controls decomposition])
    J26 --> S25[/S25 Defect controls decomposition/]
    S25 --> G28[/G28 Defect controls decomposition/]
    G28 --> A28[/A28 Defect controls decomposition/]
    A28 --> J27([J27 Defect controls decomposition])
    J27 --> S26[/S26 Defect controls decomposition/]
    S26 --> G29[/G29 Defect controls decomposition/]
    G29 --> A29[/A29 Defect controls decomposition/]
    A29 --> J28([J28 Defect controls decomposition])
    J28 --> S27[/S27 Defect controls decomposition/]
    S27 --> G30[/G30 Defect controls decomposition/]
    G30 --> A30[/A30 Defect controls decomposition/]
    A30 --> J29([J29 Defect controls decomposition])
    J29 --> S28[/S28 Defect controls decomposition/]
    S28 --> G31[/G31 Defect controls decomposition/]
    G31 --> A31[/A31 Defect controls decomposition/]
    A31 --> J30([J30 Defect controls decomposition])
    J30 --> S29[/S29 Defect controls decomposition/]
    S29 --> G32[/G32 Defect controls decomposition/]
    G32 --> A32[/A32 Defect controls decomposition/]
    A32 --> J31([J31 Defect controls decomposition])
    J31 --> S30[/S30 Defect controls decomposition/]
    S30 --> G33[/G33 Defect controls decomposition/]
    G33 --> A33[/A33 Defect controls decomposition/]
    A33 --> J32([J32 Defect controls decomposition])
    J32 --> S31[/S31 Defect controls decomposition/]
    S31 --> G34[/G34 Defect controls decomposition/]
    G34 --> A34[/A34 Defect controls decomposition/]
    A34 --> J33([J33 Defect controls decomposition])
    J33 --> S32[/S32 Defect controls decomposition/]
    S32 --> G35[/G35 Defect controls decomposition/]
    G35 --> A35[/A35 Defect controls decomposition/]
    A35 --> J34([J34 Defect controls decomposition])
    J34 --> S33[/S33 Defect controls decomposition/]
    S33 --> G36[/G36 Defect controls decomposition/]
    G36 --> A36[/A36 Defect controls decomposition/]
    A36 --> J35([J35 Defect controls decomposition])
    J35 --> S34[/S34 Defect controls decomposition/]
    S34 --> G37[/G37 Defect controls decomposition/]
    G37 --> A37[/A37 Defect controls decomposition/]
    A37 --> J36([J36 Defect controls decomposition])
    J36 --> S35[/S35 Defect controls decomposition/]
    S35 --> G38[/G38 Defect controls decomposition/]
    G38 --> A38[/A38 Defect controls decomposition/]
    A38 --> J37([J37 Defect controls decomposition])
    J37 --> S36[/S36 Defect controls decomposition/]
    S36 --> G39[/G39 Defect controls decomposition/]
    G39 --> A39[/A39 Defect controls decomposition/]
    A39 --> J38([J38 Defect controls decomposition])
    J38 --> S37[/S37 Defect controls decomposition/]
    S37 --> G40[/G40 Defect controls decomposition/]
    G40 --> A40[/A40 Defect controls decomposition/]
    A40 --> J39([J39 Defect controls decomposition])
    J39 --> S38[/S38 Defect controls decomposition/]
    S38 --> G41[/G41 Defect controls decomposition/]
    G41 --> A41[/A41 Defect controls decomposition/]
    A41 --> J40([J40 Defect controls decomposition])
    J40 --> S39[/S39 Defect controls decomposition/]
    S39 --> G42[/G42 Defect controls decomposition/]
    G42 --> A42[/A42 Defect controls decomposition/]
    A42 --> J41([J41 Defect controls decomposition])
    J41 --> S40[/S40 Defect controls decomposition/]
    S40 --> G43[/G43 Defect controls decomposition/]
    G43 --> A43[/A43 Defect controls decomposition/]
    A43 --> J42([J42 Defect controls decomposition])
    J42 --> S41[/S41 Defect controls decomposition/]
    S41 --> G44[/G44 Defect controls decomposition/]
    G44 --> A44[/A44 Defect controls decomposition/]
    A44 --> J43([J43 Defect controls decomposition])
    J43 --> S42[/S42 Defect controls decomposition/]
    S42 --> G45[/G45 Defect controls decomposition/]
    G45 --> A45[/A45 Defect controls decomposition/]
    A45 --> J44([J44 Defect controls decomposition])
    J44 --> S43[/S43 Defect controls decomposition/]
    S43 --> G46[/G46 Defect controls decomposition/]
    G46 --> A46[/A46 Defect controls decomposition/]
    A46 --> J45([J45 Defect controls decomposition])
    J45 --> S44[/S44 Defect controls decomposition/]
    S44 --> G47[/G47 Defect controls decomposition/]
    G47 --> A47[/A47 Defect controls decomposition/]
    A47 --> J46([J46 Defect controls decomposition])
    J46 --> S45[/S45 Defect controls
```



# AV Software is \_\_\_\_\_





GARFIELD COMMONS  
APARTMENT HOMES  
1110

Arrival in 19 mi  
10:28 AM



# By what Metrics can we judge AV software (e.g., safety)?



# Participation Activity

- Write name in top left hand corner of paper
- What are some ways we could measure if the code is safe?
- Come up with two metrics, and then discuss with your neighbor.
- For each metric come up with a pro and a con for that metric.

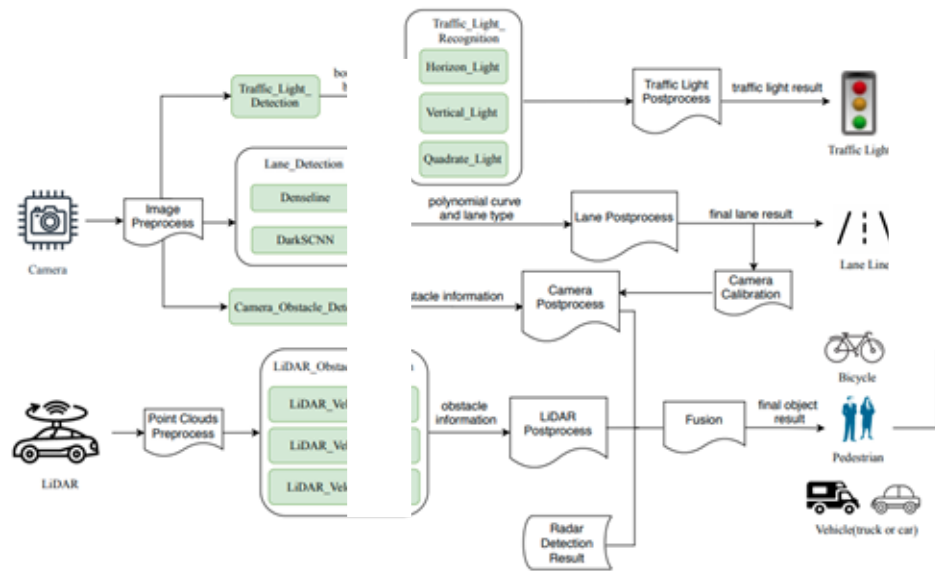
# (1) Code Coverage

- Amount of code executed during testing.
