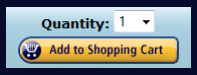
The Analytics Revolution, Apr 2010, SDForum

Online Controlled Experiments:
Listening to the Customers, not to the HiPPO

Ronny Kohavi, General Manager
Experimentation Platform, Microsoft
ronnyk@microsoft.com

2

Amazon Shopping Cart Recommendations



- Add an item to your shopping cart at a website
  - Most sites show the cart
- At Amazon, Greg Linden had the idea of showing recommendations based on cart items
- Evaluation
  - Pro: cross-sell more items (increase average basket size)
  - Con: distract people from checking out (reduce conversion)
- HiPPO (Highest Paid Person's Opinion) was: stop the project
- Simple experiment was run, wildly successful, and the rest is history

From Greg Linden's Blog: http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

3

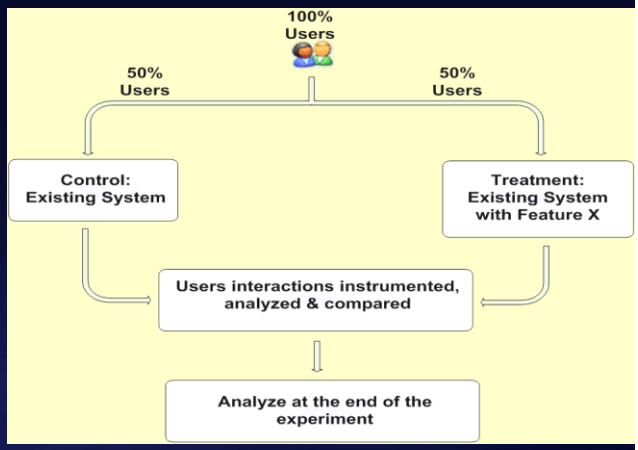
Agenda

- Controlled Experiments in one slide
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Two key messages to remember
  - It is hard to assess the value of ideas.
    Get the data by experimenting because data trumps intuition
  - Make sure the org agrees **what** you are optimizing and evolve your culture towards data-driven decisions
- Papers, examples, all the statistics, pros/cons at http://exp-platform.com (reprints of key paper available here)

4

Controlled Experiments in One Slide

- Concept is trivial
  - Randomly split traffic between two (or more) versions
    - A (Control)
    - B (Treatment)
  - Collect metrics of interest
  - Analyze



- Must run statistical tests to confirm differences are not due to chance
- Best scientific way to prove causality, i.e., the changes in metrics are caused by changes introduced in the treatment(s)

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

5

Examples

- Three experiments that ran at Microsoft
- All had enough users for statistical validity
- Game: see how many you get right
  - Everyone please stand up
  - Three choices are:
    - A wins (the difference is statistically significant)
    - A and B are approximately the same (no stat sig diff)
    - B wins
  - If you guess randomly
    - 1/3 left standing after first question
    - 1/9 after the second question

6

MSN Real Estate

- "Find a house" widget variations
- Overall Evaluation Criterion(OEC): Revenue to Microsoft generated every time a user clicks search/find button

A

B

- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if you think they're about the same

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

7

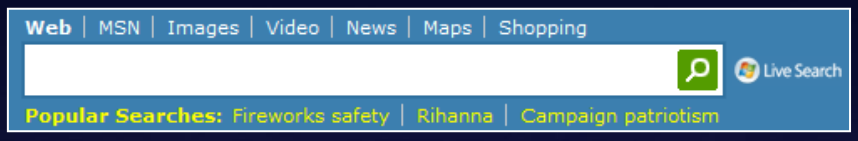
MSN Real Estate

- If you did not raise a hand, please sit down
- If you raised your left hand, please sit down
- A was 8.5% better
- Since this is the #1 monetization, it effectively raised revenues significantly
- Actual experiment had 6 variants.
  If you're going to experiment, try more variants, especially if they're easy to implement

MSN Home Page Search Box

OEC: Clickthrough rate for Search box and popular searches

A



B

Differences: A has taller search box (overall size is the same), has magnifying glass icon, "popular searches"
B has big search button

- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if they are the about the same

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

9

Search Box

- If you raised any hand, please sit down
- Insight

  Stop debating, it's easier to get the data

10

MSN US Home Page: Search Box

- A later test showed that changing the magnifying glass to an actionable word (search, go, explore) was highly beneficial.
- This:

