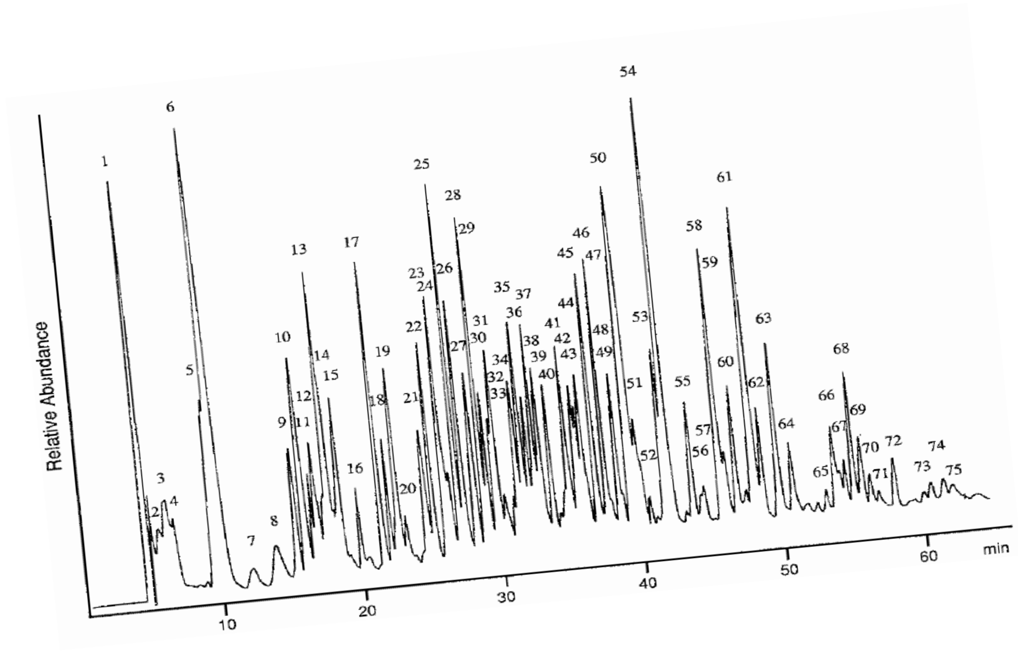
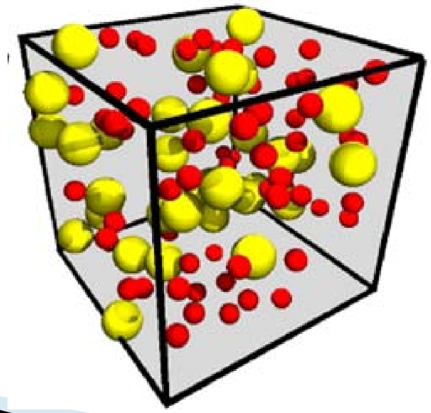
Reliable Data Processing Enabled by Program Analysis

Xiangyu Zhang

December 22, 2013 @ ISHCS, PKU

- Data processing is becoming increasing important.





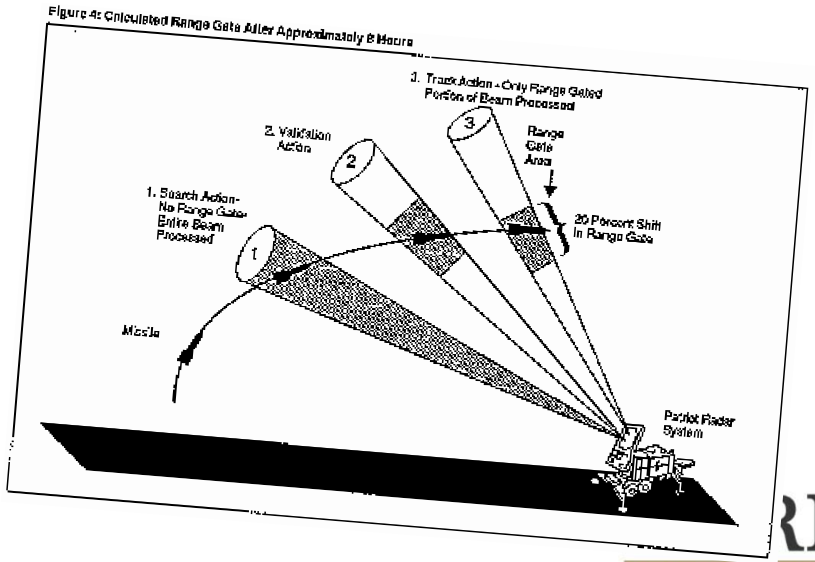
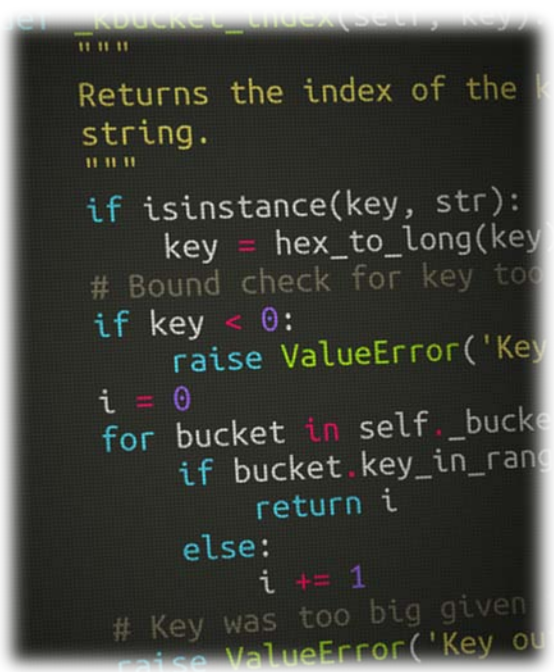


