**Doel:** Ik wil beginnen met voelen dat ik een goeie prestatie heb geleverd en dat ik doorzettingsvermogen heb getoond. Dit bestaat eruit dat ik wil dat ik mijn scriptie inlever zonder stress richting het einde. Dat wil zeggen ik wil van tevoren klaar zijn en zoals Jaap eerder heeft gezegd ‘fluitend’ een document inleveren en een Github hebben waarbij alle code uitgewerkt is. Verder wil ik de experimenten die ik nu voorstel in mijn thesis doen en hier resultaten voor hebben + analyse m.b.v. eerdere analyse uit GOGGLE (**t-SNE**). Ik wil mijzelf hebben overtroffen op technisch gebied. Concreet bestaat dit uit: technische wiskundige beheersing van wat ik heb opgeschreven in de thesis en een goeie flexibele mogelijkheid om formules uit te leggen. Een goed lopend verhaal narratief.

Mijn doelen bepaal ik door een vraag te stellen per subonderdeel. Ik begin met een lijst aan punten die de basis vormen. Ik maak hierin onderscheid tussen thesis deel en experimenten. De punten die ik erken zijn al volgt:

**Thesis – tussentijdse doelen**

* Welke conclusie wil ik trekken uit mijn onderzoek?
* Wat voor analyse kan ik doen om te analyseren of de resultaten de conclusie ondersteunen of afbreken?
* Welke experimenten kan ik doen zodat de resultaten de conclusie ondersteunen?
* Welke theorie is hiervoor relevant?
* Welke eerdere werken zijn hiervoor relevant?
* Hoe introduceer ik het onderwerp?

**Code**

* De code is in eerste instantie afhankelijk van wat ik wil testen, de code is ondergeschikt aan het paper. Het is het tweede dat ik aflever. Het is geen absolute vereiste.

Concreet vertalen de doelen zich in een aantal to-do’s. Deze to-do’s dienen dag afgevinkt te worden om de subdoelen te behalen. Ze dienen vooral concreet te zijn en samen een subdoel van de scriptie te behandelen. Laten we beginnen met het eerste tussentijdse doel:

*Welke conclusie wil ik trekken uit mijn onderzoek?*

Terug naar de conclusie is het zo dat ik graag onderzoek hoe beide modellen met elkaar te verenigen zijn. Ik zou graag beide modellen met elkaar willen vergelijken. Ik zou graag uiteindelijk willen dat ik iets kan zeggen over de definitie van fairness getest in een context van een ander model dat zich met dezelfde taak bezighoudt. Aanvullend zou ik graag een toevoeging doen aan een bestaand model en die testen op haalbaarheid.

*Wat voor analyse kan ik doen om te analyseren of de resultaten de conclusie ondersteunen of afbreken?*

Analyse zal bestaan uit de volgende aspecten.

To evaluate DECAF and to compare it’s performance against GOGGLE we split the training set into D\_train and D\_test and the GOGGLE and DECAF generative models are trained on D\_train. Then, both trained models are used to generate a synthetic dataset, that has the same size as D\_test. This results in two synthetic datasets: D\_goggle and D\_DECAF. Both are then evaluated based on the same desiderata mentioned in the experiments of GOGGLE. To measure **Quality**, we calculate three metrics alpha-precision, beta-recall and authenticity and report the average to obtain a holistic score. Secondly we measure **detection** (generated samples should be indistinguishable from the real data). The AUROC score is calculated of three post-hoc classifiers. These classifiers are: a two layer-MLP, XGB classifiers and GMM classifiers. These models are trained to distinguish real samples from synthetic samples for both D\_DECAF compared to D\_test and the difference between D\_GOGGLE and D\_test and we average the AUROC score of the three classifiers. Thirdly, to evaluate **Utility** we train three downstream prediction models on D\_DECAF and D\_GOGGLE and evaluate them on D\_test and report the change in AUROC compared to training models on D\_train.

