Human Factors: Quantitative and Qualitative Methods

Yu Huang

Vanderbilt University

yu.huang@vanderbilt.edu

Final Schedule

- Teams, paper presentations
- Due date for HW2
  - Sep 25 (current plan) or Oct 6 (after all mid-proposal presentations)?

We want to improve productivity and reduce cost in software development and maintenance.

What is software engineering?

Programs

- Testing
- Fault localization
- Static analysis
- Dynamic analysis
- Debugging
- ...

Programmers

- Will programmers use these tools? Why or why not?
- How do experts become experts?
- How to be productive?
- Biases?
- How to make a team function?
- How to estimate effort?
- ...

The Human Aspect Matters

Captain Sully

Chesley (Sully) Sullenberger clarified vividly **the significance of the “human factor”** in our digital age. After saving 155 people by landing his disabled Airbus A320 in the Hudson River in January 2009, Sully became a national hero.

Sichuan Airlines Flight 8633

At the altitude of 9 km (30,000 ft; 9,000 m), the right front segment of the windshield separated from the aircraft followed by an uncontrolled decompression. The flight control unit was damaged, and the loud external noise made spoken communications impossible. Because the flight was within a mountainous region, the pilots were unable to descend to the required 8,000 ft (2,400 m) to compensate for the loss of cabin pressure. The sudden loss of pressure in the cockpit had caused multiple instruments to fail.

The half-body of copilot was sucked out of the window and the pilot kept flown ***by manual and sight****.* The three pilots were in short sleeves and suddenly it was -40°C in the cockpit. After 35 minutes, the crew made an emergency landing. 2 crew members were injured.

"Epic-level diversion".

The Human Aspect Matters

1. The Mariner 1 Spacecraft, 1962

The first entry in our rundown goes right back to the sixties.

Before the summer of love or the invention of the lava lamp, NASA launched a space mission to fly past Venus. It did not go to plan.

The Mariner 1 space probe barely made it out of Cape Canaveral before the ro course. Worried that the rocket was heading towards a crash-landing on earth destruct command and the craft was obliterated about 290 seconds after laun

2. The Morris Worm, 1988

Not all costly software errors are worn by big companies or government organizations. In fact, most costly software bugs ever was caused by a single student. A Cornell University student cre as part of an experiment, which ended up spreading like wildfire and crashing tens of thousanc computers due to a coding error.

The computers were all connected through a very early version of the internet, making the Mor essentially the first infectious computer virus. Graduate student Robert Tappan Morris charged and convicted of criminal hacking and fined $10,000, although the cost of the estimated to be as high as $10 million.

History has forgiven Morris though, with the incident now widely credited for exposing ital security. These days, Morris is a professor at MIT and the worm's sou eum piece on a floppy disc at the University of Boston.

3. Pentium FDIV Bug, 1994

The Pentium FDIV bug is a curious case of a minor problem that

Thomas Nicely, a math professor, discovered a flaw in the Pentiu response was to offer a replacement chip to anyone who could p

The original error was relatively simple, with a problem in the loc cause tiny inaccuracies in calculations, but only very rarely. In fac

4. Bitcoin Hack, Mt. Gox, 2011

Mt. Gox was the biggest bitcoin exchange in the world in the 2010s, until they were hit by a software error that ultimately proved fatal.

The glitch led to the exchange creating transactions that could never be fully redeemed, costing up to $1.5 million in lost bitcoins.

But Mt. Gox's woes didn't end there. In 2014, they lost more than 850,000 bitcoins (valued at roughly half a billion USD at the time) in a hacking incident. Around 200,000 bitcoins were recovered, but the financial loss was still overwhelming and the exchange ended up declaring bankruptcy.

5. EDS Child Support System, 2004

Back in 2004, the UK government introduced a new and complex system to manage the operations of the Child Support Agency (CSA). The contract was awarded to IT services company Electronic Data Systems (EDS). The system was called CS2, and there were problems as soon as it went live.

A leaked internal memo at the time revealed that the system was "badly designed, badly tested and badly implemented". The agency reported that CS2 "had over 1,000 reported problems, of which 400 had no known workaround", resulting in "around 3,000 IT incidents a week". The system was budgeted to cost around £450 million, but ended up costing an estimated £768 million altogether. EDS, a Texas-based contractor, also announced a $153 million loss in their subsequent financial results.

6. Heathrow Terminal 5 Opening, 2008

Imagine prepping to jet off on your eagerly-awaited vacation or important business trip, only to find that your flight is grounded or and your luggage is nowhere to be seen.

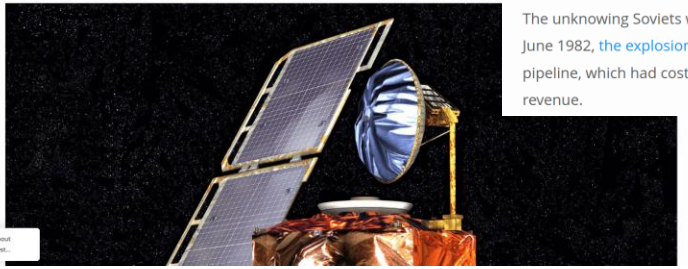
This was exactly what happened to thousands of travelers when Heathrow's Terminal 5 opened back in March 2008, and it was that performed well on malfunctioning luggage

British Airways also rev airport. Over the next 1 than £16 million.

7. NASA's Mars Climate Orbiter, 1998

Losing $20 from your wallet is probably enough to ruin your day — how would spacecraft? NASA engineers found out back in 1998 when the Mars Climate Ort too close to the surface of Mars.

It took engineers several months to work out what went wrong. It turned out tc mistake in converting imperial units to metric. According to the investigation re software produced by Lockheed Martin used imperial measurements, while the by NASA, was programmed with SI metric units. The overall cost of the failed m million.



8. Soviet Gas Pipeline Explosion, 1982

This error is a little bit different to the others, as it was deliberate (or so rumor has it). In fact, the Soviet gas pipeline explosion is alleged to be a cunning example of cyber-espionage, carried out by the CIA.

Back in 1982, at the height of the cold war tensions between the USA and USSR, the Soviet government built a gas pipeline that ran on advanced automated control software. The Soviets planned to steal from a Canadian company that specialized in this kind of programming.

According to accounts, the CIA the Canadians to place delibera Soviet pipeline.

The unknowing Soviets went ah June 1982, the explosion occurr pipeline, which had cost tens of revenue.

9. Knight's $440M in bad trades, 2012

Losing $440 million is a bad day at the office by anyone's standards. Even more so when it happens in just 30 minutes due to a software error that wipes 75% off the value of one the biggest capital groups in the world.

Knight Capital Group had invested in new trading software that was supposed to help them make a killing on the stock markets. Instead, it ended up killing their firm. Several software errors combined to send Knight on a crazy buying spree, spending more than $7 billion on 150 different stocks.

... step in to ... in a year

10. ESA Ariane 5 Flight V88, 1996

Given the complexity and expense of space exploration, it's no wonde missions on our list of all-time software errors. However, the Europea even harsher cautionary tale than the rest, as it was caused by more t

Just 36 seconds after its maiden launch, the rocket engines failed due code from Ariane 4 and a conversion error from 64-bit to 16-bit data.

The failure resulted in a $370 million loss for the ESA, and a whole hos subsequent investigation, including calls for improved software analysis and evaluation.

11. The Millennium Bug, 2000

The Millennium Bug, AKA the notorious Y2K, was a massive concern in the lead-up to the year 2000. The concern was that computer systems around the world would not be able to cope with dates after December 31, 1999, due to the fact that most computers and operating systems only used two digits to represent the year, disregarding the 19 prefix for the twentieth century. Dire predictions were made about the implosion of banks, airlines, power suppliers and critical data storage. How would systems deal with the 00 digits?

The anticlimatic answer was "pretty well, actually". The millennium bug was a bit of a non-starter and didn't cause too many real-life problems, as most systems made adjustments in advance. However, the fear caused by the potential fallout throughout late 1999 cost thousands of considerable amounts of money in contingency planning and preparations, with institutions, businesses and even families expecting the worst. The USA spent vast quantities to address the issue, with some estimates putting the cost at $100 billion.

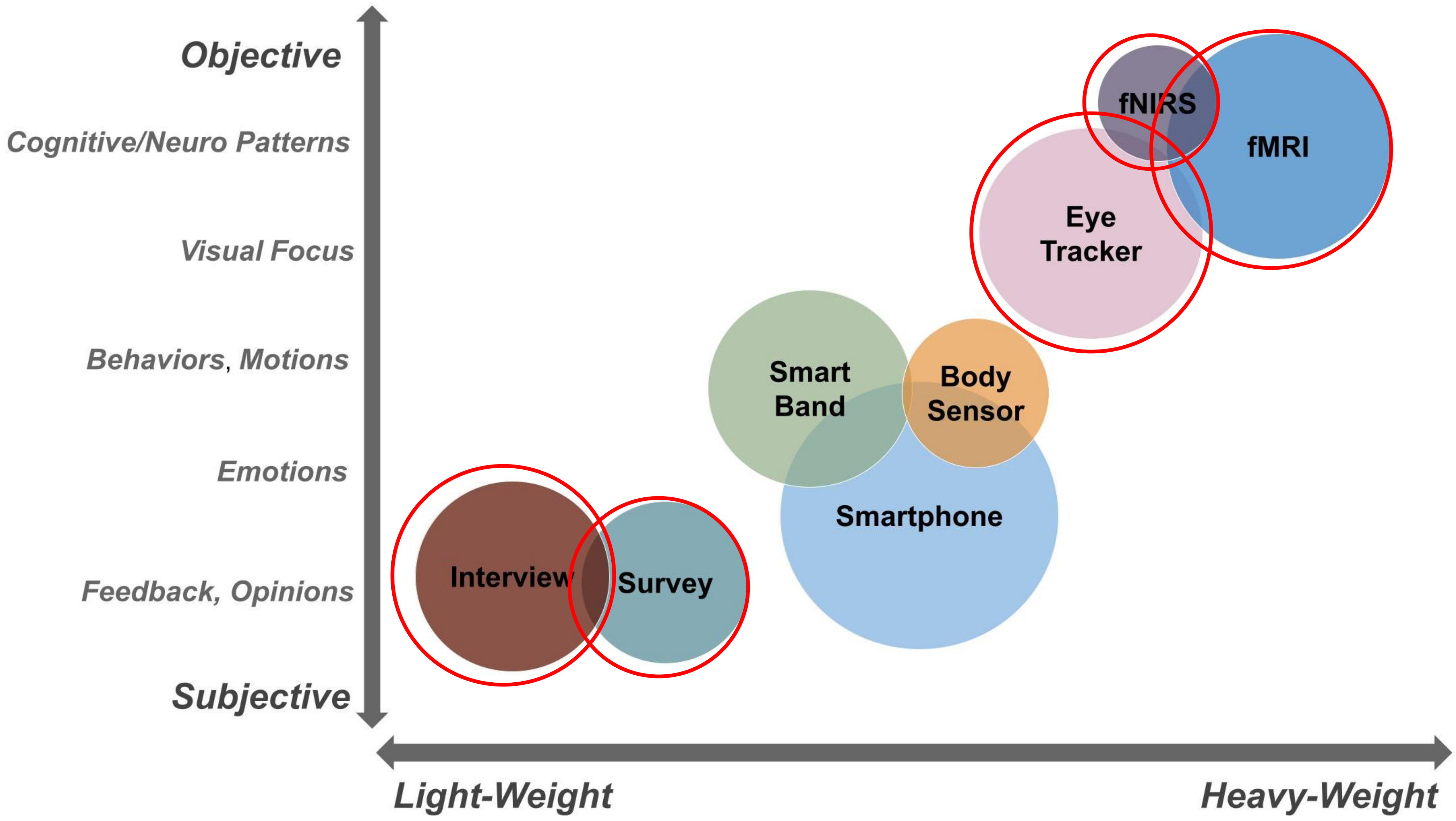
The Human Aspect Matters

- Early study of industrial developers found ***order-of-magnitude*** individual variations

Metric	Poorest	Best	Ratio
Debugging Hours Algebra	170	6	**28:1**
Debugging Hours Maze	26	1	**26:1**
CPU Seconds Algebra	3075	370	**8:1**
CPU Seconds Maze	541	50	**11:1**
Code Writing Hours Algebra	111	7	**16:1**
Code Writing Hours Maze	50	2	**25:1**
Program Size Algebra	6137	1050	**6:1**
Program Size Maze	3287	651	**5:1**
Run Time Algebra	7.9	1.6	**5:1**
Run Time Maze	8.0	0.6	**13:1**

H. Sackman, W. J. Erikson and E. E. Grant. *Exploratory Experimental Studies Comparing Online and Offline Programming Performance.* Communications of the ACM, 1968.