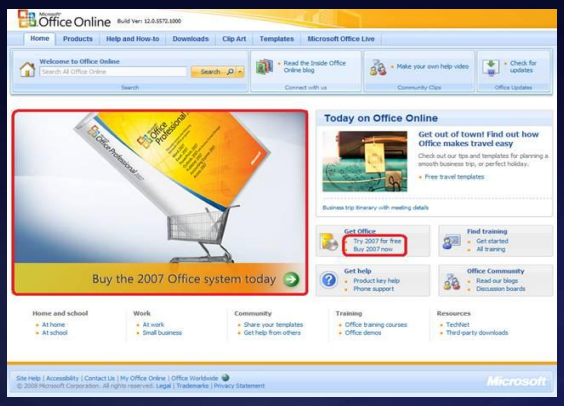
is better than

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

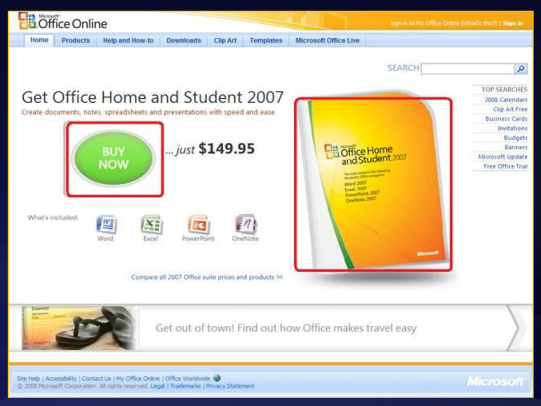
Office Online

OEC: Clicks on revenue generating links (red below)

A



B



- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if they are the about the same

12

Office Online

- If you did not raise a hand, please sit down
- If you raised your left hand, please sit down
- B was 64% worse

- What % of the audience is still standing?
- Humbling!

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

13

Twyman's Law

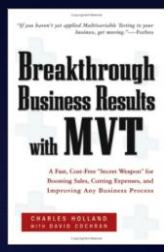
Any statistic that appears interesting is almost certainly a mistake

- If something is "amazing," find the flaw!
- Examples
  - If you have a mandatory birth date field and people think it's unnecessary, you'll find lots of 11/11/11 or 01/01/01
  - If you have an optional drop down, do not default to the first alphabetical entry, or you'll have lots jobs = Astronaut
- The previous Office example assumes click maps to revenue. Seemed reasonable, but when the results look so extreme, find the flaw (conversion rate is not the same; see why?)

14

Hard to Assess the Value of Ideas: Data Trumps Intuition

- At Amazon, half of the experiments failed to show improvement
- QualPro tested 150,000 ideas over 22 years
  - 75 percent of important business decisions and business improvement ideas either have no impact on performance or actually hurt performance…
- Based on experiments with ExP at Microsoft
  - 1/3 of ideas were positive ideas and statistically significant
  - 1/3 of ideas were flat: no statistically significant difference
  - 1/3 of ideas were negative and statistically significant
- Our intuition is poor: 2/3rd of ideas do not improve the metric(s) they were designed to improve. Humbling!



© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Key Lessons

- Avoid the temptation to try and build optimal features through extensive planning without early testing of ideas
- Experiment often
  - *To have a great idea, have a lot of them* -- Thomas Edison
  - *If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster* -- Mike Moran, Do it Wrong Quickly
- Try radical ideas. You may be surprised
  - Doubly true if it's cheap to implement (e.g., shopping cart recommendations)
  - *If you're not prepared to be wrong, you'll never come up with anything original* – Sir Ken Robinson, TED 2006

16

The OEC

- If you remember one thing from this talk, remember this point
- OEC = Overall Evaluation Criterion
  - Agree early on what you are optimizing
  - Getting agreement on the OEC in the org is a huge step forward
  - Suggestion: optimize for customer lifetime value, not immediate short-term revenue
  - Criterion could be weighted sum of factors, such as
    - Time on site (per time period, say week or month)
    - Visit frequency
  - Report many other metrics for diagnostics, i.e., to understand the why the OEC changed and raise new hypotheses

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

17

Agenda

- Controlled Experiments in one slide
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Two key messages to remember
  - It is hard to assess the value of ideas. Get the data by experimenting because data trumps intuition
  - Make sure the org agrees **what** you are optimizing and evolve your culture towards data-driven decisions

18

The Cultural Challenge

It is difficult to get a man to understand something when his salary depends upon his not understanding it.
-- Upton Sinclair

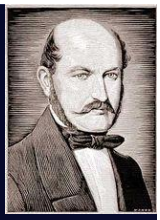
- Why people/orgs avoid controlled experiments
  - Some believe it threatens their job as decision makers
  - At Microsoft, program managers select the next set of features to develop. Proposing several alternatives and admitting you don't know which is best is hard
  - Editors and designers get paid to select a great design
  - Failures of ideas may hurt image and professional standing. It's easier to declare success when the feature launches
  - We've heard: "we know what to do. It's in our DNA," and "why don't we just do the right thing?"

