ADDRESS-BASED SAMPLING

Merits, Design & Implementation, and Review of Field Statistics

Mansour Fahimi
VP, Statistical Research Services

MARKETING | SYSTEMS | GROUP
Leadership Through Innovation

GENESYS
SAMPLING SYSTEMS

FedCASIC 2010 Workshop
March 17, 2010

Presentation Outline

- Emerging Alternatives in Survey Administrations
- Issues with the "Old" Methods
- Need for More Flexible/Innovative Methods
- Using the CDSF for Sampling Purposes
- Potential Issues with the CDSF as a Sampling Frame
- Possible Enhancements of the CDSF
- Closing Remarks
- Ad-hoc Field Statistics

Emerging Alternatives in Survey Administrations

- Reasons for the emergence of address-based sampling (ABS) methodologies:
  - Evolving coverage problems associated with telephone samples
  - Eroding rates of response to single modes of contact along with the increasing costs of remedial measures to counter nonresponse
  - ABS provides a versatile platform for peeking outside of the box to improve coverage and response rates
- Using the USPS Computerized Delivery Sequence File (CDSF) for sampling purposes

Coverage Problems for Telephone Surveys

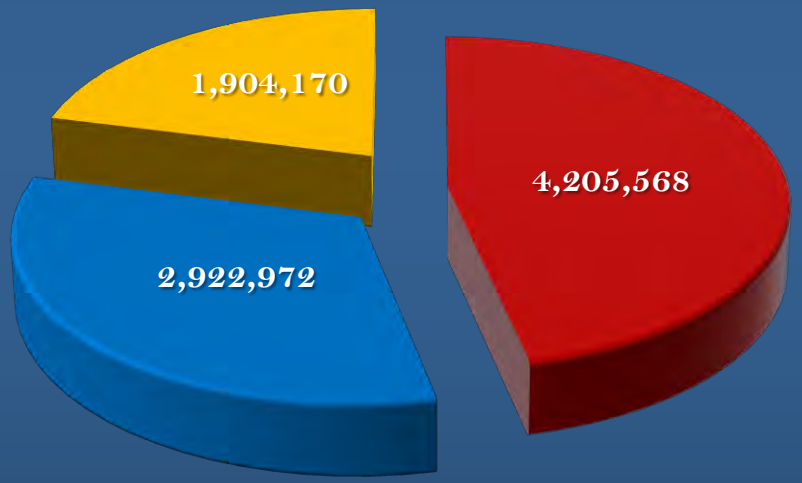
- A growing number of households are becoming cell-only or cell-mostly:
  - According to NCHS more than 3 out of 10 adults in the US receive all or nearly all calls on cell phones
  - Cell-only and cell-mostly individuals have different characteristics: younger and more mobile
  - If these individuals are not included in surveys results can be biased
- A growing number of residential telephone numbers are begin assigned outside of the traditional (list-assisted) RDD frame

Coverage Problems for RDD Surveys

Nature of the Problem

- RDD was developed nearly two decades ago under an AT&T dominated infrastructure
- List-assisted RDD simplified sampling and increased productivity
- Some of the assumptions based on which RDD was developed no longer hold

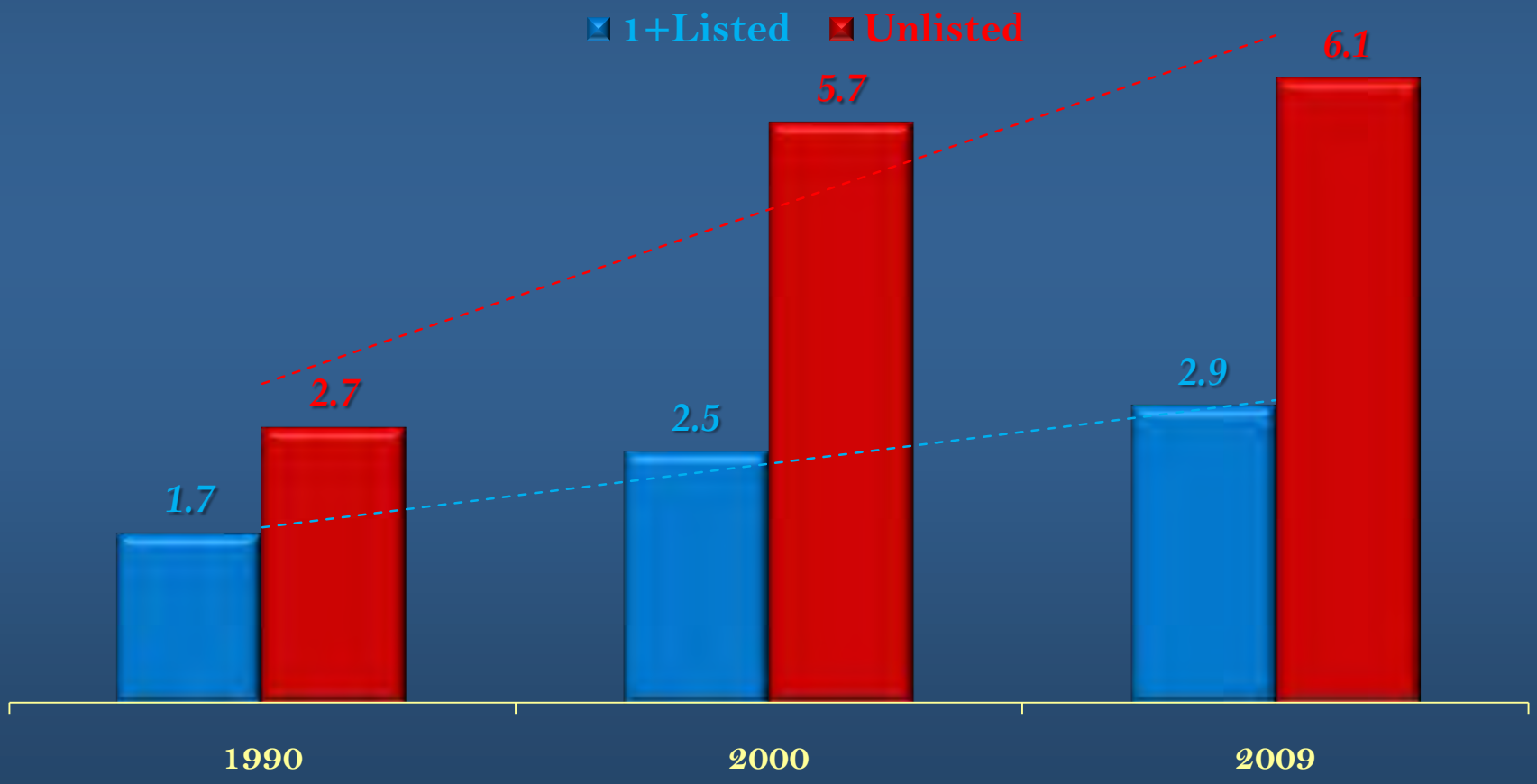
Distribution of 100-Series Bank Types
NPA-NXX-XX00 to NPA-NXX-XX99