How to measure human aspects?



fMRI vs. fNIRS

Measure brain activities by calculating the **b**lood-**o**xygen **l**evel **d**ependent **(BOLD)** signal

- **F**unctional **M**agnetic **R**esonance **I**maging
    - ***Magnets***
    - **Strong** penetration power
    - Lying down in a magnetic tube:
        - Cannot move

- **F**unctional **N**ear-**I**nfra**R**ed **S**pectroscopy
    - ***Light***
    - **Weak** penetration power
    - Wearing a specially-designed cap:
        - More freedom of movement

fMRI vs. fNIRS

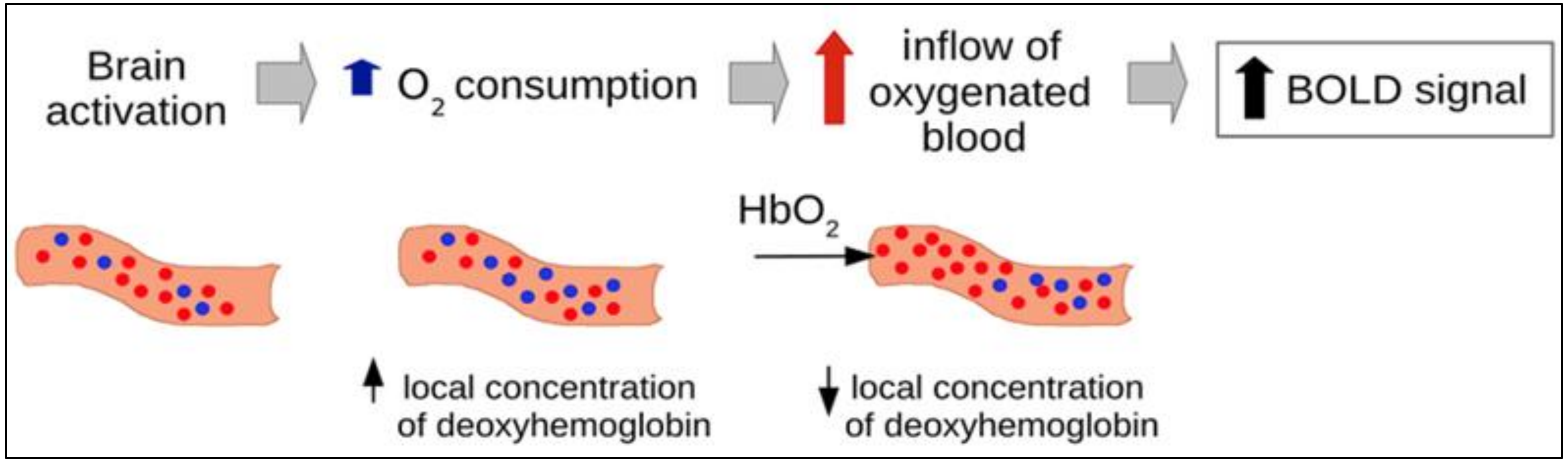
Measure brain activities by calculating the **b**lood-**o**xygen **l**evel **d**ependent **(BOLD)** signal

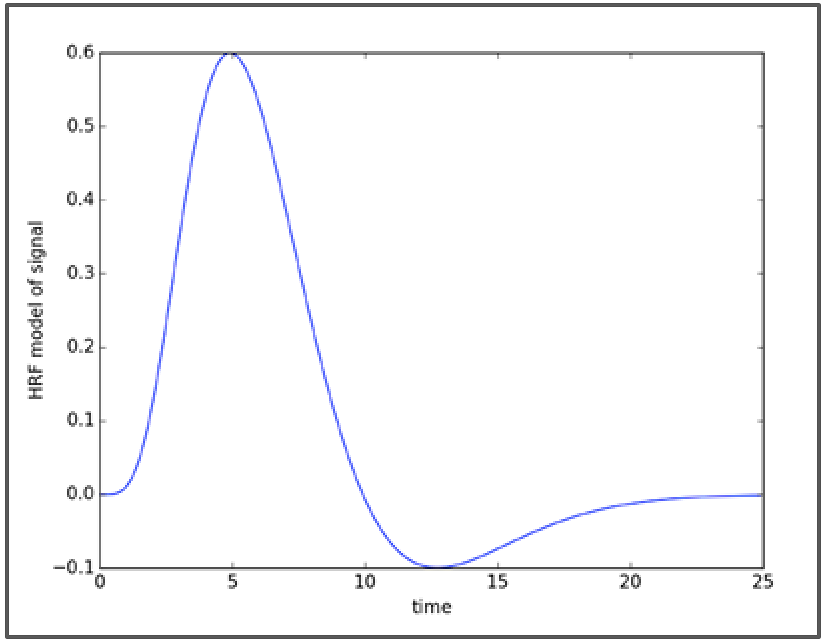
- **F**unctional **M**agnetic **R**esonance **I**maging
  - ***Magnets***
  - **Strong** penetration power
  - Lying down in a magnetic tube:
    - Cannot move

- **F**unctional **N**ear-**I**nfra**R**ed **S**pectroscopy
  - ***Light***
  - **Weak** penetration power
  - Wearing a specially-designed cap:
    - More freedom of movement

What is *BOLD* signal?

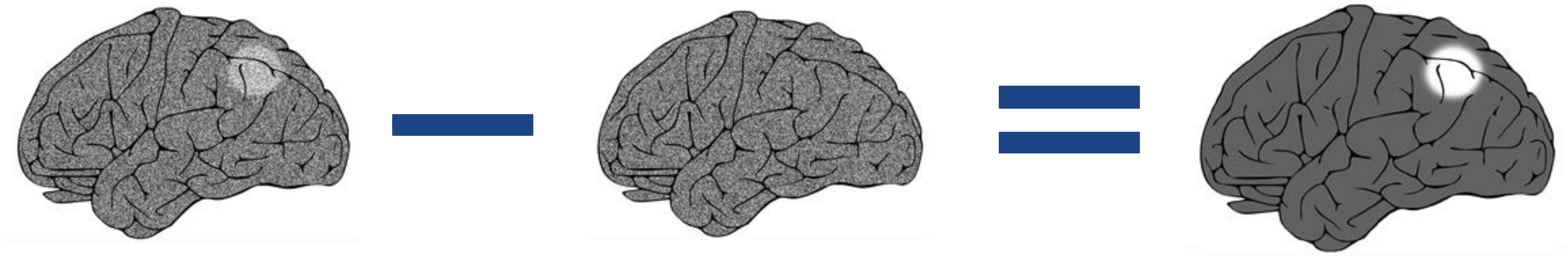
- **B**lood-**O**xygen **L**evel **D**ependent **(BOLD)** signal
- Blood flow and oxygen consumption as a **proxy** for brain activity
- Activation model: hemodynamic response function (HRF)
- Stimulus, HRF, design matrix, noise
  - Comprehensive quantitative model of BOLD signals
    - General Linear Model (GLM)





Think in Terms of Contrasts!

- Controlled experimental design
    - Task A = “**balancing trees** + nervous + …”
    - Task B = “**rotating 3D objects** + nervous + ...”
    - Contrast **A** > **B**: brain activations that vary between the tasks



Data Analysis

- We need to be ***careful***
    - 153,000 voxels or more
    - Spurious correlations due to multiple comparison: false positives

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

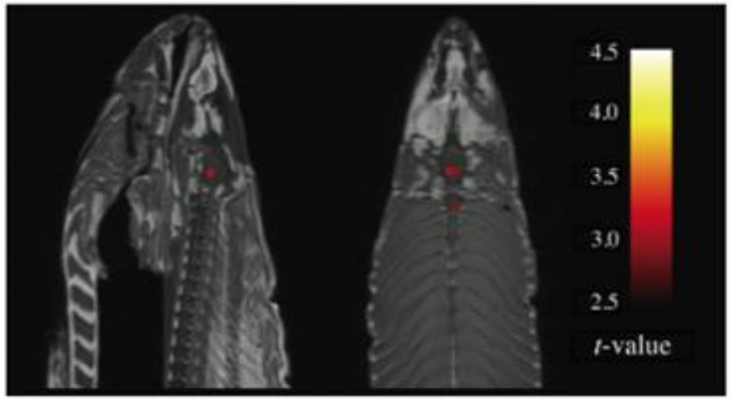
INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at

GLM RESULTS



Data Analysis

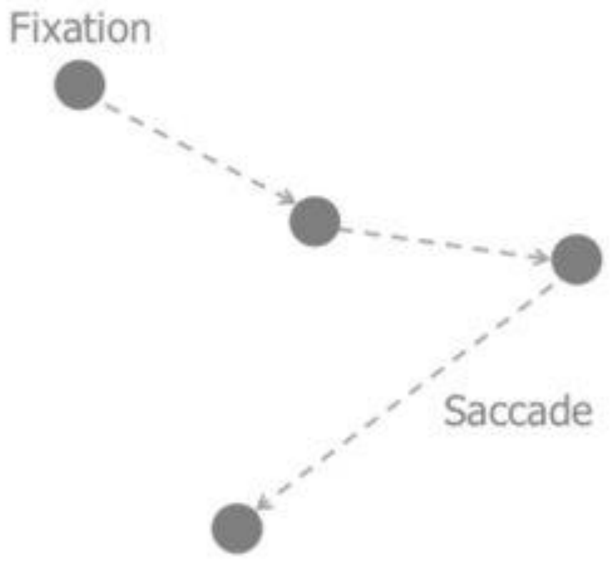
- False discovery rate (FDR) correction ($q<0.05$)

Eye-tracking

- Collect participants' visual attention by recording **eye-gaze** data: what are you looking at? How do you look at it?

Eye-tracking: how we "look"

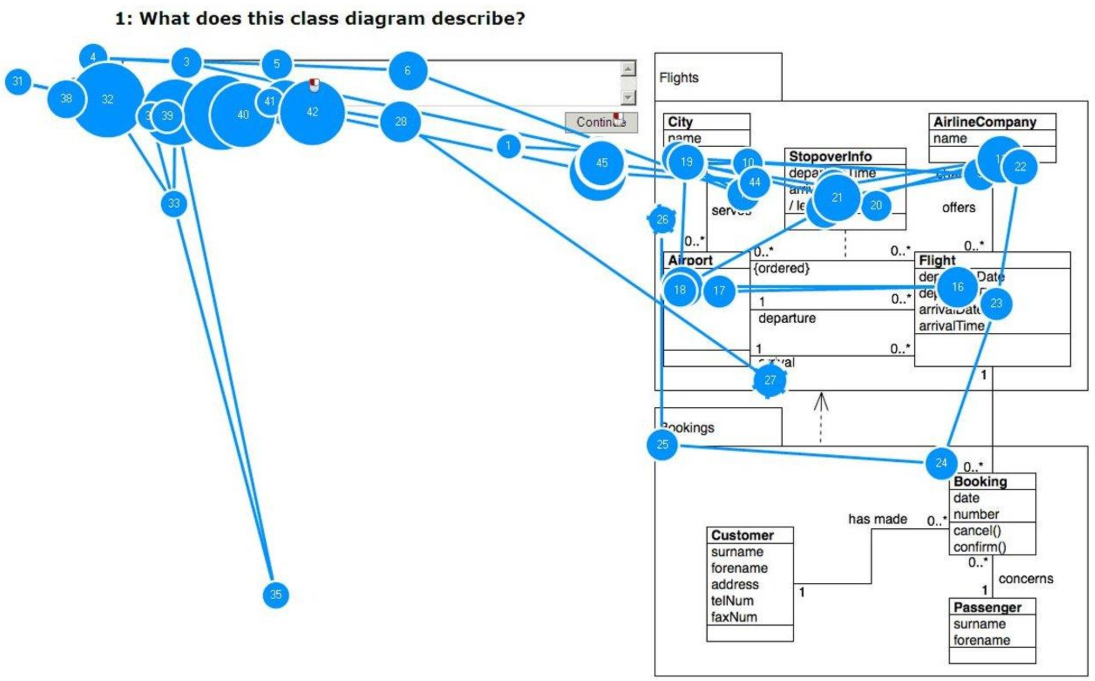
- Fixation: a spatially stable eye-gaze that lasts for approximately 100-300ms
  - Most of the information acquisition and processing occur during fixations
  - Only a small set of fixations is necessary to process a complex visual stimulus
- Saccade: continuous and extremely rapid eye movements, within 40-50ms, that occur between fixations
- Pupil size
  - Dilation is associated with cognitive work load

