UNIVERSITÀ DEGLI STUDI
DI TRENTO

Effective security management: using case control studies to measure vulnerability risk

A Tutorial

ISSRE 2014, 6th of November

Luca Allodi & Fabio Massacci

References

- Industry Talk:
  - *How CVSS is DOSsing your patching policy (and wasting your money).* BlackHat USA 2013, Las Vegas, NV.
- Paper:
  - Luca Allodi, Fabio Massacci. *Comparing Vulnerability Severity and Exploits Using Case-Control Studies.* ACM Transactions on System and Information Security (TISSEC). Aug 2014. Vol 17, Issue 1.
- These slides:
  - You can download them now:
  - http://tinyurl.com/pn776ly
  - → http://disi.unitn.it/~allodi/ISSRE-14/slides.zip

TENACE

In this tutorial

- Before coffee break: theory
  - Vulnerability risk assessment: the problem
  - CVSS and Inner workings
  - Methodology: case control study
    - CVSS as a risk factor, and other hypotheses
    - Metrics
    - Statistical background
- After coffee break: hands-on session
  - Presentation of Datasets
  - Implementation of methodology
  - Wrap-up

Vulnerabilities: a CIO Perspective

- 50k+ vulnerabilities in NVD
- My Software has a vulnerability: should I worry?
  - Published somewhere at BlackHat, DefCon, Slashdot, whatever.
- The fanatical answer is “I should, for each and every one”
- The actual answer is “For this one, I just can’t”
  - Technical Reasons
    - May not be technically fixable →integrated legacy sw may break
    - Even if expert to fix is there → she may have other tasks: relative priority?
    - Already planned upgrade in 3 months → why not just wait?
  - Budget Reasons
    - Money already allotted → again delay or stop other tasks
  - Compliance Issues
    - “It’s the law” → zilions of competing laws (e.g. Internet crimes, building safety, health insurance contribution, etc. etc.)
    - Paying a fine (later) may be cheaper than deploying a fix (now)
- → Need to Prioritize: “Worry now”, “Worry later”, “Life’s too short”
  - Need a “risk factor” with wich measure the vulnerability in order to decide

Vulnerabilities: a Compliance Perspective

- Listen to the U.S. Government….
  - US Cyber Security Order (Press release Feb'2013)
    - "NIST will work collaboratively with critical infrastructure stakeholders to develop the framework relying on existing international standards, practices, and procedures that have **proven to be effective**"
  - U.S. NIST SCAP Protocol v1.2(Draft Jan 2012)
    - "Organizations should use **CVSS base scores** to assist in **prioritizing** the remediation of known security-related software flaws based on the **relative severity of the flaws**."
  - PCI-DSS v2 (June 2012)
    - "**Risk rankings** should be based on industry best practices. For example, criteria for ranking —High risk vulnerabilities may include a **CVSS base score of 4.0 or above**"
  - U.S. Government Configuration Baseline (USGCB)
    - Supported by the industry→ Rapid7, Telos, VmWare, Symantec, Qualys, Retina etc. etc.