Coverage Problems for RDD Surveys

Dilution of the RDD Frame

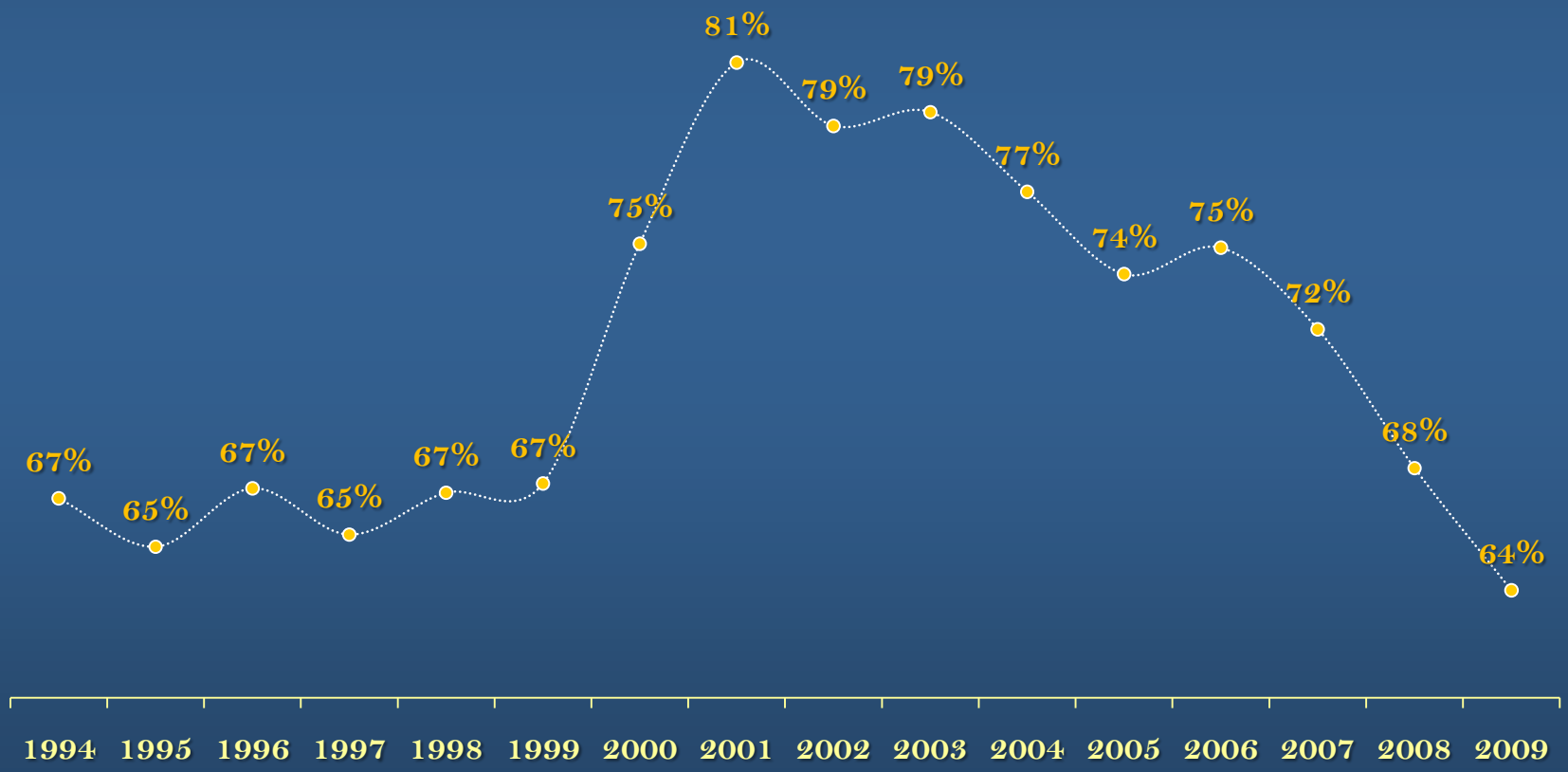
Uneven Growth of the 100-Series Banks



Coverage Problems for RDD Surveys

Decline in Directory-Listed Rates

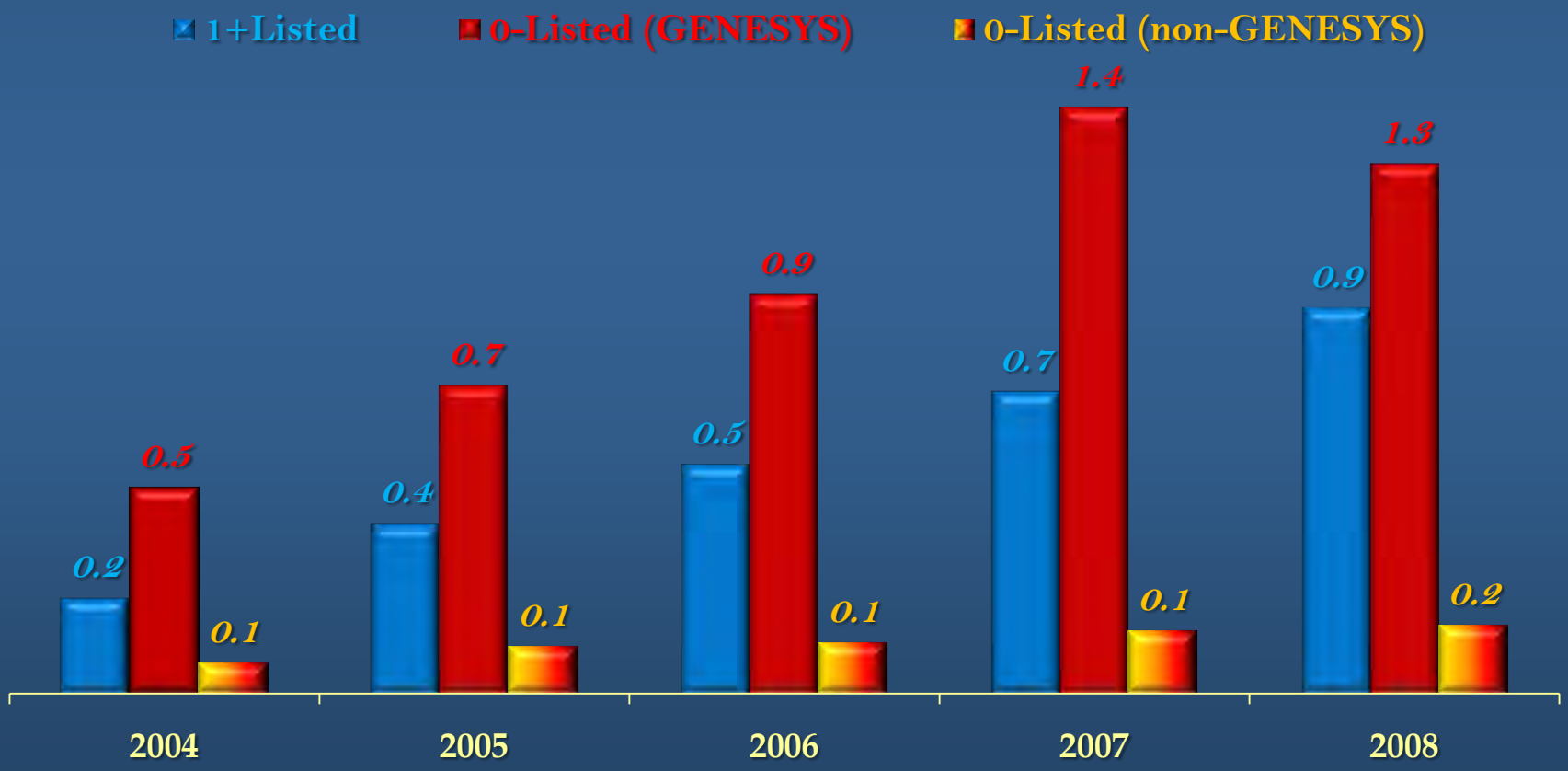
Estimated Proportion of Listed Households by Year



Coverage Problems for RDD Surveys

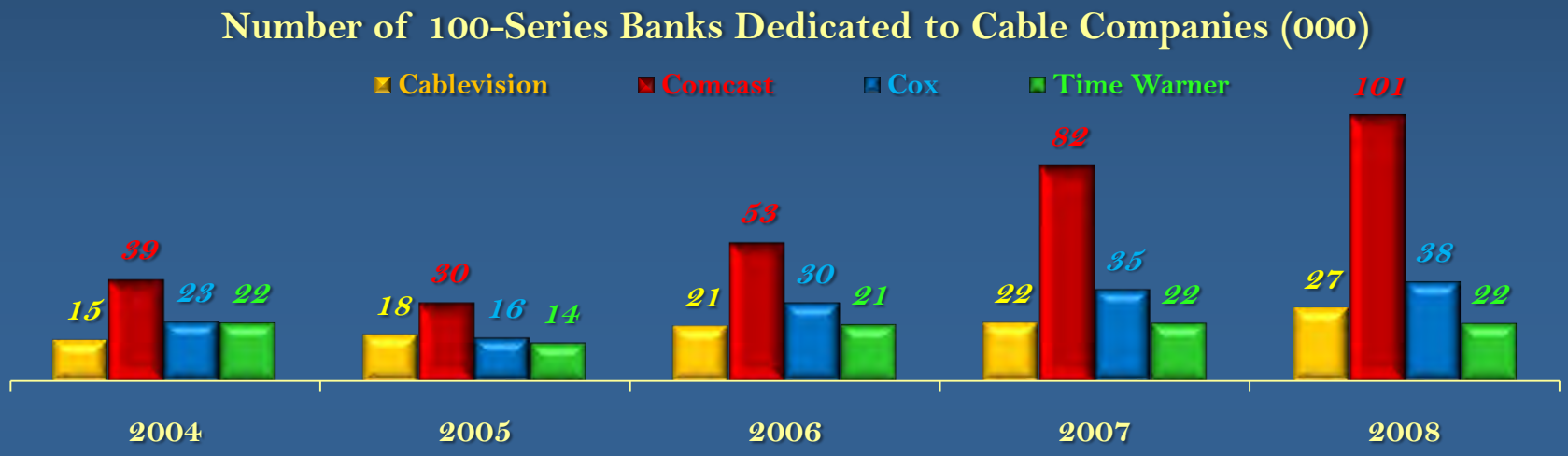
Number Porting

Estimated Ported Landline Numbers to Cellular by Bank Type (000,000)



