Source: im.edirectory.co.uk

Mastering The Dance:

Partnering Theory and Experimentation in Networking Research

Ravi Jain

jain@docomolabs-usa.com

August 30, 2005

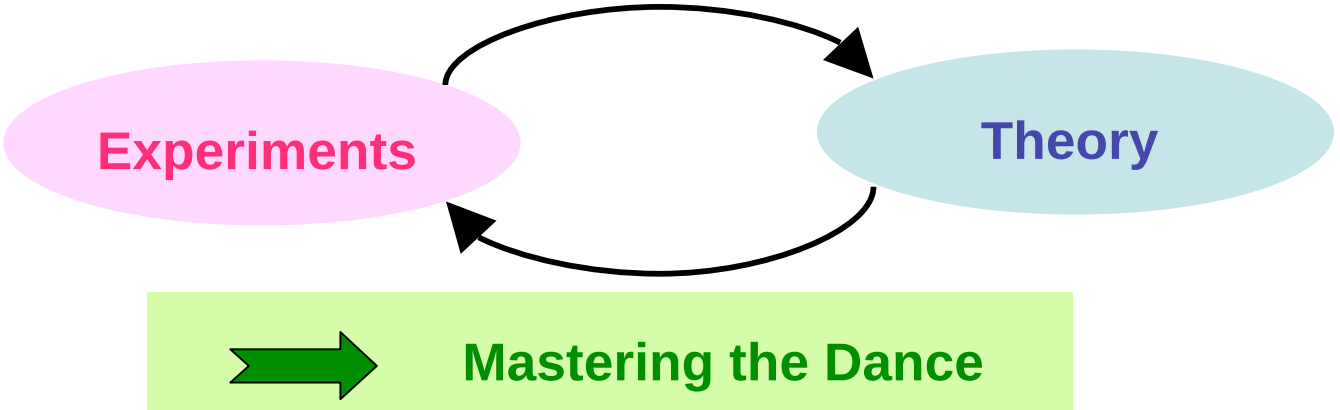
This talk represents the views of the author and not necessarily DoCoMo USA Labs