- Statement coverage, line coverage, branch coverage, etc.
- E.g., 75% branch coverage
  - 3/4 if-else outcomes have been executed

```
1698 : const TrajectoryPoint& StGraphData::init_point() const { return init_point_; }
2264 : const SpeedLimit& StGraphData::speed_limit() const { return speed_limit_; }
212736 : double StGraphData::cruise_speed() const {
212736 :     return cruise_speed_ > 0.0 ? cruise_speed_ : FLAGS_default_cruise_speed;
1698 : double StGraphData::path_length() const { return path_data_length_; }
1698 : double StGraphData::total_time_by_conf() const { return total_time_by_conf_; }
1698 : planning_internal::STGraphDebug* StGraphData::mutable_st_graph_debug() {
1698 :     return st_graph_debug_;
566 : bool StGraphData::SetSTDriableBoundary(
    const std::vector<std::tuple<double, double, double>>& a_boundary,
    const std::vector<std::tuple<double, double, double>>& v_obs_info) {
[+ -] 566 :     if (a_boundary.size() != v_obs_info.size()) {
        return false;
[+ +] 40752 :     for (size_t i = 0; i < a_boundary.size(); ++i) {
80372 :         auto st_bound_instance = st_drivable_boundary_.add_st_boundary();
160744 :         st_bound_instance->set_t(st_i.get<0>(a_boundary[i]));
120558 :         st_bound_instance->set_s_lower(st_i.get<1>(a_boundary[i]));
120558 :         st_bound_instance->set_s_upper(st_i.get<2>(a_boundary[i]));
[+ -] 40186 :         if (st_i.get<1>(v_obs_info[i]) > -kObsSpeedIgnoreThreshold) {
            0 :             st_bound_instance->set_v_obs_lower(st_i.get<1>(v_obs_info[i]));
[+ *] 40186 :         if (st_i.get<2>(v_obs_info[i]) < kObsSpeedIgnoreThreshold) {
50254 :             st_bound_instance->set_v_obs_upper(st_i.get<2>(v_obs_info[i]));
}
```

## (2) Model Accuracy

- Train machine-learning models on labelled data (sensor data + ground truth)
- Compute accuracy on a separate labelled test set
- E.g., 90% accuracy implies that object recognition is right for 90% of the test inputs.



Source: Peng et al. ESEC/FSE'20

# (3) Failure Rate

- Frequency of crashes / fatalities
- Per 1,000 rides, per million miles, per month (in the news)

TRANSP / WAYMO / TECH

## Waymo's driverless cars were involved in two crashes and 18 'minor contact events' over 1 million miles



Image: Allen J. Schaben / LA

/ The Alphabet-owned company pulls back the curtain on more stats from its public road testing. Of the 20 incidents, only two met the federal government's reporting criteria, and no one was injured.

By Andrew J. Hawkins, transportation editor with 10+ years of experience who covers EVs, public transportation, and aviation. His work has appeared in The New York Daily News and City & State.

Feb 28, 2023, 8:00 PM GMT+3 | 1 Comment | 1 News



## 'Complete meltdown': Driverless cars in San Francisco stall causing a traffic jam



By Jonathan Vukobrat, CNN Business

Updated 3:45 PM EDT, March 14, 2023



A Cruise driverless car in San Francisco



# (4) Mileage

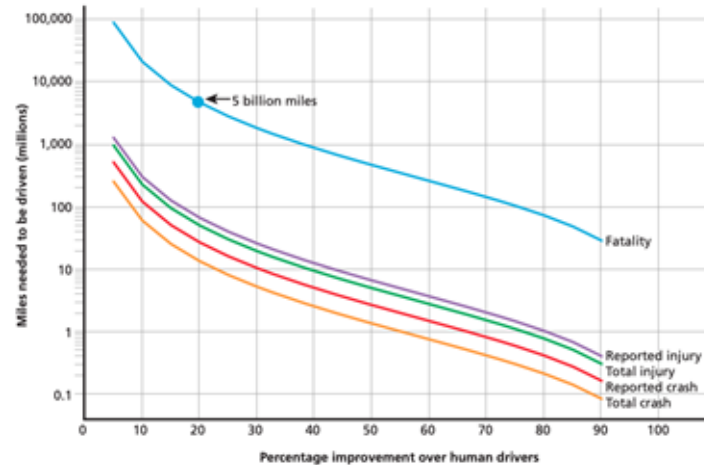


## Driving to Safety

How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?

Nidhi Kalra, Susan M. Paddock

Figure 3. Miles Needed to Demonstrate with 95% Confidence that the Autonomous Vehicle Failure Rate Is Lower than the Human Driver Failure Rate



## Building the World's Most Experienced Driver™

The Waymo Driver gains experience with every mile, in each car. →



10+

More than a Decade of Autonomous Driving in More than 10 States

5

Generations of Autonomously Driven Vehicles

15+

Billion Autonomously Driven Miles in Simulation

20+

Million Real-World Miles on Public Roads

Source: waymo.com/safety (September 2019)

# What is Measurement?

# What is Measurement?

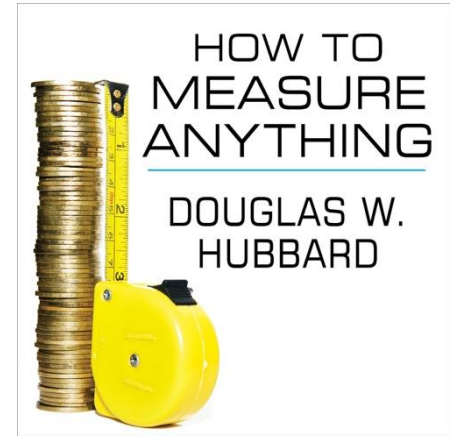
- “Measurement is the empirical, **objective assignment of numbers, according to a rule** derived from a model or theory, to **attributes of objects or events** with the intent of describing them”

Kaner, Bond, *Software Engineering Metrics: What Do They Measure and How Do We Know?*



# What is Measurement?

- “A quantitatively expressed **reduction of uncertainty** based on one or more observations.”  
Hubbard, *How to Measure Anything* ...



# What is Measurement?

- “A **software quality metric** is a function whose inputs are software data and whose output is a single numerical value that can be interpreted as the degree to which the software possesses a given attribute that affects its quality.”