Figure 4: Calculated Range Gate After Approximately 8 Hours
3. Track Action - Only Range Gated Portion of Beam Processed
2. Validation Action
1. Search Action - No Range Gate: Entire Beam Processed
Range Gate Area
20 Percent Shift In Range Gate
Missile
Patriot Radar System

Errors in Data Processing

external errors



```
def _kbucket_index(self, key):
    """
    Returns the index of the k
    string.
    """
    if isinstance(key, str):
        key = hex_to_long(key)
    # Bound check for key too
    if key < 0:
        raise ValueError('Key
    i = 0
    for bucket in self._bucke
        if bucket.key_in_rang
            return i
        else:
            i += 1
    # Key was too big given
    raise ValueError('Key ou
```

internal errors

External Errors

```
float x, z;
x = input();
z = f(x);

if (z > 0.5)
    printf ("hit");
else
    printf ("miss");
```

x = 100.0;

1. How would the program output change if input *x* is uncertain?

e.g. $x \in [50.0, 150.0]$

Internal Errors

```
float x, z;
x = input();
z = f(x);

if (z > 0.5)
    printf ("hit");
else
    printf ("miss");
```

x = 100.0;

2. Are the computed results reliable?

miss

Errors in Data Processing

- External errors
  - Also known as data uncertainty problem
  - Existing techniques
    - query-based uncertain data processing e.g. [R. Jampani, SIGMOD 2008], [S. Singh, ICDE 2008] and etc.
    - Interval analysis
    - Automatic differentiation
- Internal errors
  - Existing work include interval analysis, using high precision
- Existing solutions are hardly applicable or usable -- too expensive, too many false positives
- Errors may get propagated and magnified, leading to unreliable output.
  - We call it the *instability* problem.

PURDUE UNIVERSITY

Outline

- Overview
- External Errors
  - White-box sampling (OOPSLA 2012)
- Internal Errors
  - On-the-fly detection of instability problems (OOPSLA 2013)

PURDUE UNIVERSITY

White Box Sampling --External Errors

```
float x, z;
x = input();
z = f(x);

if (z > 0.5)
    printf (“hit”);
else
    printf (“miss”);
```

x = 100.0;

