Producing wrong data without doing anything obviously wrong!
Evaluating innovations in computer systems
Evaluating innovations in computer systems

System

System + I

< → I is useful

> → I is not useful
Evaluating innovations in computer systems

Experimental setup

System

< → I  is useful

> → I  is not useful

But what if the data are biased? Wrong conclusions!
Methodology

• SPEC CPU 2006 C programs
• Intel Core 2 (2.4GHz)
  ‒ Linux 2.6.25
  ‒ gcc 4.2
  ‒ Papi 3.5.1 / perfmon 2.8
• Best Practices
  ‒ Unloaded machine
  ‒ Multiple runs
  ‒ Confidence intervals
Example of bias in 400.perlbench

System    = gcc -O2
System + I = gcc -O3
Example of bias in 400.perlbench

System    = gcc -O2
System + I = gcc -O3

Amer:
speedup = 1.18 ± 0.0002

Conclusion: O3 is good
Example of bias in 400.perlbench

System = gcc -O2
System + I = gcc -O3

Amer:
speedup = 1.18 ± 0.0002
Conclusion: O3 is good

Todd:
speedup = 0.84 ± 0.0002
Conclusion: O3 is bad
Example of bias in 400.perlbench

System = gcc -O2
System + l = gcc -O3

Amer:
speedup = 1.18 ± 0.0002

Conclusion: O3 is good

Todd:
speedup = 0.84 ± 0.0002

Conclusion: O3 is bad

Why does this happen?
Differences in our experimental setup
Differences in our experimental setup

Amer:
HOME=/home/amerdiwan/

Todd:
HOME=/home/toddmytkowicz
Differences in our experimental setup

Amer:
HOME=/home/amerdiwan/

Todd:
HOME=/home/toddmytkowicz
Differences in our experimental setup

Amer:
HOME=/home/amerdiwan/

Todd:
HOME=/home/toddmytkowicz

Could this be the source of bias?
Bias from size of UNIX environment

400.perlbench

cycles(O2) / cycles(O3)

bytes added to empty environment
Bias from size of UNIX environment

400.perlbench

cycles(O2) / cycles(O3)

bytes added to empty environment

perlbench
Bias from size of UNIX environment

Environment size from 0-4096 bytes
Bias from size of UNIX environment

Environment size from 0-4096 bytes

The setting of irrelevant environment variables can lead to biased conclusions
By using an empty UNIX environment, Amer and I now agree.

But for perlbench, we still differ...
Other differences in our experimental setup

Amer:
$> \text{ld A.o B.o}$

Todd:
$> \text{ld B.o A.o}$
Other differences in our experimental setup

Amer:
$\text{ld } \text{A.o B.o}$

Todd:
$\text{ld } \text{B.o A.o}$
Other differences in our experimental setup

Amer:
$> \text{ld A.o B.o}$

Todd:
$> \text{ld B.o A.o}$

Could this be the source of bias?
Bias from linking order

400.perlbench

cycles(O2) / cycles(O3)

linking order
Bias from linking order

32 randomly generated linking orders

cycles (O2) / cycles (O3)

gcc libquantum perlbench bzip2 h264ref mcf gobmk hmer sjeng sphinx milc ibm

+ default × alphabetical
Bias from linking order

32 randomly generated linking orders

Order of .o files can lead to contradictory conclusions
Are we just showing you corner cases?
Are we just showing you corner cases?

No: Bias occurs on multiple microprocessors

Core 2
Are we just showing you corner cases?

No: Bias occurs on multiple microprocessors
Are we just showing you corner cases?

No: Bias occurs on multiple compilers

gcc 4.2
Are we just showing you corner cases?

No: Bias occurs on multiple compilers

gcc 4.2

icc 10.1
Are we just showing you corner cases?

No: Bias occurs in simulation
Are we just showing you corner cases?

No: Bias occurs in simulation
Can we **easily** avoid bias?
Can we easily avoid bias?

No: Bias is not predictable

![Graph showing cycles(O2) / cycles(O3) vs. bytes added to empty environment for 400.perlbench]
Can we easily avoid bias?

No: Averaging across benchmarks does not cancel it out
Where does bias come from?
Where does bias come from?

Interactions with hardware buffers

O2

Page N

Page N + 1
Where does bias come from?