Coverage Problems for RDD Surveys

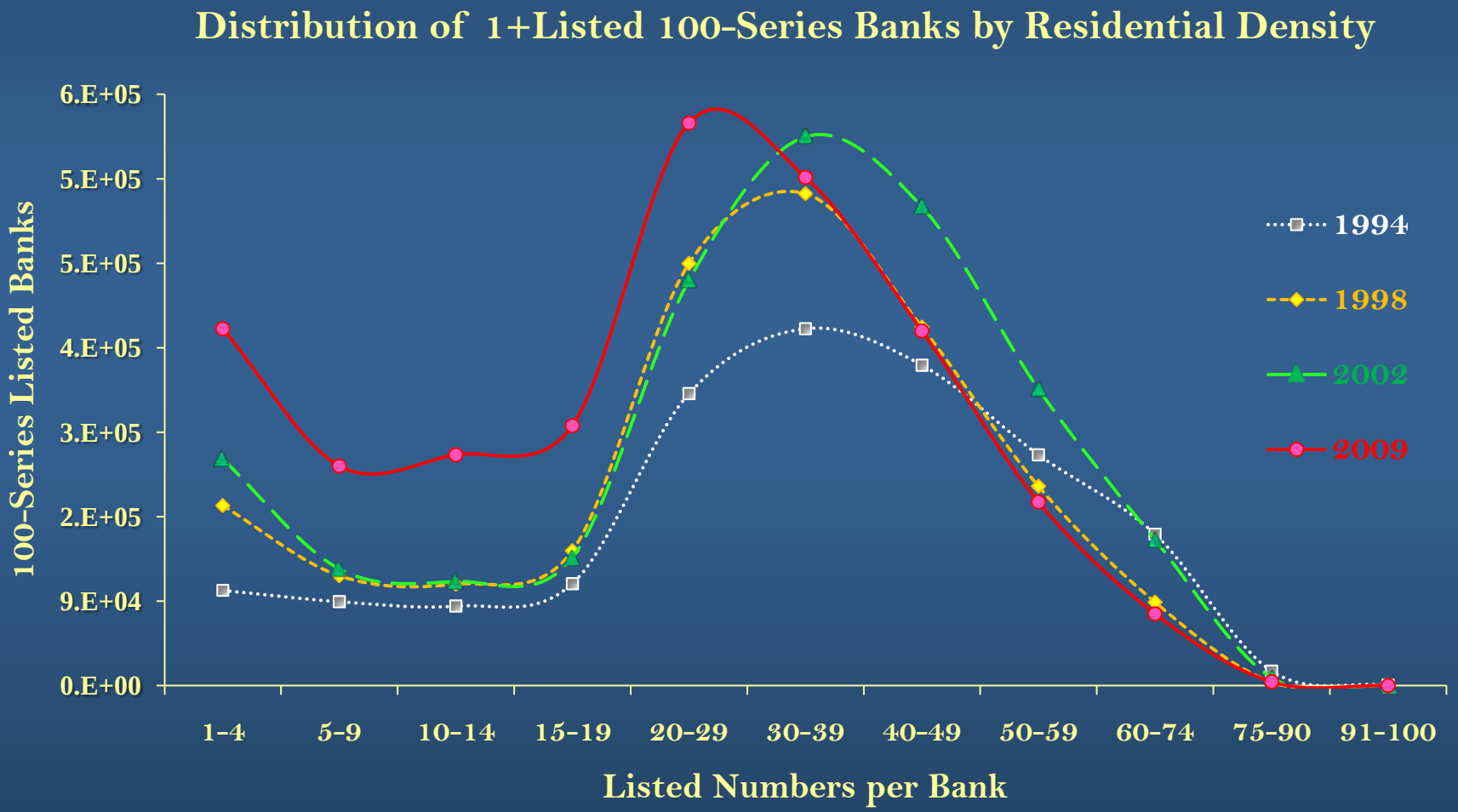
Emergence of Alternative Providers of Voice Services (Cable & VoIP)



- With VoIP one is no longer tied to local area codes and can choose a new number anywhere in the country
- www.voip-info.org shows 180+ residential VoIP providers in the US
- Magic-Jack has over 250,000 new subscribers every month ($20/year)
- "However, your telephone number will not be listed in any directory assistance or any public telephone directories."

Coverage Problems for RDD Surveys

Changes in Residential Density of 1+Listed 100-Series Banks

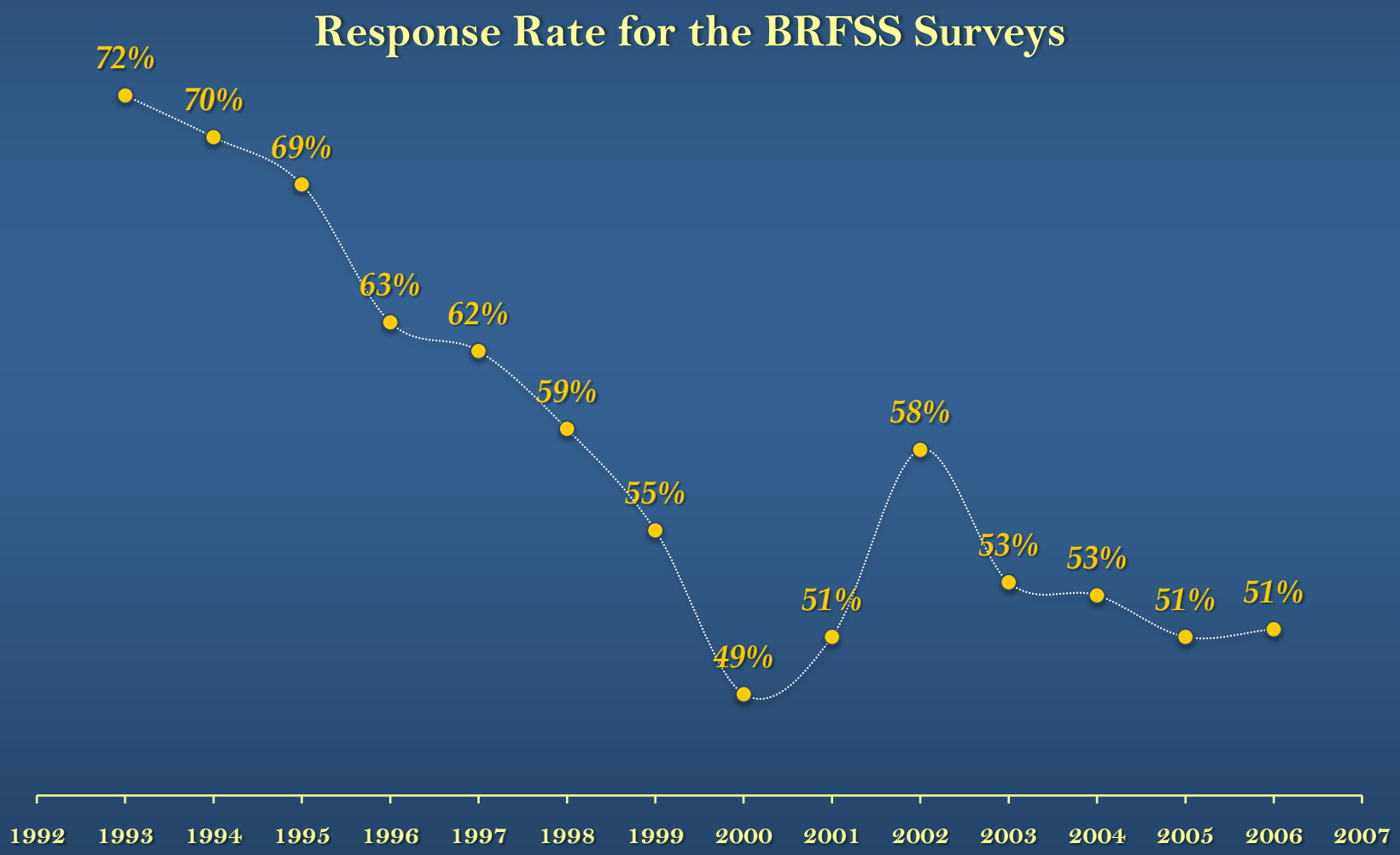


Coverage Problems for RDD Surveys

Supporting Studies

- A Reassessment of List-Assisted RDD Methodology
  - Mansour Fahimi, Dale Kulp, and J. Michael Brick
  - *Public Opinion Quarterly, Vol. 73, No. 4, 2009, pp. 751–760*
- Topology of the Landline Telephone Sampling Frame
  - Mansour Fahimi, Dale Kulp, and David Malarek
  - *Survey Practice, December 2009*

