Web Search

Rolf Michelsen

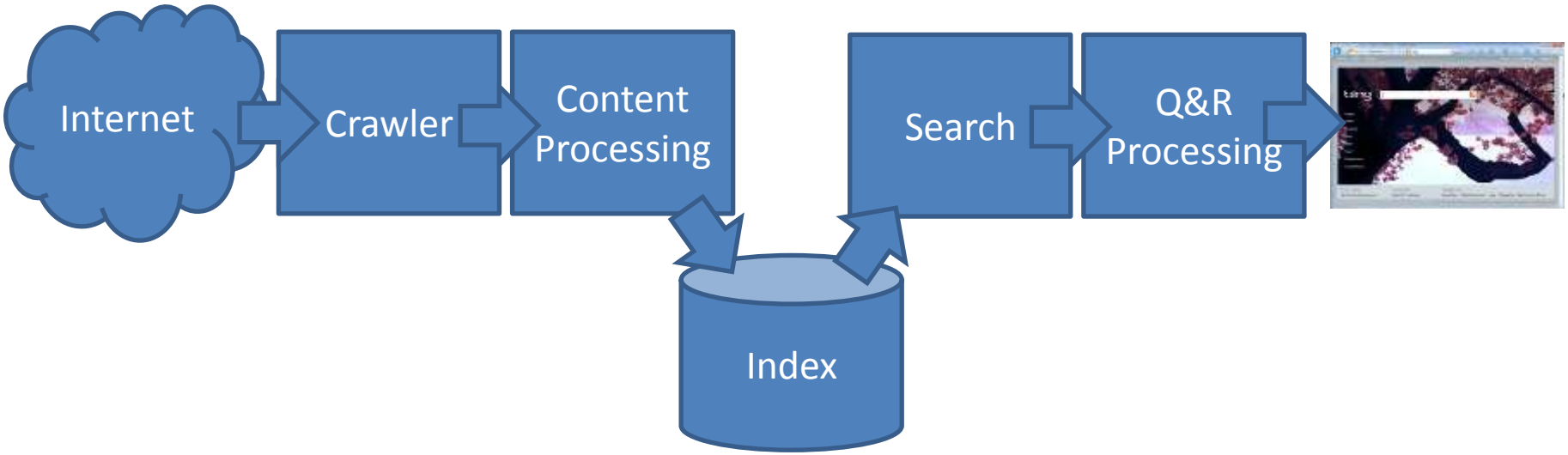
Microsoft | Development Center Norway

Background

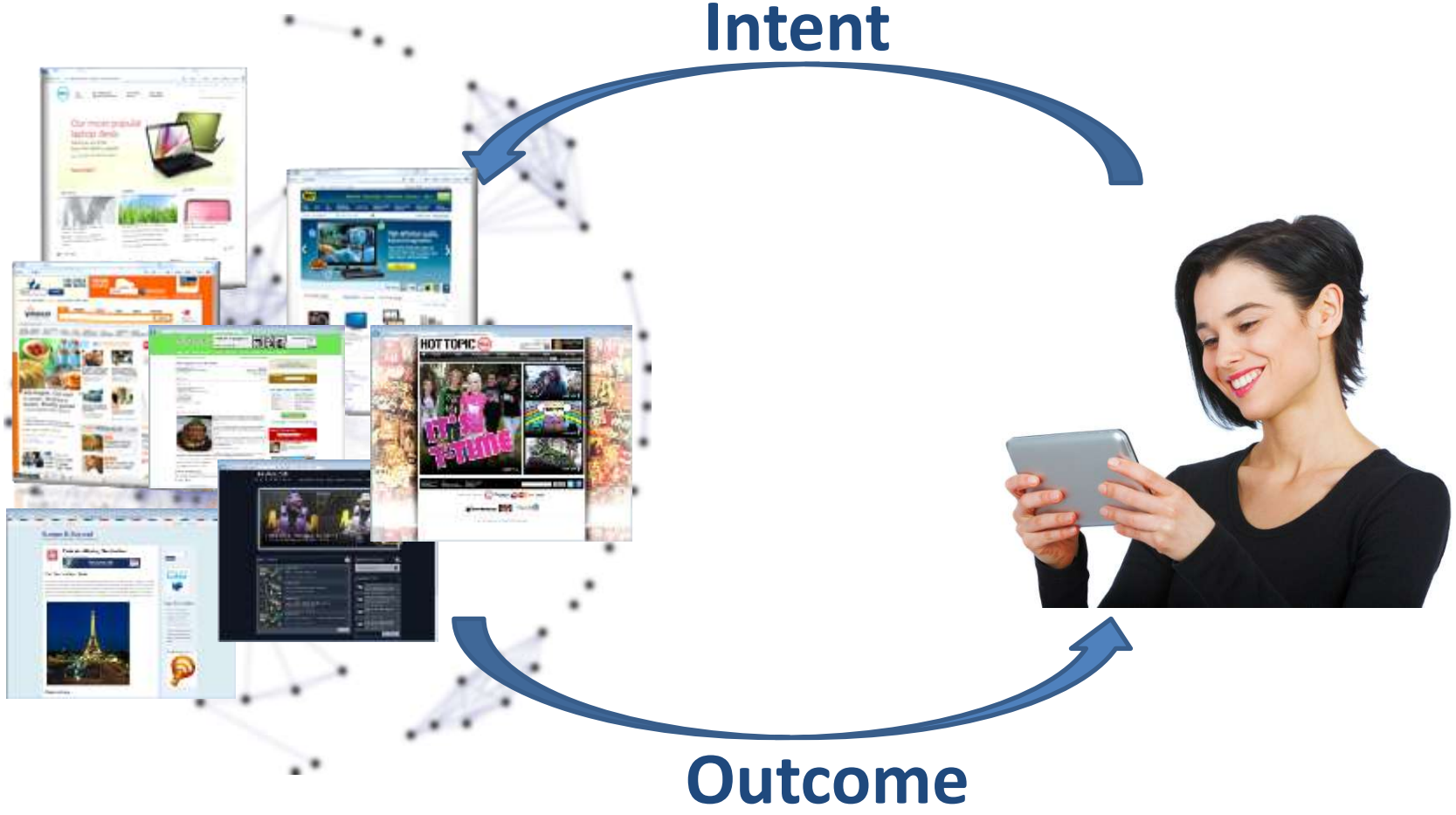

fast
YAHOO!
Microsoft
1997
1999
2003
2008
2010
alltheweb
fast ESP™
bing
Microsoft
SharePoint 2010

INTRODUCTION TO WEB SEARCH

Web Search — System



Web Search — Application



