

# The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text

**Christopher Cieri, David Miller, Kevin Walker**  
**[{ccieri,damiller,walker}@ldc.upenn.edu](mailto:ccieri@ldc.upenn.edu)**

**University of Pennsylvania**  
**Linguistic Data Consortium and Department of Linguistics**  
**3600 Market Street, Philadelphia, PA 19104 U.S.A.**

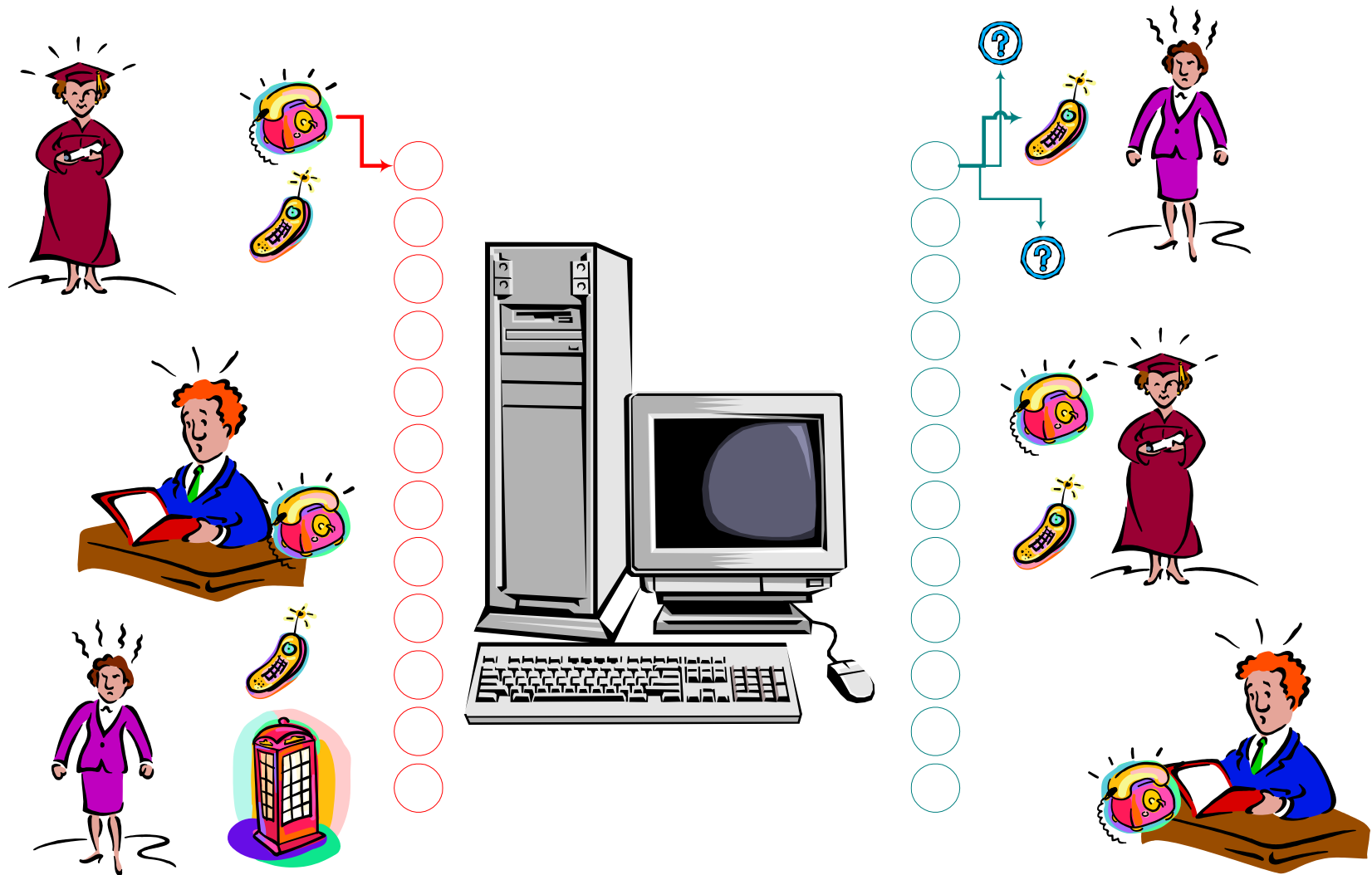
**[www.ldc.upenn.edu](http://www.ldc.upenn.edu)**

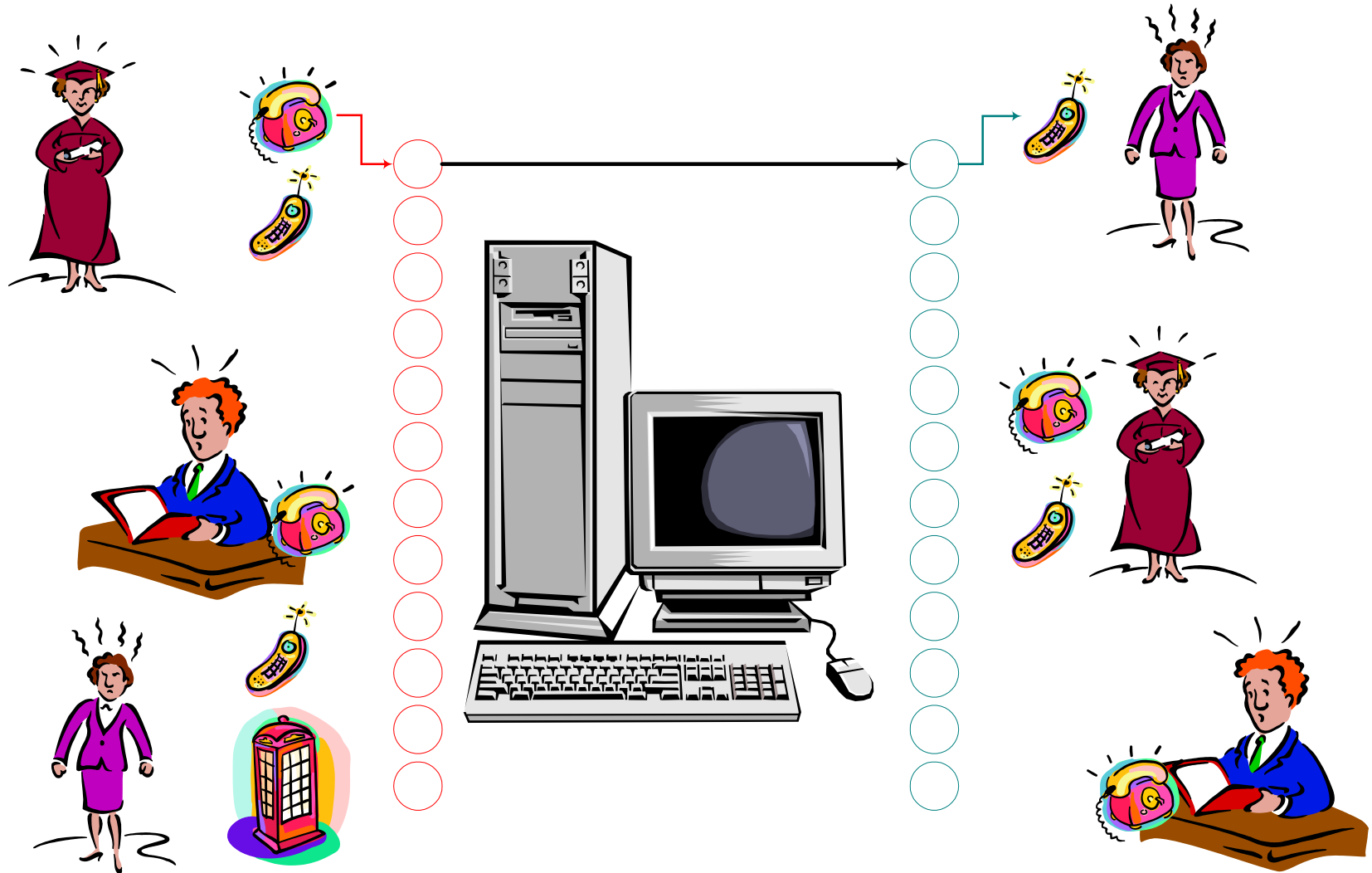
- **Corpus users and authors increasingly interested in:**
  - greater volumes of data in more languages
  - with more sophisticated annotation
  - for use in an expanding number of disciplines
  - requiring standards, tools and best practices
- **LDC addressing needs by**
  - specific projects in data collection, annotation and publications
  - incorporating annotation, research and tool development
- **Need to increase the quantity, quality and diversity of language resources**
  - more intensive collaboration between researchers and data providers
  - yielding more data creators, researchers with better appreciation for data creation and data creators with better appreciation of data uses
- **Requires more intensive resources planning (roadmaps)**
- **Need greater cooperation among international data centers which is compatible with local mandates.**
- **LDC open to cooperation with individuals and data centers around to world.**