Eroding Rates of Response to Single Modes of Contact



Need for More Flexible & Innovative Methods of Survey Administration

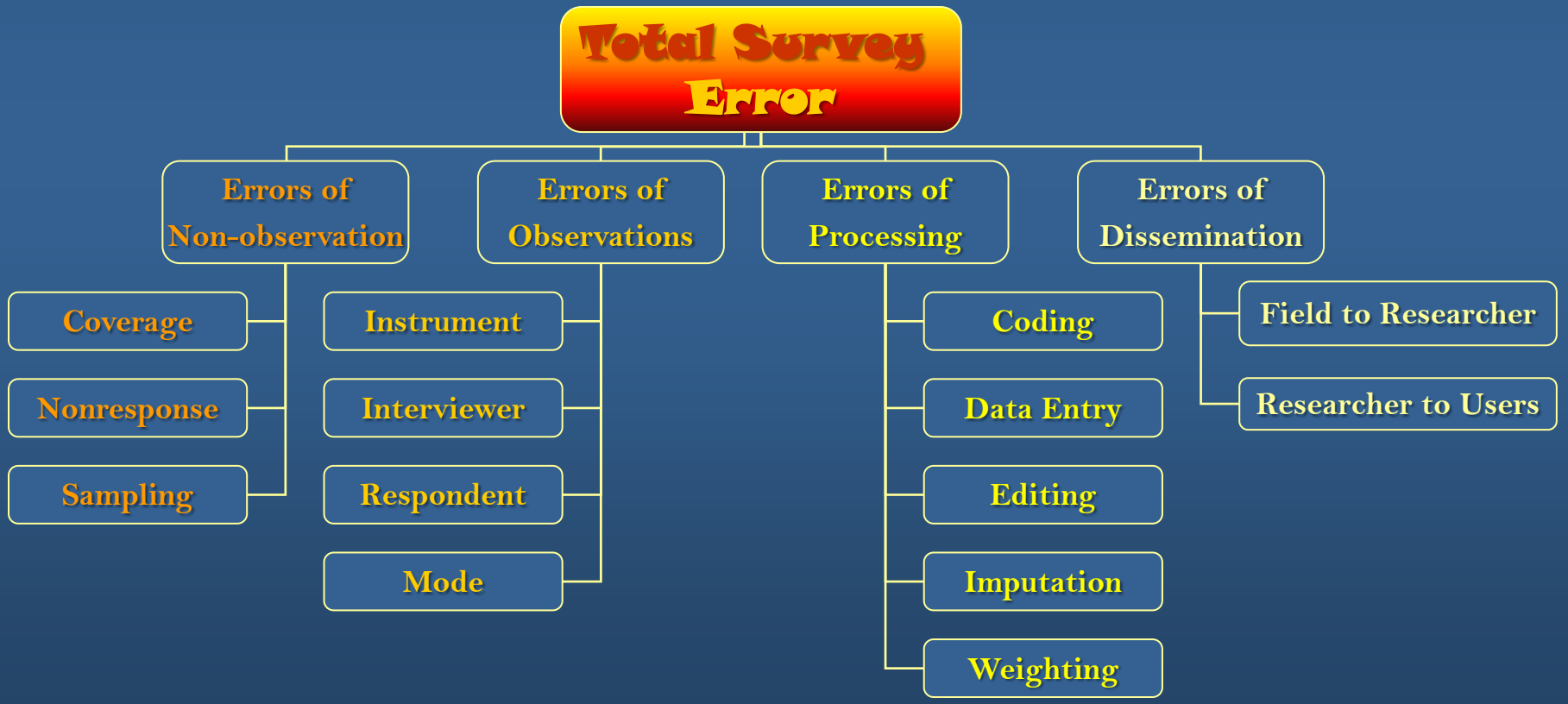
- Researchers are struggling with the "old" methods of survey administration:
  - Evolving coverage problems of telephone surveys
  - Growing rates of nonresponse to single mode methods
  - Messy weighting strategies for dual-frame methods
- Multi-mode methods are gaining popularity because different modalities can be combined effectively to:
  - Improve coverage
  - Boost response rates
  - Reduce cost
- Addressed-based sample designs provide a convenient framework for multi-mode alternatives

Pros & Cons of Multi-Mode Alternatives

- Concerns about systematic differences when collecting similar data using different modes (Dillman 1996):
- Greater likelihood for socially desirable responses to sensitive questions in interviewer-administered surveys (Aquilino 1994)
- The rate of missing data tends to be higher when surveys are self-administered (Biemer 2003)
- House effects are higher with interviewer-administered surveys (Russell 2004)

Pros & Cons of Multi-Mode Alternatives

At the end of the day, is it feasible to untangle the convoluted interactions between the mode, interviewer, respondent, and survey content? (Voogt & Saris 2005)



Improvements in Databases of Household Addresses

- With 135 million addresses the second generation of CDSF is the most complete address database
- By providing the most current delivery information and improved address hygiene this system helps reduce cost and improve efficiency by:
  - Reducing the number of undeliverable-as-addressed mailings
  - Increasing the speed of delivery
- Given daily feedback from thousands of letter carriers the database is updated on a nearly continuous basis

