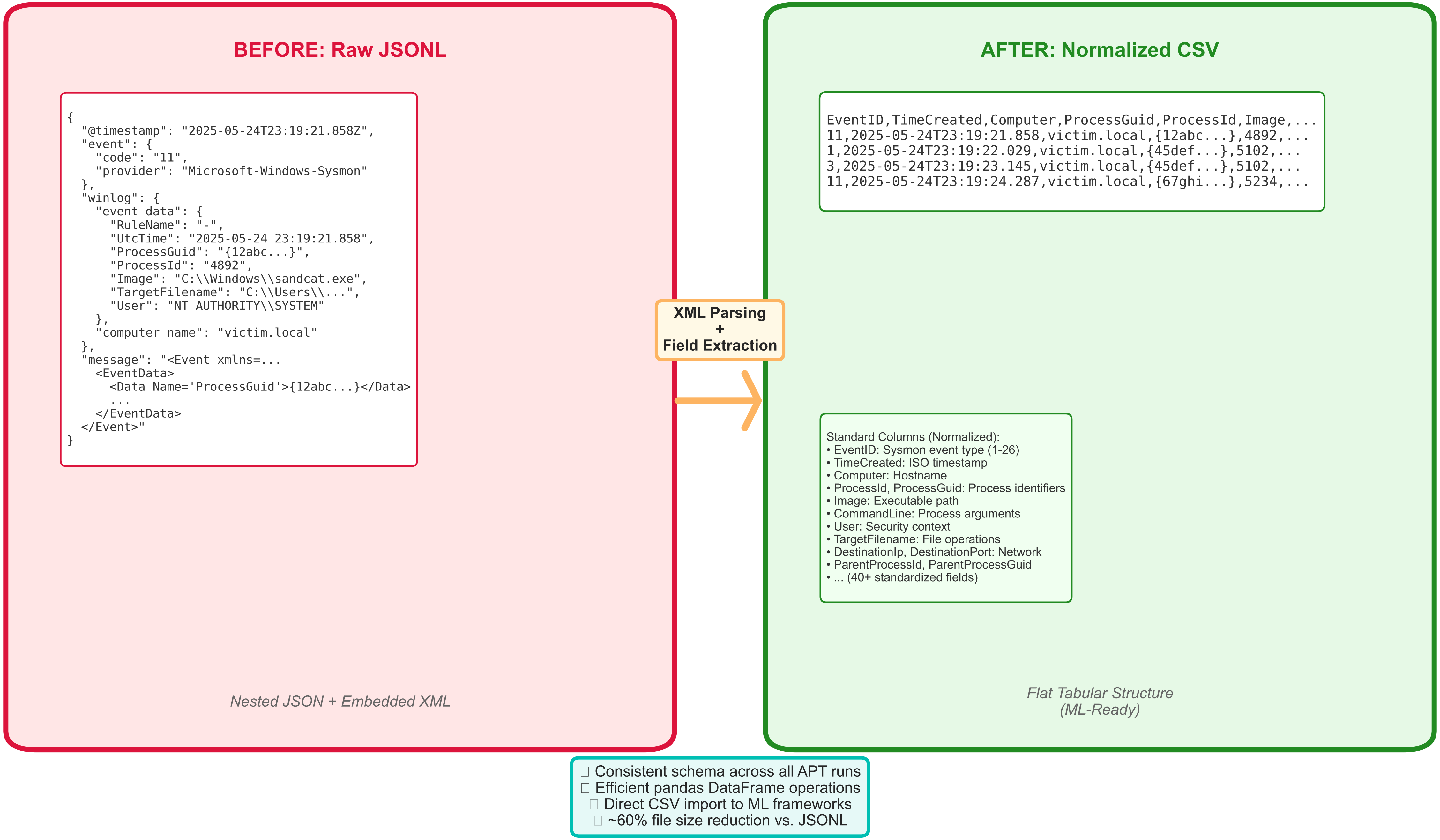
Figure 2.2: Sysmon Event Schema Transformation

JSONL (Nested XML) → CSV (Flat Normalized Schema)



BEFORE: Raw JSONL

```
{
  "@timestamp": "2025-05-24T23:19:21.858Z",
  "event": {
    "code": "11",
    "provider": "Microsoft-Windows-Sysmon"
  },
  "winlog": {
    "event_data": {
      "RuleName": "-",
      "UtcTime": "2025-05-24 23:19:21.858",
      "ProcessGuid": "{12abc...}",
      "ProcessId": "4892",
      "Image": "C:\\Windows\\sandcat.exe",
      "TargetFilename": "C:\\Users\\...",
      "User": "NT AUTHORITY\\SYSTEM"
    },
    "computer_name": "victim.local"
  },
  "message": "<Event xmlns=...
    <EventData>
      <Data Name='ProcessGuid'>{12abc...}</Data>
      ...
    </EventData>
  </Event>"
}
```

Nested JSON + Embedded XML

AFTER: Normalized CSV

```
EventID,TimeCreated,Computer,ProcessGuid,ProcessId,Image,...
11,2025-05-24T23:19:21.858,victim.local,{12abc...},4892,...
1,2025-05-24T23:19:22.029,victim.local,{45def...},5102,...
3,2025-05-24T23:19:23.145,victim.local,{45def...},5102,...
11,2025-05-24T23:19:24.287,victim.local,{67ghi...},5234,...
```

Standard Columns (Normalized):
- EventID: Sysmon event type (1-26)
- TimeCreated: ISO timestamp
- Computer: Hostname
- ProcessId, ProcessGuid: Process identifiers
- Image: Executable path
- CommandLine: Process arguments
- User: Security context
- TargetFilename: File operations
- DestinationIp, DestinationPort: Network
- ParentProcessId, ParentProcessGuid
- ... (40+ standardized fields)

Flat Tabular Structure
(ML-Ready)

□ Consistent schema across all APT runs
□ Efficient pandas DataFrame operations
□ Direct CSV import to ML frameworks
□ ~60% file size reduction vs. JSONL