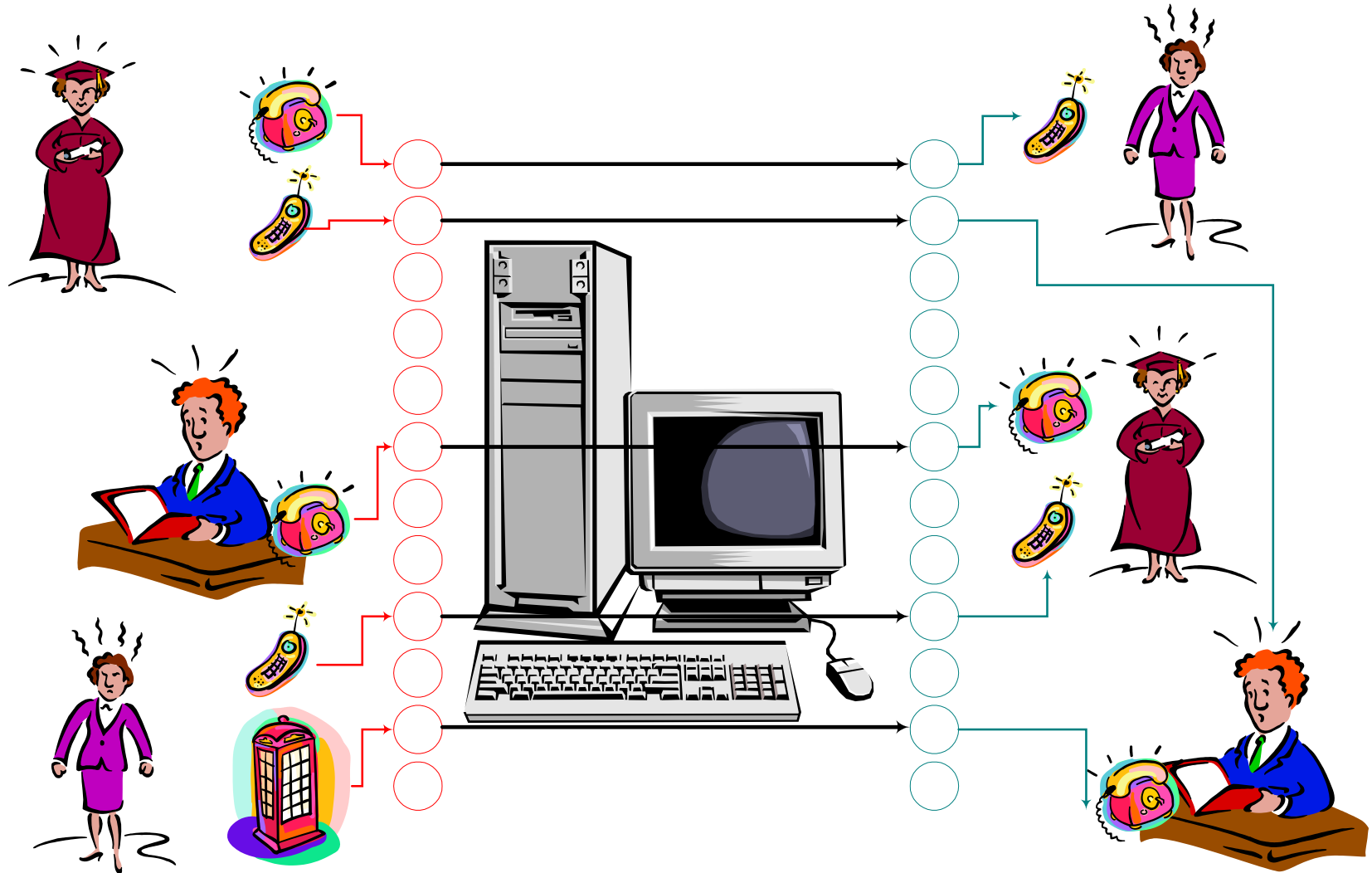
- **Effective Affordable, Reusable Speech-to-Text**
  - DARPA common task project driven by annual go/no-go criteria
  - to achieve 5 fold increase in speed, accuracy
  - generate readable transcripts adapted for downstream processing
- **Case study in resource planning where demand exceeds supply**
  - exploited existing resources: Switchboard, TDT, new TIDES collections
  - required difficult decisions RE
    - priority of different research areas, languages (effort for English > Arabic > Chinese) and volumes of data for training and testing
  - raw data collection required to supply STT & MDE, training and test corpora
  - focus on simple annotations that humans perform consistently in high volume
- **LDC provides**
  - broadcast news, conversational telephone speech, meetings
  - time aligned transcripts, annotation for metadata extraction (MDE)
  - training, development test and evaluation data
  - English, Mandarin and Arabic

- **Just one of many EARS data goals**
- **Volume**
  - 2000 hours
  - each subject makes 1-3 calls
  - maximum call length is 10 minutes
- **Assigned topics**
  - 40 original
  - 60 implemented in November
- **Demographic Goals – balanced within 10% absolute**
  - Sex: m/f
  - Age: 16-29, 30-49, 50+
  - Region: North, Midland, South, West, Canada, Other (?)
  - also monitor handset, education, occupation in collection
- **High Quality, Time-Aligned Transcripts for all speech**