Samenvattend test ik het model onder de nieuwe condities en vergelijk het model met de analyse van het ene model toegepast op het andere model. Aanvullend test ik de aanpassing van het model door prior knowledge toe te voegen in de vorm van iets wat lijkt op fairness. Hiervoor is het belangrijk dat ik specifiek begrijp hoe het model werkt zodat ik precies in de code een aanpassing kan doen die iets simuleert dat met fairness te maken heeft. Als ik een dergelijke analyse heb dan kan ik hier iets over zeggen in de conclusie.

*Welke experimenten kan ik doen zodat de resultaten de conclusie ondersteunen?*

Ik doe met name twee experimenten: 1 zal bestaan uit het vergelijken van synthetische data die gegenereerd wordt door GOGGLE en die uit DECAF voor verschillende noties van fairness. Ik zal hierin de aanpak van GOGGLE gebruiken voor het testen van verschillende modellen, specifiek zal dit bestaan uit: rapporteren van de metrics uit GOGGLE voor data quality en toepassen van deze metrics op synthetische data die gegenereerd is door DECAF met bias die eruit verwijderd is. Dan zullen we de data uit DECAF ook evalueren op basis van detection. We zullen de output testen op 3 classifiers. Die erop getraind sullen worden om synthetische van echte data te onderscheiden. Ten derde zullen we utility rapporteren, dit is bedoeld om te kijken naar hoe nuttig synthetische data is voor het trainen van modellen en of het nut lager zal zijn voor DECAF. We zullen change of AUROC score rapporteren van 3 downstream classifiers op een predictive task.

Ten tweede is het interessant om de bestaande modellen (GOGGLE) uit te breiden met een notie van fairness die ingebouwd kan worden in dit model. Dit is een lastige maar uitdagende stap omdat ik hiervoor het model goed dien te leren kennen en dat ik aanpassingen moet doen die ik nog niet begrijp. **Dit zal een groot obstakel zijn en het is nog niet duidelijk hoeveel tijd mij dit zal kosten**.

Ten derde zou het nog interessant zijn om de resultaten te vergelijken met een dataset die erkende bias heeft en die te vergelijken voor beide modellen. Bijvoorbeeld eentje die gebasseerd op de conclusie uit GOGGLE (namelijk het werkt beter voor een grote of kleine dataset) zodat we ook nog vergelijkingsmateriaal hebben op basis van die meeteenheid.

*Welke theorie is hiervoor relevant?*

Inidergeval fairness in DECAF en werking van het model. Ook is GOGGLE relevant en een totale werking van het model. Met name dient erin te worden gegaan op hoe prior knowledge erin verwerkt wordt want hierin zal een cruciale aanpassing worden gedaan. Allereerst dient onderzocht te worden hoe en of dit kan. Verder zijn studies gedaan door DECAF en GOGGLE met name interessant omdat er later een vergelijkend onderzoek gedaan wordt. Sowieso dient alle theorie ook ingebed te worden in een verhaal dat iets meer lopend is maar het is een goede eerste stap om dit iniedergeval al te hebben staan. Dan is het heel makkelijk om er iets van een lopend verhaal aan vast te breien. Van GOGGLE is de evaluatie met name belangrijk om te vermelden.

*Welke eerdere werken zijn hiervoor relevant?*

Alle werken uit GOGGLE en DECAF, hou het simpel. Kopieer de papers. Voeg naar wens nog een aantal papers die die bijvoorbeeld recenter zijn uitgekomen.

*Hoe introduceer ik het onderwerp?*

Hier dient het onderwerp geïntroduceerd te worden. Een begin is er dat vgm leuk uitlegt waar het heen gaat. Echter dient er wel een concreter stuk te worden geschreven over de punten die dit onderzoek adresseert en hoe dat wordt gedaan. Welke dingen worden hier onderzocht? Waarom is het belangrijk? Hoe draagt het bij aan de wetenschappelijke literatuur rondom dit onderwerp? Mits er antwoord wordt gegeven op deze vragen is dit voldoende. Het begin kan een leuke anecdote bevatten.

