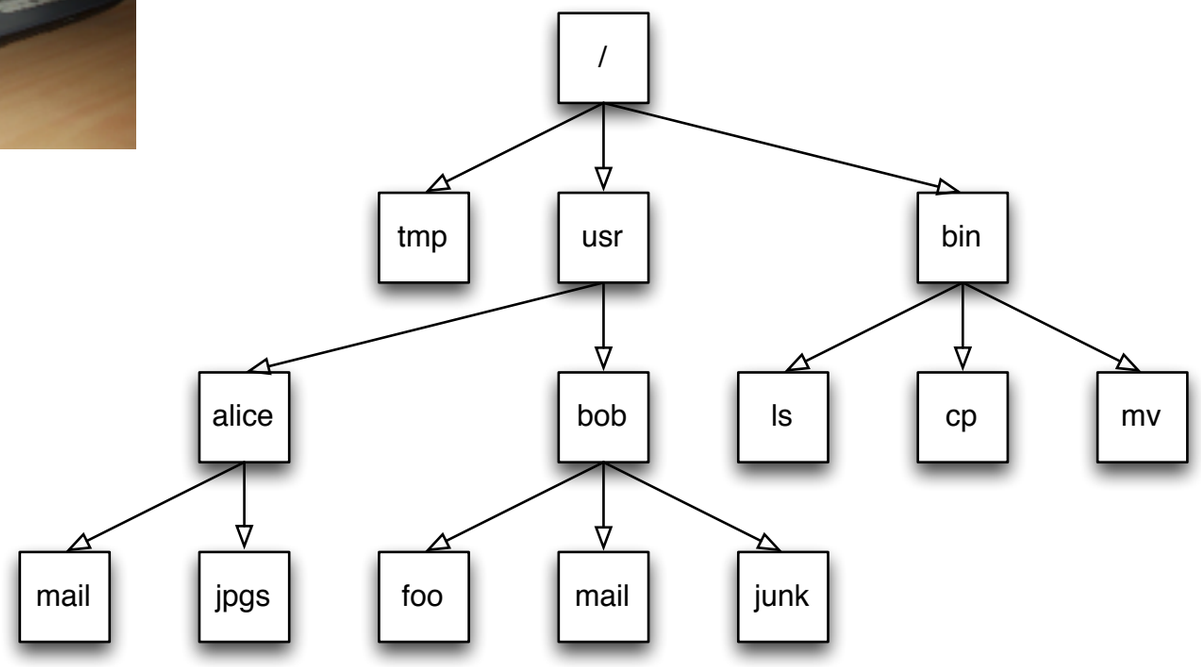
Measuring Systematic and Random Error in Digital Forensics

Alex J. Nelson, Simson L. Garfinkel
NIST

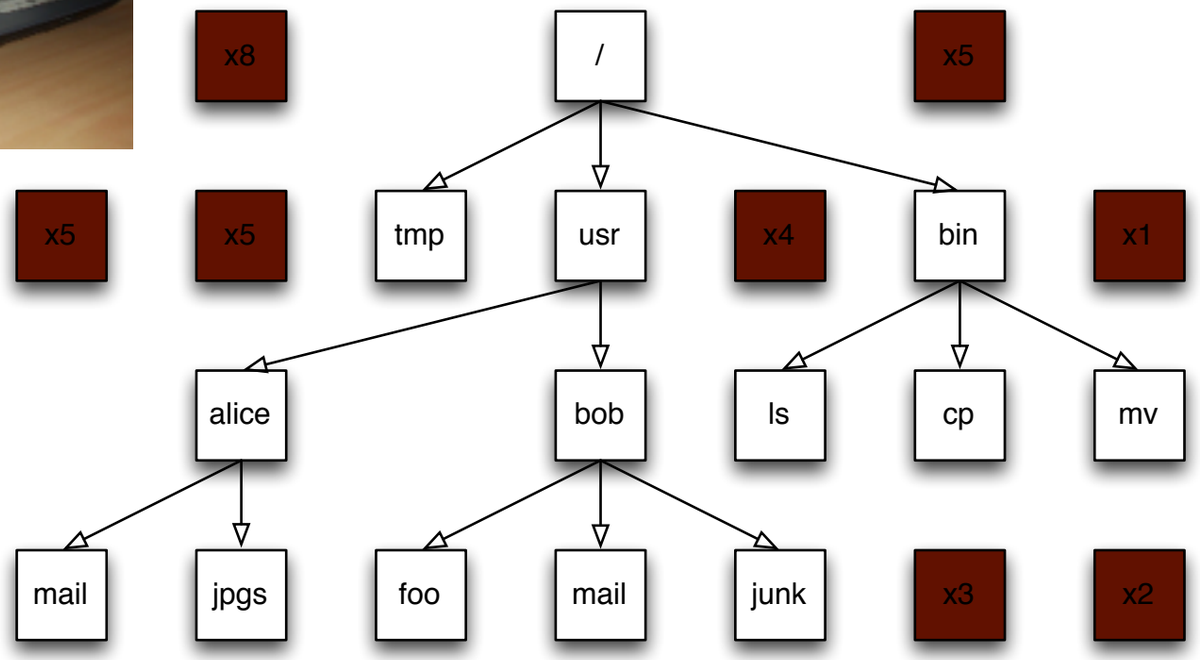
Forensic Science Error Management
July 23, 2015

Computer systems organize mass storage into files.



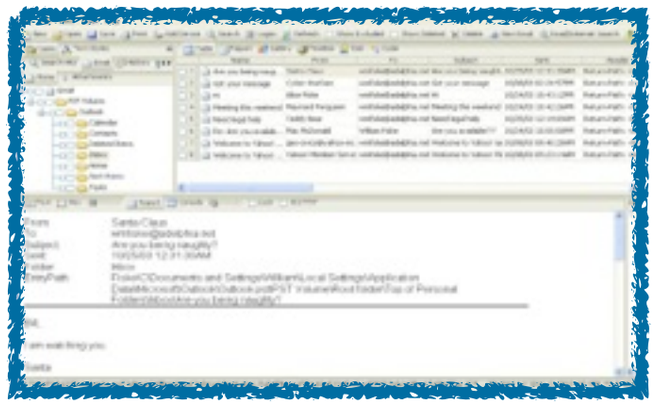
Computer systems only show *allocated files*.

When files are deleted, they remain on the computer.