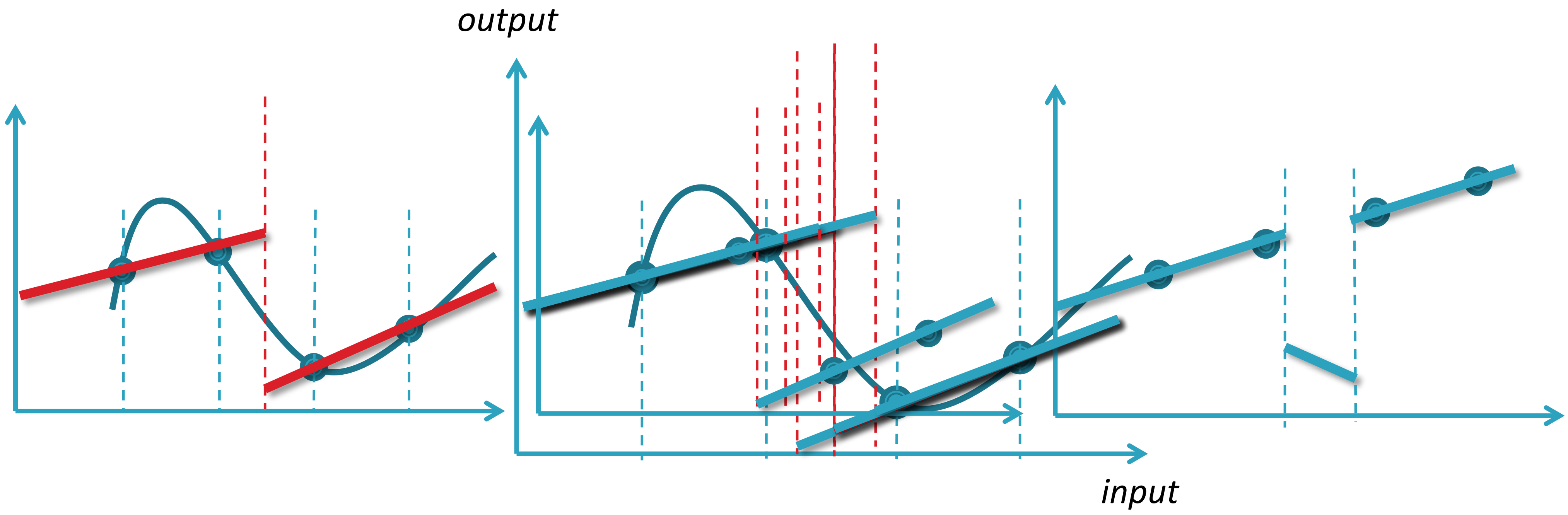
1. How would the program output change if input *x* is uncertain?