Available Data Items on the CDSF

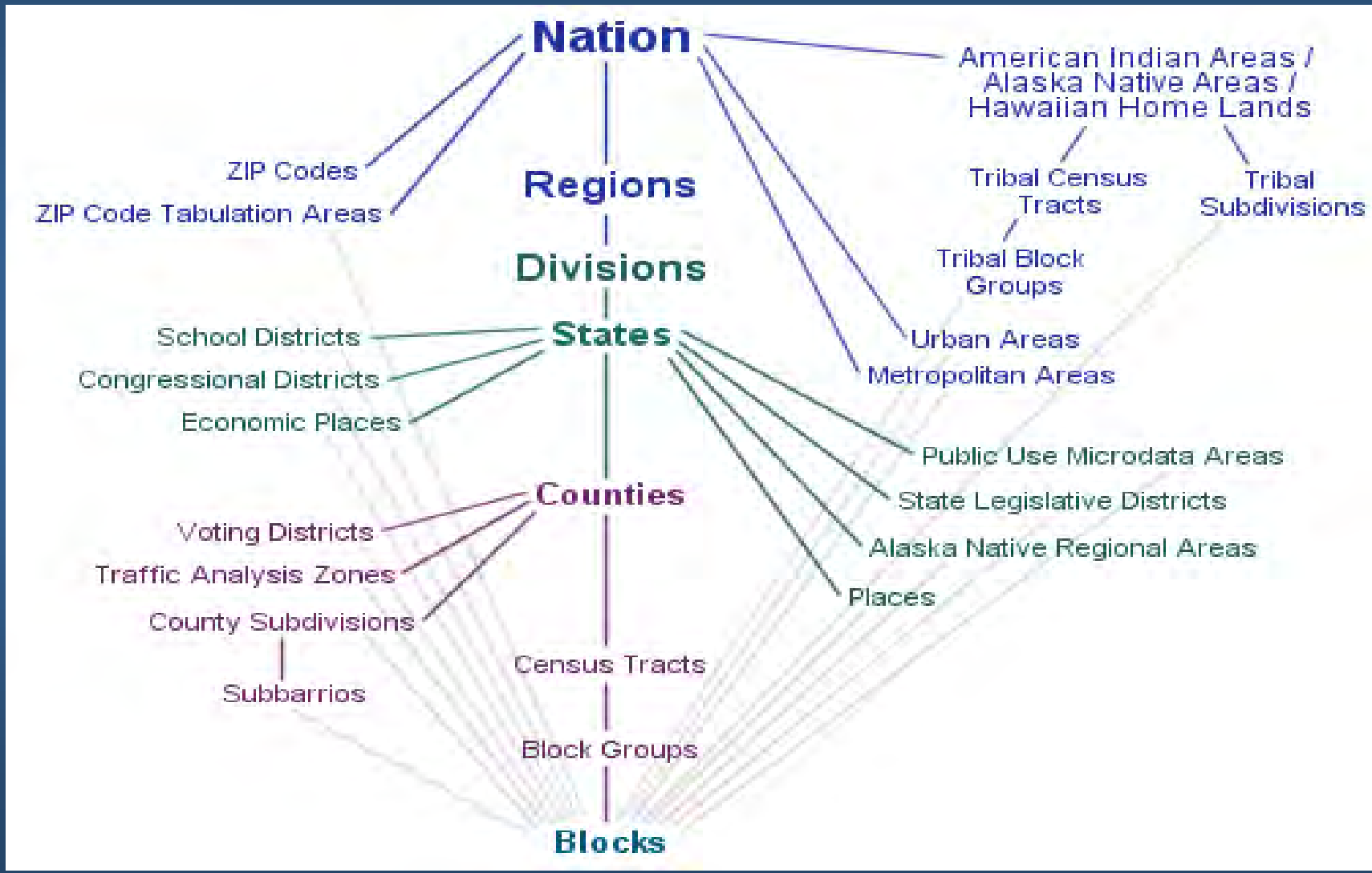
Example of Delivery Information

- House Number
- Apt Number
- Street Name
- Street Suffix: Ave and Blvd
- Directional: NE and W
- Zip
- Zip+4
- City Name
- City Code
- State Code
- State Name
- Tract
- Block
- County
- Walk Sequence Number
- Route Type
- Delivery Type Code
- Vacant Code
- Seasonal Code
- Drop Count
- PO Box
- Secondary Unit Descriptor

Using the CDSF for Sample Surveys
How?

- Start with an address-based sample down to ZIP+4:
  - Stratified or random across the entire domain
  - Clustered in an area probability fashion for efficient in-person attempts
- Initial contacts can be by phone and/or mail and include attempts for:
  - Survey administration at the point of initial contact
  - Recruitment for participation via other modes
- Once contact has been established follow-up attempts can take place in any order or combination of modes

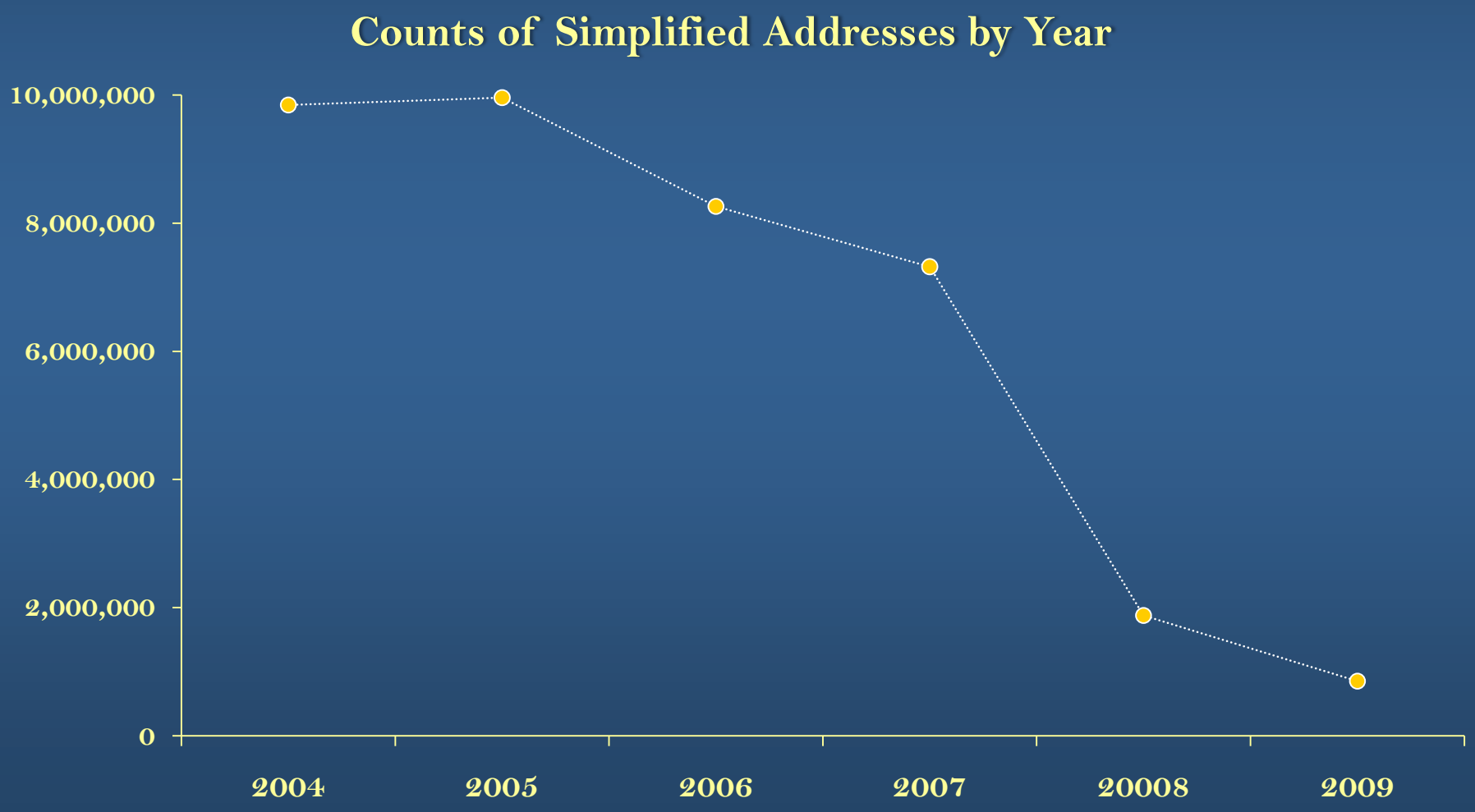
Topology of the Census Geography



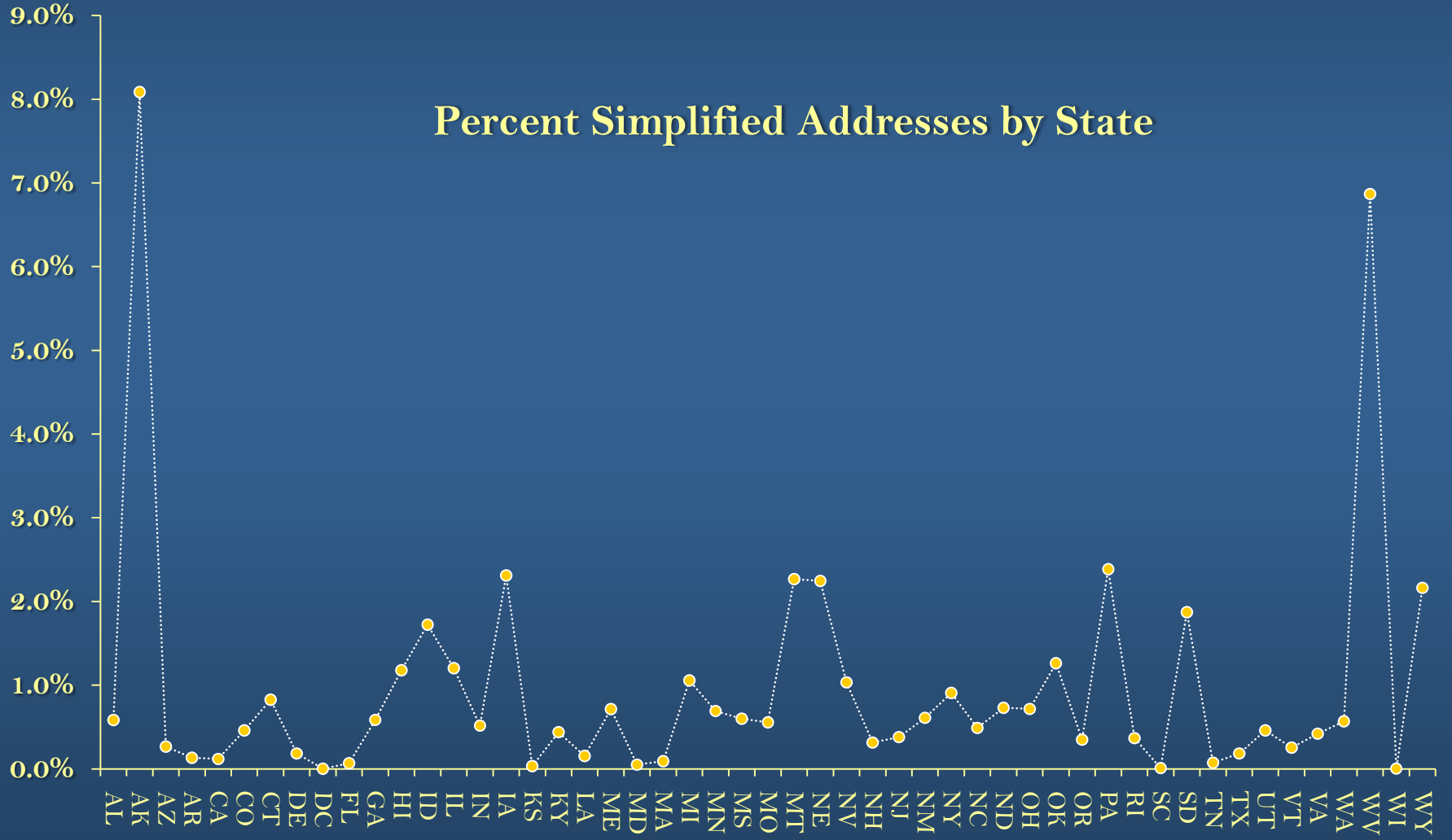
Potential Issues When Using the CDSF for Sampling