Web Search — Business Model

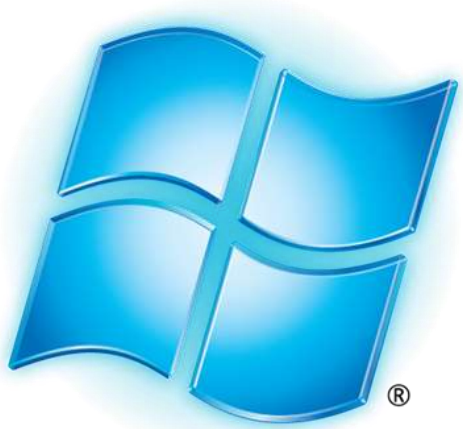
Bing Users

Publishers

Advertisers

INDUSTRIAL SOFTWARE ENGINEERING

Opportunities

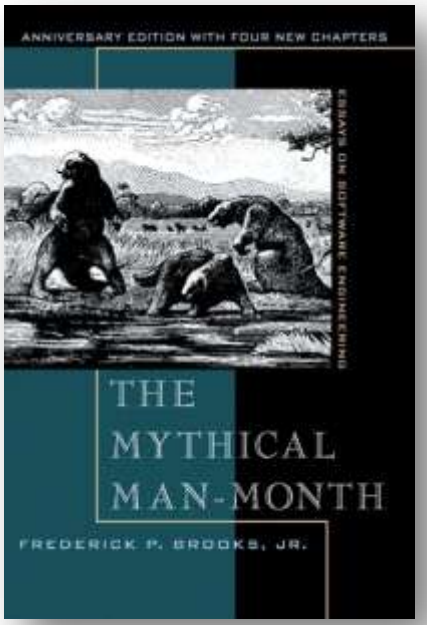


Windows Azure™

Industrial software engineering

Creating a *platform product* costs an order of magnitude more than creating a *program*.