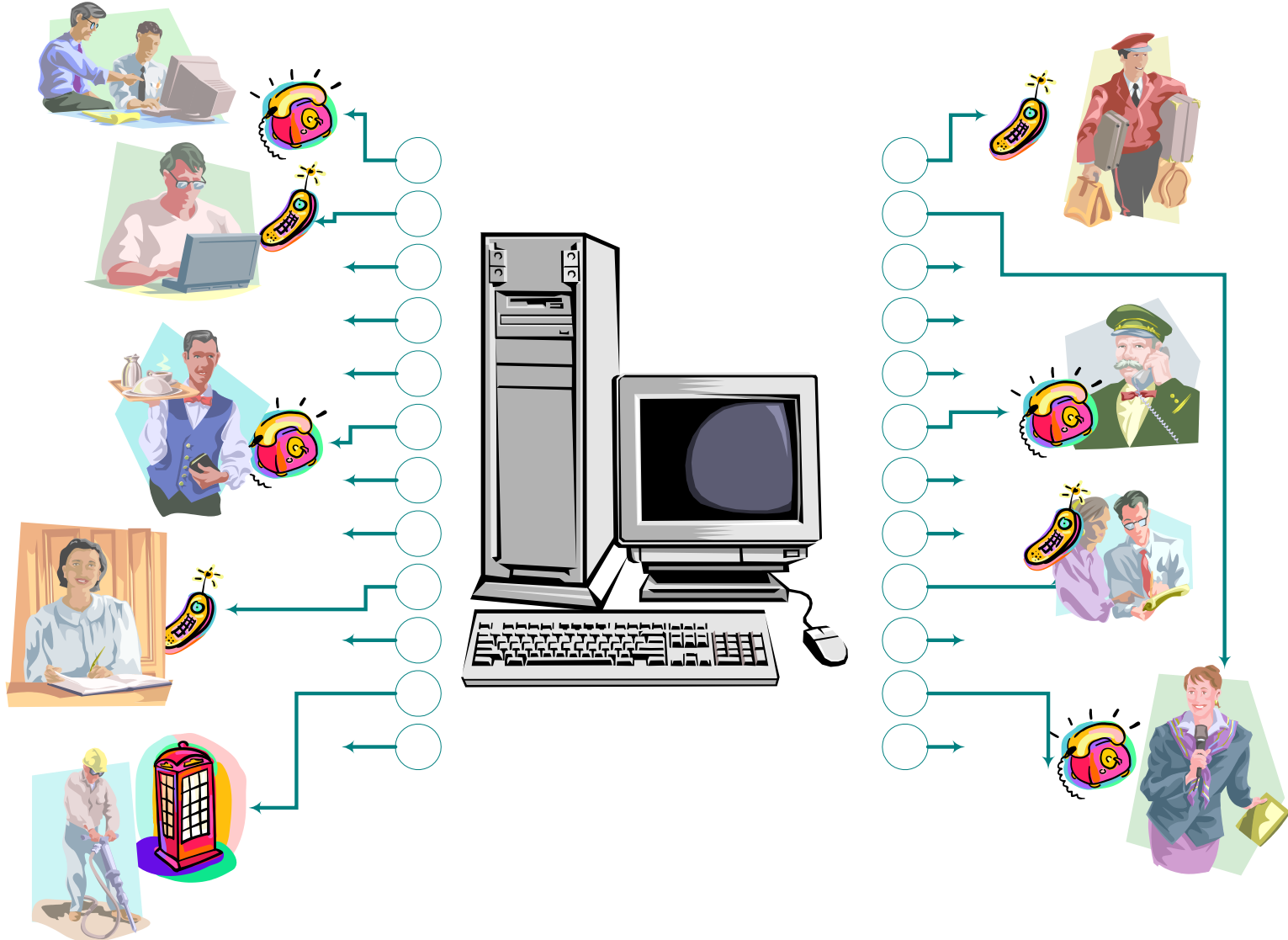
- **All LDC telephone studies**
  - follow US regulations on treatment of human subjects
  - audited annually by an Internal Review Board (IRB)
  - managed by the University of Pennsylvania Office of Regulatory Affairs
- **Main issues informed consent & risk vs. benefit**
  - all participants informed that calls recorded for research, educational purposes
  - main benefits are societal
    - » benefit to subjects is monetary compensation, free call
  - main risk is to anonymity
    - » Subjects identified by 5 digit PIN
- **New IRB protocol covers all speech collections**
  - prompted or conversational
  - human-human or human-machine
  - face-to-face or telephone