Outline

- Background
- Paradigms and examples
- Challenges
- Summary and Conclusions

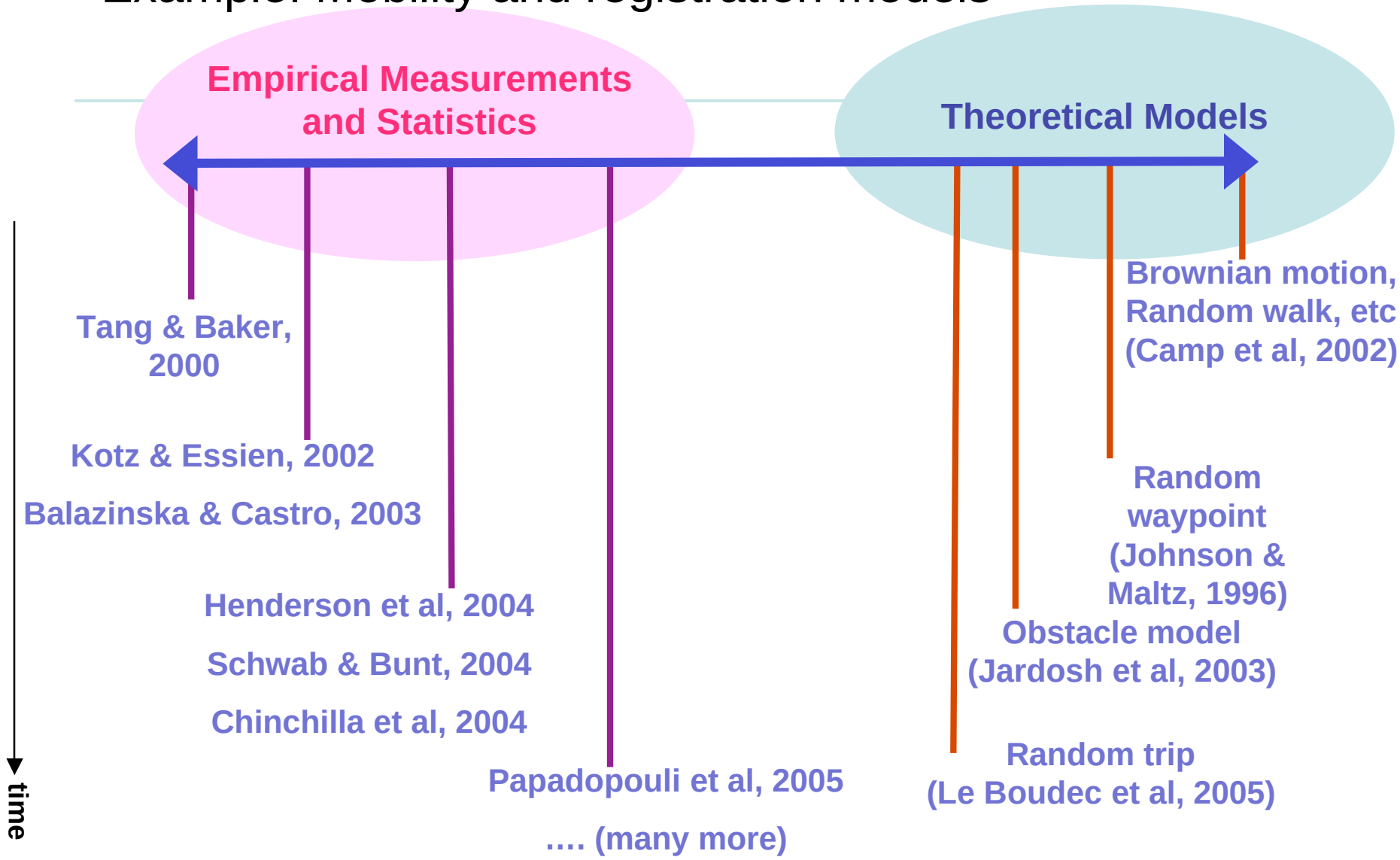
Background

- Networking research is starting to become mature
- A sign of maturity is
  - Rigorous experimentation
    - Not necessarily difficult or detailed
  - Effective theory
    - Not necessarily sophisticated or complex
  - A methodology for abstracting experiments into theory, validating theory by experiments, and iterating the process