— Frederick P. Brooks, jr.



QUALITY

What is quality?

What is quality?

Vs.

Measuring quality

«When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.»

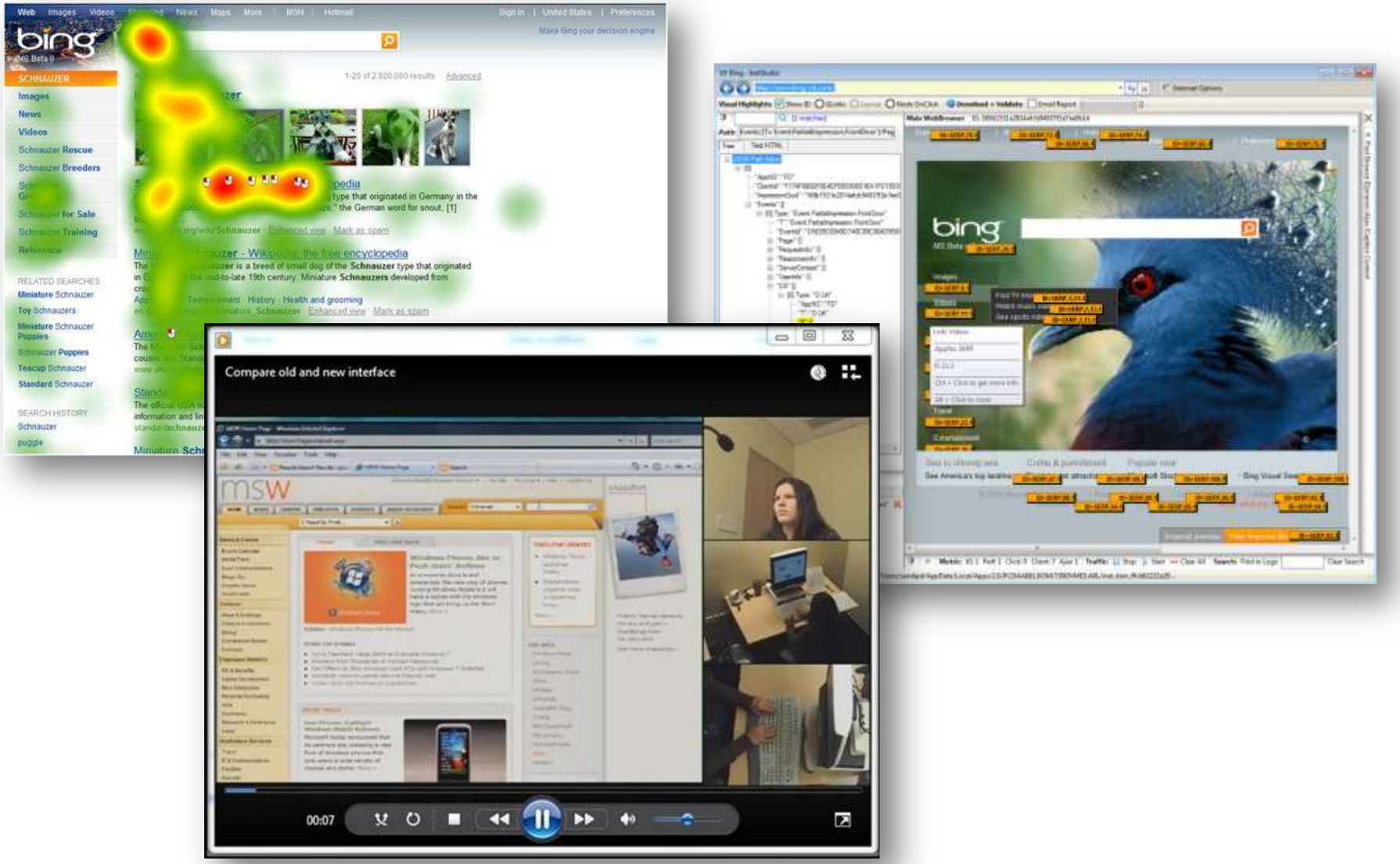
— Lord Kelvin (1824 – 1907)

