The Total Growth of Open Source

Amit Deshpande, Dirk Riehle
SAP Research, SAP Labs LLC

OSS 2008
September 8, 2008
Milan, Italy

Motivation

- Understand growth of open source
  - Growth of individual projects is known (linear, quadratic)
  - But what about the ***total growth of all of open source***

- Derive model for growth prediction
  - Determine model of growth
  - Where are we in the adoption curve?

- Were will we end up?
  - In 2006, nominally, at 0.7% of total packaged software market
  - Will open source take 10%? 30%? 70%? 100%? of the market?

- ***All of this, to gauge the significance of open source***

Expectations

- Some boundary conditions
  - Number of software developers is limited, not growing beyond limits
  - Any given software developer can only do a finite amount of work
- Thus, whatever current growth, there is an obvious limit
  - May not be relevant at current (small?) penetration of open source
  - Limit itself may be expanding
- We expect an S-curve (sigmoidal curve)
  - Not sure where on the curve
  - Expect eventual asymptotic approximation of limit
  - Not sure whether linear, quadratic, exponential, other growth

Measuring Growth

- Two measures
  - Amount of work going into open source (code contributions)
  - Number of projects being worked upon
  - Both measures should obviously be closely correlated
- Number of projects
  - Counting the number of active working open source projects
- Amount of work
  - Measured as the size of code contributions in SLoC
  - Constrained to projects that are active and working

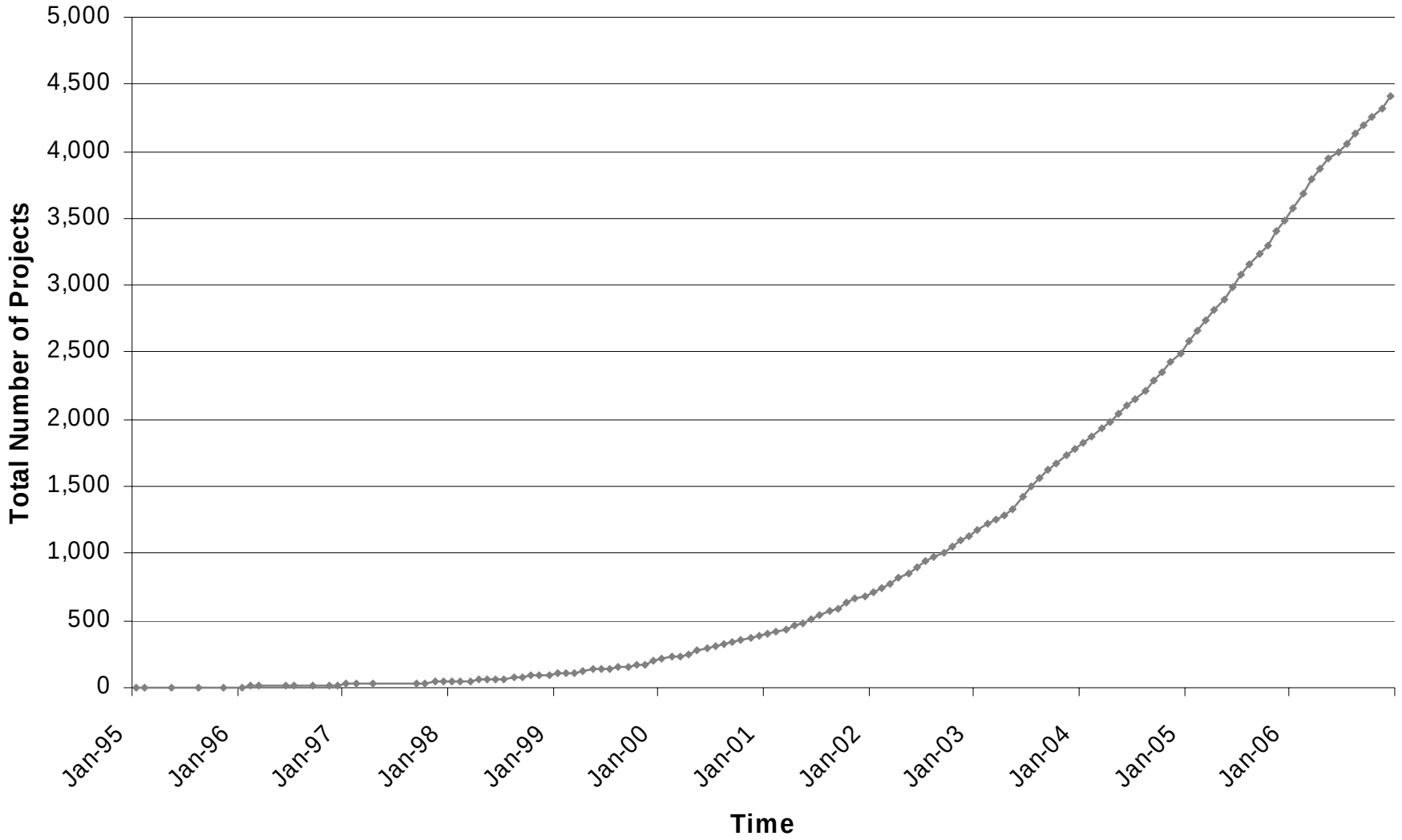
Database and Approach

- Database snapshot (May 2007) from Ohloh, Inc.
  - See http://ohloh.net, also available through API
  - Also see http://labs.ohloh.net for SLoC counter (diff)