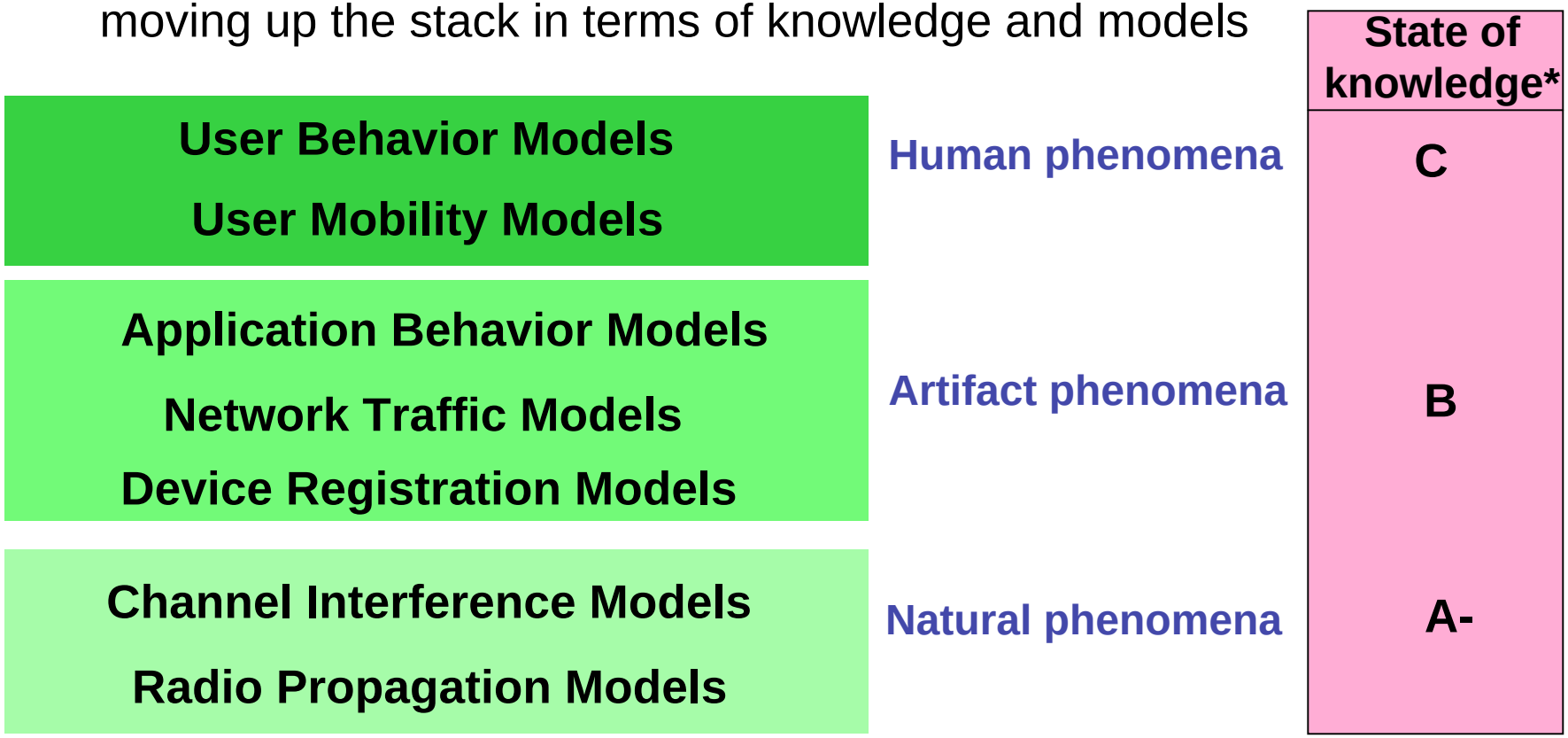
Where we are

Example: Mobility and registration models



Where we are

- Broadly speaking, (wireless) networking research is gradually moving up the stack in terms of knowledge and models



***A surely subjective and biased view**

Where we want to go
Examples

- The Locality Principle (Source: *Denning, 2005*)
  - 1959: Virtual memory introduced into the second-generation Atlas OS
  - Question: How to replace a page that was in the cache?
  - A bad replacement algorithm could
    - Result in thrashing, killing performance
    - Cost millions of dollars of wasted machine time over the system lifetime
  - 1960s: Many page replacement algorithms designed and tested
  - 1966: Denning proposes “Working set” theory
  - 1965-69: Major IBM experimental study concludes that LRU is best
  - 1976: Locality is part of the human cognitive process transmitted to programs
  - Today: locality in search, web caches, computer forensics etc

Where we want to go
Examples

- Self-similarity in network traffic
  - *Leland, Taqqu, Willinger and Wilson, 1993*
  - Long-held view: Markovian (e.g. Poisson) models describe telephone traffic, and hence, by extension, data traffic
  - Problem: unable to explain poor loss and delay performance of switches and other artifacts
  - New view: Every $@#% type of data traffic seems to indicate heavy-tailed behavior