Metrics

Hard metrics

Soft metrics

User Studies

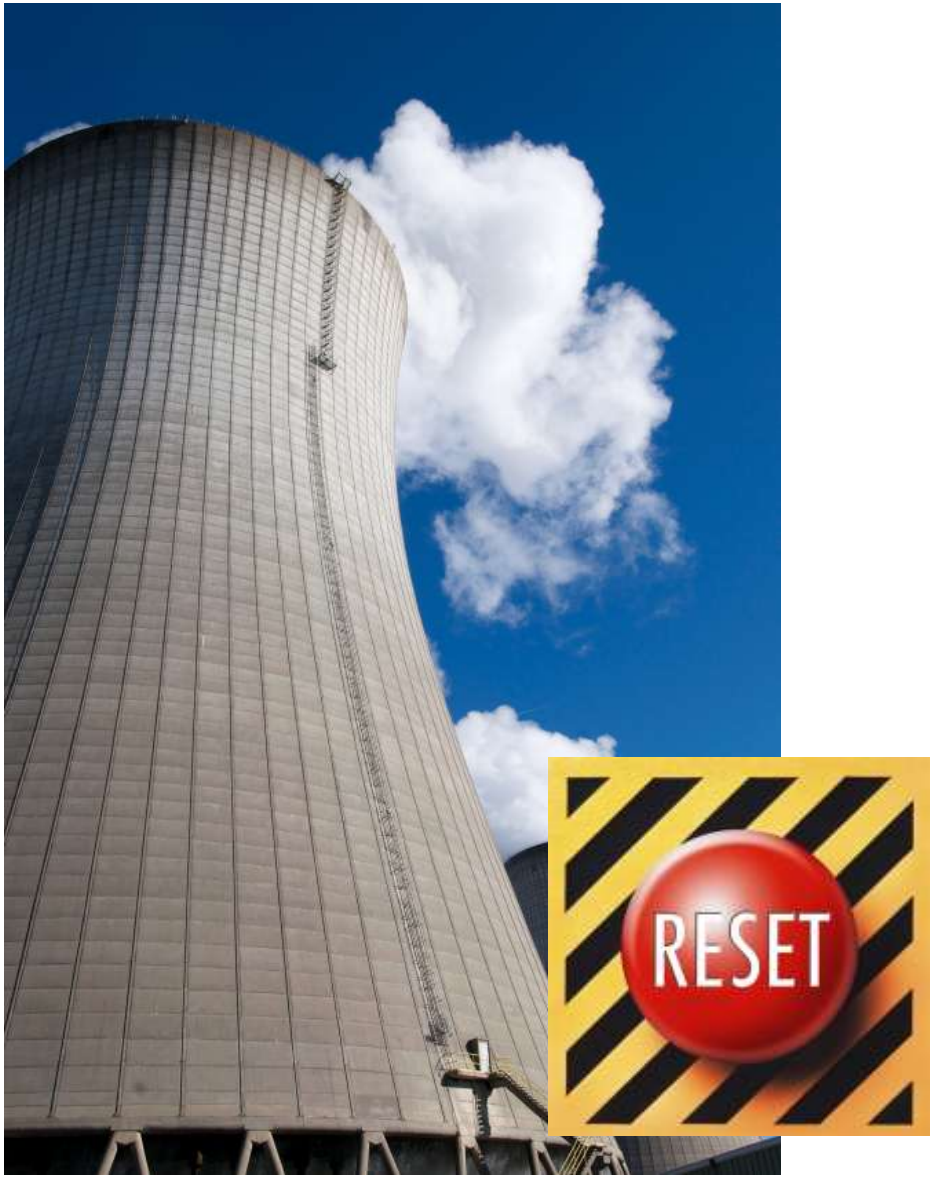


The Future

- Task completion
- Intent detection
- Personalization

ENVIRONMENT

Web standards


RESET

Robots.txt

User-agent: *
Disallow: /Private/

Please do not index this site.