Tot slot is het professioneel om een na-woord toe te voegen. Hierin wil ik Jaap Kamps met name bedanken, de auteurs van het paper, Andrew Yates, Ilya Markov en Etienne Posthummus. Ook wil ik Charlotte mijn vriendin bedanken voor het meedenken over het proces. En ik bedank mijn vrienden die mij hebben ondersteund in dit process en mij waardevolle input hebben gegeven.

**Concreet**

Nu is het belangrijk om concrete to-do’s op te stellen voor de stukken die ik hier publiceer.

* Schrijven voorlopige conclusie
  + Noem hierin met name hoe de resultaten uit de experimenten de conclusie ondersteunen. Dit is een verwachting zonder het doen van onderzoek.
  + Belicht specifiek hoe een bepaald resultaat uit een experiment (een aspect van) de conclusie ondersteund.
* Schrijven analyse en experimenten
  + Hoe worden de resultaten uit GOGGLE geanalyseerd. Benoem en belicht dit en beschrijf hoe dit in deze vernieuwde context toegepast zal worden
  + Benoem hoe het integreren van prior knowledge in deze context is onderzocht en hoe deze resultaten zijn geanalyseerd en beschrijf hoe dit met deze nieuwe prior knowledge op een soortgelijke manier gedaan zal worden.
  + (aanvullend) Beschrijf hoe we in zullen gaan op aanvullende data. En hoe de resultaten hiervan zullen geanalyseerd worden.
* Experimenten
  + Beschrijf hoe beide modellen gecombineerd zullen worden
  + Beschrijf hoe fairness geïntegreerd wordt in in GOGGLE. Maar neem aan dat dit onderbouwd is in de theorie (dit zal namelijk uitgelegd worden in de theorie)
  + Beschrijf welke experimenten er gedaan zullen worden op aanvullende biased datasets. Beschrijf hoe bepaalde sets biased zijn gemaakt (in DECAF) en hoe met deze (nieuwe) biased set er geëxperimenteerd zal worden.
* Theorie
  + Beschrijf met name hoe de regularization gedaan zal worden op de learned graph. Beschrijf hoe deze graph geleerd zal worden en hoe dit samenhangt met bestaande definities van fairness.
* Eerdere werken
  + Discussieer eerdere werken en werken gerelateerd met bestaande modellen.
* Introductie
  + Zorg voor een scherp verhaal dat leidt tot het onderzoek wat we hier doen
  + Benoem concreet welke punten onderzocht zullen worden.

**Tijdslijn**

* Ma 10: opzet experimenten en analyse en ablation
* Di 11: Herhalen experimenten, doen 1e experiment (debiasing census data)
* Woe 12: Herhalen experiment 2: hoe integreer je de methoden, uitwerken theorie voor integreren methoden
* Do 13: Experiment 3: census data (biased), runnen modellen en experimenten
* Vrij 14: schrijven van voorlopige conclusie
* Za 15: zeilen
* Zo 16: zeilen
* Ma 17: Analyse 1/3
* Di 18: Uitschrijven GOGGLE analyse
* Woe 19: Uitschrijven analyse prior knowledge, aanvullende data: analyse
* Do 20: Beschrijven experimenten
* Vrij 21: Related work
* Za 22: Introductie
* Zo 23: Doe aanvullende experimenten
* Ma 24: strak verhaal introductie
* Di 25: strak verhaal related work
* Woe 26: strak verhaal Theorie
* Do 27: strak verhaal theorie: integreren prior knowledge biases in model
* Vrij 28: Strak verhaal analyse
* Za 29: strak verhaal conclusie, nawoord
* Zo 30: puntjes op de i – layout: DRAFT finished.