- The "raw" CDSF is for delivery not suitable for complex surveys
- Virtually all list suppliers simply offer basic extracts from the CDSF without any enhancements
- CDSF does not include effective stratification variables
- Certain households have a higher likelihood of not being included as a delivery point on the CDSF (Simplified):
  - The coverage rate diminishes in areas where home delivery of mail is unavailable (Staab & Iannacchione 2003)
  - List of on-site enumerated addresses differs from those from the CDSF in rural areas (Dohrmann & Mohadjer 2006)
  - A minor source of non-coverage is due to households that request that their addresses not be sold (O'Muircheartaigh 2003)

Potential Issues When Using the CDSF for Sampling



Potential Issues When Using the CDSF for Sampling



Possible Enhancements of the CDSF
Appending Information

- Geographic Information Enactments:
  - Census geographic domains
  - Marketing and media domains
- Demographic Information Enhancements:
  - Direct household data from commercial databases
  - Molded household statistics at various levels of aggregation
- Name and Telephone Number Retrievals:
  - Append a name associated with the address
  - Retrieve listed telephone number associated with the name
- Simplified Address Resolution

Possible Enhancements of CDSF
Simplified Address Resolution

- The CDSF provides only counts of addresses in simplified routes
- There are legitimate city-style addresses in simplified carrier routes known to commercial (very expensive) databases that can be added to CDSF:
  - Experian
  - Axiom
  - infoUSA
  - Targus

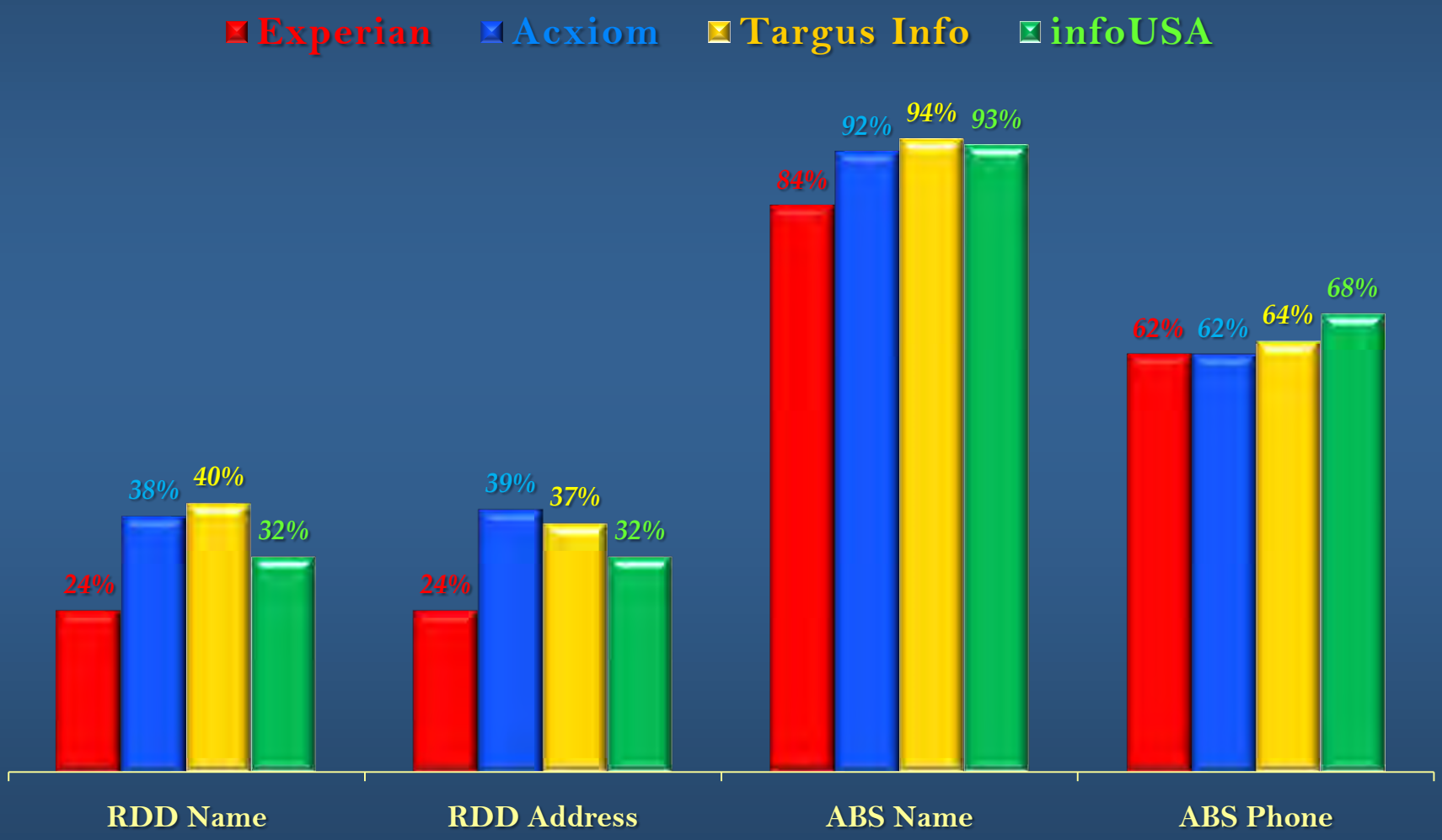
Possible Enhancements of CDSF

Resolution Summary for CDSF-Based Samples

- There are about 135 million residential addresses:
  - Simplified addresses account for 852,723 addresses
  - MSG can augment about 718,121 of simplified addresses
  - Augmented sampling frame covers over 99% of all residential addresses in the U.S.
- Percent name append on average is about 90 and more
- Percent phone append on average is about 65
- Match rates will vary with geography and inclusion of P.O. Boxes as they tend to drive down the rates

Possible Enhancements of CDSF

Match Rate Comparisons



Topology of the CDSF
Delivery Point Type Indicator

- **Business:** Indicates the delivery point is a business address
- **Central:** The delivery point is serviced at a mail receptacle located within a centralized unit
- **CMRA (Commercial Mail Receiving Agency):** A private business that acts as a mail-receiving agent for specific clients
- **Curb:** The delivery point is serviced via motorized vehicle at a mail receptacle located at the curb
- **Drop:** A delivery point or receptacle that services multiple residences such as a shared door slot or a boarding house in which mail is distributed internally by the site
- **Educational:** Identified as an educational facility such as colleges, universities, dormitories, sorority or fraternity houses, and apartment buildings occupied by students

Topology of the CDSF
Delivery Point Types

- **NDCBU (Neighborhood Delivery Collection Box Unit):** Services at a mail receptacle located within a cluster box
- **No-Stat:** Indicates address is not receiving delivery and is not counted as a possible delivery point for various reasons
- **Seasonal:** Receives mail only during a specific season and the months the seasonal addresses are occupied are identified
- **Throwback:** Address associated with this delivery point is a street address but the delivery is made to a P.O. Box address
- **Vacant:** Was active in the past, but is currently vacant (in most cases unoccupied over 90 days) and not receiving delivery