<MagicTag>


SPAM
Classic
NET WT
340g
Hormel
Foods

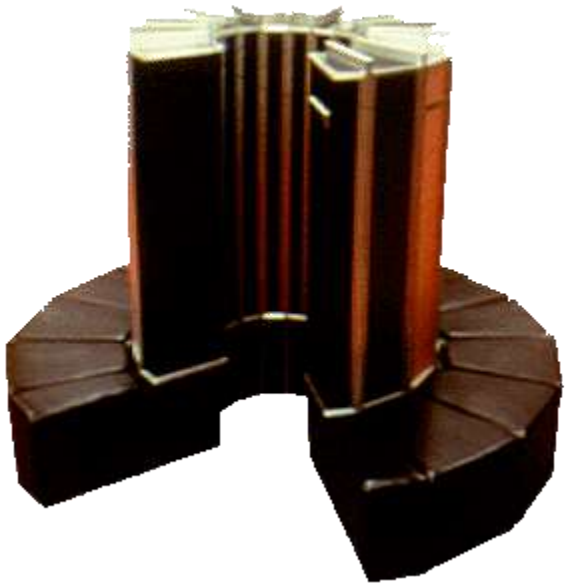

Web 2.0
Aggregators
Folksonomy
Wikis
Blogs
User Centered
Joy of Use
Participation
Six Degrees
Usability
Pagerank
XFN
Widgets
Recommendation
Social Software
FOAF
Browser
Simplicity
Sharing
Collaboration
Perpetual Beta
Videocasting
Podcasting
AJAX
Audio
IM
Video
Design
Convergence
CSS
Pay Per Click
UMTS
Mobility
Atom
XHTML
SVG
Ruby on Rails
VC
Trust
Affiliation
OpenAPIs
RSS
Semantic
Web Standards
SEO
Economy
OpenID
Remixability
REST
Standardization
The Long Tail
DataDriven
Accessibility
XML
Microformats
Syndication
Modularity
SOAP