*IEEE 1061*

**Metric**

MEASURES

**Attribute**

DESCRIBES

**Entity**

# What is Measurement?

Method to obtain a number or symbol →

**Metric**

MEASURES

Quality of interest →

**Attribute**

DESCRIBES

Object or process →

**Entity**

# Entities represent an Object or Process

## Examples

- Software Product
- Modules
- Software Development Process
- People

**Metric**

MEASURES

**Attribute**

DESCRIBES

**Entity**

# Attributes represent Qualities of Interest

## Examples for **Software Quality**

- Reliability
- Security
- Scalability
- Extensibility
- Portability
- Availability
- Safety
- Observability
- Accuracy
- Robustness
- Resilience
- Timeliness
- Responsiveness
- Intuitiveness

Metric

MEASURES

Attribute

DESCRIBES

Entity



# Attributes represent Qualities of Interest

## Examples for **Software Dev. Process**

- Development efficiency
- Meeting efficiency
- Conformance to process
- Accuracy of predictions
- Fairness in decision making
- Regulatory compliance
- On-time release
- ...

Metric

MEASURES

Attribute

DESCRIBES

Entity

# Attributes represent Qualities of Interest

## Examples for **People**

### Developers

- Productivity
- Agility
- Well-Being + Job Satisfaction
- Communication and Collaboration
- Creativity
- Morale
- Regulatory Compliance

### End Users

- Product Satisfaction
- Ease of Use
- Feature Usage
- Regulatory Compliance

**Metric**

MEASURES

**Attribute**

DESCRIBES

**Entity**



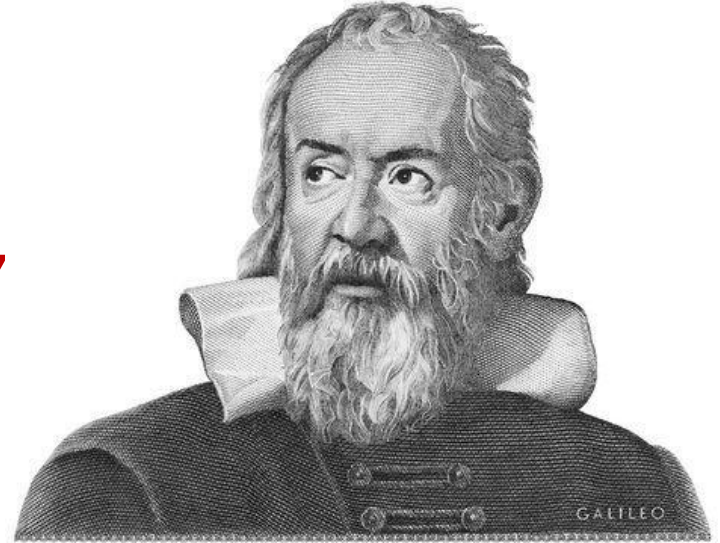
# What about Non-Trivial Qualities?

- **Software**
  - code elegance
  - code maintainability
- **Process**
  - development efficiency
  - fairness in decision making
- **Team**
  - productivity
  - collaboration
  - creativity

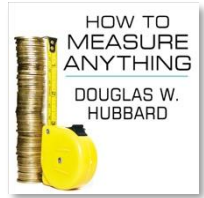


***“Measure what is  
measurable, and make  
measurable what is not so.”***

Galileo Galilei



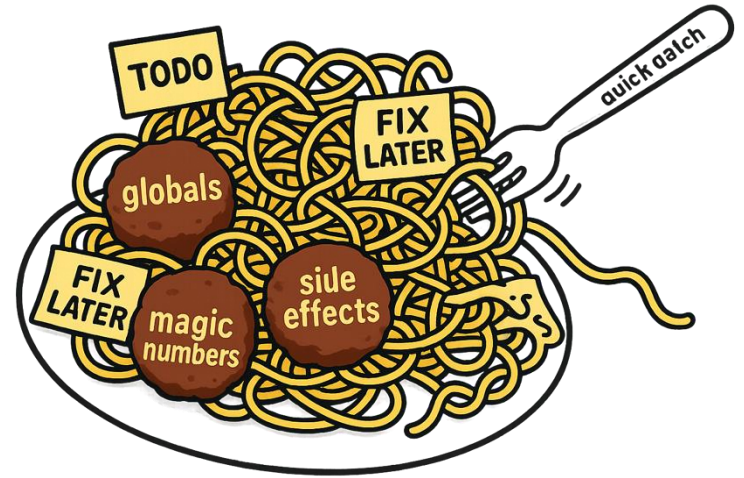
# Everything is Measurable



- If we care about X, then, by definition, X **must be detectable**
  - how could we care about things like *quality, risk, security, or public image* if these things were totally undetectable, directly or indirectly?
  - if we have reason to care about some unknown quantity, it is because we think it corresponds to desirable or undesirable results in some way
- If X is detectable, then it **must be detectable in some amount**
  - if you can observe a thing at all, you can observe more of it or less of it
- If we can observe it in some amount, then it **must be measurable**

Douglas Hubbard, How to Measure Anything, 2010

# Example: Code Complexity



thanks to ChatGPT for the terrible image

# Number of Lines

- Easy to Measure! `wc -l file1 file2...`

LOC	projects
450	Expression Evaluator
2,000	Sudoku
100,000	Apache Maven
500,000	Git
3,000,000	MySQL
15,000,000	gcc
50,000,000	Windows 10
2,000,000,000	Google (MonoRepo)

# Normalizing Lines of Code

- Ignore comments and empty lines
- Ignore lines < 2 characters
- Pretty print source code first
- Count statements (logical lines of code)
- See also: cloc

```
for (i = 0; i < 100; i += 1) printf("hello"); /* How many lines of code is this? */
```

```
/* How many lines of code is this? */
```

```
for (  
    i = 0;  
    i < 100;  
    i += 1  
) {  
    printf("hello");  
}
```



# Normalization by Language

Language	Statement factor (productivity)	Line factor
C	1	1
C++	2.5	1
Fortran	2	0.8
Java	2.5	1.5
Perl	6	6
Smalltalk	6	6.25
Python	6	6.5

Source: "Code Complete: A Practical Handbook of Software Construction", S. McConnell, Microsoft Press (2004)  
and <http://www.codinghorror.com/blog/2005/08/are-all-programming-languages-the-same.html> u.a.