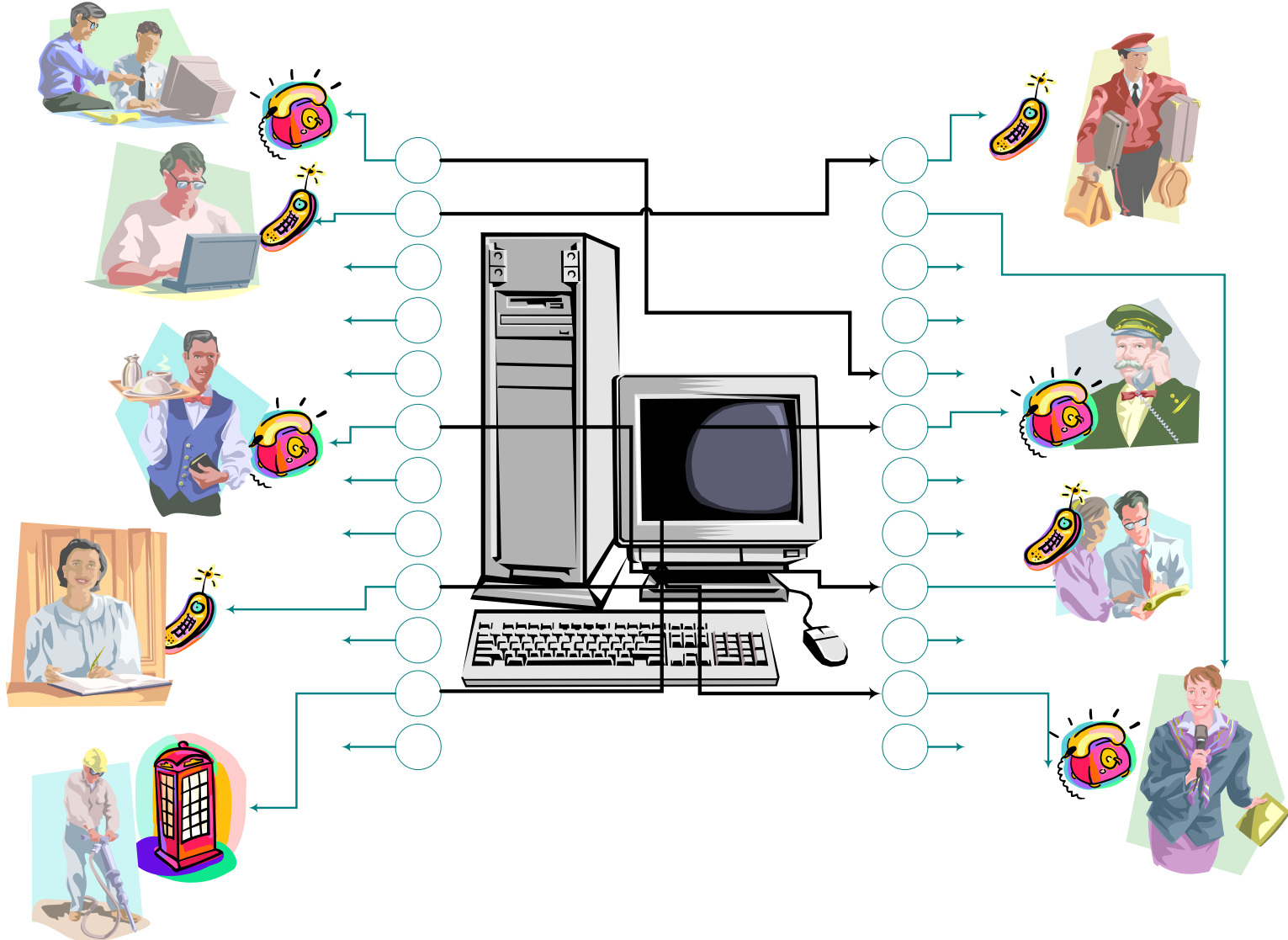


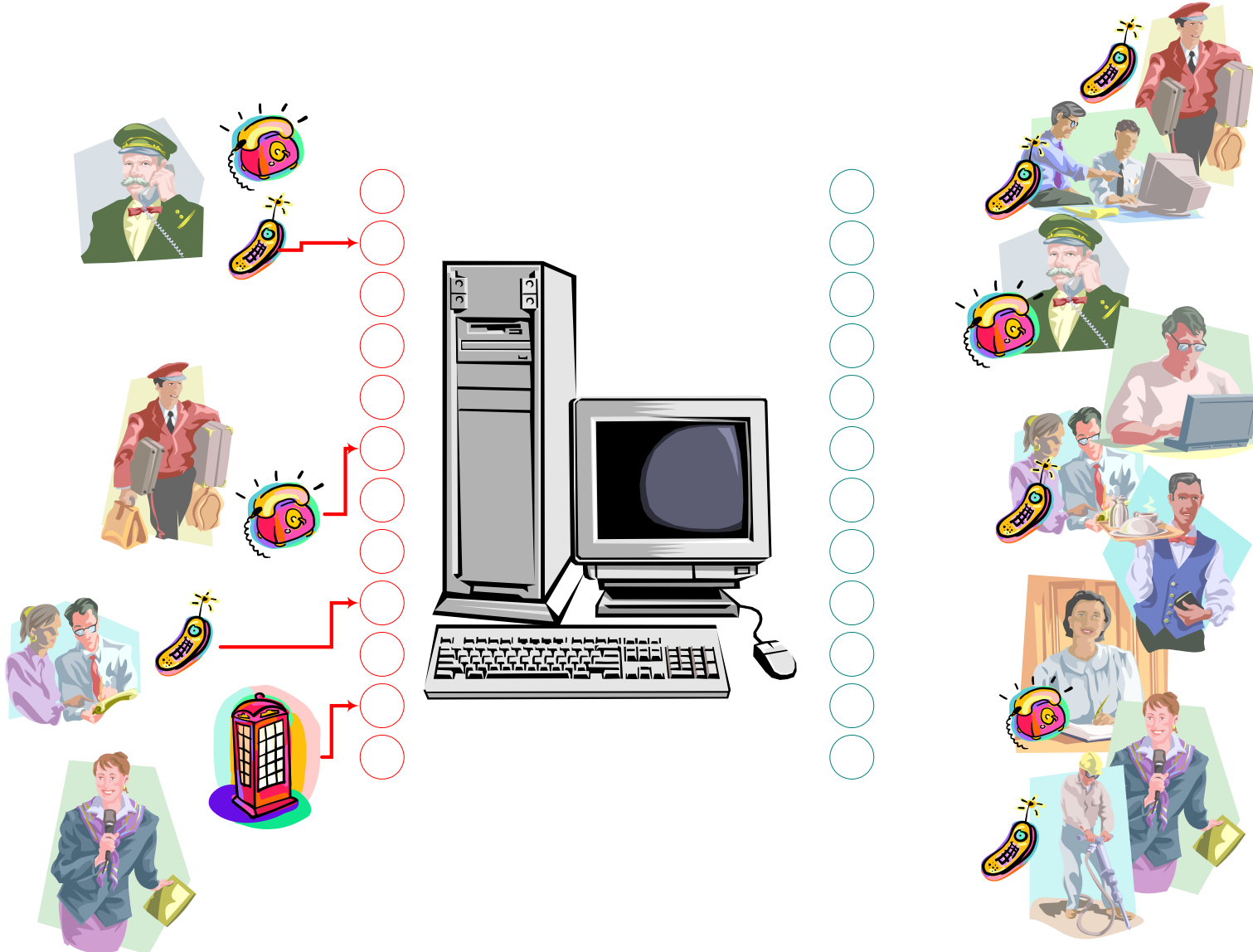


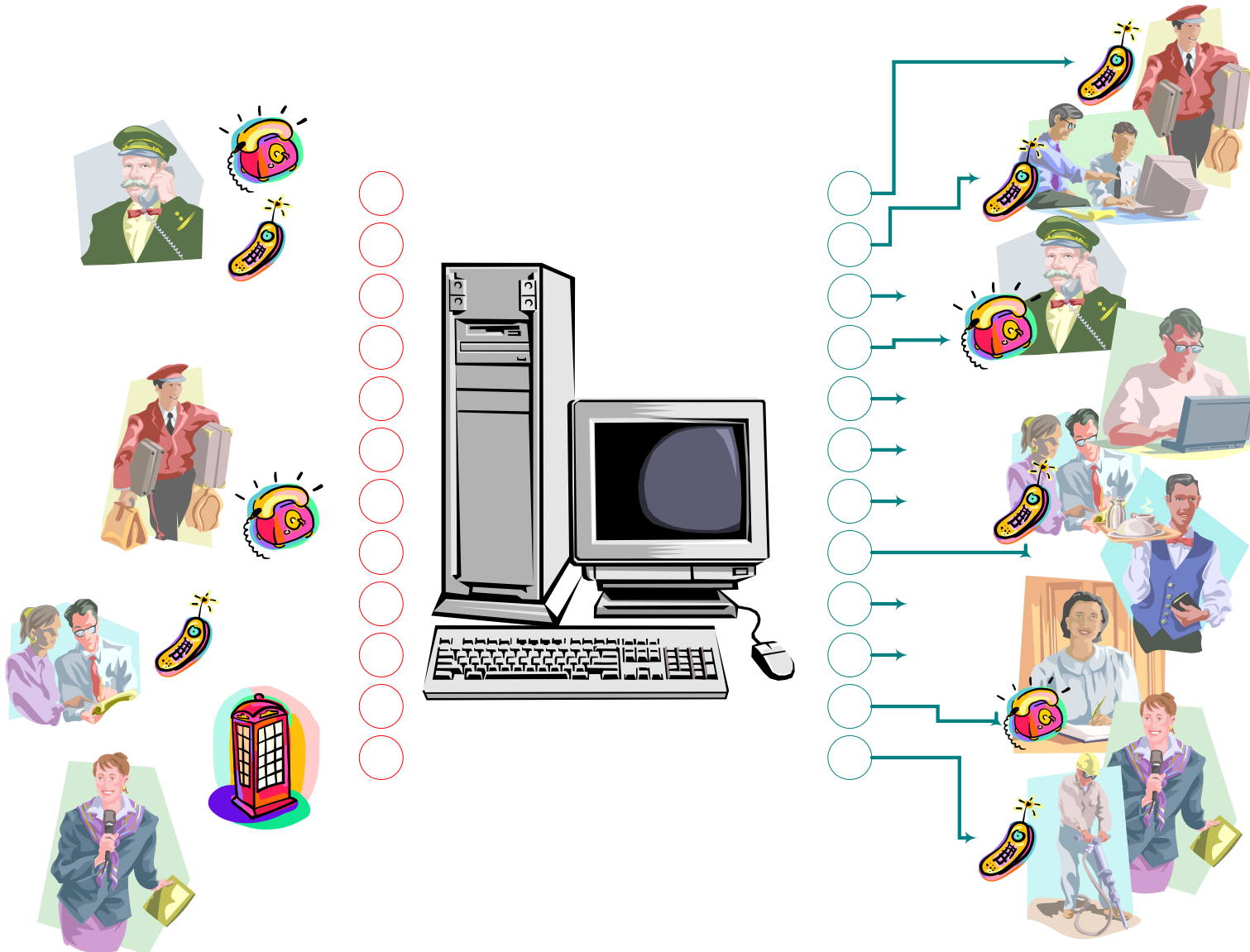


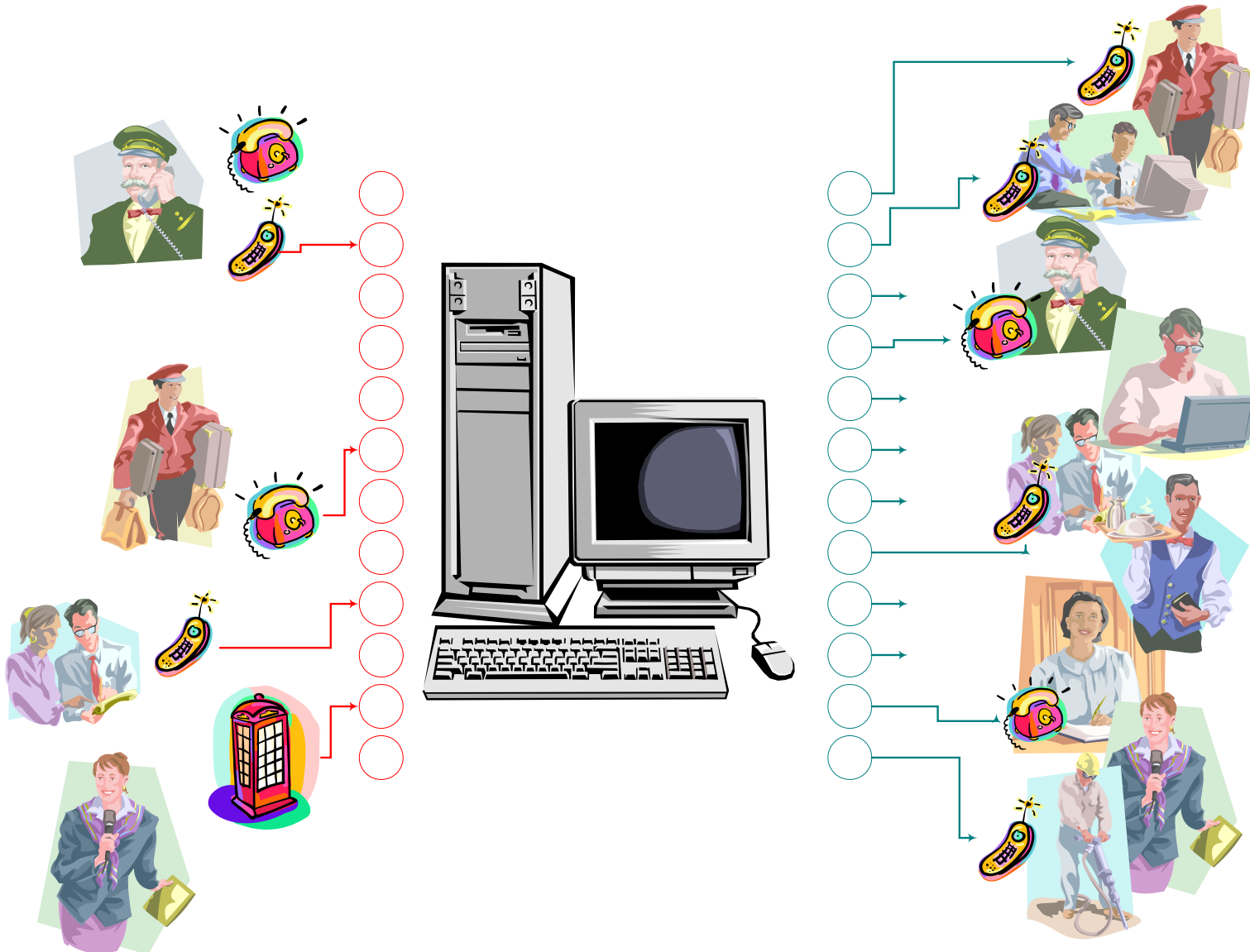


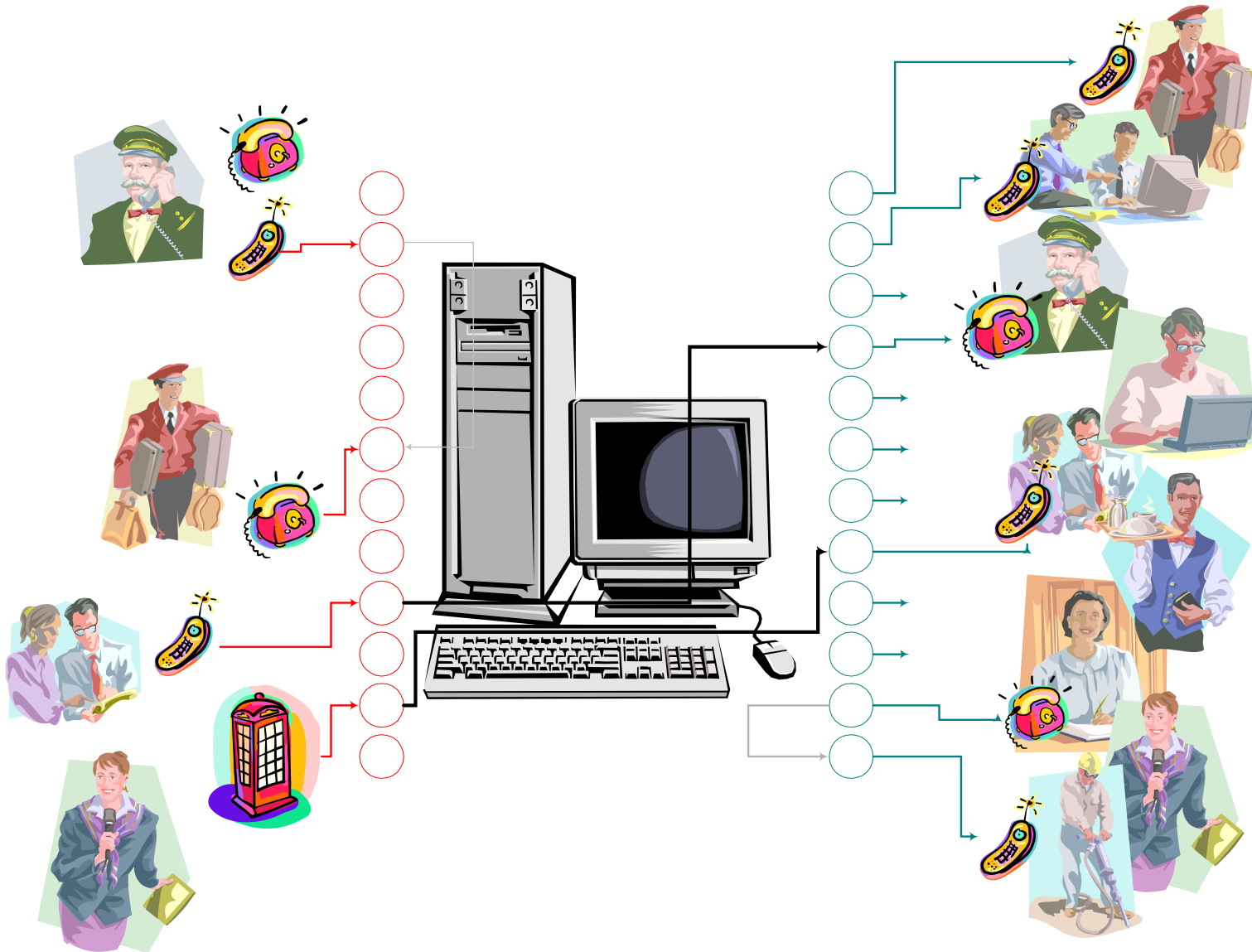


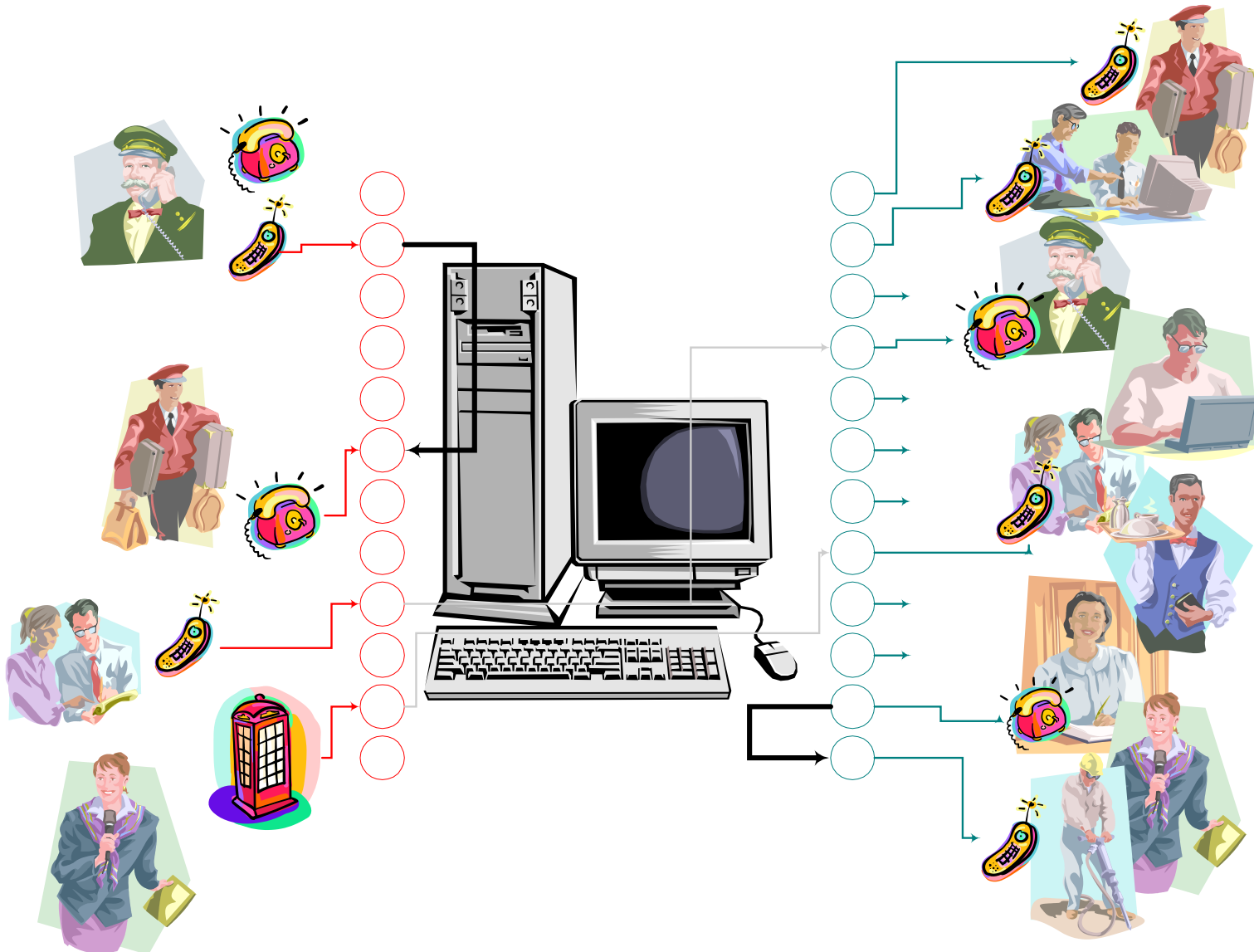


















- **Collection began 12/15/2002, continued for 1 year**
- **Platform in operation**
  - 7 days per week
  - noon (EST) > midnight (PST)
- **Call collection driven by:**