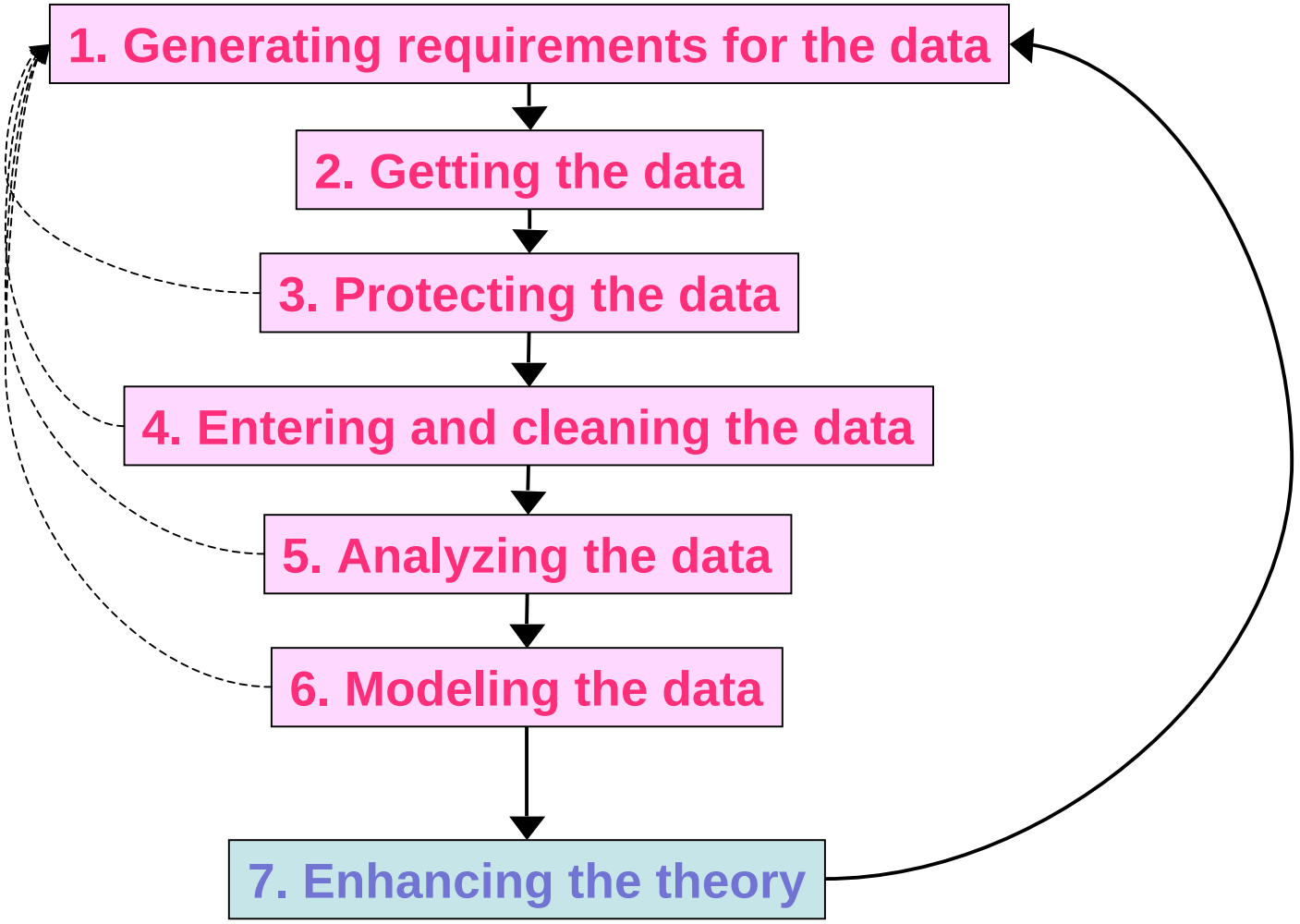
Paradigm
Mapping The Genome

- News Flash (August 22, 2005)
  - Public collections of DNA and RNA sequences reach 100 Gigabases of sequence data
    - 100,000,000,000 is ~number of stars in the Milky Way
- Late 1970s: scientists study genetic code sequences and explore sharing of data
  - Manual data entry, distribution by magnetic tape, disks etc
- Public collaborative databanks
  - Early 1980s: Two data banks (Germany, US), 1987: Japan
- Free access to scientists anywhere in the world
- Numerous free browsing and analysis tools, catalogs, sophisticated sequence matching software, etc etc
  - Vastly expands the talent pool that can work on these problems
  - Facilitates cross-disciplinary work (life sciences, CS, math people)
- A boon to bioinformatics and genomics

Base: Any basic (alkaline) compound containing nitrogen, but generally referring to one of four complex molecules (nucleotides) that form the building blocks of the nucleic acids, DNA and RNA

One level deeper: A process view

Challenges at every stage of the process



Challenges:
2. Getting the data

- First option: Get the data from someone else
  - No obvious incentive for commercial entities to release proprietary data
    - Need to show that the external research community provides leverage that exceeds competitive advantage of hoarding data
  - Unlike genomics, there is no publicly funded data collection that mandates release of cellular system data
- Second option: Get it yourself
  1. Measure someone else's system
     - Example: black-box measurements of cellular systems
     - Example: campus WLAN traces
  2. Build your own system
     - Obviously impractical for medium or large-scale systems
     - Or is it?