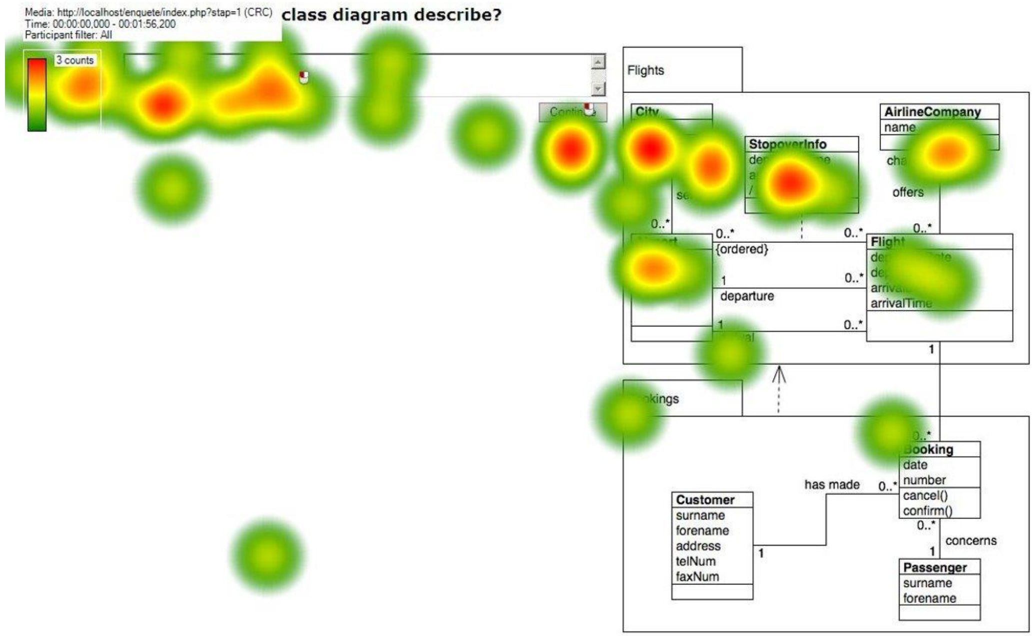


Eye-tracking: assumptions

- The immediacy assumption (Just and Carpenter, 1980):
  - The comprehension begins as soon as a participant sees a stimulus, e.g., as soon as a reader reads a word
- The eye-mind assumption:
  - The participant fixates her attention on a part of the stimulus until she understands that part

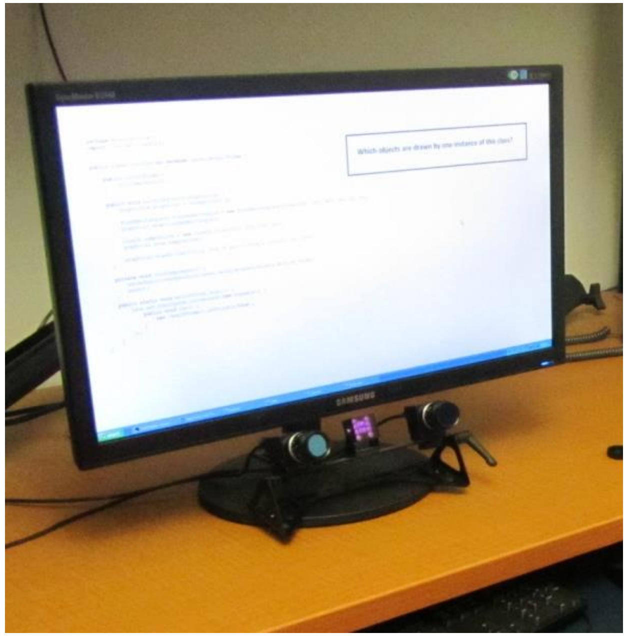
Eye-tracking: gaze plot, heat map, and raw data






```
Slide, Code, Number, x, y, Pupil X, Pupil Y, Time
Slide, Code, Number, x, y, Pupil X, Pupil Y, Start Time, End Time, Duration
C:\diagrams\1-M.png, K Space U, 3, 0, 0, 0.000000, 0.000000, 0.000, 0.000, 0.000
C:\diagrams\1-M.png, G, 1, 751, 1063, 39.000000, 38.000000, 0.297
C:\diagrams\1-M.png, G, 2, 688, 918, 39.000000, 38.000000, 0.314
C:\diagrams\1-M.png, G, 3, 688, 918, 39.000000, 38.000000, 0.331
C:\diagrams\1-M.png, G, 4, 688, 918, 39.000000, 38.000000, 0.347
C:\diagrams\1-M.png, G, 5, 684, 911, 39.000000, 38.000000, 0.364
C:\diagrams\1-M.png, G, 6, 683, 906, 39.000000, 38.000000, 0.381
C:\diagrams\1-M.png, G, 7, 683, 906, 39.000000, 38.000000, 0.397
C:\diagrams\1-M.png, G, 8, 683, 906, 39.000000, 38.000000, 0.414
C:\diagrams\1-M.png, G, 9, 681, 900, 39.000000, 38.000000, 0.431
C:\diagrams\1-M.png, G, 10, 678, 892, 38.000000, 38.000000, 0.447
```

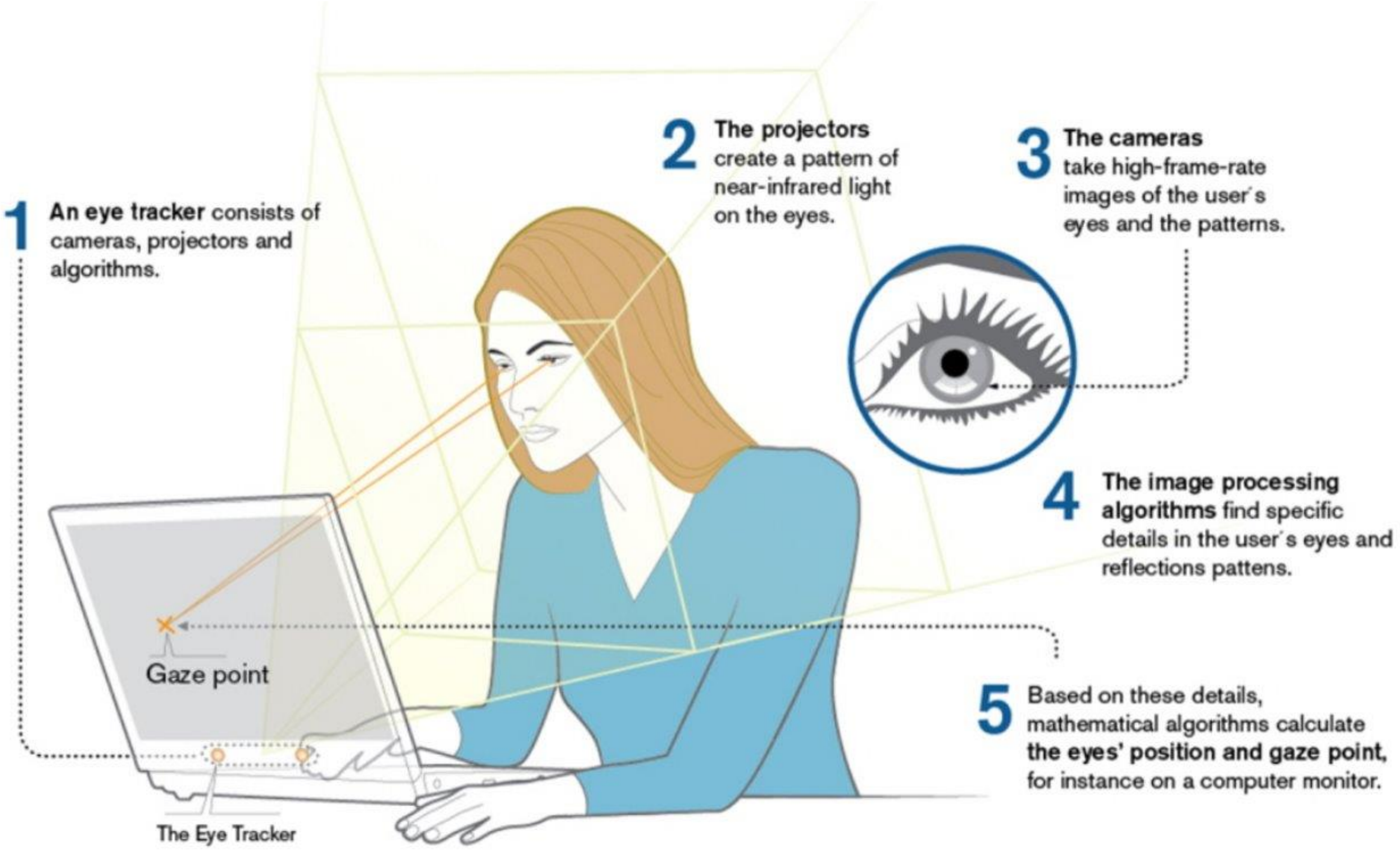
Eye-tracking: eye trackers



https://www.tobiipro.com/

https://www.tobiipro.com/

Eye-tracking: how does an eye tracker work?

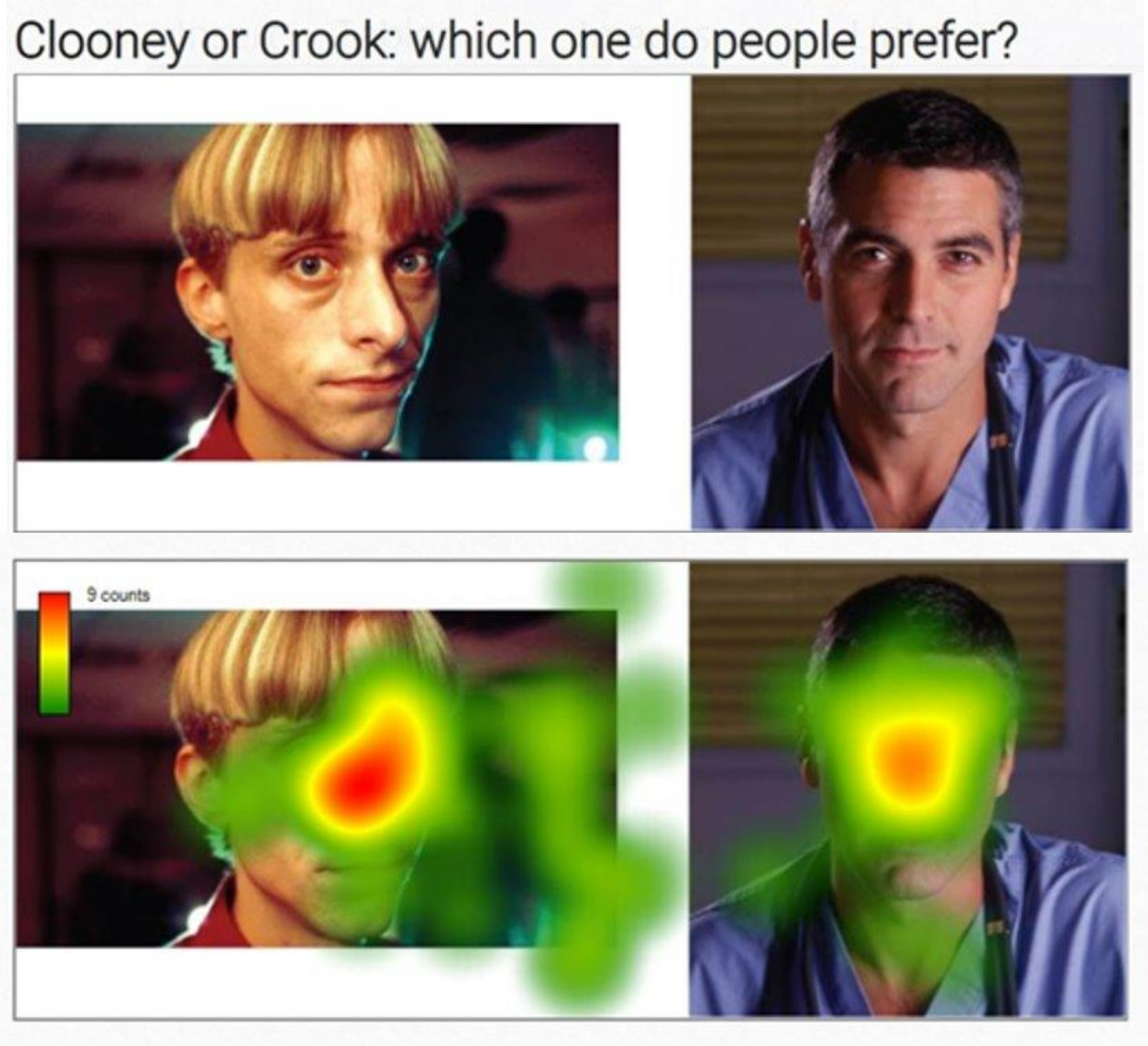


Eye-tracking: truth?