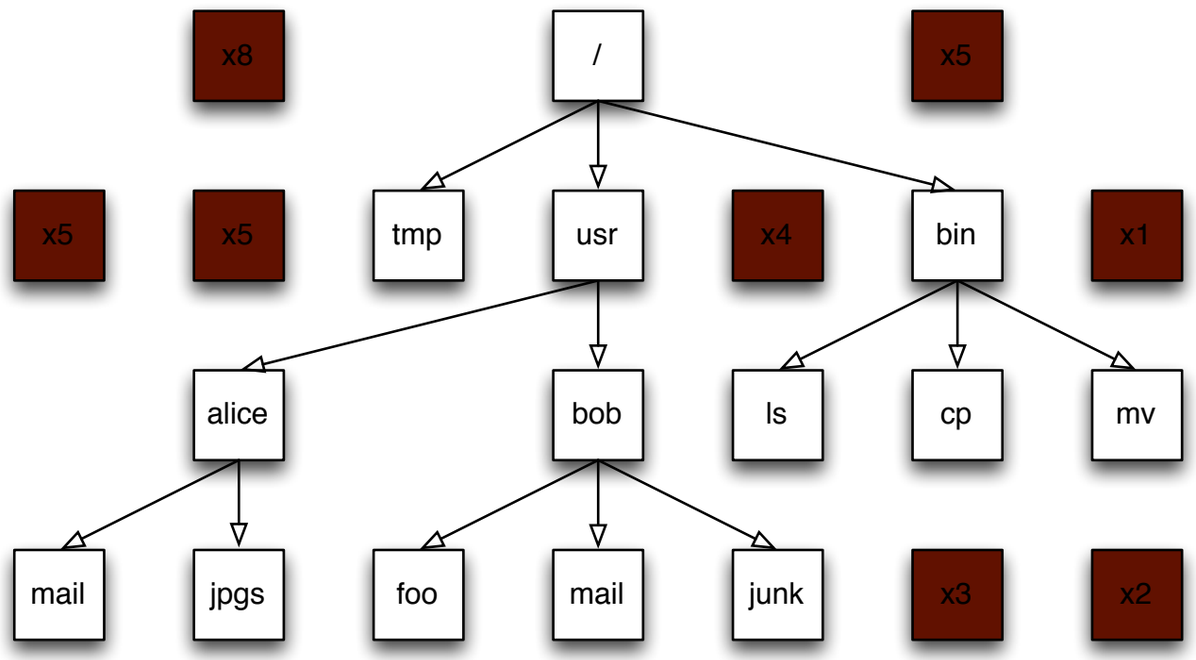


A primary task of digital forensics is recovering deleted files.

Digital forensics tools extract *allocated* and *deleted* files from mass storage device.



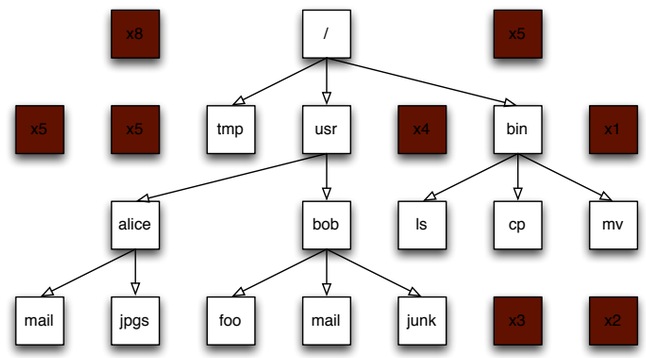
Typical computers have millions of files.



This talk discusses two sources of error when extracting files from digital media.

Problem #1: Verifying the “deleted” files extraction.

- Was the data actually in a deleted file?
- Is the extracted file complete?
- Is the extracted file corrupted?
- What was the extracted file’s name?



—Error sources: ambiguity in handling of deleted data; tool error.

—Solution: examining multiple tools for inter-tool agreement.

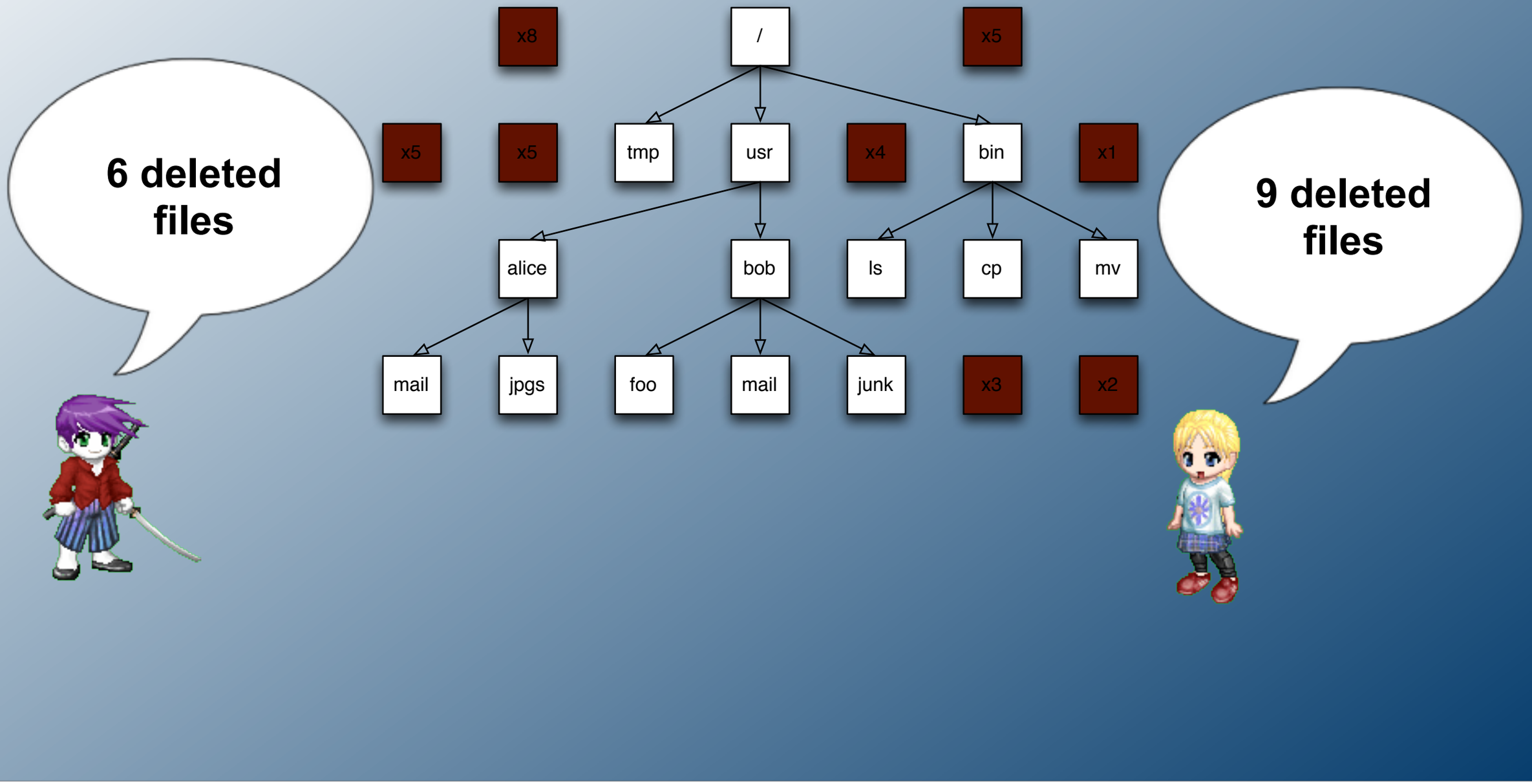
Problem #2: Determining the “owner” of deleted files.

- (If a computer was used by multiple people.)

—Error source: incomplete information for deleted files.

—Solution: statistical machine learning to create a model for the users of each drive.