e.g. $x \in [50.0, 150.0]$

PURDUE UNIVERSITY

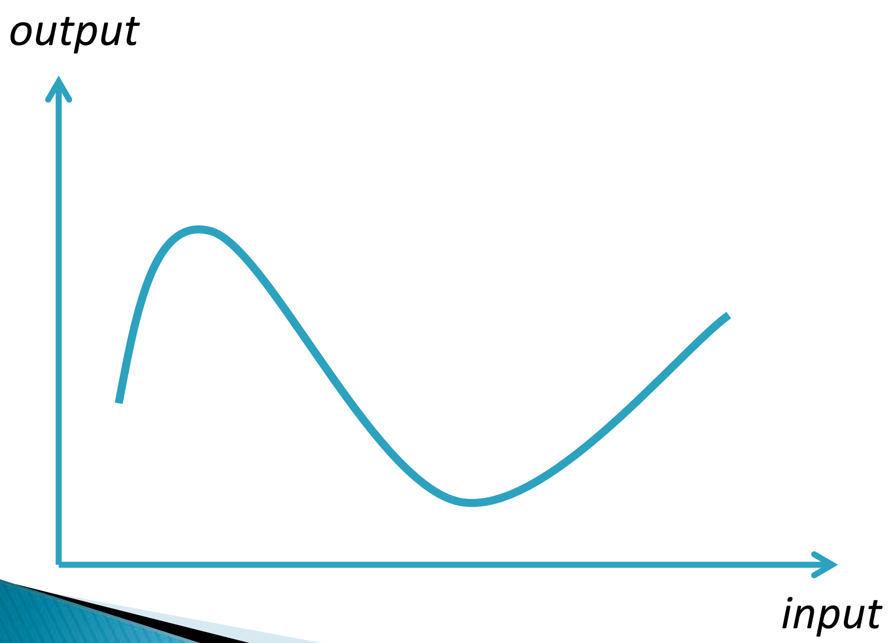
Monte Carlo (MC) methods

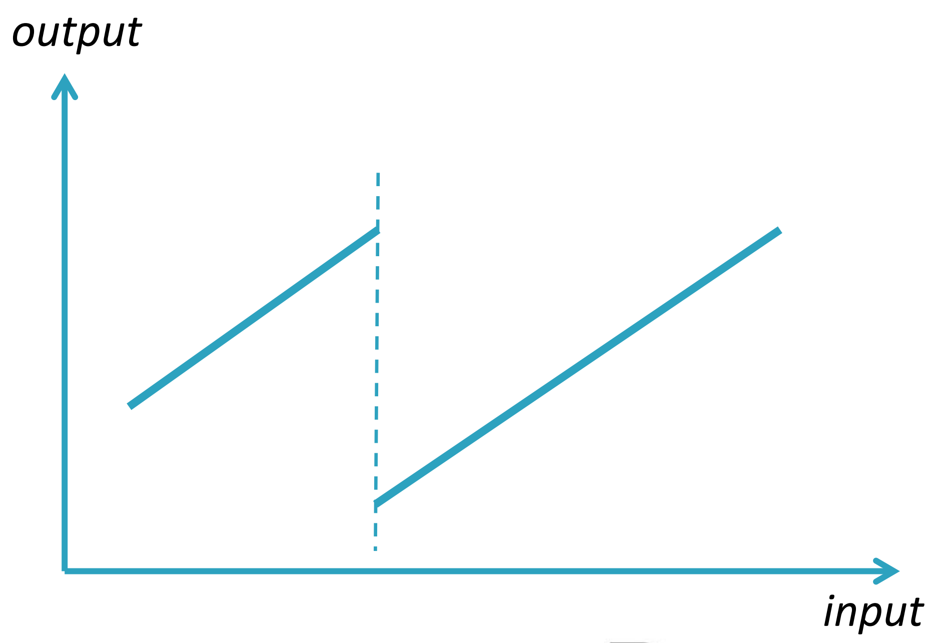
- Sampling-based *Monte Carlo method* is effective, yet imprecise.



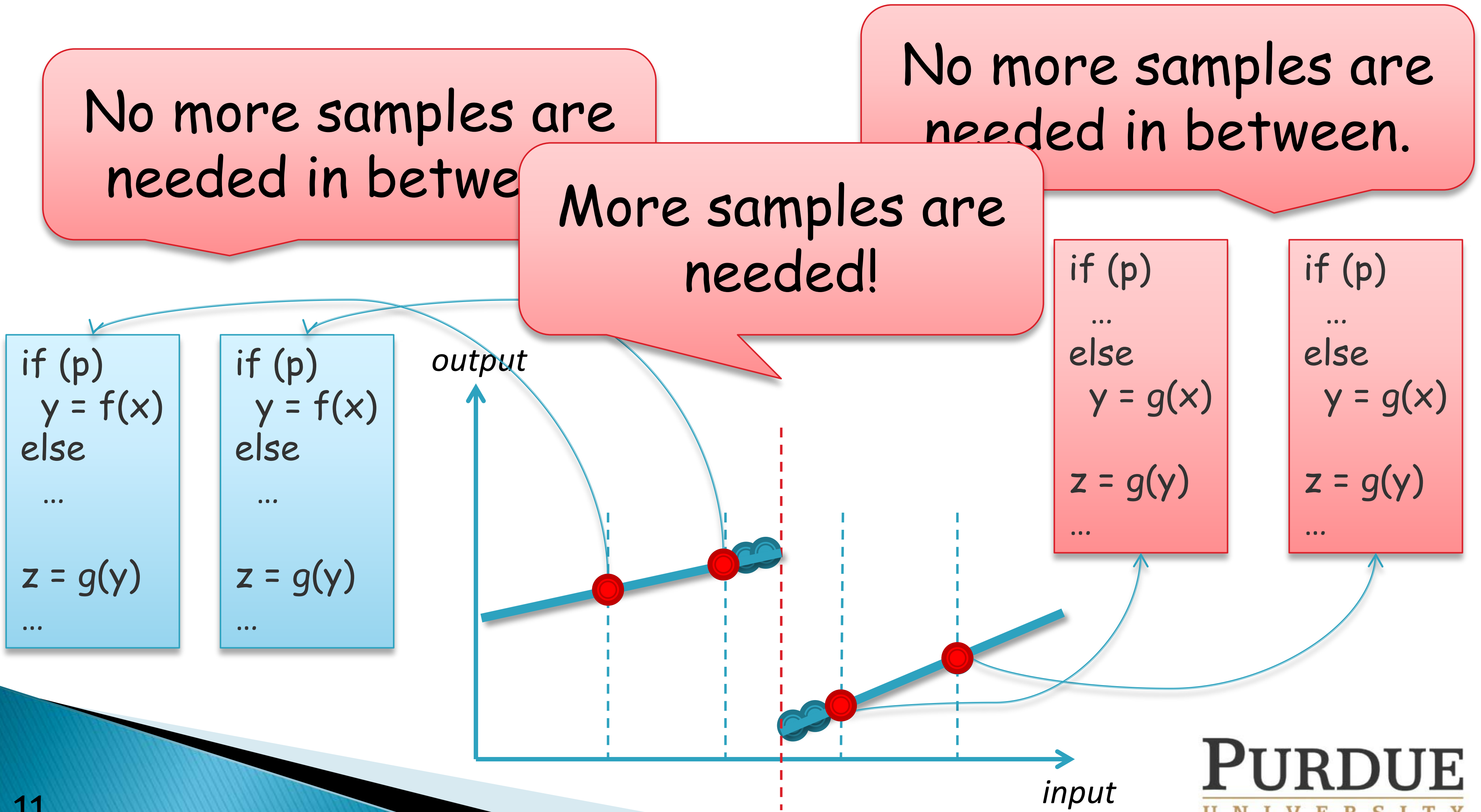
Our Idea

- MC sampling guided by program analysis (few samples and sampling at the critical places)
  - Use dynamic analysis to predict output continuity
  - Perform demand driven sampling based on continuity