SCALE

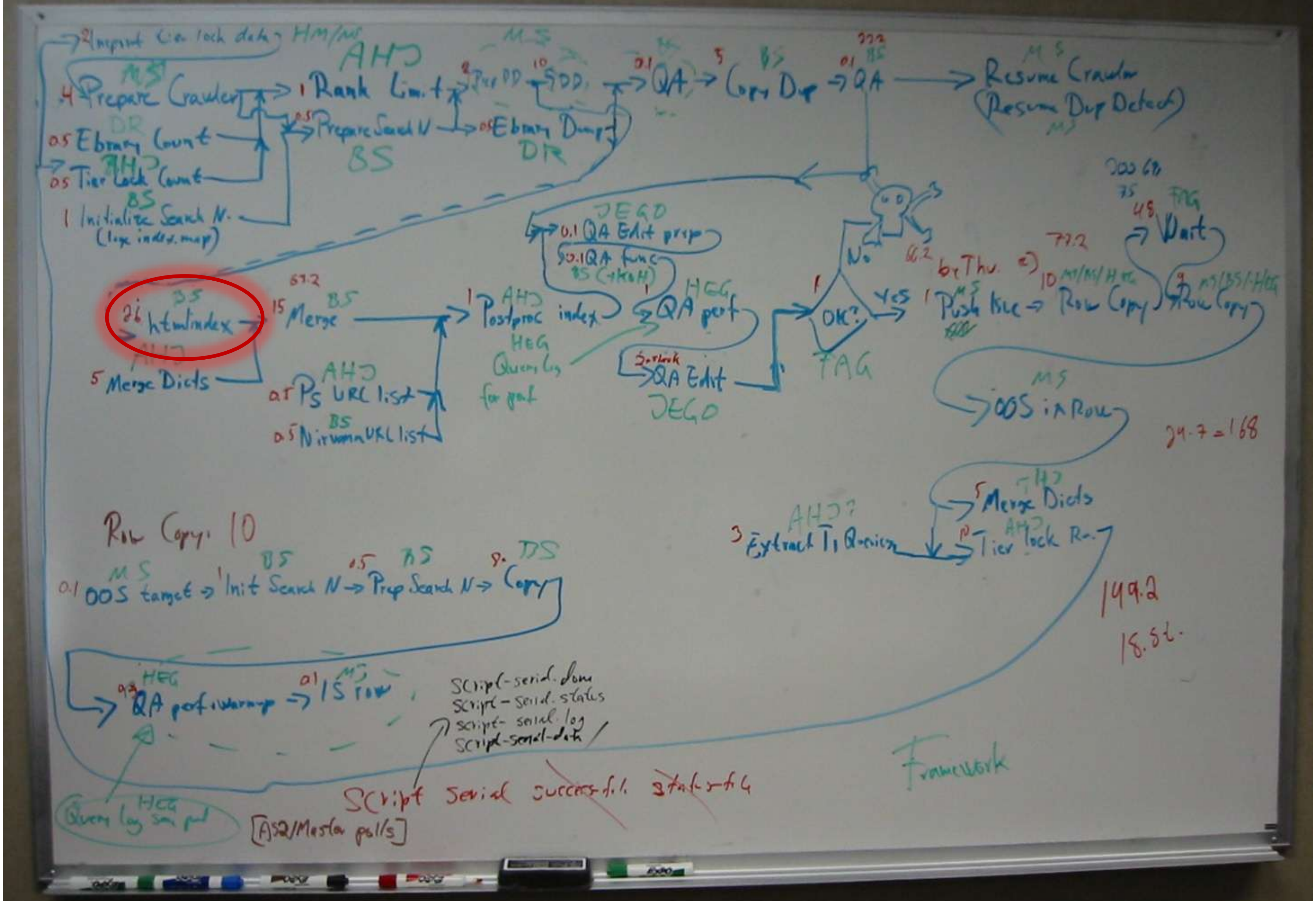
Engineering Complex Systems

«Anyone can build a fast CPU. The trick is to build a fast system.»

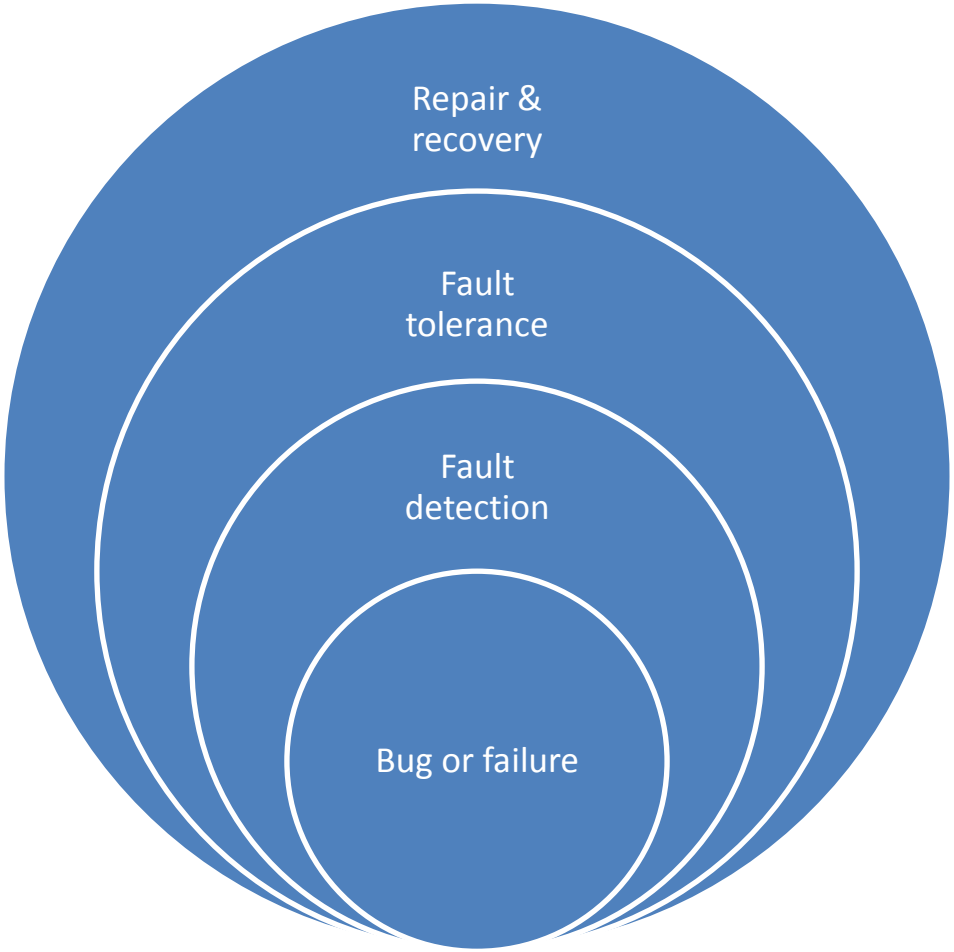
— Seymour Cray (1926-1996)