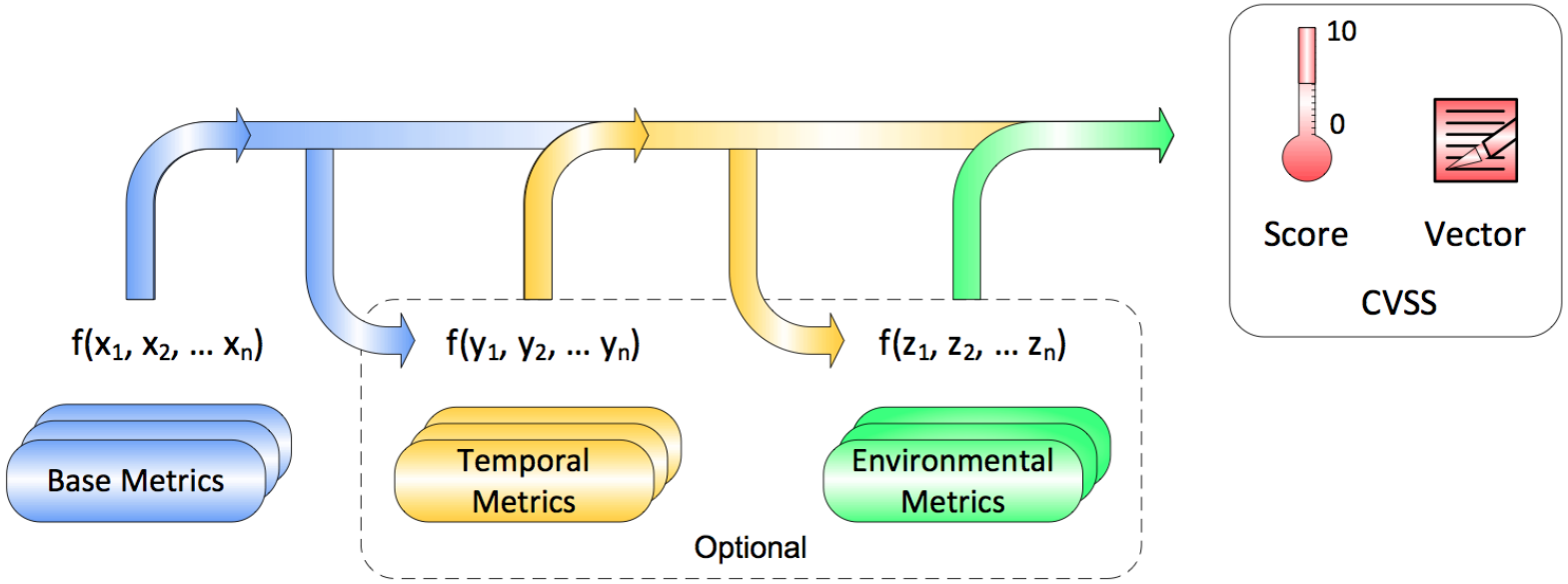
TENACE

Risk factors – Open questions

- Industry's risk factor of choice is CVSS
  - One rationale for all environments
  - Fix all vulnerabilities with high or medium CVSS score
- →Is this "proven to be effective"?
- → Can we check for the effectiveness of an arbitrary risk factor **before** sticking with it?
- → How do we evaluate if a risk factor is any good?

What is CVSS

- CVSS (2.0) is an assessment of how the vulnerability can impact the system
- Based on expert assessments to evaluate:
  - **Base Score *[0..10]***
  - Temporal Score
  - Environmental Score



Zooming in on Base Metrics

CVSS.base = f(Exploitability) x g(Impact)

- Exploitability
  - Access Vector (local, adj, network)
  - Access Complexity (high,medium,low)
  - Authentication (multiple,single,none)
- Impact (High,Medium,Low)
  - Confidentiality, Integrity, Availability
- Basically it is a “Clinical Evaluation”
  - “clinical examination is the process by which a medical professional investigates the body of a patient for signs of disease” (Wikipedia)

Risk: a clear answer

- HIGH CVSS → Patch (Best security practice)
- Double vision → Brain Surgery (Best medical practice?!)
- A clinical test must be matched to the risk it subdues
  - Binocular diplopia and no additional evidence → 42% recovered ***without*** treatment
  - Binocular diplopia AND intracranial lesion → 0% recovered without treatment
    - Nolan "Diplopia" B. J. Ophtalm. 1966
- What the CIO would like to know:
  - IF HIGH CVSS listed by Sec. Config. Manager and Metasploit finds it → fix it and decrease risk of successful attacks by +15%
  - IF fix all remaining HIGH listed by Sec. Config. Manager but no additional evidence → risk decreses only by 3%

    → Is +3% worth the extra money?

TENACE

Our goal

- CVSS in this context is just a risk factor for exploitation
  - One may consider any other risk factor (e.g. Proof-of-Concept exploit)
- Risk factor = foresight assessment of a vulnerability
- Develop a methodology that evaluates an arbitrary risk factor or measure with respect to the actual risk
- A good risk factor should mark
  - Risky vulnerabilities as high risk
  - Non-risky vulnerabilities as low risk
- The "riskier" high-risk vulnerabilities are, the better the investment in patching them
  - ❑ → Risk(high-risk vulns) – Risk(low-risk vulns) >> 0

"IF we mitigate vulnerabilities with feature Y THEN risk of attacks decreases by x%"

Linking risk factors and attacks

- We want to assess “how meaningful a risk factor is”
  - Is there a correlation between the risk factor and the existence of an attack?
- How do you link
  - Smoking to lung cancer?
  - Safety belts to reduced risk of death?
- You can’t run a “controlled” experiment
- So you run a “case control” study
  - Doll & Bradfor Hill, British Medical Journal 1950 (&1970)
    - Smoking → Lung Cancer
  - Is there a (statistical) difference in smoking habits between sick people and healthy people with the same characteristics?
- Is there a (statistical) difference in the risk factor between exploited and not exploited vulnerabilities?