- Eye tracking allows you to know what people are thinking

Eye-tracking: truth?

- Eye tracking allows you to know what people are thinking



Eye-tracking: truth?

Misconception

~~Truth~~ about eye tracking

- Eye tracking allows you to know what people are thinking

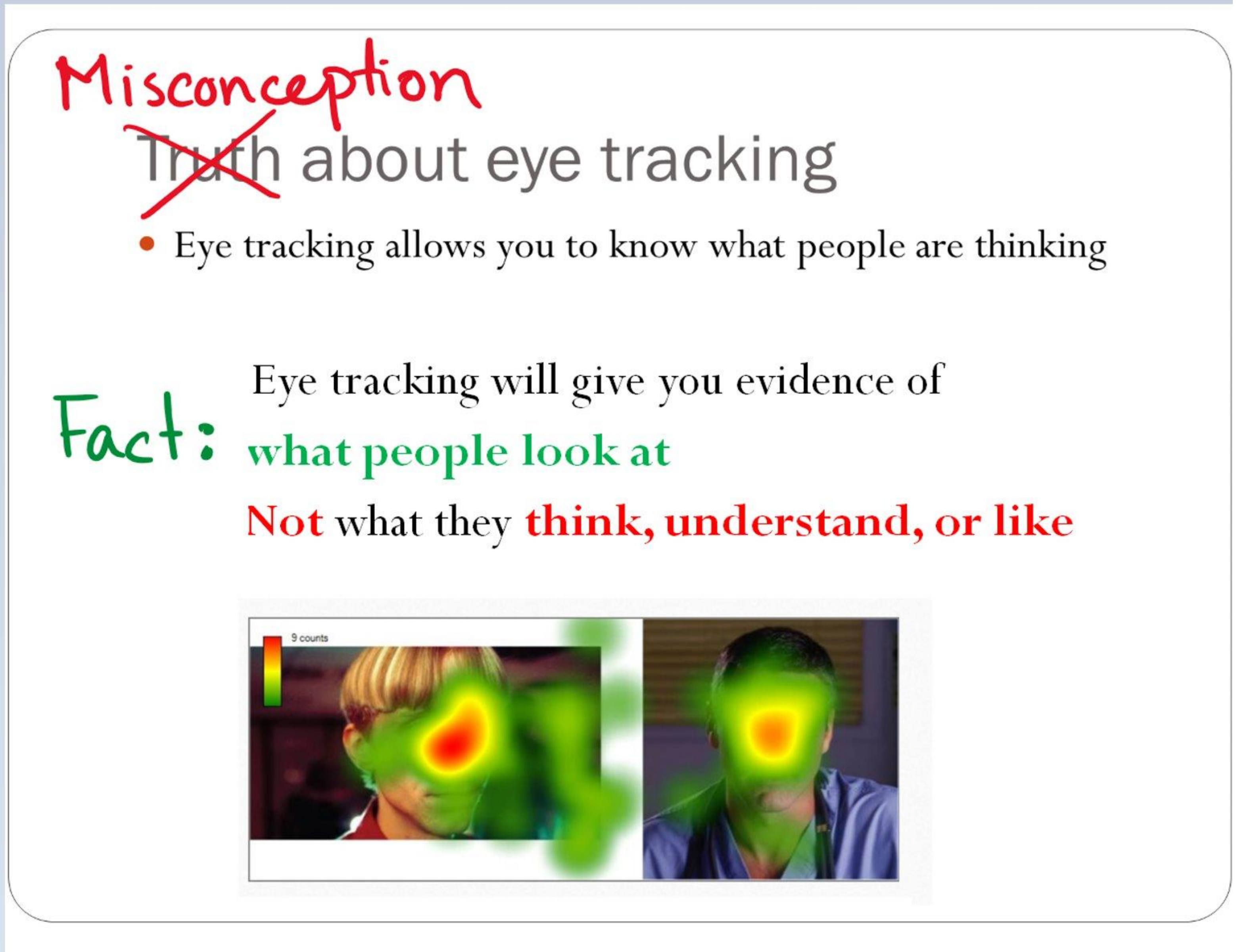
Fact:

Eye tracking will give you evidence of

what people look at

Not what they think, understand, or like

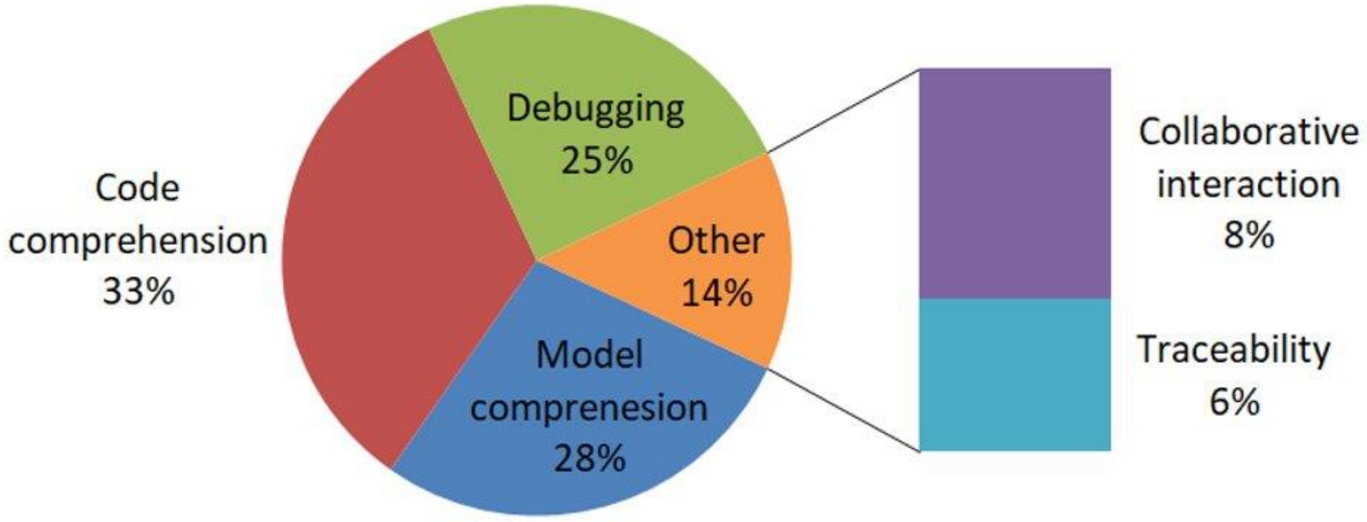
Eye-tracking: truth?



- Combination:
  - Medical imaging
  - Surveys, interviews

Eye-tracking: for software engineering

Classification of SE eye tracking papers based on category (2015)



Code					Model				English text	Other
Pascal	C/C++	Java	C♯	Python	UML	ER	Tropos	BPMN		
2	3	16	1	1	7	1	1	1	2	3 applications

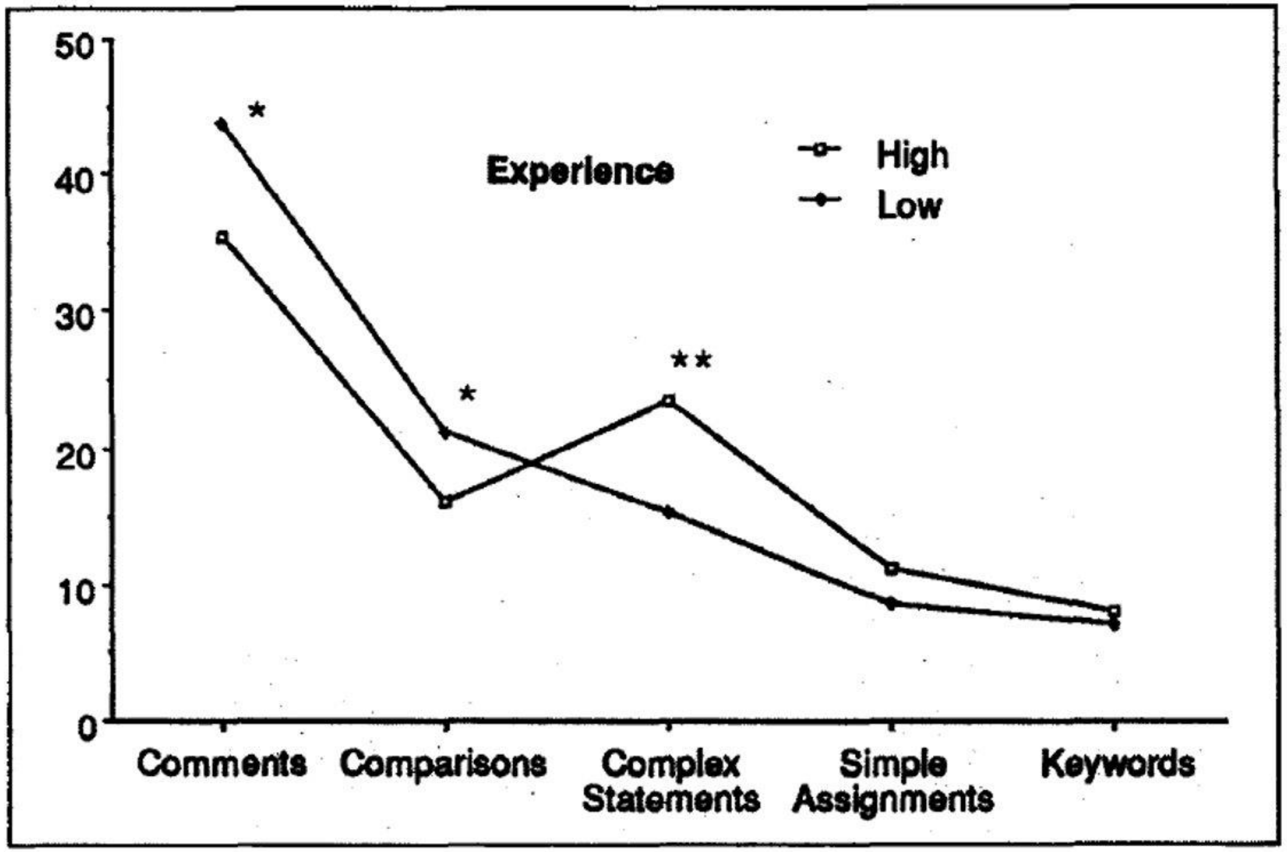
Eye-tracking: for software engineering

Types of SE questions in eye tracking experiments

Category	Type of Questions
Finding the Areas of Interest	What items or what parts of artifact (X), do participants view while performing task (Y)? Example: Does experience influence a participants focus on critical areas of the algorithm? (Crosby and Stelovsky, 1990)
Navigation Strategies	How do participants navigate through artifact/system (X) while performing task (Y)?
	Does the type of artifact (X) impact the participants' navigation strategies while they perform task (Y)?
	Do the participants' individual characteristics (Z) impact their strategies while they perform task (Y)? Example: Do the viewing patterns of experienced participants dier from those of novices?

