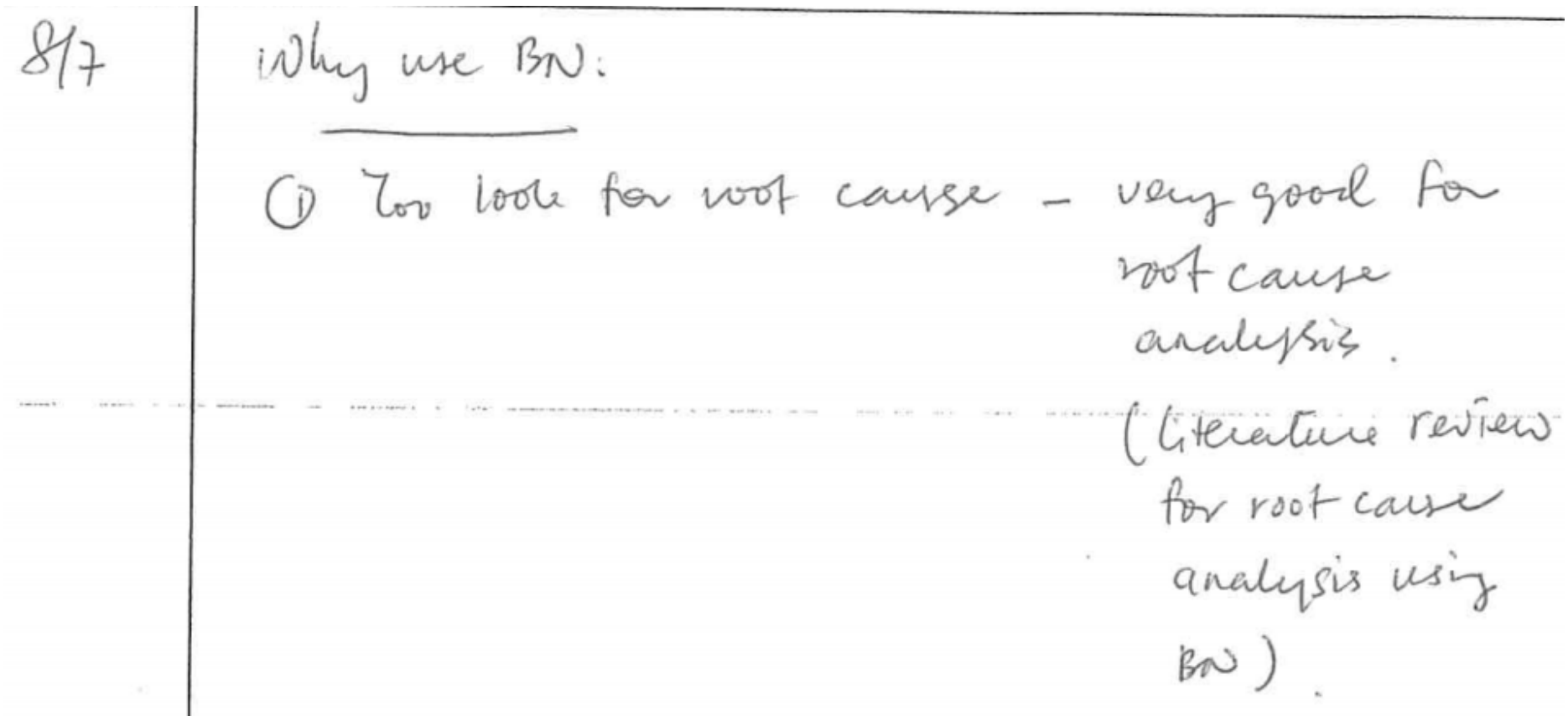
PHD - Progress update 14/10/2015

Fauzy Bin Che Yayah

October 14, 2015

- Why use Bayesian Net ?

1. To look for the root cause - the dataset is in factor type which is status not a number.
2. To find out the probabilistic relationship between the symptom error code and the resolution
3.



Citation	
TroubleMiner: Mining network trouble tickets Medem, A. ; Akodjenou, M.-I ; Teixeira, R. 20091	
Knowledge Discovery from Trouble Ticketing Reports in a Large Telecommunication Company Temprado, Y. ; Garcia, C. ; Molinero, F.J. 2009	Data Mining , Text Mining an
A Bayesian Approach To Stochastic Root Finding 2011	
A Fully Bayesian Approach For Unit Root Testing 2011	
Online Root-Cause Analysis Of Alarms In Discrete Bayesian 2014	
Documents Categorization Based On Bayesian Spanning Tree 2006	
Benefits of a Bayesian Approach to Anomaly and Failure 2009	