Interactions with hardware buffers

- Page N
- Page N + 1

Dead Code
Where does bias come from?

Interactions with hardware buffers
Where does bias come from?

Interactions with hardware buffers

Page N

Page N + 1

Hot code
Where does bias come from?

Interactions with hardware buffers

O2

O3

Page N

Page N + 1

O3 better than O2
Where does bias come from?

Interactions with hardware buffers

Page N
O2
Page N + 1
O3
O2 better than O3
O3 better than O2

Where does bias come from?
Where does bias come from?

**Interactions with hardware buffers**

Hardware buffers abound in modern systems
What can we do about bias?
What can we do about bias?

Randomized Trials

Program

Experimental Setup Generator

P1

P2

PN
What can we do about bias?

Randomized Trials

Program

Experimental Setup Generator

Link Order 1/Env 1

Link Order 2/Env 2

... (ellipsis)

Link Order N/Env N
What can we do about bias?

Randomized Trials

Effectiveness depends upon representativeness of setups
What can we do about bias?

Randomized Trials

400.perlbench

![Histogram showing frequency of cycles for O2 and O3 with x-axis representing cycles x 1e+09 and y-axis representing frequency.]

- **O2**
- **O3**
What can we do about bias?

Causality Analysis

(1) Analyze data to arrive at a hypothesis
What can we do about bias?

**Causality Analysis**

(1) Analyze data to arrive at a hypothesis

(2) Perform intervention to test hypothesis
What can we do about bias?

Causality Analysis

(1) Analyze data to arrive at a hypothesis
(2) Perform intervention to test hypothesis
(3) Validate effects of hypothesis
What can we do about bias?

Causality Analysis

(1) Analyze data to arrive at a hypothesis
(2) Perform intervention to test hypothesis
(3) Validate effects of hypothesis
What can we do about bias?

Causality Analysis

(1) Analyze data to arrive at a hypothesis
(2) Perform intervention to test hypothesis
(3) Validate effects of hypothesis

- Changing environment size
- Changing stack address
- Changing performance
What can we do about bias?

**Causality Analysis**

1. Analyze data to arrive at a hypothesis
2. Perform intervention to test hypothesis
3. Validate effects of hypothesis

Popular in sciences but difficult and manual
Related Work

- “Correctness” via microkernels
  - [Korn et al: IPCCC '01], [Maxell et al: LASCIC '02] and [Moore: ICCS '02]

- Other sources of bias
  - Heap Size for GC [Blackburn et al: OOPSLA '06]
  - Variability in multi-threaded simulation [Alameldeen and Wood: HPCA '03]
  - Input Shaking [Tsafrir et al: MASCOTS '07]

- Statistical Rigor in performance evaluations
  - [Georges et al: OOPSLA '07]
Summary

Would you believe a U.S. Census conducted in one small town?
Summary

Would you believe a U.S. Census conducted in **one** small town?

Would you believe a systems experiment conducted in **one** experimental setup?
Summary

Would you believe a U.S. Census conducted in one small town?

Would you believe a systems experiment conducted in one experimental setup?

We show bias is pervasive, unpredictable and significant
Where does bias come from?

Microkernel example

```c
static int i, j, k, inc;
int main() {
  int g;
i = j = k = 0;
inc = 1;
for (g = 0; g < 65536; g++) {
i += inc;
j += inc;
k += inc;
}
return (0);
}
```

Where does bias come from?
Where does bias come from?

**Microkernel example**

```c
static int i, j, k, inc;
int main() {
  int g;
  i = j = k = 0;
  inc = 1;
  for (g = 0; g < 65536; g++) {
    i += inc;
    j += inc;
    k += inc;
  }
  return (0);
}
```

Where does bias come from stack allocated variable
Where does bias come from?

**Microkernel example**

```c
static int i, j, k, inc;
int main() {
  int g;
i = j = k = 0;
inc = 1;
for (g = 0; g < 65536; g++) {
  i += inc;
  j += inc;
  k += inc;
}
return (0);
}
```

**global variable**
Where does bias come from?

Bias comes from sensitivity of program behavior to an experimental setup.

Microkernel example

ICC 10.1

cycles

shift in stack start address

shift in data start address

500000

600000

700000

800000

900000

1000000

1100000