White Box Sampling - Intuition


No more samples are needed in betwe
More samples are needed!
No more samples are needed in between.
output
input

```
if (p)
  y = f(x)
else
  …

z = g(y)
…
```

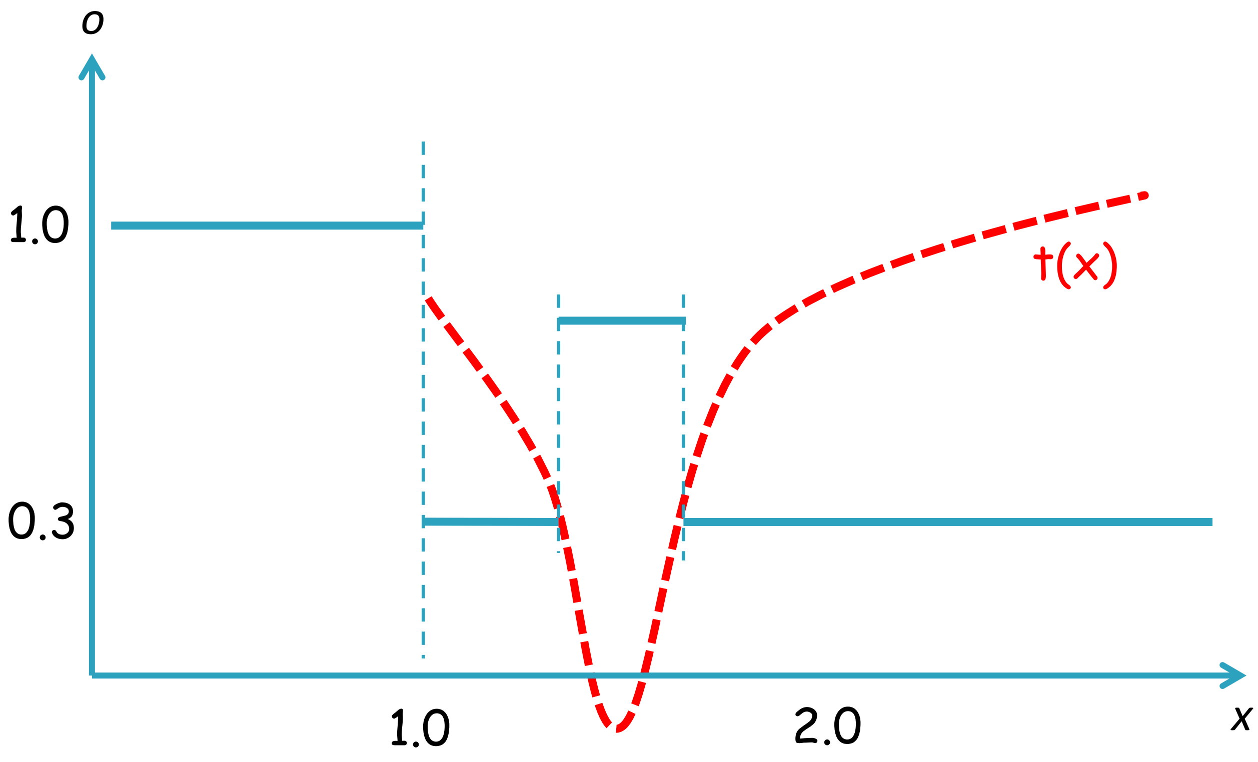
```
if (p)
  y = f(x)
else
  …

z = g(y)
…
```

```
if (p)
  …
else
  y = g(x)

z = g(y)
…
```

```
if (p)
  …
else
  y = g(x)

z = g(y)
…
```

A Running Example

```
x = sample(1.5);
y = (int) x;
if (x < 1.0)
  o = 1 + y;
else
  if (t(x) > 0.3)
    o = 0.3;
  else
    o = 0.75;
```


o
1.0
0.3
t(x)
1.0
2.0
x