Case control study

- Can link an hypothesis to its (hypothesized) consequence
  → Explanatory variable to an effect
- A case control study looks backwards at the history of what happened
- Controls the population of observed cases by assuring that controls and cases are virtually identical but for one factor
  – The remaining factor = hypothesis or risk factor
- Evaluate whether the risk factor characterizes the cases significantly more than the controls

Case control studies: disclaimer

- Causation can be established only through a controlled experiment
  - Every variable is under the control of the researcher
- Case control studies can only look backward at cases and controls the researcher can only observe
  - Historic records may be incomplete
  - Variables may be measured inconsistently between subjects
  - Some variables may be simply not reported
- → Case control studies can only highlight a correlation between a factor and an observation
  - ✓ HIGH CVSS + PoC is correlated to an attack
  - X ~~HIGH CVSS + PoC causes an attack~~

Terminology for case control studies

- **Cases:** a case is the entity/phenomenon of interest for the investigation
  - **Medical: Cancer patient**
  - **Security: Exploited vulnerability**
- **Explanatory variable(s):** the hypothesised explanation(s) for the observation of the case
  - **Medical: Smoking cigarettes**
  - **Security: High CVSS**
- **Confounding factor(s):** alternative explanations that may influence the probability of having a case
  - **Medical: Environmental pollution; Gender; Age; Wealth (others?)**
  - **Security: Type of Vuln; Type of Affected Software; Age of Vuln (others?)**
- **Control group:** a group of people with the same value for the confounding factors as the cases
  - **Medical: Pick at random from population people with same confounding values as the Cancer patients**
  - **Security: Pick at random from population vulnerabilities with same confounding as the Exploited vulnerabilities**

Methodology (1) - Cases

1. Find cases (exploited vulnerabilities)
   - Records of vulnerabilities for which you are certain there is an exploit in the wild
     - E.g. from your network's IDSs / monitors
   - For the purpose of this exercise, we will use Symantec's data:

Our Cases (the lung cancer patients, deads in accidents)	What's there
SYMANTEC's Threat Explorer Browser/Plugins 14% – Server 22% – App. 17% - Windows 13% - Other OS 5% - Developer 5% - Business 7% - Unclassified 17%	1266 vulns with at least 1 attack in the wild (~600 sw)

Methodology – Hypotheses (1)

- Hypotheses:
  - Define the risk factors you want to assess
  - Can be anything you think may be correlated with the existence of an exploit in your "dataset of cases"
    - Must be measurable **in foresight** from your data
  - In this example we hypothesize:
    - Hyp1. CVSS → correlates with exploitation
    - Hyp2. Proof-of-Concept → correlates with exploitation

Risk factors	What's there
CVSS - NATIONAL VULNERABILITY DB (the "universe")	45K+ vulns, 16K types of sw/versions etc.
Proof-of-Concept exploit is published by security researchers - EXPLOIT DB	8K+ vulns, (~6k sw)

Methodology – Confoundings

- Cannot use data as-is to draw conclusions
- There may be other reasons why you have a record about that particular case
  - Some factors may influence the probability of you observing a particular attack

  → Attack is not in your data != attack does not exist
- Examples:
  1. Some exploit you may just never see because you don't have that software deployed in your environment
  2. Some exploits you may not see because they are old and your data starts at a later point in time
  3. You are particularly good at detecting certain types of attacks (e.g. networked) but not others (e.g. local)
  4. Etc..

Example of the confounding problem

- Symatec's commercial interest in a particular software platform influences the existence of an exploit record for that platform
  - NVD/EDB may list software almost nobody uses
    - Sw in SYM → 568, sw in EDB→ 5.819, sw in NVD → 16.399
    - E.g. a Joomla expansion module
- Symatec's consumer clients may mainly suffer from particular types of attack
  - E.g. Installing malware on the machine

    →Vulnerabilities in SYM tend to have complete impacts on CIA
- Symantec's efforts in reporting attacks may change with time
  - E.g. after 2009 Symantec started the WINE initiative to share attack data with security researchers

Methodology – Our Confoundings

- The confoundings we will consider in our example:
  - **Vulnerable software**
  - **CVSS Confidentiality, Integrity, Availability impact**
  - **Year of vulnerability disclosure**