# Halstead's Metrics (1977)

- Based on number of operators (e.g., `print`, `+`, `-`) and operands (name, “maurice”)
  - $n_1$  and  $n_2$ : # of distinct operators and operands
  - $N_1$  and  $N_2$ : # of instances of operators and operands
- Derived metrics include vocabulary ( $N$ ), length ( $n$ ), volume ( $V$ ), difficulty ( $D$ ), effort ( $E$ ), time to write ( $T$ ), number of bugs ( $B$ )

$$D = \frac{n_1}{2} + \frac{N_2}{n_2}$$

$$V = N \times \log_2 n$$

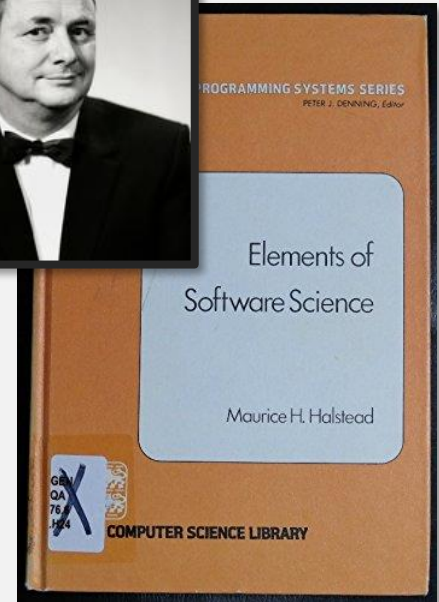
$$E = D \times V$$

$$N = N_1 + N_2$$

$$T = \frac{E}{18}$$

$$B = \frac{V}{3000}$$

$$n = n_1 + n_2$$

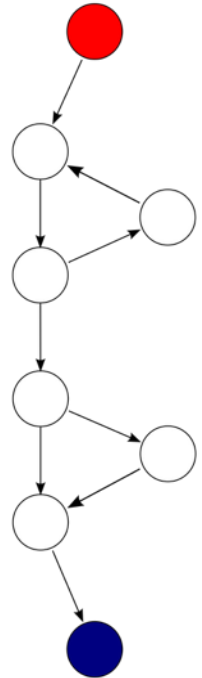


# McCabe Cyclomatic Complexity (1976)

- Computed according to **control flow** graph
  - tells you how many tests you need for full branch coverage
- Equal to **number of decision points + 1**
  - if, while, do-while, ?:, catch, switch, case
  - && and || in if condition

*"For each module, either limit cyclomatic complexity to [X] or provide a written explanation of why the limit was exceeded."*

NIST Structured Testing methodology



# Object-Oriented Metrics (1994)

- Number of Methods per Class
- Depth of Inheritance Tree
- Number of Child Classes
- Coupling between Object Classes
- Calls to Methods in Unrelated Classes
- ...

476

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 20, NO. 6, JUNE 1994

## A Metrics Suite for Object Oriented Design

Shyam R. Chidamber and Chris F. Kemerer

**Abstract**—Given the central role that software development plays in the delivery and application of information technology, managers are increasingly focusing on process improvement in the software development area. This demand has spurred the provision of a number of new and/or improved approaches to software development, with perhaps the most prominent being object-orientation (OO). In addition, the focus on process improvement has increased the demand for software measures, or metrics with which to manage the process. The need for such metrics is particularly acute when an organization is adopting a new technology for which established practices have yet to be developed. This research addresses these needs through the development and implementation of a new suite of metrics for OO design. Metrics developed in previous research, while contributing to the field's understanding of software development processes, have generally been subject to serious criticisms, including the lack of a theoretical base. Following Wand and Weber, the theoretical base chosen for the metrics was the ontology of Runge. Six design metrics are developed, and then analytically evaluated against Weyaker's proposed set of measurement principles. An automated data collection tool was then developed and implemented to collect an empirical sample of these metrics at two field sites in order to demonstrate their feasibility and suggest ways in which managers may use these metrics for process improvement.

**Index Terms**—CR categories and subject descriptors: D.2.8 [software engineering]: metrics; D.2.9 [software engineering]: management; F.2.2 [analysis of algorithms and problem complexity]: tradeoffs among complexity measures; K.6.3 [management of computing and information systems]: software management. General terms: Class, complexity, design, management, measurement, metrics, object orientation, performance.

### I. INTRODUCTION

It has been widely recognized that an important component of process improvement is the ability to measure the process. Given the central role that software development plays in the delivery and application of information technology, managers are increasingly focusing on process improvement in the software development area. This emphasis has had two effects. The first is that this demand has spurred the provision of a number of new and/or improved approaches to software development, with perhaps the most prominent being object-orientation (OO). Second, the focus on process improvement has increased the demand for software measures, or metrics with which to manage the process. The need for such metrics

is particularly acute when an organization is adopting a new technology for which established practices have yet to be developed.

This research addresses these needs through the development and implementation of a new suite of metrics for OO design. Previous research on software metrics, while contributing to the field's understanding of software development processes, have generally been subject to one or more types of criticisms. These include: lacking a theoretical basis [41], lacking in desirable measurement properties [47], being insufficiently generalized or too implementation technology dependent [45], and being too labor-intensive to collect [22].

Following Wand and Weber, the theoretical base chosen for the OO design metrics was the ontology of Runge [5], [16], [43]. Six design metrics were developed, and analytically evaluated against a previously proposed set of measurement principles. An automated data collection tool was then developed and implemented to collect an empirical sample of these metrics at two field sites in order to demonstrate their feasibility and to suggest ways in which managers may use these metrics for process improvement.

The key contributions of this paper are the development and empirical validation of a set of theoretically-grounded metrics of OO design. The rest of this paper is organized as follows. The next section presents a brief summary of the research problem, followed by a section describing the theory underlying the approach taken. Then Weyaker's list of software metric evaluation criteria is presented, along with a brief description of the empirical data collection sites. The Results section presents the metrics, their analytical evaluation, the empirical data and a managerial interpretation of the data for each metric. Some concluding remarks are presented in the final section.

### II. RESEARCH PROBLEM

There are two general types of criticisms that can be applied to current software metrics. The first category are those theoretical criticisms that are leveled at conventional software metrics as they are applied to traditional, non-OO software design and development. Kearney, et al. criticized software complexity metrics as being without solid theoretical bases and lacking appropriate properties [21]. Vessey and Weber also commented on the general lack of theoretical rigor in the structured programming literature [41]. Both Prather and Weyaker proposed that traditional software complexity metrics do not possess appropriate mathematical properties, and consequently fail to display what might be termed normal predictable behavior [34], [47]. This suggests that software

Manuscript received February 17, 1993; revised January, 1994; accepted by S. Zureick. This research was supported in part by the MIT Center for Information Systems Research (CISR), and the cooperation of two industrial organizations who supplied the data. The authors are with the Massachusetts Institute of Technology, 635-113, 30 Wadsworth Street, Cambridge, MA 02139 USA. e-mail: shyam@infocis.mit.edu or ckemerer@infocis.mit.edu. IEEE Log Number 9401707.

0895-5589/94/060476-08 © 1994 IEEE

# Takeaways: Code Complexity

- There are **lots of complexity metrics** that measure different things, which may or may not be what we care about
  - there is no single value that fully captures “code complexity”
  - most of these metrics are confounded by code size!
- We need to be **intentional** about what we are measuring
  - how hard is this code to read?
  - how difficult is this code to maintain?
  - does this function need to be refactored?

"The key result has to be measurable. But at the end you can look, and without any arguments: **Did I do that or did I not do it?** Yes? No? Simple. No judgments in it."

Andy Grove, Inventor of OKRs

## What are OKRs?

### OBJECTIVES AND KEY RESULTS

OKRs are a management methodology which helps to ensure that your company focuses efforts on the same important issues throughout the organization.