Magenta **Yellow**



Comparative Metadata Verification

Data on hard drives can be divided into three categories:

Allocated Files

Deleted Data

No Data

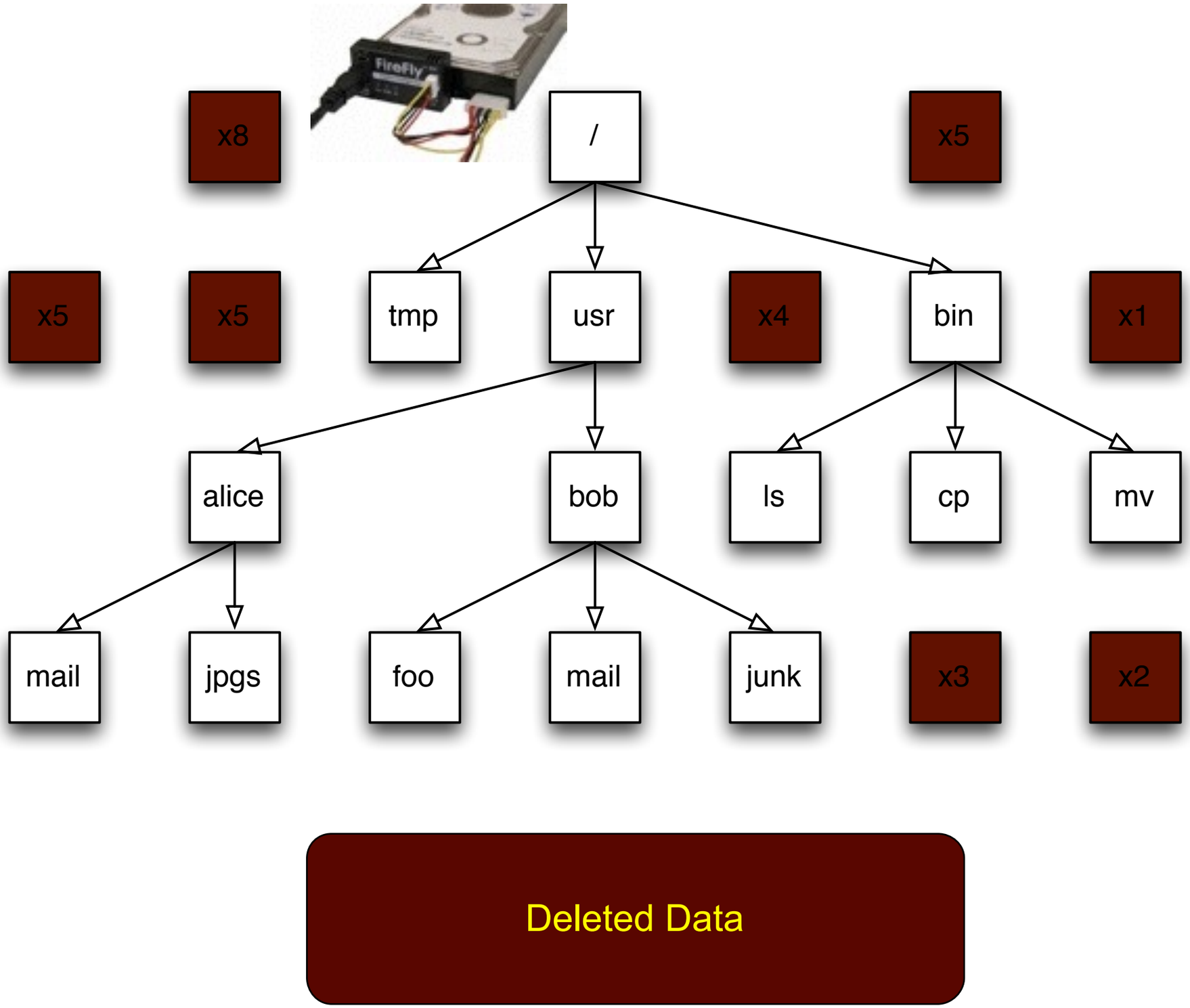
user files
email messages
[temporary files]

blank sectors

Computer *file systems* organize bytes, into disk blocks, into files.

Allocated Files

When files are deleted, their data may remain. Forensic tools can recover this data.



File systems are provided with all operating systems.

Microsoft Windows: FAT32, EXFAT, NTFS

Apple Macintosh: FAT32, EXFAT, NTFS, HFS+

Android: FAT32, YAFFS2, EXT4

XBOX: FAT32, FATX, XTAF

Challenges:

- Vendor file systems do not recover deleted files.
- Different vendors implement file systems differently.
- Some file systems are not documented.

Vendors solve these challenges with reverse engineering.

- Reverse engineering is error prone.

We analyzed an XBOX 360 hard drive with 3 forensic tools and got 3 different results.

XBOX360 hard drive

	Tool #1: SleuthKit	Tool #2: py360	Tool #3: uxtaf
Partitions processed	5	6	4
Allocated directories	65	58	56
Allocated files	293	231	231
Unallocated directories	1	14	8
Unallocated files	2	15	11

Challenge: each tool output data in a different format.

Solution: normalize output to a consistent XML representation.

Error analysis

	Tool #1: SleuthKit	Tool #2: py360	Tool #3: uxtaf
Partitions processed	5	6	4
Allocated directories	65	58	56
Allocated files	293	231	231
Unallocated directories	1	14	8
Unallocated files	2	15	11

Each tool disagreement:

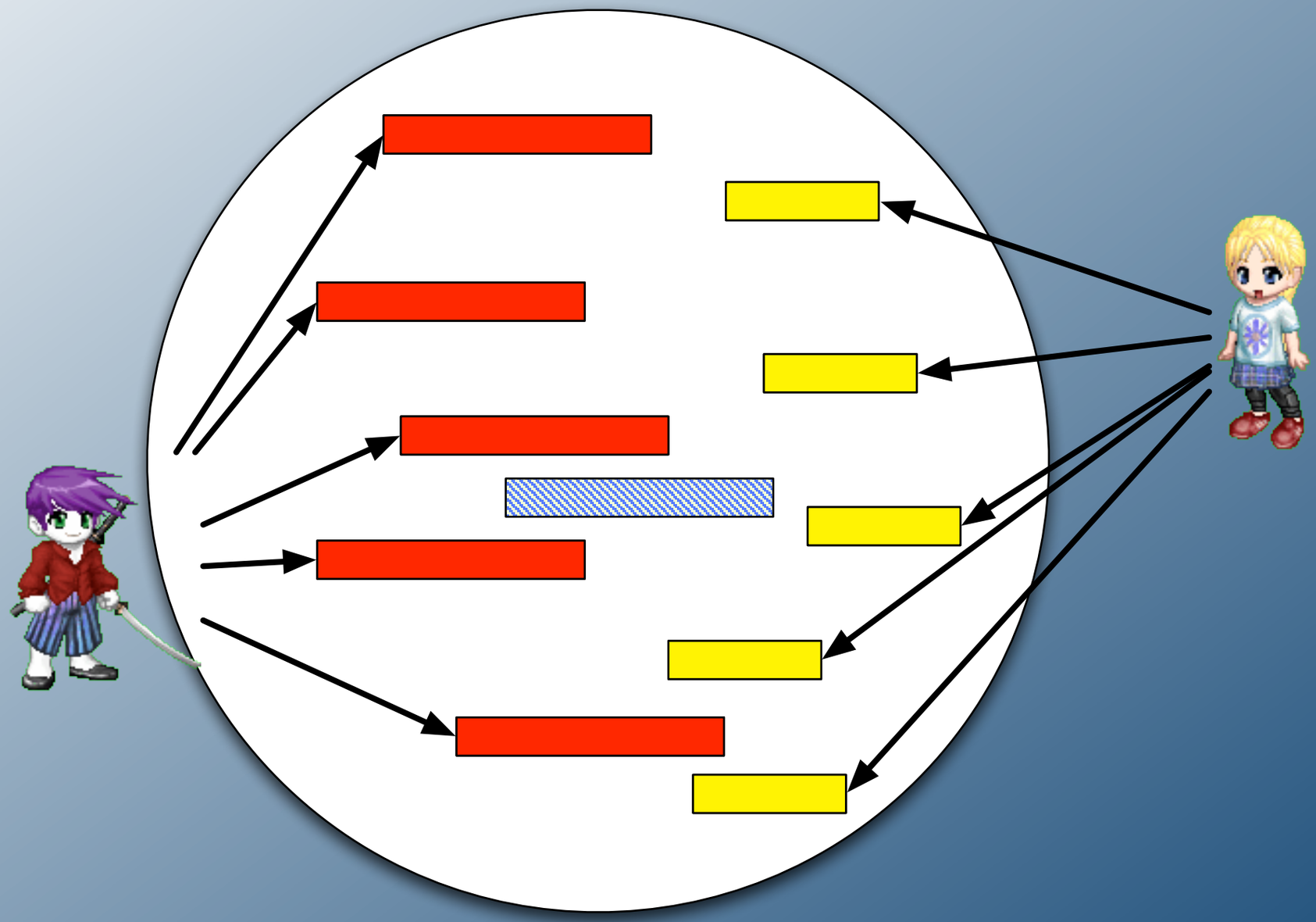
- An error in at least one tool.
- Could be an error in all tools!