  - availability schedules of participants
    - » given by day and hour
    - » robot operator called at least once in each available block
  - caller activity
    - » in Fisher, callers had little motivation to initiate calls
    - » Mixer offer incentives for call-ins and volume is much higher
    - » platform functioned well in both cases
    - » non-participation = de-selection
  - total platform activity (energy)
- **Relatively small number of calls per subject increased requirement on recruiting**

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers

So, all I have to do is **talk on my phone. I get paid \$10 each time!** Yeah, and they only need 1-3 calls per participant. But, by me chatting on the phone to support scientific research

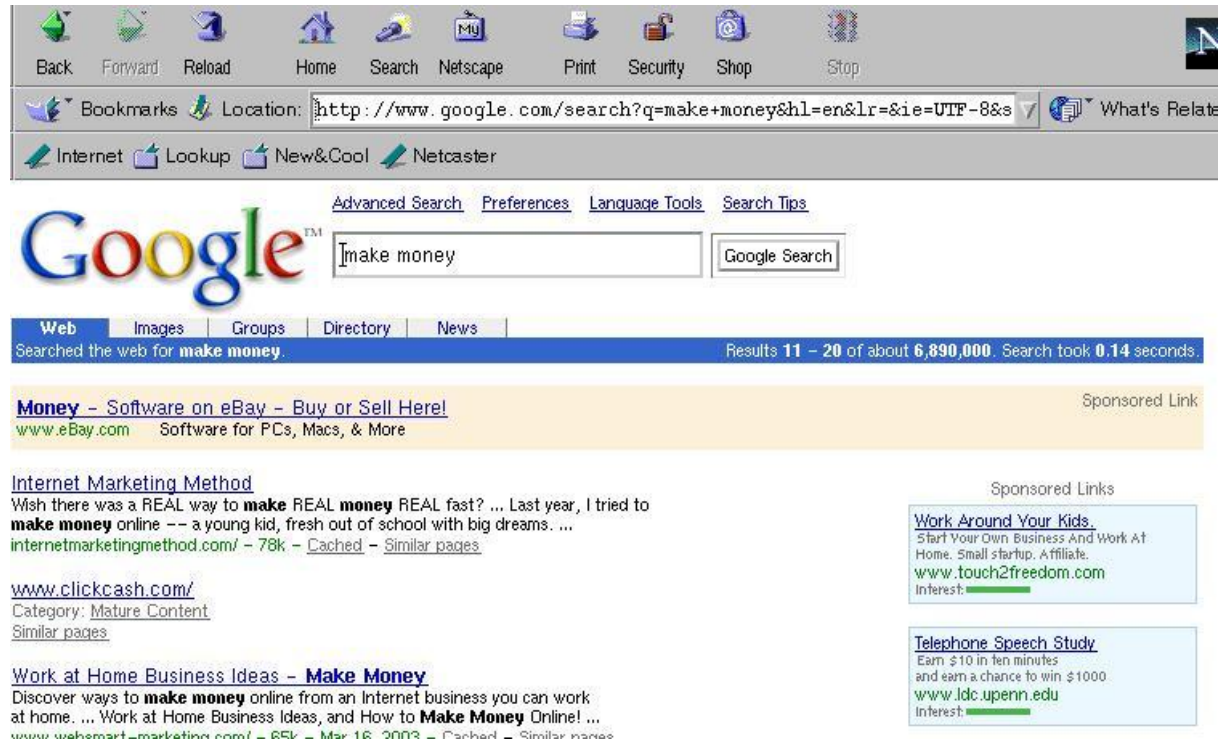
**I may have the chance to win \$1000!**  
Wow!