#### OBJECTIVES

**An Objective** is what you want to accomplish.

A good Objective is significant, concrete, action-oriented and inspirational. Can be set annually or over an even longer-term.

#### KEY RESULTS

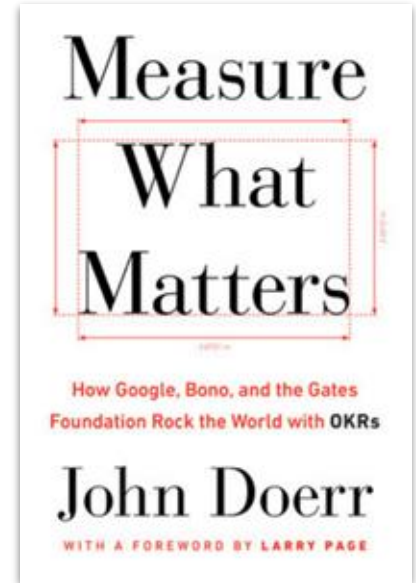
**Key Results** are how you will accomplish it.

Good Key Results are specific, timebound, aggressive yet realistic measurable and verifiable. Can be set quarterly and evolve as work progresses.

What Matters



WV

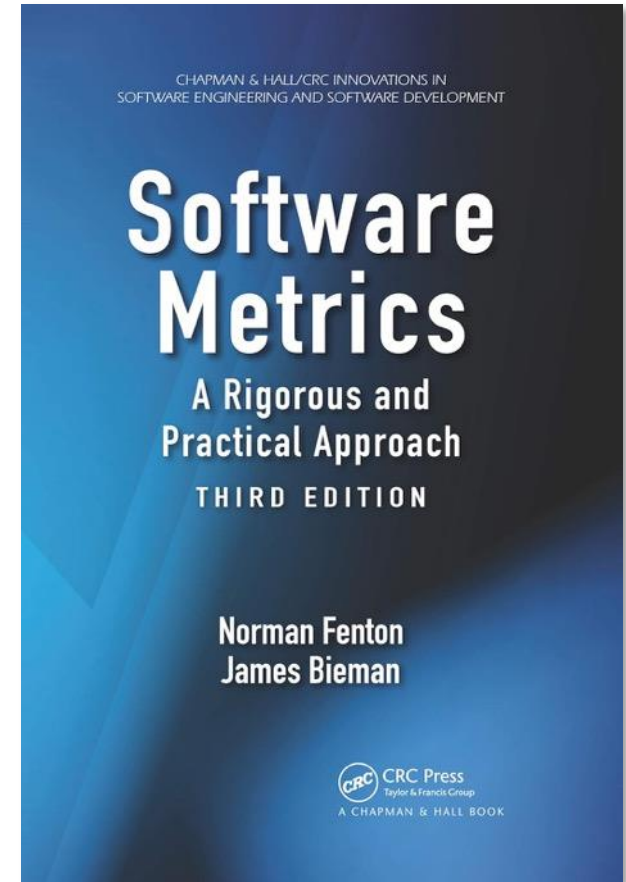


"OKRs have helped lead us to 10x growth, many times over."

Larry Page, CEO of Google

***“Every measurement action must be motivated by a particular goal or need that is clearly defined and easily understandable.”***

*Software Metrics: A Rigorous and Practical Approach*  
N.Fenton, J.Bieman



# Goal-Based Frameworks

- Objectives and Key Results (OKRs)
- **Goal-Question-Metric (GQM)**
- Assurance Cases
- ...

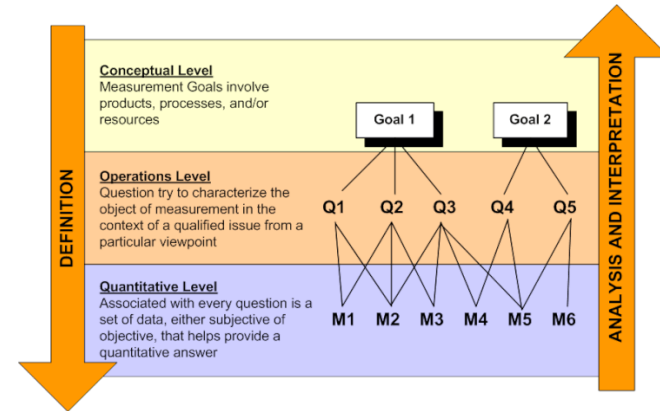


# The GQM Framework

**Goal:** What do you want to achieve?

**Questions:** What do you need to answer to know whether your goal is met?

**Metrics:** What measurements do you need to answer each question?



## THE GOAL QUESTION METRIC APPROACH

Victor R. Basili<sup>1</sup> Gianluigi Caldiera<sup>1</sup> H. Dieter Rombach<sup>2</sup>

(<sup>1</sup>) Institute for Advanced Computer Studies  
Department of Computer Science  
University Of Maryland  
College Park, Maryland

(<sup>2</sup>) FB Informatik  
Universität Kaiserslautern  
Kaiserslautern, Germany

# GQM: Defining Goals

- P: Purpose** (improve, evaluate, monitor, ...)
- I: Issue** (reliability, usability, effectiveness, ...)
- O: Object** (final product, component, process, activity)
- V: Viewpoint** (any stakeholder)

**Goal:**

**Evaluate** the **effectiveness** of the **organization's coding standard** from the **team's** perspective

**Questions:**

How comprehensible are the coding standards?

What is the impact of coding standards on the efficiency and productivity of the team?

**Metrics:**

Survey results measuring team members' understanding

Number of revisions required to achieve standard compliance

Code size:  
# of lines of code,  
# of classes,  
# of functions

# Participation Activity

- Apply the **Goal-Question-Metric** framework to explore various aspects of AV software
- Define one goal, two questions, and at least one metric per question
- Share with the class!

# Example

**Goal: Ensure energy efficiency and sustainability** from the point of view of the **organization and environmental analysts**

**Q1:** What is the energy consumption under different driving conditions?

**Metrics:** Miles per kWh  
{city, highway, mixed}  
{winter, summer, fall}

**Q2:** How efficient is the battery management system?

**Metrics:** Battery life  
number of charge cycles  
peak battery temperature

# Measurement for Decision Making

- Fund project?
- More testing?
- Fast enough? Secure enough? Safe enough?
- Code quality sufficient?
- Which feature to focus on?
- Should we refactor the code? Rewrite?
- Developer bonus?
- Time and cost estimation? Predictions reliable?