Methodology – High-level implementation

- For each case=*vuln in SYM*
  1. Measure the confoundings for that vulnerability
     - Year
     - CIA
     - Software
  2. Pick one vulnerability with the same confoundings at random from the population
     - Measure explanatory variable (CVSS; PoC)
     - Is the vulnerability also a case?
       - » Yes & Explanatory variable HIGH → True positive
       - » Yes & Explanatory variable LOW → False Negative
       - » No & Explanatory variable HIGH→ False Positive
       - » No & Explanatory variable LOW→ True Negative

Results of the Experiment

- We are interested in measuring how precise our risk factor is in identifying exploits from not exploits
- Result is a latin square:

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

- We will derive all our measures from this table

TENACE

Evaluating the risk factor

- A good risk factor should:

1. Mark dangerous vulnerabilities as risky
   → cover most of the risk
2. Mark not dangerous vulnerabilities as not risky
   → avoid unnecessary overhead by erroneously prioritizing low risk vulnerabilities
3. Aid the vuln fixing by efficiently addressing risky vulnerabilities
   → leftovers should represent much lower risk than fixed vulnerabilities

Sensitivity and Specificity

1. Sensitivity → true positives vs all exploited vulns
   - **HIGH→ the test correctly identifies exploited vulns**
     - **100% = Patching fully covers risky vulns**
   - LOW → lots of “bad vulns” undetected

2. Specificity → true negatives vs all not exploited vulns
   - **HIGH → the test correctly identifies non exploited vulns**
     - **100% = Patching has no overhead**
   - LOW → lots of “ok vulns” flagged for patching

How to calculate Sensitivity and Specificity

$$Sensitivity = P(TruePositive \mid SYM) = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$Specificity = P(TrueNegative \mid notSYM) = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

Risk Reduction

- "If I patch vulnerabilities with features X would this reduce my risk of getting attack?"
- Risk reduction = Difference in risk between "HIGH" and "LOW" levels
- Compute answer from same table but by row
  - How effective is the risk factor in addressing risky vulns

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

- The larger the risk reduction the more effective the marker

How to calculate Risk Reduction

$$Risk(MarkedHigh) = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Risk(MarkedLow) = \frac{FalseNegatives}{FalseNegatives + TrueNegatives}$$

$$Riskreduction = Risk(MarkedHigh) - Risk(MarkedLow)$$

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

Statistical background

Inferential power of the results (1)

- Our results table reports, so far, only a point estimate generated from the random sampling
- May not be good enough
  - Unrealistic by pure chance (random sampling)
  - Could be an “unlucky” selection of vulnerabilities that is biased in one direction
    - E.g. More True Positives than the marker is really able to find
      - Extreme case: picked only vulns that are exploited

Inferential power of the results (2)

- In general, one set of results may say little about the real performance of the marker
- Your cases are just a sample of the real population of exploits (that you don't know)
  - Our results are an estimate of the real value based on our sample
    - E.g. sensitivity=x → is the **real sensitivity <,=,> x?**
  - Can we understand anything about that unknown population by looking at our sample?

Solution: bootstrapping

- One sample is not good enough so..
- ..Perform the selection multiple times!
  - The probability of each sample must stay unchanged
  - → Sample n times (n>>1) with replacement
- The result of the procedure is no more a single table but n tables
- The distribution over the n tables approximates the real (unobservable) performance of the marker
- Allows to build confidence intervals and descriptive statistics for the estimate of interest

B. Efron. *Bootstrap Methods: Another Look at the Jackknife*. Annals of Statistics Vol 7, no 1 (1979).

Median

- Median is the value of a distribution below which and above which there is 50% of the population sample
- More robust to outliers than the mean