The Linguistic Data Consortium at the University of Pennsylvania ([www ldc upenn edu](http://www ldc upenn edu)) needs participants for FISHER, a new telephone speech study. The FISHER project will support linguistic research, technology development and education. FISHER participants will take part in 1 to 3 telephone calls talking for ten minutes to other participants on suggested topics. A robot operator will initiate all calls. Participants need only answer their phones at the times they specify during the registration process. To register call:

**1-800-380-PENN**

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers



The screenshot shows a Netscape browser window with the address bar displaying <http://www.google.com/search?q=make+money&hl=en&lr=&ie=UTF-8&s>. The search bar contains the text "make money" and the "Google Search" button is visible. Below the search bar, the results show "Searched the web for **make money**." and "Results 11 - 20 of about 6,890,000. Search took 0.14 seconds."

The search results include several sponsored links and organic results:

- Sponsored Link:** [Money - Software on eBay - Buy or Sell Here!](#) [www.eBay.com](#) Software for PCs, Macs, & More
- Sponsored Links:**
  - [Work Around Your Kids.](#) Start Your Own Business And Work At Home. Small startup. Affiliate. [www.touch2freedom.com](#) Interest: [www.touch2freedom.com](#)
  - [Telephone Speech Study.](#) Earn \$10 in ten minutes and earn a chance to win \$1000. [www ldc.upenn.edu](#) Interest: [www ldc.upenn.edu](#)
- Organic Results:**
  - [Internet Marketing Method](#)  
Wish there was a REAL way to **make REAL money** REAL fast? ... Last year, I tried to **make money** online -- a young kid, fresh out of school with big dreams. ... [internetmarketingmethod.com/ - 78k - Cached - Similar pages](#)
  - [www.clickcash.com/](#)  
Category: [Mature Content](#)  
[Similar pages](#)
  - [Work at Home Business Ideas - Make Money](#)  
Discover ways to **make money** online from an Internet business you can work at home. ... Work at Home Business Ideas, and How to **Make Money** Online! ... [www.wahomart-marketing.com/ - 65k - Mar 16, 2003 - Cached - Similar pages](#)

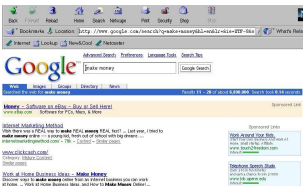



- # Make What You Say Count!

Earn some spare cash at the same time!



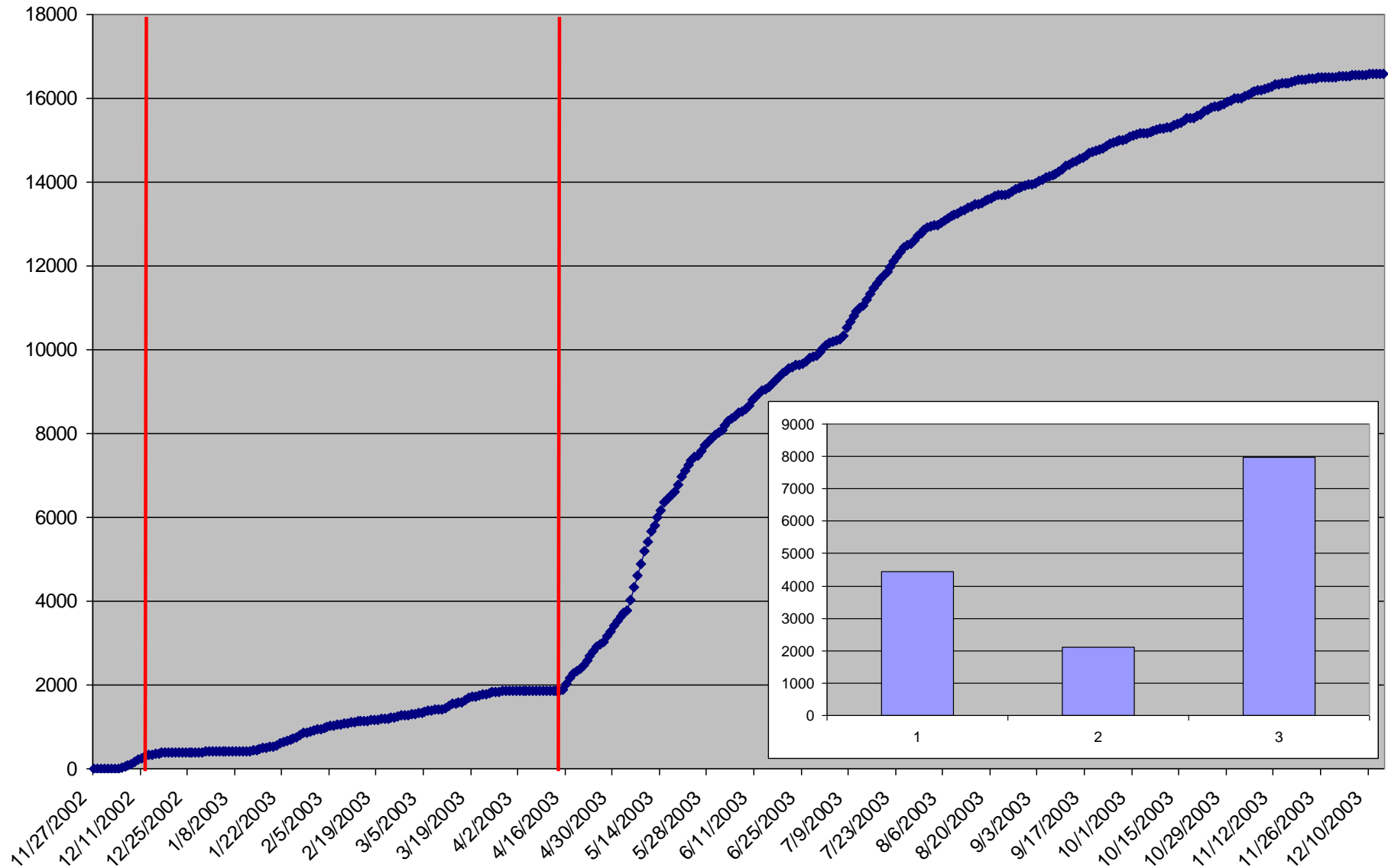
[www ldc upenn edu/Fisher](http://www ldc upenn edu/Fisher)



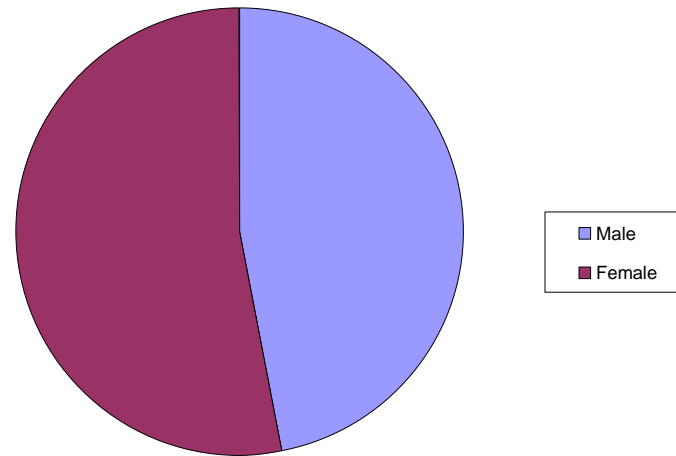
- 

[illegible]