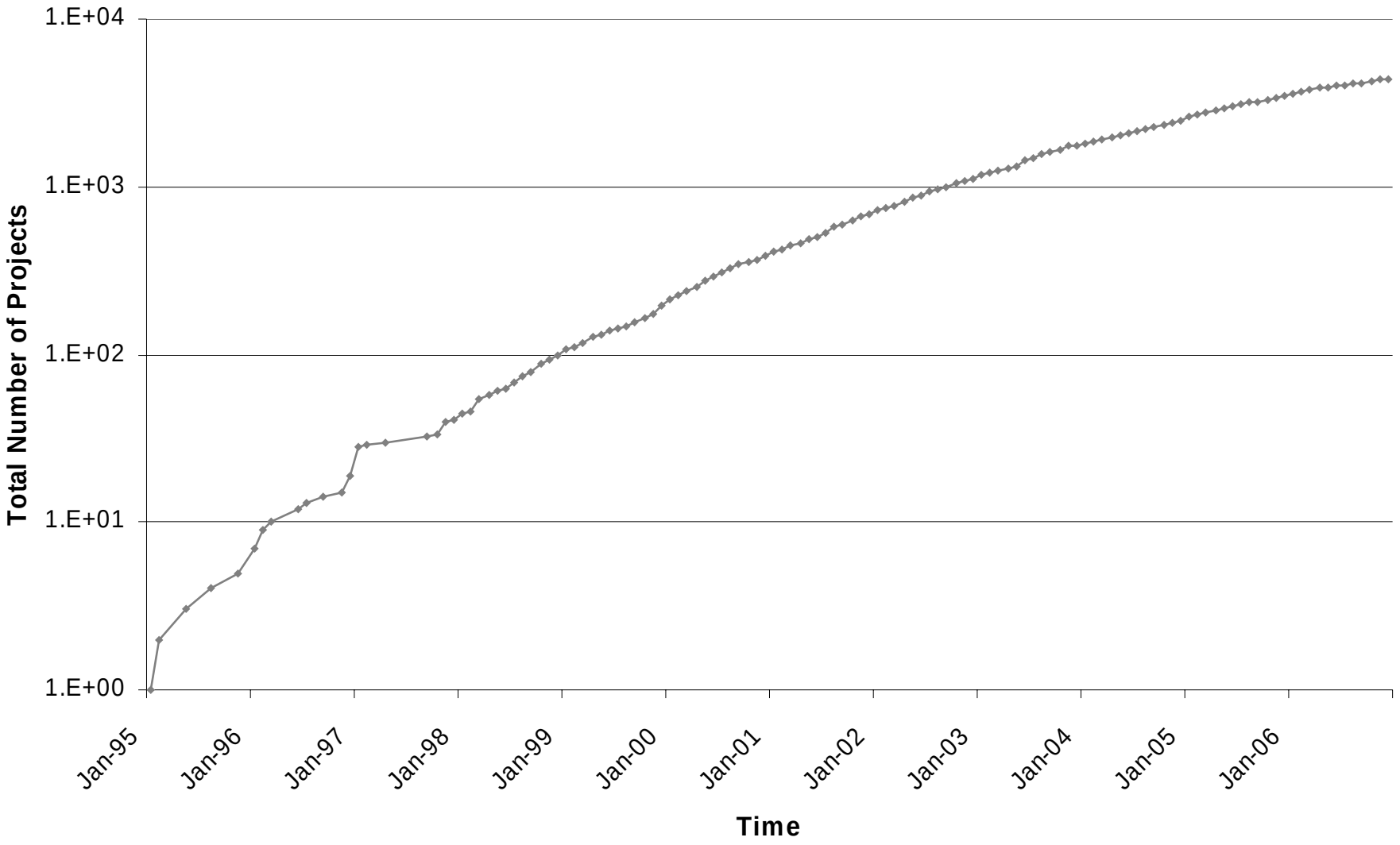
- Projects in database in this snapshot added by Ohloh
  - Pre-selected 5000 most popular projects using Yahoo! in-links
  - After this, community edited (reducing data quality)

- Provides detailed analysis of these projects
  - Including full configuration management system history (commits)
  - Project start date determined by date of first commit

Project Growth (Total)



Project Growth (Total, Semi-Log)