Large Scale Systems


AHJ
Prepare Crawler
Rank Limit
QA
Copy Dup
QA
Resume Crawler
(Resume Dup Detect)
Ebrary Count
Prepare Search N
Ebrary Dump
Tier Lock Count
Initialize Search N.
(logx index.map)
BS
DR
JEGO
QA Edit prep
QA func
htmlindex
Merge
Postproc index
QA perf
OK?
No
Yes
Push Live
Row Copy
Wait
Merge Dicts
PS URL list
Nirvana URL list
HEG
Query log
for perf
QA Edit
FAG
OOS in Row
24 · 7 = 168
Merge Dicts
Extract T1 Queries
Tier lock Ro.
149.2
18.5 d.
Row Copy: 10
OOS target
Init Search N
Prep Search N
Copy
QA perf warmup
IS row
script-serial.done
script-serial.status
script-serial.log
script-serial-data
Query log
[ASQ/Master polls]
Script Serial successful status-file
Framework

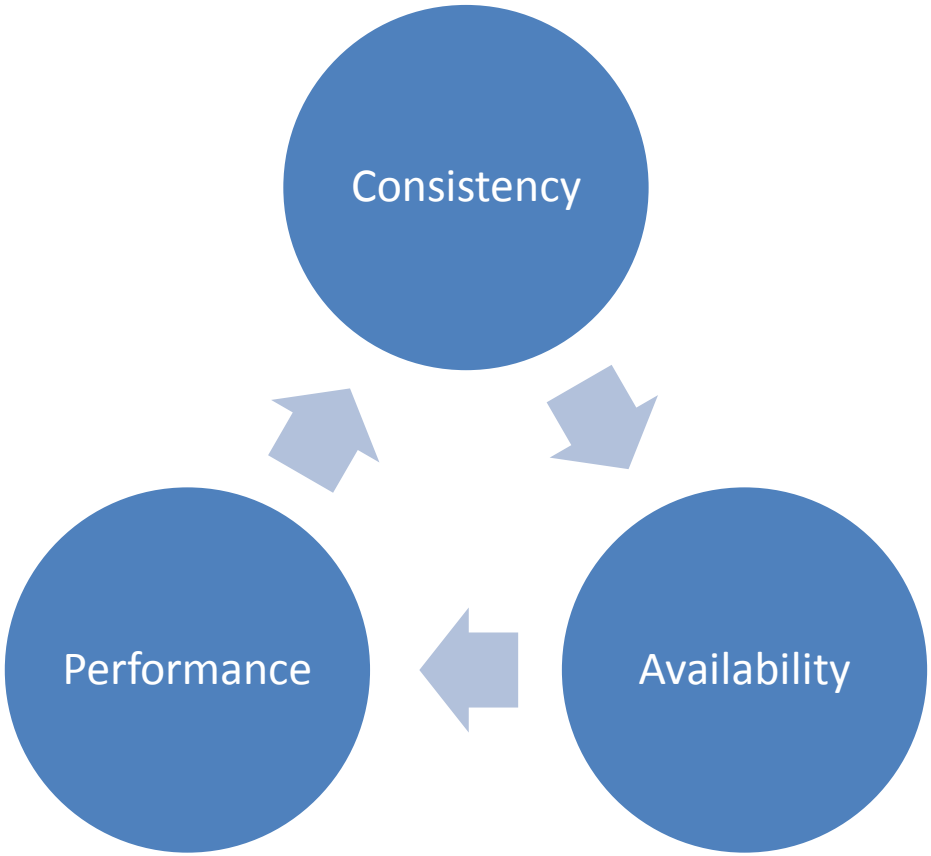
Engineering for failure



Operations

TRADE-OFFS

The CAP Theorem



Technical Debt

CONCLUSION


Plan A
Plan B


Software Engineering
Programming
Algorithms & data structures
Distributed computing
Networking
Math & statistics
Quality assurance
Project management
Operations
Domain expertize
Technology & Tools
User interaction & design
Business development

Microsoft®