Eye-tracking: for software engineering

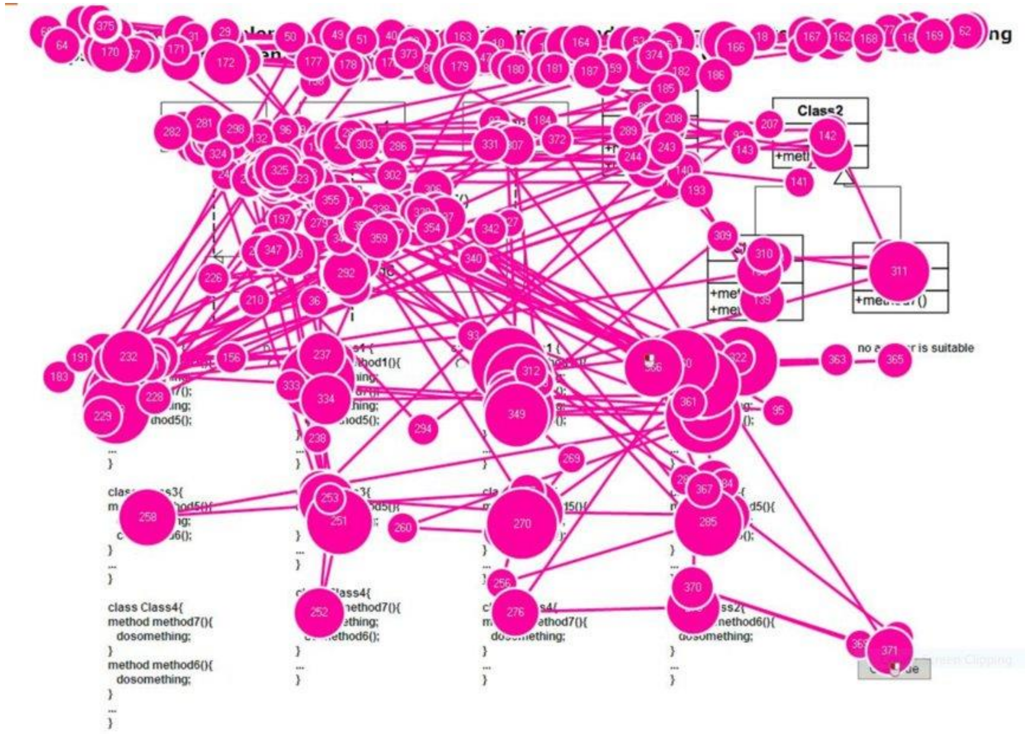
Martha Crosby 1990
Algorithm areas viewed: novices vs. experts

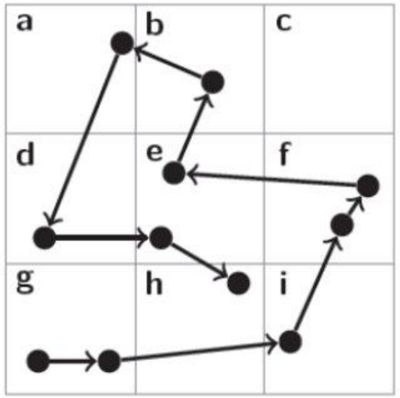


Eye-tracking: for software engineering

Scan path analysis

- A series of fixations or visited AOIs (Area of Interest) in chronological order.





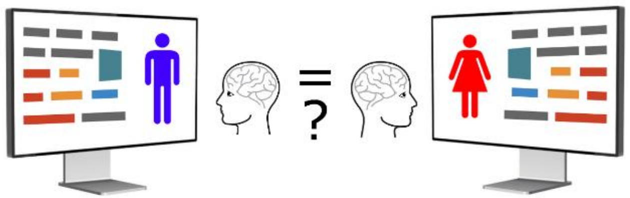
Eye-tracking: for software engineering

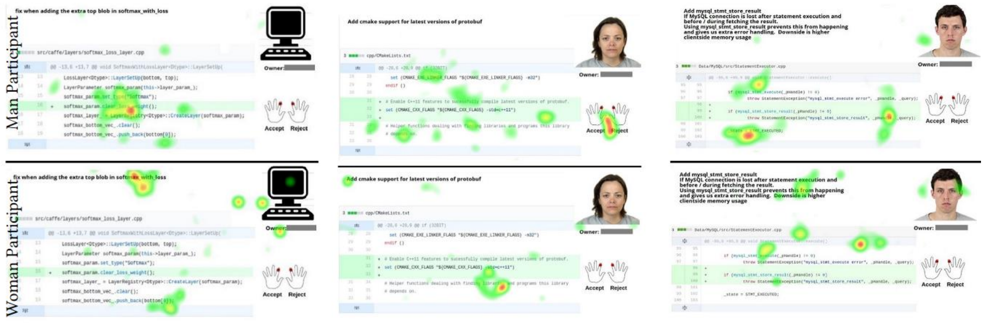
Recent work:

- combined with other measures, e.g., medical imaging
- Investigate human biases in SE activities: e.g., gender, social info

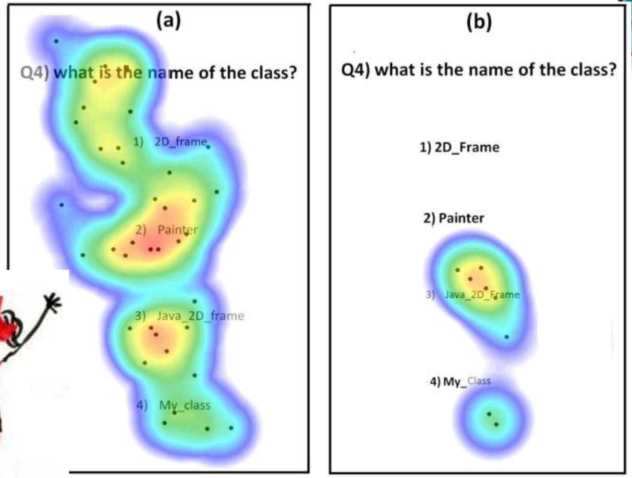
Biases and Differences in Code Review using Medical Imaging and Eye-Tracking: Genders, Humans, and Machines

Yu Huang, Univ. of Michigan, Ann Arbor, MI, USA, yhhy@umich.edu
Kevin Leach, Univ. of Michigan, Ann Arbor, MI, USA, kjleach@umich.edu
Zohreh Sharafi, Univ. of Michigan, Ann Arbor, MI, USA, zohrehsh@umich.edu
Nicholas McKay, Univ. of Michigan, Ann Arbor, MI, USA, njmckay@umich.edu
Tyler Santander, Univ. of California, Santa Barbara, Santa Barbara, CA, USA, t.santander@psych.ucsb.edu
Westley Weimer, Univ. of Michigan, Ann Arbor, MI, USA, weimerw@umich.edu

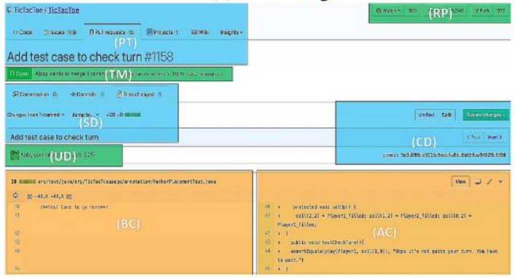




(a) A stimulus with a machine author (b) A stimulus with a woman author (c) A stimulus with a man author