2. Getting the Data: PlanetLab

- Overlay-based Network Testbed
  - World-wide, 500+ nodes, 20+ country
  - Runs over Internet
  - 450+ research projects are currently running over it
- Distributed Computing and Resource Sharing
  - Virtual machine based technology
  - Each project/service gets one or multiple resource slices to be deployed and experimented with
  - Centralized resource scheduling
- Founded in 2002 by Princeton Univerisity, UC Berkeley, and Intel Research
- http://www.planet-lab.org/

2. Getting the data: Wireless PlanetLab

- PlanetLab with wireless devices, last-hops and subnetworks
- Two aspects
  - Define, allocate and control a wireless “slice”
  - Use as a testbed to understand performance and scalability issues
- Perform experiments under real (or close to real) conditions and network traffic
  - “real”:
    - PlanetLab is over Internet
    - Routing delay is real
    - Background traffic is real
  - “not real”
    - Service node is not a dedicated node
    - Service usage pattern and traffic model may not be real
      - May need to emulate mobility or traffic based on other models

3. Protecting the data

- Protecting the user: Privacy
- Protecting the owner of the data (e.g. enterprise): Reputation, Market advantage
- Protecting the system: Attacks

User privacy:

I hate to bring this up, but …

- Press Release
- 7/8/2004
- **CONFIDENTIALITY OF GENETIC DATABASES QUESTIONED BY STANFORD RESEARCHERS**
- STANFORD, Calif. – In their exuberance over cracking the genetic code, scientists have paid too little attention to privacy issues, say researchers at the Stanford University School of Medicine.
- "I am surprised that no one has looked at this problem before and asked, 'Can we really release genome-wide information about individuals to the public,'" said Zhen Lin, a genetics graduate student who led the study. "Nobody did a careful calculation to find whether 'anonymous' patients could be identified from this data."
- Why worry? Lin said that insurance companies and employers potentially have an interest in learning whether a person is prone to certain illnesses, and that malevolent individuals might also try to seek out this type of information.

User Privacy

- **Wired magazine**
- **Issue 3.03** - Mar 1995
- **You're Not Paranoid: They Really Are Watching You**
  Surveillance in the workplace is getting digitized - and getting worse.

User privacy

- Options
  - Get data by delegated consent (e.g. campus sys admin) and satisfy them that privacy will be maintained
  - Get data by informed consent of the end user

- IMHO, in the long run the second option will be required
  - Informed consent, with opt-in/opt-out
  - Model consent forms
  - Anonymization and cryptographic protection
  - Security of data handling
  - Audit capabilities and Review boards

Protecting the owner and system

- Protecting the data owner
  - Anonymization protects the user but not the service provider
  - Example of Reputation risk: Poor coverage in certain areas
  - Example of Market risk: Leak some information that gives competitors advantage
  - IMHO, this will not be solved easily
- Protecting the system
  - Example: Does public release of the data (or a model based on the data) make DoS attacks easier
    - May also hurt the user or the data owner
- Is it worth the hassle?
  - Community effort is required to help reduce the burden

4. Entering and cleaning the data

- Press Release
- March 7, 2005
- **Genome centers combine forces to validate a gene set for biomedical research**
- The advent of online databases to access the human genome has been a boon to biomedical research, (but …)
- … the same gene may have different names in different databases. Since the data characterizing the genes come from a variety of sources, researchers are not always certain that a listed gene is real and its stated function is accurate. … Inconsistencies arise because different centers have used different methods …

Early on, identify and make explicit terminology, assumptions, scenario and testbed description method (XML?)

5. Analyzing the data

- Beyond the data itself we need a public repository of
  - Basic tools: sniffers, loggers, scripts
  - Methodologies: How-to
  - Folklore: “Measuring 802.11b management frames is really hard”

Summary and Conclusion

- A maturing field needs interplay of rigorous experimentation and effective theory
- Pooling information and resources in common scientific archives and databases can be an invaluable boon
- Getting data from companies will continue to be hard … so maybe we should “build it ourselves” (Wireless PL)
- There are serious technical and non-technical challenges
  - Protecting data, maintaining its quality, reducing the drudgery of analyzing it
- This is a lot of work
  - Yes, but think of the end result …!