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Cultural Stage 1: Hubris

- The org goes through stages in its cultural evolution
- Stage 1: we know what to do and we're sure of it
  - True story from 1849
  - John Snow claimed that Cholera was caused by polluted water
  - A landlord dismissed his tenants' complaints that their water stank
    - Even when Cholera was frequent among the tenants
  - One day he drank a glass of his tenants' water to show there was nothing wrong with it
- He died three days later
- That's hubris. Even if we're sure of our ideas, evaluate them
- Controlled experiments are a powerful tool to evaluate ideas

Cultural Stage 2: Insight through Measurement and Control



- Semmelweis worked at Vienna's General Hospital, an important teaching/research hospital, in the 1830s-40s
- In 19th-century Europe, childbed fever killed more than a million women
- Measurement: the mortality rate for women giving birth was
  - 15% in his ward, staffed by doctors and students
  - 2% in the ward at the hospital, attended by midwives

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Cultural Stage 2: Insight through Measurement and Control

- He tries to control all differences
  - Birthing positions, ventilation, diet, even the way laundry was done
- He was away for 4 months and death rate fell significantly when he was away. Could it be related to him?
- Insight:
  - Doctors were performing autopsies each morning on cadavers
  - Conjecture: particles (called germs today) were being transmitted to healthy patients *on the hands of the physicians*
- He experiments with cleansing agents
  - Chlorine and lime was effective: death rate fell from 18% to 1%

Cultural Stage 3: Semmelweis Reflex

- Success? No! Disbelief. Where/what are these particles?
  - Semmelweis was dropped from his post at the hospital
  - He goes to Hungary and reduced mortality rate in obstetrics to 0.85%
  - His student published a paper about the success. The editor wrote
    We believe that this chlorine-washing theory has long outlived its usefulness… It is time we are no longer to be deceived by this theory
- In 1865, he suffered a nervous breakdown and was beaten at a mental hospital, where he died
- Semmelweis Reflex is a reflex-like rejection of new knowledge because it contradicts entrenched norms, beliefs or paradigms
- Only in 1800s? No! A 2005 study: inadequate hand washing is one of the prime contributors to the 2 million health-care-associated infections and 90,000 related deaths annually in the United States

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Cultural Stage 4: Fundamental Understanding

- In 1879, Louis Pasteur showed the presence of Streptococcus in the blood of women with child fever
- 2008, 143 years after he died, there is a 50 Euro coin commemorating Semmelweis

Summary: Evolve the Culture

- In many areas we're in the 1800s in terms of our understanding, so controlled experiments can help
  - First in doing the right thing, even if we don't understand the fundamentals
  - Then developing the underlying fundamental theories

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

25

Summary

The less data, the stronger the opinions

1. It is hard to assess the value of ideas
   - Listen to your customers
   - Get the data by experimenting because data trumps intuition
2. Empower the HiPPO with data-driven decisions
   - Hippos kill more humans than any other (non-human) mammal (really)
   - OEC: make sure the org agrees **what** you are optimizing (long term lifetime value)
3. Compute the statistics carefully
   - Getting a number is easy. Getting a number you should trust is harder
4. Experiment often
   - Triple your experiment rate and you triple your success (and failure) rate. Fail fast & often in order to succeed
   - Accelerate innovation by lowering the cost of experimenting

http://exp-platform.com

Accelerating software Innovation through trustworthy experimentation

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Bonus True Story – Scurvy and Vitamin C

- Scurvy is a disease that results from vitamin C deficiency
- It killed over 100,000 people in the 16th-18th centuries, mostly sailors
- First known controlled experiment in 1747
  - Dr. James Lind noticed lack of scurvy in Mediterranean ships
  - Gave some sailors limes (treatment), others ate regular diet (control)
  - Experiment was so successful, British sailors are still called limeys
- But Lind didn't understand the reason
  - At the Royal Naval Hospital in England, he treated Scurvy patients with concentrated lemon juice called "rob."
  - He concentrated the lemon juice by heating it, thus destroying the vitamin C
  - He lost faith in the remedy and became increasingly reliant on bloodletting
- In 1793, a formal trial was done and lemon juice became part of the daily rations throughout the navy; Scurvy was quickly eliminated

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.
MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.