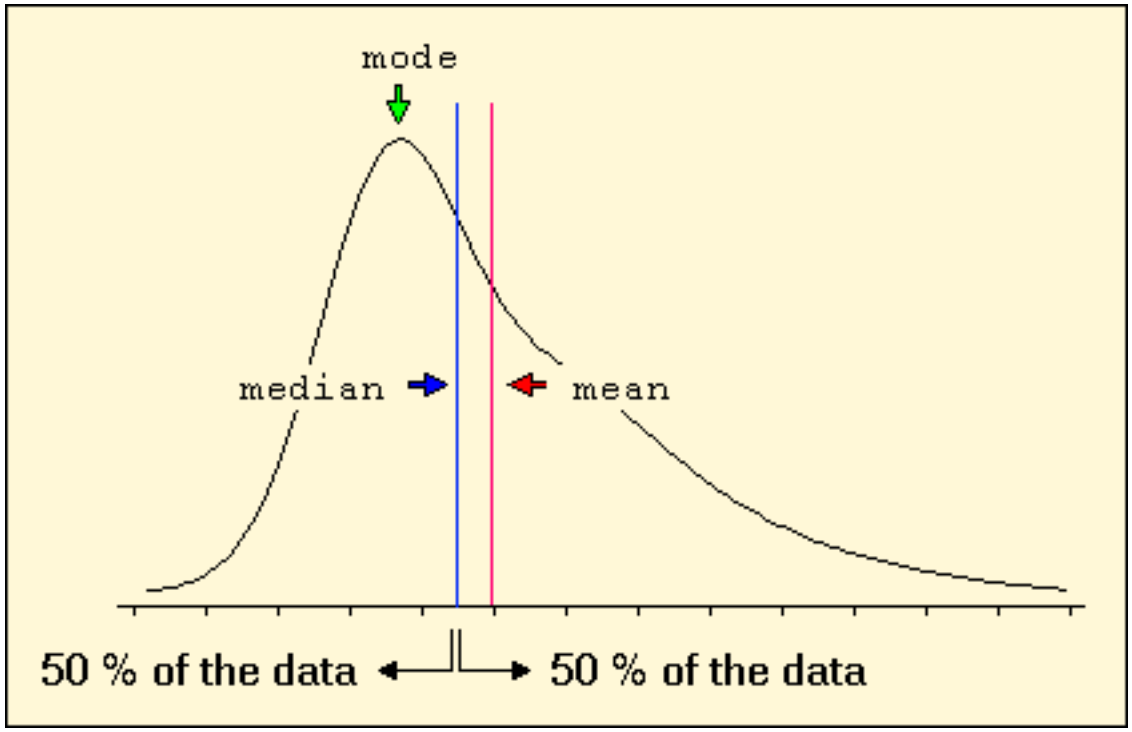


Image from: http://www.statistics4u.info/

Confidence Intervals

- We build an **estimate** of the true value of the parameter
- But where is the **real one** *(that we don't know)?*
- We can choose a confidence interval in which we are *x%* sure the *real value* is
  - *x%* usually is 95%

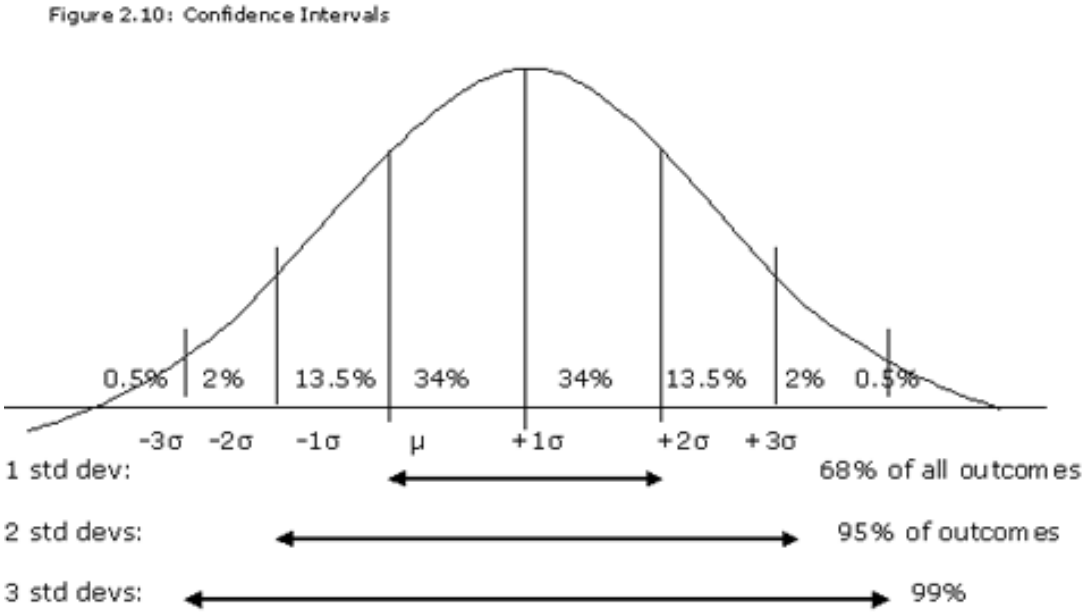


Image from: http://www.investopedia.com

TENACE

Statistical Significance

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

- Estimations in output are just numbers
  - If the marker has no positive or negative influence on the existence of the vuln in SYM, then TP,FP,FN,TN are there by pure chance
  - In this case, our estimation would not make any sense
- **A statistical test tells you whether your conclusion is distinguishable from a conclusion generated from random data**
- The table in output may contain sparse data
  - Sparse = some cells with a count *< 5*
- We use a Fisher Exact Test
- $p < 0.05$ = statistically different from random
- $p > 0.95$ = statistically identical to random