Error is *systematic.*

- Might depend on the data.
- Tool always exhibits the same error with the same input.

Error can be accounted for:

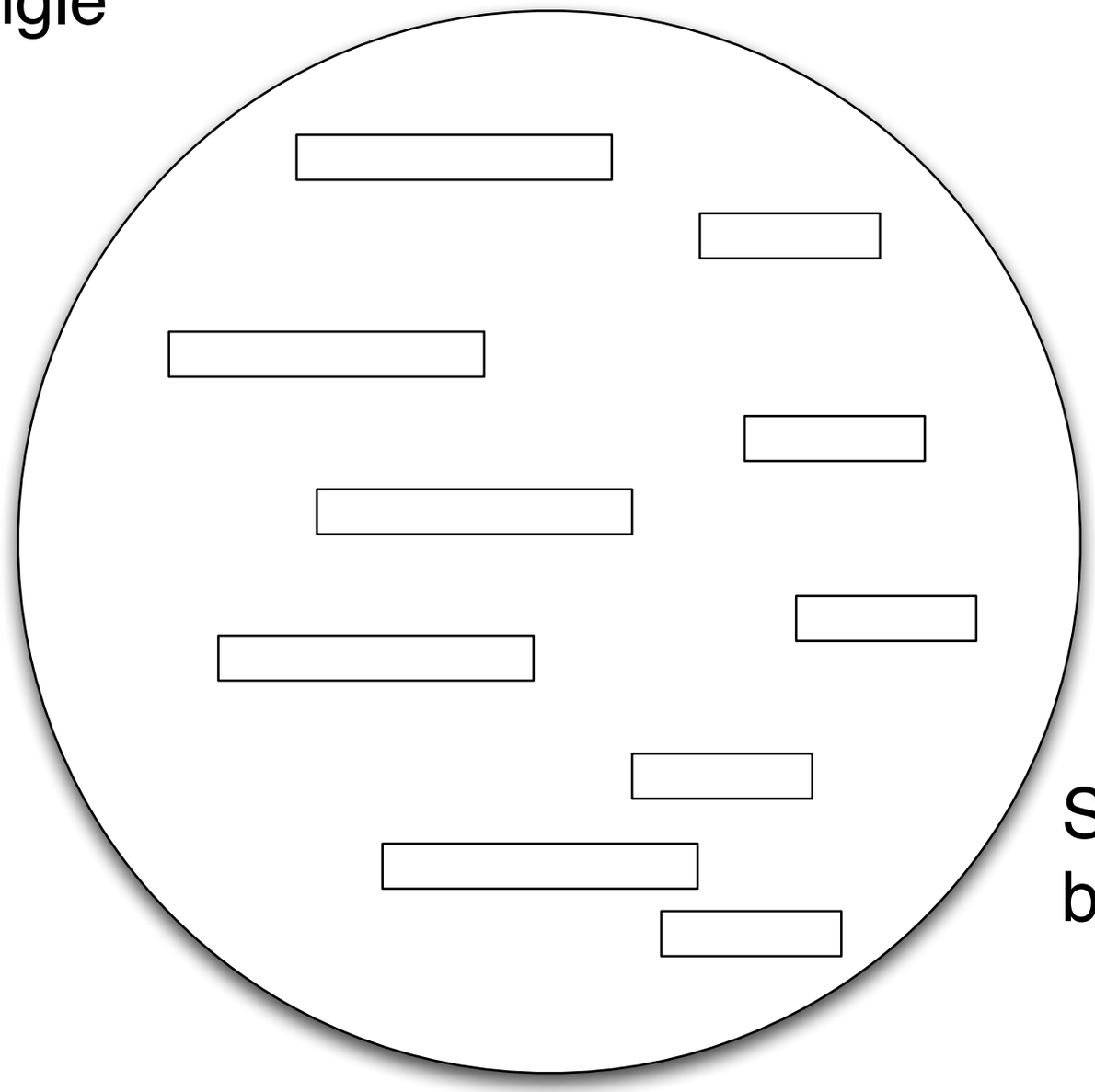
- Strong support for files found by every tool.
- Files found by a single tool may be subject to additional scrutiny.
- Examiner can always show the specific disk sector where data resides.