Project Growth (Total) Model

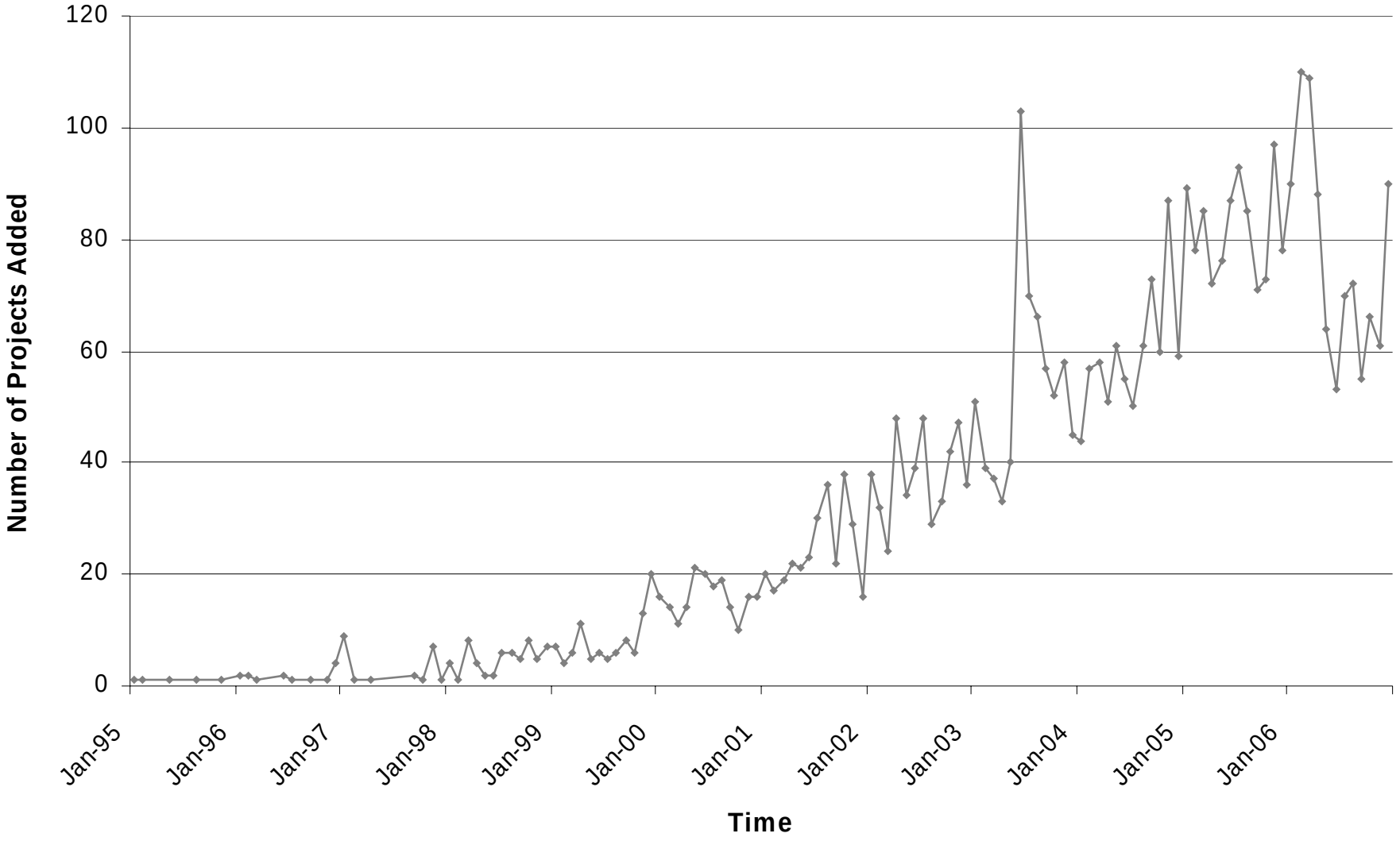
Model	R-square value
$y = 7.1511e^{0.0499x}$	0.956