# Analysis and Interpretation

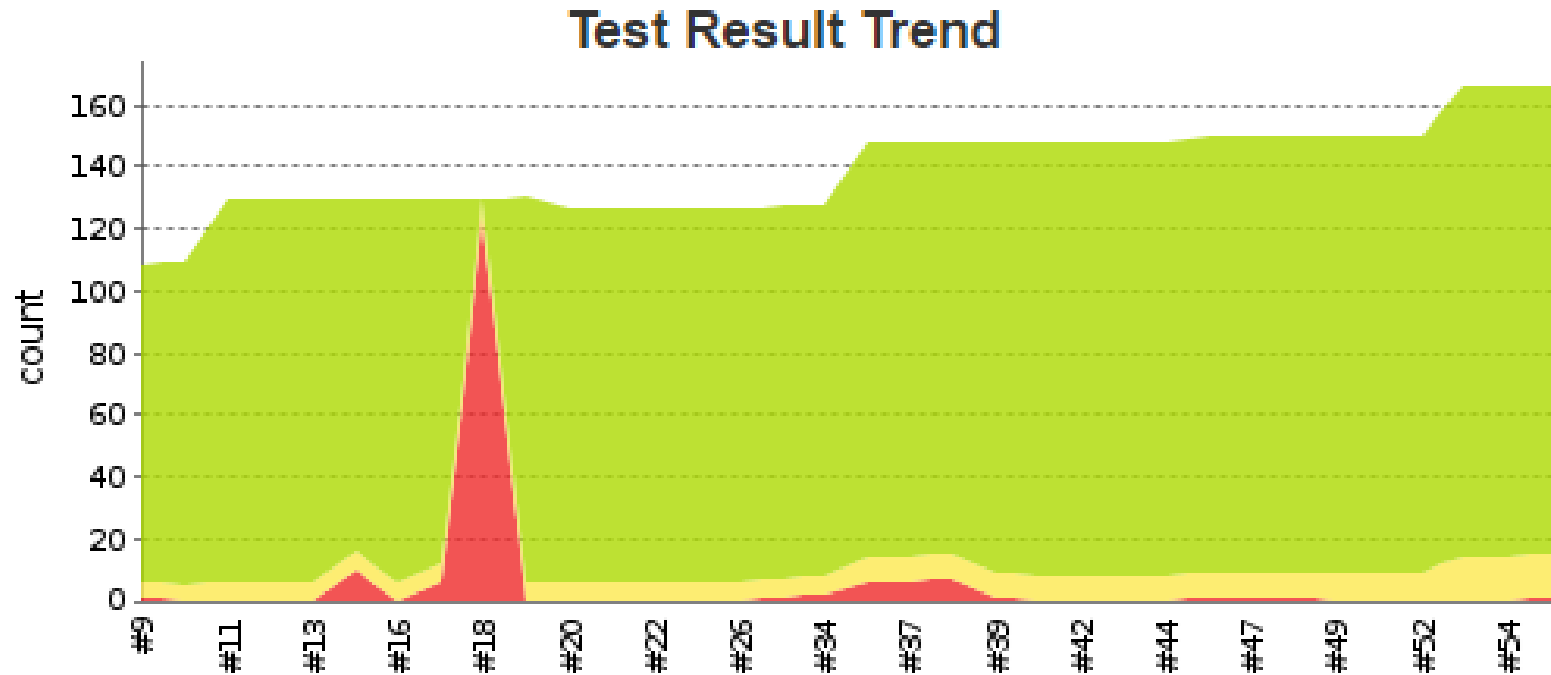
**Challenge:** How do we actually **interpret the measurements** from our metrics?

# Calibrate thresholds based on feedback

- **Problem:** "I'll know it when I see it"
  - "That function is way too complex!"
- **Solution:** Use existing code to obtain a set of thresholds
  - pick exemplars (clear pass / fail / borderline)
  - present to team, track agreement, and derive thresholds
    - E.g., cyclomatic complexity shouldn't be higher than 10
  - periodically recalibrate over time (e.g., technology changes)

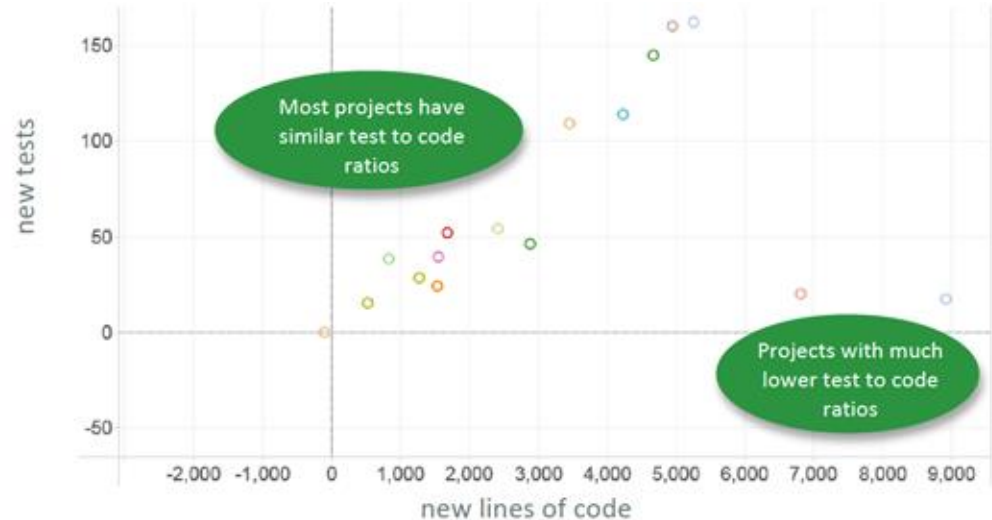


# Track metrics over time to assess trends



# Benchmark against existing standards

- Monitor **similar** products, projects, modules, teams
- Refer to **external standards** if possible
  - e.g., ISO 26262
- Record typical values for metrics of interest
- Investigate deviations



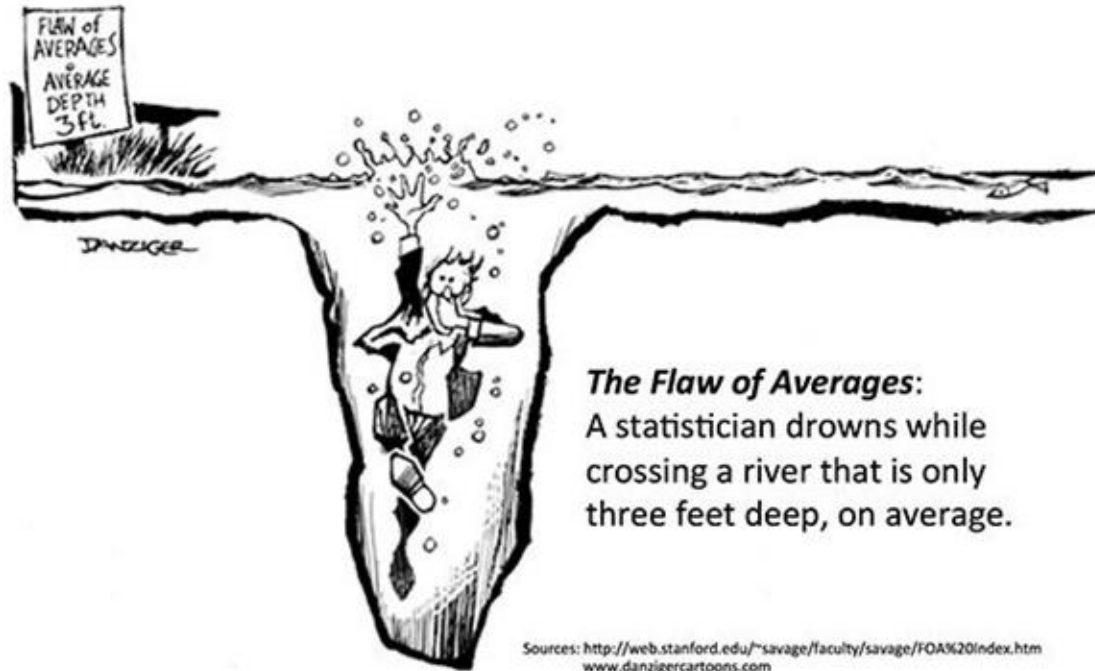


# The Streetlight Effect

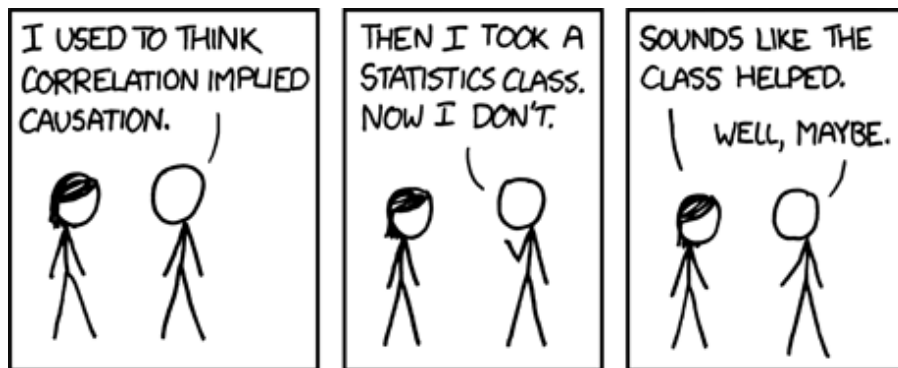


- A known **observational bias**
- People tend to look for something only where it's easiest to do so
  - If you drop your keys at night, you'll tend to look for it under streetlights

# Bad Statistics: What could possibly go wrong?

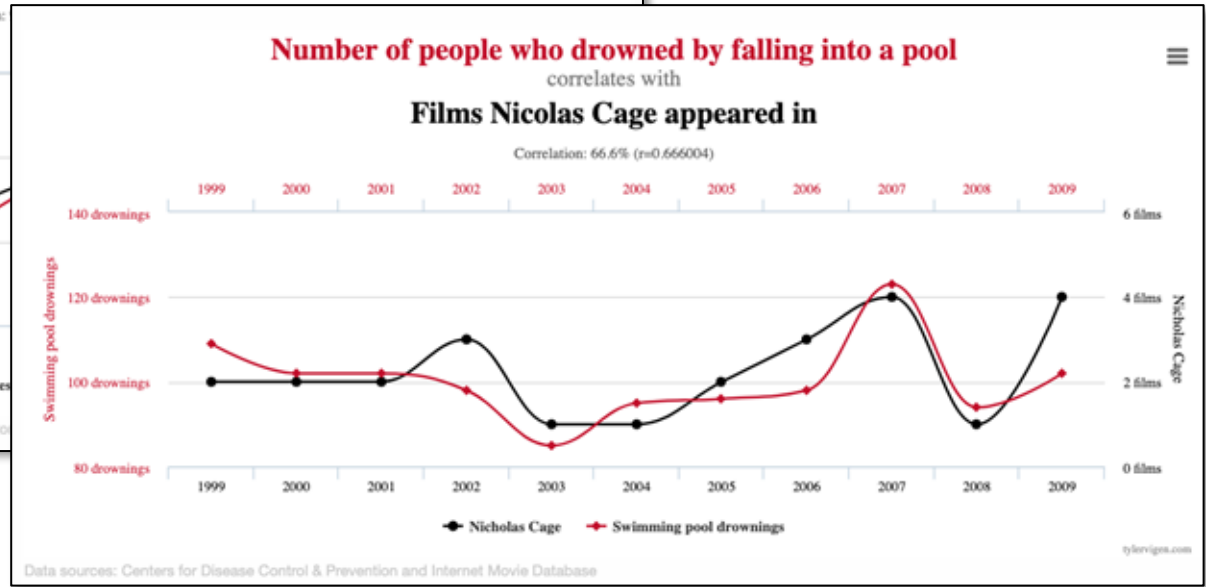
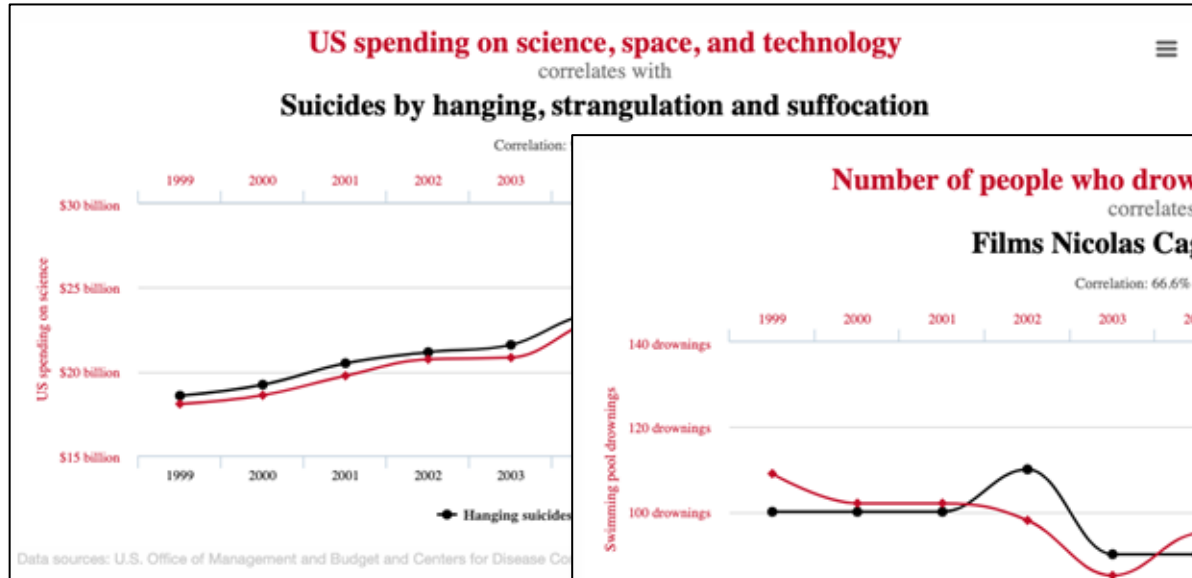


# Making Inferences

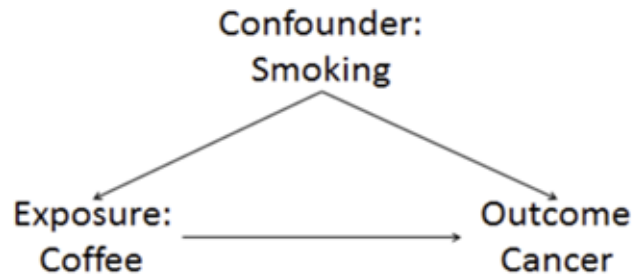
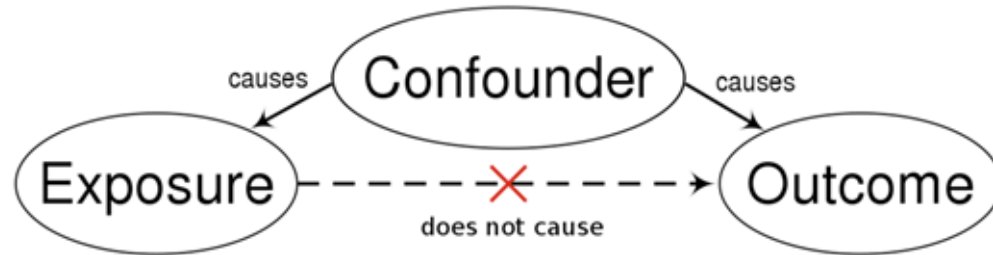


- To infer causation:
  - Provide a theory (from domain knowledge, independent of data)