Automated Ascription of Multi-User Data

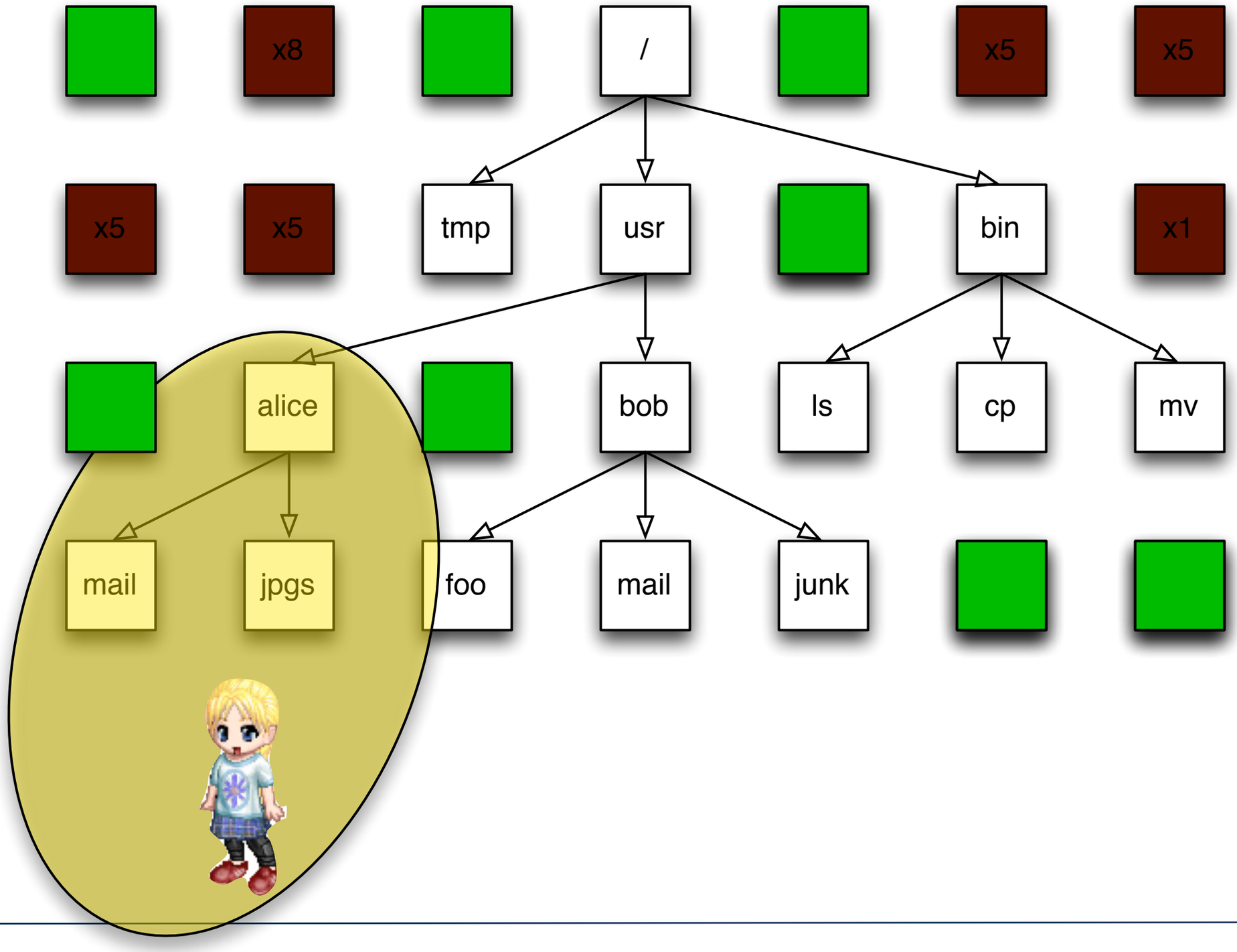
There is no "typical" hard disk.
Today's disks can have 0 — 10,000,000 files.

Some disks are used by a single person.

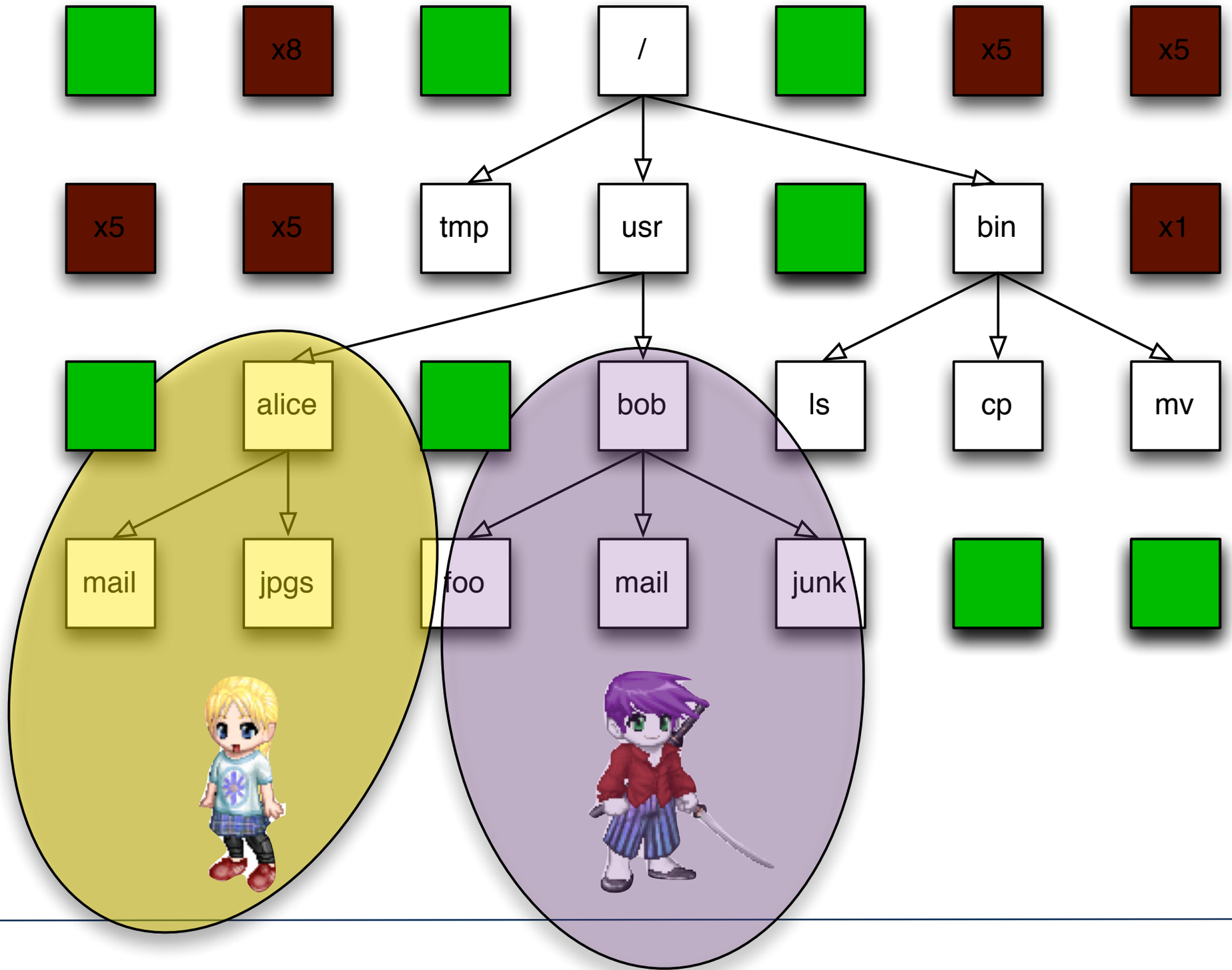


Some are used by multiple people.