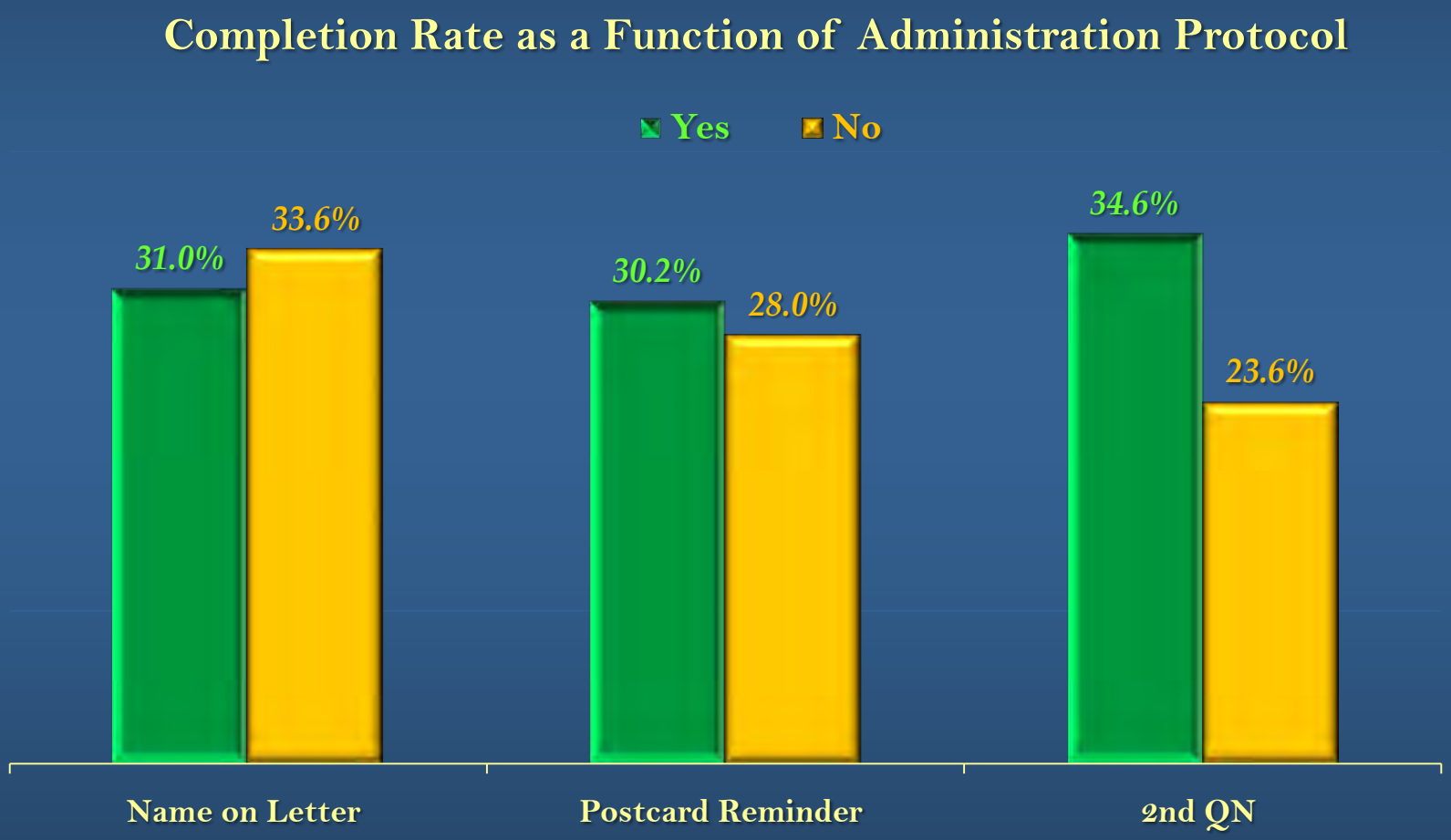
Topology of the CDSF
Delivery Point Type Counts

Residential Delivery Type	Count
Curb	51,670,749
Cluster Box Unit (CBU)	15,682,109
Central	19,952,644
Other	32,149,836
Facility Box	14,574,511
Contract Box	156,168
Detached Box	41,440
Non-Personnel Unit	33,568
Total	**134,261,025**
Seasonal	893,276
Vacant	6,972,657
Educational Off Campus Residential	112,130
Reachable Only via P.O. Boxes	5,195,679
Reachable Only via P.O. Boxes (Not Vacant)	3,734,427

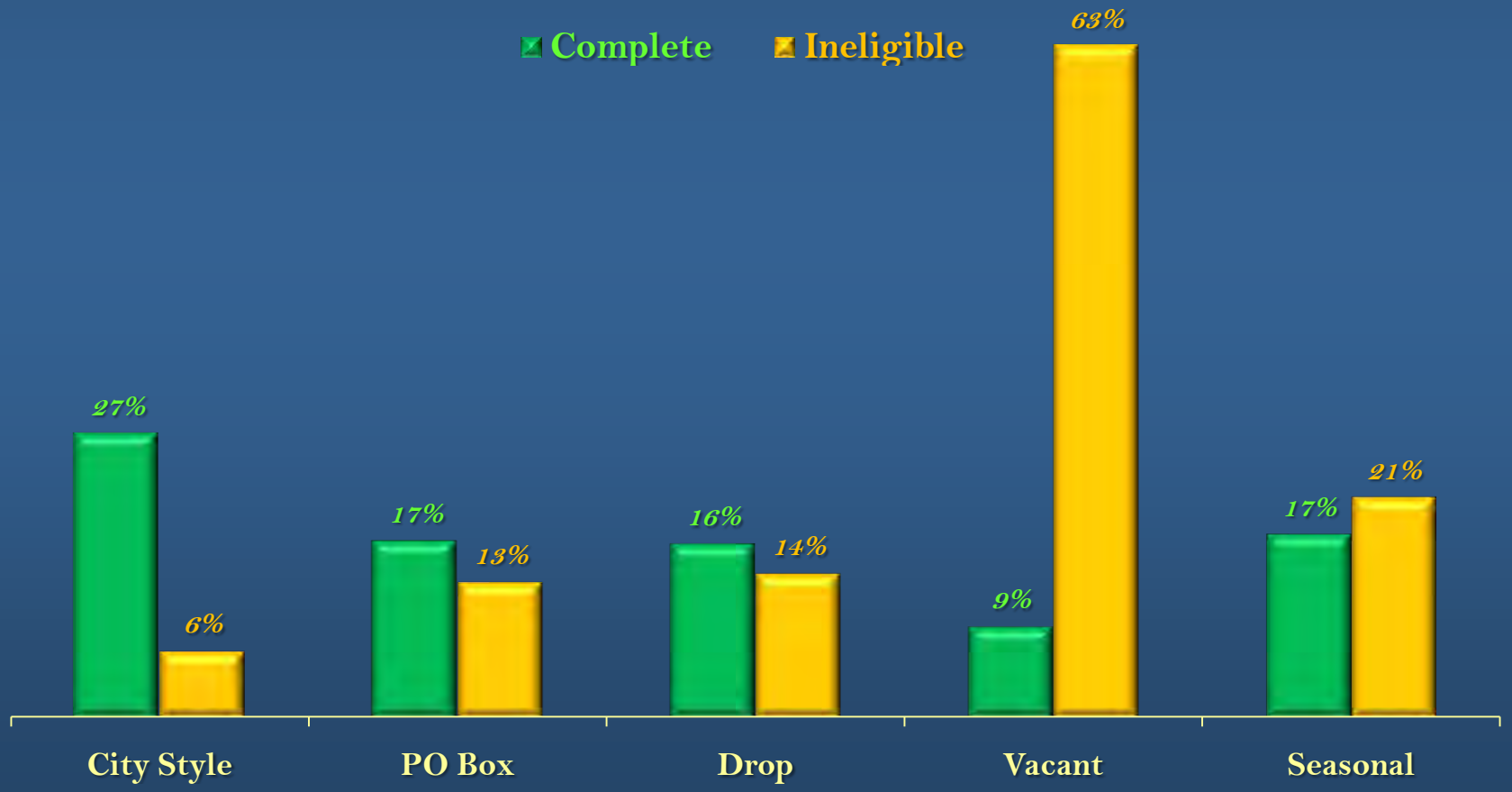
Closing Remarks

- Single-mode methods of data collection are problematic for response rate, coverage, and cost reasons
- Telephone surveys based on landline RDD samples are subject to non-ignorable coverage bias
- Multi-mode methods of data collection can reduce some of the problems associated with the conventional methods
- CDSF provides a natural and efficient framework for design and implementation of multi-mode surveys
- Enhancing the CDSF can significantly improve its coverage and expand its utility for design and analytical applications