  - Show correlation
  - Demonstrate ability to predict new cases (replicate/validate)

# Spurious Correlations



# Confounding Variables



- If you look only at the coffee consumption → cancer relationship, you can get very misleading results
- Smoking is a confounder



RESEARCH-ARTICLE

## Coverage is not strongly correlated with test suite effectiveness

**Authors:**  [Laura Inozemtseva](#),  [Reid Holmes](#) [Authors Info & Affiliations](#)

ICSE 2014: Proceedings of the 36th International Conference on Software Engineering • May 2014 • Pages 435–445 • <https://doi.org/10.1145/2568225.2568271>

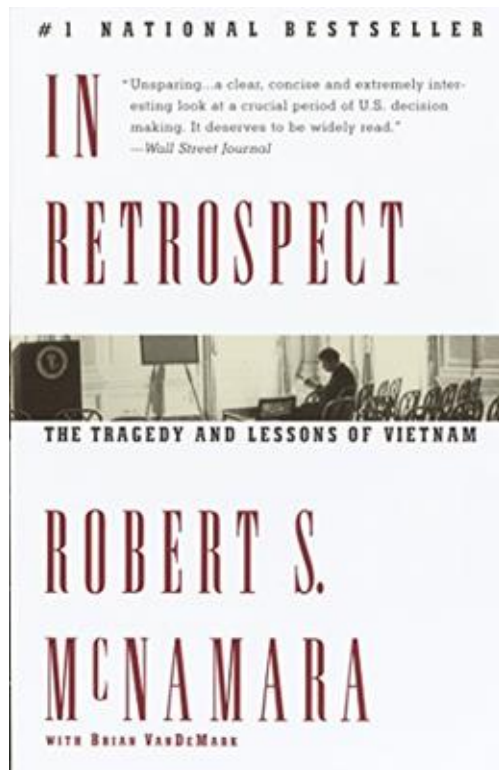
“We found that there is a low to moderate correlation between coverage and effectiveness when the number of test cases in the suite is controlled for.”

74

**Most prior studies didn't account for the confounding influence of test suite size**

# Measurement Reliability

- Extent to which a measurement yields similar results when applied multiple times
- Goal is to reduce uncertainty, increase consistency
- Example: Performance
  - Time, memory usage
  - Cache misses, I/O operations, instruction execution count, etc.
- Law of large numbers
  - Taking multiple measurements to reduce error
  - Trade-off with cost



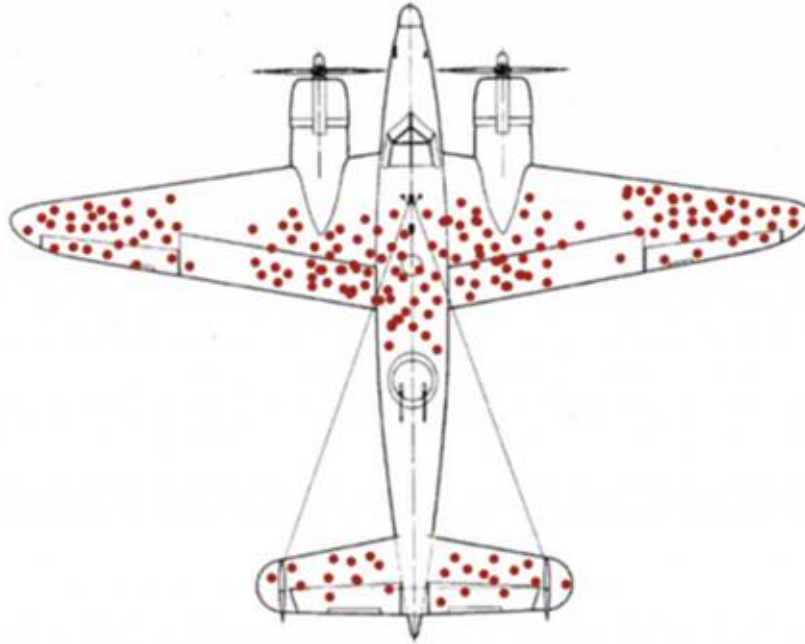
# McNamara Fallacy

- Measure whatever can be easily measured
- Disregard that which cannot be measured easily
- Presume that which cannot be measured easily is not important