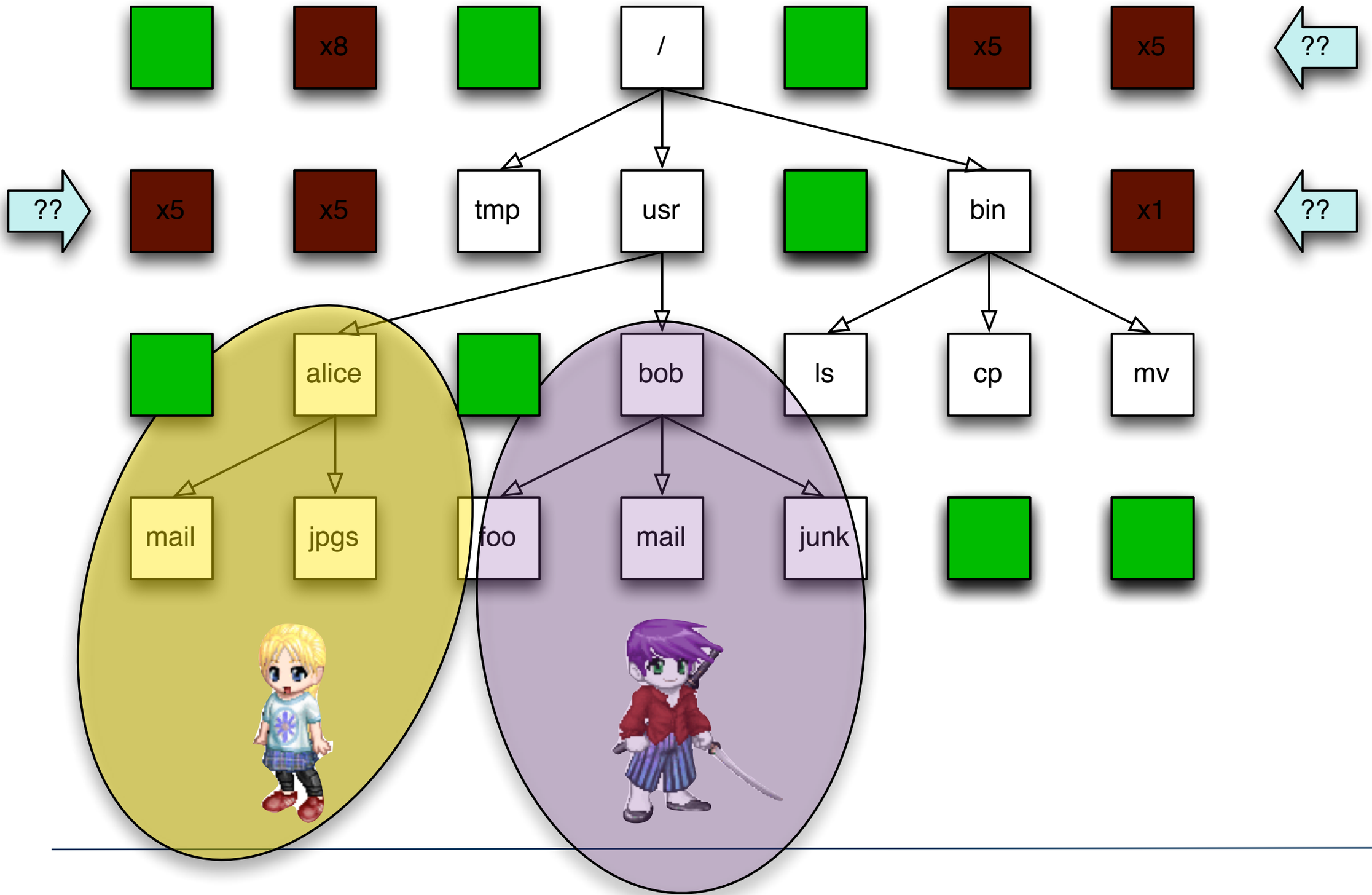
Some files were created by Yellow.



Some files were created by Magenta.



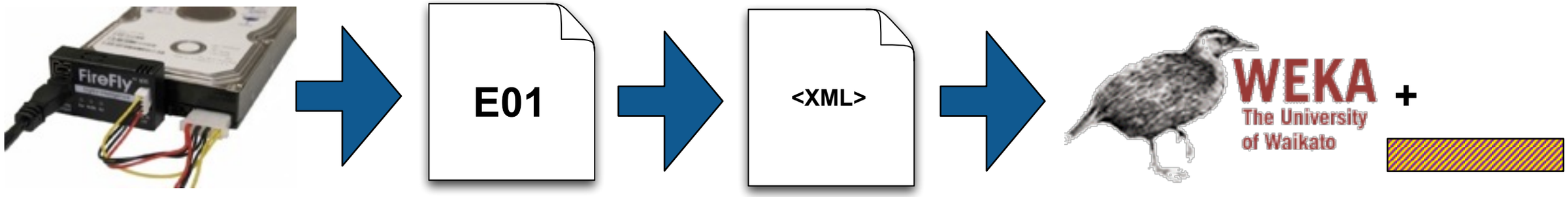
This work: Automatically identify the owner of files not in the file system.



Our approach to identifying the “owner:” Find commonalities with other files on the disk.

Magenta	Yellow		Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Print time: 9am & 10am	Print time: 5pm & 6pm	Print time: 5:30pm	
Location: 100 & 200	Location: 23,000 & 25,000	Location: 24,500	

We developed a statistical machine learning approach for attributing files to specific users.



Step 1: Extract all files and file *metadata.*

- File Owner (from filename or metadata)
- Location on disk
- JPEGs: Camera metadata
- Word Documents: Author, Last Edit Time, Print Time, etc.

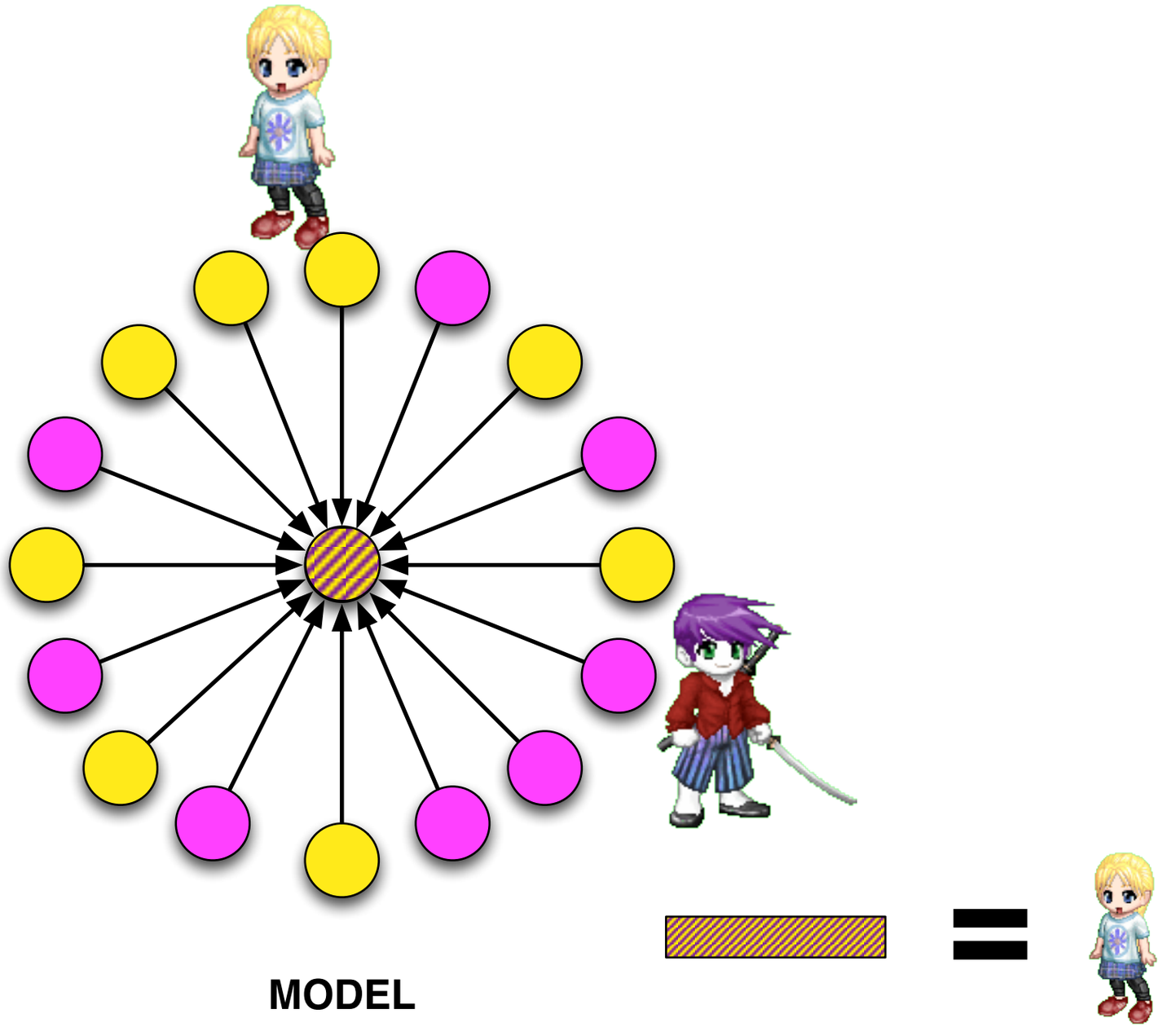
Step 2: Build a classifier using known files as exemplars.