Field Statistics
BRFSS Experience

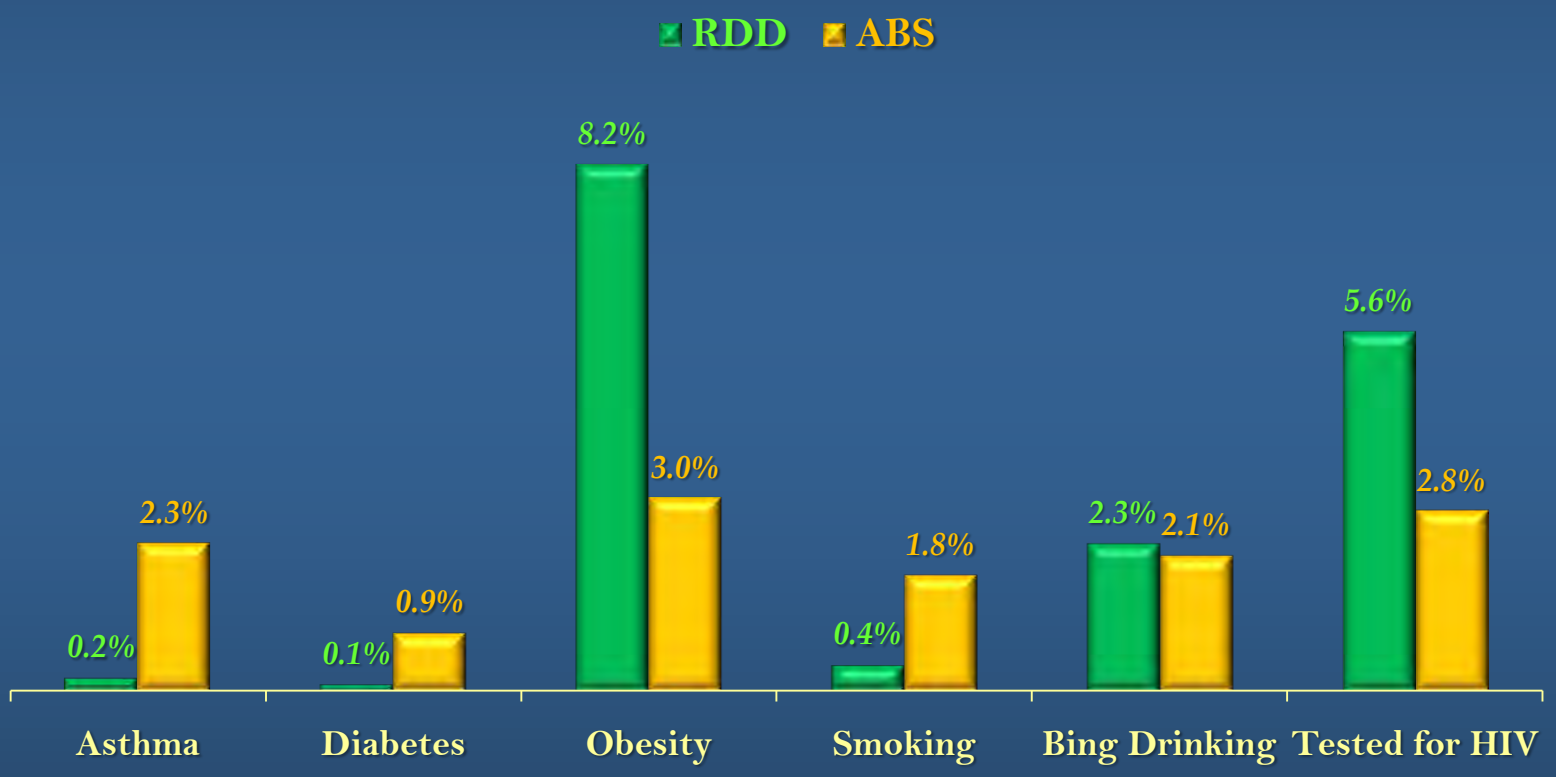


Field Statistics
BRFSS Experience



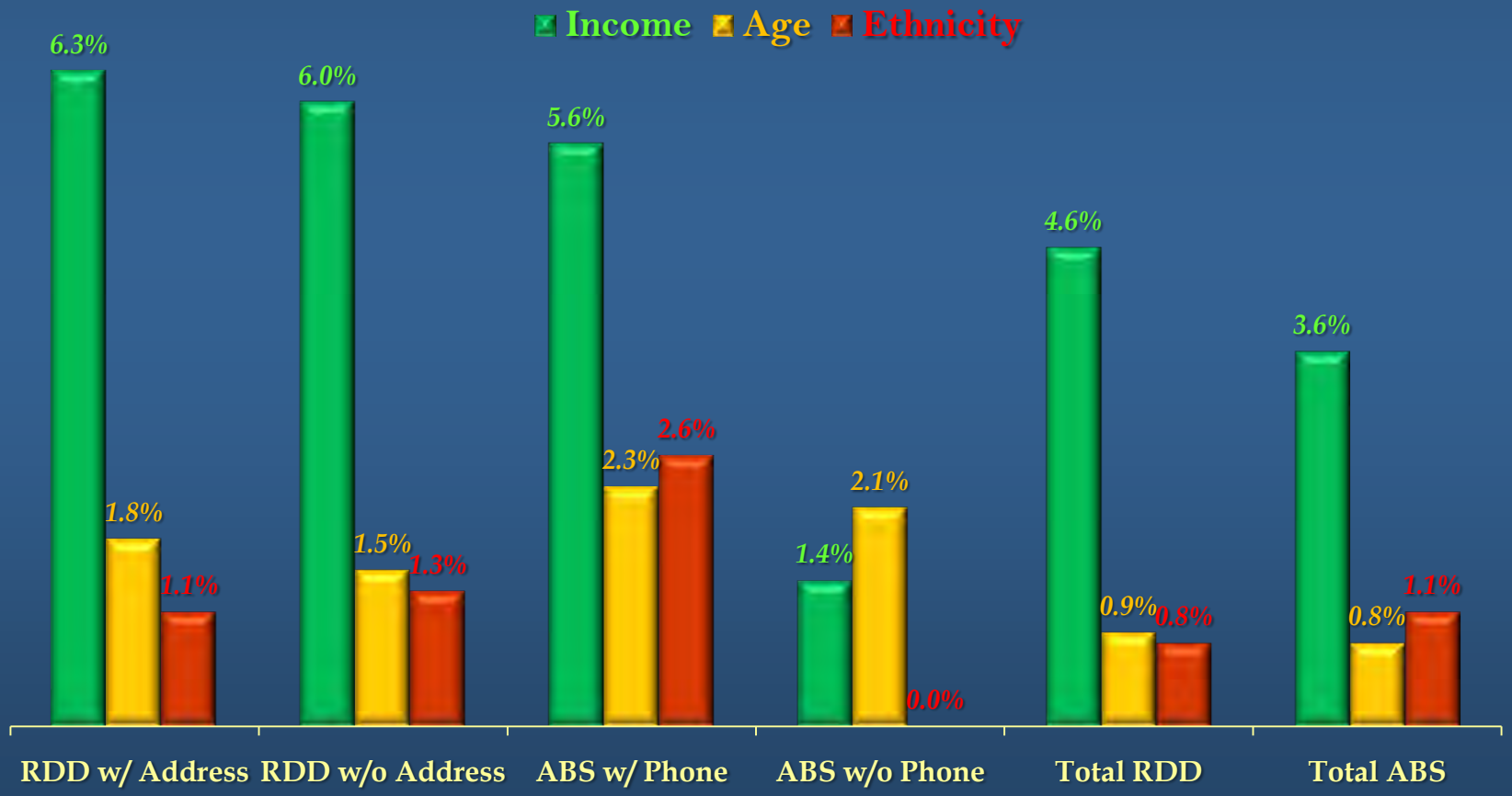
Field Statistics
BRFSS Experience

Item Nonresponse by Mode



Field Statistics
Massachusetts Health Interview Survey 2008

Item Nonresponse Rates by Household Type



Field Statistics

Massachusetts Health Interview Survey 2008