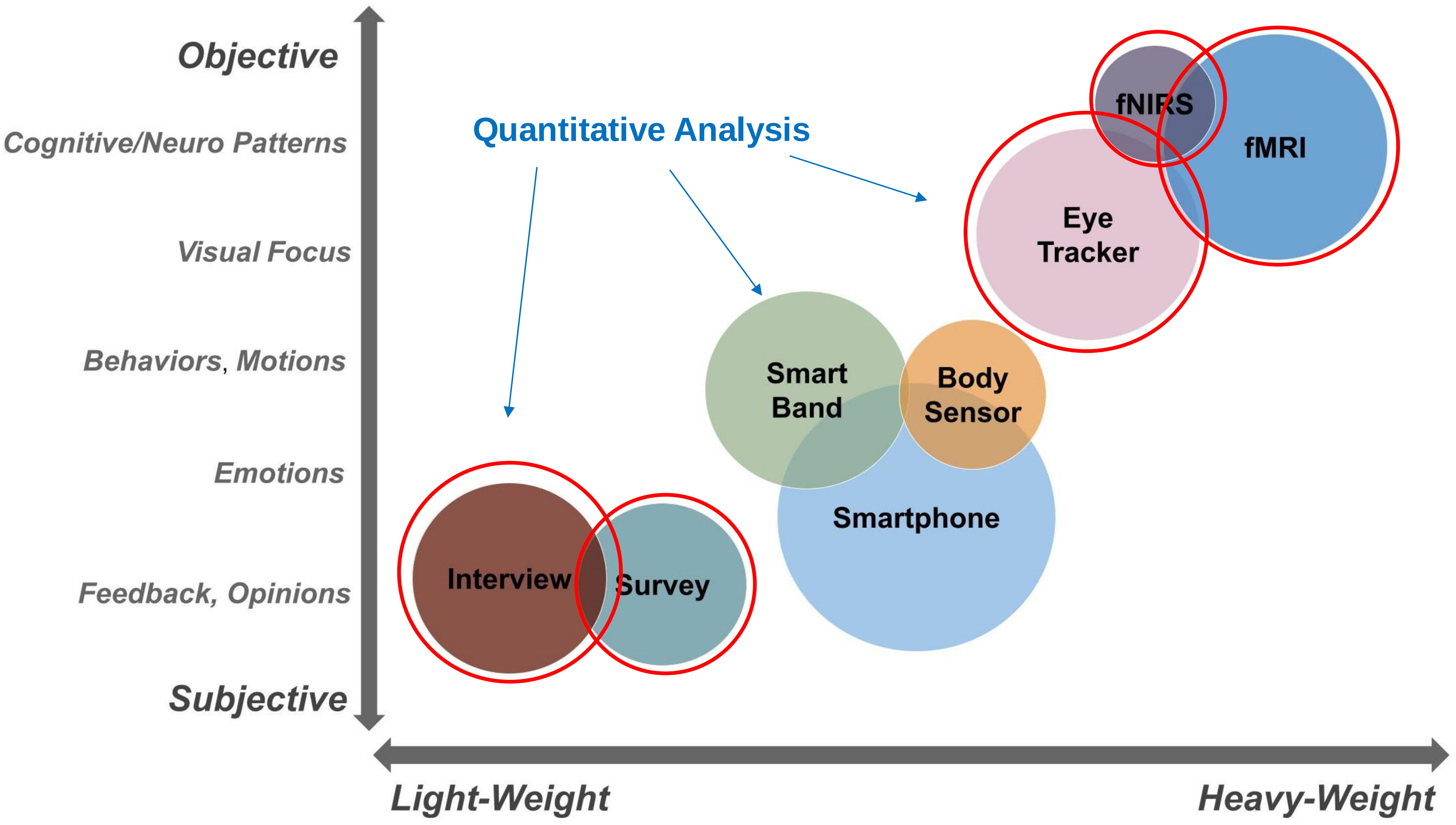
(a) Profile Page



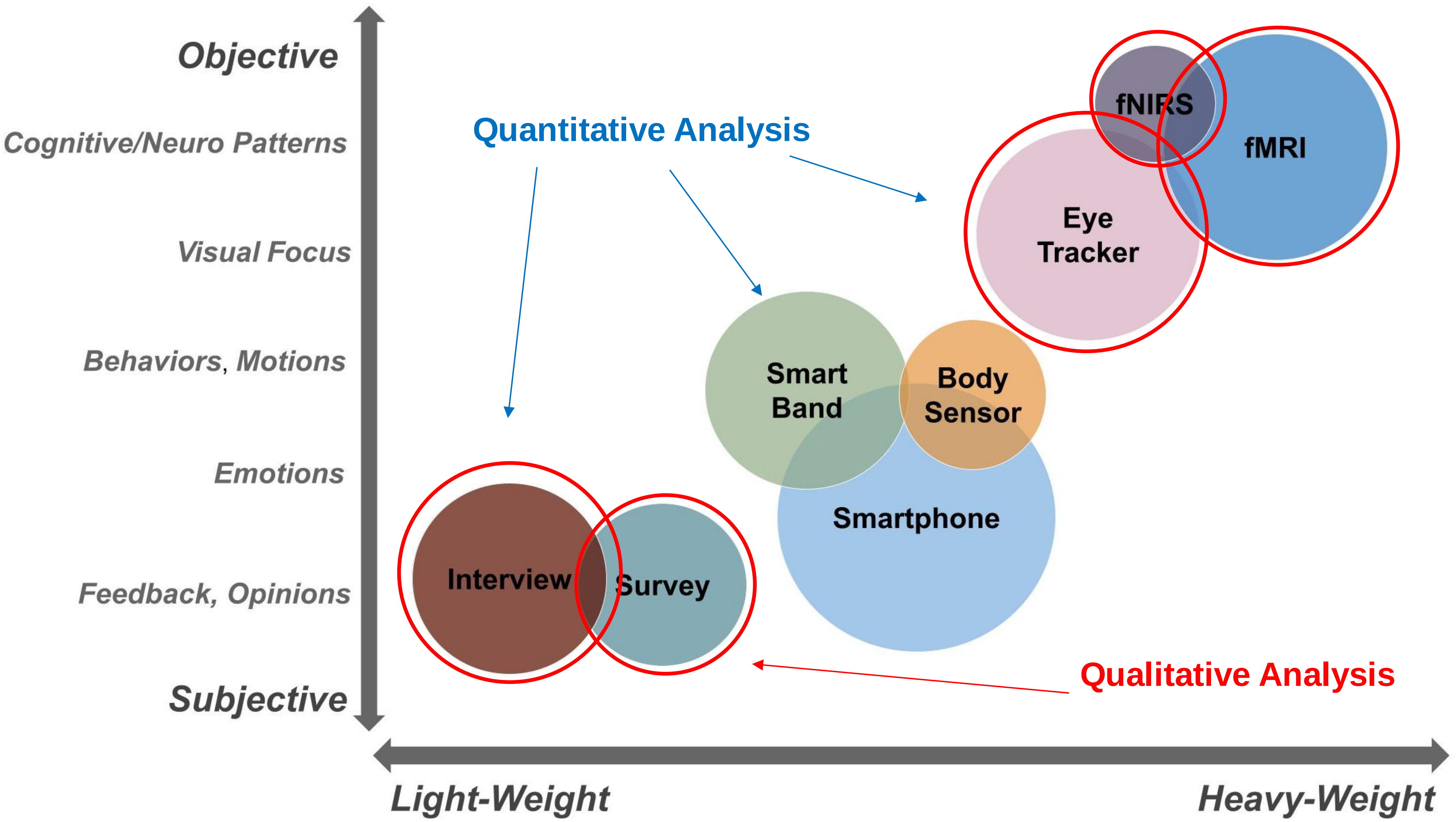
Beyond the Code Itself: How Programmers *Really* Look at Pull Requests

Denae Ford, Mahnaz Behroozi, *North Carolina State University*, Raleigh, NC, USA, {dford3, mbehroo}@ncsu.edu
Alexander Serebrenik, *Eindhoven University of Technology*, Eindhoven, The Netherlands, a.serebrenik@tue.nl
Chris Parnin, *North Carolina State University*, Raleigh, NC, USA, cjparnin@ncsu.edu

How to analyze human aspects?



How to analyze human aspects?



How to analyze human aspects: qualitative analysis

- Verbally-acquired data
  - Information that is gathered via speech, think-aloud protocol, oral retrospection, formal or informal interviews and surveys

With appropriate care in data gathering and analysis, verbal data *can* provide impactful insights in software engineering research.

How to analyze human aspects: qualitative analysis

- Verbally-acquired data
  - Information that is gathered via speech, think-aloud protocol, oral retrospection, formal or informal interviews and surveys
- Classic example: the "Sillito et al." Questions, published in FSE ‘06, cited over 350 times

them. Participants in the second study (E1...E16) were observed working on code with which they had experience. In both studies

During each session an audio recording was made of discussion between the pair of participants, a video of the screen was captured,

To structure our data collection and the analysis of our results, we have used a *grounded theory* approach which has been described as an emergent process intended to support the production of a theory that “fits” or “works” to explain a situation of interest [5, 19]. In

Questions Programmers Ask During Software Evolution Tasks

Jonathan Sillito, Gail C. Murphy and Kris De Volder
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada
{sillito,murphy,kdvolder}@cs.ubc.ca

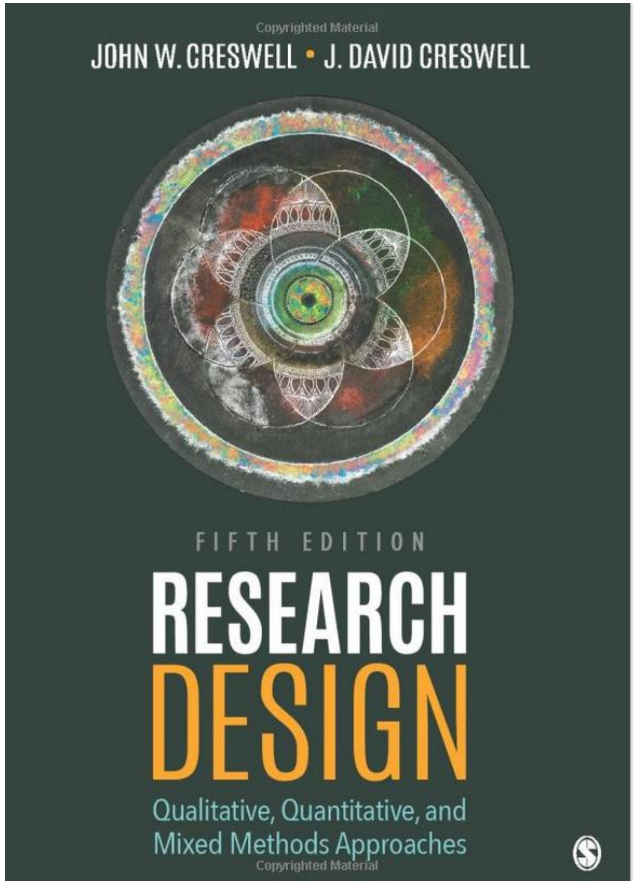
about the source code on which we observed them working. We report on 44 kinds of questions we observed our participants asking. These questions are generalized versions of the specific ques-

Results are useful directly (a structured answer to a fundamental question) and also as artifacts (re-used by later projects as indicative developer queries)

[Sillito, Murph, De Volder. Questions programmers ask during software evolution tasks. FSE 2006.]

Qualitative Analysis: Metrics

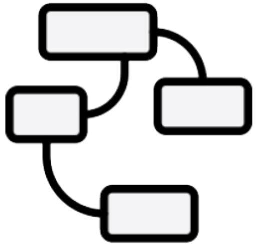
- Establishing **validity** in qualitative research
  - Using multiple validity procedures
    - Member checking
    - Clarify bias
    - Spend prolonged time in the field
  - Using qualitative reliability
    - Document your procedures (scripts, codebook, etc.)
    - No drift in the definition of codes
    - Cross-check codes developed by different researchers



Showing Prompts

Audio Recording

Transcribing



Qualitative analysis

[Chi. Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. J. Learning Sciences 1997.]

Qualitative Analysis: Useful Techniques

- Grounded theory in SE
  - Similar to socio-technical studies, qualitative research can have a lot of variance
    - How can we mitigate that variance?
- Grounded Theory is a systematic methodology for qualitative research for constructing hypotheses via inductive (not deductive) reasoning
  - Method
    - Empirical/evidence based
  - Outcome
    - Key patterns of the data
    - Relationships between patterns

"It is not in your mind; it is in your data."

[Hoda. Socio-Technical Grounded Theory for Software Engineering. IEEE Trans. Software Engineering 2021.]

Qualitative Analysis: Useful Techniques

- Grounded theory in SE
- Inductive Thematic Analysis
  - Thematic exploration (thematic coding)
    - Codes and the relationships

Category	Code	Description
motivation	motivation-helpuser	help end users
	motivation-helpdev	help developers
	motivation-longterm	how to keep yourself engaged in the project for a long time
	motivation-giveback	altruism
	motivation-impact	want to make impact
	motivation-better-programmer	want to look good in the community, improving skills, build up portofolio
	mitivation-hobby	I feel happy/fun, e.g., as a hobby.
	motivation-work	This is my job, or school projects, etc

Codebook Example

Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good

Yu Huang
University of Michigan
Ann Arbor, MI
yhhy@umich.edu

Denae Ford
Microsoft Research
Redmond, WA USA
denae@microsoft.com

Thomas Zimmermann
Microsoft Research
Redmond, WA USA
tzimmer@microsoft.com

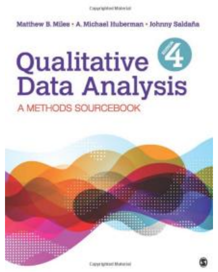


TABLE II: Themes of Motivations for Contributing to OSS for Social Good.

Theme	Description	Representative Example	Participants
To help those in need	Contributors wanted to help people who are in need but may lack the capability of solving the problems themselves.	*"I'm so much more motivated to build products that I know have a good outcome for a group of people that is generally underserved."*	P2, P3, P4, P5, P6, P7, P8, P9, P10, P12, P14, P18, P19
To become a better programmer	Contributors wanted to improve their skills, build up their portfolios, or improve their reputation in the community.	*"when I contribute to that, it can definitely give me more experience."*	P2, P3, P5, P10, P11, P12, P14, P16, P17, P20
To have an impact on society	Contributors wanted to make a difference to the society.	*"So, I think the main reason is because I want to make a difference on my life... make a fingerprint on the world."*	P1, P3, P4, P7, P13, P14, P15, P17
For emotional fulfillment	Contributors were motivated by feeling good about the impacts of the project.	*"It gives a mental satisfaction that I'm working towards something good"*	P3, P4, P10, P11, P12, P17, P20
To help fellow developers with their project	Contributors want to help the developers to achieve the accomplishment of the projects.	*"Another is to help the people in the project to help reach their goals."*	P3, P7, P10, P12, P13, P18
To give back as I received	Contributors want to give back to the society (e.g., altruism).	*"And I also feel like however much you take from something, you should give back."*	P4, P5, P9, P16, P20
To meet like-minded people	Contributors wanted to get to know more people.	*"I think it brings like-minded people together most of the time, so I get to interact with people who are working on similar project or they have similar interests."*	P11, P13, P17
As a hobby	Contributors worked in OSS4SG as a hobby or something they like doing.	*"I've moved to sales but still collaborating ... It's just as a hobby."*	P14, P15
Because I need it for work	Contributors worked on OSS4SG for their professional work projects.	*"So the direct cause that I found it is through [elided]'s little competition."*	P2

Qualitative Analysis: Useful Techniques

- Grounded theory in SE
- Inductive Thematic Analysis
  - Thematic exploration
    - Codes and the relationships
  - Evaluation metrics
    - Saturation
    - Agreement

Qualitative Analysis: Useful Techniques

- Grounded theory in SE
- Inductive Thematic Analysis
  - Thematic exploration
    - Codes and the relationships
  - Evaluation metrics
    - Saturation
    - Agreement
- Inter Rater Reliability (IRR) or Inter Rater Agreement (IRA)
  - Statistics as evidence
    - Cohen's kappa, Fleiss' kappa, etc.

"It is not in your mind; it is in your data."

Qualitative Analysis: Combining Verbal and Nonverbal Data

- Strength of verbal data
  - Richness and holism
  - Discovery
  - New ideas, hypothesis
- Weakness of verbal data
  - Hard to evaluate the analysis (i.e., no “equations”)
  - Human biases
- Combining verbal and nonverbal data makes a strong and interesting case
  - Supplement, validate, or illuminate each other
- Contrast: surprising knowledge!

Qualitative Analysis: Combining Verbal and Nonverbal Data

- What do you think about pull requests generated by machines
  - "Machine generated code is worse on readability!"

  But all pull requests were written by humans! (We deceived you!)

- Do you think women and men write pull request differently
  - "There is no difference between pull requests written by men and women"

  But there *is* a significant difference on your behavior! Both response time and final decisions are affected!