where,

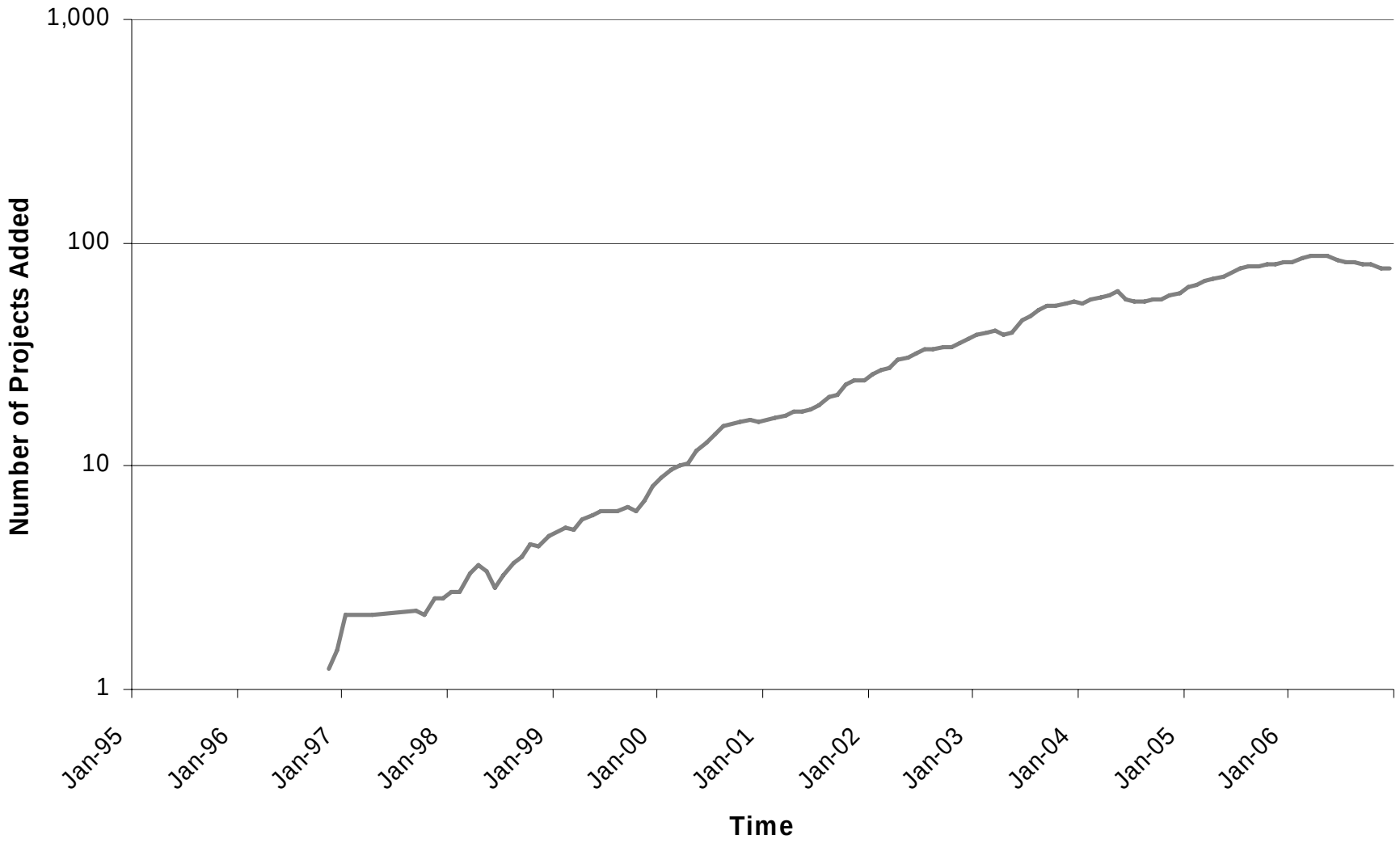
y: Total number of open source projects

x: Time from Jan 1995 to Dec 2006 in months

Project Growth (Added)



Project Growth (Added, Trend, Semi-Log)