- Ground truth: Directory path & file ownership
- Models: K-Nearest Neighbor, J48 Decision Tree

Step 3: Use classifier to ascribe deleted files.

The classifier is built from *all* of the allocated files.

To find the owner of a carved file, the file's metadata is extracted and classified with the model.

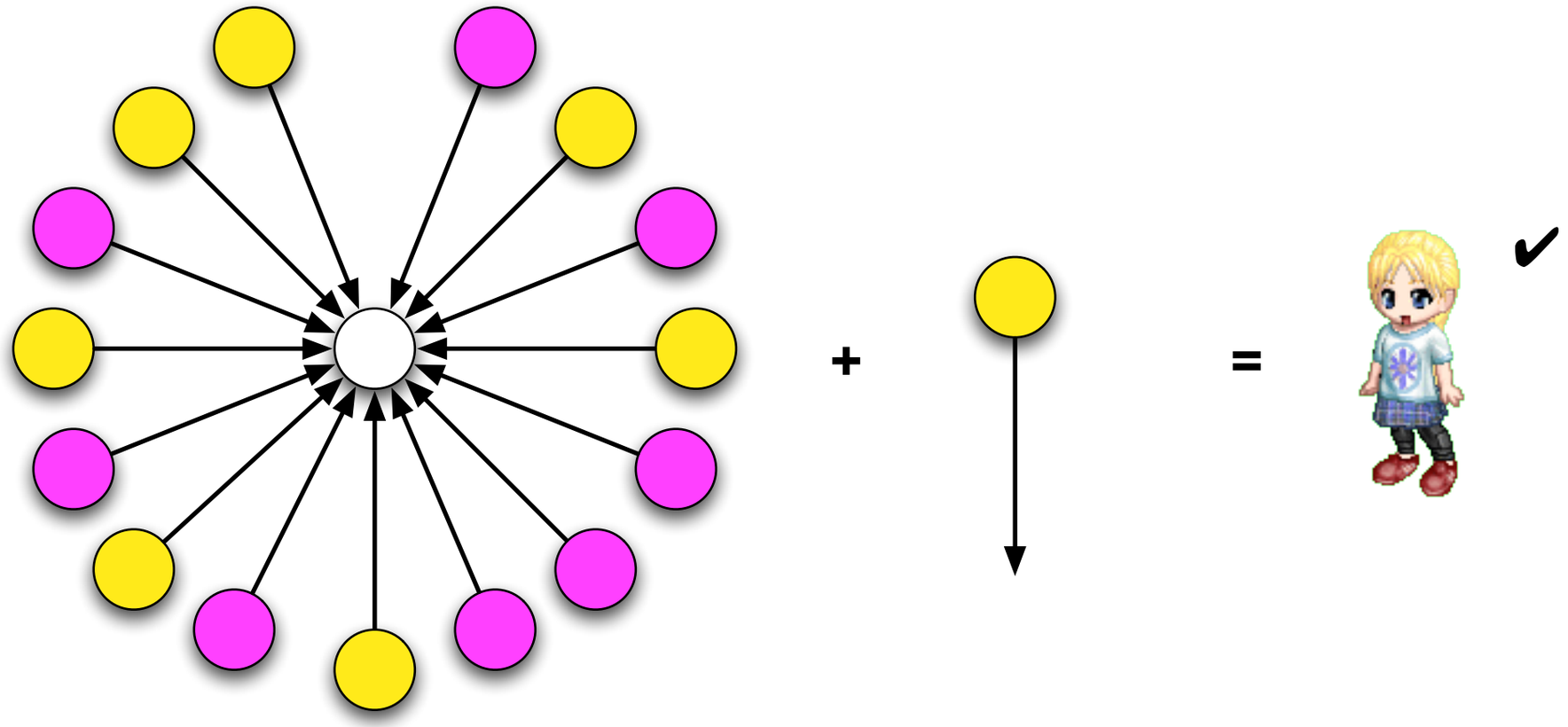


We validate the classifier with take-one-out validation.

For N files, we create N models (each missing one file).

We then classify the taken-out-file using the classifier.

Error rate = # wrong ÷ total number of files.