Biases and Differences in Code Review using Medical Imaging and Eye-Tracking: Genders, Humans, and Machines

Yu Huang
Univ. of Michigan
Ann Arbor, MI, USA
yhhy@umich.edu

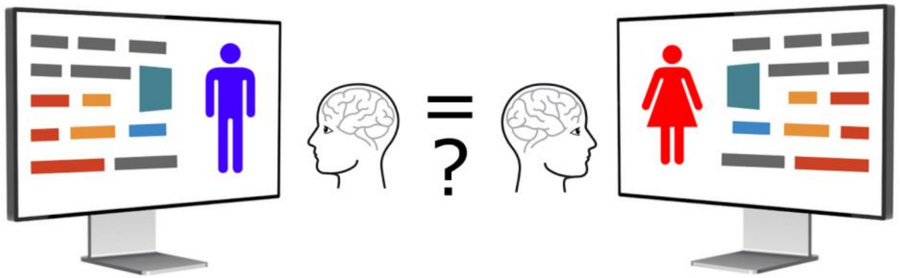
Kevin Leach
Univ. of Michigan
Ann Arbor, MI, USA
kjleach@umich.edu

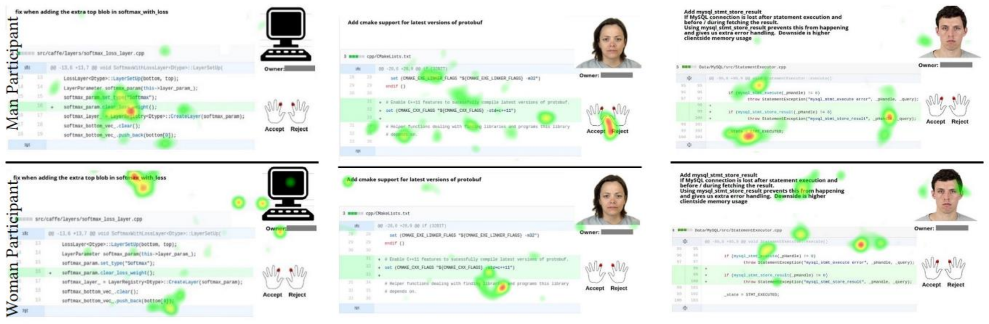
Zohreh Sharafi
Univ. of Michigan
Ann Arbor, MI, USA
zohrehsh@umich.edu

Nicholas McKay
Univ. of Michigan
Ann Arbor, MI, USA
njmckay@umich.edu

Tyler Santander
Univ. of California, Santa Barbara
Santa Barbara, CA, USA
t.santander@psych.ucsb.edu

Westley Weimer
Univ. of Michigan
Ann Arbor, MI, USA
weimerw@umich.edu





(a) A stimulus with a machine author (b) A stimulus with a woman author (c) A stimulus with a man author

Statistics: A Brief Overview

- Important but used to be overlooked in CS research
    - "The proposed system achieves a 10% higher accuracy on average compared to X in 10 runs..."
- Statistical tests
    - Is it significant?

showed notable resistance to this decline. For example, the equiprobable heuristic chose the optimal alternative almost six times as often as one would expect by change in the 8 × 2 decision situation. And five of the heuristics—E, Min, MR, ML, and P—found one of the highest two expected value alternatives over 80% of the time in the 8 × 2 decision situations. The propensity to avoid the alternatives with lowest EV decreased to well below chance for all heuristics as the number of alternatives increased. Indeed, only three heuristics

Why statistics for this class?

- A number of papers use statistical techniques, and understanding something about them will be useful.
- You may also need to run statistical tests as part of your research projects.
- Examples:
  - Is there a difference in gaze times on identifiers in Gerrit vs. GitHub?
  - Is there a relationship between how much you pay someone and how fast they complete a programming task?

Why statistics at all?

- Descriptive statistics
  - Describe or summarize the data
  - Example: What *usually* happens?
    - Mean
    - Median

- Inferential statistics
  - Intuition: Can we be confident the data is telling us the story we think it is, or did we just get lucky?
  - Does the data we have represent the data we don’t have?

Some technical terms

- Population = the items you’re interested in, e.g. all developers
- Sample = the items you’re actually looking at, e.g. 10 developers interviewed
- Distribution = the shape of the data on the plot (e.g., normal)

Hypothesis testing

- Inferential statistics can be run when you state your research problem as hypothesis, specifically using:
  - Null hypothesis (H_0) = no difference or no relationship
  - Alternative hypothesis (H_1) = a difference or relationship exists
- Example 1:
  - Null: Teams with high IQ member perform equally well as teams with high social intelligence member
  - Alternative: Two teams perform differently
- Example 2:
  - Null: No relationship exists between how much we pay someone and how quickly they complete a programming puzzle
  - Alternative: The more we pay, the faster or slower someone completes the task

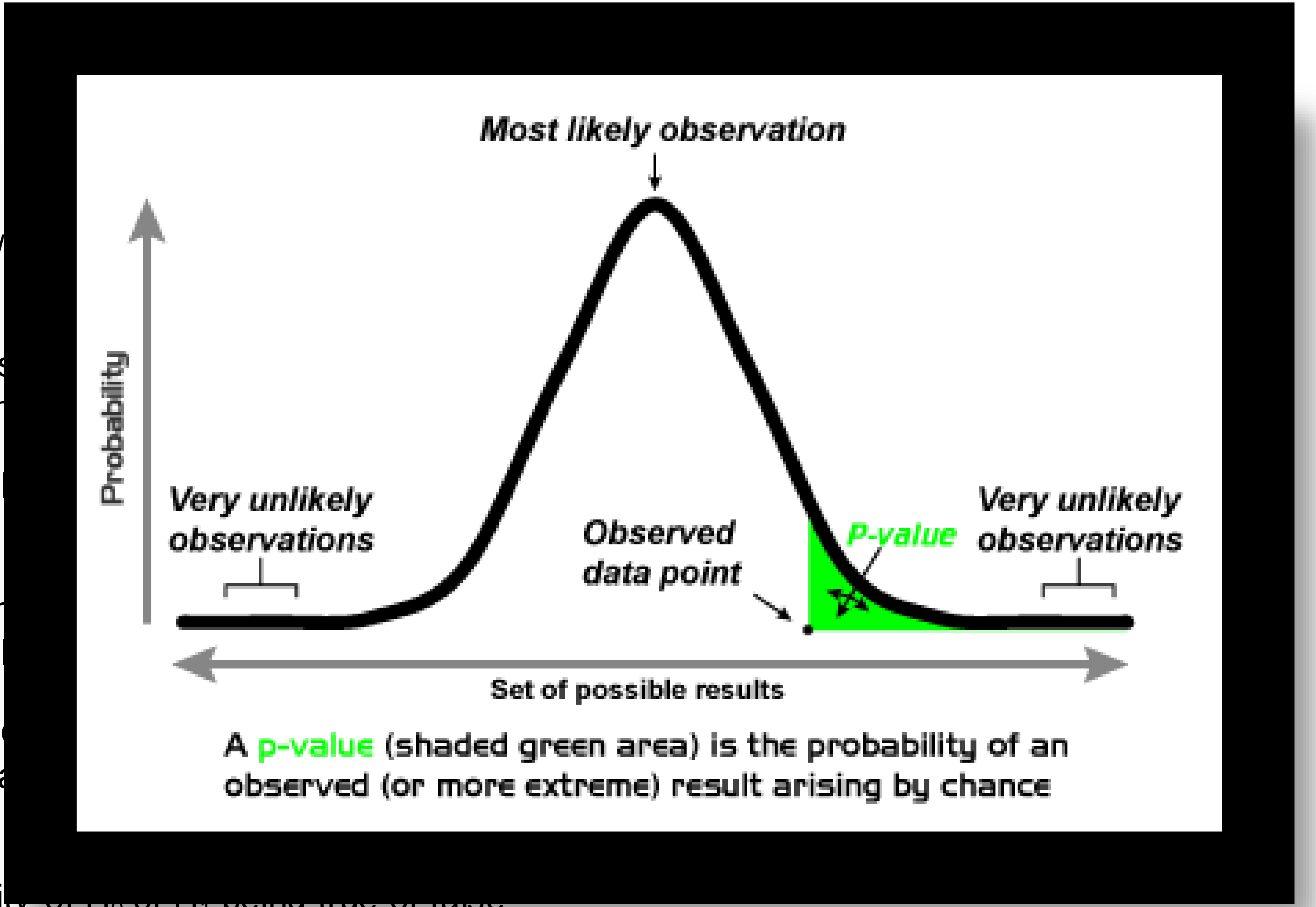
p-value: "statistically significant"

- A probability, between 0 and 1.
- Definition:
  - Technical: Assuming that the null hypothesis is true, the probability of obtaining a result this extreme or more extreme
  - Intuitive: Probability that we got this result by chance
- Use
  - We define an alpha level, below which we consider the result to be “statistically significant”. Conventionally (but for no particularly good reason) α=.05
  - If a difference or relationship appears to exist, but is not significant, we probably should not say that there is a difference at all
- What it’s not:
  - ~~The probability of H_0 or H_1 being true or false~~

https://upload.wikimedia.org/wikipedia/en/0/00/P-value_Graph.png

p-value

- A probability, betw
- Definition:
  - Technical: As
    extreme or m
  - Intuitive: Prol
- Use
  - We define ar
    Conventional
  - If a differenc
    that there is a
- What it's not:
  - The probability of H_0 or H_1 being true or false



https://upload.wikimedia.org/wikipedia/en/0/00/P-value_Graph.png

Statistical power

1. If you have an IBM developer who's 2x more productive than a Google developer, do we believe that IBM developers are more productive than Google developers?
2. What if we have 1000 IBM developers who are, on average 2x more productive than 1000 Google developers?

- Are we equally or more likely to believe (1) or (2)?
- The second situation has more **statistical power**, that is, the ability to detect a real effect
- The following affects statistical power
  - Sample size
  - Effect size
  - Statistical test (t-test, chi-square, etc)

Confidence Interval

- A range of values for which you're confident the "true" value lies
- You determine the confidence intervals, usually set at 90%, 95%, or 99%
- Similar to p-value, but integrates effect size, so more informative
- Given as x ± value

Example

- Average pulse rate = 101 bpm; Standard Deviation = 50; N = 200

- 95% Confidence Interval = (94, 108)

  We are 95% confident that the true pulse rate for our population is between 94 and 108.

  Margin of error = (108 – 94) / 2 = ± 7 bpm

  - Example: We are 95% confident that the true pulse rate for our population is between 94 and 108
- Question:

  - Does more data increase or decrease your confidence interval?

Confidence Interval

- A range of values for which you're confident the "true" value lies
- You determine the confidence intervals, usually set at 90%, 95%, or 99%
- Similar to p-value, but integrates effect size, so more informative
- Given as x ± value
- Question:
    - Does more data increase or decrease your confidence interval?
        - A larger sample size or lower variability will result in a tighter confidence interval with a smaller margin of error.
        - A smaller sample size or a higher variability will result in a wider confidence interval with a larger margin of error.
        - The level of confidence also affects the interval width. If you want a higher level of confidence, that interval will not be as tight. A tight interval at 95% or higher confidence is ideal.

Examples:

- Average Scene Time = 5.5. mins; Standard Deviation = 3 mins; N = 10 runs
- 95% Confidence Interval = (3.6, 7.4)

  Margin of Error = ±1.9 minutes

- Average Scene Time = 5.5 mins; Standard Deviation = 3 mins; N=1,000 runs
- 95% Confidence Interval = (5.4, 5.6)

  Margin of Error = ± 0.1 minutes

Two types of statistical tests

Parametric Tests

- Assume a particular distribution of data, typically normal
- Assumes differences between values are meaningful
- More statistical power
- Examples:
    - Student t-test
    - ANOVA
    - Pearson correlation

Non-parametric tests

- Does not assume a distribution
- Ignores differences between values
- Less powerful
- Examples:
    - Chi-square
    - Fisher
    - Wilcoxon and Mann-Whitney
    - Spearman

Student t-test

- "t-test"
  - Commonly used: two-sample t-test
    - test of the null hypothesis such that the means of two populations are equal.
    - Paired vs. unpaired

Student t-test

- "t-test"
    - Commonly used: two-sample t-test
        - test of the null hypothesis such that the means of two populations are equal.
        - Paired vs. unpaired
- History
    - Gets its name from William Sealy Gosset who first published it in 1908 in the scientific journal Biometrika using his pseudonym "Student", because his employer preferred staff to use pen names when publishing scientific papers instead of their real name, so he used the name "Student" to hide his identity
    - Guinness Brewery: Is beer 1 better than beer 2 using different barley? Guinness did not want their competitors to know that they were using the t-test to determine the quality of raw material

Chi-square test

- Similar to t-test
  - Frequency: categorical data
  - determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories

Wilcoxon Signed-Ranks / Rank Sum

- Non-parametric versions of paired and unpaired t-tests
- H0: for randomly selected values *X* and *Y* from two populations, the probability of *X* being greater than *Y* is equal to the probability of *Y* being greater than *X*.
- Compares medians, rather than means (so report 'em!)
- Mann-Whitney U-test = Wilcoxon rank-sum

Hypothesis 1.3: Compared to males, females make pull requests that modify fewer lines of code, modify fewer files, and contain fewer commits.