Project Growth (Added) Model

Model	R-square value
$y = 1.0641e^{0.035x}$	0.884

where,

y: Total number of open source projects

x: Time from Jan 1995 to Dec 2006 in months

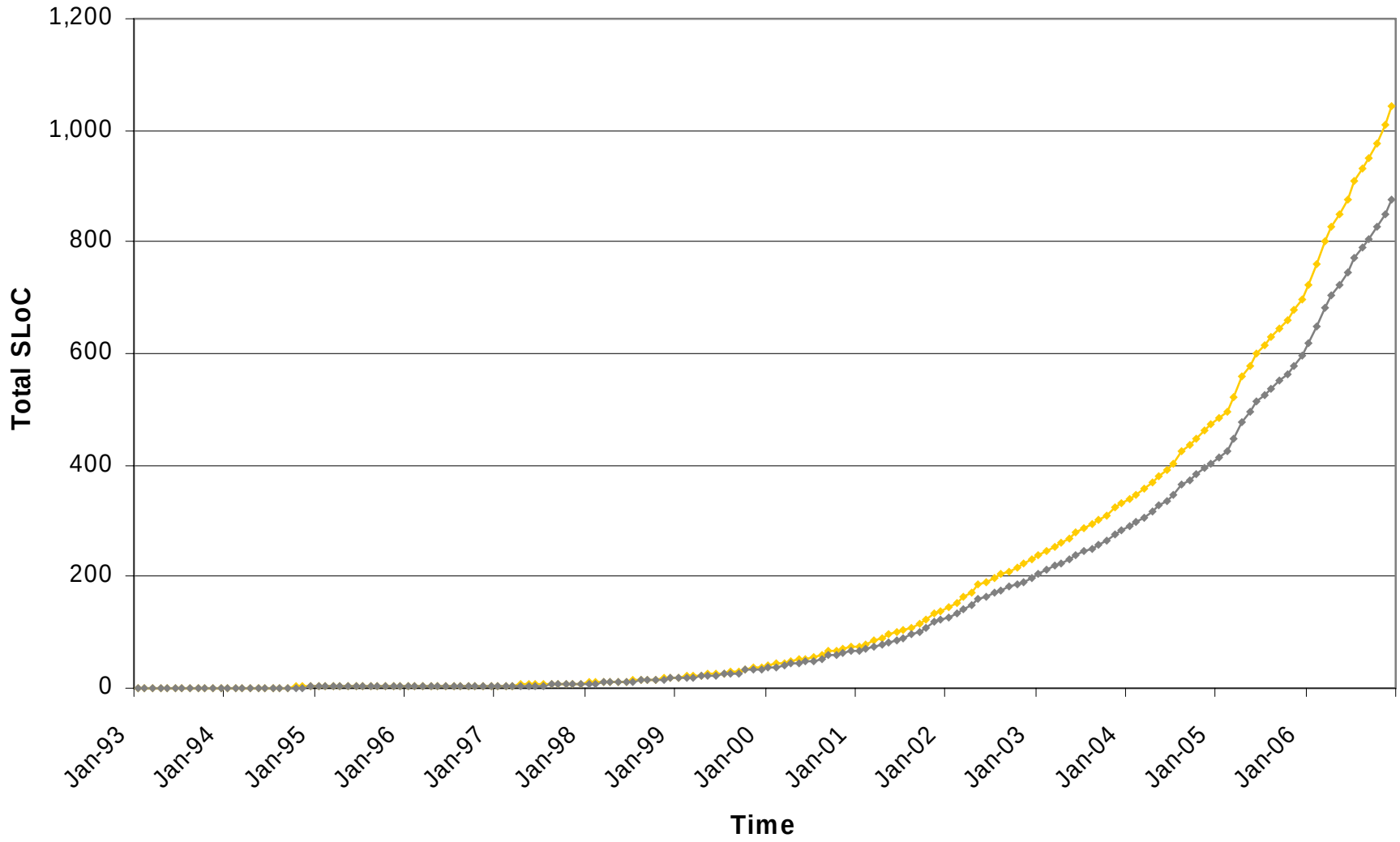
Calculating Commit Size

- Commit size to represent “amount of work performed”
  - Measured in source lines of code (no comments, empty lines)
  - Calculated using Unix/Ohloh diff tool
  - Database provides SLoC added and removed (not changed)

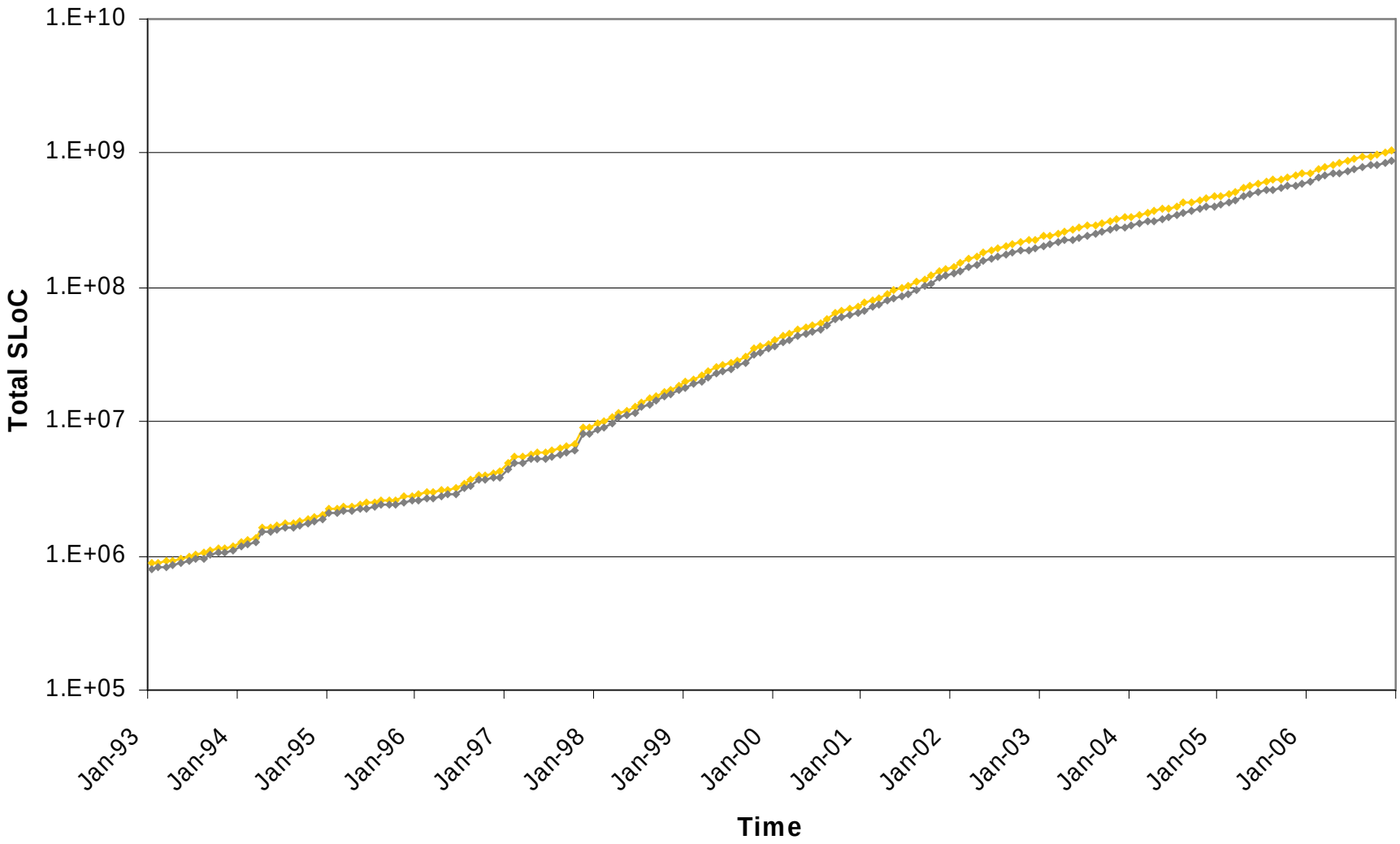
- Commit size upper limit
  - Adds SLoC added, subtracts SLoC removed, ignores copy and paste
  - Gracefully handles file moves, renames