- Presume that which cannot be measured easily does not exist



<https://chronotopeblog.com/2015/04/04/the-mcnamara-fallacy-and-the-problem-with-numbers-in-education>

# Survivorship Bias



# Goodhart's law: "When a measure becomes a target, it ceases to be a good measure."



<http://dilbert.com/strips/comic/1995-11-13>

## *The Price of Wells Fargo's Fake Account Scandal Grows by \$3 Billion*

The bank reached a settlement with federal prosecutors and the Securities and Exchange Commission after abusing customers.

Share full article



199



Wells Fargo used fraud to open up fake accounts and force customers into services

# Incentivizing Productivity

- What happens when developer bonuses are based on:
  - Lines of code per day?
  - Amount of documentation written?
  - Low number of reported bugs in their code?
  - Low number of open bugs in their code?
  - High number of fixed bugs?
  - Accuracy of time estimates?



# What You Need to Know



Metrics are important in Software Engineering



Apply goal-oriented approaches to software metrics



Provide examples of metrics for software qualities and process



Understand limitations and dangers of decisions and incentives based on measurements

# Questions to Consider (Projects)

- What properties do we care about and how do we measure them?
- What is being measured? Does it (to what degree) capture the thing you care about? What are its limitations?
- How should it be incorporated into process?
- What are potentially negative side effects or incentives?