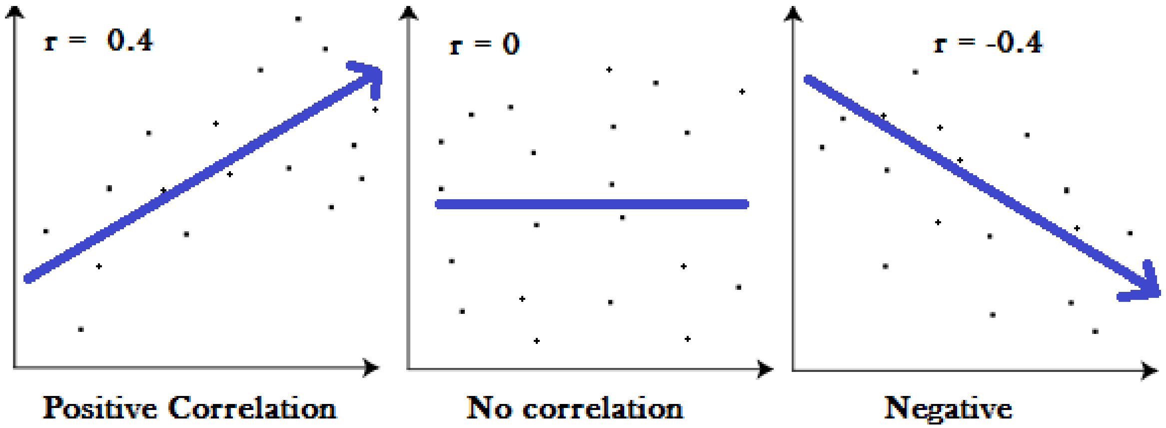
The following table lists the median and mean lines of code added (+), removed (−), files changed, and commits per pull request:

		lines +	lines −	files changed	commits
female	median	21	3	2	1
	mean	1640	617	5.4	30.4
male	median	13	2	1	1
	mean	762	299	4.1	24.5

With the exception of lines removed, all differences between females and males are significantly higher (Wilcoxon rank-sum test, $p < .001$). On threat to this analysis is that

Pearson and Spearman Correlations

- Correlation
  - Test: if there is strong association between one variable versus another.
  - Coefficient: 0-1 and p-value



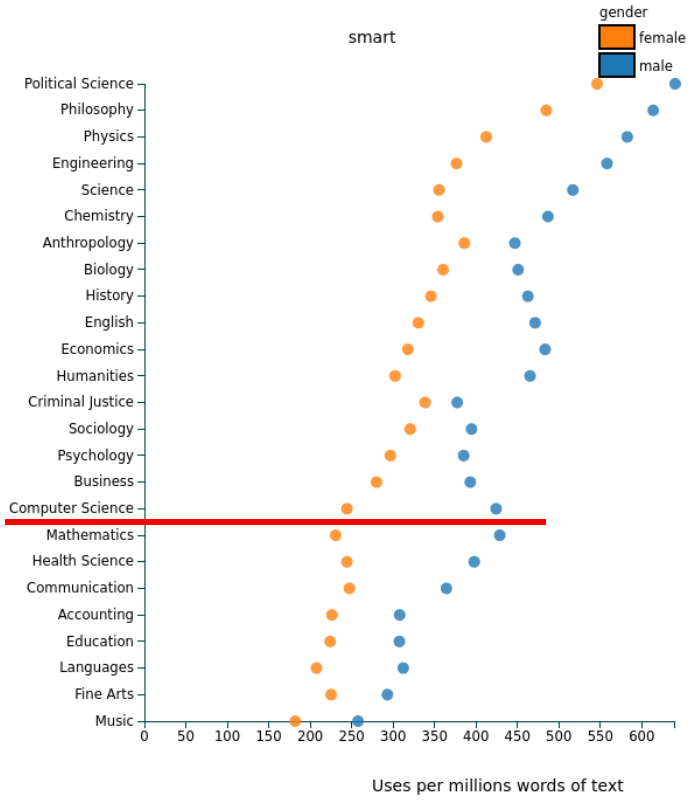
Pearson and Spearman Correlations

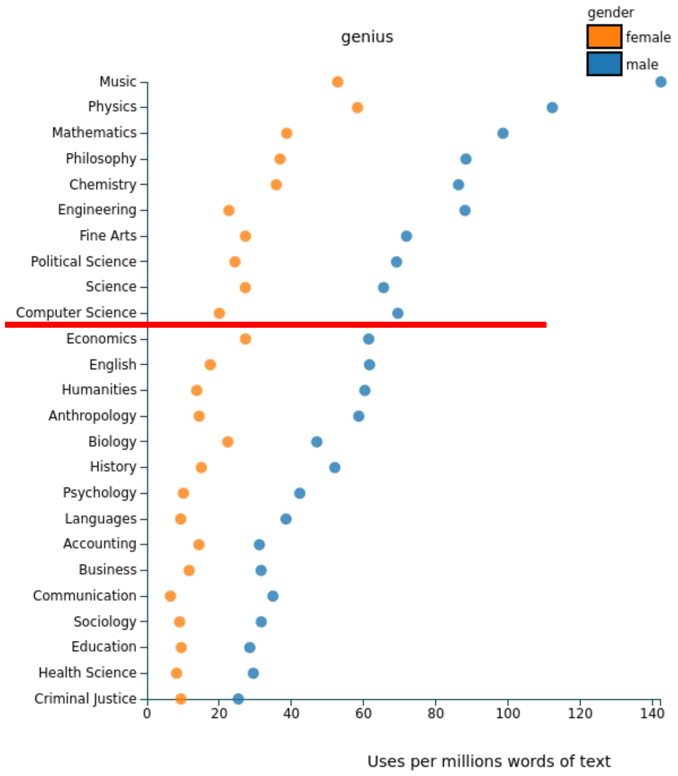
- Pearson correlation anlaysis:
  - Parametric
  - Continuous in nature: each variable is able to take on a potentially infinite number of values
  - The shape of the relationship between the variables must be linear
- If the conditions are not met: use Spearman correlations
  - Examples: likert scale (ordinal data)

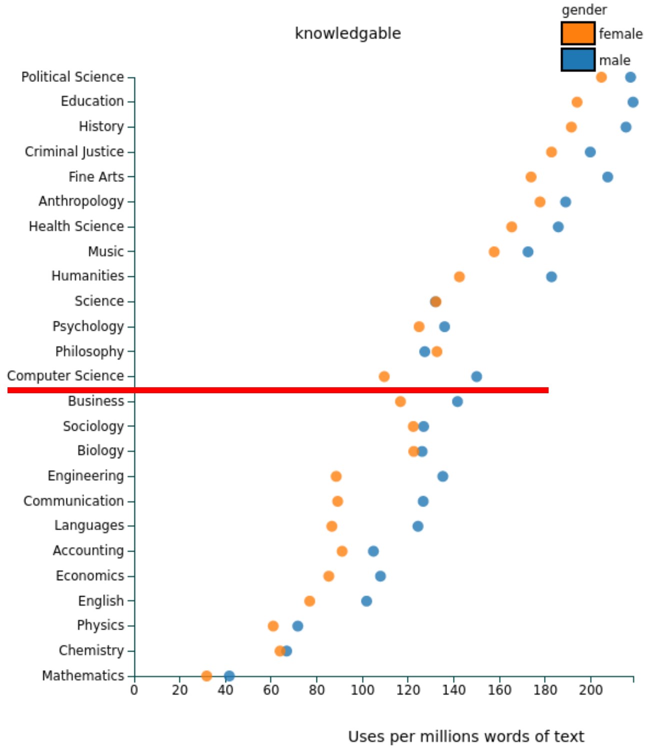
Biases and Diversity (endless)...

- Ratemyprofessors.com
- 14 million reviews
- A new tool allows those being rated (or anyone) to see the way students tend to use different words when rating male and female professors -- generally to the disadvantage of the latter.

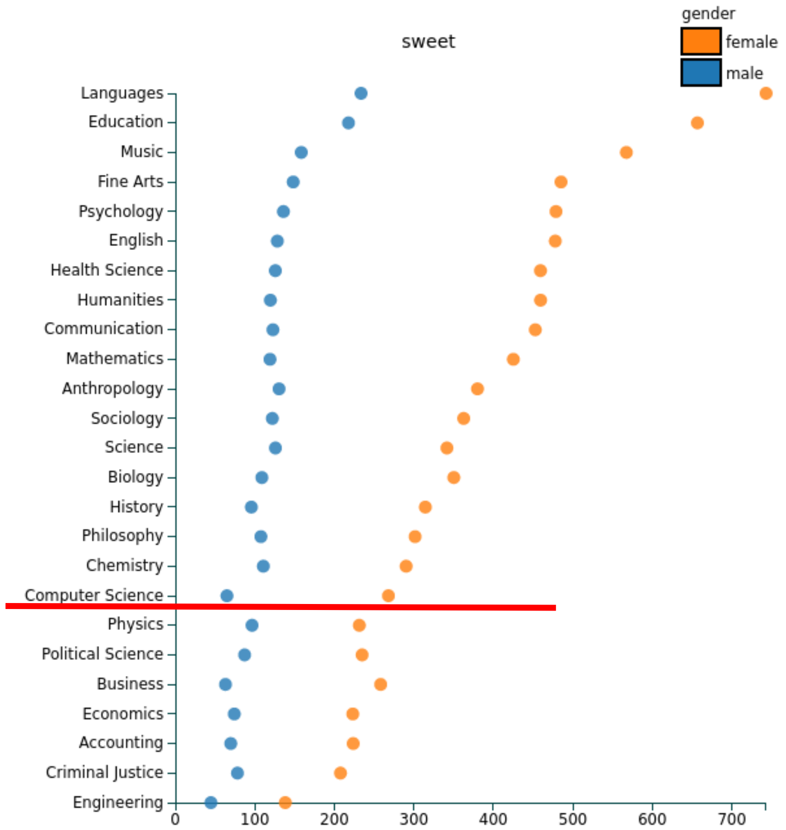
Biases and Diversity (endless)...







Biases and Diversity (endless)...



More on Biases and Diversity (endless)...

Can salience of gender identity impair math performance among 7-8 years old girls? The moderating role of task difficulty

Emmanuelle Neuville
University Blaise Pascal, Clermont-Ferrand, CNRS, France

Jean-Claude Croizet
University of Poitiers, France

Can the salience of gender identity affect the math performance of 7–8 year old girls? Third-grade girls and boys were required to solve arthmetical problems of varied difficulty. Prior to the test, one half of the participants had their gender identity activated. Results showed that activation of gender identity affected girls' performance but not boys. When their gender was activated as opposed to when it was not, girls solved more problems when the material was less difficult but underperformed on the difficult problems. Results are discussed with regard to the stereotype threat literature.

More on Biases and Diversity (endless)...

Gender, Confidence, Math: Why Aren't the Girls "Where the Boys Are?"

Caporrimo, Rosaria

Analyses were conducted to examine the relationship of standardized mathematics achievement scores, problem-solving strategies, self-report scores, and Confidence in Learning Mathematics survey scores among 122 eighth-grade students, 70 females and 52 males, representing all levels of mathematics achievement. Among the findings, no gender differences were evident on any of these scores; however, the Confidence scores functioned differently for the sexes. When consideration was focused upon average scores on the problem-solving strategies measure, males exhibited a direct relationship between routine problem scores and Confidence scores, whereas females showed an inverse relationship. (22 references) (JJK)