- Commit size lower limit
  - Filters out commits beyond three times standard deviation (3060 SLoC)
  - This is to filter out copy and paste actions, additions of libraries, etc.
  - Then adds SLoC added, subtracts SLoC removed

SLoC Growth (Total)



SLoC Growth (Total, Semi-Log)



SLoC Growth (Total) Models

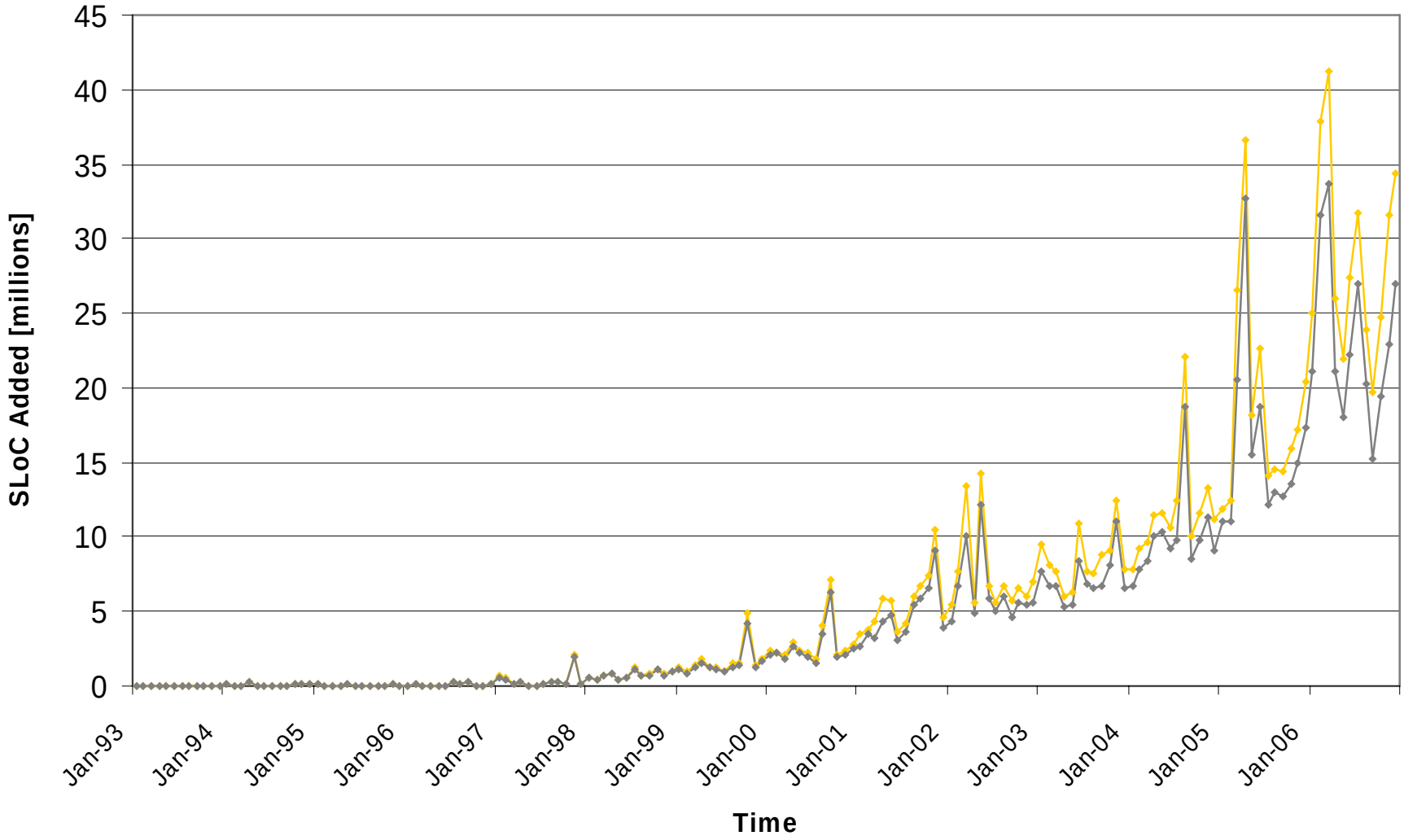
Approach	Model	R-square value
1	$y = 784098*e^{0.0555x}$	0.961
2	$y = 2E+06*e^{0.0464x}$	0.964