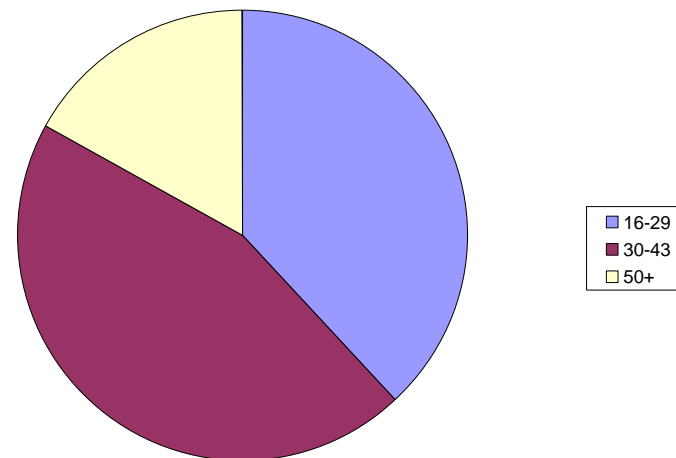
- 16,454 calls, 2742 total hours audio



- **Gender balance**
- **53% female**
- **47% male**

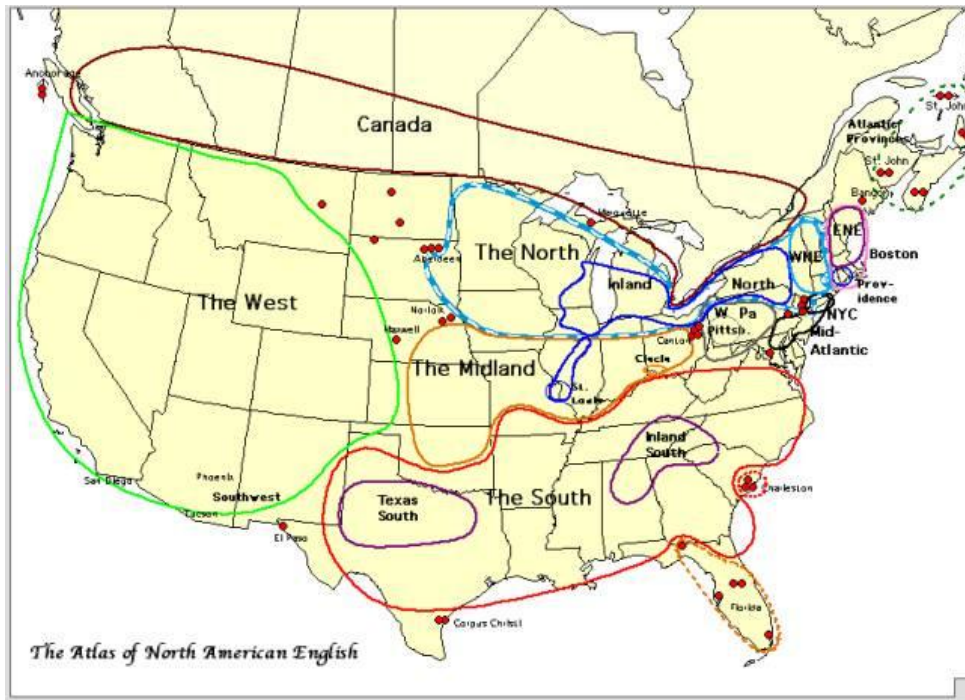
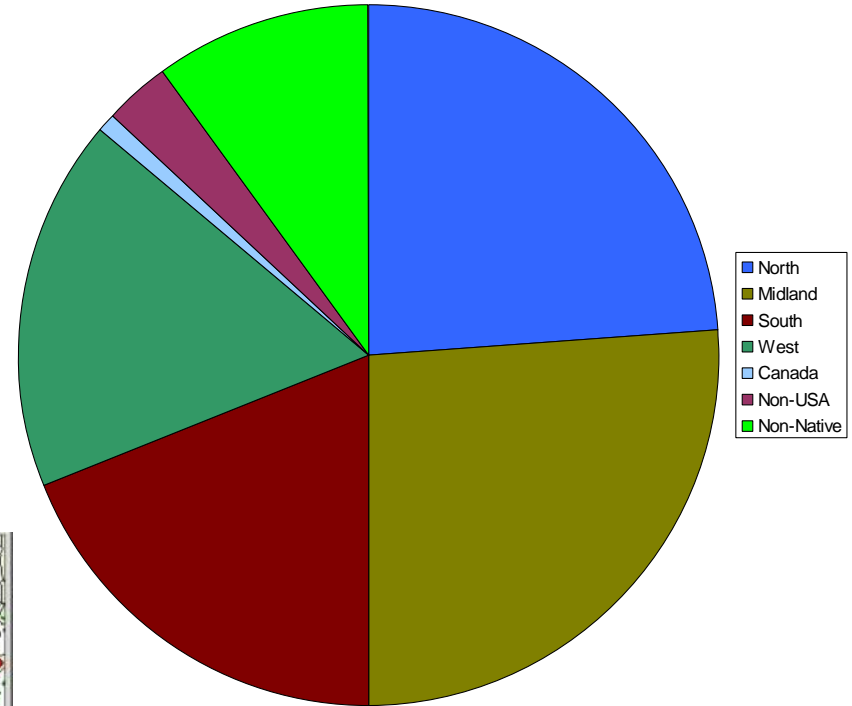


- **Distribution by Age Group**
  - 16-29 38%
  - 30-49 45%
  - 50+ 17%





- **North** **24%**
- **Midland** **26%**
- **South** **19%**
- **West** **17%**
- **Canada** **1%**
- **Non-USA** **3%**
- **Non-Native** **10%**



- **All calls receive quick human audit**
  - 160 seconds, 4 segments
  - Grade: A, C, F
- **Auditors check for:**
  - **Language:** Is it English? Is it understandable?
  - **Speaker:** Does speaker seem to belong to age, gender registered?
  - **Channel:** Do noise, echo, distortion levels interfere with comprehension
  - **Call Content:** Is discussion directed speech on assigned topic?

- **Provides order of magnitude more training data by focusing on speed of transcription**
- **Specification**
  - complete, verbatim
  - without punctuation, special symbols, talker/background noise
  - with limited interjections, non-lexemes
  - (( )) for unclear speech, – for truncated speech
  - annotators may insert other special symbols, punctuation if natural
- **Rates**
  - Segmentation: 3xRT > 0xRT (automatic or forced aligned)
  - Transcription: 5xRT
  - Post Processing 1xRT: QC on spelling, format, numbers
- **Challenges:**
  - spelled acronyms, numbers, spacing, proper names, disfluencies
- **Compared favorably with carefully transcribed training data**
  - all new EARS English and Arabic training data is QTr style
  - most English produced by WordWave under contract to BBNT.
  - LDC provides some English QTr and all Levantine Arabic

- **Fisher 2003 used in EARS; released in 2004-2005 (?)**
- **Fisher 2004 underway**
  - similar model
  - >1000 hours new collection
  - subjects allowed to make up to 20 calls
- **Collection protocol used in MMSR**
  - Multilingual, Multi-channel Speaker Recognition
  - Subjects complete 10+ six-minute calls on assigned topics
  - 400+ bilingual subjects speak in Arabic, Mandarin, Russian, Spanish
  - 200 subjects recorded on 9 different channels, sensors
  - 550 subjects completed 20+ calls
  - See the poster today at 5:00 in session 9-SE in the Laman room

Insert Buffers Files Tools Edit Search Mule Help

98.64 100.53 A: and it's not popular at all and i think

100.67 103.39 A: it just because she's not likeable and people don't like her

103.75 105.81 B: right yeah that makes sense

106.00 106.51 A: uh-huh

106.96 112.64 B: yeah i think also like with reality ~tv people can imagine themselves more easily

112.83 113.53 B: being

114.25 115.22 B: a part of that

115.59 116.30 B: you know like