List of literature review regarding Bayesian Net :-

1.A real-life application of multi-agent systems for fault diagnosis in the provision of an Internet business service

2.A Bayesian Network approach to diagnosing the root cause of failure

3.sss

Process on gathering the dataset

- Acquiring dataset for 100 records, for each zone , randomize , selective year ; ie . 2015
- Rules :-

Rules	Description
status = 'Closed'	Dataset must be closed for complete information
network_tt_id is NULL	Dataset must be not related to Network Trouble Ticket
trouble ticket type <> PASSIVE	Trouble Ticket must related to the Active elements such as routers, switches , modem , etc
installed_date is NOT NULL	This field must have value
created_date is NOT NULL	This field must have value
closed_date is NOT NULL	This field must have value

- Sample SQL to retrieve from Impala :-

Documentation - https://cran.r-project.org/web/packages/RImpala/RImpala.pdf

- Package installation

```
install.packages("RImpala")
library("RImpala")
```

- Notes

```
Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine
runs natively in Apache Hadoop
```

- Connection to Impala - Sample SQL

```
select * from nova_trouble_ticket a join active_code b on (trim(a.cause_code) = trim(b.cause_code)) join exchange_zone c ON (trim(a.exchange)=trim(
where  c.zone_name like '%ZONE KEPONG%' and a.status like '%Closed%'  and length(a.cause_category) > 1  and length(a.created_date) > 6 and length(a
and a.package_name not like '%null%' and a.product not like '%null%' and a.sub_product not like '%null%'  and  length(a.description) > 10
and network_tt_id = 'null'
order by rand() limit 10000
```

Below is the dataset column name :-

```
#a <- read.csv("table_struct.csv")
#names(a)
```

Total Zone available : 53

```
Air Itam,Bangi,Bangsar,Banting,Batu,Batu Pahat,Bayan Baru,Bintulu,Bukit Anggerik,Bukit Mertajam,
Bukit Raja,Butterworth,Cyberjaya,Gombak,Ipoh,Kajang,Kepong,Keramat,Kinrara,Kl Central,Klang,Kota Kinabalu
Selatan,Kota Kinabalu Utara,Kuching,Kulim,Langkawi,Maluri,Melaka Utara,Miri,N. Sembilan Utara,Pandan,
Pelangi,Perlis,Petaling Jaya,Puchong,Seberang Jaya,Senai,Sg Petani,Shah Alam,Sibu,Skudai
Pontian,Stampin,Subang Jaya,Taman Petaling,Tampoi,Tar,Tasek,Tasik Ampang,Tdi,Teluk Intan,Terengganu
Selatan,Teruntum
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.2.2
```

```
a <- read_excel("kepong.xls")
names(a)
```

```
##  [1] "tt_row_id"             "tt_num"
##  [3] "tt_type"               "tt_sub_type"
##  [5] "status"                "severity"
##  [7] "important_message"     "appointment_flag"
##  [9] "nova_account_name"     "nova_subscriber_num"
## [11] "nova_account_num"      "package_row_id"
## [13] "created_by"            "category"
## [15] "symptom_error_code"    "priority"
## [17] "product"               "sub_product"
## [19] "package_name"          "network_tt_id"
## [21] "swap_order_num"        "cause_category"
## [23] "cause_code"            "resolution_code"
## [25] "closure_category"      "resolution_team"
## [27] "service_affected"      "service_order_num"
## [29] "btu_type"              "owner"
## [31] "owner_name"            "group_owner"
## [33] "owner_position"        "btu_platform"
## [35] "dp_location"           "created_date"
## [37] "pending_verify_date"   "closed_by"
## [39] "closed_date"           "source"
## [41] "installed_date"        "description"
## [43] "repeat_ticket_count"   "follow_up_ticket_count"
## [45] "fdp_device_name"       "fdp_site_name"
## [47] "olt_site_name"         "exchange"
## [49] "timestamp"             "contact_id"
## [51] "contact_name"          "contact_office_phone"
## [53] "contact_mobile_phone"  "contact_home_phone"
## [55] "contact_email_addr"    "due_date"
## [57] "part_num"              "network_layer"
## [59] "network_row_id"        "asset_id"
## [61] "ptt"                   "zone"
## [63] "service_point_id"      "cause_code"
## [65] "code"                  "building_id"
## [67] "region"                "state"
## [69] "district"              "zone_name"
```

```
# correlations <- cor(a, use="pairwise.complete.obs", method="pearson")
# print(correlations)
# highlyCorrelated <- findCorrelation(correlations, 0.95 ,verbose = FALSE,names = TRUE)
# print(highlyCorrelated)
# summary(correlations)
# corrplot(correlations, method = "circle",tl.cex = 0.9)
```