where,

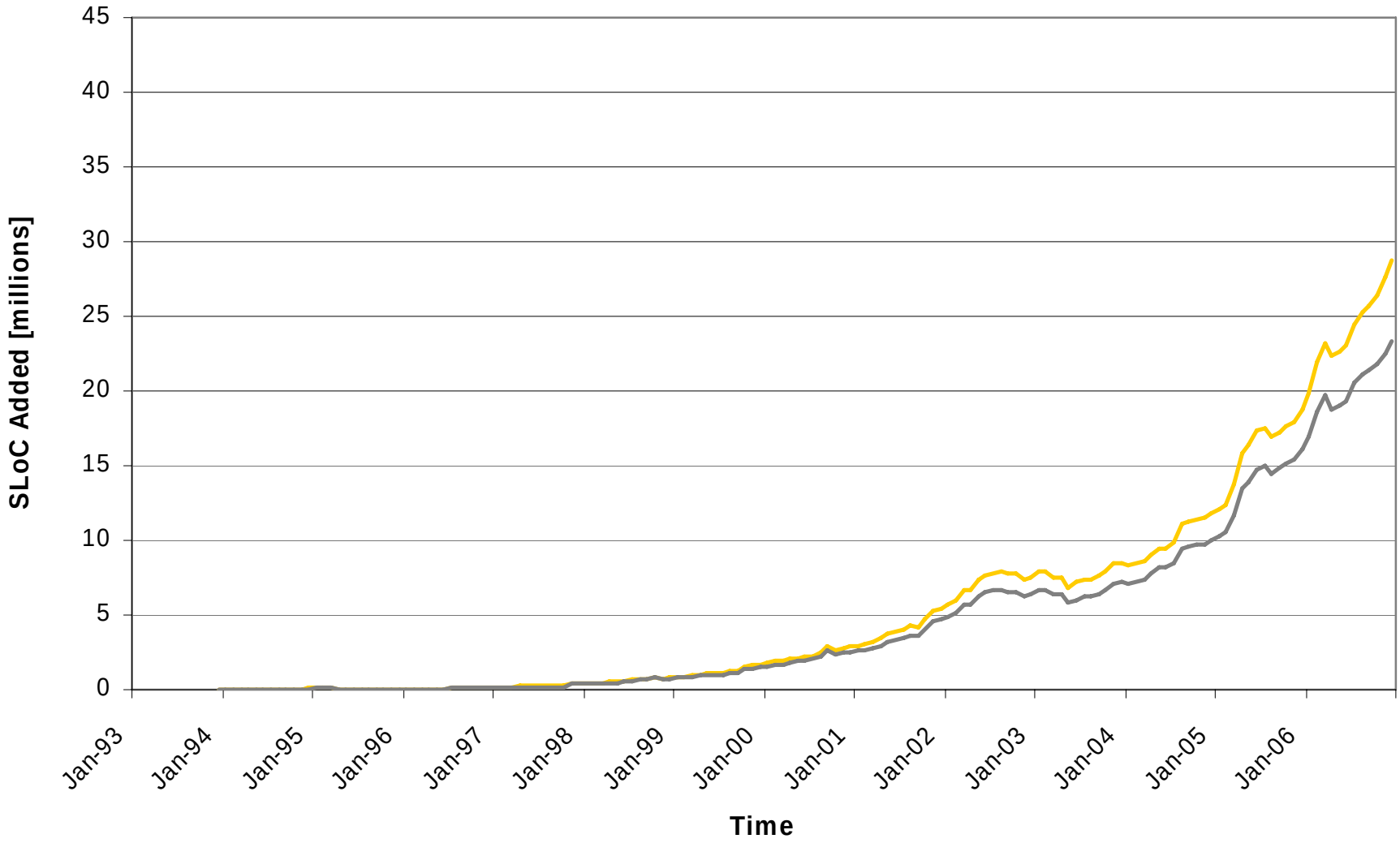
y: Total open source lines of code

x: Time from Jan 1995 to Dec 2006 in months

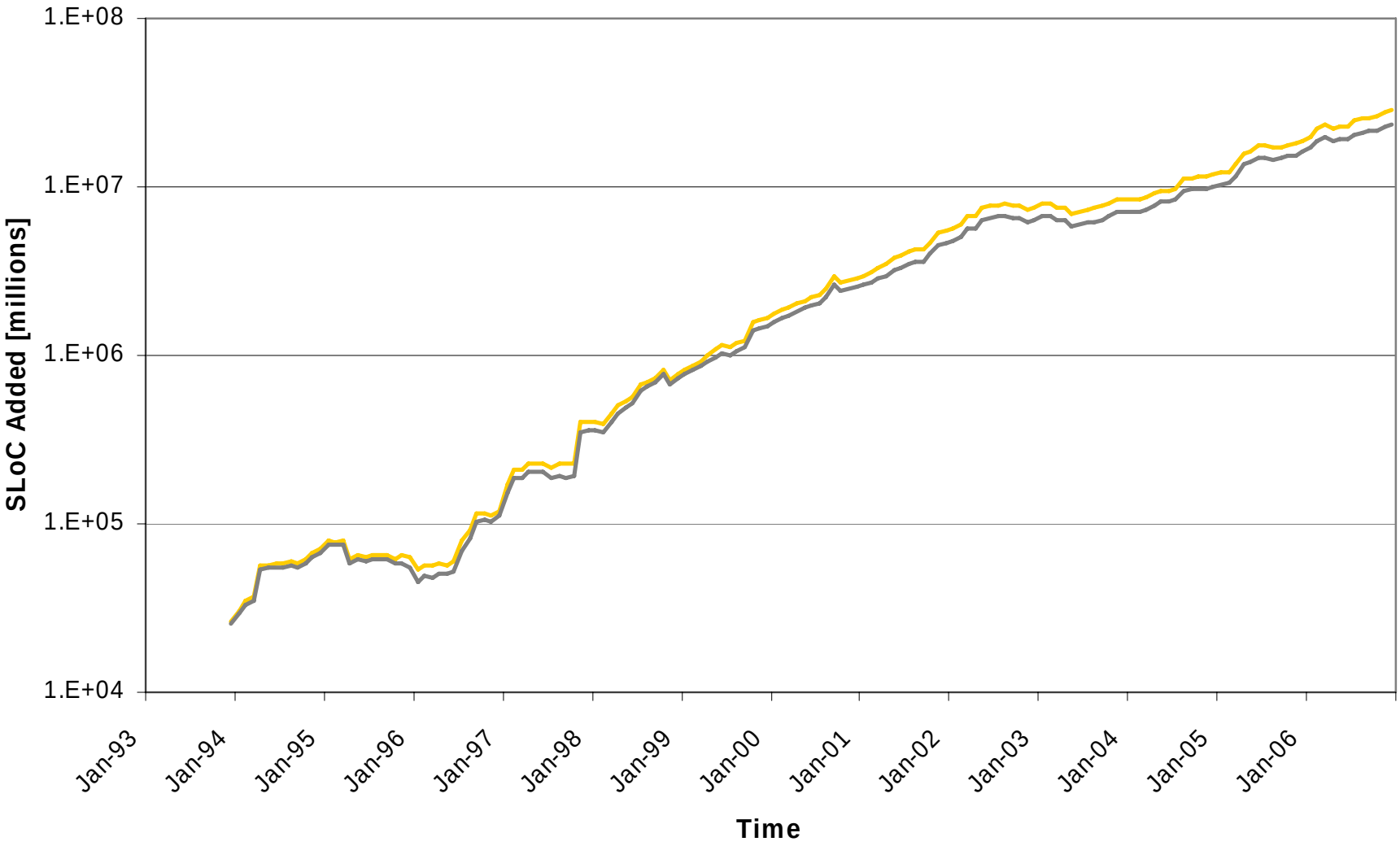
SLoC Growth (Added)



SLoC Growth (Added, Trend)



SLoC Growth (Added, Trend, Semi-Log)



SLoC Growth (Added) Models

Approach	Model	R-square value
1	$y = 70833*e^{0.0464x}$	0.901
2	$y = 64004*e^{0.046x}$	0.897

where,

y: Source lines of open source code added

x: Time from Jan 1995 to Dec 2006 in months

Threats to Validity

- Sample size? Cf. Daffara's 18,000 projects Aug 2007
- Some data has already been lost (old version history)
- Ohloh limitations, e.g. only CVS, SVN, Git
- Data quality issues with copy and paste, etc.

Conclusions

- Still early in the growth of open source
- Don’t know yet how close to s-curve inflection point
- Don’t know yet upper limit (further investigation)
- Clearly impressive continued growth in open source overall

The 2008 International Symposium on Wikis

Sept 8-10, 2008, Porto, Portugal

http://www.wikisym.org/ws2008

Thank you!