Implementation

TENACE

Implementation schema

1. Read the data
2. For each vuln in SYM
   - Build the test samples from the overall population
     - Identical to SYM's as measured by the controls
3. Evaluate hypothesis for sampled population
   - What's the risk factor level?
   - Is it in SYM?
4. Output results

Run example

- Case (attacked vulnerability):
  - CVE-2010-3962 (use-after-free vulnerability in MS IE 6,7,8)
  - Year=2010
  - Confidentiality =C, Integrity=C, Availability=C
  - Software = ie
- Control (vulnerabilities similar to attacked ones):
  - Select randomly 1 out of 37 in NVD
  - Measure explanatory variable (your hypothesis)
  - Is it a case?
- Repeat for all 1266 cases of attacked vulnerabilities

Repeat n times (bootstrapping)

Reminder: our hypotheses for today

- We will have two hypotheses:
  1. CVSS > 4 score increases the chances of exploitation
  2. Existence of PoC increases the chances of exploitation

- You can add any further hypothesis you want
  - Of course it needs to be measurable from the data you have

Reminder: Output table

- At the end of the day we are interested in measuring how precise our risk factor is in identifying exploits from not exploits

	In SYM	Not in SYM
Marked HIGH	True Positives	False Positives
Marked LOW	False Negatives	True Negatives

Hands on session

- Have you installed R?

→ http://cran.r-project.org

  - These slides
    - http://tinyurl.com/pn776ly
    - → http://disi.unitn.it/~allodi/ISSRE-14/slides.zip

- For those of you interested in having access to attack data:
  - http://www.symantec.com/about/profile/universityresearch/sharing.jsp

Hands on session – R quick guide

- In R assignments to variables are made with the “<-” operator
  - E.g. variable <- “this variable will be a string”
  - E.g. variable <- c(“first string”, “second string”)
- Variables can be atomic, vectors or tables (data frames)
  - When more than a dimension, can access cells with
    - Vectors : variable[index]
    - Data frames : variable[row, column]
  - Columns can be listed
    - Data frames : names(variable)
  - And accessed directly
    - Data frames : variable$columnName

The data

- Download from
  - http://disi.unitn.it/~allodi/ISSRE-14/tutorial_data.zip
  - http://disi.unitn.it/~allodi/ISSRE-14/code.zip
- Dataset Sample:

cve	cvss	impact	exploitability	acc.vector	acc.complexi	auth	conf	integ	avail	edb	nvd	symantec	software	pub_date	UNIX
CVE-2012-2612	5	2.9	10	N	L	N	N	N	P	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2611	9.3	10	8.6	N	M	N	C	C	C	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2514	5	2.9	10	N	L	N	N	N	P	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2513	5	2.9	10	N	L	N	N	N	P	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2512	5	2.9	10	N	L	N	N	N	P	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2511	5	2.9	10	N	L	N	N	N	P	0	1	0	netweaver	15/05/12	FALSE
CVE-2012-2450	9	10	8	N	L	S	C	C	C	0	1	0	fusion	04/05/12	FALSE
CVE-2012-2449	9	10	8	N	L	S	C	C	C	0	1	0	fusion	04/05/12	FALSE
CVE-2012-2448	7.5	6.4	10	N	L	N	P	P	P	0	1	0	esx	04/05/12	FALSE
CVE-2012-2441	8.5	10	6.8	N	M	S	C	C	C	0	1	0	ros	27/04/12	FALSE

TENACE

Hands on session

- Open your Rs

- Set your local working directories where you copied the data
  - `setwd('<directory>')`
- `Read the data`
  - `data <- read.csv('general_table.csv', sep=',', header=T,stringsAsFactors=F)`
  - `data <- subset(data, data$UNIX==F)`
  - `data$cvss <- as.numeric(as.character(data$cvss))`
  - `data$pub_date <- as.POSIXlt(data$pub_date, format="%d/%m/%Y")`

Hands-on tour

Next steps

- We would be interested in having a replication experiment
- → CVSS v3 assessment of vulnerabilities
  - -How and by what factors does a CVSS assessment vary?
- Would you be interested in participating?
- Feel free to contact me:

luca.allodi@unitn.it