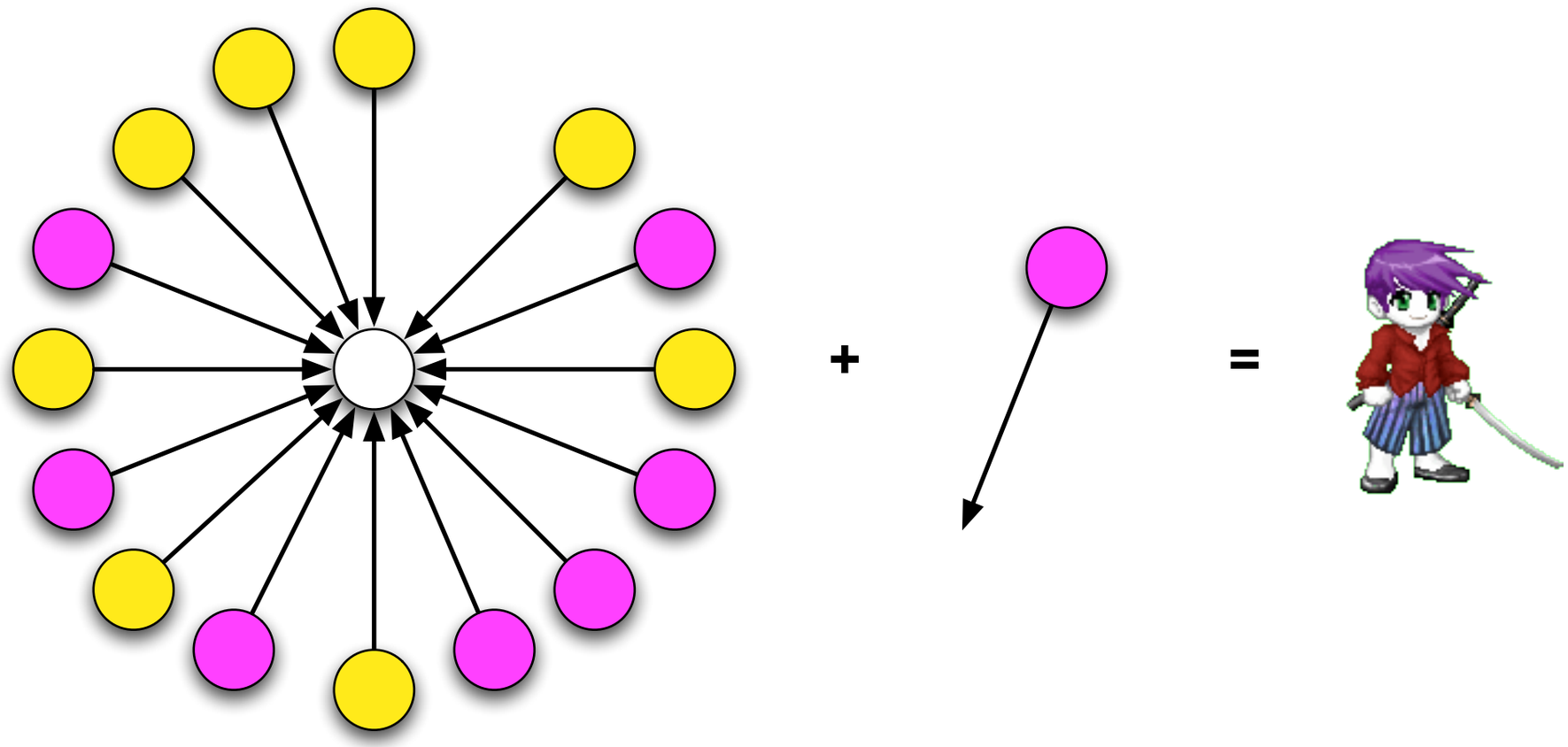


Here the classifier got it right!

Take-one-out validation produces the most accurate measure of the error rate.

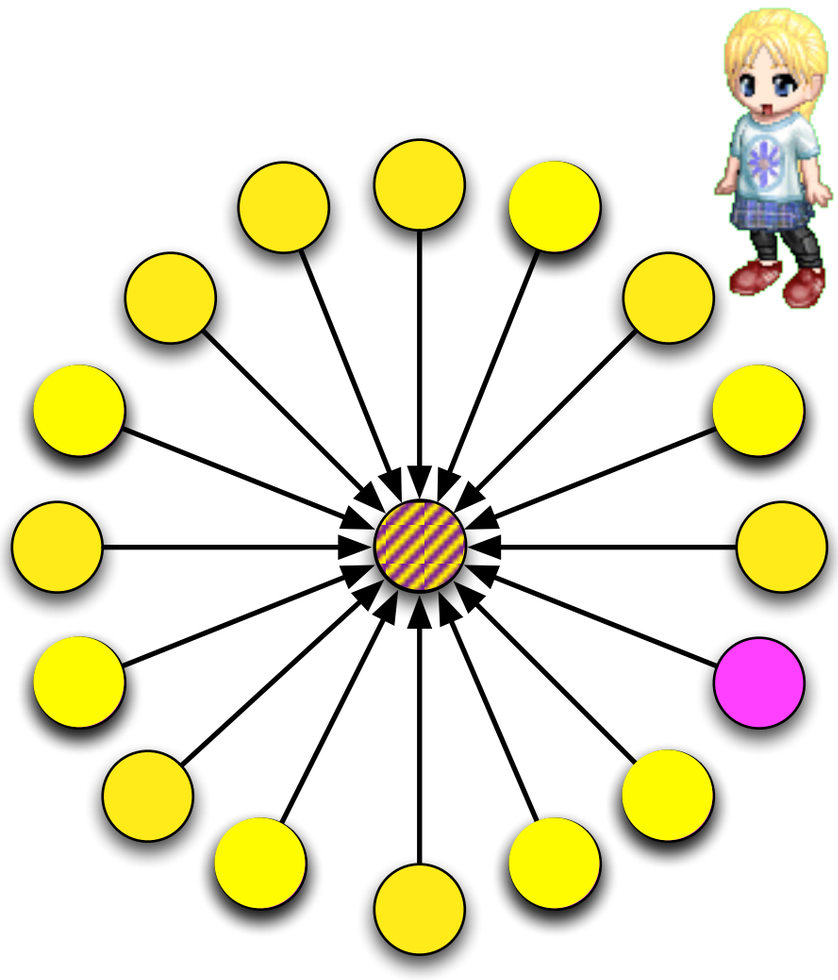
For N files, take-one-out tests N models, each *nearly identical* to the model that is used for classifying the carved data.

(More extreme version of *10-fold cross validation*.)

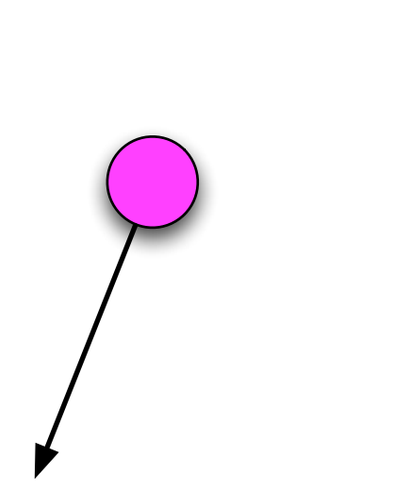


Each drive will have a different error rate.

A drive that over-represents data from Yellow may not be able to accurately classify files from Magenta:



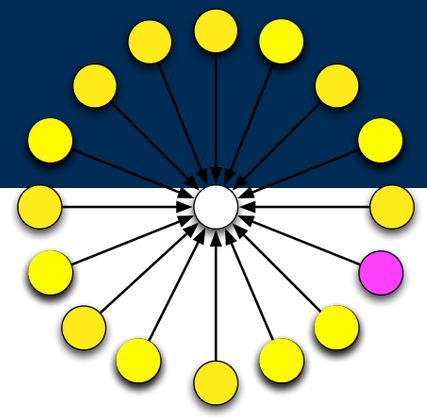
MODEL



?

Insufficient exemplars to classify file

Error is random, not systematic.



MODEL

The measurable accuracy of the model depends upon:

- Distribution of the allocated files.
- Having files that are representative of different use cases.

The accuracy of the classification depends upon:

- Similarity of the unknown files to the allocated files.

The tool reports the accuracy of classifying *known files.*

- This is assumed to be similar to the accuracy of reporting unknown files.
- We can’t know for sure!

Error rate:

- Different for each drive.
- Different for each user of each drive (some users classify better than others).

Summary

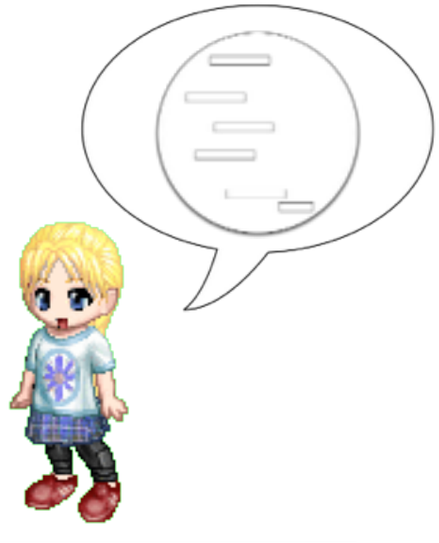
We presented two types of error measurements in storage forensics.

Random errors:

To find who owned a deleted file,
fit the file among all of the other files.

Systematic errors:

To efficiently verify file system reconstruction,
compare tool results with a descriptive and
precise language.



References

- Garfinkel, S., Parker-Wood, A., Huynh, D., and Migletz, J., *A Solution to the Multi-User Carved Data Ascription Problem*, IEEE Transactions on Information Forensics & Security, December 2010.

- Nelson, A., Steggall, E., and Long, D., *Cooperative mode: Comparative storage metadata verification applied to the Xbox 360*, in Proceedings of the DFRWS 2014 US Annual Conference, August 2014.